Strategic Behavior and Social Optimization in Markovian Vacation Queues: The Case of Heterogeneous Customers

Pengfei Guo*
Department of Logistics and Maritime Studies
Hong Kong Polytechnic University
Hung Hom, Hong Kong
Email: lgtpguo@polyu.edu.hk

Refael Hassin
Department of Statistics and Operations Research
Tel Aviv University
Tel Aviv 69978, Israel
Email: hassin@post.tau.ac.il

May 3, 2012

Abstract

We consider a single server queueing system in which service shuts down when no customers are present, and is resumed when the queue length reaches a given critical length. We assume customers are heterogeneous on delay sensitivity and analyze customers' strategic response to this mechanism and compare it to the overall optimal behavior. We provide algorithms to compute the equilibrium arrival rates and also derive the monotonicity of equilibrium and optimal arrival rates. We show that there may exist multiple equilibria in such a system and the optimal arrival rate may be larger or smaller than the decentralized equilibrium one.

Keywords: **queueing, strategic customers, balking, vacation queue, equilibrium analysis**

*Corresponding author. Tel: +852 34003623.

1. Introduction

I don't like the opening paragraph and the citations in it because they are not related to our paper. We don't say why polling is relevant. More important, as we have discussed with respect to our other paper, the association of ATC with negative externalities and FTC with positive ones is wrong. For example in Naor's model there are negative externalities but it is neither ATC not FTC because the best response is independent of what others do. Another example is the unobservable queue where for a cost one can learn the queue length. Alternatively it is possible to join without doing so (this is in my paper with Roet-Green that we cite in our duplications paper). Here we have positive externalities and also ATC

In many queueing systems, a customer joining behavior usually increases other customers' delay, imposing an effect called *negative externality* on the system. In such systems, it is common for customers to exhibit a behavior of *avoid the crowd* (ATC) when they decide to join or balk a queue. However, in another type of systems where the service rate is increasing with congestion, customers may find that more congestion could be beneficial for them. For example, a shuttle may leave only after all seats are occupied. In grocery stores and banks, a long queue can stimulate more service counters to be opened and thus the queue can move progressively faster. More examples can be found in surveys on polling systems by Boxma (1989), Levy and Sidi (1990), Takagi (1986, 2000), Tian and Zhang (2006) and Yechiali (1993). In those systems, customers' joining behavior can bring positive impacts on other customers, exhibiting an effect called *positive externality.* In such a system, *follow the crowd* (FTC) behavior may appear (see Hassin and Haviv, 2003).

Pioneered by Naor (1969), studies on customers' decentralized joining/balking decision to a queue and social optimal requirements have been carried out by many researchers; see the survey book by Hassin and Haviv (2003). In recent years, there exists an increasing trend on studying customers' strategic behavior in queues with vacations/breakdowns, in which positive externalities and FTC behavior are observed. Burnetas and Economou (2007) study a system with an exponential setup time when the server starts a new busy period. They consider customers' strategic behavior under different levels of information which may include the queue length and/or the state of the server (during setup or busy). In particular, if only the queue length is known and the setup time is considerably long, customers' FTC behavior is observed. Economou et al. (2011) extend the model to general distributions of service and vacation lengths. Economou and Kanta (2008) consider a system with breakdowns and repairs which take exponential times. If an arriving customer can observe both the queue length and the state of the server, there exists a dominant pure threshold strategy for customers. In the almost observable case where the arrival observes the queue length but not the state of the server, the authors show that it is an FTC situation and there could exist multiple equilibria. Sun et al. (2010) consider the strategic customer behavior in an M/M/1 system with closedown and setup. Guo and Hassin (2011) study the decentralized equilibrium and social optimization in a vacation queue with N-policy and exhaustive service. The server starts working when the queue reaches size N and once the server starts working, it finishes all the work in the system before taking the next 'vacation'. In that work, customers are identical. Here, we extend the analysis to heterogeneous customers. The methods used in this work are very different from those in Guo and Hassin (2011). **I think that we need also to explain that not only the methods are different but also the results. In some place we also need to compare the results of the two models, probably in the concluding section.** Recently, Dimitrakopoulos and Burnetas (2011) study an unobservable M/M/1 queue with the service rate switching between a low and high value. They show that at most three equilibria exist. Guo and Zhang (2012) study the strategic queueing behavior in a multi-server queue with some servers to be turned on or off according to the queue length. They observe multiple equilibria there.

Besides the work on vacation queues, there exists research on other systems with positive externalities associated with customer arrivals. Veeraraghavan and Debo (2008) consider the situations where queue length indicates not only congestion but also service quality. In such a situation, customers may prefer to join longer queues. Johari and Kumar (2008) study a type of network service such as an on-line gaming system where users form a club, in which both negative and positive externalities could exist.

We study both unobservable and observable queues and consider two situations regarding customers' delay sensitivity: two-type and continuously-distributed. In total, we consider four cases. We obtain equilibrium and optimal arrival rates for each case and also obtain some analytical results on the monotonicity of equilibrium and optimal arrival rates in the case with observable queues. We show that when equilibrium arrival rates are increasing with the queue length in the case of an idle server, the optimal ones are increasing too.

The paper is organized as follows: Section 2 introduces the model and assumptions. Sections 3-4 study the decentralized equilibrium and socially optimal solution with unobservable and observable queues, respectively. Section 5 provides concluding remarks. Proofs are relegated to the appendix.

2. Formulation and Preliminaries

We assume that potential customers arrive according to a Poisson process with rate Λ. There is a single server, and the service times are independent and exponentially distributed with rate μ. The server employs an N-policy: It shuts down when the system becomes empty of customers and resumes service after N arrivals.

Assume that a customer's utility consists of a *reward* for receiving service minus a *waiting cost.* This waiting cost is linear and depends on a customer-specific parameter and the expected *waiting time*. Here, the waiting time means the total sojourn time in the system. We also consider an additive social utility composed of the sum of individual utilities of all served customers.

Specifically, define

W = expected waiting time for a customer

θ = customer-type delay-sensitivity parameter, indicating his cost per time unit spent in the system.

H = cumulative distribution function of θ in the population of potential customers.

R = reward to the customer for receiving service, $R \geq 0$.

U = the utility for a customer who joins the system, equal to R minus by his waiting time multiplied by his θ value, i.e.,

$$U = R - \theta W.$$

$\nu = \frac{R\mu}{\theta}$ the upper bound on the number of service epochs that a customer is ready to wait. **we don't use the notation of ν and U so it's best to delete both of them**

The service reward R is the same for all customers. Customers differ on the delay sensitivity parameter θ. We consider both discrete and continuous distributions of θ. For the discrete distribution of θ, we consider two types of customers.

3. Unobservable Vacation Queues with an N-policy

In this section we assume that customers have no information on the server's current status and the system occupancy. For an M/M/1 queue with an N-policy and arrival rate λ ($\lambda < \mu$), we can express the waiting time explicitly (see Yadin and Naor, 1963)

$$W(\lambda) = \frac{1}{\mu - \lambda} + \frac{N-1}{2\lambda}, \tag{1}$$

where $\frac{1}{\mu-\lambda}$ is the expected waiting time in a standard M/M/1 queue and $\frac{N-1}{2\lambda}$ is the expected extra waiting time for the server to begin to work.

The function $W(\lambda)$ is strictly convex in λ, with minimum value

$$W(\tilde{\lambda}) = \frac{1}{\mu}\left(1+\sqrt{\frac{N-1}{2}}\right)^2 \quad \text{at} \quad \tilde{\lambda} = \frac{\mu\sqrt{\frac{N-1}{2}}}{1+\sqrt{\frac{N-1}{2}}}. \tag{2}$$

3.1 Two customer types

Now assume customers are distributed on two delay-sensitivity parameters θ_1 and θ_2 ($\theta_1 < \theta_2$). Assume the arrival rate for θ_j-customers is Λ_j, $j = 1, 2$. Both types of customers can use pure or mixed strategies. Denote the joining rate for θ_j-customers by λ_j. Since $\theta_1 < \theta_2$, it must be true that $\lambda_1 = \Lambda_1$ if $\lambda_2 > 0$ (if some of θ_2-customers join, θ_1-customers must all join.) We assume that $R > \theta_2 W(\tilde{\lambda})$ as, otherwise, it is impossible for a θ_2-customer to join.

3.1.1 Equilibrium

Denote by $\lambda \in [0, \Lambda_1 + \Lambda_2]$ the total arrival rate. Then, $\lambda_1 = \min\{\Lambda_1, \lambda\}$ and $\lambda_2 = \lambda - \Lambda_1$.

The following solutions define equilibria:

1. $\lambda = 0$.

2. Every $\lambda \in (0, \Lambda_1)$ such that $R = \theta_1 W(\lambda)$ (there are at most two such solutions).

3. Λ_1 if $\theta_1 W(\Lambda_1) \le R \le \theta_2 W(\Lambda_1)$.

4. Every $\lambda \in (\Lambda_1, \Lambda_1 + \Lambda_2)$ such that $R = \theta_2 W(\lambda)$ (there are at most two such solutions).

5. $\Lambda_1 + \Lambda_2$ if $R \ge \theta_2 W(\Lambda_1 + \Lambda_2)$.

The above list allows for seven solutions, but some combinations are not possible, and there may be at most five equilibria for any set of input parameters. For example, in Figure 1 we observe five equilibrium solutions. Note that the left-most positive **it seems redundant to mention left-most positive, because there is just one equilibrium in $(0, \Lambda_1)$?** equilibrium in $(0, \Lambda_1)$ is not stable as a slight increase of arrival rate will reduce the expected waiting time, which, in turn, attracts more θ_1-customers, diverting to the equilibrium Λ_1. Similarly, the equilibrium in $(\Lambda_1, \tilde{\lambda})$ is unstable as an increase in λ will reduce the expected waiting time, attracting more θ_2-customers.

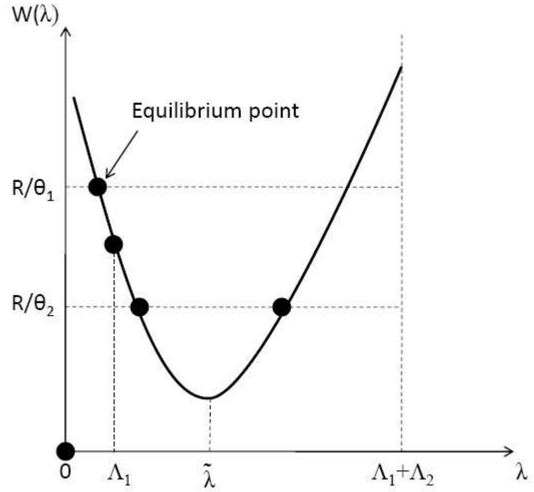

Figure 1: Equilibria with two types of customers

3.1.2 Social optimization

The social welfare function has different expressions on different ranges.

On $[0, \Lambda_1]$, the social welfare function is

$$SW_1(\lambda) = \lambda[R - \theta_1 W(\lambda)]. \tag{3}$$

On $[\Lambda_1, \Lambda_1 + \Lambda_2]$, the social welfare function is

$$SW_2(\lambda) = \lambda R - [\Lambda_1\theta_1 + (\lambda - \Lambda_1)\theta_2]W(\lambda). \tag{4}$$

We first consider **when I see 'first consider' I expect to have an alternative option after it but this seems not to be the case here** the property of social welfare function SW on the range $[0, \mu)$. On $[0, \Lambda_1]$, $SW = SW_1$ and on $[\Lambda_1, \mu), SW = SW_2$. We have the following conclusion on the property of the social welfare function.

Lemma 3.1 *Both SW_1 and SW_2 are strictly increasing on the domain where they are positive within $[0, \tilde{\lambda}]$; SW is strictly concave on $[\tilde{\lambda}, \mu)$.*

From Lemma 3.1, we conclude that the socially optimal arrival rate, denoted as $\bar{\lambda}$, must be unique and at least as large as $\tilde{\lambda}$, i.e., $\bar{\lambda} \geq \tilde{\lambda}$. **I think that this conclusion is important and justifies to be stated as a corollary. Also, we can say something about the fact that it doesn't give us a clear answer to whether $\bar{\lambda}$ is greater or smaller than $\tilde{\lambda}$ because there are equilibrium solutions of both types** This conclusion is intuitive as increasing arrival rate reduces the expected waiting time in the range of $[0, \tilde{\lambda}]$.

Consider two systems with the same total potential arrival rate, but different composition of θ_1- and θ_2-customers. Denote the parameters for the two systems by the superscripts 1 and 2, respectively. Assume that $\Lambda_1^1 + \Lambda_2^1 = \Lambda_1^2 + \Lambda_2^2$ and $\Lambda_1^1 < \Lambda_1^2$. We have the following proposition about $\bar{\lambda}$.

Proposition 3.2 $\bar{\lambda}^1 \leq \bar{\lambda}^2$.

From Proposition 3.2, the optimal arrival rate for the system with a larger fraction of sensitive customers should not be larger than the one with a smaller fraction of sensitive customers. **I suggest rephrasing: From Proposition 3.2, as more customers become highly delay-sensitive, the optimal arrival rate to the system decreases**

3.2 Continuous distribution

Suppose that the customers' delay-sensitivity parameter is continuously distributed. Clearly, $\lambda_e = 0$ **this is the first time that we use this notation and without defining it** is always an equilibrium, and sometimes $\lambda_e = \Lambda$ is also an equilibrium. Otherwise, in equilibrium, there exists a marginal customer with time value θ_e who is indifferent between joining and balking. The customers who stay are precisely those with $\theta \leq \theta_e$. Consequently, the fraction of customers who stay is $H(\theta_e)$ and the effective arrival process is Poisson with rate $\lambda(\theta_e) = \Lambda H(\theta_e)$. By the indifference of the marginal customer,

$$R - \theta_e W(\lambda(\theta_e)) = 0. \tag{5}$$

The social planner's decision problem is to set a threshold value θ^* that maximizes the social utility $SW(\theta)$, where

$$SW(\theta) = \Lambda \int_0^{\theta} [R - xW(\lambda(\theta))] dH(x). \tag{6}$$

Solving (5) yields the threshold delay-sensitivity parameter for the marginal customer. $W(\lambda)$ is convex in λ, first decreasing and then increasing. That is, customers' waiting is higher when the arrival rate is either very small or very large. Therefore the solution to (5) is not necessarily unique.

In fact the number of equilibrium solutions is not bounded or even countable. Suppose for example that for every $\lambda \leq \tilde{\lambda}$, the arrival rate of customers with delay sensitivity parameter $\theta \leq \frac{R}{W(\lambda)}$ is exactly λ [1]. Note that in this range $W(\lambda)$ is monotone decreasing and hence $\lambda(\theta)$ defined this way is monotone increasing, as required. In this case, every value of θ in this range

[1] **I think that we need to give the expression for H and not for λ here** $\lambda(\theta) = b - \sqrt{b^2 - \mu(N-1)\theta/2R}$, where $b = \frac{(N-3)\theta + 2R\mu}{4R}$.

satisfies (5). Therefore, in the rest of this section we will solve the model for a specific distribution, namely the uniform distribution on $[0,1]$.

3.2.1 Uniform distribution

Equilibrium

When H is the uniform distribution on $[0,1]$, the cost for the marginal customer with θ_e becomes

$$\theta_e W(\theta_e \Lambda) = \frac{\theta_e}{\mu - \theta_e \Lambda} + \frac{N-1}{2\Lambda}.$$

By (5), an equilibrium threshold with $0 < \lambda_e < \Lambda$ satisfies

$$R = \frac{\theta_e}{\mu - \theta_e \Lambda} + \frac{N-1}{2\Lambda}. \tag{7}$$

The right-hand side of (7) is monotone increasing in θ_e. The value of this expression is $\frac{N-1}{2\Lambda}$ at $\theta_e = 0$ and $\frac{1}{\mu-\Lambda} + \frac{N-1}{2\Lambda}$ at $\theta_e = 1$. Hence, only when $R > \frac{N-1}{2\Lambda}$, there exists a (unique) positive solution for the equilibrium threshold. The condition guarantees that if all others join when the server is idle then a customer arriving to an empty system would profit by joining.

The equilibrium solutions are summarized below:

Proposition 3.3

- *If $R \leq \frac{N-1}{2\Lambda}$, then $\lambda_e = \theta_e = 0$ is the unique equilibrium;*
- *if $\frac{N-1}{2\Lambda} < R < \frac{1}{\mu-\Lambda} + \frac{N-1}{2\Lambda}$, there exists a unique positive equilibrium $\lambda_e = \theta_e \Lambda$, where*

$$\theta_e = \frac{\mu}{\Lambda} \cdot \frac{1 + 2R\Lambda - N}{3 + 2R\Lambda - N}; \tag{8}$$

- *if $R > \frac{1}{\mu-\Lambda} + \frac{N-1}{2\Lambda}$ then $\theta_e = 1$ is an equilibrium;*
- *$\theta_e = 0$ is always an equilibrium.*

Social optimization

We now consider social welfare maximization. The problem is $\max_\theta\{SW(\theta)\}$ where by (6),

$$\begin{aligned} SW(\theta) &= \Lambda \int_0^\theta \left[R - x \left(\frac{1}{\mu - \theta\Lambda} + \frac{N-1}{2\theta\Lambda} \right) \right] dx \\ &= \Lambda R \theta - \frac{\Lambda \theta^2}{2} \left[\frac{1}{\mu - \theta\Lambda} + \frac{N-1}{2\theta\Lambda} \right] = \Lambda \left(R - \frac{N-1}{4\Lambda} \right) \theta - \frac{\Lambda\theta^2}{2(\mu - \theta\Lambda)}. \end{aligned}$$

$SW''(\theta) = -\frac{\Lambda\mu^2}{(\mu-\theta\Lambda)^3} < 0$, hence SW is concave. (Note that $\mu > \theta\Lambda$ is required to guarantee finite expected waiting time.) The first-order condition is

$$SW'(\theta) = \frac{\Lambda}{2}\left(2R - \frac{N-1}{2\Lambda} - \theta\frac{2\mu - \theta\Lambda}{(\mu-\theta\Lambda)^2}\right) = 0.$$

This equation has a unique root in $[0,1]$, at

$$\theta^* = \frac{\mu}{\Lambda}\left(1 - \frac{1}{\sqrt{2\Lambda R - \frac{N-3}{2}}}\right). \tag{9}$$

The condition for $\theta^* > 0$ is

$$R > \frac{N-1}{4\Lambda}.$$

Note that this condition is weaker than the condition for existence of a positive equilibrium arrival rate, namely $R > \frac{N-1}{2\Lambda}$ (see Proposition 3.3). Thus, it might be that under the optimal policy customers should join the queue (when it is short and the server is idle) though their expected utility while doing so is negative.

We have the following proposition that compares the decentralized and optimal thresholds.

Proposition 3.4 *If $\frac{\Lambda}{\mu-\Lambda} \leq \sqrt{\frac{N-1}{2}}$, then $\theta_e \leq \theta^*$. Otherwise, $\theta_e \leq \theta^*$ if and only if $R \leq \frac{1}{\Lambda}\left(\frac{N-1}{2} + \sqrt{\frac{N-1}{2}}\right)$.*

This conclusion shows that when the total arrival rate Λ is small or the threshold N is large, the decentralized arrival rate is smaller than the socially desired. Therefore, a social planner shall adopt a subsidy to encourage arrivals. This is in sharp contrast with the conclusion in regular queues where a tax is usually used to reduce arrival rate to the socially optimal level.

4. Observable Vacation Queues with an N-policy

In this section we assume that customers possess the information on the queue length and the server's status when they make their decision of whether to join or balk.[2] The set of the states is

$$\{0, 1^-, \ldots, (N-1)^-, 1^+, 2^+, \ldots, (N-1)^+, N, N+1, \ldots\},$$

where m^- means that the system occupancy is m and the server is idle, and m^+ means that the system occupancy is m and the server is busy.

[2] To simplify the presentation we assume that a customer who is indifferent between joining and balking chooses to join.

In general, the strategy of never joining is always an equilibrium when $N > 1$, since if this policy is adopted by the others, the expected wait for a customer who joins at state 0 is infinite, and balking at this state is a best response. In this case the server is never active.[3] We concentrate on the existence of other equilibrium strategies in which the server is busy for at least a positive fraction of the time. We refer to such a solution as a solution with *active server.*

A *threshold strategy* with threshold n is a strategy where customers join if and only if they find at most $n-1$ customers in the system upon arrival. Thus the maximum number of customers in the system at any time is n. We will see that indeed, the equilibrium solutions are threshold strategies. However, the optimal strategy may have a more general structure, though it also involves a threshold.

4.1 Two customer types

Assume customers are distributed on two delay-sensitivity parameters θ_1 and θ_2 ($\theta_1 < \theta_2$). Assume the arrival rate for θ_j-customers is Λ_j, $j = 1, 2$.

4.1.1 Equilibrium

When the server is busy and customers observe the queue length m, both θ_1-customers and θ_2-customers use a threshold strategy, joining only if the number of customers does not exceed $\frac{R\mu}{\theta_1}$ or $\frac{R\mu}{\theta_2}$, respectively. Similarly, while joining at state $(N-1)^-$, the utility, and hence strategy, is identical to state $(N-1)^+$. The more difficult case is their strategy when the server is on vacation and the system's occupancy is at most $N-2$. We assume first that θ_1-customers always join when the server is idle; otherwise, the server would never be active. We compute the resulting behavior of θ_2-customers under this assumption, check the utility of θ_1-customers, and verify that it is nonnegative. In some cases, θ_1-customers join anyway, without considering what the others do. In other cases, their readiness to join is based on joining of θ_2-customers, at least in some states where the server is idle.

Theorem 4.1 *Suppose that θ_1 customers always join when the server is idle.*
(a) Suppose that $\Lambda_1 > \mu$, and let $n_e^a = \left\lfloor \left(\frac{R}{\theta_2} - \frac{N}{\Lambda_1}\right) / \left(\frac{1}{\mu} - \frac{1}{\Lambda_1}\right) \right\rfloor$.
if $n_e^a < 0$ then θ_2-customers never join when the server is idle;
if $n_e^a > N$ they always join when the server is idle;

[3] The conditions for equilibrium are actually milder, for example it suffices that customers balk at state 1^- to ensure that balking at 0 is an optimal response. We avoid a thorough analysis of this subject and in particular a description of all equilibria and subgame perfect solutions. The reader is referred to Hassin and Haviv (2002) and §1.5 in Hassin and Haviv (2003) for examples of such analysis.

if $n_e^a \in \{1, \ldots, N\}$*,* θ_2*-customers join at states* $0, 1^-, \ldots, (n_e^a - 1)^-$*, and balk at all other states with idle server.*

(b) Suppose that $\Lambda_1 + \Lambda_2 < \mu$*, and let* $n_e^b = \left\lceil \left(\frac{R}{\theta_2} - \frac{N}{\Lambda_1+\Lambda_2} \right) / \left(\frac{1}{\mu} - \frac{1}{\Lambda_1+\Lambda_2} \right) \right\rceil - 1$.

if $n_e^b < 0$ *then* θ_2*-customers always join when the server is idle;*

if $n_e^b > N$ *they never join when the server is idle.*

If $n_e^b \in \{1, \ldots, N\}$*,* θ_2*-customers balk at states* $0, 1^-, \ldots, (n_e^b - 1)^-$*, and join at all other states with idle server.*

(c) Suppose that $\Lambda_1 \leq \mu \leq \Lambda_1 + \Lambda_2$*.*

if $\frac{N}{\mu} < \frac{R}{\theta_2}$ *then* θ_2*-customers always join when the server is idle;*

if $\frac{N}{\mu} > \frac{R}{\theta_2}$ *then* θ_2*-customers never join when the server is idle.*

We now turn to discuss the decision of θ_1 customers.

Theorem 4.2

(a) Suppose that $\Lambda_1 > \mu$*. There exists an equilibrium with active server if and only if* $\frac{N}{\mu} \leq \frac{R}{\theta_1}$*.*

(b) Suppose that $\Lambda_1 + \Lambda_2 < \mu$*. There exists an equilibrium with active server if and only if*

$$\frac{1}{\mu} + \frac{n_e^b - 1}{\Lambda_1} + \frac{N - n_e^b}{\Lambda_1 + \Lambda_2} \leq \frac{R}{\theta_1}.$$

In both cases, θ_1*-customers join if the system occupancy is at most* $\left\lfloor \frac{R\mu}{\theta_1} \right\rfloor$*.* θ_2*-customers join if the server is busy and the system occupancy is at most* $\left\lfloor \frac{R\mu}{\theta_2} \right\rfloor$*. They also join if the server is idle as described in Theorem 4.1.*

c) Suppose that $\Lambda_1 \leq \mu \leq \Lambda_1 + \Lambda_2$*.*

If $\frac{N}{\mu} < \frac{R}{\theta_2}$*, there exists an equilibrium with active server if and only if* $\frac{N}{\mu} \leq \frac{R}{\theta_1}$*.*

If $\frac{N}{\mu} > \frac{R}{\theta_2}$*, there exists an equilibrium with active server if and only if* $\frac{1}{\mu} + \frac{N-1}{\Lambda_1} \leq \frac{R}{\theta_1}$*.*

4.1.2 Social optimization

We now consider the optimal arrival rates, which can be obtained by formulating the dynamic admission problem as a Markovian Decision Process with the average reward criterion. In general, the optimal policy shall generate none-randomized solution; that is, the optimal joining rate is either 0, Λ_1 or $\Lambda_1 + \Lambda_2$ on each state.

By looking at the transition graph of a vacation queue, we observe that it has a very special structure on the states $\{1^-, \ldots, (N-1)^-\}$. On this subset, the transition is always from state m^- to $(m+1)^-$. This means, if we exchange the arrival rates for the two conjunct states in this set, the distribution for other states is unchanged. Therefore, a *swapping* of arrival rates in that set only changes the social welfare for customers who joined at the two states. Using this property, we obtain some analytical results on the monotonicity of the social optimal arrival rates.

Proposition 4.3 *If* $\frac{\Lambda_1}{\mu} < (>) \frac{\theta_2}{\theta_2-\theta_1}$*, then the optimal arrival rates are monotone increasing (decreasing) on the states* $0, 1^-, \ldots, (N-1)^-$.

Proposition 4.3 shows that, under the condition that Λ_1 is too small or θ_2 is too large, a θ_2-customer seeing an inactive server shall join only when the queue is long (close to the re-opening time of the server).

Note that we cannot do swapping on two states when the system is busy, since, if we exchange the arrival rates for two adjacent states, the distribution for other states is changed.

From Theorem 4.1 and Proposition 4.3, we obtain the following corollary.

Corollary 4.4 *If the decentralized arrival rates are monotone increasing on* $0, 1^-, \ldots, (N-1)^-$*, the optimal ones are increasing too.*

Therefore, social optimization has a relaxed requirement on increasing arrival rates than the decentralized equilibrium does, but has a stronger requirement on decreasing monotonicity.**we need to explain the source for the last assertion**

4.2 Continuous distribution

4.2.1 Equilibrium

In the decentralized system, a θ-customer joins the queue at state s, given that the expected waiting time while doing so is W_s, if his parameter θ does not exceed a threshold θ_s defined as

$$\theta_s = \frac{R}{W_s}.$$

To simplify the presentation, we assume that the arrival rate of customers, Λ, is large so that it is neither an equilibrium nor optimal that all of them join.

For states with active server and m customers already in the system the expected waiting time is $\frac{m+1}{\mu}$. Therefore, the effective arrival rate at such a state is $\Lambda H\left(\frac{R\mu}{m+1}\right)$.

When customers are informed that the state of the queue is m^-, their expected waiting time is affected by the future arrival process. Therefore, their joining strategy depends on future customers' strategies. To solve the equilibrium arrival rates, we think backwards. We first consider $(N-1)^-$. In this case, an incoming customer's strategy is not affected by future arrivals: If he joins the queue, the system occupancy will jump up to N and the service is started immediately. In fact, joining at $(N-1)^-$ is exactly equivalent to joining at $(N-1)^+$, the expected waiting time is $W_{(N-1)^-} = \frac{N}{\mu}$, and the threshold is $\theta_{(N-1)^-} = \frac{R}{W_{(N-1)^-}} = \frac{R\mu}{N}$. In particular, the joining rate is

$$\lambda_{(N-1)^-} = \Lambda H\left(\frac{R\mu}{N}\right).$$

We now consider state $(N-2)^-$. If a customer joins, his expected waiting time consists of two parts: the expected waiting time for the queue to reach N, which is $\frac{1}{\lambda_{(N-1)^-}}$, and the expected waiting time after the server starts to work, which is $\frac{N-1}{\mu}$. That is,

$$W_{(N-2)^-} = \frac{1}{\lambda_{(N-1)^-}} + \frac{N-1}{\mu} = W_{(N-1)^-} + \frac{1}{\lambda_{(N-1)^-}} - \frac{1}{\mu}.$$

The joining rate at $(N-2)^-$ is $\lambda_{(N-2)^-} = \Lambda H\left(\frac{R}{W_{(N-2)^-}}\right)$.

One can generalize this recursive relationship to general state m^-. We summarize this recursive algorithm below.

Proposition 4.5 *The equilibrium joining rates into an idle system can be solved recursively:*

$$\begin{aligned} W_{(N-1)^-} &= \frac{N}{\mu}, & (10)\\ W_{(m-1)^-} &= W_{m^-} + \frac{1}{\lambda_{m^-}} - \frac{1}{\mu}, \quad 1 \le m \le N-1, & (11) \end{aligned}$$

where $\lambda_{m^-} = \Lambda H\left(\frac{R}{W_{m^-}}\right)$.

Recursive equation (11) is explained in the following way: Compared with a customer joining at m^-, a customer joining at $(m-1)^-$ saves $\frac{1}{\mu}$ waiting time due to the position in front of him, but has to wait for one more arrival for the service to be started, which takes $\frac{1}{\lambda_{m^-}}$ time. Using the initial condition (10) and the recursive equation (11), one can solve all the expected waiting time in equilibrium and the equilibrium arrival rates.

It is easy to see that the effective arrival rate when the system is idle may increase with the system occupancy as customers expect a shorter time for the service to be started. The following proposition shows that the monotonicity of θ_m and λ_m depends only on the relationship between $\frac{\Lambda}{\mu}$ and $H\left(\frac{R\mu}{N}\right)$.

Proposition 4.6 *Suppose that* $\frac{\Lambda}{\mu}H\left(\frac{R\mu}{N}\right) > \;(=,<)\; 1$, *then* $\theta_{(m-1)^-} > \;(=,<)\; \theta_{m^-}$ *for* $m = 1, \ldots, N-1$.

4.2.2 Social optimization

We first give a proposition about the monotonicity of arrival rates.

Proposition 4.7 *Suppose that* $\frac{\Lambda}{\mu}H\left(\frac{R\mu}{N}\right) \leq 1$, *then the optimal thresholds satisfy*

$$\theta_0^* \leq \theta_{1^-}^* \leq \cdots \leq \theta_{(N-1)^-}^*.$$

Similar to Proposition 4.3, Proposition 4.7 shows that as the total arrival rate Λ is small or the threshold N is large, more sensitive customers shall join the longer queues when the server is idle.

This proposition is similar to the one with decentralized arrivals. However, we cannot say that if the decentralized arrival rates are increasing on the states $0, 1^-, \ldots, (N-1)^-$, the optimal ones are also increasing. The reason is that $\lambda_{(N-1)^-} < \mu$ need not imply that $\lambda_{(N-1)^-}^* \leq \mu$.

Define $u_m(\lambda)$, $m = \{0, 1^-, \ldots, (N-1)^-, 1^+, 2^+, \ldots, (N-1)^+, \ldots, N, \ldots\}$ to be the average rate of utility obtained from customers who arrive at the given state m and the arrival rate is λ. For example, if H is the cdf of a uniform distribution on $[0, 1]$, the threshold level for joining customers is $\theta = \frac{\lambda}{\Lambda}$ and we can expressed $u_m(\lambda)$ as follows:

$$u_m(\lambda) \quad = \quad \int_0^{\frac{\lambda}{\Lambda}} (R - \theta W_m) d\theta = R\frac{\lambda}{\Lambda} - 0.5\frac{\lambda^2}{\Lambda^2}W_m,$$

where W_m is the expected waiting time given the state m.

The social welfare optimization problem is to choose $(\lambda_m, m \in \{0, 1^-, \ldots, (N-1)^-, 1^+, 2^+, \ldots, (N-1)^+, \ldots, N, \ldots\})$ maximizing the social welfare

$$SW = \Lambda \sum_{m=0,1^-,\ldots} u_m(\lambda_m)p_m.$$

To obtain the optimal arrival rates for the above social optimization problem, we formulate the problem as a Markov Decision Problem. **I think that the next sentence and reference are redundant, we don't have to use uniformization, it's just an option** The optimality equations can be obtained by applying the uniformization method to the Markov process (see, Bertsekas, 1976).

Assume $N = 3$ and that H is the uniform distribution on $[0, 1]$. We change R over $\{1, 2, 3, 4\}$ and μ and Λ over $\{0.5, 1, 2, 3, 4, 5, 6\}$. Table 1 shows the social welfare under social optimization and in equilibrium for the cases with $R = 1$, $\mu = 2$ and $\Lambda = \{1, 2, 3, 4, 5, 6\}$.

Figure 2 presents the equilibrium and optimal arrival rates for some cases. On the X-axis, we use -2 and -1 to represent the state 2^- and 1^-, respectively. We observe that when the server is busy, the optimal arrival rate is always smaller than the equilibrium arrival rate. However, when the server is on vacation, the optimal arrival rate is larger than the equilibrium arrival rate when the potential arrival is not too heavy; and is smaller when the potential arrival rate is heavy. We observed similar phenomena for other sets of parameters. This can be explained from the externality. When the server is on vacation and arrival is light, a joining customer brings positive externality to other customers. Therefore, social optimization generates larger arrival rates than equilibrium ones. However, when the arrival is heavy, a joining customer brings negative externality to other customers even though the server is on vacation. Therefore, optimal arrival rates are smaller than the equilibrium ones.

Our numerical results also conform with Proposition 4.7: We see that when $\lambda^*_{2^-} < \mu$, the optimal arrival rates are increasing on $\{0, 1^-, 2^-\}$.

Table 1: Social welfare

Λ	Maximal SW	Equilibrium SW
1	0.2373	0.2080
2	0.5946	0.5720
3	0.8568	0.8064
4	1.0410	0.9199
5	1.1745	0.9683
6	1.2752	0.9880

5. Conclusions

To summarize, we study customers' decentralized behavior and social optimization in an M/M/1 queue with an N-policy. When the queue is unobservable, though multiple equilibria usually exist, the optimal arrival rate is unique, which could be larger than some of the equilibrium ones.

When the queue is observable, we derive the conditions for the system to be active and provide closed-form expressions for the thresholds with two-type customers and a recursive algorithm for calculating the thresholds with continuous customers. We show that when the equilibrium arrival rates are increasing with state when the server is on vacation, the optimal ones are increasing too. For the continuous customers, we show, by numerical study, that the optimal arrival rates are always smaller than the equilibrium ones when the server is busy; smaller when the server is on vacation and the arrival is very large. However, they are larger than equilibrium ones when the server is on vacation and arrival rate is not too large.

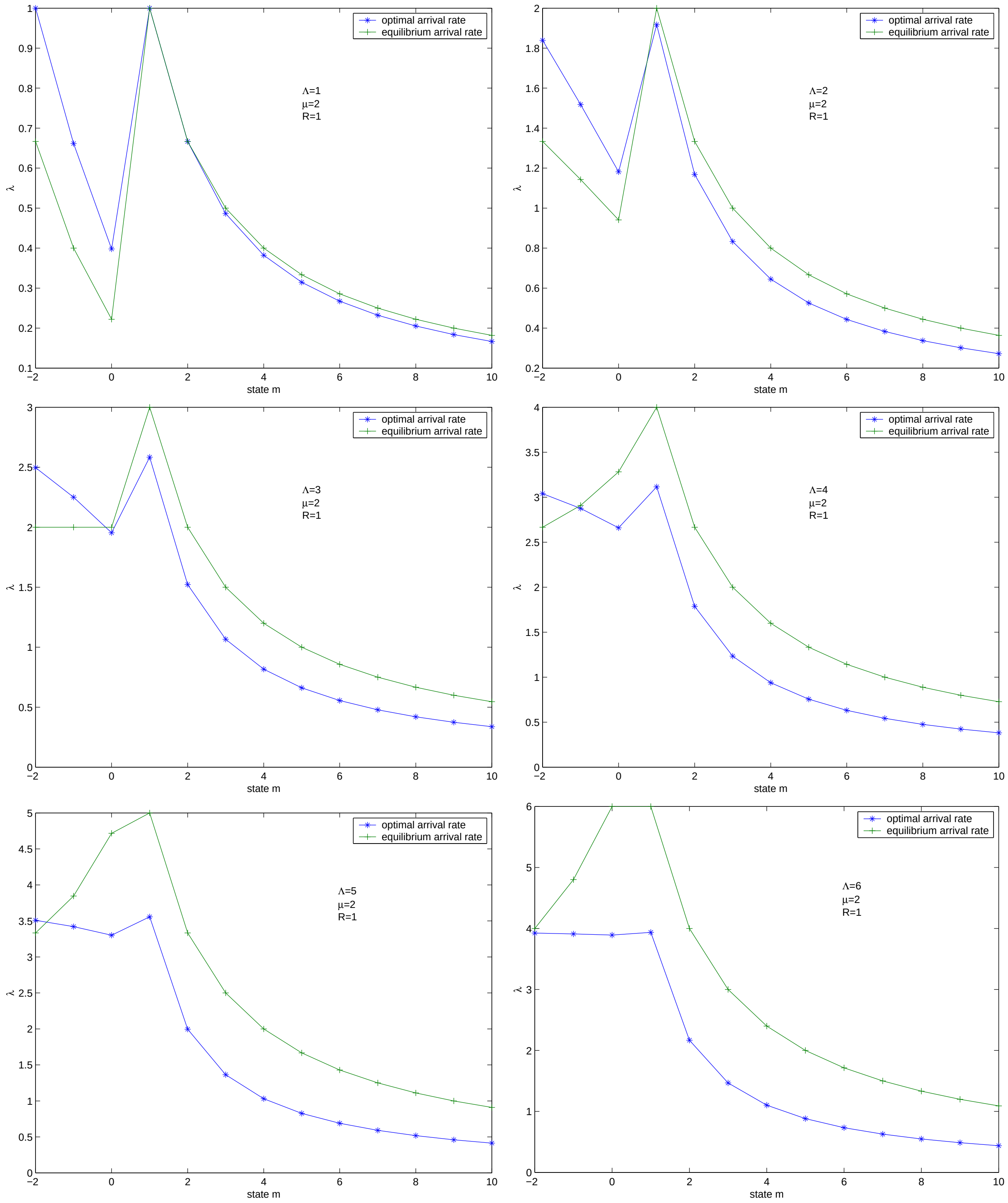

Figure 2: Equilibrium and optimal arrival rates

Acknowledgment

Guo's research was supported by HK PolyU research foundation grant no. PB0W. Hassin's research was supported by the Israel Science Foundation grant no. 526/08.

Appendix

Proof of Lemma 3.1

Consider the first- and second-order derivatives of SW_1 with respect to λ.

$$\begin{aligned} SW_1^{'}(\lambda) &= R - \theta_1 W(\lambda) - \lambda\theta_1 W^{'}(\lambda) \\ SW_1^{''}(\lambda) &= -2\theta_1 W^{'}(\lambda) - \lambda\theta_1 W^{''}(\lambda) \\ &= -\theta[2W^{'}(\lambda) + \lambda W^{''}(\lambda)] \\ &= -\theta\left[2\left(\frac{1}{(\mu-\lambda)^2} - \frac{N-1}{4\lambda^2}\right) + \lambda\left(\frac{2}{(\mu-\lambda)^3} + \frac{2(N-1)}{4\lambda^3}\right)\right] \\ &= -\theta\frac{2\mu}{(\mu-\lambda)^3} < 0. \end{aligned}$$

Now consider SW_2. We can rewrite it as

$$\begin{aligned} SW_2(\lambda) &= \lambda R - [\Lambda_1\theta_1 + (\lambda - \Lambda_1)\theta_2]W(\lambda) \\ &= \lambda R - \lambda\theta_1 W(\lambda) - (\lambda - \Lambda_1)(\theta_2 - \theta_1)W(\lambda) \\ &= SW_1(\lambda) - (\lambda - \Lambda_1)(\theta_2 - \theta_1)W(\lambda). \end{aligned}$$

It can be easily shown that $-(\lambda - \Lambda_1)(\theta_2 - \theta_1)W(\lambda)$ is strictly concave on $[\tilde{\lambda}, \mu)$. Therefore, SW_2 is the sum of two strictly concave functions and thus is also strictly concave. **It seems that the next sentence is redundant** Furthermore, it is always smaller than SW_1. The social welfare SW (which is composed first by SW_1 and then SW_2 (smaller than SW_1)) must be strictly concave on $[\tilde{\lambda}, \mu)$.

On $[0, \tilde{\lambda}]$, we know that $W^{'}(\lambda) < 0$ because $W(\lambda)$ is strictly decreasing on $[0, \tilde{\lambda}]$. Also, when λ is very small, $R - \theta_1 W(\lambda) < 0$ which implies that the social welfare is negative. Therefore, we can safely ignore such small λ and assume that $R - \theta_1 W(\lambda) \geq 0$. These conditions together imply that $SW_1^{'}(\lambda) > 0$. Similarly, we can show that SW_2 is strictly increasing in λ on the domain where it is positive within the range of $[0, \tilde{\lambda}]$. **The next sentence is unclear to me, and seems unrelated to the proof of the lemma** Therefore, we can restrict the social welfare optimization problem on $[\tilde{\lambda}, \mu)$.

Proof of Proposition 3.2

The uniqueness of $\bar{\lambda}$ follows from the strict unimodality of SW function in Lemma 3.1.

To prove part (b), consider different cases.

Case 1: $\bar{\lambda}^1 \in [0, \Lambda_1^1]$.

Since $SW^1 = SW^2 = SW_1$ on this range, $(SW^2(\bar{\lambda}^1))' = 0$. $\bar{\lambda}^1$ must be the optimal solution due to the strict unimodality of SW^2. Therefore, $\bar{\lambda}^2 = \bar{\lambda}^1$.

Case 2: $\bar{\lambda}^1 \in (\Lambda_1^1, \Lambda_1^2]$. On this range,

$$\begin{aligned} SW^2 &= SW_1 \\ SW^1 &= SW_1 - (\lambda - \Lambda_1^1)(\theta_2 - \theta_1)W(\lambda). \end{aligned}$$

Since $\bar{\lambda}^1$ is the optimal solution, $\bar{\lambda}^1 \geq \tilde{\lambda}$ as SW^1 is monotone increasing in $[0, \tilde{\lambda})$. Therefore, $W(\lambda)$ must be increasing in $[\bar{\lambda}^1, \infty)$ and so is $(\lambda - \Lambda_1^1)(\theta_2 - \theta_1)W(\lambda)$. Hence, $(\lambda - \Lambda_1^1)(\theta_2 - \theta_1)W(\lambda)$ must have positive derivative at $\bar{\lambda}^1$. We obtain that

$$(SW^2(\bar{\lambda}^1))' = SW_1'(\bar{\lambda}^1) > 0.$$

From the above inequality, we can conclude that $\bar{\lambda}^2 \geq \bar{\lambda}^1$.

Case 3: $\bar{\lambda}^1 \in (\Lambda_1^2, \Lambda_1^1 + \Lambda_2^1]$. On this range,

$$\begin{aligned} SW^2 &= SW_1 - (\lambda - \Lambda_1^2)(\theta_2 - \theta_1)W(\lambda) \\ SW^1 &= SW_1 - (\lambda - \Lambda_1^1)(\theta_2 - \theta_1)W(\lambda). \end{aligned}$$

We can rewrite SW^2 as

$$SW^2 = SW^1 + (\Lambda_1^2 - \Lambda_1^1)(\theta_2 - \theta_1)W(\lambda).$$

We know that $W(\lambda)$ is increasing on $[\bar{\lambda}^1, \infty)$. Therefore

$$(SW^2(\bar{\lambda}^1))' = (\Lambda_1^2 - \Lambda_1^1)(\theta_2 - \theta_1)W'(\bar{\lambda}^1) > 0.$$

Again, we can conclude that $\bar{\lambda}^2 \geq \bar{\lambda}^1$.

Proof of Proposition 3.4

Comparing the thresholds in the range where both are in $(0, 1)$, by (8) and (9), the condition for $\theta^* > \theta_e$ is

$$\frac{\mu}{\Lambda}\left(1 - \frac{1}{\sqrt{\Lambda R_0 + 1}}\right) \quad > \quad \frac{\mu}{\Lambda} \cdot \frac{1 + 2R\Lambda - N}{3 + 2R\Lambda - N}.$$

This can be simplified to be

$$\frac{1}{1+\Lambda R-\frac{N-1}{2}} \quad > \quad \frac{1}{\sqrt{1+2\Lambda R-\frac{N-1}{2}}}.$$

This is equivalent to

$$(\Lambda R)^2-(N-1)\Lambda R+\frac{(N-1)(N-3)}{4}<0,$$

or

$$\left(\frac{N-1}{2}-\sqrt{\frac{N-1}{2}}\right)\frac{1}{\Lambda}<R<\left(\frac{N-1}{2}+\sqrt{\frac{N-1}{2}}\right)\frac{1}{\Lambda}.$$

The left-hand side inequality follows from our assumption that $\theta_e>0$.

If $\frac{\Lambda}{\mu-\Lambda}\leq\sqrt{\frac{N-1}{2}}$ then $\frac{1}{\Lambda}\left(\frac{N-1}{2}+\sqrt{\frac{N-1}{2}}\right)>\frac{1}{\mu-\Lambda}+\frac{N-1}{2\Lambda}$, and therefore $\theta_e<\theta^*$ whenever $\theta_e<1$.

Proof of Theorem 4.1

(a) $\Lambda_1>\mu$. If θ_2-customers join at m^-,

$$W_{(m-1)^-}=W_{m^-}+\frac{1}{\Lambda_1+\Lambda_2}-\frac{1}{\mu}<W_{m^-};$$

otherwise,

$$W_{(m-1)^-}=W_{m^-}+\frac{1}{\Lambda_1}-\frac{1}{\mu}<W_{m^-}.$$

Therefore, the expected waiting time is increasing with the queue length when the server is idle. Hence, θ_2-customers use a threshold strategy to join the queue: Only when the queue is smaller than a level, they join.

Define n_e to be the minimum index such that θ_2-customers balk at n_e^-. That is, they join at all lower index states, and balk at n_e and all higher index states. The expected waiting time for a customer who joins at state $(n_e-1)^-$ is

$$\frac{N-n_e}{\lambda_1}+\frac{n_e}{\mu}\leq\frac{R}{\theta_2},$$

where the inequality is necessary for the customer to be willing to join at this state. At state n_e^- the customer refuses to join, hence the waiting time there satisfies

$$\frac{N-n_e-1}{\Lambda_1}+\frac{n_e+1}{\mu}>\frac{R}{\theta_2}.$$

These two inequalities determine the claimed value for n_e. Note that if $\frac{R}{\theta_2}<\frac{N-1}{\Lambda_1}+\frac{1}{\mu}$ then it is not worth joining for a θ_2-customer even at state 0. If, on the other hand, $\frac{R}{\theta_2}>\frac{N}{\mu}$ so that the resulting value for n_e is greater than N, it means that θ_2-customers always join when the server is idle, even at state $(N-1)^-$.

(b) $\Lambda_1 + \Lambda_2 < \mu$. Similar to the argument in (a), the expected waiting time is decreasing with the queue when the server is idle. Therefore, θ_2 customers use a threshold strategy: They join when the queue is larger than a threshold level.

Define n_e^b to be the maximum index such that θ_2-customers balk at $(n_e^b - 1)^-$. That means, when the server is idle, they balk at all lower index states, and join at $(n_e^b)^-$ and all higher index idle states. The expected waiting time for a customer who joins at state $(n_e^b - 1)^-$ satisfies

$$\frac{N - n_e^b}{\Lambda_1 + \Lambda_2} + \frac{n_e^b}{\mu} > \frac{R}{\theta_2}.$$

At state $(n_e^b)^-$, the customer joins, hence

$$\frac{N - n_e^b - 1}{\Lambda_1 + \Lambda_2} + \frac{n_e^b + 1}{\mu} \leq \frac{R}{\theta_2}.$$

These two inequalities define the claimed threshold value.

Note that if $\frac{N}{\mu} > \frac{R}{\theta_2}$ then there is no joining even at state $(N-1)^-$, whereas if $\frac{N-1}{\Lambda_1} + \frac{1}{\mu} \leq \frac{R}{\theta_2}$ then θ_2-customers always join when the server is idle, even at state 0.

(c) $\Lambda_1 \leq \mu \leq \Lambda_1 + \Lambda_2$. If θ_2-customers join at m^-,

$$W_{(m-1)^-} = W_{m^-} + \frac{1}{\Lambda_1 + \Lambda_2} - \frac{1}{\mu} < W_{m^-};$$

Then, θ_2-customers must join at $(m-1)^-$,

If θ_2-customers balk at m^-,

$$W_{(m-1)^-} = W_{m^-} + \frac{1}{\Lambda_1} - \frac{1}{\mu} > W_{m^-}.$$

Then, θ_2-customers must balk at $(m-1)^-$ too.

Therefore, we only need consider θ_2-customers' strategy at $(N-1)^-$. This yields the conclusion.

Proof of Theorem 4.2

Clearly, an equilibrium with active server requires that θ_1 customers join in all states where the server is idle. Of these states, in case (a), the one with highest expected waiting time is state $(N-1)^-$, and in case (b) it is state 0. The conditions given by the theorem state that these expected waiting times should exceed R/θ_1.

Consider case (b).

The longest waiting for θ_1 customers happens at 0 since we show in Theorem 4.1 that the expected waiting time is decreasing with the state. When $n_e^b \in \{1, 2, \ldots, N\}$, the condition becomes:

$$\frac{1}{\mu} + \frac{n_e^b - 1}{\Lambda_1} + \frac{N - n_e^b}{\Lambda_1 + \Lambda_2} \leq \frac{R}{\theta_1}.$$

Specifically, when $n_e^b \leq 0$, according to Theorem 4.1, θ_2 customers join all states with idle server. The condition becomes

$$\frac{1}{\mu} + \frac{N-1}{\Lambda_1 + \Lambda_2} \leq \frac{R}{\theta_1}.$$

Similarly, when $n_e^b \geq N$, the θ_2-customers balk when the server is idle. The condition becomes

$$\frac{1}{\mu} + \frac{N-1}{\Lambda_1} \leq \frac{R}{\theta_1}.$$

Consider case (c). We show in Theorem 4.1 that if $\frac{N}{\mu} < \frac{R}{\theta_2}$ then θ_2-customers always join if θ_1-customers join; otherwise, they never join. When θ_2-customers join the queue, since $\Lambda_1 + \Lambda_2 \geq \mu$, the longest waiting happens at state $(N-1)^-$. The condition for an active server is $\frac{N}{\mu} \leq \frac{R}{\theta_1}$. When θ_2-customers never join the queue, $\Lambda_1 \leq \mu$, therefore, the longest waiting happens at state 0. The condition an active server is $\frac{1}{\mu} + \frac{N-1}{\Lambda_1} \leq \frac{R}{\theta_1}$.

Proof of Proposition 4.3

Consider the condition $\frac{\Lambda_1}{\mu} < \frac{\theta_2}{\theta_2 - \theta_1}$. Assume decreasing arrival rates on states with idle servers: the equilibrium arrival rate on state m^- is $\Lambda_1 + \Lambda_2$ and the one on $(m+1)^-$ is Λ_1. Then, we swap the customers behavior on the two states. This swapping doesn't change the distribution for other states. Hence, we can compare the social welfare on the two states only.

Since the conditional probabilities on state m^+ and $(m+1)^+$ are proportional to the inverse of arrival rates on the two states, we have the conditional probabilities: $p_{m^+} = \frac{1/(\Lambda_1+\Lambda_2)}{1/(\Lambda_1+\Lambda_2)+1/\Lambda_1}$ and $p_{(m+1)^+} = \frac{1/\Lambda_1}{1/(\Lambda_1+\Lambda_2)+1/\Lambda_1}$. Denote the average cost on state m^+ by C_{m^+} and the one on state $(m+1)^+$ by $C_{(m+1)^+}$. When we compute the total utility or cost on these two states, we need times the probability with the corresponding arrival rate. Thus, the total cost can be expressed as

$$C = p_{m^+}(\Lambda_1 + \Lambda_2)C_{m^+} + p_{(m+1)^+}\Lambda_1 C_{(m+1)^+} = \frac{1}{1/(\Lambda_1 + \Lambda_2) + 1/\Lambda_1}(C_{m^+} + C_{(m+1)^+}).$$

We can therefore just need consider the value for the term $C_{m^+} + C_{(m+1)^+}$ before and after swapping. Denote this value before and after swapping by $C^{(1)}$ and $C^{(2)}$, respectively.

Denote the expected waiting time for the customer joining at $(m+1)^-$ to be W. For the original system,

$$C^{(1)} = \left(\frac{\Lambda_1\theta_1}{\Lambda_1 + \Lambda_2} + \frac{\Lambda_2\theta_2}{\Lambda_1 + \Lambda_2}\right)\left(W + \frac{1}{\Lambda_1} - \frac{1}{\mu}\right) + \theta_1 W.$$

The average cost after swapping is

$$C^{(2)} = \theta_1 \left(W + \frac{1}{\Lambda_1 + \Lambda_2} - \frac{1}{\mu} \right) + \left(\frac{\Lambda_1 \theta_1}{\Lambda_1 + \Lambda_2} + \frac{\Lambda_2 \theta_2}{\Lambda_1 + \Lambda_2} \right) W.$$

We can then obtain

$$C^{(1)} - C^{(2)} = \frac{\Lambda_2}{\Lambda_1 + \Lambda_2} \left(\frac{\theta_2}{\Lambda_1} - \frac{\theta_2 - \theta_1}{\mu} \right) > 0.$$

Therefore, if $\frac{\Lambda_1}{\mu} < \frac{\theta_2}{\theta_2 - \theta_1}$, the decreasing sequence shall be swapped, which generates the increasing monotonicity of the optimal arrival rates.

Proof of Corollary 4.4

If the decentralized arrival rates are monotone increasing on $\{0, 1^-, \ldots, (N-1)^-\}$, then $\Lambda_1 \geq \mu$, according to Theorem 4.1.

$$\frac{\Lambda_1}{\mu} < 1 < \frac{\theta_2}{\theta_2 - \theta_1}.$$

Hence, the optimal arrival rates must be monotone increasing on the set $\{0, 1^-, \ldots, (N-1)^-\}$ according to Proposition 4.3.

Proof of Proposition 4.6

By the recursive relationship in (11),

$$\theta_{(m-1)^-} \left(W_{m^-} + \frac{1}{\lambda_{m^-}} - \frac{1}{\mu} \right) = R.$$

By assumption, $\theta_{m^-} W_{m^-} = R$.

By comparing the above two equations, we see that: if $\lambda_{(N-1)^-} > \mu$, $\theta_{(N-2)^-} > \theta_{(N-1)^-}$ and therefore $\lambda_{(N-2)^-} > \lambda_{(N-1)^-} > \mu$. Again, using the new conclusion, we conclude that $\lambda_{(N-3)^-} > \lambda_{(N-2)^-}$ and so forth.

The proofs for other cases are similar.

Proof of Proposition 4.7

Note that $\lambda^*_{(N-1)^-} = \Lambda H(\frac{R\mu}{N})$. Therefore, the initial condition $\frac{\Lambda}{\mu} H(\frac{R\mu}{N}) \leq 1$ is equivalent to $\lambda^*_{(N-1)^-} \leq \mu$.

We now provide the conclusion by contradiction. Consider adjacent states m^- and $(m+1)^-$. Assume $\lambda^*_{m^-} > \lambda^*_{(m+1)^-}$. This implies that $\theta^*_{m^-} > \theta^*_{(m+1)^-}$. Consider swapping the arrival rates for these two states. It only affects the costs of customers joining in these states.

Let W denote the expected waiting time for a customer who joins at state $(m+1)^-$. Let $C^{(1)}$ and $C^{(2)}$ denote the costs associated with customers joining at the two states before and after the change, respectively. Then,

$$\begin{aligned} C^{(1)} &= \int_0^{\theta^*_{m^-}} \theta \left(W + \frac{1}{\lambda^*_{(m+1)^-}} - \frac{1}{\mu} \right) dH(\theta) + \int_0^{\theta^*_{(m+1)^-}} \theta W dH(\theta); \\ C^{(2)} &= \int_0^{\theta^*_{(m+1)^-}} \theta \left(W + \frac{1}{\lambda^*_{m^-}} - \frac{1}{\mu} \right) dH(\theta) + \int_0^{\theta^*_{m^-}} \theta W dH(\theta). \end{aligned}$$

From that, we obtain

$$C^{(1)} - C^{(2)} = \left(\frac{1}{\lambda^*_{(m+1)^-}} - \frac{1}{\mu} \right) \int_{\theta^*_{(m+1)^-}}^{\theta^*_{m^-}} \theta dH(\theta) + \left(\frac{1}{\lambda^*_{(m+1)^-}} - \frac{1}{\lambda^*_{m^-}} \right) \int_0^{\theta^*_{(m+1)^-}} \theta dH(\theta). \quad (12)$$

We now proceed backwards starting from $m = N-2$. When $m = N-2$, $\lambda^*_{(m+1)^-} = \lambda^*_{(N-1)^-} \le \mu$ according to the initial condition $\frac{\Lambda}{\mu} H(\frac{R\mu}{N}) \le 1$. Therefore the first term of the right-hand-side of (12) is non-negative. By the assumption that $\lambda^*_{m^-} > \lambda^*_{(m+1)^-}$, the second-term of r.h.s. of (12) is positive. Hence, it is better to do swapping to generate (weakly) increasing arrival rates on $\{N-2, N-1\}$. Consequently, $\lambda^*_{N-2^-} \le \lambda^*_{(N-1)^-} \le \mu$.

Similarly, we can show the increasing property on states $\{(N-3)^-, (N-2)^-\}$ and so on so forth for other smaller states.

References

Bertsekas, D., 1976. Dynamic Programming and Stochastic Control. Academic Press, New York.

Boxma, O., 1989. Workloads and waiting times in single-server systems with multiple customer classes. Queueing Systems 5, 185-214.

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

Dimitrakopoulos, Y. and A. Burnetas. 2011. Customer equilibrium and optimal strategies in an M/M/1 queue with dynamic service control. Working paper, Univ. of Athens.

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

Economou, A., A. Gómez-Corral, and S. Kanta, 2011. Optimal balking strategies in single-server queues with general service and vacation time. Performance Evaluation 68 967-982.

Edelson, N. and K. Hildebrand, 1975. Congestion tolls for Poisson queueing processes. Econometrica 43, 81-92.

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

Guo, P. and G. Zhang. 2012. Strategic queueing behavior and its impact on system performance in service systems with the congestion-based staffing policy. Working paper, Hong Kong Polytechnic University, Hong Kong.

Levy, H. and M. Sidi, 1990. Polling systems: applications, modeling and optimization. IEEE Transactions on Communication 38, 1750-1760.

Naor, P., 1969. The regulation of queue size by levying tolls. Econometrica 37, 15-24.

Sun, W., P. Guo, and N. Tian, 2010. Equilibrium threshold strategies in observable queueing systems with setup/closedown times. Central European Journal of Operations Research 18 241-268.

Hassin, R. and M. Haviv, 1997. Equilibrium threshold strategies: The case of queues with priorities. Operations Research 45, 966–973.

Hassin, R. and M. Haviv, 2003. To Queue Or Not To Queue: Equilibrium Behavior in Queueing Systems. Kluwer.

Hassin, R. and M. Haviv 2002 Nash equilibrium and subgame perfection: the case of observable queues. Annals of Operations Research 113, 15–26.

Johari, R. and S. Kumar, 2008. Externalities in services. Working paper, Graduate School of Business, Stanford University, Stanford, CA.

Takagi, H., 1986. Analysis of Polling Systems. The MIT Press.

Takagi, H., 2000. Analysis and application of polling models. In G. Haring, C. Lindemann, M. Reiser (Eds.), Performance Evaluation: Origins and Directions, Lecture Notes in Computer Science 1769, Springer, Berlin, 423-442.

Tian, N. and G. Zhang, 2006. Vacation Queueing Models: Theory and Applications. Springer.

Veeraraghavan, S. and L. Debo, 2009. Joining longer queues: Information externalities in queue choice. Manufacturing & Service Operations Management 11 543-562.

Yadin, M. and P. Naor, 1963. Queueing systems with a removable service station. Operations Research 14, 393–405.

Yechiali, U. 1993. Analysis and control of polling systems. In L. Donatiello and R. Nelson (Eds.) Performance Evaluation of Computer and Communication Systems. Springer-Verlag 630-650.