Queueing Models and
Service Management
Vol. 8, No. 4, page 1–26, 2025

QMSM
©AU 2023

Customer Balking Behaviors in a Multi-server Queue with Synchronous Multiple Uninterrupted Vacations Under N-policy

Wei Sun[1,2], Xumeng Xie[1,*] and Zhiyuan Zhang[1]

[1]School of Economics and Management
Yanshan University, Qinhuangdao, 066004, China

[2]Shenzhen Research Institute
Yanshan University, Shenzhen, 518063, China

(*Received February 2025; accepted August 2025*)

Abstract: This paper studies the customers' equilibrium and socially optimal balking behaviors in an $M/M/c$ queue with synchronous multiple uninterrupted vacations under N-policy. Considering four cases where the information about the system is entirely/nearly observable or nearly/entirely unobservable, we obtain and compare the customers' equilibrium and socially optimal balking strategies, set the pricing strategies and analyze the system manager's benefits, respectively. It is shown that the pricing strategy varies under different information levels. What's more, regardless of the information level, the best N-policy should be designed for social optimization. However, when the threshold N is pre-determined, disclosing the information of the servers' status to customers is a better choice no matter the queue length is observable or not.

Keywords: Balking behaviors, multi-server queue, N-policy, pricing strategy, synchronous multiple uninterrupted vacations.

1. Introduction

In the queueing system with vacation(s), it is ideally assumed that the server enters a vacation status upon completing all service tasks, employing this period for maintenance, repairs or auxiliary operations, and resumes service immediately upon a customer's arrival. While this classic vacation policy aims to eliminate server idleness and optimize resource utilization, it often induces frequent on-off switching of the server. Such operational fluctuations not only accelerate equipment wear and tear but also incur substantial switching costs. To alleviate the adverse effects, scholars introduced the N-policy —–a control mechanism that regulates server vacations by activating the server only when a predefined number N of customers have accumulated in the queue. This policy enhances both resource efficiency and economic performance by reducing unnecessary server activations, thereby striking a balance between operational stability and cost control. Consequently, N-policy has been

* Corresponding author
Email: xumengxie@stumail.ysu.edu.cn

widely integrated into various classic and economic queueing models with server vacations, demonstrating its adaptability across diverse service systems.

Generally, studies about the classic queue with vacation(s) under N-policy primarily focus on the steady performance analysis of the system. Typically, researchers first constructed a new relational model, deduced the steady-state distributions of queue length next and then computed key performance measures like the average queue length, the mean waiting time, the busy/vacation period and so on [7, 8, 21]. After analyzing the system's performance, some of them also considered the total cost per unit time of the system and attained the optimal N minimizing the cost, if anything [1, 6, 9–11, 13, 16]. In addition, some scholars still demonstrated several stochastic decomposition theorems, decomposing some performance indicators like the queue length and the waiting time into the sum of a few independent random variables [12, 20, 22].

As for the economic queueing systems, since Guo et al. [2, 3] firstly investigated homogeneous and heterogenous customers' behaviors in an $M/M/1$ queue with vacations under N-policy, with full and no information separately, there have been scholars conducting subsequent studies on different queues with vacation(s) under N-policy from the economic point. As a complementary work, Guo et al. [4] subsequently discussed the same issue in the same model as [2] in two partially observable cases. In an unobservable $M/G/1$ queue with a removable server and N-policy, besides analyzing the customers' equilibrium and socially optimal behaviors, Tian et al. [15] also compared the two strategies. And then Sun et al. [14] studied the equilibrium and optimal balking behaviors of customers in the entirely observable and unobservable $M/M/1$ queues with multiple vacations under N-policy. In terms of the nearly observable and unobservable cases, the customers' equilibrium joining strategies and the social welfare in an $M/M/1$ queue with setup time and N-policy were analyzed by Hao et al. [5]. For the retrial queues, Zhou et al. [23] obtained the customers' equilibrium and optimal arrival rates as well as the optimal social welfare in an $M/M/1$ constant retrial queue with N-policy and setup time. Considering an $M/M/1$ retrial queue with breakdowns and multiple vacations under N-policy, Wang et al. [17] analyzed the equilibrium joining strategy in the noncooperative case, the socially optimal joining strategy in the cooperative case and the pricing strategy for the system. For a constant retrial $M/M/1$ queue with multiple vacations under N-policy, the equilibrium and optimal balking strategies of customers in the entirely observable and unobservable cases were discussed by Wang et al. [18]. Subsequently, Wang et al. [19] assumed that the customer would choose to join the orbit with probability p or leave forever with probability $1-p$ in the same model with [18], explored the cooperative and non-cooperative customers' joining cases, and then set up a price for social optimization.

By sorting out the literature, we notice that the majority of the current studies about the queues with vacation(s) under N-policy still concentrate on the single-server systems. However, it is manifested that the multi-server queues are more common in our daily life, including visible systems (e.g., hospital/bank queues) and invisible systems (e.g., online service platforms/call centers). In fact, N-policy offers broader applicability and operational value in the multi-server queues. Here are some examples. Multiple production lines (multiple servers) will not be started until the orders (customers) reach a certain number (N-policy). Idle virtual machines (multiple servers) that enter low-power hibernation will not be reactivated until a certain number of computational tasks (customers) await for being processed (N-policy). Couriers (multiple servers) will not start delivering until the deliveries (customers) are accumulated to a certain amount (N-policy). Nevertheless, there are only a few studies focusing on the multi-server queues like [12, 20, 21], where just the stationary performance of the systems is analyzed, and the individual behavioral decisions as well as the overall social welfare are neglected. In other words, there are no studies about the multi-server queues with vacation(s) under N-policy from the economic viewpoint currently.

On the other hand, N-policy exhibits two distinct operational paradigms in the queueing systems with vacation(s). The first paradigm, as explored in [2–7, 9–11, 15, 16, 22, 23] involves vacation interruption: the server terminates its vacation immediately upon the queue size reaching the predefined threshold N, even if the vacation activity (e.g., maintenance, administrative tasks) is ongoing. However, this “instant reactivation” mechanism faces practical limitations: operational processes often cannot be abruptly suspended. Here are some examples. A salesperson who is counting the goods cannot stop to check out at once. It is impossible for the machine being repaired to produce the components as soon as possible. And an employee on break may be extremely reluctant to interrupt his/her vacation to work in a minute. From this perspective, the other uninterrupted way to use N-policy in [1, 8, 12–14, 17–21] that the server(s) will reactivate only when there are N or more customers in the queue after a vacation while continuing the next vacation otherwise, seems to be more rational. This approach aligns better with real-world service dynamics by respecting the continuity of non-service operations, thereby reducing activation disruptions and associated inefficiencies.

As a result, it is meaningful to explore the customers’ strategic behaviors in the multi-server queues with uninterrupted vacation(s) under N-policy, both theoretically and practically. Inspired by but different from the studies in [20] and [14], we investigate the customers’ balking behaviors in an $M/M/c$ queueing system with synchronous multiple uninterrupted vacations under N-policy. Considering four information levels, we discuss the customers’ equilibrium and socially optimal balking strategies, based on which the pricing strategy is formulated, and then analyze the system manager’s revenue, respectively. Our main contributions include the following several aspects:

(1) In terms of each information level, the system’s stationary distributions as well as the customers’ mean sojourn time are deduced, based on which we discuss the customers’ equilibrium and socially optimal behaviors from the perspectives of individual optimality and overall optimality, respectively.

(2) In view of the inconformity between customers’ equilibrium and socially optimal balking strategies, we formulate the specific pricing strategy according to the corresponding information level, where the price can be adjusted in a range when the queue length is observable, and determined to be a certain level otherwise.

(3) We notice that if the information level is determined, there exists a specific threshold N maximizing the optimal social welfare. And if the threshold N is preset, disclosing the

information of the servers' status can improve the optimal social welfare, regardless of the queue length is observable or not.

For better comprehension of our model, let us take an example to illustrate it. Considering a company with several production equipment (multiple servers), to balance the production requirements for standardized and customized products, it produces in a combination of make-to-stock and make-to-order mode. Due to the relatively small demand but complicated production process, only when a certain number of customized orders (queue length) have been accumulated (N-policy) does the make-to-order production mode begin, otherwise the company always produces in the make-to-stock mode to guarantee the relatively great demand of the standardized products as much as possible. Moreover, even if the customized orders have accumulated to a certain number, it seems impossible to stop the ongoing make-to-stock production immediately, so the make-to-order production will not be activated until the end of make-to-stock production, where the make-to-order production and make-to-stock production can be regarded as the work status and vacation status of the multiple servers respectively. Customers decide whether to place the customized order or not in this company based on the lead time (the estimation of it depends on the revealment of the number of existing orders and the production status or not).

The residual parts of this paper are presented as follows. Section 2 includes the description of the queueing model and the explanations of some notations. In Section 3 and Section 4, the customers' equilibrium and socially optimal balking strategies, the pricing strategy and the system manager's revenue are discussed theoretically and numerically, in terms of four different information levels respectively. Last but not least, we summarize the paper and give some managerial suggestions in Section 5.

2. Model Descriptions and Notations

In an $M/M/c$ queue with synchronous multiple uninterrupted vacations under N-policy, the sufficient potential customers arrive in accordance with a Poisson process with rate Λ and each server serves at an exponential service rate μ. Once there is no customer in the queue, c servers start a vacation meanwhile exponentially distributed with parameter θ. After completing a vacation, all servers will come back to work together if the number of customers in the queue is not less than the preset threshold N, otherwise continue the next vacation. To avoid resource waste caused by the fact that the number of customers stimulating the dormant servers to work again is less than the number of servers, $N > c$ is surely assumed.

Let $(L_S(t), S_S(t))$ denote the system state at time t, where $L_S(t)$ is the queue length, i.e., the number of customers in the queue, and $S_S(t)$ is the status of servers, in which $S_S(t) = 0$ means that all servers are in the dormant status and $S_S(t) = 1$ means that all servers are in the working status. There are four information levels based the availability of the queue length and the servers' status to customers, which are described in Table 1.

Each customer will acquire a reward after receiving the service, but has to pay the sojourn cost at the same time. We use the linear cost-reward structure to calculate the customer's expected utility, i.e., a joining customer's expected utility after service is a linear function

Table 1. The four information levels

	entirely observable case	nearly observable case	nearly unobservable case	entirely unobservable case
$L_S(t)$	observable	observable	unobservable	unobservable
$S_S(t)$	observable	unobservable	observable	unobservable

of the mean sojourn time. An arriving customer will estimate the personal expected utility according to the system information he or she has firstly, and join the queue only when the utility is nonnegative. Once the customer chooses to join the queue, he or she will be served on a first-come first-served basis and leave the system instantly after the service. What's more, neither queue-jumping nor reneging is allowed. In addition, each server has to bear both the service cost throughout the working period and the switching cost owing to the switch of turning on and off. Thus, we define the social welfare per unit time as the difference between the total rewards of all joining customers and the sojourn cost of them, as well as the service cost and the switching cost of all servers.

Besides the system parameters above, we also define some other notations related to the customers' behaviors as follows.

- R_c = the reward after service of a joining customer.
- C_c = the sojourn cost per unit time of a joining customer.
- T = the mean sojourn time of a joining customer.
- U = the expected utility of a joining customer.
- C_s = the service cost per unit time of a working server.
- C_t = the single switching cost of a server.
- S = the social welfare of the whole system per unit time.
- P = the fee charging a customer.
- R_s = the revenue per unit time of the system manager after charging customers.
- n = the balking threshold in the observable cases.
- λ = the effect joining rate in the unobservable cases.
- ρ = the service intensity per unit time.
- subscript i = the status of all servers, where $i = 0$ means the servers are in the dormant status, $i = 1$ means the servers are in the working status.
- subscript $o/eo/no/nu/eu$ = the observable/entirely observable/nearly observable/nearly unobservable/entirely unobservable case.
- superscript e/* = some equilibrium/socially optimal indicators.

3. The Observable Cases

Due to the acquisition of the queue length information, the customers will follow the threshold balking strategies in the observable cases, that is, they will join the queue when the queue length does not exceed a threshold, and balk otherwise. In this section, we will discuss the customers' balking behaviors in two observable cases separately.

3.1. The entirely observable case

In the entirely observable case, the customers will follow a threshold balking strategy (n_{eo0}, n_{eo1}) since they grasp both the queue length and the servers' status, i.e., when the servers are on vacation or working, they will join the queue if the queue length does not exceed the corresponding threshold, otherwise choose to balk. In the observable cases, that all customers balk does indeed represent a theoretically valid equilibrium, but servers will never be activated in this scenario, which means that all customers cannot get the service. While acknowledging the existence of this equilibrium, the study of such an equilibrium holds limited practical relevance, since it describes a system where service provision never materializes, effectively negating the fundamental purpose of queueing systems designed to serve customers. Thus, we indeed focus on characterizing the subgame perfect Nash equilibrium, where the dormant servers can be activated normally and customers can optimize their decisions at every possible system state. Firstly, we need to discuss how to make the precondition that the dormant servers can be activated successfully attainable.

Let us mark a joining customer when the system's state is (l,i) (if $i=0$, $l\geq 0$, and if $i=1$, $l\geq 1$). If $i=0$ and $0\leq l<c$, this customer will not get the service until $N-l-1$ customers join the queue and the servers finish their vacation. If $i=0$ and $c\leq l<N-1$, besides the time of $N-l-1$ customers joining the queue and the servers completing the vacation, this customer has to wait for $l-c+1$ customers finishing their service before getting service. If $i=0$ and $l\geq N-1$, this customer will be served after the completion of the servers' vacation and $l-c+1$ customers' service. If $i=1$ and $1\leq l<c$, this customer will receive service directly. And if $i=1$ and $l\geq c$, this customer will get service only after $l-c+1$ customers completing their service. As a result, this customer's mean sojourn time can be summarized as

$$T_{eo}(l,0)=\begin{cases}\dfrac{N-l-1}{\Lambda}+\dfrac{1}{\theta}+\dfrac{1}{\mu},\ 0\leq l\leq c-1;\\ \dfrac{N-l-1}{\Lambda}+\dfrac{1}{\theta}+\dfrac{l+1}{c\mu},\ c\leq l\leq N-2;\\ \dfrac{1}{\theta}+\dfrac{l+1}{c\mu},\ l\geq N-1,\end{cases} \tag{1}$$

$$T_{eo}(l,1)=\begin{cases}\dfrac{1}{\mu},\ 1\leq l\leq c-1;\\ \dfrac{l+1}{c\mu},\ l\geq c.\end{cases} \tag{2}$$

To ensure that the servers on vacation can be resumed smoothly, it is necessary to ensure that all arriving customers when the servers are on vacation and the queue length is $l(0\leq l\leq N-1)$ choose to join the queue, i.e., $n_{eo0}\geq N-1$. Therefore, the key is to satisfy that the reward of a joining customer is not less than the highest sojourn cost when the system's state is $(l,0)(0\leq l\leq N-1)$. Next, let us discuss the longest sojourn time of the customer who joins the queue when the system's state is $(l,0)(0 \leq l \leq N-1)$. When $0 \leq l \leq c-1$, the first customer has the longest sojourn time $(N-1)/\Lambda + 1/\theta + 1/\mu$. And when $c \leq l \leq N-1$ and $\Lambda \leq c\mu$, the $c+1th$ customer needs to bear the longest sojourn time $(N-c-1)/\Lambda + 1/\theta + (c+1)/(c\mu)$. Otherwise if $\Lambda > c\mu$, the Nth customer' sojourn time is $1/\theta + N/(c\mu)$, which is the longest one. To sum up, the sufficient condition to ensure the normal operation of the entirely observable system can be given in Lemma 3.1.

Lemma 3.1. *If and only if (1) $\rho_o = \Lambda/(c\mu) \leq 1$ and $R_c/C_c \geq (N-1)/\Lambda + 1/\theta + 1/\mu$, or (2) $\rho_o = \Lambda/(c\mu) > 1$ and $R_c/C_c \geq max\{(N-1)/\Lambda + 1/\theta + 1/\mu, 1/\theta + N/(c\mu)\}$ is satisfied, the normal operation of the entirely observable $M/M/c$ queueing system with synchronous multiple uninterrupted vacations under N-policy can be guaranteed.*

If Lemma 3.1 is satisfied, we can infer that $n_{eo0}^{e} \geq N-1$. And it is noticed that when $0 < l \leq N-1$, $T_{eo}(l,1) < T_{eo}(l,0)$, which manifests that the customers will definitely join the queue when the system is at state $(l,1)(1 < l \leq N-1)$, so that $n_{eo1}^{e} \geq N-1$ can be deduced. Therefore, considering the expected utility of the joining customer when the system is at state $(l,i)(l \geq N-1, i=0,1)$ and solving $U_{eo}(l,i) = R_c - C_c T_{eo}(l,i) = 0 (l \geq N-1, i=0,1)$, $\left(n_{eo0}^{e}, n_{eo1}^{e}\right)$ are obtained.

Theorem 3.2. *On the premise of Lemma 3.1, the equilibrium balking thresholds of customers in the entirely observable $M/M/c$ queueing system with synchronous multiple uninterrupted vacations under N-policy are*

$$n_{eo0}^{e} = \left\lfloor c\mu\left(\frac{R_c}{C_c} - \frac{1}{\theta}\right)\right\rfloor - 1, n_{eo1}^{e} = \left\lfloor \frac{R_c c\mu}{C_c} \right\rfloor - 1. \tag{3}$$

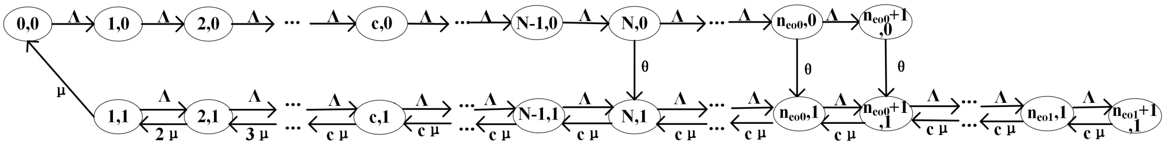

Figure 1. State transition diagram in the entirely observable case if $N-1 \leq n_{eo0}^{*} \leq n_{eo1}^{*}$.

Next, to get the socially optimal balking thresholds $\left(n_{eo0}^{*}, n_{eo1}^{*}\right)$, we need to analyze the stationary distributions of the entirely observable system firstly. Although we can find that $n_{eo0}^{e} < n_{eo1}^{e}$ from Theorem 3.2, there may theoretically exist three cases about the relationship between n_{eo0}^{*} and n_{eo1}^{*}: $N-1 \leq n_{eo0}^{*} \leq n_{eo1}^{*}$, $N-1 \leq n_{eo1}^{*} \leq n_{eo0}^{*}$ and $n_{eo1}^{*} \leq N-1 \leq n_{eo0}^{*}$. In the first place, we denote the stationary distributions of the entirely observable system as $p_{eo}^{1}(l,i)$ if $N-1 \leq n_{eo0}^{*} \leq n_{eo1}^{*}$, where $(l,i) \in \{(l,0)|0 \leq l \leq n_{eo0}+1\} \cup \{(l,1)|1 \leq l \leq n_{eo1}+1\}$, and the corresponding state transition diagram is portrayed in Figure 1.

Theorem 3.3. *If $N-1 \leq n_{eo0}^{*} \leq n_{eo1}^{*}$, the stationary distributions of the entirely observable $M/M/c$ queueing system with synchronous multiple uninterrupted vacations under N-policy are*

$$p_{eo}^{1}(l,0)=\begin{cases} Z_{eo}^{1},\ 0\le l\le N-1; \\ Z_{eo}^{1}\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{l+1-N}, N\le l\le n_{eo0}; \\ Z_{eo}^{1}\dfrac{\Lambda}{\theta}\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{n_{eo0}+1-N}, l=n_{eo0}+1, \end{cases} \tag{4}$$

$$p_{eo}^{1}(l,1)=\begin{cases} Z_{eo}^{1}\rho_o{}^{l}\alpha_l,\ 1\le l\le c; \\ Z_{eo}^{1}\left[\rho_o{}^{l}\alpha_{c-1}+\dfrac{\rho_o(1-\rho_o{}^{l+1-c})}{1-\rho_o}\right], c+1\le l\le N; \\ Z_{eo}^{1}\left\{\rho_o{}^{l}\alpha_{c-1}+\beta_l+\dfrac{\Lambda\rho_o{}^{l-N}}{\Lambda+\theta-c\mu}\left[1-\left(\dfrac{c\mu}{\Lambda+\theta}\right)^{l-N}\right]\right\}, N+1\le l\le n_{eo0}+1; \\ Z_{eo}^{1}\left\{\begin{array}{l}\rho_o{}^{n_{eo0}+2}\alpha_{c-1}-\rho_o\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{n_{eo0}+1-N}+\beta_{n_{eo0}+2} \\ +\dfrac{\Lambda\rho_o{}^{n_{eo0}+1-N}}{\Lambda+\theta-c\mu}\left[\rho_o+\dfrac{\theta-c\mu}{\Lambda+\theta}\left(\dfrac{c\mu}{\Lambda+\theta}\right)^{n_{eo0}-N}\right]\end{array}\right\}, l=n_{eo0}+2; \\ Z_{eo}^{1}\left[\begin{array}{l}\rho_o{}^{l}\alpha_{c-1}+\dfrac{\Lambda\rho_o{}^{l-N}}{\Lambda+\theta-c\mu}+\beta_l \\ +\dfrac{(\Lambda-c\mu)\rho_o{}^{l-n_{eo0}}}{(\Lambda+\theta-c\mu)(1-\rho_o)}\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{n_{eo0}+1-N}\end{array}\right], n_{eo0}+3\le l\le n_{eo1}+1, \end{cases} \tag{5}$$

where

$$\begin{cases} \alpha_l=\dfrac{c^l}{l!}\sum\limits_{i=0}^{l-1}i!\left(\dfrac{\mu}{\Lambda}\right)^{i},\ 1\le l\le c; \\ \beta_l=\dfrac{\rho_o{}^{l+1-N}(1-\rho_o{}^{N+1-c})}{1-\rho_o},N+1\le l\le n_{eo1}+1; \\ Z_{eo}^{1}=\left\{\begin{array}{l}\sum\limits_{l=1}^{c}\rho_o{}^{l}\alpha_l+\dfrac{\rho_o{}^{c+1}(1-\rho_o{}^{n_{eo1}+1-c})}{1-\rho_o}\alpha_{c-1}+\dfrac{N+\rho_o(\rho_o-c)}{1-\rho_o} \\ +\dfrac{\rho_o\beta_{n_{eo0}+1}}{\rho_o-1}+\dfrac{\Lambda\left[\Lambda-c\mu+\theta\left(1-\rho_o{}^{n_{eo1}+2-N}\right)\right]}{\theta(1-\rho_o)(\Lambda+\theta-c\mu)} \\ +\left[\dfrac{\Lambda(\Lambda-\theta\rho_o)}{\theta(\Lambda+\theta-c\mu)}+\dfrac{\rho_o{}^{3}(\Lambda-c\mu)(1-\rho_o{}^{n_{eo1}-n_{eo0}-1})}{(\Lambda+\theta-c\mu)(1-\rho_o)^2}\right]\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{n_{eo0}+1-N}\end{array}\right\}^{-1}. \end{cases} \tag{6}$$

Proof. From Figure 1, we can attain the stationary transition probability equations

$$\Lambda p_{eo}^{1}(0,0)=\mu p_{eo}^{1}(1,1), \tag{7}$$

$$\Lambda p_{eo}^{1}(l,0)=\Lambda p_{eo}^{1}(l-1,0),1\le l\le N-1, \tag{8}$$

$$(\Lambda+\theta)\,p_{eo}^{1}(l,0)=\Lambda p_{eo}^{1}(l-1,0),N\le l\le n_{eo0}, \tag{9}$$

$$\theta p_{eo}^{1}(n_{eo0}+1,0)=\Lambda p_{eo}^{1}(n_{eo0},0), \tag{10}$$

$$(\Lambda+\mu)\,p_{eo}^{1}(1,1)=2\mu p_{eo}^{1}(2,1), \tag{11}$$

$$(\Lambda+l\mu)\,p_{eo}^{1}(l,1)=\Lambda p_{eo}^{1}(l-1,1)+(l+1)\,\mu p_{eo}^{1}(l+1,1),2\le l\le c-1, \tag{12}$$

$$(\Lambda+c\mu)\,p_{eo}^{1}(l,1)=\Lambda p_{eo}^{1}(l-1,1)+c\mu p_{eo}^{1}(l+1,1),$$

$$c \le l \le N-1 \; \& \; n_{eo0}+2 \le l \le n_{eo1}, \tag{13}$$

$$(\Lambda + c\mu)\, p_{eo}^{1}(l,1) = \Lambda p_{eo}^{1}(l-1,1) + \theta p_{eo}^{1}(l,0) + c\mu p_{eo}^{1}(l+1,1), \quad N \le l \le n_{eo0}+1, \tag{14}$$

$$c\mu p_{eo}^{1}(n_{eo1}+1,1) = \Lambda p_{eo}^{1}(n_{eo1},1). \tag{15}$$

Let us solve the expression of $\left\{ p_{eo}^{1}(l,0) \,|\, 0 \le l \le n_{eo0}+1 \right\}$ firstly. Regarding $p_{eo}^{1}(0,0) = Z_{eo}^{1}$ as an undetermined constant temporarily, we can readily get $p_{eo}^{1}(l,0)\,(1 \le l \le N-1)$ from Eq.(8). Afterwards, substituting $p_{eo}^{1}(N-1,0)$ into

$$p_{eo}^{1}(l,0) = \left(\frac{\Lambda}{\Lambda+\theta} \right)^{l+1-N} p_{eo}^{1}(N-1,0), N \le l \le n_{eo0}, \tag{16}$$

which can be obtained from Eq.(9), $p_{eo}^{1}(l,0)\,(N \le l \le n_{eo0})$ is attainable. And then we can gain $p_{eo}^{1}(n_{eo0}+1,0)$ by plugging $p_{eo}^{1}(n_{eo0},0)$ into Eq.(10).

Next, let us consider the expression of $\left\{ p_{eo}^{1}(l,1) \,|\, 1 \le l \le n_{eo1}+1 \right\}$. According to Eq. (12), we can establish the recurrence relation formula

$$\left[l\mu p_{eo}^{1}(l,1) - \Lambda p_{eo}^{1}(l-1,1) \right] - \left[(l-1)\mu p_{eo}^{1}(l-1,1) - \Lambda p_{eo}^{1}(l-2,1) \right] = 0, 3 \le l \le c, \tag{17}$$

and then get

$$l\mu p_{eo}^{1}(l,1) - \Lambda p_{eo}^{1}(l-1,1) = 2\mu p_{eo}^{1}(2,1) - \Lambda p_{eo}^{1}(1,1), 3 \le l \le c. \tag{18}$$

Solving $p_{eo}^{1}(1,1)$ and $p_{eo}^{1}(2,1)$ from Eq.(7) and Eq.(11) and substituting them into Eq. (18), we can attain $p_{eo}^{1}(l,1)\,(1 \le l \le c)$. Similarly, the recurrence relation formulae

$$\left[c\mu p_{eo}^{1}(l,1) - \Lambda p_{eo}^{1}(l-1,1) \right] - \left[c\mu p_{eo}^{1}(l-1,1) - \Lambda p_{eo}^{1}(l-2,1) \right] = 0, \quad c+1 \le l \le N \; \& \; n_{eo0}+3 \le l \le n_{eo1}+1, \tag{19}$$

and

$$\left[c\mu p_{eo}^{1}(l,1) - \Lambda p_{eo}^{1}(l-1,1) \right] - \left[c\mu p_{eo}^{1}(l-1,1) - \Lambda p_{eo}^{1}(l-2,1) \right] = -\theta p_{eo}^{1}(l-1,0), \quad N+1 \le l \le n_{eo0}+2, \tag{20}$$

can be built from Eq.(13) and Eq.(14) respectively, by which we can attain

$$c\mu p_{eo}^{1}(l,1) - \Lambda p_{eo}^{1}(l-1,1) = c\mu p_{eo}^{1}(c,1) - \Lambda p_{eo}^{1}(c-1,1), c+1 \le l \le N, \tag{21}$$

$$c\mu p_{eo}^{1}(l,1) - \Lambda p_{eo}^{1}(l-1,1) = c\mu p_{eo}^{1}(N,1) - \Lambda p_{eo}^{1}(N-1,1) - \theta \sum_{i=N}^{l-1} p_{eo}^{1}(i,0), N+1 \le l \le n_{en0}+1, \tag{22}$$

and

$$c\mu p_{eo}^{1}(l,1)-\Lambda p_{eo}^{1}(l-1,1)=c\mu p_{eo}^{1}(n_{eo0}+2,1)\\ -\Lambda p_{eo}^{1}(n_{eo0}+1,1),n_{eo0}+3\leq l\leq n_{eo1}+1. \qquad (23)$$

After some derivations and simplifications, $p_{eo}^{1}(l,1)\,(c+1\leq l\leq n_{eo1}+1)$ can be attained.

At last, the undetermined constant Z_{eo}^{1} can be computed based on the normalization condition

$$\sum_{l=0}^{n_{eo0}+1}p_{eo}^{1}(l,0)+\sum_{l=1}^{n_{eo1}+1}p_{eo}^{1}(l,1)=1. \qquad (24)$$

According to Theorem 3.3, if $N-1\leq n_{eo0}^{*}\leq n_{eo1}^{*}$, the probability that customers balk is $p_{eo}^{1}(n_{eo0}+1,0)+p_{eo}^{1}(n_{eo1}+1,1)$, and the probabilities that the servers are on vacation and working are $\sum_{l=0}^{n_{eo0}+1}p_{eo}^{1}(l,0)$ and $\sum_{l=1}^{n_{eo1}+1}p_{eo}^{1}(l,1)$ respectively. The servers will be turned on with the probability $\theta/(\Lambda+\theta)$ when the system is at state $(l,0)(N\leq l\leq n_{eo0})$, while they will start working after vacation definitely when the system's state is $(n_{eo0}+1,0)$. And then we can get the switching rate between on and off of all servers, which is expressed by $\sum_{l=N}^{n_{eo0}}\theta^{2}p_{eo}^{1}(l,0)/(\Lambda+\theta)+\theta p_{eo}^{1}(n_{eo0}+1,0)$. Thus, if $N-1\leq n_{eo0}^{*}\leq n_{eo1}^{*}$, the social welfare per unit time of the entirely observable system is

$$S_{eo}^{1}(n_{eo0},n_{eo1})=R_{c}\Lambda\left[1-p_{eo}^{1}(n_{eo0}+1,0)-p_{eo}^{1}(n_{eo0}+1,1)\right]-C_{c}\left[\sum_{l=1}^{n_{eo0}+1}lp_{eo}^{1}(l,0)\right.\\ \left.+\sum_{l=1}^{n_{eo1}+1}lp_{eo}^{1}(l,1)\right]-cC_{s}\sum_{l=1}^{n_{eo1}+1}p_{eo}^{1}(l,1)-cC_{t}\left[\sum_{l=N}^{n_{eo0}}\frac{\theta^{2}}{\Lambda+\theta}p_{eo}^{1}(l,0)+\theta p_{eo}^{1}(n_{eo0}+1,0)\right]. \qquad (25)$$

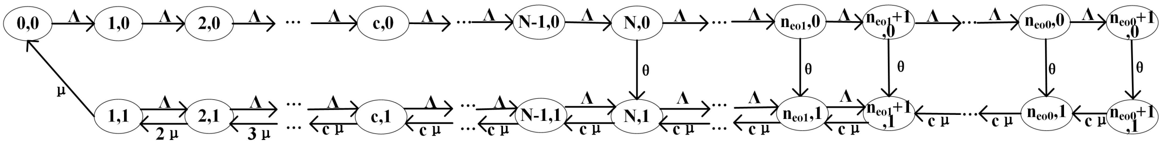

Figure 2. State transition diagram in the entirely observable case if $N-1\leq n_{eo1}^{*}\leq n_{eo0}^{*}$.

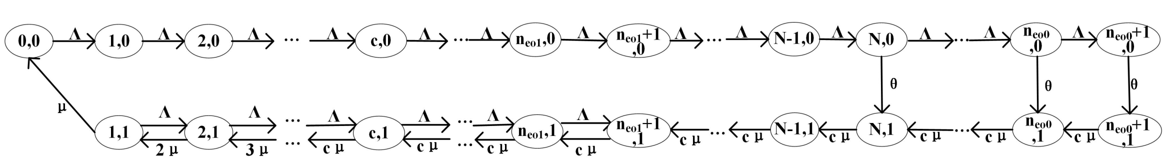

Figure 3. State transition diagram in the entirely observable case if $n_{eo1}^{*}\leq N-1\leq n_{eo0}^{*}$.

Next in the $N-1\leq n_{eo1}^{*}\leq n_{eo0}^{*}$ and $n_{eo1}^{*}\leq N-1\leq n_{eo0}^{*}$ cases, we denote the stationary distributions of the entirely observable system as $p_{eo}^{2}(l,i)$ and $p_{eo}^{3}(l,i)$ respectively, where $(l,i)\in\{(l,0)|0\leq l\leq n_{eo0}+1\}\cup\{(l,1)|1\leq l\leq n_{eo0}+1\}$, and the corresponding state transition diagrams are portrayed in Figure 2 and Figure 3 respectively. Since the derivation processes are similar with the case when $N-1\leq n_{eo0}^{*}\leq n_{eo1}^{*}$, we will not go into details and just give the final consequence in Theorem 3.4 and Theorem 3.5 here.

Theorem 3.4. *If $N-1 \leq n^*_{eo1} \leq n^*_{eo0}$, the stationary distributions of the entirely observable $M/M/c$ queueing system with synchronous multiple uninterrupted vacations under N-policy are*

$$p^2_{eo}(l,0)=\begin{cases} Z^2_{eo},\ 0\leq l\leq N-1; \\ Z^2_{eo}\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{l+1-N}, \ N\leq l\leq n_{eo0}; \\ Z^2_{eo}\dfrac{\Lambda}{\theta}\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{n_{eo0}+1-N}, \ l=n_{eo0}+1, \end{cases} \tag{26}$$

$$p^2_{eo}(l,1)=\begin{cases} Z^2_{eo}\rho_o{}^l\alpha_l,\ 1\leq l\leq c; \\ Z^2_{eo}\left[\rho_o{}^l\alpha_{c-1}+\dfrac{\rho_o(1-\rho_o{}^{l+1-c})}{1-\rho_o}\right],\ c+1\leq l\leq N; \\ Z^2_{eo}\left\{\rho_o{}^l\alpha_{c-1}+\beta_l+\dfrac{\Lambda\rho_o{}^{l-N}}{\Lambda+\theta-c\mu}\left[1-\left(\dfrac{c\mu}{\Lambda+\theta}\right)^{l-N}\right]\right\},\ N+1\leq l\leq n_{eo1}+1; \\ Z^2_{eo}\rho_o\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{l-N},\ n_{eo1}+2\leq l\leq n_{eo0}+1, \end{cases} \tag{27}$$

where

$$Z^2_{eo}=\left\{\begin{array}{l} \displaystyle\sum_{l=1}^{c}\rho_o{}^l\alpha_l+\frac{\rho_o{}^{c+1}\left(1-\rho_o{}^{n_{eo1}+1-c}\right)}{1-\rho_o}\alpha_{c-1}+\frac{N+\rho_o(\rho_o-c)}{1-\rho_o} \\ \displaystyle+\frac{\rho_o}{\rho_o-1}\beta_{n_{eo1}+1}+\frac{\Lambda\left[\Lambda-c\mu+\theta\left(1-\rho_o{}^{n_{eo1}+2-N}\right)\right]}{\theta(1-\rho_o)(\Lambda+\theta-c\mu)} \\ \displaystyle\left\{\frac{\Lambda^2}{\theta\left(\Lambda+\theta-c\mu\right)}+\frac{\rho_o\Lambda}{\theta}\left[1-\left(\frac{\Lambda}{\Lambda+\theta}\right)^{n_{eo0}-n_{eo1}}\right]\right\}\left(\frac{\Lambda}{\Lambda+\theta}\right)^{n_{eo1}+1-N} \end{array}\right\}^{-1}. \tag{28}$$

Theorem 3.5. *If $n^*_{eo1} \leq N-1 \leq n^*_{eo0}$, the stationary distributions of the entirely observable $M/M/c$ queueing system with synchronous multiple uninterrupted vacations under N-policy are*

$$p^3_{eo}(l,0)=\begin{cases} Z^3_{eo},\ 0\leq l\leq N-1; \\ Z^3_{eo}\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{l+1-N}, \ N\leq l\leq n_{eo0}; \\ Z^3_{eo}\dfrac{\Lambda}{\theta}\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{n_{eo0}+1-N}, \ l=n_{eo0}+1, \end{cases} \tag{29}$$

$$p_{eo}^3(l,1)=\begin{cases} Z_{eo}^3{\rho_o}^l\alpha_l,\ 1\le l\le c; \\ Z_{eo}^3\left[{\rho_o}^l\alpha_{c-1}+\dfrac{\rho_o\left(1-{\rho_o}^{l+1-c}\right)}{1-\rho_o}\right],\ c+1\le l\le n_{eo1}+1; \\ \rho_o Z_{eo}^3,\ n_{eo1}+2\le l\le N; \\ Z_{eo}^3\rho_o\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{l-N},\ N+1\le l\le n_{eo0}+1, \end{cases} \tag{30}$$

where

$$Z_{eo}^3=\left\{\begin{array}{l} \displaystyle\sum_{l=1}^{c}{\rho_o}^l\alpha_l+\frac{{\rho_o}^{c+1}\left(1-{\rho_o}^{n_{eo1}+1-c}\right)}{1-\rho_o}\alpha_{c-1}+\frac{{\rho_o}^3\left({\rho_o}^{n_{eo1}+1-c}-1\right)}{(1-\rho_o)^2} \\ \displaystyle+\frac{\rho_o\left[\rho_o\left(n_{eo1}+1-N\right)-c\right]+N}{1-\rho_o}+\frac{\Lambda}{\theta}\left\{\rho_o\left[1-\left(\frac{\Lambda}{\Lambda+\theta}\right)^{n_{eo0}+1-N}\right]+1\right\} \end{array}\right\}^{-1}. \tag{31}$$

On the basis of Theorem 3.4 and Theorem 3.5, the social welfare per unit time of the entirely observable system when $N-1\le n_{eo1}^*\le n_{eo0}^*$ or $n_{eo1}^*\le N-1\le n_{eo0}^*$ can be expressed by

$$\begin{aligned} S_{eo}^j\left(n_{eo0},n_{eo1}\right)=R_c\Lambda&\left[1-p_{eo}^j\left(n_{eo0}+1,0\right)-\sum_{l=n_{eo1}+1}^{n_{eo0}+1}p_{eo}^j(l,1)\right] \\ &-C_c\left[\sum_{l=1}^{n_{eo0}+1}lp_{eo}^j(l,0)+\sum_{l=1}^{n_{eo1}+1}lp_{eo}^j(l,1)\right]-cC_s\sum_{l=1}^{n_{eo0}+1}p_{eo}^j(l,1) \\ &-cC_t\left[\sum_{l=N}^{n_{eo0}}\frac{\theta^2}{\Lambda+\theta}p_{eo}^j(l,0)+\theta p_{eo}^j\left(n_{eo0}+1,0\right)\right],j=2,3. \end{aligned} \tag{32}$$

To sum up, the social welfare per unit time of the entirely observable queueing system

$$S_{eo}\left(n_{eo0},n_{eo1}\right)=\begin{cases} S_{eo}^1\left(n_{eo0},n_{eo1}\right),\ N-1\le n_{eo0}^*\le n_{eo1}^*; \\ S_{eo}^2\left(n_{eo0},n_{eo1}\right),\ N-1\le n_{eo1}^*\le n_{eo0}^*; \\ S_{eo}^3\left(n_{eo0},n_{eo1}\right),\ n_{eo1}^*\le N-1\le n_{eo0}^*. \end{cases} \tag{33}$$

is obtained. And then solving the optimization problem *max* $S_{eo}\left(n_{eo0},n_{eo1}\right)$, the customers' socially optimal balking thresholds $\left(n_{eo0}^*,n_{eo1}^*\right)$ and the optimal social welfare $S_{eo}^*\left(n_{eo0}^*,n_{eo1}^*\right)$ can be obtained.

3.2. The nearly observable case

In the nearly observable case, the customers' balking threshold is unified as n_{no} owing to the lack of the information about the servers' status. We denote the stationary distributions of the nearly observable system as $p_{no}(l,i)$, where $(l,i)\in\{(l,0)|0\le l\le n_{no}+1\}\cup\{(l,1)|1\le$

$l \leq n_{no}+1\}$, and the corresponding state transition diagram is portrayed in Figure 4. Just let $n_{eo0} = n_{eo1} = n_{no}$ in the case where $N-1 \leq n^*_{eo0} \leq n^*_{eo1}$ in the entirely observable case, we can attain the stationary distributions of the nearly observable queueing system, which is summarized in Theorem 3.6.

Theorem 3.6. *The stationary distributions of the nearly observable $M/M/c$ queueing system with synchronous multiple uninterrupted vacations under N-policy are*

$$p_{no}(l,0) = \begin{cases} Z_{no}, \ 0 \leq l \leq N-1; \\ Z_{no}\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{l+1-N}, \ N \leq l \leq n_{no}; \\ Z_{no}\dfrac{\Lambda}{\theta}\left(\dfrac{\Lambda}{\Lambda+\theta}\right)^{n_{no}+1-N}, \ l = n_{no}+1, \end{cases} \tag{34}$$

$$p_{no}(l,1) = \begin{cases} Z_{no}{\rho_o}^l\alpha_l, \ 1 \leq l \leq c; \\ Z_{no}\left[{\rho_o}^l\alpha_{c-1} + \dfrac{\rho_o\left(1-{\rho_o}^{l+1-c}\right)}{1-\rho_o}\right], \ c+1 \leq l \leq N; \\ Z_{no}\left\{{\rho_o}^l\alpha_{c-1} + \beta_l + \dfrac{\Lambda{\rho_o}^{l-N}}{\Lambda+\theta-c\mu}\left[1-\left(\dfrac{c\mu}{\Lambda+\theta}\right)^{l-N}\right]\right\}, \ N+1 \leq l \leq n_{no}+1, \end{cases} \tag{35}$$

where

$$Z_{no} = \left\{ \begin{array}{l} \displaystyle\sum_{l=1}^{c}{\rho_o}^l\alpha_l + \frac{{\rho_o}^{c+1}\left(1-{\rho_o}^{n_{no}+1-c}\right)}{1-\rho_o}\alpha_{c-1} + \frac{N+\rho_o(\rho_o-c)}{1-\rho_o} + \frac{{\rho_o}^{n_{no}+3-N}\left({\rho_o}^{N+1-c}-1\right)}{(1-\rho_o)^2} \\ \displaystyle+\frac{\Lambda\left[\Lambda-c\mu+\theta\left(1-{\rho_o}^{n_{no}+2-N}\right)\right]}{\theta(1-\rho_o)(\Lambda+\theta-c\mu)} + \frac{\Lambda^2}{\theta(\Lambda+\theta-c\mu)}\left(\frac{\Lambda}{\Lambda+\theta}\right)^{n_{no}+1-N} \end{array} \right\}^{-1}. \tag{36}$$

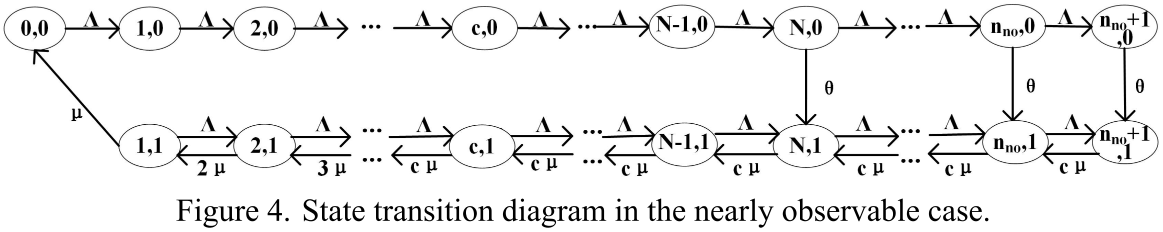

Figure 4. State transition diagram in the nearly observable case.

Denoting the probability that the queue length is $l(0 \leq l \leq n_{no}+1)$ and the conditional probability that the servers' status is $i(i=0,1)$ when there are $l(1 \leq l \leq n_{no}+1)$ customers queueing up in the nearly observable case are $p_{no}(l)$ and $p_{no}(i|l)$ respectively, we can get

$$p_{no}(i|l) = \frac{p_{no}(l,i)}{p_{no}(l)} = \frac{p_{no}(l,i)}{p_{no}(l,0)+p_{no}(l,1)}, \ 1 \leq l \leq n_{no}+1, \ i=0,1 \tag{37}$$

according to Theorem 3.6. And then from Eqs.(1) and (2), the mean sojourn time of customers joining the queue when there are already k customers in the system can be expressed by

$$T_{no}(l)=\begin{cases}T_{eo}(0,0),\ l=0;\\ T_{eo}(l,0)p_{no}(0|l)+T_{eo}(l,1)p_{no}(1|l),\ 1\leq l\leq n_{no}+1.\end{cases}\tag{38}$$

Just like the entirely observable case, that all customers balk is also an equilibrium strategy in the nearly observable case, but this case is also of no practical significance. Similarly, to analyze the other equilibrium case where the dormant servers can be activated successfully, firstly it is necessary to ensure the customers arriving when the queue length is not more than $N-1$ all join the queue, i.e., $n_{no}\geq N-1$. And then, the positive equilibrium balking threshold of customers in the nearly observable case is obtained by solving the equation $U_{no}(l)=R_c-C_cT_{no}(l)=0(l\geq N-1)$.

Theorem 3.7. *If and only if $R_c/C_c\geq max\{T_{no}(0),T_{no}(1),...,T_{no}(N-1)\}$, the normal operation of the nearly observable $M/M/c$ queueing system with synchronous multiple uninterrupted vacations under N-policy can be guaranteed, and then the equilibrium balking threshold of customers is*

$$n_{no}^{e}=\lfloor n_{no}^{'}\rfloor,\tag{39}$$

where $n_{no}^{'}$ is the solution of the equation

$$R_c-C_c\frac{T_{eo}(l,0)p_{no}(l,0)+T_{eo}(l,1)p_{no}(l,1)}{p_{no}(l,0)+p_{no}(l,1)}=0,\ l\geq N-1.\tag{40}$$

Next, the social welfare per unit time of the nearly observable system can be expressed by

$$\begin{aligned}S_{no}(n_{no})=R_c\Lambda\left[1-p_{no}\left(n_{no}+1,0\right)-p_{no}\left(n_{no}+1,1\right)\right]-C_c\sum_{l=1}^{n_{no}+1}l\left[p_{no}\left(l,0\right)+p_{no}\left(l,1\right)\right]\\ -cC_s\sum_{l=1}^{n_{no}+1}p_{no}(l,1)-cC_t\left[\sum_{l=N}^{n_{no}}\frac{\theta^2}{\Lambda+\theta}p_{no}(l,0)+\theta p_{no}(n_{no}+1,0)\right]\end{aligned}\tag{41}$$

according to Theorem 3.6. And then we can gain the socially optimal balking threshold n_{no}^{*} and the optimal social welfare $S_{no}^{*}\left(n_{no}^{*}\right)$ by solving the optimization problem $max\ S_{no}\left(n_{no}\right)$.

3.3. Numerical analysis and comparisons

Based on the entirely observable and the nearly observable cases, we numerically reveal the relationship between the customers' equilibrium and socially optimal balking behaviors, analyze the pricing strategies and the system manager's revenue according to the information level, and explore the influence of the threshold N and the amount of information on the equilibrium and socially optimal balking thresholds, the pricing strategies, the revenue of the system manager and the socially optimal welfare.

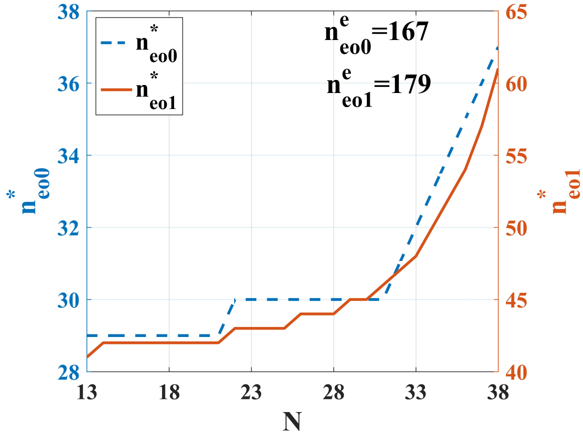

Figure 5. The customers' equilibrium and socially optimal balking thresholds in the entirely observable case.

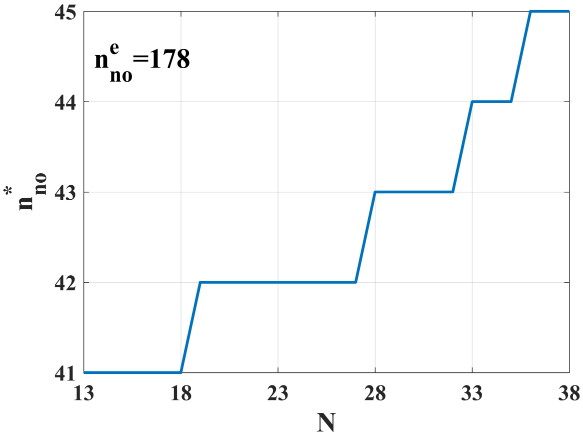

Figure 6. The customers' equilibrium and socially optimal balking thresholds in the nearly observable case.

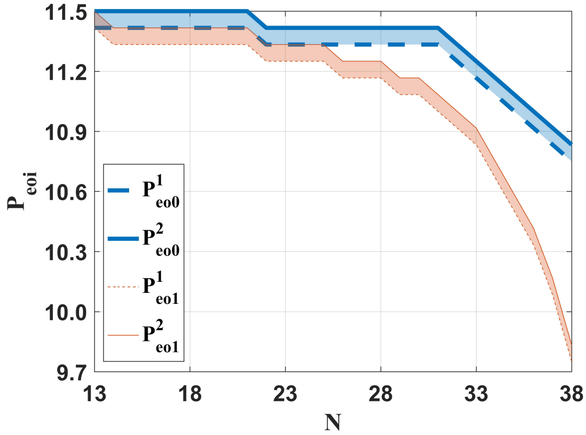

Figure 7. The pricing strategy in the entirely observable case.

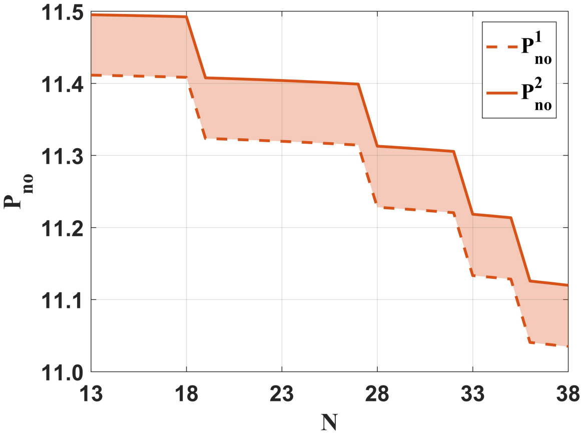

Figure 8. The pricing strategy in the nearly observable case.

Due to the relevant theoretical analysis where an arriving customer faces $l(l \leq N-1)$ customers in the queue chooses to join the queue is ensured in the observable cases, n^e_{eo0}, n^e_{eo1} and n^e_{no} are all not affected by N. And for the same queue length in the observable cases, compared with the dormant status, a joining customer can attain a better expected utility when the servers are working, so $n^e_{eo0} < n^e_{no1}$. As for the optimal balking thresholds, we can see that n^*_{eo0}, n^*_{eo1} and n^*_{no} all increase with respect to N on the whole from Figure 5 and Figure 6 [1] , which indicates that with the increase of N, more and more customers are needed to join the queue for the socially optimization when the queue length is observable, no matter the status of servers is observable or not.

Furthermore, by comparing Figure 5 and Figure 6, we can find that $n^e_{eo0} < n^e_{no} < n^e_{eo1}$ and

[1] The values of the legend parameters in Figure 5 - Figure10 are $\Lambda = 10, \mu = 1, c = 12, \theta = 1, R_c = 15, C_c = 1, C_s = 10, C_t = 8$.

$n^*_{eo0} < n^*_{no} \leq n^*_{eo1}$, and the case of $n^*_{no} = n^*_{eo1}$ is relatively rare, that is, the customers' balking behavior in the nearly observable case can be regarded as the comprehensive reflection of that in the entirely observable case when the servers are in two statuses $(i = 0 \text{ or } 1)$. On the other hand, what can be also found is that $n^*_{eo0} < n^e_{eo0}$, $n^*_{eo1} < n^e_{eo1}$ and $n^*_{no} < n^e_{no}$ from the comparison between Figure 5 and Figure 6, which implies that the customers' equilibrium and socially optimal behaviors is inconsistent in the observable cases, and their selfish behavior will lead to the over congestion of the system. And from the overall viewpoint, the social planner who intends to optimize the social welfare can make the pricing strategy based on the information level, i.e., charging customers, to hold the consistence between the customers' equilibrium and socially optimal balking behaviors. Thus, the price in the entirely observable case when the servers are in two statuses can be set as $P_{eoi} \in \left(P^1_{eoi}, P^2_{eoi}\right]$ $(i = 0$ or $1)$, where

$$\begin{cases} P^1_{eoi} = U_{eo}\left(n^*_{eoi} + 1, i\right) - U_{eo}\left(n^e_{eoi}, i\right), \\ P^2_{eoi} = U_{eo}\left(n^*_{eoi}, i\right) - U_{eo}\left(n^e_{eoi}, i\right). \end{cases} \tag{42}$$

and the price in the nearly observable case can be set as $P_{no} \in \left(P^1_{no}, P^2_{no}\right]$, where

$$\begin{cases} P^1_{no} = U_{no}\left(n^*_{no} + 1\right) - U_{no}\left(n^e_{no}\right), \\ P^2_{no} = U_{no}\left(n^*_{no}\right) - U_{no}\left(n^e_{no}\right). \end{cases} \tag{43}$$

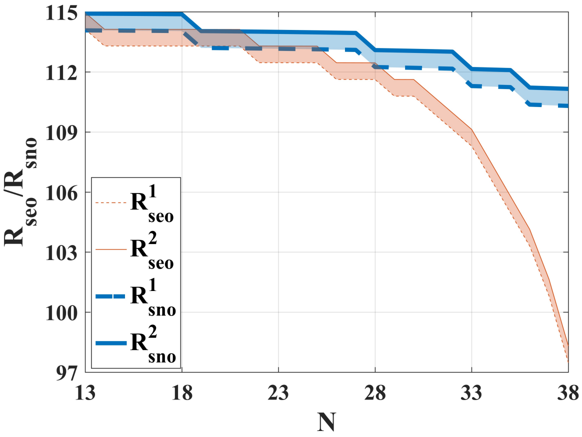

Figure 9. The system manager's revenue in the entirely/nearly observable case.

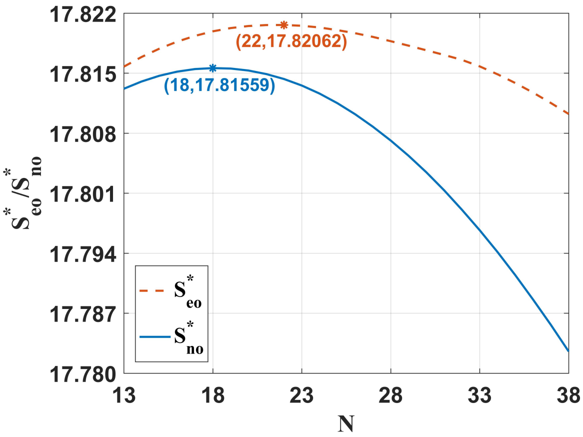

Figure 10. The optimal social welfare in the entirely/nearly observable case.

Due to the fixedness of the equilibrium balking thresholds and the increase of the socially optimal balking thresholds with respect to N, the gap of them is getting smaller and smaller, which manifests that less and less control is needed to customers in the observable cases. As a result, the prices in the observable cases all decrease with respect to N, as shown in Figures 7-8. What's more, we can also find that $P^1_{eo1} \leq P^1_{eo0}$ as well as $P^2_{eo1} \leq P^2_{eo0}$ is available in the entirely observable case.

After customers pay the fee, the system manager who is responsible for charging will get a revenue. And according to the corresponding pricing strategy and the stationary distributions, we can get the system manager's revenue in the entirely observable case as $R_{seo} \in$

$(R_{seo}^{1}, R_{seo}^{2}]$, where

$$R_{seo}^{j} = P_{eo0}^{j}\Lambda\sum_{l=0}^{n_{eo0}^{*}} p_{eo}^{h}(l,0) + P_{eo1}^{j}\Lambda\sum_{l=1}^{n_{eo1}^{*}} p_{eo}^{h}(l,1), j = 1,2, \tag{44}$$

if $max\, S_{eo}\left(n_{eo0}, n_{eo1}\right) = max\, S_{eo}^{k}\left(n_{eo0}, n_{eo1}\right)$, $h = k(k = 1,2,3)$. Similarly, the system manager's revenue in the nearly observable case is $R_{sno} \in \left(R_{sno}^{1}, R_{sno}^{2}\right]$, where

$$R_{sno}^{j} = P_{no}^{j}\Lambda\left[1 - p_{no}\left(n_{no}^{*}+1,0\right) - p_{no}\left(n_{no}^{*}+1,1\right)\right], j = 1,2. \tag{45}$$

Figure 9 shows the system manager's revenue in the observable cases with respect to N. Clearly, just like the fee charging customers, the system manager's revenue in the observable cases all decreases with respect to N overall.

At last, we also analyze the optimal social welfare in the observable cases with respect to N in Figure 10. Evidently, there exist different optimal Ns in the entirely and nearly observable cases maximizing the corresponding optimal social welfare respectively. In addition, it can be also found that $S_{eo}^{*} > S_{no}^{*}$, which manifests that when the queue length is observable, disclosing the information about the servers' status at the same time may be beneficial to improve the optimal social welfare.

4. The Unobservable Cases

In the unobservable cases, customers are incapable to get the queue length, so they will join the queue with a certain probability. Considering two unobservable cases, the customers' balking behaviors are analyzed in this section.

4.1. The nearly unobservable case

In the nearly unobservable case, customers will join the queue with different probabilities simply based on the available information of the servers' status. We denote the stationary distributions of the nearly unobservable system as $p_{nu}(l,i)$, where $(l,i) \in \{(l,0)|l \geq 0\} \cup \{(l,1)|l \geq 1\}$. The corresponding state transition diagram is portrayed in Figure 11. Referring to the derivation process of the stationary distributions in the entirely observable case, we can get the stationary distributions of the nearly unobservable system presented in Theorem 4.1.

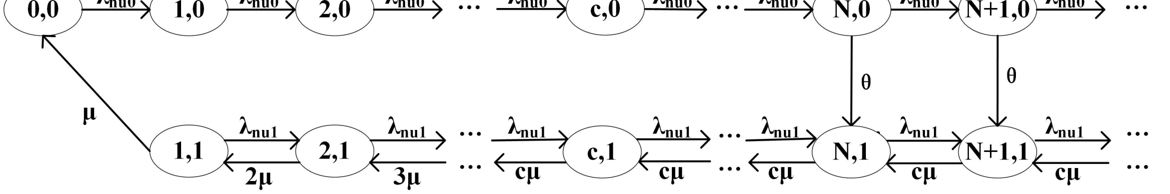

Figure 11. State transition diagram in the nearly unobservable case.

Theorem 4.1. *When $\rho_{nu} = \lambda_{nu1}/(c\mu) < 1$, the stationary distributions of the nearly unobservable $M/M/c$ queueing system with synchronous multiple uninterrupted vacations under N-policy are*

$$p_{nu}(l,0) = \begin{cases} Z_{nu}, 1 \le l \le N-1; \\ Z_{nu}\left(\dfrac{\lambda_{nu0}}{\lambda_{nu0}+\theta}\right)^{l+1-N}, l \ge N, \end{cases} \tag{46}$$

$$p_{nu}(l,1) = \begin{cases} Z_{nu}{\rho_{nu}}^{l}\gamma_l,\ 1 \le l \le c; \\ Z_{nu}\left[{\rho_{nu}}^{l}\gamma_{c-1} + \dfrac{\lambda_{nu0}\rho_{nu}\left(1-{\rho_{nu}}^{l+1-c}\right)}{\lambda_{nu1}\left(1-\rho_{nu}\right)}\right],\ c+1 \le l \le N; \\ Z_{nu}\left\{\begin{array}{l} {\rho_{nu}}^{l}\gamma_{c-1} + \dfrac{\lambda_{nu0}{\rho_{nu}}^{l+1-N}\left(1-{\rho_{nu}}^{N+1-c}\right)}{\lambda_{nu1}\left(1-\rho_{nu}\right)} + \\ \dfrac{{\lambda_{nu0}}^{2}}{c\mu\left[\lambda_{nu0}-\rho_{nu}\left(\lambda_{nu0}+\theta\right)\right]}\left[\left(\dfrac{\lambda_{nu0}}{\lambda_{nu0}+\theta}\right)^{l-N} - {\rho_{nu}}^{l-N}\right] \end{array}\right\},\ l > N, \end{cases} \tag{47}$$

where

$$\begin{cases} \gamma_l = \dfrac{\lambda_{nu0}c^l}{\lambda_{nu1}l!}\displaystyle\sum_{i=0}^{l-1} i!\left(\dfrac{\mu}{\lambda_{nu1}}\right)^{i},\ 1 \le l \le c, \\ Z_{nu} = \left\{\begin{array}{l} N + \displaystyle\sum_{l=1}^{c}{\rho_{nu}}^{l}\gamma_l + \dfrac{\lambda_{nu1}{\rho_{nu}}^{c}}{c\mu-\lambda_{nu1}}\gamma_{c-1} + \dfrac{\lambda_{nu0}(N+\rho_{nu}-c)}{c\mu-\lambda_{nu1}} \\ + \dfrac{\lambda_{nu0}\left\{(1-\rho_{nu})\left\{c\mu\left[\lambda_{nu0}-\rho_{nu}\left(\lambda_{nu0}+\theta\right)\right]+\lambda_{nu0}^{2}\right\}-\lambda_{nu0}\theta\rho_{nu}\right\}}{\theta\left(c\mu-\lambda_{nu1}\right)\left[\lambda_{nu0}-\rho_{nu}\left(\lambda_{nu0}+\theta\right)\right]} \end{array}\right\}^{-1}. \end{cases} \tag{48}$$

Supposing the probability that the servers' status is $i(i=0,1)$ and the conditional probability that the queue length is $l(l \ge 0)$ when the servers' status is $i(i=0,1)$ in the nearly unobservable case are $p_{nu}(i)$ and $p_{nu}(l \mid i)$ respectively, we can get

$$p_{nu}(l \mid 0) = \frac{p_{nu}(l,0)}{p(0)} = \frac{p_{nu}(l,0)}{\sum_{l=0}^{\infty} p_{nu}(l,0)},\ l \ge 0, \tag{49}$$

$$p_{nu}(l \mid 1) = \frac{p_{nu}(l,1)}{p(1)} = \frac{p_{nu}(l,1)}{\sum_{l=1}^{\infty} p_{nu}(l,1)},\ l \ge 1. \tag{50}$$

in accordance with Theorem 4.1. Next, the mean sojourn time of the customer who joins the queue when the system is at state (l,i) (if $i=0$, $l \ge 0$ and if $i=1$, $l \ge 1$) in the nearly unobservable case, i.e., $T_{nu}(l,i)$, is obtained by replacing Λ in Eq.(1) with λ_{nu0} when $i=0$ and equals Eq.(2) when $i=1$. And then we can express the sojourn time of a customer joining the queue when the servers' status is $i(i=0,1)$ as

$$\begin{aligned} T_{nu0} &= \sum_{l=0}^{\infty} T_{nu}\left(l,0\right)p\left(l\,|0\right) \\ &= \frac{N\left[\theta\left(N-1\right)+2\lambda_{nu0}\right]}{2\lambda_{nu0}\left(\lambda_{nu0}+\theta N\right)} + \frac{\theta\left[c\left(c-1\right)+N\left(N+1\right)\right]}{2c\mu\left(\lambda_{nu0}+\theta N\right)} + \frac{\lambda_{nu0}\left(c\mu+\theta N+\lambda_{nu0}+\theta\right)}{c\mu\theta\left(\lambda_{nu0}+\theta N\right)}, \end{aligned} \tag{51}$$

$$
\begin{aligned}
T_{nu1} &= \sum_{l=1}^{\infty} T_{nu}\left(l,1\right) p\left(l\,|1\right) \\
&= \delta \left\{
\begin{array}{l}
\displaystyle\sum_{l=1}^{c-1} \frac{{\rho_{nu}}^{l}}{\mu}\gamma_l + \frac{\lambda_{nu1}{\rho_{nu}}^{c-1}\left[c\left(1-\rho_{nu}\right)+1\right]}{\left(c\mu-\lambda_{nu1}\right)^2}\gamma_{c-1} \\
\displaystyle + \frac{\lambda_{nu0}\left(N+1-c\right)\left(N+c+2\right)}{2c\mu\left(c\mu-\lambda_{nu1}\right)} + \frac{\lambda_{nu0}\lambda_{nu1}\left(1-\rho_{nu}\right)\left(N+1-c\right)}{\left(c\mu-\lambda_{nu1}\right)^3} \\
\displaystyle + \frac{{\lambda_{nu0}}^3\left[\theta\left(N+2\right)+\lambda_{nu0}\right]}{c^2\theta^2\mu^2\left[\lambda_{nu0}-\rho_{nu}\left(\lambda_{nu0}+\theta\right)\right]} + \frac{{\lambda_{nu0}}^2\rho_{nu}\left[\left(N+1\right)\left(\rho_{nu}-1\right)-1\right]}{\left[\lambda_{nu0}-\rho_{nu}\left(\lambda_{nu0}+\theta\right)\right]\left(c\mu-\lambda_{nu1}\right)^2}
\end{array}
\right\}, \qquad (52)
\end{aligned}
$$

where γ_l is expressed by Eq.(48) and

$$
\delta = \left\{ \sum_{l=1}^{c} {\rho_{nu}}^{l}\gamma_l + \frac{\lambda_{nu1}{\rho_{nu}}^{c}}{c\mu-\lambda_{nu1}}\gamma_{c-1} + \frac{{\lambda_{nu0}}^2\left[\lambda_{nu0}\left(1-\rho_{nu}\right)-\theta\rho_{nu}\right]}{\theta\left(c\mu-\lambda_{nu1}\right)\left[\lambda_{nu0}-\rho_{nu}\left(\lambda_{nu0}+\theta\right)\right]} + \frac{\lambda_{nu0}\left(N+\rho_{nu}-c\right)}{c\mu-\lambda_{nu1}} \right\}^{-1}. \qquad (53)
$$

Evidently, in the nearly unobservable case, that all customers balk is an equilibrium strategy when the servers are on vacation. And the positive equilibrium joining rates are considered and given in Theorem 4.2.

Theorem 4.2. *When $\rho_{nu} = \lambda_{nu1}/(c\mu) < 1$, the positive equilibrium joining rates of customers in the nearly unobservable $M/M/c$ queueing system with synchronous multiple uninterrupted vacations under N-policy, i.e., $(\lambda^e_{nu0}, \lambda^e_{nu1})$, are the solutions of*

$$
U_{nu0}\left(\lambda_{nu0}\right) = R_c - C_c T_{nu0} = 0 \qquad (54)
$$

and

$$
U_{nu1}\left(\lambda_{nu0}, \lambda_{nu1}\right) = R_c - C_c T_{nu1} = 0. \qquad (55)
$$

Next let us discuss the customers' socially optimal balking strategy in the nearly unobservable case. Based on Theorem 4.1, we can get the social welfare per unit time

$$
\begin{aligned}
S_{nu}(\lambda_{nu0}, \lambda_{nu1}) = R_c &\left[\lambda_{nu0}\sum_{l=0}^{\infty} p_{nu}(l,0) + \lambda_{nu1}\sum_{l=1}^{\infty} p_{nu}\left(l,1\right)\right] - C_c \sum_{l=1}^{\infty} l\left[p_{nu}\left(l,0\right)+p_{nu}\left(l,1\right)\right] \\
&- C_s c \sum_{l=1}^{\infty} p_{nu}(l,1) - C_t c \frac{\theta^2}{\lambda_{nu0}+\theta}\sum_{l=N}^{\infty} p_{nu}(l,0). \qquad (56)
\end{aligned}
$$

And then solving the optimal problem *max* $S_{nu}(\lambda_{nu0}, \lambda_{nu1})(\lambda_{nu1} < c\mu)$, the customers' socially optimal joining rates $(\lambda^*_{nu0}, \lambda^*_{nu1})$ and the optimal social welfare $S^*_{nu}(\lambda^*_{nu0}, \lambda^*_{nu1})$ can be all obtained.

4.2. The entirely unobservable case

In the entirely unobservable case where customers can get no information about the system, they will join the queue with a uniform probability, and the corresponding state transition diagram is portrayed in Figure 12. Just letting $\lambda_{nu0} = \lambda_{nu1} = \lambda_{eu}$ in Theorem 4.1, the stationary distributions of the entirely unobservable system, i.e., $p_{eu}(l,i)$, where $(l,i) \in \{(l,0)|l \geq 0\} \cup \{(l,1)|l \geq 1\}$, can be achieved. And then substituting λ_{eu} for Λ in Eq.(1) and referring to Eq.(1), a customer's mean sojourn time in the entirely unobservable case can be expressed by

$$\begin{aligned}
T_{eu} &= \sum_{l=0}^{c-1}\left(\frac{N-l-1}{\lambda_{eu}}+\frac{1}{\theta}+\frac{1}{\mu}\right)p_{eu}(l,0)+\sum_{l=c}^{N-1}\left(\frac{N-l-1}{\lambda_{eu}}+\frac{1}{\theta}+\frac{l-c+1}{c\mu}+\frac{1}{\mu}\right)p_{eu}(l,0)\\
&\quad+\sum_{l=N}^{\infty}\left(\frac{1}{\theta}+\frac{l-c+1}{c\mu}+\frac{1}{\mu}\right)p_{eu}(l,0)+\sum_{l=1}^{c-1}\frac{1}{\mu}p_{eu}(l,1)+\sum_{l=c}^{\infty}\left(\frac{l-c+1}{c\mu}+\frac{1}{\mu}\right)p_{eu}(l,1)\\
&= Z_{eu}\left\{\begin{aligned}
&\sum_{l=1}^{c-1}\frac{\xi_l}{\mu}+\frac{\lambda_{eu}[c(1-\rho_{eu})+1]}{(c\mu-\lambda_{eu})^2}\xi_{c-1}+\frac{c(c-1)+N(N+1)}{2c\mu}+\frac{N[\theta(N-1)+2\lambda_{eu}]}{2\lambda_{eu}\theta}\\
&+\frac{\rho_{eu}[\lambda_{eu}+c\mu+\theta(N+1)]}{\theta^2}+\frac{\rho_{eu}(N+1-c)(N+c+2)}{2(c\mu-\lambda_{eu})}+\frac{\lambda_{eu}\rho_{eu}[\lambda_{eu}+\theta(N+2)]}{\theta^2(c\mu-\lambda_{eu}-\theta)}\\
&+\frac{{\lambda_{eu}}^2(1-\rho_{eu})(N+1-c)}{(c\mu-\lambda_{eu})^3}+\frac{{\rho_{eu}}^2[(N+1)(\rho_{eu}-1)-1]}{(c\mu-\lambda_{eu}-\theta)(1-\rho_{eu})^2}
\end{aligned}\right\},
\end{aligned} \tag{57}$$

where

$$\begin{cases}
\xi_l = \dfrac{(c\rho_{eu})^l}{l!}\sum\limits_{i=0}^{l-1} i!\left(\dfrac{\mu}{\lambda_{eu}}\right)^i,\ 1 \leq l \leq c,\\
Z_{eu} = \left\{\sum\limits_{l=1}^{c}\xi_l+\dfrac{{\rho_{eu}}^2}{1-\rho_{eu}}\xi_{c-1}+\dfrac{\rho_{eu}(\rho_{eu}-c)+N}{1-\rho_{eu}}+\dfrac{\lambda_{eu}[c\mu(1-\rho_{eu})-\theta]}{\theta(1-\rho_{eu})(c\mu-\lambda_{eu}-\theta)}\right\}^{-1}.
\end{cases} \tag{58}$$

Obviously, that all customers balk is also an equilibrium strategy in the entirely case. And the positive equilibrium joining rate is given in Theorem 4.3.

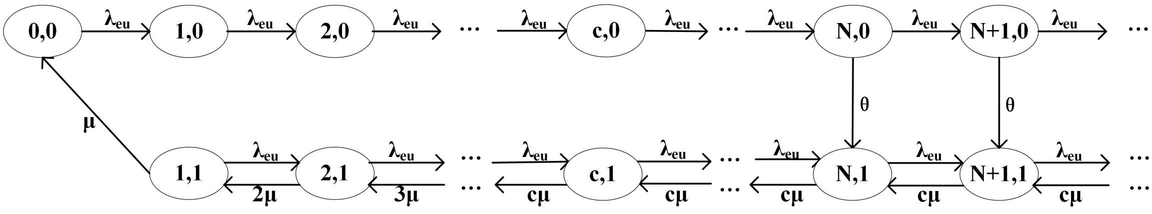

Figure 12. State transition diagram in the entirely unobservable case.

Theorem 4.3. *When $\rho_{eu} = \lambda_{eu}/(c\mu) < 1$, the positive equilibrium joining rate of customers in the entirely unobservable $M/M/c$ queueing system with synchronous multiple uninterrupted vacations under N-policy, i.e., λ_{eu}^{e}, is the solution of*

$$U_{eu}(\lambda_{eu}) = R_c - C_c T_{eu} = 0. \tag{59}$$

Afterwards, the social welfare per unit time in the entirely unobservable case is expressed by

$$S_{eu}(\lambda_{eu}) = R_c\lambda_{eu} - C_c\sum_{l=1}^{\infty} l\left[p_{eu}(l,0)+p_{eu}(l,1)\right] - C_s c\sum_{l=1}^{\infty} p_{eu}(l,1)$$
$$-C_t c\frac{\theta^2}{\lambda_{eu}+\theta}\sum_{l=N}^{\infty} p_{eu}(l,0). \tag{60}$$

And then we can gain the socially optimal joining rate λ_{eu}^* and the optimal social welfare $S_{eu}^*(\lambda_{eu}^*)$ by solving the optimization problem $max\ S_{eu}(\lambda_{eu})(\lambda_{eu} < c\mu)$.

4.3. Numerical analysis and comparisons

Similar with Subsection 3.3, based on the nearly unobservable and the entirely unobservable cases, we continue to numerically reveal the relationship between the customers' equilibrium and socially optimal balking behaviors, explore how the threshold N as well as the amount of information affects the equilibrium and socially optimal joining rates, the pricing strategies, the system manager's revenue and the optimal social welfare.

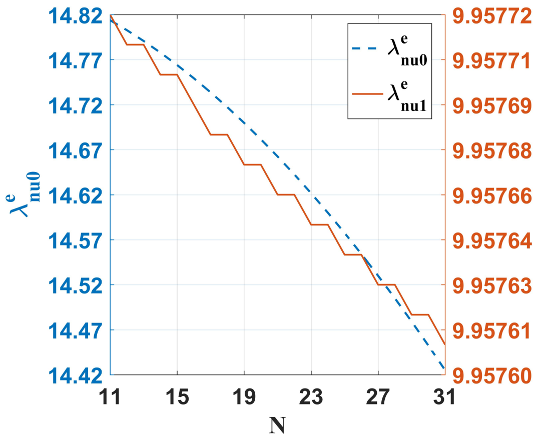

Figure 13. The customers' equilibrium joining rates in the nearly unobservable case.

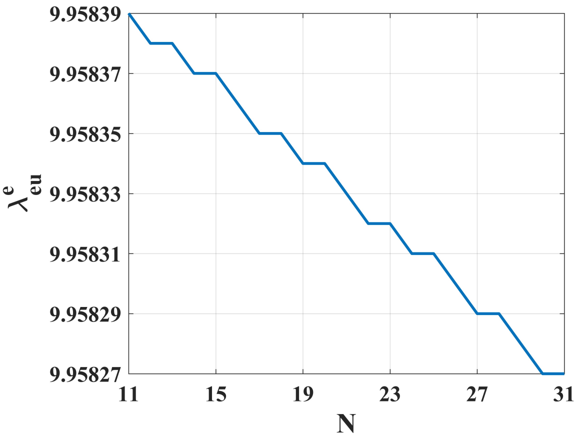

Figure 14. The customers' equilibrium joining rate in the entirely unobservable case.

Figure 13 and Figure 14[2] show that λ_{nu0}^e, λ_{nu1}^e and λ_{eu}^e all decrease with respect to N, since the greater N can result in a longer sojourn time for customers. By contrast, λ_{nu0}^e is most sensitive to N, while λ_{nu1}^e and λ_{eu}^e are less affected by N. What's more, we can find that $\lambda_{nu0}^e > \lambda_{eu}^e > \lambda_{nu1}^e$ always holds, which indicates that the customers' equilibrium balking behavior in the entirely unobservable case can be regarded as the comprehensive reflection of that in the nearly unobservable case when the servers are in the two statuses ($i = 0$ or 1). The relationship $\lambda_{nu0}^e > \lambda_{nu1}^e$ may go against the thought that the customers may be more

[2] The values of the legend parameters in Figure 13 - Figure 19 are $\mu = 1, c = 10, \theta = 0.1, R_c = 25, C_c = 1, C_s = 20, C_t = 15$.

willing to join the queue when the servers are working rather than on vacation, while that can be explained that c customers will be served instantly once the servers are turned on in the dormant queue, and the customer may feel that the dormant queue length gets shortened faster than that in the working queue so that they want to join the dormant queue more.

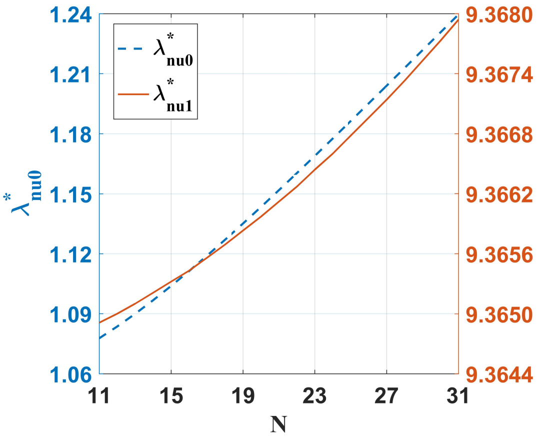

Figure 15. The customers' socially optimal joining rates in the nearly unobservable case.

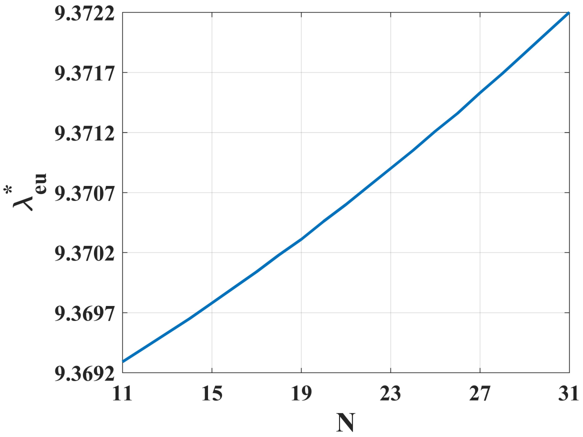

Figure 16. The customers' socially optimal joining rate in the entirely unobservable case.

Accordingly, Figure 15 and Figure 16 show the customers' socially optimal joining rates in the unobservable cases. Apparently, more and more customers are needed to join the queue with the increase of N in the unobservable cases for social optimization. Comparing Figure 15 and Figure 16, it can be found that $\lambda^*_{eu} > \lambda^*_{nu1} > \lambda^*_{nu0}$, which manifests that ensuring the optimal social welfare always needs more customers joining the queue in the entirely unobservable compared with the nearly unobservable case. Comparing Figure 13 and Figure 15, Figure 14 and Figure 16 separately, we can see that $\lambda^e_{nu0} > \lambda^*_{nu0}$, $\lambda^e_{nu1} > \lambda^*_{nu1}$ and $\lambda^e_{eu} > \lambda^*_{eu}$, that is, the pricing strategy is also needed in the unobservable cases to adjust the inconsistence between the customers' equilibrium and socially optimal behaviors. The prices in the nearly unobservable and entirely unobservable cases can be respectively expressed by

$$P_{nu0} = U_{nu0}(\lambda^*_{nu0}) - U_{nu0}(\lambda^e_{nu0}), \tag{61}$$

$$P_{nu1} = U_{nu1}(\lambda^*_{nu0}, \lambda^*_{nu1}) - U_{nu1}(\lambda^e_{nu0}, \lambda^e_{nu1}) \tag{62}$$

and

$$P_{eu} = U_{eu}(\lambda^*_{eu}) - U_{eu}(\lambda^e_{eu}). \tag{63}$$

From Figure 17 and Figure 18, we can observe that P_{nu0}, P_{nu1} and P_{eu} all decrease with the increase of N since the gap between the equilibrium and the socially optimal joining rates are smaller and smaller with the increase of N. In addition, it is also found that P_{nu0} is relatively more sensitive about N and $P_{nu1} > P_{eu} > P_{nu0}$ is attainable by comparing Figure 17 and Figure 18.

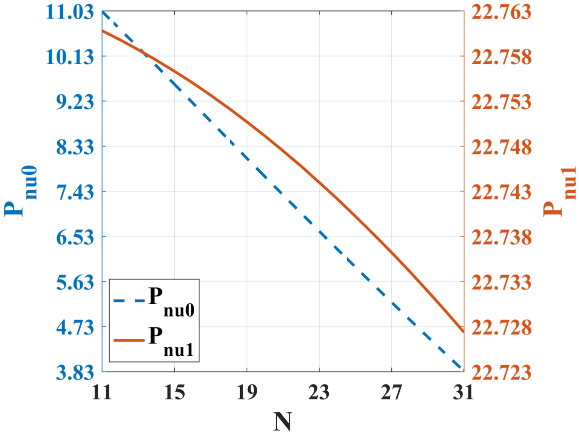

Figure 17. The pricing strategy in the nearly unobservable case.

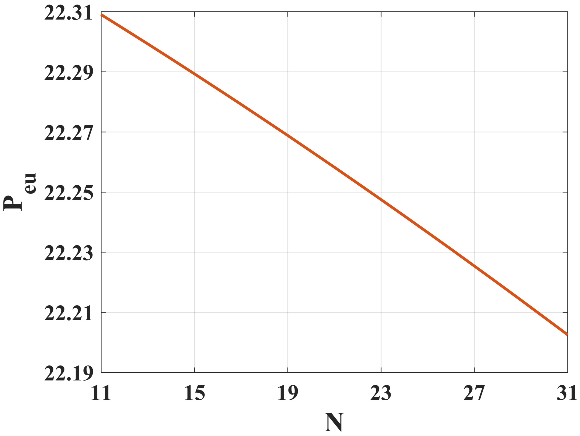

Figure 18. The pricing strategy in the entirely unobservable case.

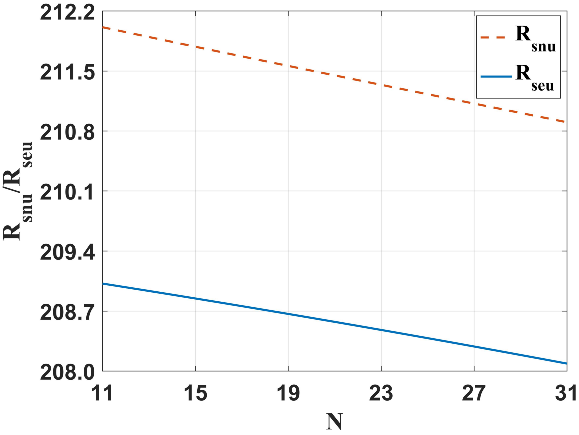

Figure 19. The system manager's revenue in the nearly/entirely unobservable case.

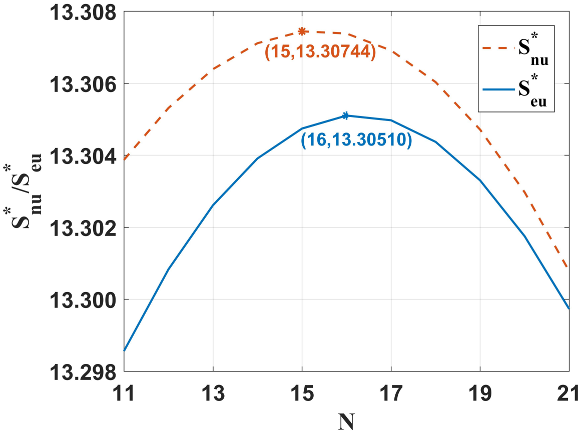

Figure 20. The optimal social welfare in the nearly/entirely unobservable case.

Subsequently, the system manager's revenue in the unobservable cases can be given as

$$R_{snu} = P_{nu0}\lambda^*_{nu0}\sum_{l=0}^{\infty} p_{nu}(l,0) + P_{nu1}\lambda^*_{nu1}\sum_{l=1}^{\infty} p_{nu}(l,1) \tag{64}$$

and

$$R_{seu} = P_{eu}\lambda^*_{eu} \tag{65}$$

according to the corresponding pricing strategy and the stationary distributions. From Figure 19, we can find that either R_{snu} or R_{seu} decreases with the increase of N because of the decrease of P_{nu0}, P_{nu1} and P_{eu} with respect to N. What's more, $R_{snu} > R_{seu}$ always holds.

The optimal social welfare in the unobservable cases is also explored, which is presented in Figure 20[3]. Obviously, just like the observable cases, there also exist different Ns

[3] The values of the legend parameters in Figure 20 are $\mu = 1, c = 10, \theta = 3, R_c = 25, C_c = 1, C_s = 20, C_t = 15$.

maximizing the corresponding optimal social welfare. Moreover, that $S^*_{nu} > S^*_{eu}$ is always available indicates that when the queue length is unobservable, disclosing the information about the servers' status can improve the optimal social welfare.

5. Conclusion and Management Inspirations

This paper mainly studies both equilibrium and socially optimal balking behaviors of customers in an $M/M/c$ queue with synchronous multiple uninterrupted vacations under N-policy in the entirely/nearly observable and nearly/entirely unobservable cases. What's more, the pricing strategies ensuring the optimal social welfare are made, and the system manager's revenue is analyzed. Based on the analysis, we can get the following two management inspirations for the social planner:

(1) The pricing strategy is suggested to be formulated according to the specific information level from the view of social optimization. When the queue length is observable, the price can be adjusted within a certain interval, otherwise it should be specific to a fixed level.

(2) No matter what information level the system is at, the best N-policy that maximizes the optimal social welfare should be evaluated and adopted. And when the threshold N is predetermined objectively or subjectively, the information about the servers' status should be disclosed to customers to improve the optimal social welfare regardless of the queue length is observable or not.

Funding

This work was supported by the National Natural Science Foundation of China (No. 71971188), the Humanity and Social Science Foundation of Ministry of Education of China (Nos. 24YJAZH135, 22YJCZH086), Hebei Natural Science Foundation (No. G2023203008).

Competing interests declaration

The authors declare that they have no competing interests related to this study.

References

[1] Begum, A., & Choudhury, G. (2022). Analysis of a bulk arrival N-policy queue with two-service genre, breakdown, delayed repair under Bernoulli vacation and repeated service policy. *RAIRO-Operations research*, 56(2), 979–1012.

[2] Guo, P., & Hassin, R. (2011). Strategic behavior and social optimization in Markovian vacation queues. *Operations Research*, 59(4), 986–997.

[3] Guo, P., & Hassin, R. (2012). Strategic behavior and social optimization in Markovian vacation queues: The case of heterogeneous customers. *European Journal of Operational Research*, 222(2), 278–286.

[4] Guo, P., & Li, Q. (2013). Strategic behavior and social optimization in partially-observable Markovian vacation queues. *Operations Research Letters*, 41(3), 277–284.

[5] Hao, Y., Wang, J., Wang, Z., & Yang, M. (2019). Equilibrium joining strategies in the $M/M/1$ queues with setup times under N-policy. *Journal of Systems Science and Systems Engineering*, 28(2), 141–153.

[6] Huang, H., Wang, T., & Ke, J. (2016). Random Policy for an Unreliable Server System with Delaying Repair and Setup Time Under Bernoulli Vacation Schedule. *Journal of Testing and Evaluation*, 44(3), 1400–1408.

[7] Kempa, W. M., & Kurzyk, D.(2022). Non-stationary departure process in a batch-arrival queue with finite buffer capacity and threshold-type control mechanism. *Kybernetika*, 58(1), 82–100.

[8] Kim, S. J., Kim, N. K., Park, H., Chae, K. C., & Lim, D. (2013). On the discrete-time $Geo^X/G/1$ queues under N-policy with single and multiple vacations. *Journal of Applied Mathematics*, 2013(2013), 1–7.

[9] Lan, S., & Tang, Y. (2019). An N-policy discrete-time $Geo/G/1$ queue with modified multiple server vacations and bernoulli feedback. *RAIRO-Operations Research*, 53(2), 367–387.

[10] Lim, D., Lee, D. H., Yang, W. S., & Chae, K. (2013). Analysis of the $GI/Geo/1$ queue with N-policy. *Applied Mathematical Modelling*, 37(7), 4643–4652.

[11] Meena, R. K., Jain, M., Assad,A., Sethi, R., & Garg, D. (2022). Performance and cost comparative analysis for $M/G/1$ repairable machining system with N-policy vacation. *Mathematics and Computers in Simulation*, 200(2022), 315–328.

[12] Shen, L., Jin, S., & Tian, N. (2004). The $M/M/c$ queue with N-policy and multiple vacations of partial servers. *Chinese Journal of Engineering Mathematics*, 21(2), 238–244.

[13] Shree, L., Singh, P., Sharma, D. C., & Jharotia, P. (2015). Mathematical modeling and performance analysis of machine repairable system with hot spares. *Proceedings of The National Academy of Sciences India Section A-physical Sciences*, 85(1), 127-135.

[14] Sun, W., Li, S., & E, C. (2016). Equilibrium and optimal balking strategies of customers in Markovian queues with multiple vacations and N-policy. *Applied Mathematical Modelling*, 40(1), 284–301.

[15] Tian, R., Yue, D., & Yue, W. (2015). Optimal balking strategies in an $M/G/1$ queueing system with a removable server under N-policy . *Journal of Industrial and Management Optimization*, 11(3), 715–731.

[16] Wang, K. H., Wang, T. Y., & Pearn, W. L. (2007). Optimal control of the N policy $M/G/1$ queueing system with server breakdowns and general startup times. *Applied Mathematical Modelling*, 31(10), 2199–2212.

[17] Wang, Z., Liu, L., Shao, Y., & Zhao, Y. (2021). Joining strategies under two kinds of games for a multiple vacations retrial queue with N-policy and breakdowns. *Aims Mathematics*, 6(8), 9075–9099.

[18] Wang, Z., Liu, L., & Zhao, Y. (2022). Equilibrium customer and socially optimal balk-

ing strategies in a constant retrial queue with multiple vacations and N-policy. *Journal of Combinatorial Optimization*, 43(4), 870–908.

[19] Wang, Z., Liu, L., Zhao, Y., Li, L., & Xu, W. (2023). Joining strategies of noncooperative and cooperative in a single server retrial queue with N-policy and multiple server vacations. *Communications in Statistics-theory and Methods*, 52(4), 1076–1100.

[20] Xia, M., & Tian, N. (1997). The $M/M/c$ queue with synchronous N-policy and multiple vacations. *Operations Research Transactions*, 1(2), 86–94.

[21] Yue, D., & Sun, Y. (2008). Performance analysis of an N-policy $M/M/R/K$ queuing system with balking, reneging and multiple synchronous vacations. *Systems Engineering-Theory* & *Practice*, 2008(11), 94-102+108.

[22] Zhang, Z., & Tian, N. (2004). The N threshold policy for the $GI/M/1$ queue. *Operations Research Letters*, 32(1), 77-84.

[23] Zhou, M., Liu, L., Chai, X., & Wang, Z. (2020). Equilibrium strategies in a constant retrial queue with setup time and the N-policy. *Communications in Statistics-Theory and Methods*, 49(7), 1695-1711.