A Model of Rational Retrials in Queues

Shiliang Cui* Xuanming Su† Senthil Veeraraghavan‡

Consumers often have to join a queue before they can obtain a service; however, they suffer disutility in waiting for the service. When consumers can self-organize the timing of their service visits, they may avoid long queues and choose to retry later. Nevertheless, the existing research on decision-making in queues has focused mainly on join or balk decisions, while acknowledging that some consumers may postpone their visits and retry later. We study an observable queue in which consumers make rational join, balk and (costly) retry decisions upon their arrival. Retrial attempts could be costly due to factors such as transportation costs, retrial hassle and visit fees. We characterize the equilibrium under such retrial behavior, and study its welfare effects. With the additional option to retry, consumer welfare could worsen compared to the welfare in a system without retrials. Surprisingly, self-interested consumers retry too little (in equilibrium compared to the socially optimal policy) when the retrial cost is low, and retry too much when the retrial cost is high. We explore the impact of customers who have emergency needs and do not have the flexibility to retry or balk. We find that the presence of these "critical" arrivals typically destroys consumer welfare, even that their welfare is increased by the presence of the "regular" consumers who make join, balk, or retry decisions.

Key words: Consumer Decisions, Retrials, Queueing Games, Equilibrium versus Social Optimum.

1. Introduction

In many service settings such as post offices or ATM machines, consumers usually have to join a queue before they can obtain the service. It is well-known that consumers do not enjoy waiting in queues. In real life, when the queues are long, consumers may not be willing to wait, but may choose to retry later (instead of balking). For instance, consider a customer who arrives at a post office to pick up a package. Upon seeing the state of the queue, i.e., the number of consumers that are already in the queue, this customer can either decide to join the queue or to leave only to return at a later time when the queue may be shorter. However, the existing queueing literature on modeling strategic consumer decisions has focused only on join and balk decisions – with customers not making retry decisions.

* McDonough School of Business, Georgetown University, Washington, DC, 20057. EML: shiliang.cui@georgetown.edu
† The Wharton School, University of Pennsylvania, Philadelphia, PA, 19104. EML: xuanming@wharton.upenn.edu
‡ The Wharton School, University of Pennsylvania, Philadelphia, PA, 19104. EML: senthilv@wharton.upenn.edu

In the package pick-up example and in many other real-life queues, such as lottery kiosks and DMV centers, retry decisions are commonly observed practices. In this paper, we build a model that allows consumers to retry for the service in future, with some retrial cost, upon seeing the status of the queue. This retrial hassle cost could be due to the additional transportation back and forth, the disutility of receiving the service at a future date, the hassle of re-planning / rescheduling, or visit fees for entering the system.

To the best of our knowledge, strategic consumer *retry decisions* in service operations settings have remained almost unaddressed, even though such actions have been acknowledged in both popular press and academic literature. We study three key issues in this paper:

1. Using a game-theoretic framework, we study how consumers make join, balk or retry decisions. We also explore how equilibrium decisions depend on retrial hassle cost which may vary based on the specific circumstance.
2. From a consumer-welfare point of view, a customer who decides to come back to the system later will generate externalities to other consumers. However, it is not clear if the net effect of the overall externalities is positive or negative. We study the impact of retrials on consumer welfare, and compare the equilibrium welfare to the social optimum.
3. We acknowledge that in some service settings, a portion of the consumers cannot afford to delay their workload or balk. Examples include patients arriving at a hospital with critical conditions, or drivers at a gas station whose cars have run out of gas. We term these consumers with emergency needs as "critical" arrivals, as opposed to "regular" arrivals who can choose to join, balk or retry. We study the impact of the critical arrivals on the queueing system.

To address these issues, in this paper we propose a new model for rational decisions of consumer retrials in queues. Our main findings include:

- With the additional option to retry, consumer welfare could be worse in equilibrium.
- Compared to the socially optimal outcome, self-interested consumers do not retry enough when the retrial hassle cost is small, but they retry too often when the retrial hassle cost is high.
- Consumer welfare decreases with the proportion of critical arrivals, although these customers are better off as the proportion of regular customers increases.

The remainder of this paper is structured as follows. We conclude this section with a review of related literature. Section 2 describes the model. Section 3 investigates consumers' behavior as a function of the retrial hassle cost. Section 4 studies welfare, and then compares individual equilibrium to the socially optimal outcome. Section 5 exploits the impact of critical arrivals. Section 6 gives concluding remarks. All technical proofs are deferred to Appendix A.

1.1. Related Literature

Mandelbaum and Yechiali (1983) considers a single strategic customer who can join the queue, balk or wait outside the queue to retry, upon seeing the current state of the system, while all other customers join the queue unconditionally. In our paper, we allow for *all* consumers to have the option to retry, therefore the future arrivals to the system are endogenously determined by consumer decisions, since each consumer's decision depends on the decisions of other people.

While Kulkarni (1983a,b), Elcan (1994) and Hassin and Haviv (1996) have studied socially optimal and equilibrium retrial frequency decisions of consumers who are forced to retry upon seeing a busy system, our paper reverses the focus to examine consumers' strategic retry versus join and balk decisions upon arrival, and does not study the retrial frequency decisions.

Besides Mandelbaum and Yechiali (1983), there seems to be no other previous research on the topic of modeling strategic consumer retry decisions, which our paper does. However, two well-established research streams are relevant. These are papers on "modeling strategic consumer decision-making without retry decisions" under the service operations literature, and papers on "retrial queues with non-decision-making consumers" under the network and call center literature.

The first literature on queueing models with strategic consumers dates back to the seminal work by Naor (1969), who studies a single-server system with an observable queue. In Naor's model, customers who are homogeneous in waiting costs observe the queue length upon arrival before making a decision to join the system or to balk. Our paper directly extends Naor's model by allowing for customers to have a third option to retry later.

Following Naor (1969), strategic consumer behavior has been studied in the context of heterogeneous service values (e.g., Larsen (1998), Miller and Buckman (1987)), time costs (e.g., Afèche and Mendelson (2004)) and many others (e.g., see the survey in Mendelson and Whang (1990) and the comprehensive review by Hassin and Haviv (2003)). Nevertheless, in all these papers above strategic consumers only make state-dependent join or balk decisions, and do not retry.

Second, retrial queues have been employed in network models with *non-decision-making* consumers, i.e., consumers or other objects are specified by the system on when and how they *should* retry in order to optimize the system efficiency. In contrast, consumers make their own strategic retry decisions in our model. The network models are also called the *orbit models*, as the consumers waiting to try again are said to be in orbit. Because of the intractable nature of the orbit models, analytic results are generally difficult to obtain. Hence, there has been a significant focus on numerical and approximation methods (e.g., Reed and Yechiali (2013)). As a special case, Kulkarni and Choi (1990), Aissani (1994) and Artalejo (1997) consider retrial queues due to unreliable servers.

We refer interested readers to Yang and Templeton (1987), Falin (1990), Falin and Templeton (1997) and Artalejo (1999, 2010) for surveys, a monograph and bibliographies on work related to orbit models.

Besides the network literature, 'retrials' have also been recognized as an important factor in the call center literature. For example, Whitt (2002) describes some research directions related to stochastic models of call centers and points out that the possibility of postponing some work, for example by using call-back options, is worth more careful study. Hoffman and Harris (1986) and Aguir et al. (2004) consider consumer abandonments and retrials in a call center setting to estimate real arrivals (i.e., first-time consumers vs. retrial consumers). Mandelbaum et al. (2002) provide approximations of system metrics of a multi-server system with abandonment and retrials. de Véricourt and Zhou (2005) consider retrials generated by quality problems but not system capacity; that is, a caller whose call has not been resolved will call the service center again with the same concern. Aguir et al. (2008) investigate the impact of disregarding retrials in the staffing of a call center. For other papers in this literature, we refer interested readers to the surveys in Gans et al. (2003) and Akşin et al. (2007). Nevertheless, these papers above typically assume that balking or abandoning consumers retry with a constant probability.

Artalejo (1995), Artalejo and Lopez-Herrero (2000) and Shin and Choo (2009) consider queues where the retrial probabilities dependent on the number of consumers in the orbit but are controlled by the system. In our model, consumers' retry decisions also vary with the state of the system but are endogenously determined by the equilibrium conditions.

Parlaktürk and Kumar (2004) consider self-routing consumers in queueing networks. They study the behavior of the system in Nash equilibrium. Our work is similar with this paper in that any consumer needs to take into consideration future arrivals when making his decisions. Also related is the paper by Hassin and Roet-Green (2011) who consider an unobservable queue where consumers may inspect a queue (to obtain queue length information) before joining or balking.

The closest work to our model contains a set of papers by Armony and Maglaras (2004a,b). In these two papers, the authors study the call center context where customers, upon calling and hearing the waiting signal, can choose to join the queue, balk or leave their numbers for a customer representative to call back. There is a guaranteed delay before which the call back would take place. Consumers would make their decisions based on this guaranteed call-back delay information with either real-time (Armony and Maglaras 2004a) or steady-state (Armony and Maglaras 2004b) waiting-time information provided by the system. The differences between our paper and Armony and Maglaras (2004a,b) are two-fold: First, the retry decisions are driven by consumers in our

paper but the call-back decisions by the server in theirs. Second, we focus on the direct analysis of the underlying Markov chain while they provide an asymptotic analysis in heavy-traffic regimes, in a more complex setting.

Kostami and Ward (2009) study an amusement park setting where there is a regular waiting line versus an off-line waiting option (e.g., the FASTPASS system at Disneyland). Similar to Armony and Maglaras (2004a,b), consumers can choose their line to join based on the waiting time information provided by the system, but the authors assume that some consumers waiting in the off-line queue will not return. The focus of this paper is the service provider's capacity allocation rather than consumer decisions.

Finally, Akşin et al. (2013) studies consumers' endogenous abandonment behavior from call center data using a structural estimation approach. They assume that callers waiting in the line make abandon or continue-to-wait decisions at the beginning of discrete time periods. In comparison, the consumers in our model make join and balk decisions, or can defer their decision-making to the following period by exercising the option to retry.

To summarize, we propose a new model for rational decisions of consumer retrials in queues. We study the welfare effects of retrial decisions as our focus. The main contribution of the paper is to fill the research gap in the existing service operations literature which has typically either omitted retrial decisions or specified it exogenously.

2. A Model of Retrials: Base Model

To introduce our base model, we first consider an observable $M/M/1$ queue as in Naor (1969). Consumers arrive to the queue with a single server according to a Poisson process. In Naor's model, consumers observe the state of the queue and make join or balk decisions, based on the service value and the cost of waiting in the queue. They join if the queue length is below a certain threshold (typically addressed as the balking threshold), and balk from the queue otherwise. In Naor's model, balking consumers do not return.

2.1. Consumer Arrivals and Service Provision

We consider a model in which all arriving consumers can choose to retry later, in addition to the options to join and balk. For this purpose, we consider an infinite horizon model discretized into time periods. In each period, all arriving customers observe the length of the queue, and make a decision whether to join the queue immediately, or to balk from the queue (and not return), or decide to come back at retrying the queue in the next period. A "period" is symbolic for the time between retrial attempts, and its length differs by service settings. For instance, a consumer who

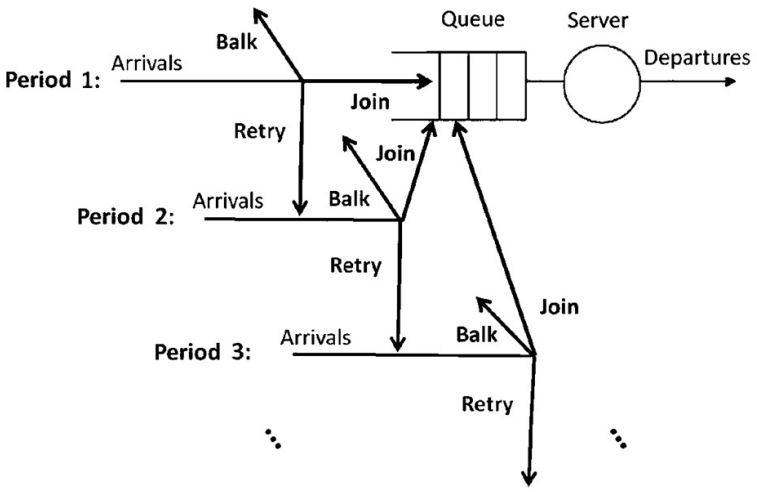


Figure 1 **Illustration of the join, balk versus retrial decisions in our model. Note that retrial is simply a "deferral" action. Eventually, every consumer has to either join the queue or balk from it.**

finds a long line at the post office in the morning will likely choose to retry much later, either in the afternoon or during the next day. A period is equivalent to half a day or a day in this context.

At any given period, the arriving population would consist of consumers who have arrived at the queue for the first time (who shall be henceforth referred to as *new* consumers) and those consumers who are retrying the service queue again (referred to simply as *old* consumers). It is likely that a fraction of the consumer population has retried multiple times in the past. A schematic representation of retrials is provided in Figure 1.

We assume that new consumers arrive to the queue according to a Poisson process with rate λ. In addition to the new consumers, there could be other old consumers re-attempting to join the queue, having decided to retry in the past. The service provider processes consumers waiting in the queue at a rate of μ per unit time and the service times are independent and exponentially distributed. We assume, as is done typically, that the values of λ and μ are common knowledge. Table 1 displays a list of notations that will be used throughout the paper.

Upon joining the queue, all consumers are served according to *first come, first served* (FCFS) discipline. We assume that the new arrivals do not exceed the system capacity, i.e., $\lambda < \mu$. Thus, the workload at the server due to new arrivals, $l \triangleq \lambda/\mu$, is bounded by 1. However, the total arrivals to the system in any period might exceed the system capacity, since the arrivals also include the consumers retrying from past attempts. While the new arrivals are exogenous in the model, old and thus total arrivals are endogenously determined, based on the consumer decisions (to retry) in the past.

		Queue Parameters
λ	$\triangleq$	Arrival rate of new consumers
μ	$\triangleq$	Service rate
l	$\triangleq$	λ/μ, work load due to new arrivals
c	$\triangleq$	Waiting cost per unit of time
v	$\triangleq$	Value of the service (net of price)
$\mathbb{N}_0$	$\triangleq$	$\{0,1,2,\ldots\}$, state space of the queue
n	$\in$	$\mathbb{N}_0$, variable for queue state / length
N	$\triangleq$	$\lfloor v/\frac{c}{\mu} \rfloor$, (equilibrium) balking threshold in Naor (1969)
N'	$\triangleq$	Socially optimal balking threshold in Naor (1969)
		Key Input Variables
α	$\triangleq$	Retrial cost
θ	$\triangleq$	Portion of critical arrivals in the population
		Strategy
$\mathcal{A}$	$\triangleq$	$\{J,R,B\}$, action set including "Join", "Retry" and "Balk" actions
σ	:	$\mathbb{N}_0 \to [0,1]\times[0,1]\times[0,1]$, variable for a strategy
π_n^σ	$\triangleq$	Steady-state probability for state $n \in \mathbb{N}_0$ under strategy σ
$\pi_{J/R/B}^\sigma$	$\triangleq$	Long-run probability that arrivals join the queue / retry / balk under strategy σ
π_{total}^σ	$\triangleq$	$\lambda/(1-\pi_R^\sigma)$, total arrival rate under strategy σ
ρ^σ	$\triangleq$	$\pi_{total}^\sigma/\mu = l/(1-\pi_R^\sigma)$, traffic intensity under strategy σ
W^σ	$\triangleq$	Expected waiting time conditional on joining the queue under strategy σ
L^σ	$\triangleq$	The average number of consumers in the system under strategy σ
JnR	$\triangleq$	A retry strategy under which consumers join the queue at states $\{0,1,2,\ldots,n-1\}$, and retry at state n or higher.
$J_{R(1-\beta)}^{J(\beta)}nR$	$\triangleq$	A join/retry strategy under which consumers join the queue at states $\{0,1,2,\ldots,n-2\}$, mix join and retry decisions at state $n-1$ with probabilities β and $1-\beta$, and retry at state n or higher.
JnB	$\triangleq$	A balk strategy under which consumers join the queue at states $\{0,1,2,\ldots,n-1\}$, and balk at state n or higher.
$Jn_{B(\gamma)}^{R(1-\gamma)}$	$\triangleq$	A retry/balk strategy under which consumers join the queue at states $\{0,1,2,\ldots,n-1\}$, and mix retry and balk decisions with probabilities $1-\gamma$ and γ at state n or higher.
		Welfare
$\chi(\alpha)$	$\triangleq$	Consumer welfare as a function of the underlying retrial cost α (when $\theta = 0$).
$\mathcal{U}_{\alpha,\theta}$	$\triangleq$	Consumer welfare *per capita* for regular arrivals under θ and α.
$\mathcal{V}_{\alpha,\theta}$	$\triangleq$	Consumer welfare *per capita* for critical arrivals under θ and α.

Table 1 Table of Notation

2.2. Consumer Actions

Every consumer arriving to the system observes the queue length, which we refer to as the state of the queue. Let $\mathbb{N}_0 \triangleq \{0, 1, 2, \ldots\}$ denote the state space. On arriving to some state n, a consumer has three choices of actions: Join the queue, Retry or Balk. We define the action set $\mathcal{A} \triangleq \{J, R, B\}$. Every *rational* consumer chooses an action that maximizes his long-run risk neutral expected utility. We explore the different actions below.

Join: Joining consumers wait in the queue and incur a cost of c per unit of time while waiting in line. After service completion, they receive a value v (net of price) from the service. We do not consider pricing decisions in the model. If there are n people ahead in the queue when a customer joins, he receives an expected net value of $v - \frac{c}{\mu}(n+1)$, on the completion of his service. We assume that there is no reneging or abandonment.

For ease of exposition, we assume that (i) $v > \frac{c}{\mu}$, i.e., consumers will join the server if it is idle (otherwise they would not be interested in consuming the service); (ii) v is not a multiple of $\frac{c}{\mu}$, i.e., the payoff from a join decision, $v - \frac{c}{\mu}(n+1)$ $\forall n \in \mathbb{N}_0$, is either positive or negative. When the retry option is not allowed (i.e., in the model of Naor (1969)), this specification implies that there exists a unique balking threshold N such that $v - \frac{c}{\mu}(N+1) < 0 < v - \frac{c}{\mu}N$.

Formally, we define this *Naor's balking threshold* as

$$N \triangleq \lceil v/\frac{c}{\mu} \rceil - 1 = \lfloor v/\frac{c}{\mu} \rfloor. \tag{1}$$

Balk: We assume that consumers who balk do not receive the service nor do they incur any waiting cost. Without loss of generality, we normalize the payoff for the balking consumers to zero.

Retry: In addition to joining the queue or balking from it, a consumer may also retry: He leaves the queue without waiting any further but will return to the server during the next period. When he returns in the next period, this consumer may again decide to join, retry, or balk from the queue, on observing the queue. We can extend our model to randomly distributed retrial intervals – a consumer may retry after a random number of periods, or after a random time length. These alternative model specifications are studied in Appendix B.

When a consumer chooses to retry, he suffers no waiting cost even as he waits in an off-line queue, as in Cachon and Feldman (2011). However, retrial attempts are costly. Each retrial attempt creates some "hassle" cost to the consumer. We denote this hassle cost as "α" which is incurred on *every* retrial attempt. This retrial cost could come from the transportation cost from consumer's work/home to the service center, or a toll to re-enter the service system, or the penalty cost for rescheduling the visit, or the opportunity cost for the time spent on the trip back and forth. We

assume that consumers are homogeneous in their waiting cost c, their retrial cost α, and in their value from the service v. In Section 5, we will consider heterogeneous consumer classes.

Consumers can retry as often as they want but have to pay hassle cost α for every retrial attempt. Consumers are rational and forward-looking, i.e., in each period they compare the expected payoffs from join, balk and retry decisions, and then choose a state-dependent action that maximizes their expected long-run payoff. Retrial costs incurred before arrival to a service are irrelevant to the decision that has to be made on arrival, i.e., retrial costs are sunk.

Finally, we assume that periods are much longer than mean service times. This assumption is identical to the frequency of repeat users in queues in prior research. For example, consider the subscription buyers who use a service multiple times in Cachon and Feldman (2011), and the multiple trip users who engage in hyperbolic discounting in Plambeck and Wang (2013). Consistent with these papers, in our model consumers who retry and come back in the following period expect to see a queue length that is independent from the current state of the queue.[1] Thus, the state they observe in the next period will be a draw (independent of the current observed states) from the distribution of queue lengths. Consider again the parcel pick-up example - a single service duration (typically, few minutes for a customer to get his parcel) is relatively short compared to the time between retrials (half a day, or a day). Hence, up on his return to the queue the customer who is retrying would see a *fresh* realization from the queue length distribution.

2.3. Consumer Strategies

A consumer's strategy on a particular service occasion specifies a probability distribution over the action space $\mathcal{A}$ for any state $n \in \mathbb{N}_0$. That is, σ is a strategy if $\sigma : \mathbb{N}_0 \to [0,1] \times [0,1] \times [0,1]$ creates a vector, on any $n \in \mathbb{N}_0$, $\sigma(n) \triangleq [\sigma_J(n) \;\; \sigma_R(n) \;\; \sigma_B(n)]^T$ such that $\sigma_J(n) + \sigma_R(n) + \sigma_B(n) = 1$. In words, a consumer who follows strategy σ arriving at the system to observe state $n \in \mathbb{N}_0$ will join the queue with probability $\sigma_J(n)$, retry with probability $\sigma_R(n)$ and balk with probability $\sigma_B(n)$.

A strategy σ is a pure strategy if $\sigma_a(n)$ is an integer $\forall a \in \mathcal{A}, \forall n \in \mathbb{N}_0$. Otherwise, σ is a mixed strategy. For example, the balking threshold strategy being used in Naor (1969) is a pure strategy where $\sigma_J(n) = 1$ for $n < N$, and $\sigma_B(n) = 1$ for $n \geq N$.

In Naor (1969), consumers also only have the options to join or to balk. The expected payoff from a join decision "$v - \frac{c}{\mu}(n+1)$" or a balk decision "0" is perfectly determined by the current

[1] It is well known that a queue converges to its steady-state characteristics, independent of the system's initial state, in a roughly exponential manner, e.g., see Morse (1955), Abate and Whitt (1988), etc. Therefore, the auto-correlation between queue lengths at two time points in a stationary $M/M/1$ queue converges exponentially to zero as the time interval in-between increases. When a period is much longer than a service cycle, queue lengths observed at two time points in two different periods can be assumed to be independent of each other. Also see Odoni and Roth (1983) for an upper bound on the time it takes for the queue length in a $M/M/1$ queue to reach steady state.

state of the queue n, but not influenced by decisions made by future customers. In contrast, in our model consumers are allowed to retry, and they choose an action to maximize their long-run expected payoff. Since the (long-run) expected payoff from a retry decision would depend on the decisions or strategies chosen by other consumers, we shall pursue an equilibrium analysis.[2]

As all consumers in our model are ex-ante symmetric, we consider symmetric equilibria under which all consumers adopt identical strategies on every service occasion. (We will relax this consideration in Section 5.) Nevertheless, consumers may arrive at different states of queue lengths and end up choosing differing actions. We define the equilibrium strategy below.

Definition 1. A (mixed) strategy σ is a symmetric equilibrium strategy when all consumers adopt σ, and no consumer can strictly improve his expected payoff by unilaterally deviating from σ along any equilibrium path that occurs with a positive probability.

Since our game is an infinite player game on states that evolve according to a Markovian process, the appropriate equilibrium solution concept is *Markov Perfect Equilibrium* due to Maskin and Tirole (2001) which specifies equilibrium actions for all states that are positive recurrent. A Markov Perfect Equilibrium is a Nash Equilibrium. Essentially, given a symmetric equilibrium strategy σ in our model, all the states at or above $\min\{n \in \mathbb{N}_0 : \sigma_B(n) + \sigma_R(n) = 1\}$ are eventually transient, and have zero probability measure.

The long-run effective workload of the system when the population adopts some strategy σ is bounded above by l which is less than 1. Thus, there exists a stationary probability distribution of the underlying birth and death process. For a given strategy σ, let us define π_n^σ to be the long-run probability that the queueing system is in state $n \in \mathbb{N}_0$, and use $\pi_J^\sigma, \pi_R^\sigma$ and π_B^σ, to represent the unconditional probabilities that a customer arriving to the queue will choose to join, retry or balk respectively, when the population adopts the strategy σ (thereafter, "under" strategy σ). Then, the rate for total arrivals in any period is given by

$$\lambda_{total}^\sigma \triangleq \lambda + \lambda\pi_R^\sigma + \lambda(\pi_R^\sigma)^2 + \ldots = \frac{\lambda}{1 - \pi_R^\sigma}. \tag{2}$$

To understand (2), we illustrate the quantity of this period's total arrival rate. Total arrivals in this period include new arrivals (i.e., λ) plus the old arrivals who came to the service provider for the first time in the previous period, but decided to retry according to σ (i.e., $\lambda\pi_R^\sigma$), plus the old arrivals who came the service provider two periods back and have retried twice in a row (i.e., $\lambda(\pi_R^\sigma)^2$), and old arrivals who have retried three times, four times, and so on.

[2] For example, if a consumer were to make a retry decision during the current period, his long-run expected payoff from this decision would depend on the arrivals in the future periods. And future arrivals are endogenously determined by retrial decisions made by consumers who show up in the current or subsequent periods.

When a period is sufficiently long, we assume, according to Lariviere and Van Mieghem (2004), that the arrival times for customers who are entering or retrying the queue for the k-th time ($\forall k = 1, 2, \ldots, \infty$) and who can strategically choose a time point to return during a period, are governed by some renewal process. The total arrival process, which is a superposition of infinitely many renewal processes, is then of the Poisson type (see Feller (1971), pp. 370-371, and Albin (1982)). Thus, we model the total arrivals including new and all old consumers as a Poisson process with rate λ^{σ}_{total} defined as in (2).

Using PASTA property, we then have

$$\pi^{\sigma}_J = \sum_{n\in\mathbb{N}_0} \pi^{\sigma}_n \cdot \sigma_J(n); \quad \pi^{\sigma}_R = \sum_{n\in\mathbb{N}_0} \pi^{\sigma}_n \cdot \sigma_R(n); \quad \pi^{\sigma}_B = \sum_{n\in\mathbb{N}_0} \pi^{\sigma}_n \cdot \sigma_B(n). \tag{3}$$

And it is clear from the definition of a strategy together with (3) that for any σ, $\pi^{\sigma}_J + \pi^{\sigma}_R + \pi^{\sigma}_B = 1$.

Finally, we denote the traffic intensity of the system under σ by

$$\rho^{\sigma} \triangleq \frac{\lambda^{\sigma}_{total}}{\mu} = \frac{\lambda/(1-\pi^{\sigma}_R)}{\mu} = \frac{l}{1-\pi^{\sigma}_R}. \tag{4}$$

Note that although the workload due to new arrivals $l < 1$, the total traffic intensity ρ can be less than, equal to, or greater than 1, which depends on the underlying strategy σ being adopted by the population.

2.4. Consumer Best Response

Now we can consider the best response strategy of a consumer j. Fix the strategy of all other consumers $i \neq j$ at σ on every service occasion.

The consumer j can retry repeatedly but has to pay for the retrial cost, namely α, every time he retries. Since consumers in our model do not suffer from sunk cost fallacy, when the consumer j makes a retry versus a join or balk decision upon a particular service occasion, only the future retrial and expected waiting costs are in consideration. Therefore, if a strategy is best for him for one service occasion in response to the population strategy σ, it will be so for all service occasions (regardless of the number of times he has visited before).

Thus, let us suppose that the consumer j adopts some strategy σ^j on every service occasion. We can now consider the conditional payoffs of this consumer arriving to the queue and observing a state n. As described earlier, the (expected) payoff for consumer j joining at a state n is the value of the service net of waiting costs, i.e., $v - \frac{c}{\mu}(n+1)$. On the other hand, the payoff for consumer j balking at any state n is 0.

Now consider the payoffs for the consumer j from choosing to retry at state n (and return during the subsequent period with a retrial cost α). In the next period, he may join the queue, balk from it, or retry again, based on the state of the queue he observes. Since all other consumers follow

strategy σ, by PASTA property, the consumer j's (unconditional) probability of joining, retrying and balking in the next period according to the strategy σ' are, respectively,

$$\pi_J^{\sigma,\sigma^j} \triangleq \sum_{n\in\mathbb{N}_0} \pi_n^{\sigma}\cdot\sigma_J^j(n);\;\; \pi_R^{\sigma,\sigma^j} \triangleq \sum_{n\in\mathbb{N}_0} \pi_n^{\sigma}\cdot\sigma_R^j(n);\;\; \pi_B^{\sigma,\sigma^j} \triangleq \sum_{n\in\mathbb{N}_0} \pi_n^{\sigma}\cdot\sigma_B^j(n). \tag{5}$$

If the consumer j retries again next period, the same decision process plays out, and he faces the same steady-state probabilities in the period after, and so on. Note that the customer j eventually balks or joins the queue.

Let W^σ denote the expected waiting time for consumer j conditional on joining the queue in a period when the system is under σ. Then his expected long-run payoff from the retry decision at state n in the current period is given by

$$\begin{aligned}
&\pi_J^{\sigma,\sigma'}\cdot(v-cW^\sigma-\alpha)+\pi_B^{\sigma,\sigma'}\cdot(0-\alpha) && \text{(Joins/Balks in the next period)}\\
&+\pi_R^{\sigma,\sigma'}\cdot[\pi_J^{\sigma,\sigma'}\cdot(v-cW^\sigma-2\alpha)+\pi_B^{\sigma,\sigma'}\cdot(0-2\alpha)] && \text{(Joins/Balks in 2 periods)}\\
&+(\pi_R^{\sigma,\sigma'})^2\cdot[\pi_J^{\sigma,\sigma'}\cdot(v-cW^\sigma-3\alpha)+\pi_B^{\sigma,\sigma'}\cdot(0-3\alpha)]+\dots && \text{(Joins/Balks in 3 periods, }\dots\text{)}\\
=&\,\pi_J^{\sigma,\sigma'}(v-cW^\sigma)+\pi_B^{\sigma,\sigma'}\cdot 0-(1-\pi_R^{\sigma,\sigma'})\alpha\\
&+\pi_R^{\sigma,\sigma'}\pi_J^{\sigma,\sigma'}(v-cW^\sigma)+\pi_R^{\sigma,\sigma'}\pi_B^{\sigma,\sigma'}\cdot 0-\pi_R^{\sigma,\sigma'}(1-\pi_R^{\sigma,\sigma'})2\alpha\\
&+(\pi_R^{\sigma,\sigma'})^2\pi_J^{\sigma,\sigma'}(v-cW^\sigma)+(\pi_R^{\sigma,\sigma'})^2\pi_B^{\sigma,\sigma'}\cdot 0-(\pi_R^{\sigma,\sigma'})^2(1-\pi_R^{\sigma,\sigma'})3\alpha+\dots\\
=&\frac{\pi_J^{\sigma,\sigma'}}{1-\pi_R^{\sigma,\sigma'}}(v-cW^\sigma)+\cancel{\frac{\pi_B^{\sigma,\sigma'}}{1-\pi_R^{\sigma,\sigma'}}\cdot 0}-\frac{1}{1-\pi_R^{\sigma,\sigma'}}\cdot\alpha.
\end{aligned} \tag{6}$$

We keep the cancelled term in (6) to better interpret the expected retrial payoff. Recall that retrial is only a deferral option, and eventually consumer j completes his "task" by either joining the queue or balking. The first term in (6) can be interpreted as the probability that consumer j completes his task by joining the queue (at some period) according to σ^j, times the conditional expected payoff that he will receive upon joining. Similarly, the second term in (6) can be interpreted as the probability that consumer j ends his task by balking (at some point), times the conditional expected payoff upon balking. The last term is interpreted as the expected number of periods it takes for consumer j to finish the task (by either joining or balking), times the retrial cost. Therefore, we can describe the expected long-run payoff from a retry decision for consumer j at state n, given by the expression (6), as the total expected end-of-task payoff less the total expected retrial cost incurred during the process.

We observe that the retrial payoff of the consumer j at state n given by (6) does not actually depend on n, and this is because the current state of the queue is independent of the queue length realization in the next period should consumer j choose to retry. Therefore, retrial payoff of a consumer does not dependent on the state of arrival n. But unlike the joining or the balking payoff, the retrial payoff depends on the underlying population strategy σ. In other words, actions of other consumers may make the retry option more or less attractive to consumer j.

Suppose σ^j is the best strategy of the consumer j in response to all other consumers adopting σ, then for every state $n \in \mathbb{N}_0$, σ^j must specify a decision or decisions that maximize the expected payoff for consumer j among (i) the balking payoff "0", (ii) the joining payoff "$v - \frac{c}{\mu}(n+1)$", and (iii) the retrial payoff given by (6).

At a symmetric equilibrium, we set σ^j to σ, and σ must belong in the best response strategy set in response to itself. It follows from (3) and (5) that when $\sigma^j = \sigma$, we have $\pi_J^{\sigma,\sigma^j} = \pi_J^\sigma$, $\pi_R^{\sigma,\sigma^j} = \pi_R^\sigma$ and $\pi_B^{\sigma,\sigma^j} = \pi_B^\sigma$, and (6) is updated. We then have the following proposition, that characterizes the conditions for a strategy to be an equilibrium.

PROPOSITION 1. *σ is an equilibrium strategy, if and only if, for all $n \in \mathbb{N}_0$,*

$$\sigma_B(n) > 0 \Rightarrow 0 \geq \max\{v - \frac{c}{\mu}(n+1), \frac{\pi_J^\sigma}{1-\pi_R^\sigma}(v - cW^\sigma) - \frac{\alpha}{1-\pi_R^\sigma}\}; \quad \text{(BALK)}$$

$$\sigma_J(n) > 0 \Rightarrow v - \frac{c}{\mu}(n+1) \geq \max\{0, \frac{\pi_J^\sigma}{1-\pi_R^\sigma}(v - cW^\sigma) - \frac{\alpha}{1-\pi_R^\sigma}\}; \quad \text{(JOIN)}$$

$$\sigma_R(n) > 0 \Rightarrow \frac{\pi_J^\sigma}{1-\pi_R^\sigma}(v - cW^\sigma) - \frac{\alpha}{1-\pi_R^\sigma} \geq \max\{0, v - \frac{c}{\mu}(n+1)\}. \quad \text{(RETRY)}$$

Now that we have introduced the model and the equilibrium concept, we are ready to study consumers' equilibrium behavior and its impact on consumer welfare. In Sections 3 and 4, we assume that the population consists of all regular consumers who can strategically choose to join the queue, retry or balk. Then, in Section 5, we extend our findings to two classes of consumers which include both regular and critical arrivals.

3. Equilibrium Strategies

Let the population be entirely made up of strategic consumers, i.e., every new and old consumer makes rational state-dependent join, balk and retry decisions, upon arrival to the system.

In (6), we showed that the payoff from a retry decision depends on the strategy being adopted by the population. However, since the queue length one sees in the current period is of no use in predicting what the queue realization will be during the following period, the payoff from a retry decision does not depend on the state at which the decision is being made in the current period.[3] This fact implies that when the population adopts any fixed strategy, the expected payoff from any retry decision is a fixed value (which is illustrated in Table 2).

Now suppose the system reaches an equilibrium when the population adopts some strategy σ. Proposition 1 indicates that, for every state $n \in \mathbb{N}_0$, σ must specify the payoff-maximizing

[3] For example, suppose that the population adopts a strategy which specifies that one should retry at seeing state $n = 3$ or $n = 8$ upon arrival. Then the payoff one expects to receive from a retry decision made at state $n = 3$ or made at state $n = 8$ will be identical, because when this consumer returns, he will observe the same stationary process and make future decisions according to the same strategy.

State | Action / $\overline{Payoffs}$	*Balk* under any strategy	*Join* under any strategy	*Retry* under some σ^1	*Retry* under some σ^2	*Retry* under some σ^3	*Retry* under some σ^4
n=0	0	55	40	25	-8	0
n=1	0	45	40	25	-8	0
n=2	0	35	40	25	-8	0
n=3	0	25	40	25	-8	0
n=4	0	15	40	25	-8	0
n=5	0	5	40	25	-8	0
n=6	0	-5	40	25	-8	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2 Illustration of the possible types of an equilibrium strategy (with $v = 65$ and $\frac{c}{\mu} = 10$).

decision(s). Since the joining payoff "$v - \frac{c}{\mu}(n+1)$" is linearly decreasing in state $n \in \mathbb{N}_0$, the balking payoff "0" is constant, and the retrial payoff is also constant for all states $n \in \mathbb{N}_0$, the equilibrium strategy σ must be one where consumers join the queue up to some state after which the joining payoff drops below the balking or the retrial payoff. As a result, only the following four types of threshold strategies can qualify for an equilibrium strategy.[4]

- Threshold *retry strategy*, denoted by "JnR" (join up to n then retry): Consumers join the queue at states $\{0, 1, 2, \ldots, n-1\}$ and retry at state n.
- Threshold *join/retry strategy*, denoted by "$J^{J(\beta)}_{R(1-\beta)}nR$": Consumers join the queue at states $\{0, 1, 2, \ldots, n-2\}$, mix join and retry decisions at state $n-1$, and retry at state n. $\beta \in [0, 1]$ denotes the probability of a join decision being made at state $n-1$, and $1-\beta$ a retry decision.
- Threshold *balk strategy*, denoted by "JnB" (join up to n then balk): Consumers join the queue at states $\{0, 1, 2, \ldots, n-1\}$ and balk at state n.[5]
- Threshold *retry/balk strategy*, denoted by "$Jn^{R(1-\gamma)}_{B(\gamma)}$": Consumers join the queue at states $\{0, 1, 2, \ldots, n-1\}$, and mix retry and balk decisions at state n. $\gamma \in [0, 1]$ denotes the probability of a balk decision being made at state n, and $1-\gamma$ a retry decision.

Note that we have specified the four types of strategies above at only states 0, 1, 2 up to some n, and this is because, when the entire population adopts any of these strategies, the resulting queueing system is $M/M/1/n$ (i.e., no state higher than n will be ever reached).

We illustrate all the possible types of an equilibrium strategy in Table 2, using values $v = 65$ and $c/\mu = 10$. The balking and joining payoffs depend only on queue parameters and the underlying state of the system, but not the strategy being adopted by the population. We display them as a function of the state n in the second and third columns of the table, namely "0" and "$n - \frac{c}{\mu}(n+1)$".

[4] By our assumption made before, the joining payoff can only assume strictly positive or negative values. This assumption eliminates some trivial equilibrium strategy types that mix between join and balk decisions.

[5] This is the type of strategy studied in Naor (1969).

Then, four *imaginary* representative strategies, σ^1, σ^2, σ^3 and σ^4, are being considered in the fourth through the eighth column, such that when the population adopts them, the retrial payoffs, given by (6), are equal to 40, 25, -8 and 0, respectively. Consistent with Proposition 1, we see that σ^1 must be a retry strategy type (i.e., $J2R$) to possibly be an equilibrium strategy, while σ^2 has to be a join/retry strategy type (i.e., $J_R^J 4R$), σ^3 a balk strategy type (i.e., $J6B$) and σ^4 a retry/balk strategy type (i.e., $J6_B^R$).

The following theorem further reduces the number of possible equilibrium strategies by imposing bounds on the threshold queue lengths. Specifically, it shows that the threshold of a retry or a join/retry strategy must be smaller than or equal to Naor's threshold, N. On the other hand, the threshold of a balk or a retry/balk threshold coincides with Naor's threshold.

THEOREM 1. *If a strategy JnR or $J_R^J nR$ is an equilibrium, then we must have $1 \le n \le N$. On the other hand, if a strategy JnB or Jn_B^R is an equilibrium, then we must have $n = N$.*

According to Theorem 1, an equilibrium strategy can only be JnR for $n \le N$, $J_R^J nR$ for $n \le N$, JNB, or JN_B^R. In the next theorem, we identify *all* equilibrium strategies as a function of the retrial cost α, and specify the unique Pareto-dominant equilibrium when multiple equilibria exist.

THEOREM 2. *For a given system (i.e., fixed λ, μ, v, c), define $\alpha_L \triangleq (1-\pi_N^{JNR})(v - cW^{JNR})$ and $\alpha_H \triangleq (1-\pi_N^{JNB})(v - cW^{JNB})$. Then, $0 < \alpha_L < \alpha_H < v$. Furthermore,*

(i) For $\alpha \le \alpha_L$, there exists one or multiple equilibrium strategies, either of the retry type $\{JnR\}_{n=1,2,\ldots,N}$ or the join/retry type $\{J_R^J nR\}_{n=1,2,\ldots,N}$, but Pareto-dominant equilibrium is unique.

Specifically, for α that falls in one of the N intervals denoted by $\{I_n\}_{n=1,2,\ldots,N}$ where $I_1 \triangleq (0, \alpha_1], I_2 \triangleq (\alpha_1, \alpha_2], \ldots, I_{N-1} \triangleq (\alpha_{N-2}, \alpha_{N-1}]$, $I_N \triangleq (\alpha_{N-1}, \alpha_L]$ for a sequence of increasing thresholds $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_{N-1} < \alpha_N$ determined by λ, μ, v, c, the corresponding retry strategy $\{JnR\}_{n=1,2,\ldots,N}$ is the unique Pareto-dominant equilibrium.

(ii) For $\alpha > \alpha_H$, Naor's strategy, i.e., JNB, is the unique equilibrium strategy.

(iii) For $\alpha \in [\alpha_L, \alpha_H]$, there exists a unique equilibrium strategy in the retry/balk form $JN_{B(\gamma)}^{R(1-\gamma)}$ for some $\gamma \in [0,1]$. Moreover, γ increases in α continuously from 0 to 1.[6]

The establishment of Theorem 2 requires a careful search for all possible four types of equilibrium strategies stated before. The complete proof consists of 10 supporting lemmas from Lemma 6 to Lemma 16, all stated and proved in Appendix A. Theorem 2 completely answers our first

[6] The bijection is given by $\alpha = (1 - \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}})(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}})$. When $\alpha = \alpha_L$, the equilibrium is given by $JN_{B(0)}^{R(1)}$ or simply the retry strategy JNR and when $\alpha = \alpha_H$, the equilibrium is given by $JN_{B(1)}^{R(0)}$ or the balk strategy JNB. Therefore, conclusions reached in part (iii) of the theorem are consistent with those under part (i) and part (ii).

important research questions, i.e., how strategic consumers make join, balk, versus retry decisions at equilibrium. It specifies the following outcomes:

Low Retrial Cost Region: When the retrial cost is low, i.e., α is smaller than some threshold α_L, any equilibrium strategy is of the retry type or the join/retry type – an arriving consumer would join the queue if it is short and retry if it is long. (The retrial cost is low enough such that a consumer would not consider a balk decision to leave the service value on the table when the queue is long.) As α increases from 0 to α_L, the Pareto-dominate equilibria are given by retry strategies, JnR, where the threshold n increases from 1 to N. In other words, when retrials become more costly, consumers would retry only upon seeing longer queues.

Let us apply the result to an example. Suppose a customer would like to pick up a parcel that is not time urgent, and he can always go to the post office that holds the parcel after his work. Imagine the first scenario that this customer walks past the post office every evening on his way home, then the additional cost for him to retry will be almost negligible, i.e., $\alpha \to 0$. Knowing that there will be no line at the post office at some point (because the work load $l < 1$), this consumer should adopt the $J1R$ strategy. That is, he will only join the queue if the server is idle. Even if there is only one other consumer in the system, it would be better for him to retry rather than joining, because retrials are free in this case, and he can encounter an idle server in the future. Now imagine another scenario where getting to the post office requires some detour. Then, Theorem 2 states that the higher the retrial cost α is (within the region that $\alpha \leq \alpha_L$), the longer queue this customer is willing to tolerate/join upon arrival, as opposed to retrying.

When it is more costly to retry (i.e., higher α), not only does the equilibrium retry strategy have a higher threshold, but consumers are also more reluctant to make retry decisions (i.e., lower retrial probability π_R and lower total traffic $\rho = l/1 - \pi_R$). On the other hand, when it is less costly to retry (i.e., lower α), we observe lower threshold for the equilibrium retry strategy and more retrials (i.e., higher π_R and higher ρ).

High Retrial Cost Region: When the retrial cost $\alpha \geq \alpha_H$, the balking strategy JNB (or Naor's strategy) is found to be the unique equilibrium strategy. Simply speaking, when a retrial attempt becomes very costly, welfare-maximizing consumers no longer consider the retry option.

For example, when retrials are substantially more costly than the net gain of the service (i.e., $\alpha >> v$), the retry decision cannot be optimal – the expected payoff from retrials is always negative, and the retry decision is thus dominated by balking). In fact, a retrial cost α can be even less than v when retrials are already not worthwhile, because a consumer also incurs some waiting cost when he returns from a retrial to join the queue in the future. As a result, when the retrial

cost α is greater than some threshold α_H (where $\alpha_H < v$), the retrial system would induce the same equilibrium as in Naor's model where only join and balk decisions are allowed. In this case, consumers join the queue when the it is short and balk when the queue is long.

Moderate Retrial Cost Region: Finally, for any retrial cost $\alpha \in (\alpha_L, \alpha_H)$, we find that a unique equilibrium retry/balk strategy (JN_B^R) exists. That is, when the queue is short, consumers still join the queue for service. However, when the queue is long, they choose between the retry and the balk options. Specifically, as the retrial cost increases on the region, consumers decide to retry less frequently and balk more frequently (when seeing the long queue). To provide intuition that this mixed type of strategy can form an equilibrium when $\alpha \in (\alpha_L, \alpha_H)$, let us consider the retrial cost at $\alpha_L + \Delta\alpha$ where $\Delta\alpha$ is infinitesimal.

Recall from part (i) of the theorem that, when the retrial cost is α_L, the population adopts a retry strategy with threshold N (i.e., JNR) at equilibrium. In the proof, it is shown that the expected payoff from a retry decision (at any state) in this case is actually equal to zero. At $\alpha_L + \Delta\alpha$, if everybody still sticks to the same strategy, the expected payoff from a retry decision becomes $-\Delta\alpha$, which makes retrials less favorable than balking. Thus, every individual has an incentive to balk rather than retrying at state N, and JNR can no longer be an equilibrium strategy.

On the other hand, when the retrial cost $\alpha_L \to \alpha_L + \Delta\alpha$, if every consumer switches to the balk strategy (i.e., JNB), then the expected waiting cost (compared to that under the equilibrium strategy JNR when the retrial cost is α_L) would suddenly drop because the retrial population all disappears at once, more than enough to cover the infinitesimal increased retrial fee ($\Delta\alpha$) to lead to a positive return for a retry option. As a result, everyone has an incentive to retry at seeing state N rather than balking, so the balk strategy JNB is also not an equilibrium strategy at $\alpha_L + \Delta\alpha$.

An equilibrium will be reached, however, if at seeing state N, some consumers choose to retry while others choose to balk such that a retry decision (at state N and all other states), with the new retrial cost $\alpha_L + \Delta\alpha$, again generates an expected payoff of zero. This leaves both consumers who retry and who balk at state N no incentive to deviate unilaterally.

4. Consumer Welfare Analysis

Throughout Section 3, we have examined all possible equilibrium strategies for our service system with rational retrials. In this section, we turn to study consumer welfare under the (Pareto-dominant) equilibrium strategies. Since different types of equilibrium strategies are formed on the regions $\alpha \leq \alpha_L$, $\alpha \in [\alpha_L, \alpha_H]$ and $\alpha \geq \alpha_H$, we will analyze consumer welfare *region by region*. Denote the consumer welfare function in α as $\chi(\alpha)$.

Recall that when $\alpha \leq \alpha_L$, the Pareto-dominant equilibrium strategy is given by JnR for some n. Under JnR, the total arrival rate given by (2) is $\frac{\lambda}{1-\pi_R^\sigma} = \frac{\lambda}{1-\pi_n^{JnR}}$, but the system's true workload, $\rho^{JnR}\pi_J^{JnR} = \rho^{JnR}(1-\pi_n^{JnR}) = \frac{l}{1-\pi_n^{JnR}}(1-\pi_n^{JnR})$, is always equal to l regardless of n, i.e., the effective joining rate is λ. This is because under JnR, we have $\pi_B^{JnR} = 0$. Every consumer never balks, and he or she receives the service eventually through a random number of retrials. Since there is no loss of consumers, the long-run effective joining rate must equal to the new arrival rate, namely λ.

Under the strategy JnR, the total arrival rate is greater than the new arrival rate due to old consumers who are retrying (i.e., $\frac{\lambda}{1-\pi_n^{JnR}} > \lambda$), but among them only a portion of the consumers (an amount that equals λ) join the queue. The rest of them (an amount equal to $\frac{\lambda}{1-\pi_n^{JnR}} - \lambda$ or $\lambda_{total}^{JnR} \cdot \pi_n^{JnR} = \frac{\lambda}{1-\pi_n^{JnR}} \cdot \pi_n^{JnR}$) will observe state n upon arrival to the system and retry accordingly.

The overall consumer welfare is thus given by $\chi(\alpha) = \lambda v - \lambda c W^{JnR} - \lambda(\frac{1}{1-\pi_n^{JnR}} - 1)\alpha$ where the first term λv is the revenue (rate), the second term $\lambda c W^{JnR}$ is the total waiting cost (rate), and the last term $\lambda(\frac{1}{1-\pi_n^{JnR}} - 1)\alpha$ indicates the total retrial cost (rate).

Let us denote L^{JnR} the average number of consumers in the system when the population adopts JnR. Applying Little's Law to the population that joins the queue, we have $\lambda W^{JnR} = L^{JnR}$. Therefore, consumer welfare is also equal to

$$\chi(\alpha) = \lambda v - cL^{JnR} - \lambda(\frac{1}{1-\pi_n^{JnR}} - 1)\alpha \text{ (for } \alpha \leq \alpha_L), \tag{7}$$

We know from Theorem 2 that as α increases on $(0, \alpha_L]$, the threshold of the equilibrium retry strategy increases from 1 to N, and that on average fewer consumers choose to retry and more choose to join at equilibrium. As a result, system congestion goes up in α, and the extra congestion will hurt consumer welfare. To understand why congestion goes up with less retrying/more joining activities, one can think of retrials as a *smoothing* mechanism. Because consumers retry only when they see long queues, their retrial activities reduce the (steady-state) probabilities of finding higher states of the system and increase the probabilities of lower states. In other words, *retrials can generate positive externalities to other consumers.*

On the other hand, when α increases on $(0, \alpha_L]$, on average fewer consumers are paying for the retrial costs at equilibrium but each pays more. Therefore, it is not clear *ex-ante* how overall consumer welfare is affected by the amount of the retrial cost α. It turns out that congestion drives consumer welfare to decrease in α on the region $(0, \alpha_L]$, as shown in the following lemma.

LEMMA 1. *In the region $\alpha \leq \alpha_L$, consumer welfare $\chi(\alpha)$ is a decreasing left-continuous piecewise linear function, with discontinuity at the thresholds $\alpha_1, \alpha_2, \ldots, \alpha_{N-1}$. The maximum welfare, $\lambda v - cl = \lambda(v - \frac{c}{\mu})$, is achieved when $\alpha \to 0$, and the minimum welfare, $\lambda v - cL^{JNR} - \lambda(\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_L$, is achieved when $\alpha = \alpha_L$. Moreover, the linear slopes $\chi'(\alpha)$ become flatter as α increases.*

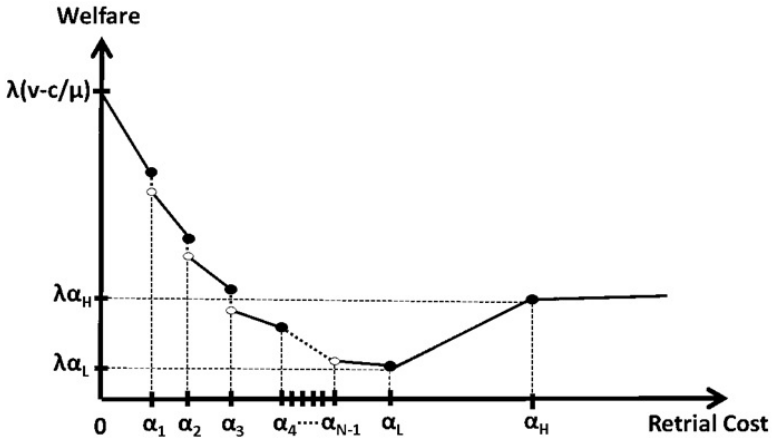


Figure 2 **Illustration of consumer welfare at equilibrium as a function of the retrial cost α over the entire region. The welfare $\chi(\alpha)$ decreases in α on $(0, \alpha_L]$, increases on $[\alpha_L, \alpha_H]$ and stays flat on $[\alpha_H, \infty)$.**

Lemma 1 states that, when the retrial cost increases on $(0, \alpha_L]$, consumer welfare drops. An illustration of the shape of the welfare curve is plotted in Figure 2. The reason we have downward-sloping linear curves for $\chi(\alpha)$ within each of the intervals $\{(0, \alpha_1], [\alpha_1, \alpha_2], \ldots, [\alpha_{N-1}, \alpha_L]\}$ is due to the facts that (i) on each individual interval, the same equilibrium holds, i.e., we observe the same system including the same amount of joining, retrial and balking activities, (ii) as α increases within any individual interval, retrial consumers incur more total retrial cost in proportion to it.

Mathematically, fixing the n-th individual interval for any $n \in \{1, 2, \ldots, N\}$, both L^{JnR} and $\lambda(\frac{1}{1-\pi_n^{JnR}} - 1)$ stay the same as α increases. So total consumer welfare, given by (7), decreases linearly in α. Moreover, the slopes get flatter as α increases, because the magnitude of the slope, defined by $\lambda(\frac{1}{1-\pi_n^{JnR}} - 1) = \rho^{JnR}\mu - \lambda$ from (7), decreases in α. In essence, fewer consumers retry as α increases, so the marginal effect of increasing α on the total consumer welfare becomes diminishing.

Although $\chi(\alpha)$ decreases over the region $\alpha \leq \alpha_L$ (i.e., less retrial hassle corresponds to higher consumer welfare), we show in the following Lemma 2 that it actually rises on $\alpha \in [\alpha_L, \alpha_H]$. To set up the result, recall from Theorem 2 that, when $\alpha \in [\alpha_L, \alpha_H]$, the equilibrium is formed under the retry/balk strategy $JN_{B(\gamma)}^{R(1-\gamma)}$ for some unique $\gamma \in [0, 1]$. Using (2) and the fact that

$$\pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}} = 1 - \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}, \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}} = (1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}, \text{ and } \pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}} = \gamma\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}},$$

the total arrival rate at equilibrium is then given by

$$\lambda_{total}^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\lambda}{1 - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}\lambda} = \frac{\lambda}{1 - (1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}. \tag{8}$$

By (8), the effective joining rate at equilibrium is

$$\lambda_{total}^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot \pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\lambda}{1 - (1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} \cdot \pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\lambda(1 - \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}})}{1 - (1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}, \tag{9}$$

and the rate at which consumers are paying for the retrial cost is equal to

$$\lambda_{total}^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\lambda}{1-(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}} \cdot \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\lambda(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}. \tag{10}$$

Consumer welfare (on $\alpha \in [\alpha_L, \alpha_H]$) is then given by total payoff less total retrial cost, so by (9) and (10), it is equal to

$$\chi(\alpha) = \frac{\lambda(1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}})}{1-(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{\lambda(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}\alpha \tag{11}$$

where α is the particular retrial cost that induces the equilibrium strategy $JN_{B(\gamma)}^{R(1-\gamma)}$.

Note that, when $\gamma = 0$, $JN_{B(0)}^{R(1)} = JNR$ and consumer welfare in (11) reduces to that in (7). Similarly, when $\gamma = 1$, $JN_{B(1)}^{R(0)} = JNB$ and consumer welfare in (11) reduces to

$$\lambda(1-\pi_N^{JNB})(v - cW_N^{JNB}) = \lambda v(1-\pi_N^{JNB}) - cL^{JNB}, \tag{12}$$

which is exactly the consumer welfare in Naor (1969).[7]

LEMMA 2. *On the region $\alpha \in [\alpha_L, \alpha_H]$, $\chi(\alpha) = \lambda\alpha$. That is, consumer welfare increases linearly in the retrial cost α with slope λ, from value $\lambda v - cL^{JNR} - \lambda(\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_L = \lambda\alpha_L$ when $\alpha = \alpha_L$ to value $\lambda v(1-\pi_N^{JNB}) - cL^{JNB} = \lambda\alpha_H$ when $\alpha = \alpha_H$.*

Lemma 2 seems counter-intuitive at first by claiming that consumer welfare as a whole increases on the region $\alpha \in [\alpha_L, \alpha_H]$ despite the increasing retrial cost, but it can be explained as follows. As the retrial cost α increases over $[\alpha_L, \alpha_H]$, consumers who arrive at state N choose to balk more frequently and retry less frequently at equilibrium (i.e., γ increases for the underlying equilibrium strategy $JN_{B(\gamma)}^{R(1-\gamma)}$). For these consumers, their expected payoffs remain the same, in fact "0" at arriving at state N. However, retrying at state N causes negative externalities (i.e., congestion) to other consumers while balking at N does not. Thus as α increases on $[\alpha_L, \alpha_H]$, there is less congestion in the system. As a consequence, the overall welfare goes up.

Note that a consumer's retry decision can impose both positive and negative externalities to others in the system. When $\alpha < \alpha_L$, consumer welfare decreases in α due to increased joining but less retrials. (Retrials, compared to joining, smooth out arrivals and reduce congestion.) When $\alpha \in [\alpha_L, \alpha_H]$ consumer welfare increases in α as a result of more balking and less retrials. (Retrials compared to balking induces congestion.)

Finally, when $\alpha \geq \alpha_H$, consumers follow the balk strategy JNB at equilibrium. There are no more retrial activities, so $\chi(\alpha)$ is a constant function. The level of the consumer welfare will be the

[7] In Naor's system, the effective joining rate is $\lambda(1-\pi_N^{JNB})$ therefore by applying Little's Law on the effective-joining population, we have $\lambda(1-\pi_N^{JNB})W^{JNB} = L^{JNB}$ which explains (12).

same as that in Naor (1969). It is then clear that the equilibrium welfare function $\chi(\alpha)$ renders a down-up-flat shape on the three regions $(0, \alpha_L]$, $[\alpha_L, \alpha_H]$ and $[\alpha_H, \infty)$, as plotted in Figure 2. However, what is lacking is the comparison among the values of $\chi(\alpha)$ as $\alpha \to 0$ versus when $\alpha = \alpha_H$. This leads to the next lemma.

LEMMA 3. *Consumer welfare $\chi(\alpha)$ is higher as $\alpha \to 0$ compared to that at $\alpha = \alpha_H$.*

Lemma 3 compares the welfare when retrials are "free" to Naor's system (i.e., when retrials are costly). Recall from Lemma 1 that consumer welfare as $\alpha \to 0$ is equal to "$\lambda v - cl = \lambda(v - \frac{c}{\mu})$". This is the largest possible consumer welfare for the population, as λ is the maximum effective joining rate, and $v - \frac{c}{\mu}$ is the largest possible welfare any individual consumer can get out from the server. Essentially, when $\alpha \to 0$, the population adopts $J1R$. Every consumer retries repeatedly (with paying zero retrial cost) until he or she sees an idle server and joins without any wait. It is thus clear that the welfare when retrials are free exceeds that when they are not free (i.e., $\alpha > 0$ or $\alpha \geq \alpha_H$). Combining Lemma 1, Lemma 2 and Lemma 3, we have the following proposition.

PROPOSITION 2. *When the retrial cost is small, the additional option to retry (comparing our model to Naor's model) improves consumer welfare at equilibrium. However, as the retrial cost increases, consumer welfare could worsen compared to welfare in Naor' model at equilibrium.*

The result is counter-intuitive because when consumers have an *additional* option, they could however be worse off. In Figure 2, this happens roughly when the retrial cost is between $\alpha = \alpha_3$ and $\alpha = \alpha_H$. This phenomenon can be explained by an argument that is similar that in Naor (1969): Consumers are self-interested. When they are presented with the additional option to retry, they will jump on it as long as exercising the option can increase their individual expected payoffs. However, at times, these extra gains in individual welfare cannot compensate for the negative externalities (i.e., congestion costs) they impose on other customers in future. Therefore, at these times overall consumer welfare would actually improve if consumers were not to have the additional retrial options.

Naturally, the next question to be asked is what would the socially optimal policy be.

4.1. Socially Optimal Policies

In queueing systems, self-interested consumers usually form an equilibrium that deviates from the socially optimal outcome. For example, Naor (1969) shows that consumers who have the options to join or balk over-congest the system when left to their devices. He then suggests that tolls or taxes can be levied to control the joining population.

Thus far we have established equilibrium strategies and welfare results for self-interested consumers with the option to retry in this paper, and we shall now consider socially optimal policies (i.e., first best solutions) within the model framework. Socially optimal policies are highly state dependent. Like Naor (1969), we will focus on policies of the threshold type. These types of policies have been used for queueing control in settings that even generalize Naor (1969), e.g., see Yechiali (1971, 1972) and other related papers surveyed in Stidham (1985). Specifically, we consider all the retry and balk strategies, i.e., the class of strategies $\{s : s = JnR \text{ or } s = JnB \text{ for some } n \geq 1\}$. We characterize the socially optimal policies (within this class) in the following theorem. Let N' be the socially optimal balking threshold in Naor (1969), and it is well-known that $N' < N$.

THEOREM 3. *For a given system λ, μ, c, v (and thus α_L and α_H), there exists a deterministic point $\alpha' \in (0, \alpha_L]$ such that the optimal policy is a retry strategy (i.e., JnR-type) with increasing thresholds for $\alpha \leq \alpha'$, and a balk strategy with threshold N' (i.e., $JN'B$) for $\alpha > \alpha'$. Moreover, for any fixed $\alpha \leq \alpha'$, the socially optimal retrial threshold is smaller than the equilibrium threshold.*

Theorem 3 provides the structure of the socially optimal policy. Comparing equilibrium outcome given in Theorem 2 to the social optimum, we find that

(i) When $\alpha \leq \alpha'$, the retrial thresholds for the equilibrium strategies are smaller than those for the socially optimal policies, and thus consumers retry too little and join too much at the self equilibrium compared to the social optimum.

(ii) When $\alpha \in (\alpha', \alpha_H)$, consumers retry and join too much, but balk too little on their own.

(iii) When $\alpha \geq \alpha_H$, consumers join too much and balk too little.

Result (iii), where consumers over-join the system, certainly is not new. Recall that when the retrial cost is high, consumers on their own disregard the option to retry, and our model becomes the one in Naor (1969). In Naor (1969), it is pointed out that self-interested consumers join the system too much as opposed to balking because they ignore the negative externalities (i.e., congestion costs to other consumers) in making the decisions.

On the other hand, results (i) and (iii) are new and counter-intuitive. For example, when the retrial cost is low, one should anticipate that consumers, when left to their own devices, would retry more often, because it is inexpensive to retry. However, our result indicates that they are not retrying enough compared to the socially optimal outcome. On the other hand, when the retrial cost rises, one would anticipate that consumers are discouraged from retrying. But we find that they still retry too much compared to social optimum.

We can explain consumers' deviation from the social optimum above using principles similar to what Naor (1969) had observed, i.e., self-selecting consumers suffer from externalities. Recall

earlier in the paper, we have identified both the positive and negative externalities of a retrial activity. When the retrial cost is low, there are no balking at both self-equilibrium and social optimum. But a consumer who retries at a long queue (compared to joining) reduces congestion costs imposed on other consumers, and self-interested consumers over-join the system and do not retry enough, because they ignore positive externalities from retrials. In contrast, as the retrial cost rises, consumers retry and join too much, but balk too little at equilibrium because they fully ignore the negative externalities from joining and retrials. If everyone were acted to maximize overall consumer welfare in this case, some of those consumers would balk.

5. Extended Model: Two Classes of Consumers

In many service settings, a portion of the consumer population cannot delay or give up their tasks. Examples include patients visiting a hospital who are in critical conditions, or consumers arriving at DMV centers who are renewing driving licenses at the very last minute. In this section, we extend the basic model to two classes of consumers, to include these "critical" arrivals.

We assume that among the new arrival rate λ, the portion $\theta \cdot \lambda$ is made up of critical arrivals who will join the queue unconditionally due to emergency needs. The rest of the population, $(1-\theta)\cdot\lambda$, includes regular consumers like in the basic model who will make rational decisions to join the queue, balk, or to retry later. We assume that $\theta \in [0,1)$ to ensure there exist some regular consumers whose equilibrium strategies are of our interest.

Note that our basic model is a special case of the two-class model with $\theta = 0$. With two classes of consumers, we can now study the impact of the critical arrivals on the decision-making and welfare of the regular consumers, as well as the impact of the regular consumers on the critical arrivals.

5.1. Equilibrium Strategies for the Regular Consumers

First we shall examine how regular consumers make retry versus join and balk decisions in the presence of the critical arrivals. For any fixed mixture θ of the population, it turns out that due to the monotonicity of balking, joining and retrial payoffs with respect to the state of the queue, the same four types of equilibrium strategies survive the equilibrium just like in the basic model except now they need to specify actions for every state of the system. (All states are now positive recurrent due to the existence of critical arrivals who join any state.)

We will keep the same notations and terminology wherever unambiguous:

• Threshold retry strategy with a threshold $n \leq N$, denoted by "JnR": Consumers join the queue at states $\{0,1,2,\ldots,n-1\}$ and retry at states $\{n, n+1, n+2, \ldots\}$.

• Threshold join/retry strategy with a threshold $n \leq N$, denoted by "$J^{J(\beta)}_{R(1-\beta)}nR$": Consumers join the queue at states $\{0,1,2,\ldots,n-2\}$, mix join and retry decisions at state $n-1$ with probabilities β and $1-\beta$, and retry at states $\{n,n+1,n+2,\ldots\}$.

• Threshold balk strategy (Naor's strategy), denoted by "JNB": Consumers join the queue at states $\{0,1,2,\ldots,N-1\}$ and balk at states $\{N,N+1,N+2,\ldots\}$.

• Threshold retry/balk strategy, denoted by "JN^R_B": Consumers join the queue at states $\{0,1,2,\ldots,N-1\}$ and mix retry and balk decisions at states $\{N,N+1,N+2,\ldots\}$.

Note that a retry/balk strategy σ, with the setting of this extended model, can actually have different retrial versus balking probabilities at states $\{N,N+1,N+2,...\}$. Let us denote γ^σ_N, γ^σ_{N+1}, γ^σ_{N+2},... the corresponding balking probabilities at these states. We will define $JN^{R(1-\gamma)}_{B(\gamma)}$ for any $\gamma \in [0,1]$ as an equivalence class of retry/balk strategies whose ratio of the steady-state balking probability over the steady-state non-joining probability is equal to γ, in other words,

$$JN^{R(1-\gamma)}_{B(\gamma)} \triangleq \{\text{Retry/balk strategy } \sigma : \frac{\pi^\sigma_B}{\pi^\sigma_B+\pi^\sigma_R} = \frac{\gamma^\sigma_N\pi^\sigma_N+\gamma^\sigma_{N+1}\pi^\sigma_{N+1}+\ldots}{\pi^\sigma_N+\pi^\sigma_{N+1}+\ldots} = \frac{\sum_{i=N}^\infty \gamma^\sigma_i\pi^\sigma_i}{\sum_{i=N}^\infty \pi^\sigma_i} = \gamma\}.$$

We show in the next lemma that all the strategies in the same equivalence class correspond to the same underlying queueing system. In other words, one cannot differentiate strategies from the same equivalence class $JN^{R(1-\gamma)}_{B(\gamma)}$ by observing the queueing system and its evolution.

LEMMA 4. *Fix $\theta \in [0,1)$. When the regular population, $(1-\theta)\cdot\lambda$, adopts any two retry/balk strategies from a given $JN^{R(1-\gamma)}_{B(\gamma)}$ class defined above, the underlying queueing systems are identical.*

As a result of Lemma 4, we will treat each equivalent class $JN^{R(1-\gamma)}_{B(\gamma)}$ as one strategy. For the two-class model, define

$$\alpha_L \triangleq (1-\pi^{JNR}_R)(v-cW^{JNR}) = (1-\sum_{i=N}^\infty \pi^{JNR}_i)(v-cW^{JNR}) = (1-\frac{\pi^{JNR}_N}{1-\theta l})(v-cW^{JNR}),$$
$$\alpha_H \triangleq (1-\pi^{JNB}_B)(v-cW^{JNB}) = (1-\sum_{i=N}^\infty \pi^{JNB}_i)(v-cW^{JNB}) = (1-\frac{\pi^{JNB}_N}{1-\theta l})(v-cW^{JNB}),$$

The next result Theorem 2' (which is analogous to Theorem 2) shows that with the presence of critical arrivals, equilibrium strategies for the regular customers follow exactly the same structure as before.

THEOREM 2'. *Fix $\theta \in [0,1)$. We have $0<\alpha_L<\alpha_H<v$. Moreover,* (i) *A Pareto-dominant equilibrium exists and is unique for any $\alpha \leq \alpha_L$: For α that falls in one of the N intervals $\{(0,\alpha_1],(\alpha_1,\alpha_2],\ldots,(\alpha_{N-2},\alpha_{N-1}],(\alpha_{N-1},\alpha_L]\}$ for an increasing sequence of thresholds $0<\alpha_1<\alpha_2<\ldots<\alpha_{N-1}<\alpha_N$ determined by the system, the retry strategies $J1R, J2R,\ldots,J(N-1)R, JNR$ represent the Pareto-dominant equilibria, respectively.* (ii) *The balk strategy JNB is the unique*

equilibrium strategy for $\alpha \geq \alpha_H$. (iii) *For* $\alpha \in [\alpha_L, \alpha_H]$, $JN_{B(\gamma)}^{R(1-\gamma)}$ *is a unique equilibrium strategy (class) with some* $\gamma \in [0,1]$, *and* γ *increases continuously from* 0 *to* 1 *in* α.[8]

In essence, regular consumers are committed to the same pattern of equilibrium strategies with or without presence of the critical arrivals, i.e., a consumer who plays strategically is always going to join the queue if it is short and retry or balk if it is long. From the regular consumers' point of view, the critical arrivals, $\theta \cdot \lambda$, can be treated simply as some *environment*. A change in the environment (i.e., a change in θ) results in changes in the values of the retrial cost thresholds, but *not the structure* of the equilibrium strategies. Henceforth, we use environment θ to denote a population λ where $\theta \cdot \lambda$ are critical arrivals, and $(1-\theta) \cdot \lambda$ are the regular consumers.

5.2. Conditional and Overall Welfare

When there was only one class of consumers in the basic model, we studied the total consumer welfare. Recall that the welfare function $\chi(\alpha)$ first decreases, then increases and finally stays constant. Scaling $\chi(\alpha)$ by a factor of $1/\lambda$, we know that consumer welfare per consumer in the basic model also follows the same down-up-flat pattern.

Now that we have two classes of consumers (with some environment θ), we can study the following three welfare quantities: (i) per-capita welfare for the regular consumers; (ii) per-capita welfare for the critical arrivals; and (iii) per-capita welfare for the whole population. Moreover, we can examine how these three welfare quantities change with θ.

Per-capita Welfare for Regular Consumers. It turns out that not only do regular consumers adopt the same structure of equilibrium strategies with or without the presence of critical arrivals, the per-capita welfare function also renders the same shape as before: (i) When $\alpha \leq \alpha_L$, per-capita welfare decreases in α, from $v - \frac{c}{\mu}$ to α_L. (ii) When $\alpha \in [\alpha_L, \alpha_H]$, it increases linearly in α, from α_L to α_H. (iii) When $\alpha \geq \alpha_H$, it remains constant at the value of α_H. However, α_L, α_H and all other thresholds on the retrial cost are functions of θ and change with θ. We demonstrate in the following lemma that all retrial cost thresholds monotonically decrease in θ.

LEMMA 5. (i) $\alpha_1, \alpha_2, \ldots, \alpha_{N-1}, \alpha_L, \alpha_H$ *all decrease in* θ. (ii) *In the region* $\alpha \in [\alpha_L, \alpha_H]$, *the unique retrial cost that induces the equilibrium retry/balk strategy* $JN_{B(\gamma)}^{R(1-\gamma)}$ *for each* $\gamma \in [0,1]$, *denoted by* $\alpha(\gamma)$, *also decreases in* θ. *Note that* $\alpha(0) = \alpha_L$ *and* $\alpha(1) = \alpha_H$. (iii) $\alpha_H - \alpha_L \to 0$ *as* $\theta \to 1$. *On the other hand, we still have* $\alpha_1 < \alpha_2 < \ldots < \alpha_{N-1} < \alpha_L$ *as* $\theta \to 1$.

[8] The new bijection is given by $\alpha = (1 - \sum_{i=N}^{\infty} \pi_i^{JN_{B(\gamma)}^{R(1-\gamma)}})(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) = (1 - \frac{\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\theta l})(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}})$.

Therefore, by focusing on the values of the retrial cost thresholds, we see that the presence of more critical arrivals makes retrials more costly for regular consumers (i.e., consumers are making the same decisions as before only with smaller retrial hassle). We will explain the intuition after presenting the next theorem, which describes how the whole welfare curve shift in θ.

Let $\mathcal{U}_{\alpha,\theta}$ denote the per-capita welfare for regular arrivals under the Pareto-dominant equilibrium strategy when the environment (i.e., the ratio of the critical arrivals in the population) is θ and the retrial cost is α. We show that, fixing any retrial cost, per-capita consumer welfare for the regular customers decreases when there are more critical arrivals in the population.

THEOREM 4. *If* $0 \le \theta_1 < \theta_2 < 1$, *then* $\mathcal{U}_{\alpha,\theta_1} \ge \mathcal{U}_{\alpha,\theta_2}$ *for all retrial cost* $\alpha \in (0,\infty)$.

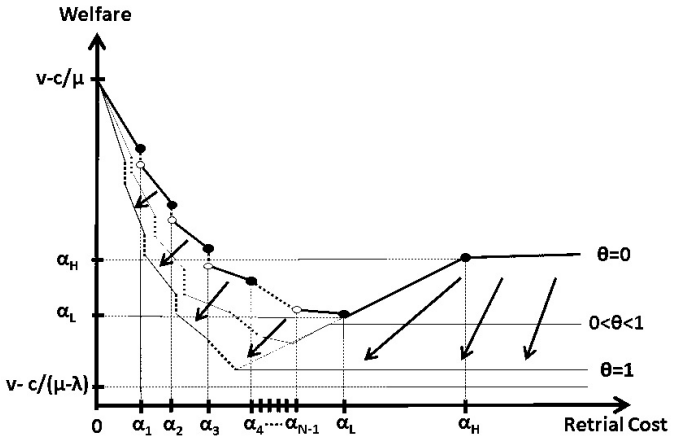


Figure 3 **Illustration of per-capita welfare for regular customers at equilibrium as the portion of critical arrivals in the population increases. Note that** $\alpha_1, \alpha_2, \ldots, \alpha_{N-1}, \alpha_L$ **and** α_H **all decrease in** θ**, and** $\alpha_H \to \alpha_L$ **as** $\theta \to 1$**. Also, for any particular** $\theta \in [0,1)$**, the welfare curve remains the down-up-flat shape.**

Theorem 4 provides us with insights on the impact of critical arrivals. An illustration is provided in Figure 3. It is indicated that the presence of critical arrivals makes every regular consumer worse off. Consider a case where one regular consumer turns himself into a critical arrival (i.e., consider the environment $\theta \to \theta + \Delta\theta$ where $\Delta\theta$ is infinitesimal). Conditional on the state under which this consumer arrives to the system, say n: (i) If this consumer's decision were to join the queue (when he was still a regular consumer), making him a critical arrival does not affect his or anybody else's welfare, because a critical arrival also joins the queue. (ii) If this consumer's decision were to balk, then forcing him join the queue not only makes this consumer worse off (receiving a negative payoff versus a zero payoff), but his join decision also causes negative externalities to all other consumers in the system. (iii) Similarly, if this consumer's decision were to retry, then forcing him join the

queue generates negative externalities and reduces his payoff because the retry decision would have been more rewarding.

Therefore, overall, when critical arrivals replace the regular consumers, the system becomes more congested. This also explains the insight we see from Lemma 5 that with the presence of critical arrivals, it is as if the retrial hassle for every regular consumer has increased. Going forward, we find that the presence of critical arrivals not only reduces regular consumers' welfare, but also their own welfare.

Per-capita Welfare for Critical Arrivals. We denote by $\mathcal{V}_{\alpha,\theta}$ the per-capita welfare for critical arrivals when the environment is θ and the retrial cost is α, given the regular population adopts their equilibrium strategies. As seen in the following theorem, per-capita welfare for the critical arrivals under any θ exhibits the down-up-flat pattern, and decreases curve-wise in θ.

THEOREM 5. *Fix a given $\theta \in [0,1)$. Per-capita welfare for the critical arrivals is a decreasing left-continuous step function over $\alpha \leq \alpha_L$. It rises over $\alpha \in [\alpha_L, \alpha_H]$, and then remains constant over $\alpha \geq \alpha_H$. Moreover, if $0 \leq \theta_1 < \theta_2 < 1$, $\mathcal{V}_{\alpha,\theta_1} \geq \mathcal{V}_{\alpha,\theta_2}$ for all retrial cost $\alpha \in (0,\infty)$.*

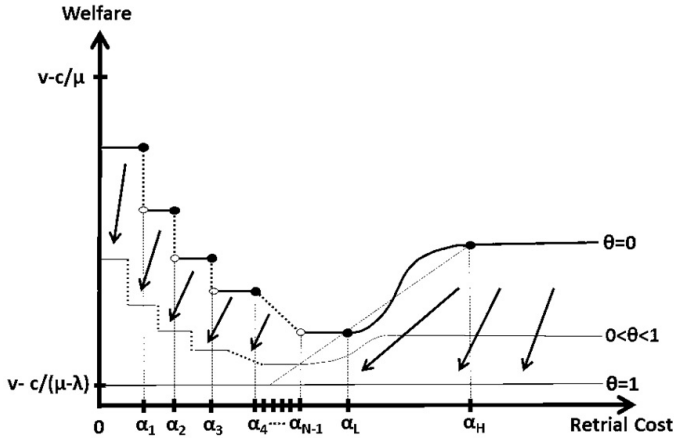


Figure 4 **Illustration of the per-capita welfare for critical arrivals when there is a higher ratio of them in the population.**

An illustration of Theorem 5 is plotted in Figure 4. The welfare curve forms a step function on the region $\alpha \leq \alpha_L$ because the underlying equilibrium strategy adopted by the regular consumers remains the same for all α that falls in one of the N intervals: $(0,\alpha_1], (\alpha_1,\alpha_2], \ldots, (\alpha_{N-2},\alpha_{N-1}], (\alpha_{N-1},\alpha_L]$. The slope for each linear piece is zero because the retrial cost is irrelevant to the welfare of critical arrivals given a fixed strategy used by the regular customers. Welfare decreases on $\alpha \leq \alpha_L$ and increases on $[\alpha_L,\alpha_H]$ is due to increased and decreased

expected waiting cost. On the other hand, the whole per-capita welfare curve decreases in θ due to the congestion created by the critical arrivals themselves, as already explained before.

Consider the extreme case when $\theta \to 1$, i.e., the whole population approaches 100% of critical arrivals. Then the underlying system would coincide with the regular $M/M/1$ queue with no balking or retrials. We show in Proposition 3 in the appendix that indeed the per-capita welfare for the critical arrivals (as well as for the whole population in this case) at any retrial cost α is equal to the per-capita welfare in a regular $M/M/1$ system, namely $v - c/(\mu - \lambda)$.

Per-capita Welfare in the Population. Following Theorem 4 and Theorem 5, we can draw the conclusion that welfare per capita in the mixed population (including both the regular consumers and the critical arrivals) still has the down-up-flat shape as a function of the retrial cost α for any given θ, and that the whole welfare curve decreases in θ. When $\theta = 0$, the curve coincides with the per-capita welfare curve for a single class of strategic consumers, namely $\chi(\alpha)/\lambda$, and when $\theta = 1$, it becomes a constant function in α with a value of $v - \frac{c}{\mu - \lambda}$, i.e., the $M/M/1$ consumer welfare.

6. Conclusions and Implications

Consumers often retry when they are faced with a long queue. However, existing service operations literature on modeling strategic consumer decisions has primarily focused on join and balk decisions. As a result, not much is known about consumer decision-making with the option to retry. Our paper seeks to fill this gap in the literature by modeling rational retrials by consumers.

We find that, with or without the presence of consumers with emergency needs (who join the queue unconditionally), consumers who play strategically act the following way when given the options to join, balk or retry: (i) When the hassle cost to retry is low, consumers follow the threshold retry strategies, i.e., they join the queue if it is short than some threshold and retry otherwise. All else being equal, when the retrial cost increases, the retrial threshold also increases. (ii) When the hassle cost to retry becomes significant, all consumers still join short queues, but some will start to balk from long queues. (iii) Eventually when the hassle cost is too high (e.g., when it exceeds the value from the service), consumers follow the threshold balking policy given in Naor (1969).

Surprisingly, while the retrial option expands consumers' choice set, it does not always increase consumer welfare compared to a system that does not allow retrials (i.e., the model in Naor (1969)). This is because the extra gain an individual benefits from retrials does not always compensate for the negative externalities (i.e., congestions costs) the retrial activities impose on other consumers in the system, when balk decisions are actually better off for the society as a whole.

We also identified positive externalities that retrials can generate. Since consumers retry at seeing longer queues, their retrial activities effectively reduce the steady-state probabilities on the higher

states of the system and increase those on the lower states. This leads to less congestion in the system, and reduced expected waiting costs for every other consumer. In this way, retrials serve as a smoothing mechanism to spread the workload over time.

Self-interested consumers ignore externalities and thus behave differently from the socially optimal point of view. In Naor (1969), consumers over-congest the system because they ignore the negative externalities of joining the queue. In our retrial model, the divergence between individual equilibrium and social optimum runs counter to intuition. We find that when the retrial cost is small, consumers on their own do not retry enough compared to the social optimum, and the deviation comes from ignorance of the positive externalities of the retrials (i.e., smoothing). On the contrary, when the retrial cost is high, consumers retry too often due to ignorance of negative externalities of the retrials (i.e., congestion).

In an extension to the basic model, we consider both regular consumers and critical arrivals in the population. The regular consumers make state-dependent join, balk and retry decisions as in the basic model, but the critical arrivals unconditionally join the queue due to emergency needs. Thus our findings can be further applied to settings such as hospitals where both types of consumers may be present. We find that, when there is a higher ratio of critical arrivals in the population, welfare decreases for both regular consumers and critical arrivals due to the extra congestion that the critical arrivals bring to the system.

Finally, in modeling retrials in this paper, we assume that the number of periods between individual retrials is fixed. However, all of our results will continue to hold if we allow the number of periods between retrials to be random. In fact, instead of using the periodic model that we currently have, our results still hold if the time between retrials is modeled by an exponential distribution as in orbit models. (We will address the details in the Appendix B.)

Our findings have several implications for queue management policies in many service settings. As consumer equilibrium deviates from the social optimum, a social planner could implement various policies to improve the overall consumer welfare. For example, we find that consumers retry too little when the retrial hassle is low; and they retry too much when it is high. As a result, the social planner should consider either subsidizing retrials at low queue lengths or charging tolls for retrials at high queue lengths.

Appendix A: Proofs

Proof of Proposition 1: Results follow from discussion in §2.4 and Definition 1. □

Proof of Theorem 1: Suppose the population adopts JnR or $J_J^R nR$ and $n > N$, then some consumers, if not all, will join the queue at state N. Such strategies cannot be equilibrium strategies because these consumers can do better if they balk at state N, rather than joining, as $0 > v - \frac{c}{\mu}(N+1)$. On the other hand, we have assumed that every consumer will join (i.e., will not retry) on seeing an idle server, so we must have $n \geq 1$. Similarly, if JnB or Jn_B^R is adopted by the population and $n \neq N$, then either some consumers are specified by the strategy to balk when they are better off not to, or specified to join the queue when better off not to join. □

Proof of Theorem 2: Complete proof of this theorem requires Lemma 6 through Lemma 16 below. We first establish in Lemma 6 that JNB is an equilibrium strategy if and only if $\alpha \geq \alpha_H$.

LEMMA 6. *JNB is an equilibrium if and only if $\alpha \geq \alpha_H \triangleq (1-\pi_N^{JNB})(v - cW^{JNB})$ where π_N^{JNB} is the steady-state probability of state N, and W^{JNB} is the expected waiting time for a consumer conditional on joining the queue, when the population adopts the balk strategy JNB, written as*

$$W^{JNB} = \frac{\pi_0^{JNB}}{\pi_J^{JNB}}\frac{1}{\mu} + \frac{\pi_1^{JNB}}{\pi_J^{JNB}}\frac{2}{\mu} + \ldots + \frac{\pi_{N-1}^{JNB}}{\pi_J^{JNB}}\frac{N}{\mu} = \frac{\pi_0^{JNB}}{1-\pi_N^{JNB}}\frac{1}{\mu} + \frac{\pi_1^{JNB}}{1-\pi_N^{JNB}}\frac{2}{\mu} + \ldots + \frac{\pi_{N-1}^{JNB}}{1-\pi_N^{JNB}}\frac{N}{\mu}. \tag{13}$$

Proof of Lemma 6: Let us suppose that everybody else is adopting JNB on every service occasion, and only one consumer is given the opportunity to change his strategy unilaterally. Then, JNB will be an equilibrium (i.e., this consumer has no incentive to retry at any state $\{1, 2, \ldots, N-1, N\}$) if and only if his expected payoff from a retry decision is less than or equal to 0. Since this consumer's retrial payoff is decreasing in the retrial cost α, there exists one unique value of α above which JNB is an equilibrium and below which it is not. We will show this value is equal to α_H, given by $(1-\pi_N^{JNB})(v - cW^{JNB})$.

At α_H, this consumer's retrial payoff is exactly 0, i.e.,

$$\begin{aligned}
\pi_0^{JNB}(v - \frac{c}{\mu}) + \pi_1^{JNB}(v - \frac{2c}{\mu}) + \ldots + \pi_{N-1}^{JNB}(v - \frac{cN}{\mu}) + \pi_N^{JNB}\cdot 0 - \alpha_H &= 0 \\
(1-\pi_N^{JNB})v - c(\pi_0^{JNB}\frac{1}{\mu} + \pi_1^{JNB}\frac{2}{\mu} + \ldots + \pi_{N-1}^{JNB}\frac{N}{\mu}) - \alpha_H &= 0 \\
(1-\pi_N^{JNB})v - (1-\pi_N^{JNB})cW^{JNB} - \alpha_H &= 0 \\
(1-\pi_N^{JNB})(v - cW^{JNB}) - \alpha_H &= 0
\end{aligned}$$

and therefore Lemma 6 is proved. □

Next, we show that when $\alpha \in (0, \alpha_L]$, where $\alpha_L \triangleq (1-\pi_N^{JNR})(v - cW^{JNR})$, the system equilibrium is given by either a retry strategy (i.e., JnR-type) or a join/retry strategy (i.e., $J_R^J nR$-type).

We first look for any equilibrium retry strategy (i.e., JnR-type), if it exists. Under the strategy JnR, any consumer who sees a state smaller than n will join the queue. Otherwise, he retries during the following

period. The underlying queueing system when the population adopts JnR is thus $M/M/1/n$. And we have $\pi_J^{JnR} = \sum_{i=0}^{n-1} \pi_i^{JnR}$, $\pi_R^{JnR} = \pi_n^{JnR}$, and $\pi_B^{JnR} = 0$.

The total arrival rate defined in (2) for the system under JnR is thus given by

$$\lambda_{total}^{JnR} = \frac{\lambda}{1-\pi_R^{JnR}} = \frac{\lambda}{1-\pi_n^{JnR}}, \tag{14}$$

and the traffic intensity defined in (4) for the system under JnR is

$$\rho^{JnR} = \frac{l}{1-\pi_R^{JnR}} = \frac{l}{1-\pi_n^{JnR}}. \tag{15}$$

For strategy JnR, we call equation (15) the *stability condition* on ρ and π_n that ensures that the steady-state probabilities are consistent. In fact, it is simply a constraint on the variables ρ and n.

From (6) and the fact that $\pi_J^{JnR} = 1-\pi_R^{JnR}$, $\pi_R^{JnR} = \pi_n^{JnR}$, the retrial payoff under JnR is

$$v - cW^{JnR} - \frac{\alpha}{1-\pi_n^{JnR}}, \tag{16}$$

where the expected waiting time for a customer conditional on joining can be written as

$$W^{JnR} = \frac{\pi_0^{JnR}}{1-\pi_n^{JnR}}\frac{1}{\mu} + \frac{\pi_1^{JnR}}{1-\pi_n^{JnR}}\frac{2}{\mu} + \ldots + \frac{\pi_{n-1}^{JnR}}{1-\pi_n^{JnR}}\frac{n}{\mu}.$$

If the retry strategy JnR indeed leads to an system equilibrium, then besides the stability condition from equation (15), it also needs to satisfy the condition in Proposition 1 to ensure that consumers would not want to deviate from it at any state, on any service occasion. In this context, we require that the retry payoff is greater than the joining or balking payoff at state n, but is less than or equal to the joining payoff at state $n-1$, i.e.,

$$\max\{0, v - \frac{c}{\mu}(n+1)\} \le v - cW^{JnR} - \frac{\alpha}{1-\pi_n^{JnR}} \le v - \frac{c}{\mu}n. \tag{17}$$

From Theorem 1, we know that $n \le N$. We can then transfer (17) into

$$\frac{c}{\mu}(n+1) \ge cW^{JnR} + \frac{\alpha}{1-\pi_n^{JnR}} \ge \frac{c}{\mu}n \text{ with } n \le N, \tag{18}$$

which we call the *indifference condition* of the equilibrium. To be more precise here, when $n = N$, we require the strategy JNR satisfies $v \ge cW^{JNR} + \frac{\alpha}{1-\pi_N^{JNR}} \ge \frac{c}{\mu}N$ instead of $\frac{c}{\mu}(N+1) \ge cW^{JNR} + \frac{\alpha}{1-\pi_N^{JNR}} \ge \frac{c}{\mu}N$ because of the max operator in (17).

We shall look for all possible n's that satisfy both the stability and the indifference conditions and then pick out the integer solutions (because the retrial threshold n is an integer). Each integer solution for n then represents a legitimate equilibrium retry strategy in JnR. It turns out that

LEMMA 7. *Fix a system (i.e., λ, μ, v, c). For any integer $n \in \{1, 2, \ldots, N\}$, there exists a unique retrial cost α_n with which n satisfies (15) and (18) at the same time. Moreover, α_n increases in n.*

Proof of Lemma 7: If $\alpha = \alpha^*$ solves the equality $\frac{c}{\mu}(n^*+1) = cW^{Jn^*R} + \frac{\alpha}{1-\pi_{n^*}^{Jn^*R}}$ for some n^*, and assuming the stability condition in (15) holds, we have from (18) that Jn^*R generates an equilibrium retry strategy for $\alpha \in [\alpha^* - \frac{c}{\mu}(1-\pi_{n^*}^{Jn^*R}), \alpha^*]$. On the other hand, if $\alpha = \alpha^{**}$ solves the equality $\frac{c}{\mu}n^{**} = cW^{Jn^{**}R} + \frac{\alpha}{1-\pi_{n^{**}}^{Jn^{**}R}}$

for some n^{**}, and if the stability condition holds, again from (18) we know that $Jn^{**}R$ is an equilibrium for $\alpha \in [\alpha^{**}, \alpha^{**} + \frac{c}{\mu}(1-\pi_{n^{**}}^{Jn^{**}R})]$. For a fixed $n = n^* = n^{**}$, the set of α that satisfies the inequality $\frac{c}{\mu}(n+1) \geq cW^{JnR} + \frac{\alpha}{1-\pi_n^{JnR}} \geq \frac{c}{\mu}n$ in (18) is $\alpha \in [\alpha^* - \frac{c}{\mu}(1-\pi_{n^*}^{Jn^*R}), \alpha^*]$ which is exactly that same as $\alpha \in [\alpha^{**}, \alpha^{**} + \frac{c}{\mu}(1-\pi_{n^{**}}^{Jn^{**}R})]$ by setting $\alpha^* - \frac{c}{\mu}(1-\pi_n^{JnR}) = \alpha^{**}$. In other words, to find all (α, n) pairs that satisfy the seemingly two inequalities in (18), we only need to find solutions (α^*, n^*) to one equality, say,

$$\frac{c}{\mu}(n+1) = cW^{JnR} + \frac{\alpha}{1-\pi_n^{JnR}} \text{ (with } n \leq N), \tag{19}$$

and then all the pairs $\{(\alpha, n^*) : \alpha \in [\alpha^* - \frac{c}{\mu}\pi_{n^*}^{Jn^*R}, \alpha^*]\}$ are solutions to the indifference condition in (18).

Since $W^{JnR} = \frac{1}{\mu}\sum_{k=0}^{n-1}\frac{\pi_k^{JnR}}{1-\pi_n^{JnR}}(k+1)$, equation (19) becomes

$$\begin{aligned}
&\frac{c}{\mu}(n+1) = \frac{c}{\mu}\sum_{k=0}^{n-1}\frac{\pi_k^{JnR}}{1-\pi_n^{JnR}}(k+1) + \frac{\alpha}{1-\pi_n^{JnR}} \Leftrightarrow (1-\pi_n^{JnR})\frac{c}{\mu}(n+1) = \frac{c}{\mu}\sum_{k=0}^{n-1}\pi_k^{JnR}(k+1) + \alpha \\
\Leftrightarrow (1-\pi_n^{JnR})\frac{c}{\mu}n &= \frac{c}{\mu}\sum_{k=0}^{n-1}\pi_k^{JnR}k + \alpha \Leftrightarrow \frac{c}{\mu}n = \frac{c}{\mu}\sum_{k=0}^{n}\pi_k^{JnR}k + \alpha \Leftrightarrow \frac{c}{\mu}(n - \sum_{k=0}^{n}\pi_k^{JnR}k) = \alpha
\end{aligned} \tag{20}$$

By letting $L^{JnR} \triangleq \sum_{k=0}^{n}\pi_k^{JnR}k$ denote the long-run average number of consumers in the $M/M/1/n$ system under the retrial strategy JnR and $r \triangleq \alpha/\frac{c}{u}$ the cost ratio of the retrial cost over the expected waiting cost for a service cycle, equation (20) becomes

$$n - L^{JnR} = r \tag{21}$$

Ignore the feasibility condition $n \leq N$ and the integer condition on n for now. Then equation (21) (the indifference condition) along with equation (15) (the stability condition) give us two equations for three unknowns (n, ρ and r) where the other parameters λ, μ, v, c are fixed system inputs. In what follows, we should find solutions (ρ^{Jn^*R}, n^*) to equations (15) and (21) in terms of r . Algebraic operations force us to separate the case when $\rho^{Jn^*R} = 1$ (so-called *the trivial solution*) with the case when $\rho^{Jn^*R} \neq 1$ (so-called *the non-trivial solution*).

Trivial Solution: We first consider the "trivial" case, i.e., when $\rho^{Jn^*R} = 1$ is part of the solution to equations (15) and (21). When $\rho^{JnR} = 1$, we have $\pi_0^{JnR} = \pi_1^{JnR} = \pi_2^{JnR} = \ldots = \pi_n^{JnR} = \frac{1}{n+1}$ and $L^{JnR} = n/2$. From equation (21), we have $n = 2r$. Plugging $n = 2r$ into (15), we then have $\rho^{JnR} = \frac{2r+1}{2r}l$. Since $\rho^{JnR} = 1$, we must have $\frac{2r+1}{2r}l = 1$, or

$$r = \frac{1}{2}\frac{l}{1-l} \tag{22}$$

Therefore, when $r = \frac{1}{2}\frac{l}{1-l}$, we have a trivial (and unique) solution to equations (15) and (21) where

$$(\rho^{Jn^*R}, n^*) = (1, 2r) = (1, \frac{l}{1-l}).$$

Non-trivial Solution: Next, we search for any possible solution, (ρ^{Jn^*R}, n^*), to equations (15) and (21) when $\rho^{Jn^*R} \neq 1$. Although $l < 1$, ρ^{JnR} can be both smaller than or greater than 1 at equilibrium (we only know $\rho^{JnR} > l$). When $\rho^{JnR} \neq 1$, we have

$$\pi_0^{JnR} = \frac{1}{1 + \rho^{JnR} + (\rho^{JnR})^2 + \ldots + (\rho^{JnR})^n} = \frac{1-\rho^{JnR}}{1-(\rho^{JnR})^{n+1}} \tag{23}$$

$$\pi_n^{JnR} = (\rho^{JnR})^n \pi_0^{JnR} = \frac{(\rho^{JnR})^n - (\rho^{JnR})^{n+1}}{1-(\rho^{JnR})^{n+1}} \tag{24}$$

$$1-\pi_n^{JnR} = 1 - \frac{(\rho^{JnR})^n - (\rho^{JnR})^{n+1}}{1-(\rho^{JnR})^{n+1}} = \frac{1-(\rho^{JnR})^n}{1-(\rho^{JnR})^{n+1}} \tag{25}$$

Therefore, equation (15), the stability condition becomes

$$\rho^{JnR}(1-\pi_n^{JnR}) = l \Leftrightarrow \rho^{JnR}[\frac{1-(\rho^{JnR})^n}{1-(\rho^{JnR})^{n+1}}] = l \text{ from (25)} \Leftrightarrow \frac{\rho^{JnR}-(\rho^{JnR})^{n+1}}{1-(\rho^{JnR})^{n+1}} = l$$

$$\Leftrightarrow \rho^{JnR} - (\rho^{JnR})^{n+1} = l - l(\rho^{JnR})^{n+1} \Leftrightarrow \rho^{JnR} - l = (1-l)(\rho^{JnR})^{n+1} \Leftrightarrow (\rho^{JnR})^{n+1} = \frac{\rho^{JnR}-l}{1-l} \tag{26}$$

$$\Leftrightarrow n+1 = \log_{\rho^{JnR}}(\frac{\rho^{JnR}-l}{1-l}) = \frac{\ln(\frac{\rho^{JnR}-l}{1-l})}{\ln \rho^{JnR}} \tag{27}$$

On the other hand, when $\rho^{JnR} \neq 1$, the average number of customers in the queue is given by

$$L^{JnR} = \sum_{k=0}^{n} \pi_k^{JnR} k = \frac{\rho^{JnR}}{1-\rho^{JnR}} - \frac{(n+1)(\rho^{JnR})^{n+1}}{1-(\rho^{JnR})^{n+1}}. \tag{28}$$

Equation (21), the indifference condition, $n - L^{JnR} = r$ is thus equivalent to

$$n - \frac{\rho^{JnR}}{1-\rho^{JnR}} + \frac{(n+1)(\rho^{JnR})^{n+1}}{1-(\rho^{JnR})^{n+1}} = r \Leftrightarrow n + \frac{(n+1)(\rho^{JnR})^{n+1}}{1-(\rho^{JnR})^{n+1}} = r + \frac{\rho^{JnR}}{1-\rho^{JnR}}$$

$$\Leftrightarrow n+1 + \frac{(n+1)(\rho^{JnR})^{n+1}}{1-(\rho^{JnR})^{n+1}} = r+1 + \frac{\rho^{JnR}}{1-\rho^{JnR}} \Leftrightarrow (n+1)(1 + \frac{(\rho^{JnR})^{n+1}}{1-(\rho^{JnR})^{n+1}}) = r+1 + \frac{\rho^{JnR}}{1-\rho^{JnR}}$$

$$\Leftrightarrow (n+1)\frac{1}{1-(\rho^{JnR})^{n+1}} = r+1 + \frac{\rho^{JnR}}{1-\rho^{JnR}} \Leftrightarrow (n+1)\frac{1-l}{1-\rho^{JnR}} = r+1 + \frac{\rho^{JnR}}{1-\rho^{JnR}} \text{ from (26)}$$

$$\Leftrightarrow (n+1)(1-l) = r(1-\rho^{JnR}) + 1 \Leftrightarrow n+1 = \frac{r(1-\rho^{JnR})+1}{1-l} \tag{29}$$

We have just transferred the two equilibrium conditions in (15) and (21), into an equivalent set of equations

$$(27') : n = \frac{\ln(\frac{\rho^{JnR}-l}{1-l})}{\ln \rho^{JnR}} - 1; \text{ And } (29') : n = \frac{r(1-\rho^{JnR})+1}{1-l} - 1.$$

in the sense that a pair of solutions (n^*, ρ^*) that solves equations (15) and (21) also solves equations (27') and (29'), and vice versa.

We observe from equations (27') and (29') that a necessary condition for $n \geq 0$ is $l < \rho \leq 1 + \frac{l}{r}$. Define

$$f_1(\rho) \triangleq \frac{\ln(\frac{\rho-l}{1-l})}{\ln \rho} - 1 \text{ from } (27'); \text{ and } f_2(\rho) \triangleq \frac{r(1-\rho)+1}{1-l} - 1 \text{ from } (29').$$

Then, any intersection of the graphs of $f_1(\rho)$ and $f_2(\rho)$ (other than at $\rho = 1$) on the feasible region $l < \rho \leq 1 + \frac{l}{r}$ will lead to a solution $(\rho^* = \rho^{Jn^*R}, n^*) = (\rho^*, f_1(\rho^*) = f_2(\rho^*))$.

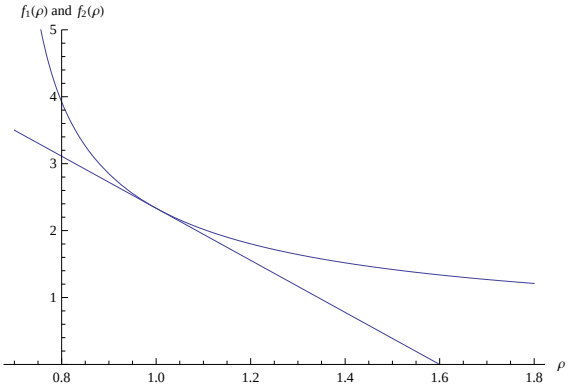
It can be shown that $f_1(\rho)$ is continuous, decreasing and convex in ρ on the region $\rho \in (l, +\infty)$ with

$$\lim_{\rho \to l} f_1(\rho) = +\infty; \lim_{\rho \to 1} f_1(\rho) = \frac{l}{1-l}; \lim_{\rho \to +\infty} f_1(\rho) = 0. \tag{30}$$

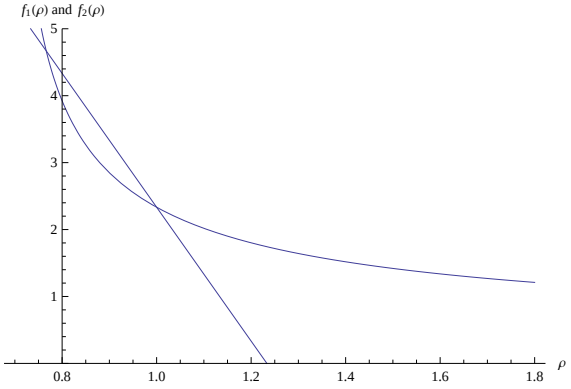
On the other hand, $f_2(\rho)$ is simply a straight line in ρ with slope $-\frac{r}{1-l}$ on the region $\rho \in [0, 1 + \frac{l}{r}]$ with

$$f_2(0) = \frac{r+l}{1-l}; f_2(1) \quad = \frac{l}{1-l}; f_2(1+\frac{l}{r}) = 0. \tag{31}$$

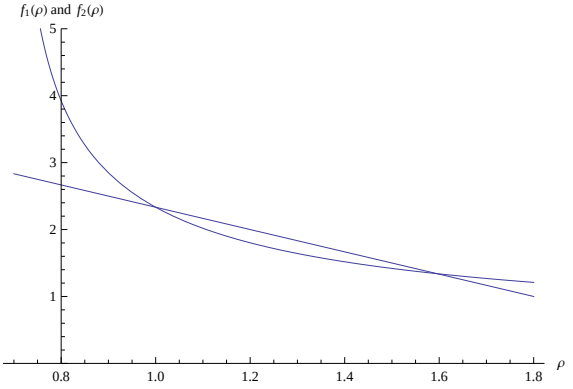
Note from (30) and (31) that $f_1(\rho)$ and $f_2(\rho)$ intersect at $\rho = 1$ with $f_1(1) = f_2(1) = \frac{l}{1-l}$. If $f_2(\rho)$ is a tangent line to $f_1(\rho)$ at $\rho = 1$, as shown in Figure 5/(a), then we will not have a solution (ρ^*, n^*) with $\rho^* \neq 1$



(a)$r = \frac{7}{6}$: $f_2(\rho)$ is tangent to $f_1(\rho)$.



(b)$r = 3$: not tangent and cross above.



(c)$r = \frac{1}{2}$: not tangent and cross below.

Figure 5 **In the three subfigures above, we let $l = 0.7$ and plot $f_1(\rho)$ and $f_2(\rho)$ w.r.t. different values of r. In every case, $f_1(\rho)$ and $f_2(\rho)$ cross the point ($\rho = 1, f(\rho) = \frac{l}{1-l}) = (1, \frac{7}{3})$. (a) When $r = \frac{7}{6} = \frac{1}{2}\frac{l}{1-l}$, $f_2(\rho)$ is a tangent line to the curve $f_1(\rho)$. (b) When $r = 3 > \frac{7}{6} = \frac{1}{2}\frac{l}{1-l}$, they cross above. (c) When $r = \frac{1}{2} < \frac{7}{6} = \frac{1}{2}\frac{l}{1-l}$, they cross below.**

(because in this case $f_1(\rho)$ and $f_2(\rho)$ will only intersect at $\rho = 1$ on the feasible region $l < \rho \leq 1 + \frac{l}{r}$). It can be verified that the slope of the curve $f_1(\rho)$ at $\rho = 1$ is $-\frac{1}{2}\frac{l}{(1-l)^2}$ so $f_2(\rho)$ coincides with the tangent line of $f_1(\rho)$ at $\rho = 1$ if and only if

$$-\frac{r}{1-l} = -\frac{1}{2}\frac{l}{(1-l)^2} \Leftrightarrow r = \frac{1}{2}\frac{l}{(1-l)} \tag{32}$$

note that condition (32) is exactly that same as condition (22), so in this case, although we do not have any solution (ρ^*, n^*) such that $\rho^* \neq 1$, we do have an unique solution found earlier during the discussion of the trivial solution, i.e., $(\rho^*, n^*) = (1, \frac{l}{1-l})$.

When $r \neq \frac{1}{2}\frac{l}{(1-l)}$, then $f_2(\rho)$ is still a straight line but no longer a tangent one to the curve $f_1(\rho)$ at $\rho = 1$. Since $f_1(\rho)$ is a smooth decreasing convex curve, $f_2(\rho)$ is a straight line, and they both have a common intersection at $(1, \frac{1}{1-l})$, they will intersect at some point ρ^* other than $\rho^* = 1$ on the region $l < \rho \leq 1 + \frac{l}{r}$ as shown in Figure 5/(b) and (c). The intersection point depends on the slope of $f_2(\rho)$, or the value of r,

or simply the retrial cost α (assuming parameters λ, μ, v, c are system inputs and fixed). Thus, solution also exists when $r \neq \frac{1}{2}\frac{l}{(1-l)}$ and it is unique due to the single-crossing property of $f_1(\rho)$ and $f_2(\rho)$. Note again that the intersection where $\rho = 1$ cannot be counted as a second non-trivial solution.

So far we have shown that given the inputs of the system (i.e., λ, μ, v, c), a particular value of the retrial fee α, which is transformed into a particular value of the cost ratio $r = \alpha/\frac{c}{\mu}$, induces a unique pair of solutions, in the threshold n^* and its corresponding traffic $\rho^* = \rho^{Jn^*R}$ among consumers, to equations (15) and (21). Furthermore, it is clear from the graph that when $\alpha \uparrow$ (i.e., $f_2(\rho)$ becomes steeper), $n^* \uparrow$ and $\rho^* \downarrow$. In fact, when $\alpha \to 0$, $\lim_{\alpha\to 0} n^* = 0$ and $\lim_{\alpha\to 0} \rho^* = \infty$. When $\alpha \to \infty$, $\lim_{\alpha\to\infty} n^* = \infty$ and $\lim_{\alpha\to\infty} \rho^* = l$.

At this moment we are ready to pick up the feasibility condition $n^* \leq N$ and the integer condition on n. (Recall that $N = \lfloor v/\frac{c}{\mu} \rfloor$ indicates the balking threshold in Naor's model, see (1).) Let us denote $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_{N-1} < \alpha_N$ the retrial costs that induce the solutions to equations (15) and (21) with $n = 1, 2, \ldots, N-1, N$, respectively. We next show existence and uniqueness of these α's.

Recall that $\lim_{\alpha\to 0} n^* = 0$, $\lim_{\alpha\to\infty} n^* = \infty$, and n^* is continuous and increases in α. For any fixed integer $n \in (0, N]$, compute the unique ρ_n such that $f_1(\rho_n) = n$, i.e., (ρ_n, n) solves equation (27'). Next, let $\alpha_n = \frac{(n+1)(1-l)-1}{1-\rho_n}\frac{c}{\mu}$, i.e., α_n is such that (ρ_n, n) solves equation (29'). Since (ρ_n, n) solves both equations (27') and (29'), α_n is the amount of the retrial cost that would induce $n^* = n$ as a solution to equations (15) and (21). Uniqueness and monotonicity of α_n were demonstrated in earlier discussion with the graph. □

From Lemma 7, we know that with retrial cost $\{\alpha_n : n \in \{1, 2, \ldots, N\}\}$, the integer n is a solution of equations (15) and (18). Thus, the associated α_n induces an equilibrium retry strategy JnR for the system. However, for α_N, it does not ensure that JNR is an equilibrium, because $0 > v - \frac{c}{\mu}(N+1) = v - cW^{JNR} - \frac{\alpha_N}{1-\pi_N^{JNR}}$, and the indifference condition in (17) is thus violated. (This issue was mentioned after condition (18).) To fix this boundary condition, let $\{\alpha_L : \alpha_{N-1} < \alpha_L < \alpha_N\}$ be the particular retrial cost that induces a zero retrial payoff when the population adopts JNR, i.e., $0 = v - cW^{JNR} - \frac{\alpha_L}{1-\pi_N^{JNR}}$, or $\alpha_L \triangleq (1-\pi_N^{JNR})(v - cW^{JNR})$. We are now at a good position to characterize *all* equilibrium retry strategies of the system in the following lemma.

LEMMA 8. *For a given system, there exist a sequence of thresholds on the retrial cost,* $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_{N-1} < \alpha_L$, *such that* (i) *when the retrial cost* $\alpha \leq \alpha_1$, *the retry strategy* $J1R$ *is an equilibrium;* (ii) *when* $\alpha \in [\alpha_n - \frac{cl}{\mu\rho^{JnR}}, \alpha_n]$ *for* $n \in \{2, \ldots, N-1\}$, *the retry strategy* JnR *is an equilibrium;* (iii) *when* $\alpha \in [\alpha_L - (v - \frac{c}{\mu}N)\frac{l}{\rho^{JNR}}, \alpha_L]$, *the retry strategy* JNR *is an equilibrium.*

Proof of Lemma 8: Fix $n \in \{1, 2, \ldots, N-1\}$. When the retrial cost $\alpha = \alpha_n$, since both stability and indifference conditions are satisfied for the solution $(n^*, \rho^*) = (n, \rho^{JnR})$, we have an equilibrium under JnR. Specifically, when $\alpha = \alpha_1$, $J1R$ is an equilibrium strategy which specifies that an arrival only joins the server if it is idle. By the construction of the game, every consumer would join an idle server no matter what the retrial cost α is (because the retrial and balking payoffs are always less than the joining payoff at seeing an idle server). Therefore, for $\alpha \leq \alpha_1$, $J1R$ remains an equilibrium strategy. On the other hand,

for $n \in \{1, 2, \ldots, N-1\}$, when $\alpha \in [\alpha_n - \frac{c}{\mu}(1-\pi_n^{JnR}), \alpha_n]$, the indifference condition of (19) still holds at n. Thus, JnR is also an equilibrium for such α. Note that, from equation (15) that $1-\pi_n^{JnR} = \frac{l}{\rho^{JnR}}$. Thus, when the retrial cost $\alpha \in [\alpha_n - \frac{c}{\mu}\frac{l}{\rho^{JnR}}, \alpha_n]$, we have an equilibrium under JnR. Finally, when $\alpha = \alpha_L$, we have a retrial payoff of zero under JNR according to the definition of α_L, so such a strategy is an equilibrium. Furthermore, when $\alpha \in [\alpha_L - (v - \frac{c}{\mu}N)\frac{l}{\rho^{JNR}}, \alpha_L]$, the retrial payoff under JNR is greater than zero and less than $v - \frac{c}{\mu}N$ thus the strategy remains an equilibrium. □

Lemma 8 states that there exist N intervals on the low retrial cost region $\alpha \in (0, \alpha_L]$ such that the retry strategies with increasing thresholds $J1R, J2R, \ldots, JNR$ form equilibrium strategies. However, Lemma 8 does not provide information on whether these N intervals overlap or if there are gaps between the intervals. Therefore, it is possible that under some retrial cost $\alpha \leq \alpha_L$, there does not exist an equilibrium retry strategy because α is not contained in any of the N intervals, or that some other retrial cost may induce more than one equilibrium. This can happen for example, if $\alpha_n - \alpha_{n-1} \leq \frac{cl}{\mu\rho_n^{JnR}}$, then α_{n-1} will induce both equilibrium retry strategies $J(n-1)R$ and JnR. We demonstrate below, via looking at equilibrium join/retry strategies, that this is indeed the case for every $n \in \{2, 3, \ldots, N\}$. Therefore, there is no gap between the N intervals stated in Lemma 8. Given the presence of multiple equilibria, we will then identify the Pereto-dominant equilibrium strategy for every $\alpha \leq \alpha_L$ in Lemma 13.

We now search for system equilibria under a join/retry strategy in the form of $J_{R(1-\beta)}^{J(\beta)}nR$, i.e., all consumers join the queue at states $\{0, 1, 2, \ldots, n-2\}$, join the queue with probability β and retry with probability $1-\beta$ at state $n-1$, and then retry at state n. To tackle down such a strategy, we realize that it is an intermediate phase between two retry strategies. Consider $J_{R(1-\beta)}^{J(\beta)}nR$: when β goes to 0, it coincides with the retry strategy with threshold $n-1$, namely $J(n-1)R$; and when β goes to 1, it coincides with the retry strategy with threshold n, namely JnR. And when $\beta \in (0,1)$, we have a non-degenerate join/retry mixed strategy. We first show in the following lemma that there exists a unique retrial cost that induces $J_{R(1-\beta)}^{J(\beta)}nR$ as an equilibrium strategy, for fixed n.

LEMMA 9. *Fix $1 \leq n \leq N$. There exists a unique retrial cost α, for any particular $\beta \in (0,1)$, such that $J_{R(1-\beta)}^{J(\beta)}nR$ is an equilibrium strategy. Therefore, we can write α which induces the equilibrium as a function as β. Moreover, α is continuous in β.*

Proof of Lemma 9: Fix $1 \leq n \leq N$. Choose any $\beta \in (0,1)$. For $J_{R(1-\beta)}^{J(\beta)}nR$ to be an equilibrium, upon arriving at state $n-1$, a consumer is indifferent between join and retry decisions. Therefore, the joining and the retrial payoffs at state $n-1$ (under $J_{R(1-\beta)}^{J(\beta)}nR$) are identical and both should be greater than 0 (the balking payoff), i.e.,

$$v - \frac{c}{\mu}(n) = v - cW^{J_{R(1-\beta)}^{J(\beta)}nR} - \frac{\alpha}{1-\pi_R^{J_{R(1-\beta)}^{J(\beta)}nR}} \tag{33}$$

If and only if the retrial cost α satisfies equation (33), the policy $J_{R(1-\beta)}^{J(\beta)}nR$ becomes an equilibrium. Since

$W^{J^{J(\beta)}_{R(1-\beta)}nR}$ and $\pi_R^{J^{J(\beta)}_{R(1-\beta)}nR}$ are just some fixed quantities given that the population adopts the join/retry strategy $J^{J(\beta)}_{R(1-\beta)}nR$, it is clear from equation (33) that α exists and is unique, where

$$\alpha = (1-\pi_R^{J^{J(\beta)}_{R(1-\beta)}nR})[\frac{c}{\mu}n - cW^{J^{J(\beta)}_{R(1-\beta)}nR}] \tag{34}$$

is a function of β. Finally, when β increases from 0 to 1, the underlying queueing system, where everyone adopts the strategy $J^{J(\beta)}_{R(1-\beta)}nR$, evolves continuously, and thus the conditional waiting time $W^{J^{J(\beta)}_{R(1-\beta)}nR}$ and the steady-state retrial probability $\pi_R^{J^{J(\beta)}_{R(1-\beta)}nR}$ are both continuous quantities in β. It follows that $\alpha(\beta)$, given in (34), is also continuous in β. □

Lemma 9 tells us that α is continuous in β. Lemma 10 further tells us that α decreases in β.

LEMMA 10. *Fix $n \in \{2,\dots,N-1,N\}$.* (i) $\alpha_n - \frac{cl}{\mu\rho^{J nR}} < \alpha_{n-1}$. (ii) *For any $\beta \in [0,1]$, there exists a unique $\alpha \in [\alpha_n - \frac{cl}{\mu\rho^{J nR}}, \alpha_{n-1}]$ such that the strategy $J^{J(\beta)}_{R(1-\beta)}nR$ is an equilibrium.* (iii) *Moreover, α decreases in β.*

To prove Lemma 10, we first prove two supporting lemmas (Lemmas 11 and 12).

LEMMA 11. *For fixed $n : 1 \le n \le N$, vary α such that $J^{J(\beta)}_{R(1-\beta)}nR$ is an equilibrium policy for $\beta \in (0,1)$. Then, the partial derivative of consumer welfare under the equilibrium join/retry strategy $J^{J(\beta)}_{R(1-\beta)}nR$ with respect to β is negative. Mathematically, if we denote U as consumer welfare, then $\frac{\partial U^{J^{J(\beta)}_{R(1-\beta)}nR}}{\partial\beta} < 0$.*

Proof of Lemma 11: Suppose $J^{J(\beta)}_{R(1-\beta)}nR$ is an equilibrium strategy. We still use

$$\pi_0^{J^{J(\beta)}_{R(1-\beta)}nR}, \pi_1^{J^{J(\beta)}_{R(1-\beta)}nR}, \dots, \pi_{n-1}^{J^{J(\beta)}_{R(1-\beta)}nR}, \pi_n^{J^{J(\beta)}_{R(1-\beta)}nR}$$

as the steady-state probabilities for states $\{0,1,\dots,n-1,n\}$, respectively. Let

$$U_0^{J^{J(\beta)}_{R(1-\beta)}nR}, U_1^{J^{J(\beta)}_{R(1-\beta)}nR}, \dots, U_{n-1}^{J^{J(\beta)}_{R(1-\beta)}nR}, U_n^{J^{J(\beta)}_{R(1-\beta)}nR}$$

denote a consumer's (expected) payoff arriving to state $x \in \{0,1,\dots,n-1,n\}$ (when the population adopts $J^{J(\beta)}_{R(1-\beta)}nR$). Then,

$$\begin{aligned}
U_0^{J^{J(\beta)}_{R(1-\beta)}nR} &= v - \frac{c}{\mu}(1) \text{ for joining;}\\
U_1^{J^{J(\beta)}_{R(1-\beta)}nR} &= v - \frac{c}{\mu}(2) \text{ for joining;}\\
\vdots &= \quad \vdots\\
U_{n-2}^{J^{J(\beta)}_{R(1-\beta)}nR} &= v - \frac{c}{\mu}(n-1) \text{ for joining;}\\
U_{n-1}^{J^{J(\beta)}_{R(1-\beta)}nR} &= v - \frac{c}{\mu}(n) \text{ no matter choosing to join or retry;}\\
U_n^{J^{J(\beta)}_{R(1-\beta)}nR} &= v - \frac{c}{\mu}(n) \text{ for retrying.}
\end{aligned}$$

On the other hand, consumer welfare (rate) is given by

$$U^{J^{J(\beta)}_{R(1-\beta)}nR} = \lambda\sum_{x=0}^{n}\pi_x^{J^{J(\beta)}_{R(1-\beta)}nR}U_x^{J^{J(\beta)}_{R(1-\beta)}nR}. \tag{35}$$

The retrial probability under $J^{J(\beta)}_{R(1-\beta)}nR$ is

$$\pi_R^{J^{J(\beta)}_{R(1-\beta)}nR} = (1-\beta)\pi_{n-1}^{J^{J(\beta)}_{R(1-\beta)}nR} + \pi_n^{J^{J(\beta)}_{R(1-\beta)}nR}.$$

According to (15), we have $\rho^{J^{J(\beta)}_{R(1-\beta)}nR} = \frac{l}{1-\pi_R^{J^{J(\beta)}_{R(1-\beta)}nR}}$ and

$$\begin{aligned}
\pi_1^{J^{J(\beta)}_{R(1-\beta)}nR} &= \rho^{J^{J(\beta)}_{R(1-\beta)}nR}\pi_0^{J^{J(\beta)}_{R(1-\beta)}nR};\\
\pi_2^{J^{J(\beta)}_{R(1-\beta)}nR} &= \rho^{J^{J(\beta)}_{R(1-\beta)}nR}\pi_1^{J^{J(\beta)}_{R(1-\beta)}nR};\\
\vdots &= \vdots\\
\pi_{n-2}^{J^{J(\beta)}_{R(1-\beta)}nR} &= \rho^{J^{J(\beta)}_{R(1-\beta)}nR}\pi_{n-3}^{J^{J(\beta)}_{R(1-\beta)}nR};\\
\pi_{n-1}^{J^{J(\beta)}_{R(1-\beta)}nR} &= \rho^{J^{J(\beta)}_{R(1-\beta)}nR}\pi_{n-2}^{J^{J(\beta)}_{R(1-\beta)}nR};\\
\pi_n^{J^{J(\beta)}_{R(1-\beta)}nR} &= \beta\rho^{J^{J(\beta)}_{R(1-\beta)}nR}\pi_{n-1}^{J^{J(\beta)}_{R(1-\beta)}nR}.
\end{aligned}$$

It follows that for fixed $n: 1 \le n \le N$, $\rho^{J^{J(\beta)}_{R(1-\beta)}nR}$ decreases in $\beta: 0 \to 1$ due to (i)

$$\begin{aligned}
&\pi_0^{J^{J(\beta)}_{R(1-\beta)}nR} + \pi_1^{J^{J(\beta)}_{R(1-\beta)}nR} + \ldots + \pi_n^{J^{J(\beta)}_{R(1-\beta)}nR}\\
=&\pi_0^{J^{J(\beta)}_{R(1-\beta)}nR}[\rho^{J^{J(\beta)}_{R(1-\beta)}nR} + (\rho^{J^{J(\beta)}_{R(1-\beta)}nR})^2 + \ldots + (\rho^{J^{J(\beta)}_{R(1-\beta)}nR})^{n-1} + \beta(\rho^{J^{J(\beta)}_{R(1-\beta)}nR})^n] = 1,
\end{aligned}$$

and (ii) $\pi_0^{J^{J(\beta)}_{R(1-\beta)}nR} \equiv 1-l$ no matter what β is (and in fact what n is), because in a non-balking system (i.e., $\pi_B^{J^{J(\beta)}_{R(1-\beta)}nR} = 0$), the server's long-run idle probability is always equal to one minus the utility rate. As a result, consumer welfare in (35) decreases in β because the payoffs $U_x^{J^{J(\beta)}_{R(1-\beta)}nR}$'s are decreasing in state x, i.e., $U_0^{J^{J(\beta)}_{R(1-\beta)}nR} > U_1^{J^{J(\beta)}_{R(1-\beta)}nR} > \ldots > U_{n-2}^{J^{J(\beta)}_{R(1-\beta)}nR} > U_{n-1}^{J^{J(\beta)}_{R(1-\beta)}nR} = U_n^{J^{J(\beta)}_{R(1-\beta)}nR}$.

As a side result, we also notice that the steady-state retrial probability

$$\pi_R^{J^{J(\beta)}_{R(1-\beta)}nR} = (1-\beta)\pi_{n-1}^{J^{J(\beta)}_{R(1-\beta)}nR} + \pi_n^{J^{J(\beta)}_{R(1-\beta)}nR} = [(1-\beta)(\rho^{J^{J(\beta)}_{R(1-\beta)}nR})^{n-1} + \beta(\rho^{J^{J(\beta)}_{R(1-\beta)}nR})^n]\pi_0^{J^{J(\beta)}_{R(1-\beta)}nR}$$

decreases in β. This is because for fixed $\rho^{J^{J(\beta)}_{R(1-\beta)}nR}$, $(1-\beta)(\rho^{J^{J(\beta)}_{R(1-\beta)}nR})^{n-1} + \beta(\rho^{J^{J(\beta)}_{R(1-\beta)}nR})^n$ decreases in β. And now since $\rho^{J^{J(\beta)}_{R(1-\beta)}nR}$ decreases in β, so $(1-\beta)(\rho^{J^{J(\beta)}_{R(1-\beta)}nR})^{n-1} + \beta(\rho^{J^{J(\beta)}_{R(1-\beta)}nR})^n$ further decreases as β increases. This can be formally proved by taking derivatives to show $\frac{\partial \pi_R^{J^{J(\beta)}_{R(1-\beta)}nR}}{\partial\beta} < 0$. □

LEMMA 12. *Fix $n: 1 \le n \le N$. Consumer welfare under the equilibrium join/retry strategy $J^{J(\beta)}_{R(1-\beta)}nR$, i.e., $U^{J^{J(\beta)}_{R(1-\beta)}nR}$ in Lemma 11, is also equal to*

$$\lambda v - \lambda c W^{J^{J(\beta)}_{R(1-\beta)}nR} - \lambda\left(\frac{1}{1-\pi_R^{J^{J(\beta)}_{R(1-\beta)}nR}} - 1\right)\alpha \tag{36}$$

where $\alpha = \alpha(\beta)$ is the quantity that induces the equilibrium join/retry strategy $J^{J(\beta)}_{R(1-\beta)}nR$.

Proof of Lemma 12: We show below that (36) equals $U^{J^{J(\beta)}_{R(1-\beta)}nR}$ via equation (35). Since α induces the equilibrium policy $J^{J(\beta)}_{R(1-\beta)}nR$, we have from (33) that $v - \frac{c}{\mu}(n) = v - cW^{J^{J(\beta)}_{R(1-\beta)}nR} - \frac{\alpha}{1-\pi_R^{J^{J(\beta)}_{R(1-\beta)}nR}}$. Using $\sigma = J^{J(\beta)}_{R(1-\beta)}nR$ in the rest of the proof, it then follows that

$$U^\sigma = \lambda\sum_{x=0}^{n}\pi_x^\sigma U_x^\sigma$$

$$
\begin{aligned}
&=\lambda\{\pi_0^\sigma[v-\frac{c}{\mu}(1)]+\ldots+\pi_{n-2}^\sigma[v-\frac{c}{\mu}(n-1)]+\beta\pi_{n-1}^\sigma[v-\frac{c}{\mu}(n)]+\pi_R^\sigma[v-\frac{c}{\mu}(n)]\}\\
&=\lambda\{\pi_0^\sigma[v-\frac{c}{\mu}(1)]+\ldots+\pi_{n-2}^\sigma[v-\frac{c}{\mu}(n-1)]+\beta\pi_{n-1}^\sigma[v-\frac{c}{\mu}(n)]+\pi_R^\sigma[v-cW-\frac{\alpha}{1-\pi_R^\sigma}]\}\\
&=\lambda\{(\pi_0^\sigma+\ldots+\pi_{n-1}^\sigma+\beta\pi_{n-1}^\sigma+\pi_R^\sigma)v\\
&\quad-[\pi_0^\sigma\frac{c}{\mu}(1)+\ldots+\pi_{n-2}^\sigma\frac{c}{\mu}(n-1)+\beta\pi_{n-1}^\sigma\frac{c}{\mu}(n)+\pi_R^\sigma cW^\sigma]-\frac{\pi_R^\sigma}{1-\pi_R^\sigma}\alpha\}\\
&=\lambda v-\lambda[(1-\pi_R^\sigma)cW^\sigma+\pi_R cW^\sigma]-\lambda\frac{\pi_R}{1-\pi_R}\alpha\\
&\quad(\text{because } W^\sigma=\frac{\pi_0^\sigma}{1-\pi_R^\sigma}\frac{1}{\mu}+\frac{\pi_1^\sigma}{1-\pi_R^\sigma}\frac{2}{\mu}+\ldots+\frac{\pi_{n-2}^\sigma}{1-\pi_R^\sigma}\frac{n-1}{\mu}+\frac{\beta\pi_{n-1}^\sigma}{1-\pi_R^\sigma}\frac{n}{\mu})\\
&=\lambda v-\lambda cW^\sigma-\lambda\frac{\pi_R^\sigma}{1-\pi_R^\sigma}\alpha
\end{aligned}
$$

which is equal to the expression in (36). □

Proof of Lemma 10: As an immediate consequence to Lemmas 11 and 12 proved above, at an equilibrium strategy $J_{R(1-\beta)}^{J(\beta)}nR$, we must have

$$\frac{\partial[v-cW^{J_{R(1-\beta)}^{J(\beta)}nR}-\frac{\alpha}{1-\pi_R^{J_{R(1-\beta)}^{J(\beta)}nR}}]}{\partial\beta}+\frac{\partial\alpha}{\partial\beta}<0. \tag{37}$$

This is because the partial derivative of $\lambda v-\lambda cW^{J_{R(1-\beta)}^{J(\beta)}nR}-\lambda\frac{1}{1-\pi_R^{J_{R(1-\beta)}^{J(\beta)}nR}}\alpha+\lambda\alpha$ in (36) with respect to β is negative and λ is only a constant.

Fix $n\in\{2,\ldots,N-1,N\}$. Let $\alpha(\beta)$ be the retrial cost that induces the equilibrium strategy $J_{R(1-\beta)}^{J(\beta)}nR$. Under any equilibrium strategy $J_{R(1-\beta)}^{J(\beta)}nR$, the retrial payoff must be the same as the joining payoff at state $n-1$, i.e.,

$$v-\frac{c}{\mu}(n)=v-cW^{J_{R(1-\beta)}^{J(\beta)}nR}-\frac{\alpha}{1-\pi_R^{J_{R(1-\beta)}^{J(\beta)}nR}}. \tag{38}$$

It follows from (38) that

$$\frac{\partial[v-cW^{J_{R(1-\beta)}^{J(\beta)}nR}-\frac{\alpha}{1-\pi_R^{J_{R(1-\beta)}^{J(\beta)}nR}}]}{\partial\beta}=0. \tag{39}$$

Comparing (39) to (37) gives us that $\frac{\partial\alpha}{\partial\beta}<0$, i.e., when β increases from 0 to 1, the unique retrial cost that induces the equilibrium policy $J_{R(1-\beta)}^{J(\beta)}nR$ decreases continuously in β.

On the other hand, we have showed in Lemma 8 that α_{n-1} induces $J(n-1)R$ or simply $J_{R(1)}^{J(0)}nR$, and $\alpha_n-\frac{cl}{\mu\rho^{JnR}}$ induces JnR or simply $J_{R(0)}^{J(1)}nR$ with binding indifference conditions, i.e.,

$$
\begin{aligned}
v-\frac{c}{\mu}(n)&=v-cW^{J_{R(1)}^{J(0)}nR}-\frac{\alpha_{n-1}}{1-\pi_{n-1}^{J_{R(1)}^{J(0)}nR}}\\
v-\frac{c}{\mu}(n)&=v-cW^{J_{R(0)}^{J(1)}nR}-\frac{\alpha_n-\frac{cl}{\mu\rho^{JnR}}}{1-\pi_n^{J_{R(0)}^{J(1)}nR}}
\end{aligned}
$$

Therefore, when β increase from 0 to 1, we have an equilibrium strategy $J_{R(1-\beta)}^{J(\beta)}nR$ for some unique retrial cost α that is decreasing from α_{n-1} to $\alpha_n-\frac{cl}{\mu\rho^{JnR}}$. It follows that $\alpha_n-\frac{cl}{\mu\rho^{JnR}}<\alpha_{n-1}$. □

A direct consequence of Lemma 10 is that the N intervals of the retrial cost constructed in Lemma 8, on which $J1R, J2R,\ldots,JNR$ form equilibria, actually overlap with each other. Therefore, for any retrial

cost $\alpha \le \alpha_L$, it induces at least one retry strategy (according to Lemma 8). On the other hand, for $n \in \{2,3,\ldots,N\}$, both retry strategies $J(n-1)R$ and $J(n)R$ are equilibrium strategies on the overlapped region $[\alpha_n - \frac{cl}{\mu\rho^{JnR}}, \alpha_{n-1}]$, plus there exist equilibrium join/retry strategies given by Lemma 10. Fortunately, in the following lemma we can establish the uniqueness of a Pareto-dominant equilibrium strategy for every $\alpha \in (0, \alpha_L]$, i.e., the equilibrium strategy that gives the highest consumer welfare as a whole or per capita.

LEMMA 13. *For any retrial cost $\alpha \le \alpha_L$, we have one or more symmetric equilibria that are of the retry or join/retry types. However, there exists a unique Pareto-dominant equilibrium: For α that falls in one of the N intervals denoted by $\{I_n\}_{n=1,2,\ldots,N}$ where $I_1 \triangleq (0,\alpha_1], I_2 \triangleq (\alpha_1,\alpha_2], \ldots, I_{N-1} \triangleq (\alpha_{N-2},\alpha_{N-1}]$ and $I_N \triangleq (\alpha_{N-1},\alpha_L]$, the corresponding retry strategy $\{JnR\}_{n=1,2,\ldots,N}$ is the Pareto-dominant equilibrium.*

Proof of Lemma 13: From Lemmas 8 and 10, we see that for any $\alpha \le \alpha_L$, there exists at least one equilibrium strategy in the forms of a retry or a join/retry strategy. The lemmas tell us that on the region of I_n, the equilibrium strategies certainly include retry strategy JnR and join/retry strategies $J^{J(\beta)}_{R(1-\beta)}nR$. If there are any other equilibrium strategies of a retry or a join/retry strategy on this region, say JmR or $J^{J(\beta)}_{R(1-\beta)}mR$, then it must be the case that $m > n$.

Therefore, to show the retry strategy JnR is the Pareto-dominant equilibrium for all $\alpha \in I_n$, it is equivalent to show that JnR generates the highest welfare among the family of strategies $\{J^{J(\beta)}_{R(1-\beta)}mR : m > n, 0 \le \beta \le 1\}$. With Lemma 11 in mind, it suffices to show that $U^{J^{J(0)}_{R(1)}(n+1)R} > U^{J^{J(0)}_{R(1)}(n+2)R} > U^{J^{J(0)}_{R(1)}(n+3)R}, \ldots$, or equivalently $U^{JnR} > U^{J(n+1)R} > U^{J(n+2)R}, \ldots$. We will show next that for all $k \in \{n, n+1, \ldots\}$, $U^{JkR} > U^{J(k+1)R}$. Recall

$$U^{JkR} = \lambda \sum_{x=0}^{k} \pi_x^{JkR} U_x^{JkR}$$
$$U^{J(k+1)R} = \lambda \sum_{x=0}^{k+1} \pi_x^{J(k+1)R} U_x^{J(k+1)R}$$

where

$$\begin{array}{ccccccccc} U_0^{JkR} & > U_1^{JkR} & > U_2^{JkR} & > \ldots > & U_{k-1}^{JkR} & \ge U_k^{JkR} \\ || & || & || & || & \vee \\ U_0^{J(k+1)R} > U_1^{J(k+1)R} > U_2^{J(k+1)R} > \ldots > U_{k-1}^{J(k+1)R} > U_k^{J(k+1)R} \ge U_{k+1}^{J(k+1)R}. \end{array}$$

Since $\pi_0^{JkR} = \pi_0^{J(k+1)R} = 1-l$ and $\rho^{JkR} > \rho^{J(k+1)R}$, we can conclude that $U^{JkR} > U^{J(k+1)R}$. □

As the final key missing step to complete the proof to Theorem 2, we show in Lemma 14 that $\alpha_L < \alpha_H$, and characterize the equilibrium strategies for $\alpha \in [\alpha_L, \alpha_H]$ in Lemma 15 and Lemma 16.

LEMMA 14. $\alpha_L < \alpha_H$

Proof of Lemma 14: Recall by definition that

$$\alpha_L = (1 - \pi_N^{JNR})(v - cW^{JNR}), \tag{40}$$

$$\alpha_H = (1-\pi_N^{JNB})(v - cW^{JNB}). \tag{41}$$

where W^{JNR} and W^{JNB} are the expected waiting time conditional on joining, given by

$$W^{JNR} = \frac{\pi_0^{JNR}}{1-\pi_N^{JNR}}\frac{1}{\mu} + \frac{\pi_1^{JNR}}{1-\pi_N^{JNR}}\frac{2}{\mu} + \ldots + \frac{\pi_{N-1}^{JNR}}{1-\pi_N^{JNR}}\frac{N}{\mu} \tag{42}$$

$$W^{JNB} = \frac{\pi_0^{JNB}}{1-\pi_N^{JNB}}\frac{1}{\mu} + \frac{\pi_1^{JNB}}{1-\pi_N^{JNB}}\frac{2}{\mu} + \ldots + \frac{\pi_{N-1}^{JNB}}{1-\pi_N^{JNB}}\frac{N}{\mu} \tag{43}$$

The underlying queueing system under both JNR and JNB is $M/M/1/N$. The birth rate is bigger in the system under JNR than that under JNB, i.e., $\lambda^{JNR} = \frac{\lambda}{1-\pi_N^{JNR}} > \lambda = \lambda^{JNB}$. The death rates are the same in both systems, namely μ. Therefore, $\rho^{JNR} > \rho^{JNB}$. As a result,

$$\pi_0^{JNR} = \frac{1}{1+\rho^{JNR}+(\rho^{JNR})^2+\ldots+(\rho^{JNR})^N} < \frac{1}{1+\rho^{JNB}+(\rho^{JNB})^2+\ldots+(\rho^{JNB})^N} = \pi_0^{JNB}.$$

That is, the server's long-run idle rate is higher with a system that has balking consumers compared to one that does not. Now suppose $\pi_n^{JNR} > \pi_n^{JNB}$ for some $n \in \{1,2,\ldots,N\}$, then

$$\pi_x^{JNR} = \pi_n^{JNR}(\rho^{JNR})^{x-n} > \pi_n^{JNB)}(\rho^{JNB})^{x-n} = \pi_x^{JNB}$$

for all $x \in \{n, n+1, \ldots, N\}$. Since $\sum_{x=1}^{N}\pi_x^{JNR} = \sum_{x=1}^{N}\pi_x^{JNB} = 1$, we must have $\pi_N^{JNR} > \pi_N^{JNB}$.

We observe that the conditional probabilities in equations (42) and (43) sum up to one and form geometric progressions with ratios ρ^{JNR} and ρ^{JNB}, respectively, i.e., for $s \in \{JNR, JNB\}$,

$$\frac{\pi_0^s}{1-\pi_N^s} + \frac{\pi_1^s}{1-\pi_N^s} + \ldots + \frac{\pi_{N-1}^s}{1-\pi_N^s} = 1$$

$$\frac{\pi_{N-1}^s}{1-\pi_N^s} = \rho^s\frac{\pi_{N-2}^s}{1-\pi_N^s} = (\rho^s)^2\frac{\pi_{N-2}^s}{1-\pi_N^s} = \ldots = (\rho^s)^{(N-1)}\frac{\pi_0^s}{1-\pi_N^s}$$

It then follows from $\rho^{JNR} > \rho^{JNB}$ and the structure of W that $W^{JNR} > W^{JNB}$.

Since $\pi_N^{JNR} > \pi_N^{JNB}$ and $W^{JNR} > W^{JNB}$, we can tell from (40) and (41) that $\alpha_H > \alpha_L$. □

Next, we claim that the retry/balk strategies (i.e., JN_B^R-type) will form equilibria for any retrial cost $\alpha \in (\alpha_L, \alpha_H)$. In other words, when the retrial hassle is moderate, at equilibrium consumers join the queue at states $\{0,1,2,\ldots,N-1\}$, then mix retry and balk decisions at state N. Note that, when the population adopts the retry/balk strategy $JN_{B(\gamma)}^{R(1-\gamma)}$, the steady-state joining, retrial and baling probabilities are given by

$$\pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}} = 1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}, \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}} = (1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}, \text{ and } \pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}} = \gamma\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}. \tag{44}$$

LEMMA 15. *The retry/balk strategy $JN_{B(\gamma)}^{R(1-\gamma)}$ is an equilibrium strategy for the unique retrial cost*

$$\alpha = (1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}})(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) \tag{45}$$

where $\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}$ is the steady-state probability on state N or the steady-state not-joining probability (balking or retrial) under $JN_{B(\gamma)}^{R(1-\gamma)}$, and $W^{JN_{B(\gamma)}^{R(1-\gamma)}}$ is the expected waiting time for a consumer conditional on joining the system under $JN_{B(\gamma)}^{R(1-\gamma)}$. $W^{JN_{B(\gamma)}^{R(1-\gamma)}}$ can be written as

$$W^{JN_{B(\gamma)}^{R(1-\gamma)}} = \frac{\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}\frac{1}{\mu} + \frac{\pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}\frac{2}{\mu} + \ldots + \frac{\pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}\frac{N}{\mu}.$$

Proof of Lemma 15: We note that the underlying system is $M/M/1/N$ when the retry/balk strategy $JN_{B(\gamma)}^{R(1-\gamma)}$ is adopted by the population. With retrial cost α, the expected retrial payoff is given by

$$\begin{aligned}
&\pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}}(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}}-\alpha)-\pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}}\alpha+\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}\pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}}(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}}-2\alpha)\\
&-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}\pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}}2\alpha+(\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}})^2\pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}}(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}}-3\alpha)-(\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}})^2\pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}}3\alpha+\ldots\\
=&\frac{\pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}}(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}})-\frac{\pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}}}{(1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}})^2}\alpha-\frac{\pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}}}{(1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}})^2}\alpha\\
=&\frac{\pi_J^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}}(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}})-\frac{1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}}{(1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}})^2}\alpha\\
=&\frac{1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}}(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}})-\frac{\alpha}{1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}} \qquad (46)
\end{aligned}$$

where

$$W^{JN_{B(\gamma)}^{R(1-\gamma)}}=\frac{\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}\frac{1}{\mu}+\frac{\pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}\frac{2}{\mu}+\ldots+\frac{\pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}\frac{N}{\mu}.$$

The strategy $JN_{B(\gamma)}^{R(1-\gamma)}$ will be an equilibrium of the system if and only the retrial payoff under $JN_{B(\gamma)}^{R(1-\gamma)}$ is exactly zero, i.e., expression (46) is zero. Therefore, the unique retrial cost that induces $JN_{B(\gamma)}^{R(1-\gamma)}$ as an equilibrium is a solution to

$$\frac{1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}}(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}})-\frac{\alpha}{1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}}=0,$$

or simply $\alpha=(1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}})(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}})$ which is given by (45). □

Since there exists a unique retrial cost α that induces the equilibrium strategy $JN_{B(\gamma)}^{R(1-\gamma)}$ for every $\gamma\in[0,1]$, we can write α as a function of γ. Moreover, since $\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}$ and $W^{JN_{B(\gamma)}^{R(1-\gamma)}}$ are both continuous in γ, by (45) we see that $\alpha(\gamma)$ is a continuous function in γ. It then can be shown in the proof of Lemma 16 below that $\frac{\partial\alpha}{\partial\gamma}>0$. We can then conclude that γ, for which $JN_{B(\gamma)}^{R(1-\gamma)}$ forms an equilibrium strategy for some α, is also a function of and increases in α.

LEMMA 16. *For any retrial cost $\alpha\in(\alpha_L,\alpha_H)$, there exists a unique equilibrium strategy for the system which is the retry/balk strategy $JN_{B(\gamma)}^{R(1-\gamma)}$. Moreover, when the retrial cost α increases from α_L to α_H, the corresponding γ in the equilibrium strategy increases from 0 to 1.*

Proof of Lemma 16: Suppose the consumer population adopts the retry/balk strategy $JN_{B(\gamma)}^{R(1-\gamma)}$ and γ increases from 0 to 1. The underlying system remains $M/M/1/N$. By a similar proof to "$\alpha_L<\alpha_H$", it is easy to see that $\rho^{JN_{B(\gamma)}^{R(1-\gamma)}}$, $\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}$ and $W^{JN_{B(\gamma)}^{R(1-\gamma)}}$ all continuously decrease in γ. Therefore, by (45), the retrial cost $\alpha(\gamma)$ that induces the equilibrium strategy $JN_{B(\gamma)}^{R(1-\gamma)}$ increases continuously in γ. Furthermore, when $\gamma=0$, the retry/balk strategy $JN_{B(0)}^{R(1)}$ is equivalent to the retrial strategy JNR and equation (45) is simply equation (40), so $\alpha(0)=\alpha_L$. Similarly, when $\gamma=1$, the retry/balk strategy $JN_{B(1)}^{R(0)}$ is equivalent to the balk strategy JNB and equation (45) becomes equation (41), and so $\alpha(1)=\alpha_H$.

We have argued that for each $\gamma \in (0,1)$, there exists a unique retrial cost $\alpha \in (\alpha_L, \alpha_H)$ such that $JN_{B(\gamma)}^{R(1-\gamma)}$ is an equilibrium strategy. Due to the continuity and strict monotonicity of $\alpha(\gamma)$ in γ, we can conclude that for each retrial cost $\alpha \in (\alpha_L, \alpha_H)$, there exists a unique equilibrium retry/balk strategy $JN_{B(\gamma)}^{R(1-\gamma)}$ where γ increases in α (from 0 to 1). Note that there does not exist other types of equilibrium strategy when $\alpha \in (\alpha_L, \alpha_H)$ because equilibrium retry and join/retry strategies appear for $\alpha \leq \alpha_L$ and equilibrium balk strategy appears for $\alpha \geq \alpha_H$. Therefore, the equilibrium retry/balk strategy is unique. □

Lemma 16 fills in the last missing piece of our equilibrium analysis. It states that (i) the unique equilibrium strategy for any retrial cost $\alpha \in (\alpha_L, \alpha_H)$ is given by a retry/balk strategy. (ii) when $\alpha \to \alpha_L$, consumers adopt the retry strategy JNR (or $JN_{B(0)}^{R(1)}$) at equilibrium; when $\alpha \to \alpha_H$, consumers adopt the balking strategy JNB (or $JN_{B(1)}^{R(0)}$); (iii) when $\alpha \in (\alpha_L, \alpha_H)$, consumers adopt the retry strategy JNR with probability $1-\gamma$ and the balk strategy JNB with probability γ at equilibrium, but the likelihood of using the balk strategy (i.e., γ) increases in the retrial cost. We have now completed the proof of Theorem 2.

Proof of Lemma 1: By the proof of Lemma 13, we know that when $\alpha \leq \alpha_L$, consumer welfare (of the Pareto-dominant retry strategy JnR) decreases in the retrial cost α. The welfare is linearly decreasing on each region of $(0, \alpha_1], (\alpha_1, \alpha_2], \ldots, (\alpha_{N-1}, \alpha_L]$ with flatter and flatter slope because $\frac{1}{1-\pi_n^{JnR}}$ decreases in n. Therefore, the highest welfare is achieved when $\alpha \to 0$ and the lowest when $\alpha \to \alpha_L$. Recall from (28) that the average number of consumers in the system is

$$L = \frac{\rho}{1-\rho} - \frac{(n+1)\rho^{n+1}}{1-\rho^{n+1}} = \frac{\rho}{1-\rho} - [\log_\rho(\frac{\rho - l}{1-l})]\frac{\rho - l}{1-\rho}. \tag{47}$$

It can be verified that (47) decreases in ρ on $\rho > l$. In fact, $\lim_{\rho \to l} L = \frac{l}{1-l}$ and $\lim_{\rho \to \frac{l}{1-l}} L = l$.

When α goes from 0 to α_L, the threshold of the equilibrium retry strategy JnR increases from 1 to N, and ρ^{JnR} decreases from $\rho^{J1R} = \frac{l}{1-l}$ to some $\rho^{JNR} > l$. (When $J1R$ is an equilibrium strategy, $\rho^{J1R} = \frac{l}{1-\pi_1^{J1R}} = \frac{l}{1-l}$.) Therefore, when $\alpha \to 0$, the equilibrium strategy is $J1R$ with $L^{J1R} = \pi_1^{J1R} = l$ and $\pi_0^{J1R} = 1-l$. It then follows from (7) that consumer welfare equals $\lambda v - cl$. On the other hand, when $\alpha = \alpha_L$, consumer welfare is $\lambda v - cL^{JNR} - \lambda(\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_L$. □

Proof of Lemma 2: By (11), consumer welfare equals

$$\begin{aligned}
&\frac{\lambda(1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}})}{1-(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{\lambda(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}\alpha \\
=&\frac{\lambda}{1-(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}(1-\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}})(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{\lambda(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}\alpha
\end{aligned}$$

which by (45) is equal to

$$\frac{\lambda}{1-(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}\alpha - \frac{\lambda(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-(1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}\alpha = \lambda\alpha. \tag{48}$$

As (48) clearly increases in α and by considering the two extreme scenarios, our claim is proved.

Recall from (40) and (41) that $\alpha_L = (v - cW^{JNR})(1 - \pi_N^{JNR})$ and $\alpha_H = (v - cW^{JNB})(1 - \pi_N^{JNB})$, it is then a trivial verification that

$$\lambda \cdot \alpha_L = \lambda v - \lambda c W^{JNR} - \lambda(\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_L = \lambda v - cL^{JNR} - \lambda(\frac{1}{1-\pi_N^{JNR}} - 1)\alpha_L,$$
$$\lambda \cdot \alpha_H = \lambda v(1 - \pi_N^{JNB}) - \lambda(1 - \pi_N^{JNB})cW^{JNB} = \lambda v(1 - \pi_N^{JNB}) - cL^{JNB}.$$

Here's an alternative proof. Since $\alpha(\beta)$ always adjust to make the retrial payoff of the strategy $JN_{B(\gamma)}^{R(1-\gamma)}$ zero, the welfare can also be given as

$$\begin{aligned}
&U^{JN_{B(\gamma)}^{R(1-\gamma)}} \\
=&\lambda[\sum_{x=0}^{N} \pi_x^{JN_{B(\gamma)}^{R(1-\gamma)}} U_x^{JN_{B(\gamma)}^{R(1-\gamma)}}] \\
=&\lambda[\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{c}{\mu}) + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{2c}{\mu}) + \ldots + \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{Nc}{\mu}) \\
&+ \gamma\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot \text{ balking payoff} + (1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot \text{ retrial payoff}] \\
=&\lambda[\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{c}{\mu}) + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{2c}{\mu}) + \ldots + \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{Nc}{\mu}) + \gamma\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot 0 + (1-\gamma)\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot 0] \\
=&\lambda[\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{c}{\mu}) + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{2c}{\mu}) + \ldots + \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{Nc}{\mu})] \\
=&\lambda[(1 - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}})v - (1 - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}})cW^{JN_{B(\gamma)}^{R(1-\gamma)}}] \\
=&\lambda\alpha
\end{aligned}$$

Proof of Lemma 3: We have provided an argument in the paper. Here, we provide a more formal proof to show that $\lambda v - cl$ is greater than Naor's consumer welfare given in (12), i.e., $\lambda v - cl > \lambda v(1 - \pi_N^{JNB}) - cL^{JNB}$, or equivalently, $\pi_N^{JNB}\lambda v > cl - cL^{JNB}$. Recall that $v > N\frac{c}{\mu}$, therefore it is sufficient to show that

$$\begin{aligned}
&\pi_N^{JNB}\lambda N\frac{c}{\mu} > cl - cL^{JNB} \Leftrightarrow \ \pi_N^{JNB} lN > l - L^{JNB} \Leftrightarrow \ \frac{(1-l)l^N}{1-l^{N+1}} lN > l - \frac{l}{1-l} + (N+1)\frac{l^{N+1}}{1-l^{N+1}} \\
\Leftrightarrow \ &\frac{l}{1-l} - l > [(N+1) - (1-l)N]\frac{l^{N+1}}{1-l^{N+1}} \Leftrightarrow \ \frac{l^2}{1-l} > (1+lN)\frac{l^{N+1}}{1-l^{N+1}} \qquad (49)
\end{aligned}$$

Since the RHS of expression (49) is decreasing in N and equals to $\frac{l^2}{1-l}$ when $N = 1$ (the smallest possible value), we know (49) holds and therefore the claim is proved. □

Proof of Proposition 2: Results follow from Lemma 1, Lemma 2 and Lemma 3. □

Proof of Theorem 3: We take the graph on the results of the equilibrium welfare in Figure 2, and extend all the N line segments on $\alpha \in (0, \alpha_L]$ into N lines, and denote them as $L_1, L_2, \ldots, L_N$, respectively, see Figure 6. The original N line segments correspond to the welfare under the retry strategy $J1R$ for $\alpha \in (0, \alpha_1]$, welfare under the retry strategy $J2R$ for $\alpha \in (\alpha_1, \alpha_2]$,..., and welfare under the retry strategy JNR for $\alpha \in (\alpha_{N-1}, \alpha_L]$, etc. With the extension, the full line L_n for $n = 1, 2, \ldots, N$ would represent the welfare under retry strategies JnR for all $\alpha \in (0, \infty)$.

On the other hand, if we denote L the welfare under $JN'B$, i.e., the socially optimal policy in Naor (1969), we know it is a straight line describing a constant function of the retrial cost α. Naor (1969) showed that

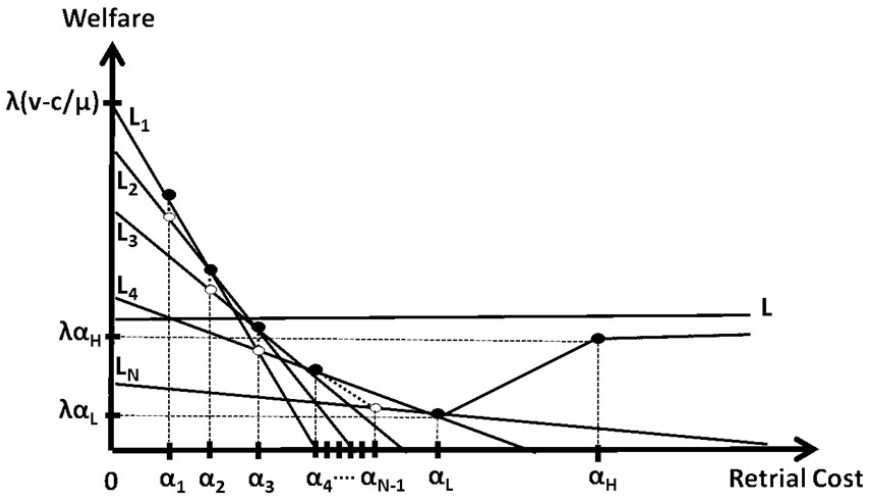


Figure 6 Illustration of the welfare under socially optimal policies (the upper envelope of $L_1, L_2, \ldots, L_N$ and L).

$N' < N$ and consumer welfare under $JN'B$ exceeds that under JNB which is equal to $\lambda\alpha_H$. In fact, welfare curve under any balk strategy (i.e., JnB-type) will be a horizontal line in the graph, because there are no retrials at equilibrium. Since $JN'B$ generates the highest welfare among all balk strategies, $\{JnB : n \geq 1\}$, the upper envelope of $L_1, L_2, \ldots, L_N$ and L will give the socially optimal welfare as a function of the retrial cost α among the class of pure threshold policies that we study, namely $\{s : s = JnR$ or $s = JnB$ for some $n \geq 1\}$. Result then follows. □

Proof of Lemma 4: The queueing system under a retry/balk strategy σ will have the following birth and death rates:

$$\begin{aligned}
\pi_1^\sigma &= \rho^\sigma \pi_0^\sigma; \\
\pi_2^\sigma &= \rho^\sigma \pi_1^\sigma; \\
\pi_3^\sigma &= \rho^\sigma \pi_2^\sigma; \\
\vdots &= \vdots \\
\pi_N^\sigma &= \rho^\sigma \pi_{N-1}^\sigma; \\
\pi_{N+1}^\sigma &= \theta l \pi_N^\sigma; \\
\pi_{N+2}^\sigma &= \theta l \pi_{N+1}^\sigma; \\
\vdots &= \vdots
\end{aligned}$$

where

$$\rho^\sigma = \theta l + (1-\theta)\frac{l}{1-\pi_R^\sigma} \tag{50}$$

$$\pi_B^\sigma + \pi_R^\sigma = \pi_N^\sigma + \pi_{N+1}^\sigma + \pi_{N+2}^\sigma + \ldots = \frac{\pi_N^\sigma}{1-\theta l} \tag{51}$$

If the underlying queueing systems when the population adopts retry/balk strategy σ_1 or σ_2, share the same steady-state balking and retrial probabilities, i.e., if $\pi_B^{\sigma_1} = \pi_B^{\sigma_2}$ and $\pi_R^{\sigma_1} = \pi_R^{\sigma_2}$, then by (50) and (51),

$\rho_R^{\sigma_1}=\rho_R^{\sigma_2}$ and $\pi_N^{\sigma_1}=\pi_N^{\sigma_2}$. It will be true then $\pi_x^{\sigma_1}=\pi_x^{\sigma_2}$ for all $x\in\mathbb{N}_0$. Therefore, given two distinct strategies $\sigma_1,\sigma_2\in JN_{B(\gamma)}^{R(1-\gamma)}$ where $\frac{\pi_B^{\sigma_1}}{\pi_B^{\sigma_1}+\pi_R^{\sigma_1}}=\frac{\pi_B^{\sigma_2}}{\pi_B^{\sigma_2}+\pi_R^{\sigma_2}}=\gamma\Rightarrow\pi_B^{\sigma_1}\pi_R^{\sigma_2}=\pi_B^{\sigma_2}\pi_R^{\sigma_1}$, we will only need to show that $\pi_B^{\sigma_1}=\pi_B^{\sigma_2}$ and $\pi_R^{\sigma_1}=\pi_R^{\sigma_2}$.

WLOG, assume that $\pi_B^{\sigma_1}<\pi_B^{\sigma_2}$ and $\pi_R^{\sigma_1}<\pi_R^{\sigma_2}$. (Signs need to be the same for $\pi_B^{\sigma_1}\pi_R^{\sigma_2}=\pi_B^{\sigma_2}\pi_R^{\sigma_1}$ to hold.) According to (50) and (51), $\rho_R^{\sigma_1}<\rho_R^{\sigma_2}$ and $\pi_N^{\sigma_1}<\pi_N^{\sigma_2}$. Also, since

$$\pi_0^{\sigma}=1-\frac{1-\pi_R^{\sigma}-\pi_B^{\sigma}}{1-\pi_R^{\sigma}}l=1-(1-\frac{\pi_B^{\sigma}}{1-\pi_R^{\sigma}})l,$$

we must have $\pi_0^{\sigma_1}<\pi_0^{\sigma_2}$. Then, $\pi_x^{\sigma_1}<\pi_x^{\sigma_2}$ for all $x\in\mathbb{N}_0$. And this is a contradiction to the fact that $\sum_{x\in\mathbb{N}_0}\pi_x^{\sigma_1}=\sum_{x\in\mathbb{N}_0}\pi_x^{\sigma_2}=1$. □

Proof of Theorem 2': Proof is similar to that of Theorem 2 for the one-class case. Except for retry strategies, the new stability condition is

$$\rho^{JnR}=\theta l+(1-\theta)\frac{l}{1-\pi_R^{JnR}}\Leftrightarrow\rho^{JnR}-\theta l=\frac{(1-\theta l)(l-\theta l)}{1-\theta l-(1-l)(\rho^{JnR})^n} \tag{52}$$

$$\text{since }\pi_R^{JnR}=\frac{\pi_0^{JnR}(\rho^{JnR})^n}{1-\theta l}=\frac{(1-l)(\rho^{JnR})^n}{1-\theta l}\Leftrightarrow(1-l)(\rho^{JnR})^n=\frac{(1-\theta l)(\rho^{JnR}-l)}{\rho^{JnR}-\theta l}$$

$$\Leftrightarrow(\rho^{JnR})^n=\frac{(1-\theta l)(\rho^{JnR}-l)}{(\rho^{JnR}-\theta l)(1-l)}\Leftrightarrow n=\frac{\ln\frac{(1-\theta l)(\rho^{JnR}-l)}{(\rho^{JnR}-\theta l)(1-l)}}{\ln\rho^{JnR}} \tag{53}$$

The new indifference condition is

$$\frac{c}{\mu}(n+1)=\frac{c}{\mu}\sum_{k=0}^{n-1}\frac{\pi_k^{JnR}}{1-\pi_R^{JnR}}(k+1)+\frac{\alpha}{1-\pi_R^{JnR}}\Leftrightarrow(1-\pi_R^{JnR})\frac{c}{\mu}(n+1)=\frac{c}{\mu}\sum_{k=0}^{n-1}\pi_k^{JnR}(k+1)+\alpha$$

$$\Leftrightarrow(1-\pi_R^{JnR})\frac{c}{\mu}n=\frac{c}{\mu}\sum_{k=0}^{n-1}\pi_k^{JnR}k+\alpha\Leftrightarrow\frac{c}{\mu}n=\frac{c}{\mu}\sum_{k=0}^{n}\pi_k^{JnR}k+\frac{c}{\mu}\pi_R^{JnR}n+\alpha$$

$$\Leftrightarrow\sum_{k=0}^{n}\pi_k^{JnR}(n-k)=r\Leftrightarrow(1-\rho^{JnR})r=\pi_0^{JnR}n-\pi_0^{JnR}\rho(1+\rho^{JnR}+\rho^{JnR2}+\ldots+(\rho^{JnR})^n)$$

$$\Leftrightarrow(1-\rho^{JnR})r=\pi_0^{JnR}n-\pi_0^{JnR}\rho\frac{1-(\rho^{JnR})^n}{1-\rho^{JnR}}\Leftrightarrow(1-\rho^{JnR})r=(1-l)n-\frac{(1-\theta)\rho^{JnR}l}{\rho^{JnR}-\theta l} \tag{54}$$

$$\text{since (53) }\Rightarrow\frac{1-(\rho^{JnR})^n}{1-\rho^{JnR}}=\frac{(1-\theta)l}{(\rho^{JnR}-\theta l)(1-l)}\Leftrightarrow n=\frac{(1-\rho^{JnR})(\rho^{JnR}-\theta l)r+(1-\theta)\rho^{JnR}l}{(1-l)(\rho^{JnR}-\theta l)} \tag{55}$$

Note that (53) and (55) reduce to (27) and (29) when $\theta=0$. Define

$$f_3(\rho)\triangleq\frac{\ln\frac{(1-\theta l)(\rho-l)}{(\rho-\theta l)(1-l)}}{\ln\rho}\text{ from (53); and }f_4(\rho)\triangleq\frac{(1-\rho)(\rho-\theta l)r+(1-\theta)\rho l}{(1-l)(\rho-\theta l)}\text{ from (55).}$$

It can be shown that $f_3(\rho)$ and $f_4(\rho)$ always intercept at $\rho=1$ with a function value of $\frac{l}{1-l}\cdot\frac{1-\theta}{1-\theta l}$ regardless of r, but intercept at only one point, say n, other than $\rho\neq1$. As $r\uparrow$ (i.e., $\alpha\uparrow$), $n\uparrow$ and $\rho\downarrow$. Similar as before, we can choose $\alpha_1,\alpha_2,\ldots,\alpha_N$ such that for $n=1,2,\ldots,N$ and for $\alpha\in[\alpha_n-\frac{c}{\mu}(1-\pi_R^{JnR}),\alpha_n]$, the pair $(n,\rho)=(n,\rho^{JnR})$ satisfies both conditions (53) and (55). □

Per-capita Welfare for Regular Arrivals: Using an argument similar to Lemma 13, it can be shown that total consumer welfare for the regular consumers decreases in α when $\alpha\leq\alpha_L$, from $(1-\theta)(\lambda v-cl)$ to $(1-\theta)\lambda(v-cW^{JNR}-(\frac{1}{1-\pi_N^{JNR}}-1)\alpha_L)$. Dividing the quantities by $(1-\theta)\lambda$ gives per-capita welfare, going from $v-\frac{c}{\mu}$ to $v-cW^{JNR}-(\frac{1}{1-\pi_N^{JNR}}-1)\alpha_L=\alpha_L$.

Per-capita welfare for regular consumers under retrial cost $\alpha \in [\alpha_L, \alpha_H]$ is given by

$$\begin{aligned}
&\frac{1}{(1-\theta)\lambda}\Big[\frac{(1-\theta)\lambda(1-\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}})}{1-(1-\gamma)\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}}(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{(1-\theta)\lambda(1-\gamma)\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-(1-\gamma)\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}}\alpha\Big] \\
=&\frac{(1-\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}})}{1-(1-\gamma)\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}}(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{(1-\gamma)\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-(1-\gamma)\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}}\alpha \\
=&\frac{1}{1-(1-\gamma)\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}}(1-\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}})(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) - \frac{(1-\gamma)\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-(1-\gamma)\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}}\alpha \\
=&\frac{1}{1-(1-\gamma)\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}}\alpha - \frac{(1-\gamma)\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-(1-\gamma)\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}}\alpha \\
&\text{since } \alpha = (1-\sum_{n=N}^{\infty}\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}})(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) \\
=&\alpha
\end{aligned}$$

Therefore, the welfare per strategic consumer increases linearly in α from α_L to α_H.

When $\alpha \geq \alpha_H$, consumer welfare will stay constant as the retrial cost varies (because no consumer retries and pays for the retrial cost). Total consumer welfare for the regular consumers is $(1-\theta)\lambda(1-\frac{\pi_N^{JNB}}{1-\theta l})(v-cW^{JNB})$, and therefore per-capita welfare equals $(1-\frac{\pi_N^{JNB}}{1-\theta l})(v-cW^{JNB}) = \alpha_H$.

Finally, it is easy to see that per-capita welfare on $\alpha \geq \alpha_H$ is less than the welfare when α approaches 0, because the total welfare reaches the maximum possible at the value of $(1-\theta)\lambda(v-\frac{c}{\mu})$ when α approaches 0. If the objective function were changed to maximize per-capita welfare with respect to the retrial cost, the maximizer remains at $\alpha = 0$ regardless of the value of θ. □

Proof of Lemma 5:

We only prove part (ii) here. Proofs to part (i) and (iii) are given in the proofs to Theorem 4 and Proposition 3. For part (ii), fix $\gamma \in [0,1]$. Since

$$\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}} = 1-[\theta+(1-\theta)\frac{1-\pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}}-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}}]l; \text{ and } \rho^{JN_{B(\gamma)}^{R(1-\gamma)}} = (\theta+\frac{1-\theta}{1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}})l,$$

and $\frac{1-\pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}}-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}} < 1$, $\frac{1}{1-\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}} > 1$, we have by taking derivatives that $\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}$ and $\rho^{JN_{B(\gamma)}^{R(1-\gamma)}}$ both decrease in θ. As a result, $\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}, \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}, \ldots, \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}$ all decrease in θ. Recall that

$$\begin{aligned}
\alpha(\gamma) =&(1-\sum_{i=N}^{\infty}\pi_i^{JN_{B(\gamma)}^{R(1-\gamma)}})(v-cW^{JN_{B(\gamma)}^{R(1-\gamma)}}) \\
=&(\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}+\pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}+\ldots+\pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}})v-(\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}\frac{c}{\mu}+\pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}\frac{2c}{\mu}+\ldots+\pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}\frac{Nc}{\mu}),
\end{aligned}$$

we can conclude $\alpha(\gamma)$ decreases in θ because $v > \frac{Nc}{\mu} > \frac{(N-1)c}{\mu} > \ldots > \frac{c}{\mu}$ and $\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}, \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}, \ldots, \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}$ all decrease in θ. □

Proof of Theorem 4:

We denote $\mathcal{U}_{\alpha,\theta}^{\sigma}$ the per-capita welfare for regular arrivals when the environment is θ, the retrial cost is α and the strategy being adopted by the regular consumers is σ. The idea is to show that the down-up-flat per-capita welfare curve for regular arrivals (given by the Pareto-dominant equilibrium strategies) with environment θ_1 dominates that with environment θ_2. Note that when the retrial cost $\alpha \to 0$, the Pareto-dominant strategy (either with environment θ_1 or θ_2) is for every regular consumer to join an idle server and retry whenever the server is busy (i.e., the retrial strategy $J1R$). Therefore,

$$\lim_{\alpha\to 0} \mathcal{U}_{\alpha,\theta_1} = \lim_{\alpha\to 0} \mathcal{U}_{\alpha,\theta_2} = v - \frac{c}{\mu}$$

To show $\mathcal{U}_{\alpha,\theta_1} \geq \mathcal{U}_{\alpha,\theta_2}$ for all α, we will use the fact that the welfare curve between α_L and α_H is a 45-degree straight line. It suffices to prove that

(i) the slope of each piecewise welfare curve on $\alpha \leq \alpha_L$ becomes steeper as θ increases, i.e., $\frac{1}{1-\pi_R^{JnR}}$ increases in θ for $n = 1, 2, \ldots, N$;

(ii) The values of $\alpha_1, \alpha_2, \ldots, \alpha_{N-1}, \alpha_L$ decrease in θ.

(iii) Value on the left end of each of the piecewise welfare curves on $\alpha \leq \alpha_L$ decreases in θ, i.e., $\mathcal{U}_{\alpha_{n-1},\theta}^{JnR}$ decreases in θ for $n = 2, \ldots, N$;

(iv) Value on the right end of each of the piecewise welfare curves on $\alpha \leq \alpha_L$ decreases in θ, i.e., $\mathcal{U}_{\alpha_n,\theta}^{JnR}$ decreases in θ for $n = 1, 2, \ldots, N-1$ and $\mathcal{U}_{\alpha_L,\theta}^{JNR}$ decreases in θ;

(v) Finally, $\mathcal{U}_{\alpha_H,\theta}^{JNB}$ decreases in θ.

For (i), fix $n \in \{1, 2, \ldots, N\}$. The steady-state probabilities of the underlying system under JnR satisfy

$$\begin{aligned}
\pi_1^{JnR} &= \rho^{JnR}\pi_0^{JnR};\\
\pi_2^{JnR} &= \rho^{JnR}\pi_1^{JnR};\\
\pi_3^{JnR} &= \rho^{JnR}\pi_2^{JnR};\\
&\vdots = \vdots\\
\pi_n^{JnR} &= \rho^{JnR}\pi_{n-1}^{JnR};\\
\pi_{n+1}^{JnR} &= \theta l\pi_n^{JnR};\\
\pi_{n+2}^{JnR} &= \theta l\pi_{n+1}^{JnR};\\
&\vdots = \vdots
\end{aligned}$$

Since $\pi_0^{JnR} \equiv 1 - l$ (none of the regular or critical consumer balks), ρ^{JnR} must decrease when θ increases. Therefore, $\pi_0^{JnR}, \pi_1^{JnR}, \ldots, \pi_{n-1}^{JnR}$ all decrease in θ, which implies that $\pi_R^{JnR} = 1 - (\pi_0^{JnR}, \pi_1^{JnR}, \ldots, \pi_{n-1}^{JnR})$ increases. And thus, $\frac{1}{1-\pi_R^{JnR}}$ increases in θ.

For (ii), recall from (40) that

$$\alpha_L = (1 - \pi_R^{JNR})(v - cW^{JNR})$$

$$
\begin{aligned}
&=(1-\pi_R^{JNR})v-(1-\pi_R^{JNR})(cW^{JNR})\\
&=(1-\pi_R^{JNR})v-(\pi_0^{JNR}\frac{c}{\mu}+\pi_1^{JNR}\frac{2c}{\mu}+\ldots+\pi_{N-1}^{JNR}\frac{Nc}{\mu})
\end{aligned}
\tag{56}
$$

From (i), we know that $\pi_0^{JNR},\pi_1^{JNR},\ldots,\pi_{N-1}^{JNR}$ all decrease in θ while π_R^{JNR} increases in θ. And the total reduction by $\pi_0^{JNR},\pi_1^{JNR},\ldots,\pi_{N-1}^{JNR}$ must be equal to the reduction of $1-\pi_R^{JNR}$, as they sum up to 1. Since $v>\frac{Nc}{\mu}>\frac{(N-1)c}{\mu}>\ldots>\frac{2c}{\mu}>\frac{c}{\mu}$, it is clear from (56) that α_L decreases in θ.

Now fix $n\in\{1,2,\ldots,N-1\}$. The retrial payoff when the population adopts JnR with the retrial cost α_n equals the joining payoff at state n, i.e.,

$$
v-cW^{JnR}-\frac{\alpha_n}{1-\pi_R^{JnR}}=v-\frac{(n+1)c}{\mu}
\tag{57}
$$

$$
\begin{aligned}
\Leftrightarrow \alpha_n =&(1-\pi_R^{JnR})(\frac{(n+1)c}{\mu}-cW^{JnR})\\
=&(1-\pi_R^{JnR})\frac{(n+1)c}{\mu}-(1-\pi_R^{JnR})(cW^{JnR})\\
=&(1-\pi_R^{JnR})\frac{(n+1)c}{\mu}-(\pi_0^{JnR}\frac{c}{\mu}+\pi_1^{JnR}\frac{2c}{\mu}+\ldots+\pi_{n-1}^{JnR}\frac{nc}{\mu})
\end{aligned}
\tag{58}
$$

From (i), we know that $\pi_0^{JnR},\pi_1^{JnR},\ldots,\pi_{n-1}^{JnR}$ all decrease in θ while π_R^{JnR} increases in θ. And the total reduction by $\pi_0^{JnR},\pi_1^{JnR},\ldots,\pi_{n-1}^{JnR}$ must be equal to the reduction of $1-\pi_R^{JnR}$. Since $\frac{(n+1)c}{\mu}>\frac{(n)c}{\mu}>\ldots>\frac{2c}{\mu}>\frac{c}{\mu}$, it is clear from (58) that α_n decreases in θ.

For (iii), fix $n\in\{2,\ldots,N\}$. The retrial payoff when the population adopts JnR with the retrial cost α_{n-1} equals the joining payoff at state $n-1$, i.e., the indifference condition is binding at the smaller side:

$$
v-cW^{JnR}-\frac{\alpha_{n-1}}{1-\pi_R^{JnR}}=v-\frac{nc}{\mu}
\tag{59}
$$

Therefore,

$$
\begin{aligned}
\mathcal{U}^{JnR}_{\alpha_{n-1},\theta} =&v-cW^{JnR}-(\frac{1}{1-\pi_R^{JnR}}-1)\alpha_{n-1}\\
=&v-[(1-\pi_R^{JnR})cW^{JnR}+\pi_R^{JnR}cW^{JnR}]-\frac{\pi_R^{JnR}}{1-\pi_R^{JnR}}\alpha_{n-1}\\
=&v-[\pi_0^{JnR}\frac{c}{\mu}+\pi_1^{JnR}\frac{2c}{\mu}+\ldots+\pi_{n-1}^{JnR}\frac{nc}{\mu}+\pi_R^{JnR}cW^{JnR}]-\frac{\pi_R^{JnR}}{1-\pi_R^{JnR}}\alpha_{n-1}\\
=&\pi_0^{JnR}(v-\frac{c}{\mu})+\pi_1^{JnR}(v-\frac{2c}{\mu})+\ldots+\pi_{n-1}^{JnR}(v-\frac{nc}{\mu})+\pi_R^{JnR}(v-cW^{JnR}-\frac{\alpha_{n-1}}{1-\pi_R^{JnR}})\\
=&\pi_0^{JnR}(v-\frac{c}{\mu})+\pi_1^{JnR}(v-\frac{2c}{\mu})+\ldots+\pi_{n-1}^{JnR}(v-\frac{nc}{\mu})+\pi_R^{JnR}(v-\frac{nc}{\mu}) \text{ due to (59)}
\end{aligned}
\tag{60}
$$

As θ increases, $\pi_0^{JnR},\pi_1^{JnR},\ldots,\pi_{n-1}^{JnR}$ all decrease and π_R^{JnR} increases. It is clear from (60) that $\mathcal{U}^{JnR}_{\alpha_{n-1},\theta}$ decreases in θ.

For (iv), fix $n\in\{1,2,\ldots,N-1\}$. Then,

$$
\begin{aligned}
\mathcal{U}^{JnR}_{\alpha_n,\theta} =&v-cW^{JnR}-(\frac{1}{1-\pi_R^{JnR}}-1)\alpha_n\\
=&v-[(1-\pi_R^{JnR})cW^{JnR}+\pi_R^{JnR}cW^{JnR}]-\frac{\pi_R^{JnR}}{1-\pi_R^{JnR}}\alpha_n\\
=&v-[\pi_0^{JnR}\frac{c}{\mu}+\pi_1^{JnR}\frac{2c}{\mu}+\ldots+\pi_{n-1}^{JnR}\frac{nc}{\mu}+\pi_R^{JnR}cW^{JnR}]-\frac{\pi_R^{JnR}}{1-\pi_R^{JnR}}\alpha_n
\end{aligned}
$$

$$
\begin{aligned}
=&\pi_0^{JnR}(v-\frac{c}{\mu})+\pi_1^{JnR}(v-\frac{2c}{\mu})+\ldots+\pi_{n-1}^{JnR}(v-\frac{nc}{\mu})+\pi_R^{JnR}(v-cW^{JnR}-\frac{\alpha_n}{1-\pi_R^{JnR}})\\
=&\pi_0^{JnR}(v-\frac{c}{\mu})+\pi_1^{JnR}(v-\frac{2c}{\mu})+\ldots+\pi_{n-1}^{JnR}(v-\frac{nc}{\mu})+\pi_R^{JnR}(v-\frac{(n+1)c}{\mu}) \text{ due to (57)} \qquad (61)
\end{aligned}
$$

As θ increases, $\pi_0^{JnR},\pi_1^{JnR},\ldots,\pi_{n-1}^{JnR}$ all decrease and π_R^{JnR} increases. And $\mathcal{U}_{\alpha_n,\theta}^{JnR}$ decreases in θ due to (61).

When $n=N$. The retrial payoff when the population adopts JNR with retrial cost α_L equals 0 because we recall from (40) that $v-cW^{JNR}-\frac{\alpha_L}{1-\pi_R^{JNR}}=0$. Then,

$$
\begin{aligned}
\mathcal{U}_{\alpha_L,\theta}^{JNR}=&v-cW^{JNR}-(\frac{1}{1-\pi_R^{JNR}}-1)\alpha_L\\
=&v-[(1-\pi_R^{JNR})cW^{JNR}+\pi_R^{JNR}cW^{JNR}]-\frac{\pi_R^{JNR}}{1-\pi_R^{JNR}}\alpha_L\\
=&v-[\pi_0^{JNR}\frac{c}{\mu}+\pi_1^{JNR}\frac{2c}{\mu}+\ldots+\pi_{N-1}^{JNR}\frac{Nc}{\mu}+\pi_R^{JNR}cW^{JNR}]-\frac{\pi_R^{JNR}}{1-\pi_R^{JNR}}\alpha_L\\
=&\pi_0^{JNR}(v-\frac{c}{\mu})+\pi_1^{JNR}(v-\frac{2c}{\mu})+\ldots+\pi_{N-1}^{JNR}(v-\frac{Nc}{\mu})+\pi_R^{JNR}(v-cW^{JNR}-\frac{\alpha_L}{1-\pi_R^{JNR}})\\
=&\pi_0^{JNR}(v-\frac{c}{\mu})+\pi_1^{JNR}(v-\frac{2c}{\mu})+\ldots+\pi_{N-1}^{JNR}(v-\frac{Nc}{\mu}) \qquad (62)
\end{aligned}
$$

As θ increases, $\pi_0^{JNR},\pi_1^{JNR},\ldots,\pi_{N-1}^{JNR}$ all decrease. It is clear from (62) that $\mathcal{U}_{\alpha_L,\theta}^{JNR}$ decreases in θ.

For (v). The steady-state probabilities of the underlying system under JNB satisfy

$$
\begin{aligned}
\pi_1^{JNB}&=l\pi_0^{JNB};\\
\pi_2^{JNB}&=l\pi_1^{JNB};\\
\pi_3^{JNB}&=l\pi_2^{JNB};\\
\vdots&=\vdots\\
\pi_N^{JNB}&=l\pi_{N-1}^{JNB};\\
\pi_{N+1}^{JNB}&=\theta l\pi_N^{JNB};\\
\pi_{N+2}^{JNB}&=\theta l\pi_{N+1}^{JNB};\\
\vdots&=\vdots
\end{aligned}
$$

When θ increases, π_0^{JNB} must decrease. Therefore, $\pi_0^{JNB},\pi_1^{JNB},\ldots,\pi_{N-1}^{JNB}$ all decrease in θ. Since

$$\mathcal{U}_{\alpha_H,\theta}^{JNB}=\pi_0^{JNB}(v-\frac{c}{\mu})+\pi_1^{JNB}(v-\frac{2c}{\mu})+\ldots+\pi_{N-1}^{JNB}(v-\frac{Nc}{\mu}),$$

$\mathcal{U}_{\alpha_H,\theta}^{JNB}$ clearly decreases in θ. □

Proof of Theorem 5:

Per-capita welfare for critical arrivals under an equilibrium strategy σ (played by the regular consumers) is given by

$$\pi_0^{\sigma}(v-\frac{c}{\mu})+\pi_1^{\sigma}(v-\frac{2c}{\mu})+\ldots+\pi_{n-1}^{\sigma}(v-\frac{nc}{\mu})+\pi_n^{\sigma}(v-\frac{(n+1)c}{\mu})\ldots \qquad (63)$$

It forms a step function over $\alpha\leq\alpha_L$ because equilibrium strategy remains the same for regular consumers for all α that falls in one of the N intervals: $(0,\alpha_1],(\alpha_1,\alpha_2],\ldots,(\alpha_{N-2},\alpha_{N-1}],(\alpha_{N-1},\alpha_L]$. When the retrial cost increases from some value in one interval to the next interval, i.e., when the equilibrium strategy jumps from JnR to $J(n+1)R$, we have $\rho^{JnR}>\rho^{J(n+1)R}$. As a result,

$$\pi_0^{JnR}=\pi_0^{J(n+1)R}=1-l$$

$$\pi_x^{JnR} > \pi_x^{J(n+1)R} \text{ for } x = 1, 2, \ldots, n$$
$$\pi_x^{JnR} < \pi_x^{J(n+1)R} \text{ for } x = n+1, n+2, \ldots$$

As the weights of these steady-state probabilities shift to the right, it is easy to see from (63) that per-capita welfare for the critical arrivals decreases on $\alpha \le \alpha_L$.

When $\alpha \in [\alpha_L, \alpha_H]$, the equilibrium strategy is $JN_{B(\gamma)}^{R(1-\gamma)}$ for some γ. As α increases within this region, γ also increases. As a result, $\pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}}$ decreases and $\pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}}$ increases in α. However, $\frac{\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}}{1-\theta l} = \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}} + \pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}}$ must decrease in α. Because otherwise both $\pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}$ and γ would increase in α, then $\pi_B^{JN_{B(\gamma)}^{R(1-\gamma)}} = \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}} \cdot \gamma$ increases and generates a contradiction.

According to (63), per-capita welfare for critical arrivals is given by

$$\begin{aligned}
&\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{c}{\mu}) + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{2c}{\mu}) + \ldots + \pi_{n-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{(n)c}{\mu}) + \pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{(n+1)c}{\mu}) \ldots \\
=&[\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{c}{\mu}) + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{2c}{\mu}) + \ldots + \pi_{n-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{(n)c}{\mu})] + \pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{(n+1)c}{\mu}) \ldots \\
=&[(1 - \pi_R^{JN_{B(\gamma)}^{R(1-\gamma)}})(v - cW^{JN_{B(\gamma)}^{R(1-\gamma)}})] + \pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{(n+1)c}{\mu}) + \pi_{n+1}^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{(n+2)c}{\mu}) \ldots \\
=&\alpha + \pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{(n+1)c}{\mu}) + \pi_{n+1}^{JN_{B(\gamma)}^{R(1-\gamma)}}(v - \frac{(n+2)c}{\mu}) \ldots
\end{aligned} \tag{64}$$

Since $\pi_n^{JN_{B(\gamma)}^{R(1-\gamma)}}, \pi_{n+1}^{JN_{B(\gamma)}^{R(1-\gamma)}}, \pi_{n+2}^{JN_{B(\gamma)}^{R(1-\gamma)}}, \ldots$ all decreases in α, and $0 > v - \frac{(n+1)c}{\mu} > v - \frac{(n+2)c}{\mu} > \ldots$, per-capita welfare for critical arrivals given by (64) increases in α.

Finally, per-capita welfare for critical arrivals remains the same when $\alpha \ge \alpha_H$ because regular consumers adopt the balk strategy over this region.

Next, let us denote $\mathcal{V}_{\alpha,\theta}^{\sigma}$ the per-capita welfare for critical arrivals when the environment is θ, the retrial cost is α and the strategy being adopted by the regular consumers is σ. We have already proved in Lemma 5 and Theorem 4 that the values of $\alpha_1, \alpha_2, \ldots, \alpha_{N-1}, \alpha_L$, $\alpha(\gamma)$ for each $\gamma \in [0,1]$ (including α_L and α_H) all decrease in θ, meanwhile the welfare forms a step function on $\alpha \le \alpha_L$ and is increasing between α_L and α_H *for any fixed* θ. To prove $\mathcal{V}_{\alpha,\theta_1} \ge \mathcal{V}_{\alpha,\theta_2}$ for all α, it then suffices to show

(i) Value on the right end of each of the piecewise welfare curves on $\alpha \le \alpha_L$ decreases in θ, i.e., $\mathcal{V}_{\alpha_n,\theta}^{JnR}$ decreases in θ for $n = 1, 2, \ldots, N-1$ and $\mathcal{V}_{\alpha_L,\theta}^{JNR}$ decreases in θ;

(ii) Fix $\gamma \in [0,1]$. $\mathcal{V}_{\alpha(\gamma),\theta}^{JN_{B(\gamma)}^{R(1-\gamma)}}$ decreases in θ.

For (i), fix $n \in \{1, 2, \ldots, N-1\}$. The steady-state probabilities of the underlying system under JnR satisfy

$$\begin{aligned}
\pi_1^{JnR} &= \rho^{JnR}\pi_0^{JnR}; \\
\pi_2^{JnR} &= \rho^{JnR}\pi_1^{JnR}; \\
\pi_3^{JnR} &= \rho^{JnR}\pi_2^{JnR}; \\
&\vdots = \vdots \\
\pi_n^{JnR} &= \rho^{JnR}\pi_{n-1}^{JnR}; \\
\pi_{n+1}^{JnR} &= \theta l \pi_n^{JnR}; \\
\pi_{n+2}^{JnR} &= \theta l \pi_{n+1}^{JnR};
\end{aligned}$$

$$\vdots = \vdots$$

Since $\pi_0^{JnR} \equiv 1-l$ (none of the regular or critical arrivals balks), ρ^{JnR} must decrease when θ increases. Therefore, the weights of the steady-state probabilities shift to the left, and as a result, per-capita welfare for critical arrivals,

$$\mathcal{V}_{\alpha_n,\theta}^{JnR} = \pi_0^{JnR}(v-\frac{c}{\mu}) + \pi_1^{JnR}(v-\frac{2c}{\mu}) + \ldots + \pi_{n-1}^{JnR}(v-\frac{nc}{\mu}) + \pi_n^{JnR}(v-\frac{(n+1)c}{\mu})\ldots,$$

decreases in θ. Now at α_L,

$$\mathcal{V}_{\alpha_L,\theta}^{JNR} = \pi_0^{JNR}(v-\frac{c}{\mu}) + \pi_1^{JNR}(v-\frac{2c}{\mu}) + \ldots + \pi_{N-1}^{JNR}(v-\frac{Nc}{\mu}) + \pi_N^{JNR}(v-\frac{(N+1)c}{\mu})\ldots,$$

decreases in θ for the same reason.

For (ii), fix $\gamma \in [0,1]$,

$$\mathcal{V}_{\alpha(\gamma),\theta}^{JN_{B(\gamma)}^{R(1-\gamma)}} = \pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}(v-\frac{c}{\mu}) + \pi_1^{JN_{B(\gamma)}^{R(1-\gamma)}}(v-\frac{2c}{\mu}) + \ldots + \pi_{N-1}^{JN_{B(\gamma)}^{R(1-\gamma)}}(v-\frac{ic}{\mu}) + \pi_N^{JN_{B(\gamma)}^{R(1-\gamma)}}(v-\frac{(N+1)c}{\mu})\ldots.$$

Since both $\pi_0^{JN_{B(\gamma)}^{R(1-\gamma)}}$ and $\rho^{JN_{B(\gamma)}^{R(1-\gamma)}}$ decrease in θ, the steady-state probabilities shift to the left, and as a result, $\mathcal{V}_{\alpha(\gamma),\theta}^{JN_{B(\gamma)}^{R(1-\gamma)}}$ decreases in θ. □

PROPOSITION 3. *As $\theta \to 1$, $\mathcal{V}_{\alpha_n,\theta}^{JnR} \downarrow (v-\frac{c}{\mu-\lambda})$ for all $n=1,2,\ldots,N-1$. Moreover, $\mathcal{V}_{\alpha_L,\theta}^{JNR}$ and $\mathcal{V}_{\alpha_H,\theta}^{JNB} \downarrow (v-\frac{c}{\mu-\lambda})$. Since the welfare curve is a step function on $\alpha \le \alpha_L$ with jumps at $\alpha_1,\alpha_2,\ldots,\alpha_{N-1}$ for every θ, and $\alpha_H \to \alpha_L$ as $\theta \to 1$, we conclude that per-capita welfare becomes $(v-\frac{c}{\mu-\lambda})$ as $\theta \to 1$ for any retrial cost $\alpha \in (0,\infty)$.*

Proof of Proposition 3: Let $\pi_0^M, \pi_1^M, \pi_2^M, \ldots,$ denote the steady-state probabilities of a regular $M/M/1$ system (no balking and no retrials). Recall from (63) that the per-capita consumer welfare for the critical arrivals when the regular customers follow the strategy σ equals

$$\pi_0^\sigma(v-\frac{c}{\mu}) + \pi_1^\sigma(v-\frac{2c}{\mu}) + \ldots + \pi_{n-1}^\sigma(v-\frac{nc}{\mu}) + \pi_n^\sigma(v-\frac{(n+1)c}{\mu})\ldots \tag{65}$$

As $\theta \to 1$, e.g., imagine there is only 1 regular consumer among the new arrivals in each period, what strategy this regular consumer or the regular consumers adopt would have no impact on the steady-state probabilities of the underlying queueing system. That is, $\pi_x^\sigma \to \pi_x^M$ as $\theta \to 1$ for all $x \in \mathbb{N}_0$.

Therefore, for $n=1,2,\ldots,N-1$, we have

$$\mathcal{V}_{\alpha_n,\theta}^{JnR} \to \pi_0^M(v-\frac{c}{\mu}) + \pi_1^M(v-\frac{2c}{\mu}) + \ldots + \pi_{n-1}^M(v-\frac{nc}{\mu}) + \pi_n^M(v-\frac{(n+1)c}{\mu})\ldots = v-\frac{c}{\mu-\lambda}.$$

Similarly, we have

$$\mathcal{V}_{\alpha_L,\theta}^{JNR} = \mathcal{V}_{\alpha_H,\theta}^{JNB} \to \pi_0^M(v-\frac{c}{\mu}) + \pi_1^M(v-\frac{2c}{\mu}) + \ldots + \pi_{N-1}^M(v-\frac{Nc}{\mu}) + \pi_N^M(v-\frac{(N+1)c}{\mu})\ldots = v-\frac{c}{\mu-\lambda}.$$

Also as $\theta \to 1$, we have

$$\alpha_L = (1-\frac{\pi_N^{JNR}}{1-\theta l})(v-cW^{JNR}) \to (1-\frac{\pi_N^M}{1-l})(v-cW^M) = \sum_{x=0}^{N-1}\pi_x^M[v-\frac{(x+1)c}{\mu}];$$

$$\alpha_H = (1 - \frac{\pi_N^{JNB}}{1-\theta l})(v - cW^{JNB}) \to (1 - \frac{\pi_N^M}{1-l})(v - cW^M) = \sum_{x=0}^{N-1} \pi_x^M [v - \frac{(x+1)c}{\mu}]$$

where W^M is the limiting conditional waiting time defined by

$$W^M \triangleq \frac{\pi_0^M}{\sum_{x=0}^{N-1} \pi_x^M} \frac{c}{\mu} + \frac{\pi_1^M}{\sum_{x=0}^{N-1} \pi_x^M} \frac{2c}{\mu} + \frac{\pi_2^M}{\sum_{x=0}^{N-1} \pi_x^M} \frac{3c}{\mu} + \ldots + \frac{\pi_{N-1}^M}{\sum_{x=0}^{N-1} \pi_x^M} \frac{Nc}{\mu}.$$

It becomes clear that $\alpha_H \to \alpha_L$ as $\theta \to 1$, although both α_H and α_L decrease in θ.

On the other hand, it should be noted that not any two points in the set $\{\alpha_1, \alpha_2, \ldots \alpha_{N-1}\}$ will converge to one point as $\theta \to 1$. Since $v - cW^{JnR} - \frac{\alpha_i}{1-\pi_R^{JnR}} = v - \frac{(n+1)c}{\mu}$ for any $\theta \in [0,1)$ (e.g., see (57)), as $\theta \to 1$, we have for any $n \in \{1, 2, \ldots N-1\}$,

$$\begin{aligned}\alpha_n \to & (\sum_{x=0}^{n-1} \pi_x^M) \cdot [(\frac{(n+1)c}{\mu}) - c(\frac{\pi_0^M}{\sum_{x=0}^{n-1} \pi_x^M} \frac{c}{\mu} + \frac{\pi_1^M}{\sum_{x=0}^{n-1} \pi_x^M} \frac{2c}{\mu} + \frac{\pi_2^M}{\sum_{x=0}^{n-1} \pi_x^M} \frac{3c}{\mu} + \ldots + \frac{\pi_{n-1}^M}{\sum_{x=0}^{n-1} \pi_x^M} \frac{nc}{\mu})] \\ = & \sum_{x=0}^{n-1} \pi_x^M [\frac{(n+1)c}{\mu} - \frac{(x+1)c}{\mu}] = \sum_{x=0}^{n-1} \pi_x^M [\frac{(n-x)c}{\mu}].\end{aligned}$$

Therefore, as $\theta \to 1$, we still have $\alpha_1 < \alpha_2 < \ldots < \alpha_{N-1}$. □

Appendix B: Alternative modeling on the time between retrials

In the model presented in the paper, a retrial consumer always returns in the next period. We show here that all the results of the paper will still hold if the time to return after a consumer has made a retry decision is either X number of periods where X is a positive finite integer random variable, or T amount of time where T is exponentially distributed with some rate t, like in orbit models. (Note that when X assumes infinity, it represents a balk decision instead, so we do not consider it.)

The key is to show under both cases, the total arrival rate is still $\lambda^\sigma_{total} = \frac{\lambda}{1-\pi^\sigma_R}$, and the long-run idle probability of the server is still given by $\pi^\sigma_0 = 1 - \frac{\pi^\sigma_J}{1-\pi^\sigma_R} l$, when the population adopts some strategy σ at equilibrium. Then, all the steady-state probabilities of the underlying queue will be calculated the same way as before.

First, we assume that each retrial consumer returns in X number of periods, where X takes on values in $\{x_1, x_2, \ldots, x_k\}$ with probabilities $p_1, p_2, \ldots, p_k$, respectively. Then the total arrival rate is given by

$$\lambda^\sigma_{total} = \lambda + \lambda\pi^\sigma_R \sum_{i=1}^{k} p_k + \lambda(\pi^\sigma_R)^2 (\sum_{i=1}^{k} p_k)^2 + \lambda(\pi^\sigma_R)^3 (\sum_{i=1}^{k} p_k)^3 + \ldots = \frac{\lambda}{1-\pi^\sigma_R},$$

and long-run idle probability of the server is $\pi^\sigma_0 = 1 - \frac{\pi^\sigma_J \cdot \lambda^\sigma_{total}}{\mu} = 1 - \frac{\pi^\sigma_J}{1-\pi^\sigma_R} l$.

In the case with exponential returning time, let L_t denote the average number of consumers in the orbit, i.e., the average number of consumers waiting to retry. (Note that L_t depends on t.) At equilibrium, equating the rates customers enter and leave the system, we have

$$\begin{aligned} \lambda\pi^\sigma_R + tL_t\pi^\sigma_R &= tL_t \\ \lambda\pi^\sigma_R &= tL_t(1-\pi^\sigma_R) \\ tL_t &= \lambda\frac{\pi^\sigma_R}{1-\pi^\sigma_R} \end{aligned}$$

Therefore the total arrival rate equals $\lambda + tL_t = \lambda(1 + \frac{\pi^\sigma_R}{1-\pi^\sigma_R}) = \frac{\lambda}{1-\pi^\sigma_R}$.

Now, equating the rates customers enter and leave the orbit, we have

$$\begin{aligned} \lambda\pi^\sigma_J + tL_t\pi^\sigma_J &= \mu(1-\pi^\sigma_0) \\ \lambda\frac{\pi^\sigma_J}{1-\pi^\sigma_R} &= \mu(1-\pi^\sigma_0) \\ \pi^\sigma_0 &= 1 - \frac{\pi^\sigma_J}{1-\pi^\sigma_R} l \end{aligned}$$

Our claim is thus proved. □

References

Abate, Joseph, Ward Whitt. 1988. The correlation functions of RBM and M/M/1. *Stochastic Models* **4**(2) 315–359.

Afèche, Philipp, Haim Mendelson. 2004. Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science* **50**(7) 869–882.

Aguir, Salah, Zeynep Akşin, Fikri Karaesmen, Yves Dallery. 2008. On the interaction between retrials and sizing of call centers. *European Journal of Operational Research* **191**(2) 398–408.

Aguir, Salah, Fikri Karaesmen, Zeynep Akşin, Fabrice Chauvet. 2004. The impact of retrials on call center performance. *OR Spectrum* **26**(3) 353–376.

Aissani, Amar. 1994. A retrial queue with redundancy and unreliable server. *Queueing Systems* **17**(3-4) 431–449.

Akşin, Zeynep, Mor Armony, Vijay Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* **16**(6) 665–688.

Akşin, Zeynep, Baris Ata, Seyed Emadi, Che-Lin Su. 2013. Structural estimation of callers' delay sensitivity in call centers. *Management Science* **59**(12) 2727–2746.

Albin, SL. 1982. On poisson approximations for superposition arrival processes in queues. *Management Science* **28**(2) 126–137.

Armony, Mor, Constantinos Maglaras. 2004a. Contact centers with a call-back option and real-time delay information. *Operations Research* **52**(4) 527–545.

Armony, Mor, Constantinos Maglaras. 2004b. On customer contact centers with a call-back option: customer decisions, routing rules, and system design. *Operations Research* **52**(2) 271–292.

Artalejo, Jesús. 1995. A queueing system with returning customers and waiting line. *Operations Research Letters* **17**(4) 191–199.

Artalejo, Jesús. 1997. Analysis of an M/G/1 queue with constant repeated attempts and server vacations. *Computers & Operations Research* **24**(6) 493–504.

Artalejo, Jesús. 1999. Accessible bibliography on retrial queues. *Mathematical and Computer Modelling* **30**(3) 1–6.

Artalejo, Jesús. 2010. Accessible bibliography on retrial queues: progress in 2000–2009. *Mathematical and Computer Modelling* **51**(9) 1071–1081.

Artalejo, Jesús, MJ Lopez-Herrero. 2000. On the single server retrial queue with balking. *INFOR. Information Systems and Operational Research* **38**(1) 33–50.

Cachon, Gérard, Pnina Feldman. 2011. Pricing services subject to congestion: Charge per-use fees or sell subscriptions? *Manufacturing & Service Operations Management* **13**(2) 244–260.

de Véricourt, Francis, Yong-Pin Zhou. 2005. Managing response time in a call-routing problem with service failure. *Operations Research* **53**(6) 968–981.

Elcan, Amie. 1994. Optimal customer return rate for an M/M/1 queueing system with retrials. *Probability in the Engineering and Informational Sciences* **8**(4) 521–539.

Falin, Gennadij. 1990. A survey of retrial queues. *Queueing Systems* **7**(2) 127–167.

Falin, Gennadij, James Templeton. 1997. *Retrial Queues*, vol. 75. CRC Press.

Feller, William. 1971. *An Introduction to Probability Theory and Its Applications*, vol. 2. Wiley.

Gans, Noah, Ger Koole, Avishai Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141.

Hassin, Refael, Moshe Haviv. 1996. On optimal and equilibrium retrial rates in a queueing system. *Probability in the Engineering and Informational Sciences* **10**(02) 223–227.

Hassin, Refael, Moshe Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behaviour in Queueing Systems*, vol. 59. Kluwer Academic Pub.

Hassin, Refael, Ricky Roet-Green. 2011. Equilibrium in a two dimensional queueing game: When inspecting the queue is costly. *Working paper, Tel Aviv University, Israel* .

Hoffman, Karla, Carl Harris. 1986. Estimation of a caller retrial rate for a telephone information system. *European Journal of Operational Research* **27**(2) 207–214.

Kostami, Vasiliki, Amy Ward. 2009. Managing service systems with an offline waiting option and customer abandonment. *Manufacturing & Service Operations Management* **11**(4) 644–656.

Kulkarni, Vidyadhar. 1983a. A game theoretic model for two types of customers competing for service. *Operations Research Letters* **2**(3) 119–122.

Kulkarni, Vidyadhar. 1983b. On queueing systems with retrials. *Journal of Applied Probability* **20** 380–389.

Kulkarni, Vidyadhar, Bong Dae Choi. 1990. Retrial queues with server subject to breakdowns and repairs. *Queueing Systems* **7**(2) 191–208.

Lariviere, Martin A, Jan A Van Mieghem. 2004. Strategically seeking service: How competition can generate poisson arrivals. *Manufacturing & Service Operations Management* **6**(1) 23–40.

Larsen, Christian. 1998. Investigating sensitivity and the impact of information on pricing decisions in an M/M/1 queueing model. *International Journal of Production Economics* **56** 365–377.

Mandelbaum, Avishai, William Massey, Martin Reiman, Alexander Stolyar, Brian Rider. 2002. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems* **21**(2-4) 149–171.

Mandelbaum, Avishai, Uri Yechiali. 1983. Optimal entering rules for a customer with wait option at an M/G/1 queue. *Management Science* **29**(2) 174–187.

Maskin, Eric, Jean Tirole. 2001. Markov perfect equilibrium: I. observable actions. *Journal of Economic Theory* **100**(2) 191–219.

Mendelson, Haim, Seungjin Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research* **38**(5) 870–883.

Miller, Bruce, AG Buckman. 1987. Cost allocation and opportunity costs. *Management Science* **33**(5) 626–639.

Morse, Philip M. 1955. Stochastic properties of waiting lines. *Operations Research* **3**(3) 255–261.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37**(1) 15–24.

Odoni, Amedeo R, Emily Roth. 1983. An empirical investigation of the transient behavior of stationary queueing systems. *Operations Research* **31**(3) 432–455.

Parlaktürk, Ali, Sunil Kumar. 2004. Self-interested routing in queueing networks. *Management Science* **50**(7) 949–966.

Plambeck, Erica, Qiong Wang. 2013. Implications of hyperbolic discounting for optimal pricing and scheduling of unpleasant services that generate future benefits. *Management Science* **59**(8) 1927–1946.

Reed, Josh, Uri Yechiali. 2013. Queues in tandem with customer deadlines and retrials. *Queueing Systems* **73**(1) 1–34.

Shin, Yang Woo, Taek Sik Choo. 2009. M/M/s queue with impatient customers and retrials. *Applied Mathematical Modelling* **33**(6) 2596–2606.

Stidham, Shaler. 1985. Optimal control of admission to a queueing system. *Automatic Control, IEEE Transactions on* **30**(8) 705–713.

Whitt, Ward. 2002. Stochastic models for the design and management of customer contact centers: Some research directions. *Department of IEOR, Columbia University* .

Yang, Tao, James Templeton. 1987. A survey on retrial queues. *Queueing Systems* **2**(3) 201–233.

Yechiali, Uri. 1971. On optimal balking rules and toll charges in the GI/M/1 queuing process. *Operations Research* **19**(2) 349–370.

Yechiali, Uri. 1972. Customers' optimal joining rules for the GI/M/s queue. *Management Science* **18**(7) 434–443.