Strategic queueing behavior for individual and social optimization in managing discrete time working vacation queue with Bernoulli interruption schedule

Miaomiao Yu[a,b], Attahiru Sule Alfa[b,c]

[a] *School of Science, Sichuan University of Science and Engineering, 643000, Zigong, Sichuan, China*
[b] *Department of Electrical & Computer Engineering, University of Manitoba, R3T 5V6, Winnipeg, Manitoba, Canada*
[c] *Department of Electrical, Electronic & Computer Engineering, University of Pretoria, 0002, Pretoria, South Africa*

Abstract

In this paper, we consider a discrete time working vacation queue with a utility function for the reward of receiving the service and the cost of waiting in the system. A more flexible switching mechanism between low and regular service states is introduced to enhance the practical value of the working vacation queue. Under different precision levels of the system information, namely observable, almost unobservable and fully unobservable cases, the utility function is studied from both the individual customer's and the system administrator's points of view. By analyzing the steady-state behavior of the system, the associated optimal joining decisions under different information scenarios are obtained. We find that for the fully observable queue, the joining threshold for individual optimization may be less than the one for social optimization in working vacation period. A similar situation also appears in almost unobservable case. Such phenomenon is not possible for the classic first come first served queue due to the fact that there is no vacation time and thus will not cause large fluctuations in customers' conditional waiting time. Additionally, we also conduct some numerical comparisons to demonstrate the effect of the information levels as well as system parameters on customer joining behavior.

Keywords: Queueing; Working vacation; Bernoulli interruption schedule; Joining strategy; Conditional sojourn time

1. Introduction

In a queueing system, customers arrive at the service facility to get a certain benefit from the service, but they often encounter the annoyance from waiting. Under normal circumstances, upon arrival at the system, customers usually observe the existing queue length and the service fee, and then decide whether or not to join the queue based on the perceptions of personal benefits. Thus, during the last decades, there exists a widespread tendency to investigate decision making problems in the waiting line system from an economic viewpoint. Some natural reward-cost structures which incorporate customers' desires for service and their unwillingness to wait are imposed on the system. Such an economic analysis for decision making in the queues can be traced to the pioneering work of Naor [1] who studied the situation where arriving customers are admitted or not based on the observed queue length. By establishing a queueing cost model which envisages the imposition of tolls on newly arriving customers, Naor showed that levying tolls is an effective strategy that might attain social welfare optimization. In the several years following this article, a number of authors have addressed related issues. Yechiali [2, 3] extended Naor's results to $GI/M/1$ and $GI/M/s$ queues with one customer type and linear holding cost. Stidham [4] introduced a fixed reward and a waiting cost for each job passing through the system, and considered the optimal control of admission to a queueing system. Mendelson [5] studied optimal pricing and capacity decisions for a service facility in a microeconomic framework, and also investigated the effects of queueing delays and customer's related costs on the management of computing systems. These studies invariably assumed that customers can balk if the expected waiting cost of their jobs is too high. At the same time, they also demonstrated the inconsistency between the individually and socially optimal joining rules. Additionally, Edelson and Hildebrand [6] reexamined the work of Naor and introduced a balking model in which customers do not observe the system state before making an unchangeable joining decision. They further revealed that revenue maximization and social optimization occur simultaneously under such a situation.

In recent times, there has been an upsurge of interest towards economic analysis for decision making in the waiting line system, especially those classified as optimal design and control of queues. For different precision levels of system information, Burnetas and E-conomou [7] considered a Markovian single-server queueing system with setup times. They derived the equilibrium balking strategies for the customers and analyzed the stationary behavior of the system. Based on the above work, Economou and Kanta [8] further studied the equilibrium joining behavior in an $M/M/1$ repairable queue. Following the idea of Economou and Kanta, by assuming that repair is not provided immediately, Wang and Zhang [9] reconsidered the customer's balking strategy in fully and partially observable queues. Moreover, equilibrium analysis is also conducted on an observable queueing system with setup and closedown times by Sun et al. [10], and on a clearing queueing system in alternating environment by Economou and Manou [11]. Meanwhile, accomplishing the development of vacation queueing theory, the economic analysis for decision making in vacation queues has drawn the attention of numerous researchers. Guo and Hassin [12, 13] studied a Markovian vacation queue with N-policy and exhaustive service. They presented individually and socially optimal strategies for unobservable and observable queues. An essential extension to queue with general service and vacation times appeared in the recent paper written by Economou et al.[14]. Mean value approach was employed by them for the derivation of the main performance measures. Liu et al.[15] were the first to study customer's strategic queueing behavior in discrete time vacation queue. Shortly after the publication of this seminal paper, Ma et al.[16] developed this research topic by considering the $Geo/Geo/1$ queue with multiple vacation policy, in which customers' equilibrium balking strategies were discussed under four different information scenarios. More recently, inspired by the working vacation mechanism, equilibrium analysis for the $M/M/1$ queue with multiple working vacations are conducted simultaneously by Sun and Li [17] and Zhang et al.[18]. Both individually and socially optimal joining rules are obtained and compared by them. More interesting extensions about the effect of information on the strategic behavior in queueing systems can be found in papers Economou and Kanta [19], Boudali and Economou [20, 21], Wang and Zhang [22] and Li et al.[23], etc.

Based on the above brief literature review, we note that significant progress has been made in the continuous time queues with customers' strategic queueing behaviors. But it still seems that little more than a beginning has been made in their discrete time counterparts. Except a limited number of studies done by several Chinese scholars (see, e.g. [15, 16]), no work in this direction has come to our notice. Many economic analysis and decision making problems for discrete time queues have been left unexplored. More importantly, for practical measurement related purpose, time is sometimes considered as a discrete quantity although it is continuous. We often hear people say a system is observed every minute, every second, every half a second, etc. Meanwhile, with the development of digital communication technology, many modern communication systems are operated based on a time-slot basis, which are naturally and appropriately modeled by discrete time queues. Thus, the main purpose of this paper is to develop an analytical model that allows us extensively analyze and explore the strategic queueing behavior arising in $Geo/Geo/1$ working vacation queue with Bernoulli interruption schedule.

For evaluating the performance measures of gateway router in fiber communication networks, the concept of working vacation policy was first introduced by Servi and Finn [24] in 2002. A major difference between working vacation queue and classical vacation queue is that during a vacation period, customers in the former can be served in a lower service rate; however, customers in the latter can impossibly be served and depart the system. Due to the strong application background in optimal design of stochastic service systems, working vacation queues have received considerable attention in the past ten years. Many fruitful theoretical results and interesting applications are presented in this area (see, e.g. [25-37]). On the other hand, in both single and multiple working vacation policies, server resumes fast service rate only when the system is non-empty at the end of a vacation. Obviously, such assumption might not be tenable in many occasions. For example, when a batch of patients who are injured badly in a car accident need surgery, idle surgeons may temporarily interrupt their vacations and return to the hospital for doing emergency surgery. In addition to the above, it is generally known that virus scan is an important maintenance activity for the terminal server which helps keep its functioning well. This type of maintenance could be performed when the system load is relatively light. Although running a virus scan will consume some system resources and result in a slow processing speed, the terminal server could still provide his service with a lower rate.

Further, if there are some tasks that require special treatment, the process of virus scan can be interrupted, and the server can resume his original service speed depending on the time-delay sensitivity of these tasks. Thus, it seems that the introduction of the Bernoulli vacation interruption schedule into working vacation queue is very reasonable and necessary for some practical situations. So, investigating individually and socially optimal joining rules for this system would be an interesting and significant research topic. At the same time, it is worth mentioning here that one of the complexities in the analysis of discrete time queues is the occurrences of simultaneous arrivals and departures at the boundary epochs of a slot. The analysis becomes further complicated in the case of working vacation queue with Bernoulli interruption schedule as there exist many directly accessible states for each system state.

The rest of this paper is organized as follows. The next section describes the mathematical model. Sections 3, 4, and 5 are devoted to the fully observable, the almost unobservable and the fully unobservable queues, respectively. We demonstrate how to obtain the individually and socially optimal joining strategies for each type of queues. Some numerical results are also presented and discussed in these sections. This paper ends with Section 6 where conclusions and future scope are given.

2. Model description

We consider a discrete time multiple working vacation queue with Bernoulli interruption schedule, whose service will not completely remain inactive during the server vacation period. In our model, the inter-arrival times $\{T_r, r \geq 1\}$ of customers are independent and identically distributed random variables with probability mass function (p.m.f.) $\Pr\{T_r = k\} = \lambda\bar{\lambda}^{k-1}$, $k \geq 1$, where we use symbol $\bar{x} = 1-x$, for any real number $x \in (0, 1)$. The service time S_b in a regular busy period follows geometric distribution with parameter μ_b, namely, $\Pr\{S_b = k\} = \mu_b\bar{\mu}_b^{k-1}$, $k \geq 1$. The server commences a working vacation of random length at the epoch when the system becomes empty. The working vacation time V is geometrically distributed with p.m.f. $\Pr\{V = k\} = \theta\bar{\theta}^{k-1}$, $k \geq 1$. It is an operating period with a lower service rate, the service time S_v in working vacation period follows a geometric distribution with parameter $\mu_v (0 < \mu_v < \mu_b < 1)$, namely, $\Pr\{S_v = k\} = \mu_v\bar{\mu}_v^{k-1}$, $k \geq 1$. After serving a customer in working vacation period, if the server finds any customer waiting in the queue, the vacation either is interrupted with probability $p (0 \leq p \leq 1)$ or continues with probability $\bar{p}$. If there are no customers in the queue, the working vacation continues. Furthermore, after completing a working vacation, if the system is non-empty at that moment, the server switches its service rate from μ_v to μ_b and starts a regular busy period immediately; otherwise, the server takes another working vacation. The low speed service interrupted at the end of working vacation restarts from the beginning. In order to give an economical sense to the queueing management, we also assume that each customer who joins the queue receives a reward R from service and experiences a delay cost of C per unit time. To avoid triviality, we impose two conditions on the current model:

(1) Customers are risk neutral and maximize their expected net benefit in equilibrium. Under favorable circumstance, namely when the server stays idle, a customer will desire to queue up for receiving the completion of service. This also means that we always restrict ourselves in the non-trivial case where some customers have an incentive to join the queue. This assumption ensures that the reward for service exceeds the expected cost for a customer who finds the system empty; (2) To facilitate the analysis, we suppose that various stochastic processes involved in the system are independent of each other.

In discrete time queueing system, the time axis is divided into equal intervals called slots and all queueing activities occur at the slot boundaries. Traditionally, there are two types of systems in the discrete time case (see, e.g.[38, 39]), one is the late arrival with delayed access (LAS-DA) and the other is the early arrival system (EAS). In this paper, we consider the model for the early arrival system and therefore, a potential arrival occurs in (t, t^+), and a potential departure takes place in (t^-, t), for $t = 0, 1, 2, \ldots$. To make it clear, the various time epochs at which events occur are shown in a self-explanatory figure (see Figure 1).

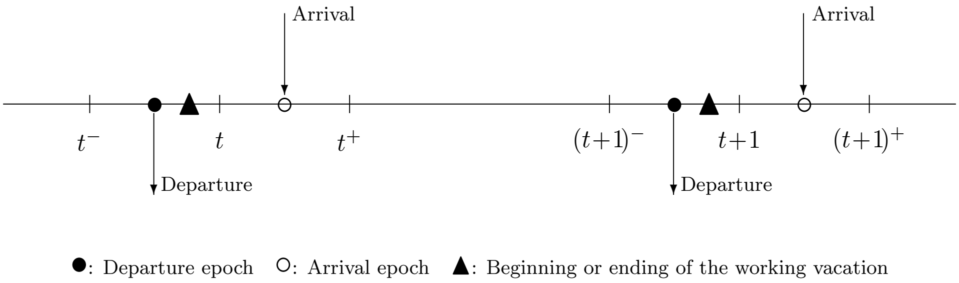
Arrival
Arrival
t^-
t
t^+
$(t+1)^-$
$t+1$
$(t+1)^+$
Departure
Departure
●: Departure epoch ○: Arrival epoch ▲: Beginning or ending of the working vacation

Figure 1. Various time epochs in an early arrival system.

To obtain the state space of the queueing system, in the sequel we use the following random variables. Let $N(t)$ be the number of customers in the system at time t. The different states of the server at time t are defined as follows:

$$Y(t) = \begin{cases} 0, \text{ if the server is on working vacation,} \\ 1, \text{ if the server is in regular busy period.} \end{cases}$$

It is readily seen that the process $\{(N(t), Y(t)), t \geq 0\}$ is a Discrete Time Markov Chain (DTMC) whose state space $\Omega = \Omega_0 \cup \Omega_1$, where $\Omega_0 = \{(i, j) | i = 0, 1, 2, \ldots, j = 0\}$ and $\Omega_1 = \{(i, j) | i = 1, 2, \ldots, j = 1\}$.

3. Analysis of the fully observable queue

In this section, we focus our attention on the fully observable case, where the state of the server and the existing queue length is communicated to the customers upon their arrival. Under such situation each customer makes a decision based on the number of customers in the system. We will prove that a threshold type individual optimal strategy exists, in the sense that this strategy maximizes the expected net reward of a customer. Next, to determine the pure threshold strategy, the analysis is carried out from two perspectives, namely individual and social optimization.

3.1. Optimal joining threshold for individual optimization

Let us take an arbitrary arriving customer as the tagged one. If the tagged customer enters the system, the total delay cost is determined by his mean sojourn time. Then, his expected net benefit after service completion can be expressed as $U = R - C\mathrm{E}\left[W_{i,j}\right]$, where $\mathrm{E}\left[W_{i,j}\right]$ $((i,j) \in \Omega)$ is the tagged customer's mean conditional sojourn time given that he finds the queueing system at state (i,j) just before his arrival. Thus the tagged customer will join the queue if and only if $U = R - C\mathrm{E}\left[W_{i,j}\right] \geq 0$. The above linear utility function allows us to compute exactly the individual net benefit of the tagged customer for any value of (i,j) and to finally get the optimal joining rule for individual optimization. From the cost structure, we may see that the expected conditional sojourn time plays an important role in solving the optimal joining threshold under individual optimization. It is clear that if the tagged customer finds the queueing system is in state $(i,1)$ just before his arrival epoch, then his sojourn time is the sum of $i+1$ regular service times, namely, for $i = 1, 2, \dots,$

$$\mathrm{E}\left[W_{i,1}\right] = \frac{i+1}{\mu_b}. \tag{1}$$

Since after each transition there exist many directly accessible states for state $(i,0)$, the calculation of $\mathrm{E}\left[W_{i,0}\right]$ is much more difficult than $\mathrm{E}\left[W_{i,1}\right]$. Here, we will first use a first-step argument to derive the generating function of $W_{i,0}$. Then, differentiating the generating function and doing some algebraic manipulations we can directly obtain the mean value of $W_{i,0}$. Let $\tilde{V}$ denote the residual working vacation time. By conditioning on the events that may occur in the next step, for $i \geq 1$, $\mathrm{E}\left[z^{W_{i,0}}\right]$ can be decomposed as

$$\begin{aligned}\mathrm{E}\left[z^{W_{i,0}}\right] =& \mathrm{E}\left[z^{W_{i,0}}\middle|\,\tilde{V} > S_v\right]\Pr\left\{\tilde{V} > S_v\right\} + \mathrm{E}\left[z^{W_{i,0}}\middle|\,\tilde{V} < S_v\right]\Pr\left\{\tilde{V} < S_v\right\}\\ &+ E\left[z^{W_{i,0}}\middle|\,\tilde{V} = S_v\right]\Pr\left\{\tilde{V} = S_v\right\}\end{aligned} \tag{2}$$

The first term on the right hand side of Eq.(2) means that the remaining working vacation time is greater than a residual service time with low service rate. According to the Bernoulli vacation interruption schedule, upon completion of the current service, the server will terminate his vacation and resume regular busy period with probability p or continue his vacation and remain low service state with probability $\bar{p}$. Thus, we have

$$\begin{aligned}&\mathrm{E}\left[z^{W_{i,0}}\middle|\,\tilde{V} > S_v\right]\Pr\left\{\tilde{V} > S_v\right\}\\ =& \sum_{k=1}^{\infty}\sum_{n=k+1}^{\infty}\mathrm{E}\left[z^{k+W_{i-1,0}}\right]\bar{p}\theta\bar{\theta}^{n-1}\mu_v\bar{\mu}_v^{k-1} + \sum_{k=1}^{\infty}\sum_{n=k+1}^{\infty}\mathrm{E}\left[z^{k+W_{i-1,1}}\right]p\theta\bar{\theta}^{n-1}\mu_v\bar{\mu}_v^{k-1}\end{aligned}$$

where we define $W_{0,1} = S_b$. The second term on the right hand side of Eq.(2) indicates that the remaining working vacation time is less than a residual service time with low service rate. Further noting that the remaining working vacation time is at least one time slot and the residual service time is at least two time slots, we obtain

$$\mathrm{E}\left[z^{W_{i,0}}\middle|\, \tilde{V}<S_v\right]\mathrm{Pr}\left\{\tilde{V}<S_v\right\}=\sum_{k=2}^{\infty}\sum_{n=1}^{k-1}\mathrm{E}\left[z^{\tilde{V}+W_{i,1}}\right]\theta\bar{\theta}^{n-1}\mu_v\bar{\mu}_v^{k-1}.$$

Since the probability of occurrence of the event $\tilde{V} = S_v$ is not equal to zero in discrete time queueing model, we must not ignore the discussion of this case. Applying some simple probabilistic arguments, the third term on the right hand side of Eq.(2) can be expressed as follows

$$\mathrm{E}\left[z^{W_{i,0}}\middle|\, \tilde{V}=S_v\right]\mathrm{Pr}\left\{\tilde{V}=S_v\right\}=\sum_{n=1}^{\infty}\mathrm{E}\left[z^{n+W_{i-1,1}}\right]\theta\bar{\theta}^{n-1}\mu_v\bar{\mu}_v^{n-1}.$$

As an immediate consequence of these results, after some algebraic manipulations, we have

$$\mathrm{E}\left[z^{W_{i,0}}\right]=\mathrm{E}\left[z^{W_{i-1,0}}\right]\frac{\mu_v\bar{p}\bar{\theta}z}{1-\bar{\mu}_v\bar{\theta}z}+\mathrm{E}\left[z^{W_{i-1,1}}\right]\frac{\mu_v p\bar{\theta}z+\theta\mu_v z}{1-\bar{\mu}_v\bar{\theta}z}+\mathrm{E}\left[z^{W_{i,1}}\right]\frac{\theta\bar{\mu}_v z}{1-\bar{\mu}_v\bar{\theta}z},\quad i\geq 1.$$

Similarly, for $i = 0$,

$$\begin{aligned}\mathrm{E}\left[z^{W_{0,0}}\right]&=\mathrm{E}\left[z^{W_{0,0}}\middle|\,\tilde{V}\geq S_v\right]\mathrm{Pr}\left\{\tilde{V}\geq S_v\right\}+\mathrm{E}\left[z^{W_{0,0}}\middle|\,\tilde{V}<S_v\right]\mathrm{Pr}\left\{\tilde{V}<S_v\right\}\\&=\sum_{k=1}^{\infty}\sum_{n=k}^{\infty}z^k\theta\bar{\theta}^{n-1}\mu_v\bar{\mu}_v^{k-1}+\sum_{k=2}^{\infty}\sum_{n=1}^{k-1}\mathrm{E}\left[z^{\tilde{V}+S_b}\right]\theta\bar{\theta}^{n-1}\mu_v\bar{\mu}_v^{k-1}\\&=\frac{z\mu_v}{1-z\bar{\mu}_v\bar{\theta}}+\sum_{k=2}^{\infty}\sum_{n=1}^{k-1}\mathrm{E}\left[z^{S_b}\right]z^n\theta\bar{\theta}^{n-1}\mu_v\bar{\mu}_v^{k-1}\\&=\frac{z\mu_v}{1-z\bar{\mu}_v\bar{\theta}}+\frac{z\mu_b}{1-z\bar{\mu}_b}\frac{z\theta\bar{\mu}_v}{1-z\bar{\theta}\bar{\mu}_v}.\end{aligned}\tag{3}$$

Employing Eq.(1) and evaluating $\frac{\mathrm{dE}\left[z^{W_{i,0}}\right]}{\mathrm{d}z}$ at $z = 1$, we can obtain

$$\mathrm{E}\left[W_{0,0}\right]=\frac{\mu_b+\theta\bar{\mu}_v}{\mu_b\left(1-\bar{\mu}_v\bar{\theta}\right)},\tag{4}$$

$$\mathrm{E}\left[W_{i,0}\right]=\frac{1}{1-\bar{\mu}_v\bar{\theta}}+\frac{\mu_v\bar{p}\bar{\theta}}{1-\bar{\mu}_v\bar{\theta}}\mathrm{E}[W_{i-1,0}]+\frac{i}{\mu_b}\left(\frac{\mu_v p\bar{\theta}+\theta\mu_v}{1-\bar{\mu}_v\bar{\theta}}\right)+\frac{i+1}{\mu_b}\frac{\theta\bar{\mu}_v}{1-\bar{\mu}_v\bar{\theta}},\ i=1,2,\ldots.\tag{5}$$

From Eq.(5), we should be able to derive the following relationship

$$\mathrm{E}\left[W_{i,0}\right]-\mathrm{E}\left[W_{i-1,0}\right]-\frac{1}{\mu_b}=\frac{\mu_v\bar{p}\bar{\theta}}{1-\bar{\theta}\bar{\mu}_v}\left(\mathrm{E}\left[W_{i-1,0}\right]-\mathrm{E}\left[W_{i-2,0}\right]-\frac{1}{\mu_b}\right),\ i=2,3,\ldots.\tag{6}$$

Using Eqs.(4) and (5) in Eq.(6), we have

$$\mathrm{E}\left[W_{i,0}\right]-\mathrm{E}\left[W_{i-1,0}\right]=\frac{\mu_v\bar{p}\bar{\theta}\left(\mu_b-\mu_v\right)}{\mu_b\left(1-\bar{\mu}_v\bar{\theta}\right)^2}\left(\frac{\mu_v\bar{p}\bar{\theta}}{1-\bar{\mu}_v\bar{\theta}}\right)^{i-1}+\frac{1}{\mu_b},\ i=1,2,\ldots.\tag{7}$$

Noting that $\mathrm{E}\left[W_{i,0}\right]-\mathrm{E}\left[W_{0,0}\right]=\sum\limits_{k=1}^{i}\left(\mathrm{E}\left[W_{k.0}\right]-\mathrm{E}\left[W_{k-1,0}\right]\right)$, we finally obtain

$$\mathrm{E}\left[W_{i,0}\right]=\frac{\mu_v\bar{\theta}\bar{p}\left(\mu_b-\mu_v\right)}{\mu_b\left(1-\bar{\mu}_v\bar{\theta}\right)\left(1-\bar{\mu}_v\bar{\theta}-\mu_v\bar{p}\bar{\theta}\right)}\left[1-\left(\frac{\mu_v\bar{p}\bar{\theta}}{1-\bar{\mu}_v\bar{\theta}}\right)^{i}\right]+\frac{i}{\mu_b}+\frac{\mu_b+\theta\bar{\mu}_v}{\mu_b\left(1-\bar{\theta}\bar{\mu}_v\right)},\ i=0,1,2,\ldots. \tag{8}$$

From Eqs.(1) and (8), we see that in the fully observable queue, the optimal joining strategy for individual optimization can be established in a fashion which finds out a pair of integers $(I_e(0), I_e(1))$, satisfies the following four inequalities,

$$R-C\left\{\frac{\mu_v\bar{\theta}\bar{p}\left(\mu_b-\mu_v\right)}{\mu_b\left(1-\bar{\mu}_v\bar{\theta}\right)\left(1-\bar{\mu}_v\bar{\theta}-\mu_v\bar{p}\bar{\theta}\right)}\left[1-\left(\frac{\mu_v\bar{p}\bar{\theta}}{1-\bar{\mu}_v\bar{\theta}}\right)^{I_e(0)}\right]+\frac{I_e(0)}{\mu_b}+\frac{\mu_b+\theta\bar{\mu}_v}{\mu_b\left(1-\bar{\theta}\bar{\mu}_v\right)}\right\}\geq 0, \tag{9}$$

$$R-C\left\{\frac{\mu_v\bar{\theta}\bar{p}\left(\mu_b-\mu_v\right)}{\mu_b\left(1-\bar{\mu}_v\bar{\theta}\right)\left(1-\bar{\mu}_v\bar{\theta}-\mu_v\bar{p}\bar{\theta}\right)}\left[1-\left(\frac{\mu_v\bar{p}\bar{\theta}}{1-\bar{\mu}_v\bar{\theta}}\right)^{I_e(0)+1}\right]+\frac{I_e(0)+1}{\mu_b}+\frac{\mu_b+\theta\bar{\mu}_v}{\mu_b\left(1-\bar{\theta}\bar{\mu}_v\right)}\right\}<0, \tag{10}$$

$$R-C\left(\frac{I_e(1)+1}{\mu_b}\right)\geq 0, \tag{11}$$

$$R-C\left(\frac{I_e(1)+2}{\mu_b}\right)<0. \tag{12}$$

Here, inequalities (9) and (11) correspond to the cases where the customer is supposed to join the system and then receives the service to reap the reward. On the other hand, when inequalities (10) and (12) hold, arriving customer balks without joining. Incorporating the inequalities (11) and (12) in one expression, we have

$$I_e(1)=\left\lfloor\frac{R\mu_b}{C}\right\rfloor-1,$$

where the symbol $\lfloor x\rfloor$ denotes the largest integer not exceeding x.

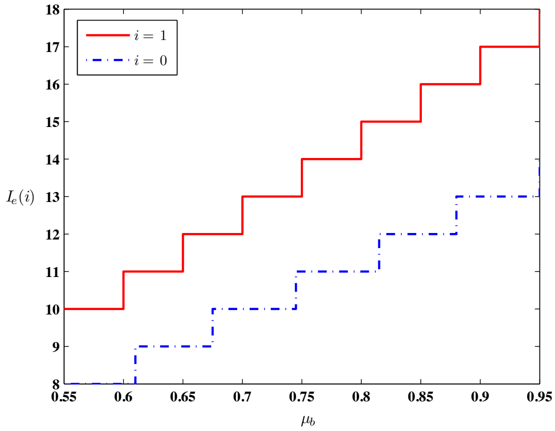

Figure 2. Optimal joining threshold under individual optimization for the different values of μ_b.

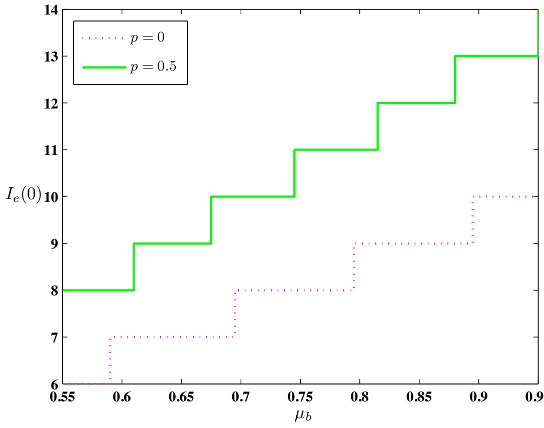

Figure 3. Compare the threshold value $I_e(0)$ for $p=0$ with $p=0.5$.

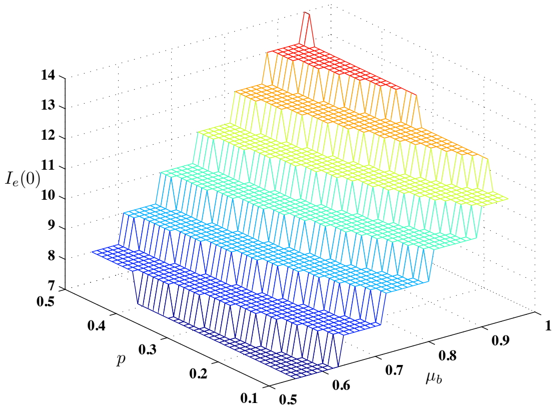

Figure 4. Dependence of individually optimal joining threshold $I_e(0)$ on μ_b and p.

Due to the high non-linear and complex nature of inequalities (9) and (10), it is extremely difficult to develop an analytic solution of $I_e(0)$ symbolically. However, we can get the value of $I_e(0)$ through some effective numerical methods. To illustrate the application of theoretical results, we investigate several numerical examples by selecting $R = 30$, $C = 1.5$, $\mu_v = 0.2$, $\theta = 0.1$, $p = 0.5$ and letting μ_b vary from 0.55 to 0.95. From Figure 2, we observe that an increase in the parameter μ_b implies an increase in two kinds of joining thresholds, and $I_e(1)$ is always greater than $I_e(0)$ with the increase of μ_b. This numerical result is entirely consistent with the actual situation since an arriving customer is more likely to enter the system when the server is in regular busy period. Moreover, we notice that if $p = 0$ this model becomes $Geo/Geo/1$ multiple working vacation queue without vacation interruptions. The values of $I_e(0)$ for $p = 0$ and $p = 0.5$ are graphically presented in Figure 3. We can see that the joining threshold value for the working vacation queue with Bernoulli interruption schedule is greater than the one without vacation interruptions. This is exactly what we had expected, since vacation interruption mechanism can attract more customers to join the queue. Additionally, letting p vary from 0.1 to 0.5, the effect of Bernoulli schedule control probability p on the joining threshold $I_e(0)$ is depicted together with the change of parameter μ_b in Figure 4. It can be seen that when the value of μ_b is smaller, the change of $I_e(0)$ is not sensitive to the value of p.

3.2. Optimal joining threshold for social optimization

In this subsection, we change our viewpoint from the individual customer to the whole system, where the queueing system is composed of a server, all potential customers and an administrator who has the authority to control the queue length. If the control of administrator is exercised by setting an entrance fee imposed on the entering customers, then the form of the social optimal policy is the same as the individual optimal policy and only the values of specific parameters change. Since the individual optimal policies are threshold policies, it suffices to seek social optimal policies within the family of threshold polices. Thus, By restricting the number of entering customers, we will discuss how to achieve the maximal system expected profit, namely the so-called social welfare optimization. Here, we assume that when the server is on working vacation, administrator restricts $I(0)+1$ customers in the system. That is to say, the service facility has a limited waiting room of size $I(0)+1$ in working vacation period. Similarly, if the number of customers has already reached $I(1)+1$ in regular busy period, then any newly arriving customers will simply be rejected. In addition, since the server works at a lower rate than in a regular busy period, it is reasonable to assume that $I(0)$ is smaller than $I(1)$. On the other hand, if the administrator charges an entrance fee for customer who joins the queue, then the reward of customer is reduced and everything else still remains the same, so we only have imposed restrictions on balking threshold. According to the above assumptions, it is easily seen that the two-dimensional process $\{N(t), Y(t), t \geq 0\}$ is a finite DTMC with state space

$$\Omega_{ob} = \{(i,0) : i \in \{0, 1, \ldots, I(0)+1\}\} \cup \{(i,1) : i \in \{1, \ldots, I(1)+1\}\}.$$

In order to calculate the social welfare per unit time, we first need to obtain the steady-state queue length distribution at time t. Toward this end, let us define the following stationary probability distributions for the DTMC:

$$\begin{aligned} P_{i,0} &= \lim_{t\to\infty} P_{i,0}(t) = \lim_{t\to\infty} \Pr\{N(t)=i, Y(t)=0\},\ i = 0,1,\ldots,I(0)+1, \\ P_{i,1} &= \lim_{t\to\infty} P_{i,1}(t) = \lim_{t\to\infty} \Pr\{N(t)=i, Y(t)=1\},\ i = 1,2,\ldots,I(1)+1. \end{aligned}$$

Relating the state of the system at time t and $t+1$, and using simple probabilistic arguments, we can get a set of Kolmogorov-type difference equations as follows:

$$P_{0,0}\lambda\bar{\mu}_v = P_{1,0}\bar{\lambda}\mu_v + P_{1,1}\bar{\lambda}\mu_b, \tag{13}$$

$$P_{i,0}\left[1-\bar{\theta}\big(\bar{\lambda}\bar{\mu}_v+\lambda\mu_v\bar{p}\big)\right] = P_{i-1,0}\lambda\bar{\mu}_v\bar{\theta} + P_{i+1,0}\bar{\lambda}\mu_v\bar{p}\bar{\theta},\ i = 1,2,\ldots,I(0)-1, \tag{14}$$

$$P_{I(0),0}\left[1-\bar{\theta}\big(\bar{\lambda}\bar{\mu}_v+\lambda\mu_v\bar{p}\big)\right] = P_{I(0)-1,0}\lambda\bar{\mu}_v\bar{\theta} + P_{I(0)+1,0}\mu_v\bar{p}\bar{\theta}, \tag{15}$$

$$P_{I(0)+1,0}\left(1-\bar{\mu}_v\bar{\theta}\right) = P_{I(0),0}\lambda\bar{\mu}_v\bar{\theta}, \tag{16}$$

$$P_{1,1}\left(\bar{\lambda}\mu_b+\lambda\bar{\mu}_b\right) = P_{2,1}\bar{\lambda}\mu_b + P_{0,0}\lambda\bar{\mu}_v\theta + P_{1,0}\left[\bar{\lambda}\bar{\mu}_v\theta+\lambda\mu_v\left(p\bar{\theta}+\theta\right)\right] + P_{2,0}\bar{\lambda}\mu_v\big(\theta+p\bar{\theta}\big), \tag{17}$$

$$\begin{aligned} P_{i,1}\left(\bar{\lambda}\mu_b+\lambda\bar{\mu}_b\right) = {} & P_{i+1,1}\bar{\lambda}\mu_b + P_{i-1,1}\lambda\bar{\mu}_b + P_{i-1,0}\lambda\bar{\mu}_v\theta + P_{i,0}\left[\bar{\lambda}\bar{\mu}_v\theta+\lambda\mu_v\left(p\bar{\theta}+\theta\right)\right] \\ & + P_{i+1,0}\bar{\lambda}\mu_v\big(\theta+p\bar{\theta}\big),\quad i = 2,3,\ldots,I(0)-1, \end{aligned} \tag{18}$$

$$\begin{aligned} P_{I(0),1}\left(\bar{\lambda}\mu_b+\lambda\bar{\mu}_b\right) = {} & P_{I(0)+1,1}\bar{\lambda}\mu_b + P_{I(0)-1,1}\lambda\bar{\mu}_b + P_{I(0)-1,0}\lambda\bar{\mu}_v\theta + P_{I(0),0}\big[\bar{\lambda}\bar{\mu}_v\theta+\lambda\mu_v\left(p\bar{\theta}+\theta\right)\big] \\ & + P_{I(0)+1,0}\mu_v\big(\theta+p\bar{\theta}\big), \end{aligned} \tag{19}$$

$$P_{I(0)+1,1}\big(\bar{\lambda}\mu_b+\lambda\bar{\mu}_b\big) = P_{I(0)+2,1}\bar{\lambda}\mu_b + P_{I(0),1}\lambda\bar{\mu}_b + P_{I(0),0}\lambda\bar{\mu}_v\theta + P_{I(0)+1,0}\bar{\mu}_v\theta, \tag{20}$$

$$P_{i,1}\left(\bar{\lambda}\mu_b+\lambda\bar{\mu}_b\right) = P_{i+1,1}\bar{\lambda}\mu_b + P_{i-1,1}\lambda\bar{\mu}_b,\ i = I(0)+2, I(0)+3,\ldots,I(1)-1, \tag{21}$$

$$P_{I(1),1}\left(\bar{\lambda}\mu_b+\lambda\bar{\mu}_b\right) = P_{I(1)+1,1}\mu_b + P_{I(1)-1,1}\lambda\bar{\mu}_b, \tag{22}$$

$$P_{I(1)+1,1}\mu_b = P_{I(1),1}\lambda\bar{\mu}_b. \tag{23}$$

We note that Eq.(14) is a second order linear homogeneous difference equation with coefficients independent of i. So, the stationary probability $P_{i,0}\,(i = 0,1,\ldots,I(0))$ can be determined by finding the general solution of the following difference equation

$$\bar{\lambda}\mu_v\bar{p}\bar{\theta}x_{i+1}-\left[1-\bar{\theta}\left(\bar{\lambda}\bar{\mu}_v+\lambda\mu_v\bar{p}\right)\right]x_i+\lambda\bar{\mu}_v\bar{\theta}x_{i-1}=0,\ i=1,2,\dots,I(0)-1. \tag{24}$$

According to the theory of linear difference equation with constant coefficients (see [40]), the general solution of Eq.(24) may be written as

$$x_i^{\mathrm{LH}}=A_1\sigma_1^i+B_1\sigma_2^i,\ i=0,1,\dots,I(0), \tag{25}$$

where A_1 and B_1 are constants to be determined, σ_1 and σ_2 are the roots of quadratic characteristic equation

$$\bar{\lambda}\mu_v\bar{p}\bar{\theta}x^2-\left[1-\bar{\theta}\left(\bar{\lambda}\bar{\mu}_v+\lambda\mu_v\bar{p}\right)\right]x+\lambda\bar{\mu}_v\bar{\theta}=0.$$

Clearly,

$$\sigma_{1,2}=\frac{\left[1-\bar{\theta}\left(\bar{\lambda}\bar{\mu}_v+\lambda\mu_v\bar{p}\right)\right]\pm\sqrt{\left[1-\bar{\theta}\left(\bar{\lambda}\bar{\mu}_v+\lambda\mu_v\bar{p}\right)\right]^2-4\bar{\lambda}\mu_v\bar{p}\lambda\bar{\mu}_v\bar{\theta}^2}}{2\bar{\lambda}\mu_v\bar{p}\bar{\theta}}.$$

To find the constants A_1 and B_1, substituting Eq.(25) into Eqs.(13), (15) and (16) gives

$$\begin{cases} A_1\left(\lambda\bar{\mu}_v-\sigma_1\bar{\lambda}\mu_v\right)+B_1\left(\lambda\bar{\mu}_v-\sigma_2\bar{\lambda}\mu_v\right)=P_{1,1}\bar{\lambda}\mu_b, \\ A_1\left(\sigma_1^{I(0)}-\psi\sigma_1^{I(0)-1}\right)+B_1\left(\sigma_2^{I(0)}-\psi\sigma_2^{I(0)-1}\right)=0, \end{cases}$$

where $\psi=\dfrac{\lambda\bar{\mu}_v\bar{\theta}\left(1-\bar{\mu}_v\bar{\theta}\right)}{1-\bar{\mu}_v\bar{\theta}-\bar{\theta}\left(\bar{\lambda}\bar{\mu}_v+\lambda\mu_v\bar{p}\right)+\bar{\lambda}\bar{\mu}_v^2\bar{\theta}^2}$. After solving the above system of equations, we can obtain A_1 and B_1, respectively. Having calculated the constants A_1 and B_1, $P_{i,0}$ can be expressed in terms of $P_{1,1}$ as follows

$$P_{i,0}=\begin{cases} A_1\sigma_1^i+B_1\sigma_2^i, & i=0,1,\dots,I(0), \\ \dfrac{\lambda\bar{\mu}_v\bar{\theta}}{1-\bar{\mu}_v\bar{\theta}}\left(A_1\sigma_1^{I(0)}+B_1\sigma_2^{I(0)}\right), & i=I(0)+1. \end{cases} \tag{26}$$

Here, we do not list the expressions for A_1 and B_1 due to the space restrictions. Further, employing Eq.(26), we can rewrite Eq.(18) as

$$\begin{aligned}\bar{\lambda}\mu_bP_{i+1,1}-\left(\bar{\lambda}\mu_b+\lambda\bar{\mu}_b\right)P_{i,1}+\lambda\bar{\mu}_bP_{i-1,1}=&-A_1\left[\bar{\lambda}\mu_v\left(\theta+p\bar{\theta}\right)\sigma_1^2+\left(\bar{\lambda}\bar{\mu}_v\theta+\lambda\mu_v\left(\theta+p\bar{\theta}\right)\right)\sigma_1+\lambda\bar{\mu}_v\theta\right]\sigma_1^{i-1}\\ &-B_1\left[\bar{\lambda}\mu_v\left(\theta+p\bar{\theta}\right)\sigma_2^2+\left(\bar{\lambda}\bar{\mu}_v\theta+\lambda\mu_v\left(\theta+p\bar{\theta}\right)\right)\sigma_2+\lambda\bar{\mu}_v\theta\right]\sigma_2^{i-1},\\ &i=2,3,\dots,I(0)-1.\end{aligned} \tag{27}$$

Eq.(27) indicates that the stationary probability $P_{i,1}(i=0,1,\dots,I(0))$ is the solution of the following linear nonhomogeneous difference equation

$$\bar{\lambda}\mu_by_{i+1}-\left(\bar{\lambda}\mu_b+\lambda\bar{\mu}_b\right)y_i+\lambda\bar{\mu}_by_{i-1}=-A_1\left[\bar{\lambda}\mu_v\left(\theta+p\bar{\theta}\right)\sigma_1^2+\left(\bar{\lambda}\bar{\mu}_v\theta+\lambda\mu_v\left(\theta+p\bar{\theta}\right)\right)\sigma_1+\lambda\bar{\mu}_v\theta\right]\sigma_1^{i-1}$$

$$-B_1\left[\bar{\lambda}\mu_v\left(\theta+p\bar{\theta}\right)\sigma_2^2+\left(\bar{\lambda}\bar{\mu}_v\theta+\lambda\mu_v\left(\theta+p\bar{\theta}\right)\right)\sigma_2+\lambda\bar{\mu}_v\theta\right]\sigma_2^{i-1},$$
$$i=2,3,\ldots,I(0)-1. \tag{28}$$

The general solution of Eq.(28), denoted by y_i^{LNH}, has the following structure

$$y_i^{\mathrm{LNH}}=y_i^{\mathrm{LH}}+y_i^{\mathrm{P}},\ i=1,2,\ldots,I(0),$$

where $y_i^{\mathrm{LH}}=A_2 1^i+B_2\alpha^i$, $\alpha=\frac{\lambda\bar{\mu}_b}{\bar{\lambda}\mu_b}$ is the general solution of the associated homogeneous equation $\bar{\lambda}\mu_b y_{i+1}-\left(\bar{\lambda}\mu_b+\lambda\bar{\mu}_b\right)y_i+\lambda\bar{\mu}_b y_{i-1}=0$, and y_i^{P} represents a particular solution of Eq.(28). Since the nonhomogeneous term is not a solution of the associated homogeneous equation, we set $y_i^{\mathrm{P}}=C_1\sigma_1^i+D_1\sigma_2^i$. Substituting y_i^{P} into Eq.(28), the values of the constants C_1 and D_1 can be determined accordingly. Since it is only a routine and straightforward calculation, the explicit expressions for C_1 and D_1 are also omitted here.

The next step is to find the values of A_2 and B_2 in the general solution y_i^{LNH}. Taking into account Eq.(17) and noting that $P_{1,1}=y_1^{\mathrm{LNH}}$, we have

$$\begin{cases} P_{1,1}\left(\bar{\lambda}\mu_b+\lambda\bar{\mu}_b\right)=\left(A_2+B_2\alpha^2+C_1\sigma_1^2+D_1\sigma_2^2\right)\bar{\lambda}\mu_b+\left(A_1+B_1\right)\lambda\bar{\mu}_v\theta \\ \qquad +\left(A_1\sigma_1+B_1\sigma_2\right)\left(\bar{\lambda}\bar{\mu}_v\theta+\lambda\mu_v\left(p\bar{\theta}+\theta\right)\right)+\left(A_1\sigma_1^2+B_1\sigma_2^2\right)\bar{\lambda}\mu_v\left(p\bar{\theta}+\theta\right), \\ P_{1,1}=A_2+B_2\alpha+C_1\sigma_1+D_1\sigma_2. \end{cases}$$

Solving the two simultaneous equations algebraically gives A_2 and B_2. However, their expressions are too cumbersome and lengthy, detailed results are not shown here due to page limitation. Also, employing A_2, B_2, C_1 and D_1, $P_{i,1}$ is given as

$$P_{i,1}=A_2+B_2\alpha^i+C_1\sigma_1^i+D_1\sigma_2^i,\ i=1,2,\ldots,I(0). \tag{29}$$

With the help of Eqs.(26) and (29), we can obtain $P_{I(0)+1,1}$ from Eq.(19)

$$\begin{aligned} P_{I(0)+1,1}=&A_2+B_2\alpha^{I(0)+1}+C_1\sigma_1^{I(0)}\left[1+\alpha\left(1-\sigma_1^{-1}\right)\right]+D_1\sigma_2^{I(0)}\left[1+\alpha\left(1-\sigma_2^{-1}\right)\right] \\ &-\left(A_1\sigma_1^{I(0)}\gamma_1+B_1\sigma_2^{I(0)}\gamma_2\right). \end{aligned}$$

where

$$\gamma_i=\frac{\lambda\bar{\mu}_v\theta}{\bar{\lambda}\mu_b\sigma_i}+\frac{\left(\bar{\lambda}\bar{\mu}_v+\lambda\mu_v\right)\theta+\lambda\mu_v p\bar{\theta}-\bar{\lambda}\bar{\mu}_v^2\theta\bar{\theta}}{\bar{\lambda}\mu_b\left(1-\bar{\mu}_v\bar{\theta}\right)},\ i=1,2.$$

Then from Eq.(20), we get

$$\begin{aligned} P_{I(0)+2,1}=&A_2+B_2\alpha^{I(0)+2}+C_1\sigma_1^{I(0)}\left[\alpha\left(1-\sigma_1^{-1}\right)(1+\alpha)+1\right]+D_1\sigma_2^{I(0)}\left[\alpha\left(1-\sigma_2^{-1}\right)(1+\alpha)+1\right] \\ &-A_1\sigma_1^{I(0)}\left[\gamma_1(1+\alpha)+\frac{\lambda\bar{\mu}_v\theta}{\bar{\lambda}\mu_b\left(1-\bar{\mu}_v\bar{\theta}\right)}\right]-B_1\sigma_2^{I(0)}\left[\gamma_2(1+\alpha)+\frac{\lambda\bar{\mu}_v\theta}{\bar{\lambda}\mu_b\left(1-\bar{\mu}_v\bar{\theta}\right)}\right]. \end{aligned}$$

We find that the Eq.(21) is also a linear homogenous difference equation of order two, having for its characteristic equation $\bar{\lambda}\mu_b u_{i+1}-\left(\bar{\lambda}\mu_b+\lambda\bar{\mu}_b\right)u_i+\lambda\bar{\mu}_b u_{i-1}=0$. Thus, for $i=$

$I(0)+3, I(0)+4, \ldots, I(1)$, $P_{i,1}$ can be expressed as $P_{i,1} = A_3 + B_3\alpha^i$. Using Eqs.(21)-(23) and noting that $P_{I(0)+3,1} = (1+\alpha)P_{I(0)+2,1} - \alpha P_{I(0)+1,1}$, we have $A_3 = 0$, and B_3 is also a constant expressed in terms of A_1, B_1, A_2, B_2, C_1 and D_1. Consequently, for $i = I(0)+3, \ldots, I(1)+1$, $P_{i,1}$ be presented in the following simplified form:

$$P_{i,1} = \begin{cases} B_3\alpha^i, & i = I(0)+3, I(0)+4, \ldots, I(1), \\ \dfrac{\lambda\bar{\mu}_b}{\mu_b} B_3\alpha^{I(1)}, & i = I(1)+1. \end{cases}$$

Up to now we get the stationary probability $P_{i,j}((i,j) \in \Omega_{ob})$ in terms of $P_{1,1}$. Using the normalization condition $\sum_{(i,j)\in\Omega_{ob}} P_{i,j} = 1$, we can find $P_{1,1}$ so that all $P_{i,j}$ are completely determined. Once the stationary probabilities are calculated, the social welfare per unit time under threshold policy $(I(0), I(1))$ can be given based on BASTA property (i.e. Bernoulli arrivals see time averages)

$$U_s\left(I(0), I(1)\right) = \lambda R\left(1 - P_{I(0)+1,0} - P_{I(1)+1,1}\right) - C\left(\sum_{i=1}^{I(0)+1} iP_{i,0} + \sum_{i=1}^{I(1)+1} iP_{i,1}\right).$$

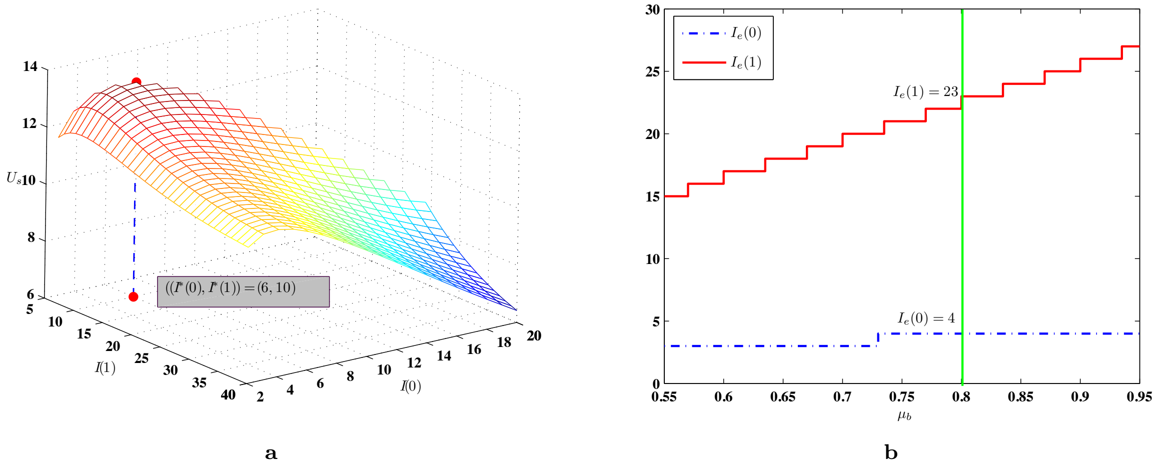

Figure 5. Comparisons of socially and individually optimal joining thresholds for the fully observable case.

In practice, for the system administrator, the decision problem is to impose a socially optimal threshold strategy, denoted by $(I^*(0), I^*(1))$, so that the social welfare can be maximized. Since $I(0)$ and $I(1)$ are discrete decision variables, we may use the direct search method to find the joint optimal values $(I^*(0), I^*(1))$. In Figure 5, the numerical results on $(I_e(0), I_e(1))$ and $(I^*(0), I^*(1))$ are presented for $R = 30$, $C = 1$, $\lambda = 0.78$, $\mu_b = 0.8$, $\mu_v = 0.12$, $\theta = 0.01$ and $p = 0.1$. Here, an interesting phenomenon is worthy of our attention. As for classic $M/M/1$ queue, an essential point illustrated by Naor is that the threshold for individual optimization is larger than the one for social optimization. However, through the above numerical experiments we may find that Naor's conclusion does not still hold in server's working vacation period. From Figure 5, we see that when the working vacation period is relatively long and vacation interruption probability is relatively small, $I^*(0)$ will be greater than $I_e(0)$. Some intuitive explanations of this phenomenon are given as follows. When a working vacation ends, if there are no customers in the queue, the server will take another working vacation. That is to say, the server will continue to provide services at a lower speed. This is abound to affect the sojourn time of future customers. Thus, in order to reduce the expected delay of future customers, the administrator should actively encourage customers to join the queue and try to avoid repeated working vacations, and resume normal working level as soon as possible. On the other hand, for creating additional service completion epoch in working vacation period and ending the low-speed period as quickly as possible, socially optimal strategy accepts more customers than those who join individually. In other words, under such a situation, the type of externalities that a joining customer brings to the system is positive.

4. Analysis of the almost unobservable queues

We now turn our interest to the almost unobservable case in which only the state of the server $Y(t)$ is communicated to the customers upon their arrival, and the information about the queue length is not being told. Since all customers are assumed indistinguishable, an equivalent way to describe customer's strategic queueing behavior is by a pair of probabilities (q_0, q_1), where $q_i(i = 0, 1)$ denotes the customer's joining probability when the server is in state i. Thus, in this section our main objective is to obtain the mixed strategies (q_0, q_1) under individual and social optimization, respectively.

4.1. Mixed strategy for individual optimization

To determine the mixed strategy for individual optimization, denoted by (q_0^e, q_1^e), we first try to get the stationary queue length distribution of the system. Because each customer can balk with probability $\bar{q}_i(i = 0, 1)$, the effective arrival rates under such a situation are Bernoulli with parameter λq_i. Arranging the elements of Ω in lexicographic order, the transition probability matrix of the vector-valued DTMC $\{(N(t), Y(t)), t \geq 0\}$ has the block tridiagonal matrix form in which three diagonal blocks repeat after a certain level. We write the matrix as

$$\mathbb{P} = \begin{pmatrix} \boldsymbol{A}_{0,0} & \boldsymbol{A}_{0,1} & & & \\ \boldsymbol{B}_{1,0} & \boldsymbol{A}_1 & \boldsymbol{A}_0 & & \\ & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \\ & & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 \\ & & & \ddots & \ddots & \ddots \end{pmatrix},$$

where

$$\boldsymbol{A}_{0,0} = \left(\lambda q_0 \mu_v + \lambda \bar{q}_0 + \bar{\lambda}\right), \boldsymbol{A}_{0,1} = \left(\, \lambda q_0 \bar{\mu}_v \bar{\theta} \;\, \lambda q_0 \bar{\mu}_v \theta \,\right), \boldsymbol{B}_{1,0} = \begin{pmatrix} (1-\lambda q_0)\,\mu_v \\ (1-\lambda q_1)\,\mu_b \end{pmatrix},$$

$$\boldsymbol{A}_0 = \begin{pmatrix} \lambda q_0 \bar{\mu}_v \bar{\theta} & \lambda q_0 \bar{\mu}_v \theta \\ 0 & \lambda q_1 \bar{\mu}_b \end{pmatrix},$$

$$\boldsymbol{A}_1 = \begin{pmatrix} (1-\lambda q_0)\bar{\mu}_v \bar{\theta} + \lambda q_0 \mu_v p \bar{\theta} & (1-\lambda q_0)\,\bar{\mu}_v \theta + \lambda q_0 \mu_v \left(p\bar{\theta} + \theta\right) \\ 0 & (1-\lambda q_1)\bar{\mu}_b + \lambda q_1 \mu_b \end{pmatrix},$$

$$\boldsymbol{A}_2 = \begin{pmatrix} (1-\lambda q_0)\mu_v p \bar{\theta} & (1-\lambda q_0)\mu_v\!\left(\theta + p\bar{\theta}\right) \\ 0 & (1-\lambda q_1)\mu_b \end{pmatrix}.$$

Moreover, because the one step transitions from a state are restricted to the same level or to the two adjacent levels, vector-valued DTMC $\{(N(t), Y(t)), t \geq 0\}$ is called a quasi birth and death chain. In order to have a stable system, the DTMC should be positive recurrent. The condition for the stability is represented by the following relationship (see [41,42]):

$$\boldsymbol{\pi} \boldsymbol{A}_0 \mathbf{e} < \boldsymbol{\pi} \boldsymbol{A}_2 \mathbf{e},$$

where $\mathbf{e}$ denotes a column vector of 1's of appropriate dimension and $\boldsymbol{\pi}$ is the invariant probability vector of matrix $\boldsymbol{A}$, i.e. $\boldsymbol{\pi}\boldsymbol{A} = \boldsymbol{\pi}$, $\boldsymbol{\pi}\mathbf{e} = 1$ and $\boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{A}_1 + \boldsymbol{A}_2$. In other words, the rate of moving down one level in the DTMC must exceed the rate of moving up one level. Since the vector $\boldsymbol{\pi}$ can be obtained explicitly, the stability condition of this queueing system is also simplified as follows:

$$\rho = \frac{\lambda q_1 \bar{\mu}_b}{(1-\lambda q_1)\,\mu_b} < 1.$$

Let $\tilde{\boldsymbol{P}}$, partitioned as $\tilde{\boldsymbol{P}} = \left(\tilde{P}_{0,0}, \tilde{\boldsymbol{P}}_1, \tilde{\boldsymbol{P}}_2, \ldots\right)$ and $\tilde{\boldsymbol{P}}_i = \left(\tilde{P}_{i,0}, \tilde{P}_{i,1}\right)$, denote the steady state probability vector of $\mathbb{P}$. That is to say, $\tilde{\boldsymbol{P}}$ satisfies $\tilde{\boldsymbol{P}}\left(\mathbb{P} - \boldsymbol{I}\right) = \mathbf{0}$ and $\tilde{\boldsymbol{P}}\mathbf{e} = 1$, where $\boldsymbol{I}$ and $\mathbf{0}$ are identity matrix and zero column vector respectively. According to the matrix geometric method, we see that under the stability condition, $\tilde{\boldsymbol{P}}_i$ is obtained as

$$\tilde{\boldsymbol{P}}_k = \left(\tilde{P}_{k,0}, \tilde{P}_{k,1}\right) = \tilde{\boldsymbol{P}}_1 \boldsymbol{R}^{k-1} = \left(\tilde{P}_{1,0}, \tilde{P}_{1,1}\right) \boldsymbol{R}^{k-1}, \quad k = 1, 2, \ldots. \tag{30}$$

where $\boldsymbol{R}$ is the minimal nonnegative solution to the matrix-quadratic equation (see [42]). Clearly, in order to obtain the stationary probability vector $\tilde{\boldsymbol{P}}$, one should thus determine the rate matrix $\boldsymbol{R}$. In most applications, $\boldsymbol{R}$ needs to be computed by using an iterative algorithm. However, in our current model, the rate matrix $\boldsymbol{R}$ can be determined explicitly. Based on the structures of matrices, $\boldsymbol{A}_0$, $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ which are upper triangular matrices, the matrix solution $\boldsymbol{R}$ is also an upper triangular matrix. So we assume that

$$\boldsymbol{R} = \begin{pmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{pmatrix}.$$

Substituting $\boldsymbol{R}$ into the matrix-quadratic equation and after some algebraic manipulation, the explicit expression for rate matrix $\boldsymbol{R}$ is given by

$$\boldsymbol{R} = \begin{pmatrix} r & \frac{\Delta}{\mu_b(1-\lambda q_1)(1-r)} \\ 0 & \rho \end{pmatrix},$$

and

$$\boldsymbol{R}^k = \begin{pmatrix} r^k & \frac{\Delta}{\mu_b(1-\lambda q_1)(1-r)} \sum\limits_{j=0}^{k-1} r^j \rho^{k-1-j} \\ 0 & \rho^k \end{pmatrix}, \quad k = 1, 2, \ldots,$$

where

$$r = \frac{[\xi + \lambda q_0 \bar{\mu}_v + (1-\lambda q_0)\mu_v \bar{p}] - \sqrt{[\xi + \lambda q_0 \bar{\mu}_v + (1-\lambda q_0)\mu_v \bar{p}]^2 - 4\,(1-\lambda q_0)\mu_v \bar{\mu}_v \bar{p} \lambda q_0}}{2\,(1-\lambda q_0)\,\mu_v \bar{p}},$$

$$\xi = \frac{\theta + \bar{\theta}\mu_v p}{\bar{\theta}},$$

$$\Delta = \bar{p}^{-1}\big(\theta + p\bar{\theta}\big)\{r\xi + r[\lambda q_0 \bar{\mu}_v + (1-\lambda q_0)\mu_v \bar{p}] - \lambda q_0 \bar{\mu}_v\} + r\big[\lambda q_0 \mu_v\big(\theta + p\bar{\theta}\big) + (1-\lambda q_0)\bar{\mu}_v \theta\big] + \lambda q_0 \bar{\mu}_v \theta.$$

To start the recursive relation Eq.(30), the following boundary equation must be solved to obtain $\left(\tilde{P}_{0,0}, \tilde{P}_{1,0}, \tilde{P}_{1,1}\right)$ under the usual normalization condition $\tilde{P}_{0,0} + \sum\limits_{k=1}^{\infty} \tilde{P}_{k,0} + \sum\limits_{k=1}^{\infty} \tilde{P}_{k,1} = 1$, namely

$$\left(\tilde{P}_{0,0}, \tilde{P}_{1,0}, \tilde{P}_{1,1}\right) \boldsymbol{B}\,[\boldsymbol{R}] = \left(\tilde{P}_{0,0}, \tilde{P}_{1,0}, \tilde{P}_{1,1}\right) \begin{pmatrix} \boldsymbol{A}_{0,0} & \boldsymbol{A}_{0,1} \\ \boldsymbol{B}_{1,0} & \boldsymbol{R}\boldsymbol{A}_2 + \boldsymbol{A}_1 \end{pmatrix} = \left(\tilde{P}_{0,0}, \tilde{P}_{1,0}, \tilde{P}_{1,1}\right),$$

where

$$\boldsymbol{B}[\boldsymbol{R}] = \begin{pmatrix} \lambda q_0 \mu_v + \bar{\lambda} + \lambda \bar{q}_0 & \lambda q_0 \bar{\mu}_v \bar{\theta} & \lambda q_0 \bar{\mu}_v \theta \\ (1-\lambda q_0)\mu_v & \bar{\theta}[1-\mu_v(1-\lambda q_0 \bar{p})] - \frac{\xi\bar{\theta}r}{1-r} & \theta\big[1-\mu_v(1-\lambda q_0 \bar{p}) + \mu_v \lambda q_0 \frac{p}{\theta}\big] + \frac{\xi\bar{\theta}r}{1-r} \\ (1-\lambda q_1)\,\mu_b & 0 & 1-(1-\lambda q_1)\mu_b \end{pmatrix}.$$

Taking $\tilde{P}_{0,0}$ as constant, from above equation, $\tilde{P}_{1,0}$ and $\tilde{P}_{1,1}$ can be conveniently expressed in terms of $\tilde{P}_{0,0}$,

$$\tilde{P}_{1,0} = \frac{\lambda q_0 \bar{\mu}_v \bar{\theta}\,(1-r)}{\theta + \bar{\theta}\mu_v(1-\lambda q_0 \bar{p})(1-r) + \bar{\theta}\mu_v p r}\tilde{P}_{0,0}, \tag{31}$$

$$\tilde{P}_{1,1} = \frac{\lambda q_0 \bar{\mu}_v}{(1-\lambda q_1)\mu_b} \frac{\theta + \bar{\theta}\mu_v p\,[\lambda q_0 + (1-\lambda q_0)r]}{\theta + \bar{\theta}\mu_v(1-\lambda q_0 \bar{p})(1-r) + \bar{\theta}\mu_v p r}\tilde{P}_{0,0}. \tag{32}$$

Substituting Eqs.(31), (32) and $\boldsymbol{R}^{k-1}$ into Eq.(30), we have

$$\left(\tilde{P}_{k,0}, \tilde{P}_{k,1}\right) = \frac{\tilde{P}_{0,0}}{\theta + \bar{\theta}\mu_v\,(1-\lambda q_0 \bar{p})\,(1-r) + \bar{\theta}\mu_v p r} \times$$

$$\left(\lambda q_0\bar{\theta}\bar{\mu}_v(1-r)r^{k-1},\ \tfrac{\lambda q_0\bar{\mu}_v}{\mu_b(1-\lambda q_1)}\left\{\bar{\theta}\Delta\sum_{j=0}^{k-2}r^j\rho^{k-2-j}+\rho^{k-1}\big[\theta+\bar{\theta}\mu_v p[\lambda q_0+(1-\lambda q_0)r]\big]\right\}\right).$$

Employing the usual normalization condition and performing some algebraic manipulation, we get

$$\begin{cases} \tilde{P}_{0,0}=H\left[\theta+\bar{\theta}\mu_v(1-\lambda q_0\bar{p})(1-r)+\bar{\theta}\mu_v pr\right], \\ \tilde{P}_{k,0}=H\lambda q_0\bar{\theta}\bar{\mu}_v\,(1-r)\,r^{k-1},\ k=1,2,\ldots, \\ \tilde{P}_{k,1}=H\tfrac{\lambda q_0\bar{\mu}_v}{\mu_b(1-\lambda q_1)}\left\{\bar{\theta}\Delta\sum\limits_{j=0}^{k-2}r^j\rho^{k-2-j}+\rho^{k-1}\big[\theta+\bar{\theta}\mu_v p\,[\lambda q_0+(1-\lambda q_0)\,r]\big]\right\},\ k=1,2,\ldots, \end{cases} \tag{33}$$

where

$$H=\mu_b(1-\lambda q_1)(1-r)(1-\rho)\big\{\mu_b(1-\lambda q_1)(1-r)(1-\rho)\big[\theta+\bar{\theta}\mu_v(1-\lambda q_0\bar{p})(1-r)+\bar{\theta}\mu_v pr+\lambda q_0\bar{\theta}\bar{\mu}_v\big]$$
$$+\lambda q_0\bar{\theta}\bar{\mu}_v\Delta+\lambda q_0\bar{\mu}_v(1-r)\big[\theta+\bar{\theta}\mu_v p[\lambda q_0+(1-\lambda q_0)r]\big]\big\}^{-1}.$$

Let $\mathrm{E}\,[L_s|i]$ be the average number of customers in the system seen by an arrival, given that the server is at state $i(i=0,1)$. By the definition of conditional expectation, we have

$$\mathrm{E}\,[L_s|\,0]=\frac{\sum\limits_{k=1}^{\infty}k\tilde{P}_{k,0}}{\Pr\{\text{The server is on working vacation}\}}=\frac{\sum\limits_{k=1}^{\infty}k\tilde{P}_{k,0}}{\sum\limits_{k=0}^{\infty}\tilde{P}_{k,0}},$$

$$\mathrm{E}\,[L_s|\,1]=\frac{\sum\limits_{k=1}^{\infty}k\tilde{P}_{k,1}}{\Pr\{\text{The server is in regular busy period}\}}=\frac{\sum\limits_{k=1}^{\infty}k\tilde{P}_{k,1}}{\sum\limits_{k=1}^{\infty}\tilde{P}_{k,1}}.$$

Thus, the explicit expressions for $\mathrm{E}\,[L_s|\,0]$ and $\mathrm{E}\,[L_s|\,1]$ can be derived from Eq.(33)

$$\mathrm{E}\,[L_s|\,0]=\frac{\lambda q_0\bar{\theta}\bar{\mu}_v}{\big[\theta+\bar{\theta}\,(1-\lambda q_0\bar{p})(1-r)+\bar{\theta}\mu_v pr+\lambda q_0\bar{\theta}\bar{\mu}_v\big](1-r)}, \tag{34}$$

$$\mathrm{E}\,[L_s|\,1]=\frac{\bar{\theta}\Delta(2-\rho-r)+(1-r)^2\big[\theta+\bar{\theta}\mu_v p\,[\lambda q_0+(1-\lambda q_0)\,r]\big]}{(1-\rho)(1-r)\big\{\bar{\theta}\Delta+(1-r)\big[\theta+\bar{\theta}\mu_v p\,[\lambda q_0+(1-\lambda q_0)\,r]\big]\big\}}. \tag{35}$$

Then, from Eq.(35), the expected conditional sojourn time of a customer who arrives during regular busy period is given by

$$\mathrm{E}\,[W_1]=\frac{\bar{\theta}\Delta(2-\rho-r)+(1-r)^2\big[\theta+\bar{\theta}\mu_v p\,[\lambda q_0+(1-\lambda q_0)\,r]\big]}{\mu_b(1-\rho)(1-r)\big\{\bar{\theta}\Delta+(1-r)\big[\theta+\bar{\theta}\mu_v p\,[\lambda q_0+(1-\lambda q_0)\,r]\big]\big\}}+\frac{1}{\mu_b}. \tag{36}$$

Similarly, let W_0 denote the conditional sojourn time of a customer who arrives during working vacation period. In order to determine the expected value of W_0, we need to use Eq.(8) described in Subsection 3.1. Clearly,

$$\mathrm{E}\left[W_0\right] = \frac{\sum\limits_{k=0}^{\infty} \tilde{P}_{k,0}\mathrm{E}\left[W_{k,0}\right]}{\Pr\left\{\text{The server is on working vacation}\right\}}. \tag{37}$$

Substituting Eq.(8) into Eq.(37), $\mathrm{E}\left[W_0\right]$ can be expressed as

$$\mathrm{E}\left[W_0\right] = \frac{\mu_b+\bar{\mu}_v\theta}{\mu_b\left(1-\bar{\mu}_v\bar{\theta}\right)} + \frac{\lambda q_0\bar{\theta}\bar{\mu}_v}{\mu_b\left[\theta+\bar{\theta}\mu_v(1-\lambda q_0\bar{p})(1-r)+\bar{\theta}\mu_v pr+\lambda q_0\bar{\theta}\bar{\mu}_v\right]} \times \left[\frac{\xi\mu_v\bar{\theta}^2\bar{p}(\mu_b-\mu_v)}{\left(1-\bar{\mu}_v\bar{\theta}\right)\left(1-\bar{\mu}_v\bar{\theta}-\mu_v\bar{p}\bar{\theta}\right)\left(1-\bar{\mu}_v\bar{\theta}-\mu_v\bar{p}\bar{\theta}r\right)} + \frac{1}{1-r}\right]. \tag{38}$$

Let $U_{au}^{e}\left(0;q_0\right)$ and $U_{au}^{e}\left(1;q_0,q_1\right)$ be the utility function of a joining customer in state 0 and 1, respectively. With the notations introduced above, we can write

$$U_{au}^{e}\left(0;q_0\right) = R - C\mathrm{E}\left[W_0\right] \quad \text{and} \quad U_{au}^{e}\left(1;q_0,q_1\right) = R - C\mathrm{E}\left[W_1\right].$$

In light of the complexity of the expressions for $\mathrm{E}\left[W_0\right]$ and $\mathrm{E}\left[W_1\right]$, we only perform some numerical calculations to illustrate the existence and uniqueness of the mixed strategy for individual optimization. In all the numerical experiments considered below, several system parameters are taken as $\mu_v = 0.2$, $\mu_b = 0.6$, $p = 0.2$, $\lambda = 0.5$, and all the calculation results are reported here in the form of tables. By changing the values of the parameters R, C and θ, we demonstrate six possible cases for the mixed strategy (q_0^e, q_1^e) in Table 1. As can be seen from the table, there is an interesting phenomenon worthy of our attention. In working vacation queue, we always suppose that $\mu_v < \mu_b$. Such assumption usually causes us to mistakenly believe that q_0^e should be smaller than q_1^e. However, Table 1 shows that in some cases q_0^e is not definitely smaller than q_1^e (see Cases a, d and e in Table 1). Actually, if a tagged customer is given the information that the server is on working vacation, then he knows that he must go through a semi-dormant period (i.e. server provides service at a lower rate during vacation). On the other hand, he expects that few customers are ahead of him, because the mean vacation time is small and the vacation can be interrupted at a service completion instant in semi-dormant period. Cases a, d and e of Table 1 show that the first factor prevails, thus it is optimal for the tagged customer to enter. Furthermore, when the mean vacation time is moderate (Case b) or high (Case c), things are totally different.

Table 1. Individually optimal joining probabilities for different values of R, C and θ.

	q_0^e	q_1^e		q_0^e	q_1^e
Case a $(R = 20, C = 5.5, \theta = 0.3)$	0.3287	0	Case d $(R = 32, C = 9, \theta = 0.5)$	1	0
Case b $(R = 20, C = 3, \theta = 0.1)$	0.5489	0.8423	Case e $(R = 30, C = 3.8, \theta = 0.15)$	1	0.8837
Case c $(R = 35, C = 4, \theta = 0.015)$	0.3334	1	Case f $(R = 30, C = 3, \theta = 0.15)$	1	1

4.2. Mixed strategy for social optimization

On the other hand, employing Eq.(33), the unconditional expected number of customers in the system equals

$$\mathrm{E}\left[L_s\right]=\frac{H\lambda q_0\bar{\mu}_v}{(1-r)}\left\{\bar{\theta}+\frac{(2-\rho-r)\,\bar{\theta}\Delta+(1-r)^2\left[\theta+\bar{\theta}\mu_v p[\lambda q_0+(1-\lambda q_0)\,r]\right]}{\mu_b(1-\lambda q_1)(1-\rho)^2(1-r)}\right\}. \tag{39}$$

Applying Eq.(39), the social benefit per unit time for the joining strategy (q_0, q_1) can now be easily computed as

$$U_{au}^{s}\left(q_0,q_1\right)=\lambda_{\text{eff}}R-C\mathrm{E}\left[L_s\right], \tag{40}$$

where

$$\begin{aligned}\lambda_{\text{eff}}&=\lambda\left(q_0\sum_{k=0}^{\infty}\tilde{P}_{k,0}+q_1\sum_{k=1}^{\infty}\tilde{P}_{k,1}\right)\\&=\lambda q_0 H\Bigg[\theta+\bar{\theta}\mu_v(1-\lambda q_0\bar{p})(1-r)+\bar{\theta}\mu_v pr+\lambda q_0\bar{\theta}\bar{\mu}_v\\&\quad+\frac{\lambda q_1\bar{\mu}_v\left\{\bar{\theta}\Delta+\left[\theta+\bar{\theta}\mu_v p[\lambda q_0+(1-\lambda q_0)\,r]\right](1-r)\right\}}{\mu_b(1-\lambda q_1)(1-\rho)(1-r)}\Bigg].\end{aligned} \tag{41}$$

Here, our main purpose is to find the optimal mixed strategy, denoted by (q_0^*, q_1^*), that will yield the system's maximal expected profit, to reach the so-called social (or overall) optimization. However, when we substitute Eqs.(39) and (41) into the right hand side of Eq.(40), we may see that the form of the function $U_{au}^{s}\left(q_0,q_1\right)$ becomes too complicated. Thus, trying to get the optimal values of q_0 and q_1 analytically would have been an extremely difficult task. In spite of that, we can still obtain the optimal solution to meet the precise requirements by steepest descent algorithm. To demonstrate the validity and effectiveness of the steepest descent method in our optimization problem, a numerical example is provided by considering the following system parameters:

$$R=30, C=1, \theta=0.01, \mu_v=0.1, \mu_b=0.8, p=0.05, \lambda=0.79.$$

Further, with the same parameter settings, Figure 6 compares the mixed strategies for individual and social optimization. From the numerical results, we may reveal that for the almost unobservable working vacation queue the joining probability for social optimization is not always lower than the one for individual optimization. Thus, the positive externalities that a joining customer brings to the system still exist in such case. These results are entirely consistent with the conclusions of Subsection 3.2.

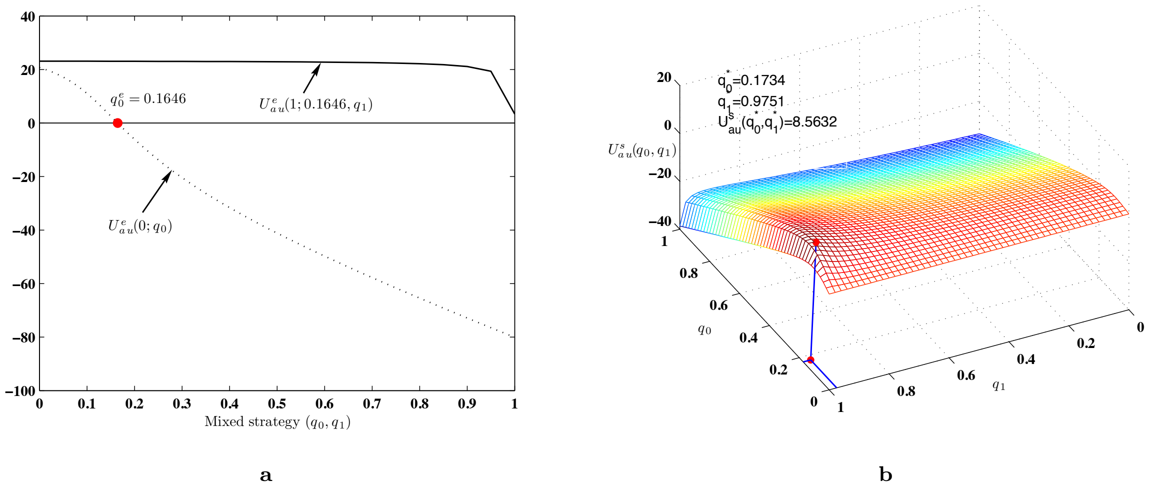

Figure 6. Comparisons of individually and socially optimal mixed strategies under the partially observable case.

5. Analysis of the fully unobservable queues

Finally, employing the results in the previous section, we analyze the customer's optimal response in the case of unobservable, where arriving customers do not observe the state of the server and the actual length of the queue, but system parameters are common knowledge to all arriving customers. Applying these parameters, customers are able to postulate the average sojourn time and waiting cost. Based on these limited available information, they may choose to balk or join with an appropriate probability. In other words, their decision is equivalent to select a joining probability $q(0 \le q \le 1)$ that satisfies some specified conditions. Thus, our main purpose in this section is to determine the joining probability q, so as to optimize the individual and social utilization. Toward this end, we first derive the expected unconditional sojourn time of a joining customer, denoted by $\mathrm{E}\left[W(q)\right]$. Replacing q_0 and q_1 by q in Eqs.(33), (36), (38), and using the following relationship

$$\lambda q\bar{\mu}_v - (1-\lambda q)\mu_v\bar{p}r' = \frac{\xi r'}{(1-r')},$$

we have

$$\mathrm{E}\left[W(q)\right] = \frac{1-\lambda q}{\mu_b - \lambda q} + H^* \frac{\lambda q\bar{\mu}_v\bar{\theta}(\mu_b - \mu_v)(1-r')}{r'} \left\{ \frac{(1-\lambda q)(1-\rho')}{1-\bar{\mu}_v\bar{\theta} - \mu_v\bar{p}\bar{\theta}r'} + \frac{\lambda q}{\mu_b} + \frac{r'}{\mu_b(1-r')} \right\}, \quad (42)$$

where

$$\rho' = \frac{\lambda q\bar{\mu}_b}{(1-\lambda q)\,\mu_b},$$

$$r' = \frac{[\xi + \lambda q\bar{\mu}_v + (1-\lambda q)\mu_v\bar{p}] - \sqrt{[\xi + \lambda q\bar{\mu}_v + (1-\lambda q)\mu_v\bar{p}]^2 - 4\,(1-\lambda q)\mu_v\bar{\mu}_v\bar{p}\lambda q}}{2\,(1-\lambda q)\,\mu_v\bar{p}},$$

$$
\begin{aligned}
H^{*}=&\big\{\mu_b(1-\lambda q)(1-r')(1-\rho')\big[\theta+\bar{\theta}\mu_v(1-\lambda q\bar{p})(1-r')+\bar{\theta}\mu_v p r'+\lambda q_0\bar{\theta}\bar{\mu}_v\big]+\lambda q\bar{\theta}\bar{\mu}_v\Delta' \\
&+\lambda q\bar{\mu}_v(1-r')\big[\theta+\bar{\theta}\mu_v p[\lambda q+(1-\lambda q)r']\big]\big\}^{-1}, \\
\Delta'=&\,\bar{p}^{-1}\big(\theta+p\bar{\theta}\big)\{r'\xi+r'[\lambda q\bar{\mu}_v+(1-\lambda q)\mu_v\bar{p}]-\lambda q\bar{\mu}_v\}+r'\big[\lambda q\mu_v\big(\theta+p\bar{\theta}\big)+(1-\lambda q)\bar{\mu}_v\theta\big]+\lambda q\bar{\mu}_v\theta.
\end{aligned}
$$

Actually, Eq.(42) gives the stochastic decomposition structure of the expected sojourn time which indicates the relationship with that of $Geo/Geo/1$ queue without working vacations. Obviously, the first term on the right hand side of Eq.(42) is the sojourn time of a joining customer in a corresponding classic $Geo/Geo/1$ queue.

Once the expected sojourn time is calculated, the utility functions for individual customer and whole system can be established, respectively. Further, the optimal values of the decision variable can also be found by numerical analysis of utility functions.

5.1. Mixed strategy for individual optimization

With a mixed strategy, an arriving customer joins the waiting line with probability q and the equilibrium value of q, denoted by q^e, is such that if all customers follow it all are indifferent. Here, it is necessary to point out when we evaluate the customers' strategic response in equilibrium, we should guarantee that an equilibrium strategy can make the queueing system stable. That is to say, we should limit our search for equilibrium strategies in the interval $[0, \mu_b/\lambda]\cap[0,1]$. Employing $\mathrm{E}\left[W(q)\right]$, the expected net gain for a customer entering the system is given by

$$
U_{fu}^{e}=R-C\left[\frac{1-\lambda q}{\mu_b-\lambda q}+H^{*}\frac{\lambda q\bar{\mu}_v\bar{\theta}(\mu_b-\mu_v)(1-r')}{r'}\left(\frac{(1-\lambda q)(1-\rho')}{1-\bar{\mu}_v\bar{\theta}-\mu_v\bar{p}\bar{\theta}r'}+\frac{\lambda q}{\mu_b}+\frac{r'}{\mu_b(1-r')}\right)\right].
$$

Since the explicit expression of $\mathrm{E}\left[W(q)\right]$ is complicated enough, by varying the value of parameter R, we numerically demonstrate some possible cases for the mixed strategy q^e in Figure 7. Here, we take the system parameters as $C=1, \theta=0.1, \mu_v=0.1, \mu_b=0.5, p=0.5$ and $\lambda=0.45$. Figure 7 shows the behavior of U_{fu}^{e} with respect to the joining probability q for different values of reward R. The following conclusions can be drawn from the graph:

(i) U_{fu}^{e} decreases with the increase in the values of q;

(ii) If $R\in\left(\dfrac{C(\mu_b+\bar{\mu}_v\theta)}{\mu_b\big(1-\bar{\mu}_v\bar{\theta}\big)}, C\mathrm{E}\left[W(q)\right]_{q=1}\right)$, namely R belongs to $(6.2105, 17.0726)$, there exists a point $\tilde{q}$ in the interval $(0,1)$ such that $U_{fu}^{e}=0$. For example, when $R=12$, the value of U_{fu}^{e} is positive or zero or negative according as q less than or equal to or greater than 0.8897 and hence $\tilde{q}=q^e=0.8897$;

(iii) If $R\in(17.0726,+\infty)$, U_{fu}^{e} is always positive for any $q\in(0,1)$. This indicates that customer will be profitable as long as he enters the system and receives the service. Therefore, the optimal response for individual customer is to join the queue with probability one, namely $q^e=1$.

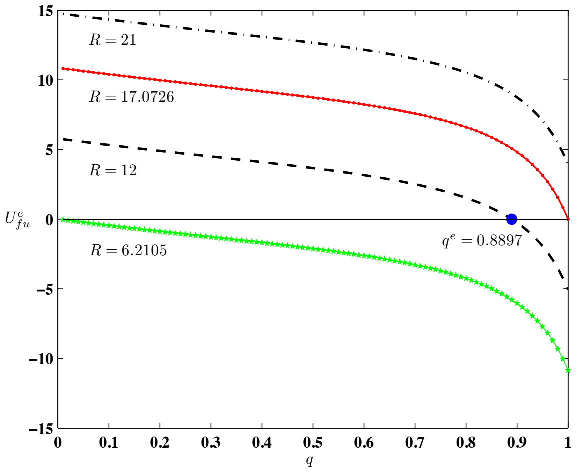

Figure 7. U^e_{fu} versus joining probability q for different values of reward R.

Moreover, giving $R = 10$ and letting p vary from 0.1 to 0.9, we present some numerical results in Table 2 for comparison of individual customer's optimal joining rules in partially observable and fully unobservable cases. As is to be expected q^e_0 and q^e are increasing functions of p when all other parameters are fixed. However, as p increases, q^e_1 shows a decreasing trend. Furthermore, from Table 2 we also see that the rate of increase of q^e becomes smaller for higher values of p, and q^e is always locating between q^e_0 and q^e_1. This means that if customers are not informed about the state of the server, they will adopt an intermediate strategy to decide whether or not to join the queue.

Table 2. Individual customer's optimal joining probabilities for different values of p.

p	0.1	0.3	0.5	0.7	0.9
(q^e_0, q^e_1)	(0.4690, 0.8649)	(0.5600, 0.8476)	(0.6657, 0.8274)	(0.7851, 0.8045)	(0.9172, 0.7788)
q^e	0.6158	0.6998	0.7578	0.7972	0.8249

5.2. Mixed strategy for social optimization

On the other hand, from the system administrator's perspective, the social benefit per unit time for the joining strategy q can now be easily computed as

$$U^s_{fu} = \lambda q \left\{ R - C \left[\frac{1-\lambda q}{\mu_b - \lambda q} + H^* \frac{\lambda q \bar{\mu}_v \bar{\theta}(\mu_b - \mu_v)(1-r')}{r'} \left(\frac{(1-\lambda q)(1-\rho')}{1-\bar{\mu}_v\bar{\theta} - \mu_v \bar{p}\bar{\theta} r'} + \frac{\lambda q}{\mu_b} + \frac{r'}{\mu_b(1-r')} \right) \right] \right\}.$$

Since customers tend to lack self-control, higher access probability for individual interests can cause queue congestion so that the expected waiting time of future customers will severely increase, and eventually lead to a decline in net profit for the whole system. Hence, we discuss how to achieve the maximal system's expected profit by controlling the probability of joining the queue.

We note that showing concavity of the utility function U^s_{fu} and computation of its derivatives are non-trivial tasks. For the reasons mentioned above, we employ quadratic fit

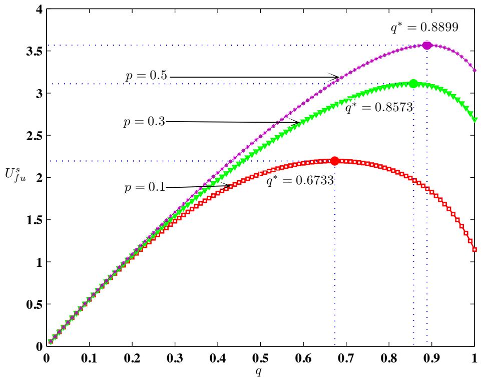

Figure 8. U_{fu}^s versus q for varying p.

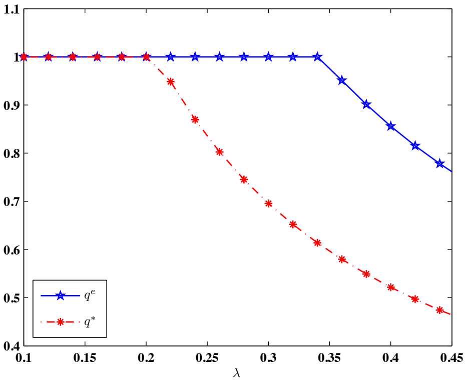

Figure 9. Optimal joining probability versus λ in fully unobservable case.

line search method to obtain the joining probability q^* for social optimization. In Figure 8, we display the graph of U_{fu}^s as a function of q for various values of p by fixing $R = 17$, $C = 0.7$, $\theta = 0.025$, $\mu_v = 0.2$, $\mu_b = 0.5$ and $\lambda = 0.42$. From Figure 8 we observe that for fixed p, U_{fu}^s increases initially and then decreases as q increases. That is to say, beyond a certain value of q, any further increase in its value will only result in the system's overall profits to decline. So, this is the fundamental reason why we implement access control to reject some customers to join the queue. Taking $R = 10$, $C = 1.5$, $\theta = 0.1$, $\mu_v = 0.25$ and $\mu_b = 0.5$, in Figure 9, the joining probabilities for individual and social optimization are plotted against λ for $p = 0.7$. We could visually observe that the socially optimal mixed strategy q^* is less than or equal to the individually mixed strategy q^e. This indicates that if every customer acts selfishly, the system may be over-congested and is impossible to achieve maximum social welfare. Next, we also conduct numerical comparisons between the socially optimal joining rules for partially observable and fully unobservable queues. In Table 3, the computation results of the optimal mixed strategies are given for various vacation interruption probabilities with the same system parameters $R = 17$, $C = 0.7$, $\theta = 0.025$, $\mu_v = 0.2$, $\mu_b = 0.5$ and $\lambda = 0.42$. We reveal that when the state of the server is not communicated to the customer upon his arrival, q^* is still inside the interval formed by two socially optimal joining probabilities for partially observable case. Furthermore, it is worth noting that as vacation interruption probability increases up to a certain value, q_0^* will always be one. Thus, there is no need to take a large value of p for increasing the socially optimal joining probability in a working vacation period. Since larger value of p often means frequent switching of service rate, the more the server switches its service rate, the more additional cost it has to face.

Table 3. Socially optimal joining probabilities for different values of p.

p	0.1	0.3	0.5	0.7	0.9
(q_0^*, q_1^*)	(0.5471, 0.9757)	(0.7223, 0.9350)	(0.8893, 0.9021)	(1, 0.8731)	(1, 0.8462)
q^*	0.6733	0.8573	0.8899	0.9046	0.9126

6. Conclusions

From an economic point of view, the customer's individually and socially optimal joining rules in a discrete time working vacation queue have been extensively analyzed in this paper. Unlike classic multiple and single working vacation queues, Bernoulli vacation interruption schedule has been incorporated into our current model. Such mechanism makes the server state transitions have greater flexibility and can potentially capture the correlations between service time and vacation duration. The analysis of the model is performed under three different information levels. For the full information case, it is shown that the symmetric equilibrium strategies are determined by two balking thresholds $I(0)$ and $I(1)$. These are established by deriving iterative relationships on the generating function and the expected value of the conditional waiting time. As for the social optimization model, which is notably more difficult, it is assumed that the central planner employs a similar threshold policy and the optimal threshold values are also computed by analyzing the steady state distribution of the resulting Markov process under the threshold policy. Also, we compare the equilibrium and optimal threshold numerically. During regular service the optimal threshold is always lower than that under equilibrium, which implies that the joining customer impose negative externalities, while during working vacation period the situation may be reversed, and the central planner may need to subsidize the arriving customers and encourage them to join the queue. This also means that Naor's classic conclusion is not necessarily true for discrete time working vacation queue. In addition, by developing the expressions for the waiting function which allow for numerically deriving the equilibrium and socially optimal strategy, we further find that the same situation also occurs in almost unobservable case. That is to say, the joining probability for social optimization is not always lower than the one for individual optimization. By entering in a vacation state, especially when the expected length of the vacation time is relatively long, joining customers can create more additional service completion epochs and increase the probability of a vacation interruption to reduce the delay of other present or future customers. Thus, in some cases, the externalities that a joining customer brings to the system are positive. Therefore, it is unclear whether the social planner wants a tax to discourage arrivals or a subsidy to encourage arrivals. Furthermore, it would be a good topic to analyze the problem discussed in this paper where the inter-arrival time is generally distributed. Naturally, with the inclusion of generally distributed inter-arrival time, the problem will be more challenging.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper. This research was partially supported by grant from NSERC DAS programs, National Natural Science Foundation of China (Nos.71301111, 71571127, 71402072) and the FSUSE (No.2012RC23).

References

[1] Naor P. The regulation of queue size by levying tolls. Econometrica 1969; 37: 15-24.

[2] Yechiali U. On optimal balking rules and toll charges in the $GI/M/1$ queueing process. Operations Research 1971; 19: 349-370.

[3] Yechiali U. Customers' Optimal joining rules for the $GI/M/s$ queue. Management Science 1972; 18: 434-443.

[4] Stidham S Jr. Optimal control of admissions to a queueing system. IEEE Transactions on Automatic Control 1985; 30: 705-713.

[5] Mendelson H, Whang S. Optimal incentive-compatible priority pricing for the $M/M/1$ queue. Operations Research 1990; 38: 870-883.

[6] Edelson N, Hildebrand K. Congestion tolls for Poisson queueing processes. Econometrica 1975; 43: 81-92.

[7] Burnetas A, Economou A. Equilibrium customer strategies in a single server Markovian queue with setup times. Queueing Systems 2007; 56: 213-228.

[8] Economou A, Kanta S. Equilibrium balking strategies in the observable single-server queue with breakdowns and repairs. Operations Research Letters 2008; 36: 696-699.

[9] Wang J, Zhang F. Equilibrium analysis of the observable queues with balking and delayed repairs. Applied Mathematics and Computation 2011; 218: 2716-2729.

[10] Sun W, Guo P, Tian N. Equilibrium threshold strategies in observable queueing systems with setup/closedown times. Central European Journal of Operational Research 2010; 18: 241-268.

[11] Economou A, Manou A. Equilibrium balking strategies for a clearing queueing system in alternating environment, Annals of Operations Research, 2013, 208, 489-514.

[12] Guo P, Hassin R. Strategic behavior and social optimization in Markovian vacation queues. Operations Research 2011; 59: 986-997.

[13] Guo P, Hassin R. Strategic behavior and social optimization in Markovian vacation queues: The case of heterogeneous customers. European Journal of Operational Research 2012; 222: 278-286.

[14] Economou A, Gómez-Corral A, Kanta S. Optimal balking strategies in single-server queues with general service and vacation times. Performance Evaluation 2011; 68: 967-982.

[15] Liu W, Ma Y, Li J. Equilibrium threshold strategies in observable queueing systems under single vacation policy. Applied Mathematical Modelling 2012; 36: 6186-6202.

[16] Ma Y, Liu W, Li J. Equilibrium balking behavior in the $Geo/Geo/1$ queueing system with multiple vacations. Applied Mathematical Modelling 2013; 37: 3861-3878.

[17] Sun W, Li S. Equilibrium and optimal behavior of customers in Markovian queues with multiple working vacations. TOP 2014; 22: 694-715.

[18] Zhang F, Wang J, Liu, B. Equilibrium balking strategies in Markovian queues with working vacations. Applied Mathematical Modelling 2013; 37: 8264-8282.

[19] Economou A, Kanta S. Optimal balking strategies and pricing for the single server Markovian queue with compartmented waiting space. Queueing Systems 2008; 59: 237-269.

[20] Boudali O, Economou A. Optimal and equilibrium balking strategies in the single server Markovian queue with catastrophes. European Journal of Operational Research 2012; 218: 708-715.

[21] Boudali O, Economou A. The effect of catastrophes on the strategic customer behavior in queueing systems, Naval Research Logistics 2013; 60: 571-587.

[22] Wang J, Zhang F. Strategic joining in $M/M/1$ retrial queues. European Journal of Operational Research 2013; 240: 76-87.

[23] Li L, Wang J, Zhang F. Equilibrium customer strategies in Markovian queues with partial breakdowns. Computers & Industrial Engineering 2013; 66: 751-757.

[24] Servi L, Finn S. $M/M/1$ queues with working vacations ($M/M/1/WV$). Performance Evaluation 2002; 50: 41-52.

[25] Baba Y. Analysis of a $GI/M/1$ queue with multiple working vacations. Operations Research Letters 2005; 33: 201-209.

[26] Wu D, Takagi H. $M/G/1$ queue with multiple working vacations. Performance Evaluation 2006; 63: 654-681.

[27] Li J, Tian N, Liu W. Discrete-time $GI/Geo/1$ queue with multiple working vacations. Queueing Systems 2007; 56: 53-63.

[28] Li J, Tian N. The discrete time $GI/Geo/1$ queue with working vacations and vacation interruption. Applied Mathematics and Computation 2007; 185: 1-10.

[29] Tian N, Ma Z, Liu M. The discrete-time $Geom/Geom/1$ queue with multiple working vacations. Applied Mathematical Modelling 2008; 32: 2941-2953.

[30] Chae K, Lim D, Yang W. The $GI/M/1$ queue and the $GI/Geo/1$ queue both with single working vacation. Performance Evaluation 2009; 66: 356-367.

[31] Goswami C, Selvaraju N. The discrete-time $MAP/PH/1$ queue with multiple working vacations. Applied Mathematical Modelling 2010; 34: 931-946.

[32] Yu M, Tang Y, Fu Y, Pan L. $GI/Geom/1/N/MWV$ queue with changeover time and searching for the optimum service rate in working vacation period. Journal of Computational and Applied Mathematics 2011; 235: 2170-2184.

[33] Gao S, Wang J, Zhang D. Discrete-time $GI^X/Geo/1/N$ queue with negative customers and multiple working vacations. Journal of the Korean Statistical Society 2013; 42: 515-528.

[34] Gao S, Liu Z. An $M/G/1$ queue with single working vacation and vacation interruption under Bernoulli schedule. Applied Mathematical Modelling 2013; 37: 1564-1579.

[35] Gao S, Wang J. Discrete-time $Geo^X/G/1$ retrial queue with general retrial times, working vacations and vacation interruption. Quality Technology & Quantitative Management 2013; 10: 493-510.

[36] Gao S, Liu Z, Du Q. Discrete-time $GI^X/Geo/1/N$ queue with working vacations and vacation interruption. Asia-Pacific Journal of Operational Research 2014; 31: 1450003.

[37] Li T, Zhang L, Xu X, Gao S. The $GI/Geo/1$ queue with Bernoulli-schedule-controlled vacation and vacation interruption. Computers & Operations Research 2013; 40: 1680-1692.

[38] Hunter J. Mathematical techniques of applied probability, discrete time models: techniques and applications. Vol. 2. New York: Academic Press; 1983.

[39] Takagi H. Queueing Analysis: A Foundation of Performance Evaluation. vol. 3. Amsterdam: North-Holland; 1993.

[40] Elaydi S. An Introduction to Difference Equations. Third ed. New York: Springer-Verlag; 2005.

[41] Neuts M. Matrix-Geometric Solutions in Stochastic Models. Baltimore: Johns Hopkins University Press; 1981.

[42] Alfa A. S. Queueing Theory for Telecommunications. New York: Springer-Verlag; 2010.