RAIRO-Oper. Res. 55 (2021) 2807–2825
https://doi.org/10.1051/ro/2021130

RAIRO Operations Research
www.rairo-ro.org

EQUILIBRIUM ANALYSIS OF CLOUD USER REQUEST BASED ON THE MARKOV QUEUE WITH VARIABLE VACATION AND VACATION INTERRUPTION

Yitong Zhang[1] and Xiuli Xu[2,*]

Abstract. This paper considers the equilibrium balking behavior of customers in a single-server Markovian queue with variable vacation and vacation interruption, where the server can switch across four states: vacation, working vacation, idle period, and busy period. Once the queue becomes empty, the server commences a working vacation and slows down its service rate. However, this period may be interrupted anytime by the vacation interruption. Upon the completion of a working vacation, the server takes a vacation in a probability-based manner and stops service if the system is empty. The system stays idle after a vacation until a new customer arrives. The comparisons between the equilibrium balking strategy of customers and the optimal expected social benefit per time unit for each type of queue are elucidated and the inconsistency between the individual optimization and the social optimization is revealed. Moreover, the sensitivity of the expected social benefit and the equilibrium threshold with respect to the several parameters as well as diverse precision levels is illustrated through numerical examples in a competitive cloud environment.

Mathematics Subject Classification. 60K25, 90B22.

Received January 11, 2021. Accepted August 14, 2021.

1. Introduction

Queueing economics has received increasing attention in recent decades. It was firstly introduced by Naor [12], who proved that the decisions of individual customers often deviate from the whole interests in that system. Subsequently, Economou and Kanta [3] examined the equilibrium strategy behavior of customers in a fully observable queue with service failures. Zhang and Wang [23] considered the equilibrium strategy in a queue with fault and repairable server and obtained the system's performance indices.

In order to increase profits, the servers may do some auxiliary or part-time jobs, which temporarily prevent the servers from providing services for some period, therefore, the vacation policy is introduced. Vacation queueing systems have been spotlighted due to their versatility and applicability. Readers can refer to the monographs of Takagi [18] and Tian and Zhang [19] for the fundamental results and comprehensive surveys. Burnetas and Economou [1] introduced the economic analysis of a vacation queue and compared four precision levels of a

Keywords. Variable vacation, vacation interruption, equilibrium strategy, the expected social benefit.

[1] School of Economics and Management, Yanshan University, Hebei Qinhuangdao 066004, P.R. China
[2] School of Science, Yanshan University, Hebei Qinhuangdao 066004, P.R. China
*Corresponding author: xxl-ysu@163.com

Markov chain with setup times. Wang *et al.* [21] analyzed the equilibrium strategies of customers in terms of the overall welfare in an M/M/k queue with asynchronous or synchronous multiple vacations. Motivated by modeling the data transmission in telecommunication networks, Peng and Wu [14] studied a Lévy-driven stochastic queueing system where the server may be subject to breakdowns and repairs. Because introducing a vacation policy into the inventory system has more realistic significance, Padmavathi *et al.* [13] considered two different vacation policies in the inventory queueing system, where the number of demands in the orbit and the joint probability distribution of the inventory level are obtained by using the method of matrix-geometric solution. Melikov *et al.* [11] proposed a three-dimensional Markov vacation system with perishable inventory and developed an approximate method based on the hierarchical merging of the states, which can be applied to the asymptotic analysis of any dimension.

For most of the management systems, the vacation of servers sometimes results in the excessive load of the system. To address this issue, Servi and Finn [15] presented the working vacation policy where the server does not completely deactivate. Zhang *et al.* [24] illustrated four equilibrium strategies and provided the optimal thresholds in a multiple working vacation queue. Li [8] demonstrated an exact steady-state analysis in a discrete-time Geo/G/1 queue with working vacation and provided an application to the network scheduling in the wavelength division multiplexed (WDM) system. Recently, Sun *et al.* [17] dynamically compared the performance and the threshold policy of the exhaustive or non-exhaustive M/M/1/N queue with working vacations. Do *et al.* [2] investigated the M/M/1 retrial queue with working vacation and the optimal strategies for joining the system were obtained. Tian and Wang [20] combined multiple classical vacations with working vacation and made the economic analysis of this model in the unobservable cases.

Despite these considerable advantages, the modern management system is still far from a sensitive response. In practice, the service agencies and network systems must have the ability to deal with emergencies and require the servers to return to work rather than remain on leave, which is called the vacation interruption strategy, that is, if there are still a certain number of customers in the system after a service, the vacation may be suspended immediately and a normal busy period starts in some cases. Lee [7] explored the economic analysis of an M/M/1 queueing system with single vacation or multiple vacations and vacation interruption. Li *et al.* [9] discussed a model with working vacation and vacation interruption under four precision levels. Shekhar *et al.* [16] concerned a randomized arrival control policy for prospective customers in a queueing system with working vacation and vacation interruption.

The growing demands of cloud computing have contributed to a broadened scale in cloud data centers. Balancing the computing proliferation of cloud users and energy consumption in cloud platforms has inspired a hotspot field of cloud computing [4, 10, 22]. Because idle virtual machines falling into deep dormancy effectively reduce energy consumption in the traditional cloud computing platforms, Jin *et al.* [6] proposed an allocation strategy for the clustered virtual machine and constructed a cost function to balance different system performances. Furthermore, a mechanism of semi dormancy is considered in [5], where a virtual machine provides a lower service rate for cloud user requests, which significantly enhances the energy efficiency in the cloud environment.

However, to the best of our knowledge, none of the above models incorporated the classical vacation, working vacation and vacation interruption together in a cloud computing platform. Motivated by the allocation strategy for cloud data centers, this study attempts to analyze a variable dormancy cloud strategy with sleep-wake control. The virtual machine only enters the semi dormancy state when the cloud platform has no user request. If the buffer has a waiting cloud user request, the virtual machine immediately changes from the semi dormancy state to the wake-up state and provides the normal service for cloud users. Otherwise, it continues to stay in the semi dormancy. When a semi dormancy state is completed and there exist any requests, the virtual machine participates in the wake-up state. Otherwise, it joins the deep dormancy state for a certain period with a specified probability. The cloud users and platforms display distinct behavior to avoid the crowd and maximize their own profits. Based on the above, we derive the expected social benefit using a linear revenue and expenditure structure and discuss the optimal benefit per time unit for the cloud system in the fully or partly

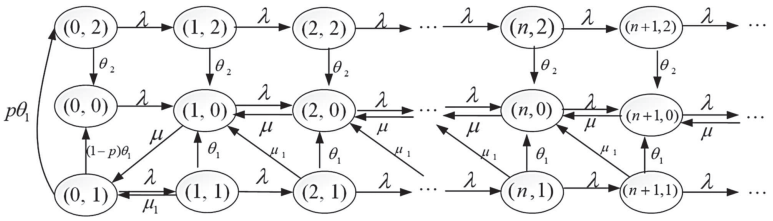

FIGURE 1. State transition diagram in an M/M/1 queue with variable vacation and vacation interruption.

observable cases, respectively. Our model effectively provides a theoretical basis and potential application for optimizing user access control in a cloud computing environment.

The rest of this paper is organized as follows: Section 2 describes the queueing model and different information precision levels. Sections 3–4 focus on the fully observable and partly observable cases, respectively. For each type of situation, the equilibrium balking strategies of customers and the expected social benefits per time unit are discussed. Section 5 illustrates the sensitivity analysis on performance characteristics in different situations. Section 6 summarizes the findings and conclusions of this paper.

2. Model descriptions

This paper introduces a variable vacation and vacation interruption policy in the classical M/M/1 queue with the arrival rate λ and the normal service rate μ. When the system becomes empty, it begins a working vacation and the working vacation time is assumed to be exponentially distributed with parameter θ_1. During this period, a new arrival is served at a low rate $\mu_1(\mu_1 < \mu)$. When a service is completed in the working vacation and customs are waiting in line, the working vacation ends and the server comes back to the normal service level. Otherwise, it takes another working vacation. Once a working vacation ends, the server immediately switches to a regular busy period if the queue is non-empty, otherwise, the system enters a vacation period with probability $p(0 < p \leq 1)$, or enters an idle state with probability $1 - p$, and the vacation time follows an exponential distribution with parameter θ_2. After the vacation finishes, a regular busy period starts if there are customers in the system. Otherwise, the server stays idle until a new customer arrives.

Assume that the interarrival times, the service times, the vacation times and the working vacation times are mutually independent. In addition, the service discipline is first in and first out (FIFO). Denote the number of customers in the system at time t by $N(t)$, and define

$$I(t) = \begin{cases} 0, & \text{the server is busy or stays idle;} \\ 1, & \text{the server is taking a working vacation;} \\ 2, & \text{the server is taking a vacation.} \end{cases}$$

Obviously, the state space of the Markov chain $\{N(t), I(t)\}$ is $\Omega = \{(n, i): n \geq 0, i = 0, 1, 2\}$. The state transition diagram is depicted in Figure 1.

Suppose that each customer can receive a reward of R utility units after service completion but has to pay a waiting cost of C utility units for waiting a time unit in the system. Assume that $R > C/\mu$ hereafter to ensure customers are attracted to join the system. Customers are risk neutral and maximize their expected net benefit. Moreover, the decisions are irrevocable in that retrials of balking customers and reneging of queueing behavior are not allowed.

There are two situations according to the information level acquired by an arrival.

(1) Fully observable case: Arriving customers are informed about the queue length $N(t)$ and the server state $I(t)$;

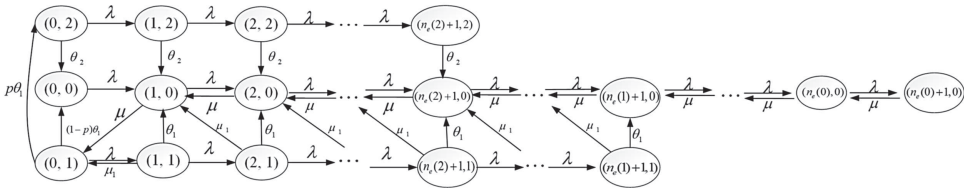

FIGURE 2. State transition diagram in the fully observable case.

(2) Partly observable case: Arriving customers are informed about the queue length $N(t)$ only;

For convenience, denote

$$\rho = \frac{\lambda}{\mu}, \quad \sigma_1 = \frac{\lambda}{\lambda + \mu_1 + \theta_1}, \quad \sigma_2 = \frac{\lambda}{\lambda + \theta_2}.$$

3. Equilibrium strategy in fully observable case

3.1. Equilibrium threshold strategy

Suppose that an arriving customer encounters the system state (n, i) and decides to enter, his expected net benefit after the service completion is $S_{\text{fo}}(n, i) = R - CT_{\text{fo}}(n, i)$, where $T_{\text{fo}}(n, i)$ represents his mean sojourn time in the system. There exists a balk threshold $n_e(i)$, meaning that an arriving customer enters the system in state i $(i = 0, 1, 2)$ if the number of customers is less than the given threshold. Therefore, a pure threshold strategy is denoted by the tuple $(n_e(0), n_e(1), n_e(2))$ and the balking strategy has the following form: "While arriving at time t, observe: enter if $N(t) \le n_e(I(t))$ and balk otherwise." An arrival who finds the same number of customers in the system prefers to join the queue during the busy or idle period than during the vacation period or working vacation period. Therefore, the thresholds in different states satisfy $n_e(2) \le n_e(1) \le n_e(0)$. The state transition diagram is depicted in Figure 2.

Let $\lfloor x \rfloor$ represent the largest integer up to x. Then we get the following result.

Theorem 3.1. *In the fully observable M/M/1 queue with variable vacation and vacation interruption, there exist thresholds*

$$(n_e(0), n_e(1), n_e(2)) = \left(\left\lfloor \mu \frac{R}{C} \right\rfloor - 1, \left\lfloor \mu \left(\frac{R}{C} - \frac{\mu + \theta_1}{\mu(\mu_1 + \theta_1)} \right) \right\rfloor - 1, \left\lfloor \mu \left(\frac{R}{C} - \frac{1}{\theta_2} \right) \right\rfloor - 1 \right),$$

which satisfy the unique Nash equilibrium strategy that a customer enters if $N(t) \le n_e(I(t))$ and balks otherwise when he observes the system is in state $(N(t), I(t))$.

Proof. From assumption, we have

$$\begin{aligned}
T_{\text{fo}}(0,0) &= \frac{1}{\mu}, \\
T_{\text{fo}}(0,1) &= \frac{1}{\mu_1 + \theta_1} + \frac{\theta_1}{\mu_1 + \theta_1} T_{\text{fo}}(0,0), \\
T_{\text{fo}}(0,2) &= \frac{1}{\theta_2} + \frac{1}{\mu}, \\
T_{\text{fo}}(n,0) &= \frac{1}{\mu} + T_{\text{fo}}(n-1,0), \quad n = 1, 2, \ldots \\
T_{\text{fo}}(n,1) &= \frac{1}{\mu_1 + \theta_1} + \frac{\mu_1}{\mu_1 + \theta_1} T_{\text{fo}}(n-1,0) + \frac{\theta_1}{\mu_1 + \theta_1} T_{\text{fo}}(n,0), \quad n = 1, 2, \ldots \\
T_{\text{fo}}(n,2) &= \frac{1}{\theta_2} + T_{\text{fo}}(n,0), \quad n = 1, 2, \ldots
\end{aligned}$$

After manipulating, we get

$$T_{\text{fo}}(n,0)=\frac{n+1}{\mu},T_{\text{fo}}(n,2)=\frac{n+1}{\mu}+\frac{1}{\theta_2},T_{\text{fo}}(n,1)=\frac{\mu+\theta_1}{\mu(\mu_1+\theta_1)}+\frac{n}{\mu},\quad n=0,1,\ldots,$$

According to the definition, $S_{\text{fo}}(n,i)$ is a monotonically decreasing function of n as $T_{\text{fo}}(n,i)$ is strictly increasing of n. Besides, customers prefer to enter when $S_{\text{fo}}(n,i)>0$. The thresholds are obtained by substituting $T_{\text{fo}}(n,i)$ into $S_{\text{fo}}(n,i)$. Theorem 3.1 is proved. □

Specially, if $p\to 0$, we can get the thresholds $(n_e(0),n_e(1))$, which tend to the results derived by Li *et al.* [9].

3.2. Stationary queue length distribution

Denote the steady state probabilities in the fully observable situation by $p_{\text{fo}}(n,i)=\lim\limits_{t\to+\infty}p_{\text{fo}}\{N(t)=n, I(t)=i\},\ (n,i)\in\Omega$, then we have

Theorem 3.2. *In the fully observable M/M/1 queue with variable vacation and vacation interruption where customers enter the system according to a pure threshold strategy $(n_e(0),n_e(1),n_e(2))$, the stationary distribution is as follows*

$$p_{\text{fo}}(n,1)=\frac{\lambda}{\theta_1(1-p\sigma_2)}\sigma_1^n p_{\text{fo}}(0,0),\quad n=0,1,\ldots,n_e(1),\tag{3.1}$$

$$p_{\text{fo}}(n_e(1)+1,1)=\frac{\lambda\sigma_1}{\theta_1(1-p\sigma_2)(1-\sigma_1)}\sigma_1^{n_e(1)}p_{\text{fo}}(0,0),\tag{3.2}$$

$$p_{\text{fo}}(n,2)=\frac{p\sigma_2}{1-p\sigma_2}\sigma_2^n p_{\text{fo}}(0,0),\quad n=0,1,\ldots,n_e(2),\tag{3.3}$$

$$p_{\text{fo}}(n_e(2)+1,2)=\frac{p\sigma_2^2}{(1-p\sigma_2)(1-\sigma_2)}\sigma_2^{n_e(2)}p_{\text{fo}}(0,0),\tag{3.4}$$

$$p_{\text{fo}}(n,0)=B_{\text{fo}}\rho^n+C_{\text{fo}}\sigma_1^n+D_{\text{fo}}\sigma_2^n,\quad n=0,1,\ldots,n_e(0),\tag{3.5}$$

$$\begin{aligned}p_{\text{fo}}(n_e(0)+1,0)&=\rho\left(B_{\text{fo}}\rho^{n_e(0)}+C_{\text{fo}}\sigma_1^{n_e(0)}+D_{\text{fo}}\sigma_2^{n_e(0)}\right)\\&\quad+\left(\frac{\lambda\theta_1\sigma_1^{n_e(1)+2}}{\mu\theta_1(1-p\sigma_2)(1-\sigma_1)}+\frac{p\theta_2\sigma_2^{n_e(2)+1}}{\mu(1-p\sigma_2)}\right)p_{\text{fo}}(0,0),\end{aligned}\tag{3.6}$$

where

$$\begin{aligned}p_{\text{fo}}(0,0)=\Bigg\{&1+\frac{(\mu_1\sigma_1+\theta_1)\sigma_1\rho}{\theta_1(1-p\sigma_2)(\rho-\sigma_1)(1-\sigma_1)}\left(\frac{\rho\sigma_1^{n_e(0)+1}-\sigma_1}{1-\sigma_1}+\frac{(\rho-\sigma_1)(\lambda+\theta_1)\rho\sigma_1^{n_e(1)+2}}{(\mu_1\sigma_1+\theta_1)\rho\sigma_1}\right)\\&+\frac{p\rho\sigma_2}{(1-p\sigma_2)(\rho-\sigma_2)}\left(\frac{\rho\sigma_2^{n_e(0)+1}-\sigma_2}{1-\sigma_2}+\frac{1}{\rho}+\sigma_2^{n_e(2)}(\rho-\sigma_2)\right)+\left(\frac{\rho(\lambda+\theta_1)\sigma_1}{\theta_1(1-p\sigma_2)(\rho-\sigma_1)}\right.\\&\left.+\frac{\rho-\sigma_2+p\sigma_2^2}{(1-p\sigma_2)(\rho-\sigma_2)}\right)\frac{\rho(1-\rho^{n_e(0)+1})}{1-\rho}\Bigg\}^{-1},\end{aligned}$$

$$B_{\text{fo}}=\left[\frac{\rho(\lambda+\theta_1)\sigma_1}{\theta_1(1-p\sigma_2)(\rho-\sigma_1)}+\frac{\rho-\sigma_2+p\sigma_2^2}{(1-p\sigma_2)(\rho-\sigma_2)}\right]p_{\text{fo}}(0,0)=p_{\text{fo}}(0,0)-C_{\text{fo}}-D_{\text{fo}},$$

$$C_{\text{fo}}=-\frac{\rho(\lambda+\theta_1)\sigma_1}{\theta_1(1-p\sigma_2)(\rho-\sigma_1)}p_{\text{fo}}(0,0),\quad D_{\text{fo}}=-\frac{p\sigma_2\rho}{(1-p\sigma_2)(\rho-\sigma_2)}p_{\text{fo}}(0,0).$$

Proof. The corresponding stationary distribution can be obtained in Figure 2.

$$\lambda p_{\mathrm{fo}}(0,0) = (1-p)\theta_1 p_{\mathrm{fo}}(0,1) + \theta_2 p_{\mathrm{fo}}(0,2), \tag{3.7}$$

$$\begin{aligned}(\lambda+\mu)p_{\mathrm{fo}}(n,0) &= \mu p_{\mathrm{fo}}(n+1,0) + \mu_1 p_{\mathrm{fo}}(n+1,1) + \theta_1 p_{\mathrm{fo}}(n,1)\\ &\quad +\lambda p_{\mathrm{fo}}(n-1,0) + \theta_2 p_{\mathrm{fo}}(n,2), \quad n = 1,2,\ldots,n_e(0),\end{aligned} \tag{3.8}$$

$$\mu p_{\mathrm{fo}}(n_e(0)+1,0) = \theta_1 p_{\mathrm{fo}}(n_e(1)+1,1) + \lambda p_{\mathrm{fo}}(n_e(0),0) + \theta_2 p_{\mathrm{fo}}(n_e(2)+1,2), \tag{3.9}$$

$$(\lambda+\theta_1)p_{\mathrm{fo}}(0,1) = \mu_1 p_{\mathrm{fo}}(1,1) + \mu p_{\mathrm{fo}}(1,0), \tag{3.10}$$

$$(\lambda+\mu_1+\theta_1)p_{\mathrm{fo}}(n,1) = \lambda p_{\mathrm{fo}}(n-1,1), \quad n = 1,2,\ldots,n_e(1), \tag{3.11}$$

$$(\mu_1+\theta_1)p_{\mathrm{fo}}(n_e(1)+1,1) = \lambda p_{\mathrm{fo}}(n_e(1),1), \tag{3.12}$$

$$(\lambda+\theta_2)p_{\mathrm{fo}}(0,2) = p\theta_1 p_{\mathrm{fo}}(0,1), \tag{3.13}$$

$$(\lambda+\theta_2)p_{\mathrm{fo}}(n,2) = \lambda p_{\mathrm{fo}}(n-1,2), \quad n = 1,2,\ldots,n_e(2), \tag{3.14}$$

$$\theta_2 p_{\mathrm{fo}}(n_e(2)+1,2) = \lambda p_{\mathrm{fo}}(n_e(2),2). \tag{3.15}$$

According to (3.7) and (3.13), we have

$$p_{\mathrm{fo}}(0,1) = \frac{\lambda}{\theta_1(1-p\theta_2)}p_{\mathrm{fo}}(0,0), \quad p_{\mathrm{fo}}(0,2) = \frac{p\sigma_2}{1-p\sigma_2}p_{\mathrm{fo}}(0,0). \tag{3.16}$$

Substituting (3.16) into (3.11) and (3.14), by iterating we have

$$p_{\mathrm{fo}}(n,1) = \left(\frac{\lambda}{\lambda+\mu_1+\theta_1}\right)^n \frac{\lambda}{\theta_1(1-p\sigma_2)}p_{\mathrm{fo}}(0,0), \quad n = 1,2,\ldots,n_e(1),$$

$$p_{\mathrm{fo}}(n,2) = \left(\frac{\lambda}{\lambda+\theta_2}\right)^n \frac{p\sigma_2}{1-p\sigma_2}p_{\mathrm{fo}}(0,0), \quad n = 1,2,\ldots,n_e(2),$$

Combining σ_1 and σ_2, we can get (3.1) and (3.3).

Substituting $n = n_e(1)$ and $n = n_e(2)$ into (3.12) and (3.15), respectively, we obtain

$$p_{\mathrm{fo}}(n_e(1)+1,1) = \left(\frac{\lambda}{\lambda+\mu_1+\theta_1}\right)^{n_e(1)} \frac{\lambda\sigma_1}{\theta_1(1-p\sigma_2)(1-\sigma_1)}p_{\mathrm{fo}}(0,0),$$

$$p_{\mathrm{fo}}(n_e(2)+1,2) = \left(\frac{\lambda}{\lambda+\theta_2}\right)^{n_e(2)} \frac{p\sigma_2^2}{(1-p\sigma_2)(1-\sigma_2)}p_{\mathrm{fo}}(0,0).$$

Manipulating above equations, we get 3.2 and (3.4).

Next, we calculate the formulas of $p_{\mathrm{fo}}(n_e(0)+1,0)$ and $p_{\mathrm{fo}}(n,0), n = 0,1,\ldots,n_e(0)$. From (3.8), we obtain

$$\begin{aligned}&\mu p_{\mathrm{fo}}(n+1,0) - (\lambda+\mu)p_{\mathrm{fo}}(n,0) + \lambda p_{\mathrm{fo}}(n-1,0)\\ &= \left(-\frac{\lambda(\mu_1\sigma_1+\theta_1)}{\theta_1(1-p\sigma_2)}\sigma_1^n - \frac{p\sigma_2\theta_2}{1-p\sigma_2}\sigma_2^n\right)p_{\mathrm{fo}}(0,0), \quad n = 1,2,\ldots,n_e(0).\end{aligned}$$

From above expressions, it follows that $p_{\mathrm{fo}}(n,0), n = 1,2,\ldots,n_e(0)$ are solutions of the following nonhomogeneous linear difference equation with constant coefficients

$$\begin{aligned}\mu x_{n+1} - (\lambda+\mu)x_n + \lambda x_{n-1} = \left(-\frac{\lambda(\mu_1\sigma_1+\theta_1)}{\theta_1(1-p\sigma_2)}\sigma_1^n - \frac{p\sigma_2\theta_2}{1-p\sigma_2}\sigma_2^n\right)p_{\mathrm{fo}}(0,0),\\ n = 1,2,\ldots,n_e(0).\end{aligned} \tag{3.17}$$

Considering their corresponding characteristic equation

$$\mu x^2 - (\lambda + \mu)x + \lambda = 0,$$

which has two roots, 1 and ρ. The general solution of the homogeneous form of (3.17) is $x_n^{\text{hom}} = A_{\text{fo}} + B_{\text{fo}}\rho^n$ due to $\lambda < \mu$, where A_{fo} and B_{fo} are constants. The general solution x_n^{gen} is given by $x_n^{\text{gen}} = x_n^{\text{hom}} + x_n^{\text{spec}}$, where x_n^{spec} is a specific solution of (3.17). The standard method can be used to find a specific solution because the nonhomogeneous part of (3.17) is the sum of the geometric distributions with parameter σ_1 and σ_2, respectively, and $\sigma_1, \sigma_2 \neq \rho, 1$. Then the specific solution of (3.17) has the form $C_{\text{fo}}\sigma_1^n + D_{\text{fo}}\sigma_2^n$, where C_{fo} and D_{fo} are constants. Substituting the specific solution x_n^{spec} into (3.17), we obtain

$$C_{\text{fo}} = -\frac{\rho(\lambda + \theta_1)\sigma_1}{\theta_1(1 - p\sigma_2)(\rho - \sigma_1)} p_{\text{fo}}(0,0), \quad D_{\text{fo}} = -\frac{p\sigma_2\rho}{(1 - p\sigma_2)(\rho - \sigma_2)} p_{\text{fo}}(0,0).$$

Hence, the general solution of (3.17) is given as

$$x_n = A_{\text{fo}} + B_{\text{fo}}\rho^n + C_{\text{fo}}\sigma_1^n + D_{\text{fo}}\sigma_2^n, \quad n = 1, 2, \ldots, n_e(0). \tag{3.18}$$

Specially, taking $n = 1, n = 2$ in (3.18), respectively, we have

$$\begin{aligned} A_{\text{fo}} + B_{\text{fo}}\rho + C_{\text{fo}}\sigma_1 + D_{\text{fo}}\sigma_2 &= p_{\text{fo}}(1,0), \\ A_{\text{fo}} + B_{\text{fo}}\rho^2 + C_{\text{fo}}\sigma_1^2 + D_{\text{fo}}\sigma_2^2 &= p_{\text{fo}}(2,0). \end{aligned} \tag{3.19}$$

From (3.8) to (3.10), we obtain

$$\begin{aligned} p_{\text{fo}}(1,0) &= \frac{[(\lambda + \theta_1)\sigma_1 + \theta_1]\rho}{\theta_1(1 - p\sigma_2)} p_{\text{fo}}(0,0), \\ p_{\text{fo}}(2,0) &= \frac{\rho[(\rho + \sigma_1)(\lambda + \theta_1)\sigma_1 + \theta_1(\rho + p\sigma_2^2)]}{\theta_1(1 - p\sigma_2)} p_{\text{fo}}(0,0). \end{aligned}$$

Furthermore, substituting the above expressions into (3.19), we get

$$\begin{aligned} A_{\text{fo}} &= 0, \\ B_{\text{fo}} &= \left[\frac{\rho(\lambda + \theta_1)\sigma_1}{\theta_1(1 - p\sigma_2)(\rho - \sigma_1)} + \frac{\rho - \sigma_2 + p\sigma_2^2}{(1 - p\sigma_2)(\rho - \sigma_2)}\right] p_{\text{fo}}(0,0), \end{aligned}$$

then

$$p_{\text{fo}}(n,0) = B_{\text{fo}}\rho^n + C_{\text{fo}}\sigma_1^n + D_{\text{fo}}\sigma_2^n, \quad n = 1, 2, \ldots, n_e(0).$$

Similarly, we obtain (3.5) by substituting $n = 0$ into above expressions. Through some algebraic simplifications, we obtain $p_{\text{fo}}(n_e(0) + 1, 0)$ based on (3.2), (3.4) and (3.9). Finally, $p_{\text{fo}}(0,0)$ can be solved by the normalization equation

$$\sum_{n=0}^{n_e(0)+1} p_{\text{fo}}(n,0) + \sum_{n=0}^{n_e(1)+1} p_{\text{fo}}(n,1) + \sum_{n=0}^{n_e(2)+1} p_{\text{fo}}(n,2) = 1.$$

□

3.3. Analysis of social benefits

Denote the expected social benefit per time unit in the fully observable situation by $SW_{\text{fo}} = R\lambda(1 - p_{\text{balk}}) - CL_{\text{fo}}$, where p_{balk} and L_{fo} respectively represent the balking probability and the expected queue length of customers in the fully observable case, then we have

Theorem 3.3. *In the fully observable M/M/1 queue with variable vacation and vacation interruption, the expected social benefit per time unit $SW_{\rm fo}$ is given as follows*

$$\begin{aligned} SW_{\rm fo} = R\lambda\Bigg[1 - \frac{\lambda\sigma_1^{n_e(1)+1}(\theta_1\sigma_1+\mu)}{\mu\theta_1(1-p\sigma_2)(1-\sigma_1)}p_{\rm fo}(0,0) - \frac{p\sigma_2^{n_e(2)+1}\rho(\mu+\theta_2)}{\theta_2(1-p\sigma_2)}p_{\rm fo}(0,0) - \rho\Big(B_{\rm fo}\rho^{n_e(0)} \\ + C_{\rm fo}\sigma_1^{n_e(0)} + D_{\rm fo}\sigma_2^{n_e(0)}\Big)\Bigg] - C\Bigg[\frac{\lambda\sigma_1(1-\sigma_1^{n_e(1)+1})}{\theta_1(1-p\sigma_2)(1-\sigma_1)^2}p_{\rm fo}(0,0) + \frac{p\sigma_2^2(1-\sigma_2^{n_e(2)+1})}{(1-p\sigma_2)(1-\sigma_2)^2} \\ \times p_{\rm fo}(0,0) + B_{\rm fo}\frac{\rho-(n_e(0)+2)\rho^{n_e(0)+2}+(n_e(0)+1)\rho^{n_e(0)+3}}{(1-\rho)^2} + D_{\rm fo}\Bigg(\rho\sigma_2^{n_e(0)}(n_e(0) \\ + 1) + \frac{\sigma_2-(n_e(0)+1)\sigma_2^{n_e(0)+1}+n_e(0)\sigma_2^{n_e(0)+2}}{(1-\sigma_2)^2} - \frac{\sigma_2^{n_e(2)}}{\rho-\sigma_2}\Bigg) + C_{\rm fo}\Bigg((n_e(0)+1)\rho\sigma_1^{n_e(0)} \\ + \frac{\sigma_1-(n_e(0)+1)\sigma_1^{n_e(0)+1}+n_e(0)\sigma_1^{n_e(0)+2}}{(1-\sigma_1)^2} - \frac{(\rho-\sigma_1)\theta_1\sigma_1^{n_e(1)+1}}{\mu_1\sigma_1+\theta_1}\Bigg)\Bigg]. \end{aligned}$$

Proof. Arrivals will balk when they find the system at state $(n_e(0)+1,0),(n_e(1)+1,1)$ and $(n_e(2)+1,2)$. Therefore, the balking probability is $p_{\rm balk} = p_{\rm fo}(n_e(0)+1,0) + p_{\rm fo}(n_e(1)+1,1) + p_{\rm fo}(n_e(2)+1,2)$, the effective arrival rate is $\lambda(1-p_{\rm balk})$ and the expected queue length equals

$$L_{\rm fo} = \sum_{n=0}^{n_e(0)+1} np_{\rm fo}(n,0) + \sum_{n=0}^{n_e(1)+1} np_{\rm fo}(n,1) + \sum_{n=0}^{n_e(2)+1} np_{\rm fo}(n,2).$$

The expression of the expected social benefit can be obtained by Theorem 3.2. □

3.4. Numerical analysis

This section demonstrates the impacts of several parameters on the expected social benefit in a variable dormancy cloud strategy with sleep-wake control.

Assuming that $\lambda = 0.5, \mu_1 = 0.2, \theta_1 = 0.5, \theta_2 = 0.4, p = 0.5, C = 2$, the relationship between the expected social benefit $SW_{\rm fo}$ and the service rate μ of the virtual machine for different revenues R is shown in Figure 3. It shows a gently increasing trend of the expected social benefit with respect to the increase of the service rate. Moreover, the cloud user requests prefer to enter the system when enhancing the expected social benefit by raising the revenue R.

Assuming that $R = 25, \mu_1 = 0.2, \theta_1 = 0.5, \theta_2 = 0.4, p = 0.5, C = 2$, the relationship between the expected social benefit $SW_{\rm fo}$ and the service rate μ of the virtual machine for different arrival rates λ is shown in Figure 4. It displays that the expected social benefit is not necessarily increasing with the parameter λ which is contrary to our common sense. This mainly results from the general increasing trend of considerable waiting costs. Specially, the expected social benefit $SW_{\rm fo}$ is negative at parameter $\lambda = 0.4, \mu = 1.5$.

Assuming that $\mu = 5, \mu_1 = 1.5, \theta_1 = 0.8, \theta_2 = 1, p = 0.5, C = 2$, the relationship between the expected social benefit $SW_{\rm fo}$ and the arrival rate λ of cloud user requests for different revenues R is shown in Figure 5. It presents that the expected social benefit rises with respect to the increasing arrival rate when the revenue is large enough and the expected social benefit boosts remarkably with raising revenue. Besides, the vertex of the expected social benefit is $SW_{\rm fo} = 3.0$ for the arrival rate $\lambda = 0.4$ at parameter $R = 15$.

Assuming that $R = 15, \mu_1 = 1.5, \theta_1 = 0.8, \theta_2 = 1, p = 0.5, C = 2$, the relationship between the expected social benefit $SW_{\rm fo}$ and the arrival rate λ of cloud user requests for different service rates μ is shown in Figure 6. It reveals that the expected social benefit first gradually increases and achieves a maximum at about $\lambda = 0.4$, then decreases as parameter λ continues to increase at parameter $\mu = 4.5$ and presents a fluctuating tendency. The maximum and minimum of the expected social benefit are $SW_{\rm fo} = 2.0$ and $SW_{\rm fo} = -0.9$ respectively for the arrival rate $\lambda = 0.3$ and $\lambda = 0.8$ at parameter $\mu = 3.5$. Moreover, the cloud user requests prefer to enter the system with the increasing service rate which raises the expected social benefit as well.

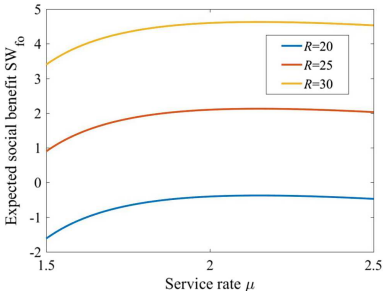

FIGURE 3. The expected social benefit *versus* parameter μ with the revenue R.

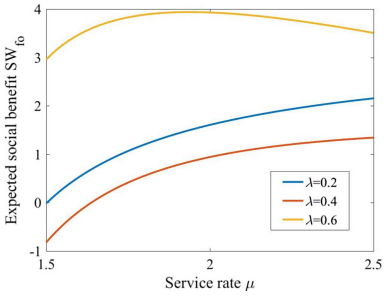

FIGURE 4. The expected social benefit *versus* parameter μ with the arrival rate λ.

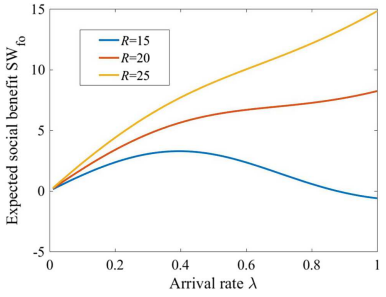

FIGURE 5. The expected social benefit *versus* parameter λ with the revenue R.

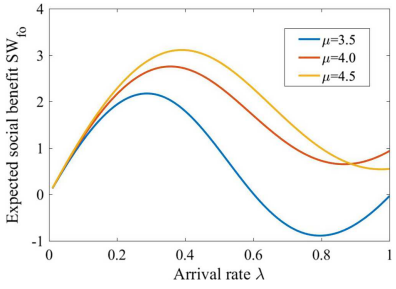

FIGURE 6. The expected social benefit *versus* parameter λ with the service rate μ.

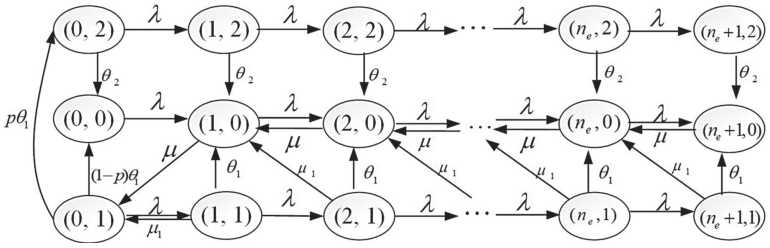

FIGURE 7. State transition diagram in the partly observable case.

4. Equilibrium strategy in partly observable case

4.1. Stationary queue length distribution

This section proceeds to analyze the partly observable case that arriving customers can only observe the number $N(t)$ of customers presented. To obtain the pure equilibrium threshold strategy, we should compute the steady-state distribution in this situation. Denote the stationary distribution in the partly observable queue by $p_{\mathrm{ao}}(n,i) = \lim_{t\to+\infty} p_{\mathrm{ao}}\{N(t)=n, I(t)=i\},\ (n,i)\in\Omega$. The state transition diagram is illustrated in Figure 7.

Theorem 4.1. *In the partly observable M/M/1 queue with variable vacation and vacation interruption where customers enter the system according to the pure threshold strategy n_e, the stationary distribution is as follows*

$$p_{\mathrm{ao}}(n,1) = \frac{\lambda}{\theta_1(1-p\sigma_2)}\sigma_1^n p_{\mathrm{ao}}(0,0), \quad n=0,1,\dots,n_e, \tag{4.1}$$

$$p_{\mathrm{ao}}(n_e+1,1) = \frac{\lambda\sigma_1}{\theta_1(1-p\sigma_2)(1-\sigma_1)}\sigma_1^{n_e} p_{\mathrm{ao}}(0,0), \tag{4.2}$$

$$p_{\mathrm{ao}}(n,2) = \frac{p\sigma_2}{1-p\sigma_2}\sigma_2^n p_{\mathrm{ao}}(0,0), \quad n=0,1,\dots,n_e, \tag{4.3}$$

$$p_{\mathrm{ao}}(n_e+1,2) = \frac{p\sigma_2^2}{(1-p\sigma_2)(1-\sigma_2)}\sigma_2^{n_e} p_{\mathrm{ao}}(0,0), \tag{4.4}$$

$$p_{\mathrm{ao}}(n,0) = B_{\mathrm{ao}}\rho^n + C_{\mathrm{ao}}\sigma_1^n + D_{\mathrm{ao}}\sigma_2^n, \quad n=0,1,\dots,n_e, \tag{4.5}$$

$$\begin{aligned} p_{\mathrm{ao}}(n_e+1,0) &= \rho(B_{\mathrm{ao}}\rho^{n_e} + C_{\mathrm{ao}}\sigma_1^{n_e} + D_{\mathrm{ao}}\sigma_2^{n_e}) \\ &\quad + \left(\frac{\lambda\theta_1\sigma_1^{n_e+2}}{\mu\theta_1(1-p\sigma_2)(1-\sigma_1)} + \frac{p\theta_2\sigma_2^{n_e+1}}{\mu(1-p\sigma_2)}\right) p_{\mathrm{ao}}(0,0), \end{aligned} \tag{4.6}$$

where

$$\begin{aligned} p_{\mathrm{ao}}(0,0) = \Bigg\{ & 1 + \frac{\rho-\rho^{n_e+2}}{1-\rho} + \frac{\rho(\lambda+\theta_1)\sigma_1}{\theta_1(1-p\sigma_2)(\rho-\sigma_1)}\left(\frac{\rho-\rho^{n_e+2}}{1-\rho} + \sigma_1^{n_e}\left(\frac{\theta_1}{(\lambda+\theta_1)(1-\sigma_1)} - \rho\right)\right. \\ & \left. + \frac{\lambda-\sigma_1\mu}{(\lambda+\theta_1)(1-\sigma_1)\sigma_1} - \frac{\sigma_1-\sigma_1^{n_e+1}}{1-\sigma_1}\right) + \frac{p\rho\sigma_2}{(1-p\sigma_2)(\rho-\sigma_2)}\left(\frac{\rho-\rho^{n_e+2}}{1-\rho} + \sigma_2^{n_e}(1-\rho)\right. \\ & \left. + \frac{1}{\rho} - \frac{\sigma_2-\sigma_2^{n_e+2}}{1-\sigma_2}\right)\Bigg\}^{-1}, \end{aligned}$$

$$B_{\mathrm{ao}} = \left[\frac{\rho(\lambda+\theta_1)\sigma_1}{\theta_1(1-p\sigma_2)(\rho-\sigma_1)} + \frac{\rho-\sigma_2+p\sigma_2^2}{(1-p\sigma_2)(\rho-\sigma_2)}\right] p_{\mathrm{ao}}(0,0) = p_{\mathrm{ao}}(0,0) - C_{\mathrm{ao}} - D_{\mathrm{ao}},$$

$$C_{\mathrm{ao}} = -\frac{\rho(\lambda+\theta_1)\sigma_1}{\theta_1(1-p\sigma_2)(\rho-\sigma_1)} p_{\mathrm{ao}}(0,0), \quad D_{\mathrm{ao}} = -\frac{p\sigma_2\rho}{(1-p\sigma_2)(\rho-\sigma_2)} p_{\mathrm{ao}}(0,0).$$

Proof. The corresponding stationary distribution can be obtained in Figure 7.

$$\lambda p_{\mathrm{ao}}(0,0) = (1-p)\theta_1 p_{\mathrm{ao}}(0,1) + \theta_2 p_{\mathrm{ao}}(0,2), \tag{4.7}$$

$$\begin{aligned}(\lambda+\mu)p_{\mathrm{ao}}(n,0) &= \mu p_{\mathrm{ao}}(n+1,0) + \mu_1 p_{\mathrm{ao}}(n+1,1) + \theta_1 p_{\mathrm{ao}}(n,1)\\ &\quad +\lambda p_{\mathrm{ao}}(n-1,0) + \theta_2 p_{\mathrm{ao}}(n,2), \quad n = 1,2,\ldots,n_e,\end{aligned} \tag{4.8}$$

$$\mu p_{\mathrm{ao}}(n_e+1,0) = \theta_1 p_{\mathrm{ao}}(n_e+1,1) + \lambda p_{\mathrm{ao}}(n_e,0) + \theta_2 p_{\mathrm{ao}}(n_e+1,2), \tag{4.9}$$

$$(\lambda+\theta_1)p_{\mathrm{ao}}(0,1) = \mu_1 p_{\mathrm{ao}}(1,1) + \mu p_{\mathrm{ao}}(1,0), \tag{4.10}$$

$$(\lambda+\mu_1+\theta_1)p_{\mathrm{ao}}(n,1) = \lambda p_{\mathrm{ao}}(n-1,1), \quad n = 1,2,\ldots,n_e, \tag{4.11}$$

$$(\mu_1+\theta_1)p_{\mathrm{ao}}(n_e+1,1) = \lambda p_{\mathrm{ao}}(n_e,1), \tag{4.12}$$

$$(\lambda+\theta_2)p_{\mathrm{ao}}(0,2) = p\theta_1 p_{\mathrm{ao}}(0,1), \tag{4.13}$$

$$(\lambda+\theta_2)p_{\mathrm{ao}}(n,2) = \lambda p_{\mathrm{ao}}(n-1,2), \quad n = 1,2,\ldots,n_e, \tag{4.14}$$

$$\theta_2 p_{\mathrm{ao}}(n_e+1,2) = \lambda p_{\mathrm{ao}}(n_e,2). \tag{4.15}$$

According to (4.7) and (4.13), we have

$$p_{\mathrm{ao}}(0,1) = \frac{\lambda}{\theta_1(1-p\theta_2)}p_{\mathrm{ao}}(0,0), \quad p_{\mathrm{ao}}(0,2) = \frac{p\sigma_2}{1-p\sigma_2}p_{\mathrm{ao}}(0,0). \tag{4.16}$$

Substituting (4.16) into (4.11) and (4.14), by iterating we obtain

$$p_{\mathrm{ao}}(n,1) = \left(\frac{\lambda}{\lambda+\mu_1+\theta_1}\right)^n \frac{\lambda}{\theta_1(1-p\sigma_2)}p_{\mathrm{ao}}(0,0), \quad n = 1,2,\ldots,n_e,$$

$$p_{\mathrm{ao}}(n,2) = \left(\frac{\lambda}{\lambda+\theta_2}\right)^n \frac{p\sigma_2}{1-p\sigma_2}p_{\mathrm{ao}}(0,0), \quad n = 1,2,\ldots,n_e,$$

Combining σ_1 and σ_2, we get (4.1) and (4.3).

Substituting $n = n_e$ into (4.12) and (4.15), respectively, we get

$$p_{\mathrm{ao}}(n_e+1,1) = \left(\frac{\lambda}{\lambda+\mu_1+\theta_1}\right)^{n_e} \frac{\lambda\sigma_1}{\theta_1(1-p\sigma_2)(1-\sigma_1)}p_{\mathrm{ao}}(0,0),$$

$$p_{\mathrm{ao}}(n_e+1,2) = \left(\frac{\lambda}{\lambda+\theta_2}\right)^{n_e} \frac{p\sigma_2^2}{(1-p\sigma_2)(1-\sigma_2)}p_{\mathrm{ao}}(0,0).$$

Manipulating above equations, we get (4.2) and (4.4).

Next, we calculate the formulas of $p_{\mathrm{ao}}(n,0), n = 0,1,\ldots,n_e$ and $p_{\mathrm{ao}}(n_e+1,0)$. Similar to the proof for Theorem 3.2, $p_{\mathrm{ao}}(n,0)$ are solutions of the following nonhomogeneous linear difference equation with constant coefficients

$$\begin{aligned}\mu x_{n+1} - (\lambda+\mu)x_n + \lambda x_{n-1} = \left(-\frac{\lambda(\mu_1\sigma_1+\theta_1)}{\theta_1(1-p\sigma_2)}\sigma_1^n - \frac{p\sigma_2\theta_2}{1-p\sigma_2}\sigma_2^n\right)p_{\mathrm{ao}}(0,0),\\ n = 1,2,\ldots,n_e.\end{aligned} \tag{4.17}$$

As the nonhomogeneous part of (4.17) is the sum of geometric distributions with parameter σ_1 and σ_2 and $\sigma_1,\sigma_2 \neq \rho,1$, Equation (4.17) have specific solutions of the form $C_{\mathrm{ao}}\sigma_1^n + D_{\mathrm{ao}}\sigma_2^n$, where C_{ao} and D_{ao} are constants. Substituting the specific solution into (4.17), we obtain

$$C_{\mathrm{ao}} = -\frac{\rho(\lambda+\theta_1)\sigma_1}{\theta_1(1-p\sigma_2)(\rho-\sigma_1)}p_{\mathrm{ao}}(0,0), \quad D_{\mathrm{ao}} = -\frac{p\sigma_2\rho}{(1-p\sigma_2)(\rho-\sigma_2)}p_{\mathrm{ao}}(0,0).$$

Hence, general solutions of (4.17) are given as

$$x_n^{\text{gen}} = A_{\text{ao}} + B_{\text{ao}}\rho^n + C_{\text{ao}}\sigma_1^n + D_{\text{ao}}\sigma_2^n, \quad n = 1, 2, \ldots, n_e,$$

where $A_{\text{ao}}, B_{\text{ao}}, C_{\text{ao}}$ and D_{ao} are constants. Specially, taking $n = 1$ and $n = 2$ in the general solution, we have

$$\begin{aligned} &A_{\text{ao}} + B_{\text{ao}}\rho + C_{\text{ao}}\sigma_1 + D_{\text{ao}}\sigma_2 = p_{\text{ao}}(1,0), \\ &A_{\text{ao}} + B_{\text{ao}}\rho^2 + C_{\text{ao}}\sigma_1^2 + D_{\text{ao}}\sigma_2^2 = p_{\text{ao}}(2,0). \end{aligned} \tag{4.18}$$

From (4.8) to (4.10), we obtain

$$\begin{aligned} p_{\text{ao}}(1,0) &= \frac{[(\lambda+\theta_1)\sigma_1 + \theta_1]\rho}{\theta_1(1-p\sigma_2)} p_{\text{ao}}(0,0), \\ p_{\text{ao}}(2,0) &= \frac{\rho[(\rho+\sigma_1)(\lambda+\theta_1)\sigma_1 + \theta_1(\rho + p\sigma_2^2)]}{\theta_1(1-p\sigma_2)} p_{\text{ao}}(0,0). \end{aligned}$$

Furthermore, substituting above expressions into (4.18), we get

$$\begin{aligned} A_{\text{ao}} &= 0, \\ B_{\text{ao}} &= \left[\frac{\rho(\lambda+\theta_1)\sigma_1}{\theta_1(1-p\sigma_2)(\rho-\sigma_1)} + \frac{\rho - \sigma_2 + p\sigma_2^2}{(1-p\sigma_2)(\rho-\sigma_2)}\right] p_{\text{ao}}(0,0), \end{aligned}$$

then

$$p_{\text{ao}}(n,0) = B_{\text{ao}}\rho^n + C_{\text{ao}}\sigma_1^n + D_{\text{ao}}\sigma_2^n, \quad n = 1, 2, \ldots, n_e.$$

Similarly, substituting $n = 0$ into above expressions, we obtain (4.5). Through some algebraic simplifications, we obtain $p_{\text{ao}}(n_e+1, 0)$ based on (4.2), (4.4) and (4.9). Finally, the probability $p_{\text{ao}}(0,0)$ can be solved using the normalization equation

$$\sum_{n=0}^{n_e+1} p_{\text{ao}}(n,0) + \sum_{n=0}^{n_e+1} p_{\text{ao}}(n,1) + \sum_{n=0}^{n_e+1} p_{\text{ao}}(n,2) = 1.$$

□

4.2. Benefit analysis of customers

The expected net benefit of an arriving customer who finds n customers ahead and decides to enter is

$$S(n) = R - CT_{\text{ao}}(n) \tag{4.19}$$

where $T_{\text{ao}}(n) = E_{\text{ao}}[S|N = n]$ represents the mean sojourn time of this new arrival.

For convenience, we introduce the following notations

$$a = \frac{(\mu+\theta_1)(1-\sigma_2)}{\mu(1-\sigma_1)}, b = \frac{\rho(1-\sigma_1)}{\sigma_2}, c = \frac{\rho(1-\sigma_2)}{\sigma_1}, d = \frac{(\lambda+\theta_2)(\mu+\theta_2)(1-\sigma_1)}{\mu\theta_2}.$$

Theorem 4.2. *In the partly observable M/M/1 queue with variable vacation and vacation interruption, if an arriving customer finds n customers waiting in line and other customers follow the policy n_e, if he decides to enter, his expected net benefit per time unit is*

$$\begin{aligned} S(n) = R - C\Bigg[&\frac{n+1}{\mu} + \frac{(\mu-\mu_1)\sigma_1^n p_{\text{ao}}(0,1)}{\mu(\mu_1+\theta_1)[B_{\text{ao}}\rho^n + (C_{\text{ao}} + p_{\text{ao}}(0,1))\sigma_1^n + (D_{\text{ao}} + p_{\text{ao}}(0,2))\sigma_2^n]} \\ &+ \frac{\sigma_2^n p_{\text{ao}}(0,2)}{\theta_2[B_{\text{ao}}\rho^n + (C_{\text{ao}} + p_{\text{ao}}(0,1))\sigma_1^n + (D_{\text{ao}} + p_{\text{ao}}(0,2))\sigma_2^n]}\Bigg], \quad n = 0, 1, \ldots, n_e. \end{aligned}$$

$$S(n_e+1)=R-C\Bigg[\frac{n_e+2}{\mu}+\frac{\frac{\mu-\mu_1}{\mu(\mu_1+\theta_1)}\sigma_1^{n_e+1}p_{\mathrm{ao}}(0,1)}{(1-\sigma_1)B_{\mathrm{ao}}\rho^{n_e+1}+\left[\frac{\mu_1+\theta_1}{\mu}C_{\mathrm{ao}}+\frac{\mu+\theta_1}{\mu}p_{\mathrm{ao}}(0,1)\right]\sigma_1^{n_e+1}+[bD_{\mathrm{ao}}+dp_{\mathrm{ao}}(0,2)]\,\sigma_2^{n_e+1}}+\frac{\sigma_2^{n_e+1}p_{\mathrm{ao}}(0,2)}{\theta_2\left((1-\sigma_2)B_{\mathrm{ao}}\rho^{n_e+1}+(cC_{\mathrm{ao}}+ap_{\mathrm{ao}}(0,1))\,\sigma_1^{n_e+1}+\left(\frac{\theta_2}{\mu}D_{\mathrm{ao}}+\frac{\mu+\theta_2}{\mu}p_{\mathrm{ao}}(0,2)\right)\sigma_2^{n_e+1}\right)}\Bigg].$$

Moreover, $S(n)$ is monotonically decreasing with respect to n, $0\le n\le n_e+1$.

Proof. Using the conditional expectation formula, we obtain

$$\begin{aligned}T_{\mathrm{ao}}(n)=&\,T_{\mathrm{ao}}(n,0)p_{\mathrm{ao}}(i=0|N=n)+T_{\mathrm{ao}}(n,1)p_{\mathrm{ao}}(i=1|N=n)\\&+T_{\mathrm{ao}}(n,2)p_{\mathrm{ao}}(i=2|N=n),\end{aligned}\tag{4.20}$$

where $p_{\mathrm{ao}}(i|n)$ is the conditional probability that the server is in state i when there are n customers in the line.

By the definition of the conditional probability, we have

$$p_{\mathrm{ao}}(i|n)=\frac{p_{\mathrm{ao}}(n,i)}{p_{\mathrm{ao}}(n,0)+p_{\mathrm{ao}}(n,1)+p_{\mathrm{ao}}(n,2)},\ 0\le n\le n_e+1.$$

Calculating $p_{\mathrm{ao}}(0|n), p_{\mathrm{ao}}(1|n)$ and $p_{\mathrm{ao}}(2|n)$ using Theorem 4.1 and substituting $T_{\mathrm{fo}}(n,0), T_{\mathrm{fo}}(n,1)$ and $T_{\mathrm{fo}}(n,2)$ into (4.20), we obtain

$$\begin{aligned}T_{\mathrm{ao}}(n)=&\left(\frac{\mu+\theta_1}{\mu(\mu_1+\theta_1)}+\frac{n}{\mu}\right)\frac{\sigma_1^n p_{\mathrm{ao}}(0,1)}{B_{\mathrm{ao}}\rho^n+(C_{\mathrm{ao}}+p_{\mathrm{ao}}(0,1))\sigma_1^n+(D_{\mathrm{ao}}+p_{\mathrm{ao}}(0,2))\sigma_2^n}\\&+\left(\frac{n+1}{\mu}+\frac{1}{\theta_2}\right)\frac{\sigma_2^n p_{\mathrm{ao}}(0,2)}{B_{\mathrm{ao}}\rho^n+(C_{\mathrm{ao}}+p_{\mathrm{ao}}(0,1))\sigma_1^n+(D_{\mathrm{ao}}+p_{\mathrm{ao}}(0,2))\sigma_2^n}\\&+\frac{n+1}{\mu}\frac{B_{\mathrm{ao}}\rho^n+C_{\mathrm{ao}}\sigma_1^n+D_{\mathrm{ao}}\sigma_2^n}{B_{\mathrm{ao}}\rho^n+(C_{\mathrm{ao}}+p_{\mathrm{ao}}(0,1))\sigma_1^n+(D_{\mathrm{ao}}+p_{\mathrm{ao}}(0,2))\sigma_2^n},n=0,1,\ldots,n_e,\end{aligned}$$

$$T_{\mathrm{ao}}(n_e+1)=\frac{n_e+2}{\mu}+\frac{\frac{\mu-\mu_1}{\mu(\mu_1+\theta_1)}\sigma_1^{n_e+1}p_{\mathrm{ao}}(0,1)}{(1-\sigma_1)B_{\mathrm{ao}}\rho^{n_e+1}+\left[\frac{\mu_1+\theta_1}{\mu}C_{\mathrm{ao}}+\frac{\mu+\theta_1}{\mu}p_{\mathrm{ao}}(0,1)\right]\sigma_1^{n_e+1}+[bD_{\mathrm{ao}}+dp_{\mathrm{ao}}(0,2)]\,\sigma_2^{n_e+1}}+\frac{\frac{\sigma_2^{n_e+1}}{\theta_2}p_{\mathrm{ao}}(0,2)}{(1-\sigma_2)B_{\mathrm{ao}}\rho^{n_e+1}+[cC_{\mathrm{ao}}+ap_{\mathrm{ao}}(0,1)]\,\sigma_1^{n_e+1}+\left[\frac{\theta_2}{\mu}D_{\mathrm{ao}}+\frac{\mu+\theta_2}{\mu}p_{\mathrm{ao}}(0,2)\right]\sigma_2^{n_e+1}}.$$

Substituting $T_{\mathrm{ao}}(n)$ and $T_{\mathrm{ao}}(n_e+1)$ into (4.19), we obtain the expression of $S(n)$ and $S(n_e+1)$.

A tagged customer decides to queue when he sees $j(1\le j\le n_e+1)$ customers ahead, then his residual sojourn time equals the sum of the residual waiting time and the service time. Therefore, $S(j)<S(j-1)$ and $S(n)$ declines monotonically with respect to n. Theorem 4.2 is proved. □

4.3. Equilibrium threshold strategy

New arrivals will balk when the system is empty if $S(0)<0$, hence, we suppose $S(0)>0$ hereafter. n_e is identified as the optimal threshold when $S(n_e)>0$ and $S(n_e+1)<0$. For convenience, referring to the

method used in Burnetas and Economou [1] and Li *et al.* [9], and considering the form of $S(n)$ and $S(n_e+1)$, we introduce a set of functions $S(x,n), x \in \mathbb{R}; n = 0,1,\ldots$.

$$S(x,n) = R - C\Bigg[\frac{n+1}{\mu} + \frac{\frac{\mu-\mu_1}{\mu(\mu_1+\theta_1)}\sigma_1^n p_{\text{ao}}(0,1)}{(1-\sigma_1 x)B_{\text{ao}}\rho^n + \left[\frac{\mu+(\mu_1+\theta_1-\mu)x}{\mu}C_{\text{ao}} + \frac{\mu+\theta_1 x}{\mu}p_{\text{ao}}(0,1)\right]\sigma_1^n + [(1+h_2(x))D_{\text{ao}} + (1+h_1(x))p_{\text{ao}}(0,2)]\,\sigma_2^n} + \frac{\frac{\sigma_2^n}{\theta_2}p_{\text{ao}}(0,2)}{(1-\sigma_2 x)B_{\text{ao}}\rho^n + [(1+h_3(x))C_{\text{ao}} + (1+h_4(x))p_{\text{ao}}(0,1)]\,\sigma_1^n + \left[\frac{\mu+(\theta_2-\mu)x}{\mu}D_{\text{ao}} + \frac{\mu+\theta_2 x}{\mu}p_{\text{ao}}(0,2)\right]\sigma_2^n}\Bigg],$$

where

$$h_1(x) = \left(\frac{(1-\sigma_1)[\lambda(\mu+\theta_2)+\theta_2^2]}{\mu\theta_2} - \sigma_1\right)x, \quad h_2(x) = \frac{\rho(1-\sigma_1)-\sigma_2}{\sigma_2}x,$$
$$h_3(x) = \frac{\rho(1-\sigma_2)-\sigma_1}{\sigma_1}x, \quad h_4(x) = \frac{\theta_1(1-\sigma_2)+\mu(\sigma_1-\sigma_2)}{\mu(1-\sigma_1)}x.$$

Next we will certify the existence of the equilibrium threshold strategies of customers and explore the corresponding thresholds using $S(x,n)$. By calculating, we have

$$S(n) = S(0,n), n = 0,1,\ldots,n_e, \quad S(n_e+1) = S(1,n_e+1).$$

Let us mark an arriving customer and suppose that other customers enter the queue following the threshold n_e. Then $S(0,n), n = 0,1,\ldots,n_e$ and $S(1,n_e+1)$ respectively represent the expected net benefit of this marked customer who observes there are n or n_e+1 customers ahead and decides to enter.

Denote $S^{'}(x,n) = \frac{\partial S(x,n)}{\partial x}$ and take the derivative of $S(x,n)$ with respect to x for a given n, we obtain

$$S^{'}(x,n) = \frac{\frac{\mu-\mu_1}{\mu(\mu_1+\theta_1)}C_{\text{ao}}\sigma_1^n p_{\text{ao}}(0,1)\left[-B_{\text{ao}}\rho^n + \left(\frac{\mu_1+\theta_1-\mu}{\mu}C_{\text{ao}} + \frac{\theta_1}{\mu}p_{\text{ao}}(0,1)\right)\sigma_1^n + [h_2(1)D_{\text{ao}} + h_1(1)p_{\text{ao}}(0,2)]\,\sigma_2^n\right]}{\left[(1-\sigma_1 x)B_{\text{ao}}\rho^n + \left[\frac{\mu+(\mu_1+\theta_1-\mu)x}{\mu}C_{\text{ao}} + \frac{\mu+\theta_1 x}{\mu}p_{\text{ao}}(0,1)\right]\sigma_1^n + [(1+h_2(x))D_{\text{ao}} + (1+h_1(x))p_{\text{ao}}(0,2)]\sigma_2^n\right]^2}$$
$$+ \frac{\frac{\sigma_2^n}{\theta_2}C_{\text{ao}}p_{\text{ao}}(0,2)\left[-B_{\text{ao}}\sigma_2\rho^n + [h_3(1)C_{\text{ao}} + h_4(1)p_{\text{ao}}(0,1)]\,\sigma_1^n + \left(\frac{\theta_2-\mu}{\mu}D_{\text{ao}} + \frac{\theta_2}{\mu}p_{\text{ao}}(0,2)\right)\sigma_2^n\right]}{\left[(1-\sigma_2 x)B_{\text{ao}}\rho^n + [(1+h_3(x))C_{\text{ao}} + (1+h_4(x))p_{\text{ao}}(0,1)]\,\sigma_1^n + \left[\frac{\mu+(\theta_1-\mu)x}{\mu}D_{\text{ao}} + \frac{\mu+\theta_2 x}{\mu}p_{\text{ao}}(0,2)\right]\sigma_2^n\right]^2}.$$

Because the denominator of $S^{'}(x,n)$ is greater than zero and the numerator of $S^{'}(x,n)$ only relates to n, the positive and negative properties of $S^{'}(x,n)$ doesn't depend on x for a given n. If $S^{'}(x,n) < 0$, that is, $S(x,n)$ is monotonically decreasing with respect to x for a given n, then $S(0,n) > S(1,n)$; Otherwise, $S(0,n) < S(1,n)$. Therefore, we obtain Theorem 4.3.

Theorem 4.3. *There exist finite non-negative integers $n_L \le n_U$, such that the following holds.*

Case 1. If $S(x,n)$ is monotonically decreasing in x given n, we have

$$(a) \qquad S(0,0), S(0,1), \ldots, S(0,n_L), \ldots, S(0,n_U) > 0, S(0,n_U+1) \le 0, \tag{4.21}$$
$$(b) \qquad S(1,n_U+1), S(1,n_U)\ldots, S(1,n_L+1) \le 0, S(1,n_L) > 0, \tag{4.22}$$
$$\text{or } S(1,n_U+1), S(1,n_U), \ldots, S(1,n_L), \ldots, S(1,1), S(1,0) \le 0. \tag{4.23}$$

The equilibrium threshold strategy in the partly observable M/M/1 queue with variable vacation and vacation interruption is "While arriving at time t, observe: enter if $n \le n_e$ and balk otherwise" for $n_e \in \{n_L, n_L + 1, \ldots, n_U\}$.

Case 2. If $S(x,n)$ is monotonically increasing in x given n, we have

$$(c) \qquad S(0,0), S(0,1), \ldots, S(0,n_L-1) > 0, S(0,n_L), \ldots, S(0,n_U) \le 0,$$

$$(d) \qquad S(1,n_L), S(1,n_L+1), \ldots, S(1,n_U+1) \ge 0, S(1,n_U+2) < 0,$$
$$\text{or } S(1,n_L), S(1,n_L+1), \ldots, S(1,n_U), S(1,n_U+1) \ge 0.$$

The equilibrium threshold strategy in the partly observable M/M/1 queue with variable vacation and vacation interruption is "While arriving at time t, observe: enter if $n > n_e$ and balk otherwise" for $n_e \in \{n_L, n_L + 1, \ldots, n_U\}$.

Proof. We prove case 1 in two steps.

Step 1. We prove that $n_e \in \{n_L, n_L + 1, \ldots, n_U\}$.

(a) We can easily get $S(0,0) > 0$ and $\lim_{n \to +\infty} S(0,n) = -\infty$, so there exists a positive integer n_U which satisfies $S(0, n_U + 1) \le 0$ and $S(0, n_U) > 0$ when $S(x,n)$ is monotonically decreasing, therefore, (4.21) is obtained.

(b) Similarly, due to $S(1,n) < S(0,n)$, we can easily find the positive integer n_L satisfying (4.22). Otherwise, if all items of $S(1,n)$ are non-positive, we get (4.23).

Step 2. We prove that n_e is the threshold.

On one hand, for an arriving customer who decides to enter the system when there are $n (n \le n_e)$ customers ahead of him, his expected net benefit is $S(n)$, and $S(n) \ge S(n_e) = S(0, n_e) > 0$ based on Theorem 4.2 and (4.21). Hence, he prefers to enter the system. On the other hand, for an arriving customer who decides to enter the system when there are $n = n_e + 1$ customers ahead of him, his expected net benefit is $S(n_e + 1)$, and $S(n_e + 1) = S(1, n_e + 1) \le 0$ based on Theorem 4.2, (4.22) and (4.23). Therefore, this arrival is reluctant to queue.

Case 2 can be similarly proved. □

It is interesting to note that there is a Follow-The-Crowd (FTC) situation in the partly observable queue when $S(x,n)$ is monotonically decreasing. Otherwise, there is an Avoid-The-Crowd (ATC) [9] situation in the partly observable queue.

4.4. Analysis of social benefits

Denote the expected social benefit per time unit in the partly observable situation as $SW_{\text{ao}} = R\lambda(1 - p_{\text{ao}}(n_e+1)) - CL_{\text{ao}}$, where $p_{\text{ao}}(n_e+1)$ and L_{ao} represent the balking probability and the expected queue length of customers in the partly observable case. Then we have

Theorem 4.4. *In the partly observable M/M/1 queue with variable vacation and vacation interruption, if an arriving customer finds other customers follow the threshold n_e and decides to enter, his expected social benefit per time unit is*

$$\begin{aligned} SW_{\text{ao}} = R\lambda \Bigg[& 1 - \frac{\lambda \sigma_1^{n_e+1} p_{\text{ao}}(0,0)}{\theta_1 (1 - p\sigma_2)(1 - \sigma_1)} - \frac{p\sigma_2^{n_e+2} p_{\text{ao}}(0,0)}{(1 - p\sigma_2)(1 - \sigma_2)} - B_{\text{ao}} \rho^{n_e+1} - C_{\text{ao}} \sigma_1^{n_e+1} \frac{\mu_1 \rho + \theta_1}{\mu_1 \sigma_1 + \theta_1} \\ & - D_{\text{ao}} \sigma_2^{n_e+1} \Bigg] - C \Bigg[\frac{p\sigma_2^2 \left(1 - \sigma_2^{n_e+1}\right) p_{\text{ao}}(0,0)}{(1 - p\sigma_2)(1 - \sigma_2)^2} + B_{\text{ao}} \frac{\rho - (n_e+2)\rho^{n_e+2} + (n_e+1)\rho^{n_e+3}}{(1-\rho)^2} \\ & + C_{\text{ao}} \frac{(\mu_1 \rho + \theta_1)(n_e+1)\sigma_1^{n_e+1}}{\mu_1 \sigma_1 + \theta_1} + C_{\text{ao}} \frac{\sigma_1 - (n_e+1)\sigma_1^{n_e+1} + n_e \sigma_1^{n_e+2}}{(1-\sigma_1)^2} \\ & + \frac{\lambda \sigma_1 \left(1 - \sigma_1^{n_e+1}\right) p_{\text{ao}}(0,0)}{\theta_1 (1 - p\sigma_2)(1 - \sigma_1)^2} + D_{\text{ao}} \frac{\sigma_2 - (n_e+2)\sigma_2^{n_e+2} + (n_e+1)\sigma_2^{n_e+3}}{(1-\sigma_2)^2} \Bigg]. \end{aligned}$$

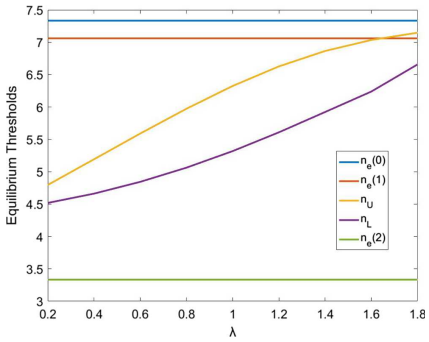

FIGURE 8. The equilibrium thresholds *versus* parameter λ.

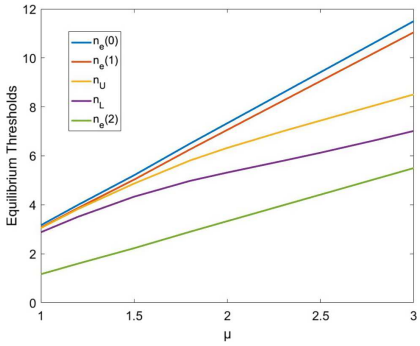

FIGURE 9. The equilibrium thresholds *versus* parameter μ.

Proof. Arrivals are reluctant to enter when they find the system at state $(n_e+1,0),(n_e+1,1)$ and $(n_e+1,2)$. Therefore, the balking probability is $p_{\text{ao}}(n_e+1)=p_{\text{ao}}(n_e+1,0)+p_{\text{ao}}(n_e+1,1)+p_{\text{ao}}(n_e+1,2)$, the effective arrival rate is $\lambda(1-p_{\text{ao}}(n_e+1))$ and the expected queue length can be calculated using the formula

$$L_{\text{ao}}=\sum_{n=0}^{n_e+1} n(p_{\text{ao}}(n,0)+p_{\text{ao}}(n,1)+p_{\text{ao}}(n,2)).$$

The expression of the expected social benefit can be derived from Theorem 4.1. □

5. Numerical comparison of equilibrium thresholds

This section presents numerical experiments to demonstrate the impact of information levels and several parameters on the equilibrium threshold strategy of customers.

Assuming that $R=5,\mu=2,\mu_1=0.5,\theta_1=5,\theta_2=0.5,p=0.5,C=1.2$, the relationship between the equilibrium threshold strategy and the arrival rate λ of cloud user requests is shown in Figure 8. The equilibrium threshold strategy remains unaltered since the arrival rate λ is irrelevant to the decisions of customers when the cloud user requests master the state information completely, which is derived from Theorem 3.1. In addition, the equilibrium threshold increases with the parameter λ in the partly observable case, meaning that the customer prefers to join the system in this situation.

Assuming that $R=5,\lambda=1,\mu_1=0.5,\theta_1=5,\theta_2=0.5,p=0.5,C=1.2$, Figure 9 shows that all types of the equilibrium thresholds increase along with the service rate μ resulted by accelerating the delivery of the virtual

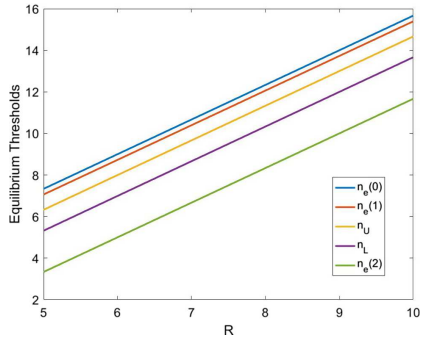

FIGURE 10. The equilibrium thresholds *versus* parameter R.

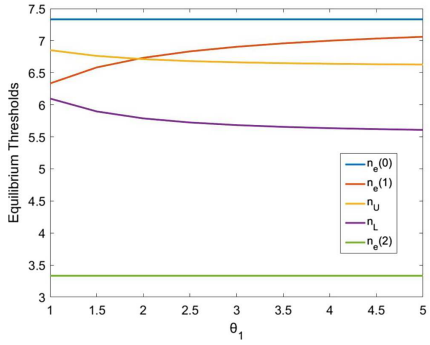

FIGURE 11. The equilibrium thresholds *versus* parameter θ_1.

machine and lessening the expected sojourn time of customers. In other words, the cloud user requests have a greater incentive to enter the system whether they acquire the state information or not when the service rate increases.

When $\lambda = 1, \mu = 2, \mu_1 = 0.5, \theta_1 = 5, \theta_2 = 0.5, p = 0.5, C = 1.2$, the result shown in Figure 10 indicates that the whole equilibrium threshold policies linearly increase with parameter R, which is consistent with the results of Theorems 3.1 and 4.2. It is intuitive that the more revenues after customers are served, the more customers prefer to enter the system.

When $R = 5, \lambda = 1.2, \mu = 0.5, \mu_1 = 0.5, \theta_2 = 0.5, p = 0.5, C = 1.2$, the relationship between the equilibrium threshold strategy and parameter θ_1 in semi dormancy state is shown in Figure 11. It is observed that the maximum entrance thresholds $n_e(0)$ and $n_e(2)$ for cloud user requests in the fully observable case remain constant, while the threshold $n_e(1)$ is more sensitive. Moreover, the thresholds n_U and n_L display a gently reducing trend in the partly observable case. When shortening the duration of semi dormancy, new arrivals predict that the system will be congested and overloaded, therefore, they enter reluctantly.

More importantly, Figures 8–11 reveal that the range of thresholds $\{n_L, n_{L+1}, \ldots, n_U\}$ in the partly observable case always contains inside the range of thresholds $n_e(0)$ and $n_e(2)$ in the fully observable case. In other words, customers prefer to queue if they master the information levels completely, which means that there exists an intermediate value in the partly observable case between two separate thresholds in the fully observable case.

6. Conclusions

This paper describes the individual and social strategic behaviors of customers in a single server Markov queueing system that incorporates the classical vacation and working vacation. We discuss and numerically compare the equilibrium thresholds of customers and the corresponding expected social benefits based on a reward-cost structure for both fully observable and partly observable cases. The research outcomes can help the managers to make appropriate strategic decisions utilizing the available information. Further, we conduct the sensitivity analysis of the expected social benefit as well as the equilibrium threshold for various parameters and different information levels. This study can reduce expenses by setting suitable indicators in the virtual technology and automatic management system. Further extensions of this work may explore the equilibrium behaviors under the unobservable cases in view of the application.

Acknowledgements. The authors are grateful to the anonymous referees for their detailed comments and valuable suggestions, and would like to thank the support by the National Natural Science Foundation (No. 6217012029), the Hebei Province Natural Science Foundation (No. A2019203313) and the Science Research Project of Education Department of Hebei Province (No. ZD2019079), China.

References

[1] A. Burnetas and A. Economou, Equilibrium customer strategies in a single server Markovian queue with setup times. *Queue. Syst.* **56** (2007) 213–228.

[2] N.H. Do, T.V. Do and A. Melikov, Equilibrium customer behavior in the M/M/1 retrial queue with working vacations and a constant retrial rate. *Oper. Res.* **20** (2020) 627–646.

[3] A. Economou and S. Kanta, Equilibrium balking strategies in the observable single-server queue with breakdowns and repairs. *Oper. Res. Lett.* **36** (2008) 696–699.

[4] S. Jin, S. Hao, X. Qie and W. Yue, A virtual machine scheduling strategy with a speed switch and a multi-sleep mode in cloud data centers. *J. Syst. Sci. Syst. Eng.* **28** (2019) 194–210.

[5] S. Jin, S. Hao and B. Wang, Virtual machine scheduling strategy based on dual-speed and work vacation mode and its parameter optimization. *J. Commun.* **38** (2017) 10–20.

[6] S. Jin, X. Qie, W Zhao, W. Yue and Y. Takahashi, A clustered virtual machine allocation strategy based on a sleep-mode with wake-up threshold in a cloud environment. *Ann. Oper. Res.* **293** (2020) 193–212.

[7] D.H. Lee, Equilibrium balking strategies in Markovian queues with a single working vacation and vacation interruption. *Qual. Technol. Quant. Manage.* **16** (2019) 355–376.

[8] J. Li, Analysis of the discrete-time Geo/G/1 working vacation queue and its application to network scheduling. *Comput. Indus. Eng.* **65** (2013) 594–604.

[9] K. Li, J. Wang, Y. Ren and J. Chang, Equilibrium joining strategies in M/M/1 queues with working vacation and vacation interruptions. *RAIRO – OR* **50** (2016) 451–471.

[10] R. Marek and K. Hoon, Cognitive systems and operations research in big data and cloud computing. *Ann. Oper. Res.* **265** (2018) 183–186.

[11] A.Z. Melikov, A.M. Rustamov and L.A. Ponomarenko, Approximate analysis of a queueing-inventory system with early and delayed server vacations. *Autom. Remote Cont.* **78** (2017) 1991–2003.

[12] P. Naor, The regulation of queue size by levying tolls. *Econometrica* **37** (1969) 15–24.

[13] I. Padmavathi, B. Sivakumar and G. Arivarignan, A retrial inventory system with single and modified multiple vacation for server. *Ann. Oper. Res.* **233** (2015) 335–364.

[14] Y. Peng and J. Wu, A Lévy-Driven stochastic queueing system with server breakdowns and vacations. *Mathematics* **8** (2020) 13–29.

[15] L. Servi and S. Finn, M/M/1 queues with working vacations (M/M/1/WV). *Perform. Eval.*, **50** (2002) 41–52.

[16] C. Shekhar, S. Varshney, and A. Kumar, Optimal and sensitivity analysis of vacation queueing system with F-policy and vacation interruption. *Syst. Eng.* **45** (2020) 7091–7107.

[17] W. Sun, S. Li, Y. Wang and N. Tian, Comparisons of exhaustive and non exhaustive M/M/1/N queues with working vacation and threshold policy. *J. Syst. Sci. Syst. Eng.* **28** (2019) 154–167.

[18] H. Takagi, *Queueing analysis, a foundation of performance evaluation, Vol. 1: Vacation and Priority Systems*. North-Holland, New York (1991).

[19] N. Tian and Z.G. Zhang, *Vacation queueing models: Theory and applications*. Springer-Verlag, New York, Inc (2006).

[20] R. Tian and Y. Wang, Optimal strategies and pricing analysis in M/M/1 queues with a single working vacation and multiple vacations. *RAIRO – OR*, **54** (2020) 1593–1612.

[21] J. Wang, Y. Zhang and Z.G. Zhang, Strategic joining in an M/M/k queue with asynchronous and synchronous multiple vacations. *J. Oper. Res. Soc.* **2** (2019) 1–19.

[22] F. Yang, Y. Jiang and Q. Li. Mean-field macro computation in large-scale cloud service systems with resource management and jobs scheduling. *J. Syst. Sci. Syst. Eng.* **28** (2019) 238–261.
[23] F. Zhang and J. Wang, Equilibrium analysis of the observable queue with balking and delayed repairs. *Appl. Math. Comput.* **218** (2011) 2716–2729.
[24] F. Zhang, J. Wang, and B. Liu, Equilibrium balking strategies in Markovian queues with working vacations. *Appl. Math. Model.* **37** (2013) 8264–8282.