Journal of Systems Science and Information
Jun., 2020, Vol. 8, No. 3, pp. 273–290
DOI: 10.21078/JSSI-2020-273-18

Strategic Behavior and Optimization in an Unobservable Constant Retrial Queue with Balking and Set-Up Time

Linlin WANG
School of Science, Nanjing University of Science and Technology, Nanjing 210094*, China*
*E-mail: wanglinlin*502*@*163*.com*

Liwei LIU
School of Science, Nanjing University of Science and Technology, Nanjing 210094*, China*
E-mail: lwliu@njust.edu.cn

Zhen WANG
School of Science, Nanjing University of Science and Technology, Nanjing 210094*, China*
E-mail: 1721157913*@qq.com*

Xudong CHAI
School of Science, Nanjing University of Science and Technology, Nanjing 210094*, China*
E-mail: 690752860*@qq.com*

Abstract An $M/M/1$ constant retrial queue with balking customers and set-up time is considered. Once the system becomes empty, the server will be turned down to reduce operating costs, and it will be activated only when there is a customers arrives. In this paper, the almost unobservable case is studied, in which the information of the queue length is unavailable, whereas the state of the server can be obtained. Firstly, the steady state solutions are derived and the individual equilibrium strategies are analyzed. In addition, social optimization problems, including cost analysis and social welfare maximization are investigated by using the PSO algorithm. Finally, by appropriate numerical examples, the sensitivity of some main system parameters is shown.

Keywords retrial queue; set-up time; unobservable; equilibrium; PSO algorithm

1 Introduction

In recent years, retrial queueing systems have been widely adopted to model LAN (local area networking) systems. Because in a LAN, a job that can not be disposed at its arrival instant will be translated again in a later time, which consistents with the key thought of retrial queues. Assuming that the waiting space is finite, when the server is idle, new arriving customer will take up the server and accept service immediately. Otherwise if the server is in other states, then arriving customers have to leave the server, but they can join a virtual retrial orbit and wait for retry. So far, there has been vast literature on retrial queuing system. Falin and

Received June 19, 2019, accepted April 24, 2020
Supported by the National Natural Science Foundation of China (61773014), Postgraduate Research & Practice Innovation Program of Jiangsu Province China (SJKY19_0277)

Templeton[1] summarized the main methods and results of previous research on retrial queuing system. For interested readers, more details can be found in Artalejo[2], Gomez-Corral[3] and Artalejo[4]. It is worth noting that in the retrial queuing literatures, most of the articles assume that the retrial intervals of customers are independent and follow the exponential distribution with the parameter of θ. Under this strategy, the total retrial rate changes with the number of customers in the orbit. While in our model, customers in the orbit follow the first come, first served discipline (FCFS), only the head customer can retry for service after a random time, therefore the retrial rate becomes a constant and is independent of the queue length. Fayolle[5] first came up with the constant retrial policy to model a telephone exchange system, Later researches can be seen in Falin[6], Artalejo, Gomez-Corral[7].

During the past two decades, the game theory has been introduced into the queueing systems. Various queueing models have been studied from the perspective of economics. In equilibrium literature, Naor[8] first investigated an observable M/M/1 queue under the linear reward-cost structure. He gave the equilibrium balking strategy and proposed that customers could enter the system with social optimal strategy if charged tolls. Edelson and Hildebrand[9] studied the same model under the assumption that the new arrival customers could not observe the system queue length. Subsequently, many researchers devoted to study and generalize the queueing model, and achieved more results.

In addition, for the set-up time, it was first studied from an economic view by Burnetas and Economou[10] as a vacation policy. For the sake of costs reduction and energy saving, when the system becomes empty, the server should be turned down and won't be restarted until a customer arrives. Considering to the reality, the opening time of the server should not be ignored. Therefore, regarding the set-up time as a random time subjected to exponential distribution is widely accepted. Zhang and Wang[11] studied an M/G/1 retrial queue with reserved idle time and setup time. They derived the optimal pricing strategies from the view of the social planner and the server, respectively. However, they take consideration of the general retrial rate, and assume that the information of the queue length and the server's state are unobservable, which are quite different from this paper. Recent results of queueing systems with set-up time can be seen in Yutaka, Yoshitaka, Yutaka, et al.[12].

The main contribution of this article is the introduction of the set-up time to a constant retrial queue. After the set-up time policy is added, the model becomes more realistic and is firstly studied from an economic view. Closed server will only be activated by an arriving customer and will go through a period of set-up time. Under the almost unobservable condition, we not only derive the customers' equilibrium strategies, but also investigate the social optimization problems respectively from the standpoints of the service provider and the social manager, so that each party in the game can take specific measures to achieve their own goals.

The rest content of this paper can be summarized as follows. Section 2 presents the investigated model and corresponding assumptions. In Section 3, we study the steady state solutions and derive some main system performance measures. In Section 4, we get the individual equilibrium arrival rates by analyzing the properties of the customers' utility function. Section 5 gives the expressions of the social welfare and the cost function. Due to the complexity of their formulas, the Particle Swarm Optimization algorithm (PSO) is introduced to find the socially optimal arrival rates and the cost-optimal arrival rates. Section 6 shows the results of the numerical examples and Section 7 concludes this paper.

2 Model Description

In this section, we consider a constant retrial queue with balking and set-up time. Potential primary customers arrive in a Poisson process with rate Λ. Assuming that there is no waiting space in front of the server, an arriving customer who finds the server idle will take up it and receive service immediately, otherwise he could choose to enter the "virtual" retrial orbit waiting for retry or to balk, depending on his expected net payoff. Once a service is finished, only the head customer in the orbit can begin to request for service. The retrial time is exponentially distributed with rate θ. If a new customer arrives during the retrial time, he will interrupt the process and receive service at once. The service times for customers (both external and repeated) are independent and exponentially distributed with rate μ.

Every time the system becomes empty, the server will be turned down and won't be restarted until there is a customer arrives. Once activated, the server will go through a random opening time subjected to exponential distribution with parameter α. The customer who activated the server will automatically enter the retrial orbit and become the head of the retrial queue. It is assumed that the interarrival times, service times, retrial times and set-up times are mutually independent.

Assuming that every customer gains a profit of R units after completing the service, while has to pay for a waiting cost of C per unit time for remaining in the system. All customers are indistinguishable and rational to pursue the maximization of their own profits, and they have the rights to determine whether or not to join at their arrival instants, which forms a game among them. Explicitly, customers tend to enter when they gain more from the service than they pay for the cost, otherwise they prefer to balk. On the other hand, if the reward equals to the cost, it doesn't matter for customers whether to enter or to stop. To ensure that a customer who arrives during the idle period will always choose to enter, we assume that

$$R > \frac{C}{\mu}. \tag{2.1}$$

In this paper, the almost unobservable case is investigated. We use the pair $\{(I(t), N(t)), t \geq 0\}$ to describe the system at time t, where $I(t)$ indicates the state of the server and $N(t)$ shows the number of customers in the orbit. Explicitly,

$$I(t) = \begin{cases} 0, & \text{if the server is idle,} \\ 1, & \text{if the server is busy,} \\ 2, & \text{if the server is in set-up time.} \end{cases} \tag{2.2}$$

Then $\{(I(t), N(t)), t \geq 0\}$ is a two-dimensional continuous time Markov chain, and its state space is $\{(i,j), i = 0,1,2; j \geq 0\}$. According to the above analysis, the joining probabilities of arriving customers depend on $I(t)$. Define $\lambda_i = \Lambda q_i, i = 0,1,2$, where q_i is the joining probability that specifies a general strategy, and λ_i becomes the effective arrival rate at state i, which reflects the customer's real demand rate for the service.

Because of the assumption (2.1), we have that $\lambda_0 = \mathit{\Lambda}$. In the following analysis, we only need to study how do customers make decision at state $I(t) = 1, 2$, and we focus on the corresponding effective arrival rates λ_1 and λ_2. According to Economou & Kanta[13], the steady-state condition of this model is given by

$$\rho = \frac{\lambda_1(\mathit{\Lambda} + \theta)}{\mu\theta} < 1. \tag{2.3}$$

The transition rate diagram of this model is shown in Figure 1.

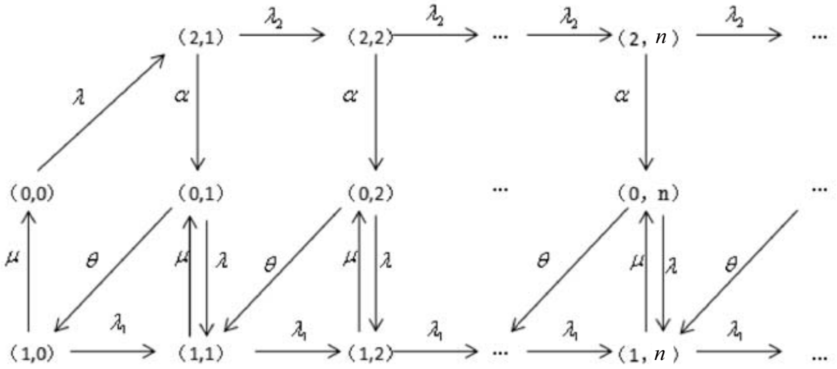


Figure 1 Transition rate diagram of a retrial queue with balking and set-up time

3 Steady State Solutions

For the convenience to analyze the equilibrium strategies, we first study the steady state solutions. Suppose that the system is stable and let $p(i, j)$ be the steady-state probability of state (i, j), then we can easily get the balance equations:

$$(\lambda_2 + \alpha)p(2, 1) = \mathit{\Lambda}p(0, 0), \tag{3.1}$$

$$(\lambda_2 + \alpha)p(2, n) = \lambda_2 p(2, n-1), \quad n \geq 2, \tag{3.2}$$

$$\mathit{\Lambda}p(0, 0) = \mu p(1, 0), \tag{3.3}$$

$$(\mathit{\Lambda} + \theta)p(0, n) = \mu p(1, n) + \alpha p(2, n), \quad n \geq 1, \tag{3.4}$$

$$(\lambda_1 + \mu)p(1, 0) = \theta p(0, 1), \tag{3.5}$$

$$(\lambda_1 + \mu)p(1, n) = \mathit{\Lambda}p(0, n) + \lambda_1 p(1, n-1) + \theta p(0, n+1), \quad n \geq 1. \tag{3.6}$$

The generating function technique is used to solve the equations. Define the partial generating functions as:

$$P_0(z) = \sum_{n=0}^{\infty} z^n p(0, n), \quad P_1(z) = \sum_{n=0}^{\infty} z^n p(1, n), \quad P_2(z) = \sum_{n=1}^{\infty} z^n p(2, n). \tag{3.7}$$

Then we can get the following results.

Theorem 1 *For the almost unobservable $M/M/1$ retrial queue with balking and set-up time, the probabilities that the server is idle, busy or in set-up period are, respectively given by*

$$P_0(1) = \frac{\mathit{\Lambda}\lambda_2\mu + \alpha\mu\theta - \mathit{\Lambda}\lambda_1\alpha - \lambda_1\alpha\theta}{\alpha\mu\theta - \mathit{\Lambda}\lambda_1\alpha - \lambda_1\alpha\theta} p(0, 0), \tag{3.8}$$

$$P_1(1) = \frac{\Lambda\lambda_2(\Lambda+\theta)}{\alpha\mu\theta - \Lambda\lambda_1\alpha - \lambda_1\alpha\theta}p(0,0), \tag{3.9}$$

$$P_2(1) = \frac{\Lambda}{\alpha}p(0,0), \tag{3.10}$$

where

$$p(0,0) = \frac{\alpha\mu\theta - \Lambda\lambda_1\alpha - \lambda_1\alpha\theta}{\Lambda\mu\theta - \Lambda^2\lambda_1 - \Lambda\lambda_1\theta + \Lambda\lambda_2\mu + \alpha\mu\theta - \Lambda\lambda_1\alpha - \lambda_1\alpha\theta + \Lambda^2\lambda_2 + \Lambda\lambda_2\theta}. \tag{3.11}$$

Proof Multiplying Equations (3.1)∼(3.2) by z^n and summing up over all n, we derive the following equation:

$$(\lambda_2+\alpha)P_2(z) = \lambda_2 z P_2(z) + \Lambda p(0,0). \tag{3.12}$$

Similarly, from Equations (3.3)∼(3.6), we have that

$$(\Lambda+\theta)P_0(z) - \theta p(0,0) = \mu P_1(z) + \alpha P_2(z), \tag{3.13}$$

$$(\Lambda+\mu)P_1(z) = \Lambda P_0(z) - \Lambda p(0,0) + \lambda_1 z P_1(z) + \frac{\theta}{z}P_0(z) - \frac{\theta}{z}p(0,0). \tag{3.14}$$

By some algebraic manipulations, we use $p(0,0)$ to express $P_i(z), i = 0,1,2$ as:

$$P_2(z) = \frac{\Lambda p(0,0)}{\alpha + \lambda_2(1-z)}, \tag{3.15}$$

$$P_1(z) = \frac{\Lambda + \frac{\theta}{z}}{\mu + \lambda_1(1-z)}[P_0(z) - p(0,0)], \tag{3.16}$$

$$P_0(z) = \frac{\alpha}{\Lambda + \theta - \frac{\mu(\Lambda+\theta/z)}{\mu+\lambda_1(1-z)}}P_2(z) + \frac{\theta - \frac{\mu(\Lambda+\theta/z)}{\mu+\lambda_1(1-z)}}{\Lambda + \theta - \frac{\mu(\Lambda+\theta/z)}{\mu+\lambda_1(1-z)}}p(0,0). \tag{3.17}$$

Taking (3.15) into (3.17), then we have,

$$\begin{aligned} &P_0(z) \\ &= \frac{\Lambda\alpha[\mu+\lambda_1(1-z)] - \mu(\Lambda+\theta/z)[\alpha+\lambda_2(1-z)] + \theta[\alpha+\lambda_2(1-z)][\mu+\lambda_1(1-z)]}{[\alpha+\lambda_2(1-z)]\{(\Lambda+\theta)[\mu+\lambda_1(1-z)] - \mu(\Lambda+\theta/z)\}}p(0,0). \end{aligned} \tag{3.18}$$

Noticed that, $P_0(z)$ and $P_1(z)$ are both indeterminate forms when $z = 1$, so we use the L' Hospital rule to solve them and we have (3.8)∼(3.10). Taking them into the normalization condition: $P_0(1) + P_1(1) + P_2(1) = 1$, we derive $p(0,0)$ as given by (3.11). ▮

Based on the above proof process, we can immediately derive $P_i'(z), i = 0,1,2$ by differentiating (3.15), (3.16) and (3.18). Letting $z = 1$, we find that

$$P_0'(1) = \frac{\mu}{\Lambda+\theta}P_1'(1) + \frac{\Lambda\lambda_2}{\alpha(\lambda+\theta)} \cdot p(0,0), \tag{3.19}$$

$$P_1'(1) = \left[\frac{-\Lambda\lambda_2\theta}{(\alpha\mu\theta - \Lambda\lambda_1\alpha - \lambda_1\alpha\theta)} + \frac{(\Lambda+\theta)(\Lambda\lambda_2^2\mu\theta + \Lambda\lambda_2\alpha\mu\theta - \Lambda\lambda_1\lambda_2^2\theta - \Lambda^2\lambda_1\lambda_2^2)}{(\alpha\mu\theta - \Lambda\lambda_1\alpha - \lambda_1\alpha\theta)^2}\right]p(0,0), \tag{3.20}$$

$$P_2'(1) = \frac{\Lambda\lambda_2}{\alpha^2}p(0,0). \tag{3.21}$$

$P_i'(1), i = 0,1,2$ can be considered as the contribution of state i to the expected number of customers in the orbit. As the three states are connected, customers accumulating in a single state will affect the other states.

Next we turn our attention to investigate the mean sojourn time(including service time) of a singled out customer. Assuming that an arriving customer decides to enter the system, and becomes the jth customer in the orbit. We denote the expected sojourn time of the specific customer at different states as $T_i(j), i = 0, 1, 2$. Then we obtain $T_i(j), i = 0, 1, 2$ in the following analysis.

Lemma 1 *For the almost unobservable $M/M/1$ retrial queue with balking and set-up time, the mean sojourn times of the jth customer in the orbit at steady state $I(t) = 0, 1, 2$ are, respectively given by*

$$T_0(j) = j\left(\frac{1}{\theta} + \frac{\Lambda + \theta}{\mu\theta}\right), \tag{3.22}$$

$$T_1(j) = j\left(\frac{1}{\theta} + \frac{\Lambda + \theta}{\mu\theta}\right) + \frac{1}{\mu}, \tag{3.23}$$

$$T_2(j) = j\left(\frac{1}{\theta} + \frac{\Lambda + \theta}{\mu\theta}\right) + \frac{1}{\alpha}. \tag{3.24}$$

Proof Through state transition analysis, we have the following equations:

$$T_1(0) = \frac{1}{\mu}, \tag{3.25}$$

$$T_0(j) = \frac{1}{\Lambda + \theta} + \frac{\Lambda}{\Lambda + \theta}T_1(j) + \frac{\theta}{\Lambda + \theta}T_1(j-1), \tag{3.26}$$

$$T_1(j) = \frac{1}{\lambda_1 + \mu} + \frac{\lambda_1}{\lambda_1 + \mu}T_1(j) + \frac{\mu}{\lambda_1 + \mu}T_0(j), \tag{3.27}$$

$$T_2(j) = \frac{1}{\lambda_2 + \alpha} + \frac{\lambda_2}{\lambda_2 + \alpha}T_2(j) + \frac{\alpha}{\lambda_2 + \alpha}T_0(j). \tag{3.28}$$

Combining equations (3.26) and (3.27), we have the recursive form of $T_1(j), j \geq 1$,

$$T_1(j) = T_1(j-1) + \frac{1}{\theta} + \frac{\Lambda + \theta}{\mu\theta}, \quad j \geq 1. \tag{3.29}$$

Considering to (3.25), we have that,

$$T_1(j) = T_1(0) + j\left(\frac{1}{\theta} + \frac{\Lambda + \theta}{\mu\theta}\right) = j\left(\frac{1}{\theta} + \frac{\Lambda + \theta}{\mu\theta}\right) + \frac{1}{\mu}, \quad j \geq 0. \tag{3.30}$$

Taking the expression of $T_1(j)$ into (3.26), we derive $T_0(j)$ as in (3.22). Furthermore, we have $T_2(j)$ by substituting $T_0(j)$ in (3.28). ▌

For a non-specific new arriving customer, the situation is different, but we can also get the mean sojourn time with the help of Lemma1. Denote $W_i, (i = 0, 1, 2)$ as the mean sojourn times of a new arriving customer who finds that $I(t) = 0, 1, 2$. We have the following conclusion.

Theorem 2 *For the almost unobservable $M/M/1$ retrial queue with balking and set-up time, the mean sojourn times of a new arriving customer who finds that the server is in idle, busy or set-up period are, respectively given by*

$$W_0 = 1/\mu, \tag{3.31}$$

$$W_1 = \frac{\Lambda + \mu + \theta}{\mu\theta}\left(\frac{\lambda_2}{\alpha} + \frac{\mu\theta}{\mu\theta - \Lambda\lambda_1 - \lambda_1\theta}\right) + \frac{\Lambda + \theta}{\mu\theta} + \frac{\Lambda}{\theta(\Lambda + \theta)}, \tag{3.32}$$

$$W_2 = \frac{\Lambda + \mu + \theta}{\mu\theta} \cdot \frac{\lambda_2}{\alpha} + \frac{\Lambda\alpha + \alpha\mu + \alpha\theta + \mu\theta}{\alpha\mu\theta}. \tag{3.33}$$

Proof First of all, any customer arrives at idle state will enter the system and immediately be served, so the sojourn time equals to the mean service time $\frac{1}{\mu}$.

What needs to be concerned is that in busy or set-up period cases. Assume that the server is busy and there are already k customers in the orbit, if a new customer decides to enter the system, he will be the $(k+1)$th customer in the orbit, whose mean sojourn time is $T_1(k+1)$. Using the PASTA property, we denote $P(k|1) = \frac{p(1,k)}{\sum_{n=0}^{\infty} p(1,k)} = \frac{p(1,k)}{P_1(1)}, k \geq 0$ as the conditional probability that there are k customers in the orbit, given that the server is busy. According to the total probability formula, we derive that,

$$\begin{aligned}
W_1 &= \sum_{k=0}^{\infty} T_1(k+1)P(k|1) \\
&= \frac{\sum_{k=0}^{\infty} T_1(k+1)p(1,k)}{P_1(1)} \\
&= \frac{\sum_{k=0}^{\infty}[(k+1)(\frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta}) + \frac{1}{\mu}] \cdot p(1,k)}{P_1(1)} \\
&= \frac{\sum_{k=0}^{\infty} k(\frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta})p(1,k)}{P_1(1)} + \frac{\sum_{k=0}^{\infty}(\frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta} + \frac{1}{\mu}) \cdot p(1,k)}{P_1(1)} \\
&= \left(\frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta}\right) \cdot \frac{P_1'(1)}{P_1(1)} + \frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta} + \frac{1}{\mu},
\end{aligned}$$

where $\frac{P_1'(1)}{P_1(1)}$ can be computed by (3.9) and (3.20) as

$$\frac{P_1'(1)}{P_1(1)} = -\frac{\theta}{\Lambda+\theta} + \frac{\lambda_2\mu\theta + \alpha\mu\theta - \lambda_1\lambda_2\theta - \Lambda\lambda_1\lambda_2}{\alpha\mu\theta - \Lambda\lambda_1\alpha - \lambda_1\alpha\theta}.$$

Combining the two equations and by some algebraic computation, we get the expression of W_1 as in (3.32). Meanwhile, using the same argument, we obtain that,

$$\begin{aligned}
W_2 = \sum_{k=1}^{\infty} T_2(k+1)P(k|2) &= \left(\frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta}\right) \cdot \frac{P_2'(1)}{P_2(1)} + \frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta} + \frac{1}{\alpha} \\
&= \left(\frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta}\right) \cdot \frac{\lambda_2}{\alpha} + \frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta} + \frac{1}{\alpha},
\end{aligned}$$

which can be converted to (3.33). ▮

Theorem 2 reflects an important fact that W_2 only depends on λ_2. Although W_1 is related to the effective arrival rates both in busy and set-up time. As long as λ_2 is determined, W_1 is considered to be subject to λ_1 only. This can be helpful for the following equilibrium analysis.

4 Individual Equilibrium

In this section, we aim to derive the individual equilibrium strategies in different states. Since the self-optimizing arriving customers are indistinguishable, and have the right to decide whether to join or not, it is reasonable to regard them as players in a symmetric game, whose strategies are balking or joining. For simplicity, a Nash equilibrium strategy is the one that the tagged customer has to adopt to maximize his benefit when all others take the same strategy. According to the game theory, there exists an equilibrium joining strategy. That is to say, the equilibrium arrival rates can be specified. According to the assumption of the reward-cost structure, we have the expression of the customers' benefit functions in state $I(t) = 1, 2$ as:

$$\begin{aligned} S_1(\lambda_1, \lambda_2) &= R - CW_1 \\ &= R - C\left[\frac{\Lambda+\mu+\theta}{\mu\theta}\left(\frac{\lambda_2}{\alpha} + \frac{\mu\theta}{\mu\theta - \Lambda\lambda_1 - \lambda_1\theta}\right) + \frac{\Lambda+\theta}{\mu\theta} + \frac{\Lambda}{\theta(\Lambda+\theta)}\right], \qquad (4.1) \\ S_2(\lambda_2) &= R - CW_2 = R - C\left[\frac{\Lambda+\mu+\theta}{\mu\theta} \cdot \frac{\lambda_2}{\alpha} + \frac{\Lambda\alpha + \alpha\mu + \alpha\theta + \mu\theta}{\alpha\mu\theta}\right]. \qquad (4.2) \end{aligned}$$

Then we have the following results.

Theorem 3 *For the almost unobservable $M/M/1$ retrial queue with balking and set-up time,*

1) *when the server is in set-up period, the individual equilibrium arrival rate is given by*

$$\lambda_2^e = \begin{cases} 0, & if\ \ \dfrac{R}{C} \le \dfrac{1}{\theta} + \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{1}{\alpha}, \\ \lambda_2', & if\ \ \dfrac{1}{\theta} + \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{1}{\alpha} < \dfrac{R}{C} \le \dfrac{\Lambda+\mu+\theta}{\mu\theta} \cdot \dfrac{\Lambda}{\alpha} + \dfrac{1}{\theta} + \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{1}{\alpha}, \\ \Lambda, & if\ \ \dfrac{R}{C} > \dfrac{\Lambda+\mu+\theta}{\mu\theta} \cdot \dfrac{\Lambda}{\alpha} + \dfrac{1}{\theta} + \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{1}{\alpha}. \end{cases} \qquad (4.3)$$

2) *when the server is busy, there are three cases, and the individual equilibrium arrival rates are given as:*

(a) *When* $\frac{R}{C} \le \frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta} + \frac{1}{\alpha}$, *i.e.*, $\lambda_2^e = 0$,

$$\lambda_1^e = \begin{cases} 0, & if\ \ \dfrac{R}{C} \le \dfrac{\Lambda+\mu+\theta}{\mu\theta} + \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{\Lambda}{\theta(\Lambda+\theta)}, \\ \lambda_{11}', & if\ \ \dfrac{\Lambda+\mu+\theta}{\mu\theta} + \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{\Lambda}{\theta(\Lambda+\theta)} < \dfrac{R}{C} \le \dfrac{\Lambda+\mu+\theta}{\mu\theta - \Lambda^2 - \Lambda\theta} + \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{\Lambda}{\theta(\Lambda+\theta)}, \\ \Lambda, & if\ \ \dfrac{R}{C} > \dfrac{\Lambda+\mu+\theta}{\mu\theta - \Lambda^2 - \Lambda\theta} + \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{\Lambda}{\theta(\Lambda+\theta)}. \end{cases} \qquad (4.4)$$

(b) *When* $\frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta} + \frac{1}{\alpha} < \frac{R}{C} \le \frac{\Lambda+\mu+\theta}{\mu\theta} \cdot \frac{\Lambda}{\alpha} + \frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta} + \frac{1}{\alpha}$, *i.e.*, $\lambda_2^e = \lambda_2'$,

$$\lambda_1^e = \begin{cases} 0, & if\ \ \dfrac{1}{\alpha} \le \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{\Lambda}{\theta(\Lambda+\theta)}, \\ \lambda_{12}', & if\ \ \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{\Lambda}{\theta(\Lambda+\theta)} < \dfrac{1}{\alpha} \le \dfrac{\Lambda+\mu+\theta}{\mu\theta - \Lambda^2 - \Lambda\theta} + \dfrac{\Lambda}{\theta(\Lambda+\theta)} - \dfrac{1}{\theta}, \\ \Lambda, & if\ \ \dfrac{1}{\alpha} > \dfrac{\Lambda+\mu+\theta}{\mu\theta - \Lambda^2 - \Lambda\theta} + \dfrac{\Lambda}{\theta(\Lambda+\theta)} - \dfrac{1}{\theta}. \end{cases} \qquad (4.5)$$

(c) *When* $\frac{R}{C} > \frac{\Lambda+\mu+\theta}{\mu\theta} \cdot \frac{\Lambda}{\alpha} + \frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta} + \frac{1}{\alpha}$, *i.e.,* $\lambda_2^e = \Lambda$,

$$\lambda_1^e = \begin{cases} 0, & if \quad \dfrac{R}{C} \le \dfrac{\Lambda^2+\Lambda\mu+\Lambda\theta}{\alpha\mu\theta} + \dfrac{\Lambda+\mu+\theta}{\mu\theta} + \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{\Lambda}{\theta(\Lambda+\theta)}, \\ \lambda'_{13}, & if \quad \dfrac{\Lambda^2+\Lambda\mu+\Lambda\theta}{\alpha\mu\theta} + \dfrac{\Lambda+\mu+\theta}{\mu\theta} + \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{\Lambda}{\theta(\Lambda+\theta)} < \dfrac{R}{C}, \\ & and \quad \dfrac{R}{C} \le \dfrac{\Lambda(\Lambda+\mu+\theta)}{\alpha\mu\theta} + \dfrac{\Lambda+\mu+\theta}{\mu\theta-\Lambda^2-\Lambda\theta} + \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{\Lambda}{\theta(\Lambda+\theta)}, \\ \Lambda, & if \quad \dfrac{R}{C} > \dfrac{\Lambda(\Lambda+\mu+\theta)}{\alpha\mu\theta} + \dfrac{\Lambda+\mu+\theta}{\mu\theta-\Lambda^2-\Lambda\theta} + \dfrac{\Lambda+\theta}{\mu\theta} + \dfrac{\Lambda}{\theta(\Lambda+\theta)}. \end{cases} \tag{4.6}$$

Specifically, λ'_2 *is the unique root of the equation* $R - CW_2(\lambda_2) = 0$, *and is given by*

$$\lambda'_2 = \left(\frac{R}{C} - \frac{1}{\theta} - \frac{\Lambda+\theta}{\mu\theta} - \frac{1}{\alpha}\right) \cdot \frac{\alpha\mu\theta}{\Lambda+\mu+\theta}. \tag{4.7}$$

$\lambda'_{1i}, i = 1, 2, 3$, *are respectively the roots of the three formulas :* $S_1(\lambda_1, 0) = 0$*;* $S_1(\lambda_1, \lambda'_2) = 0$*;* $S_1(\lambda_1, \Lambda) = 0$, *and are given by*

$$\lambda'_{11} = \frac{\mu\theta}{\Lambda+\theta} - \frac{\Lambda+\mu+\theta}{\Lambda+\theta}\left(\frac{R}{C} - \frac{\Lambda+\theta}{\mu\theta} - \frac{\Lambda}{\theta(\Lambda+\theta)}\right)^{-1}, \tag{4.8}$$

$$\lambda'_{12} = \frac{\Lambda\mu\theta + \mu\theta^2 - \Lambda\alpha\mu - \Lambda^2\alpha - 2\Lambda\alpha\theta - \alpha\theta^2}{(\Lambda+\theta)(\Lambda+\alpha+\theta)}, \tag{4.9}$$

$$\lambda'_{13} = \frac{\mu\theta}{\Lambda+\theta} - \frac{\Lambda+\mu+\theta}{\Lambda+\theta}\left[\frac{R}{C} - \frac{\Lambda(\Lambda+\mu+\theta)}{\alpha\mu\theta} - \frac{\Lambda+\theta}{\mu\theta} - \frac{\Lambda}{\theta(1+\theta)}\right]^{-1}. \tag{4.10}$$

Proof Similar to the above discussion, customers have to measure the losses and gains to make the decision on whether to join or not.

1) If a customer observes that the server is in set-up period upon arriving, then his joining strategy is dependent on his benefit function $S_2(\lambda_2)$. It can be intuitively found from (4.2) that $S_2(\lambda_2)$ is decreasing with λ_2, which leads an ATC(avoid the crowd) behavior of the customer, thus there exists a unique equilibrium arrival rate λ_2^e (see Hassin and Haviv[14]). To derive λ_2^e, we have the following discuss:

(i) When $S_2(0) \le 0$, i.e., $\frac{R}{C} \le \frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta} + \frac{1}{\alpha}$, then $S_2(\lambda_2) \le 0$ in every $\lambda_2 \in [0, \Lambda]$. In this case, customers' benefits are always negative, hence nobody wants to enter the system, so there forms a equilibrium $\lambda_2^e = 0$, which is the first branch of (4.3).

(ii) When $S_2(0) > 0$ and $S_2(\Lambda) \le 0$, i.e., $\frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta} + \frac{1}{\alpha} < \frac{R}{C} \le \frac{\Lambda+\mu+\theta}{\mu\theta} \cdot \frac{\Lambda}{\alpha} + \frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta} + \frac{1}{\alpha}$, since $S_2(\lambda_2)$ strictly decreases with λ_2, there is a unique root of $S_2(\lambda_2) = 0$ in $(0, \Lambda]$, denoted by λ'_2, which is exactly the equilibrium point. Because if $\lambda_2 > \lambda'_2$, new arriving customers will all choose to balk since their benefits are negative, thus their strategies are equal to zero. This is against to the equilibrium property that all customers adopt the same strategy. On the other hand, the arrival rate which is less than λ'_2 will lead an "all joining" strategy and finally converge to λ'_2, too. Therefore, λ'_2 is the unique equilibrium point of this branch.

(iii) When $S_2(\Lambda) > 0$, i.e., $\frac{R}{C} > \frac{\Lambda+\mu+\theta}{\mu\theta} \cdot \frac{\Lambda}{\alpha} + \frac{1}{\theta} + \frac{\Lambda+\theta}{\mu\theta} + \frac{1}{\alpha}$, the situation is completely opposite to (a), and all customers choose to enter since their benefits are always positive, which leads to a equilibrium of $\lambda_2^e = \Lambda$. This forms the last branch of (4.3).

2) Theorem 2 tells that, once λ_2 is determined, W_1 is only dependent with λ_1, so as the customers' benefit function in busy state, which can be remarked as $S_1(\lambda_1)$. Similarly, it can be found from (4.1) that $S_1(\lambda_1)$ is decreasing with λ_1 when λ_2 is settled. Thus, there exists three cases according to the value of λ_2 in (4.3), which respectively correspond to (a), (b), (c) in Theorem 3. Proceeding the analysis as before, the proof of (4.4), (4.5) and (4.6) can be done, and λ'_{11}, λ'_{12}, λ'_{13} are, respectively derived by solving the equations: $S_1(\lambda_1, 0) = 0$; $S_1(\lambda_1, \lambda'_2) = 0$; $S_1(\lambda_1, \varLambda) = 0$. $\blacksquare$

There is one thing to notice in the above analysis: The process of deriving λ_1^e has preconditions in the three cases. In each case, we discuss three branches and present λ_1^e as in (4.4)$\sim$(4.6) for brevity and clarity. However, when there is a conflict between the precondition and the branch condition, that branch will be naturally canceled.

5 Social Optimization

In this section, we seek for social optimization for the social planner and the service provider, respectively, since they are out of different considerations. For the former, the goal is to maximize social welfare, while for the latter, reducing the total cost is the emphases.

5.1 Social Welfare

First, we investigate the social welfare, which is the sum of all the customers' expected net benefits. The expression of the social welfare is given by

$$\begin{aligned}
S(\lambda_1,\lambda_2) &= \varLambda\left(R-\frac{C}{\mu}\right)P_0(1)+\lambda_1(R-CW_1)P_1(1)+\lambda_2(R-CW_2)P_2(1)\\
&= \varLambda\left(R-\frac{C}{\mu}\right)\cdot\frac{\varLambda\lambda_2\mu+\alpha\mu\theta-\varLambda\lambda_1\alpha-\lambda_1\alpha\theta}{\varLambda\mu\theta-\varLambda^2\lambda_1-\varLambda\lambda_1\theta+\varLambda\lambda_2\mu+\alpha\mu\theta-\varLambda\lambda_1\alpha-\lambda_1\alpha\theta+\varLambda^2\lambda_2+\varLambda\lambda_2\theta}\\
&\quad+\lambda_1\left[R-C\left(\frac{\varLambda+\mu+\theta}{\mu\theta}\cdot\frac{\lambda_2}{\alpha}+\frac{\varLambda+\mu+\theta}{\mu\theta-\varLambda\lambda_1-\lambda_1\theta}+\frac{(\varLambda+\theta)^2+\varLambda\mu}{\mu\theta(\varLambda+\theta)}\right)\right]\\
&\quad\times\frac{\varLambda\lambda_2(\varLambda+\theta)}{\varLambda\mu\theta-\varLambda^2\lambda_1-\varLambda\lambda_1\theta+\varLambda\lambda_2\mu+\alpha\mu\theta-\varLambda\lambda_1\alpha-\lambda_1\alpha\theta+\varLambda^2\lambda_2+\varLambda\lambda_2\theta}\\
&\quad+\lambda_2\left[R-C\left(\frac{\varLambda+\mu+\theta}{\mu\theta}\cdot\frac{\lambda_2}{\alpha}+\frac{\varLambda\alpha+\alpha\mu+\alpha\theta+\mu\theta}{\alpha\mu\theta}\right)\right]\\
&\quad\times\frac{\varLambda\mu\theta-\varLambda^2\lambda_1-\varLambda\lambda_1\theta}{\varLambda\mu\theta-\varLambda^2\lambda_1-\varLambda\lambda_1\theta+\varLambda\lambda_2\mu+\alpha\mu\theta-\varLambda\lambda_1\alpha-\lambda_1\alpha\theta+\varLambda^2\lambda_2+\varLambda\lambda_2\theta}.
\end{aligned}\tag{5.1}$$

The purpose of the social planner is to maximize $S(\lambda_1, \lambda_2)$ by seeking the optimal arrival rates λ_1^*, λ_2^*, thus we can model the socially optimal problem as $\max_{\lambda_1,\lambda_2\in[0,\varLambda]} S(\lambda_1, \lambda_2)$. Unfortunately, the expression of $S(\lambda_1, \lambda_2)$ seems especially complex. Through trial and explore, we have to admit that it's too difficult to find the analytic solution. However, after consulting the literatures and analyzing our own model, we find it appropriate to adopt the Particle Swarm Optimization (PSO) algorithm to derive the numerical solutions.

At the mention of PSO algorithm, the most significant advantage is that it does not require too much analyticity of the objective function. It is an optimization algorithm based on swarm intelligence theory. In each iteration search process, Particles in swarm can dynamically adjust their position and speed by tracking two extremes of swarm: The best solution found by the particle itself, namely P-best, and the best solution found by the swarm, namely G-best.

Through multiple iterations, the global optimal solution can be found. The specific process is presented in numerical examples.

5.2 Cost Analysis

In this subsection, we carry out the cost analysis from the service provider's point. The expenses come from many aspects, but can be simplified into the following four items:

C_k= cost of the server per unit time for keeping a customer in the orbit;

C_b= cost of the server per unit time for providing service;

C_s= cost of the server per unit time during the set-up period;

C_θ= cost of the server per unit time for retry.

Therefore we set up the expected total cost function per unit time as

$$C(\lambda_1,\lambda_2)=C_k[P_0'(1)+P_1'(1)+P_2'(1)]+C_b\mu+C_s\alpha+C_\theta\theta, \tag{5.2}$$

where $P_0'(1)$, $P_1'(1)$ and $P_2'(1)$ are given by (3.19)$\sim$(3.21). Aiming to minimize the total cost, service providers seek for the cost-optimal arrival rates. First we investigate the situation that customers are not allowed to balk. In this case, we denote that $\lambda_1=\lambda_2=\lambda$, and the cost function is written as:

$$\begin{aligned} &C(\lambda)\\ &=C_b\mu+C_s\alpha+C_\theta\theta+C_k\cdot\frac{\alpha\mu\theta-\lambda^2\alpha-\lambda\alpha\theta}{\lambda^2(\mu-\alpha)+\lambda\theta(\mu-\alpha)+\alpha\mu\theta}\\ &\times\left[\frac{-\lambda^2\theta(\lambda+\mu+\theta)}{(\lambda+\theta)(\alpha\mu\theta-\lambda^2\alpha-\lambda\alpha\theta)}+\frac{\lambda^2(\lambda+\mu+\theta)(\lambda\mu\theta+\mu\theta-\lambda^2\theta-\lambda^3)}{(\alpha\mu\theta-\lambda^2\alpha-\lambda\alpha\theta)^2}+\frac{\lambda^2(\lambda+\alpha+\theta)}{\alpha^2(\lambda+\theta)}\right]. \end{aligned} \tag{5.3}$$

Due to the highly nonlinearity and complexity of the expression, we adopt the quadratic interpolation method to give the numerical solution. The basic idea of quadratic interpolation method is the continuous use of quadratic polynomials to approximate objective function $C(s)$ in search intervals, and the minimum point of objective function is gradually approximated by the minimum point of interpolation polynomial. The main steps are as follows:

Step 0 Set the initial three points $s_0<s_1<s_2$, which satisfy $C(s_1)<C(s_0)$, and $C(s_2)<C(s_0)$. Choose the stopping tolerance as $\epsilon=10^{-4}$.

Step 1 If $|s_2-s_0|\le\epsilon$, stop and output that $s^*\approx s_1$.

Step 2 According to the interpolation formula $\overline{s}=\frac{(s_1+s_2)C(s_0)-2(s_0+s_2)C(s_1)+(s_0+s_1)C(s_2)}{2(C(s_0)-2C(s_1)+C(s_2))}$, compute $\overline{s}$ and $C(\overline{s})$. If $C(s_1)<C(\overline{s})$, turn to Step 4; otherwise go to Step 3.

Step 3 If $s_1>\overline{s}$, update $s_2=s_1$, $s_1=\overline{s}$, $C(s_2)=C(s_1)$, $C(s_1)=C(\overline{s})$, and turn to Step 1; if not, update $s_0=s_1$, $s_1=\overline{s}$, $C(s_0)=C(s_1)$, $C(s_1)=C(\overline{s})$, and then turn to Step 1.

Step 4 If $s_1<\overline{s}$, let $s_2=\overline{s}$, $C(s_2)=C(\overline{s})$ and go to Step 1; otherwise let $s_0=\overline{s}$, $C(s_0)=C(\overline{s})$, and then turn to Step 1.

Table 1 The quadratic interpolation method in the constant retrial queue with set-up time

Iterations	s_0	s_1	s_2	$\overline{s}$	$f(\overline{s})$	Tolerance
0	0.2000	0.8000	1.4000	0.2000	44.9939	1.2000
1	0.7572	0.8000	0.8428	0.7572	44.9172	0.0855
2	0.0637	0.8000	1.5363	0.0637	44.9994	1.4725
3	0.5711	0.8000	1.0289	0.5711	44.9506	0.4579
4	0.3757	0.8000	1.2243	0.3757	44.8764	0.8486
5	1.0166	1.1204	1.2243	1.1204	44.8764	0.2077
6	1.0311	1.0758	1.0758	1.0758	44.8752	0.0447
7	1.0534	1.0758	1.0981	1.0981	44.8754	0.0447
8	1.0758	1.0804	1.0850	1.0804	44.8751	0.0092
9	1.0804	1.0808	1.0812	1.0804	44.8751	0.0008
10	1.0808	1.0808	1.0808	**1.0808**	**44.8751**	0.0000

Here we assume that $C_k = 1$, $C_b = 2$, $C_s = 7$, $C_\theta = 2$. Using the software Matlab, after several iterations, the cost-optimal arrival rate is shown in Table 1 with the error controlled by $\epsilon = 10^{-4}$. It is clearly that the solution converges to $\lambda_c^* = 1.0808$, and the minimal cost is $C(\lambda_c^*) = 44.8751$.

Now, we turn our attention to the more general case, in which customers are allowed to balk, then the cost function is given by

$$C(\lambda_1, \lambda_2) = C_b\mu + C_s\alpha + C_\theta\theta + C_k \cdot p(0,0) \cdot \left[\frac{-\varLambda\lambda_2\theta(\varLambda + \mu + \theta)}{(\varLambda + \theta)(\alpha\mu\theta - \varLambda\lambda_1\alpha - \lambda_1\alpha\theta)} \right. \\ \left. + \frac{\varLambda\lambda_2(\varLambda + \mu + \theta)(\lambda_2\mu\theta + \alpha\mu\theta - \lambda_1\lambda_2\theta - \varLambda\lambda_1\lambda_2)}{(\alpha\mu\theta - \varLambda\lambda_1\alpha - \lambda_1\alpha\theta)^2} + \frac{\varLambda\lambda_2}{\alpha(\varLambda + \theta)} + \frac{\varLambda\lambda_2}{\alpha^2}\right]. \quad (5.4)$$

Similar to the analysis of the social welfare, we can also use the PSO algorithm to solve the optimization problem. The difference is that the cost-optimal arrival rates we seek are to minimize the objective function. To do this, we just need to add a minus sign before the objective function $C(\lambda_1, \lambda_2)$ in the corresponding program. Section 6 gives the specific analysis process.

6 Numerical Examples

In this section, we use numerical experiments to intuitively reflect the impact of main parameters, i.e., R, α, μ, θ, on the discussed arrival rates respectively. The expressions of the individual equilibrium arrival rate (λ_1^e, λ_2^e) are clearly given by Theorem 3. As for the socially optimal arrival rate (λ_1^*, λ_2^*) and the cost-optimal arrival rate (λ_{1c}^*, λ_{2c}^*), we have to adopt the PSO algorithm to analyze.

6.1 To the Individual Equilibrium Arrival Rate

We choose one of the cases in Theorem 3 to explain the influence of the related parameters on $(\lambda_1^e, \lambda_2^e)$. The graphical representations of the main results are shown in Figures 2∼4.

Firstly, it can be known from Figure 2 that there exists the opposite tendency of λ_1^e and λ_2^e with regard to α. It's reasonable that with the decreasing of the set-up time, customers finding the server at state 2 are more willing to enter the system. On the other hand, as long as the net benefit is positive, selfish customers will always choose to enter, which results in the congestion of the system. Therefore, the customer's arrival rate at busy state decreases correspondingly.

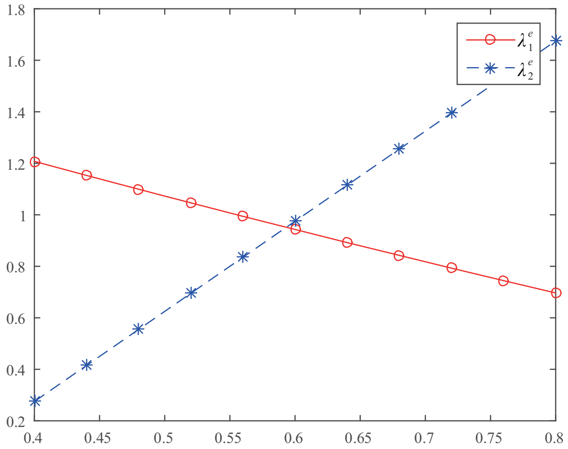


Figure 2 The individual equilibrium arrival rates λ_1^e, λ_2^e with respect to α when $R = 8, C = 2, \Lambda = 2, \mu = 3, \theta = 3$

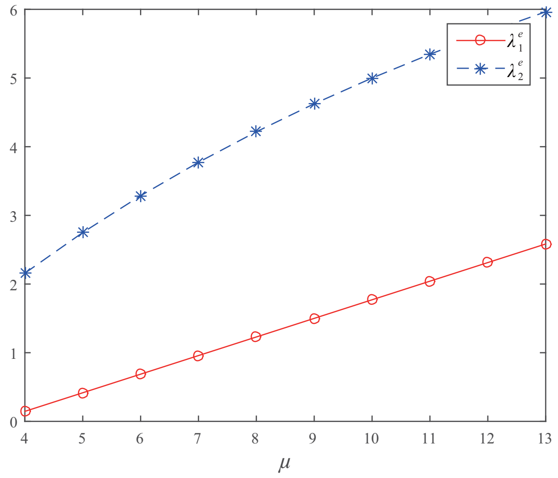


Figure 3 The individual equilibrium arrival rates λ_1^e, λ_2^e with respect to μ when $R = 8, C = 2, \Lambda = 10, \alpha = 1, \theta = 5$

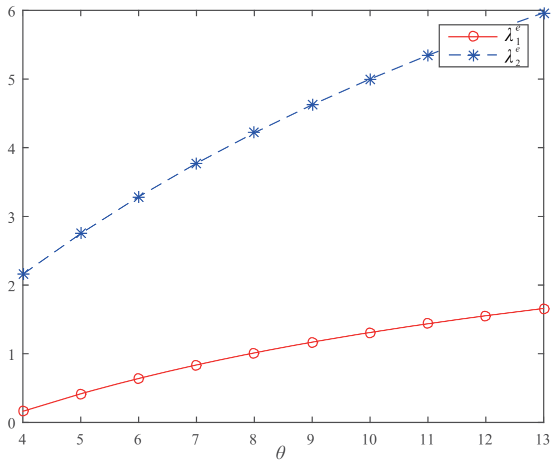


Figure 4 The individual equilibrium arrival rates λ_1^e, λ_2^e with respect to θ when $R = 8, C = 2, \Lambda = 10, \alpha = 1, \mu = 5$

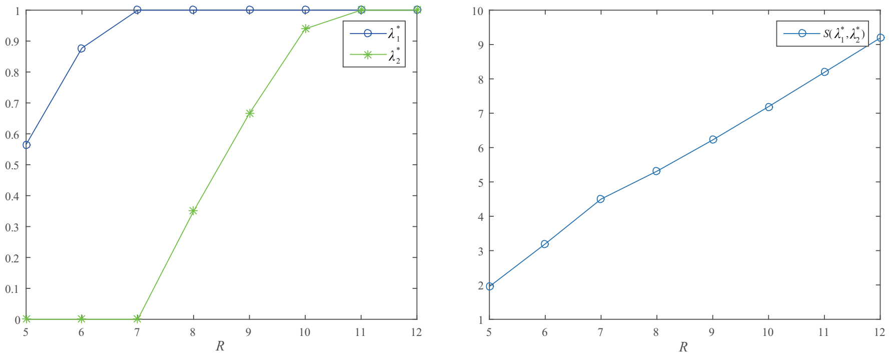
Secondly, it is easily understood that the service rate μ and the retrial rate θ have the same influence on λ_1^e and λ_2^e. Because the expected waiting time decreases as the decline of the mean service time or the retrial interval, customers spend less time and expenses in queue, surely the equilibrium arrival rates will increase. Figures 3∼4 also support this point.

6.2 To the Socially Optimal Arrival Rate

In this section, we adopt the PSO algorithm to investigate the effect of parameters R, α, μ, θ on the socially optimal arrival rates $(\lambda_1^*, \lambda_2^*)$ and social welfare $S(\lambda_1^*, \lambda_2^*)$, respectively.

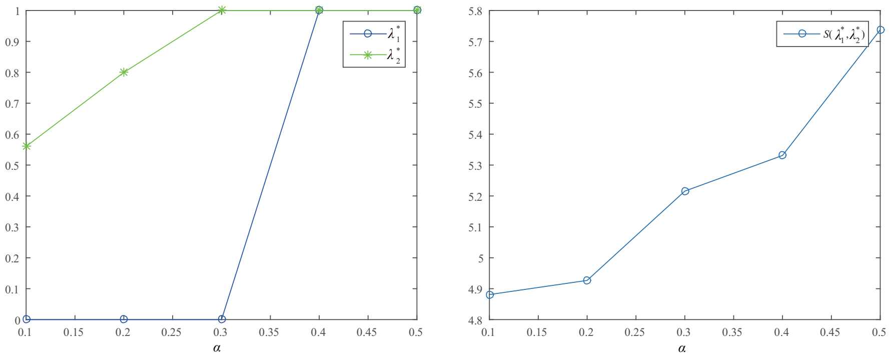
Firstly, it is obvious that the socially optimal arrival rates and the social welfare are all increase with reward R when $R > 7$. When $R \leq 7$, the socially optimal arrival rate at state 2 is zero. The reason is that the net payoff of the latter arrivals would result in the loss of the earlier. From the social manager's point of view, since the reward is less than the loss, he would rather there is no customer entering the orbit during set-up time. However, when R is large enough, the total benefits brought by the customers are always higher than the cost, to maximize the social welfare, social manager would encourage people to join the system. Therefore, λ_1^* and λ_2^* achieve the maximization Λ gradually. Noticed that, $S(\lambda_1^*, \lambda_2^*)$ is almost increasing linearly with R due to its structure.

Secondly, Figure 6 shows that both λ_1^*, λ_2^* and $S(\lambda_1^*, \lambda_2^*)$ are increasing with parameter α in general, for the mean sojourn time decreases with set-up time, and the net benefits increases correspondingly. But compared with λ_2^*, λ_1^* reached its maximum earlier. The reason is that the server doesn't provide service during set-up time. When $\alpha < 0.3$, the cost incurred by the customers entering at state 2 is more than the benefit they brought, so at this situation, social manager prefers there is no enter when the server is setting up. However, since $\alpha \geq 0.3$, λ_2^* increases rapidly to its maximum. Because from then on, the set up time is sufficiently short, customers arriving at state 2 don't have to wait so long, and the net benefits exceed their waiting cost.



(a) (b)

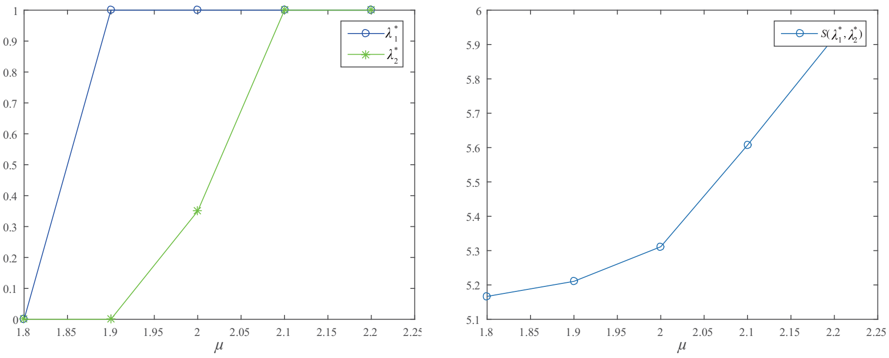
Figure 5 The effect of R on the socially optimal arrival rates λ_1^*, λ_2^* and social welfare $S(\lambda_1^*, \lambda_2^*)$ for $\Lambda = 1, \mu = 2, \alpha = 3, \theta = 3, C = 2$



(a) (b)

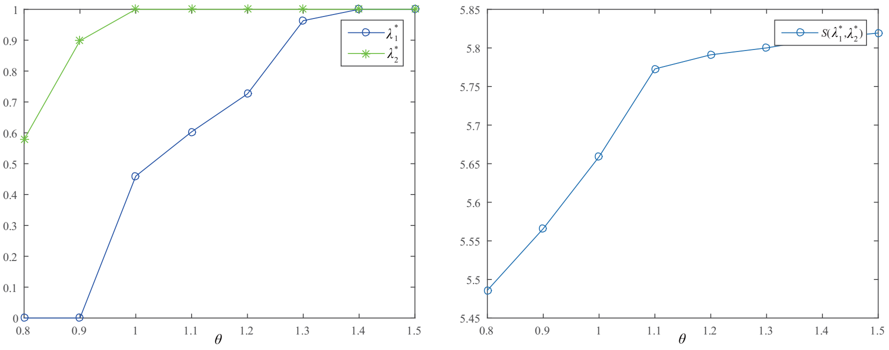
Figure 6 The effect of α on the socially optimal arrival rates λ_1^*, λ_2^* and social welfare $S(\lambda_1^*, \lambda_2^*)$ for $\Lambda = 1, \mu = 3, \theta = 3, R = 8, C = 2$

Finally, similar to the analysis in Section 6.1, λ_1^* and λ_2^* have the same trends with respect to μ and θ, and Figures 7∼8 also show the evidence. So, we only study the impact of the service rate μ on λ_1^*, λ_2^* and $S(\lambda_1^*, \lambda_2^*)$. From Figure 7(a) we can see that when $\mu < 1.8$, both λ_1^* and λ_2^* are equal to 0. Intuitively, the smaller service rate results in the negative expected benefit of all the customers. When $1.8 \leq \mu < 1.9$, λ_1^* begin to increase while λ_2^* is still equal to 0. Because the positive net payoff of customers arriving after the system activated can offset the negative effect of customers arriving at set-up period. As $\mu > 1.9$, the service rate is large enough to avoid the congestion of the system, therefore λ_2^* also increases gradually to its maximum. Figure 7(b) shows that the social welfare is strictly increasing with respect to μ, since the larger the service rate, the shorter the mean service time. Therefore customers spend less time in queue and incur less cost.



(a) (b)

Figure 7 The effect of μ on the socially optimal arrival rates λ_1^*, λ_2^* and social welfare $S(\lambda_1^*, \lambda_2^*)$ for $\Lambda = 1, \alpha = 3, \theta = 3, R = 8, C = 2$



(a) (b)

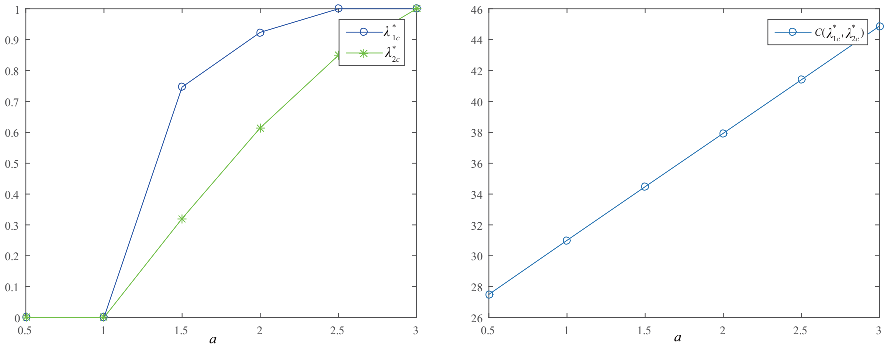
Figure 8 The effect of θ on the socially optimal arrival rates λ_1^*, λ_2^* and social welfare $S(\lambda_1^*, \lambda_2^*)$ for $\Lambda = 1, \mu = 3, \alpha = 3, R = 8, C = 2$

6.3 To the Cost-Optimal Arrival Rate

Now we focus on the impact of parameters α, μ, θ from the perspective of the service provider. Through a large number of numerical experiments, we found that compared with other parameters, when θ takes a smaller value, the total cost $C(\lambda_1, \lambda_2)$ always reaches its minimum at $\lambda_2 = 0$. That is because the retrial interval is too large for the server that will cause congestion of the system. Meanwhile, customers arriving at set-up time will only aggravate this problem since the server can't provide service during this period. Therefore, from the service provider's point, they would better prevent customers from entering at the set-up time. But when θ is large enough, it is shown that λ_{1c}^*, λ_{2c}^* and $C(\lambda_{1c}^*, \lambda_{2c}^*)$ have the following tendencies.

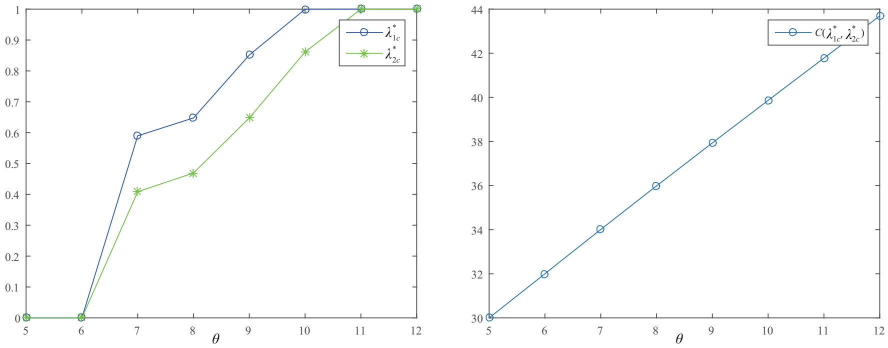
Figure 9(a) shows that when $\alpha < 1$, due to the long set-up time, the set-up cost is too high for the service provider, so the system is better shuttled down. As α increases, the cost-optimal arrival rates and the total cost all increase accordingly. Because the set-up cost decreases, and the service provider needs the reward brought by customers to offset the cost. Thus customers are encouraged to enter the system and λ_{1c}^*, λ_{2c}^* increase. Figure 9(b) shows that as the number of customers increase, the cost for remaining customers in the orbit rises, and becomes the dominant player. Therefore, although the set-up cost decrease, the total cost continues to rise.

Figure 10 shows the tendency of λ_{1c}^*, λ_{2c}^* and $C(\lambda_{1c}^*, \lambda_{2c}^*)$ about parameter θ (μ has similar effect). As analyzed before, when $\theta \leq 6$ the cost-optimal arrival rates are $\lambda_1 \in [0, \Lambda]$, $\lambda_2 = 0$. As the increase of θ, the retrial interval gets shorter and the retrial cost decreases. Relatively the set-up cost is higher. Therefore, to avoid frequent set-up periods, the service provider should encourage potential customers to enter the system. Accordingly, the trends of λ_{1c}^*, λ_{2c}^* are upward curves. Meanwhile, more customers in the orbit means more keeping cost, so $C(\lambda_{1c}^*, \lambda_{2c}^*)$ increases correspondingly.



(a) (b)

Figure 9 The effect of α on the socially optimal arrival rates λ_{1c}^*, λ_{2c}^* and total cost $C(\lambda_{1c}^*, \lambda_{2c}^*)$ for $\Lambda = 1, \mu = 3, \theta = 9, C_b = 2, C_s = 7, C_\theta = 2, C_k = 1$



(a) (b)

Figure 10 The effect of θ on the socially optimal arrival rates λ_{1c}^*, λ_{2c}^* and total cost $C(\lambda_{1c}^*, \lambda_{2c}^*)$ for $\Lambda = 1, \mu = 3, \alpha = 2, C_b = 2, C_s = 7, C_\theta = 2, C_k = 1$

7 Conclusion

In this paper, we studied the M/M/1 constant retrial queue with balking and set-up time. When the server's state is observable while the queue length is unknown, the individual equilibrium strategies and social optimization problems are considered. Firstly, under the stability condition, we derived the customers' mean sojourn times for different states, and obtained the individual equilibrium arrival rates by analyzing the benefit function of arriving customers. After that, the expression of the social welfare is given. From the point of the social manager, the optimal arrival rates are presented by PSO algorithm. Furthermore, we carried out cost analysis from the point of the service provider, and derived the cost-optimal arrival rates in two cases that whether customers are allowed to balk or not, respectively. Finally, the numerical examples were given to show the sensitivity of main system performance measures.

To further study, this work can be generalized in different directions. One extension is to consider the general service times. In addition, the optimal pricing strategies that equate the customers' equilibrium arrival rates to the socially optimal arrival rates can be studied. Furthermore, the equilibrium analysis in other information levels could also be considered.

Acknowledgements The authors gratefully acknowledge the editor and two anonymous referees for their insightful comments and helpful suggestions that led to a marked improvement of the article.

References

[1] Falin G, Templeton J. Retrial queues. Monographs on Statistics & Applied Probability, 1997.

[2] Artalejo J R. A queueing system with returning customers and waiting line. Operations Research Letters, 1995, 17: 191–199.

[3] Artalejo J R, Gomez-Corral A. Steady state solution of a single-server queue with linear repeated requests. Journal of Applied Probability, 1997, 34: 223–233.

[4] Artalejo J R, Gomez-Corral A. Analysis of a stochastic clearing system with repeated attempts. Stochastic Models, 1998, 14: 623–645.

[5] Fayolle G. A simple telephone exchange with delayed feedbacks. Proceedings of the International Seminar on Teletraffic Analysis and Computer Performance Evaluation, 1986: 245–253.

[6] Falin G I. The M/M/1 retrial queue with retrials due to server failures. Queueing Systems, 2008, 58: 155–160.

[7] Artalejo J R, Gomez-Corral A. Retrial Queueing Systems: A Computational Approach. Springer, Berlin, 2008.

[8] Naor P. The regulation of queue size by levying tolls. Econometrica, 1969, 37: 15–24.

[9] Edelson Noel M, Hilderbrand D K. Congestion tolls for Poisson queuing processes. Econometrica, 1975, 43(1): 81–92.

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

[11] Zhang Y, Wang J. Equilibrium pricing in an M/G/1 retrial queue with reserved idle time and setup time. Applied Mathematical Modelling, 2017, 49: 514–530.

[12] Yutaka S, Yoshitaka T, Yutaka T, et al. A composite queue with vacation/set-up/close-down times for svcc in IP over atm networks. Journal of the Operations Research Society of Japan, 1998, 41(1): 68–80.

[13] Economou A, Kanta S. Equilibrium customer strategies and social-profit maximization in the single-server constant retrial queue. Naval Research Logistics, 2011, 58(2): 107–122.

[14] Hassin R, Haviv M. To Queue or not to queue: Equilibrium behavior in queueing systems. Springer US, 2003.