ISSN 1866-3494

October 1, 2020

ESMT Working Paper 18-04 (R1)

Queueing systems with rationally inattentive customers

Caner Canyakmaz, ESMT Berlin

Tamer Boyaci, ESMT Berlin

Revised version

Queueing Systems with Rationally Inattentive Customers

Caner Canyakmaz
ESMT Berlin, caner.canyakmaz@esmt.org

Tamer Boyacı
ESMT Berlin, tamer.boyaci@esmt.org

Problem description: Classical models of queueing systems with rational and strategic customers assume queues to be either fully visible or invisible while service parameters are known with certainty. In practice, however, people only have "partial information" on the service environment in the sense that they are not able to fully discern prevalent uncertainties. This is because assessing possible delays and rewards is costly as it requires time, attention, and cognitive capacity which are all limited. On the other hand, people are also adaptive and endogenously respond to information frictions. *Methodology:* We develop an equilibrium model for a single-server queueing system with customers having limited attention. Following the theory of rational inattention, we assume that customers optimize their learning strategies by deciding the type and amount of information to acquire and act accordingly while internalizing the associated costs. *Results:* We establish the existence and uniqueness of a customer equilibrium and delineate the impact of service characteristics and information costs. We numerically show that when customers allocate their attention to learn uncertain queue length, limited attention of customers improves throughput in a congested system that customers value reasonably highly, while it can be detrimental for less popular services that customers deem rather unrewarding. This is also reflected in social welfare if the firm's profit margin is high enough, although customer welfare always suffers from information costs. *Managerial implications:* Our results shed light on optimal information provision and physical design strategies of service firms and social planners by identifying service settings where they should be most cautious for customers' limited attention. *Academic/practical relevance:* We propose a microfounded framework for strategic customer behavior in queues that links beliefs, rewards, and information costs. It offers a holistic perspective on the impact of information prevalence (and information frictions) on operational performance and can be extended to analyze richer customer behavior and complex queue structures, rendering it a valuable tool for service design.

Key words: service operations, rational inattention, strategic customers, rational queueing, information costs, system throughput, social welfare

1. Introduction

People are inattentive. Even though in today's digitally connected era a vast amount of information is at our disposal when we make choices, we selectively acquire information. This stems mainly from the fact that we all have limited time and attention, which prompts us to acquire and pay more attention to information that we deem important and ignore the rest. In addition, we have limited cognitive capacities, which constrain our ability to process the acquired information. When decisions are made

based on partial information, wrong choices and economic losses are inevitable. Thankfully, people are also adaptive. Depending on the extent of potential losses, we try to allocate time and attention appropriately, and adjust the amount of information acquired (and hence reduce uncertainty). While inattention is omnipresent, our focus in this paper is on service settings which are characterized by the need to execute "production" on demand, tightly coupling customer decisions and resulting demand with supply capacity. Mismatches in demand and supply result in queues and delays, directly impacting perceived service quality. How does limited time and attention of customers impact customer behavior, throughput, and social welfare? These questions form the crux of our paper.

There are different elements of a queueing environment that customers may desire to actively acquire information about, such as service speed, service quality (reward), and expected delays. Perhaps the most prominent element, where the extant literature has mostly concentrated on, is expected delays upon arrival. In many settings, customers do not have perfect information about the queue length and associated delays before they decide to join or balk. For example, in supermarkets or ticketing booths for events, physical obstructions such as shelves, walls, or pillars may make it difficult or even impossible to judge the extent of the queue. For many other services like call centers or health services, information about the queue length may not be readily available, and customers must engage in time-consuming searches to obtain some information about potential delays. Service providers (hereon, firms) may take strategic actions to make it easier or more difficult for customers to acquire this information. Disney is known for having serpentine queues that are partially hidden for popular rides. Yet, it also offers a mobile app that provides wait time information for rides in its theme parks. Likewise, many hospitals in Canada and the US post emergency room (ER) wait times online. Even then, the provided information does not completely resolve customers' uncertainty about queue length and delays. First, the delay information may be deemed as not necessarily accurate, prodding customers to privately learn and validate it. For example, in the context of ER wait times, Ang et al. (2015) empirically show that hospital-posted wait times are extremely unreliable and can be off by as much as an hour and a half for much of the time. Second, and perhaps more importantly, customers have limited time and attention to devote to acquiring such data, and processing the obtained data into useful information. Since information acquisition and processing is "costly", rational customers should trade off the benefits of acquiring better information with the cost associated with it. Our paper is based on this premise of customer behavior.

In classical models of queueing behavior, customers are assumed to be rational. That is, they maximize the expected utility, which is negatively influenced by expected delays. They are also strategic, implying that they consider the actions of other customers when deciding to join or to balk. The equilibrium analysis of these models mainly differs depending on queue visibility, termed as *visible* (observable) and *invisible* (unobservable) queues. The visible and invisible queues are canonical and

orthogonal representations. Visible queues capture the case where customers freely obtain and process all information to accurately assess the length of the queue and expected waiting costs. Invisible queues, on the other hand, represent the other extreme scenario under which customers cannot observe or process any information about the length of the queue and expected waiting costs. In this case, customers act completely based on their prior beliefs on queue length distribution. However, as emphasized above, from the customer's perspective, in practice most queues are neither perfectly visible nor invisible, but they are rather *opaque*. In the end, determining the exact number of customers ahead in the queue requires time, attention, and cognitive capacity, which are all limited in quantity. It may simply turn out to be impossible to make these calculations with certainty and in some cases it may not make rational sense to spend any effort to determine it.

We capture the salient characteristics of limited attention and cognitive capabilities of customers through a model based on the rapidly growing theory of *rational inattention*[1] in economics. Following the seminal work of Sims (2003), this theory quantifies information as reduction in Shannon entropy and assumes that utility-maximizing customers optimally select the *type* and *quantity* of information they need, ignoring the information that is not worth obtaining. In other words, information acquisition process is completely endogenized. Rationally inattentive customers know that they are not going to be able to resolve all uncertainties and make perfect queueing decisions, but they are able to decide (optimally) on what to learn and to what detail. Naturally, this selection depends on all key factors: how much time and attention customers have (i.e., information costs), prior beliefs, as well the nature of uncertainties faced (i.e., what is at stake). We embed the rationally inattentive behavior of customers in a strategic queueing framework, where customers arriving to the queue are not able to discern the exact number of customers ahead of them and hence the expected delay cost. Utilizing this framework, we seek answers to the following fundamental research questions:

- Which equilibrium behavior will prevail (if any) if customers have limited attention and optimally acquire costly information about queue length prior to joining or balking? How is the equilibrium shaped by service characteristics such as service rates, delay costs, rewards, and information costs?
- How does limited attention and information costs impact throughput, firm revenues and pricing, and social welfare?
- Can the firm benefit from customers' limited attention? When does it have the most detrimental effects? What are the implications on the firm's queue information provision strategy?

Rational inattention models are known to be analytically challenging, even in the absence of any strategic queueing considerations. Nevertheless, we are able to show that there is a unique equilibrium that emerges, and establish the directional properties with respect to service characteristics. We find that impact of information costs is more involved, resulting in non-trivial and possibly non-monotone

[1] Throughout the paper we use limited attention and rational inattention interchangeably.

queueing behavior in equilibrium. Elaborating more on such cases numerically, we find that when customers attach reasonably high value to a service that is also popular (high demand), the firm can benefit from the limited attention of customers. The throughput of an opaque queue is higher than fully visible or invisible cases. In sharp contrast, when the firm faces relatively low demand from customers who do not particularly value the service, limited attention of customers can be detrimental. In this case, the firm is better off by either making the queue fully visible or by completely obstructing it. Interestingly, these insights remain valid when the firm sets prices optimally. In particular, customer inattentiveness accentuates effects in this case. This is because the firm can benefit from inattention by overcharging customers when they are more willing to join the queue. Pricing also enables the firm to moderate the throughput losses when customers are less willing to join the queue.

From a welfare perspective, customer surplus suffers from limited attention. However, when firm surplus is taken into account, social welfare can exactly mimic throughput behavior when service price is exogenous. This suggests that obstructing queue information partially can result in win-win outcomes for *both* the consumer and the firm when firm profitability (margin) and congestion (demand) are high, and customer's reward from service is moderate or reasonably high. In contrast, when the service is moderated optimally through pricing by a social planner, social welfare always suffers from limited attention of customers. Furthermore, the welfare losses arising from a service firm's revenue-maximizing pricing behavior are highest when the queue is opaque.

As noted before, limited attention and costly information acquisition do not need to be confined to queue length or delays. Customers can spend time and effort to learn additional factors that may be uncertain, such as service speed or customer reward (service quality). It could also be the case that inattention is only towards these factors, but queue length is visible or invisible to the customers. We extend our baseline model to capture these cases and demonstrate the throughput implications of information costs. Comparing different scenarios, we generate additional insights on the information provision strategy of the firm. For example, we find that throughput is higher when the firm discloses service speed information and lets customers learn queue length privately.

Our contribution to the literature is threefold. First, from a theoretical perspective, we develop a microfounded, tractable framework for service systems that accounts for customers' limited attention and information processing capabilities. When information frictions are related to queue length, our framework bridges the classical visible and invisible queues, covering also the entire intermediate spectrum. Furthermore, it can be adapted to deal with uncertainty in other service characteristics or physical queue capacity, rendering it a versatile tool for service design. Second, utilizing our framework, we provide descriptive results on rational customers' queueing decisions in the presence of information frictions. Through a unified lens, we demonstrate the effects of information costs on operational performance, firm profitability, and social welfare. Third, we translate our descriptive results into prescriptive

insights for practicing managers. We can identify settings where service firms should be most concerned about customers' limited attention and offer practical insights regarding information provision.

2. Literature Review

There is a long-standing literature on strategic behavior in queueing, starting with the pioneering works of Naor (1969) and Edelson and Hilderbrand (1975) for the cases of visible and invisible queues respectively. Hassin and Haviv (2003) and Hassin (2016) present excellent coverage of various extensions and a comprehensive review of related literature. Within this vast literature, the papers that are most closely related to our study can be broadly categorized into two groups: 1) Customer behavior and information acquisition; 2) Firm policies and information provision.

Our work is more closely connected to the first group of papers. One stream in this group examines customers' bounded rationality, which postulates that customers are not always able to make perfect rational choices and therefore make errors. Huang et al. (2013) adapts this to a service context by assuming that customers are not able to perfectly estimate their expected waiting times and investigate the revenue, pricing, and welfare implications for both visible and invisible queues. The degree of bounded rationality is taken as an exogenous parameter. Along similar lines, Huang and Chen (2015) and Ren et al. (2018) assume customers resort to a heuristic in an invisible queue, which involves sampling experiences of previous customers (referred to as anecdotal reasoning). Huang and Chen (2015) assume that customers sample the experience (utility net of delay costs) of a single customer from past generations, and elaborate on pricing and welfare implications. In Ren et al. (2018), customers estimate the expected service *quality* (reward) by taking average of k sampled anecdotes from past customers. As customers gather more samples (i.e., as k increases), they become less boundedly rational, and their quality estimate converges to the true mean. In our model, customers can also make mistakes, but there are crucial differences. First, information acquisition is completely endogenous and is not restricted to particular structures (e.g., sampling past behavior). Second, customers are rational; they optimally decide on what and how much to learn, effectively controlling the type and extent of mistakes they make. This is analogous to customers forming optimal heuristics (Maćkowiak et al. 2018).

Another stream in this group investigates the impact of additional queue length or delay information on queue joining behavior. In Hu et al. (2017), an exogenously specified proportion of customers have perfect information about queue length, while the rest are completely uninformed about it (but they know the fraction of informed customers). They examine how heterogeneity impacts equilibrium throughput and social welfare. In contrast, customers are homogeneous in our model and we focus on the equilibrium that emerges from a microfounded private learning efforts of customers. Closer to our work is the idea that customers can inspect queues at a predetermined cost, upon which queue length is fully revealed, as in Hassin and Roet-Green (2017). They consider an invisible queue where customers make three distinct decisions: join, balk, or inspect. The existence and uniqueness of an equilibrium

strategy is proven and the effects of inspection cost on throughput and social revenue are discussed. We go a step further and allow the customers to determine how much they should improve their knowledge on the state of the queue, which in optimality is never fully informative (some uncertainty always remains). Customers can also decide *not* to acquire any information if processing is deemed to be not useful or too costly. Interestingly, this approach also avoids the complications in equilibrium analysis when customers make a choice between three distinct actions as in Hassin and Roet-Green (2017); once customers acquire information optimally, their eventual decision is only to either join or balk.

We remark that despite the fundamental structural differences noted above, there are some connections with this first body of research, especially regarding the impact of the level of "customer information" (broadly defined) on throughput and social welfare. We establish these links wherever possible, and draw parallels and contrasts. We highlight the complementary insights generated by our unifying framework, which combines customer inattention, endogenous information acquisition, and strategic queueing behaviour in a natural and consistent manner.

The second group of papers investigate queue information disclosure policies and explore the revenue and welfare implications for the service firms. Some works (e.g., Allon and Bassamboo 2011, Yu et al. 2016) focus on the impact of delay announcements on consumer behavior. Others aim to determine optimal queue disclosure policies. For instance, Simhon et al. (2016) show that disclosing the queue length when it is shorter than some threshold and concealing it otherwise, is never optimal when customers are aware of the policy. In contrast, Cui et al. (2017) show that it is indeed optimal when customers are not aware of the policy. There are also other papers that examine the effect of disclosing some form of delay information from the service firm's perspective. For a comprehensive survey, we refer to Ibrahim (2018). We differ fundamentally from these works since it is the *customer* who decides to obtain information about the service environment; there is no predetermined information disclosure strategy of the firm. Nevertheless, the "cost of information" in our model captures the ease at which customers can obtain queue information, and this can be influenced by firm choices related to physical infrastructure and technology. Recognizing this, we comment on how information provision strategies of the service firm may impact equilibrium behavior. This is the only (high-level) connection we have with this second group of literature.

Finally, we note that our paper also contributes to the literature on rational inattention. With recent advances made in both theoretical and empirical grounds, there is a surge in the interest on rational inattention. Applications include consumer (discrete) choice (Matějka and McKay 2015, Hüttner et al. 2019), pricing (Boyacı and Akçay 2017, Matějka 2015), energy efficiency (Sallee 2014), among others. To our knowledge, our paper is the first study that incorporates rational inattention in a service/queueing setting with strategic customers.

3. Baseline Model

Consider a service system modeled as a basic single-server queue operating under FCFS (first come, first served) discipline, with Poisson arrival rate λ and exponentially distributed service times with mean $1/\mu$. Let R denote the unit reward a customer obtains upon being served and p denote the price charged by the firm. A customer arriving to the queue incurs a delay cost of C per unit time for waiting in queue and during service. Suppose that a customer arrives when there are n customers in the system. Then the payoff, expected reward net of the delay cost, is given as

$$v_n = R - p - \frac{C}{\mu}(n+1). \tag{1}$$

Let the value of the outside option of balking (not joining) from the queue be normalized to 0. Clearly, if n is known, the customer will only join if $v_n \geq 0$. Alternatively, if all customers can observe the queue length freely, then they will only join if $n+1 \leq n_e$ where

$$n_e = \left\lfloor \frac{R-p}{C}\mu \right\rfloor \tag{2}$$

and there can only be at most n_e customers in the system. This is exactly the threshold in Naor (1969) for visible queues. Let us assume that $R - p - C/\mu \geq 0$ so that $v_0 \geq 0$, ruling out the uninteresting case where customers have no incentive to join the queue. In our setting, customers are not able to use this threshold policy because they are not able to discern the queue length precisely due to limited attention and cognitive capacity. We first describe how such customers would optimally acquire information and decide to join the queue or not. Subsequently, we characterize the equilibrium.

3.1. Join or Balk Decisions Under Limited Attention

Customers know that the number of customers ahead in the queue is a random variable and have a prior belief about its distribution (common to all). Let us denote customers' prior belief as G and suppose for now that it is specified. One can view this as the *anticipated* queue length distribution, which in equilibrium will coincide with the actual distribution.

Rationally inattentive customers can ask questions and receive signals $\mathbf{s}$ to update their beliefs. Let $\omega \in \Omega$ denote the unknown state of the system (here, queue length) at any time. The customer is free to select an *information processing strategy*, which is represented as the joint distribution $F(\mathbf{s}, \omega)$ of signals and states. The only requirement is that the marginal distribution over the states equals the prior distribution, so that the customers' posterior beliefs are consistent with their priors. The customer chooses this distribution to maximize her *ex ante* expected payoff minus the total cost of information $\hat{c}(F)$ associated with generating signals of different precision levels. Information costs are quantified by the reduction in uncertainty, measured by the Shannon entropy. More specifically, let $H(B)$ denote the uncertainty of belief B measured by entropy. For a discrete distribution,

$$H(B) = -\sum_{\omega} P_\omega \log(P_\omega)$$

where P_ω is the probability of the state of the world $\omega \in \Omega$. Then the total cost of information associated with the information strategy F is given as

$$\hat{c}(F) = \theta(H(G) - \mathbb{E}_{\mathbf{s}}[H(F(\cdot|\mathbf{s})]). \tag{3}$$

Here, $\theta > 0$ is the marginal cost of acquiring and processing information that the customer deems useful (simply referred to as cost of information hereon), and $F(\cdot|\mathbf{s})$ is the posterior belief about state after receiving the signal $\mathbf{s}$. Note that the total cost of information is defined as the *mutual information* between the signal and the state, multiplied by the marginal information cost parameter θ. This cost function is well supported by information theory, since from Shannon's coding theorem it relates to the expected number of questions needed to be asked for implementing a particular information strategy (see Matějka and McKay 2015, Cover and Thomas 2012). The cost of information θ can also be viewed as the *shadow price* of a constraint on the information processing *capacity* of the customer. A stable θ implies that the customer has sufficiently more total information processing capability (and attention) than to the amount she is allocating to the decision task at hand (Sims 2010).

In the context of our queueing system, a customer has two discrete choices, either to "*join*" or to "*balk*" and the state space is $\Omega = \mathbb{N}$, i.e., natural numbers. Let $G = \{g_n; n \in \mathbb{N}\}$ where g_n is the customer's prior belief for the scenario where there are n customers in the system. Given prior belief G, the customer (referred to as "she" hereon) solves a two-stage problem. In the first stage, she selects an information strategy F to refine her beliefs and in the second stage she selects the best option given her posterior belief. Let $V(B)$ denote the expected payoff from choosing the best option given some belief B. Then, a rationally inattentive customer's decision-making problem can be formally stated as:

$$\begin{aligned} &\max_F \sum_{\omega \in \mathbb{N}} \int_s V\left(F\left(\cdot \mid s\right)\right) F\left(ds, \omega\right) - \hat{c}\left(F\right) \\ &s.t. \int_s F\left(ds, \omega\right) = g_\omega \text{ for } \omega \in \mathbb{N}. \end{aligned} \tag{4}$$

The first term in (4) is the ex ante expected payoff from selecting the best option based on the generated posterior belief and the second term is the total cost of information given by (3). According to this model, the customer is optimally choosing (i) what and how much information to process (what to pay attention to, how much attention to pay) and (ii) what action to select given the information gathered.

A central result in rational inattention theory that helps to simplify the customer's problem is that each action can be selected in at most one posterior belief (Matějka and McKay 2015). This means that receiving distinct signals that lead to the same posterior is suboptimal, since it implies the acquisition of ample information (which is costly) that is not acted upon. The immediate consequence is that choosing signals is equivalent to choosing actions. As such, total information cost can be written using the mutual information between the chosen actions and state. Thanks to this property, it becomes

possible to write an alternative maximization problem for the customer that uses state-dependent choices as decision variables, without any referencing to signals. Specifically, let S^J denote the set of signals that lead to joining decision. Then the induced *conditional* joining probability when there are n customers in the system can be represented as

$$\pi_n = \int_{s \in S^J} F\left(ds|\omega = n\right)$$

where $F(\cdot|\omega)$ is the conditional distribution of signals given state ω. Let $\Pi = \{\pi_n; n \geq 0\}$ denote the collection of conditional joining probabilities; i.e. customer's joining policy (which also implies customer's balking policy). Based on this, conditional joining probabilities averaged over all states, i.e., the *unconditional* joining probability is

$$\overline{\pi} = \sum_{n \geq 0} \pi_n g_n. \tag{5}$$

Then, for our queueing system, the customer's equivalent optimization problem can be reformulated as

$$\begin{aligned} &\max_{\Pi = \{\pi_n; n \geq 0\}} \sum_{n \geq 0} v_n \pi_n g_n - c\left(\Pi, G\right) \\ &s.t. \quad \pi_n \in [0,1] \ \ \forall n \geq 0. \end{aligned} \tag{6}$$

Here $\pi_n g_n$ is simply the joint probability that the customer joins and there are n people in the system. Since the utility of balking is 0, the first term is the total expected payoff obtained under joining policy Π. The second term is the total cost of information quantifying the reduction in entropy, i.e. mutual information between the action and state (due to the equivalence of signals and actions) scaled by information cost θ:[2]

$$c\left(\Pi, G\right) = \theta \left(-\overline{\pi} \log \overline{\pi} - (1 - \overline{\pi}) \log\left(1 - \overline{\pi}\right) + \sum_{n \geq 0} g_n \left(\pi_n \log \pi_n + (1 - \pi_n) \log\left(1 - \pi_n\right) \right) \right). \tag{7}$$

It is established in the rational inattention literature that the optimal information processing strategy for any $\theta > 0$ results in conditional choice that follows a generalized multinomial logit (GMNL) formula (Matějka and McKay 2015). In particular, when the choice is about joining a queue or not, as described above, the conditional probability π_n of joining the queue when there are n customers satisfy

$$\pi_n = \frac{\overline{\pi} e^{v_n/\theta}}{\overline{\pi} e^{v_n/\theta} + 1 - \overline{\pi}} \quad \text{almost surely for } \theta > 0 \tag{8}$$

where $\overline{\pi}$ is the *unconditional* probability of joining the queue that needs to satisfy equation (5) for consistency. Here it is worth noting that the customer is *not* randomizing her choice. Rather, her choice

[2] Due to the symmetry of mutual information $H(G) - H(G|\Pi) = H(\Pi) - H(\Pi|G)$ where the left hand side is analogous to (3).

is uncertain ex ante as she does not know what her assessment about the queue length (i.e., signals) will reveal. If $\theta = 0$, then the customer joins or balks deterministically, depending on which one yields the higher payoff in that state.

The conditional joining probabilities characterized by the GMNL equation (8) capture the intricate relationship between three central drivers of a customer's decision, namely the payoffs, beliefs, and information costs. Observe first that according to (8), if the customer has a positive probability of joining in at least one state (i.e., $\overline{\pi} > 0$), then she has a positive probability to join in all other states of the system. However, the higher her payoff (v_n), the more likely she will join. Accordingly, the state-dependent joining probabilities increase as n decreases (i.e., v_n increase). The impacts of prior beliefs are captured through the unconditional probability $\overline{\pi}$. It is crucial to note that $\overline{\pi}$ is *not* an exogenous parameter; rather it is part of the customer's endogenous decision-making process. One needs to simultaneously solve (8) (for all $n \geq 0$) together with the consistency equation (5) to arrive at a complete explicit solution. Rewriting (8) as $\pi_n = \left(e^{v_n/\theta + ln(\overline{\pi})}\right) / \left(e^{v_n/\theta + ln(\overline{\pi})} + e^{0 + ln(1-\overline{\pi})}\right)$, it is evident that unconditional probability $\overline{\pi}$ effectively shifts the customer's payoff. Hence, her joining decision is swayed by how "attractive" it is a priori to join the queue. Information costs play a strong role in how much emphasis the customer puts on the beliefs. When θ is low, the customer can acquire more information about each state and the payoff. In the extreme case when $\theta \downarrow 0$, the queue length is fully visible to the customer, and she deterministically makes the best choice in each state. In contrast, as θ increases, she acquires less information and relies more on her belief. In the extreme case when $\theta \uparrow \infty$, the customer deterministically joins or balks based on her ex ante beliefs.

It is worthwhile at this point to make a connection with models that use the *standard* MNL choice to model customers' bounded rationality. In particular, Huang et al. (2013) adopt MNL choice in their separate analysis of visible and invisible queues. The corresponding parameter in the MNL specification (standard deviation of an additive noise) captures customers' degree of bounded rationality. At one extreme, customers are fully rational, recovering separately the Naor (1969) and Edelson and Hilderbrand (1975) models for visible and invisible queues respectively. At the other extreme, customers are fully irrational and join or balk with equal probability. In contrast, there is no degree of rationality in our framework (customers are always rational). The GMNL formula captures rational customers' endogenous response to information frictions, effectively bridging the invisible and visible queues via the information cost θ.

In order to fully solve the problem, we plug the conditional joining probability π_n given by (8) into the customer's optimization problem in (6), which yields a simplified representation of

$$\max_{\overline{\pi} \in [0,1]} \theta \mathbb{E}_G \left[\log \left(\overline{\pi} e^{v_Q/\theta} + 1 - \overline{\pi} \right) \right] \tag{9}$$

where Q denotes the uncertain queue length with prior distribution G. Hence, in effect, the customer is choosing the unconditional probability $\overline{\pi}$. It is clear that the problem in (9) is concave in $\overline{\pi}$ with linear

constraints and can be easily solved. Once the unconditional joining probability $\overline{\pi}$ is found, conditional joining probabilities are calculated using (8).

Note that equation (8) constitutes the necessary conditions for optimality but they are not sufficient. For instance, the queueing policy $\Pi_1 = \{\pi_n = 1; n \geq 0\}$ where everyone joins and the policy $\Pi_0 = \{\pi_n = 0; n \geq 0\}$ where everyone balks at each state automatically satisfy these conditions. Hence, it holds trivially when one of the two actions (join or balk) is not chosen at all, but does not specify when this may occur. The next lemma presents the complete characterization of the customer's optimal joining policy, including both the necessary and sufficient conditions. The proofs for the next and subsequent results are relegated to the Appendix.

LEMMA 1 **(Necessary and Sufficient Conditions)**. *The policy $\Pi = \{\pi_n; n \geq 0\}$ is optimal if and only if the implied unconditional choice probabilities $\overline{\pi} = \sum_{n\geq 0} \pi_n g_n$ for actions "join" and "balk" satisfy*

$$\sum_{n\geq 0} \frac{e^{v_n/\theta} g_n}{\overline{\pi} e^{v_n/\theta} + 1 - \overline{\pi}} \leq 1 \quad \textit{(for "joining")} \quad \textit{and} \quad \sum_{n\geq 0} \frac{g_n}{\overline{\pi} e^{v_n/\theta} + 1 - \overline{\pi}} \leq 1 \quad \textit{(for "balking")}, \tag{10}$$

and both equations have to hold with equality if $0 < \overline{\pi} < 1$. Otherwise, the sufficient conditions are

$$\sum_{n\geq 0} e^{v_n/\theta} g_n \leq 1 \textit{ for } \overline{\pi} = 0 \quad \textit{and} \quad \sum_{n\geq 0} e^{-v_n/\theta} g_n \leq 1 \textit{ for } \overline{\pi} = 1. \tag{11}$$

Lemma 1 establishes that there are cases that yield join or balk decisions with certainty, i.e., without the need for processing any information. For instance, it is possible that the first condition in (11) is satisfied when the customer's prior belief towards low states (i.e., when there are few customers in the queue) is very weak and it is optimal for the customer to balk with certainty, i.e., $\overline{\pi} = 0$. A similar effect may also take place when the customer attaches a very low value to the service provided, i.e., a low service reward. On the contrary, when the customer strongly believes that there will be few customers in the queue and/or her reward from service is high enough, then she may decide to join with certainty without obtaining further information, i.e., $\overline{\pi} = 1$ and the second condition in (11) is satisfied.

As a final remark, we highlight the link with more traditional search/inspection models. A prominent example is the model by Hassin and Roet-Green (2017), where arriving customers have the opportunity to buy perfect information at a fixed cost. In our framework, receiving perfect information is equivalent to reducing the posterior entropy to zero. This is tantamount to customers paying a fixed cost of $\theta H(\Pi)$ and deciding to join based on the current queue length (or not pay and act on the basis of prior beliefs only). However, this is a suboptimal strategy when information strategy is endogenized, since a marginal deviation from generating perfect signals is always beneficial as per the objective function in (6).[3] This is why the optimal decision is always probabilistic if the customer chooses to acquire and process some information. Nevertheless, the customer can also choose *not* to process any information and act solely based on her prior belief (see Lemma 1).

[3] Technically speaking, this is because the slope of entropy is infinite when signals fully resolve the uncertainty (i.e., $\pi_n = 0$ or $\pi_n = 1$). Therefore, leaving some uncertainty drastically reduces the information cost compared to the corresponding finite gain in expected utility due to more information.

3.2. Joining Behaviour in Equilibrium

Until this point, we have assumed that customers have exogenously specified prior beliefs, representing the anticipated queue distribution. In fact, customers are strategic in our framework; they are aware of the other rationally inattentive customers and anticipate their actions. As a result, customer beliefs are formed by queueing behavior in equilibrium. Queueing behavior itself is shaped by the optimal information acquisition and joining decisions of rationally inattentive customers in each state. Next, we define such an equilibrium. To this end, let $\tilde{G} = \{\tilde{g}_n; n \geq 0\}$ denote the queue length distribution in equilibrium and $\rho = \lambda/\mu$ be the utilization factor.

DEFINITION 1. In the queueing system with rationally inattentive customers with information cost $\theta > 0$, the equilibrium probability of joining the queue when there are n customers present is

$$\widetilde{\pi}_n = \frac{\widetilde{\pi} e^{v_n/\theta}}{1 - \widetilde{\pi} + \widetilde{\pi} e^{v_n/\theta}} \tag{12}$$

where $\widetilde{\pi} = \sum_{n\geq 0} \widetilde{\pi}_n \tilde{g}_n$ is the unconditional joining probability. Equilibrium queue length distribution is

$$\tilde{g}_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \rho^k \widetilde{\pi}_0 \widetilde{\pi}_1 ... \widetilde{\pi}_{k-1}} \quad \text{and} \quad \tilde{g}_n = \tilde{g}_0 \rho^n \widetilde{\pi}_0 \widetilde{\pi}_1 ... \widetilde{\pi}_{n-1} \quad \text{for } n \geq 1. \tag{13}$$

THEOREM 1. *There exists a unique equilibrium satisfying Definition 1.*

Although our framework involves nontrivial customer behaviour in terms of queueing, Theorem 1 proves a strong result that a unique equilibrium exists despite the complexity of the model. Finding this equilibrium requires solving a fixed point equation since it requires a consistency between the joining probabilities (which is a solution of the rational inattentive customer's optimization problem) and resulting queue length distribution (which is an input as customer's prior belief to the same optimization problem).

There still remains the question of whether rationally inattentive customers can form the correct belief about the queue length distribution, that is whether the equilibrium can be attained. We show in the Appendix that such an equilibrium can be attained via adaptive learning, in a setting where customers can observe and take averages of the joining fractions of past customers.

An immediate corollary of Theorem 1 presents the limiting cases of the information cost, which have strong connections with extant literature on strategic queueing.

COROLLARY 1. *(i) (Visible queues) When $\theta \downarrow 0$, customers can determine the queue length exactly and decide to join only if the number of people in the system is strictly less than the threshold $n_e = \left\lfloor \frac{(R-p)\mu}{C} \right\rfloor$. Corresponding equilibrium effective joining fraction is $\widetilde{\pi} = \frac{1-\rho^{n_e}}{1-\rho^{n_e+1}}$.*

(ii) (Invisible queues) When $\theta \uparrow \infty$, customers base their queueing decisions on their prior beliefs only, resulting in the following outcomes (when $\lambda < \mu$)

(a) *(Always join) If $R - p - \frac{C}{\mu - \lambda} \geq 0$, then all customers join the queue.*

(b) *(Mixed strategy) Otherwise, equilibrium joining fraction is* $\widetilde{\pi} = \min\left\{\frac{\mu}{\lambda} - \frac{C}{(R-p)\lambda}, 1\right\}$.

The two extreme cases of our framework covered in Corollary 1 retrieve the classical models and results in the literature. The first case is precisely the scenario with the visible queue model of Naor (1969), where the equilibrium is a threshold policy. The latter case is precisely the invisible queue model of Edelson and Hilderbrand (1975). In particular, if the benefit of joining the queue is positive even if everyone else joins, customers join with probability 1. Otherwise, customers join with a fixed probability. Given our assumption $R-p-C/\mu>0$, the case where all customers balk does not occur.

Next, we investigate the impact of salient characteristics of the service system, namely service reward R, price p, delay cost rate C, and service rate μ on customer joining fraction $\widetilde{\pi}$ (or throughput $\lambda\widetilde{\pi}$).

PROPOSITION 1. *Equilibrium joining fraction* $\widetilde{\pi}$ *increases in* R *and* μ *and decreases in* p *and* C.

The rationale for this result is quite intuitive. Rationally inattentive customers cater their learning and ultimate actions towards what is most beneficial for them. In this sense, given an exogenous prior belief, they choose to join more if doing so brings more payoff at each state. This is surely the case when the reward from service R is higher, and the price p and/or waiting cost per unit time C are lower. Proposition 1 establishes that this also holds in equilibrium. To account for the unilateral impact of these service characteristics, we define $\overline{R} = \frac{R-p}{C}$ as the attractiveness of service to the customers.[4] According to Proposition 1, equilibrium joining probability increases in service attractiveness $\overline{R}$. The effect of service rate μ, on the other hand, is not as straightforward. Although a faster service may potentially mean less congestion, it also incentivizes customers to join which in turn may increase congestion. It turns out that the first affect is always stronger, and a faster service rate always yields more joining customers in equilibrium.

The impact of information cost on joining behavior is more convoluted and usually more difficult to analytically ascertain. We take a deeper look into it and its implications in the next section.

4. Impact of Information Cost on Throughput and Social Welfare

Information costs predominantly affect the extent of learning customers can afford to (or be able to) undertake regarding queue length prior to joining. This has a critical impact on customers' joining behaviour in equilibrium, and consequently on system throughput (which proportionally impacts the service provider's profitability) and social welfare. We start with the former.

4.1. Throughput in Equilibrium

Although more information about the queue length always leads to better joining and balking decisions on the consumers' side (and hence higher expected payoffs), it does not always benefit the service firm. Indeed, considering the two extreme cases of queue visibility, it is known that throughput is

[4] Note that when unit waiting cost C is normalized to 1, $\overline{R}$ becomes $R-p$ which is customer's net value from the service.

higher for invisible queues when arrival rate is lower than a certain unique threshold (Chen and Frank 2004). This is because customers in an invisible queue blindly choose to join due to low congestion while customers in visible queues may still face long queues to deter them from joining. As we bridge these two extremes, at a first glance, it seems quite plausible that the effects of information cost θ on throughput should be monotonic, and its direction should depend on whether visible or invisible queues have higher joining rates in equilibrium. However, this intuition turns out to be only partially correct in the case of rationally inattentive customers as more convoluted forces start to impact customers' equilibrium joining behaviour. To illustrate this, in the following discussion, we use $\widetilde{\pi}^{(\theta)}$ to denote the equilibrium joining fraction when the information cost is θ.

We find that the impact of information cost on throughput is largely governed by both the attractiveness of the service to the customer and the level of demand (congestion). We first show that for any demand level, there is a range for service attractiveness in which throughput can potentially be non-monotone in information cost θ. In particular, outside of this interval, throughput is provably either monotonically increasing or decreasing.

PROPOSITION 2. *For $\lambda < \mu$ there exists two critical thresholds $\overline{R}_L$ and $\overline{R}_H$ for service attractiveness such that throughput ($\lambda\widetilde{\pi}^{(\theta)}$) is monotonically decreasing in θ if $\overline{R} < \overline{R}_L$ and monotonically increasing if $\overline{R} > \overline{R}_H$.*

In order to see the intuition behind Proposition 2, note that when service is quite attractive to the customers, they naturally have a tendency to join even if they process no information and act only based on their prior beliefs. Indeed, the sufficient conditions in Lemma 1 confirm that for sufficiently large $\overline{R}$ (hence large v_n) customers may choose to always join without processing any information. However, as they start to discern the queue length more precisely, some customers will choose to balk due to congestion. In such cases, $\widetilde{\pi}^{(0)}$ is relatively low compared to $\widetilde{\pi}^{(\infty)}$, and equilibrium joining fraction $\widetilde{\pi}^{(\theta)}$ is monotonically increasing in θ. In contrast, when service is quite unattractive, customers are unlikely to join, simply based on their prior beliefs and $\widetilde{\pi}^{(\infty)}$ will be low. As they learn the queue length, however, some (lucky) customers will still be able to find the queue short enough to warrant joining. In such cases, $\widetilde{\pi}^{(0)}$ is relatively higher compared to $\widetilde{\pi}^{(\infty)}$, and equilibrium joining probability $\widetilde{\pi}^{(\theta)}$ is monotonically decreasing in θ. When the equilibrium joining fractions under visible and invisible queues are not very distinct, then we numerically observe a different, and possibly non-monotone behavior with respect to information costs. This occurs within an intermediate range of service attractiveness $\overline{R}$. Furthermore, we find that the non-monotone range increases in demand level (congestion). However, the behavior inside this region can be quite different depending on the level of demand and service attractiveness. In what follows we investigate two distinct throughput patterns via two examples for low demand and high demand scenarios, respectively, to explain the main factors that drive the non-monotone joining behaviour.

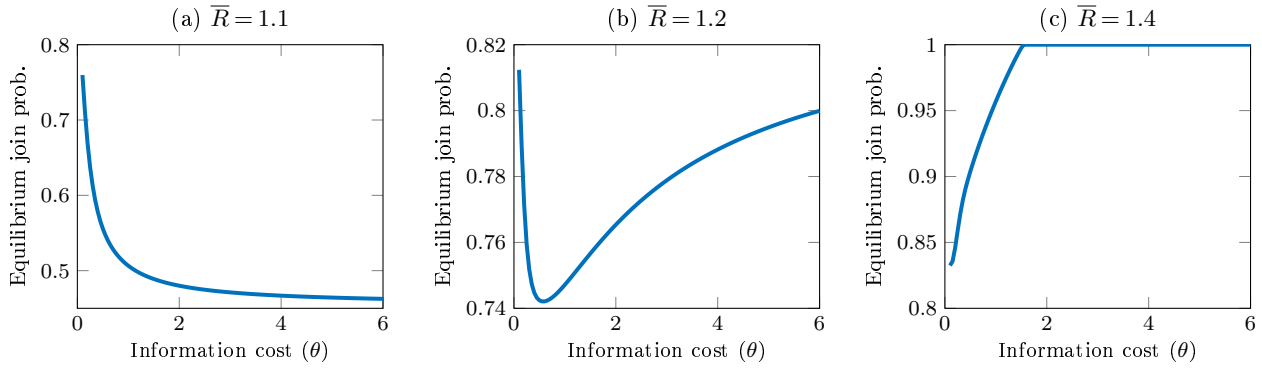

Figure 1 **Impact of information cost on equilibrium joining fraction and throughput: low demand ($\lambda=0.2,\mu=1,C=1$)**

Figure 1 illustrates three different cases when the demand rate is low. Evidently, when service attractiveness $\overline{R}$ is also sufficiently low, higher information costs lead to lower throughput as per Proposition 2. Similarly, when service attractiveness is sufficiently high, throughput is monotonically increasing in θ. Interestingly, when $\overline{R}$ is in the intermediate range (but still relatively low), throughput may first decrease and then increase. The main rationale for this behaviour is as follows. When $\theta=0$, customers can observe the queue length and deterministically make the best decision. On the other hand, when information cost θ is slightly increased, customers start to process information and due to their limited attention, they may not be able to discern queue length perfectly when it is in fact relatively short, and balk erroneously. This has a negative impact on throughput. Arguably, for the same reason, customers may not be able to discern queue length when it is longer, and erroneously join instead of balking. In the case of low demand, however, this is less likely to happen as low states are more likely to be observed. Hence, throughput initially decreases in θ. When information cost is further increased and customers start to rely on their beliefs more, they join with a higher probability because they believe that the system is not congested and despite the low reward, it makes sense to join instead of balking. As a result, throughput starts increasing in θ.

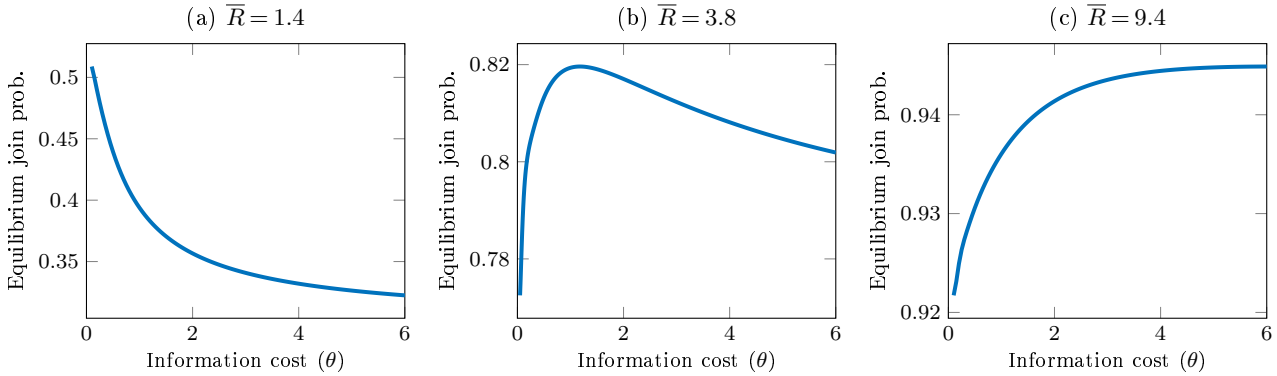

Figure 2 **Impact of information cost on equilibrium joining fraction and throughput: high demand ($\lambda=0.95,\mu=1,C=1$)**

Figure 2 illustrates three different cases when the demand rate is high. As in the case of low demand, when service attractiveness is dominating in either direction, throughput becomes monotone in θ accordingly. Similarly, there is an intermediate level of service attractiveness that potentially makes throughput non-monotone in information costs. Strikingly, here the effect is predominantly opposite; throughput first increases and then decreases. The intuition follows a logic very similar to the low demand case, but the effects are reversed. This is because when demand is high, longer queue lengths are more likely, and therefore (erroneous) joining decisions at higher queue lengths due to limited attention outnumber (erroneous) balking decisions at lower queue lengths. Hence, we see an initial increase in throughput. As information cost further increases, the impact of customers' prior beliefs kick in and they start to join less due to high congestion, in which case, throughput starts decreasing (Figure 2b).

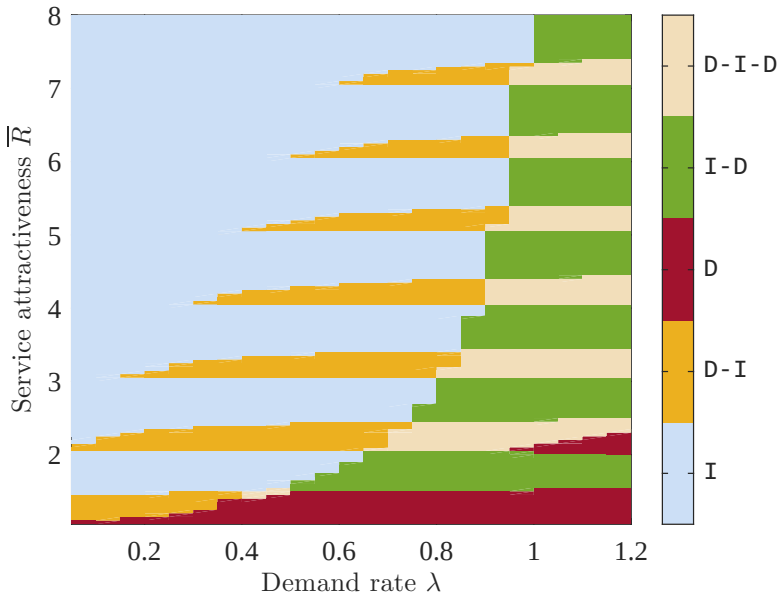

Figure 3 **Spectrum of different throughput behaviour with respect to information cost (I: incr., D: decr., $\mu = 1$)**

The two distinct unimodal structures illustrated in Figures 1b and 2b, are not the only patterns that can be observed inside the non-monotone range of service attractiveness. In fact, a combination of the two is also possible and can be observed for a wide range of parameters along with the two monotone structures. Figure 3 depicts the full spectrum of possible patterns in $\lambda - \overline{R}$ space and hence provides a comprehensive picture of the impact of information cost on throughput. Here, I and D respectively denote *increasing* and *decreasing* patterns, and their collections define how throughput changes as information cost increases (i.e., I-D denotes first increasing, then decreasing behaviour as illustrated in Figure 2b). As Proposition 2 shows, for any given λ, throughput is monotonically decreasing when service is sufficiently unattractive (red region). Conversely, for $\lambda < \mu$, throughput

is monotonically increasing when service is sufficiently attractive[5] (blue region). In between these two thresholds, throughput is possibly non-monotone. Furthermore, for higher demand levels, this intermediate range is wider, as noted before. This is because relative attractiveness of service should be much higher to dominate the negative impact of increased congestion so that the less customers are informed about the queue length, the more they decide to join.

The full parameter range of the two distinct non-monotone patterns *D-I* and *I-D* are shown by the orange and green regions in Figure 3, respectively. Note that the *I-D* pattern is usually observed when demand is high. On the other hand, *D-I* pattern emerges for all $\lambda < \mu$ depending on the value of $\overline{R}$. In fact, it also manifests itself in the region where the *I-D* pattern is more prominent, which creates the combination pattern *D-I-D* as shown by the light yellow region in Figure 3.

Figure 3 generates additional insights on non-monotone throughput behavior. In particular, note that both *D-I* and *D-I-D* structures repeatedly emerge when $\overline{R}\mu$ (equivalently $\overline{R}$ since μ is normalized to 1) is close to its integer part as defined by the Naor threshold n_e in (2). The rationale for this behaviour is as follows. When $\overline{R}\mu$ is close to the Naor threshold, joining brings only marginally more payoff to customers than balking at state $n_e - 1$; that is v_{n_e-1} is slightly above zero. Accordingly, it becomes less important for the customers to discern this threshold state and make the correct joining decision. Yet, $n_e - 1$ is the longest queue length that makes joining desirable, and hence it has the most bearing on the errors made by the customers. On one hand, when customers are perfectly informed about the queue length (i.e., a visible queue), they will always join, even though it is only slightly more rewarding. On the other hand, when customers face some information frictions, they will be prompted to spend less time and effort, and this would lead to more erroneous balking, albeit not to significant reductions in payoffs. This is why an initial decline in throughput is observed as θ is increased from zero. This effect is present as long as service attractiveness is not dominating (otherwise throughput is increasing by Proposition 2). When demand is high, the same effects discussed before for Figure 2b are at play. The initial decline in throughput is quickly followed by an increase as high demand leads to higher states being more likely, which in turn yields more erroneous joining behaviour. This is further followed by a decrease in throughput as the impact of prior beliefs become more dominant, which leads to the *D-I-D* pattern.

It is important to note that although the initial decline in throughput noted above can be observed in multiple regions, its magnitude and impact becomes greater when *both* $\overline{R}$ and λ are lower. This is because as $\overline{R}$ (and equivalently n_e) is smaller, there is a smaller number of states where joining is preferred to balking, and the impact of erroneously balking at state $n_e - 1$ becomes more critical. When λ reduces, low states (shorter queue lengths) are more likely to be realized, further amplifying

[5] For $\lambda \geq \mu$, there is no entrance to the system when the queue is invisible, i.e., $\theta \to \infty$. Therefore, throughput cannot be monotonically increasing.

this effect. Consequently, the most detrimental impact, and hence the biggest decline in throughput, is realized when $n_e = 1$ and customers need to distinguish only one state from the others (i.e., to learn whether the queue is empty or not). Coupled with a low congestion level, this results in the *D-I* pattern demonstrated in 1b and represented by the lower-left orange region in Figure 3. We substantiate this point in Figure 4. In both Figure 4a and 4b we observe the *D-I* pattern when $\overline{R}$ is close to its integer part and the extent of initial throughput reduction is greater for lower $\overline{R}$. As $\overline{R}$ moves up and away from its integer part, throughput becomes monotonically increasing. Finally, Figure 4c represents the high demand case where we observe the *D-I-D* pattern when $\overline{R}$ is close to n_e. Otherwise, the throughput is first increasing then decreasing.

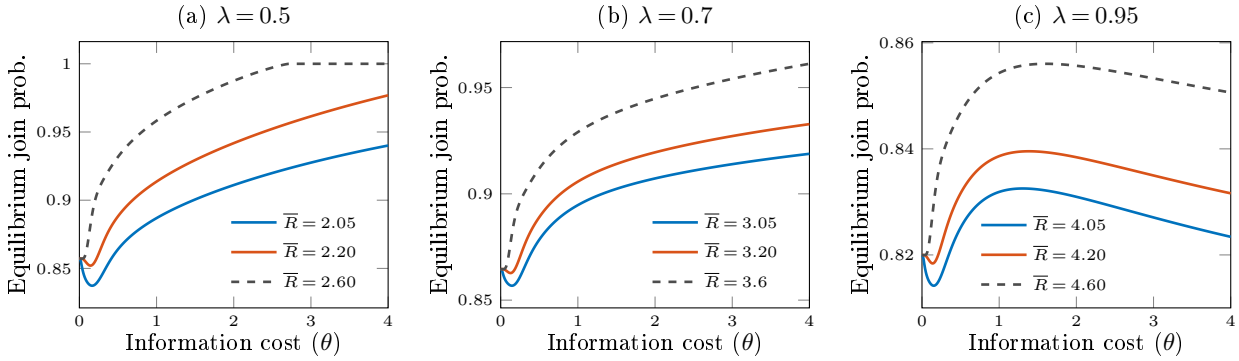

Figure 4 **Impact of information cost on throughput ($C = 1, \mu = 1$)**

To summarize, limited attention and information costs have involved effects on throughput. When the service is significantly attractive (resp., unattractive), the firm always benefits from information frictions and ideally should make the queue invisible (resp., visible). In the intermediate range of service attractiveness, a more complex and non-monotone behavior consistently prevails, and this range widens as demand increases. In particular, the *D-I* pattern is most distinctly observed and has significant impact in low demand systems with reasonably low service attractiveness, while the *I-D*/*D-I-D* patterns are predominantly observed in high demand systems. These results offer a compounding insight on the effect of information prevalence on customer behaviour and hence have significant managerial implications. It establishes that when customers value a service that is already on-demand (e.g., a popular restaurant) the firm can benefit from an opaque queue and find it in its best interest to make information acquisition difficult or even to deliberately obstruct it to some extent, rather than providing a completely visible or invisible system. The opposite is true when the firm faces low demand from customers who do not value the service much (e.g., a public office call center). In this case, it may be optimal for the firm to employ an "all-or-nothing" information provisioning strategy.

We remark that there are some other papers that find that throughput might be unimodal in terms of information prevalence. Hassin and Roet-Green (2017) conclude that a positive and finite inspection

cost might achieve a higher throughput. Similarly, Hu et al. (2017) finds that having a portion of uninformed customers in the society might be better in terms of throughput. Our framework is able to explain these results using a behavioral model that is rooted in the first principles and systematically links the main drivers of human decision making such as beliefs, payoffs, and information costs. More importantly, it is also able to capture counter cases where throughput suffers from limited attention alluding to the potential dangers of deliberate obstruction of information acquisition. To the best of our knowledge, this is not noted in the strategic queueing literature.

In a similar vein, Huang et al. (2013) and Huang and Chen (2015) capture the impact of customers having more information on expected delays in invisible queues. They show that, as customers improve their anecdotal reasoning or become less boundedly rational, throughput improves in high reward (low price) systems and deteriorates in low reward (high price systems). We complement these results by demonstrating what happens when rational customers start discerning *realized* queue length and associated expected delays. Interestingly, when customers are informed beyond the fully rational benchmarks of Huang et al. (2013) and Huang and Chen (2015), throughput is strictly worse in reasonably high reward systems (Figures 1c, 2c, and the blue region in Figure 3) and strictly better in reasonably low reward systems (Figures 1a, 2a, and the red region in Figure 3).

4.2. Social Welfare in Equilibrium

We now investigate how information cost impacts social welfare. For a given service fee p and information cost θ, social welfare is defined as the expected net utility to the society (customers plus the service firm) per unit time. In other words, it is the sum of customer surplus and firm surplus. Since information cost is a real cost that customers take into account in their decisions, it is also part of their surplus, and hence social welfare. Service fee p, on the other hand, is merely a transfer payment between customer and service firm, and only indirectly influences social welfare through the joining behaviour it induces. More specifically, social welfare W_s is

$$W_s = \lambda(\mathbb{E}_{\widetilde{G}}[(R - C(Q+1)/\mu)\widetilde{\pi}_Q] - c(\widetilde{\Pi}, \widetilde{G})) \tag{14}$$

where the former term inside the parentheses is the total expected service reward net of waiting cost, and the latter is total information cost given by (7) in equilibrium. Social welfare per unit time is this difference scaled by the demand rate λ.

A more convenient way of writing the social welfare is as the sum of customer and firm surpluses. In particular, social welfare in (14) can be written as $W_s = W_c + R_I$ where customer surplus

$$W_c = \lambda(\mathbb{E}_{\widetilde{G}}[v_Q\widetilde{\pi}_Q] - c(\widetilde{\Pi}, \widetilde{G}))$$

is customers' optimal net expected utility (9) in equilibrium, and firm surplus

$$R_I = \lambda p \widetilde{\pi}$$

is simply the throughput times profit margin p of the firm. Writing social welfare this way helps to disentangle the impact of system parameters on its two constituents.

We first focus on the case where equilibrium social welfare consists only of customer surplus (equivalently, $p = 0$). When $W_s = W_c$, social welfare is decreasing in information cost θ. This is intuitive as customers benefit from making more informed decisions due to lower cost, and capacity and demand are better matched. Furthermore, θ scales the information cost in customers' optimization problem in (6) and it is natural that in optimality their surplus is lower (for an anticipated queue length distribution). Next proposition shows that this also holds in equilibrium.

PROPOSITION 3. *W_c is decreasing in θ.*

Proposition 3 is largely consistent with related literature that investigates the impact of information prevalence on social welfare when it is measured by customer surplus only. In particular, Hassin and Roet-Green (2017) find that social welfare decreases as the cost of inspecting the queue length increases. Similarly, in Hu et al. (2017), social welfare increases as the proportion of informed customers in the population increases (as long as some uninformed customers still remain). In the bounded rationality model of Huang et al. (2013), however, customer surplus *can* increase when customers are less rational. In other words, by making suboptimal decisions customers might improve their surplus. We show that this is no longer the case when customers endogenize and optimize their learning, while internalizing the associated costs.

When $p \neq 0$ and firm surplus is taken into account, social welfare may not always decrease in information cost. Indeed, we already know that the service provider may benefit from limited attention and information costs, especially when demand is high where throughput is increasing or increasing-decreasing in information costs. In such cases, the welfare loss on the customer side due to increased information cost may not offset the increase in firm surplus. Mathematically speaking, social welfare is the sum of a decreasing function (W_c) and a possibly increasing or non-monotone function (R_I) as per the last section. Furthermore, while customer surplus is affected by the level of service reward net of price (i.e., $R - p$), firm's surplus is additionally affected by the magnitude of p. Accordingly, by changing p while keeping $R - p$ constant, it is possible to scale the two components and observe different social welfare behavior in terms of information cost θ. In fact, social welfare would fully mimic the throughput behavior summarized in Figure 3, provided that firm surplus is sufficiently large in scale compared to consumer surplus. This is demonstrated in Figures 5a and 5b for low and high demand systems, respectively. Note that in both cases, customer surplus W_c is monotonically decreasing in θ as Proposition 3 indicates. The structure of social welfare function W_s, on the other hand, mimics the non-monotone throughput behavior. In fact, customer joining behavior (and throughput) in these two cases are identical to those demonstrated in Figures 1b and 2b. Here, the high service fee p renders the

firm surplus to dominate the customer surplus. Conversely, Figure 5c illustrates the case where firm surplus is not particularly dominating to the customer surplus, and both customer surplus and social welfare are decreasing in information cost θ. Note that the throughput in this case is the same as in Figure 2b; the only difference is the scaled-down profit margin of the firm.

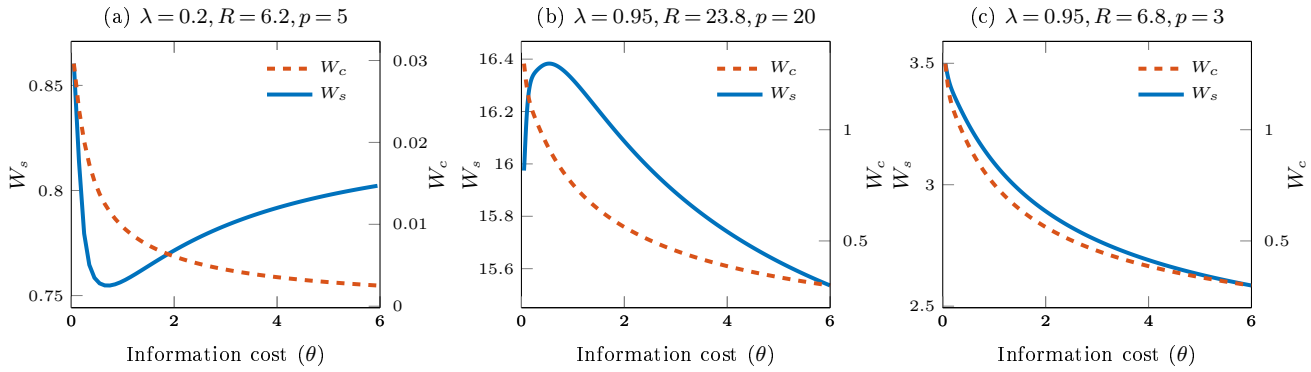

Figure 5 **Impact of information cost on social welfare ($C = 1, \mu = 1$)**

Evidently, if a popular (high demand) service that is well valued by customers (reasonably high $R-p$) is also very profitable for the firm (high p), social welfare might benefit from information frictions. It might even be optimal from the society's perspective to have an opaque queue. The opposite can also occur, though is perhaps less likely. For a low-demanded and reasonably low-valued service that is still relatively profitable to the firm, social welfare might suffer the most at an intermediate level of information friction. In the moderate price regime, customer welfare and social welfare are generally aligned, promoting visible queues and easing of information acquisition for customers.

5. Pricing Implications

Our analysis thus far assumed that service fees are exogenously given. In this section we focus on the case when the service provider has the ability to moderate the service through pricing, and analyze two distinct cases where the service provider is a revenue maximizer or a social planner.

5.1. Revenue Maximization

Suppose now that the service provider aims to maximize the expected revenue $R_I = \lambda p \widetilde{\pi}$. Unfortunately, the expected firm revenue is not necessarily unimodal in price and can admit multiple optimal prices, especially for low information cost θ. To see this, it is important to revisit the limiting cases of visible and invisible queues. It is well-known that for visible queues, the revenue function displays a sawtooth pattern, but a unique optimal Naor threshold and an associated price can be found. For invisible queues, the revenue function is strictly concave, ensuring a unique optimal price. The information cost θ in our opaque queue effectively bridges these two cases, which may result in a multi-modal revenue function and multiple optimal prices. Furthermore, there could be discontinuities in optimal prices (i.e.,

price jumps) when information cost varies. Figure 6 presents an illustrative case in detail. Observe first how the expected revenue function in Figure 6a smooths out as θ increases, starting from the sawtooth shape for the visible queues and eventually becoming concave (here $\theta = 15$ is effectively identical to the invisible queue case). Figure 6b is a close-up version of the same revenue curve for two very close θ values, which clearly illustrates how a marginal deviation in θ may cause a significant jump in the optimal price (the dots refer to the highest revenue values). Finally, Figure 6c illustrates how the firm-optimal price changes with respect to θ for the same service setting, where an initial decline in optimal price is followed by a drop, which is then followed by a unimodal pattern.

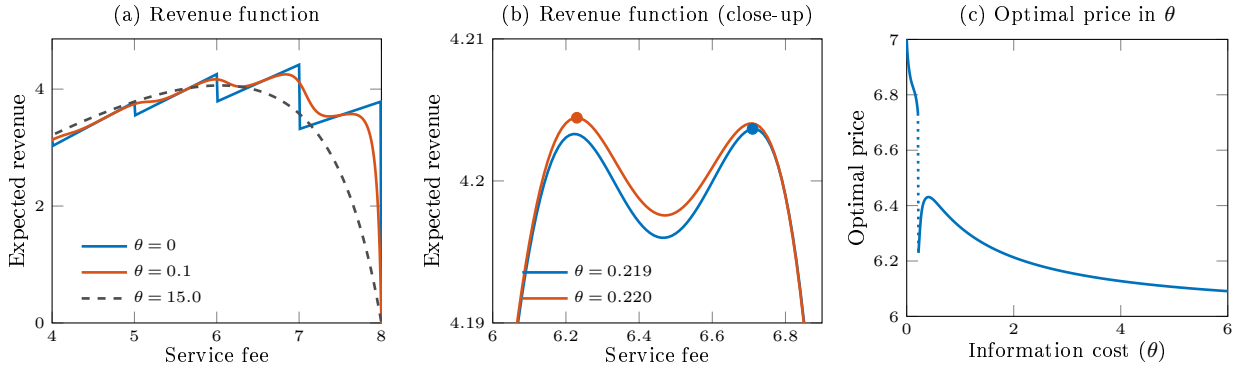

Figure 6 **Impact of information cost on revenue maximizing prices ($R = 9.0, \lambda = 0.9$)**

Despite the occasional price jumps at low θ levels, we find that the nuanced impact of information cost on customer joining behaviour and throughput remains prominent when the firm optimally sets the price. Especially when customers are still very able to discern queue length (i.e., when θ is very low), the firm takes advantage of charging as much as possible with minimal effects on customers' joining behaviour. This follows from the results on visible queues: customers' joining behaviour does not change as long as the Naor threshold $n_e = \lfloor \overline{R}\mu \rfloor$ remains the same, and hence it is optimal for the firm to charge the highest price that leads to the optimal n_e. This means that for very low θ, when the firm sets the optimal price, the resulting $\overline{R}\mu$ will be very close to its integer part (equals it when $\theta = 0$). Therefore, when θ is slightly increased, as per our throughput analysis in §4.1 (see *D-I* and *D-I-D* regions in Figure 3), there is an initial decline in the optimal price to compensate for the decreased throughput. This is also reflected on the optimal firm revenue, which faces an initial decline as θ is increased from zero. The precise evolution of the optimal price and the associated firm revenue when θ is further increased hinges on both demand and service reward. In particular, when both of them are high, customers join more often when the queue is opaque. Given this tendency, the firm can afford to increase the price and extract a premium from the customers. In contrast, when both demand is low and service is not rewarding, the firm is prompted to reduce the price and thereby moderate the customer losses due to limited attention. Nevertheless, the optimal price and revenue display similar

non-monotone behavior of the throughput in both cases (albeit in narrower ranges due to pricing), as illustrated in Figures 7a and 7b. Note that both optimal price and firm revenue exhibit the *D-I-D* pattern in the former case, while exhibiting the *D-I* pattern in the latter.

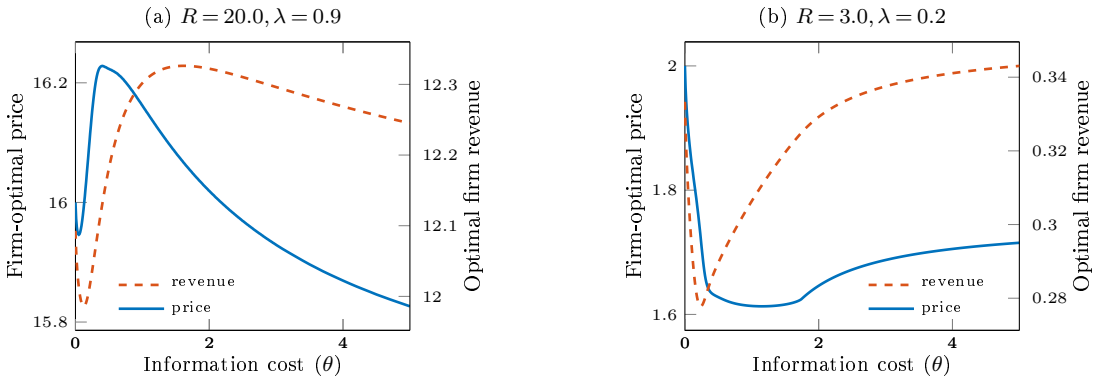

Figure 7 **Impact of information cost on revenue maximizing prices ($R = 9.0, \lambda = 0.9$)**

5.2. Social Welfare Maximization and Welfare Gaps

A social planner aims to maximize the social welfare W_s given in (14). We find that the social planner's ability to set prices has a profound impact on the social welfare. In particular, our numerical experiments reveal without exception two distinct behaviors in comparison to the case when the firm sets prices to maximize revenue. First, there is no more non-monotonicity, and the optimal social welfare is decreasing in information cost θ. In other words, information frictions caused by limited attention are always detrimental for society when the service is optimally regulated via prices. This is consistent with the well-known result that social welfare is higher in a visible queue compared to an invisible queue under socially optimal fees (Hassin and Haviv 2003). Second, optimal price no longer behaves erratically with respect to the information cost; it always increases. The social planner raises the price (hence the margin) as information cost increases to compensate for the corresponding decline in customers' surplus as established in Proposition 3.

Next, we compare the optimal prices set by the social planner with that of the revenue-maximizing firm. We know from literature that in a visible queue with $\theta = 0$, a revenue maximizer always charges customers more than what is socially optimal. Differently, in an invisible queue with $\theta = \infty$, they both charge the same price as customer surplus vanishes in equilibrium and hence revenue and social welfare functions become identical (Edelson and Hilderbrand 1975). Our model sheds light on the entire intermediate range of information costs. We find that the firm-optimal price remains higher than the socially-optimal price for any finite information cost $\theta \geq 0$. This follows the logic that a revenue maximizer only considers his individual portion of the social welfare, while a social planner has to also consider the consumer surplus which is always decreasing in price. The social planner needs to compensate for this negative impact by lowering prices from what is ideal for a revenue maximizer. As

θ approaches infinity, the two prices converge to the same value, retrieving the equivalence result for invisible queues. Figures 8a and 9a illustrate the impact of information cost on socially-optimal and firm-optimal prices for low- and high-demand systems respectively. Note that socially-optimal prices are monotonically increasing while firm-optimal prices are non-monotone. Figures 8b and 9b depict the resulting social welfare. As noted before, socially optimal welfare decreases in θ. Furthermore, optimal prices, and hence social welfare functions, converge as $\theta \to \infty$.[6]

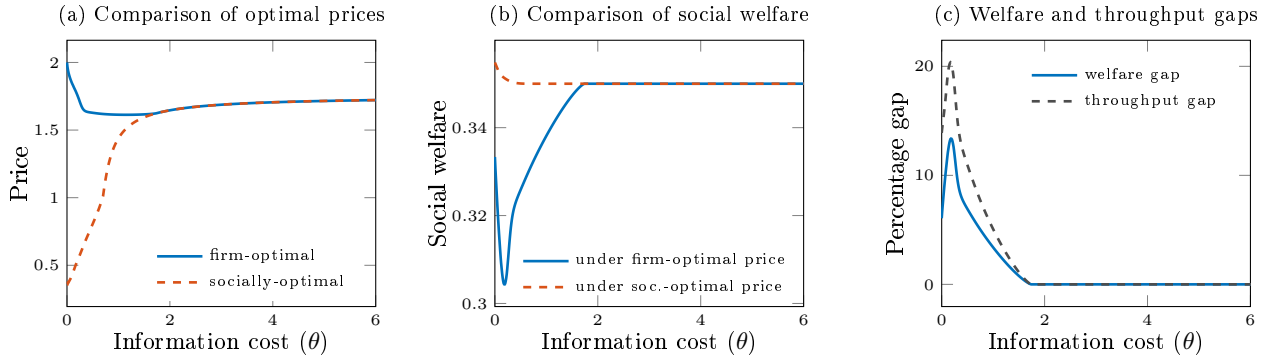

Figure 8 **Socially optimal and revenue-maximizing prices and resulting welfare gap ($R = 3.0, \lambda = 0.2, C = 1, \mu = 1$)**

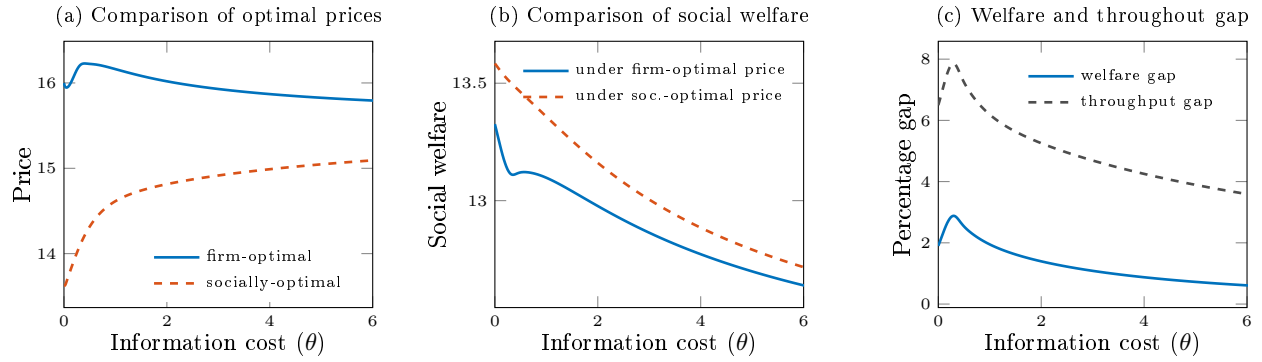

Figure 9 **Socially optimal and revenue-maximizing prices and resulting welfare gap ($R = 20, \lambda = 0.9, C = 1, \mu = 1$)**

Finally, note that when the service firm is a revenue-maximizer, the resulting social welfare is lower than the socially-optimal level by default, that is there is a natural social welfare gap. We numerically investigate the extent of this gap. Interestingly, our results unequivocally reveal that the welfare gap is highest at an intermediate information cost value. This suggests that although an opaque queue might be favorable for a revenue-maximizing firm, it causes the largest detriment on society's overall welfare. Furthermore, we find that gap is directly linked to customers' joining behaviour induced by the firm's pricing policies. This follows from the fact that the service fee is merely a transfer between the customer and the firm, so it has only an indirect effect on welfare. Even when the firm moderates

[6] To provide a more clear illustration in Figure 9a, θ is cut at 6.0. However, it is apparent that the two curves are going to converge to each other as θ is further increased.

social welfare through pricing, the real impact on welfare comes from its implications on throughput. Accordingly, welfare gap mimics the difference in effective joining probabilities and hence throughput. This is clearly observed in Figure 8c and Figure 9c.[7] Again, welfare gap starts to diminish after a certain θ value and eventually vanishes.

6. Finite Queue Capacity

In our analysis so far, we assume that any customer who decides to join the queue can effectively do so. In practice, this may not be possible when the queue has a finite waiting room/capacity. Such limits on queue capacity result in throughput losses in standard queueing models. Consider now a queue with finite waiting capacity N, serving rationally inattentive customers. This means that an arriving customer is not admitted to the system when it is full, and to reflect this inconvenience, let there be a rejection cost $T \geq 0$. The state space (total number of customers) is then $\{0,1,..,N\}$, with payoffs v_n given as in our baseline model by (1), except for state N, $v_N = -T$. The equilibrium of the finite capacity queueing system can be defined analogous to Definition 1, with the only change being the limited state space $\{0,1,..,N\}$. Note that this framework can be readily adapted for a service system with multiple servers. In both cases, it can be verified that a unique equilibrium strategy exists.

The baseline model we analyzed corresponds to the limiting case with $N \to \infty$. In order to shed light on the impact of finite queue capacity, we elaborate on the other limiting scenario with zero waiting capacity, $N = 1$. This implies that arriving customers cannot join when the server is busy. In this model, customers aim to learn whether the server is busy or not in order to make a "joining" decision.

THEOREM 2. *In a queueing system with no waiting capacity and rationally inattentive customers with information cost $\theta > 0$, the unique equilibrium unconditional joining (trying) probability $\widetilde{\pi}$ is*

$$\widetilde{\pi} = \min\left\{\frac{\left(e^{v_0/\theta}-1\right)}{\left(1-e^{v_1/\theta}\right)\left(\rho e^{v_0/\theta}+e^{v_0/\theta}-1\right)}, 1\right\}. \tag{15}$$

The conditional joining probabilities in equilibrium are $\widetilde{\pi}_n = \left(\widetilde{\pi} e^{v_n/\theta}\right) / \left(\widetilde{\pi} e^{v_n/\theta} + 1 - \widetilde{\pi}\right)$, for $n \in \{0,1\}$ and the equilibrium steady-state probability of the server being idle is $\widetilde{g}_0 = \left(1+\rho\widetilde{\pi}_0\right)^{-1}$.

Throughput in equilibrium is defined as $\lambda \sum_{n=0}^{N-1} \widetilde{\pi}_n \widetilde{g}_n$ for a queue with capacity N. It is clear that throughput is *not* proportional to the equilibrium joining fraction since there is no entrance in state N. The next proposition characterizes the effect of information cost on both equilibrium joining probability and throughput for the zero-capacity case.

PROPOSITION 4. *(i) If $R - p - \frac{C}{\mu} \geq T$, the equilibrium joining fraction $\widetilde{\pi}$ is increasing in information cost θ. Otherwise, $\widetilde{\pi}$ takes its maximum value when $\theta = 0$.*

[7] Percentage welfare (resp., throughput) gap is the difference in welfare (resp., throughput) under socially-optimal and revenue-maximizing prices normalized by the welfare (resp., throughput) under socially-optimal price.

(ii) Throughput takes its maximum value when $\theta = 0$.

Proposition 4 states that in a service system with no waiting room, the firm should make the system completely visible to maximize throughput, regardless of the rejection cost to the customers. The rationale is that obstructing information acquisition for customers can only deter them joining the system when it is empty, which is definitely undesirable for the firm. At the other extreme with infinite waiting room, we already know that throughput may exhibit non-monotone behavior. Putting these together, it is evident that queue capacity can be an important design consideration when customers have limited attention. This is corroborated in Figure 10a, which depicts the impact of information costs on throughput for different queue capacity levels ($N = 20$ is practically identical to our baseline case). Most notably, this extended framework elucidates the possibility for the firm to garner throughput *gains* from *limiting* waiting room capacity. As illustrated in Figure 10b, throughput may be maximized at an intermediate, finite waiting room capacity for a given information cost level θ. This might explain why some high-end restaurants do not increase their capacity and prefer taking reservations and asking people to wait.

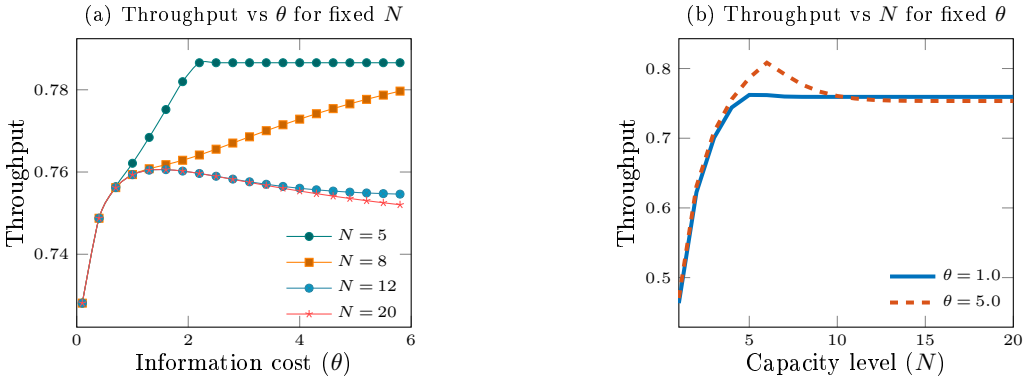

Figure 10 **Impact of finite capacity level on customer behaviour** ($R = 3.8, p = 0, C = 1, \mu = 1, \lambda = 0.9, T = 2$)

7. Learning Beyond the Queue Length

In our baseline framework, rationally inattentive customers aim to learn only the uncertain queue length. In practice, however, there could be other aspects of the service environment that are also not easily discernible by customers, such as service speed (rate) and reward (quality), and customers may find it desirable to obtain information about these elements as well. Furthermore, it may even be the case that queue length is fully revealed at no cost (resp., obstructed), but customers spend time and attention to learn service rate and/or quality. In this section, we incorporate such multi-dimensional learning of the service environment into our framework. We provide equilibrium definitions and results for each case, and reveal insights regarding customers' strategic queueing behaviour.

7.1. Learning Queue Length, Service Rate and/or Quality

Consider the case where customers optimally allocate their attention to *simultaneously* learn about queue length, service rate (μ) and/or reward (R). Without loss of generality assume that R and μ are discrete random variables with finite support and state space $\Omega_{R,\mu}$ with customers' common prior $h_{R,\mu}$. State of the service system is then the triplet $\Omega = \{(n, \omega_R, \omega_\mu) : \in \mathbb{N} \times \Omega_{R,\mu}\}$ with common prior h_Ω. Let $\widetilde{\pi}_n(\omega_R, \omega_\mu)$ denote the conditional joining probability in equilibrium when service reward is ω_R, service rate is ω_μ, and there are n customers in the system. Similarly, let $\widetilde{G}(\omega_R, \omega_\mu) = \{\widetilde{g}_n(\omega_R, \omega_\mu); n \geq 0\}$ denote the conditional steady-state queue length distribution in equilibrium. State-dependent utility of joining is

$$v_n(\omega_R, \omega_\mu) = \omega_R - p - \frac{C}{\omega_\mu}(n+1)$$

and value of balking is normalized to zero. For a given prior h_Ω, rationally inattentive customers solve the extended version of the optimization problem in (9):

$$\max_{\overline{\pi} \in [0,1]} \left[\theta \sum_{\Omega} h_\Omega(n, \omega_R, \omega_\mu) \log\left(\overline{\pi} e^{v_n(\omega_R, \omega_\mu)/\theta} + 1 - \overline{\pi} \right) \right]$$

to arrive at the optimal unconditional joining probability. The difference is that the expectation is now taken with respect to the joint distribution h_Ω. Let $\rho_{\omega_\mu} = \lambda/\omega_\mu$ denote the utilization parameter given service rate ω_μ.

Definition 2. In a queueing system with rationally inattentive customers with information cost $\theta > 0$, the equilibrium probability of joining the queue is

$$\widetilde{\pi}_n(\omega_R, \omega_\mu) = \frac{\widetilde{\pi} e^{v_n(\omega_R, \omega_\mu)/\theta}}{1 - \widetilde{\pi} + \widetilde{\pi} e^{v_n(\omega_R, \omega_\mu)/\theta}}$$

where

$$\widetilde{\pi} = \sum_{\Omega} \widetilde{\pi}_n(\omega_R, \omega_\mu) \tilde{g}_n(\omega_R, \omega_\mu) h_{R,\mu}(\omega_R, \omega_\mu)$$

is the unconditional joining probability in equilibrium. Conditional queue length distribution is

$$\begin{aligned} \tilde{g}_0(\omega_R, \omega_\mu) &= \left(1 + \sum_{k=1}^{\infty} \rho_{\omega_\mu}^k \widetilde{\pi}_0(\omega_R, \omega_\mu) ... \widetilde{\pi}_{k-1}(\omega_R, \omega_\mu) \right)^{-1} \\ \tilde{g}_n(\omega_R, \omega_\mu) &= \tilde{g}_0(\omega_R, \omega_\mu) \rho_{\omega_\mu}^n \widetilde{\pi}_0(\omega_R, \omega_\mu) ... \widetilde{\pi}_{n-1}(\omega_R, \omega_\mu) \text{ for } n \geq 1. \end{aligned} \tag{16}$$

Theorem 3. *There exists a unique equilibrium satisfying Definition 2.*

Theorem 3 generalizes our baseline model results, establishing its validity for a multi-dimensional learning environment. This testifies to the versatility of our framework, and signifies its potential use as a design tool in a variety of different service settings. Note that an implicit assumption here is that the information cost is the same for each uncertain element of the environment the customer is learning about. It is plausible that some aspects of the service system may be easier to learn than others. In what follows, we explore two special cases with this flavor.

7.2. Invisible Queues: Learning Service Rate and/or Quality

Suppose now that queue is invisible and rationally inattentive customers learn the service rate and/or service reward. Clearly, a customer's payoff for joining in any state depends on the strategy of other customers (as it impacts the expected waiting time). In particular, for given realizations ω_R, ω_μ, and conditional joining strategy of other customers $\pi(\omega_R,\omega_\mu)$, expected net payoff of joining is

$$v(\omega_R,\omega_\mu) = \omega_R - p - CW(\omega_R,\omega_\mu)$$

where

$$W(\omega_R,\omega_\mu) = \frac{1}{(\omega_\mu - \lambda\pi(\omega_R,\omega_\mu))^+}$$

is the expected waiting time in the system. Anticipating $\pi(\omega_R,\omega_\mu)$, customers optimize their information processing strategy and act accordingly.

Definition 3. In an invisible queueing system with rationally inattentive customers with information cost $\theta > 0$, the conditional probability of joining in equilibrium satisfies

$$\widetilde{\pi}(\omega_R,\omega_\mu) = \frac{\widetilde{\pi} e^{v(\omega_R,\omega_\mu)/\theta}}{\widetilde{\pi} e^{v(\omega_R,\omega_\mu)/\theta} + 1 - \widetilde{\pi}}$$

where

$$\widetilde{\pi} = \sum_{\Omega_{R,\mu}} \widetilde{\pi}(\omega_R,\omega_\mu) h_{R,\mu}(\omega_R,\omega_\mu).$$

is the unconditional joining probability.

Theorem 4. *There exists a unique equilibrium that satisfies Definition 3.*

The impact of limited attention and associated information costs on the equilibrium performance of invisible queues depend critically on whether the customer is acquiring information on service reward, service rate, or both. When service reward is known and customers aim to learn only the service rate μ, we find that throughput is monotonically increasing (resp., decreasing) in information cost if demand is low (resp., high) enough. This is rather intuitive. When customers anticipate low congestion (due to low demand), they are inclined to join the queue, but as they start to distinguish realized service rates, especially slow ones, they may choose to balk more often. The opposite occurs when demand is sufficiently high. In the moderate demand regime, however, throughput is unimodal, implying again that some degree of information frictions (regarding service speed) is beneficial for the firm.

When customers know the service rate but are uncertain about service reward (quality) R, the level of demand, and hence congestion, in the system does not play a significant role in throughput behavior. We observe that, in general, throughput is monotonically increasing in information cost. Notice that as information cost decreases, customers are able to discern the actual service quality better, leading to higher (resp., lower) joining rates for high (resp., low) reward states. Interestingly, the combined

effect is typically such that having better-informed customers is detrimental for throughput. The only exception we find to this is when expected reward (quality) is very low. In such cases, having customers who are only partially informed (intermediate θ) can yield higher throughput (i.e., throughput is unimodal in θ). We remark that these insights are in stark contrast with those in Ren et al. (2018), where customers in an invisible queue estimate *expected* service reward by sampling anecdotes from earlier customers. They find that customers sampling more (and obtaining more information) improves (resp., deteriorates) throughput for high-quality (resp., low-quality) services. In their model, customers are most informed when they can perfectly estimate the expected reward. This "rational" benchmark corresponds to the case when $\theta \to \infty$ in our model. What we add is, when customers learn beyond the mean and start distinguishing the realized service quality (as θ decreases), the accretion of customer information has the opposite impact on throughput.

When both service reward and service rate are uncertain, the impact of learning service rate is reflected more on throughput. As illustrated in Figure 11, the implied throughput when customers acquire information on the service rate (service reward known) is in line with the case where customers aim to learn both. As such, for high demand systems, disclosing full service rate information and obstructing the acquisition of service reward information yields the highest throughput for the firm.

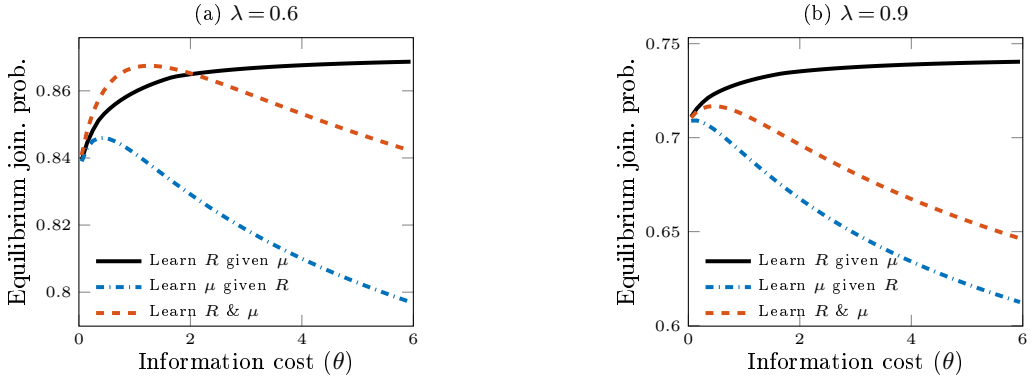

Figure 11 **Impact of learning service rate and/or quality in invisible queues** ($C = 1, p = 0, R \sim U(1.8, 5.8), \mu \sim U(0.5, 1.5)$)

7.3. Visible Queues: Learning Service Rate and/or Quality

Finally, consider a visible queue setting. Note that when queue length is freely observable, it is possible for customers to infer the distribution of service reward and/or service rate, even before engaging in information acquisition. Indeed, regardless of customers' prior beliefs, a long queue is more likely due to high service reward and/or low service rate. In particular, given the (correct) prior $\widetilde{G}$ on queue length (as in 16), customers can construct Bayesian posterior beliefs after observing the actual queue length:

$$h_{R,\mu|Q}\left(\omega_R, \omega_\mu | n\right) = \frac{\widetilde{g}_n(\omega_R, \omega_\mu) h_{R,\mu}(\omega_R, \omega_\mu)}{\sum_{\Omega_{R,\mu}} \widetilde{g}_n(\omega_R, \omega_\mu) h_{R,\mu}(\omega_R, \omega_\mu)}. \tag{17}$$

Then, customers process information only to reduce the uncertainty on their "prior" $h_{R,\mu|Q}$.

Naturally, when the queue is visible, for certain queue lengths, customers do not need to process information at all. Let $\overline{n} = \left\lfloor \frac{(\overline{\omega}_R - p)\overline{\omega}_\mu}{C} \right\rfloor$ and $\underline{n} = \left\lfloor \frac{(\underline{\omega}_R - p)\underline{\omega}_\mu}{C} \right\rfloor$ denote the maximum and minimum "Naor" thresholds based on the state space $\Omega_{R,\mu}$. Clearly, customers join with probability 1 if queue length is strictly less than $\underline{n}$ since the payoff of joining is positive even if service reward and rate are realized at their lowest values. Similarly, customers balk with certainty if queue length is strictly greater than $\overline{n}$.

Definition 4. In a visible queueing system with rationally inattentive customers with information cost $\theta > 0$, the equilibrium conditional probability of joining is

$$\widetilde{\pi}_n(\omega_R, \omega_\mu) = \frac{\widetilde{\pi}_n e^{v_n(\omega_R,\omega_\mu)/\theta}}{\widetilde{\pi}_n e^{v_n(\omega_R,\omega_\mu)/\theta} + 1 - \widetilde{\pi}_n} \text{ for } \underline{n} \le n \le \overline{n}$$

where

$$\widetilde{\pi}_n = \sum_{\Omega_{R,\mu}} \widetilde{\pi}_n(\omega_R, \omega_\mu) h_{R,\mu|Q}\left(\omega_R, \omega_\mu | n\right)$$

is the unconditional joining probability when there are n customers in the system. $\widetilde{\pi}_n(\omega_R, \omega_\mu) = 1$ if $n \le \underline{n} - 1$ and $\widetilde{\pi}_n(\omega_R, \omega_\mu) = 0$ if $n \ge \overline{n} + 1$.

Definition 4 is quite different than previous equilibrium definitions, and as such, the underlying equilibrium model. In previous cases, it is sufficient for a customer to anticipate only the unconditional joining fraction ($\widetilde{\pi}$), as it uniquely determines conditional joining fractions, which, in turn, uniquely define the queue length distribution. Therefore, a simple fixed point search on a single variable is sufficient to find the unique equilibrium. In the visible queue case, however, customers need to anticipate unconditional joining fractions for each queue length state (between $\underline{n}$ and $\overline{n}$) to form the correct belief in (17). Accordingly, finding the equilibrium requires solving a system of nonlinear fixed point equations. Due to this, only the existence of an equilibrium is guaranteed.

Theorem 5. *There exists an equilibrium that satisfies Definition 4.*

In a visible queue, customer behavior is more convoluted, and hence it is hard to draw unequivocal throughput and welfare conclusions. Nevertheless, the construct is still instrumental in generating new insights through comparative analysis. For instance, it is possible to compare equilibria when i) both service rate and queue length are uncertain, ii) only service rate is uncertain, and iii) only queue length is uncertain. Such a comparison would shed light on whether it is better for the firm to provide information on service speed or queue length. We give a flavor of these insights in Figure 12. First, observe that "knowing more" (i.e., less uncertainty) does not necessarily lead to higher throughput. Also, we consistently find that throughput is higher when customers learn queue length instead of service speed, implying that providing visibility on service speed may be more effective for the firm.

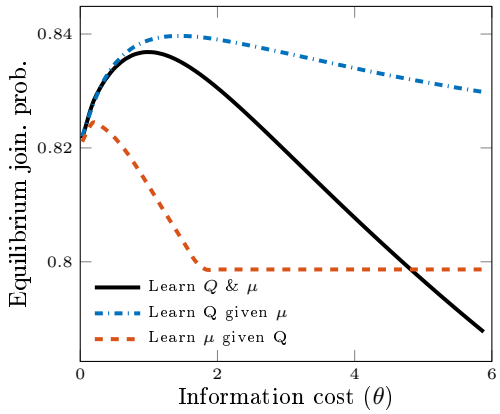

Figure 12 **Impact of learning service speed and/or queue length** ($R = 5.8, p = 0, C = 1, \lambda = 0.9, \Omega_\mu = \{0.5, 0.6, ..., 1.5\}$)

8. Concluding Remarks

Limited attention is ubiquitous and learning is costly. In a queueing system, customers have to spend time and cognitive resources to determine uncertain aspects of the environment, estimate associated delays, and translate this information into decisions. As information is costly, rational customers need to trade off the benefits of information against the costs and have to make joining decisions based on partial information. In this paper, we propose a framework that integrates these salient features. At the core of our framework is rationally inattentive choice, linking customer beliefs, service rewards, and information costs. We incorporate this into different strategic queueing models, and establish existence and uniqueness of an equilibrium for each case. Utilizing this framework, we provide a unified perspective and a comprehensive view on the effect of information cost (information prevalence) on throughput, pricing and social welfare. Instead of replicating the descriptive results here, it is perhaps better to summarize the managerial prescriptions they translate into.

When inattentive customers acquire information about uncertain queue length and expected delay, our framework naturally connects the canonical visible and invisible queues studied in the extant literature. In this case, our results suggest that firms should be most cautious about customers' limited attention and their information provision strategies when attractiveness of the service is not excessively high or low, but in a moderate regime. In particular, when customers value a service reasonably well and there is robust demand for it, service firms should intentionally leave some uncertainty around queue length, but not completely obstruct the information acquisition process. In this sense, Disney's deployment of special layouts that partially disguise queue length prevails as a reasonable practice. If firm profits are relatively more significant compared to customer surplus, then this might even be beneficial from a total social welfare perspective. In stark contrast, for less congested firms offering a service that is not highly valued, partial hindrance of information acquisition is precisely what the firm should try to avoid. To the best of our knowledge, this has not been identified and noted in the extant literature. It is therefore in the interest of a public service office with little congestion or

a drive-through fast-food restaurant to deploy a completely visible queueing system. It may even be better to completely obstruct the observation of the queue (if the physical environment allows).

These qualitative insights remain valid when the service is moderated by prices by a revenue-maximizing firm. The resulting social welfare, however, is lower than the optimal social welfare that can be generated by a social planner. Interestingly, even though a revenue-maximizing service firm may still want to partially obstruct the observation of the queue length, the opaque queue that ensues from this choice may result in the greatest social welfare losses.

Throughput benefits of limited attention and information frictions extend beyond the case where customers aim to discern queue length upon arrival. It may also hold true when customers acquire information on multiple dimensions of the service environment, as well as when the waiting room is capacitated. Even when the queue is invisible (e.g., as in a call center), providing customers some assistance on delineating service speed or service quality might be beneficial. Our framework can identify required conditions perspicuously, making it a useful tool for service design.

References

Allon G, Bassamboo A (2011) The impact of delaying the delay announcements. *Operations research* 59(5):1198–1210.

Ang E, Kwasnick S, Bayati M, Plambeck EL, Aratow M (2015) Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management* 18(1):141–156.

Boyacı T, Akçay Y (2017) Pricing when customers have limited attention. *Management Science* 64(7):2973–3468.

Caplin A, Dean M, Leahy J (2016) Rational inattention, optimal consideration sets and stochastic choice, NYU working paper.

Chen H, Frank M (2004) Monopoly pricing when customers queue. *IIE Transactions* 36(6):569–581.

Cover TM, Thomas JA (2012) *Elements of information theory* (John Wiley & Sons, Hoboken, NJ).

Cui S, Li K, Wang J (2017) On the optimal disclosure of queue length information. *Available at SSRN* .

Edelson NM, Hilderbrand DK (1975) Congestion tolls for poisson queuing processes. *Econometrica: Journal of the Econometric Society* 43(1):81–92.

Hassin R (2016) *Rational queueing* (CRC press, Boca Raton, FL).

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

Hassin R, Roet-Green R (2017) The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Operations Research* 65(3):804–820.

Hu M, Li Y, Wang J (2017) Efficient ignorance: Information heterogeneity in a queue. *Management Science* 64(6):2473–2972.

Huang T, Allon G, Bassamboo A (2013) Bounded rationality in service systems. *Manufacturing & Service Operations Management* 15(2):263–279.

Huang T, Chen YJ (2015) Service systems with experience-based anecdotal reasoning customers. *Production and Operations Management* 24(5):778–790.

Hüttner F, Boyacı T, Akçay Y (2019) Consumer choice under limited attention when alternatives have different information costs. *Operations Research* Forthcoming.

Ibrahim R (2018) Sharing delay information in service systems: a literature survey. *Queueing Systems* 89(1-2):49–79.

Maćkowiak B, Matějka F, Wiederholt M (2018) Rational inattention: A disciplined behavioral model, CERGE-EI working paper.

Matějka F (2015) Rationally inattentive seller: Sales and discrete pricing. *The Review of Economic Studies* 83(3):1125–1155.

Matějka F, McKay A (2015) Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review* 105(1):272–98.

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

Ren H, Huang T, Arifoglu K (2018) Managing service systems with unknown quality and customer anecdotal reasoning. *Production and Operations Management* 27(6):1038–1051.

Sallee JM (2014) Rational inattention and energy efficiency. *The Journal of Law and Economics* 57(3):781–820.

Simhon E, Hayel Y, Starobinski D, Zhu Q (2016) Optimal information disclosure policies in strategic queueing games. *Operations Research Letters* 44(1):109–113.

Sims CA (2003) Implications of rational inattention. *Journal of monetary Economics* 50(3):665–690.

Sims CA (2010) Rational inattention and monetary economics. *Handbook of monetary economics*, volume 3, 155–181 (Elsevier).

Yu Q, Allon G, Bassamboo A (2016) How do delay announcements shape customer behavior? an empirical study. *Management Science* 63(1):1–20.

Appendix

Proof of Lemma 1 Note that both equations in (10) can be obtained by a direct application of KKT conditions, as in Proposition 1 in Caplin et al. 2016) which provides necessary and sufficient conditions for a general rationally inattentive discrete choice problem with a given prior. Plugging $\overline{\pi}=0$ and $\overline{\pi}=1$ in the first and second expressions in (10), we obtain (11). $\square$

Proof of Theorem 1 We first show that the equilibrium in Definition 1 is stable. Assume that $\widetilde{\pi}<1$ and $\theta\in[0,\infty)$. Then note that steady-state queue distribution is well-defined as the series $\sum_{k=1}^{\infty}\rho^{k}\widetilde{\pi}_{0}\widetilde{\pi}_{1}...\widetilde{\pi}_{k-1}=\sum_{k=1}^{\infty}a_{k}$ converges. By ratio test, note that $\lim_{k\to\infty}\left|\frac{a_{k+1}}{a_{k}}\right|=\lim_{k\to\infty}|\rho\widetilde{\pi}_{k}|=\lim_{k\to\infty}\left|\rho\frac{\widetilde{\pi}e^{v_{k}/\theta}}{1-\widetilde{\pi}+\widetilde{\pi}e^{v_{k}/\theta}}\right|=0$ as $v_{k}\to-\infty$ as $k\to\infty$. This is true for any finite $\lambda>0$. Note that when $\lambda\geq\mu$, equilibrium joining probability can not be 1 as expected waiting time for a customer is infinite. When $\theta=\infty$, the series is convergent if $\lambda<\mu$. Now we prove the existence and uniqueness.

Existence: Let us define conditional joining probability and prior queue length distribution as a function of other customers' unconditional joining probability q:

$$\pi_{n}(q)=\frac{qe^{v_{n}/\theta}}{1-q+qe^{v_{n}/\theta}} \tag{18}$$

$$g_{0}(q)=\left(1+\sum_{k=1}^{\infty}\rho^{k}\pi_{0}(q)\,\pi_{1}(q)...\pi_{k-1}(q)\right)^{-1},$$

$$g_{n}(q)=g_{0}(q)\rho^{n}\pi_{0}(q)\,\pi_{1}(q)...\pi_{n-1}(q)\quad\text{for }n\geq1. \tag{19}$$

Our aim is to show that there exists at least a point satisfying $\overline{\pi}(q)=q$ where

$$\overline{\pi}(q):=\underset{\pi\in[0,1]}{\arg\max}\theta\sum_{n\geq0}g_{n}(q)\log\left(\pi e^{v_{n}/\theta}+1-\pi\right)=\underset{\pi\in[0,1]}{\arg\max}f(\pi;q) \tag{20}$$

Let $h(q)=\overline{\pi}(q)-q$. Note that $\pi_{n}(0)=0$ for all $n\geq0$ and consequently $g_{0}(0)=1$ and $g_{n}(0)=0$ for all $n\geq1$. By (20), it is clear that $\overline{\pi}(0)=1$ since $v_{0}>0$ by assumption. This gives $h(0)=\overline{\pi}(0)=1>0$, which is strictly positive. On the other hand, $h(1)=\overline{\pi}(1)-1\leq0$ since $\overline{\pi}(1)\in[0,1]$. Since $h(q)$ is continuous in $[0,1]$, by the intermediate value theorem, there exists a point $q\in[0,1]$ satisfying $h(q)=0$.

Uniqueness: To prove uniqueness, we show that $h(q)$ is decreasing in q in the interval $[0,1]$. Note that for given $q\in[0,1]$, f is concave in π. As shown in Matějka and McKay (2015), there always exists a solution to (20) and if the vectors $e^{v_{n}/\theta}$ are linearly independent the solution is unique, which is exactly our case since v_{n} is strictly monotone in n. Taking partial derivative of f with respect to π gives $\frac{\partial}{\partial\pi}f(\pi;q)=\theta\sum_{n}g_{n}(q)\,\varphi_{n}(\pi;\theta)=\theta E_{G(q)}\varphi_{Q}(\pi)$ where $\varphi_{Q}(\pi)=\frac{e^{v_{Q}/\theta}-1}{\pi e^{v_{Q}/\theta}+1-\pi}$ and $Q\in\mathbb{N}$ is the uncertain queue length. Note that $\varphi_{Q}(\pi)$ is decreasing in Q. Since $Q(q)$ (with distribution $G(q)$) is stochastically increasing in q by Lemma 2 and $\varphi_{Q}(\pi)$ is decreasing in Q, $E_{G(q)}\varphi_{Q}(\pi)$ is also decreasing in q. This means $\bar{\pi}(q)$, the maximizer of (20), is also decreasing in q. Therefore, there is a unique $q^{*}\in[0,1]$ that satisfies $\overline{\pi}(q^{*})=q^{*}$. $\square$

LEMMA 2. *$Q(q,\theta)$ is stochastically increasing in q for given θ.*

Proof. Let $\theta > 0$ be fixed. Then note that $\pi_n(q,\theta)$ is increasing in q for any $n \geq 0$ and consequently $g_0(q,\theta)$ is strictly decreasing in q. Additionally, if $g_n(q,\theta)$ is increasing in q for any $n \geq 1$, $g_k(q,\theta)$ is also increasing in q for all $k \geq n$ due to the recursive relationship between steady-steady probabilities of the queueing system, i.e., $g_{n+1}(q,\theta) = \rho g_n(q,\theta)\pi_n(q,\theta)$. Let $\widehat{k} \geq 1$ be the first integer such that $g_{\widehat{k}}(q,\theta)$ is increasing in q. Then $g_n(q,\theta)$ is decreasing in q for all $n \leq \widehat{k}-1$ and increasing in q for all $n \geq \widehat{k}$. The existence of $\widehat{k}$ is guaranteed as $g_0(q,\theta)$ is strictly decreasing and $\sum_{n=0}^{\infty} g_n(q,\theta) = 1$. Let us fix $n \geq 0$. If $n \leq \widehat{k}$, then $P(Q \geq n) = 1 - \sum_{i=0}^{n-1} g_i(q,\theta)$ is increasing in q. Similarly, if $n \geq \widehat{k}+1$, $P(Q \geq n) = \sum_{i=n}^{\infty} g_i(q,\theta)$ *is* increasing n. Thus, $Q(q,\theta)$ is stochastically increasing in q. □

Proof of Corollary 1

1. By (12), as $\theta \to 0$, $\pi_n \to 1$ if $v_n > 0$ and $\pi_n \to 0$ if $v_n < 0$. This is precisely the visible queue model of Naor (1969) which results in a capacitated $M/M/1/n_e$ system. Hence, $\widetilde{\pi} = (1-\rho^{n_e})\left(1-\rho^{n_e+1}\right)^{-1}$.

2. As $\theta \to \infty$, (12) implies that in equilibrium $\widetilde{\pi}_n = \widetilde{\pi}$ for all $n \geq 0$ and by (6), the problem is the same as in that of invisible queues. More specifically, let us assume that other customers join at each state with probability q. Then, a tagged customer's optimization problem in (6) reduces to $\max_{\pi} \pi\left(R - p - C/(\mu - \lambda q)\right)$ where $1/(\mu-\lambda q)$ is the expected waiting time for an $M/M/1$ system with arrival rate λq. In this case, the unique equilibrium is the value of q that makes the expected utility zero, i.e., $(R-p-C/(\mu-\lambda q)) = 0$, hence $\pi = q$. This is precisely the equilibrium in invisible queues. Furthermore, if $(R-p-C/(\mu-\lambda)) > 0$, then clearly $\pi = 1$ and the result follows. □

Proof of Proposition 1 We prove this by contradiction. Let $R_1 < R_2$ and assume that there exists a $k \geq 0$ such that $\widetilde{\pi}_k(R_1) > \widetilde{\pi}_k(R_2)$. Then, it is easy to verify that $\widetilde{\pi}_n(R_1) > \widetilde{\pi}_n(R_2)$ for all $n \geq 0$ and consequently $\widetilde{G}(R_1) >_{st} \widetilde{G}(R_2)$ since arrival rates at all states are greater when service reward is R_1. Then it follows that $\widetilde{\pi}(R_1) > \widetilde{\pi}(R_2)$ since $E_{\widetilde{G}(R_1)}\widetilde{\pi}_Q(R_1) > E_{\widetilde{G}(R_2)}\widetilde{\pi}_Q(R_2)$ due to the stochastic monotonicity and $\widetilde{\pi}_Q$ being decreasing in Q. Let $\widetilde{\varphi}_Q(R) = \frac{e^{vQ(R)/\theta}-1}{\widetilde{\pi}(R)e^{vQ(R)/\theta}+1-\widetilde{\pi}(R)}$. Then note that, $E_{\widetilde{G}(R_1)}\widetilde{\varphi}_Q(R_1) = 0$ due to the optimality of $\widetilde{\pi}(R_1)$ (which is actually the equilibrium and satisfies the first order condition). Since $\widetilde{\varphi}_Q(R)$ is decreasing in $\widetilde{\pi}(R)$ and increasing in R, we have $0 = E_{\widetilde{G}(R_1)}\widetilde{\varphi}_Q(R_1) < E_{\widetilde{G}(R_1)}\widetilde{\varphi}_Q(R_2)$. Lastly, since $\widetilde{\varphi}_Q(R)$ is decreasing in Q it follows that $0 < E_{\widetilde{G}(R_1)}\widetilde{\varphi}_Q(R_2) < E_{\widetilde{G}(R_2)}\widetilde{\varphi}_Q(R_2)$ which is a contradiction since $\widetilde{\pi}(R_2)$ is the equilibrium when service reward is R_2 and $E_{\widetilde{G}(R_2)}\widetilde{\varphi}_Q(R_2) = 0$. This means $\widetilde{\pi}_n(R_1) < \widetilde{\pi}_n(R_2)$ for all $n \geq 0$. It is still unclear whether this result implies $\widetilde{\pi}(R_1) < \widetilde{\pi}(R_2)$. To show this we use the rate-in equals rate-out principle for birth and death processes. Note that in equilibrium $\lambda\widetilde{\pi}(R) = \mu(1-\widetilde{g}_0(R))$ must be satisfied. Since $\widetilde{\pi}_n(R_1) < \widetilde{\pi}_n(R_2)$ for all $n \geq 0$, $\widetilde{g}_0(R_1) > \widetilde{g}_0(R_2)$ by (13). Then it must be that $\widetilde{\pi}(R_1) < \widetilde{\pi}(R_2)$. The same proof arguments follow for the other parameters.

Proof of Proposition 2 Since $\widetilde{\pi}$ is an equilibrium, it satisfies the fixed point equation $\widetilde{\pi}(\theta) = \overline{\pi}(\widetilde{\pi}^{(\theta)}, \theta)$. Using implicit differentiation on $\bar{\pi}$, we have

$$\frac{d}{d\theta}\widetilde{\pi}^{(\theta)} = \frac{\partial \overline{\pi}\left(q,\theta\right)/\partial\theta_{|q=\widetilde{\pi}(\theta)}}{1-\partial\overline{\pi}\left(q,\theta\right)/\partial q_{|q=\widetilde{\pi}^{(\theta)}}}. \tag{21}$$

Note that $\partial\overline{\pi}(q,\theta)/\partial q_{|q=\widetilde{\pi}^{\theta}} < 0$ (check proof of Theorem 1) and consequently the denominator is positive. Therefore, it is enough to look at sign of the numerator. Let $f(\pi, q, \theta)$ be the objective function in (20) and $f_x(x,q,\theta) = E_G(q,\theta)\frac{e^{vQ/\theta}-1}{xe^{vQ/\theta}+1-x}$ denote the partial derivative with respect to x. Using implicit differentiation on $f_x\left(\overline{\pi}\left(q,\theta\right),q,\theta\right) = 0$, we have

$$\frac{\partial}{\partial\theta}\overline{\pi}\left(q,\theta\right) = -\frac{\partial f_x\left(x,q,\theta\right)/\partial\theta_{|x=\overline{\pi}(q,\theta)}}{\partial f_x\left(x,q,\theta\right)/\partial x_{|x=\overline{\pi}(q,\theta)}}.$$

Note that, since f is concave in x, we have $\partial f_x\left(x,q,\theta\right)/\partial x < 0$ and the sign of $\frac{\partial}{\partial\theta}\overline{\pi}\left(q,\theta\right)$ is the same as the sign of $\partial f_x\left(x,q,\theta\right)/\partial\theta_{|x=\overline{\pi}(q,\theta)}$ which can be written as

$$\frac{\partial f_x(x,q,\theta)}{\partial\theta}_{|x=\overline{\pi}(q,\theta),q=\widetilde{\pi}^{(\theta)}} = \frac{\partial}{\partial\theta}E_{G(q,\theta)}\left(\frac{e^{vQ/\theta}-1}{qe^{vQ/\theta}+1-q}\right)_{|q=\widetilde{\pi}^{(\theta)}} = \lim_{h\to 0}\frac{E_{G(\widetilde{\pi}^{(\theta)},\theta+h)}\left(\frac{e^{vQ/(\theta+h)}-1}{\widetilde{\pi}^{(\theta)}e^{vQ/(\theta+h)}+1-\widetilde{\pi}^{(\theta)}}\right)}{h}$$

since $f_x\left(\widetilde{\pi}^{(\theta)},\widetilde{\pi}^{(\theta)},\theta\right) = 0$. Let us look at the sign of this limit for a fixed θ. Assuming $\lambda < \mu$, the case where $\widetilde{\pi}^{(\theta)} = 1$ induces stochastically the largest distribution among all possible prior distributions (which yields an $M/M/1$-type queue). Then we have

$$1-E_{G_{M/M/1}}\left(\frac{1}{e^{v_n/(\theta+h)}}\right) \le E_{G_{M/M/1}}\left(\frac{e^{v_n/(\theta+h)}-1}{\widetilde{\pi}^{(\theta)}e^{v_n/(\theta+h)}+1-\widetilde{\pi}^{(\theta)}}\right) \le E_{G(\widetilde{\pi}^{(\theta)},\theta+h)}\left(\frac{e^{v_n/(\theta+h)}-1}{\widetilde{\pi}^{(\theta)}e^{v_n/(\theta+h)}+1-\widetilde{\pi}^{(\theta)}}\right)$$

where the first inequality is due to the expression inside the expectation being decreasing in $\widetilde{\pi}^{(\theta)}$, and the second inequality is due to $M/M/1$ being stochastically greater than $G(\widetilde{\pi}^{(\theta)},\theta+h)$ for any θ. Clearly, there exists an $R_H(\theta) > p + C/\mu$ value such that $E_{G_{M/M/1}}\left(\frac{1}{e^{v_n/(\theta+h)}}\right) \le 1$ for all $R > R_H(\theta)$ and consequently $E_{G(\widetilde{\pi}^{(\theta)},\theta+h)}\left(\frac{e^{v_n/(\theta+h)}-1}{\widetilde{\pi}^{(\theta)}e^{v_n/(\theta+h)}+1-\widetilde{\pi}^{(\theta)}}\right) \ge 0$. Now consider the case where $n_e = 1$. In this case, stochastically smallest distribution is given by an $M/M/1/2$-type queueing system where $g_0 = 1/(1+\rho)$ and $g_1 = \rho/(1+\rho)$ we have,

$$E_{G(\widetilde{\pi}^{(\theta)},\theta+h)}\left(\frac{e^{v_n/(\theta+h)}-1}{\widetilde{\pi}^{(\theta)}e^{v_n/(\theta+h)}+1-\widetilde{\pi}^{(\theta)}}\right) \le \frac{1}{(1+\rho)}\frac{e^{v_0/(\theta+h)}-1}{\widetilde{\pi}^{(\theta)}e^{v_0/(\theta+h)}+1-\widetilde{\pi}^{(\theta)}} + \frac{\rho}{(1+\rho)}\frac{e^{v_1/(\theta+h)}-1}{\widetilde{\pi}^{(\theta)}e^{v_1/(\theta+h)}+1-\widetilde{\pi}^{(\theta)}}.$$

Similarly, there must exist an $R_L(\theta) > p + C/\mu$ value such that for all $R < R^L(\theta)$, the last term is negative, hence $E_{G(\widetilde{\pi}^{(\theta)},\theta+h)}\left(\frac{e^{v_n/(\theta+h)}-1}{\widetilde{\pi}^{(\theta)}e^{v_n/(\theta+h)}+1-\widetilde{\pi}^{(\theta)}}\right) \le 0$. This is true for any $\lambda > 0$. Since we have these thresholds for given θ, there must also exists thresholds that work for all θ.

Proof of Proposition 3 To show that customer utility in equilibrium is decreasing in information cost, we use an following equivalent maximization problem that gives the equilibrium. Let $W_c(\Pi,\theta) = E_{G_\Pi}[\pi_Q v_Q] - c(\Pi,G,\theta)$ where $c(\Pi,G,\theta)$ is as in (7). Equilibrium joining probability is found by solving

$$\max_{\Pi=\{\pi_n;n\geq 0\}} W_c(\Pi,\theta) \quad \text{s.to } \pi_n \in [0,1] \text{ where } G \text{ is given in (13).} \tag{22}$$

This is seen by noting that an equilibrium to the model in Definition 1 is a feasible point to (22). Since the equilibrium is unique, there is no way to improve the objective function and solving (22) gives the equilibrium. The customer welfare in equilibrium is then denoted by $W_c(\widetilde{\Pi}^{(\theta)},\theta)$. Let $\theta_1 < \theta_2$. Since $W_c(\Pi,\theta)$ is decreasing in θ, we have $W_c(\widetilde{\Pi}^{(\theta_2)},\theta_2) = E_{G_{\widetilde{\Pi}^{(\theta_2)}}}\pi_n v_n - c(\widetilde{\Pi}^{(\theta_2)},G_{\widetilde{\Pi}^{(\theta_2)}},\theta_2) < E_{G_{\widetilde{\Pi}^{(\theta_2)}}}\pi_n v_n - c(\widetilde{\Pi}^{(\theta_2)},G_{\widetilde{\Pi}^{(\theta_2)}},\theta_1)$. By optimality of $\widetilde{\Pi}^{(\theta_1)}$, we have $E_{G_{\widetilde{\Pi}^{(\theta_2)}}}\pi_n v_n - c(\widetilde{\Pi}^{(\theta_2)},G_{\widetilde{\Pi}^{(\theta_2)}},\theta_1) \leq E_{G_{\widetilde{\Pi}^{(\theta_1)}}}\pi_n v_n - c(\widetilde{\Pi}^{(\theta_1)},G_{\widetilde{\Pi}^{(\theta_1)}},\theta_1)$. Hence, $W_c(\widetilde{\Pi}^{(\theta_2)},\theta_2) \leq W_c(\widetilde{\Pi}^{(\theta_1)},\theta_1)$.

Proof of Theorem 2 Using necessary and sufficient conditions in Lemma 1, for a given belief g_0 (steady-state probability of zero customers in the system), the solution to the rational inattention problem with $v_0 = R - p - \frac{C}{\mu} > 0$ and $v_1 = -T$ is

$$\overline{\pi} = \begin{cases} 0 & \text{if } \frac{g_0}{1-e^{v_1/\theta}} - \frac{1-g_0}{e^{v_0/\theta}-1} < 0 \\ 1 & \text{if } \frac{g_0}{1-e^{v_1/\theta}} - \frac{1-g_0}{e^{v_0/\theta}-1} > 1 \\ \frac{g_0}{1-e^{v_1/\theta}} - \frac{1-g_0}{e^{v_0/\theta}-1} & \text{otherwise.} \end{cases}. \tag{23}$$

Noting $\widetilde{g}_0 = (1+\rho\widetilde{\pi}_0)^{-1}$, the first element in (15) is the solution to the fixed point equation

$$\frac{\frac{1}{1+\rho\frac{\overline{\pi}^* e^{v_0/\theta}}{1-\overline{\pi}^*+\overline{\pi}^* e^{v_0/\theta}}}}{1-e^{v_1/\theta}} - \frac{1-\frac{1}{1+\rho\frac{\overline{\pi}^* e^{v_0/\theta}}{1-\overline{\pi}^*+\overline{\pi}^* e^{v_0/\theta}}}}{e^{v_0/\theta}-1} = \overline{\pi}^*.$$

Note that this point can not be negative since v_0 is assumed positive. However, it can be greater than one and in this case the unique equilibrium point is in the boundary, i.e., $\overline{\pi}^* = 1$. $\square$

Proof of Proposition 4

(1) Let us rewrite the first element in $\widetilde{\pi}$ in (15) as

$$\widetilde{\pi} = \frac{\left(e^{v_0/\theta}-1\right)}{\left(e^{T/\theta}-1\right)\left(\rho e^{(v_0-T)/\theta} + e^{(v_0-T)/\theta} - e^{-T/\theta}\right)}.$$

When $v_0 \geq T$, it is clear that $\left(e^{v_0/\theta}-1\right)/\left(e^{T/\theta}-1\right)$ is increasing in θ. Furthermore, the denominator is decreasing in θ, which makes $\widetilde{\pi}$ increasing. When $v_0 < T$, on the other hand, $\widetilde{\pi}$ may not be monotone, yet, $\widetilde{\pi}$ is maximum when $\theta = 0$. To show this, let us first denote $\widetilde{\pi}(\theta)$ as a function of information cost. Assume to the contrary that for some $\theta > 0$, $\widetilde{\pi}(0) < \widetilde{\pi}(\theta)$, i.e.

$$\frac{1}{1+\rho} < \frac{\left(e^{v_0/\theta}-1\right)}{\left(1-e^{-T/\theta}\right)\left(\rho e^{v_0/\theta} + e^{v_0/\theta} - 1\right)}.$$

After some manipulation, it reduces to

$$\rho < \frac{e^{(v_0-T)/\theta} - e^{-T/\theta}}{(1-e^{(v_0-T)/\theta})}.$$

Note that for $v_0 < T$, the right hand side is increasing in θ and in the limit

$$\lim_{\theta\to\infty} \frac{e^{(v_0-T)/\theta} - e^{-T/\theta}}{(1-e^{(v_0-T)/\theta})} = \lim_{\theta\to\infty} \frac{(v_0-T)/\theta e^{(v_0-T)/\theta} + T/\theta e^{-T/\theta}}{-(v_0-T)/\theta e^{(v_0-T)/\theta}} = \frac{v_0}{T-v_0}.$$

However, $\widetilde{\pi}(0) > \widetilde{\pi}(\infty)$ when $\rho > v_0/(T-v_0)$ which is a contradiction.

(2) For any $\theta \geq 0$, throughput is $\lambda\widetilde{\pi}_0/(1+\rho\widetilde{\pi}_0) = \lambda/(1/\widetilde{\pi}_0+\rho)$. When $\theta = 0$, throughput is $\lambda/(1+\rho)$. Since $\widetilde{\pi}_0 \in [0,1]$, $\lambda/(1+\rho) \geq \lambda/(1/\widetilde{\pi}_0+\rho)$. □

Proof of Theorem 3 Let us define conditional joining probabilities as a function of queue length n, service reward ω_R, service rate ω_μ and unconditional joining strategy q as $\overline{\pi}_n(q;\omega_R,\omega_\mu) = \frac{qe^{v_n(i,j)/\theta}}{1-q+qe^{v_n(\omega_R,\omega_\mu)/\theta}}$ with corresponding queue length distribution

$$g_0(q;\omega_R,\omega_\mu) = \left(1+\sum_{n=1}^{\infty} \rho_j^n \overline{\pi}_0(q;\omega_R,\omega_\mu)\overline{\pi}_1(q;\omega_R,\omega_\mu)...\overline{\pi}_{n-1}(q;\omega_R,\omega_\mu)\right)^{-1},$$
$$g_n(q;\omega_R,\omega_\mu) = \overline{g}_0(q;\omega_R,\omega_\mu)\rho_{\omega_\mu}^n \overline{\pi}_0(q;\omega_R,\omega_\mu)\overline{\pi}_1(q;\omega_R,\omega_\mu)...\overline{\pi}_{n-1}(q;\omega_R,\omega_\mu) \text{ for } n \geq 1.$$

Note that $\overline{\pi}_n(q;\omega_R,\omega_\mu)$ is increasing in q for all $n \geq 0$. Then, exactly as in Lemma 2, $G(q;\omega_R,\omega_\mu)$ is stochastically increasing in q as a distribution. Now, consider the following maximization problem:

$$\overline{\pi}(q) = \arg\max_{\pi\in[0,1]} \left[\theta \sum_{(n,\omega_R,\omega_\mu)\in\Omega} g_n(q;\omega_R,\omega_\mu) h_{R,\mu}(\omega_R,\omega_\mu) \log\left(\pi e^{v_n(\omega_R,\omega_\mu)/\theta} + 1 - \pi\right)\right] \tag{24}$$

whose first order derivative with respect to decision variable π is

$$\theta \sum_{(\omega_R,\omega_\mu)\in\Omega_{R,\mu}} h_{R,\mu}(\omega_R,\omega_\mu) \sum_{n=0}^{\infty} g_n(q;\omega_R,\omega_\mu) \left[\frac{e^{v_n(\omega_R,\omega_\mu)/\theta} - 1}{\pi e^{v_n(\omega_R,\omega_\mu)/\theta} + 1 - \pi}\right]. \tag{25}$$

Since the term inside the brackets is decreasing in n and $G(q;\omega_R,\omega_\mu)$ is stochastically increasing in q, its expectation with respect to $G(q;\omega_R,\omega_\mu)$ is decreasing. This in turn implies that (25) and consequently $\overline{\pi}(q)$ is decreasing in q. Then it follows that there exists a unique $q^* \in [0,1]$ that satisfies $\overline{\pi}(q^*) = q^*$. □

Proof of Theorem 4 Note that given unconditional joining probability q, one needs to solve

$$\pi(q,\omega_R,\omega_\mu) = \frac{qe^{v(q,\omega_R,\omega_\mu)/\theta}}{qe^{v(q,\omega_R,\omega_\mu)/\theta} + 1 - q} \tag{26}$$

where $v(q,\omega_R,\omega_\mu) = \omega_R - p - \frac{C}{(\omega_\mu - \lambda\pi(q,\omega_R,\omega_\mu))^+}$ for $\pi(q,\omega_R,\omega_\mu)$ to arrive at consistent conditional joining probabilities $\overline{\pi}(q,\omega_R,\omega_\mu)$. Since $v(q,\omega_R,\omega_\mu)$ is decreasing in $\pi(q,\omega_R,\omega_\mu)$, so the right hand side of (26). Since the left hand side is increasing and both sides are in $[0,1]$, the solution is unique.

Then we can compute the utilities $v(q,\omega_R,\omega_\mu)$ consistent with q. Finally, we find the corresponding rationally inattentive behaviour (implied unconditional joining probability) by solving $\overline{\pi}(q) = \arg\max_{\pi\in[0,1]} \sum_{(\omega_R,\omega_\mu)\in\Omega_{R,\mu}} h_{R,\mu}(\omega_R,\omega_\mu)\log\left[\pi e^{v(q,\omega_R,\omega_\mu)}+1-\pi\right]$. Then it is easy to see that $\overline{\pi}(q)$ is decreasing in q since the first order derivative of the objective function is decreasing in q which is a direct consequence of the consistent utilities being decreasing in q. Since $\pi(q)$ maps $[0,1]$ to $[0,1]$, there exists a unique point where $\overline{\pi}(q)=q$ which is the equilibrium point. $\square$

Proof of Theorem 5 Let $q=\{q_k;\underline{n}\le k\le\overline{n}\}$ be a queue-dependent *unconditional* joining strategy. Since $\widetilde{\pi}_n=1$ for $n<\underline{n}$ and $\widetilde{\pi}_n=0$ for $n>\overline{n}$, we define conditional queue length distribution as a function of q as

$$\begin{aligned}\tilde{g}_0(q,\omega_R,\omega_\mu) &= \left(1+\sum_{i=1}^{\overline{n}+1}\rho^i_{\omega_\mu}q_{\underline{n}}q_{\underline{n}+1}...q_{i-1}\right)^{-1}\\ \tilde{g}_k(q,\omega_R,\omega_\mu) &= \tilde{g}_0(q,\omega_R,\omega_\mu)\rho^n_{\omega_\mu}q_{\underline{n}}q_{\underline{n}+1}...q_{k-1} \text{ for } 1\le k\le\overline{n}\end{aligned}$$

and the corresponding prior belief given q as $h_{R,\mu|Q}(q,\omega_R,\omega_\mu\mid\cdot)$ which is given in (17). We also define,

$$\overline{\pi}_k(q) = \arg\max_{\pi\in[0,1]}\theta\sum_{\Omega_{R,\mu}} h_{R,\mu|Q}(q,\omega_R,\omega_\mu\mid k)\log\left(\pi e^{v_k(\omega_R,\omega_\mu)/\theta}+1-\pi\right) \quad \text{for} \quad \underline{n}\le k\le\overline{n}.$$

Let $m=\overline{n}-\underline{n}+1$ and define $\overline{\pi}:[0,1]^m\to[0,1]^m$ such that $\overline{\pi}(q)=(\overline{\pi}_{\underline{n}}(q),\overline{\pi}_{\underline{n}+1}(q),...,\overline{\pi}_{\overline{n}}(q))$. Then note that finding equilibrium joining strategies is equivalent to finding the fixed points of the function $\overline{\pi}$. It can be shown that $\overline{\pi}_k:[0,1]^m\to[0,1]$ is a continuous function. Hence, $\overline{\pi}$ is also continuous. Furthermore, $\overline{\pi}$ maps points from a convex compact set to the points in the same set. Therefore, by Brouwer Fixed Point Theorem, it must have a fixed point in $[0,1]^m$ $\square$

Stability of the Equilibrium

We now show how the equilibrium in Definition 1 can be attained in an adaptive way. We use time periods indexed as $t\in\{0,1,2,...\}$ and assume that each period is long enough for the system to reach steady-state. At $t=1$ we assume that customers start with an arbitrary belief about percentage of joining customers q_0 by which they form their prior belief $G(q_0)$ on queue length which is defined in (19). Customers' best response in period $t=1$ given $G(q_0)$ is then $q_1=\overline{\pi}(q_0)$ which is defined in (20). At any period $t\ge1$, customers use the average of joining fractions of previous periods to form their belief about queue length distribution. More specifically, their prior at period t is $G\left(\sum_{k=0}^{t-1}q_k/t\right)$ and the resulting unconditional joining probability is $q_t=\overline{\pi}\left(\sum_{k=0}^{t-1}q_k/t\right)$. In the following proposition, we show that customer behaviour q_t converges to the equilibrium joining probability q^* that satisfies $\overline{\pi}(q^*)=q^*$, i.e., equilibrium given in Definition 1. Here we also remark that the equilibrium can also be achieved if customers weight observations from more recent generations more heavily using a geometric distribution (as in anecdotal reasoning model of Huang and Chen (2015)).

PROPOSITION 5. *The sequence* $q=\{q_t; t\geq 0\}$ *converges to* q^* *which satisfies* $\overline{\pi}(q^*)=q^*$.

Proof Let $x_t=\sum_{k=0}^{t-1} q_k/t$. Then $q_t=\overline{\pi}(x_t)$ and

$$q_t=\overline{\pi}(x_t)=\overline{\pi}\left(\frac{1}{t}\sum_{k=0}^{t-2}q_k+\frac{1}{t}q_{t-1}\right)=\overline{\pi}\left(\left(\frac{t-1}{t}\right)x_{t-1}+\frac{1}{t}q_{t-1}\right)$$
$$=\overline{\pi}\left(x_{t-1}+\frac{(q_{t-1}-x_{t-1})}{t}\right)=\overline{\pi}\left(x_{t-1}+\frac{(\overline{\pi}(x_{t-1})-x_{t-1})}{t}\right)$$

This means that if $q_{t-1}=\overline{\pi}(x_{t-1})>x_{t-1}$, $q_t<q_{t-1}$ and if $q_{t-1}=\overline{\pi}(x_{t-1})<x_{t-1}$, $q_t>q_{t-1}$. Note that if $\overline{\pi}(x_{t-1})=x_{t-1}$ for some $t^*\geq 1$, then $q_t=\overline{\pi}(x_t)=\overline{\pi}(x_{t-1})$, for all $t\geq t^*$ and fixed point is achieved, i.e., $q^*=x_{t^*-1}$. Otherwise, $\frac{(\overline{\pi}(x_{t-1})-x_{t-1})}{t}\to 0$ as $t\to\infty$ and $\overline{\pi}(x_t)\to\overline{\pi}(x_{t-1})$, i.e., x converges to the fixed point q^*.

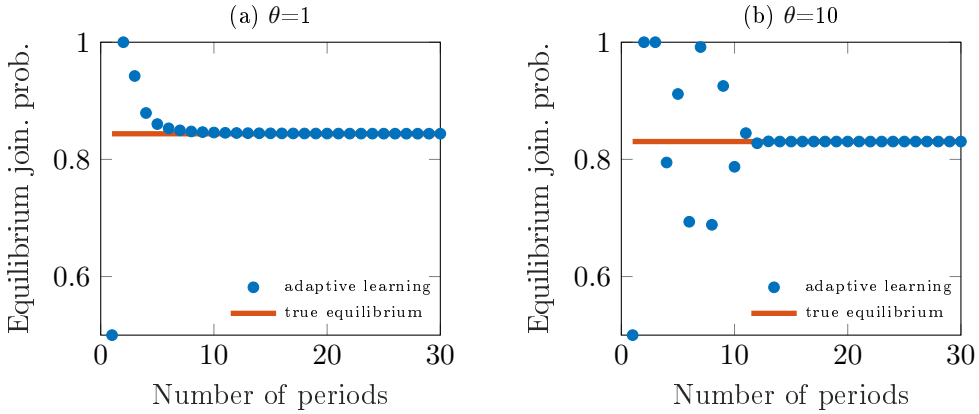

Figure 13 **Convergence to the system equilibrium** ($R=3.8, p=0, C=1, \mu=1, \lambda=0.9$)

In Figure 13, we provide an illustrative example on convergence to the equilibrium for two different information cost values. We arbitrarily assume that customers at period $t=1$ start constructing their beliefs using $q_0=0.5$. The horizontal line represents the true equilibrium value in these figures. Note that customers construct the true belief and hence equilibrium is reached very quickly. As θ gets higher, convergence speed gets slightly slower.

Recent ESMT Working Papers

	ESMT No.
The impact of EU cartel policy reforms on the timing of settlements in private follow-on damages disputes: An empirical assessment of cases from 2001 to 2015	**19-03 (R1)**
Hans W. Friederiszick, ESMT Berlin and E.CA Economics Linda Gratz, E.CA Economics Michael Rauber, E.CA Economics	
Contracting, pricing, and data collection under the AI flywheel effect	**20-01**
Francis de Véricourt, ESMT Berlin Huseyin Gurkan, ESMT Berlin	
Beyond retail stores: Managing product proliferation along the supply chain	**19-02 (R1)**
Işık Biçer, Schulich School of Business, York University Florian Lücker, Cass Business School, City, University of London Tamer Boyaci, ESMT Berlin	
Marginality, dividends, and the value in games with externalities	**19-01**
Frank Huettner, ESMT Berlin André Casajus, HHL Leipzig Graduate School of Management	
Consumer choice under limited attention when alternatives have different information costs	**16-04 (R3)**
Frank Huettner, ESMT Berlin Tamer Boyacı, ESMT Berlin Yalçın Akçay, Melbourne Business School	