國立政治大學應用數學系
碩士學位論文

即時雙線服務系統之等候模型
Modeling on a Real-time Two-tier Service System

碩士班學生：黃賴均　撰
指導教授：陸　行　博士
中 華 民 國　108　年　10　月

DOI:10.6814/NCCU201901241

致謝

本論文得以順利完成，首先應感謝陸行老師的悉心教導與督促，以及各位同學、學長姊弟妹與朋友們的幫助。此外感謝在論文口試時，陳政輝教授以及洪一薰教授的指導並提供寶貴的意見，使本論文更加完善，也使不才在研究領域上獲益良多，在此獻上最誠摯的感謝。最後，也感謝父母的支持，給予我無後顧之憂的學習環境，謹將此論文與所有支持我、幫助我的人分享。

DOI:10.6814/NCCU201901241

中文摘要

等待時間一直是服務品質的重要指標，例如減少在醫療保健，公共服務和各種重點服務 (VIP) 系統的等待時間。本論文考慮由兩個不同的服務站組成的雙線服務系統，包含一個免費服務站，和一個付費服務站，每個服務站都有隊列和服務提供者，據此建立數學等候模型。兩個服務站提供相同的服務內容。假設其中付費服務站的隊列具有長度限制，該服務站為了減少客戶等待時間維持服務質量而採取溢價服務。溢價服務意指系統通過收取額外費用提供另一服務選擇的機制。

由於有一些客戶會根據自己的時間價值做出決策，我們在這種雙線服務系統中研究隊列長度信息對顧客行為的影響，我們發現向客戶提供即時隊列長度信息可以顯著地減少總等待成本。此外，從最小化所有客戶的總等待成本和最大化付費服務提供者的利潤的角度，我們利用數學模型分析提供即時隊列長度信息與否之影響。

在論文中，我們展示此模型能夠反映減輕客戶等待之負擔的信息效應，同時也揭示價格策略和服務保障對雙線服務系統服務指標的影響。

關鍵詞：即時信息，雙線服務系統，類生死過程，矩陣幾何解法

DOI:10.6814/NCCU201901241

Abstract

Waiting time has been an unavoidable concern for service such as healthcare, public provision and VIP systems of various services. We address this issue for considering a two-tier service system which is composed of two different service stations: a gratis station and a toll station. Each service station is set up by a queue and a service provider. The service providers of service stations provide the same service. In the thesis, we study a queueing model that one of the service stations charges a premium in order to guarantee a maximum expected waiting time and the queue of this service station has a length limit.

We study the effects of the queue length information on the performance of such a two-tier service system with customers who make decisions based on their own time value. We show that offering the real-time queue length information to customers can effectively enhance the performances of both services in the system.

Furthermore, for both with and without real-time queue length information scenarios, we analyze the problem from two perspectives. There are the perspectives of minimizing the expected social waiting cost for customers and maximizing the expected profit for the manager. We show that this model can obviously reflect the information effects of alleviating the burden of waiting for customers, and it also reveals the impact of service guarantee and price discrimination on the performance of the two-tier service system.

Keywords: Real-time information, Two-tier service system, QBD process, Matrix geometric method

DOI:10.6814/NCCU201901241

Contents

致謝 i

中文摘要 ii

Abstract iii

Contents iv

List of Tables vii

List of Figures viii

1 Introduction 1

1.1 Research Background 1

1.2 Literature Review on Modeling 3

1.3 The Objective of This Study 6

2 A Two-tier Service System 8

2.1 Definitions and Assumptions 8

2.2 A No Real-time Information Scenario 9

2.3 A Real-Time Information Scenario 15

3 An Optimization Model 24

3.1 The Perspective of the Society 24

3.1.1 The Perspective of The Society in No Real-time Information Scenario . 24

3.1.2 The Perspective of the Society in Real-time Information Scenario . . . 27

3.2 The Perspective of the Manager 29

DOI:10.6814/NCCU201901241

3.2.1 The Perspective of the Manager in No Real-time Information Scenario . 29
3.2.2 The Perspective of the Manager in Real-time Information Scenario . . 31

4 Numerical Examples and Discussion **33**
4.1 Parameters 33
4.2 Real-time and No Real-time Information Scenario 34
4.3 The Perspective of the Society and the Manager 38

5 Conclusion **42**

Bibliography **44**

Appendix A The Proofs and Background Informations **49**
A.1 Matrix Geometric Method 49
A.2 Algorithm for Computing the Rate Matrix 50
A.3 Newton's Method 51
A.4 The Distribution Function of the Waiting Time 51
A.5 The Stability Condition for Real-time information 52
A.6 The Convex of the Expected Social Waiting Cost for no Real-time 53
A.7 The Distributions 53
A.7.1 The Uniform Distribution 54
A.7.2 The Exponential Distribution 55
A.7.3 The Pareto Distribution 55

Appendix B MATLAB Codes **57**
B.1 Program for Distributions of Customers' Time Values 57
B.1.1 Parameters of Distributions 57
B.1.2 The Cumulative Distribution Function of Θ 58
B.1.3 The Function of Expected Value of $f(\theta)$ 59
B.1.4 Inverse Function of the Distribution Cumulative Function of Θ 60
B.2 Taylor Series of Exponential Function 61
B.2.1 Taylor Expansion of Exponential Function 61
B.2.2 Error of Taylor Expansion of Exponential Function 61
B.3 Program for a No Real-time Information Scenario 61

DOI:10.6814/NCCU201901241

B.3.1 The Main Program 62
B.3.2 The Planned Arrival Rate 63
B.3.3 The Effective Arrival Rate 64
B.3.4 The Expected Waiting Time 64
B.3.5 The Balanced Function of θ 65
B.3.6 The Search Algorithm of Computing $\overline{\theta}$ 65
B.3.7 The Complementary Cumulative Distribution Function of the Waiting Time 66
B.3.8 The Expected Social Waiting Cost 67
B.3.9 The Expected Profit of Manager 68
B.3.10 The Constraints of Optimization 68
B.4 Program for a Real-time Information Scenario 69
B.4.1 The Main Program 70
B.4.2 The Planned Arrival Rate 71
B.4.3 The Transfer Matrix 72
B.4.4 The Eigenvalue of K-Matrix 72
B.4.5 The Stationary Probability 73
B.4.6 The Expected Queue length 75
B.4.7 The Expected Social Waiting Cost 76
B.4.8 The Expected Profit of Manager 77
B.4.9 The Complementary Cumulative Distribution Function of The Waiting Time 78
B.4.10 The Constraints of Optimization 80

DOI:10.6814/NCCU201901241

List of Tables

4.1 Optimal solutions minimizing of the expected social waiting cost with $\Lambda = 1, K = 10, c = 0.5$. 40
4.2 Optimal solutions maximizing of the expected profit of the manager with $\Lambda = 1, K = 10, c = 0.5$. 41

DOI:10.6814/NCCU201901241

List of Figures

2.1 A two-tier service system 8

4.1 A two-tier service system with $p = 10$, $\Theta \sim U(0, 20)$ 34
4.2 A two-tier service system with $p = 10$, $\Theta \sim Exp(1/10)$ 35
4.3 A two-tier service system with $p = 10$, $\Theta \sim Pa(1, 10/9)$ 35
4.4 A two-tier service system with $K = 10$, $\Theta \sim U(0, 20)$ 36
4.5 A two-tier service system with $K = 10$, $\Theta \sim Exp(1/10)$ 36
4.6 A two-tier service system with $K = 10$, $\Theta \sim Pa(1, 10/9)$ 36
4.7 A comparison of $E[L_1]$ and $E[L_2]$ with K between with and without information. 37
4.8 A comparison of $E[L_1]$ and $E[L_2]$ with p between with and without information. 38

A.1 A distribution of customers' time value 54

DOI:10.6814/NCCU201901241

Chapter 1

Introduction

In this thesis, we consider a system with two service stations for customers to choose. One station offers a gratis service and the other offers a toll service but with shorter waiting time. Customers are heterogeneous in view of their individual time value. Customers select the station based on either the expected waiting time or the real-time queue length information of two stations. We develop two stochastic models for these two information scenarios, and give an optimized price of the toll station for the perspective of the expected social waiting cost of customers as well as the expected profit of the manager.

We study the effects of the queue length information on the performance of the two-tier service system where customers make decisions based on their own time value. These results not only benefit the practitioners to manage the actual service system that fits the model, but also facilitate a computational model used by researchers with queueing theories to simplify the analysis.

1.1 Research Background

A two-tier healthcare system usually consists of a public station and a private station. The former is supplied by the government and the latter by private investors. This system is used by many countries, for example, in Singapore, France and U.K. etc. In a two-tier healthcare system, the price for providing the service in the public station often is set low by the government. This is to provide services to everyone, but usually leads to long queues. Thus, when customers with higher opportunity costs use the public service, he/she may has a high

DOI:10.6814/NCCU201901241

waiting cost. Johar and Savage [20] provide a comparison of waiting times between Australian public and private hospitals and show that the patients in public hospitals typically wait longer than in private hospitals. This is where private services can take place, although it charges an additionally higher price. If the waiting cost saved, compared with public services exceed the price, customers with higher waiting costs are suitable to use private services.

Waiting costs have been one of the important issues for services. This is reflected in the design of two-tier healthcare in various countries [27]. No matter how rich the country is, each health care system has limited resources. In another aspect, demand is increasing due to an aging population and rising levels of medical services.

A healthcare service system should find a balance between the patient's waiting cost and the cost of service capacity. Besides direct expenditure and lost productivity, the patient's waiting cost also includes the cost of increased service demand due to service delays. For healthcare services, customers or patients differ in terms of socioeconomic status, and even in terms of their medical conditions. The former usually means the difference in the opportunity costs of their waiting time, and the latter probably causes the patient to be in danger when the treatment is delayed. It causes that some customers need shorter waiting time than the other because of their ability or willingness to wait for the service. Given this situation, it is important to be able to lower the patient's waiting cost without increasing service capacity.

Service schemes generated by price discrimination is a common way to solve this problem. This strategy effectively allocates resources between heterogeneous and selfish customers. For increasing the efficiency of the healthcare service system, the strategy minimizing waiting costs can be motivated by exploiting the heterogeneity of the customers or patients [12]. In this case, distinguishing customers by price and leadtime can increase systemefficiency or profit [31].

Many of the literature in the past often assumes that customers are provided with long-term statistical information rather than the real-time queue information. Of course, there are some practical systems that support the assumption the customers make their decisions according to the long-term statistics or unobservable queues in the two-tier system. For example, the customer will decide to join a public or private health insurance plan based on long-term statistical information. At another example, the customer decides to go to a restaurant based on past experience, where experience represents a kind of long-term statistical information. It is reasonable to assume that customers use long-term statistics such as expected waits and service

DOI:10.6814/NCCU201901241

time, rather than real-time queue length information. In addition, by decomposing the two-tier system into two subsystems, the analysis of system performance can be simplified. This is another important reason for making this assumption in previous research.

There are some realities where customers choose services based on the real-time queue lengths of the two service providers. Real-time queue length information will make the arrival process depending on the system state which complicates the analysis, and some analysis methods are mathematically difficult to handle. Because it is not easy to compute the stationary probability at the model with real-time information, we used the matrix geometric method to analyze the system performance of customers with real-time queue information. The matrix geometric method was developed around 1975 [3]. It is a method for the analysis of quasi-birth–death processes (QBD, see Appendix A.1). Computationally, this type of problem can be modeled as a QBD process that the levels are countably infinite. It can be solved by the matrix geometric method, but the calculation of the matrices with large size makes the solution complex. Most importantly, we consider the optimization of the perspectives of the society and the manager. From a social perspective view, the government may want to differentiate customers by price differences, so that customers with urgency or high time costs can get services as soon as possible. Therefore, tolls are only a means of separating customers with the objective of minimizing the expected social waiting cost of the customers. From the perspective of the manager, the toll price is adopted for profit. Their objective is naturally to maximize the profits. However, very high price may make customers reluctant to enter and lose more than the manager's gain. Therefore, managers need to strike a balance between the arrival rate and the toll price.

1.2 Literature Review on Modeling

This paper is related to two streams of literature. First, it is related to the information effects on performance of queueing systems. There are some literatures on the information effects on performance of queueing systems.

Naor [28] study that strategic customers observe either real-time queue length information or the long-term statistics to make the decision whether to join or balk on an M/M/1 queue. He showed that the self-interest customer's choice does not maximize social welfare, but the

DOI:10.6814/NCCU201901241

socially optimization can be attained by using a toll price. However, his study is based on first come first served (FCFS) queueing rules and homogeneous delay costs. Hassin and Haviv [17] summarize more research findings in this area. See [17] and literature therein.

Armony and Maglaras [2] [1] analyze a system that offers two modes of service, in which there are real-time service and postponed services with a delayed guarantee. In Armony and Maglaras [2], customer's decisions are based on the real-time delay, or long-term statistics. In Armony and Maglaras [1], each arriving customer is given an estimation of delay based on the system occupancy. Comparing the results with the other systems, they show that more information improves performance on several aspects.

Guo and Zipkin [16] study an M/M/1 service queueing system. Based on the three kind of information: no information, partial information (the system occupancy), and full information (the exact waiting time), and the expected waiting cost, they show that the customer decides whether to join or reject with more accurate information can improve performance. However, in other cases, the information may actually harm the service provider or customers. Guo and Zhang [15] investigate two-tier service systems motivated by healthcare and border-crossing systems where customers can choose to join a free system or a toll system. Both partial real-time information and no real-time information cases are considered. They show that the system performance is more robust by setting a relatively high than a relatively low price.

Chen et al. [5] conduct an empirical analysis on the two-tier system based on the real dataset verified by a two-queue model. Hua et al. [18] assume that the customers make their decisions according to long-term statistics and study the competition and coordination in a two-tier service system where the free system strives for maximizing its expected total customer utility with limited capacity, while the toll system is aimed at maximizing its profit. Qian et al. [32] analyze subsidy schemes for reducing waiting times under the assumption that public healthcare service has no user fee but an observable delay and private healthcare service has a fee but no delay.

Zhang and Luh [37] consider a two-tier service system including one free and one toll stations. The free station provides free service and the toll station provides paid service but a guaranteed maximum expected waiting time. The waiting cost rate of strategic heterogeneous customers follows a uniform distribution. They investigate how the toll price and service guarantee affect the performance of the two-tier service system and indicate the inconsistency between the profit of the toll station. On this basis, we study more general distributions of

DOI:10.6814/NCCU201901241

the waiting cost rate of strategic heterogeneous customers in this thesis. In addition, we find optimized prices for both the social and managerial perspectives.

The second stream of research is pricing and priority decisions in queueing systems. It focuses on reducing delayed cost. To address the self-optimization behavior of customers, Kleinrock [22] considers a problem that a relative position in queue is determined according to the size of a customers' bribe while weighing the delay cost and bribery under the general queueing system.

Mendelson and Whang [26] characterize user classes by its expected service time and waiting cost rate and consider an M/M/1 queueing system with these classes. They raised a priority pricing mechanism in which each user decides whether to enter the system with certain priority level or not. These decisions make the objective function value of the entire system maximized.

Edelson and Hildebrand [10] study when customers have the homogeneous or heterogeneous value for waiting time, expanding the number of servers and charging premium are methods for segmenting the market to implicate the welfare. Schroeter [33] considers heterogeneous customers with uniformly distributed unit-time waiting costs. Stidham [34] discusses more comprehensively the price in priority queues with one or more classes of customers.

Nazerzadeh and Randhawa [29] show that for M/M/1 systems with heterogeneous waiting costs of customers, it is asymptotically optimal to provide two service levels with appropriate prices. Gavirneni and Kulkarni [12] study a pay-for-priority mechanism named concierge option with heterogeneous waiting cost rates. They show that the concierge option is a valid method to be beneficial to each customer, and these benefits are greater when the system utilization is high and the variance of the customers' waiting costs is large.

In particular, we have noticed the study of Wan and Wang [36]. They consider a setting of two service providers, and the objective of minimizing the expected waiting cost per patent and maximizing the social gain. The arrival rates of customers follows a general probability distribution. Customers with the waiting cost rate choose a service provider according to their own preferences. To compare with their study, we consider the more general probability distribution of customers' waiting cost rates and considers the impact of with and without real-time information.

DOI:10.6814/NCCU201901241

1.3 The Objective of This Study

Our work involves three blocks. We built the model based on two-tier service system. In this model, one of the service stations charge each customer a fee for the service. First, we explore how the real-time queue length information affects the system performance. In fact, there are some systems where customers make decisions based on long-term statistics when the two-tier system has unobservable queues. It is reasonable to assume that customers use long-term statistics rather than real-time queue length information. In addition, another important reason for making this assumption in previous studies is that the analysis can be simplified by decomposing the two-tier system into two subsystems. Because of the dependence of real-time queue length information on the system state, the analysis will be complicated. Nevertheless, there are circumstances for customers to make selections based on the real-time queue lengths of the two service providers.

We show that the real-time queue information makes the queue length always shorter when the two service rates are not different. We compare the impact on queue lengths with and without real-time queue length information when the service rates, the toll price, and the buffer size change, respectively.

Second, we consider the toll pricing from two different perspectives in the two-tier system. From the perspective of the society, the toll price is just a transfer payment. Thus, to make the expected social waiting cost for the customers minimized is its objective. By contrast, from the perspective of the manager, the toll price is the source of income. Thus, the goal is to maximize the manager's profit. We show that real-time information can reduce the expected social waiting cost and enhance the social welfare. Moreover, we indicate that the two-tier system with the price discrimination strategy has a lower expected social waiting cost when the customer's time value is more diverse. Additionally, in the perspective of the manager, we show that the optimal price chosen by the manager of the toll station is affected by the service rate of the gratis station, and this decision will make the expected social waiting cost relatively high. Therefore, in this case, the government's control for the toll price is usually necessary.

Finally, by providing real-time information in a two tier service system, the expected social waiting cost per customer is significantly reduced. This does not mean that the queue lengths will always be shorter in real-time information scenarios. By adjusting the service rate ratios of the two stations μ_1/μ_2, we find that when the gratis station has a higher service rate, the queue

DOI:10.6814/NCCU201901241

length of the gratis station in real-time information scenario is shorter than that in no real-time information scenario. This is because customers will overestimate the service efficiency of the gratis station without real-time information. When the toll station has a higher service rate, the similar situation that customers overestimate the service efficiency of the toll station will also occur. Therefore, we find an interval of service rate ratios where the queue lengths of the gratis and toll station are improved. Furthermore, in this interval, we speculate that the expected social waiting cost in the real-time information seems to be a convex upward function of μ_1/μ_2 where there is a global minimum by given μ_1/μ_2. That is to say, when the government designs their two-tier healthcare system, if μ_1/μ_2 is a decision variable, there is a best division so that the expected social waiting cost is minimized.

The thesis is organized as follows. In Chapter 2, we formulate a basic model for a two-tier service system with and without the real-time information. We develop a search algorithm to compute the price under no real-time information scenario. Moreover, we formulate a computational quasi-birth-death (QBD) process model under the real-time information scenario. In Chapter 3, we analyze the optimization model that minimizes the expected social waiting cost for customers or maximizes the profit for managers with or without real-time information. In Chapter 4, we compare the performances of two perspective optimization in these two scenarios. We discuss the management implications of the results. Finally, we conclude our study in Chapter 5.

DOI:10.6814/NCCU201901241

Chapter 2

A Two-tier Service System

We consider a service system consisting of two service stations. Each service station is composed of a queue and a service provider. These two service providers offer the same service, but one of service providers charges their customers while the queue has a finite buffer. The time value of customers is subject to a distribution. Due to his/her own time value, the customer selects a service station to minimize his/her expected waiting cost. Our intention in this chapter is to characterize the system behavior under the optimal choice of customers.

2.1 Definitions and Assumptions

We build a queueing model that is composed of a gratis service station and a toll service station which is shown in Figure 2.1 for the two-tier service system. The service times for

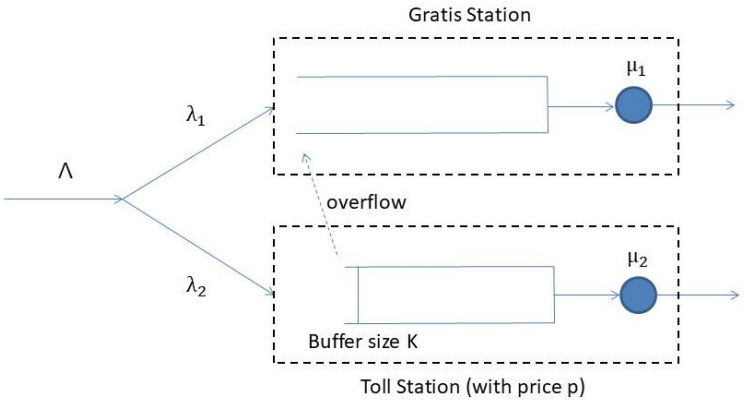


Figure 2.1: A two-tier service system

DOI:10.6814/NCCU201901241

the gratis station and the toll station are independent and identically distributed following exponential probability distributions with the service rates μ_1 and μ_2. In general, they may represent the service capacity of the service providers. The gratis service station denoted by station 1 offers free service with mean service time $1/\mu_1$, and the toll service station with a finite buffer denoted by station 2 offers the service with mean service time $1/\mu_2$ and with presumably shorter waiting time but for a toll fee of p.

For example, at a two-tier healthcare system, the toll station represents a private hospital which usually emphasizes the service efficiency. In order to ensure a shorter service time for the patient who pays, the waiting time of the toll service needs an upper bound that the patient may bear. We assume that the queue of the toll station has a finite buffer of size K, and as long as the buffer is full, the customer must join the queue of the gratis station. In addition, a finite buffer is able to make the service efficient as it can avoid excessive waiting time that usually leads to the high "no-show" rate or low service efficiency [13] [19].

Customers arrive according to a Poisson process with rate Λ and should be served on first-come-first-served (FCFS) basis. Customers may be considered differently in their time values. Denote time value by a random variable $\Theta \geq 0$ in terms of the waiting cost per unit time for a customer. $F(\theta)$ and $f(\theta)$ denote the cumulative distribution function and the probability density function of θ, respectively. Let the function of expected value of $f(\theta)$ be $G(x) = \int_0^x \theta f(\theta) d\theta$ and $\xi = \int_0^\infty \theta f(\theta) d\theta$. The existence of ξ depends on the condition of $f(\theta)$. See Appendix A.7 for more details of selected probability distributions in this thesis.

Based on the difference of the expected waiting cost of two queues, each customer wisely select a service station. Assume customers are split into two flows to enter the gratis or the toll station with arrival rates λ_1 or λ_2, where $\lambda_1 + \lambda_2 = \Lambda$. Further assume that λ_1 and λ_2 are functions of Θ. In view to whether the customer knows the queue lengths of the two stations at their arrival instants, there are different scenarios about choosing the gratis station or the toll station, which we discuss in the following sections.

2.2 A No Real-time Information Scenario

In this section, we consider the case where the customer does not know the actual queue length of the two stations when they arrive. Customers will choose a station to enter based on

DOI:10.6814/NCCU201901241

the expected queue lengths of the gratis and toll stations. Intuitively, because of the positive toll price, every customer would want to enter the gratis station at the beginning. But when more and more customers join the gratis station, the waiting time of the gratis station will increase, and the waiting time of the toll station is relatively short. This will force customers with high time value to think to join the toll station for a shorter expected waiting time.

Given a distribution of the time value of customers, suppose there exists the balanced waiting time cost rate $\overline{\theta}$ satisfies the following equation:

$$\overline{\theta}E[W_1] = p + \overline{\theta}E[W_2], \tag{2.2.1}$$

where W_1 and W_2 is the waiting time of the gratis and toll stations. In this equation, because of the positive p, it implies the expected waiting time at the gratis station is no less than that at the toll station, namely, $E[W_1] \geq E[W_2]$.

Base on their time value θ and the expected waiting time of both stations, each customer will compare the cost when entering the gratis station $\theta E[W_1]$ and the cost of entering the toll station $p + \theta E[W_2]$, and choose the lower one to enter.

Ideally, customers with time value lower than $\overline{\theta}$ consider the left hand side of the equation and they will choose the gratis service station because of free of entrance fee. On the other hand, those who bear high time value will choose the toll service station for the shorter expected waiting time. Then, $\overline{\theta}$ is the watershed between this two groups of customers, such that we may compute the planned arrival rates by $\lambda_1(\overline{\theta}) = \Lambda F(\overline{\theta})$ and $\lambda_2(\overline{\theta}) = \Lambda(1 - F(\overline{\theta}))$. In fact, the expected waiting times of the gratis and toll stations depend on the arrival rates which depends on $\overline{\theta}$. Therefore the expected waiting time is affected by $\overline{\theta}$, namely denoted as $E[W_1(\overline{\theta})]$ and $E[W_2(\overline{\theta})]$. In this way, as long as p is given from the equation (2.2.1), it will find the watershed $\overline{\theta}$. However, the calculation of $E[W_1]$ and $E[W_2]$ is quite complicated, we cannot directly solve $\overline{\theta}$. Therefore, we need to develop an algorithm to compute $\overline{\theta}$.

Since we have assumed that when the buffer of the toll station is full, the customers are forced to join the gratis station, it is necessary to distinguish the planned arrival rate and the effective arrival rate in the system. As explained in the previous section, λ_1 and λ_2 represent the planned arrival rates for the gratis and toll stations, the effective arrival rates for the gratis and toll stations are $\lambda_1^{\text{eff}}(\theta) = \lambda_1(\theta) + \pi_{\bullet K}\lambda_2(\theta)$ and $\lambda_2^{\text{eff}}(\theta) = (1 - \pi_{\bullet K})\lambda_2(\theta)$, respectively, where $\pi_{\bullet K}$ is the probability that there are exactly K customers in the toll station.

DOI:10.6814/NCCU201901241

The toll station clearly is an M/M/1/K queue, since its arrival rate for the non-full buffer states is λ_2 subject to the Poisson arrival process and 0 as the buffer is full. As a result of λ_1^{eff} superimposing the overflow from the toll station and the Poisson arrival process with rate λ_1, the aggregate arrival process for the gratis station can be modeled as an M/M/1 queuing system.

We can easily obtain the stability condition for the two-tier service system under no real-time queue length information.

Proposition 2.2.1. With no real-time queue length information, the two-tier service system reaches the steady state if $\mu_1 > \lambda_1^{\text{eff}} = \lambda_1 + \pi_{\bullet K}\lambda_2$.

Even though the stability condition is easy to understand, we are still unable to estimate customer's behavior in terms of θ from λ_1 and λ_2, satisfying Proposition (2.2.1).

For the toll station with a given $\overline{\theta}$, then there are $\lambda_2 = \Lambda(1 - F(\overline{\theta}))$ and $\rho_2 = \lambda_2/\mu_2$, we have the expected waiting time and the stationary probability [14]

$$E[W_2] = \begin{cases} \frac{1}{\lambda_2^{\text{eff}}}\left\{\frac{1}{1-\rho_2} - \frac{1+K\rho_2^{K+1}}{1-\rho_2^{K+1}}\right\} & \text{if } \rho_2 \neq 1, \\ \frac{K}{2} & \text{if } \rho_2 = 1, \end{cases} \tag{2.2.2}$$

$$\pi_{\bullet m} = \begin{cases} \frac{(1-\rho_2)\rho_2^m}{1-\rho_2^{K+1}} & \text{if } \rho_2 \neq 1, \\ \frac{1}{K+1} & \text{if } \rho_2 = 1, \end{cases} \quad m = 0, 1, ..., K. \tag{2.2.3}$$

Since the number of customers arriving is subject to the Poisson distribution, λ_1 and λ_2 are still parameters of the Poisson distribution. Furthermore, $\pi_{\bullet K}\lambda_2$ and $(1-\pi_{\bullet K})\lambda_2$ are parameters of the Poisson distribution, which means λ_1^{eff} is also a parameter of the Poisson distributions. Then the queuing system of the gratis station can be regarded as the M/M/1 queuing system with an arrival rate of λ_1^{eff}. So we can get the expected waiting time and the stationary probability for the gratis station as follows:

$$E[W_1] = \frac{1}{\mu_1 - \lambda_1^{\text{eff}}} \tag{2.2.4}$$

$$\pi_{n\bullet} = (1-\rho_1)\rho_1^i, n = 0, 1, ... \tag{2.2.5}$$

where $\rho_1 = \lambda_1^{\text{eff}}/\mu_1$.

Proposition 2.2.2. For a stable two-tier service system without real-time queue length

DOI:10.6814/NCCU201901241

information, and a fixed p there exists a unique $\overline{\theta}$ satisfying the equation (2.2.1).

Proof. For the balanced waiting time cost rate $\overline{\theta}$ satisfies (2.2.1) which can be re-written as

$$\overline{\theta} = \frac{p}{E[W_1(\overline{\theta})] - E[W_2(\overline{\theta})]}. \tag{2.2.6}$$

For any θ and a distribution of θ, $F(\theta)$, satisfying (2.2.6), we have $\lambda_1 = \Lambda F(\theta)$ and $\lambda_2 = \Lambda(1 - F(\theta))$. Thus as θ increases, λ_1 increases and λ_2 decreases. For the toll station with a given λ_2, we have the stationary probability when the queue length of the toll station is K by (2.2.3). Thus, we get

$$\lambda_2^{\text{eff}} = (1 - \pi_{\bullet K})\lambda_2 = \begin{cases} (1 - \frac{(1-\rho_2)\rho_2^K}{1-\rho_2^{K+1}})\lambda_2 & \text{if } \rho_2 \neq 1, \\ (1 - \frac{1}{K+1})\lambda_2 & \text{if } \rho_2 = 1. \end{cases} \tag{2.2.7}$$

It is easy to see that λ_2^{eff} increases with λ_2 for the case of $\rho_2 = 1$. Then, we consider the case of $\rho \neq 1$. Define $g(x) = \sum_{i=1}^{K} x^{-(i+1)}\mu_2^i$, we rewrite (2.2.7) as

$$\lambda_2^{\text{eff}}(\lambda_2) = (1 - \frac{(1-\rho_2)\rho_2^K}{1-\rho_2^{K+1}})\lambda_2 = \frac{\lambda_2 g(\lambda_2)}{g(\lambda_2) + \lambda_2^{-1}} \text{ for } \rho_2 \neq 1$$

Given $\delta \in \mathbb{R}^+$, consider the ratio of λ_2^{eff} between λ_2 and $\lambda_2 + \delta$,

$$\begin{aligned} ratio\_\lambda_2^{\text{eff}}(\lambda_2, \lambda_2 + \delta) &= \frac{\lambda_2^{\text{eff}}(\lambda_2)}{\lambda_2^{\text{eff}}(\lambda_2 + \delta)} \\ &= \frac{\lambda_2 g(\lambda_2)}{g(\lambda_2) + \lambda_2^{-1}} (\frac{(\lambda_2 + \delta)g(\lambda_2 + \delta)}{g(\lambda_2 + \delta) + (\lambda_2 + \delta)^{-1}})^{-1} \\ &= \frac{g(\lambda_2)(\lambda_2 g(\lambda_2 + \delta) + \lambda_2(\lambda_2 + \delta)^{-1})}{g(\lambda_2)(\lambda_2 g(\lambda_2 + \delta) + 1) + \delta g(\lambda_2 + \delta)(g(\lambda_2) + \lambda_2^{-1})} \end{aligned} \tag{2.2.8}$$

We know that the function value of (2.2.8) is less than 1, since $\lambda_2(\lambda_2 + \delta)^{-1} < 1$ and $\delta g(\lambda_2 + \delta)(g(\lambda_2) + \lambda_2^{-1}) > 0$. It implies λ_2^{eff} is increasing with λ_2 when $\rho_2 \neq 1$. Furthermore, λ_2^{eff} is decreasing with θ. On the other hand, because of $\lambda_2^{\text{eff}} + \lambda_1^{\text{eff}} = \Lambda$, we know λ_1^{eff} increases with θ.

It is intuitive to see that $E[W_1(\theta)]$ increases with λ_1^{eff} and $E[W_2(\theta)]$ increases with λ_2^{eff}. Therefore, $E[W_1(\theta)] - E[W_2(\theta)]$ is positive and increasing with θ since p in (2.2.1) is positive such that $E[W_1(\theta)] > E[W_2(\theta)]$. Then, when the left hand side of (2.2.6) θ increases, $p/\ (E[W_1(\theta)] - E[W_2(\theta)])$, the right hand side of (2.2.6) decreases. Note that

DOI:10.6814/NCCU201901241

$p/\ (E[W_1(\theta)] - E[W_2(\theta)])$ is positive and as θ increases, whether the denominator tends to infinity or not, causing it to converge to a finite number.

This implies that there exists a unique solution to (2.2.6) which in turn determines the arrival rates of $\lambda_1 = \Lambda F(\overline{\theta})$ and $\lambda_2 = \Lambda(1 - F(\overline{\theta}))$. □

Although we have an explicit expression for $E[W_1]$ and $E[W_2]$, they are quite complex. This makes it difficult to obtain the balanced waiting time cost rate directly by equations. Define $Eq(\theta) := p/([E[W_1(\theta)] - E[W_2(\theta)]]) - \theta$, we compute the balanced performance measures by the following search algorithm.

DOI:10.6814/NCCU201901241

A Search Algorithm of Computing $\overline{\theta}$:

- Step 1: Initialization - Given positive ϵ, δ and let $n = 0$. Select an initial $\theta_0 = \frac{F^{-1}(\mu_1/\Lambda)+F^{-1}(1-\mu_2/\Lambda)}{2}$.
- Step 2: Compute $\lambda_1 = \Lambda F(\theta_0)$, $\lambda_2 = \Lambda(1 - F(\theta_0))$, $E[W_1(\theta_0)]$ and $E[W_2(\theta_0)]$ from (2.2.4) and (2.2.2).
- Step 3: Let $\theta_{n+1} = \theta_n - \frac{\delta\epsilon Eq(\theta_n)}{Eq(\theta_n+\epsilon)-Eq(\theta_n)}$.
- Step 4: Compute $\lambda_1 = \Lambda F(\theta_{n+1})$, $\lambda_2 = \Lambda(1 - F(\theta_{n+1}))$, $E[W_1(\theta_{n+1})]$ and $E[W_2(\theta_{n+1})]$ from (2.2.4) and (2.2.2).
- Step 5: Check the condition $E[W_1(\theta_{n+1})] > E[W_2(\theta_{n+1})]$ (*). If it is false, then keep reducing δ and compute (*) until it is satisfied.
- Step 6: If $|Eq(\theta_{n+1})| > \epsilon$, replace θ_n by θ_{n+1} for $n \leftarrow n+1$, and go to Step 3. Otherwise, go to step 7.
- Step 7: Check the stability condition $\lambda_1 + \lambda_2\pi_{\bullet K} < \mu_1$. If it is false, then this system is not stable. Otherwise, the algorithm will reach the balanced waiting time cost rate $\overline{\theta}$.

In fact, this is an algorithm based on Newton's method (see Appendix A.3). The choice of θ_0 in accordance with $E[W_1] > E[W_2]$ is crucial. Because even if we are sure that $Eq(\theta)$ has a unique solution and it is strictly decreasing in the area of $E[W_1] \geq E[W_2]$, the $Eq(\theta)$ is singular at $E[W_1] = E[W_2]$. In this case, we need to check $E[W_1] > E[W_2]$ before the arrival rates are computed. We calculate the θ_0 as an initial solution by composing a half of $F^{-1}(\mu_1/\Lambda)$ and a half of $F^{-1}(1 - \mu_2/\Lambda)$ which are gained from $\mu_1 = \Lambda(F(\theta_0)$ and $\mu_2 = \Lambda(1 - F(\theta_0))$, respectively. This is because if choosing $\mu_1 = \lambda_1$ and $\mu_2 = \lambda_2$, it will critically make the gratis and toll stations unstable. Once $\overline{\theta}$ is calculated, the price can be determined by (2.2.6), say $\overline{p}$.

On the other hand, the stability condition for the gratis station, $\lambda_1 + \lambda_2\pi_{\bullet K} < \mu_1$, avoids large θ leading to a oversized arrival rate of the gratis station. Moreover, it actually provides a right boundary of θ. However, no matter how large θ is, $Eq(\theta)$ is always strictly decremented when $E[W_1] > E[W_2]$, the stability condition does not hinder convergence from the proceeding of the algorithm. Despite this, it is still necessary to note that the final $\overline{\theta}$ must match the stability condition.

DOI:10.6814/NCCU201901241

After we obtain the balanced $\overline{\theta}$, we can calculate the stationary probability and the average waiting time of both stations. We give the detailed expressions of the distribution function of the waiting time (see Appendix A.4). Finally, we get the distribution of the waiting time of the gratis and toll stations, respectively.

$$\begin{aligned} P(W_1 > t) &= e^{(\lambda_1^{\text{eff}} - \mu_1)t} \\ P(W_2 > t) &= \sum_{m=0}^{K} \pi_{\bullet m} T_m(\mu_2 t) e^{-\mu_2 t} \end{aligned} \tag{2.2.9}$$

where $T_i(x) = \sum_{j=0}^{i} x^j / j!$.

Although we know the expected waiting time, we may actually wait longer than expected. According to (2.2.9), we can answer some of these questions regarding risk if p is given. We will give a pricing strategy at next chapter, and using (2.2.9) as a constraint such that the risk is tolerable.

2.3 A Real-Time Information Scenario

Under the same assumptions of the previous section, except every customer will be informed the queue length information for both queues when he/she arrives, we consider real-time information scenarios in this section. Unlike the previous section, customers make decisions based on information without relying on long-term statistics. In this section, similar to (2.2.1), there is an immediate judgement at the arrival instant t, the arrival customer will compare the price p with his/her acceptable price $\overline{p}$ in the following equation.

$$\overline{p} = \theta \frac{L_1(t)}{\mu_1} - \theta \frac{L_2(t)}{\mu_2}, \tag{2.3.1}$$

where $L_1(t)$ and $L_2(t)$ are the queue lengths of the gratis and toll stations, respectively. Naturally, we have $L_1(t) \in \mathbb{N}$ and $L_2(t) \in \{0, 1, 2, \cdots, K\}$.

Based on their time value θ and the queue lengths of both stations, L_1 and L_2, each customer will compare the cost of entering the gratis station $\theta \frac{L_1(t)}{\mu_1}$ and the cost of entering the toll station $p + \theta \frac{L_2(t)}{\mu_2}$, and choose the line with a lower value to enter.

This is a reasonable assumption. If an arrival customer whose acceptable price $\overline{p}$ is higher than p, then he/she will possibly choose to enter the toll system. Otherwise, he/she will choose

DOI:10.6814/NCCU201901241

to enter the gratis system. It is worth noting that when an arrival customer whose time value leads two values in (2.3.1) indifferent, we assume that he/she will enter the gratis station. When the system reaches stability, we have $lim_{t\to\infty}P(L_1(t)=n, L_2(t)=m)=\pi_{nm}$. We consider a QBD process $X(t)$ with system states in a countable set $\mathcal{X}$. Let the system states in $\mathcal{X}$ be written as (n,m), where in the state description the first entry $n=0,1,\cdots$ indicates the queue length of the gratis station and the second entry $m=0,1,2,\cdots,K$ indicates the queue length of the toll station. We rewrite (2.3.1) as

$$\theta(n,m)=p/(n/\mu_1-m/\mu_2) \tag{2.3.2}$$

at state (n,m) with $m<K$. Then, we have

$$\begin{aligned}\lambda_1(n,m)&=\Lambda F(\frac{p}{n/\mu_1-m/\mu_2})\\ \lambda_2(n,m)&=\Lambda(1-F(\frac{p}{n/\mu_1-m/\mu_2})).\end{aligned} \tag{2.3.3}$$

If the queue of the gratis station is too long, this will result in that customers are not willing to join. In other words, as the queue length n increases and the queue length m is less than K, λ_1 becomes smaller and smaller. Given that ϵ is a small positive value, there exists a queue length n_0 of the gratis station such that $\lambda_1<\epsilon$ at $n>n_0$ and $m<K$. By (2.3.3), we estimate the lowest n_0 as

$$n_0=\lceil\frac{p\mu_1}{F^{-1}(\epsilon/\Lambda)}+K\frac{\mu_1}{\mu_2}\rceil.$$

Therefore, we assume that when the queue length of the gratis station is greater than n_0 and the queue length of the toll station is less than K, all customers will preferentially enter the toll station if the price is paid. In addition, there is an advantage of using n_0 as a threshold point to speedup the computation of the stationary probability when its convergence is slow. Thus, the arrival rate of this system can be completely written

DOI:10.6814/NCCU201901241

$$\lambda_1(n,m) = \begin{cases} \Lambda & \text{if } n \le m(\mu_1/\mu_2) \text{ or } m = K, \\ \Lambda F(\frac{p}{n/\mu_1 - m/\mu_2}) & \text{if } m(\mu_1/\mu_2) < n < n_0 \text{ and } m < K, \\ 0 & \text{if } n = n_0 \text{ and } m < K. \end{cases}$$

$$\lambda_2(n,m) = \Lambda - \lambda_1(\theta).$$

In this system, we consider an advantage of the toll station in which there is a guaranteed expected waiting time. The following argument can be clearly understood. First, when the queue length of the gratis station is short or the toll station is full, costumers must choose a gratis station. Second, when the queue length of the gratis station is slightly longer and the toll station is not full, costumers will make choices based on their own time value. Finally, when the queue length of the gratis station is too long and the toll station is still not full, costumers paying p only selects the toll station, as mentioned above. As in the previous section, we can describe the steady state conditions as follows.

Proposition 2.3.1. With real-time queue length information, the two-tier service system reaches the steady state if

$$\mu_1 > \frac{\left(1 - \frac{\Lambda}{\mu_2}\right)\left(\frac{\Lambda}{\mu_2}\right)^K}{1 - \left(\frac{\Lambda}{\mu_2}\right)^{K+1}}\Lambda. \tag{2.3.4}$$

Proof. (see Appendix A.5) □

Consider the stability condition of the gratis station, when the queue length of the gratis station becomes very long, if all customers try to enter the toll station, then the gratis station must be able to accommodate the spilled customers to under the complete service assumption. In this case, the arrival rate of the toll station λ_2 get close to Λ, and the probability that the toll station is full is obtained by (2.2.3).

Let $\boldsymbol{Q}$ denote the infinitesimal generator matrix. The process $X(t)$ is referred to as a QBD process if the transitions allowed are unchanging or increasing or decreasing only one person in

DOI:10.6814/NCCU201901241

the queue. Here, we form the infinitesimal generator matrix $\boldsymbol{Q}$ as

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{B}_{00} & \boldsymbol{C}_{00} & & & \\ \boldsymbol{A}_{00} & \boldsymbol{B} & \boldsymbol{C} & & \\ & \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{C} & \\ & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{2.3.5}$$

$$\boldsymbol{B}_{00} = \begin{bmatrix} \boldsymbol{B}_0 & \boldsymbol{C}_0 & & \\ \boldsymbol{A} & \boldsymbol{B}_1 & \boldsymbol{C}_1 & \\ & \ddots & \ddots & \vdots \\ & & \boldsymbol{A} & \boldsymbol{B}_{n_0-1} \end{bmatrix}_{n_0(K+1)\times n_0(K+1)},$$

$$\boldsymbol{C}_{00} = \begin{bmatrix} \vdots \\ \boldsymbol{C}_0 \end{bmatrix}_{n_0(K+1)\times(K+1)}, \boldsymbol{A}_{00} = \begin{bmatrix} \cdots & \boldsymbol{A} \end{bmatrix}_{(K+1)\times n_0(K+1)}.$$

The matrix $\boldsymbol{B}_n$ $n = 0, 1, \cdots, n_0 - 1$ and $\boldsymbol{B}$ indicate that the system state with the queue length of the gratis station unchange. The matrices $\boldsymbol{C}_n$ $n = 0, 1, \cdots, n_0 - 1$ and $\boldsymbol{C}$ indicate the case of decreasing one person in queue length of the gratis station, and $\boldsymbol{A}$ indicate increasing.

The matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and all elements of $\boldsymbol{A}_{00}$, $\boldsymbol{B}_{00}$ and $\boldsymbol{C}_{00}$ are $(K+1) \times (K+1)$ matrices, and details are explained as follows.

DOI:10.6814/NCCU201901241

$$\boldsymbol{B}_0 = \begin{bmatrix} -\Lambda & \lambda_2(n,0) & & & & \\ \mu_2 & -(\Lambda+\mu_2) & \lambda_2(n,1) & & & \\ & \ddots & \ddots & \ddots & & \\ & & \mu_2 & -(\Lambda+\mu_2) & \lambda_2(n,K-1) & \\ & & & \mu_2 & -(\Lambda+\mu_2) & \end{bmatrix},$$

$$\boldsymbol{C}_0 = \begin{bmatrix} \lambda_1(0,0) & & & & \\ & \lambda_1(0,1) & & & \\ & & \ddots & & \\ & & & \lambda_1(0,K-1) & \\ & & & & \Lambda \end{bmatrix}.$$

DOI:10.6814/NCCU201901241

For $n = 1, \cdots, n_0 - 1$,

$$\boldsymbol{B}_n = \begin{bmatrix} -(\mu_1+\Lambda) & \lambda_2(n,0) & & & & \\ \mu_2 & -(\Lambda+\mu_2+\mu_1) & \lambda_2(n,1) & & & \\ & \ddots & \ddots & \ddots & & \\ & & \mu_2 & -(\Lambda+\mu_2+\mu_1) & \lambda_2(n,K-1) & \\ & & & \mu_2 & -(\Lambda+\mu_2+\mu_1) \end{bmatrix},$$

$$\boldsymbol{C}_n = \begin{bmatrix} \lambda_1(n,0) & & & \\ & \lambda_1(n,1) & & \\ & & \ddots & \\ & & & \lambda_1(n,K-1) \\ & & & \Lambda \end{bmatrix}.$$

DOI:10.6814/NCCU201901241

For $n \geq n_0$,

$$\boldsymbol{A} = \mu_1 \boldsymbol{I} = \begin{bmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_1 \end{bmatrix},$$

$$\boldsymbol{B} = \begin{bmatrix} -(\mu_1+\Lambda) & \Lambda & & & \\ \mu_2 & -(\Lambda+\mu_2+\mu_1) & \Lambda & & \\ & \ddots & \ddots & \ddots & \\ & & & \mu_2 & -(\Lambda+\mu_2+\mu_1) \end{bmatrix},$$

$$\boldsymbol{C} = \begin{bmatrix} \ddots & \\ & \Lambda \end{bmatrix}.$$

The elements of the matrices $\boldsymbol{A}$, $\boldsymbol{B}_n$, $\boldsymbol{B}$, $\boldsymbol{C}_n$ and $\boldsymbol{C}$ on the main diagonal, $n = 0, 1, \cdots, n_0 - 1$, indicate the rate of unchanging the queue length at the toll station. In addition, the elements on the first diagonal below and above the main diagonal indicate the rate of increasing and decreasing one person in the queue at the toll station, respectively.

For the convenience of calculation, we define the stationary probability vector as

$$\begin{aligned} \boldsymbol{\pi}_0 &= [\boldsymbol{\pi}_0(0), \boldsymbol{\pi}_0(1), ..., \boldsymbol{\pi}_0(n_0)], \\ &= [\pi_{00}, \pi_{01}, ..., \pi_{0K}, \pi_{11}, ..., \pi_{1K}, ..., \pi_{n_0 K}], \\ \boldsymbol{\pi}_k &= [\pi_{(n_0+k)0}, \pi_{(n_0+k)1}, ..., \pi_{(n_0+k)K}], \end{aligned}$$

where $k = 0, 1, \cdots$. The size of these vectors correspond exactly to the matrix (2.3.5). By the matrix geometric method, solutions are given by

$$\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_k \boldsymbol{R}, k \geq 2, \tag{2.3.6}$$

where $\boldsymbol{R}$ is the rate matrix. Note that the matrix $\boldsymbol{C}$ has a value only at the end of the diagonal,

DOI:10.6814/NCCU201901241

which causes the rate matrix $\boldsymbol{R}$ to be a special form. Let

$$\boldsymbol{P} = \begin{bmatrix} 0 & \cdots & \mu_1 \\ & \ddots & \vdots \\ & & \mu_1 \end{bmatrix}.$$

More precisely, the rate matrix $\boldsymbol{R}$ can be replaced by an eigenvalue σ of a matrix $\boldsymbol{K}$ [24], where

$$\boldsymbol{K} = -[\boldsymbol{B} + \boldsymbol{P}]\boldsymbol{A}^{-1} \tag{2.3.7}$$

and $0 < \sigma < 1$.

The probability vectors $\boldsymbol{\pi}_0$, $\boldsymbol{\pi}_1$, $\boldsymbol{\pi}_2$ and $\boldsymbol{\pi}_3$ can be obtained by solving following equation.

$$\begin{bmatrix} \boldsymbol{\pi}_0 & \boldsymbol{\pi}_1 & \boldsymbol{\pi}_2 & \boldsymbol{\pi}_3 \end{bmatrix} \boldsymbol{Q}' = \begin{bmatrix} 1, & 0, & \cdots, & 0 \end{bmatrix}_{1\times(n_0K+3(K+1)+1)}, \tag{2.3.8}$$

where

$$\boldsymbol{Q}' = \begin{bmatrix} \widehat{\boldsymbol{e}} & \boldsymbol{B}_{00} & \boldsymbol{C}_{00} & 0 & 0 \\ \widehat{\boldsymbol{e}} & \boldsymbol{A}_{00} & \boldsymbol{B} & \boldsymbol{C} & 0 \\ \widehat{\boldsymbol{e}} & 0 & \boldsymbol{A} & \boldsymbol{B} & diag(\sigma) \\ (1-\sigma)^{-1}\widehat{\boldsymbol{e}} & 0 & 0 & \boldsymbol{A} & diag(-1) \end{bmatrix},$$

where $diag(x)$ is a diagonal matrix with nonzero are x and $\widehat{\boldsymbol{e}}$ is a column vector which all elements are 1. In order to solve the stationary probability, $\boldsymbol{Q}'$ can ignore a column within it which is linear dependence, and rewrite the equation (2.3.8).

For given $p > 0$, the queue length of the gratis and toll stations are denoted by L_1 and L_2,

DOI:10.6814/NCCU201901241

respectively. The expected queue lengths of the gratis and toll stations are given by

$$E[L_1] = \left[\sum_{n=0}^{n_0-1} n\boldsymbol{\pi}_0(n) + n_0\boldsymbol{\pi}_1 + (n_0(1-\sigma)+1)(1-\sigma)^{-2}\boldsymbol{\pi}_2\right]\widehat{\boldsymbol{e}},$$

$$E[L_2] = \left[\sum_{n=0}^{n_0-1} \boldsymbol{\pi}_0(n) + \boldsymbol{\pi}_1 + \boldsymbol{\pi}_2(1-\sigma)^{-1}\right]\boldsymbol{L}, \text{ where } \boldsymbol{L} = [\,0,\ 1,\ 2,\ \cdots,\ K\,]^T. \tag{2.3.9}$$

In the real-time case, since the arrival rate is state-dependent, the average effective arrival rate is difficult to calculate. This makes it strait to use the Little's law to calculate the average waiting time. However, with stationary probability, we can still get the distribution of each waiting time.

$$\begin{aligned} P(W_1 > t) = &\sum_{n=0}^{n_0-1} \boldsymbol{\pi}_0(n)T_n(\mu_1 t)e^{-\mu_1 t}\widehat{\boldsymbol{e}} + \sum_{n=1}^{2} \boldsymbol{\pi}_n T_{n_0-1+n}(\mu_1 t)e^{-\mu_1 t}\widehat{\boldsymbol{e}} \\ &+ \boldsymbol{\pi}_2\left(T_{n_0+1}(\mu_1 t)\sigma(1-\sigma)^{-1} + R_{n_0+1}(\mu_1 t\sigma)\sigma^{-(n_0+1)}(1-\sigma)^{-1}\right)e^{-\mu_1 t}\widehat{\boldsymbol{e}} \end{aligned}$$

$$P(W_2 > t) = \left[\sum_{n=0}^{n_0-1} \boldsymbol{\pi}_0(n) + \boldsymbol{\pi}_1 + \boldsymbol{\pi}_2(1-\sigma)^{-1}\right]\begin{bmatrix} T_0(\mu_2 t) & T_1(\mu_2 t) & \cdots & T_K(\mu_2 t) \end{bmatrix}^T e^{-\mu_2 t} \tag{2.3.10}$$

where $T_i(x) = \sum_{j=0}^{i} x^j/j!$, $R_i(x) = e^x - \sum_{j=0}^{i} x^j/j!$.

Then, we use this to calculate the expected waiting time. So far, we have used the matrix geometric method to calculate the stationary probabilities for real time and non-real time separately, which well describes the system probabilistic behavior at steady state. In the next chapter, we will consider optimizing the profitability of the toll station and the expected waiting cost of the society as whole.

DOI:10.6814/NCCU201901241

Chapter 3

An Optimization Model

We discuss two optimization problems in this chapter, considering two different perspectives with respect to no real-time and real-time information scenarios. First, if the toll station is a government agency or social welfare such as at the emergency room or on the highway, the purpose is to enable passengers with urgent or high waiting cost to be quickly serviced, thereby reducing the expected social waiting cost. To minimize the expected social waiting cost for customers is the objective of the perspective of the society. If the toll station is a private company, to maximize his gain naturally is the objective of the perspective of the manager.

3.1 The Perspective of the Society

From the perspective of the society, the extra pay as part of the service fee provides a faster service for the customer. This type of payment is a price discrimination strategy that distinguishes customers with urgency or high time costs and it expects to minimize the total waiting cost.

3.1.1 The Perspective of The Society in No Real-time Information Scenario

For no real-time information scenario, each customer will compare the cost when entering the gratis station $\theta E[W_1]$ and the cost of entering the toll station $p + \theta E[W_2]$, and choose the lower one to enter. Therefore, customers will be divided into two groups based on the balanced waiting time cost rate satisfies (2.2.1). The customers with the greater time value will choose the toll station and generate the expected waiting cost $\theta E[W_2]$ and the customers with the lower time

DOI:10.6814/NCCU201901241

value will choose the gratis system and generate the expected waiting cost $\theta E[W_1]$. Consider the expected value function of the distribution function of customers' time value $G(x)$, we get the expected social waiting cost for each customer in the system, expressed as $S_1(\overline{\theta})$, where

$$S_1(\overline{\theta}) = E[W_1]G(\overline{\theta}) + E[W_2](\xi - G(\overline{\theta})). \tag{3.1.1}$$

To simplify the analysis, we use $\overline{\theta}$ instead of p as the control variable. As shown below, there is a unique correspondence between $\overline{\theta}$ and p, they can be interchanged as control variables for each other, denoted by $\overline{\theta}(p)$ or $p(\overline{\theta})$.

Proposition 3.1.1. For no real-time information scenario, in the steady state, there is a unique correspondence between $\overline{\theta}$ and $p > 0$.

Proof. For the balanced waiting time cost rate $\overline{\theta}$ satisfies (2.2.1) which can be re-written as

$$p = \overline{\theta}(E[W_1(\overline{\theta})] - E[W_2(\overline{\theta})]). \tag{3.1.2}$$

According to the second last paragraph of proof of Proposition 2.2.2, we knew that $E[W_1(\overline{\theta})] - E[W_2(\overline{\theta})]$ is positive and increasing with $\overline{\theta}$. Therefore, when $\overline{\theta}$ increases, $E[W_1(\overline{\theta})] - E[W_2(\overline{\theta})]$ cannot be reduced, then p must be strictly increasing.

It shows that price p is a strictly increasing function of $\overline{\theta}$, and so do its inverse function. That is, $\overline{\theta}$ is also a strictly increasing function of price p. □

This means that as price p increase, more customers will be assigned to the gratis station. Inversely, in order to make customers to join the toll station, the price p must be decreased.

Mentioned in the the previous chapter, the condition $E[W_1(\overline{\theta})] > E[W_2(\overline{\theta})]$ and condition $\mu_1 \geq \lambda_1^{\text{eff}}$ provide a left and right boundary of θ, respectively. The $\overline{\theta}$ must be in this interval. In steady state, we show that $S_1(\overline{\theta})$ has a global minimum in this interval.

Proposition 3.1.2. In the steady state, $S_1(\overline{\theta})$ is a convex function of $\overline{\theta}$ when ξ exists. And the feasible region of $\overline{\theta}$ is what satisfies $E[W_1(\overline{\theta})] > E[W_2(\overline{\theta})]$ and $\mu_1 \geq \lambda_1^{\text{eff}}$.

Proof. (see Appendix A.6) □

It implies that there exists a unique watershed $\overline{\theta}$ that minimizes the expected social waiting cost for the customers. If ξ does not exist, then

DOI:10.6814/NCCU201901241

$$
\begin{aligned}
S_1(\overline{\theta}) &\geq \min(E[W_1], E[W_2])G(\overline{\theta}) + \min(E[W_1], E[W_2])(\xi - G(\overline{\theta})) \\
&= \min(E[W_1], E[W_2])\xi \\
&= \infty.
\end{aligned}
$$

Therefore the existence of ξ is necessary. And $\mu_1 \geq \lambda_1^{\text{eff}}$ is the stability condition. In addition, although the queue of the toll station has a finite buffer for a service quality guarantee, as the perspective of the society, we do not want the queue of the gratis station to be too long. Therefore, for given an irritating waiting time t and the acceptable risk c based on t, as a constraint, we consider

$$
P(W_1(\overline{\theta}) > t) < c.
$$

This means that the probability of the gratis station waiting for more than t does not exceed c. In other words, although customers entering the gratis station may wait longer than expected, but it is guaranteed that this will not happen often. Besides, it can be calculated by (2.2.9). We may choose $t = E[W_1]$ and $c = 0.5$, so that the probability of waiting time of customers more than the expected waiting time do not exceed a half. Based on the above discussion, the optimization problem we considered can be written as given $c > 0$

$$
\begin{aligned}
\min_{p} \quad & Z = S_1(\overline{\theta}(p)) \\
\text{subject to} \quad & \mu_1 \geq \lambda_1^{\text{eff}}, \\
& E[W_1] > E[W_2], \\
& P(W_1 > t) < c, \\
\text{and } & p > 0.
\end{aligned}
\tag{3.1.3}
$$

For this optimization problem at the perspective of the society in no real-time information scenario, we want to minimize the expected social waiting cost of the customer. The first constraint is the stability condition. The second constraint is caused by (2.2.1) and the positive price p. Therefore, the solution of this optimization problem must comply with $E[W_1] > E[W_2]$. Finally, the third constraint is just discussed above and the toll price is positive.

We compute this optimal balanced $\overline{\theta}$ by "fmincon" of Matlab code (see Appendix B.3.1). It uses the interior-point approach to constrained minimization to solve a sequence of approximate minimization problems.

DOI:10.6814/NCCU201901241

Finally, we get the optimal price p^* for the perspective of the society in no real-time information scenario by computing $p^* = \overline{\theta}(E[W_1(\overline{\theta})] - E[W_2(\overline{\theta})])$ from (2.2.1).

3.1.2 The Perspective of the Society in Real-time Information Scenario

For a real-time information scenario, the customer will be provided the queue length information for both queues when he/she arrives. There are different watershed $\overline{\theta}$ depending on the system states (n, m), so we need to calculate them individually according to different system states, and consider the fallowing function of the expected social waiting cost

$$S_2(p) = \sum_{n=0}^{\infty}\sum_{m=0}^{K}\pi_{nm}\left(\frac{n}{\mu_1}G(\overline{\theta}) + \frac{m}{\mu_2}(\xi - G(\overline{\theta}))\right) \tag{3.1.4}$$

where $\overline{\theta}(n, m) = p/(n/\mu_1 - m/\mu_2)$ at state (n, m).

Proposition 3.1.3. In steady state, $S_2(p)$ are finite when given $p > 0$ and ξ exists.

Proof. We have

$$\begin{aligned}
S_2(p) &= \sum_{n=0}^{\infty}\sum_{m=0}^{K}\pi_{nm}\left(n/\mu_1 G(\overline{\theta}) + m/\mu_2(\xi - G(\overline{\theta})\right) \\
&\leq \sum_{n=0}^{\infty}\sum_{m=0}^{K}(\pi_{nm}\frac{n}{\mu_1}\xi) + \sum_{n=0}^{\infty}\sum_{m=0}^{K}(\pi_{nm}\frac{m}{\mu_2}\xi) \\
&= \frac{\xi}{\mu_1}\sum_{n=0}^{\infty}(\pi_{n\bullet}n) + \frac{\xi}{\mu_2}\sum_{m=0}^{K}(\pi_{\bullet m}m) \\
&= \frac{\xi}{\mu_1}E[L_1] + \frac{\xi}{\mu_2}E[L_2] \\
&< \infty.
\end{aligned}$$

The second and third sign comes from the $\xi \geq G(\overline{\theta})$ and the positive of all components. Finally, $E[L_1]$ and $E[L_2]$ must be finite in the steady state. Therefore, $S_2(p)$ are finite. □

This shows the existence of $S_2(p)$, but it is not enough. According to the choice of $F(\theta)$, it affects π_{nm} such that the critical point of the first-order differential function value of $S_2(p)$ is not unique. Therefore, we can only retreat to explain that the global minimum of $S_2(p)$ takes place at finite $p > 0$.

DOI:10.6814/NCCU201901241

When p approaches infinity, it will result in customers becoming more and more reluctant to pay for join the toll station, which makes the whole system regarded as an M/M/1 queueing system. In this case, the expected social waiting cost of the customer is $\sum_{n=0}^{\infty}(1-\rho)\rho^n \frac{n}{\mu_1}\xi$, where $\rho = \lambda_1^{\text{eff}}/\mu_1$. On the other hand, when $p = 0$, customers can freely choose to join the gratis or toll station according to the lengths of the queues and do not have to pay any fee, which makes both stations fully utilized. That is to say, compared with an M/M/1 queueing system, the service rate is greater and the length of queue is divided into two, but the arrival rate is unchanged. In this case, the expected social waiting cost of customers must be smaller than that in M/M/1 queueing system which is the case of p approaching infinity. According to the continuity of $S_2(p)$, its minimum value must takes place at the finite p.

The optimization problem we considered can be written as given $c > 0$

$$\begin{aligned} \min_{p} \quad & Z = S_2(p) \\ \text{subject to} \quad & P(W_1 > t) < c, \\ \text{and } & p > 0. \end{aligned} \tag{3.1.5}$$

For this optimization problem at the perspective of the society in a real-time information scenario, we want to minimize the expected social waiting cost of customers $S_2(p)$. We need to check the stability condition in advance. From (2.3.4), we find that the stability condition is actually irrelevant to p. Similarly, we don't have to check $E[W_1(\overline{\theta})] > E[W_2(\overline{\theta})]$ because the decision variable is already the positive price p. Finally, the third constraint is some as previous section.

As mentioned above, depending on $F(\theta)$, the local minimum may exist more than once. We can't use Newton's method, which usually doesn't converge to the global minimum if there is more than one local minimum. Therefore, we use Matlab code (see Appendix B.4.1) to calculate the global minimum value of this optimization problem, which is based on the global search algorithm proposed by Ugray et al [35]. The problem we consider is a public healthcare system with a parallel private department, for example, a public hospital and a private hospital. The scale of such a problem, the number of customers in queue is only a few dozen implying $n_0 < 200$ in average. This processor in Matlab is enough to solve.

DOI:10.6814/NCCU201901241

3.2 The Perspective of the Manager

From the perspective of managers, it must provide better service to compete with other lower toll stations. If the manager increase the toll, it will reduce the arrival rate. On the contrary, if the manager want to increase the arrival rate, he/she must lower the toll. The manager must get a balance between price and the arrival rate to get the maximum the expected profit.

In this section, the assumption of $\Lambda < \mu_1$ is necessary. If the gratis station can not serve all the customers and reach a stable state, then there are some customers will be forced to choose to join the toll station. Therefore the toll station will be able to keep raising the price in order to alleviate the wait in the toll queue.

3.2.1 The Perspective of the Manager in No Real-time Information Scenario

For no real-time information scenario, each customer will compare the cost when entering the gratis station $\theta E[W_1]$ and the cost of entering the toll station $p + \theta E[W_2]$, and choose the lower one to enter. Therefore, customers will be divided into two groups based on the balanced waiting time cost rate satisfies (2.2.1). The customer with the greater time value will choose the toll station and the customer with the lower time value will choose the gratis system. For the perspective of the manager, we only focus on the arrival rate of the toll station and our price p. We get the expected revenue per unit time of the toll system, expressed as $O_1(\overline{\theta})$, where

$$O_1(\overline{\theta}) = \lambda_2^{\text{eff}}(\overline{\theta})p(\overline{\theta}) \tag{3.2.1}$$

By Proposition 3.1.1, we use $\overline{\theta}$ instead of p as the control variable.

Proposition 3.2.1. For no real-time information scenario, in the steady state, $O_1(\overline{\theta})$ is finite for a given finite $\overline{\theta}$. In addition, if $(1 - F(\overline{\theta}))\overline{\theta}$ converges when $\overline{\theta} \to \infty$, so does $O_1(\overline{\theta})$.

Proof. We have

$$\begin{aligned} O_1(\overline{\theta}) &= \lambda_2^{\text{eff}} p \\ &= (1 - \pi_{\bullet K})(1 - F(\overline{\theta}))\overline{\theta}(E[W_1] - E[W_2]). \end{aligned}$$

In the steady state, both $E[W_1]$ and $E[W_2]$ are finite, such that $E[W_1] - E[W_2]$ is finite. For (3.2.1), we see that λ_2^{eff} and $p = \overline{\theta}(E[W_1(\overline{\theta})] - E[W_2(\overline{\theta})])$ are finite at a finite $\overline{\theta}$.

DOI:10.6814/NCCU201901241

In addition, when $\overline{\theta} \to \infty$, $(1 - \pi_{\bullet K})$ is bounded by 1, if $(1 - F(\theta))\theta$ is convergent to zero, so is $O_1(\overline{\theta})$. On the other hand, when $\overline{\theta} \to \infty$, λ_2 approaches to zero, so do $\pi_{\bullet K}$ and $E[W_2]$. If $(1 - F(\theta))\theta$ is divergent, $O_1(\overline{\theta})$ is divergent. □

That implies that the maximum of $O_1(\overline{\theta})$ exists in a finite interval of $\overline{\theta}$. If there is an $F(\theta)$ such that $(1-F(\theta))\theta$ approaches infinity as θ approaches infinity, which implies a situation that some people are extremely rich, then $O_1(\theta)$ diverges. Although the gratis station is sufficient to meet the demand, it is foreseeable that the toll station can continue to raise the toll price to please those rich. In this case, a strategy for the manager will be considering $p = \infty$. However, it is obviously not in line with social interests as well as in a healthcare system. The intervention of government might be necessary.

In addition, although there are already guarantees for the expected waiting time, manager may hope to provide better service to customers. Therefore, for given an irritating waiting time t and the acceptable risk c based on t as a constraint, we consider

$$P(W_2(\overline{\theta}) > t) < c.$$

This means that the probability of the toll station waiting time more than t does not exceed c. If let $t = E[W_2]$, more than $(1 - c)$ percentages of the customers who enter the toll station will not wait longer than the expected waiting time. Besides, it can be calculated by (2.2.9).

The optimization problem we considered can be written as given $c > 0$

$$\begin{aligned} \max_{p} \quad & Z = O_1(\overline{\theta}(p)) \\ \text{subject to} \quad & E[W_1] > E[W_2], \\ & P(W_2 > t) < c, \\ \text{and } & p > 0. \end{aligned} \tag{3.2.2}$$

For this optimization problem at the perspective of the manager in no real-time information scenario, we want to maximize the expected revenue per unit time of the toll system. Under the premise of $\Lambda < \mu_1$, we don't need to check the stability condition because the system is always stable. The first constraint is caused by (2.2.1) and the positive price p. Therefore, the solution of this optimization problem must comply with $E[W_1] > E[W_2]$. Finally, the toll price p is

DOI:10.6814/NCCU201901241

positive.

As mentioned above, depending on $F(\theta)$, a local maximum may exist. Therefore, we also use Matlab code (see Appendix B.3.1) to calculate the global minimum value of this optimization problem by the global search algorithm.

3.2.2 The Perspective of the Manager in Real-time Information Scenario

For a real-time information scenario, the customer will be provided the queue length information for both queues when he/she arrives. There are different watershed $\overline{\theta}$ of θ depending on the system states (n, m). We need to calculate them individually according to different system states, and have the expected revenue

$$O_2(p) = \sum_{n=0}^{\infty} \sum_{m=0}^{K} \pi_{nm} \lambda_2^{\text{eff}}(n, m) p. \tag{3.2.3}$$

Proposition 3.2.2. In the steady state, for given $p > 0$, $O_2(p)$ is finite.

Proof. Since $\lambda_2^{\text{eff}}(n, m)$ is bounded by Λ and π_{nm} can be treated as weights, we have

$$\begin{aligned} O_2(p) &= \sum_{n=0}^{\infty} \sum_{m=0}^{K} \pi_{nm} \left(\lambda_2^{\text{eff}}(n, m) p \right) \\ &\leq \sum_{n=0}^{\infty} \sum_{m=0}^{K} \pi_{nm} \Lambda p \\ &= \Lambda p \end{aligned}$$

□

This shows the existence of $O_2(p)$, but it is not enough. According to the choice of $F(\theta)$, it affects π_{nm} such that the critical point of the first-order differential function value of $O_2(p)$ is not unique. In particular, when $F(\theta)$ converges to 1 very slowly, that is, in a situation in which some people are extremely rich, a strategy of the toll station might keep raising the toll price to please those rich. However, when p approaches infinity, n_0 approaches to ∞, and it causes the difficult of analysis. Therefore, we can only find the global maximum of $O_2(p)$ taking place at finite $p > 0$, and the optimal solution is considered to be divergent when it reaches a large number of p.

DOI:10.6814/NCCU201901241

The optimization problem we considered can be written as

$$\begin{aligned} \max_{p} \quad & Z = O_2(p) \\ \text{subject to} \quad & P(W_2 > t) < c, \\ \text{and } & p > 0. \end{aligned} \tag{3.2.4}$$

For this optimization problem at the perspective of the manager in a real-time information scenario, we want to maximize the expected revenue per unit time of the toll system $O_2(p)$. We need to check the stability condition in advance. From (2.3.4), we find that the stability condition is actually irrelevant to p. Similarly, we don't have to check $E[W_1(\overline{\theta})] > E[W_2(\overline{\theta})]$ because the decision variable is already the positive price p.

As mentioned above, depending on $F(\theta)$, a local maximum may exist. Therefore, we also use Matlab code (see Appendix B.4.1) to calculate the global maximum value of this optimization problem.

DOI:10.6814/NCCU201901241

Chapter 4

Numerical Examples and Discussion

Given the solutions earlier, we can calculate and describe the performance of this two-tier service system well. In this section, we will compare the performance of the two information scenarios. In fact, the gratis station usually represents government-funded services, and the toll station represents privately-provided services for profit. Therefore, the toll price and the buffer size of the toll station are more likely to be important decision variables in the two-tier service system.

4.1 Parameters

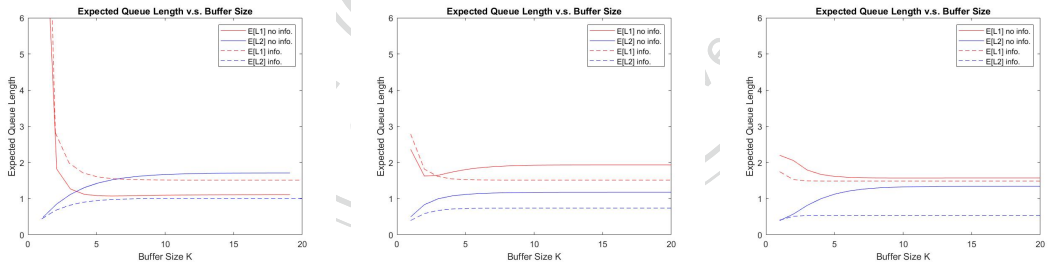
We use the following parameters for the numerical studies. We standardized the customer arrival rate Λ to be 1. And the service rates of the stations are $(\mu_1, \mu_2) = (7/6, 3/6), (5/6, 5/6), (3/6, 7/6)$. Note that we set the values of μ_1 and μ_2 to study the effects of symmetry between the two stations. Additionally, we consider three different distributions for the customers' waiting cost rate θ. There are uniform distributions in the interval $[0, 20]$, exponential distributions with mean 10, and Pareto distributions with $k = 10/9$ and $xm = 1$. The summary of selected probability distributions are given in Appendix A.7. For comparison purposes, we set the expected values for all three distributions to be 10. The variances of the three distributions are 33.33, 100 and ∞, respectively.

DOI:10.6814/NCCU201901241

4.2 Real-time and No Real-time Information Scenario

We find that the trend of the expected queue length are similar under the two information scenarios, i.e., real-time information and no real-time information, denoted by info. and no info.. Figure 4.1, Figure 4.2 and Figure 4.3 show how the expected queue length varies with the buffer size of the two-tier service system in different distribution of time value random variable Θ. As shown in Figure 4.1, Figure 4.2 and Figure 4.3, for a given price, as the buffer size increases, the toll station gradually has enough queue length to serve customers, which results in that the expected queue length of the gratis and toll stations become stable.

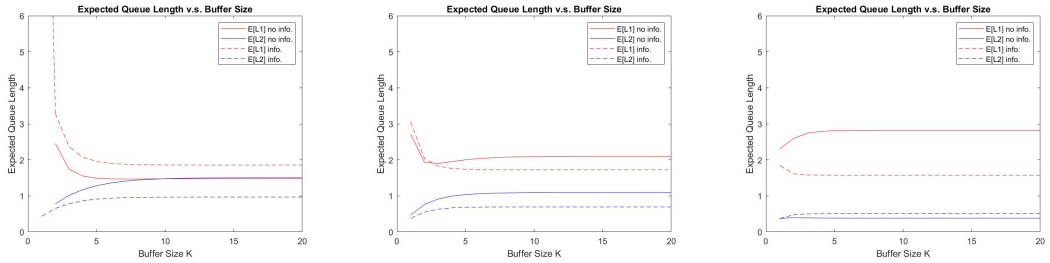
To compare the expected queue length under the different distributions of time value random variables at the same service rate, we try more than 100 numerical experiments. No matter what service rates in real-time or no real-time information scenario, the expected queue length of the gratis stations in the cases of $\Theta \sim U(0, 20)$ are shorter than the case of $\Theta \sim Pa(1, 10/9)$ and $\Theta \sim Exp(1/10)$. The expected queue length of $\Theta \sim Pa(1, 10/9)$ are the longest among all cases of these distributions. On the other hand, the expected queue length of the toll stations in the cases of $\Theta \sim U(0, 20)$ becomes the longest and that in the cases of $\Theta \sim Pa(1, 10/9)$ becomes the shortest among all experimental tests.



(a) $\mu_1 = 3/6, \mu_2 = 7/6$ (b) $\mu_1 = 5/6, \mu_2 = 5/6$ (c) $\mu_1 = 7/6, \mu_2 = 3/6$

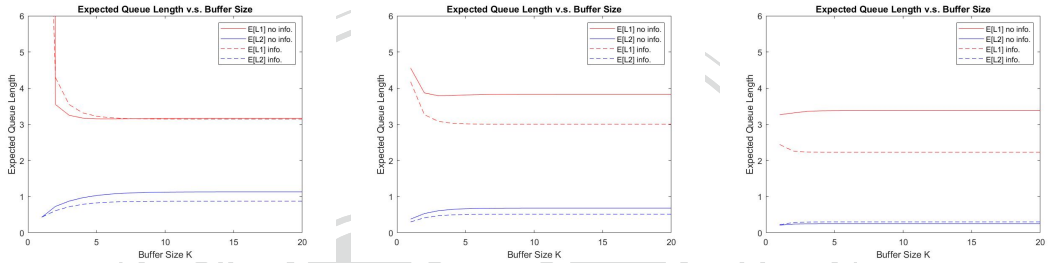
Figure 4.1: A two-tier service system with $p = 10$, $\Theta \sim U(0, 20)$

DOI:10.6814/NCCU201901241



(a) $\mu_1 = 3/6, \mu_2 = 7/6$ (b) $\mu_1 = 5/6, \mu_2 = 5/6$ (c) $\mu_1 = 7/6, \mu_2 = 3/6$

Figure 4.2: A two-tier service system with $p = 10$, $\Theta \sim Exp(1/10)$



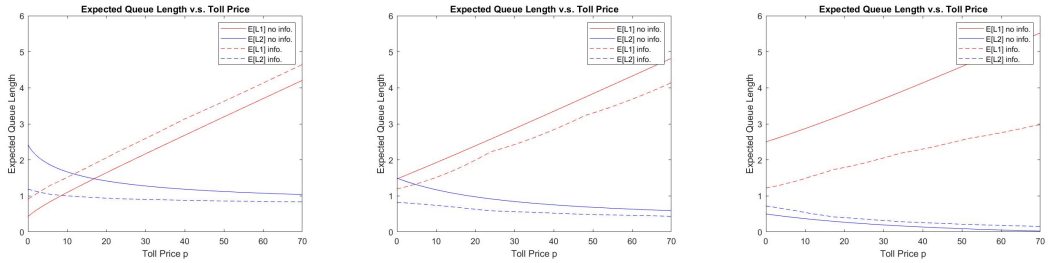
(a) $\mu_1 = 3/6, \mu_2 = 7/6$ (b) $\mu_1 = 5/6, \mu_2 = 5/6$ (c) $\mu_1 = 7/6, \mu_2 = 3/6$

Figure 4.3: A two-tier service system with $p = 10$, $\Theta \sim Pa(1, 10/9)$

Figure 4.4, Figure 4.5 and Figure 4.6 show the changes between the expected queue length and the toll price for the two-tier service system in different distributions of time value random variables. These results are consistent with the results presented by Chen et al. [5]. As shown in Figure 4.4, Figure 4.5 and Figure 4.6, for a given buffer size, when the toll price increases, customers become unwilling to pay and gradually turn to the gratis station. In this case, it causes the expected queue length difference between the gratis and toll station becomes more significant.

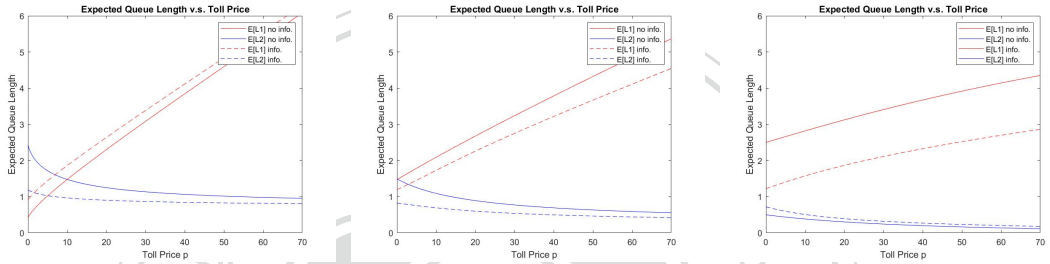
To compare the expected queue length under the different distributions of time value random variables at the same service rate, in the cases of $\Theta \sim U[0, 20]$, we find the change of the expected queue length for the gratis station is relatively flat, and the cases of $\Theta \sim Pr(1, 10/9)$ are most severe. This is because p is proportional to θ, and the cumulative distribution function of $\Theta \sim U(0, 20)$ is linear of θ. But there is a drastic change in the cumulative distribution function of $\Theta \sim Pa(1, 10/9)$.

DOI:10.6814/NCCU201901241



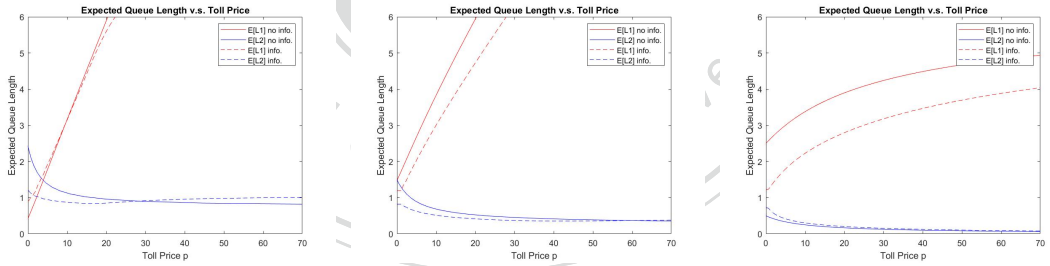
(a) $\mu_1 = 3/6, \mu_2 = 7/6$ (b) $\mu_1 = 5/6, \mu_2 = 5/6$ (c) $\mu_1 = 7/6, \mu_2 = 3/6$

Figure 4.4: A two-tier service system with $K = 10, \Theta \sim U(0, 20)$



(a) $\mu_1 = 3/6, \mu_2 = 7/6$ (b) $\mu_1 = 5/6, \mu_2 = 5/6$ (c) $\mu_1 = 7/6, \mu_2 = 3/6$

Figure 4.5: A two-tier service system with $K = 10, \Theta \sim Exp(1/10)$



(a) $\mu_1 = 3/6, \mu_2 = 7/6$ (b) $\mu_1 = 5/6, \mu_2 = 5/6$ (c) $\mu_1 = 7/6, \mu_2 = 3/6$

Figure 4.6: A two-tier service system with $K = 10, \Theta \sim Pa(1, 10/9)$

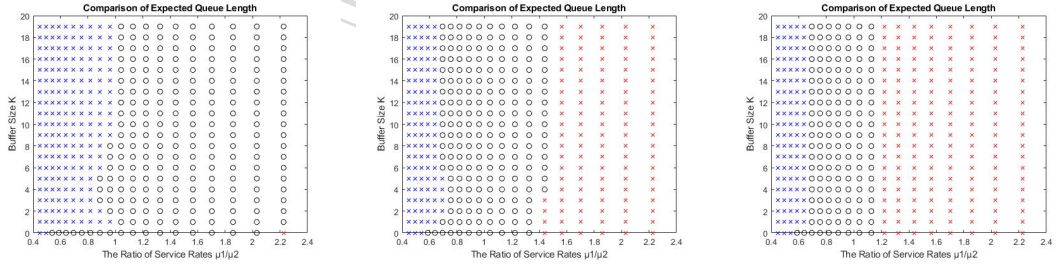
In short, in the same distribution of customers' time value of the case of $(\mu_1, \mu_2) = (7/6, 3/6)$, the system is always stable, and the queue length changes rather smoothly. Conversely, in the case of $(\mu_1, \mu_2) = (3/6, 7/6)$, the queue length changes relatively large. In addition, although the distributions are different, there still are some similar places in the case of the same service rate. In the case of $(\mu_1, \mu_2) = (5/6, 5/6)$, no matter how buffer size and toll price change, the expected queue length of the gratis and toll stations in real-time information scenario are always shorter than that in no real-time information scenario.

DOI:10.6814/NCCU201901241

In the case of $(\mu_1, \mu_2) = (3/6, 7/6)$, the expected queue length of the toll station in real-time information scenario are usually shorter than that in no real-time information scenario, but the expected queue length of the gratis stations in real-time information scenario are usually longer than that in no real-time information scenario. In the case of $(\mu_1, \mu_2) = (7/6, 3/6)$, the expected queue length of the gratis station in real-time information scenarios are shorter than in no real-time information scenario, but the expected queue length of the toll station in real-time information scenario are usually longer than that in no real-time information scenarios.

This situation arises from the asymmetry of μ_1 and μ_2. When μ_1 is similar to μ_2, the real-time information makes the customer to choose the service station wisely, which reduces their waiting cost effectively. When μ_1 and μ_2 are not symmetrical, one of the stations is easily crowded. If there is no real-time information, the customers will overestimate or underestimate the respective waiting costs. In the no real-time information scenario, taking $(\mu_1, \mu_2) = (3/6, 7/6)$ as an example, insufficient service rates of the gratis station often result in congestion.

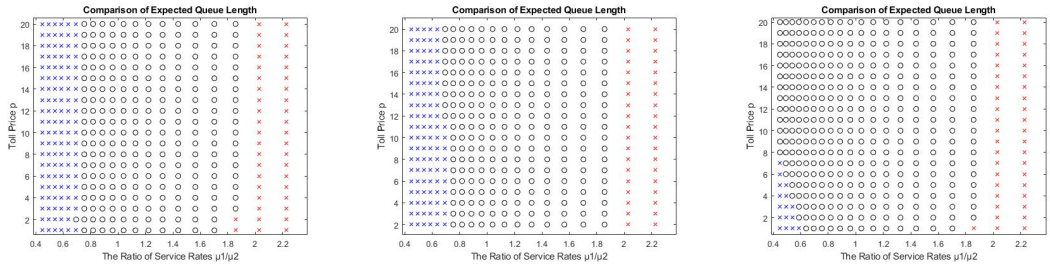
From Figure 4.7 and Figure 4.8, we make more comparison of these queue lengths between with and without information. In Figure 4.7 and Figure 4.8, "blue x" in the left block stands for the area of "$E[L_1]$ info.>$E[L_1]$ no info. but $E[L_2]$ info.<$E[L_2]$ no info.", "red x" in the right block stands for the area of "$E[L_1]$ info.<$E[L_1]$ no info. but $E[L_2]$ info.> $E[L_2]$ no info.", and "black o" in the middle block stands for the area of "both $E[L_1]$ info. and $E[L_2]$ info. are less than or equal to $E[L_1]$ no info. and $E[L_2]$ no info.", respectively.



(a) $\Theta \sim U(0, 20)$ (b) $\Theta \sim Exp(1/10)$ (c) $\Theta \sim Pa(1, 10/9)$

Figure 4.7: A comparison of $E[L_1]$ and $E[L_2]$ with K between with and without information.

DOI:10.6814/NCCU201901241



(a) $\Theta \sim U(0, 20)$ (b) $\Theta \sim Exp(1/10)$ (c) $\Theta \sim Pa(1, 10/9)$

Figure 4.8: A comparison of $E[L_1]$ and $E[L_2]$ with p between with and without information.

We find that when the ratio of μ_1 and μ_2 is small, providing real-time information will make the queue of the gratis station longer and the queue of the toll system shorter. When μ_1/μ_2 increase, $E[L_1]$ info. will gradually be exceeded by $E[L_1]$ no info. and $E[L_2]$ info. will gradually exceed $E[L_2]$ no info.. We observe the changes in these intervals from the figures. It is worth noting that the middle block, which is the part of black o, indicates that the queue length of the gratis or toll station are shorter in the scenario of real-time information.

In addition, regardless of the circumstances, the expected social waiting cost of real-time information scenario is lower than the expected social waiting cost of no real-time information scenario, $S_1(\frac{p}{E[W_1]-E[W_2]})$. Therefore, when a government is designing their two-tier healthcare system, it can make that the service rate of the gratis station and the toll station be similar, so that the queue lengths of both stations are significantly shorter.

4.3 The Perspective of the Society and the Manager

According to previous discussion, $S_1(\overline{\theta})$ in (3.1.1), $S_2(p)$ in (3.1.4), $O_1(\overline{\theta})$ in (3.2.1) and $O_2(p)$ in (3.2.3) are continuous functions, respectively. By choosing proper t and c, the optimal problems given in (3.1.3), (3.1.5), (3.2.2) and (3.2.4) contain nonempty feasible and compact sets.

Table 4.1 reports the optimal solution that minimize the expected social waiting cost. We compute the optimal balanced waiting time cost rate $\overline{\theta}$ and toll price p^* in each case. Subsequently, given $\overline{\theta}$ and p^*, the other measures are computed.

We observe that regardless of distributions of customers' time value, $E[L_1]$ and $E[L_2]$ are always incremented and decremented with the increment of μ_1/μ_2, respectively. This is reasonable, because when the (μ_1, μ_2) changes, the customers will naturally choose the station

DOI:10.6814/NCCU201901241

with a high service rate to reduce their waiting cost. Similarly, in no real-time information scenario, p^* is incremented when μ_1/μ_2 is raised. Because the price must be raised to prevent too many customers from entering the toll station with the decreasing service rates.

In addition, we find that with or without real-time information, the expected social waiting cost in uniform distribution is always the highest, while that in the pareto distribution is the lowest. This is because when the customers is more diversified (the variance in the distribution of the waiting cost of the customers becomes larger), the two-tier system based on the price discrimination strategy can more effectively segment the customers.

An interesting finding is that the expected social waiting cost in the real-time information scenario is lowest at $(\mu_1, \mu_2) = (5/6, 5/6)$ in the cases of $\Theta \sim U(0, 20)$ and $\Theta \sim Exp(1/10)$, and is decremented in the cases of $\Theta \sim Pa(1, 10/9)$. It is reasonable to suspect that the expected social waiting cost in the real-time information seems to be a convex upward function of μ_1/μ_2 where there is a global minimum for μ_1/μ_2. That is to say, when the government designs their two-tier healthcare system, if μ_1/μ_2 is a decision variable, there is a best division so that the expected social waiting cost is minimized.

The Table 4.2 reports the optimal solution that maximize the expected profit of the manager. We compute the optimal balanced waiting time cost rate $\overline{\theta}$ and the toll price p^* in each case. Subsequently, given $\overline{\theta}$ and p^*, the other measures are computed.

In the perspective of the manager, if there is not $\mu_1 > \lambda_1^{\text{eff}}$, this result in the optimal strategy for the toll station is to set a extreme expensive toll price and service barely a small part of customers. In this situation, the expected waiting time of the gratis station will be infinitely long without the toll station. Nevertheless, no matter how much the toll price is, there will always be a portion of customers who want to join the toll station. As such, to maximize the profits, the toll station will charge a very high toll price. This observation has an important signification the price regulations for the private operators is necessary for a public healthcare system.

In the case of $(\mu_1, \mu_2) = (7/6, 3/6)$ where the service rate of the gratis station is adequate to satisfy the demand, there is an maximized profit in each case except the case of $\Theta \sim Pa(1, 10/9)$ in real-time information scenario. The situations in Pareto distribution models where many customers have very high waiting cost in the real-time information scenario, customers will strategically choose the station to entry. This makes it necessary to impose more stringent

DOI:10.6814/NCCU201901241

Table 4.1: Optimal solutions minimizing of the expected social waiting cost with $\Lambda = 1, K = 10$, $c = 0.5$.

Distribution	(μ_1, μ_2)	info.	$\overline{\theta}$	p^*	$S_1(\overline{\theta})$, N $S_2(p^*)$, Y	$P(W_i > E[W_i])$ $i = 1, 2$	$(E[L_1], E[L_2])$
Uniform	(3/6, 7/6)	N	5.8514	16.0646	23.5454	(0.3679, 0.3753)	(1.4325, 1.4954)
		Y		3.3253	8.6617	(0.4023, 0.3735)	(1.1375, 1.0969)
Uniform	(5/6, 5/6)	N	10.8973	9.3363	28.7006	(0.3679, 0.3710)	(1.8937, 1.1892)
		Y		3.3810	8.6522	(0.3960, 0.3779)	(1.2862, 0.7981)
Uniform	(7/6, 3/6)	N	16.6663	0.0000	29.9984	(0.3679, 0.3679)	(2.4998, 0.5000)
		Y		11.4286	10.2117	(0.3877, 0.3805)	(1.5389, 0.5112)
Exponential	(3/6, 7/6)	N	3.9225	14.6571	22.2193	(0.3679, 0.3730)	(1.8681, 1.3484)
		Y		2.9864	8.5750	(0.4042, 0.3724)	(1.2509, 1.0696)
Exponential	(5/6, 5/6)	N	8.5792	12.5235	27.3897	(0.3679, 0.3696)	(2.2409, 1.0295)
		Y		3.7770	8.5602	(0.3981, 0.3764)	(1.3896, 0.7713)
Exponential	(7/6, 3/6)	N	17.9164	0.0000	29.9984	(0.3679, 0.3679)	(2.4998, 0.5000)
		Y		11.9368	9.8131	(0.3893, 0.3782)	(1.6358, 0.4798)
Pareto	(3/6, 7/6)	N	1.4983	8.1001	21.0884	(0.3679, 0.3711)	(2.6392, 1.1927)
		Y		2.0000	8.3410	(0.4048, 0.3721)	(1.1807, 1.0361)
Pareto	(5/6, 5/6)	N	2.5691	8.6623	24.0461	(0.3679, 0.3682)	(3.5338, 0.7251)
		Y		4.4563	8.2757	(0.4122, 0.3700)	(1.9269, 0.6419)
Pareto	(7/6, 3/6)	N	8.8727	13.2233	27.5071	(0.3679, 0.3679)	(3.5734, 0.2149)
		Y		15.6230	7.3948	(0.3968, 0.3681)	(2.5787, 0.2370)

DOI:10.6814/NCCU201901241

Table 4.2: Optimal solutions maximizing of the expected profit of the manager with $\Lambda = 1, K = 10, c = 0.5$.

Distribution (μ_1, μ_2)	info.	$\overline{\theta}$	p^*	$O_1(\overline{\theta})$, N $O_2(p^*)$, Y	$S_1(\overline{\theta})$, N $S_2(p^*)$, Y	$P(W_i > E[W_i])$ $i = 1, 2$	$(E[L_1], E[L_2])$
Uniform	N	18.5319	33.7518	2.4775	39.0785	(0.3679, 0.3679)	(3.8597, 0.1721)
(7/6, 3/6)	Y		60.6231	4.4189	17.9729	(0.3819, 0.3865)	(2.7681, 0.1813)
Exponential	N	28.5951	63.0855	3.6149	39.7715	(0.3679, 0.3679)	(4.2092, 0.1294)
(7/6, 3/6)	Y		108.1559	5.3497	20.6034	(0.3845, 0.3836)	(3.3851, 0.1196)
Pareto	N	63.1105	228.4382	2.2838	33.7667	(0.3679, 0.3679)	(5.6039, 0.0204)
(7/6, 3/6)	Y		1.5e+23	1.1e+20	32.0976	(0.3735, 0.3673)	(5.8406, 0.0037)

conditions so that the toll station cannot lift the price indefinitely. Furthermore, the social waiting cost in Table 4.2 are much higher than those in Table 4.1. This is especially true in the case of real-time information. This observation shows that although the gratis station has sufficient capacity to service all customers, price regulation may still be required.

We observe that $P(W_1 > E[W_1])$ without real-time information is always the same in Table 4.1 and Table 4.2. This is because the gratis station without real-time information is an M/M/1 waiting system. When we let $t = E[W_1]$, it makes $P(W_1 > E[W_1])$ constant and equal to e^{-1}.

In short, real-time information scenarios always perform better no matter what the circumstances. This observation is not surprising. When customers use the real-time information to decide the queue to enter, it is more accurate than the long-term statistical results. Consider the ratio of the expected social waiting costs of scenarios with real time information and no real time information $1 - S_2(p^*)/S_1(\overline{\theta})$ in Table 4.1. By providing real-time information in a two tier service system, we can reduce the expected social waiting cost per customer by more than 60%. This is a strong support to provide real-time information to public health systems with parallel private sectors that can improve the social welfare.

DOI:10.6814/NCCU201901241

Chapter 5

Conclusion

The study recognizes the impact of information on a two-tier service system and considers toll prices for heterogeneous strategic customers to reduce the expected social waiting costs or increase the expected profit of the manager. There is no charge for the gratis station. The toll station charges the price p but guarantees a faster service, where there is a limit on the number of customers in the toll system. Since the buffer of the toll system is limited, we assume that the overflowing customer will be transferred to the gratis station. Customers are heterogeneous in terms of their time value. We take the example of uniform distributions, exponential distributions and Pareto distributions to demonstrate the price effect by our analytic approach.

The customers select the service based on the their time value. There are two kinds of information scenarios to consider in the modeling, i.e., the real-time information and no real-time information. For the real-time information, customers knows the real-time queue length for each line. We model this scenario as a QBD process and calculate the stationary probability of the system under stable conditions.

For no real-time information, the customer makes a decision based on long-term statistics for both lines. Customers can be groups by the choice of their selection between the gratis line and the toll line. Therefore, we divide the system into two subsystems for analysis, an M/M/1 queueing system for the gratis station and an M/M/1/K queueing system for the toll station, then we calculate the stationary probability of those subsystems under stability conditions, respectively.

We developed a search algorithm based on the Newton's method to calculate the balanced

DOI:10.6814/NCCU201901241

performance measures. We find that the real-time queue length information help people to make the queue length always shorter when the two line service rates are no different. Moreover, we discover the dynamic behavior on queue length changes that are similar between the two information scenarios when the system parameter changes accordingly, such as the service rates, the toll price, and the buffer size.

We consider the toll price from the perspectives of the society and the manager. From the perspective of the society, the objective is to make the expected social waiting cost for the customers minimized. On another aspect, from the perspective of the manager, since the toll price is the source of income, the objective is to maximize the profit.

We established the optimization models with the two information scenarios for managerial perspectives. It also discusses the risk that the customer waits longer than the expected waiting time in the case of an optimal strategy. We show that real-time information can reduce the expected social waiting cost and enhance the social welfare. Moreover, we indicate that two-tier systems present a lower social waiting cost when customers have significant differences in the waiting time costs.

On the other hand, we show that the optimal decision of the toll station in the perspective of the manager is not only affected by the service rate of the gratis station, but also the difference of the waiting time costs of customers. When the gap between customers' time values is too large, the optimal decisions of the manager often leads a very high price, serving only those small groups of customers with high time values. It causes the expected social waiting cost relatively high. Therefore, in this case, the government's control for the toll price is usually necessary.

Finally, by providing real-time information in a two tier service system, the expected social waiting cost per customer is reduced by more than 60%. The improvement for providing real-time information is amazing, but this does not mean that the queue lengths will always be shorter in real-time information scenario. By adjusting the ratio of μ_1 and μ_2, we find that when the gratis station has a higher service rate, the queue length of the gratis station in real-time information scenario is shorter than that in no real-time information scenario. This is because customers will overestimate the service efficiency of the gratis station without real-time information. When the toll station has a higher service rate, the similar situation that customers overestimate the service efficiency of the toll station will also occur. Therefore, we find an interval where the queue lengths of the gratis and toll station are improved. Furthermore, in this

DOI:10.6814/NCCU201901241

interval, we suspect that the expected social waiting cost in the real-time information seems to be a convex upward function of μ_1/μ_2 where there is a global minimum for μ_1/μ_2. That is to say, when the government designs their two-tier healthcare system, if μ_1/μ_2 is a decision variable, there is a best division so that the expected social waiting cost is minimized.

There are some meaningful extensions to present research. First, we suppose that customers always choose the queue wisely based on the information, but this ignores the risk attitude. In our case, customers are assumed risk-neutral. It would be interesting to discuss how different levels of risk appetite and risk averse customers can invest in a pay station to save time. Second, we consider only one service with a fixed price. If there are limited resources, when we explore more services such as service quality, price and different service providers, this would be an interesting competition issue of between the gratis and toll stations. Third, we only made two perspectives of optimization. Whether there is a trade-off between different objectives will also be an important issue. This affects how the government should balance social costs and private profits when intervening in private profits. Furthermore, the government should adopt which strategies such as subsidies or restrictions to ensure that this balance will be most effective.

DOI:10.6814/NCCU201901241

Bibliography

[1] M. Armony and C. Maglaras. Contact centers with a call-back option and real-time delay information. *Operations Research*, 52(4):527–545, 2004.

[2] M. Armony and C. Maglaras. On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research*, 52(2):271–292, 2004.

[3] S. Asmussen. Random walks. *Applied Probability and Queues*, 51:220–243, 2003.

[4] D. Bini and B. Meini. On the solution of a nonlinear matrix equation arising in queueing problems. *Matrix Analysis and Applications*, 17(4):906–926, 1996.

[5] H. Chen, Qian, Q., and A. Zhang. Would allowing privately funded health care reduce public waiting time? theory and empirical evidence from canadian joint replacement surgery data. *Production and Operations Management*, 24(4):605–618, 2015.

[6] Wikipedia contributors. Exponential distribution. https://en.wikipedia.org/wiki/Exponential_distribution, 2002. Online; accessed 1-October-2019.

[7] Wikipedia contributors. Newton's method. https://en.wikipedia.org/wiki/Newton%27s_method, 2002. Online; accessed 4-September-2019.

[8] Wikipedia contributors. Pareto distribution. https://en.wikipedia.org/wiki/Pareto_distribution, 2002. Online; accessed 25-September-2019.

[9] Wikipedia contributors. Uniform distribution (continuous). https://en.wikipedia.org/wiki/Uniform_distribution_(continuous), 2005. Online; accessed 27-July-2019.

DOI:10.6814/NCCU201901241

[10] N. M. Edelson and D. K. Hilderbrand. Congestion tolls for poisson queuing processes. *Journal of the Econometric Society*, 43(1):81–92, 1975.

[11] S. H. Fredrick and J. L. Gerald. *Introduction to Operation Research.* McGraw-Hill Education, 1995.

[12] S. Gavirneni and V. G. Kulkarni. Self-selecting priority queues with burr distributed waiting costs. *Production and Operations Management*, 25(6):979–992, 2016.

[13] L Green and S. Savin. Reducing delays for medical appointments: a queueing approach. *Operations Research*, 56:1526–1538, 2008.

[14] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory, 3rd ed*. John Wiley & Sons, 1998.

[15] P. Guo and Z. G. Zhang. Strategic queueing behavior and its impact on system performance in service systems with the congestion-based staffing policy. *Manufacturing and Service Operations Management*, 15(1):118–131, 2015.

[16] P. Guo and P. Zipkin. Analysis and comparison of queues with different levels of delay information. *Management Science*, 53(6):962–970, 2007.

[17] R. Hassin and M. Haviv. *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems.*, volume 59. Springer Science and Business Media, 2003.

[18] Z. Hua, W. Chen, and Z. G. Zhang. Competition and coordination in two-tier public service systems under government fiscal policy. *Production and Operations Management*, 25(8): 1430–1448, 2016.

[19] H. Jiang, Pang Z., and S. Savin. Performance-based contracts for outpatient medical services. *Manufacturing and Service Operations Management*, 14(4):654–669, 2012.

[20] M. Johar and E. Savage. Do private patients have shorter waiting times for elective surgery? evidence from new south wales public hospitals. *Economic Papers: A Journal of Applied Economics and Policy*, 29(2):128–142, 2010.

[21] S. Kapodistria and Z. Palmowski. Matrix geometric approach for random walks: Stability condition and equilibrium distribution. *Stochastic Models*, 33(4):572–597, 2017.

DOI:10.6814/NCCU201901241

[22] L. Kleinrock. Optimum bribing for queue position. *Operations Research*, 15(2):304–318, 1967.

[23] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* SIAM, 1999.

[24] H. P. Luh and P. C. Song. Matrix analytic solutions for m/m/s retrial queues with impatient customers. *International Conference on Queueing Theory and Network Applications*, 11688:16–33, 2019.

[25] MathWorks MATLAB. Matlab r2018b. *The MathWorks: Natick, MA, USA*, 2018.

[26] H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations Research*, 38(5):870–883, 1990.

[27] R. T. Meulen and F. Jotterand. Individual responsibility and solidarity in european health care: further down the road to two-tier system of health care. *Journal of Medicine and Philosophy*, 33(3):191–197, 2008.

[28] P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37(1):15–23, 1969.

[29] H. Nazerzadeh and R. S. Randhawa. Asymptotic optimality of two service grades for customer differentiation in queueing systems. *working paper, University of Southern California*, 2014.

[30] M. F. Neuts. *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach.* Courier Corporation, 1994.

[31] E.L. Plambeck. Optimal leadtime differentiation via diffusion approximations. *Operations Research*, 52(2):213–228, 2004.

[32] Q. Qian, Guo P., and Lindsey R. Comparison of subsidy schemes for reducing waiting times in healthcare systems. *Production and Operations Management*, 26(11):2033–2049, 2017.

[33] R. Schroeter. The costs of concealing the customer queue. *working paper, Bureau of Business and Economic Research, Arizona State*, 1982.

DOI:10.6814/NCCU201901241

[34] S. Stidham Jr. *Optimal Design of Queueing Systems*. Chapman and Hall, 2009.

[35] Z. Ugray, L. Lasdon, J. Plummer, F. Glover, J. Kelly, and R. Martí. Scatter search and local nlp solvers: A multistart framework for global optimization. *INFORMS Journal on Computing*, 19(3):328–340, 2007.

[36] G. Wan and Q. Wang. Two-tier healthcare service systems and cost of waiting for patients. *Applied Stochastic Models in Business and Industry*, 33(2):167–183, 2017.

[37] G. Z. Zhang and H. P. Luh. Information effects on performance of two-tier service systems with strategic customers. *working paper*, 2013.

DOI:10.6814/NCCU201901241

Appendix A

The Proofs and Background Informations

In this appendix, we cite some of the proofs mentioned in this article and some background information. In addition, the definition of the symbol is the same as the previous content.

A.1 Matrix Geometric Method

This is extracted from [21]. Consider a QBD with finite phases and a infinitesimal generator

$$\overline{\boldsymbol{Q}} = \begin{bmatrix} \overline{\boldsymbol{B}_0} & \boldsymbol{C} & & & \\ \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{C} & & \\ & \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{C} & \\ & & \ddots & \ddots & \ddots \end{bmatrix}, \tag{A.1.1}$$

where the square matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ are non-negative, and the square matrices $\overline{\boldsymbol{B}_0} = \boldsymbol{B} - \boldsymbol{A}$ and $\boldsymbol{B}$ have strictly negative diagonals and non-negative other elements. $\boldsymbol{C}$, $\boldsymbol{B}$, $\boldsymbol{A}$ are the transition rate matrices turn the states to a higher level, within the same level and to a lower level, respectively. The row sums of $\overline{\boldsymbol{Q}}$ are equal to zero.

The stability condition of such a QBD can be obtained by the drift condition, Theorem 1.7.1 in [30],

$$\boldsymbol{x}\boldsymbol{C}\widehat{\boldsymbol{e}} < \boldsymbol{x}\boldsymbol{A}\widehat{\boldsymbol{e}}$$

where $\boldsymbol{x}$ is a row vector that is the unique solution to $\boldsymbol{x}(\boldsymbol{C} + \boldsymbol{B} + \boldsymbol{A}) = 0$ and $\widehat{\boldsymbol{e}}$ is a column

DOI:10.6814/NCCU201901241

vector of ones.

To compute the stationary distribution $\boldsymbol{\pi}$ writing $\boldsymbol{\pi}\overline{\boldsymbol{Q}} = \mathbf{0}$, the equations are considered for sub-vectors $\boldsymbol{\pi}_i$.

$$\begin{aligned}
\boldsymbol{\pi}_0\overline{\boldsymbol{B}_0} + \boldsymbol{\pi}_1\boldsymbol{A} &= \mathbf{0} \\
\boldsymbol{\pi}_0\boldsymbol{C} + \boldsymbol{\pi}_1\boldsymbol{B} + \boldsymbol{\pi}_2\boldsymbol{A} &= \mathbf{0} \\
\boldsymbol{\pi}_1\boldsymbol{C} + \boldsymbol{\pi}_2\boldsymbol{B} + \boldsymbol{\pi}_3\boldsymbol{A} &= \mathbf{0} \\
&\vdots \\
\boldsymbol{\pi}_{i-1}\boldsymbol{C} + \boldsymbol{\pi}_i\boldsymbol{B} + \boldsymbol{\pi}_{i+1}\boldsymbol{A} &= \mathbf{0} \\
&\vdots
\end{aligned}$$

Observe the relationship

$$\boldsymbol{\pi}_{i+1} = \boldsymbol{\pi}_i\boldsymbol{R} \Rightarrow \boldsymbol{\pi}_i = \boldsymbol{\pi}_1\boldsymbol{R}^{i-1}$$

where $\boldsymbol{R}$ is the rate matrix and it can be computed using cyclic reduction [4] or (appendix A.2). Then, we write

$$\begin{bmatrix} \boldsymbol{\pi}_0 & \boldsymbol{\pi}_1 \end{bmatrix} \begin{bmatrix} \overline{\boldsymbol{B}_0} & \boldsymbol{C} \\ \boldsymbol{A} & \boldsymbol{B} + \boldsymbol{R}\boldsymbol{A} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

We can solve these equations to find $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$. Therefore, we iteratively find all the $\boldsymbol{\pi}_i$, $i \in \mathbb{N}$.

A.2 Algorithm for Computing the Rate Matrix

With the $\overline{\boldsymbol{Q}}$ matrices defined in (A.1.1), we need to solve the following quadratic matrix equation for the rate matrix of the QBD process for the queue:

$$\boldsymbol{C} + \boldsymbol{R}\boldsymbol{B} + \boldsymbol{R}^2\boldsymbol{A} = \mathbf{0},$$

where $\boldsymbol{C}, \boldsymbol{B}, \boldsymbol{A}$ are same with Appendix A.1. Note that the stationary probability vector for the modulating Markov process is the stationary distribution of the queue length (including the customer in service). We can use either the simple functional iteration to compute $\boldsymbol{R}$ as follows

$$\boldsymbol{R}_{n+1} = -(\boldsymbol{C} + \boldsymbol{R}_n^2\boldsymbol{A})\boldsymbol{B}^{-1},$$

DOI:10.6814/NCCU201901241

starting with $\boldsymbol{R}_0 = \mathbf{0}$ or other iteration algorithms reported in Latouche and Ramaswami [23] such as quadratically convergent algorithms.

A.3 Newton's Method

These descriptions are excerpted from wikipedia [7]. Newton's method starts with an initial guess which is close to the root of the function. Then, the derivative of the function is calculated at the initial guess by using calculus, which forms a tangent through the point of the initial guess. The x-intercept of the tangent is calculated by elementary algebra. This x-intercept generally better approximates the root of the function, and the method can be iterated.

Suppose $Eq : (a, b) \to \mathbb{R}$ is a differentiable function defined on the interval (a, b) where $a, b \in \mathbb{R}$, and we have a current approximation x_n. Then, the equation of the tangent line to the curve $y = Eq\,(x)$ at $x = x_n$ is

$$y = Eq'(x_n)(x - x_n) + Eq(x_n),$$

where Eq' denotes the derivative function of Eq. The x-intercept of this line is taken as the next approximation, x_{n+1}. That is, x_{n+1} satisfy $0 = Eq'(x_n)(x_{n+1} - x_n) + Eq(x_n)$. Solving for x_{n+1}, there is

$$x_{n+1} = x_n - \frac{Eq(x_n)}{Eq'(x_n)}.$$

We start this process with an arbitrary initial value x_0. If the initial guess is close enough to the root of the function and its derivative is nonzero, the method will usually converge to the point where the function value is zero.

A.4 The Distribution Function of the Waiting Time

This is extracted from chapter 17 of Fredrick and Gerald [11]).

Assuming that the queueing system described as in Appendix A.1 reaches the stability condition, we now can derive the probability distribution of the waiting time in the system (so including service time) for a random arrival when the queue discipline is first-come-first-served. Same as the previous content, let the system states be written as (n, m), where in the state

DOI:10.6814/NCCU201901241

description the first entry $n = 0, 1, \cdots$ and the second entry $m = 0, 1, 2, \cdots, K$. If the arrival finds k customers already in the subsystem (the state $(k, \bullet)$), then the arrival will have to wait through $k + 1$ exponential service times, including his/her own. Therefore, let T_1, T_2, $\cdots$ be independent service time random variables having an exponential distribution with parameter μ and let

$$S_{k+1} = T_1 + T_2 + \cdots + T_{k+1} \text{ for } k \in \mathbb{N},$$

so that S_{k+1} represents the conditional waiting time given k customers already in the subsystem. The S_{k+1} is known to have an Erlang distribution. The probability that the random arrival will find k customers in the subsystem is $\boldsymbol{\pi}_k$, and given $\overline{n}$ is the first positive number i satisfied $\boldsymbol{\pi}_{i+1} = \boldsymbol{\pi}_i \boldsymbol{R}$, it follows that

$$\begin{aligned} P(W > t) &= \sum_{n=0}^{\infty} \boldsymbol{\pi}_n P(S_{n+1} > t)\widehat{\boldsymbol{e}} \\ &= \sum_{n=0}^{\overline{n}} \boldsymbol{\pi}_n P(S_{n+1} > t)\widehat{\boldsymbol{e}} + \sum_{n=\overline{n}+1}^{\infty} \boldsymbol{\pi}_{\overline{n}} \boldsymbol{R}^{n-\overline{n}} P(S_{n+1} > t)\widehat{\boldsymbol{e}} \\ &= \sum_{n=0}^{\overline{n}} \boldsymbol{\pi}_n T_n(\mu t) e^{-\mu t}\widehat{\boldsymbol{e}} + \boldsymbol{\pi}_{n_0} \left(T_{\overline{n}}(\mu t)\boldsymbol{R}(\boldsymbol{I} - \boldsymbol{R})^{-1} + R_{\overline{n}}(\mu t \boldsymbol{R})\boldsymbol{R}^{-\overline{n}}(\boldsymbol{I} - \boldsymbol{R})^{-1} \right) e^{-\mu t}\widehat{\boldsymbol{e}}, \end{aligned}$$

where $\boldsymbol{R}$ is the rate matrix, $T_i(x) = \sum_{j=0}^{i} x^j/j!$, $R_i(x) = e^x - \sum_{j=0}^{i} x^j/j!$, $P(S_{i+1} > t) = \sum_{j=0}^{i} (\mu t)^j e^{-\mu t}/j!$ and $i \in \mathbb{N}$.

A.5 The Stability Condition for Real-time information

It is proved by Proposition 1 of Zhang and Luh [37].We use a case of $K = 3$ to prove this proposition. Letting $\overline{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{B} + \boldsymbol{C}$, we have

$$\overline{\boldsymbol{A}} = \begin{bmatrix} -\Lambda & \Lambda & & \\ \mu_2 & -(\Lambda + \mu_2) & \Lambda & \\ & \mu_2 & -(\Lambda + \mu_2) & \Lambda \\ & & \mu_2 & -\mu_2 \end{bmatrix}.$$

DOI:10.6814/NCCU201901241

Denote the stationary vector for $\overline{\boldsymbol{A}}$ by $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$. Solving $\boldsymbol{\alpha}\overline{\boldsymbol{A}} = \mathbf{0}$, we obtain $\alpha_0 = 1/\sum_{i=0}^{3}(\Lambda/\mu_2)^i, \alpha_1 = (\Lambda/\mu_2)/\sum_{i=0}^{3}(\Lambda/\mu_2)^i, \alpha_2 = (\Lambda/\mu_2)^2/\sum_{i=0}^{3}(\Lambda/\mu_2)^i$, and $\alpha_3 = (\Lambda/\mu_2)^3/\sum_{i=0}^{3}(\Lambda/\mu_2)^i$. Based on the drift stability condition of $\boldsymbol{\alpha}\boldsymbol{C}\widehat{\boldsymbol{e}} < \boldsymbol{\alpha}\boldsymbol{A}\widehat{\boldsymbol{e}}$, we have

$$\mu_1 > \alpha_3\Lambda = \frac{(\Lambda/\mu_2)^3\Lambda}{\sum_{i=0}^{3}(\Lambda/\mu_2)^i} = \frac{\left(1-\frac{\Lambda}{\mu_2}\right)\left(\frac{\Lambda}{\mu_2}\right)^3\Lambda}{1-\left(\frac{\Lambda}{\mu_2}\right)^4}.$$

For a general case with buffer size K, we get

$$\mu_1 > \alpha_K\Lambda = \left(1-\frac{\Lambda}{\mu_2}\right)\left(\frac{\Lambda}{\mu_2}\right)^K\Lambda\left(1-\left(\frac{\Lambda}{\mu_2}\right)^{K+1}\right)^{-1}.$$

A.6 The Convex of the Expected Social Waiting Cost for no Real-time

It is proved by Theorem 1 of Wan and Wang [36]. By differentiating $S_1(\overline{\theta})$ with respect to $\overline{\theta}$ twice, we obtain

$$S_1''(\overline{\theta}) = E[W_1]''G(\overline{\theta}) + 2[E[W_1]' - E[W_2]']G'(\overline{\theta}) + (E[W_1] - E[W_2])G''(\overline{\theta}) + E[W_2]''(\xi - G(\overline{\theta})). \tag{A.6.1}$$

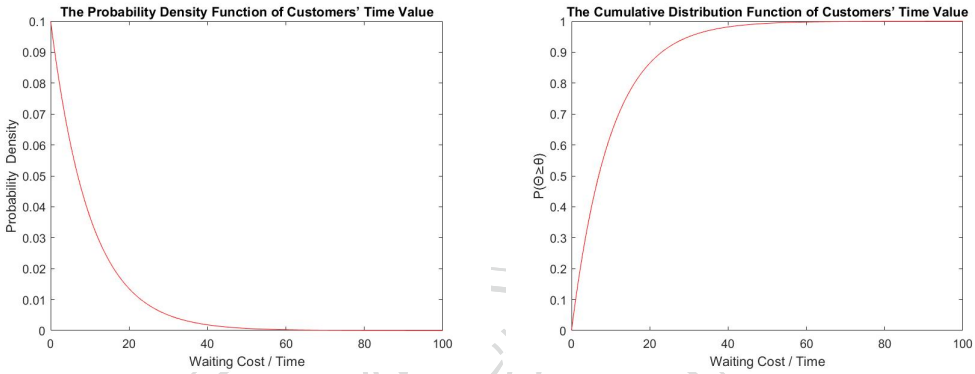
First, $E[W_2]$ is convex upward, because it is based on an M/M/1/K queue model. Same for $E[W_1]$, it is convex upward by λ_1^{eff}, because it is based on an M/M/1 queue model. Second, $E[W_1] - E[W_2]$ is increasing and positive. The queue length of toll station is bounded and it charge a premium, those cause $E[W_1] \geq E[W_2]$. Third, $G(\overline{\theta})$ is the integral function of the expected value of a probability distribution, so it is positive and increasing. Using integration by part $G''(\overline{\theta}) = 1/f(\overline{\theta}) > 0$. Finally, for any given $\overline{\theta}$, $\xi \geq G(\overline{\theta}) \geq 0$. Applying these conditions to (A.6.1), we obtain $S_1''(\overline{\theta}) \geq 0$.

A.7 The Distributions

These descriptions of selected probability distributions are excerpted from wikipedia [9] [6] [8]. In this thesis, we consider that a random variable Θ which is the waiting cost per unit time for a customer has the probability density function $f(\theta)$. This function shows the probabilities

DOI:10.6814/NCCU201901241

of customers' time value when a customer enters the system. For example, in the probability density function shown in Figure A.1, when a customer enters the system, the probability which his/her time value less than twenty per unit time is about 0.86.



(a) The probability density function (b) The cumulative distribution function

Figure A.1: A distribution of customers' time value

A.7.1 The Uniform Distribution

The uniform distribution is a symmetric probability distribution. Each interval with the same length has the same probability on the distribution's support. The distribution is defined by the two parameters, a and b, which are its minimum and maximum values. If a random variable X has uniform distribution, it can be written as $X \sim U(a, b)$. The probability density function of an uniform distribution is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b. \end{cases}$$

The cumulative distribution function is given by

$$F(x) = \begin{cases} \frac{x-a}{b-a} & \text{for } a \leq x \leq b, \\ 1 & \text{for } x > b. \end{cases}$$

The expected value of an uniform distributed random variable X is given by $E[X] = \frac{1}{2}(a+b)$. The variance of X is given by $Var[X] = \frac{1}{12}(b-a)^2$.

DOI:10.6814/NCCU201901241

A.7.2 The Exponential Distribution

The exponential distribution is a continuous probability distribution. The exponential distribution can be used to represent the time interval at which an independent random event occurs. The distribution is defined by a non-negative parameter λ, which is the number of times the event occurs per unit of time. The distribution is denoted by $Exp(\lambda)$. The probability density function of an exponential distribution is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

The cumulative distribution function of an exponential random variable is

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

The expected value of an exponential distribution is $\frac{1}{\lambda}$. The variance is $\frac{1}{\lambda^2}$.

A.7.3 The Pareto Distribution

The Pareto distribution is a power-law probability distribution that is used in description of social, geophysical, and many other types of observable phenomena. The Pareto distribution is characterized by a scale parameter xm and a shape parameter α where xm is the minimum possible value of random variable X, and α is a positive parameter. The distribution is denoted by $Pa(xm, \alpha)$. The probability density function of Pareto distribution is

$$f(x) = \begin{cases} \frac{\alpha x_{\mathrm{m}}^{\alpha}}{x^{\alpha+1}} & x \geq x_{\mathrm{m}}, \\ 0 & x < x_{\mathrm{m}}. \end{cases}$$

DOI:10.6814/NCCU201901241

The cumulative distribution function of a random variable following a Pareto distribution is

$$f(x) = \begin{cases} 1 - \left(\frac{x_{\mathrm{m}}}{x}\right)^{\alpha} & x \geq x_{\mathrm{m}}, \\ 0 & x < x_{\mathrm{m}}. \end{cases}$$

The expected value of a random variable following a Pareto distribution is

$$E[X] = \begin{cases} \infty & \alpha \leq 1, \\ \frac{\alpha x_{\mathrm{m}}}{\alpha - 1} & \alpha > 1. \end{cases}$$

The variance of a random variable following a Pareto distribution is

$$Var[X] = \begin{cases} \infty & \alpha \in (1, 2], \\ \left(\frac{x_{\mathrm{m}}}{\alpha - 1}\right)^2 \frac{\alpha}{\alpha - 2} & \alpha > 2. \end{cases}$$

Note that if $\alpha \leq 1$, the variance does not exist.

DOI:10.6814/NCCU201901241

Appendix B

MATLAB Codes

In this study, we use Matlab as a computing tool [25].

B.1 Program for Distributions of Customers' Time Values

We consider the variable $type \in \{1,\ 2,\ 3\}$ denotes the type of the distributions where represents the uniform distribution when $type = 1$, represents the exponential distribution when $type = 2$, and represents the Pareto distribution when $type = 3$.

B.1.1 Parameters of Distributions

Output a vector of parameters of uniform distribution when the variable $type$ is 1. Output the parameter of exponential distribution when the variable is 2. Output the parameters of Pareto distribution when the variable is 3.

```
function Z=parameter_distribution_(type)
if type==1
    min=0;
    max=20;
    Z=[min max];
end
if type==2
    kappa=0.1;
```


DOI:10.6814/NCCU201901241

```
    Z=[kappa 0];
end
if type==3
    alpha=10/9;
    xm=1;
    Z=[alpha xm];
end
end
```

B.1.2 The Cumulative Distribution Function of Θ

Let the variable $type \in \{1, 2, 3\}$ represents the type of the distributions and θ is the variable of the distribution. Output the value of the cumulative distribution function.

```
function Z=F(type,theta)
parameter_distribution=parameter_distribution_(type);
if type==1
    min=parameter_distribution(1);
    max=parameter_distribution(2);
    Z=1;
    if theta<=max
        Z=(theta-min)/(max-min);
        if theta<min
            Z=0;
        end
    end
end
if type==2
    kappa=parameter_distribution(1);
    if theta>0
       Z=1-exp(-kappa*theta);
    else
        Z=0;
```


DOI:10.6814/NCCU201901241

```
    end
end
if type==3
    alpha=parameter_distribution(1);
    xm=parameter_distribution(2);
    if theta>=xm
        Z=1-(xm/theta)^alpha;
    else
        Z=0;
    end
end
end
```

B.1.3 The Function of Expected Value of $f(\theta)$

Let $type \in \{1, 2, 3\}$ represents the type of the distributions and θ is the variable of the distribution. Output the value of expected value function of the probability density function.

```
function Z=G(type,theta)
parameter_distribution=parameter_distribution_(type);
if type==1
    min=parameter_distribution(1);
    max=parameter_distribution(2);
    Z=(max^2-min^2)/(max-min)/2;
    if theta<=max
        Z=(theta^2-min^2)/(max-min)/2;
        if theta<min
            Z=0;
        end
    end
end
if type==2
    kappa=parameter_distribution(1);
```


DOI:10.6814/NCCU201901241

```
    if theta<0
        Z=0;
    else
        Z=-exp(-kappa*theta)*theta-kappa*exp(-kappa*theta) ...
        /(kappa^2)+(1/kappa);
    end
end
if type==3
    alpha=parameter_distribution(1);
    xm=parameter_distribution(2);
    if theta>=xm
        if alpha==1
            Z=xm*log(abs(theta))-xm*log(abs(xm));
        else
            Z=alpha*xm^alpha*theta^(1-alpha)/(1-alpha) ...
            -(alpha*xm^alpha*xm^(1-alpha)/(1-alpha));
        end
    else
        Z=0;
    end
end
end
```

B.1.4 Inverse Function of the Distribution Cumulative Function of Θ

Let $type \in \{1, 2, 3\}$ represents the type of the distributions and pro is the value of the cumulative distribution function. Output the value of corresponding θ.

```
function Z=F_inverse(type,pro)
if pro>=1
    pro=0.99;
else
    if pro<0
```


DOI:10.6814/NCCU201901241

```
        pro=0;
    end
end
Z=fzero(@(x) F(type,x)-pro,0.5);
end
```

B.2 Taylor Series of Exponential Function

B.2.1 Taylor Expansion of Exponential Function

Let n is the number of terms. Output the value of Taylor expansion of the exponential function of the first n terms.

```
function Z=Taylor_T(n,x)
a=0;
for i=0:n
    a=a+x^i/prod(1:i);
end
Z=a;
end
```

B.2.2 Error of Taylor Expansion of Exponential Function

Let n is the number of terms. Output the value of error of Taylor expansion of the exponential function of the first n terms.

```
function Z=Taylor_R(n,x)
Z=expm(x)-Taylor_T(n,x);
end
```

B.3 Program for a No Real-time Information Scenario

We consider the variable $type \in \{1,\ 2,\ 3\}$ represents the type of the distributions where represents the uniform distribution when $type = 1$, represents the exponential distribution when

DOI:10.6814/NCCU201901241

$type = 2$, and represents the Pareto distribution when $type = 3$. In addition, p is the toll price, K is the buffer size of toll station, Λ is total arrival rate, and μ is a vector consisting of service rates μ_1 and μ_2.

B.3.1 The Main Program

The main program inputs some basic parameters and calls the subprogram to calculate various performances, then it optimizes the expected social waiting costs by "fmincon" and the expected profit of manager by "GlobalSearch", respectively.

```
clear;

%Basic parameter
type=1;
p=10;
K=10;
Lambda=1;
mu_1=5/6;mu_2=5/6;
mu=[mu_1 mu_2];
epsilon=10^(-5);
delta=0.01;
%The constraints of optimization
lb=[0];
ub=[];
A=[];
b=[];
Aeq=[];
beq=[];
%The initial theta
theta0=(F_inverse(type,mu_1/Lambda) ...
+F_inverse(type,Lambda-mu_2/Lambda))/2;

theta=
```


DOI:10.6814/NCCU201901241

```
Newton_formula_theta(type,p,K,Lambda,mu,epsilon,delta,theta0);

lambda=lambda_(type,Lambda,theta);
lambda_eff=formula_eff(type,K,Lambda,mu,theta);
formula_W=formula_(type,K,Lambda,mu,theta);
formula_S=S(type,formula_W,theta);
formula_O=O(lambda_eff,formula_W,theta);

min_theta_society=fmincon(@(x) ...
S(type,formula_(type,K,Lambda,mu,x),x), ...
theta0,A,b,Aeq,beq,lb,ub,@(x) mycon(type,K,Lambda,mu,0.5,x))

fun=@(x) -O(formula_eff(type,K,Lambda,mu,x), ...
formula_(type,K,Lambda,mu,x),x),theta0,A,b,Aeq,beq,lb,ub, ...
@(x) mycon(type,K,Lambda,mu,0.5,x))
rng default
opts=optimoptions(@fmincon,'Algorithm','sqp');
problem=createOptimProblem('fmincon','objective',fun,'x0',theta0, ...
'lb',lb,'ub',ub,'nonlcon',@(x) mycon(type,K,Lambda,mu,0.5,x) ...
,'options',opts);
gs=GlobalSearch;
[max_theta_manager,f]=run(gs,problem);
```

B.3.2 The Planned Arrival Rate

Output a vector consisting of the planned arrival rates λ_1 and λ_2, where $type \in \{1,\ 2,\ 3\}$ represents the type of the distributions and θ is the variable of the distribution.

```
function Z=lambda_(type,Lambda,theta)
lambda_1=Lambda*F(type,theta);
lambda_2=Lambda*(1-F(type,theta));
Z=[lambda_1 lambda_2];
end
```


DOI:10.6814/NCCU201901241

B.3.3 The Effective Arrival Rate

Let $type \in \{1, 2, 3\}$ represents the type of the distributions and θ is the variable of the distribution. Output a vector consisting of the effective arrival rates λ_1^{eff} and λ_2^{eff}, where K, Λ, and μ are inputs.

```
function Z=formula_eff(type,K,Lambda,mu,theta)
lambda=lambda_(type,Lambda,theta);
xi=lambda(2)/mu(2);
if xi==1
    piK=1/(K+1);
else
    piK=(1-xi)*xi^K/(1-xi^(K+1));
end
lambda_eff_1=lambda(1)+piK*lambda(2);
lambda_eff_2=(1-piK)*lambda(2);
Z=[lambda_eff_1 lambda_eff_2];
end
```

B.3.4 The Expected Waiting Time

Let $type \in \{1, 2, 3\}$ represents the type of the distributions and θ is the variable of the distribution. Output a vector consisting of the expected waiting time $E[W_1]$ and $E[W_2]$, where K, Λ, and μ are inputs.

```
function Z=formula_(type,K,Lambda,mu,theta)
lambda=lambda_(type,Lambda,theta);
xi=lambda(2)/mu(2);
if xi==1
    piK=1/(K+1);
else
    piK=(1-xi)*xi^K/(1-xi^(K+1));
end
lambda_eff_1=lambda(1)+piK*lambda(2);
```


DOI:10.6814/NCCU201901241

```
lambda_eff_2=(1-piK)*lambda(2);
lambda_eff=[lambda_eff_1 lambda_eff_2];

if xi==1
    W2=K/2/mu(2);
else
    W2=(1/(1-xi)-(1+K*xi^(K+1))/(1-xi^(K+1)))/lambda_eff(2);
    if lambda_eff(2)==0
        W2=0;
    end
end
W1=1/(mu(1)-lambda_eff(1));
Z=[W1 W2];
end
```

B.3.5 The Balanced Function of θ

Output a function value where the function is defined by $Eq(\theta) := p/([E[W_1(\theta)] - E[W_2(\theta)]]) - \theta$.

```
function Z=Eq_(p,theta,W)
 Z=p/(W(1)-W(2))-theta;
end
```

B.3.6 The Search Algorithm of Computing $\overline{\theta}$

Let $type \in \{1, 2, 3\}$ represents the type of the distributions and θ is the variable of the distribution. Output the balanced waiting time cost rate $\overline{\theta}$, where p, K, Λ, μ, a positive number ϵ, and a positive number δ are inputs.

```
function ...
Z=Newton_formula_theta(type,p,K,Lambda,mu,epsilon,delta,theta)

W=formula_(type,K,Lambda,mu,theta);
```


DOI:10.6814/NCCU201901241

```
for i=1:100
    if abs(Eq_(p,theta(i),W))<=epsilon
        break;
    else
        if W(1)-W(2)<0
            i=i-1;
            delta=delta/2;
        end
        W=formula_(type,K,Lambda,mu,theta(i));
        Eq=Eq_(p,theta(i),W);

        W_epsilon=formula_(type,K,Lambda,mu,theta(i)+epsilon);
        Eq_epsilon=Eq_(p,theta(i)+epsilon,W_epsilon);

        theta=[theta theta(i)-(Eq*epsilon*delta/(Eq_epsilon-Eq))];
    end
end
lambda_eff=formula_eff(type,K,Lambda,mu,theta(i));
if lambda_eff(1)<=mu(1)
    Z=0;
else
    Z=theta(i);
end
end
```

B.3.7 The Complementary Cumulative Distribution Function of the Waiting Time

Let $type \in \{1, 2, 3\}$ represents the type of the distributions and θ is the variable of the distribution. Output a vector consisting of the function values $P(W_1 > t)$ and $P(W_2 > t)$, which defined as (2.2.9). Where K, Λ, μ, and a positive number t are inputs.

```
function ...
```


DOI:10.6814/NCCU201901241

```
Z=formula_probability_waiting_time(type,Lambda,K,mu,t,theta)

mu_1=mu(1);mu_2=mu(2);
pw=0;
lambda=lambda_(type,Lambda,theta);
xi_2=lambda(2)/mu(2);

for i=0:K
    if xi_2==1
        pi=1/(K+1);
    else
        pi=(1-xi_2)*xi_2^i/(1-xi_2^(K+1));
    end
    pw=pw+pi*Taylor_T(i,mu_2*t);
end
lambda_eff_1=lambda(1)+pi*lambda(2);
xi_1=lambda_eff_1/mu(1);

probability_waiting_time_1=exp((xi_1-1)*mu_1*t);
probability_waiting_time_2=pw*exp(-mu_2*t);
Z=[probability_waiting_time_1 probability_waiting_time_2];
end
```

B.3.8 The Expected Social Waiting Cost

Let $type \in \{1, 2, 3\}$ represents the type of the distributions and θ is the variable of the distribution. Output the expected social waiting cost $S_1(\overline{\theta})$, where W is the output of Appendix B.3.4.

```
function Z=S(type,W,theta)
parameter_distribution=parameter_distribution_(type);

if type==1
```


DOI:10.6814/NCCU201901241

```
    min=parameter_distribution(1);
    max=parameter_distribution(2);
    xi=(max^2-min^2)/(max-min)/2;
end
if type==2
    kappa=parameter_distribution(1);
    xi=1/kappa;
end
if type==3
    alpha=parameter_distribution(1);
    xm=parameter_distribution(2);
    xi=alpha*xm/(alpha-1);
end

EF=[G(type,theta) xi-G(type,theta)];
Z=EF*W';
end
```

B.3.9 The Expected Profit of Manager

Output the value of the expected revenue per unit time of the toll system $O_1(\bar{\theta})$, which defined as (3.2.1).

```
function Z=O(lambda_eff,W,theta)
p=theta*(W(1)-W(2));
Z=lambda_eff(2)*p;
end
```

B.3.10 The Constraints of Optimization

Let $type \in \{1, 2, 3\}$ represents the type of the distributions and θ is the variable of the distribution. Output a vector consisting of the values ciq and ceq, which is used to nonlinear inequality $ciq < 0$ and nonlinear equation $ceq = 0$. Where K, Λ, μ, and a positive number c are

DOI:10.6814/NCCU201901241

inputs. This is based on the constraints of the optimization problem to choose whether to use the annotation part.

```
function [ciq,ceq]=mycon(type,K,Lambda,mu,c,theta)
lambda=lambda_(type,Lambda,theta);
xi=lambda(2)/mu(2);
if xi==1
    piK=1/(K+1);
else
    piK=(1-xi)*xi^K/(1-xi^(K+1));
end
lambda_eff_1=lambda(1)+piK*lambda(2);

W=formula_(type,K,Lambda,mu,theta);
pwt=formula_probability_waiting_time(type,Lambda,K,mu,W(1),theta);
%pwt=formula_probability_waiting_time(type,Lambda,K,mu,W(2),theta);
ciq=[W(2)-W(1) pwt(1)-c];
%ciq=[W(2)-W(1) pwt(2)-c];
ceq=[];
end
```

B.4 Program for a Real-time Information Scenario

We consider the variable $type \in \{1,\ 2,\ 3\}$ represents the type of the distributions where represents the uniform distribution when $type = 1$, represents the exponential distribution when $type = 2$, and represents the Pareto distribution when $type = 3$. In addition, p is the toll price, K is the buffer size of toll station, Λ is total arrival rate, μ is a vector consisting of service rates μ_1 and μ_2, $n = 0, 1, \cdots$ indicates the queue length of the gratis station, $m = 0, 1, 2, \cdots, K$ indicates the queue length of the toll station, and n_0 is the queue length of the gratis station such that all customers try to enter the gratis station.

DOI:10.6814/NCCU201901241

B.4.1 The Main Program

The main program inputs some basic parameters, tests the stability condition and calls the subprogram to calculate various performances, then it optimizes the expected social waiting costs or the expected profit of manager by ”GlobalSearch”.

```
clear;

%Basic parameter
type=1;
p=10;
K=10;
Lambda=1;
mu_1=5/6;mu_2=5/6;
mu=[mu_1 mu_2];
epsilon=10^(-3);
n0=fix(abs(p)*mu_1/(F_inverse(type,epsilon/Lambda))+K*mu_1/mu_2);
%The constraints of optimization
lb=[0];
ub=[];
A=[];
b=[];
Aeq=[];
beq=[];

sigma=sigma_(K,Lambda,mu);
pi=pi_(type,p,K,Lambda,mu,n0,sigma);
EL=EL_(K,n0,sigma,pi)
S_=S(type,K,n0,p,mu,sigma,pi,epsilon)
O_=O(type,K,Lambda,n0,p,mu,sigma,pi)

fun=@(x) S(type,K,n0,x,mu,sigma_(K,Lambda,mu), ...
pi_(type,x,K,Lambda,mu,n0,sigma_(K,Lambda,mu)),epsilon);
```


DOI:10.6814/NCCU201901241

```
% fun=@(x) -O(type,K,Lambda,n0,x,mu,sigma_(K,Lambda,mu), ...
pi_(type,x,K,Lambda,mu,n0,sigma_(K,Lambda,mu)));

stability_condition=mu(1)-Lambda*(1-Lambda/mu(2)) ...
*((Lambda/mu(2))^K)/(1-(Lambda/mu(2))^(K+1));

rng default
opts=optimoptions(@fmincon,'Algorithm','sqp');
problem=createOptimProblem('fmincon','objective',fun,'x0',50,...
'lb',lb,'ub',ub,'nonlcon', ...
@(x) mycon(type,K,Lambda,n0,x,mu,0.5),'options',opts);
gs=GlobalSearch;
[x,f]=run(gs,problem)
```

B.4.2 The Planned Arrival Rate

Output a vector consisting of the planned arrival rates λ_1 and λ_2 at the system state $(n,\ m)$, where $type \in \{1,\ 2,\ 3\}$ represents the type of the distributions.

```
function Z=lambda_(type,p,K,Lambda,mu,n0,n,m)
mu_1=mu(1);mu_2=mu(2);

if n<=m*(mu_1/mu_2) || m==K
    lambda_1=Lambda;
end
if n<n0 && n>m*(mu_1/mu_2) && m<K
    lambda_1=Lambda*F(type,p/(n/mu_1-m/mu_2));
end
if n>=n0 && m<K
    lambda_1=0;
end

lambda_2=Lambda-lambda_1;
```


DOI:10.6814/NCCU201901241

```
Z=[lambda_1 lambda_2];
end
```

B.4.3 The Transfer Matrix

Output the matrix B_n and C_n, which defined as (2.3.6).

```
function Z=Bn(type,p,K,Lambda,mu,n0,n)
mu_1=mu(1);mu_2=mu(2);
lambda_2_n=[];
for i=0:K-1
    lambda_n=lambda_(type,p,K,Lambda,mu,n0,n,i);
    lambda_2_n=[lambda_2_n lambda_n(2)];
end
    Z=mu_2*([1 zeros(1,K)]')*[1 zeros(1,K)] ...
    -(Lambda+mu_2+mu_1)*diag(ones(K+1,1)) ...
    +mu_2*diag(ones(K,1),-1)+diag(lambda_2_n,1);
end

function Z=Cn(type,p,K,Lambda,mu,n0,n)
lambda_1_n=[];
    for i=0:K-1
        lambda_n=lambda_(type,p,K,Lambda,mu,n0,n,i);
        lambda_1_n=[lambda_1_n lambda_n(1)];
    end
    lambda_1_n=[lambda_1_n Lambda];
Z=diag(lambda_1_n);
end
```

B.4.4 The Eigenvalue of K-Matrix

Output the eigenvalue of K-matrix, which defined as (2.3.7).

```
function Z=sigma_(K,Lambda,mu)
```


DOI:10.6814/NCCU201901241

```
mu_1=mu(1);mu_2=mu(2);
B=mu_2*([1 zeros(1,K)]')*[1 zeros(1,K)] ...
-(Lambda+mu_2+mu_1)*diag(ones(K+1,1)) ...
+mu_2*diag(ones(K,1),-1)+Lambda*diag(ones(K,1),1);
A=mu_1*eye(K+1);

P=A*ones(1,size(A,2))'*[zeros(1,K) 1];
K=-(B+P)/A;
eigK_=eig(K);

for i=1:size(K,2)
    if eigK_(i)>0 && eigK_(i)<1
        sigma=eigK_(i);
    end
end
Z=sigma;
end
```

B.4.5 The Stationary Probability

Let $type \in \{1, 2, 3\}$ represents the type of the distributions. Output a vector of the stationary probability $[\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2]$, where p, K, Λ, μ, n_0 and the eigenvalue of K-matrix σ are inputs.

```
function Z=pi_(type,p,K,Lambda,mu,n0,sigma)
mu_1=mu(1);mu_2=mu(2);
C0=Cn(type,p,K,Lambda,mu,n0,0);
B0=Bn(type,p,K,Lambda,mu,n0,0)+mu_1*eye(K+1);

A=mu_1*eye(K+1);

C00=[zeros((K+1)*(n0-1),K+1);Cn(type,p,K,Lambda,mu,n0,n0-1)];
B0_=[B0 C0 zeros(K+1,(K+1)*(n0-2))];
```


DOI:10.6814/NCCU201901241

```
for i=0:n0-3
    B0_=[B0_;zeros(K+1,(K+1)*i) A ...
    Bn(type,p,K,Lambda,mu,n0,i+1) Cn(type,p,K,Lambda,mu,n0,i+1) ...
    zeros(K+1,(K+1)*(n0-3-i))];
end
B00=[B0_;zeros(K+1,(K+1)*(n0-2)) A ...
     Bn(type,p,K,Lambda,mu,n0,n0-1)];
A00=[zeros(K+1,(K+1)*(n0-1)) A];

C=([zeros(1,K) Lambda]')*[zeros(1,K) 1];
B=mu_2*([1 zeros(1,K)]')*[1 zeros(1,K)] ...
  -(Lambda+mu_2+mu_1)*diag(ones(K+1,1)) ...
  +mu_2*diag(ones(K,1),-1)+Lambda*diag(ones(K,1),1);

Q_=[ones(n0*(K+1),1) zeros(n0*(K+1),K+1) B00 C00
    zeros(n0*(K+1),K+1);
    ones(K+1,1) zeros(K+1,K+1 A00 B C;
    ones(K+1,1) diag(sigma*ones(1,K+1)) zeros(K+1,n0*(K+1))
    A B;
    [1/(1-sigma)*ones(K+1,1)] diag(-ones(1,K+1))
    zeros(K+1,n0*(K+1)) zeros(K+1,K+1) A];

pi=(Q_'\[1;zeros(size(Q_',1)-1,1)])';
pi0=pi(1:n0*(K+1));
pi1=pi(n0*(K+1)+1:(n0+1)*(K+1));
pi2=pi((n0+1)*(K+1)+1:(n0+2)*(K+1));
Z=[pi0 pi1 pi2];
end
```


DOI:10.6814/NCCU201901241

B.4.6 The Expected Queue length

Output a vector consisting of the expected waiting time $E[L_1]$ and $E[L_2]$, where K, n_0, σ, and the stationary probability $[\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2]$ calculating above are inputs.

```
function Z=EL_(K,n0,sigma,pi)
pi0=pi(1:n0*(K+1));
pi1=pi(n0*(K+1)+1:(n0+1)*(K+1));
pi2=pi((n0+1)*(K+1)+1:(n0+2)*(K+1));

L=[];
for i=0:n0-1
    L=[L i*ones(1,K+1)];
end
EL1=pi0*L'+(n0)*pi1*ones(K+1,1)+(n0*(1-sigma)+1)/(1-sigma)/ ...
(1-sigma)*pi2*ones(K+1,1);

L=[];
for i=0:K
    L=[L i];
end

pi=[zeros(1,K+1)];
for i=0:n0-1
    pi=pi+pi0(1+i*(K+1):(i+1)*(K+1));
end
pi=pi+pi1+pi2/(1-sigma);
EL2=pi*L';

Z=[EL1 EL2];
end
```


DOI:10.6814/NCCU201901241

B.4.7 The Expected Social Waiting Cost

Let $type \in \{1,\ 2,\ 3\}$ represents the type of the distributions. Output the expected social waiting cost $S_2(p)$, where p, K, n_0, σ, a positive ϵ, and the stationary probability $[\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2]$ calculating above are inputs.

```
function Z=S(type,K,n0,p,mu,sigma,pi,epsilon)
parameter_distribution=parameter_distribution_(type);
if type==1
    min=parameter_distribution(1);
    max=parameter_distribution(2);
    xi=(max^2-min^2)/(max-min)/2;
end
if type==2
    kappa=parameter_distribution(1);
    xi=1/kappa;
end
if type==3
    alpha=parameter_distribution(1);
    xm=parameter_distribution(2);
    xi=alpha*xm/(alpha-1);
end

mu_1=mu(1);mu_2=mu(2);
pi0=pi(1:n0*(K+1));
pi1=pi(n0*(K+1)+1:(n0+1)*(K+1));
pi2=pi((n0+1)*(K+1)+1:(n0+2)*(K+1));

a=0;
for i=1:n0*K
    EF=[];
    for j=0:K
        theta=p/(i/mu_1-j/mu_2);
```


DOI:10.6814/NCCU201901241

```
        if theta <=0
            theta=0;
        end
        if theta==inf
            EF=[EF [i/mu_1 j/mu_2]*[0;xi]];
        else
            EF=[EF [i/mu_1 j/mu_2]*[G(type,xita);xi-G(type,xita)]];
        end
    if i<=n0-1
        pin=pi0(i*(K+1)+1:(i+1)*(K+1));
    end
    if i==n0
        pin=pi1;
    end
    if i>=n0+1
        pin=pi2*(sigma^(i-n0-1));
    end
    if pin*EF'< epsilon
        break;
    else
        a=a+pin*EF';
    end
end
Z=a;
end
```

B.4.8 The Expected Profit of Manager

Output the value of the expected revenue per unit time of the toll system $O_2(p)$, which defined as (3.2.3). Where p, K, Λ, n_0, μ, σ, and the stationary probability $[\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2]$ calculating above are inputs.

```
function Z=O(type,K,Lambda,n0,p,mu,sigma,pi)
```


DOI:10.6814/NCCU201901241

```
pi0=pi(1:n0*(K+1));
pi1=pi(n0*(K+1)+1:(n0+1)*(K+1));
pi2=pi((n0+1)*(K+1)+1:(n0+2)*(K+1));

a=0;
for i=1:n0*K
    EP=[];
    for j=0:K
        lambda=lambda_(type,p,K,Lambda,mu,n0,i,j);
        EP=[EP p*lambda(2)];
    end
    if i<=n0-1
        pin=pi0(i*(K+1)+1:(i+1)*(K+1));
    end
    if i==n0
        pin=pi1;
    end
    if i>=n0+1
        pin=pi2*(sigma^(i-n0-1));
    end
        a=a+pin*EP';
end
Z=a;
end
```

B.4.9 The Complementary Cumulative Distribution Function of The Waiting Time

Output a vector consisting of the function values $P(W_1 > t)$ and $P(W_2 > t)$, which defined as (2.3.10). Where K, n_0, μ, σ, $[\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2]$ and a positive number t are inputs.

```
function Z=probability_waiting_time(K,n0,mu,sigma,pi,t)
```


DOI:10.6814/NCCU201901241

```
mu_1=mu(1);mu_2=mu(2);
pi0=pi(1:n0*(K+1));
pi1=pi(n0*(K+1)+1:(n0+1)*(K+1));
pi2=pi((n0+1)*(K+1)+1:(n0+2)*(K+1));
pi_t=[zeros(1,K+1)];
for i=0:n0-1
    pi_t=pi_t+pi0(1+i*(K+1):(i+1)*(K+1));
end
pi_t=pi_t+pi1+pi2/(1-sigma);
PW_=[];
    a=0;
    for i=0:K
        a=a+pi_t(1+i)*Taylor_T(i,mu_2*t);
    end
probability_waiting_time_2=a*exp(-mu_2*t);

pi_g=[];
for i=0:n0-1
    pi_g=[pi_g sum(pi0(1+i*(K+1):(i+1)*(K+1)))];
end
pi_g=[pi_g sum(pi1) sum(pi2)];
PW_=0;
for i=0:n0+1
    PW_=PW_+pi_g(i+1)*Taylor_T(i,mu_1*t)*exp(-mu_1*t);
end
probability_waiting_time_1=PW_ ...
+sum(pi2*(Taylor_T(n0+1,mu_1*t)*sigma/(1-sigma) ...
+Taylor_R(n0+1,mu_1*t*sigma)/(sigma^(n0+1)) ...
/(1-sigma))*exp(-mu_1*t));
Z=[probability_waiting_time_1 probability_waiting_time_2];
end
```


DOI:10.6814/NCCU201901241

B.4.10 The Constraints of Optimization

Let $type \in \{1,\ 2,\ 3\}$ represents the type of the distributions. Output a vector consisting of the values ciq and ceq, which is used to nonlinear inequality $ciq < 0$ and nonlinear equation $ceq = 0$. Where p, K, Λ, n_0, μ, and a positive number c are inputs. This is based on the constraints of the optimization problem to choose whether to use the annotation part.

```
function [ciq,ceq]=mycon(type,K,Lambda,n0,p,mu,c)
sigma=sigma_(K,Lambda,mu);
pi=pi_(type,p,K,Lambda,mu,n0,sigma);
EL=EL_(K,n0,sigma,pi);
fun_=@(t) probability_waiting_time(K,n0,mu,sigma,pi,t);
intigral_W=integral(fun_,0,50,'ArrayValued',true);
pwt=probability_waiting_time(K,n0,mu,sigma,pi,intigral_W(1));
%pwt=probability_waiting_time(K,n0,mu,sigma,pi,intigral_W(2));

ciq=[pwt(1)-c];
%ciq=[pwt(2)-c];
ceq=[];
end
```


DOI:10.6814/NCCU201901241