Bounded Rationality in Service Systems

Tingliang Huang
Department of Management Science & Innovation, University College London, United Kingdom, t.huang@ucl.ac.uk
Gad Allon
Kellogg School of Management, Northwestern University, Evanston, IL, g-allon@kellogg.northwestern.edu
Achal Bassamboo
Kellogg School of Management, Northwestern University, Evanston, IL, a-bassamboo@kellogg.northwestern.edu

The traditional operations management and queueing literature typically assume that customers are fully rational. In contrast, in this paper, we study canonical service models with boundedly rational customers. We capture bounded rationality using a model in which customers are incapable of accurately estimating their expected waiting time. We investigate the impact of bounded rationality from both a profit-maximizing firm and a social planner's perspective. For visible queues with the optimal price, a little bit of bounded rationality results in revenue and welfare loss; with a fixed price, a little bit of bounded rationality can lead to strict social welfare improvement. For invisible queues, bounded rationality benefits the firm when customers are sufficiently boundedly rational. Ignoring bounded rationality, when present yet small, can result in significant revenue and welfare loss.

Key words: behavioral operations; service operations; bounded rationality; queueing; consumer behavior

1. Introduction and Literature Review

When a customer calls a call center or goes to a fast food restaurant, a café or an ATM, and has to queue for service, does he always *accurately* and *perfectly* calculate the benefits and costs of joining before making his decisions? The traditional economics and queueing literature have assumed that he does, while anecdotal evidence and experimental studies point to the contrary. In this paper, we study queueing or service systems without making this "perfect rationality" assumption on the part of customers. Our research questions are: how should one model customer bounded rationality in service systems? What are the implications of bounded rationality, e.g., on firm revenue, social welfare, and pricing?

Naor (1969) appears to be the first to incorporate customer decisions into a queueing model. Naor (1969) and subsequent researchers following his work assume customers to be fully rational and able to *perfectly* estimate their expected waiting time, and thus expected utility of joining. Naor (1969) shows that self-interested customers would join a more congested system than what the social planner prescribes, and proposes "levying tolls" (i.e., pricing) as a way to maximize social welfare. In Naor's model, customers are assumed to be able to compute with great precision

the expected waiting time, and thus expected utility they are about to obtain from making a decision about whether to join or renege. One may ask, are customers fully rational? Specifically, does a customer necessarily have the capability to perfectly estimate his expected waiting time and utility? Ariely (2009) claims that irrationality is the real invisible hand that drives human decision making. Indeed, there is abundant empirical evidence that people are boundedly rational. In this work, we study the effects and implications of bounded rationality in canonical queueing or service systems.

Our study is related to several branches of the literature: economics of queues, bounded rationality in economics, and behavioral operations.

Economics of queues. Naor (1969) studies the economics of queueing systems when customers are fully rational. Yechiali (1971, 1972) extends Naor's model to allow for $GI/M/1$ queues. Knudsen (1972) extends Naor's model to allow for a multi-server queueing system in which arriving customers' net benefits are heterogeneous. Lippman and Stidham (1977) extend the Naor model to the finite-horizon and discounted cases showing that, in these settings, the economic notion of an external effect has a precise quantitative interpretation. Hassin (1986) considers a revenue maximizing server who has the opportunity to suppress information on actual queue length, leaving customers to decide on joining the queue on the basis of the known distribution of waiting times. For other extensions, see Van Mieghem (2000), Hassin and Haviv (2003), Afèche (2004) and Hsu et al. (2009) for a comprehensive literature review. Although various models along this line are studied, one common theme in this literature is that full rationality is always assumed.

Bounded rationality in economics. Traditional economic theory postulates that decision makers are "rational," i.e., they have sufficient abilities to do perfect optimization in their choices. Simon (1955) seems to be the first to propose an alternative way to model decision-making behavior: rather than optimizing perfectly, agents search over the alternatives until they find "satisfactory" solutions. Simon (1957) coins the term "bounded rationality" to describe such human behavior.

Bounded rationality refers to a variety of behavioral phenomena in the literature. For a description of systematic errors made by experimental subjects, see Arkes and Hammond (1985), Hogarth (1980), Kahneman et al. (1981), Nisbett and Ross (1980), and the survey papers by Payne et al. (1992) and by Pitz and Sachs (1984). Tversky and Kahnemann (1974) show that people rely on a limited number of heuristic principles which, in general, are useful but sometimes lead to severe and systematic errors. On the basis of the evidence, Conlisk (1996) offers four convincing reasons for incorporating bounded rationality in economic models. Geigerenzer and Selten (2001) adopt heuristics or rules of thumb to model bounded rationality. Thurstone (1927) and Luce (1959)

appear to be the first to develop the framework for *stochastic* choice rules capturing that better options are chosen more often. This approach has attracted considerable attention and has been adopted in a variety of settings. For example, following this approach, McKelvey and Palfrey (1995) and Chen et al. (1997) develop a new equilibrium concept quantal response equilibrium (QRE) in game theory. Others include Bajari and Hortacsu (2001, 2003) in auctions, Cason and Reynolds (2005) in bargaining, Su (2008) in newsvendor, Basov (2009) in monopolistic screening, Waksberg et al. (2009) in natural environments and so on. We also adopt this approach in the paper, in the context of expected waiting time estimation of service systems.

Behavioral operations. Gino and Pisano (2008) survey the literature on modeling bounded rationality in economics, finance, and marketing and argue that operations management scholars should incorporate departures from the rationality assumption into their models and theories. There is an emerging literature on behavioral operations management: Lim and Ho (2007) and Ho and Zhang (2008) conduct experiments on designing pricing contracts for boundedly rational customers; Davis (2011) investigates pull contracts in controlled experiments; Kremer et al. (2011) analyze how individuals make forecasts based on time series data and find that forecasting behavior systematically deviates from normative predictions. We refer the readers to Bendoly et al. (2006) and Bendoly et al. (2008) for this stream of research. We point out two papers that are closely related to our work: Su (2008) is among the first papers to study bounded rationality in operational settings. He applies the logit choice framework to the classic newsvendor model and characterizes the ordering decisions made by a boundedly rational (i.e., noisy) decision maker. He identifies systematic biases and investigates the impact of these biases on several operational settings. We apply a similar framework in this paper, but we interpret bounded rationality as the incapability of estimating expected waiting time in a service setting. Different from Su (2008) where there are prior empirical and experimental studies of the newsvendor model that allow for statistical tests, our study is theoretical, and aims to obtain testable theoretical predictions that stimulate future empirical and experimental work in service systems. Recently, Kremer and Debo (2011) have presented experimental findings along this line. Plambeck and Wang (2010) study implications of hyperbolic discounting in service systems. Although the research setting (i.e., service systems) is similar, the research focus and approach are quite different. In their paper, customers lack the self control to undergo an unpleasant experience that would be in their long-run self interest, which is modeled by psychologists in terms of a hyperbolic discount rate for utility. Our model of bounded rationality focuses on customer ability to compute the expected waiting time.

Another stream of research related to ours is experimental study in queueing. This literature does not support that individuals are fully rational. Rapoport et al. (2004) study a class of queueing problems with endogenous arrival times formulated as non-cooperative n-person games in normal form. Results from their experimental study cannot be fully explained by rational behavior; see also Bearden et. al (2005) and Seale et al. (2005) along this line.

Traditional models in operations assumed that customers are rational both in maximizing their utility and in their ability to compute the anticipated expected waiting time in high precision, regardless of the complexity of the system. In this paper, our model of bounded rationality focuses on this ability to predict expected waiting times. In particular, we assume that customers may lack the capability to accurately estimate their expected waiting time (including the service time) before making their join or balk decisions, and that this capability may vary across different queue configurations (e.g., visible and invisible). We thus introduce a random error term into customer's expected waiting time estimation, which reflects their inability to accurately assess the utility obtained from each action. From both a revenue-maximizing firm's perspective and a social planner's perspective, we study the impact of bounded rationality for both visible queues (such as a fast food restaurant, a café or an ATM) and invisible queues (such as a call center).

The three main contributions of our study are: (i) ours is among the first to model bounded rationality in service systems, for both visible and invisible queues; (ii) our study provides insights on how service systems should be managed in the presence of boundedly rational customers; and (iii) we provide a framework to stimulate future empirical and experimental work in service systems and behavioral operations.

The remainder of this paper is organized as follows. §2 presents a model of bounded rationality in service systems. We study revenue maximization and social welfare maximization in §3 and §4. We provide discussion in the last section. Proofs of the results are relegated to the Appendix.

2. A Model of Bounded Rationality in Service Systems

Consider a customer who has to decide whether to join a service system or not. If he joins, he will obtain expected utility $U_1 \equiv R - p - C\mathbb{E}w$ where $R > 0$ is the reward on completion of service, p is the price, $\mathbb{E}w$ is the expected waiting and service time, and $C > 0$ is the average waiting and service cost per unit of time. If he balks, he will get utility $U_2 = 0$ (from an outside option). The existing queueing literature typically assume that he is perfectly rational: if $U_1 \geq 0$, he will join the system; otherwise, he will balk. However, accurately computing expected waiting and service time is typically not an easy task for customers. We thus depart from this literature by incorporating

a more realistic assumption: customers lack the capability to accurately estimate their expected waiting time. As a result, customers cannot *guarantee* that the best choice is *always* chosen and they may make mistakes. To formally model this noisy waiting time estimation, we introduce a random error term ε into customer's expected waiting time estimation. If $V_1 \equiv R - p - C(\mathbb{E}w + \varepsilon) \geq 0$, he will join the system, and balk otherwise. As outside observers, we then obtain the customer joining probability $\varphi \equiv \mathbb{P}(\varepsilon \leq \frac{U_1}{C})$, which should be interpreted as the *fraction* of customers who will join. For analytical tractability, we assume that the error term ε follows a logistic distribution $F(x) = \frac{1}{1+e^{-\frac{x}{\theta}}}$ for some $\theta > 0$. The logistic distribution provides a good approximation to the normal distribution, but has heavier tails (cf. Talluri and van Ryzin 2004, p. 305-306). Following McFadden (1974) and Anderson et al. (1992), we thus obtain the customer joining probability:

$$\varphi = \frac{e^{\frac{U_1}{C\theta}}}{1 + e^{\frac{U_1}{C\theta}}}.$$

It is important to note that, in our model, customers do *not* choose to play mixed strategies. It is the noisy estimation that drives this behavior, and φ denotes the fraction of customers that join the system.

To interpret the meaning of θ, note that the standard deviation of ε is $\sigma \equiv \sqrt{Var(\varepsilon)} = \frac{\pi}{\sqrt{3}}\theta \approx 1.8\theta$. Hence, the parameter θ is proportional to the standard deviation of the error term ε. Thus, the parameter θ measures the *error level* of customer expected waiting time estimation.[1] Furthermore, the standard deviation of customer expected utility V_1 is $\sigma_{V_1} \equiv C\sigma \approx 1.8C\theta$. For convenience, we define $\beta \equiv C\theta$ which measures the error level of customer expected utility estimation. This error level β reflects customer bounded rationality in the sense that customers have limited computational capability in perfectly estimating the expected waiting time (and as a result, the expected utility of joining) in the queueing setting. Hence, we interpret the parameter β as the extent customers are incapable of implementing the optimal decision due to their incapability in estimating expected waiting time. As $\beta \to 0$, the joining behavior converges to *full rationality*. At the other extreme, as $\beta \to \infty$, customers join or balk with equal fractions. Therefore, we can refer to the magnitude of β as the *level of bounded rationality*.

The interpretation of the level of bounded rationality β also follows from the well-known interpretation of the coefficients of logit regressions in that it captures the idea that better options are chosen more often. One can rewrite the joining probability: $\log\left(\frac{\varphi}{1-\varphi}\right) = \frac{1}{\beta}U_1$. The left-hand side (LHS) is the "log odds" of joining the system, so β is the inverse of the difference in the log odds

[1] See Hey and Orme (1994, p. 1301) for a similar approach and explanations; alternatively, one could normalize the utility value while using the variance of the error term to capture the level of bounded rationality.

for any one unit increase in the expected utility of joining the system. For example, when $\beta = 0.5$, then the log odds doubles for any one unit increase in the expected utility of joining the system; when $\beta = 2$, then the log odds decreases by half for any one unit increase in the expected utility of joining the system.

Given that β is the standard deviation of customer utility, it has the same unit as the reward R. We expect the magnitude of β to depend on the context. For example, McKelvey and Palfrey (1995) estimated $1/\beta$ in their game settings, which suggests that β ranges from 0.3 to 6.7 in 1982 penny. Bajari and Hortacsu (2005, Table 7) report the estimates for $1/\beta$: 15.68, 17.36, or 11.32 in 1989 dollar. In another auction setting, Goeree and Holt (2002) found out that the parameter β is around 0.09, 0.16, or 0.26 (see Table 4, p. 258) in 2001 dollar. In the queueing setting, Kremer and Debo (2011) recently ran experiments in laboratories to test bounded rationality for visible queues. They found that β is statistically significantly different from zero: from their estimation, $1/\beta$ is 0.296, and the standard deviation is 0.025. They use the expected reward $R = \$10.2$.

Given that β has the same unit as R, one can use β/R to measure the level bounded rationality relative to the reward. Note that this ratio is dimensionless and thus allows a unified comparison across studies without converting monetary units. Straightforward computation yields the magnitude of β/R in the literature: $\beta/R \in [0.06, 5.68]$ in McKelvey and Palfrey (1995); $[0.0038, 0.0059]$ in Bajari and Hortacsu (2005); $[0.018, 0.052]$ in Goeree and Holt (2002); around 0.33 in Kremer and Debo (2011). In our numerical study, β/R is in the range $[0, 0.5]$ or $[0, 5]$ for the low-reward and high-reward case respectively, which is comparable to the magnitude observed in the literature.

We believe that people are typically not capable of accurately computing expected waiting time, and we interpret bounded rationality in terms of incapabilities in accurately estimating it in this paper. However, this is not the only possible interpretation. There are several closely related interpretations in the literature that could be adopted here. First, Chen et al. (1997) propose the boundedly rational Nash equilibrium where agents are not utility maximizers, but, instead choose randomly in a fashion that is influenced by a *subconscious* utility function. In our setting, one can interpret U_1 as the subconscious utility since it is not explicitly known to the customer due to his bounded rationality. Moreover, Chen et al. (1997) point out that "any mathematical structure commonly used with the noise or random utility interpretations can also be interpreted as a model of bounded rationality." Along this line, one can interpret our model in several ways by rewriting the customer utility function: $V_1 = R - p - C\mathbb{E}w + \varepsilon' = U_1 + \varepsilon'$. The noise term does not have to come from expected waiting time. It can stem from customer noisy perception of the waiting cost

C or heterogeneity in customer preference for a particular queueing environment (cf. Maister 1985 and Larson 1987): $\varepsilon' = -\varepsilon \mathbb{E}w$.

The second different interpretation comes from Mattsson and Weibull (2002): Agents are utility maximizers. However, they have to make some effort in order to implement any desired outcome, and the disutility of this effort enters their utility function. In our setting, one may think of customers having to make some effort to implement the optimal joining decision, perhaps due to the challenging task of estimating the expected waiting time. These different interpretations provide different justifications to our model. We follow the interpretation of incapability of accurately estimating the expected waiting time, because we believe that it captures best the main computational limitation customers face in service systems.

Our model has both usefulness and limitations. The model is useful since: (i) it provides a systematic way to capture bounded rationality in service systems using a single parameter β; (ii) this single parameter β is endowed with a concrete meaning that includes both conventional full rationality and purely random behavior (i.e., full bounded rationality) and allows us to investigate the impact of bounded rationality (e.g., in terms of revenue, welfare and pricing); (iii) this model is flexible to accommodate several interpretations; and (iv) it can be readily used for further empirical or experimental tests due to its close connection to commonly used logit regressions. However, this model has limitations: flexibility comes with both advantages and disadvantages. The fact that the model itself cannot be pinned down to a unique interpretation without lab experiments calls for further empirical or experimental studies. Indeed, that is one of the goals of the paper: to stimulate a desirable theoretical-empirical feedback loop in service operations management research. Also, the logit model is a special case of the general class of the 'Fechner' approach of modeling the stochastic element in decision making, and we refer the reader to Loomes et al. (2002) for other models that can potentially be useful.

Both of the expected waiting and service time $\mathbb{E}w$ and the level of bounded rationality depend on the configuration of the service system: whether the queue length is observable to customers or not. Hereafter, we distinguish the visible and invisible queues, and denote the level of bounded rationality β and β_I for them respectively.

2.1. The Visible Queue

Consider a single-server queueing system that is observable. Customers arrive to the system according to a Poisson process with rate λ. Upon arrival, each customer decides whether or not to join the queue after observing the queue length and based on his estimate of the waiting time. Service times are assumed to be independently, identically, and exponentially distributed with mean $\frac{1}{\mu}$. Denote

the utilization of the system $\rho \equiv \lambda/\mu$ if all customers join. Customers are served on a first-come first-served basis. Upon arrival, observing n customers in the system, the fraction of customers that join the system is given by

$$\varphi_n \equiv \frac{e^{\frac{R-p-\frac{(n+1)C}{\mu}}{\beta}}}{1+e^{\frac{R-p-\frac{(n+1)C}{\mu}}{\beta}}}, \tag{1}$$

for $n = 0, 1, 2, ...$ In order to compute the expected waiting time, each customer needs perfect information about the number of customers when he joins, the service rate, and the cognitive capability to transform this information into an estimate of the expected waiting time. Thus, if customers lack either of the information or the *cognitive capability*, then the waiting time estimate will be inaccurate. Further, the accuracy of the expected waiting time is severely impacted by the fact that they need to do this in real-time.

For ease of exposition, we let $\lambda_n \equiv \lambda\varphi_n$, $n = 0, 1, 2, ...$, be the state-dependent queue-joining rates. Then, we can treat the number of customers in the system as a birth-death process with birth rate λ_n and death rate μ. Although customers are boundedly rational, we first show that the *stability* of the system is guaranteed as long as the level of bounded rationality is finite, as stated in the following proposition.

PROPOSITION 1. *The visible queueing system with boundedly rational customers is stable for $\beta < \infty$, and the probability distribution in steady-state is as follows:*

$$P_0 = \frac{1}{1+\Sigma_{k=1}^{\infty}\frac{\lambda_0\lambda_1...\lambda_{k-1}}{\mu^k}}$$

is the probability that the system is in state 0, and

$$P_n = \frac{\lambda_0\lambda_1...\lambda_{n-1}}{\mu^n(1+\Sigma_{k=1}^{\infty}\frac{\lambda_0\lambda_1...\lambda_{k-1}}{\mu^k})}$$

is the probability in state n, $n \geq 1$.

Note that the queue length distribution becomes unbounded from above for $\beta > 0$, but the system is always stable by Proposition 1.

We numerically observe that both utilization, i.e., $\rho(p, \beta) \equiv \sum_{n=0}^{\infty} \lambda_n P_n/\mu$, and expected queue length, i.e., $q(p, \beta) \equiv \sum_{n=0}^{\infty} nP_n$, have an intricate relationship with the level of bounded rationality β. In particular, neither of them is monotonic or unimodal as a function of the level of bounded rationality β. To understand why this happens, we first define $n_s = [\frac{R\mu}{C}]$ as the threshold queue length used by fully rational customers in deciding to join the queue or not. We then divide the states of the system into two regions using the threshold n_s: Region 1 comprises of the states when

the number of customers in the system is less than n_s, Region 2 comprises of all the other states, i.e., those when the number of customers in the system is greater than n_s. When customers are fully rational, customers will join with probability 1 in Region 1 and and 0 in Region 2. For every strictly positive level of bounded rationality, the joining probability will be between 0 and 1. Hence, bounded rationality in Region 1 lowers utilization (and expected queue length) while increases utilization (and expected queue length) in Region 2. As customers become more boundedly rational, these two effects take place simultaneously. It turns out that it is not clear which effect dominates.

2.2. The Invisible Queue

We now turn to an invisible queueing system using the same model setup as §2.1. The only difference is that the queue length is *invisible* to customers. Potential customers arrive to this system according to a Poisson process with rate λ. Since customers cannot observe the state of the system, they have to make a decision a priori whether to arrive to the queue or not. Different from the visible-queue setting, each customer has to form beliefs about the state of the system which is determined by other customers' strategies. We assume that each customer knows all the underlying parameters of the system such as R, C, p, λ and μ. He knows that he is boundedly rational, and all the other customers are also boundedly rational. In other words, he is able to form the correct *subconscious* belief about the state of the system, which is used to estimate his expected waiting time.

In investigating the system, we are initially interested in the fraction of customers that join the system $\varphi(p,\beta_I) \in [0,1]$ in equilibrium. Again, customers do not choose to play mixed strategies. Only from the point of view of an outside observer, customer decisions are probabilistic. A customer's net benefit or utility of joining is $U_1 = R - p - C\mathbb{E}w = R - p - \frac{C}{(\mu-\varphi(p,\beta_I)\lambda)^+}$, where we used the fact that the *thinning* of a Poisson process with arrival rate λ is still a Poisson process with rate $\varphi(p,\beta_I)\lambda$ and $\frac{C}{(\mu-\varphi(p,\beta_I)\lambda)^+}$ is the customer's expected waiting cost. The arrival rate $\varphi(p,\beta_I)\lambda$ of the queue will be referred to as *effective* demand. According to our model of bounded rationality in §2, each customer cannot perfectly estimate his expected waiting time, and hence joins the system with probability $\varphi_I \equiv \frac{e^{U_1/\beta_I}}{1+e^{U_1/\beta_I}}$. In equilibrium, *consistency* requires that the system effective arrival rate is consistent with customer behavior: $\varphi(p,\beta_I) = \varphi_I$. Hence, we define the equilibrium of the invisible queueing system as follows.

Definition 1. **(Equilibrium Joining Fraction)**. We say that $\varphi(p,\beta_I)$ is an equilibrium joining fraction if it satisfies the following

$$\varphi(p,\beta_I) = \frac{e^{\frac{R-p-\frac{C}{(\mu-\varphi(p,\beta_I)\lambda)^+}}{\beta_I}}}{1+e^{\frac{R-p-\frac{C}{(\mu-\varphi(p,\beta_I)\lambda)^+}}{\beta_I}}}, \tag{2}$$

for $\beta_I > 0$, and

$$\varphi(p, 0) = \min\{\varphi_0, 1\},$$

where φ_0 satisfies

$$R - p - \frac{C}{\mu - \varphi_0 \lambda} = 0, \tag{3}$$

for $\beta_I = 0$.

When $\beta_I > 0$, equation (2) yields a fixed-point problem given that the logit expression in the RHS includes the equilibrium joining fraction (i.e., the LHS).

When $\beta_I = 0$, i.e., customers are fully rational, then the definition is precisely Hassin (1986)'s equilibrium condition (equation (4.1) on Page 1189). It is possible that there is no $\varphi_0 \in [0, 1]$ satisfying equation (3) and the actual arrival rate then is λ since, even if all customers decide to join, each customer's expected utility is still strictly positive. According to this definition, we have $\varphi(p, 0) = \min\{\frac{\mu}{\lambda} - \frac{C}{\lambda(R-p)}, 1\}$.

The assumption that customers are boundedly rational in their strategies but not in subconscious beliefs is consistent with the economics literature of modeling bounded rationality (cf. Chen et al. 1997 and references therein). This approach is certainly restrictive, as decision makers are capable of correctly calculating the expected error-prone actions of the other players, which are certainly non-trivial cognitive tasks (Mallard 2011). But McKelvey and Palfrey (1995) and Chen et al. (1997) show that such quantal response equilibria (QRE) emerge from learning. Hence, in the short-run or transient states, customers may neglect others' bounded rationality, but eventually, they would be able to form the correct belief about others' strategies so that the fixed-point outcome according to Definition 1 below would emerge. We relax Definition 1 in Appendix D by allowing customers to have incorrect subconscious beliefs. There the model is closely related to the "level-k thinking" (Stahl and Wilson 1995): the closed-loop fixed-point type of equation (2) would become open-loop. The resulting analysis for both revenue and social welfare maximization is straightforward. We refer the reader to Appendix D for detailed discussion. In Appendix F, we prove that for a given range of the level of bounded rationality, the path of customer joining decisions over time, when the customers adaptively learn the expected waiting time, converges to the equilibrium in Definition 1. Also, focusing on bounded rationality as incapability of accurately predicting the expected waiting time allows a fair comparison of the visible queue versus invisible queue.

Next, we investigate whether an equilibrium always exists. Proposition 2 shows that there always exists a *unique* equilibrium.

PROPOSITION 2. *There always exists a unique equilibrium for the invisible queue, for any finite price p and level of bounded rationality $\beta_I > 0$.*

We are now interested in how the (unique) equilibrium $\varphi(p,\beta_I)$ behaves as a function of the price p and the level of bounded rationality β_I. For convenience, we let $\overline{p} \equiv R - \frac{2C}{2\mu-\lambda}$ denote the price under which each customer receives exactly zero utility so that the equilibrium joining fraction is half *regardless of* the level of bounded rationality. The following proposition characterizes the equilibrium joining fraction.

PROPOSITION 3. *(i) If $p<\overline{p}$, equilibrium joining fraction $\varphi(p,\beta_I)$ is strictly decreasing in β_I.*
(ii) If $p>\overline{p}$, equilibrium joining fraction $\varphi(p,\beta_I)$ is strictly increasing in β_I.
(iii) If $p=\overline{p}$, equilibrium joining fraction $\varphi(p,\beta_I)=\frac{1}{2}$ for any β_I.
(iv) For any fixed β_I, equilibrium joining fraction $\varphi(p,\beta_I)$ is strictly decreasing in p.

We offer the following intuition: when the price is so low that each customer receives strictly positive utility, the initial joining fraction is above half. As the level of bounded rationality increases, better decisions are made less often, and thus the joining fraction decreases as customers are more boundedly rational. Interestingly, if the price is set so that each customer receives exactly zero utility in equilibrium, then increasing the level of bounded rationality has no effect on the joining fraction since customers join or balk with equal fractions regardless of the level of bounded rationality.

It is intuitively clear that $\varphi(p,\beta_I)$ is strictly decreasing in price p by equality (2), i.e., a larger price always results in a lower joining fraction, regardless of the level of bounded rationality, which is the "law of demand" in this service setting.

It is useful to note that the invisible queue with boundedly rational customers is essentially an $M/M/1$ system with arrival rate $\varphi(p,\beta_I)\lambda$ and service rate μ. The server utilization is $\rho_I(p,\beta_I) \equiv \rho\varphi(p,\beta_I)$. Thus the utilization behaves the same as the joining fraction as a function of p and β_I (which is already characterized in Proposition 3). Using similar logic, we can characterize the expected queue length $q_I(p,\beta_I) \equiv \frac{\varphi(p,\beta_I)\lambda}{\mu-\varphi(p,\beta_I)\lambda}$ as a function of p and β_I.

3. Revenue Maximization

In the previous section, our attention has been focused on the system equilibrium or dynamics. In this section, we focus our attention on the revenue generated from such systems. In this sense, we are looking from a revenue-maximizing firm's perspective.

3.1. The Visible Queue

The revenue as a function of price p and level of bounded rationality $\beta>0$ is

$$\Pi(p,\beta) \equiv \sum_{n=0}^{\infty} \lambda_n P_n p = \sum_{n=0}^{\infty} \frac{e^{\frac{R-p-\frac{(n+1)C}{\mu}}{\beta}}}{1+e^{\frac{R-p-\frac{(n+1)C}{\mu}}{\beta}}} \lambda P_n p.$$

Note that we normalize the cost of serving customers to zero without loss of generality.

When customers are fully rational, i.e., $\beta = 0$, we naturally define $\Pi(p, 0) \equiv \lim_{\beta \to 0} \Pi(p, \beta)$, for any price p (one can show such a limit exists). In the setting with fully rational customers, Naor (1969) shows that choosing the revenue-maximizing price boils down to choosing the optimal integer n to maximize the revenue function $\Pi_n = \lambda \frac{1-\rho^n}{1-\rho^{n+1}}(R - \frac{Cn}{\mu})$, so that $p(n) = R - \frac{Cn}{\mu}$. Let n_r be the maximizer, and Π_{n_r} be the maximized revenue.

We are interested in comparing the optimal revenue $\Pi(p^*(\beta), \beta)$ when the revenue-maximizing price $p^*(\beta)$ is set, if customers are slightly boundedly rational, and the optimal revenue $\Pi_{n_r} \equiv \sup_p \lim_{\beta \to 0} \Pi(p, \beta)$ if customers are fully rational. For convenience, let $p^* \equiv p^*(0) = R - \frac{Cn_r}{\mu}$ be the revenue-maximizing price under full rationality.

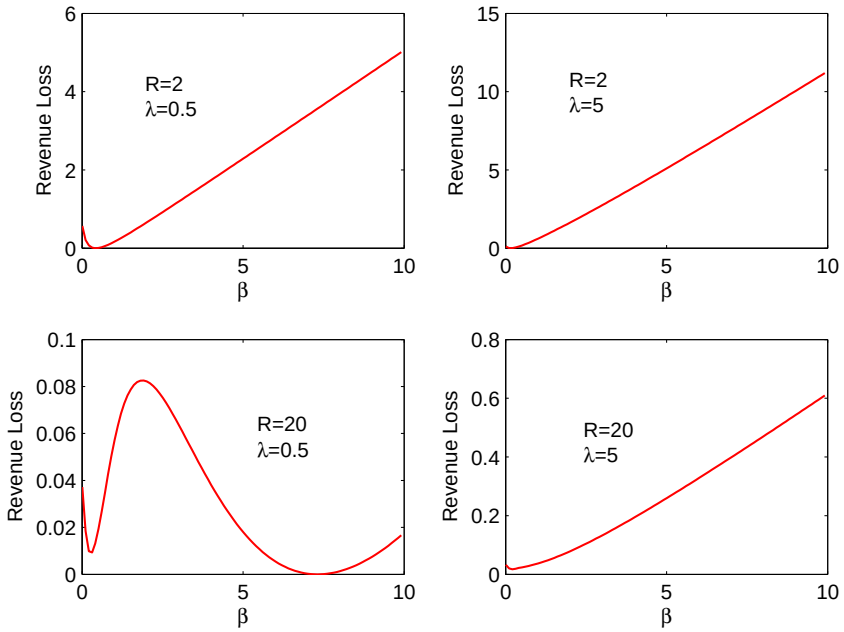
PROPOSITION 4. *$p^*(\beta) < p^*$ and $\Pi(p^*(\beta), \beta) < \Pi_{n_r}$ when β is strictly positive but sufficiently small.*

Proposition 4 does not extend to situations when the level of bounded rationality becomes high. (As customers become fully boundedly rational, the optimal revenue goes to infinity.) The intuition behind Proposition 4 is as follows: a little bit of bounded rationality forces the revenue-maximizing firm to strictly lower its price compared to full rationality, which in turn brings strictly lower revenue for the firm. In other words, a little bit of bounded rationality makes revenue-collecting less profitable. The reason is that a little bit of bounded rationality strictly reduces the effective demand of the system, ceteris paribus. A strictly lower price is necessary to increase the effective demand to maximize revenue.

From our numerical studies, we found that the optimal revenue is not necessarily monotone with respect to the level of bounded rationality, and the revenue maximizing price as a function of the level of bounded rationality can increase or decrease. The explanation is similar to why the utilization is not necessarily monotonic in the level of bounded rationality, since the revenue directly depends on the utilization: $\Pi(p, \beta) = \rho(p, \beta)\mu p$.

Impact of Ignoring Bounded Rationality Without taking into account customer bounded rationality, the revenue-maximizing firm will rationally charge price $p^*(0) = R - \frac{Cn_r}{\mu}$. We are interested in the revenue loss due to bounded rationality. We carried out a numerical study. We approximate the steady-state distribution by truncating the birth-death process to a finite state space: gradually increasing the number of states until the revenue function is no longer sensitive to the truncation level. To investigate the impact of the level of bounded rationality β, the reward-to-cost ratio R/C, and the traffic intensity λ/μ on the revenue loss, we normalize $C = 1$ and $\mu = 1$, and

Figure 1 **Revenue loss when revealing the queue if bounded rationality is ignored ($C=1, \mu=1$)**



vary R and λ. Our systematic numerical study is intended to include high/low utilization crossed with high/low reward. Figure 1 shows a representative example where $R \in \{2, 20\}$ and $\lambda \in \{0.5, 5\}$. From Figure 1, we have the following observations: (i) The revenue loss can be significant (more than 200% for instance) depending on the parameters, and can be arbitrarily large as β goes to infinity. (ii) The revenue loss is not necessarily monotone with respect to β. This is likely due to the fact that the revenue-maximizing price is not necessarily monotone in β. To understand this fact, recall that neither the utilization nor expected queue length are necessarily monotone, and our explanation in §2.1 applies here given that these basic measures drives the revenue.

3.2. The Invisible Queue

The firm's objective is to choose a price p to maximize the expected revenue $\Pi^I(p, \beta_I) \equiv p\varphi(p, \beta_I)\lambda$, where $\varphi(p, \beta_I)\lambda$ is the effective demand rate to the system.

To investigate the firm's revenue maximization problem, we first study the behavior of the revenue $\Pi^I(p, \beta_I)$ as a function of price p and level of bounded rationality β_I. Note that $\Pi^I(p, \beta_I)$ is simply a linear transformation of $\varphi(p, \beta_I)$, hence Proposition 3 characterizes how $\Pi^I(p, \beta_I)$ behaves as a function of β_I for any fixed p.

We next investigate how revenue $\Pi^I(p, \beta_I)$ behaves as a function of price p, for any fixed level of bounded rationality β_I. To state Proposition 5, we denote $\beta_0 \equiv \frac{R}{2} - \frac{2C\mu}{(2\mu-\lambda)^2}$, which is the level of

bounded rationality at which the optimal price $p^*(\beta_0) = \overline{p}$. Hence, at the level of bounded rationality β_0, each customer receives zero expected utility of joining under the revenue-maximizing price.

PROPOSITION 5. *(i) For any fixed level of bounded rationality β_I, $\Pi^I(p, \beta_I)$ is unimodal in p, and thus there exists a unique price $p^*(\beta_I)$ that maximizes $\Pi^I(p, \beta_I)$.*
(ii) The optimal price $p^*(\beta_I)$ is strictly increasing in the level of bounded rationality β_I for $\beta_I \in [\max\{\beta_0, 0\}, \infty)$.

From this proposition, we obtain that the revenue-maximizing price $p^*(\beta_I)$ is monotonically increasing in $\beta_I \in [0, \infty)$ if R is sufficiently small.

When the optimal price induces each customer to receive strictly negative expected utility in equilibrium, a higher level of bounded rationality would induce the firm to increase its price. The reason is that higher bounded rationality leads to higher joining fractions for a fixed price in this case, according to Proposition 3 (ii). Hence, when the optimal price is sufficiently high (so that each customer receives strictly negative expected utility in equilibrium), then increasing the level of bounded rationality leads to even higher optimal prices. However, when the optimal price is low so that each customer receives strictly positive utility, then increasing the level of bounded rationality *can* lead to lower optimal prices, where the firm's tradeoff is about the benefit of higher prices versus the loss of lower effective demand. Proposition 5 above partially characterizes which one of these effects dominates.

We are now ready to state the result on the effect of the level of bounded rationality on the *optimal* revenue $\Pi^I(p^*(\beta_I), \beta_I)$. Using the envelope theorem, we obtain the following immediate corollary to Proposition 3.

COROLLARY 1. *(i) If $p^*(\beta_I) > \overline{p}$, then optimal revenue $\Pi^I(p^*(\beta_I), \beta_I)$ strictly increases in β_I.*
(ii) If $p^*(\beta_I) < \overline{p}$, then optimal revenue $\Pi^I(p^*(\beta_I), \beta_I)$ strictly decreases in β_I.
(iii) $p^*(\beta_I) = \overline{p}$, then optimal revenue $\Pi^I(p^*(\beta_I), \beta_I) = \frac{1}{2}\lambda\overline{p}$ is constant in β_I.

By Proposition 5 and Corollary 1, we know that the optimal revenue $\Pi^I(p^*(\beta_I), \beta_I)$ strictly increases in β_I as β_I is sufficiently large. Therefore, the revenue-maximizing firm can exploit the bounded rationality when customers are sufficiently boundedly rational.

Finally, we are also interested in how the arrival rate affects revenue, as it would later be useful. Recall that in Hassin (1986) where customers are fully rational, there exists some λ_0, when $\lambda > \lambda_0$, the revenue function is independent of λ. Interestingly, in our case with boundedly rational customers, we have that higher arrival rate λ always leads to *strictly* higher revenue.

PROPOSITION 6. *For any fixed price p and level of bounded rationality $\beta_I > 0$, the equilibrium joining fraction is strictly decreasing and the revenue is strictly increasing in the arrival rate λ.*

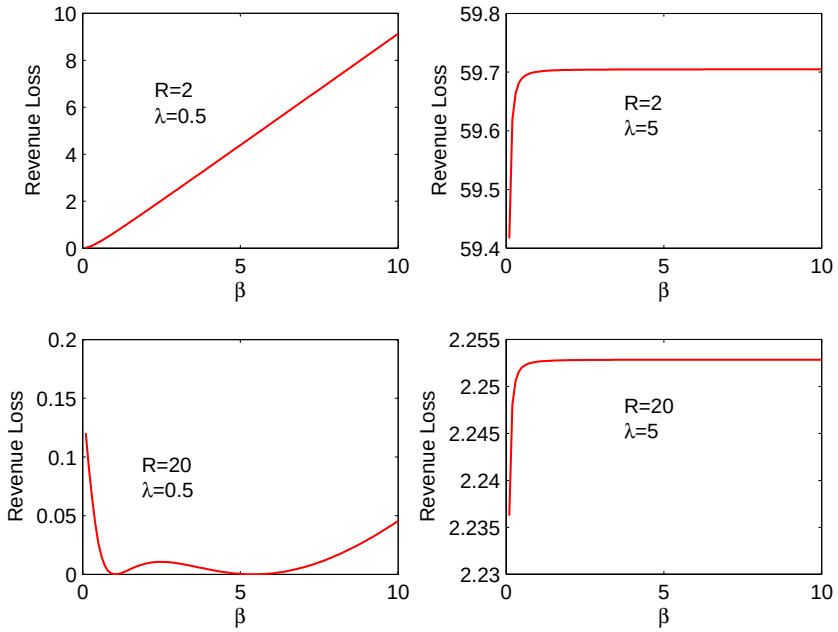
The result that higher arrival rates lead to lower equilibrium joining fractions is not surprising since more congestion forces each customer to lower his joining probability. However, the result that more arrivals always lead to more revenue may appear to be surprising. The key insight is that the marginal revenue increment has to be proportional to the marginal joining fraction decrement given the equilibrium condition (2). Hence, the effective demand $\varphi(p,\beta_I)\lambda$ increases in λ. This proposition implies that for any price p, not necessarily the optimal price, higher arrival rates lead to higher revenue. In particular, higher arrival rates lead to higher *optimal* revenue. Such finding is in stark contrast to Hassin (1986)'s full-rationality case.

Impact of Ignoring Bounded Rationality. Finally, we are interested in the consequence of ignoring bounded rationality while customers are actually boundedly rational. Without taking into account customer bounded rationality, the revenue-maximizing firm will rationally charge price $p^*(0)$ which is generally different from the revenue-maximizing price $p^*(\beta_I)$. Hence, $\Pi^I(p^*(0),\beta_I) \equiv p^*(0)\varphi(p^*(0),\beta_I)\lambda \le \Pi^I(p^*(\beta_I),\beta_I)$. We are interested in the revenue loss $\Delta\Pi^I(\beta_I) \equiv \frac{\Pi^I(p^*(\beta_I),\beta_I)-\Pi^I(p^*(0),\beta_I)}{\Pi^I(p^*(0),\beta_I)}$ as a result of this ignorance of bounded rationality. As an example, we use the same parameters as before. Figure 2 shows that the revenue loss can be nontrivial (e.g., more than 200%). In general, the revenue loss is not necessarily monotone with respect to the level of bounded rationality β. Similar to the visible queue, this observation could be driven by the fact that the revenue-maximizing price is not necessarily monotone in β (Proposition 5). Hence, when β increases, the revenue-maximizing price $p^*(\beta_I)$ and $p^*(0)$ may become closer, which would result in lower revenue loss; the case when $R=20, \lambda=0.5$ and $\beta=5$ in Figure 2 illustrates this point. However, as β is larger than a certain threshold, the loss is significant. In fact, $\lim_{\beta_I\to\infty}\Delta\Pi^I(\beta_I)=\infty$, i.e., the revenue loss can be arbitrarily large as customers are sufficiently boundedly rational. This directly follows from the fact that $\lim_{\beta_I\to\infty}p^*(\beta_I)=\infty$ (See the first-order condition in the proof of Proposition 5) and $\lim_{\beta_I\to\infty}\varphi(p,\beta_I)=0.5$.

4. Social Welfare Maximization

We now turn to study the problem from a social planner's perspective. The social planner is interested in maximizing social welfare. In this section, we study the impact of bounded rationality on the social welfare, both when the price is exogenously given and when the social planner charges the welfare-maximizing price.

Figure 2 **Revenue loss when hiding the queue if bounded rationality is ignored ($C=1, \mu=1$)**



4.1. The Visible Queue

In many settings, the price is set or influenced by other considerations such as market conditions, competition, or price being set by a third party. There are settings where optimizing over the price or even charging a price may not be feasible. We first study how bounded rationality affects social welfare for a given price. Observe that the fixed price p always appears as $R-p$ in equation (1). For the ease of comparing with the results in the classic paper Naor (1969) and for mathematical convenience, we assume $p=0$. However, the findings extend to the setting where the price is non-zero. We can derive the social welfare function as follows:

$$W(\beta) \equiv W(p,\beta)|_{p=0} = \sum_{n=0}^{\infty} \lambda_n P_n R - \sum_{n=0}^{\infty} n P_n C. \tag{4}$$

The first term in equation (4) is the (long run) average reward and the second term is the average waiting cost. Notice that if customers are fully rational, i.e., $\beta=0$, then $P_n=1$ for $n=0,1,...,[\frac{R\mu}{C}]-1$ and $P_n=0$ for $n>[\frac{R\mu}{C}]-1$, in which case our model reduces to Naor (1969)'s model.

To compare social welfare $W(\beta)$ with $W(0)$, we first recall $n_s=[\frac{R\mu}{C}]$ as the threshold queue length used by fully rational customers in deciding to join the queue or not, and n_0, the equivalent threshold from a social planner's point of view. Naor (1969) shows that $n_s \geq n_0$, i.e., self-interested customers typically make the system more congested than the socially optimal level.

Intuitively, bounded rationality can create two effects for the social welfare: a positive (i.e., welfare-improving) effect and negative (i.e., welfare-diminishing) effect. To understand how these two effects come into play, we divide the states of the system into three regions using the two thresholds n_s and n_0: Region 1 comprises of the states when the number of customers in the system is less than n_0, Region 2 comprises of the states when the number of customers in the system is greater than n_0 but less than n_s, and Region 3 comprises of all the other states, i.e., those when the number of customers in the system is greater than n_s. When customers are fully rational, customers will join with probability 1 in Region 1 and Region 2, and 0 in Region 3. Recall that to maximize social welfare, customers should join with probability 1 in Region 1 and 0 in Region 2 and 3. However, for any strictly positive level of bounded rationality, the joining fraction is between 0 and 1. Hence, social welfare will decrease in Region 1 and 3, but increase in Region 2 compared to full rationality. As customers become more boundedly rational, these effects take place simultaneously, and it seems unclear a priori which effect would dominate.

While equation (4) presents a complete characterization of the social welfare in terms of the level of bounded rationality β, the dependence is quite intricate. Thus, we begin by analyzing the social welfare $W(\beta)$ in the neighborhood of zero. We are interested in whether a little bit of bounded rationality increases or decreases the social welfare, i.e., the relationship between $W(\beta)$ and $W(0)$ when β is sufficiently small. It turns out we are able to completely characterize the conditions in which one effect dominates the other. We have the following simple inequalities showing when the social welfare increases or decreases as the customers become slightly boundedly rational.

PROPOSITION 7. *If any one of the following three conditions is satisfied:*
(1) $n_s < \frac{R\mu}{C} - \frac{1}{2}$*;*
(2) $n_s = n_0$*;*
(3) $n_s = \frac{R\mu}{C} - \frac{1}{2}$ *and* $\rho > 1$*,*
then $W(\beta) < W(0)$ *when* $\beta > 0$ *is sufficiently small. Otherwise,* $W(\beta) > W(0)$ *when* $\beta > 0$ *is sufficiently small.*

According to Proposition 7, if either of the following two conditions is satisfied:
(a) $n_s \neq n_0$, and $n_s > \frac{R\mu}{C} - \frac{1}{2}$,
(b) $n_s \neq n_0$, $n_s = \frac{R\mu}{C} - \frac{1}{2}$ and $\rho \leq 1$,
then a little bit of bounded rationality **strictly improves** the social welfare.[2]

[2] Finding a simple sufficient and necessary condition for $n_s \neq n_0$ is difficult. However, Lemma EC.5 in the Appendix shows that, either of the two conditions is sufficient for $n_s \neq n_0$: (1) $\rho > 1$ and $n_s > 1$, (2) $\sqrt{2} - 1 < \rho < 1$ and $n_s > 2$.

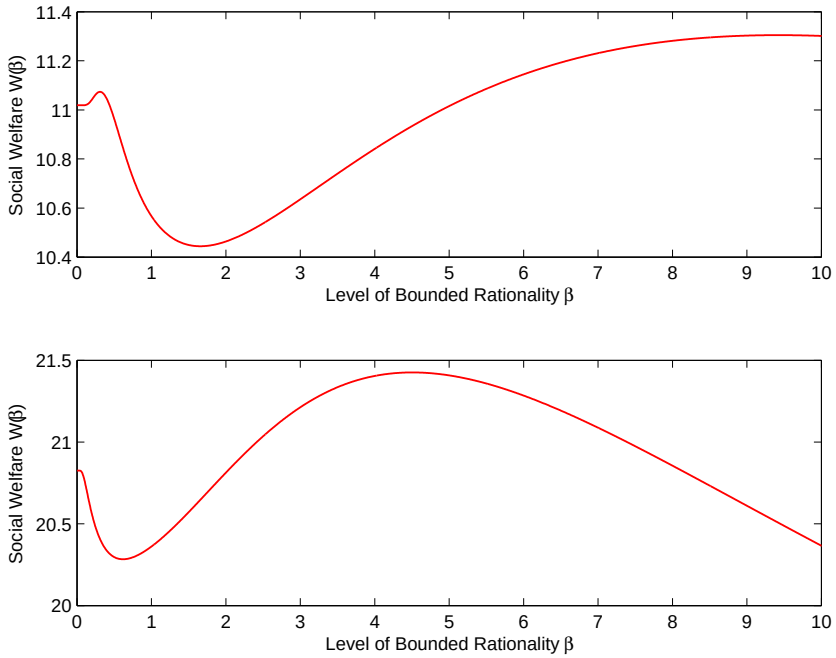
According to equation (1), as $\beta \to 0$, the customers who have strictly positive expected utility of joining will join the queue with probability converging to 1, and those who have strictly negative expected utility will join the queue with probability converging to 0. For the sake of a thought experiment, we assume that there is a single state in which customers join with non-degenerate probabilities. If this were the case, it must be the "marginal state" either "on the positive side," i.e., the state where the customer who observes $n_s - 1$ customers in the system, or "on the negative side," i.e., the state where the customer who observes n_s customers in the system. However, in the true system with boundedly rational customers, as long as the customers are slightly boundedly rational, there are multiple states in which customers join the system with non-degenerate probabilities. Considering that the level of bounded rationality is close to zero, it is intuitively clear that it is the *joint effect* of the customer behavior in the marginal state on the positive side and the customer behavior on the negative side that determines the direction of the social welfare change. The effect of the customer behavior in the marginal state on the positive side improves the social welfare, while the effect of the customer behavior in the marginal state on the negative side is detrimental to the social welfare. We have to characterize which effect dominates the other, i.e., to disentangle the joint effect.

The scenario when $n_s = n_0$, i.e., self-interested customers bring the system to the social optimality, is rare. However, in this setting, a little bit of bounded rationality causes each customer to join the system with non-degenerate probabilities that can only decrease the social welfare. In contrast, when $n_s \neq n_0$, it is the relative location of n_s and $\frac{R\mu}{C} - \frac{1}{2}$ that determines the result. If $n_s > \frac{R\mu}{C} - \frac{1}{2}$, then it is the effect of the customer behavior on the positive side of the marginal state that dominates. Hence, the social welfare is improved. If $n_s < \frac{R\mu}{C} - \frac{1}{2}$, the opposite effect would decrease the social welfare.

The case when $n_s = \frac{R\mu}{C} - \frac{1}{2}$ is more delicate since both effects come into play simultaneously. It turns out that when $\rho > 1$, the congestion level is so high that the negative effect dominates, and we obtain strictly lower social welfare. When $\rho \leq 1$, the congestion level is low enough to allow the positive effect to dominate, and we obtain strictly higher social welfare.

In other words, compared to fully rational customers, boundedly rational customers err on both sides, joining a more congested system and balking when congestion is low. While the former is detrimental to the social welfare, the latter can be beneficial. The social welfare can thus be improved depending on which of these effects dominates. In Proposition 7, we provide a simple characterization of this dichotomy. This result appears to be striking: while bounded rationality is

Figure 3 **Global Behavior of Social Welfare: Example 1 ($R=14.93, C=7, \mu=3, \lambda=5, n_s=6>\frac{R\mu}{C}-\frac{1}{2}=5.8986$) and Example 2 ($R=16, C=7, \mu=3, \lambda=2.6, v_s=6.8571, n_s=6<\frac{R\mu}{C}-\frac{1}{2}=6.3571$)**



usually associated with suboptimal decisions, it might yield better outcomes for the society overall. This is due to the externality present among the boundedly rational customers in the system.

As we discussed before, characterizing the social welfare as a function of the level of bounded rationality is difficult because of the intricate joint effects coming from the three regions simultaneously as customers become more boundedly rational. To understand the social welfare as a function of the level of bounded rationality, we carried out a numerical study. To demonstrate that the social welfare function is not necessarily unimodal in a reasonable range of bounded rationality, we intentionally use different sets of parameters. In the first example, the parameters are $R=14.93, C=7, \mu=3, \lambda=5$, so that $n_s=6>\frac{R\mu}{C}-\frac{1}{2}=5.8986$. As shown in the graph in the upper panel of Figure 3, the social welfare strictly increases initially as predicted by Proposition 7, however, it decreases and then increases again when the customers become more boundedly rational. In the second example, the parameters are $R=16, C=7, \mu=3, \lambda=2.6$, so that $v_s=6.8571, n_s=6<\frac{R\mu}{C}-\frac{1}{2}=6.3571$. As illustrated in the graph in the lower panel of Figure 3, the social welfare initially decreases as predicted by Proposition 7, however, it increases as the level of bounded rationality becomes larger, and decreases again as the level of bounded rationality further increases. Thus, even though the social welfare is well behaved for a little bit of bounded rationality, it does not possess global properties such as convexity/concavity or even unimodality.

This result stands in contrast to the invisible queue where the social welfare function is unimodal in the level of bounded rationality β.

We have analyzed the impact of bounded rationality on social welfare for a given price. However, the social planner may be able to freely charge a price to maximize the social welfare. We now investigate the implication of bounded rationality for the social welfare if the social planner can regulate the system by pricing optimally. We are interested in whether bounded rationality increases or decreases the social welfare. We denote the social welfare function $W(p,\beta)$ when the social planner charges price p and customers' level of bounded rationality is β. Obviously, the social welfare $W(p,\beta)$ can be expressed in a similar fashion as equation (4).

Naor (1969) shows that, by levying tolls, the social planner can achieve the social optimum when customers are fully rational. In particular, if any price $p^* \in (R-\frac{C(n_0+1)}{\mu}, R-\frac{Cn_0}{\mu}]$ is charged by the social planner, then the maximum social welfare $W^*(0) \equiv \sup_p W(p,0)$ can be achieved. We study whether the optimal social welfare $W^*(0)$ can be achieved by adding bounded rationality on the part of customers.

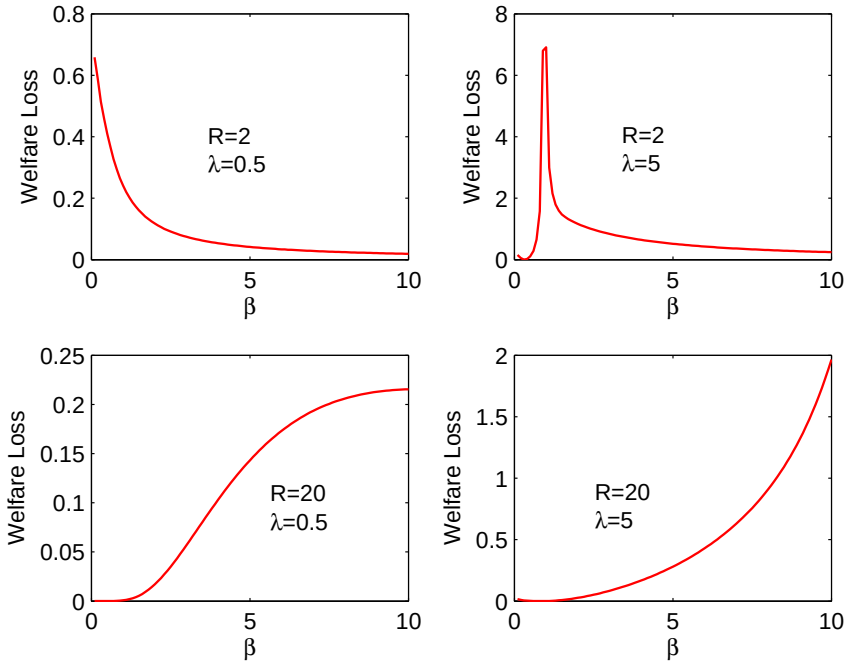
We show that when facing boundedly rational customers, the first-best social welfare can never be achieved, as stated in the following proposition.

PROPOSITION 8. *For any price $p \in \mathbb{R}$ charged to customers, the social welfare $W(p,\beta)$ is strictly lower than the social optimum when β is strictly positive, i.e., $W(p,\beta) < W^*(0)$ for $\beta > 0$.*

This proposition proves that bounded rationality always results in social welfare losses compared to the full-rationality case. This is in contrast to: (a) Naor (1969), where levying tolls achieves the socially optimal welfare; (b) The result in Proposition 7 that a little bit of bounded rationality can increase the social welfare when an arbitrary price is charged (when the firm charges the optimal price, then only case (2) in Proposition 7 arises); and (c) The result in Proposition 10 of the invisible queue where there may not be any welfare loss due to bounded rationality. This stems from the following: when customers are fully rational, the social planner can always regulate the service system by charging prices to achieve the social optimality $W^*(0)$. However, each boundedly rational customer randomizes with non-degenerate probabilities to join or balk. In this case, the social planner loses the *precise* control over the customers' joining decisions, and thus bounded rationality dilutes the effectiveness of the price regulation. Of course, if the firm can implement state-dependent pricing, then the social planner can achieve the first-best, even in the presence of bounded rationality.

From our numerical studies, we found that there can be multiple prices that maximize the welfare for a given level of bounded rationality. Second, the optimal welfare is not necessarily monotone

Figure 4 **Welfare loss when revealing the queue if bounded rationality is ignored ($C=1, \mu=1$)**



with respect to the level of bounded rationality. The explanation for the non-monotonicity behavior is similar to that for the global non-monotonicity behavior of the social welfare function with respect to β: bounded rationality has the positive and negative effect on welfare simultaneously.

Impact of Ignoring Bounded Rationality. Without taking into account customer bounded rationality, the social planner will pick one price in the range $(R-\frac{C(n_0+1)}{\mu}, R-\frac{Cn_0}{\mu}]$. We are interested in the welfare loss due to bounded rationality. Using the same example as in the revenue-maximization case, Figure 4 shows the welfare loss when the firm charges price $R-\frac{Cn_0}{\mu}$. Again, we observe non-trivial welfare loss (more than 60% for instance). The intuition from the non-monotonicity behavior is similar to the explanation we provided for the global non-monotonicity behavior of the social welfare function when $p=0$: as β increases, the positive and negative effect on welfare occurs simultaneously.

4.2. The Invisible Queue

For any price p and level of bounded rationality β_I, the social welfare function is denoted as

$$W^I(\varphi(p,\beta_I)) \equiv \varphi(p,\beta_I)\lambda R - \frac{\varphi(p,\beta_I)\lambda}{\mu-\varphi(p,\beta_I)\lambda}C. \tag{5}$$

For mathematical convenience, we may drop the dependence over $\varphi(p,\beta_I)$ and write $W^I(p,\beta_I)$.

The first term of equation (5) is the average benefit the customers receive from the system, and the second term is the average waiting cost incurred by the customers. Note that price p affects social welfare only indirectly through the equilibrium joining fraction $\varphi(p,\beta_I)$.

First, observe that the social welfare $W^I(\varphi(p,\beta_I))$ is strictly concave in $\varphi(p,\beta_I)$ (Lemma EC.2 in the Appendix). Combining this fact with the characterization of the equilibrium joining fraction $\varphi(p,\beta_I)$, we can characterize how the social welfare behaves as a function of the level of bounded rationality in Proposition 9.

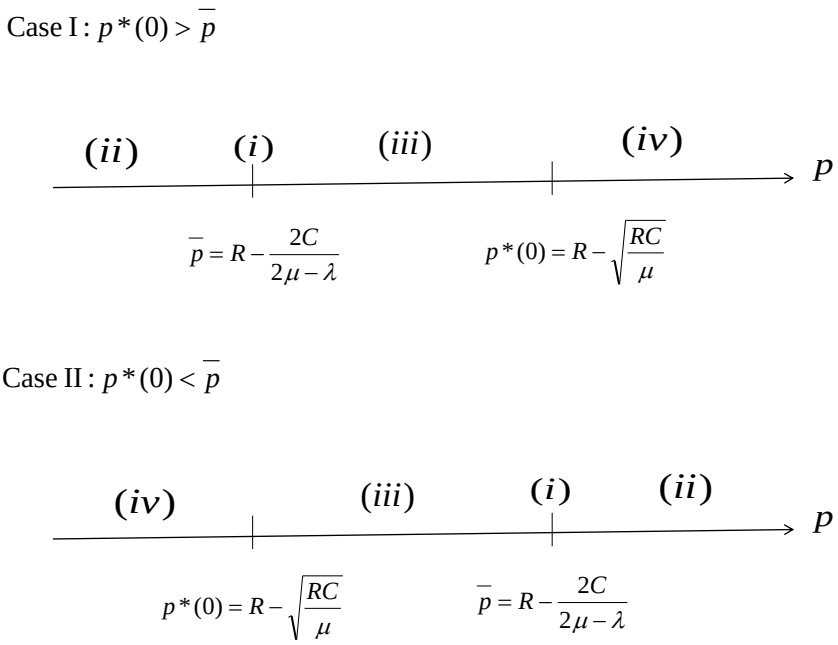
From Proposition 5, one can obtain that $p^*(0)=R(1-\sqrt{\frac{C}{\mu R}})$. Note that when customers are fully rational, the welfare-maximizing price and the revenue-maximizing price coincide.

PROPOSITION 9. *(i) If $p=\overline{p}$, then social welfare $W^I(\varphi(p,\beta_I))$ is constant for $\beta_I\geq 0$.*
(ii) If $[p^*(0)-\overline{p}][p-\overline{p}]\leq 0$ and $p\neq\overline{p}$, then social welfare $W^I(\varphi(p,\beta_I))$ strictly increases for $\beta_I\geq 0$.
(iii) If $p\in(\min\{p^*(0),\overline{p}\},\max\{p^*(0),\overline{p}\})\bigcup\{p^*(0)\}$ and $p^*(0)\neq\overline{p}$, then social welfare $W^I(\varphi(p,\beta_I))$ strictly decreases for $\beta_I\geq 0$.
(iv) If $[p^*(0)-\overline{p}][p-p^*(0)]>0$, then social welfare $W^I(\varphi(p,\beta_I))$ strictly increases in $[0,\beta_w(p)]$ and strictly decreases in $(\beta_w(p),\infty)$, where

$$\beta_w(p)=\frac{R-p-\sqrt{\frac{RC}{\mu}}}{\ln\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda-\mu+\sqrt{\frac{C\mu}{R}}}}.$$

This proposition fully characterizes the social welfare as a function of the level of bounded rationality. Figure 5 depicts the different scenarios in Proposition 9 based on the relative magnitude of $p^*(0)$ and $\overline{p}$. By comparing the magnitude of $p^*(0)$ and $\overline{p}$, we discuss two cases. For each case, we depict the range of price p that falls into the various scenarios numbered by the roman numerals (i), (ii), (iii), and (iv) in Proposition 9. (The case when $p^*(0)=\overline{p}$ is simple and hence not illustrated in the Figure.) For the first scenario, the joining fraction at price $p=\overline{p}$ is precisely half and it is independent of the level of bounded rationality. Thus the social welfare in (i) is not impacted by the level of bounded rationality. For the second scenario, the fraction 0.5 lies *between* the joining fraction induced by the welfare-maximizing price when customers are fully rational and the joining fraction induced by price p when customers' level of bounded rationality is β_I. In this case, increasing the level of bounded rationality makes their "distance" smaller. Thus, the social welfare strictly increases as customers are more boundedly rational. For the third scenario, the joining fraction induced by the welfare-maximizing price when customers are fully rational is either too high or too low compared to the joining probability induced by price p when customers' level of

Figure 5 Illustration of the Various Scenarios in Proposition 9



bounded rationality is β_I, so that increasing the level of bounded rationality can only make their "distance" further apart. Therefore, the social welfare strictly decreases in the level of bounded rationality β_I. For the last scenario, the joining probability induced by the welfare-maximizing price when customers are fully rational can be achieved (in the interior). Hence, as the level of bounded rationality increases from zero, the social welfare is "closer" to the optimal social welfare. In this case, the social welfare function is unimodal in the level of bounded rationality, and the first-order condition yields the level of bounded rationality $\beta_w(p)$.

Proposition 9 implies that the social welfare function is unimodal in the level of bounded rationality, as stated in Corollary 2.

COROLLARY 2. *Social welfare $W^I(\varphi(p, \beta_I))$ is unimodal in the level of bounded rationality β_I for any price p.*

We have demonstrated that the impact of bounded rationality on social welfare depends on the magnitude of the fixed price charged and the welfare function is unimodal in the level of bounded rationality. The next question we are interested in is: what is the welfare-maximizing price and how does the welfare behave under such a price? We first prove that the social welfare $W^I(\varphi(p, \beta_I))$ is unimodal in price p for any level of bounded rationality β_I (see Lemma EC.4 in the Appendix for a rigorous justification). Finding the welfare-maximizing price boils down to finding the optimal

joining probability φ_w^*. To derive the welfare-maximizing price, we use the first-order condition $\frac{\partial W^I(\varphi(p,\beta_I))}{\partial \varphi(p,\beta)} = 0$ and obtain

$$\varphi_w^* = \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\lambda},$$

which is the optimal equilibrium joining fraction that induces the optimal social welfare. Suppose this equilibrium point can be achieved in the interior, then it is required that $R \in (\frac{C}{\mu}, \infty)$ if $\mu < \lambda$; and $R \in (\frac{C}{\mu}, \frac{C\mu}{(\mu-\lambda)^2})$ if $\mu > \lambda$. For cases when the equilibrium point is on the boundary, the problem becomes trivial: If $R \leq \frac{C}{\mu}$, then it is socially optimal to keep everybody out of the system; if $R \geq \frac{C\mu}{(\mu-\lambda)^2}$ when $\mu > \lambda$, then it is socially optimal to let everyone join the system.

We are now ready to state the welfare-maximizing price that maximizes the social welfare. To state the result, we first substitute the joining fraction φ_w^* into the equilibrium condition, i.e., equation (1), and then we obtain the "unconstrained" optimal price by

$$p_w^*(\beta_I) = R - \sqrt{\frac{CR}{\mu}} - \beta_I \ln \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\lambda - \mu + \sqrt{\frac{C\mu}{R}}} = p^*(0) - \beta_I \ln \frac{\varphi_w^*}{1-\varphi_w^*},$$

which can be negative. The welfare-maximizing price is thus $\max\{0, p_w^*(\beta_I)\}$. Proposition 10 characterizes this welfare-maximizing price and the corresponding social welfare.

PROPOSITION 10. *(i) If $R > \frac{4C\mu}{(2\mu-\lambda)^2}$, when $\beta < \beta_w(0)$ where $\beta_w(0) = \frac{R - \sqrt{\frac{RC}{\mu}}}{\ln \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\lambda - \mu + \sqrt{\frac{C\mu}{R}}}}$, the price $p = p_w^*(\beta_I)$ is the unique price that maximizes the social welfare, $p_w^*(\beta_I)$ strictly decreases in β_I, and the optimal social welfare is $W^I(p_w^*, \beta_I) = \mu R + C - 2\sqrt{\mu RC}$; when $\beta_I \geq \beta_w(0)$, the price $p = 0$ is the unique price that yields the maximum social welfare $W^I(0, \beta_I)$.*
(ii) If $R \leq \frac{4C\mu}{(2\mu-\lambda)^2}$, the price $p = p_w^*(\beta_I)$ is the unique price that maximizes the social welfare, $p_w^*(\beta_I)$ strictly increases in β_I, and the optimal social welfare is $W^I(p_w^*, \beta_I) = \mu R + C - 2\sqrt{\mu RC}$.

We discuss the implications of this proposition as follows:

If $\varphi_w^* = \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\lambda} > \frac{1}{2}$, then $\ln \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\lambda - \mu + \sqrt{\frac{C\mu}{R}}} > 0$, which implies that the price $p_w^*(\beta_I)$ is strictly decreasing in level of bounded rationality β_I. In particular, when customers are slightly boundedly rational, the optimal price strictly decreases. The intuition is that the equilibrium joining fraction decreases as the level of bounded rationality increases. To achieve the desired optimal joining fraction φ_w^*, the social planner has to lower the price as the level of bounded rationality increases.

Similarly, if $\varphi_w^* = \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\lambda} < \frac{1}{2}$, then $\ln \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\lambda - \mu + \sqrt{\frac{C\mu}{R}}} < 0$, which implies that the price $p_w^*(\beta_I)$ is strictly increasing in level of bounded rationality β_I.

The key insight from this proposition is that the first-best social welfare (which is independent of the level of bounded rationality and the arrival rate) can be achieved when either (i) the optimal

joining fraction for social welfare maximization is strictly above half and the level of bounded rationality is not too high or (ii) the optimal joining fraction for social welfare maximization is below half.

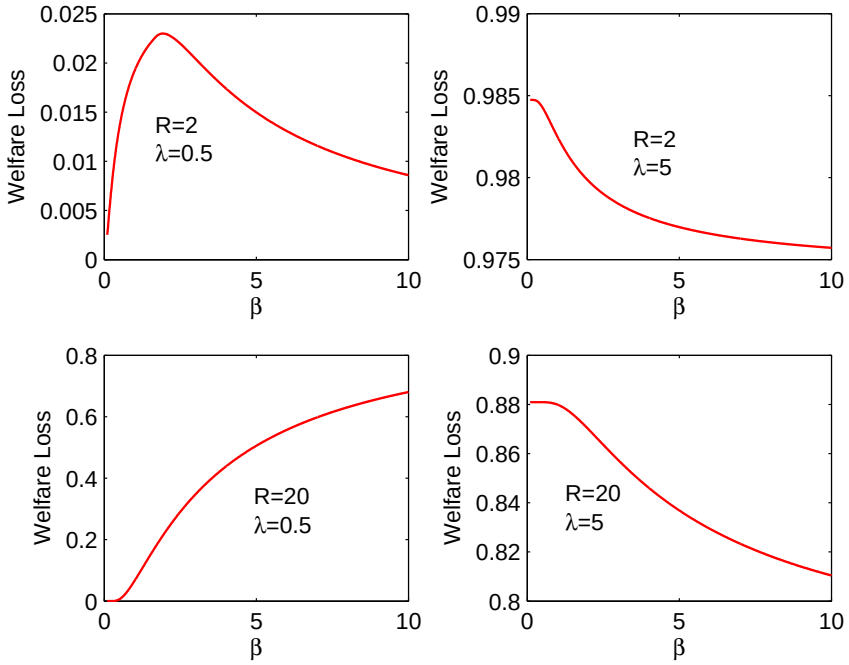
The intuition for this key insight is the following: In these settings, the social planner can always correct for the boundedly rationality on the part of customers, i.e., the social planner still achieves the same optimal social welfare, by charging *appropriate* prices. However, when the optimal joining fraction for social welfare maximization is strictly above half and the level of bounded rationality is sufficiently high, the first-best social welfare cannot be achieved. In other words, when the desired joining fraction is high, to achieve this, the customers have to join with this fraction in equilibrium. However, customers joining fraction would be much lower if they are too boundedly rational, and too low even if the firm does not charge any price. In this case, higher bounded rationality leads to more social welfare losses. This result stands in contrast to (i) the case of revenue maximization, and (ii) the result in Proposition 8 of the visible queue.

Impact of Ignoring Bounded Rationality. Without taking into account customer bounded rationality, the social planner will rationally charge price $p^*(0)$ which is generally different from the welfare-maximizing price $\max\{0, p_w^*(\beta_I)\}$. Similarly, we are interested in the welfare loss $\Delta W^I(\beta_I) \equiv \frac{W^I(\max\{0,p_w^*(\beta_I)\},\beta_I)-W^I(p^*(0),\beta_I)}{W^I(p^*(0),\beta_I)}$. For the same example as the revenue-maximization case, Figure 6 shows that the welfare loss can be significant (more than 80% for instance) and is not necessarily monotone with respect to the level of bounded rationality. The non-monotonicity behavior and its intuitive explanation stems from Proposition 9, where for a fixed price, the social welfare may be non-monotone in β.

5. Discussion

The quantal choice paradigm in the behavioral economics literature posits that people are more likely to select better choices than worse ones but do not necessarily succeed in selecting the very best choice. In this paper, we adopted this framework to model bounded rationality in service systems in the sense that customers lack the capability to perfectly estimate their expected waiting time. We investigated the impact of bounded rationality on revenue of a profit maximizing firm, social welfare, and pricing for both invisible and visible queues. From the firm's perspective, higher bounded rationality can lead to lower optimal prices, but it always leads to higher optimal prices and higher revenue when customers are sufficiently boundedly rational for invisible queues. With the optimal price, a little bit of bounded rationality always results in revenue losses for visible

Figure 6 **Welfare loss when hiding the queue if bounded rationality is ignored ($C = 1, \mu = 1$)**



queues. From the social planner's perspective, there may be strictly positive social welfare losses when customers are sufficiently boundedly rational for invisible queues. For visible queues with a fixed price, we prove that a little bit of bounded rationality can lead to strict social welfare improvement, and we provide a simple inequality under which this improvement happens. With the optimal prices, however, bounded rationality always decreases social welfare. We demonstrated that ignoring bounded rationality may result in significant revenue and social welfare loss. Our study contributes to the behavioral operations management literature by demonstrating the impact of behavioral factors in service settings.

Experimental Design. We hope our theoretical study helps develop a desirable theoretical-empirical feedback loop in behavioral operations management. To capture bounded rationality on the customers' part, in our model the customers have imperfect estimate of the expected waiting time. Loosely speaking, the standard deviation of the estimate of expected waiting time is used as a proxy for the level of bounded rationality.

The first step for such a study will be to estimate the level of bounded rationality β for visible and invisible queues in carefully controlled experiments. This could be done by assigning the experimental parameters to subjects and then observing their joining fractions. Standard maximum likelihood estimation would yield the estimate for β, and we can test whether it is significantly different from zero. As we mentioned in Section 2, McKelvey and Palfrey (1995) actually followed

this approach in bi-matrix game settings, and Bajari and Hortacsu (2003) in auction settings. Kremer and Debo (2011) have already carried out experimental studies in service systems and they found that the model of bounded rationality fits the experimental data very well.

There are a variety of interesting conjectures of customer behavior that could be tested or explored along the line of Kremer and Debo (2011). It is of interest to estimate β for different queue configurations. For example Larson (1987) and Maister (1985) discuss the impact of the queue environment on the customer waiting time perception. Specifically, they conjecture that eliminating empty time significantly reduces customer perception of the length of waiting time and that explained waits are shorter than unexplained waits. We conjecture that customers may have different levels of bounded rationality for these different queue environments. Some conjectures worth exploring are: Are people prone to estimation mistakes/errors (and hence bounded rationality) depending on the queue structure, for instance, with one long queue versus many short queues? Are more-educated (or more-knowledgeable) people less boundedly rational? Answering these questions not only deepens our understanding of customer behavior, but also helps managers better operate service systems.

It is interesting yet challenging to empirically quantify the loss that the firm incurs when it disregards bounded rationality. While the arrival rate λ, service rate μ, and price p are easy to estimate, the reward R, level of bounded rationality β, and cost C are a little more difficult to identify, given the possible heterogeneity and interaction among these parameters. These challenges can be overcome using instrumental analysis or other structural estimation methods. Once these parameters are estimated, we can compute the loss that the firm incurs.

Finally, recall that we have theoretically proven that a little bit of bounded rationality can improve social welfare. Searching for empirical evidence that in systems where bounded rationality exists indeed improves the social welfare would be highly valuable.

Managerial Insights. There are several implications for how service systems should be managed. First, the study demonstrates the importance of accounting for bounded rationality in pricing the service. In particular, as β increases above a certain threshold, the revenue and social welfare loss of not accounting for it is large. Second, there are settings where the system manager can reduce the ambiguity (or difficulty) associated with the process of estimating the waiting time, and potentially reduce the variance in the estimation. It is interesting to note that since customers are self-interested, when it comes to welfare maximization (e.g. public systems such as DMV and traffic on a road network, etc.) reducing β to zero might not be the right step since bounded rationality can actually improve social welfare. However, when the service provider can optimize the price as

well as reduce the level of bounded rationality significantly, then the system performance (both in terms of revenue and social welfare) can be dramatically improved.

References

Afèche, P. 2004. Incentive-compatible revenue management in queueing systems: optimal strategic delay and other delay tactics. Working paper, University of Toronto.

Anderson, S.P., A. de Palma, J.-F. Thisse. 1992. *Discrete Choice Theory of Product Differentiation*. MIT Press, Cambridge, MA.

Ariely, D. 2009. The end of rational economics. *Harvard Business Review*. July-August.

Arkes, H. R., K. R. Hammond, eds. 1985. *Judgment and Decision Making: An Interdisciplinary Reader*. Cambridge: Cambridge U. Press.

Bajari, P., A. Hortacsu. 2001. Auction models when bidders make small mistakes: Consequences for theory and estimation. *Stanford Economics Working Paper*, CA.

Bajari, P., A. Hortacsu. 2005. Are structural estimates of auction models reasonable? Evidence from experimental data, *Journal of Political Economy* 113(4) 703-741.

Basov, S. 2009. Monopolistic screening with boundedly rational consumers. Working paper.

Bearden, J. N., A. Rapoport, D. A. Seale. 2005. Entry times in queues with endogenous arrivals: Dynamics of play on the individual and aggregate levels. *Experimental Business Research*, Vol. II, 201-221.

Bendoly, E., R. Croson, P. Goncalves, K. Schultz. 2008. Domains of knowledge useful for behavioral operations management. Working paper, Emory University, Atlanta, GA.

Bendoly, E., K. Donohue, K. Schultz. 2006. Behavioral operations management: Assessing recent findings and revisiting old assumptions. *J. Oper. Management* 24(6) 737-752.

Cason, T.N., S.S. Reynolds. 2005. Bounded rationality in laboratory bargaining with asymmetric information. *Economic Theory* 25(3) 553-574.

Chen, H. C., J. W. Friedman, J. F. Thisse. 1997. Boundedly rational Nash equilibrium: A probabilistic choice approach. *Games Econom. Behav.* 18 32-54.

Conlisk, J. 1996. Why bounded rationality? *J. Econom. Literature* 34(2) 669-700.

Davis, A.M. 2011. An experimental investigation of pull contracts, Working Paper, Penn State University.

Geigerenzer, G., R. Selten. 2001. *Bounded Rationality: The Adaptive Toolbox*. MIT Press, Cambridge, MA.

Gino, F., G. Pisano. 2008. Toward a theory of behavioral operations. *Manufacturing Service Oper. Management* 10(4) 676-691.

Goeree, J., C. Holt. 2002. Quantal response equilibrium and overbidding in private-value auctions. *Journal of Economic Theory* 104 247-272.

Hassin, R. 1986. Consumer information in markets with random products quality: The case of queues and balking. *Econometrica* 54 1185-1195.

Hassin, R., M. Haviv. 2003. *To queue or not to queue: Equilibrium behavior in queueing systems.* Kluwer Academic Publishers

Haviv, M., M. L. Puterman. 1998. Bias optimality in controlled queueing systems. *Journal of Applied Probability* 35 136-150.

Hey, J., C. Orme. 1994. Investigating generalizations of expected utility theory using experimental data. *Econometrica* 62(6) 1291-1326.

Ho, T., J. Zhang. 2008. Designing pricing contracts for boundedly rational customers: Does the framing of the fixed fee matter? *Management Sci.* 54(4) 686-700.

Hofbauer, J., W. H. Sandholm. 2007. Evolution in games with randomly disturbed payoffs. *Journal of Economic Theory* 132 47-69.

Hogarth, R.. 1980. *Judgment and Choice: Psychology of Decision.* New York, Wiley.

Hsu, V., S. H. Xu, B. Jukic. 2009. Optimal scheduling and incentive compatible pricing for a service system with quality of service guarantees. *Manufacturing Service Oper. Management* 11(3) 375-396.

Kahneman, D., P. Slovic, A. Tversky, eds. 1981. *Judgment under Uncertainty: Heuristic and Biases.* Cambridge, Cambridge U. Press.

Knudsen, N.C. 1972. Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica* 40 515-528.

Kremer, M., L. Debo. 2011. Joining (and leaving) observable and unobservable queues - An experimental investigation, Presentation at INFORMS 2011 Annual Meeting at Charlotte, North Carolina.

Kremer, M., B. Moritz, E. Siemsen. 2011. Demand forecasting behavior: System neglect and change detection, *Management Sci.*, Forthcoming.

Larson, R. C. 1987. Perspectives on queues: Social justice and the psychology of queueing. *Oper. Res.* 35(6) 895-905.

Lim, N., T.-H. Ho. 2007. Designing price contracts for boundedly rational customers: Does the number of blocks matter? *Marketing Sci.* 26(3) 312-326.

Lippman, S. A., S. Stidham Jr. 1977. Individual versus social optimization in exponential congestion systems. *Operations Research* 25(2) 233-247.

Loomes, G., P. Moffatt, R. Sudgen. 2002. A microeconometric test of alternative stochastic theories of risky choice. *Journal of Risk and Uncertainty* 24(2) 103-130.

Luce, R. D. 1959. *Individual choice behavior: A theoretical analysis.* Wiley, New York.

Maister, D. 1985. The psychology of waiting lines, Chapter 8 in *The Service Encounter*, edited by John a Czepiel, Michael R. Solomon and Carol Suprenant, D.C. Heath and Company, Lexington Books.

Mallard, G. 2011. Modelling cognitively bounded rationality: An evaluative taxonomy. *Journal of Economic Surveys*

Mattsson, L.-G., J.W. Weibull. 2000. Probabilistic choice as a result of mistakes. Research Institute of Industrial Economics, Stockholm, Working paper No. 544, and Royal Institute of Technology, Stockholm, Working paper TRITA-IP AR 01-91.

Mattsson, L.-G., J.W. Weibull. 2002. Probabilistic choice and procedurally bounded rationality. *Games and Economic Behavior* 41(1) 61-78.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105-142. Academic Press, New York.

McKelvey, R. D., T. R. Palfrey. 1995. Quantal response equilibria for normal form games. *Games Econom. Behav.* 10 6-38.

Nadarajah, S. 2007. Linear combination of Gumbel random variables. *Stochastic Environmental Research and Risk Assessment* 21:283-286.

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

Nisbett, R., L. Ross. 1980. *Human inference: Strategies and shortcomings in the social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Payne, J. W., J. R. Bettman, E. J. Johnson. 1992. Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology* 43 87-131.

Pitz, G., N. J. Sachs. 1984. Judgment and decision: Theory and application. *Annual Review of Psychology* 35 139-62.

Plambeck, E. L., Q. Wang. 2010. Implications of hyperbolic discounting for optimal pricing and information management in service systems. Working paper, Stanford University, California.

Rapoport, A., W. E. Stein, J. E. Parco, D. A. Seale. 2004. Equilibrium play in single-server queues with endogenously determined arrival times. *Journal of Economic Behavior & Organization* 55 67-91.

Rubinstein, A. 1998. *Modeling Bounded Rationality.* MIT Press, Cambrige, MA.

Seale, D. A., J. E. Parco, W. E. Stein, A. Rapoport. 2005. Joining a queue or staying out: Effects of information structure and service time on arrival and staying out decisions. *Experimental Economics* 8 117-144.

Simon, H. A. 1955. A behavioral model of rational choice. *Quart. J. Econom.* 69(1) 99-118.

Simon, H. 1957. *Models of Man.* John Wiley and Sons, New York.

Stahl, D., P. Wilson. 1995. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior* 10 218-254.

Su, X. 2008. Bounded rationality in newsvendor models. *Manufacturing Service Oper. Management* 10(4) 566-589.

Talluri, K., G. J. van Ryzin. 2004. *The Theory and Practice of Revenue Management*, Springer-Verlag/Kluwer Academic Publishers.

Thurstone, L. L. 1927. A law of comparative judgment. *Psych. Rev.* 34 273-286.

Tversky, A., D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185 1124-1131.

Tversky, A., D. Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertainty* 5(4) 297-323.

Van Mieghem, J. A. 2000. Price and service discrimination in queuing systems: Incentive compatibility of G$c\mu$ scheduling. *Management Sci.* 46(9) 1249-1267.

Waksberg, A., A. Smith, M. Burd. 2009. Can boundedly rational behaviour maximise fitness? *Behav. Ecol. Sociobiol.* 63 461-471

Yechiali, U. 1971. On optimal balking rules and toll charges in a $GI/M/1$ queuing process. *Operations Research* 19 349-370.

Yechiali, U. 1972. Customers' optimal joining rules for the $GI/M/s$ queue. *Management Sci.* 18 434-443.

Appendix for

"Bounded Rationality in Service Systems"

This Appendix has six parts: Appendix A contains the proofs of the propositions in the paper. Appendix B contains all supporting lemmas and proofs of them. Appendix C contains a generalization result of Proposition 7. In Appendix D, we offer some discussion on neglecting other customers' bounded rationality, and explain why Definition 1 in the paper is appropriate. Appendix E includes a numerical study to investigate utilization and expected queue length. Appendix F shows the global stability of the equilibrium in Definition 1.

Appendix A: Proofs of Propositions

Proof of Proposition 1. Let $\lambda_n \equiv \lambda\varphi_n$, $\mu_n \equiv \mu$, then we can treat the number of customers in the queueing system as a birth-death process with birth rate λ_n and death rate μ_n. We have the balance equations: $\lambda_0 P_0 = \mu_1 P_1$, $(\lambda_n + \mu)P_n = \mu P_{n+1} + \lambda_{n-1}P_{n-1}, n \geq 1$. Solving these equations, we have the limiting probabilities: $P_0 = \frac{1}{1+\Sigma_{k=1}^{\infty}\frac{\lambda_0\lambda_1\ldots\lambda_{k-1}}{\mu^k}}, P_n = \frac{\lambda_0\lambda_1\ldots\lambda_{n-1}}{\mu^n(1+\Sigma_{k=1}^{\infty}\frac{\lambda_0\lambda_1\ldots\lambda_{k-1}}{\mu^k})}, n \geq 1$.

The necessary and sufficient condition for the existence of limiting probabilities is: $\Sigma_{k=1}^{\infty}\frac{\lambda_0\lambda_1\ldots\lambda_{k-1}}{\mu^k} < \infty$. Let $a_k = \frac{\lambda_0\lambda_1\ldots\lambda_{k-1}}{\mu^k}$, using the *ratio test*, we have $\frac{a_{k+1}}{a_k} = \frac{\lambda_k}{\mu} = \rho\varphi_k \to 0, k \to \infty$. The series converges, hence, the condition is always satisfied for $\beta \in (0, \infty)$. ■

Proof of Proposition 2. Define $g(\varphi(p,\beta_I)) \equiv \frac{e^{\frac{R-p-\frac{C}{\mu-\varphi(p,\beta_I)\lambda}}{\beta_I}}}{1+e^{\frac{R-p-\frac{C}{\mu-\varphi(p,\beta_I)\lambda}}{\beta_I}}} - \varphi(p,\beta_I)$, then $g(0) > 0$ and $g(d) = -d < 0$, where $d \equiv \min\{1, \frac{1}{\rho}\}$. $g(\varphi(p,\beta_I))$ is continuous in $\varphi(p,\beta_I)$. By the intermediate value theorem, there exists at least one $\varphi^*(p,\beta_I) \in (0,d)$ such that $g(\varphi^*(p,\beta_I)) = 0$. Furthermore, $g(\varphi(p,\beta_I))$ is strictly decreasing in $\varphi(p,\beta_I)$. Therefore, the solution is unique. ■

Proof of Proposition 3. (i) To prove this part, we need Lemma EC.1 in Appendix B, which states that the equilibrium joining fraction is monotone decreasing in the level of bounded rationality if the joining fraction is above one half. The reason we need Lemma EC.1 is the following: The fact that $\varphi(p,\beta_I) > \frac{1}{2}$ is equivalent to $R - p - \frac{C}{\mu-\varphi(p,\beta_I)\lambda} > 0$, which is equivalent to $R - p - \frac{C}{\mu-\frac{1}{2}\lambda} > 0$, i.e., $p < \overline{p}$. Parts (ii) and (iii) can be shown similarly.

(iv) For fixed β_I, denote $F(p,\varphi(p)) \equiv \frac{e^{\frac{R-p-\frac{C}{\mu-\varphi(p)\lambda}}{\beta_I}}}{1+e^{\frac{R-p-\frac{C}{\mu-\varphi(p)\lambda}}{\beta_I}}} - \varphi(p)$, then equation (2) is equivalent to $F(p,\varphi(p)) = 0$. For convenience, we denote $f \equiv f(p,\varphi(p)) = \frac{R-p-\frac{C}{\mu-\varphi(p)\lambda}}{\beta_I}$. Using the implicit function theorem, we take first derivative w.r.t. p in equation $F(p,\varphi(p)) = 0$. We have

$$\frac{e^f}{\beta_I(1+e^f)^2} + \frac{e^f\lambda C\varphi'(p)}{\beta_I(1+e^f)^2(\mu-\varphi(p)\lambda)^2} + \varphi'(p) = 0.$$

Simplifying the equation above, we obtain

$$\varphi'(p) = -\frac{e^f(\mu-\varphi(p)\lambda)^2}{\beta_I(1+e^f)^2(\mu-\varphi(p)\lambda)^2 + e^f\lambda C} < 0.$$

This completes the proof. ■

Proof of Proposition 4. We first show that $p^*(\beta) = p^*(0) - \epsilon_\beta$ for some $\epsilon_\beta > 0$ when β is strictly positive but sufficiently small. We exhaust all candidates to prove this result. First, we show that $p^*(\beta)$ cannot be equal to $p^*(0)$ when β is strictly positive but sufficiently small. By Lemma EC.13, $\Pi(p^*(0), \beta) < \Pi_{n_r}$, and $\Pi(p^*(0), 0) = \lim_{\beta \to 0} \Pi(p^*(0), \beta) < \Pi_{n_r}$, for $\beta \in (0, \bar{\beta})$ for some $\bar{\beta} > 0$. Hence, we have $\Pi(p^*(0), \beta)$ in the neighborhood of $\Pi(p^*(0), 0)$ when β is small by continuity. Now if we charge price $p = p^* - \epsilon = R - \frac{Cn_r}{\mu} - \epsilon$ for some small $\epsilon > 0$, then under full rationality, the customer who sees $n_r - 1$ customers in front of him will join the queue with ϵ utility. With the level of bounded rationality β, his joining probability would be $\varphi_{n_r - 1} = \frac{e^{\frac{\epsilon}{\beta}}}{1 + e^{\frac{\epsilon}{\beta}}} < 1$ but can be sufficiently close to 1, which implies less congestion and thus lower revenue. We have $\Pi(p^* - \epsilon, \beta) < \Pi(p^* - \epsilon, 0) < \Pi_{n_r}$, for $\beta \in (0, \bar{\beta}_\epsilon)$ for some $\bar{\beta}_\epsilon > \bar{\beta} > 0$. However, we know that $\lim_{\beta \to 0} \Pi(p^* - \epsilon, \beta) = \Pi(p^* - \epsilon, 0)$, and $\lim_{\epsilon \to 0} \Pi(p^* - \epsilon, 0) = \Pi_{n_r}$. Hence, $\Pi(p^* - \epsilon, \beta)$ can be made arbitrarily close to Π_{n_r} when β is small and ϵ is also small. Hence, we have $\Pi(p^*(0), \beta) < \Pi(p^* - \epsilon, \beta)$, when β is small and ϵ is also small. This shows that $p^*(0)$ cannot be the optimal price when customers are slightly boundedly rational.

Next we show that any price taking this form $p = p^* + \epsilon = R - \frac{Cn_r}{\mu} + \epsilon$ for some fixed small $\epsilon > 0$ cannot be the optimal price either. Under full rationality, the customer who observes $n_r - 1$ customers in front of him will not join the queue, and the revenue is strictly higher if the price $p_1 = R - \frac{C(n_r+1)}{\mu}$ is charged instead since this modification will still induce the same number of customers to join and the revenue per customer is strictly higher. We have $\Pi(p^* + \epsilon, 0) < \Pi_{n_r+1} \leq \Pi_{n_r}$. Furthermore, we have $\Pi(p^* + \epsilon, \beta) < \Pi_{n_r+1}$ for $\beta \in (0, \bar{\beta}_{\epsilon 1})$ for some $\bar{\beta}_{\epsilon 1} > 0$. Hence, $\Pi(p^* + \epsilon, \beta) < \Pi(p^* - \epsilon_1, \beta)$ when β is small and ϵ_1 is also small.

Other prices "faraway from" $p^*(0)$ clearly cannot be the optimal price when β is small. Therefore, $p^*(\beta) = p^*(0) - \epsilon_\beta$ for some $\epsilon_\beta > 0$ when β is small.

Finally, we need to verify the existence of the optimal price $p^*(\beta)$. For any $\beta > 0$, we know $\Pi(p, \beta)$ is continuous over the closed interval $[p^*(0) - \frac{C}{\mu}, p^*(0)]$. Hence, there exists some $p^*(\beta) \in [p^*(0) - \frac{C}{\mu}, p^*(0)]$ to maximize $\Pi(p, \beta)$.

We have shown that the optimal price under a little bit of bounded rationality is strictly lower than the optimal price under full rationality. Now we can prove the second part of the proposition by noting the following: $\Pi(p^*(\beta), \beta) = \Pi(p^* - \epsilon_\beta, \beta) < \Pi(p^* - \epsilon_\beta, 0) < \Pi_{n_r}$, for $\beta \in (0, \bar{\beta}_{\epsilon_\beta})$ for some $\bar{\beta}_{\epsilon_\beta} > 0$. ■

Proof of Proposition 5. (i) Recall that $\Pi^I(p, \beta_I) = p\varphi(p, \beta_I)\lambda$, where $\varphi(p, \beta_I)$ is the unique solution to equation (2). For any fixed β_I, we shall abuse notation by writing $\Pi^I(p)$ and $\varphi(p)$ for brevity. Taking first derivative, we have $\Pi^{I'}(p) = \lambda[\varphi(p) + p\varphi'(p)]$, where $'$ denotes the derivative. We have shown that $\varphi'(p) < 0$. Let $R_h'(p) = 0$, we have $p = -\frac{\varphi(p)}{\varphi'(p)} > 0$. Now we investigate whether this necessary FOC has multiple solutions. Substituting $\varphi'(p)$, we have

$$p = \frac{\lambda C \varphi(p)}{(\mu - \varphi(p)\lambda)^2} + \frac{\beta_I}{1 - \varphi(p)}.$$

We are interested in whether this equation has a unique solution. For exposition convenience, we denote $g(p) \equiv \frac{\lambda C \varphi(p)}{(\mu - \varphi(p)\lambda)^2} + \frac{\beta_I}{1 - \varphi(p)} - p$, then the question is whether $g(p) = 0$ has a unique solution. We claim that $g(p)$ is strictly decreasing in p. It is clear that the first term in the RHS $\frac{\lambda C \varphi(p)}{(\mu - \varphi(p)\lambda)^2}$ is strictly decreasing in p

and so are the second and third terms. Hence $g(p)$ is strictly decreasing in p. Note $g(0)>0$, and $g(\infty)=-\infty$, so there exists a unique p^* such that $g(p^*)=0$. Finally, one can verify the second-order condition is satisfied.

(ii) We know $p^*(\beta_I)$ solves the following equation

$$p^*(\beta_I)=\frac{\lambda C\varphi(p^*(\beta_I),\beta_I)}{(\mu-\varphi(p^*(\beta_I),\beta_I)\lambda)^2}+\frac{\beta_I}{1-\varphi(p^*(\beta_I),\beta_I)}.$$

Using the implicit function theorem and after simplifying, we have

$$p^{*'}(\beta_I)=\frac{A\frac{\partial\varphi(p,\beta_I)}{\partial\beta_I}+(\mu-\varphi(p,\beta_I)\lambda)^3(1-\varphi(p,\beta_I))}{(\mu-\varphi(p,\beta_I)\lambda)^3(1-\varphi(p,\beta_I))^2-A\frac{\partial\varphi(p,\beta_I)}{\partial p}}.$$

Note that $A>0$ and $\frac{\partial\varphi(p,\beta_I)}{\partial p}<0$, we know the denominator is strictly positive. Hence, $p^{*'}(\beta_I)$ has the same sign as the numerator. If $\frac{\partial\varphi(p,\beta_I)}{\partial\beta_I}\geq 0$, i.e., $p^*(\beta_I)\geq\overline{p}$ (by Proposition 3), then $p^{*'}(\beta_I)>0$. Otherwise, the numerator can be negative depending on the parameters. Hence, $p^{*'}(\beta_I)>0$ if $p^*(\beta_I)\geq\overline{p}$; Otherwise, $p^{*'}(\beta_I)$ has the same sign as $A\frac{\partial\varphi(p,\beta_I)}{\partial\beta_I}+(\mu-\varphi(p,\beta_I)\lambda)^3(1-\varphi(p,\beta_I))$, where $A\equiv\lambda C(\mu+\lambda\varphi(p,\beta_I))(1-\varphi(p,\beta_I))^2+\beta_I(\mu-\varphi(p,\beta_I)\lambda)^3$.

To further simplify the result in terms of primitives such as β_I, we want to know whether the price $\overline{p}$ (which leads to equilibrium joining fraction 0.5 regardless of the level of bounded rationality) can be optimal. Suppose it were, then we have the condition by plugging it into the condition the optimal price has to satisfy,

$$R-\frac{2C}{2\mu-\lambda}=\frac{2\lambda C}{(2\mu-\lambda)^2}+2\beta_I.$$

Simplifying yields

$$\beta_0=\frac{1}{2}R-\frac{2\mu C}{(2\mu-\lambda)^2}.$$

If it is positive, i.e., $R\geq\frac{4\mu C}{(2\mu-\lambda)^2}$, we know $\frac{dp^*(\beta_0)}{d\beta}>0$, since $\frac{\partial\varphi(\overline{p},\beta_0)}{\partial\beta_I}=0$. Then, it is clear that for any $\beta_I>\beta_0$, we have $\frac{dp^*(\beta_I)}{d\beta_I}>0$. On the other hand, if $R<\frac{4\mu C}{(2\mu-\lambda)^2}$, so that $\beta_0<0$, one can easily compute $p^*(0)=R(1-\sqrt{\frac{C}{\mu R}})$. If $p^*(0)\geq\overline{p}$, which is equivalent to $R\leq\frac{4C\mu}{(2\mu-\lambda)^2}$, we have $\frac{dp^*(0)}{d\beta_I}>0$. This then implies that $\frac{dp^*(\beta_I)}{d\beta_I}>0$ for any $\beta_I\in[0,\infty)$. ■

Proof of Corollary 1. Using the envelope theorem, we have

$$\frac{d\Pi^I(p^*(\beta_I),\beta_I)}{d\beta_I}=p^*(\beta_I)\lambda\frac{\partial\varphi(p^*(\beta_I),\beta_I)}{\partial\beta_I}.$$

Then all the results (i), (ii) and (iii) follow directly from Proposition 3 (i)-(iii). ■

Proof of Proposition 6. By definition, $\Pi^{I'}(\lambda)=p[\varphi'(\lambda)\lambda+\varphi(\lambda)]$. To determine its sign, we want to first determine the sign of $\varphi'(\lambda)$. There are at least two ways to do this. First, observe the equilibrium condition, equation (2). Suppose $\varphi(\lambda)$ is increasing in λ, then the LHS is decreasing while the RHS is increasing, a contradiction. Hence, $\varphi^{*'}(\lambda)<0$. The other way is to derive this derivative using the implicit function theorem. For convenience, denote $f\equiv f(\lambda)=\frac{R-p-\frac{C}{\mu-\varphi(\lambda)\lambda}}{\beta_I}$, and $F(\lambda,\varphi(\lambda))\equiv\frac{e^f}{1+e^f}-\varphi(\lambda)$. The equilibrium condition amounts to $F(\lambda,\varphi(\lambda))=0$. Taking first derivative and simplifying, we have

$$\frac{Ce^f}{\beta_I(1+e^f)^2}\frac{\varphi'(\lambda)\lambda+\varphi(\lambda)}{(\mu-\varphi(\lambda)\lambda)^2}+\varphi'(\lambda)=0,$$

which clearly implies

$$\frac{d\varphi^*(\lambda)}{d\lambda}\frac{d\Pi^I(\lambda)}{d\lambda}<0,$$

and $\frac{d\varphi^*(\lambda)}{d\lambda}<0$. ■

Proof of Proposition 7. To show this proposition, we study the social welfare function $W(\beta)$ as β is strictly greater than but arbitrarily close to 0, and compare it with $W(0)$.

We start from the case when only the customers on the two marginal states on the positive and the negative side randomize. Let $\sigma(\beta)$ be the probability of joining for the customer who sees n_s-1 customers in the queue in front of him, and $\delta(\beta)$ be the probability of joining for the customer who observes n_s customers in the queue. We omit β for simplicity. Let $u_0 \equiv U_{n_s-1}$ and $u_1 \equiv U_{n_s}$ be their expected utilities of joining respectively.

If $n_s > \frac{R\mu}{C} - \frac{1}{2}$, we have $u_0 \geq 0, u_1 \leq 0, u_0+u_1 \leq 0$. By Lemma EC.10 which gives conditions under which less congestion implies more welfare holds for any number of customers joining using logit probabilities, we need to show

$$(1-\sigma)\rho^{n_s} + \sigma(1-\delta)\rho^{n_s+1} + \sigma\delta\rho^{n_s+2} < \rho^{n_s+1}, \tag{EC.1}$$

when β is small. Some algebra tells us that it suffices to show

$$M(\beta) \equiv \sigma(\delta\rho+1) = \frac{e^{\frac{u_0}{\beta}}}{1+e^{\frac{u_0}{\beta}}}(\rho\frac{e^{\frac{u_1}{\beta}}}{1+e^{\frac{u_1}{\beta}}}+1) < 1.$$

We want to know the sign of $M'(\beta)$ when β is small. After lengthy algebra, we have

$$\frac{M'(\beta)(1+e^{\frac{u_0}{\beta}})^2(1+e^{\frac{u_1}{\beta}})^2\beta^2}{e^{\frac{u_0}{\beta}}} = -\rho u_1 e^{\frac{u_0+u_1}{\beta}} - (\rho+1)u_0 e^{\frac{2u_1}{\beta}} - [(\rho+2)u_0+\rho u_1]e^{\frac{u_1}{\beta}} - u_0 < 0,$$

when $\beta \in (0, \beta_1^*)$, where $M'(\beta_1^*) = 0$.

Observing this inequality, we can see that, if $u_0+u_1>0, u_0>0, u_1<0$, i.e., $n_s < \frac{R\mu}{C} - \frac{1}{2}$, then $\lim_{\beta\to 0} M'(\beta) > 0$. Hence, the social welfare will decrease in this case. If $u_0+u_1=0, u_0>0, u_1<0$, i.e., $n_s = \frac{R\mu}{C} - \frac{1}{2}$, then $\lim_{\beta\to 0} M'(\beta) > 0$. If $u_0 = 0, u_1 < 0$, then $M'(\beta)$ has the same sign as $-2\rho u_1 e^{\frac{u_1}{\beta}}$ when β is close to 0, i.e., $\lim_{\beta\to 0} M'(\beta) > 0$.

Next, let us consider the case when the customers in the two marginal states on the positive side and two marginal states on the negative side join with some positive probabilities. Let $\sigma_1(\beta), \sigma(\beta), \delta(\beta), \delta_1(\beta)$ be the probabilities of joining for the customers who observe n_s-2, n_s-1, n_s, n_s+1 customers respectively. We will also omit their dependence on β for brevity hereafter. We want to show

$$(1-\sigma_1)\rho^{n_s-1} + \sigma_1(1-\sigma)\rho^{n_s} + \sigma_1\sigma(1-\delta)\rho^{n_s+1} + \sigma_1\sigma\delta(1-\delta_1)\rho^{n_s+2} + \sigma_1\sigma\delta\delta_1\rho^{n_s+3} < \rho^{n_s+1}, \tag{EC.2}$$

when β is small.

One obvious way to show this inequality is to use the similar technique, differentiation, as for the case where two customer join with non-degenerate probabilities. But it turns out to be untractable. Here is the technique we follow: We already know when $\beta \in (0, \beta_1^*)$, Inequality (EC.1) holds. Then, it suffices to show

$$(1-\sigma_1) + \sigma_1\rho^2 - \sigma_1\sigma\delta\delta_1\rho^3 + \sigma_1\sigma\delta\delta_1\rho^4 < \rho^2$$

which is equivalent to

$$\frac{\sigma_1\sigma\delta\delta_1}{1-\sigma_1} < \frac{\rho^2-1}{\rho^3(\rho-1)}.$$

This inequality can clearly be satisfied for $\beta \in (0, \beta_2^*)$, where β_2^* makes the inequality above equal. Hence, when $\beta \in (0, \min\{\beta_1^*, \beta_2^*\})$, inequality (EC.2) is satisfied.

Before we generalize our result, let us also consider the case when the customers in the three marginal states on the positive side and three marginal states on the negative side join with non-degenerate probabilities. Let $\sigma_2(\beta), \sigma_1(\beta), \sigma(\beta), \delta(\beta), \delta_1(\beta), \delta_2(\beta)$ be the probabilities of joining for those who see $n_s - 3, n_s - 2, ..., n_s + 2$ customers respectively. We need to show

$$\begin{aligned}&(1-\sigma_2)\rho^{n_s-2} + \sigma_2(1-\sigma_1)\rho^{n_s-1} + \sigma_2\sigma_1\sigma(1-\sigma)\rho^{n_s} + \sigma_2\sigma_1\sigma(1-\delta)\rho^{n_s+1} + \sigma_2\sigma_1\sigma\delta(1-\delta_1)\rho^{n_s+2} + \\ &\sigma_2\sigma_1\sigma\delta\delta_1(1-\delta_2)\rho^{n_s+3} + \sigma_2\sigma_1\sigma\delta\delta_1\delta_2\rho^{n_s+4} < \rho^{n_s+1},\end{aligned} \tag{EC.3}$$

when β is small. We know, when $\beta \in (0, \min\{\beta_1^*, \beta_2^*\})$, inequality (EC.2) is satisfied. Hence, it suffices to show

$$1 - \sigma_2 + \sigma_2\rho^3 - \sigma_2\sigma_1\sigma\delta\delta_1\delta_2\rho^5 + \sigma_2\sigma_1\sigma\delta\delta_1\delta_2\rho^6 < \rho^3.$$

which is equivalent to

$$\frac{\sigma_2\sigma_1\sigma\delta\delta_1\delta_2}{1-\sigma_2} < \frac{\rho^3 - 1}{\rho^5(\rho - 1)}.$$

This inequality can be satisfied for $\beta \in (0, \beta_3^*)$, where β_3^* makes the inequality above be equality. Hence, when $\beta \in (0, \min\{\beta_1^*, \beta_2^*, \beta_3^*\})$, inequality (EC.3) is satisfied.

Clearly, we can proceed as this until the first arrival customer joins with a positive probability, i.e., $2n_s$ customers join with positive probabilities. For the case when any $2n + 2$ customers join with positive probabilities, $n \leq n_s - 1$, we have the inequality to be satisfied

$$\frac{\sigma_n ... \sigma_2\sigma_1\sigma\delta\delta_1\delta_2 ... \delta_n}{1-\sigma_n} < \frac{\rho^{n+1} - 1}{\rho^{2n+1}(\rho - 1)}. \tag{EC.4}$$

This inequality can be satisfied for $\beta \in (0, \beta_n^*)$, where β_n^* makes the inequality above be equality. Hence, when $\beta \in (0, \min\{\beta_1^*, \beta_2^*, \beta_3^*, ..., \beta_n^*\})$, we are done.

Now, we need to consider the cases, when the customers in the $2n$ marginal states join with some positive probabilities, where $n > n_s$. For example, for $n = n_s + 1$, we need to show

$$\begin{aligned}&(1-\sigma_{n_s-1})\rho + \sigma_{n_s-1}(1-\sigma_{n_s-2})\rho^2 + \sigma_{n_s-1}\sigma_{n_s-2}\sigma(1-\sigma_{n_s-3})\rho^3 + ... + \sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...(1-\delta_{n_s-1})\rho^{2n_s} \\ &+ \sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...\delta_{n_s-1}(1-\delta_{n_s})\rho^{2n_s+1} + \sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...\delta_{n_s-1}\delta_{n_s}\rho^{2n_s+2} < \rho^{n_s+1}\end{aligned} \tag{EC.5}$$

when β is small.

Let $x(\beta) \in (\sigma_1, 1)$ be such that $x(\beta) = y + (1-y)\sigma_1$, where $y \in (0, 1)$, then it is easy to verify that inequality (EC.3) can be modified to

$$\begin{aligned}&(1-\sigma_2)\rho^{n_s-2} + \sigma_2(1-\sigma_1)\rho^{n_s-1} + \sigma_2\sigma_1\sigma(1-\sigma)\rho^{n_s} + \sigma_2\sigma_1\sigma(1-\delta)\rho^{n_s+1} + \sigma_2\sigma_1\sigma\delta(1-\delta_1)\rho^{n_s+2} + \\ &\sigma_2\sigma_1\sigma\delta\delta_1(1-\delta_2)\rho^{n_s+3} + \sigma_2\sigma_1\sigma\delta\delta_1\delta_2\rho^{n_s+4} < x(\beta)\rho^{n_s+1}\end{aligned} \tag{EC.6}$$

when $\beta \in (0, \beta_1)$ for some $\beta_1 > 0$. In general, inequality (EC.4) can be modified to

$$\frac{\sigma_n ... \sigma_2\sigma_1\sigma\delta\delta_1\delta_2 ... \delta_n}{1-\sigma_n} < \frac{y\rho^{n+1} - 1}{\rho^{2n+1}(\rho - 1)}. \tag{EC.7}$$

Then, it suffices to show

$$\delta_{n_s}\rho^{2n_s+1}(\rho - 1) < (1-y)\rho^{n_s+1}, \tag{EC.8}$$

which is equivalent to

$$\delta_{n_s} < \frac{1-y}{\rho_s^n(\rho-1)}, \tag{EC.9}$$

which can clearly be satisfied for $\beta \in (0, \beta_y)$ for some $\beta_y > 0$.

When the number of marginal states in which the customers randomize goes to infinity (by Proposition 1, the system is always stable), then we need to show the following summable series is less than ρ^{n_s+1}:

$$\begin{aligned}&(1-\sigma_{n_s-1})\rho + \sigma_{n_s-1}(1-\sigma_{n_s-2})\rho^2 + \sigma_{n_s-1}\sigma_{n_s-2}\sigma(1-\sigma_{n_s-3})\rho^3 + ... + \sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...(1-\delta_{n_s-1})\rho^{2n_s}\\&+ [\sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...\delta_{n_s-1}(1-\delta_{n_s})\rho^{2n_s+1} + \sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...\delta_{n_s-1}\delta_{n_s}(1-\delta_{n_s+1})\rho^{2n_s+2}\\&+ \sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...\delta_{n_s}\delta_{n_s+1}(1-\delta_{n_s+2})\rho^{2n_s+3} + ...] < \rho^{n_s+1}.\end{aligned} \tag{EC.10}$$

We can show that the part in [.] can be made less than $(1-y)(1-\sigma_1)\rho^{n_s+1}$ as $\beta \in (0, \varepsilon^*)$ as follows:

$$\begin{aligned}&\sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...\delta_{n_s-1}(1-\delta_{n_s})\rho^{2n_s+1} + \sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...\delta_{n_s-1}\delta_{n_s}(1-\delta_{n_s+1})\rho^{2n_s+2}\\&+ \sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...\delta_{n_s}\delta_{n_s+1}(1-\delta_{n_s+2})\rho^{2n_s+3} + ...\\&= \rho^{2n_s+1}\sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...\delta_{n_s-1}[(1-\delta_{n_s}) + \delta_{n_s}(1-\delta_{n_s+1})\rho + \delta_{n_s}\delta_{n_s+1}(1-\delta_{n_s+2})\rho^2 + ...]\\&\le \rho^{2n_s+1}(\sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...\delta_{n_s-1})[1 + \delta_{n_s}\rho + \delta_{n_s}^2\rho^2 + ...]\\&= \rho^{2n_s+1}(\sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...\delta_{n_s-1})\frac{1}{1-\delta_{n_s}\rho} < (1-y)(1-\sigma_1)\rho^{n_s+1}.\end{aligned} \tag{EC.11}$$

The last inequality comes from

$$\frac{\sigma_{n_s-1}\sigma_{n_s-2}...\sigma\delta\delta_1...\delta_{n_s-1}}{(1-\delta_{n_s}\rho)(1-\sigma_1)} < \frac{\delta\delta_1...\delta_{n_s-1}}{(1-\delta_{n_s}\rho)(1-\sigma_1)} < \frac{1-y}{\rho^{n_s}} \tag{EC.12}$$

which can easily be satisfied as β is small.

The case when $n_s < \frac{R\mu}{C} - \frac{1}{2}$ or $n_s = \frac{R\mu}{C} - \frac{1}{2}$ can be shown by similar arguments using Lemma EC.8 which states the equivalent results for the case when only the customers on the two marginal states join with logit probabilities, Lemma EC.10, and Lemma EC.11 which give conditions under which more congestion implies less welfare holds for any number of customers joining using logit probabilities. The proofs are omitted for brevity. Hence, we have completed the proof. ■

Proof of Proposition 8. Given Lemma EC.12 in Appendix B which shows that for any price charged in the interval $(R - \frac{C(n_0+1)}{\mu}, R - \frac{Cn_0}{\mu}]$, the conclusion holds, we only need to show when p is outside of the interval $(R - \frac{C(n_0+1)}{\mu}, R - \frac{Cn_0}{\mu}]$, the conclusion continues to hold.

We use the same argument as Lemma EC.12. We know that W_{n_0} is the optimal social welfare by Yechiali (1971). However, we cannot rule out the case that $W(p,\beta) = W_{n_0}$ for some p from Yechiali (1971)'s results. To rule out the case, we use Haviv and Puterman (1998), who show that the only average optimal stationary policies are of control limit type, that there are at most two and, if there are two, they occur consecutively. This implies that the only gain optimal randomized stationary policies should randomize over the two control limit states if they exist. The argument is simple: for any randomized policy to be optimal, the deterministic policies it has strictly positive probabilities should yield the same average reward. In our setting with randomization using logit probabilities, their result implies that W_{n_0} is strictly larger than any $W(p,\beta)$ when $\beta > 0$ since the logit joining probabilities are in the interval (0,1). ■

Proof of Proposition 9. (i) According to Lemma EC.3 which gives conditions under which the social welfare is increasing or decreasing in the level of bounded rationality in Appendix B: if $\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda}=\frac{1}{2}$ and $p=\overline{p}$, then $W^I(\beta_I)$ is constant for $\beta_I\geq 0$. Simplifying the conditions yields result (i).

(ii) According to Lemma EC.3: If $(\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda}-\frac{1}{2})(p-R+\frac{2C}{2\mu-\lambda})>0$, then $W^I(\beta_I)$ strictly increases for all $\beta_I\geq 0$. Combining and simplifying these states in terms of $p^*(0)$ and $\overline{p}$ yields the results.

(iii) According to Lemma EC.3: If either $\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda}>\frac{1}{2}$ and $p\in[R-\sqrt{\frac{CR}{\mu}},R-\frac{2C}{2\mu-\lambda})$, or $\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda}<\frac{1}{2}$ and $p\in(R-\frac{2C}{2\mu-\lambda},R-\sqrt{\frac{CR}{\mu}}]$, then $W^I(\beta_I)$ strictly decreases for all $\beta_I\geq 0$. Combining these cases together yields the result in (iii).

(iv) According to Lemma EC.3: If either $\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda}>\frac{1}{2}$ and $p<\min\{R-\frac{2C}{2\mu-\lambda},R-\sqrt{\frac{CR}{\mu}}\}$, or $\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda}<\frac{1}{2}$ and $p>\max\{R-\frac{2C}{2\mu-\lambda},R-\sqrt{\frac{CR}{\mu}}\}$, then $W^I(\beta_I)$ strictly increases in $[0,\beta_w)$ then strictly decreases in (β_w,∞). Again, combining and simplifying these states in terms of $p^*(0)$ and $\overline{p}$ yields the results. ■

Proof of Proposition 10. (i) Lemma EC.4 in Appendix B shows that the social welfare function is unimodal in the price, which allows us to invoke the first-order condition to find the optimal price. For any fixed $\lambda>0$ and level of bounded rationality $\beta_I>0$, to achieve social optimality, the equilibrium joining fraction $\varphi(p,\beta_I)=\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda_I}$, plugging which into the equilibrium condition equation (2), we have

$$\frac{e^{\frac{R-p-\sqrt{\frac{CR}{\mu}}}{\beta_I}}}{1+e^{\frac{R-p-\sqrt{\frac{CR}{\mu}}}{\beta_I}}}=\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda}.$$

Solving this equation, we have the unique solution

$$p_w^*(\beta_I)=R-\sqrt{\frac{CR}{\mu}}-\beta_I\log\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda-\mu+\sqrt{\frac{C\mu}{R}}}.$$

We can easily calculate the optimal social welfare at the optimal price $p_w^*(\beta)$ if it is positive (which is satisfied when the conditions stated in part (i) on R and β are satisfied). Otherwise, we have to let price be zero to maximize the social welfare. Part (ii) follows similarly. ■

Appendix B: Lemmas and Proofs

LEMMA EC.1. *For any fixed price p, $\varphi(p,\beta_I)$ is strictly decreasing in β_I when $\varphi(p,\beta_I)>\frac{1}{2}$, strictly increasing in β_I when $\varphi(p,\beta_I)<\frac{1}{2}$, and constant in β_I when $\varphi(p,\beta_I)=\frac{1}{2}$.*

Proof of Lemma EC.1. When $\varphi(p,\beta_I)>\frac{1}{2}$, we have $R-p-\frac{C}{\mu-\varphi(p,\beta_I)\lambda}>0$. We prove the conclusion by contradiction. Suppose $\Pi^I(p,\beta_I)$ were increasing in β_I, then $\varphi(p,\beta_I)$ and thus the LHS of equation (2) in the paper increases in β_I. Note that, $R-p-\frac{C}{\mu-\varphi(p,\beta_I)\lambda}$ decreases in $\varphi(p,\beta_I)$, and thus $\frac{R-p-\frac{C}{\mu-\varphi(p,\beta_I)\lambda}}{\beta_I}$ decreases as β_I increases. Hence the RHS of equation (2) decreases, while the LHS $\varphi(p,\beta_I)$ increases, which is a contradiction. Therefore, $\varphi(p,\beta_I)$ decreases in β_I when $\varphi(p,\beta_I)>\frac{1}{2}$. Similar arguments for other two cases.

Another way to prove it is taking first derivative w.r.t. β_I directly in equation (2) using the implicit function theorem. In fact, we have

$$[1+\frac{\lambda Ce^f}{\beta_I(1+e^f)^2(\mu-\varphi(p,\beta_I)\lambda)^2}]\frac{\partial\varphi(p,\beta_I)}{\partial\beta_I}=-\frac{fe^f}{\beta_I(1+e^f)^2},$$

where $f \equiv f(p,\beta_I,\varphi(p,\beta_I)) = \frac{R-p-\frac{C}{\mu-\varphi(p,\beta_I)\lambda}}{\beta_I}$. Hence, $\frac{\partial\varphi(p,\beta_I)}{\partial\beta_I}$ has the same sign as $-f$:

$$sign\frac{\partial\varphi(p,\beta_I)}{\partial\beta_I} = -signf.$$

Indeed, after simplifying, we have

$$\frac{\partial\varphi(p,\beta_I)}{\partial\beta_I} = \frac{-\ln\frac{\varphi(p,\beta_I)}{1-\varphi(p,\beta_I)}(\mu-\varphi(p,\beta_I)\lambda)^2\varphi(p,\beta_I)(1-\varphi(p,\beta_I))}{\beta_I(\mu-\varphi(p,\beta_I)\lambda)^2+\lambda C\varphi(p,\beta_I)(1-\varphi(p,\beta_I))},$$

where we used the fact that $f = \ln\frac{\varphi(p,\beta_I)}{1-\varphi(p,\beta_I)}$. ■

LEMMA EC.2. *$W^I(\varphi(p,\beta_I))$ is strictly concave in $\varphi(p,\beta_I)$, i.e., $\frac{\partial^2 W^I(\varphi(p,\beta_I))}{\partial(\varphi(p,\beta_I))^2} < 0$.*

Proof of Lemma EC.2. Taking first-order derivative, we have

$$\frac{\partial W^I(\varphi(p,\beta_I))}{\partial\varphi(p,\beta_I)} = \lambda R - \frac{C\lambda\mu}{(\mu-\varphi(p,\beta_I)\lambda)^2}.$$

Taking second-order derivative, we have

$$\frac{\partial^2 W^I(\varphi(p,\beta_I))}{\partial(\varphi(p,\beta_I))^2} = -\frac{2\lambda^2\mu C}{(\mu-\varphi(p,\beta_I)\lambda)^3} < 0,$$

which completes the proof. ■

LEMMA EC.3. *$\frac{dW^I(\beta_I)}{d\beta_I} > 0$ if $\varphi(p,\beta_I) > \max\{\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda}, \frac{1}{2}\}$ or $\varphi(p,\beta_I) < \min\{\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda}, \frac{1}{2}\}$; $\frac{dW^I(\beta_I)}{d\beta_I} < 0$ if $\varphi(p,\beta_I) < \max\{\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda}, \frac{1}{2}\}$ and $\varphi(p,\beta_I) > \min\{\frac{\mu-\sqrt{\frac{C\mu}{R}}}{\lambda}, \frac{1}{2}\}$; $\frac{dW^I(\beta_I)}{d\beta_I} = 0$ otherwise.*

Proof of Lemma EC.3. Using the chain rule, we have

$$\frac{dW^I(\beta_I)}{d\beta_I} = \frac{dW^I(\varphi(\beta_I))}{d\varphi(\beta_I)}\frac{\varphi(\beta_I)}{d\beta_I} = [\lambda R - \frac{C\lambda\mu}{(\mu-\varphi(p,\beta_I)\lambda)^2}]\frac{\varphi(\beta_I)}{d\beta_I},$$

which has the same sign as $-[\lambda R - \frac{C\lambda\mu}{(\mu-\varphi(p,\beta_I)\lambda)^2}]f$, where $f \equiv f(\beta_I) = \frac{R-p-\frac{C}{\mu-\varphi(\beta_I)\lambda}}{\beta_I}$. Then, the lemma follows from determining the sign by discussing all possible cases. ■

LEMMA EC.4. *$W^I(p)$ is unimodal in the price p.*

Proof of Lemma EC.4. By the chain rule, we have $W^{I'}(p) = W^{I'}(\varphi(p))\varphi'(p)$, but $\varphi'(p) < 0$, hence, the FOC $W^{I'}(p) = 0$ is equivalent to $W^{I'}(\varphi(p)) = 0$. By Lemma EC.2 in Appendix B, a unique solution $\varphi(p)$ exists. Again since $\varphi'(p) < 0$, a unique solution p^* exists. ■

LEMMA EC.5. *If $n_s \neq n_0$ and all the customers are fully rational except the one who observes $n_s - 1$ customers in the queue and who joins with the fraction 0.5, then the social welfare would be strictly improved. Moreover, either of the two conditions is sufficient for $n_s \neq n_0$: (1) $\rho > 1$ and $n_s > 1$; (2) $\sqrt{2}-1 < \rho < 1$ and $n_s > 2$.*

Proof of Lemma EC.5. In this case where only the customer who observes $n_s - 1$ customers in the queue is boundedly rational, we can derive an explicit expression for the social welfare function as follows. For the birth-death process, we have the birth rates, $\lambda_n = \lambda$, when $n = 0, 1, ..., n_s - 2$; $\lambda_n = \frac{1}{2}\lambda$, when $n = n_s - 1$; $\lambda_n = 0$, when $n \geq n_s$. Then, we have:

$$a_n = \begin{cases} \rho^n \text{ if } n < n_s \\ \frac{1}{2}\rho^n \text{ if } n = n_s \\ 0 \text{ if } n > n_s. \end{cases}$$

$$\begin{aligned}
A &\equiv \sum_{n=1}^{\infty} a_n = (a_1 + a_2 + ... + a_{n_s-1}) + a_{n_s} \\
&= (\rho + \rho^2 + ... + \rho^{n_s-1}) + \frac{1}{2}\rho^{n_s} \\
&= \frac{\rho(1-\rho^{n_s-1})}{1-\rho} + \frac{1}{2}\rho^{n_s}. \\
P_0 &= \frac{1}{1+A} = \frac{2(1-\rho)}{2-\rho^{n_s}-\rho^{n_s+1}}. \\
P_n &= a_n P_0, n \geq 1. \\
F &\equiv \sum_{n=0}^{\infty} \lambda_n P_n = \sum_{n=0}^{n_s} \lambda_n P_n \\
&= \lambda P_0[(1 + \rho + ... + \rho^{n_s-2}) + \frac{1}{2}\rho^{n_s-1}] \\
&= \lambda \frac{2-\rho^{n_s-1}-\rho^{n_s}}{2-\rho^{n_s}-\rho^{n_s+1}}. \\
G &\equiv \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{n_s} nP_n \\
&= (\sum_{n=1}^{n_s} n\rho^n - \frac{1}{2}n_s\rho^{n_s})P_0 \\
&= [\frac{\rho^2(1-\rho^{n_s})}{(1-\rho)^2} + \frac{\rho[1-(n_s+1)\rho^{n_s}]}{1-\rho} - \frac{1}{2}n_s\rho^{n_s}]P_0 \\
&= [\frac{\rho^2(1-\rho^{n_s})}{(1-\rho)^2} + \frac{\rho[1-(n_s+1)\rho^{n_s}]}{1-\rho} - \frac{1}{2}n_s\rho^{n_s}]\frac{2(1-\rho)}{2-\rho^{n_s}-\rho^{n_s+1}} \\
&= \frac{\rho}{1-\rho} - \frac{(n_s+1)\rho^{n_s+1}+n_s\rho^{n_s}}{2-\rho^{n_s}-\rho^{n_s+1}}.
\end{aligned}$$

Then, we have the social welfare in this case

$$W_2 = FR - GC = \lambda \frac{2-\rho^{n_s-1}-\rho^{n_s}}{2-\rho^{n_s}-\rho^{n_s+1}}R - [\frac{\rho}{1-\rho} - \frac{(n_s+1)\rho^{n_s+1}+n_s\rho^{n_s}}{2-\rho^{n_s}-\rho^{n_s+1}}]C.$$

We want to show $W_2 - W(0) > 0$, which is equivalent to $\lambda RK - CT > 0$, where

$$K \equiv \frac{2-\rho^{n_s-1}-\rho^{n_s}}{2-\rho^{n_s}-\rho^{n_s+1}} - \frac{1-\rho^{n_s}}{1-\rho^{n_s+1}},$$

and

$$T \equiv \frac{(n_s+1)\rho^{n_s+1}}{1-\rho^{n_s+1}} - \frac{(n_s+1)\rho^{n_s+1}+n_s\rho^{n_s}}{2-\rho^{n_s}-\rho^{n_s+1}}.$$

Basic algebra yields

$$K = \frac{-(\rho-1)^2\rho^{n_s-1}}{(\rho^{n_s+1}-1)(\rho^{n_s}+\rho^{n_s+1}-2)} < 0,$$

and

$$T = \frac{(n_s+1)\rho^{n_s+1} - n_s\rho^{n_s} - \rho^{2n_s+1}}{(\rho^{n_s+1}-1)(\rho^{n_s}+\rho^{n_s+1}-2)}.$$

Then

$$\frac{T}{K} = \frac{(n_s+1)\rho^{n_s+1} - n_s\rho^{n_s} - \rho^{2n_s+1}}{-(\rho-1)^2\rho^{n_s-1}} = \frac{\rho^{n_s+2} + n_s\rho - (n_s+1)\rho^2}{(\rho-1)^2}. \tag{EC.13}$$

But $\lambda RK - CT > 0 \Leftrightarrow \frac{T}{K} > \frac{\lambda R}{C} = \frac{R\mu}{C}\rho = (n_s + \epsilon)\rho$, where $\epsilon \in [0,1)$. Equivalently, we need to show

$$\frac{\rho^{n_s+1} + n_s - (n_s+1)\rho}{(\rho-1)^2} > \frac{R\mu}{C}. \tag{EC.14}$$

We claim that inequality (EC.14) is equivalent to $n_s \neq n_0$. We show this claim as follows.

If inequality (EC.14) holds, we want to show that $n_s \neq n_0$. By definition, n_0 satisfies the two inequalities (18) and (19) on page 20, Naor (1969), which can be transformed into equivalent inequalities (20) and (21) on page 20. If $n_s = n_0$, then inequality (21) in Naor (1969) contradicts with inequality (EC.14) here. Hence, inequality (EC.14) implies that $n_s \neq n_0$.

On the other hand, if $n_s \neq n_0$, we want to show inequality (EC.14) is true. We know $n_s > n_0$, i.e., $n_s \in \{n_0+1, n_0+2, ...\}$. Naor (1969) shows that $W(0) = P(n)$ as a function of n is "discretely unimodal." Hence, we have $P(n_s - 1) > P(n_s)$. Simplification yields

$$\frac{\rho^{n_s+1} + n_s - (n_s+1)\rho}{(\rho-1)^2} > \frac{R\mu}{C},$$

which is precisely inequality (EC.14). Hence, $n_s \neq n_0$ implies inequality (EC.14).

Therefore, we have shown that inequality (EC.14) is equivalent to $n_s \neq n_0$ for $\rho \neq 1$.

For $\rho = 1$, we can compute the social welfare directly, $W(0) = \frac{n_s}{n_s+1}\lambda R - \frac{n_s}{2}C$. Then $n_s \neq n_0$ is equivalent to

$$\frac{n_s - 1}{n_s}\lambda R - \frac{n_s - 1}{2}C > \frac{n_s}{n_s+1}\lambda R - \frac{n_s}{2}C.$$

Simplifying the above, we obtain

$$\frac{n_s(n_s+1)}{2} > \frac{\lambda R}{C}.$$

Taking limits of the LHS of inequality (EC.14) when $\rho \to 1$, we know

$$\lim_{\rho \to 1} \frac{\rho^{n_s+1} + n_s - (n_s+1)\rho}{(\rho-1)^2} = \frac{n_s(n_s+1)}{2}$$

using the L'Hospital rule. Hence, our results hold when $\rho = 1$.

Furthermore, we are interested in sufficient conditions on primitives to ensure that $n_s \neq n_0$ holds. To show inequality (EC.14), it is *sufficient* to show that $\frac{T}{K} \geq (n_s+1)\rho$, which is equivalent to

$$\frac{\rho^{n_s+2} + n_s\rho - (n_s+1)\rho^2}{(\rho-1)^2} \geq (n_s+1)\rho$$

$\Leftrightarrow$

$$\frac{\rho^{n_s+1} - 1}{\rho - 1} \geq (n_s+1)\rho$$

when $\rho > 1$. But

$$\frac{\rho^{n_s+1} - 1}{\rho - 1} = \rho^{n_s} + \rho^{n_s-1} + ... + \rho + 1 > (n_s - 1)\rho + (\rho^2 + 1) > (n_s+1)\rho$$

as long as $n_s > 1$.

Now we prove the case when $\rho < 1$, it is equivalent to show

$$\frac{\rho^{n_s+1} + n_s - (n_s+1)\rho}{(\rho-1)^2} > \frac{R\mu}{C} = n_s + \epsilon.$$

Comparing this inequality with equation (21) on page 20 in Naor 1969, we know that it is necessary that $n_s \neq n_0$. And by the same argument as above, we know it is also sufficient that $n_s \neq n_0$.

Again, to find sufficient conditions on the primitives, we offer the following discussion. It is sufficient to show

$$\frac{\rho^{n_s+1}-1}{\rho-1} = \rho^{n_s} + \rho^{n_s-1} + ... + \rho + 1 \leq (n_s+1)\rho.$$

Clearly, if ρ is too small, for example, $\rho < \frac{1}{n_s+1}$, the inequality above cannot hold. Note that we assumed $n_s \geq 2$. We can use induction to show the inequality above. When $n_s = 2$, we are showing $\rho^2 + \rho + 1 \leq 3\rho$, i.e., $(\rho-1)^2 \leq 0$, which cannot be satisfied if $\rho < 1$. Hence, the social welfare decreases if $n_s = 2$ and $\rho < 1$. Now, we consider the case when $n_s = 3$, then the inequality simplifies to $(\rho-1)[\rho-(\sqrt{2}-1)](\rho+1+\sqrt{2}) \leq 0$. When $\sqrt{2}-1 \leq \rho < 1$, the inequality is satisfied. Then, we can use induction to show this inequality holds for any $n_s \geq 3$, if $\sqrt{2}-1 \leq \rho < 1$. This completes our proof. ■

The next lemma generalizes the above result.

LEMMA EC.6. *If $n_s \neq n_0$, and the customer who observes $n_s - 1$ customers already in the queue has any level of bounded rationality, i.e., he joins with probability $\sigma \in [0,1)$, then the social welfare would be strictly improved.*

Proof of Lemma EC.6. The proof is similar to that for Lemma EC.5. For this case, we have

$$P_0 = \frac{1-\rho}{1-(1-\sigma)\rho^{n_s} - \sigma\rho^{n_s+1}},$$

$$F = \lambda \frac{1-(1-\sigma)\rho^{n_s-1} - \sigma\rho^{n_s}}{1-(1-\sigma)\rho^{n_s} - \sigma\rho^{n_s+1}},$$

$$G = \frac{\rho}{1-\rho} - \frac{\sigma(n_s+1)\rho^{n_s+1} + (1-\sigma)n_s\rho^{n_s}}{1-(1-\sigma)\rho^{n_s} - \sigma\rho^{n_s+1}}.$$

After some algebra, we find that $\frac{T}{K}$ in this case is the same as equation (EC.13). Hence, we have showed the conclusion using the same argument as Lemma EC.5. ■

It is clear that, if the customers "on the negative side," i.e., the customers whose expected utility is strictly less than zero, have some degree of bounded rationality of joining the queue with some strictly positive probabilities, then the social welfare will deteriorate. We state this simple result formally.

LEMMA EC.7. *If the customer who observes n_s customers in front of him, joins the queue with some positive probability $\delta \in (0,1]$, then the social welfare will decrease for any $\rho \neq 1$.*

Proof of Lemma EC.7. The argument is intuitively simple: If such a customer joins the queue, his *net effect* on the social welfare is strictly negative, thus making the social welfare worse. This loose argument can be made rigorous using a similar technique as before as follows.

In this case, we have $\lambda_n = \lambda$, for $n = 0, 1, ..., n_s - 1$; $\lambda_{n_s} = \delta\lambda$, where $\delta \in (0,1]$; and $\lambda_n = 0$, for $n \geq n_s + 1$. Then,

$$a_n = \begin{cases} \rho^n \text{ if } n < n_s + 1 \\ \delta\rho^n \text{ if } n = n_s + 1 \\ 0 \text{ if } n > n_s + 1. \end{cases}$$

Long algebra gives us

$$F=\lambda\frac{1-(1-\delta)\rho^{n_s}-\delta\rho^{n_s+1}}{1-(1-\delta)\rho^{n_s+1}-\delta\rho^{n_s+2}},$$

$$G=\frac{\rho}{1-\rho}-\frac{(1-\delta)(n_s+1)\rho^{n_s+1}+\delta(n_s+2)\rho^{n_s+2}}{1-(1-\delta)\rho^{n_s+1}-\delta\rho^{n_s+2}}.$$

Then, we have

$$K=\frac{\delta\rho^{n_s}(\rho-1)^2}{(1-\rho^{n_s+1})(1-(1-\delta)\rho^{n_s+1}-\delta\rho^{n_s+2})}>0$$

and

$$T=\frac{\delta\rho^{n_s+1}[(n_s+1)-(n_s+2)\rho+\rho^{n_s+2}]}{(1-\rho^{n_s+1})(1-(1-\delta)\rho^{n_s+1}-\delta\rho^{n_s+2})}.$$

First, we consider the case when $\rho>1$. To show $W_2<W(0)$, it is sufficient to show $\frac{T}{K}\geq(n_s+1)\rho$. But, $\frac{T}{K}\geq(n_s+1)\rho\Longleftrightarrow(n_s+1)-(n_s+2)\rho+\rho^{n_s+2}\geq(n_s+1)(\rho^2-2\rho+1)$, which is in turn equivalent to $\frac{\rho^{n_s+1}-1}{\rho-1}=\rho^{n_s}+\rho^{n_s-1}+...+\rho+1\geq n_s+1$, which is true, since $\rho>1$.

Let us consider the case when $\rho<1$. To show $W_2<W(0)$, it is sufficient to show $\frac{\rho^{n_s+1}-1}{\rho-1}=\rho^{n_s}+\rho^{n_s-1}+...+\rho+1\leq n_s+1$, which is clearly true since $\rho<1$. ■

Now, we consider the case where both the customers "on the positive side" and "on the negative side" join with logit probabilities. We denote $W_2(\beta)$ as the social welfare when these customers join with logit probabilities.

LEMMA EC.8. *Let both of the customer who observes n_s-1 customers in front of him and the customer who observes n_s customers in front of him join the queue with logit probabilities specified in equation (1) while all other customers are fully rational, and the level of bounded rationality β be sufficiently small. If any one of the following three conditions is satisfied: (1) $n_s<\frac{R\mu}{C}-\frac{1}{2}$; (2) $n_s=n_0$; (3) $n_s=\frac{R\mu}{C}-\frac{1}{2}$ and $\rho>1$, then $W_2(\beta)<W(0)$. Otherwise, $W_2(\beta)>W(0)$.*

Proof of Lemma EC.8. We sketch the main steps here. We know, $\lambda_n=\lambda$, for $n=0,1,...,n_s-2$; $\lambda_{n_s-1}=\sigma\lambda$, where $\sigma\in[0,1)$; $\lambda_{n_s}=\delta\lambda$, where $\delta\in(0,1]$; and $\lambda_n=0$, for $n\geq n_s+1$.Then, we have

$$a_n=\begin{cases}\rho^n \text{ if } n<n_s\\ \sigma\rho^n \text{ if } n=n_s\\ \sigma\delta\rho^n \text{ if } n=n_s+1\\ 0 \text{ if } n>n_s+1.\end{cases}$$

Long algebra yields

$$F=\lambda\frac{1-(1-\sigma)\rho^{n_s-1}-\sigma(1-\delta)\rho^{n_s}-\sigma\delta\rho^{n_s+1}}{1-(1-\sigma)\rho^{n_s}-\sigma(1-\delta)\rho^{n_s+1}-\sigma\delta\rho^{n_s+2}},$$

$$G=\frac{\rho}{1-\rho}-\frac{(1-\sigma)n_s\rho^{n_s}+\sigma(1-\delta)(n_s+1)\rho^{n_s+1}+\sigma\delta(n_s+2)\rho^{n_s+2}}{1-(1-\sigma)\rho^{n_s}-\sigma(1-\delta)\rho^{n_s+1}-\sigma\delta\rho^{n_s+2}}.$$

Then, we have

$$K\equiv\frac{1-(1-\sigma)\rho^{n_s-1}-\sigma(1-\delta)\rho^{n_s}-\sigma\delta\rho^{n_s+1}}{1-(1-\sigma)\rho^{n_s}-\sigma(1-\delta)\rho^{n_s+1}-\sigma\delta\rho^{n_s+2}}-\frac{1-\rho^{n_s}}{1-\rho^{n_s+1}},$$

$$T\equiv\frac{(n_s+1)\rho^{n_s+1}}{1-\rho^{n_s+1}}-\frac{(1-\sigma)n_s\rho^{n_s}+\sigma(1-\delta)(n_s+1)\rho^{n_s+1}+\sigma\delta(n_s+2)\rho^{n_s+2}}{1-(1-\sigma)\rho^{n_s}-\sigma(1-\delta)\rho^{n_s+1}-\sigma\delta\rho^{n_s+2}}.$$

To determine the sign of $W_2(\beta)-W(0)=\lambda RK-CT$, we first need to know the sign of K. Some algebra tells us, when $\sigma=\sigma^*\equiv\frac{1}{1+\delta\rho}$, $K=0$; when $\sigma>\sigma^*$, $K>0$; when $\sigma<\sigma^*$, $K<0$. Then, we can discuss which cases are possible under different assumptions.

First, we assume that $n_s>\frac{R\mu}{C}-\frac{1}{2}$. Under the assumption that β is sufficiently small, if $n_s>\frac{R\mu}{C}-\frac{1}{2}$, i.e., $U_{n_s-1}-0<0-U_{n_s}$, where $U_{n_s-1}=R-\frac{n_sC}{\mu}$, and $U_{n_s}=R-\frac{(n_s+1)C}{\mu}$, using the logit probability specification in equation (1), we can rule out the cases when $\sigma\in[\sigma^*,1]$ by contradiction. Suppose it were possible that $\sigma\geq\sigma^*$, when β is small. After substituting the utility functions and simplifying, we have

$$\rho e^{\frac{U_{n_s-1}+U_{n_s}}{\beta}}\geq 1+\rho e^{\frac{U_{n_s}}{\beta}}. \tag{EC.15}$$

As β goes to zero, the LHS goes to zero while the RHS goes to 1, which is a contradiction. Then, we have $K<0$, to show $W_2(\beta)>W(0)$, it is equivalent to show $\frac{T}{K}>\frac{\lambda R}{C}$. After lengthy algebra, we have

$$\frac{T}{K}=\frac{\rho^{n_s+2}+n_s\rho-(n_s+1)\rho^2}{(\rho-1)^2}+\frac{\sigma\delta\rho^2(\rho^{n_s+1}-1)}{(\sigma\delta\rho+\sigma-1)(\rho-1)}. \tag{EC.16}$$

For convenience, denote $g(\beta)\equiv\frac{\sigma\delta\rho^2(\rho^{n_s+1}-1)}{(\sigma\delta\rho+\sigma-1)(\rho-1)}$, where σ and δ are the logit probabilities as functions of β, and we have abused notations by omitting β. Since $\sigma<\sigma^*$ when β is small, we know $\sigma\delta\rho+\sigma-1<0$ when β is small. Now, we claim that

$$\lim_{\beta\to 0}g(\beta)=0.$$

To show this claim, it is sufficient to show

$$\lim_{\beta\to 0}\frac{\sigma-1}{\sigma\delta}=-\infty.$$

Indeed, we have

$$\frac{\sigma-1}{\sigma\delta}=-\frac{1+e^{\frac{U_{n_s-1}}{\beta}}+e^{\frac{U_{n_s}}{\beta}}+e^{\frac{U_{n_s-1}+U_{n_s}}{\beta}}}{e^{\frac{U_{n_s-1}+U_{n_s}}{\beta}}+e^{\frac{2U_{n_s-1}+U_{n_s}}{\beta}}}.$$

Multiplying the RHS of this equation by $e^{\frac{-U_{n_s-1}}{\beta}}$ for both of the numerator and the denominator, and taking limit, we know it goes to $-\infty$, since the numerator goes to 1 and the denominator goes to 0 as β goes to 0.

Comparing inequality (EC.14) and (EC.16), we know when β is sufficiently small, inequality (EC.16) holds if inequality (EC.14) holds which is equivalent to $n_s\neq n_0$. Hence, under the assumption that $n_s>\frac{R\mu}{C}-\frac{1}{2}$, if and only if $n_s\neq n_0$, $W_2(\beta)>W(0)$, when β is sufficiently small.

If $n_s<\frac{R\mu}{C}-\frac{1}{2}$, we can similarly show that the social welfare will decrease as follows. We have $U_{n_s-1}+U_{n_s}>0$, which implies that $\sigma>\sigma^*$, which further implies that $K>0$. We claim that $W_2(\beta)<W(0)$, which is equivalent to $\frac{T}{K}>\frac{\lambda R}{C}$. Studying equation (EC.16) again, now we have $g(\beta)>0$ since $\sigma>\sigma^*$, and we know

$$\lim_{\beta\to 0}\frac{\sigma-1}{\sigma\delta}=0,$$

which further implies that

$$\lim_{\beta\to 0}g(\beta)=\frac{\rho(\rho^{n_s+1}-1)}{\rho-1}.$$

By Lemma EC.5, one can show that this inequality holds as long as $\rho\neq 1$.

The last case is when $n_s = \frac{R\mu}{C} - \frac{1}{2}$, which implies that $U_{n_s-1} + U_{n_s} = 0$. Then the inequality $\sigma \geq \sigma^*$ is equivalent to

$$\rho \geq 1 + \rho e^{\frac{U_{n_s}}{\beta}}.$$

If $\rho < 1$, then this inequality cannot hold. Therefore, it has to be the case that $\sigma < \sigma^*$ when β is sufficiently small, which implies that $K < 0$. We claim that $W_2(\beta) > W(0)$, which is equivalent to $\frac{T}{K} > \frac{\lambda R}{C}$, which holds if and only if $n_s \neq n_0$.

If $\rho > 1$, when β is sufficiently small, we have $\sigma > \sigma^*$, which implies that $K > 0$. We claim that $W_2(\beta) < W(0)$, which is equivalent to $\frac{T}{K} > \frac{\lambda R}{C}$, which holds by Lemma EC.7. Hence, we have completed the proof. ■

LEMMA EC.9. *Function* $W_0(x) \equiv \lambda \frac{1-\rho^x}{1-\rho^{x+1}} R - [\frac{\rho}{1-\rho} - \frac{\rho \frac{d\rho^{x+1}}{d\rho}}{1-\rho^{x+1}}]C = \lambda \frac{1-\rho^x}{1-\rho^{x+1}} R - [\frac{\rho}{1-\rho} - \frac{(x+1)\rho^{x+1}}{1-\rho^{x+1}}]C$, $x \in [0, +\infty)$, *is unimodal.*

Proof of Lemma EC.9. Taking the first-order derivative and simplifying, we have

$$W_0'(x) = \frac{1}{(1-\rho^{x+1})^2}[\lambda R(\rho-1)\rho^x \log\rho + C\rho^{x+1}(1-\rho^{x+1} + (x+1)\log\rho)].$$

We want to show that the equation $W'(x) = 0$ has at most one solution, which implies that $W_0(x)$ is unimodal. Indeed, $W_0'(x) = 0$ is equivalent to

$$\rho^{x+1} = (x+1)\log\rho + \frac{\lambda R(\rho-1)\log\rho}{C\rho} + 1,$$

which clearly has at most one solution. ■

LEMMA EC.10. *Assume* $n_s \neq n_0$. *Let* $\rho^k \equiv f(\rho) = (1-\sigma)\rho^{n_s} + \sigma(1-\delta)\rho^{n_s+1} + \sigma\delta\rho^{n_s+2}$, $V \equiv \rho f'(\rho) = (1-\sigma)n_s\rho^{n_s} + \sigma(1-\delta)(n_s+1)\rho^{n_s+1} + \sigma\delta(n_s+2)\rho^{n_s+2}$. *If* $k \in (n_s, n_s+1)$, *then* $W_0(k-1) \equiv \lambda \frac{1-\rho^{k-1}}{1-\rho^k} R - [\frac{\rho}{1-\rho} - \frac{V}{1-\rho^k}]C > W_0(n_s) \equiv \lambda \frac{1-\rho^{n_s}}{1-\rho^{n_s+1}} R - [\frac{\rho}{1-\rho} - \frac{(n_s+1)\rho^{n_s+1}}{1-\rho^{n_s+1}}]C$. *In general, the property that "less congestion" implies more welfare holds for any number of customers joining using logit probabilities.*

Proof of Lemma EC.10. This lemma follows directly from Lemma EC.9. ■

Similarly, we have the following lemma.

LEMMA EC.11. *Assume* $n_s \neq n_0$. *Let* $\rho^k \equiv f(\rho) = (1-\sigma)\rho^{n_s} + \sigma(1-\delta)\rho^{n_s+1} + \sigma\delta\rho^{n_s+2}$, $V \equiv \rho f'(\rho) = (1-\sigma)n_s\rho^{n_s} + \sigma(1-\delta)(n_s+1)\rho^{n_s+1} + \sigma\delta(n_s+2)\rho^{n_s+2}$. *If* $k \in (n_s+1, n_s+2)$, *then* $W_0(k-1) \equiv \lambda \frac{1-\rho^{k-1}}{1-\rho^k} R - [\frac{\rho}{1-\rho} - \frac{V}{1-\rho^k}]C < W_0(n_s) \equiv \lambda \frac{1-\rho^{n_s}}{1-\rho^{n_s+1}} R - [\frac{\rho}{1-\rho} - \frac{(n_s+1)\rho^{n_s+1}}{1-\rho^{n_s+1}}]C$. *In general, the property that "more congestion" implies less welfare holds for any number of customers joining using logit probabilities.*

Proof of Lemma EC.11. This lemma follows directly from Lemma EC.9. ■

LEMMA EC.12. *If the price* $p^* \in (R - \frac{C(n_0+1)}{\mu}, R - \frac{Cn_0}{\mu}]$ *is charged to the customers, then the social welfare* $W(p^*, \beta)$ *is lower than the social optimum, i.e.,* $W(p^*, \beta) < W^*(0)$ *for* $\beta > 0$.

Proof of Lemma EC.12. For convenience, we write W_{n_0} and $W^*(0)$ interchangeably for the first-best social welfare. We discuss two cases when the optimal prices when full rationality is assumed by the social planner are charged. The first case is that $n_0+\epsilon_2$ is the global maxima of the function $W_0(x)$ among the continuous interval $[1,n_s]$ (and n_0 is the global maximum among the discrete candidates $\{1,2,...,n_s\}$ as assumed throughout). Let $\delta(\beta)\in(0,1)$ be the probability the customer who sees n_0 will join the queue (all others are fully rational). To show $W_{n_0}(p^*,\beta)<W_{n_0}$, where $W_{n_0}(p^*,\beta)$ is the social welfare when only the customer who sees n_0 customers in front of him is boundedly rational, using the result of equation (EC.14) in Lemma EC.5 (but using n_0 instead of n_s), we know, it is equivalent to show

$$\frac{T}{K}>\frac{\lambda R}{C}=\rho v_s.$$

But

$$\frac{T}{K}=\frac{\rho[(n_0+1)-(n_0+2)\rho+\rho^{n_0+2}]}{(\rho-1)^2}.$$

Hence, we want to show

$$\frac{(n_0+1)-(n_0+2)\rho+\rho^{n_0+2}}{(\rho-1)^2}>v_s=\frac{R\mu}{C},$$

which is precisely the RHS of inequality (22), page 20, Naor (1969). To show $W(p^*,\beta)<W_{n_0}$, simply note that $W_{n_0}(p^*,\beta)\approx W(p^*,\beta)$ when β is sufficiently small by similar arguments in Proposition 8. Finally, it is clear that, if only the customer who sees n_0-1 customers randomize (all others are fully rational), then the social welfare will decrease.

The second case is that $n_0-\epsilon_1$ is the global maxima among the continuous interval $[1,n_s]$. Let $\sigma(\beta)\in(0,1)$ be the probability the customer who sees n_0-1 will join the queue. Similar to the first case, to show $W(p^*,\beta)<W_{n_0}$, using the results in Lemma EC.8 and EC.11, we know, it is equivalent to show

$$\frac{n_0-(n_0+1)\rho+\rho^{n_0+1}}{(\rho-1)^2}<\frac{R\mu}{C},$$

which is precisely the LHS of inequality (22), page 20, Naor (1969), if $\epsilon_1\neq 0$ (the interesting case).

So far we have shown that for small levels of bounded rationality, the result holds, i.e., $W(p^*,\beta)<W_{n_0}$ for $\beta\in(0,\bar{\beta}_{p^*})$ for some $\bar{\beta}_{p^*}>0$.

Now we show the result when β is large in which case the argument above does not apply. However, we know that W_{n_0} is the optimal social welfare by Yechiali (1971). However, we cannot rule out the case that $W(p,\beta)=W_{n_0}$ for some p from Yechiali (1971)'s results. To rule out the case, we use Haviv and Puterman (1998), who show that the only average optimal stationary policies are of control limit type, that there are at most two and, if there are two, they occur consecutively. This implies that the only gain optimal randomized stationary policies should randomize over the two control limit states if they exist. The argument is simple: For any randomized policy to be optimal, the deterministic policies it has strictly positive probabilities should yield the same average reward. In our setting with randomization using logit probabilities, their result implies that W_{n_0} is strictly larger than any $W(p,\beta)$ when $\beta>0$ since the logit joining probabilities are in the interval (0,1). ■

LEMMA EC.13. *For the revenue-maximizing price $p^* = R - \frac{Cn_r}{\mu}$ charged to the customers, the maximum revenue when customers are fully rational cannot be achieved when there is a little bit of bounded rationality among the customers in general, for $\beta \in (0, \bar{\beta})$ for some $\bar{\beta} > 0$.*

Proof of Lemma EC.13. When the optimal price is charged, under full-rationality assumption, the customer who observes $n_r - 1$ customers in front of him will join the queue yet with zero utility. However, when there is a little bit of bounded rationality, he will join with probability $1/2$. Such change will make the system less congested compared to the fully-rational case. Recall the optimal revenue under full rationality is

$$\Pi_{n_r} = \lambda \frac{1-\rho^{n_r}}{1-\rho^{n_r+1}}(R - \frac{Cn_r}{\mu}).$$

Note that the function $f(x) \equiv \frac{1-\rho^x}{1-\rho^{x+1}}$ is strictly increasing in x when $\rho \neq 1$, so less congestion implies less revenue. We have $\Pi(p^*, \beta) < \Pi_{n_r}$, for $\beta \in (0, \bar{\beta})$ for some $\bar{\beta} > 0$. ■

Appendix C: Generalization of Proposition 7

In this section, we provide a generalization for Proposition 7 where the price is assumed to be zero. For any price p, we have a result similar to Proposition 7. To state this result, we define

$$\varphi(p, n) = \frac{e^{\frac{R-p-\frac{(n+1)C}{\mu}}{\beta}}}{1 + e^{\frac{R-p-\frac{(n+1)C}{\mu}}{\beta}}},$$

a customer's probability of joining the system when price p is charged and there are already n customers in the system ahead of him. Clearly, we have $\varphi(0, n) = \varphi_n$. We define $n(p) = [\frac{(R-p)\mu}{C}]$, then we have $n(0) = n_s$. When customers are fully rational, i.e., $\beta = 0$, we have the social welfare function

$$W(p, 0) = \lambda \frac{1-\rho^{n(p)}}{1-\rho^{n(p)+1}} R - [\frac{\rho}{1-\rho} - \frac{(n(p)+1)\rho^{n(p)+1}}{1-\rho^{n(p)+1}}]C.$$

We are interested in comparing $W(p, 0)$ and $W(p, \beta)$ when β is small. As before, we focus on the interesting case when $n_s \neq n_0$. It is clear that we have to compare $n(p)$ with n_0 noting that $n(p) \leq n_s$.

PROPOSITION EC.1. *If $n(p) > n_0$, we have the following result: If any one of the following two conditions is satisfied: (1) $n(p) < \frac{(R-p)\mu}{C} - \frac{1}{2}$; (2) $n(p) = \frac{(R-p)\mu}{C} - \frac{1}{2}$ and $\rho > 1$, then $W(p, \beta) < W(p, 0)$ when $\beta > 0$ is sufficiently small. Otherwise, $W(p, \beta) > W(p, 0)$ when $\beta > 0$ is sufficiently small.*
If $n(p) < n_0$, we have the following result: If any one of the following two conditions is satisfied: (1) $n(p) < \frac{(R-p)\mu}{C} - \frac{1}{2}$; (2) $n(p) = \frac{(R-p)\mu}{C} - \frac{1}{2}$ and $\rho > 1$, then $W(p, \beta) > W(p, 0)$ when $\beta > 0$ is sufficiently small. Otherwise, $W(p, \beta) < W(p, 0)$ when $\beta > 0$ is sufficiently small.

The proof of this result is similar to the proof of Proposition 7, and we omit it for brevity. If $n(p) = n_0$, then p has to be one of the optimal prices $p^* \in (R - \frac{C(n_0+1)}{\mu}, R - \frac{Cn_0}{\mu}]$ and the analysis is in Section 4.1.

Appendix D: Discussion on Neglecting Other Customers' Bounded Rationality

In this section, we relax Definition 1 in the paper. Here, we assume that each customer simply treat the system as an $M/M/1$ queue with arrival rate λ. The explanation is that each customer is not informed that others are actually boundedly rational, although he himself is boundedly rational in making his decisions according to the logit probability formula. The resulting queue would be an $M/M/1$ queue with arrival rate

$$\lambda_b \equiv \lambda \varphi_b(p, \beta_I),$$

where

$$\varphi_b(p, \beta_I) = \frac{e^{\frac{R-p-\frac{C}{\mu-\lambda}}{\beta_I}}}{1+e^{\frac{R-p-\frac{C}{\mu-\lambda}}{\beta_I}}}.$$

If $\lambda_b \geq \mu$, then the system is unstable. From now on, we are interested in the case when $\lambda_b < \mu$, i.e., $\varphi_b(p, \beta_I) < \frac{\mu}{\lambda}$ so that the system is stable. One can characterize the behavior of the joining fraction $\varphi_b(p, \beta_I)$, in a similar albeit much simpler way as in Section 2.2. We omit it for brevity. Note that, when $p = p_b \equiv R - \frac{C}{\mu-\lambda}$, we have $\varphi_b(p, \beta_I) = 0.5$ so that the joining fraction is independent of the level of bounded rationality. One observation is that $p_b < p_0$. Denote the revenue function $\Pi_b(p, \beta_I) \equiv p\lambda\varphi_b(p, \beta_I)$. One can show that $\Pi_b(p, \beta_I) < \Pi^I(p, \beta_I)$ since the joining fraction is *strictly lower* when the fact that other customers are boundedly rational is neglected by each customer.

For the social welfare function

$$W_b(p, \beta_I) = \varphi_b(p, \beta_I)\lambda R - \frac{\varphi_b(p, \beta_I)\lambda}{\mu - \varphi_b(p, \beta_I)\lambda} C,$$

one can analyze similarly as in Section 4.2. It turns out the welfare-maximizing price is

$$p^*_{bw}(\beta_I) = R - \frac{C}{\mu-\lambda} - \beta_I \ln \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\lambda - \mu + \sqrt{\frac{C\mu}{R}}} = p^*_b(0) - \beta_I \ln \frac{\varphi^*_w(p, \beta_I)}{1 - \varphi^*_w(p, \beta_I)}$$

and we have $p^*_{bw}(\beta_I) < p^*_w(\beta_I)$ if $\mu < \lambda$, and $R \in (\frac{C}{\mu}, \frac{C\mu}{(\mu-\lambda)^2})$ if $\mu > \lambda$. Hence, the level of bounded rationality above which there is a welfare loss is

$$\beta_{bw}(0) = \frac{R - \frac{C}{\mu-\lambda}}{\ln \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\lambda - \mu + \sqrt{\frac{C\mu}{R}}}} < \beta_w(0),$$

if $\varphi^*_w > \frac{1}{2}$. This implies that the social welfare loss happens in a wider range of the level of bounded rationality.

In the analysis above, each customer believes others always join the system with probability one. What if each customer believes others join the system with probability 0.5, i.e., others are fully boundedly rational? This is related to the "level-k thinking" (Stahl and Wilson 1995) since each customer treats the other customers as *level-0* customers (who simply randomize), and each customer himself behaves as a *level-1* customer, yet without the perfect capability of estimating his expected waiting time. The resulting queue would be an $M/M/1$ queue with arrival rate $\lambda_B \equiv \lambda\varphi_B(p, \beta_I)$, where

$$\varphi_B(p, \beta_I) = \frac{e^{\frac{R-p-\frac{C}{\mu-\frac{1}{2}\lambda}}{\beta_I}}}{1+e^{\frac{R-p-\frac{C}{\mu-\frac{1}{2}\lambda}}{\beta_I}}}.$$

Similar as before,we focus on the setting where the system is stable. We have $p_B \equiv R - \frac{2C}{2\mu-\lambda} = p_0$, at which price the equilibrium joining fraction is independent of the level of bounded rationality. The revenue function $\Pi_B(p,\beta_I) \equiv p\lambda\varphi_B(p,\beta_I)$. We want to compare $\Pi_B(p,\beta_I)$ with $\Pi^I(p,\beta_I)$. We discuss three cases: (1) If $p > p_0$, then $\varphi(p,\beta_I) < 0.5$ and $\varphi_B(p,\beta_I) < \varphi(p,\beta_I)$. Hence, $\Pi_B(p,\beta_I) < \Pi^I(p,\beta_I)$. (2) If $p < p_0$, then $\varphi(p,\beta_I) > 0.5$ and $\varphi_B(p,\beta_I) > \varphi(p,\beta_I)$. Hence, $\Pi_B(p,\beta_I) > \Pi^I(p,\beta_I)$. (3) If $p = p_0$, then $\Pi_B(p,\beta_I) = \Pi^I(p,\beta_I)$.

For the social welfare function

$$W_B(p,\beta_I) \equiv \varphi_B(p,\beta_I)\lambda R - \frac{\varphi_B(p,\beta_I)\lambda}{\mu - \varphi_B(p,\beta_I)\lambda}C,$$

we can analyze similarly as before. The welfare-maximizing price is

$$p^*_{Bw}(\beta_I) = R - \frac{2C}{2\mu-\lambda} - \beta_I \ln\frac{\mu - \sqrt{\frac{C\mu}{R}}}{\lambda - \mu + \sqrt{\frac{C\mu}{R}}} = p^*_B(0) - \beta_I \ln\frac{\varphi^*_w(p,\beta_I)}{1-\varphi^*_w(p,\beta_I)}$$

and we want to compare $p^*_{Bw}(\beta_I)$ with $p^*_w(\beta_I)$. It turns out either one can be larger than the other, depending on whether $R < \frac{C\mu}{(\mu-\frac{1}{2}\lambda)^2}$. If $R < \frac{C\mu}{(\mu-\frac{1}{2}\lambda)^2}$ holds, then $p^*_{Bw}(\beta_I) < p^*_w(\beta_I)$. If $R > \frac{C\mu}{(\mu-\frac{1}{2}\lambda)^2}$, then $p^*_{Bw}(\beta_I) > p^*_w(\beta_I)$. If $R = \frac{C\mu}{(\mu-\frac{1}{2}\lambda)^2}$, then $p^*_{Bw}(\beta_I) = p^*_w(\beta_I)$.

In sum, in the case when each customer treats others as level-0 customers, the optimal revenue and social welfare can be higher or lower compared to the correct-belief case analyzed in the main paper.

Finally, we discuss why the correct-belief case presented in the paper is more appropriate than the scenarios analyzed here. Note that we are studying the queueing systems in *steady state*, i.e., the long-run behavior. This implies that customers have opportunities to repeatedly learn the system state. Hence, in the short-run or transient states, customers may neglect others' bounded rationality, but eventually, they would be able to form the correct belief about others' strategies so that the fixed-point outcome according to Definition 1 would emerge. For general discussions, we refer readers to McKelvey and Palfrey (1995), Chen et al. (1997) and references therein. In the main paper, we focus on bounded rationality as incapability of accurately predicting the expected waiting time, which also allows a fair comparison of the visible queue versus invisible queue.

Appendix E: Additional Numerical Examples

We conduct a numerical study to investigate utilization $\rho(p,\beta) \equiv \sum_{n=0}^{\infty}\lambda_n P_n/\mu$ and expected queue length $q(p,\beta) \equiv \sum_{n=0}^{\infty} nP_n$ as functions of the level of bounded rationality β.

We use a "high utilization" scenario and a "low utilization" scenario, crossed with a "high revenue" and a "low revenue" scenario, while holding waiting cost C constant. Figure EC.1-EC.4 present several examples, which suggest that neither utilization nor expected queue length are necessarily well behaved. In particular, neither of them is monotonic or unimodal as a function of the level of bounded rationality β.

To understand why this happens, we divide the states of the system into two regions using the threshold n_s: Region 1 comprises of the states when the number of customers in the system is less than n_s, Region 2 comprises of all the other states, i.e., those when the number of customers in the system is greater than n_s. When customers are fully rational, customers will join with probability 1 in Region 1 and and 0 in Region 2.

For every strictly positive level of bounded rationality, the joining probability will be between 0 and 1. Hence, bounded rationality in Region 1 lowers utilization (and expected queue length) while increases utilization (and expected queue length) in Region 2. As customers become more boundedly rational, these two effects take place simultaneously. It turns out that it is not clear which effect dominates. For example, in Figure EC.1, utilization is decreasing in the level of bounded rationality β for a low arrival rate λ while increasing for a high arrival rate. Hence, when the arrival rate is low, the effect from Region 1 dominates. Whereas, the effect from Region 2 dominates when the arrival rate is high. Figure EC.2 shows that utilization can increase or decrease for different reward R. The intuition is that, for the low reward scenario $R=6$, the threshold n_s is small, whereas for the high reward scenario $R=60$, the threshold n_s is big. Hence, Region 1 is small in the former while big in the latter. This implies that the effect of bounded rationality from Region 1 tends to be small in the former while big in the latter. Therefore, it turns out that utilization increases in the former scenario since the effect from Region 2 dominates, and decreases in the latter scenario since the effect from Region 1 dominates. Figure EC.3 and EC.4 present similar patterns for expected queue length. The intuition will be similar, so we omit any discussion for brevity.

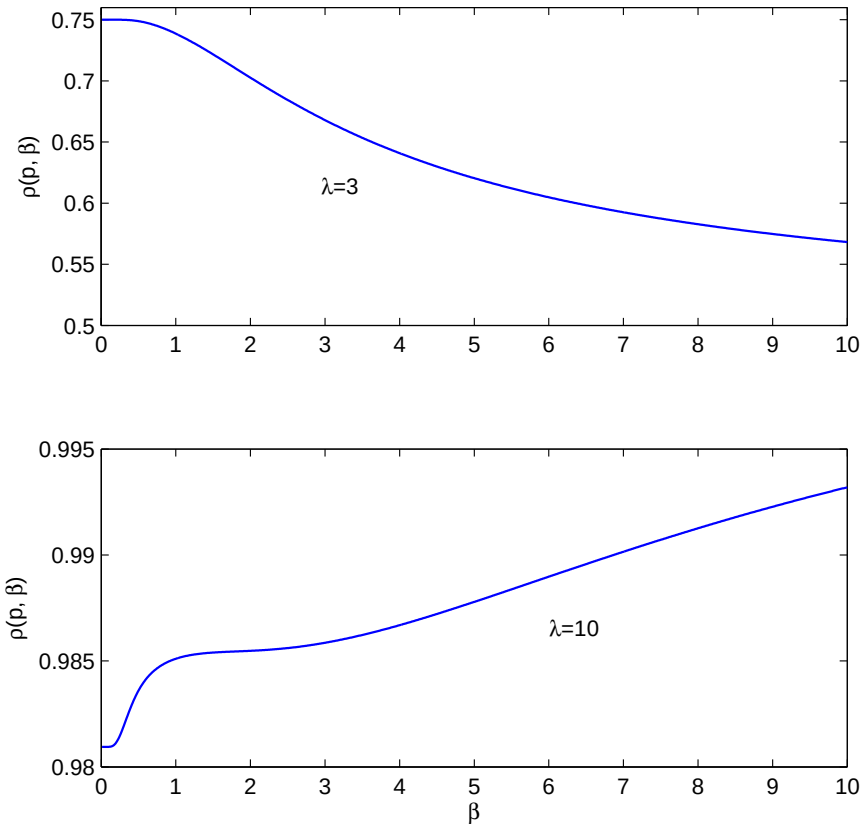
Although utilization or expected queue length is unimodal within the reasonable range of bounded rationality $[0,10]$ for some of the examples presented here. It worth noting that this is not generally true since in our extensive numerical studies, we did observe that utilization or expected queue length is not unimodal in the level of bounded rationality as latter is outside this region.

Appendix F: Global Stability of the Equilibrium for Invisible Queue

In this section, we relax the assumption in §2.2 that each customer knows the arrival rate and the service rate. We shall show, under certain settings, how Definition 1 can emerge from customer behavior in an adaptive manner. We will next describe a model in which customers do not know the arrival rate and the service rate, and take their actions based on their past experience.

We will index time period as $t \in T \equiv \{0,1,2,3,...\}$. In period 0, each customer has no information about the service system (e.g., in terms of the expected waiting time, arrival rate, and service rate), and joins the system with probability 0.5 (or any arbitrary probability). For each period $t \in T$, we assume that the period is sufficiently long so that the system reaches its steady-state. We also assume that, within each period, using the same strategy, each customer interacts with the firm repeatedly. Let $\mathbb{E}W_t$ denote the *actual* expected waiting time in period t. Therefore, $\mathbb{E}W_t = \frac{1}{(\mu-\lambda\varphi_t)^+}$, where φ_t is the fraction of the customers that join the system. However, each customer is boundedly rational in the sense that he does not have the perfect capability to compute $\mathbb{E}W_t$. Each customer thus obtains his expected waiting time estimate $\widehat{\mathbb{E}W}_t \equiv \mathbb{E}W_t + \varepsilon_t$, where ε_t is assumed to follow the logistic distribution with parameter θ as defined in §2 of the main paper.

Figure EC.1 Utilization for different λ ($R=19, p=13, C=6, \mu=3$)



In each period $t=1,2,...$, each customer decides whether to arrive to the system or not based on his noisy estimate of the expected waiting time $\widehat{\mathbb{E}W}_{t-1}$ in the period $t-1$. Following the same argument in §2, the fraction of customers that join the system is

$$\varphi_t = \frac{e^{\frac{R-p-C\mathbb{E}W_{t-1}}{\beta_I}}}{1+e^{\frac{R-p-C\mathbb{E}W_{t-1}}{\beta_I}}}$$

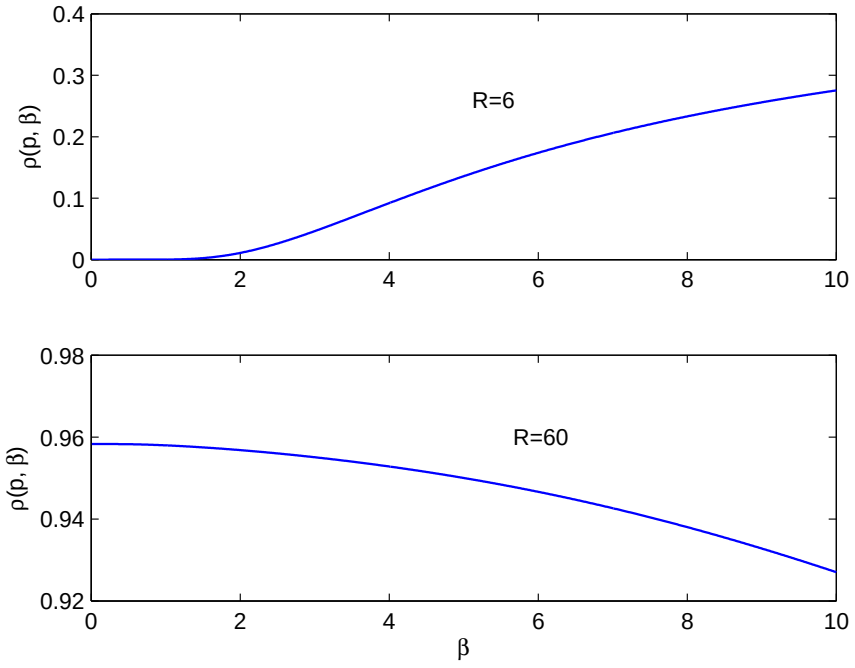
which can be interpreted as the joining probability for each customer in period $t=1,2,....$

In the proposition below, we shall prove that, under certain conditions, customer behavior will converge to the unique equilibrium in Definition 1. To state this proposition, we substitute the actual expected waiting time $\mathbb{E}W_t = \frac{1}{(\mu-\lambda\varphi_t)^+}$ and define the mapping $\widehat{\psi}$ from $[0,1]$ to $[0,1]$ so that $\varphi_t = \widehat{\psi}(\varphi_{t-1})$. We then focus on the iterative equation:

$$\varphi_t = \widehat{\psi}(\varphi_{t-1}) = \frac{e^{\frac{R-p-\frac{C}{(\mu-\lambda\varphi_{t-1})^+}}{\beta_I}}}{1+e^{\frac{R-p-\frac{C}{(\mu-\lambda\varphi_{t-1})^+}}{\beta_I}}}.$$

PROPOSITION EC.2. *Suppose $\mu > \lambda\widehat{\psi}(0)$. If $\beta_I \geq \beta_C \equiv \frac{\lambda C}{(\mu-\lambda)^2}$, customer behavior from adaptive learning converges to the unique equilibrium in Definition 1.*

Figure EC.2 **Utilization for different** R **(**$p=13, C=6, \lambda=3, \mu=3$**)**



Proof. If $\beta_I = \infty$, then the fraction of the customers that arrive to the system is 0.5 independent of customer learning. It trivially converges to the equilibrium 0.5 of Definition 1. Hereafter, we focus on the case when $\beta_I < \infty$.

There are two cases to consider.

Case 1: we suppose that in period 0, $\mu \leq \lambda\varphi_0$ so that the system is unstable. A customer will not arrive to the system in period 1 since the estimated expected waiting time is infinity when β_I is finite. In period 2, each customer frequents/interacts with the firm with an extremely small frequency to learn the expected waiting time, so that each customer arrives to the system with probability $\varphi_2 = \widehat{\psi}(0)$. If $\mu - \lambda\widehat{\psi}(0) > 0$, then the system is stable in period 2. In period 3, each customer arrives to the system with probability $\varphi_3 = \widehat{\psi}(\varphi_2) = \widehat{\psi}(\widehat{\psi}(0)) < \widehat{\psi}(0)$ since $\widehat{\psi}(.)$ is strictly decreasing by the definition of $\widehat{\psi}(.)$. Hence, we have $\mu - \lambda\varphi_t > 0$ for $t = 2, 3, 4, ...$, which suggests that we can focus on the mapping ψ:

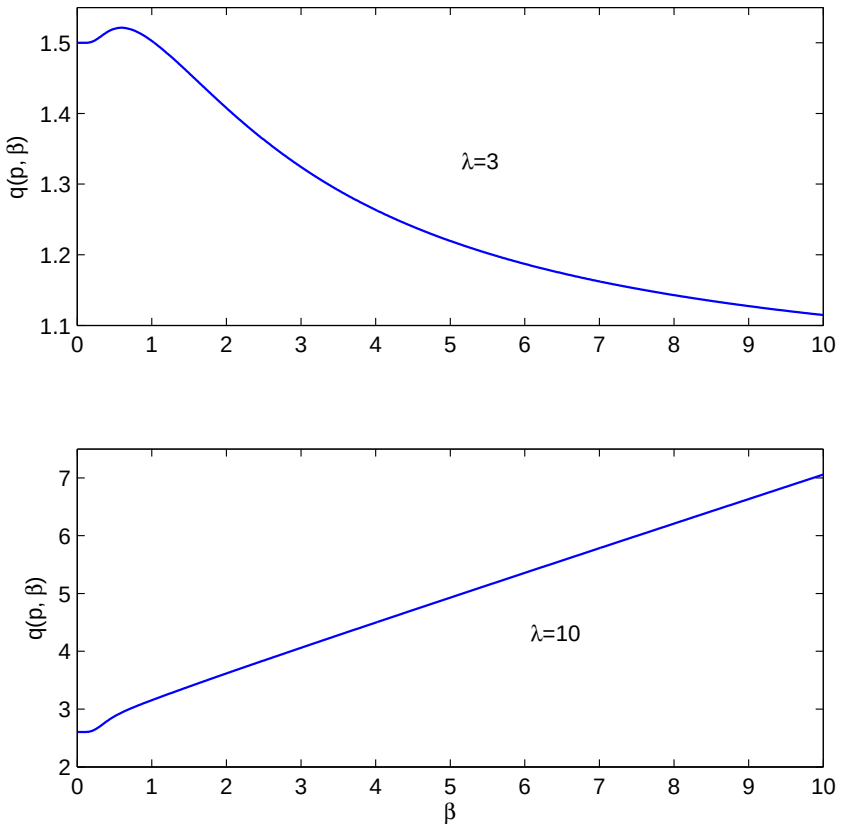
$$\varphi_t = \psi(\varphi_{t-1}) = \frac{e^{\frac{R-p-\frac{C}{\mu-\lambda\varphi_{t-1}}}{\beta_I}}}{1+e^{\frac{R-p-\frac{C}{\mu-\lambda\varphi_{t-1}}}{\beta_I}}}$$

for $t = 2, 3, 4, ...$ In lemma EC.14 below, we prove that ψ is contraction mapping for $\beta_I \geq \beta_C$. Hence, let $t \to \infty$, then the unique equilibrium $\lim_{t\to\infty} \varphi_t = \widehat{\varphi}$ in Definition 1 will emerge.

Case 2: we suppose in period $t_0 = 1, 2, ...$, $\mu \leq \lambda\varphi_{t_0}$ so that the system is unstable, then one can re-label the time period so that $t_0 = 0$ and apply the same argument above. □

LEMMA EC.14. *The function*

Figure EC.3 Expected queue length for different λ ($R=19, p=13, C=6, \mu=3$)



$$\psi(\varphi_{t-1}) = \frac{e^{\frac{R-p-\frac{C}{\mu-\lambda\varphi_{t-1}}}{\beta_I}}}{1+e^{\frac{R-p-\frac{C}{\mu-\lambda\varphi_{t-1}}}{\beta_I}}}$$

is a contraction mapping for $\beta_I \geq \beta_C$.

Proof. Denote $u_1(\varphi_t) = R-p-\frac{C}{\mu-\lambda\varphi_t}$. Then it is easy to show that $u_1'(\varphi_t) = \frac{\lambda C}{(\mu-\lambda\varphi_t)^2} < \frac{\lambda C}{(\mu-\lambda)^2}$ since $\varphi_t < 1$. Note that ψ is continuously differentiable and

$$\psi'(\varphi_t) = \frac{e^{u_1(\varphi_t)/\beta_I} u_1'(\varphi_t)}{\beta_I(1+e^{u_1(\varphi_t)/\beta_I})^2} < \frac{\lambda C}{\beta_I(\mu-\lambda)^2}.$$

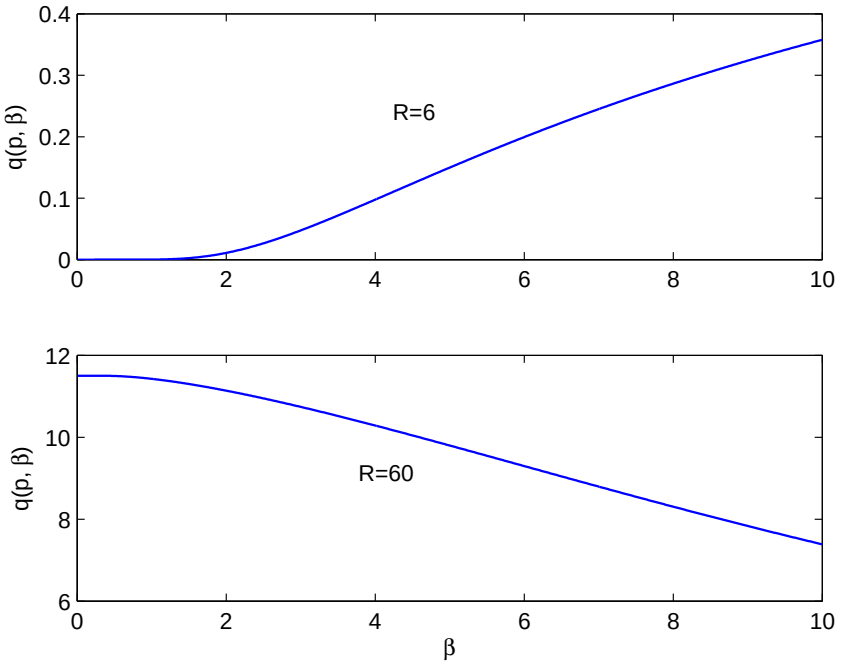
Hence, for $\beta_I \geq \beta_C = \frac{\lambda C}{(\mu-\lambda)^2}$, we have

$$\psi'(\varphi_t) < 1.$$

Therefore, there exists $\theta \in [0,1)$ such that, for any $x, y \in [0,1]$, $x < y$, we have $|\psi(x)-\psi(y)| = |\psi'(\xi)(x-y)| \leq \theta|x-y|$, where $\xi \in [x,y]$. Hence, ψ is a contraction mapping for $\beta_I \geq \beta_C$. □

To demonstrate the convergence, we provide a numerical example in Figure EC.5. We use the following parameters: $C=1, \mu=1, \lambda=0.5, R=2, p=0.6, \beta=0.3$. These parameters are also used in Figure 1, §3.1. In this figure, the horizontal line denotes the equilibrium in Definition 1 which is 0.53 in this case. The

Figure EC.4 **Expected queue length for different R ($p = 13, C = 6, \lambda = 3, \mu = 3$)**



dots illustrate the convergence path as a function of the time period t from the starting joining probability $\varphi_0 = 0.5$, while the circles for a different starting point $\varphi_0 = 0.7$. We can see that the customer behavior quickly converges to the equilibrium for $\beta = 0.3$. Note that $\beta_C = 2$ for this numerical example.

Figure EC.5 Convergence to the Equilibrium ($C=1, \mu=1, \lambda=0.5, R=2, p=0.6, \beta=0.3$)