EQUILIBRIUM AND OPTIMAL BALKING STRATEGIES FOR LOW-PRIORITY CUSTOMERS IN THE M/G/1 QUEUE WITH TWO CLASSES OF CUSTOMERS AND PREEMPTIVE PRIORITY

BIAO XU
School of Economics and Management, Beihang University
Beijing 100191, China

XIULI XU
College of Science, Yanshan University
Hebei, Qinhuangdao 066004, China

ZHONG YAO
School of Economics and Management, Beihang University
Beijing 100191, China
Institute of Economics and Business, Beihang University
Beijing 100191, China

(Communicated by Yutaka Takahashi)

Abstract. This paper investigates the low-priority customers strategic behavior in the single-server queueing system with general service time and two customer types. The priority system is preemptive resume, which means that if a high-priority customer enters the system that are serving a low-priority customer, the arriving customer preempts the service facility and the preempted customer returns to the head of the queue for his own class. The customer who is preempted resumes service at the point of interruption upon reentering the system. The low-priority customers dilemma is whether to join or balk based on a linear reward-cost structure. Two cases are distinguished based on the different levels of information that the low-priority customers acquire before joining the system. The equilibrium threshold strategy in the observable case and the equilibrium balking strategy as well as the socially optimal balking strategy in the unobservable case for the low-priority customers are derived finally.

1. Introduction. Due to the wide applications in service system, production management, electronic commerce and so forth, a growing number of papers that study customers strategic behavior in queueing models from an economic viewpoint have emerged. In such models, customers have the rights to make decisions about their own behaviors based on the available information and a natural reward-cost structure. Traditionally, queueing systems are mainly classified into two categories: the observable model and the unobservable model, based on whether the information of

2010 Mathematics Subject Classification. Primary: 60K25; Secondary: 90B22.
Key words and phrases. M/G/1 queue, two classes, equilibrium, optimization, priority, preemption.
Corresponding author: Xiuli Xu.
the system is acquired by customers upon their arrivals. The most influential early study on the observable queueing system can went back to Naor [22], which investigated the equilibrium and social optimal strategies in an M/M/1 queueing model with a linear reward-cost structure. On the other hand, the pioneering literature on the unobservable queue was firstly provided by Edelson and Hildebrand [9] which reexamined the same model as Naor [22] and examined the basic properties of an unobservable M/M/1 queueing system. Recently, many authors had further studied this fundamental model of Naor [22] (i.e., Boudali and Economou [4], Li, Wang and Zhang [19], Zhang, Wang and Liu [31], Chen and Zhou [7], Sun and Li [26], Xu and Xu [29], Takagi [27]). Our work considers the strategic decisions of customers in an M/G/1 queue, and thus contributes to the literature on the non-Markovian queueing systems. There were also several studies investigating customers behaviors in variants of the M/G/1 queue in the literature. For example, Altman and Hassin [1] argued that no equilibrium threshold strategy existed in an observable M/G/1 queue. The work of Haviv and Kerner [14] partially extended the result of Altman and Hassin [1] and presented the existence of a Nash equilibrium policy by assuming that customers could observe the server state upon their arrivals. Kerner [16] analyzed the (possibly mixed) Nash equilibrium strategy in an observable M/G/1 queue where queue length could be observed. Economou, Gmez-Corral and Kanta [8] studied the optimal balking strategies in M/G/1 queues with vacation times. Recently, Zhang, Wang and Liu [32] analyzed the equilibrium joining strategy in an M/G/1 queue with general setup times. Tian, Yue and Yue [28] examined customers strategic behavior in an M/G/1 queue under N-policy, where the server was removable. Hassin [13] offered a comprehensive summary and classification of the fundamental methods and theoretical conclusions in the field of strategy study in economics of queues.

The multiclass queueing models are often met in some service systems where different priorities distinguish different classes. The strategic behavior of customers in service systems with multiclass customers and priorities had been widely studied in the last decades. The first work was presented by Lee and Park [18], which proposed an efficient management strategy in the queueing system with (n+1) class calls. Afterwards, Altmann et al. [2] numerically analyzed the problem of priority purchase in a service model with two classes and Poisson arrivals of homogeneous customers. Mandjes [21] examined the performance of a Markovian queue with two user types and nonlinear waiting costs. Kim and Mannino [17] studied the profit maximization in an M/G/1 queue with non-preemptive priorities and N different classes. Berg, Mandjes and Núñez-Queija [3] dealt with the revenue maximization in an observable M/G/1 queueing system with two classes and two optional priorities. Printezis and Burnetastehe [23] analyzed the optimal option pricing policy in an M/M/m queue, where some priority options were provided for customers. Gilland and Warsing [12] investigated the equilibrium strategy in a Markovian queue with two priority classes. Hsu, Xu, and Jukic [15] proposed an economic analysis for a service model with multiple customer types and priorities. Sun, Guo and Tian [25] considered several queueing models with a single server, two classes of customers and different reward-cost structures. Erlichman and Hassin [10] studied an observable M/M/1 queue in which an arriving customer was allowed to purchase priority and overtake the present customers in front of him. Gavirneni and Kulkarni [11] examined a two-class M/G/1 queue with non-preemptive priorities and the Burr type waiting-cost rates.
Our work inspects preemption-based prioritization between the two user types. This model has many broad applications in practical service systems. For instance, in cognitive radio, to promote the efficient utilization of the radio electromagnetic spectrum, the primary user is able to preempt a spectrum hole occupied by the secondary user, when he requires the resource. In the Multi-Protocol Label Switching (MPLS) capable networks, a higher-priority Label Switched Path (LSP) can preempt a lower-priority LSP from a given path, when there is a competition for a finite number of resources. GSM (Global System for Mobile communications) voice calls are allowed to preempt GPRS (General Packet Radio Service) data packets when they are transmitted in cellular networks. In wireless medium access control (MAC) protocols, transmissions of voice and media traffic have preemptive access priority over transmissions of web and email traffic. Several scholars had devoted oneself to the economic analysis of the two-class preemptive priority queue. Brouns and Wal [5] studied the optimal threshold policies for two essential online decisions in a two-class $M^{1,2}|M|1$ queue with preemptive priority. Chen and Kulkarni [6] considered an $M/M/1$ queue with two customer classes and preemptive-resume priority. Liu and Berry [20] analyzed the price competition and social welfare in an unobservable priority $M/G/1$ queue with two classes and preemptive priorities. Xu, Xu and Wang [30] researched the equilibrium and socially optimal balking behavior of the high-priority customers in the unobservable $M/G/1$ queue, in which the equilibrium and socially optimal balking strategies for the high-priority customers are presented.

Based on the prior literatures, we examine the low-priority customers strategic behavior in $M/G/1$ queue serving two customer classes with preemption policy. Certain properties of this model (e.g., two classes, general service time, preemptive-resume priority) may suggest an overlap with some previous jobs in the literature (i.e., Xu, Xu and Wang [30]). However, our work conducts a novel study, and the theoretical and numerical results obtained in our study have also some important managerial insights. Xu, Xu and Wang [30] mainly focused on the high-priority customers strategic behavior in the unobservable case, whereas we investigate the optimal policy of low-priority customers in the observable case as well as the unobservable case. The closed analytic forms of the optimal joining thresholds and the optimal joining probabilities are both derived, while the social optimal joining probability is only numerically approximate in Xu, Xu and Wang [30]. In addition, as will be seen in Remark 2, several novel conclusions are proposed.

The paper is organized as follows. In Section 2, we describe the dynamics of the queueing model and the reword-cost structure. Section 3 is mainly focus on the observable queue. The equilibrium threshold strategies for the low-priority customers are presented in this section. Section 4 deals with the unobservable queue. The equilibrium strategies as well as the socially optimal strategies for the low-priority customers are both derived in this section. In Section 5, some numerical results are given to illustrate the effects of several parameters on the customers strategies. Finally, some conclusions are summarized in Section 6.

2. Model description. In this paper, we consider a single-sever queue with two customer types in which the class-$i$ ($i = 1, 2$) customer arrives according to a Poisson process with a mean arrival rate $\lambda_i$. The same customer class receives service according to the discipline of first-come first-served (FCFS). The service time of class $i$ is assumed to be mutually independent and identically distributed. It follows
a general distribution and the cumulative distribution function is denoted by \( G_i(t) \) with finite first and second moment \( E[G_i] \) and \( E[G_i^2] \). In addition, the queueing system has an infinite waiting room for each class, and class 2 has priority over class 1. The server serves the arriving customers one by one. As usual, we assume that the inter-arrival times and the service times of all customers are mutually independent.

Based on above assumptions, the service process of a class-\( i \) (\( i = 1, 2 \)) customer is illustrated in the following two figures:

**Figure 1.** The service process of an arbitrary arriving class-1 customer.

**Figure 2.** The service process of an arbitrary arriving class-2 customer.

The sojourn time of the class-\( i \) customers, denoted by \( S_i \), consists of the waiting time and the residence time, which are denoted by \( W_i \) and \( R_i \) respectively. In Figure 1, we note that there are \( N \) preemptions triggered by the class-2 customers in a residence time of the class-1 customer. Each of these preemption periods can be regarded as a busy period initiated by a class-2 customer. With a higher-priority, the class-2 customers are served without any preemption, as Figure 2 shows.

We model this resulting decision framework by assuming that a class-\( i \) customer receives a reward of \( K_i \) units after service completion but bears a waiting cost of \( C_i \) units per time unit when remaining in the system. Before making their decisions, the class-1 customers are assumed to be risk neutral and aim to maximize their expected net benefit. From now on, we further assume the condition

\[
K_i > C_i E[R_i],
\]

which ensures that the reward for receiving the service exceeds the expected cost in an expected residence period for a class-1 customer who finds the queueing system empty. Otherwise, after the system becomes empty for the first time no class-1 customers will ever enter the model because of the negative net benefit.
For all low-priority customers, they have the options to decide whether to join or balk based on the available information they acquire at their arrival instants. In this paper, we mainly consider two different cases: (1) the observable case: the arriving low-priority customer can observe the queue length of both classes only at the service completion instant; (2) the unobservable case: the arriving low-priority customer is not informed of any information about the queue length. Finally, we stress that the decisions of customers are irrevocable: retrials of balking customers and reneging of entering customers are not allowed.

3. The observable queue. Let \((n_1, n_2)\) be the exact numbers of both classes in the observable case. Concretely, \(n_1\) denotes the number of the class-1 customers who have not been served and \(n_2\) denotes the number of the class-2 customers in the queue. In the observable case, we impose a special entering policy that is only effective for the low-priority customers in the service system: the information of \((n_1, n_2)\) is only observable for a random class-1 arrival at the service completion instant. If any class-1 customer arrives at a working server, he will wait outside the system until the current service finishes. When there is more than one class-1 arrival, they will form a line temporarily and make their decisions whether to join or balk at the forthcoming service completion instants, according to the FCFS discipline. Next, some efforts are made to derive the equilibrium balking strategies of threshold type in the observable case. The fundamental results are given as follows.

**Theorem 3.1.** In the observable M/G/1 queue with two classes of customers and preemptive priority, there exists a unique equilibrium threshold \(n_e\), which is given by

\[
n_e = \left\lfloor \frac{K_1(1 - \rho_2) - C_1(n_2E[G_2] + E[G_1])}{C_1E[G_1]} \right\rfloor,
\]

such that a class-1 customer who observes the queue length \((n_1, n_2)\) at a service completion instant enters if \(n_1 \leq n_e\) and balks otherwise.

**Proof.** Mark a random arriving class-1 customer who decides to enter when he is informed of the queue length \((n_1, n_2)\) at a service completion instant, his expected benefit is

\[
U(n_1, n_2) = K_1 - C_1S(n_1, n_2),
\]

where \(S(n_1, n_2)\) denotes the expected sojourn time of the tagged customer who observes the queue length \((n_1, n_2)\) and decides to enter.

We put forward a service rule for the high-priority customers in front of this tagged customer as follows: the system will serve the first class-2 customer in the queue head along with all future class-2 customers arriving at the system during this period until it begins to serve the second one. This period, denoted by \(B\), can be regard as a busy period initiated by a class-2 customer who arrives at an idle system. Let \(Z(x)\) represent distribution function of a random variable \(Z\) and \(Z^*(s)\) denote Laplace-Stieltjes transform (LST) of \(Z\) with respect to parameter \(s\), that is,

\[
Z(x) = P(Z < x), \quad Z^*(s) = \int_{-\infty}^{+\infty} e^{-st}dZ(t).
\]

According to the Takacs Equation (or by using the similar method in Section 2.3.1 of Ross [24]), we can obtain Laplace-Stieltjes transform of the busy period \(B\)
with respect to parameter $s$

$$B^*(s) = G_2^* (s + \lambda_2 (1 - B^*(s))). \tag{4}$$

For any class-1 customer in the observable queue (including this tagged customer), there may be some preemptions triggered by the high-priority customers during his service time $G_1$. Each of these preemption periods can also be viewed as a busy period $B$. Then we can get Laplace-Stieltjes transform of the residence time $R_1$ with respect to parameter $s$ for any class-1 customer in the observable queue as follows:

$$R_1^*(s) = G_1^* (s + \lambda_2 (1 - B^*(s))). \tag{5}$$

Differentiating (4) and (5) with respect to $s$ respectively, we can obtain that

$$E[B] = E[G_2] \frac{1 - \rho_2}{1 - \rho_2}, \tag{6}$$

$$E[R_1] = E[G_1] \frac{1 - \rho_2}{1 - \rho_2}. \tag{7}$$

Since the same class is assumed indistinguishable, we have that the expected sojourn time $S(n_1, n_2)$ equals to

$$S(n_1, n_2) = (n_1 + 1) \frac{E[G_1]}{1 - \rho_2} + n_2 \frac{E[G_2]}{1 - \rho_2}. \tag{8}$$

It follows that by substituting (8) into (3)

$$U(n_1, n_2) = K_1 - C_1 \left( (n_1 + 1) \frac{E[G_1]}{1 - \rho_2} + n_2 \frac{E[G_2]}{1 - \rho_2} \right). \tag{9}$$

It can be easily checked that $U(n_1, n_2)$ strictly decreases with $n_1$. For this tagged class-1 customer, if $U(n_1, n_2) \geq 0$, he has no reason to balk. By considering (1), we can derive that $U(0, 0) > 0$ when $n_1 = n_2 = 0$. Therefore, by solving equation $U(n_1, n_2) \geq 0$ for $n_1$ in (9), the unique equilibrium threshold given by (2) can be obtained.

**Remark 1.** Since the service times of each class are generally distributed, the queueing system has only Markov property at some special instants. With the help of Markov Chain Imbedding Approach, we formulate a Markovian queue and make our discussion. It must be stressed that a preemption caused by class 2 may bring an uncompleted service to the class-1 customer in the head of the queue. The expected sojourn time of the preempted customer cannot be evaluated because the cumulative distribution function of the rest service time is unknown. Thus, we introduce a special joining mechanism that is analytically tractable into the observable case. Furthermore, it can be found that the equilibrium thresholds of the low-priority customer in the observable case are independent of the arrival rate $\lambda_1$. It is caused by the fact that an arriving customers decision is not influenced by the strategies of the future arrivals in a FCFS queue.

4. The unobservable queue. In the unobservable queueing system, we suppose that the class-1 customers follow a mixed strategy with a joining probability $q$. Then, the system behaves as the original, but with an arrival rate $\lambda_1 q$ instead of $\lambda_1$. Moreover, we suppose that this unobservable queueing system is stable on condition that

$$q \rho_1 + \rho_2 < 1, \tag{10}$$
where \( \rho_i (i = 1, 2) \) denotes the traffic intensity of the class-\( i \) customers and \( \rho_i = \lambda_i E[G_i] \).

### 4.1. Equilibrium

We firstly consider the class-1 customers equilibrium balking behavior in the unobservable queue. Then, applying the results obtained in Xu, Xu, and Wang [30], we readily obtain the following lemma.

**Lemma 4.1.** In the unobservable M/G/1 queue with two classes of customers and preemptive priority, if all arriving class-1 customers follow a common strategy \( q \), the mean sojourn time of a class-\( i \) customer \((i = 1, 2)\) is given as follows:

\[
E[S_1] = \frac{\lambda_1 q E[G_1^2] + \lambda_2 E[G_2^2]}{2(1 - \rho_2)(1 - q\rho_1 - \rho_2)} + \frac{E[G_1]}{1 - \rho_2},
\]

(11)

\[
E[S_2] = \frac{\lambda_2 E[G_2^2]}{2(1 - \rho_2)} + E[G_2].
\]

(12)

Based on the reward-cost structure, the expected net benefit of an arriving class-1 customer deciding to enter is given by

\[
U_e(q) = K_1 - C_1 \left( \frac{\lambda_1 q E[G_1^2] + \lambda_2 E[G_2^2]}{2(1 - \rho_2)(1 - q\rho_1 - \rho_2)} + \frac{E[G_1]}{1 - \rho_2} \right).
\]

(13)

Then we can proceed to determine the mixed equilibrium strategy of a class-1 customer in the unobservable case.

**Theorem 4.2.** In the unobservable M/G/1 queue with two classes of customers and preemptive priority, there exists a unique equilibrium balking strategy for the low-priority customers, with joining probability \( q_e \) given as follows:

\[
q_e = \begin{cases} 
0, & \frac{E[G_1]}{1 - \rho_2} < \frac{\lambda_2 E[G_2^2] + 2(1 - \rho_2)E[G_1]}{2(1 - \rho_2)} \\
\min \{ q_e^*, 1 \}, & \frac{\lambda_1 q E[G_1^2]}{2(1 - \rho_2)(1 - q\rho_1 - \rho_2)} + \frac{E[G_1]}{1 - \rho_2} > \frac{\lambda_2 E[G_2^2] + 2(1 - \rho_2)E[G_1]}{2(1 - \rho_2)^2}
\end{cases}
\]

(14)

where \( q_e^* \) is given by

\[
q_e^* = \frac{2(1 - \rho_2)(K_1 - K_1\rho_2 - C_1 E[G_1]) - C_1\lambda_2 E[G_2^2]}{C_1\lambda_1 E[G_1^2] + 2\rho_1(K_1 - K_1\rho_2 - C_1 E[G_1])}.
\]

(15)

**Proof.** Suppose that all class-1 customers enter with a probability \( q \) and consider a tagged arriving class-1 customer. Then, the tagged customer prefers to enter if \( U_e(q) \geq 0 \), he is indifferent between joining and balking if \( U_e(q) = 0 \) and he prefers to balk if \( U_e(q) < 0 \). We can easily check that the function \( U_e(q) \) is strictly decreasing with \( q \in [0, (1 - \rho_2)/\rho_1) \cap [0, 1] \) since

\[
\frac{dU_e(q)}{dq} = -\frac{C_1 \left( \lambda_1 (1 - \rho_2) E[G_1^2] + \lambda_2 \rho_1 E[G_2^2] \right)}{2(1 - \rho_2)(1 - q\rho_1 - \rho_2)^2} < 0.
\]

(16)

It follows that by considering (1) and taking (7) into account

\[
K_1 > \frac{E[G_1]}{1 - \rho_2} C_1.
\]

(17)
Then, we discuss the result as follows.

**Case 1.** When \( \frac{E[G_1]}{\rho_1} < \frac{K_1}{\rho_1} \leq \frac{\lambda_2 E[G_2^2] + 2(1 - \rho_2) E[G_1]}{2(1 - \rho_2)^2} \), \( U_e(q) \) is non-positive for each \( q \), the best response of a customer upon arrival is balking and the unique equilibrium probability is \( q = 0 \), which gives the first branch of (14).

**Case 2.** When \( \frac{K_1}{\rho_1} > \frac{\lambda_2 E[G_2^2] + 2(1 - \rho_2) E[G_1]}{2(1 - \rho_2)^2} \), we can obtain that \( U_e(0) > 0 \). Considering

\[
U_e \left( \frac{1 - \rho_2}{\rho_1} \right) = \lim_{q \to \frac{1 - \rho_2}{\rho_1}} U_e(q) = -\infty. \tag{18}
\]

we have that there exists a unique solution \( q^*_e \) of the equation for \( q \in [0, (1 - \rho_2)/\rho_1) \)

\[
U_e(q) = K_1 - C_1 \left( \frac{\lambda_1 q E[G_1^2] + \lambda_2 E[G_2^2]}{2(1 - \rho_2)(1 - q \rho_1 - \rho_2)} + \frac{E[G_1]}{1 - \rho_2} \right) = 0. \tag{19}
\]

Solving (19) gives (15). Noting that \( q^*_e \) is limited in the interval \([0, 1]\), we derive the second branch of (14).

**Remark 2.** The tagged customers best response to the strategies followed by other customers is a decreasing function with respect to the joining probability. That is to say, the more customers enter the queueing model, the less willing is the tagged customer to do that, which can be served as an illustration of an avoid the crowd (ATC) situation.

### 4.2. Social optimization

Next, we consider the low-priority customers socially optimal balking behavior in the unobservable queue and have the following lemma.

**Lemma 4.3.** In the unobservable \( M/G/1 \) queue with two classes of customers and preemptive priority, if all arriving class-1 customers follow a common strategy \( q \), the expected net social benefit per time unit is given by

\[
U_s(q) = q \lambda_1 K_1 + \lambda_2 K_2 - C_2 \rho_2 - C_1 \frac{q \rho_1}{1 - \rho_2} - C_2 \frac{\lambda_2^2 E[G_2^2]}{2(1 - \rho_2)} - C_1 \frac{q^2 \lambda_1^2 E[G_1^2] + q \lambda_1 \lambda_2 E[G_1^2]}{2(1 - \rho_2)(1 - q \rho_1 - \rho_2)}, \tag{20}
\]

which is doubly differentiable with respect to \( q \) in the interval of \([0, (1 - \rho_2)/\rho_1) \cap [0, 1] \).

**Proof.** The net social benefit per time unit, if the arriving class-1 customers follow a common strategy \( q \), is

\[
U_s(q) = (q \lambda_1 K_1 - C_1 E[L_1]) + (\lambda_2 K_2 - C_2 E[L_2]). \tag{21}
\]

where \( E[L_i] (i = 1, 2) \) denotes the mean queue length of the class-\( i \) customers.

It follows that by Little’s law

\[
\begin{aligned}
E[L_1] &= q \lambda_1 E[S_1]; \\
E[L_2] &= \lambda_2 E[S_2].
\end{aligned} \tag{22}
\]

By substituting (11), (12) and (22) into (21) and making some calculations, we derive (20). By considering (1), we can easily check that \( U_s(q) \) is doubly differentiable with respect to \( q \) in the interval of \([0, (1 - \rho_2)/\rho_1) \cap [0, 1] \).

The decision problem is to select a mixed strategy \( q \) for the class-1 customers to maximize the average net profit per unit time earned by the collective of all customers, thus we have the following theorem.
Theorem 4.4. In the unobservable M/G/1 queue with two classes of customers and preemptive priority, a unique socially optimal balking strategy for the low-priority customers exists, with joining probability $q_s$, which is given as follows

$$q_s = \begin{cases} 
0, & \frac{E[G_1]}{1 - \rho_2} < \frac{K_1}{C_1} \\
\min\{q_s^*, 1\}, & \frac{K_1}{C_1} > \frac{1}{1 - \rho_2} \left( E[G_1] + \frac{\lambda_2 E[G_2]}{2(1 - \rho_2)} \right), \\
\frac{1}{1 - \rho_2} \left( E[G_1] + \frac{\lambda_2 E[G_2]}{2(1 - \rho_2)} \right), & \text{Case 1.}
\end{cases}$$

(23)

where $q_s^*$ is given by

$$q_s^* = \frac{b - \sqrt{b^2 - 4ac}}{2a}.$$  

(24)

where

$$a = \rho_1 \left( 2\rho_1 (\lambda_1 K_1 (1 - \rho_2) - C_1 \rho_1) + C_1 \lambda_1^2 E[G_2] \right),$$

(25)

$$b = 2(1 - \rho_2) \left( 2\rho_1 (\lambda_1 K_1 (1 - \rho_2) - C_1 \rho_1) + C_1 \lambda_1^2 E[G_2] \right),$$

(26)

$$c = (1 - \rho_2) \left( 2 (\lambda_1 K_1 (1 - \rho_2) - C_1 \rho_1) (1 - \rho_2) - C_1 \lambda_1 \lambda_2 E[G_2] \right).$$

(27)

Proof. Differentiating (20) with respect to $q$, we obtain the first derivative of $U_s(q)$ as follows

$$U_s'(q) = \lambda_1 K_1 - \frac{C_1}{1 - \rho_2} \left( \rho_1 + q \lambda_1^2 (2(1 - \rho_2) - q \rho_1) E[G_2] + \lambda_1 \lambda_2 (1 - \rho_2) E[G_2] \right) \frac{2(1 - q \rho_1 - \rho_2)^2}{2(1 - q \rho_1 - \rho_2)^2}.$$  

(28)

Differentiating (28) with respect to $q$, we obtain the second derivative of $U_s(q)$

$$U_s''(q) = -\frac{C_1}{1 - \rho_2} \left( \frac{\lambda_1^2 E[G_2]}{1 - q \rho_1 - \rho_2} + \rho_1 \left( q ((1 - q \rho_1 - \rho_2) + (1 - \rho_2)) \lambda_1^2 E[G_2] + \lambda_1 \lambda_2 (1 - \rho_2) E[G_2] \right) \right) \frac{1 - q \rho_1 - \rho_2}{(1 - q \rho_1 - \rho_2)^3}.$$  

(29)

We can easily check that $U_s''(q)$ and that $U_s'(q)$ is strictly decreasing with respect to $q$ in the interval of $[0, (1 - \rho_2)/\rho_1)$. Similarly with Section 3.1, we discuss the result as follows

**Case 1.** When

$$\frac{E[G_1]}{1 - \rho_2} < \frac{K_1}{C_1}$$

(30)

is satisfied, we can easily check that $U_s'(0) \leq 0$. Then we have that $U_s'(q)$ is non- positive for each $q \in [0, (1 - \rho_2)/\rho_1) \cap [0, 1]$ and thus the socially optimal balking strategy for the low-priority customers is $q = 0$, which gives the first branch of (23).

**Case 2.** When

$$\frac{K_1}{C_1} > \frac{1}{1 - \rho_2} \left( E[G_1] + \frac{\lambda_2 E[G_2]}{2(1 - \rho_2)} \right)$$

(31)

is satisfied, we can easily check that $U_s'(0) > 0$. From

$$U_s' \left( \frac{1 - \rho_2}{\rho_1} \right)^- = \lim_{q \to \frac{1 - \rho_2}{\rho_1}} U_s'(q) = -\infty,$$

(32)
we can find that the following equation has a unique positive solution $q^*_s$ in the interval of $[0, (1 - \rho_2)/\rho_1)$

$$U'_s(q) = \lambda_1 K_1 - \frac{C_1}{1 - \rho_2} \left( \rho_1 + \frac{q\lambda_1^2 (2(1 - \rho_2) - q\rho_1) E[G_1^2] + \lambda_1 \lambda_2 (1 - \rho_2) E[G_2^2]}{2(1 - q\rho_1 - \rho_2)^2} \right) = 0. \tag{33}$$

After some straightforward calculations, Eq. (33) can be rewritten as

$$aq^2 - bq + c = 0. \tag{34}$$

where $a$, $b$ and $c$ are given by (25), (26) and (27) respectively.

From (17), it follows that $\lambda_1 K_1 (1 - \rho_2) - C_1 \rho_1 > 0$, which yields that $a > 0$ and $b > 0$. From (31), we can derive that $c > 0$. Considering (25) and (27), we obtain that

$$c < \frac{(1 - \rho_2)^2}{\rho_1} \left( 2\rho_1 (\lambda_1 K_1 (1 - \rho_2) - C_1 \rho_1) + C_1 \lambda_2^2 E[G_1^2] \right) = \left( \frac{1 - \rho_2}{\rho_1} \right)^2 a. \tag{35}$$

which yields that by taking into account (26)

$$\Delta = (-b)^2 - 4ac > \left( -\frac{2(1 - \rho_2) a}{\rho_1} \right)^2 - 4a \left( \frac{1 - \rho_2}{\rho_1} \right)^2 = 0. \tag{36}$$

Thus, we obtain that Eq. (34) has two positive roots, $q_1$ and $q_2$, which are given by

$$q_{1,2} = \frac{b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{37}$$

Considering that the quadratic function $f(q) = aq^2 + bq + c$ is a convex function and that it has a symmetry axis

$$q_0 = -\frac{(b)}{2a} = \frac{1 - \rho_2}{\rho_1}, \tag{38}$$

we can obtain that the unique positive solution of Eq. (34) in the interval of $[0, (1 - \rho_2)/\rho_1)$ is $q^*_s = q_2$. Noting that the joining probability $q_s$ is limited in the interval of $[0, 1]$, we derive the second branch of (23).

**Remark 3.** The socially optimal strategy for the low-priority customers is independent of the reward and the waiting cost of the high-priority customers (i.e., $K_2, C_2$) in the unobservable queue. It is an interesting result that is different from that obtained by Xu, Xu and Wang [30] under the same information assumption. In that study, the low-priority customers reward and waiting cost have some certain effects on the socially optimal strategy of the customers with high priority.

5. **Numerical examples.** In this section, by presenting a set of numerical experiments that are based on the analytical results obtained in this paper, we will investigate the effects of several parameters on the behavior of the class-1 customers. In particular, we illustrate the sensitivity of the equilibrium threshold with respect to several parameters. We also make some comparisons between the equilibrium strategy and the socially optimal strategy for the low-priority customers in the unobservable models with respect to several main indicators of the system.

Firstly, Figures 3-5 mainly illustrate the effects of the service reward $R_1$, the waiting cost $C_1$ and the expected service time $E[G_1]$ on the equilibrium thresholds of the low-priority customers. When $n_2$ is fixed, we can find that the equilibrium threshold $n_e$, which is up to the integrality requirement, increases with $R_1$ in Figure
3, decreases with $C_1$ in Figure 4 and with $E[G_1]$ in Figure 5. In addition, when $R_1$, $C_1$ and $E[G_1]$ are fixed respectively, the equilibrium threshold $n_e$ decreases with $n_2$ in three figures. Some intuitive interpretations can be presented in this situation. If the reward the customers gain after service completion/the waiting cost is higher, they are more/less likely to enter. The more high-priority customers the system accepts, the longer sojourn time the low-priority arrivals bear, and the fewer low-priority customers the system will serve.

![Figure 3](image)

Figure 3. The equilibrium thresholds of the class-1 customers vs. $K_1$ for $\lambda_2 = 0.5, E[G_1] = E[G_2] = 1, C_1 = 8$

Secondly, we obtain four concave curves with respect to the joining probability $q$ by setting $K_1 = 30, 40, 50$ and $60$ respectively in Figure 6. The socially optimal balking probability $q_s$ are expected to increase with respect to an increasing reward $K_1$. In Figure 7, we get three concave curves with a varying $q$ by setting $C_1 = 10, 15$ and $20$ respectively. The socially optimal balking probability $q_s$ are expected to decrease with respect to an increasing waiting cost $C_1$. In addition, as it is shown in Figures 6 and 7, the corresponding net profit function in each curve is increasing firstly to its peak value and then decreasing in $q$ with a higher speed. This situation indicates that when the social profit reaches the maximum value, the new class-1 arrivals will decrease the social profit rapidly.

Thirdly, Figures 8-10 are focused on the comparison between the equilibrium strategy and the socially optimal strategy. We can easily find that the equilibrium and socially optimal joining probabilities both increase with respect to the reward $K_1$ in Figure 8 while decrease in the waiting cost $C_1$ and the arrival rate $\lambda_1$ in Figures 9 and 10 respectively. From Figures 8-10, we can conclude that the equilibrium and socially optimal joining probabilities have the same varying interval. Besides, the equilibrium joining probability $q_e$ has a stronger sensitivity with respect to the varying parameters (i.e., $\lambda_1, K_1, C_1$) than the socially optimal balking probability $q_s$ in the unobservable queue. Besides, the typical inequality $q_s < q_e$ can be concluded from Figures 8-10, which is also demonstrated in the prior literatures (e.g.,
Figure 4. The equilibrium thresholds of the class-1 customers vs. $C_1$ for $\lambda_2 = 0.5, E[G_1] = E[G_2] = 1, K_1 = 100$

Figure 5. The equilibrium thresholds of the class-1 customers vs. $E[G_1]$ for $\lambda_2 = 0.5, E[G_2] = 1, K_1 = 100, C_1 = 5$

Noar [22], Edelson and Hildebrand [9]). This conclusion indicates that, compared with the social optimization strategy, the individual optimization strategy generally causes a more congested system. The self-centered customers usually ignore negative externalities that they impose on the later arrivals and tend to overuse the server. Therefore, system manager who aims to social benefit maximization must take this factor into account.
Figure 6. The expected net social benefit vs. $q$ for $\lambda_1 = \lambda_2 = 0.5$, $E[G_1] = E[G_2] = 1$, $E[G_1^2] = E[G_2^2] = 1.2$, $C_1 = 10$, $K_1 = 100$, $C_2 = 50$

Figure 7. The expected net social benefit vs. $q$ for $\lambda_1 = \lambda_2 = 0.5$, $E[G_1] = E[G_2] = 1$, $E[G_1^2] = E[G_2^2] = 1.2$, $K_1 = 10$, $K_2 = 100$, $C_2 = 50$

6. Conclusions. In this paper we analyzed the low-priority customers strategic behavior in a single server unobservable queueing system with general service time and two classes of customers where arriving customers decide whether to join the
system or balk based on different levels of information upon their arrivals. Specifically, we derived the equilibrium threshold strategy in the observable case as well as the equilibrium and socially optimal strategies in the unobservable case for the low-priority customers. We investigated the sensitivity of the equilibrium thresholds
Figure 10. Equilibrium and socially optimal joining probabilities of the class-1 customers vs. $\lambda_1$ for $\lambda_2 = 0.2$, $E[G_1] = E[G_2] = 1$, $E[G_1^2] = E[G_2^2] = 1.2$, $K_1 = 15$, $C_1 = 10$

with respect to some main parameters. We also make some comparisons between the equilibrium joining probability and the socially optimal joining probability with respect to several main indicators through several numerical examples.

This paper mainly focus on a single-sever preemptive queue with two consumer classes, future research can extend our model from several directions. One extension can consider the social optimization problem that maximizes the overall welfare of both customer classes and attempt to estimate the difference between the equilibrium solution and the social optimal solution. Another extension can relax the assumed one server and two classes, and extend to $m$ servers or $n$ customer classes ($m > 1, n > 2$). Our work can provide some beneficial suggestions on the analytical method when examining the performance of this generalized model. Additionally, a comparison between our results and the findings obtained from the generalized model will give more valuable managerial insights.

Acknowledgments. The authors would like to express sincere thanks to the anonymous reviewers, the Associate Editor and the Editor for very detailed and helpful comments and suggestions on this paper. This work was supported by the National Natural Science Foundation of China (Project Nos. 71271012, 71671011, 71332003, 11201408) and Natural Science Foundation of Hebei Province, China (Project No. A2013203148).

REFERENCES


Received October 2016; revised June 2018.

E-mail address: xubiao3613@163.com
E-mail address: tiannsh@sina.com
E-mail address: iszhyao@buaa.edu.cn