Impatience and Learning in Queues

Senthil Veeraraghavan
The Wharton School, University of Pennsylvania, Philadelphia, PA, 19104.
senthilv@wharton.upenn.edu
Li Xiao
Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, China.
xiaoli@sz.tsinghua.edu.cn
Hanqin Zhang
Business School, National University of Singapore, Singapore, 117592.
bizzhq@nus.edu.sg

November 2018

Abstract

Customers often abandon waiting in queues when they get impatient. Prior literature on Markovian queues shows that it is not rational for customers to quit "midway": Customers should either quit immediately on arrival (balk) or wait till the completion of their service. We show how in-queue abandonment behavior can be rational in queues because of learning. We compare how rational Bayesian customers make abandonment decisions under different information disclosures. Our paper reveals interesting features in waiting behavior, showing that customers can be (rationally) more patient in slower shorter queues, than in faster longer queues. Using stochastic comparisons, we demonstrate that customers who anticipate a congested system can become more impatient. Finally, we show that Bayesian customers may exhibit a more conservative threshold joining behavior compared to myopic customers with the same priors.

1 Introduction

Customers value time, and often exhibit limited patience when they are delayed waiting for a service. In services and many operational contexts, such impatience is often observed through the actions of customers "abandoning" or "reneging" from waiting in a queue. In a call center, for instance, customers hang up after waiting for a while, and such behavior has been addressed as *abandonment* in the call center literature. In hospital emergency departments, patients may leave without receiving any treatment as their waiting room is too crowded or after waiting too long, and this action has been termed as *leaving without been seeing* in the healthcare literature.

In operations research, the most common characterization of customer impatience is to assume an exogenously provided abandonment threshold or patience threshold. Specifically, each customer will abandon waiting in a queue when the time she has waited in queue exceeds her patience threshold. Otherwise, the customer will continue to stay in the queue. Naturally, customers have different patience thresholds, which generate a distribution of patience thresholds in the customer population. Using such a distribution for patience thresholds facilitates the modeling of system-level dynamics and performance.

To wit, in theoretical literature, abandonments are modeled as driven by *ex ante* commitments to thresholds. A customer abandons the queue when the time elapsed exceeds her patience threshold; Otherwise, she stays in the system till her service is completed. Thus, such an approach sheds little light on *why* and *how* customers abandon during their time waiting in queues.

In practice, abandonment decisions are affected by the announcements, timely information, and dynamics of operational services such as evolution of queue length and the nature of service flows. Empirical research provides more supportive evidence. Batt and Terwiesch (2015) find that in an emergency department at a hospital, patients abandon queues based on information such as queue length, service completions and additional arrivals. Bolandifar et al. (2016) find that observed service rate also affects abandonment behavior in emergency departments. In a call-center setting, Zohar et al. (2002) find that customers adapt their patience based on system congestion. In service settings such as in the emergency room visits, patients arrive with limited information, and react to new information in making their decision to abandon. Thus, it is imperative to model customer abandonment based on queue observations. The main goal of this paper is to study the effect of such *observational learning* and *accrual of information* on patience to wait for a service.

We consider the case of a customer arriving to a system without full information about the service rate. As the customer learns more about the service while waiting, she updates her prior belief using the evolution of information, until she is able to make a better informed abandonment decision. Such focal customer approach has been employed to elicit new findings by Mandelbaum and Yechiali (1983) on server re-entry decisions and Wang (2016) on social welfare in queues.

Our paper combines research related to two topics: customer abandonment in queues, and customer learning in queues.

Models of Abandonment: The full-information queue joining model was introduced by Naor (1969). The evolution of literature that emerged from this model is succinctly summarized in the book by Hassin and Haviv (2003). Baccelli et al. (1984) derive the formula of virtual waiting time distribution, given arrival and patience distributions. However, incorporating rational abandonment into queueing dynamics has remained an open and challenging area due to the complex interactions. Rational abandonments are triggered by queueing dynamics observed. In turn, queueing dynamics emerge from such rational abandonments.

In the seminal paper, Mandelbaum and Shimkin (2000) show that the optimal rational abandonment decision in an unobservable $M/M/m$ queue, is a “degenerate decision”: customer should abandon upon arrival if not immediately serviced, or should never abandon. This conclusion holds since the virtual waiting time has an increasing hazard rate, i.e., the longer the customer is waiting, the more likely she will get service.

However, as argued earlier, we commonly observe abandonments after waiting a finite time in queue in empirical settings. To explain such abandonment behavior, theoretical research has adopted few approaches. One approach is to assume unreliable servers: Mandelbaum and Shimkin (2000) propose an $M/M/m(q)$ model, where customers will enter a fault position and never get service with probability $1-q$. Another approach is to assume the marginal waiting cost to be increasing in waiting time (as done in Shimkin and Mandelbaum (2004), Aksin et al. (2013), Hassin and Haviv (1995), Haviv and Ritov (2001)). A third approach is to examine priorities and multi-class comparisons to characterize abandonment decisions, and corresponding policies as considered in Ata and Peng (2018), Bassamboo and Randhawa (2016) and Kim et al. (2018). Unlike the aforementioned papers, we adopt an individual consumer-level *learning while waiting* approach to characterize abandonment decisions.

Models of Learning: Our paper is related to recent papers on limited information in queues: Cui and Veeraraghavan (2016) model customers arriving with fixed, but incorrect beliefs on service rates, and Debo and Veeraraghavan (2014) model customers arriving with noisy priors on service rates. Both the aforementioned papers assume that customers do not learn service rate during waiting. In our model, we consider customer learning service rate during waiting. Such in-queue learning models are new and emerging.

In general, rational abandonment behaviors have been calibrated numerically or through simulation. Altman and Shimkin (1998) propose a dynamic learning process, which is shown, using numerical approach, to converge to the equilibrium. Zohar et al. (2002) test how simulated customers learn their waiting time such that the system will converge to the equilibrium in the numerical experiment. Parkan and Warren (1978) adopt a focal consumer approach to learning but impose that the prior belief follows a gamma distribution. The research that is most relevant to ours is the work by Yao et al. (2016) who study how consumers use last service completion and compare it with their deterministic patience time.

Our model generalizes the approach in at least three respects. First, we do not impose any assumption on the prior beliefs. Second, our exact analysis is applied to both unobservable and observable queues. Finally, we can calibrate how different information orders affect the rational abandonment decisions.

We analyze an $M/M/1$ queueing system with homogeneous service value and unit waiting cost among all customers. We investigate customer abandonment behavior using a learning perspective for both unobservable queues (waiting on a call) and observable queues (lining up at grocery checkout lines) and various informational frameworks. We find that more information disclosure can sometimes decrease the willingness to join. A summary of our findings is as follows.

- We propose a rational model of customer impatience, and demonstrate abandonment could occur even in an $M/M/1$ system due in-queue customer learning in both observable and unobservable systems.

- We theoretically demonstrate based on the likelihood ratio order, that customers who anticipate a congested system situation become less patient, and the remaining sojourn time has a decreasing hazard rate for unobservable systems. This finding contrasts with some previous theoretical hypotheses.

- We show that customers are (rationally) more patient in shorter lines when only queue length is observable (even if the shorter queue is slow). Customers can be more patient in shorter, slower queues than in longer but faster queues. Thus, a long queue can rationally discourage customers, even when service is fast.

- We find that abandonments can be randomly distributed even with identical waiting costs and priors. This conclusion emerges from a comparison of joining thresholds with and without learning the queue length.

2 The Model

Consider a first-come first-served queue in which customer service time follows an exponential distribution with rate μ, and customer arrivals follow the Poisson process with rate λ. We define traffic intensity $\rho = \lambda/\mu < 1$.

We focus on one uninformed customer, who does not know the service rate μ but has a prior belief $f(\cdot)$ about the service rate μ. Information about the other system primitives such as the arrival rate and service discipline is assumed to be known. Thus we can, without loss of generality, assume that the prior belief $f(\cdot)$ has support on interval $(\lambda,\ \infty)$. Hence, for each given value of $y \in (\lambda,\ \infty)$, the uninformed customer will believe that the service time follows an exponential distribution, $(1 - e^{-yt})$, with probability $f(y)$.

The uninformed customer is rational and Bayesian. In particular, one of the following three types of information becomes available for the uninformed customer:

- The time length she has stayed in the system (elapsed sojourn time);

- The number of the customers in the system upon her arrival;

- The number of the customers in the system upon her arrival and the first service completion time after her arrival.

The uninformed customer first uses one of above three types of information to make the Bayesian updating on her prior belief, and then calculates her utility given by the difference of her service value (denoted by V) and the cost incurred by the total time she spends in the system (the total time has two parts: one part is the time period she already spent, the other part is the expected remaining sojourn time with respect to the posterior belief), and finally decides whether to stay in the system to obtain service or to leave the system without receiving service. We use c to represent the per unit time cost for the uninformed customer to stay in the system to obtain her service.

The rest of the paper is organized as follows: In the following Section 3, we consider abandonment in unobservable queues, and then consider observable queues in Section 4. We compare observable and unobservable queues in Section 5. Finally, we conclude our paper by Section 6. All technical proofs are presented in the Appendix.

3 Unobservable Queue

When the queue length and the customer service completions are not observable to the uninformed customer, she can only use the time spent in the system to update prior beliefs on the service rate. Let τ be the time the uninformed customer has already spent in the system. That is, the service of the uninformed customer has not been completed by $(s+\tau)$ when her arrival time was s. The uninformed customer's posterior belief about the service rate denoted by $f(\cdot|\tau)$ after she has already spent time period τ in the system can be derived from Bayes rule (for e.g., see page 182 in Barber (2012)). Note that in the stationary $M/M/1$ queue, the sojourn time in the system (waiting time plus service time) is an exponential random variable with the parameter determined by the difference of service rate and arrival rate (see Gross et al. (2008)). Thus, the uninformed customer's posterior belief $f(\cdot|\tau)$ on service rate can be formally written as

$$f(y|\tau) = \frac{f(y)e^{-(y-\lambda)\tau}}{\int_{\lambda}^{\infty} f(x)e^{-(x-\lambda)\tau}\mathsf{d}x}. \tag{1}$$

In the following analysis, we use $\tilde{\mu}$ to denote the random variable corresponding the prior belief $f(\cdot)$, and $\tilde{\mu}^p$ to represent the random variable corresponding to the posterior belief $f(y|\tau)$.

Here $e^{-(y-\lambda)\tau}$ is the likelihood that the total time in the system exceeds τ when prior belief about the service rate is y, and $\int_{\lambda}^{\infty} f(x)e^{-(x-\lambda)\tau}\mathsf{d}x$ is the normalizing constant. Based on her posterior belief $f(\cdot|\tau)$, the uninformed customer's utility, denoted by $\mathcal{U}_u(\tau)$ after she has spent τ time units in the system, can be written as

$$\mathcal{U}_u(\tau) \triangleq V - c\tau - c\int_{\lambda}^{\infty} \frac{f(y|\tau)}{y-\lambda}\mathsf{d}y. \tag{2}$$

The second term on the right-hand side represents the sunk cost of time that has already been spent in waiting, and the third term on the right-hand side is the residual expected sojourn time cost based on the posterior belief about the service rate. Plugging (1) into (2) yields that

$$\begin{aligned}
\mathcal{U}_u(\tau) &= V - c\tau - c\int_{\lambda}^{\infty} \frac{1}{y-\lambda} \cdot \frac{f(y)e^{-(y-\lambda)\tau}}{\int_{\lambda}^{\infty} f(x)e^{-(x-\lambda)\tau}\mathsf{d}x}\mathsf{d}y \\
&= V - c\tau - \frac{c}{\int_{\lambda}^{\infty} f(x)e^{-(x-\lambda)\tau}\mathsf{d}x}\int_{\lambda}^{\infty} \frac{f(y)e^{-(y-\lambda)\tau}}{y-\lambda}\mathsf{d}y \\
&= V - c\tau - c\cdot\left(\mathsf{E}_{\tilde{\mu}}\left[\frac{e^{-(\tilde{\mu}-\lambda)\tau}}{\tilde{\mu}-\lambda}\right]\Big/\mathsf{E}_{\tilde{\mu}}\left[e^{-(\tilde{\mu}-\lambda)\tau}\right]\right).
\end{aligned} \tag{3}$$

Note that the expectation operator is over the uninformed customer's initial beliefs. To simplify notation, we define

$$\varphi(\tau) \triangleq \mathsf{E}_{\tilde{\mu}}\left[\frac{e^{-(\tilde{\mu}-\lambda)\tau}}{\tilde{\mu}-\lambda}\right]\Big/\mathsf{E}_{\tilde{\mu}}\left[e^{-(\tilde{\mu}-\lambda)\tau}\right].$$

Then $\mathcal{U}_u(\cdot)$ can be simplified as

$$\mathcal{U}_u(\tau) = V - c\tau - c\varphi(\tau). \tag{4}$$

In order to characterize the abandonment decision of the uninformed customer, we will first establish the monotonicity of $\varphi(\tau)$ with respect to sunk time τ in Lemma 1. The proof of the lemma is presented in the Appendix.

Lemma 1. *$\varphi(\tau)$ is continuous and strictly increasing in τ.*

Using the definition of $\varphi(\tau)$, we know that

$$\varphi(\tau) = \int_{\lambda}^{\infty} \frac{f(y|\tau)}{y-\lambda} \mathsf{d}y.$$

This indicates that $\varphi(\tau)$ can be considered as the residual expected sojourn time but based on the posterior belief about the service rate. Of course, when τ is set to be zero, we get

$$\varphi(0) = \int_{\lambda}^{\infty} \frac{f(y)}{y-\lambda} \mathsf{d}y.$$

$\varphi(0)$ is simply the expected sojourn time based only on the prior belief about the service rate.

The monotone increasing property of $\varphi(\tau)$ reveals an observation of interest, related to learning. The posterior or updated residual expected sojourn time is always larger than the prior expected sojourn time.

Now, consider the abandonment decision of the uninformed customer. The abandonment decision relates to how long time a customer can stay in the system while expecting a non-negative reward from her service completion. We solve for values of τ that would ensure that $U_t(\tau)$ is non-negative. To this end, let τ^* be a solution to the equation,

$$\mathcal{U}_u(\tau) = V - c\tau^* - c\varphi(\tau^*) = 0 \tag{5}$$

when its solution exists. With the help of Lemma 1, we know that $c\tau + c\varphi(\tau)$ is continuous and increasing with respect to τ. Moreover, $\lim_{\tau\to\infty}(c\tau + c\varphi(\tau)) = \infty$. Thus, if

$$V > \Big(c\tau + c\varphi(\tau)\Big)\Big|_{\tau=0} = c\varphi(0),$$

then (5) admits a unique solution in interval $(0, \frac{V}{c})$, and if $V < c\varphi(0)$, then (5) does not exist a solution. Moreover, if $V > c\varphi(0)$, then

$$\mathcal{U}_u(\tau) > 0 \ \text{ for } \tau \in [0,\ \tau^*), \ \text{ and } \ \mathcal{U}_u(\tau) < 0 \ \text{ for } \tau \in (\tau^*, \infty).$$

If $V < c\varphi(0)$, then

$$\mathcal{U}_u(\tau) < 0 \ \text{ for } \tau \in [0, \infty);$$

and if $V = c\varphi(0)$, then

$$\mathcal{U}_u(0) = 0, \ \text{ and } \ \mathcal{U}_u(\tau) < 0 \ \text{ for } \tau \in (0, \infty).$$

Therefore, when $V \geq c\varphi(0)$, the uninformed customer has a patience time of τ^* to stay in the system to receive her service. When $V < c\varphi(0)$, the uninformed customer will balk from the system on arrival. Again, by the increasing property of $c\tau + c\varphi(\tau)$ established in Lemma 1, we know that the solution of (5) is increasing in the service value (V), but decreasing in waiting cost (c). We summarize these findings in Theorem 1.

Theorem 1. *Consider the decision of the uninformed customer. When* $V \geq c \cdot \mathsf{E}_{\tilde{\mu}}\left[\frac{1}{\tilde{\mu}-\lambda}\right]$, *the customer waits* τ^* *for service; When* $V < c \cdot \mathsf{E}_{\tilde{\mu}}\left[\frac{1}{\tilde{\mu}-\lambda}\right]$, *customer balks from the system. Furthermore, her patience time* τ^* *after joining the system is monotone increasing in service value* (V)*, and decreasing in waiting cost* (c)*.*

Theorem 1 on the abandonment decision of the uninformed customer is quite intuitive. As V represents her service value (value obtained after service completion), and c is per time unit cost incurred by her staying in the system, the ratio of V to c indicates the *maximal time period* that the uninformed customer can "stay and learn" in the system. Lemma 1 has shown that the residual expected sojourn time of the uninformed customer will become longer as she spends more time waiting.

Thus $\frac{V}{c} < \mathsf{E}_{\tilde{\mu}}\left[\frac{1}{\tilde{\mu}-\lambda}\right]$ means that the maximal time period that the uninformed customer can spend in the system is smaller than her residual expected sojourn time even when elapsed time is zero. So the uninformed customer should leave the system immediately upon arrival (i.e., balk). Similarly, $\frac{V}{c} \geq \mathsf{E}_{\tilde{\mu}}\left[\frac{1}{\tilde{\mu}-\lambda}\right]$ means that the uninformed customer can join the system to receive service and may abandon from the system if she is not served by time τ^*.

To visualize the patience time τ^* (and in order to extend our analysis for the abandonment behavior to other cases), we introduce

$$g(\tau) := V - c\tau, \text{ and } h(\tau) := c\varphi(\tau). \tag{6}$$

$g(\tau)$ is the remaining service value after the uninformed customer has already spent time τ in the system. $h(\tau)$ is the expected remaining sojourn time cost based on the posterior belief about service rate. Thus, τ^* is the intersection point of $g(\tau) = h(\tau)$. The existence of τ^* will depend on the ordering of $g(0)$ and $h(0)$, or priors on value and expected wait when entering the queue.

Specifically, when $g(0) \geq h(0)$, the remaining expected sojourn time cost curve will cross the remaining service value (straight line). See the left-hand-side part of Figure 1. The crossing point must be below the horizontal line of $g(0)$. When $g(0) < h(0)$, the curve and straight line do not cross each other, see the right-hand-side part of Figure 1.

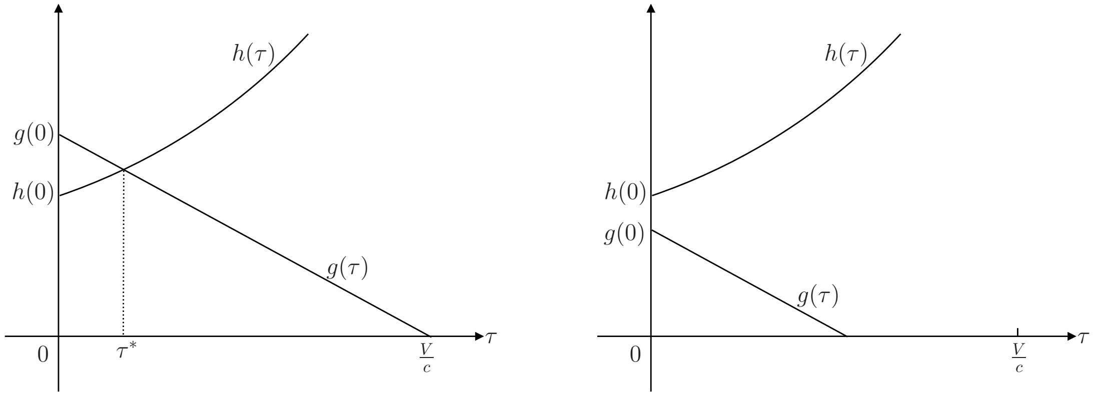
3.1 Patience and Prior Beliefs

We now analyze the dependence of the patience time τ^* with respect to the prior beliefs about the service rate. As the prior beliefs are given by the probability density function, a natural way to order those beliefs is to use the stochastic ordering of the distributions.

Following Shaked and Shanthikumar (2007), for two random variables taking values in $(-\infty, +\infty)$, say X and Y, Y is larger than X in the usual stochastic order, denote $X \leq_{\mathsf{st}} Y$, if $\mathsf{Pr}(X > x) \leq \mathsf{Pr}(Y > x)$ for all $x \in (-\infty, +\infty)$. This further implies $\mathsf{E}(X) \leq \mathsf{E}(Y)$.

Suppose we have two different beliefs on service rate μ, say $f_1(\cdot)$ and $f_2(\cdot)$. Let $\tilde{\mu}_i$ be the random variable corresponding the probability density function $f_i(\cdot)$, and τ_i^* be the patience time from Theorem 1 based on the prior belief $f_i(\cdot)$, $i = 1, 2$.

To answer the question on how the prior belief affects the abandonment decision, we first look whether there exists an order relationship between τ_1^* and τ_2^* when $\tilde{\mu}_1 \leq_{\mathsf{st}} \tilde{\mu}_2$. Let



(a) $\frac{V}{c} \geq \mathsf{E}_{\tilde{\mu}}(\frac{1}{\tilde{\mu}-\lambda})$

(b) $\frac{V}{c} < \mathsf{E}_{\tilde{\mu}}(\frac{1}{\tilde{\mu}-\lambda})$

Figure 1

$\tilde{\mu}_i^p$ denote Bayesian posterior beliefs on service rate after the elapsed sojourn time τ based on the prior belief $\tilde{\mu}_i$.

However, stochastically ordered prior beliefs do not imply the stochastically ordered posterior beliefs. To see this, consider the following Example 1.

Example 1. Consider one pair of random variables $\tilde{\mu}_1$ and $\tilde{\mu}_2$ given by

$$\tilde{\mu}_1 = \begin{cases} \mu_1, & \text{with prob. } p, \\ \mu_2, & \text{with prob. } 1-p, \end{cases} \tag{7}$$

$$\tilde{\mu}_2 = \begin{cases} \mu_1, & \text{with prob. } p, \\ \bar{\mu}_2, & \text{with prob. } 1-p, \end{cases} \tag{8}$$

where $\mu_1 < \mu_2 < \bar{\mu}_2$. Clearly $\tilde{\mu}_1 \leq_{\mathsf{st}} \tilde{\mu}_2$. Using (1), their posteriors are given by $\tilde{\mu}_1^p$ and $\tilde{\mu}_2^p$.

$$\tilde{\mu}_1^p = \begin{cases} \mu_1, & \text{with prob. } \frac{pe^{-(\mu_1-\lambda)\tau}}{pe^{-(\mu_1-\lambda)\tau}+(1-p)e^{-(\mu_2-\lambda)\tau}}, \\ \mu_2, & \text{with prob. } \frac{(1-p)e^{-(\mu_2-\lambda)\tau}}{pe^{-(\mu_1-\lambda)\tau}+(1-p)e^{-(\mu_2-\lambda)\tau}}, \end{cases} \tag{9}$$

$$\tilde{\mu}_2^p = \begin{cases} \mu_1, & \text{with prob. } \frac{pe^{-(\mu_1-\lambda)\tau}}{pe^{-(\mu_1-\lambda)\tau}+(1-p)e^{-(\bar{\mu}_2-\lambda)\tau}}, \\ \bar{\mu}_2, & \text{with prob. } \frac{(1-p)e^{-(\bar{\mu}_2-\lambda)\tau}}{pe^{-(\mu_1-\lambda)\tau}+(1-p)e^{-(\bar{\mu}_2-\lambda)\tau}}. \end{cases} \tag{10}$$

It is straightforward to verify that $\tilde{\mu}_1^p \leq_{\mathsf{st}} \tilde{\mu}_2^p$ does not hold. Especially, since the function $\frac{(1-p)e^{-(x-\lambda)\tau}}{pe^{-(\mu_1-\lambda)\tau}+(1-p)e^{-(x-\lambda)\tau}}$ is decreasing in x and $\mu_2 < \bar{\mu}_2$, then

$$\mathsf{Pr}(\tilde{\mu}_2^p \geq \bar{\mu}_2) > \mathsf{Pr}(\tilde{\mu}_1^p \geq \bar{\mu}_2) \quad \text{but} \quad \mathsf{Pr}(\tilde{\mu}_2^p \geq \mu_2) < \mathsf{Pr}(\tilde{\mu}_1^p \geq \mu_2).$$

Thus, neither $\tilde{\mu}_1^p \leq_{\mathsf{st}} \tilde{\mu}_2^p$ nor $\tilde{\mu}_2^p \leq_{\mathsf{st}} \tilde{\mu}_1^p$ holds. In other words, there is no usual stochastic order relation between $\tilde{\mu}_1^p$ and $\tilde{\mu}_2^p$. We can demonstrate the different orderings of τ_1^* and τ_2^* with numerical examples. Let $\lambda = 1,\ c = 1,\ V = 2.4,\ p = 0.4$. Consider two sets of μ_1, μ_2, and $\bar{\mu}_2$. We compute τ_1^* and τ_2^* in Table 1.

$\mu_1 = 2,\ \mu_2 = 3$	$\tau_1^* = 1.524$
$\mu_1 = 2,\ \bar{\mu}_2 = 15$	$\tau_2^* = 1.4$
$\mu_1 = 1.5,\ \mu_2 = 1.6$	$\tau_1^* = 0.595$
$\mu_1 = 1.5,\ \bar{\mu}_2 = 1.7$	$\tau_2^* = 0.723$

Table 1: Stochastic orders in posterior beliefs may not be preserved: Mixed results of the relation between τ_1^* and τ_2^*

□

Example 1 shows that when the Bayesian updating on the service rate is based on the elapsed sojourn time, the above stochastic order between the posterior beliefs ($\tilde{\mu}_1^p$ and $\tilde{\mu}_2^p$) on the service rate may not be preserved when the prior beliefs ($\tilde{\mu}_1$ and $\tilde{\mu}_2$) have a stochastic order relationship.

Hence, to order τ_1^* and τ_2^*, we require the stronger condition on prior beliefs than $\tilde{\mu}_1 \leq_{\mathsf{st}} \tilde{\mu}_2$. In the following, we restrict our attention to the likelihood ratio order, and show that Bayesian updating can preserve the likelihood ratio order in unobservable queue when using elapsed waiting time τ to update prior belief. Following Shaked and Shanthikumar (2007) (page 42), suppose X and Y are two random variables with densities $f(\cdot)$ and $g(\cdot)$, respectively. If $g(t)/f(t)$ is increasing in t over the union of the supports of X and Y (with the convention that $a/0 = \infty$ when $a > 0$), or equivalently, $f(x)g(y) \geq f(y)g(x)$ for all $x \leq y$, then X is said to be smaller than Y in the likelihood ratio order, denoted by $X \leq_{\mathsf{lr}} Y$. In Theorem 2, we first show that if prior beliefs $\tilde{\mu}_1$ and $\tilde{\mu}_2$ follow the relationship $\tilde{\mu}_1 \leq_{\mathsf{lr}} \tilde{\mu}_2$, then the posterior beliefs follow $\tilde{\mu}_1^p \leq_{\mathsf{lr}} \tilde{\mu}_2^p$. We subsequently use this result to show $\tau_1^* \leq \tau_2^*$.

Theorem 2. *The patience threshold will be shorter when the prior belief on service rate stochastically decreases in the sense of likelihood ratio order. That is, if $\tilde{\mu}_1 \leq_{\mathsf{lr}} \tilde{\mu}_2$, then $\tau_1^* \leq \tau_2^*$.*

Proof. Let $\tilde{\mu}_i^p$ be the random variables corresponding to the posterior beliefs on service rate after using τ to do Bayesian updating. Let f_i be the density function of prior belief $\tilde{\mu}_i$, and f_i^p be the density function of posterior belief $\tilde{\mu}_i^p$, $i = 1, 2$. Then

$$f_i^p(t) = \frac{f_i(t)e^{-(t-\lambda)\tau}}{\int_\lambda^\infty f_i(x)e^{-(x-\lambda)\tau}\mathsf{d}x}. \tag{11}$$

To show $\tilde{\mu}_1^p \leq_{\mathsf{lr}} \tilde{\mu}_2^p$, it's sufficient to show for any $x \leq y$, we have

$$f_1^p(x)f_2^p(y) \geq f_1^p(y)f_2^p(x). \tag{12}$$

Using (11), (12) is equivalent to

$$f_1(x)f_2(y) \geq f_1(y)f_2(x),$$

which is true because $\tilde{\mu}_1 \leq_{\mathsf{lr}} \tilde{\mu}_2$. We can write $\varphi(\tau)$ using either prior belief or posterior belief, i.e., for $i = 1, 2$,

$$\mathsf{E}_{\tilde{\mu}_i}\left[\frac{e^{-(\tilde{\mu}_i-\lambda)\tau}}{\tilde{\mu}_i - \lambda}\right] \Big/ \mathsf{E}_{\tilde{\mu}_i}\left[e^{-(\tilde{\mu}_i-\lambda)\tau}\right] = \mathsf{E}_{\tilde{\mu}_i^p}\left[\frac{1}{\tilde{\mu}_i^p - \lambda}\right].$$

By Theorem 1.C.8. in Shaked and Shanthikumar (2007) (page 46), if $\tilde{\mu}_1^p \leq_{\mathsf{lr}} \tilde{\mu}_2^p$ and $1/(x-\lambda)$ is decreasing in x, then $1/(\tilde{\mu}_1^p - \lambda) \geq_{\mathsf{lr}} 1/(\tilde{\mu}_2^p - \lambda)$, which implies

$$\mathsf{E}_{\tilde{\mu}_1^p}\Big[\frac{1}{\tilde{\mu}_1^p - \lambda}\Big] \geq \mathsf{E}_{\tilde{\mu}_2^p}\Big[\frac{1}{\tilde{\mu}_2^p - \lambda}\Big].$$

Therefore, according to the definition of τ_i^*, we have $\tau_1^* \leq \tau_2^*$. □

Theorem 2 shows that when the uninformed customer believes the system is more likely to be congested, i.e., a smaller service rate in the likelihood ratio order, then she is less patient while using elapsed sojourn time to learning the service rate. This result differs from the notion that more experienced callers may be more patient to system performance in call centers (Zohar et al., 2002). This inconsistency is caused by the fact that the Bayesian updated service rates preserve the order between the prior beliefs on the service rate when the likelihood ratio order is applied, and the faster Bayesian updated service rate implies the shorter expected residual sojourn time.

For Example 1, noting that $\mu_1 < \mu_2 < \bar{\mu}_2$, we know that neither $\tilde{\mu}_1 \leq_{\mathsf{lr}} \tilde{\mu}_2$ nor $\tilde{\mu}_2 \leq_{\mathsf{lr}} \tilde{\mu}_1$ holds. Now we demonstrate using another example to show that the prior beliefs have the likelihood ratio order, then the order relationship between τ_1^* and τ_2^* is preserved.

Example 2. Consider two random variables $\tilde{\mu}_1$ and $\tilde{\mu}_2$, with $\tilde{\mu}_1 \leq_{\mathsf{st}} \tilde{\mu}_2$,

$$\tilde{\mu}_1 = \begin{cases} \mu_1, & \text{with prob } p, \\ \mu_2, & \text{with prob } 1-p, \end{cases} \tag{13}$$

$$\tilde{\mu}_2 = \begin{cases} \mu_1, & \text{with prob } \bar{p}, \\ \mu_2, & \text{with prob } 1-\bar{p}, \end{cases} \tag{14}$$

where $\mu_1 < \mu_2$ and $p \geq \bar{p}$. It is straightforward to verify the likelihood ratio order, i.e., $\tilde{\mu}_1 \leq_{\mathsf{lr}} \tilde{\mu}_2$. Furthermore, similar to (9)-(10),

$$\tilde{\mu}_1^p = \begin{cases} \mu_1, & \text{with prob. } \frac{pe^{-(\mu_1-\lambda)\tau}}{pe^{-(\mu_1-\lambda)\tau}+(1-p)e^{-(\mu_2-\lambda)\tau}}, \\ \mu_2, & \text{with prob. } \frac{(1-p)e^{-(\mu_2-\lambda)\tau}}{pe^{-(\mu_1-\lambda)\tau}+(1-p)e^{-(\mu_2-\lambda)\tau}}, \end{cases} \tag{15}$$

$$\tilde{\mu}_2^p = \begin{cases} \mu_1, & \text{with prob. } \frac{\bar{p}e^{-(\mu_1-\lambda)\tau}}{\bar{p}e^{-(\mu_1-\lambda)\tau}+(1-\bar{p})e^{-(\mu_2-\lambda)\tau}}, \\ \mu_2, & \text{with prob. } \frac{(1-\bar{p})e^{-(\mu_2-\lambda)\tau}}{\bar{p}e^{-(\mu_1-\lambda)\tau}+(1-\bar{p})e^{-(\mu_2-\lambda)\tau}}. \end{cases} \tag{16}$$

By Theorem 2 we have $\tilde{\mu}_1^p \leq_{\mathsf{lr}} \tilde{\mu}_2^p$ and $\tau_1^* \leq \tau_2^*$. □

3.2 Rational Abandonment and Hazard Rate

Mandelbaum and Shimkin (2000) investigate customers' rational abandonment behavior from invisible queues under equilibrium setting. They prove that the hazard rate of the customer waiting time to be increasing in time. This increasing hazard-rate monotonicity property makes the customer abandonment threshold degenerate into extremal cases, either abandon immediately on arrival ($t = 0$) or never ($t = \infty$) in their model setting. They argue (correctly) this to be unrealistic, since "it implies that customers who wait for a long time become more and more optimistic about the opportunity of obtaining service in the near future." Inspired by this reasoning, we explore the monotonicity property of the hazard rate of residual sojourn times.

Given that the time length (τ) the uninformed customer has already spent in the system, denote her residual sojourn time by W_u (W_u depends on τ, but we will use W_u for notational simplicity). The cumulative distribution function of W_u based on the posterior belief about the service rate is

$$\mathsf{Pr}(W_u \le t) = \int_{\lambda}^{\infty} \left(1 - e^{-(y-\lambda)t}\right) f(y|\tau)\mathsf{d}y = 1 - \frac{\mathsf{E}_{\tilde{\mu}}\left[e^{-(\tilde{\mu}-\lambda)(t+\tau)}\right]}{\mathsf{E}_{\tilde{\mu}}\left[e^{-(\tilde{\mu}-\lambda)\tau}\right]}.$$

Let $H_u(t)$ be the hazard rate of W_u. Then,

$$H_u(t) = \frac{\mathsf{dPr}(W_u \le t)/\mathsf{d}t}{1 - \mathsf{Pr}(W_u \le t)} = \frac{\mathsf{E}_{\tilde{\mu}}\left[(\tilde{\mu}-\lambda)e^{-(\tilde{\mu}-\lambda)(t+\tau)}\right]}{\mathsf{E}_{\tilde{\mu}}\left[e^{-(\tilde{\mu}-\lambda)(t+\tau)}\right]}.$$

Theorem 3. *The hazard-rate of the residual sojourn time given by the posterior belief of a Bayesian customer is decreasing in time, t.*

Proof. To get the monotonicity of the hazard rate $H_u(\cdot)$, take the derivative with respect to t, and obtain

$$\begin{aligned}\frac{\mathsf{d}H_u(t)}{\mathsf{d}t} \quad &= \quad \frac{1}{\left[\mathsf{E}_{\tilde{\mu}}\left(e^{-(\tilde{\mu}-\lambda)(t+\tau)}\right)\right]^2}\left(\left[\mathsf{E}_{\tilde{\mu}}\left((\tilde{\mu}-\lambda)e^{-(\tilde{\mu}-\lambda)(t+\tau)}\right)\right]^2\right.\\ &\qquad \left. -\mathsf{E}_{\tilde{\mu}}\left[(\tilde{\mu}-\lambda)^2 e^{-(\tilde{\mu}-\lambda)(t+\tau)}\right]\cdot \mathsf{E}_{\tilde{\mu}}\left[e^{-(\tilde{\mu}-\lambda)(t+\tau)}\right]\right). \qquad (17)\end{aligned}$$

We will show that (17) is negative. Note that

$$\begin{aligned}&\mathsf{E}_{\tilde{\mu}}\left[(\tilde{\mu}-\lambda)^2 e^{-(\tilde{\mu}-\lambda)(t+\tau)}\right]\cdot \mathsf{E}_{\tilde{\mu}}\left[e^{-(\tilde{\mu}-\lambda)(t+\tau)}\right]\\ &\quad = \mathsf{E}_{\tilde{\mu}}\left[(\tilde{\mu}-\lambda)\cdot e^{-\frac{1}{2}(\tilde{\mu}-\lambda)(t+\tau)}\right]^2 \cdot \mathsf{E}_{\tilde{\mu}}\left[e^{-\frac{1}{2}(\tilde{\mu}-\lambda)(t+\tau)}\right]^2\\ &\quad > \left(\mathsf{E}_{\tilde{\mu}}\left(\left[(\tilde{\mu}-\lambda)\cdot e^{-\frac{1}{2}(\tilde{\mu}-\lambda)(t+\tau)}\right]\cdot\left[e^{-\frac{1}{2}(\tilde{\mu}-\lambda)(t+\tau)}\right]\right)\right)^2\\ &\quad = \left(\mathsf{E}_{\tilde{\mu}}\left[(\tilde{\mu}-\lambda)\cdot e^{-(\tilde{\mu}-\lambda)(t+\tau)}\right]\right)^2,\end{aligned}$$

where the inequality is given by the Schwarz inequality (see Corollary 3 on page 105 in Chow and Teicher (1997)) and the fact that $\tilde{\mu}-\lambda$ is random and $e^{-\frac{1}{2}(\tilde{\mu}-\lambda)(t+\tau)} \neq 0$. Hence, by (17), we have that $\mathsf{d}H_u(t)/\mathsf{d}t < 0$. This proves the theorem. □

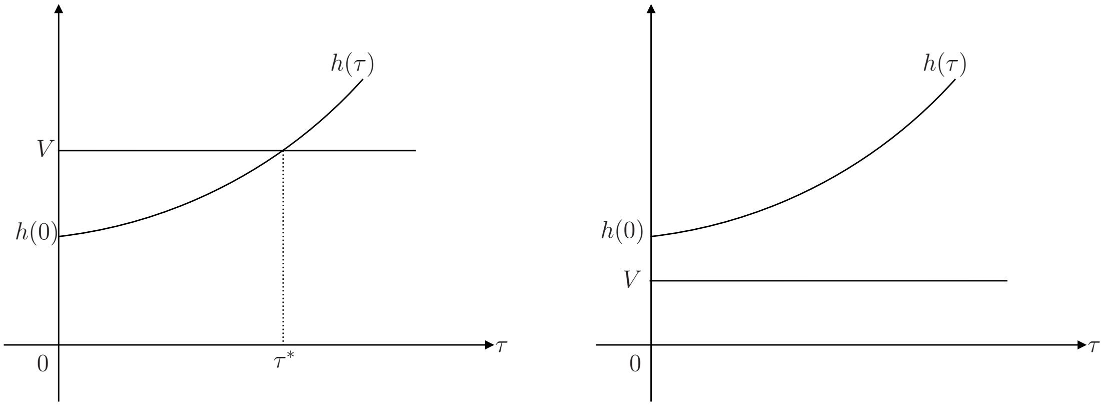
Bayesian learning in queues resolves the increasing hazard-rate monotonicity of the waiting time issue noted by Mandelbaum and Shimkin (2000). Learning leads to the decreasing hazard-rate monotonicity of the residual sojourn time. This property, in turn, offers a theoretical explanation to the non-degenerative patience time often observed in practice.

3.3 The Role of Sunk Costs

Suppose that the utility of the uninformed customer is modified into the difference of the service value (V) and the remaining expected sojourn time cost without taking account of the sunk cost incurred by the time period she has already spent in the system. If the time spent is treated as a sunk cost, $\mathcal{U}_u(\tau)$ from (2) is modified as

$$V - c\int_{\lambda}^{\infty} \frac{f(u|\tau)}{u-\lambda}\mathsf{d}u = V - h(\tau). \tag{18}$$

The above equation can be reformulated as $g(0) - h(\tau)$. Theorem 1 still holds. Furthermore, Theorem 3 also continues to hold with this modified utility (as can be verified from the proof). We modify Figure 1 as Figure 2 below using sunk cost notions. It can be seen that the structure of our main findings continues to hold.



(a) $V \geq c \cdot \mathsf{E}_{\tilde{\mu}}(\frac{1}{\tilde{\mu}-\lambda})$

(b) $V < c \cdot \mathsf{E}_{\tilde{\mu}}(\frac{1}{\tilde{\mu}-\lambda})$

Figure 2: Threshold abandonment without sunk costs (a), and balking without joining (b).

4 Observable Queues

Unlike in unobservable queues, the visibility of number of people in an observable queue provides a customer with the opportunity to observe service completion times. Using Bayesian rule, the uninformed customer updates her belief about the service rate using the available information, and then makes her abandonment (or balking) decision, if necessary. In Section 4.1, we consider the balking decision based *only* on the information about the number of the customers in the system upon arrival. In Section 4.2, we investigate the abandonment decision of a customer who observes the number of the customers in the system *and* notes the very first service completion time.

4.1 Updating Prior Beliefs Based on the Number of the Customers in the System

Upon arrival, the uninformed customer observes the number of customers in the system, which she uses to update her prior belief about the service rate by the Bayes rule, and to make the decision whether or not to balk from the system. Suppose that the number of customers in system upon arrival is n. Note that the number of customers in the stationary $M/M/1$ system follows a geometric distribution with parameter given by the ratio of arrival rate to the service rate. Similar to equation (1) for the case of unobservable queue, the posterior belief about the service rate will be given by

$$f(y|n) = \Big[\Big(\frac{\lambda}{y}\Big)^n\Big(1-\frac{\lambda}{y}\Big)f(y)\Big]\Big/\int_\lambda^\infty \Big(\frac{\lambda}{x}\Big)^n\Big(1-\frac{\lambda}{x}\Big)f(x)\mathsf{d}x. \tag{19}$$

Based on the state at her arrival, the uninformed customer's utility written by $\mathcal{U}_o(n)$, can be written as:

$$\begin{aligned}
\mathcal{U}_o(n) &= V - c(n+1)\int_\lambda^\infty \frac{1}{y} f(y|n)\mathsf{d}y \\
&= V - \frac{c(n+1)}{\int_\lambda^\infty \Big(\frac{\lambda}{x}\Big)^n\Big(1-\frac{\lambda}{x}\Big)f(x)\mathsf{d}x}\int_\lambda^\infty \frac{1}{y}\Big(\frac{\lambda}{y}\Big)^n\Big(1-\frac{\lambda}{y}\Big)f(y)\mathsf{d}y \\
&= V - c(n+1)\frac{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}}(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\Big]}. \qquad (20)
\end{aligned}$$

From the above analysis, we can see that

$$\int_\lambda^\infty \frac{1}{y} f(y|n)\mathsf{d}y = \frac{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}}(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\Big]}.$$

The right-hand-side term in the above equation is the posterior belief on the average service time. When the number of the customers in the system upon her arrival is n, the uninformed customer will join the system if and only if her utility is nonnegative, that is, $\mathcal{U}_o(n) \geq 0$. To characterize the balking decision, we first establish the monotonicity of posterior beliefs on average service time (Proof provided in the Appendix.).

Lemma 2. *The posterior belief on average service time is increasing in the number of customers in queue observed upon arrival. i.e.,* $\frac{\mathsf{E}_{\tilde{\mu}}\big[\frac{1}{\tilde{\mu}}(\frac{1}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\big]}{\mathsf{E}_{\tilde{\mu}}\big[(\frac{1}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\big]}$ *is strictly increasing in* n*.*

$$\text{Let } n_o^* = \max\Big\{n : V - c(n+1)\frac{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}}(\frac{1}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{1}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\Big]} \geq 0\Big\}. \tag{21}$$

Lemma 2 implies that a Bayesian customer correctly believes that longer queue to be slower. It follows from Lemma 2 that for $n \leq n_o^*$, $\mathcal{U}_o(n) \geq 0$, and for $n > n_o^*$, $\mathcal{U}_o(n) < 0$. Hence we have,

Theorem 4. *There exists a threshold* n_o^* *such that the uninformed customer will join the system if the number of the customers in the system upon arrival is below* n_o^**. Otherwise, she will balk from the system.*

To understand impact of Bayesian updating, we look at balking actions that occur without any updating of beliefs. We call such balking customers as "myopic". Without updating, the balking decision has to be made based on the prior beliefs. Let $\mathcal{U}(n)$ be the expected utility at state n, without any Bayesian updating.

When the uninformed customer observes the number of the customers in the system to be n upon arrival, her utility based on the prior belief $f(\cdot)$ about the service rate is given by

$$\mathcal{U}(n) = V - c(n+1) \cdot \mathsf{E}_{\tilde{\mu}}\Big(\frac{1}{\tilde{\mu}}\Big) = V - c(n+1) \int_{\lambda}^{\infty} \frac{f(x)}{x} \mathsf{d}x. \tag{22}$$

The uninformed customer will join the system if and only if $\mathcal{U}(n) \geq 0$. Let

$$n^* = \max \Big\{ n : V - c(n+1)\mathsf{E}_{\tilde{\mu}}\Big(\frac{1}{\tilde{\mu}}\Big) \geq 0 \Big\}. \tag{23}$$

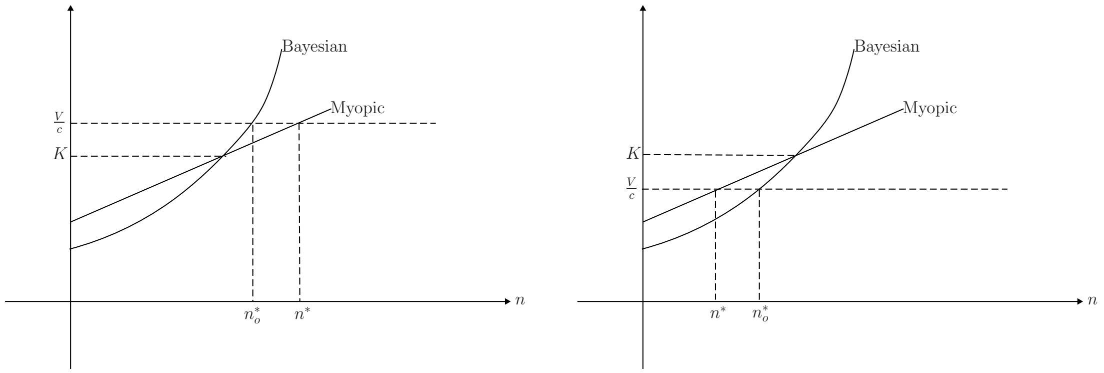
Since $V - c(n+1)\mathsf{E}_{\tilde{\mu}}\left(\frac{1}{\tilde{\mu}}\right)$ is decreasing in n, the uninformed customer will join if $n \leq n^*$, and balk from the system if $n > n^*$.

Now we contrast the myopic threshold n^* with Bayesian threshold n_o^*. We label customers who use n^* as myopic customers, and customers who use n_o^* as Bayesian customers. To study the differences in myopic and Bayesian customers, we recast (21) and (23) as

$$n_o^* = \max \Bigg\{ n : (n+1) \frac{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}}(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\Big]} \leq \frac{V}{c} \Bigg\}, \tag{24}$$

$$n^* = \max \Big\{ n : (n+1)\mathsf{E}_{\tilde{\mu}}\Big(\frac{1}{\tilde{\mu}}\Big) \leq \frac{V}{c} \Big\}. \tag{25}$$

These definitions are equivalent to (21) and (23), but help demonstrate how the relationship between thresholds n^* and n_o^* critically depends on V/c. We capture the relationship in Proposition 1, whose proof is presented in Appendix. We depict Proposition 1 graphically in Figure 3 by plotting equations (24) and (25).



(a) $n^* > n_o^*$

(b) $n^* \leq n_o^*$

Figure 3: Bayesian joining thresholds may be lower (Panel (a)) or higher (Panel (b)) than myopic thresholds.

Proposition 1. *There exists some threshold K such that when $V/c \geq K$ then $n^* \geq n_o^*$, and when $V/c < K$ then $n^* < n_o^*$.*

The proposition is intriguing. When the ratio of value to waiting cost is large, Proposition 1 shows that the Bayesian decision maker becomes *increasingly conservative* in joining the system compared to a myopic customer. In other words, the Bayesian customer will balk from queue lengths in which myopic customers may willingly join (i.e., $n_o^* < n^*$, see the left panel of Figure 3). On the other hand, the service reward is low compared to waiting costs, the Bayesian queue joiners become more "aggressive" about joining queues at queue-lengths at which myopics may balk (i.e., $n^* \leq n_o^*$). This can be seen in the right panel of Figure 3.

The principal reason for such a result is that the posterior belief about the average service time is increasing with the number of customers in the system (i.e., the queue length of the system). This monotone increasing property directly follows from the fact that the random variable with the density $f(\cdot|n)$ given by (19) is larger than the random variable with the density function $f(\cdot|(n+1))$ (replacing n by $(n+1)$ in (19)) in the likelihood ratio order.

So, more people there are in the queue, the slower the expected service rate (and longer the expected wait time). Myopics are driven by the prior belief about the average service time, which is a constant, and expected wait times grow linearly in length of the queue. As such, compared to Bayesians, myopics overestimate the wait times in short queues, and underestimate the wait times when faced with long queues.

Threshold Ordering and Prior Beliefs. Now we examine the Bayesian balking threshold n_o^* given by Theorem 4. Also, recall the result on the comparing patience time thresholds (τ_1^* and τ_2^*) for two different prior beliefs, $\tilde{\mu}_1$ and $\tilde{\mu}_2$ established in Theorem 2. Similar to the unobservable case, the usual stochastic order between the prior beliefs is not enough to determine the threshold order. We continue with the same numerical Example 1 to illustrate the result.

Example 1 (continued): Choose

$$\lambda = 1,\ c = 1,\ V = 2.4,\ p = 0.4,\ \mu_1 = 2. \tag{26}$$

If $\mu_2 = 2.1$ and $\bar{\mu}_2 = 4$, then $n_{o1}^* = 3 < n_{o2}^* = 4$. However, if $\mu_2 = 4$ and $\bar{\mu}_2 = 15$, then $n_{o1}^* = 4 > n_{o2}^* = 3$. □

While focusing on likelihood ratio order, we can show that the Bayesian updating preserves the likelihood ratio order when using queue length to update. This result is consistent with Theorem 2.

Theorem 5. *The balking threshold n_o^* becomes smaller when the prior belief on service rate stochastically decreases in the sense of likelihood ratio order. That is, if $\tilde{\mu}_1 \leq_{\mathsf{lr}} \tilde{\mu}_2$, then $n_{o1}^* \leq n_{o2}^*$.*

Proof. Let $\tilde{\mu}_i^p$ be the random variables corresponding to the posterior beliefs on service rate after using n to do Bayesian updating. Let f_i be the density function of prior belief $\tilde{\mu}_i$, and f_i^p be the density function of posterior belief $\tilde{\mu}_i^p$, $i = 1, 2$. Then for $i = 1, 2$,

$$f_i^p(t) = \left[\left(\frac{\lambda}{t}\right)^n \left(1 - \frac{\lambda}{t}\right) f_i(t)\right] \Big/ \int_\lambda^\infty \left(\frac{\lambda}{x}\right)^n \left(1 - \frac{\lambda}{x}\right) f_i(x) \mathsf{d}x. \tag{27}$$

To show $\tilde{\mu}_1^p \leq_{\mathsf{lr}} \tilde{\mu}_2^p$, it's sufficient to show for any $x \leq y$,

$$f_1^p(x) f_2^p(y) \geq f_1^p(y) f_2^p(x). \tag{28}$$

Using (27), (28) is equivalent to

$$f_1(x) f_2(y) \geq f_1(y) f_2(x),$$

which is true because $\tilde{\mu}_1 \leq_{\mathsf{lr}} \tilde{\mu}_2$. Note that for $i = 1, 2$,

$$\frac{\mathsf{E}_{\tilde{\mu}_i}\Big[\frac{1}{\tilde{\mu}_i}(\frac{\lambda}{\tilde{\mu}_i})^n(1-\frac{\lambda}{\tilde{\mu}_i})\Big]}{\mathsf{E}_{\tilde{\mu}_i}\Big[(\frac{\lambda}{\tilde{\mu}_i})^n(1-\frac{\lambda}{\tilde{\mu}_i})\Big]} = \mathsf{E}_{\tilde{\mu}_i^p}\Big[\frac{1}{\tilde{\mu}_i^p}\Big].$$

By Theorem 1.C.8. in Shaked and Shanthikumar (2007), if $\tilde{\mu}_1^p \leq_{\mathsf{lr}} \tilde{\mu}_2^p$ and $1/x$ is decreasing in x, then $1/\tilde{\mu}_1^p \geq_{\mathsf{lr}} 1/\tilde{\mu}_2^p$, which implies

$$\mathsf{E}_{\tilde{\mu}_1^p}\Big[\frac{1}{\tilde{\mu}_1^p}\Big] \geq \mathsf{E}_{\tilde{\mu}_2^p}\Big[\frac{1}{\tilde{\mu}_2^p}\Big].$$

Therefore, according to the definition of n_{oi}^*, we have $n_{o1}^* \leq n_{o2}^*$. □

Theorem 5 is intuitive. Faster service rate in the likelihood ratio order means increased likelihood of shorter average service time. An shorter average service time makes the Bayesian customer even with a longer queue length upon arrival to expect a shorter average sojourn time.

Hazard Rates Similar to the parallel discussion for unobservable queues in subsection 3.2, we now examine Theorem 3 and hazard rate of the residual sojourn time given by the posterior beliefs in observable queues. After observing the number of the customers in the system to be n, the uninformed customer updates her prior belief about the service rate into $f(y|n)$ given by (19). Her sojourn time denoted by $W_o(n+1)$ is the summation of $(n+1)$ exponential random variables with parameter y.

$W_o(n+1)$ follows the Erlang distribution with shape-parameter $(n+1)$ and rate-parameter $\tilde{\mu}$ with density $f(y|n)$. Its cumulative distribution function is

$$\begin{aligned}\mathsf{Pr}(W_o(n+1) \leq t) &= \int_\lambda^\infty \Big(1 - \sum_{k=0}^n \frac{e^{-yt}(yt)^k}{k!}\Big) f(y|n)\mathsf{d}y \\ &= \int_\lambda^\infty \Big(1 - \sum_{k=0}^n \frac{e^{-yt}(yt)^k}{k!}\Big) \frac{(\frac{\lambda}{y})^n(1-\frac{\lambda}{y})f(y)}{\int_\lambda^\infty(\frac{\lambda}{x})^n(1-\frac{\lambda}{x})f(x)\mathsf{d}x}\mathsf{d}y \\ &= \frac{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}^n}(1-\frac{\lambda}{\tilde{\mu}})\big(1-\sum_{k=0}^n \frac{e^{-\tilde{\mu}t}(\tilde{\mu}t)^k}{k!}\big)\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}^n}(1-\frac{\lambda}{\tilde{\mu}})\Big]}. \end{aligned} \tag{29}$$

Then the hazard rate denoted by $H_o(t|n)$ is

$$H_o(t|n) = \frac{\mathsf{d}\mathsf{Pr}(W_o(n+1) \leq t)/\mathsf{d}t}{1 - \mathsf{Pr}(W_o(n+1) \leq t)} = \frac{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}^n}(1-\frac{1}{\tilde{\mu}})\frac{\tilde{\mu}e^{-\tilde{\mu}t}}{n!}(\tilde{\mu}t)^n\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}^n}(1-\frac{1}{\tilde{\mu}})\sum_{k=0}^n \frac{e^{-\tilde{\mu}t}}{k!}(\tilde{\mu}t)^k\Big]}. \tag{30}$$

For $n = 0$,

$$\frac{\mathsf{d}H_o(t|n=0)}{\mathsf{d}t} = \frac{-\mathsf{E}_{\tilde{\mu}}[(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}^2 e^{-\tilde{\mu}t}] \cdot \mathsf{E}_{\tilde{\mu}}[(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}t}] + \big(\mathsf{E}_{\tilde{\mu}}[(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}e^{-\tilde{\mu}t}]\big)^2}{\big(\mathsf{E}_{\tilde{\mu}}[(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}t}]\big)^2} \leq 0. \tag{31}$$

The inequality is given by the Schwarz inequality. Therefore the hazard rate is decreasing when $n = 0$.

For $n \geq 1$,

$$\begin{aligned}\frac{\mathsf{d}H_o(t|n)}{\mathsf{d}t} \quad = \quad & \frac{1}{\Big[\mathsf{E}_{\tilde{\mu}}\big[\frac{1}{\tilde{\mu}^n}(1-\frac{\lambda}{\tilde{\mu}})\sum_{k=0}^{n}\frac{e^{-\tilde{\mu}t}}{k!}(\tilde{\mu}t)^k\big]\Big]^2}\Bigg(\Big(\mathsf{E}_{\tilde{\mu}}\big[(1-\frac{\lambda}{\tilde{\mu}})\frac{\tilde{\mu}e^{-\tilde{\mu}t}}{n!}(\tilde{\mu}t)^n\big]\Big)^2 \\ & +\mathsf{E}_{\tilde{\mu}}\big[(1-\frac{\lambda}{\tilde{\mu}})\frac{\tilde{\mu}^2 e^{-\tilde{\mu}t}}{(n-1)!}(\tilde{\mu}t)^{n-1}(1-\frac{\tilde{\mu}t}{n})\big] \cdot \mathsf{E}_{\tilde{\mu}}\big[(1-\frac{\lambda}{\tilde{\mu}})\sum_{k=0}^{n}\frac{e^{-\tilde{\mu}t}}{k!}(\tilde{\mu}t)^k\big]\Bigg) \end{aligned} \tag{32}$$

As $\mathsf{d}H_o(t|n)/\mathsf{d}t$ is continuous in t, and $\mathsf{d}H_o(t|n)/\mathsf{d}t\Big|_{t=0} > 0$. Thus we can see that there exists a t_n such that $\mathsf{d}H_o(t|n)/\mathsf{d}t > 0$ for $t \in [0, t_n)$.

Hence, unlike the case of the unobservable queue (see Theorem 3), unless the system is not empty upon arrival, the hazard rate of the sojourn time given by the posterior beliefs is not always decreasing in time, t.

4.2 Updating Belief through Service Completion Time and the Number of the Customers in the System

In Subsection 4.1, we considered queues in which a Bayesian customer updates her information based on the number of customers waiting in the queue when she arrives.

Now we consider a case when the uninformed customer uses (a) the number of customers when she arrives, and (b) the time to first service completion after she joined. First using the information about the number of the customers in the system, the posterior belief about the service rate is $f(\cdot|n)$ given by (19). Let the service completion time denoted by s. Using Bayes rule, the posterior belief about the service rate is

$$f(y|(s,n)) = \frac{f(y|n)ye^{-sy}}{\int_{\lambda}^{\infty} f(x|n)xe^{-sx}\mathsf{d}x}. \tag{33}$$

We plug (19) into (33), and rewrite (33) as

$$f(y|(s,n)) = \frac{(\frac{\lambda}{y})^n(1-\frac{\lambda}{y})ye^{-sy}f(y)}{\int_{\lambda}^{\infty}(\frac{\lambda}{x})^n(1-\frac{\lambda}{x})xe^{-sx}f(x)\mathsf{d}x}. \tag{34}$$

Let $\mathcal{U}_o(s,n)$ be the expected utility from joining the system, for a learning customer arriving at queue length n and who observes the first service completion s time units after joining.

$$\begin{aligned}\mathcal{U}_o(s,n) &= V - cs - cn\int_{\lambda}^{\infty}\frac{f(y|(s,n))}{y}\mathsf{d}y \\ &= V - cs - cn\int_{\lambda}^{\infty}\frac{1}{y}\cdot\frac{(\frac{\lambda}{y})^n(1-\frac{\lambda}{y})ye^{-sy}f(y)}{\int_{\lambda}^{\infty}(\frac{\lambda}{x})^n(1-\frac{\lambda}{x})xe^{-sx}f(x)\mathsf{d}x}\mathsf{d}y\end{aligned}$$

$$= \quad V - cs - cn\frac{\mathsf{E}_{\tilde{\mu}}\left[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})e^{-s\tilde{\mu}}\right]}{\mathsf{E}_{\tilde{\mu}}\left[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}e^{-s\tilde{\mu}}\right]}. \tag{35}$$

To characterize the abandonment decision, we prove the following lemma. The proof is presented in the Appendix.

Lemma 3. *$\mathcal{U}_o(s,n)$, is strictly decreasing in both n and s.*

The utility of the uninformed customer is continuous and strictly decreasing with respect to the observed service completion time, and strictly decreasing with the respect to the number of the customers in the system upon her arrival. By the fact that $\lim_{s\to\infty}\mathcal{U}_o(s,n) = -\infty$, it follows from Lemma 3 that for each n, there exists a unique $s^*(n)$ such that

$$\mathcal{U}_o(s,n) \geq 0 \text{ for } s < s^*(n); \text{ and } \mathcal{U}_o(s,n) < 0 \text{ for } s \geq s^*(n). \tag{36}$$

Furthermore, $s^*(n)$ is decreasing in n. Hence, we obtain the following result on the abandonment decision.

Theorem 6. *Let n be the number of the customers in the system observed by the customer upon her arrival. Then there exists a threshold $s^*(n)$ such that the customer will abandon waiting if she observes a service completion time larger than or equal to $s^*(n)$. Moreover, the threshold $s^*(n)$ is decreasing in n.*

Theorem 6 describes how the threshold decision for a customer's abandonment depends on the queue length. If she joins a longer queue, she is less likely to wait long, i.e., she will abandon earlier, $s^*(n)$ decreasing in n.

The difference $(s^*(n) - s^*(n+1))$ can be considered as the marginal decreases in abandonment threshold by one more customer in the system observed by the uninformed customer upon arrival. The example provided below demonstrates that the monotonicity of $(s^*(n) - s^*(n+1))$ depends on the service value V.

Example 3. Choose $\lambda = 0.1$, $c = 1$, and consider discrete prior distribution

$$\tilde{\mu} = \begin{cases} 0.2 & \text{with probability } 0.2, \\ 0.25 & \text{with probability } 0.3, \\ 0.3 & \text{with probability } 0.5. \end{cases} \tag{37}$$

The patience till first service time observation $s^*(n)$ can be concave or convex in n. The results of $s^*(n)$ for different values of V and n are summarized in Table 2. For small V, for example, $V = 8$ and $V = 10$, $s^*(n)$ is concave decreasing in n, and for large V, for example, $V = 20$ and $V = 30$, $s^*(n)$ is convex decreasing in n. For some V, such as $V = 15$, $s^*(n)$ can be linear in n. (Note, unfilled values in the Table implies that $s^*(n) = 0$.)

	$V=8$	$V=10$	$V=15$	$V=20$	$V=30$
$s^*(0)$	8	10	15	20	30
$s^*(1)$	4.102	6.052	10.92	15.784	25.53
$s^*(2)$	0.197	2.098	6.84	11.573	21.075
$s^*(3)$			2.76	7.368	16.632
$s^*(4)$				3.17	12.202
$s^*(5)$					7.785
$s^*(6)$					3.378

Table 2: Value of $s^*(n)$

Table 2 demonstrates the "first-event" patience time for learning customers. They abandon if they observe service time longer than $s^*(n)$. The Table clearly demonstrates that they are more patient in shorter queues. It is also evident from the table (by taking differences), that the marginal patience, i.e., $s^*(n) - s^*(n+1)$ can increase or decrease as the queue length grows.

4.3 Sunk Costs

Similar to the discussion developed at the end of Section 3, the utility function can be modified by taking out the sunk cost cs of time s that was already spent in system, from (35). Namely, the utility is modified into

$$V - cn\frac{\mathsf{E}_{\tilde{\mu}}\left[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})e^{-s\tilde{\mu}}\right]}{\mathsf{E}_{\tilde{\mu}}\left[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}e^{-s\tilde{\mu}}\right]} := \hat{\mathcal{U}}_o(s,n). \tag{38}$$

Similar to Lemma 3, we have the following Lemma.

Lemma 4. *$\hat{\mathcal{U}}_o(s,n)$ is strictly decreasing in s and n.*

Similar to the analysis of balking threshold n_o^* in Subsection 4.2, the ordering of the priors does not imply the same ordering for the posterior beliefs. Consider priors $\tilde{\mu}_1$ and $\tilde{\mu}_2$. If $\tilde{\mu}_1 \leq_{\mathsf{st}} \tilde{\mu}_2$, then the order relationship between the abandonment thresholds denoted by $s_1^*(n)$ and $s_2^*(n)$ respectively can be reversed. We adapt Examples 1 and 2 to show this property.

Example 2 (continued): Note that

$$\begin{aligned}&\frac{\mathsf{E}_{\tilde{\mu}_1}\left[\frac{1}{\tilde{\mu}_1^n}(1-\frac{\lambda}{\tilde{\mu}_1})e^{-s\tilde{\mu}_1}\right]}{\mathsf{E}_{\tilde{\mu}_1}\left[\frac{1}{\tilde{\mu}_1^{n-1}}(1-\frac{\lambda}{\tilde{\mu}_1})e^{-s\tilde{\mu}_1}\right]}\\ &\quad= \frac{\frac{1}{\mu_1^n}(1-\frac{\lambda}{\mu_1})e^{-s\mu_1} + (1-p)\left[\frac{1}{\mu_2^n}(1-\frac{\lambda}{\mu_2})e^{-s\mu_2} - \frac{1}{\mu_1^n}(1-\frac{\lambda}{\mu_1})e^{-s\mu_1}\right]}{\frac{1}{\mu_1^{n-1}}(1-\frac{\lambda}{\mu_1})e^{-s\mu_1} + (1-p)\left[\frac{1}{\mu_2^{n-1}}(1-\frac{\lambda}{\mu_2})e^{-s\mu_2} - \frac{1}{\mu_1^{n-1}}(1-\frac{\lambda}{\mu_1})e^{-s\mu_1}\right]}.\end{aligned}$$

Write the above expression as $\xi(n,p)$. Take the first order derivative of $\xi(n,p)$.

$$\frac{\mathsf{d}\xi(n,p)}{\mathsf{d}p} = -\frac{(1-\frac{\lambda}{\mu_1})(1-\frac{\lambda}{\mu_2})(\frac{1}{\mu_2}-\frac{1}{\mu_1})\frac{1}{\mu_1^{n-1}}\frac{1}{\mu_2^{n-1}}e^{-(\mu_1+\mu_2)s}}{\Big(\frac{1}{\mu_1^{n-1}}(1-\frac{\lambda}{\mu_1})e^{-s\mu_1}+(1-p)\big[\frac{1}{\mu_2^{n-1}}(1-\frac{\lambda}{\mu_2})e^{-s\mu_2}-\frac{1}{\mu_1^{n-1}}(1-\frac{\lambda}{\mu_1})e^{-s\mu_1}\big]\Big)^2} \geq 0. \tag{39}$$

The inequality is given by $\mu_1 < \mu_2$. Based on (39) and $p \geq \bar{p}$, we have $\xi(n,p) \geq \xi(n,\bar{p})$ for each n. Consequently,

$$V - cs - cn\frac{\mathsf{E}_{\tilde{\mu}_1}\big[(\frac{1}{\tilde{\mu}_1})^n(1-\frac{\lambda}{\tilde{\mu}_1})e^{-s\tilde{\mu}_1}\big]}{\mathsf{E}_{\tilde{\mu}_1}\big[(\frac{1}{\tilde{\mu}_1})^{n-1}(1-\frac{\lambda}{\tilde{\mu}_1})e^{-s\tilde{\mu}_1}\big]} \leq V - cs - cn\frac{\mathsf{E}_{\tilde{\mu}_2}\big[(\frac{1}{\tilde{\mu}_2})^n(1-\frac{\lambda}{\tilde{\mu}_2})e^{-s\tilde{\mu}_2}\big]}{\mathsf{E}_{\tilde{\mu}_2}\big[(\frac{1}{\tilde{\mu}_2})^{n-1}(1-\frac{\lambda}{\tilde{\mu}_2})e^{-s\tilde{\mu}_2}\big]}, \tag{40}$$

which implies that $s_1^*(n) \leq s_2^*(n)$. □

Example 1 (continued):

	$n=0$	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n\geq 6$
$s_1^*(n)$	2.4	1.9372	1.4775	1.0213	0.5676	0.1165	0
$s_2^*(n)$	2.4	1.9	1.4	0.9	0.4	0	0

Table 3: Value of $s_1^*(n)$ and $s_2^*(n)$

In Table 3, it can be seen that $s_1^*(n) \geq s_2^*(n)$ even as $\tilde{\mu}_1 \leq_{\mathsf{st}} \tilde{\mu}_2$. □

4.4 Abandonment and Hazard Rates

Finally, we consider the hazard rate of the residual sojourn time given by the posterior beliefs about the service rate, when information is updated by the service completion time s, and the number of the customers in the system on arrival n, and the underlying conditional distribution $f(\cdot|(s,n))$ (see (34)). The customer's residual sojourn time denoted by $W_o(s,n)$ is the summation of n exponential random variables with parameter u with density $f(\cdot|(s,n))$. Hence, its cumulative distribution function is

$$\begin{aligned}\mathsf{Pr}(W_o(s,n) \leq t) &= \int_\lambda^\infty \big(1-\sum_{k=0}^{n-1}\frac{e^{-yt}(yt)^k}{k!}\big)f(y|(s,n))\mathsf{d}y \\ &= \frac{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}^{n-1}}(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}\tau}\big(1-\sum_{k=0}^{n-1}\frac{e^{-\tilde{\mu}t}(\tilde{\mu}t)^k}{k!}\big)\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}^{n-1}}(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}\tau}\Big]},\end{aligned} \tag{41}$$

and the corresponding hazard rate is

$$H_o(t|(s,n)) = \frac{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}^{n-1}}(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}\tau}\frac{\tilde{\mu}e^{-\tilde{\mu}t}(\tilde{\mu}t)^{n-1}}{(n-1)!}\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}^{n-1}}(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}\tau}\sum_{k=0}^{n-1}\frac{e^{-\tilde{\mu}t}(\tilde{\mu}t)^k}{k!}\Big]}. \tag{42}$$

Following the same procedure given by (31)-(32), we can draw a similar conclusion on the hazard rates in observable queues. For empty queue, for $n=0$, $H_o(t|(s,n))$ is decreasing in t as customer is immediately in service. For $n \geq 1$, there exists a t_n such that $H_o(t|(s,n))$ is increasing for $t \in [0, t_n)$.

5 The Effect of Information Disclosures on Abandonment

From the analysis in Sections 3 and 4, it is evident that information structures in queues affect customer beliefs and actions quite perceptively. We now consider three different information disclosures: (i) Disclose no information. (Unobservable Queues) (ii) Disclose the number of the customers in the system upon her arrival, and (iii) Disclose the number of the customers in the system upon arrival first, and then a service completion time after her arrival.

In this section, we will explore how the above information disclosures influence abandonments. Let μ be the true rate of the server unknown to the customer. Let λ be the mean arrival rate of customers. We examine, the probability of an uninformed, learning customer to stay in line and consume service eventually, without balking or abandoning under different information disclosures.

No Information Disclosure. No information is often observed in an unobservable queue. Under these conditions, customers spend time waiting in the queue, and then abandon based on their own patience thresholds. The findings under information non-disclosure are consistent to the cases when the announcement information is unverifiable, meaningless, or ignored by the customers. See extensive literature on intentional vagueness and cheap talk in queue announcements beginning with Allon et al. (2011).

In Section 3, we showed in an unobservable queue, the Bayesian customer abandons the queue if the time she spent exceeds some threshold τ^*. In a Markovian queue, the expected sojourn time is exponential. Combining these two facts, for the unobservable queue, we can write the probability of abandoning for an uninformed customer.

We define P_{au} to the $\underline{a}$bandonment probability in an $\underline{u}$nobservable queue for the uninformed customer when no information is disclosed.

$$P_{au} \triangleq \mathsf{Pr}(\text{sojourn time} \geq \tau^*) = e^{-(\mu-\lambda)\tau^*}, \tag{43}$$

where τ^* is given by (5).

Disclosure of Queue Length Only. In some observable queues, the arriving customers can observe the length of the queue to make their joining decisions. Such systems in which queue length is disclosed, has been extensively analyzed since Naor (1969). The information structure in an observable queue whose queue length is disclosed, is also similar to the scenarios in which the queue is only partially observable or not observable at all (but the expected waiting time information is available). This similarity occurs because the customer decisions are essentially threshold-type join/balk decisions based on expected waiting times, even the queue length information is also converted to expected waiting time to make the join/balk decisions.

The key distinct feature of our paper is eliciting the additional value of learning contained in the queue length information. The length of the queue is an outcome of service speed, and hence is indicative of the service speed of the firm. The Bayesian customers use the queue length information to update their prior information to make their join/balk decisions.

Let P_{bq} be the balking probability of an arriving customer, when only queue information is disclosed.

$$P_{bq} \triangleq \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \cdot p_{bq}(n) \tag{44}$$

where $p_{bq}(n)$ indicates the conditional probability of balking when arriving at state n. Note that the number of the customers in the system under the steady-state follows the geometric distribution with parameter λ/μ. Clearly, the balking probability depends on the state of arrival or the number of customers in the queue n. In fact, from Theorem 4, we have the threshold joining probability.

$$p_{bq}(n) = \begin{cases} 0, & \text{if } n \leq n_o^*, \\ 1, & \text{if } n > n_o^*. \end{cases} \tag{45}$$

Plugging threshold result (45) in Equation (44), we get

$$P_{bq} = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \cdot p_{bq}(n) = \left(\frac{\lambda}{\mu}\right)^{n_o^*+1}. \tag{46}$$

Disclosure of Queue Length and Service Completion. Finally, we consider the abandonment probability when such decisions are made based on queue length on arrival and the first service completion observation.

Let P_{as} be the abandonment probability for the uninformed customer if the number of the customers in the system is first disclosed upon her arrival, and then followed by an observation of service completion time.

$$P_{as} \triangleq \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \cdot p_{as}(n) \tag{47}$$

where $p_{as}(n)$ is conditional probability of a Bayesian consumer eventually leaving the queue (either through balking or eventual abandonment), when seeing queue length n on arrival and staying in the system for at least s time units.

From Theorem 6, we have the customer abandons after time spent in system exceeds $s^*(n)$. Therefore,

$$p_{as}(n) = \mathsf{Pr}\Big(\text{the exponential random variable with parameter } \mu \geq s^*(n)\Big) = e^{-\mu s^*(n)}. \tag{48}$$

In the previous sections we explored the definition of balking and patience thresholds, n_o^* and $s^*(n)$ respectively. Using Equations (24) and (36), we have

$$s^*(n) = 0 \;\; \text{for } n > n_o^* + 1. \tag{49}$$

Plugging $s^*(n)$ from (49) into (48), we have

$$p_{as}(n) = \begin{cases} e^{-\mu s^*(n)}, & \text{if } n \leq n_o^* + 1, \\ 1, & \text{if } n > n_o^* + 1. \end{cases} \tag{50}$$

Now, we can plug the conditional probabilities from (50) into Equation (47), and get

$$P_{as} = \sum_{n=0}^{n_o^*+1} \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n e^{-s^*(n)\mu} + \left(\frac{\lambda}{\mu}\right)^{n_o^*+2}. \tag{51}$$

Comparison of Disclosure Structures

First, note that τ^* and n_o^* are independent of the true value of service rate μ. This property helps us compare the probabilities of customers abandoning when there is no disclosure vs. when the queue lengths are disclosed.

Lemma 5. *$\exists \mu_0$ such that $P_{bq} \leq P_{au}$ if $\mu \leq \mu_0$ and $P_{bq} > P_{au}$ if $\mu > \mu_0$.*

Lemma 5 implies that if customers are Bayesian, it is preferable to disclose queue length information when the service is slow. While this result appears counter-intuitive at first, this is related to how a Bayesian customer uses the information. Absent queue length information, Bayesian customers, stay in unobservable queues for a time period before abandoning. For a slow server, this might mean some customers may be quitting short waits. Note that a Bayesian customer factors the queue length information to update his waiting costs. Shorter queues provide more surplus to Bayesian learners as longer queues are more likely to form at a slow server. When the information is disclosed, it strengthens posterior on service speed, when queues are short. When queues are long, Bayes customer would have abandoned anyway.

Therefore, it is worthwhile to compare the abandonment actions of a Bayesian customer who joins with queue length information, and waits to learn more by seeing service completions. To this end, we represent a visualization of analytical results in Figure 4. We consider four different regions with μ increasing from (a) to (d).

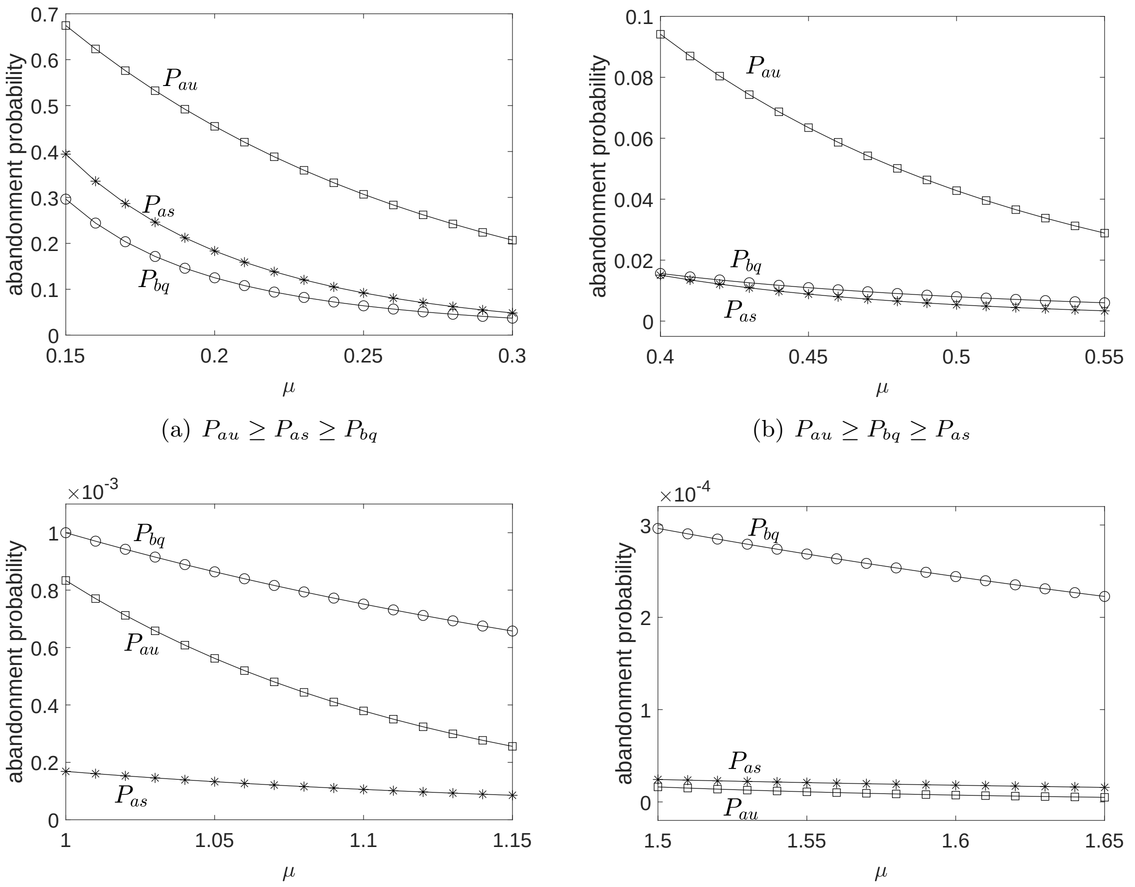
When service is slow (i.e., μ is small in Figures 4 (a) and (b)), P_{au} is the largest. This visualization underlines the finding of Lemma 5, that in slow queues, not disclosing information leads to high abandonments. As μ increases to sub-figures (c)-(d), the abandonment rates under all disclosures decrease significantly, but no-disclosure abandonment (P_{au}) is the lowest (especially when the server is fast). Fast services can see increased abandonments by disclosing more information: Customers who see long queues turn away unaware of the fast nature of the service.

In Figures 4(c) and 4(d), $P_{bq} > P_{au}$, which follows from Lemma 5 on fast servers. However, $P_{bq} > P_{as}$:[1] Just announcing queue information alone leads to higher abandonment than disclosing queue length *and* first service completion. This observation seems to imply that some consumers who balk on seeing a queue length, *must be joining* at the queue length, if some other future information is available. In fact, such reversal is precisely what occurs.

Corollary 1. *$p_{bq}(n) < p_{as}(n)$ for $n \leq n_o^*$ and $p_{bq}(n) \geq p_{as}(n)$ for $n \geq n_o^* + 1$.*

This result can be intuited by examining the conditional probabilities in Corollary 1. A Bayesian customer balking the queue $n_0^* + 1$ on queue length information alone, may decide to join with some probability, *if* there were future information that might overcome their priors. If the server is sufficiently fast, indeed first service completion can overcome their priors, and the consumers may go on to receive service.

[1] As μ increases, P_{au} and P_{as} are both small as in Figure 4 (d). The difference nevertheless exists. When $\mu >> \lambda$, the elapsed sojourn time in unobservable queue equals one service completion with almost probability one. The probability that the elapsed sojourn time equals to the service time is given by $(1 - \frac{\lambda}{\mu}) + (1 - \frac{\lambda}{\mu})\frac{\lambda}{\mu} \geq 1 - \frac{1}{225} \approx 1$ for $\mu \in [1.5, 1.65]$.



(a) $P_{au} \geq P_{as} \geq P_{bq}$

(b) $P_{au} \geq P_{bq} \geq P_{as}$

(c) $P_{bq} \geq P_{au} \geq P_{as}$

(d) $P_{bq} \geq P_{as} \geq P_{au}$

Figure 4: $\lambda = 0.1, c = 1, V = 15$, Prior belief from Eqn. (37): $\tau^* = 7.87$ and $n_o^* = 2$

5.1 Traffic and Abandonment

In the following, we test how the arrival rate λ affects the relation between P_{au}, P_{bq}, and P_{as}. In Table 4, we choose $c = 1$, $V = 15$, and use prior distribution (37). We consider different cases of true service rate μ.

In Table 4 when $\mu = 0.2$, we have $P_{au} \geq P_{as} \geq P_{bq}$. This implies in a congested queueing system, disclosing information is always better than hiding information. Learning through the information about the number of the customers in the system is more efficient than learning through service completions. In this case, the true service rate is 0.2, but the uninformed customer believes the average service rate is 0.265 (the mean of prior distribution (37) is 0.265). Thus if the uninformed customer is allowed to update her belief using service completion observation, she is more likely to abandon. The uninformed customer will realize that the true service rate is slower than expected. Therefore, disclosing only information about the number of the customers in the system is better than disclosing service completion time in this case.

In Table 4 when $\mu = 0.5$, we have $P_{au} \geq P_{as}$, $P_{au} \geq P_{bq}$. Again, disclosing information is better than hiding information. However, the efficiency of disclosing different information depends on the value of λ. For small λ, $P_{as} \geq P_{bq}$, while for large λ, $P_{as} \leq P_{bq}$. Disclosing service completion is more effective for large λ, since the system is very congested and a large number of the customers in the system may trigger more abandonment. In this case, letting customer know service speed through completions is crucial and will reduce abandonment probability.

λ	$\mu = 0.2$			$\mu = 0.5$		
	P_{bq}	P_{as}	P_{au}	P_{bq}	P_{as}	P_{au}
0.04	0.0080	0.0718	0.2016	0.0005	0.0012	0.0100
0.07	0.0429	0.1101	0.3051	0.0027	0.0026	0.0197
0.10	0.1250	0.1834	0.4549	0.0080	0.0054	0.0428
0.13	0.2746	0.3119	0.6601	0.0176	0.0104	0.1113
0.16	0.5120	0.5221	0.9064	0.0328	0.0182	0.4336

Table 4

When the service rate becomes public information, that is, it is not necessary for customers to learn about the service rate, Hassin and Roet-Green (2017) consider the probabilities for an arriving customer to join the system, balk from the system, or access the queue length information before deciding whether to join under the equilibrium situation when maximizing the service value minus the waiting cost and inspection cost.

6 Conclusion: Discussion of Customer Abandonments

In this paper, we have analyzed customers impatience and learning through an $M/M/1$ queue. Unlike the prior literature, we show that customers may abandon during their waiting as long as they keep learning the service rate. At the same time, we obtain that customers can be more patient in slower shorter queues than faster longer queues for observable queues. We prove, for both unobservable and observable queues, that customers who anticipate congested system situation become less patient by the stochastic comparison approach. To illustrate the value of customer learning, furthermore, we make some comparisons between the abandonment probabilities for different learning. Before concluding the paper, we discuss some modeling issues mainly from the customer abandonment literature point of view.

Costs: Waiting cost may be reasonably divided into two components: alternative waiting cost, reflecting the actual value of time and being approximately linear; psychological waiting cost, subjective feeling of impatience and being strictly convex. We consider linear waiting cost.

Consistency: Customers may never learn perfectly due to limited experiences, variation in time, prior belief, experience with other systems, etc. Thus, Mandelbaum and Shimkin

(2000) proposed *partially consistent equilibrium*, which is influenced by the actual system dynamics in some specific manner. Our model, in this perspective, considers a partially consistent, subjective virtual waiting time, that follows the true waiting time distribution (exponential distribution) but whose rate is influenced by customer's prior belief.

Hazard rate in the queueing model: Approaches including variable number of servers, varying arrival rates, priorities and randomly failing servers have been suggested for decreasing hazard rate during waiting time. Especially, a heavy-tailed service distribution ($M/G/m$ model) would lead to decreasing hazard rate. Our model relaxes the assumption on service rate information and allows uninformed customer to learn it during waiting. This information relaxation and learning process also lead to decreasing hazard rate of remaining waiting time.

Retrials: The decision to retry will affect customer abandonment decisions. Our model (like many one-shot models) assumes no individual memory in retrials. See Cui et al. (2018) for a model of rational retrial behavior.

Demand elasticity: Related to retrial and future decision, the virtual waiting time will affect not only customer abandonment decisions but also whether or not they are willing to approach the system in future. This could be accommodated within Mandelbaum and Shimkin (2000)'s model by appending some arrival cost, which leads to a new stabilized arrival rate.

Real-time decisions: In the classical models, assume customer abandonment time is determined upon arrival. We reinterpreted this as real time decisions. Specifically, while waiting, customers continuously consider whether to abandon immediately or wait further, provided that she is temporally consistent (cost parameters are not modified, but remaining virtual waiting time is obtained via the Bayesian rule). Our model studies the real-time decision. We study a customer who uses her elapsed waiting time to obtain the remaining virtual waiting time via Bayes' rule and then decides abandon or stay.

Although the abandonments are often formally modeled theoretically and empirically in the queueing and operations literature, not much is known on the effect of learning on customer abandonments. In this respect, our model can be viewed as a theoretical step to understand customer impatience and abandonments from information accrual and learning perspective. The next research steps could be a validation of the learning effects in abandonments, through data and laboratory analysis.

References

Akşin Z, Ata B, Emadi S M, Su C-L (2013) Structural estimation of callers' delay sensitivity in call centers. *Management Sci.* 59(12):2727–2746.

Allon G, Bassamboo A, Gurvich, I (2011). We will be right with you: Managing customer expectations with vague promises and cheap talk. *Oper. Res.* 59(6):1382–1394.

Altman E, Shimkin N (1998) Individual equilibrium and learning in processor sharing systems. *Oper. Res.* 46(6):776–784.

Ata B, Peng X (2018) An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. *Oper. Res.* 66(1):163-183.

Baccelli F, Boyer P, Hebuterne G (1984) Single-server queues with impatient customers. *Adv. Appl. Probab.* 16(3):887–905.

Barber D (2012) *Bayesian Reasoning and Machine Learning.* Cambridge University Press.

Bassamboo A and Randhawa R (2016) Scheduling homogeneous impatient customers. *Management Sci.* 62(7):2129–2147.

Batt R, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Sci.* 61(1):39–59.

Bolandifar E, DeHoratius N, Olsen T, Wiler J (2016) Modeling the behavior of patients who leave the ED without being seen. *Chicago Booth Research Paper.*

Chow YS, Teicher H (1997)*Probability Theory: Independence, Interchangeability, Martingales.* Springer, New York.

Cui SL, Veeraraghavan S (2016) Blind queues: The impact of consumer beliefs on revenues and congestion. *Management Sci.*, 62(12):3656-3672.

Cui SL, Su X, Veeraraghavan S (2018) A model of rational retrials in queues. *Working paper.*

Debo L, Veeraraghavan S (2014) Equilibrium in queues under unknown service times and service value. *Oper. Res.* 62(1):38–57.

Gross D, Shortle J, Thompson J, and Harris C (2008) *Fundamentals of Queueing Theory.* John Wiley & Sons.

Hassin R, Haviv M (1995) Equilibrium strategies for queues with impatient customers. *Oper. Res. Lett.* 17(1):41–45.

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems.* Kluwer Academic Publisher, Boston.

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

Haviv M, Ritov Y (2001) Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions. *Queueing Syst.* 38(4):495–508.

Kim J, Randhawa R, Ward A R (2018) Dynamic scheduling in a many-server, multiclass system: The role of customer impatience in large systems. *Manufacturing & Service Operations Management* 20(2):285–301

Mandelbaum A, Shimkin N (2000) A model for rational abandonments from invisible queues. *Queueing Syst.*. 36:141–173.

Mandelbaum A., Yechiali U (1983) Optimal entering rules for a customer with wait option at an M/G/1 queue. *Management Sci.* 29: 174–187.

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

Parkan C, Warren EH (1978) Optimal reneging decisions in a $G/M/1$ queue. *Decision Sciences.* 9(1):107–119.

Shaked M, Shanthikumar J (2007) *Stochastic Orders.* Springer, New York.

Shimkin N, Mandelbaum A (2004) Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Syst.* 47(1):117–146.

Wang C L (2016) On socially optimal queue length. *Management Sci.* 62(3):899-903.

Yao J, Maglaras C, Zeevi A (2016) Observational learning and abandonment in congested systems. *Working paper.*

Zohar E, Mandelbaum A, Shimkin N (2002) Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Sci.* 48(4):566–583.

Appendix: Proofs

Proof. **[of Lemma 1]:** Taking the first-order derivative of $\varphi(\tau)$, we have

$$\frac{\mathsf{d}\varphi(\tau)}{\mathsf{d}\tau} = -1 + \frac{1}{\left(\mathsf{E}_{\tilde{\mu}}\left[e^{-(\tilde{\mu}-\lambda)\tau}\right]\right)^2} \cdot \mathsf{E}_{\tilde{\mu}}\left[\frac{e^{-(\tilde{\mu}-\lambda)\tau}}{\tilde{\mu}-\lambda}\right] \cdot \mathsf{E}_{\tilde{\mu}}\left[(\tilde{\mu}-\lambda)e^{-(\tilde{\mu}-\lambda)\tau}\right]. \tag{A-1}$$

Note that

$$\begin{aligned}
&\mathsf{E}_{\tilde{\mu}}\left[\frac{e^{-(\tilde{\mu}-\lambda)\tau}}{\tilde{\mu}-\lambda}\right] \cdot \mathsf{E}_{\tilde{\mu}}\left[(\tilde{\mu}-\lambda)e^{-(\tilde{\mu}-\lambda)\tau}\right] \\
&= \mathsf{E}_{\tilde{\mu}}\left[\left(\sqrt{\frac{e^{-(\tilde{\mu}-\lambda)\tau}}{\tilde{\mu}-\lambda}}\right)^2\right] \cdot \mathsf{E}_{\tilde{\mu}}\left[\left(\sqrt{(\tilde{\mu}-\lambda)e^{-(\tilde{\mu}-\lambda)\tau}}\right)^2\right] \\
&\geq \left(\mathsf{E}_{\tilde{\mu}}\left[\sqrt{\frac{e^{-(\tilde{\mu}-\lambda)\tau}}{\tilde{\mu}-\lambda}} \cdot \sqrt{(\tilde{\mu}-\lambda)e^{-(\tilde{\mu}-\lambda)\tau}}\right]\right)^2 \\
&= \left(\mathsf{E}_{\tilde{\mu}}\left[e^{-(\tilde{\mu}-\lambda)\tau}\right]\right)^2,
\end{aligned} \tag{A-2}$$

where the Schwarz inequality (see Corollary 3 on page 105 in Chow and Teicher (1997)) is applied to obtain the inequality. Furthermore note that the Schwarz inequality becomes equality if and only if two random variables

$$\sqrt{\frac{e^{-(\tilde{\mu}-\lambda)\tau}}{\tilde{\mu}-\lambda}} \quad \text{and} \quad \sqrt{(\tilde{\mu}-\lambda)e^{-(\tilde{\mu}-\lambda)\tau}}$$

have a linear relationship. That is, only if there exists a constant C such that with probability one,

$$\sqrt{(\tilde{\mu}-\lambda)e^{-(\tilde{\mu}-\lambda)\tau}} = C\sqrt{\frac{e^{-(\tilde{\mu}-\lambda)\tau}}{\tilde{\mu}-\lambda}}, \tag{A-3}$$

then the inequality in (A-2) becomes equality. It follows from $e^{-(\tilde{\mu}-\lambda)\tau} \neq 0$ that (A-3) holds if and only if with probability one,

$$\tilde{\mu} = C + \lambda. \tag{A-4}$$

As $\tilde{\mu}$ is the uninformed customer's prior belief about the service rate, we know that (A-4) does not hold. Hence, (A-2) can be strengthened as

$$\mathsf{E}_{\tilde{\mu}}\left[\frac{e^{-(\tilde{\mu}-\lambda)\tau}}{\tilde{\mu}-\lambda}\right] \cdot \mathsf{E}_{\tilde{\mu}}\left[(\tilde{\mu}-\lambda)e^{-(\tilde{\mu}-\lambda)\tau}\right] > \left(\mathsf{E}_{\tilde{\mu}}\left[e^{-(\tilde{\mu}-\lambda)\tau}\right]\right)^2. \tag{A-5}$$

Combining (A-1) and (A-5) yields that $\frac{\mathsf{d}\varphi(\tau)}{\mathsf{d}\tau} > 0$. Hence, we have the strictly increasing property of $\varphi(\cdot)$. □

Proof. **[of Lemma 2]:** To prove the lemma, it is sufficient to show the following inequality

$$\frac{\mathsf{E}_{\tilde{\mu}}\left[\frac{1}{\tilde{\mu}}(\frac{\lambda}{\tilde{\mu}})^{n+1}(1-\frac{\lambda}{\tilde{\mu}})\right]}{\mathsf{E}_{\tilde{\mu}}\left[(\frac{\lambda}{\tilde{\mu}})^{n+1}(1-\frac{\lambda}{\tilde{\mu}})\right]} > \frac{\mathsf{E}_{\tilde{\mu}}\left[\frac{1}{\tilde{\mu}}(\frac{\lambda}{\tilde{\mu}})^{n}(1-\frac{\lambda}{\tilde{\mu}})\right]}{\mathsf{E}_{\tilde{\mu}}\left[(\frac{\lambda}{\tilde{\mu}})^{n}(1-\frac{\lambda}{\tilde{\mu}})\right]} \quad \text{for any } n \geq 0. \tag{A-6}$$

Because each numerator in the above two fractions is positive, (A-6) is equivalent to

$$\begin{aligned} &\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}}\Big(\frac{\lambda}{\tilde{\mu}}\Big)^{n+1}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big]\cdot\mathsf{E}_{\tilde{\mu}}\Big[\Big(\frac{\lambda}{\tilde{\mu}}\Big)^{n}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big] \\ &\quad > \mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}}\Big(\frac{\lambda}{\tilde{\mu}}\Big)^{n}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big]\cdot\mathsf{E}_{\tilde{\mu}}\Big[\Big(\frac{\lambda}{\tilde{\mu}}\Big)^{n+1}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big]. \end{aligned} \tag{A-7}$$

We divide both sides of (A-7) by λ^{2n+1}. Then (A-7) will be further simplified to

$$\mathsf{E}_{\tilde{\mu}}\Big[\Big(\frac{1}{\tilde{\mu}}\Big)^{n+2}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big]\cdot\mathsf{E}_{\tilde{\mu}}\Big[\Big(\frac{1}{\tilde{\mu}}\Big)^{n}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big] > \Big(\mathsf{E}_{\tilde{\mu}}\Big[\Big(\frac{1}{\tilde{\mu}}\Big)^{n+1}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big]\Big)^2. \tag{A-8}$$

By again the Schwarz inequality, we have

$$\begin{aligned} &\mathsf{E}_{\tilde{\mu}}\Big[\Big(\frac{1}{\tilde{\mu}}\Big)^{n+2}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big]\cdot\mathsf{E}_{\tilde{\mu}}\Big[\Big(\frac{1}{\tilde{\mu}}\Big)^{n}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big] \\ &= \mathsf{E}_{\tilde{\mu}}\left[\left(\sqrt{\Big(\frac{1}{\tilde{\mu}}\Big)^{n+2}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)}\right)^2\right]\cdot\mathsf{E}_{\tilde{\mu}}\left[\left(\sqrt{\Big(\frac{1}{\tilde{\mu}}\Big)^{n}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)}\right)^2\right] \\ &\geq \Big(\mathsf{E}_{\tilde{\mu}}\Big[\sqrt{\Big(\frac{1}{\tilde{\mu}}\Big)^{n+2}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)}\cdot\sqrt{\Big(\frac{1}{\tilde{\mu}}\Big)^{n}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)}\Big]\Big)^2 \\ &= \Big(\mathsf{E}_{\tilde{\mu}}\Big[\Big(\frac{1}{\tilde{\mu}}\Big)^{n+1}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big]\Big)^2. \end{aligned} \tag{A-9}$$

Similar to the argument in proving the strict inequality in (A-2), we can show

$$\mathsf{E}_{\tilde{\mu}}\Big[\Big(\frac{1}{\tilde{\mu}}\Big)^{n+2}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big]\cdot\mathsf{E}_{\tilde{\mu}}\Big[\Big(\frac{1}{\tilde{\mu}}\Big)^{n}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big] > \Big(\mathsf{E}_{\tilde{\mu}}\Big[\Big(\frac{1}{\tilde{\mu}}\Big)^{n+1}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big]\Big)^2.$$

Hence (A-6) is proved. □

Proof. **[of Proposition 1]:** Let

$$X(\tilde{\mu}) := \frac{1}{\tilde{\mu}} \text{ and } Y_n(\tilde{\mu}) := \Big(\frac{1}{\tilde{\mu}}\Big)^n\Big(1-\frac{\lambda}{\tilde{\mu}}\Big). \tag{A-10}$$

Consider the following difference

$$\begin{aligned} &V - c(n+1)\mathsf{E}_{\tilde{\mu}}\Big(\frac{1}{\tilde{\mu}}\Big) - \Big(V - c(n+1)\frac{\mathsf{E}_{\tilde{\mu}}\left[\frac{1}{\tilde{\mu}}(\frac{\lambda}{\tilde{\mu}})^{n}(1-\frac{\lambda}{\tilde{\mu}})\right]}{\mathsf{E}_{\tilde{\mu}}\left[(\frac{\lambda}{\tilde{\mu}})^{n}(1-\frac{\lambda}{\tilde{\mu}})\right]}\Big) \\ &= c(n+1)\frac{\mathsf{E}_{\tilde{\mu}}\left[(\frac{1}{\tilde{\mu}})^{n+1}(1-\frac{\lambda}{\tilde{\mu}})\right] - \mathsf{E}_{\tilde{\mu}}\Big(\frac{1}{\tilde{\mu}}\Big)\cdot\mathsf{E}_{\tilde{\mu}}\left[(\frac{1}{\tilde{\mu}})^{n}(1-\frac{\lambda}{\tilde{\mu}})\right]}{\mathsf{E}_{\tilde{\mu}}\left[(\frac{1}{\tilde{\mu}})^{n}(1-\frac{\lambda}{\tilde{\mu}})\right]} \end{aligned}$$

$$= c(n+1)\frac{\mathsf{E}_{\tilde{\mu}}[X(\tilde{\mu})Y_n(\tilde{\mu})] - \mathsf{E}_{\tilde{\mu}}[X(\tilde{\mu})] \cdot \mathsf{E}_{\tilde{\mu}}[Y_n(\tilde{\mu})]}{\mathsf{E}_{\tilde{\mu}}[Y_n(\tilde{\mu})]}. \tag{A-11}$$

Note that

$$\begin{aligned}
&\mathsf{E}_{\tilde{\mu}}[X(\tilde{\mu})Y_n(\tilde{\mu})] - \mathsf{E}_{\tilde{\mu}}[X(\tilde{\mu})] \cdot \mathsf{E}_{\tilde{\mu}}[Y_n(\tilde{\mu})] \\
&\quad = \mathsf{E}\Big[\Big(X(\tilde{\mu}) - \mathsf{E}_{\tilde{\mu}}[X(\tilde{\mu})]\Big)\Big(Y_n(\tilde{\mu}) - \mathsf{E}_{\tilde{\mu}}[Y_n(\tilde{\mu})]\Big)\Big] \\
&\quad = \mathsf{cov}(X(\tilde{\mu}), Y_n(\tilde{\mu})).
\end{aligned}$$

To prove the proposition, it suffices to show that

[A] There do not exist two nonnegative integer numbers n_1 and n_2 with $n_1 < n_2$ such that $\mathsf{cov}\big(X(\tilde{\mu}), Y_{n_1}(\tilde{\mu})\big) = \mathsf{cov}\big(X(\tilde{\mu}), Y_{n_2}(\tilde{\mu})\big) = 0$;

[B] If $\mathsf{cov}\big(X(\tilde{\mu}), Y_{n_0}(\tilde{\mu})\big) > 0$, then for any $n > n_0$, $\mathsf{cov}\big(X(\tilde{\mu}), Y_n(\tilde{\mu})\big) > 0$;

[C] If $\mathsf{cov}\big(X(\tilde{\mu}), Y_{n_0}(\tilde{\mu})\big) < 0$, then for any $n < n_0$, $\mathsf{cov}\big(X(\tilde{\mu}), Y_n(\tilde{\mu})\big) < 0$.

To prove **[A]**, suppose contrariwise that there exist n_1 and n_2 with $n_1 < n_2$ such that $\mathsf{cov}\big(X(\tilde{\mu}), Y_{n_1}(\tilde{\mu})\big) = \mathsf{cov}\big(X(\tilde{\mu}), Y_{n_2}(\tilde{\mu})\big) = 0$. This is equivalent to

$$\frac{\mathsf{E}_{\tilde{\mu}}[X(\tilde{\mu})Y_{n_1}(\tilde{\mu})]}{\mathsf{E}_{\tilde{\mu}}[Y_{n_1}(\tilde{\mu})]} = \frac{\mathsf{E}_{\tilde{\mu}}[X(\tilde{\mu})Y_{n_2}(\tilde{\mu})]}{\mathsf{E}_{\tilde{\mu}}[Y_{n_2}(\tilde{\mu})]}. \tag{A-12}$$

By the definition of $X(\tilde{\mu})$ and $Y_{n_1}(\tilde{\mu})$, the above equation is same as

$$\frac{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{1}{\tilde{\mu}})^{n_1+1}(1-\frac{\lambda}{\tilde{\mu}})\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{1}{\tilde{\mu}})^{n_1}(1-\frac{\lambda}{\tilde{\mu}})\Big]} = \frac{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{1}{\tilde{\mu}})^{n_2+1}(1-\frac{\lambda}{\tilde{\mu}})\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{1}{\tilde{\mu}})^{n_2}(1-\frac{\lambda}{\tilde{\mu}})\Big]}. \tag{A-13}$$

However, (A-13) contradicts with Lemma 2. Thus we get **[A]**.

Now prove **[B]**. Note that by $\mathsf{cov}\big(X(\tilde{\mu}), Y_{n_0}(\tilde{\mu})\big) > 0$,

$$\frac{\mathsf{E}_{\tilde{\mu}}[X(\tilde{\mu})Y_{n_0}(\tilde{\mu})]}{\mathsf{E}_{\tilde{\mu}}[Y_{n_0}(\tilde{\mu})]} = \frac{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{1}{\tilde{\mu}})^{n_0+1}(1-\frac{\lambda}{\tilde{\mu}})\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{1}{\tilde{\mu}})^{n_0}(1-\frac{\lambda}{\tilde{\mu}})\Big]} > \mathsf{E}_{\tilde{\mu}}[X(\tilde{\mu})]$$

Hence **[B]** directly follows from Lemma 2. The proof of **[C]** is similar. Here we omit it.

For $n = 0$,

$$\begin{aligned}
\mathsf{cov}(X(\tilde{\mu}), Y_0(\tilde{\mu})) &= \mathsf{E}_{\tilde{\mu}}[X(\tilde{\mu})Y_0(\tilde{\mu})] - \mathsf{E}_{\tilde{\mu}}[X(\tilde{\mu})] \cdot \mathsf{E}_{\tilde{\mu}}[Y_0(\tilde{\mu})] \\
&= \mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}}\Big(1-\frac{\lambda}{\tilde{\mu}}\Big)\Big] - \mathsf{E}_{\tilde{\mu}}\frac{1}{\tilde{\mu}} \cdot \mathsf{E}_{\tilde{\mu}}\Big[1-\frac{\lambda}{\tilde{\mu}}\Big] \\
&= \lambda\Big[\Big(\mathsf{E}_{\tilde{\mu}}\frac{1}{\tilde{\mu}}\Big)^2 - \mathsf{E}_{\tilde{\mu}}\Big(\frac{1}{\tilde{\mu}}\Big)^2\Big] \\
&= -\lambda \cdot \mathsf{Var}\Big(\frac{1}{\tilde{\mu}}\Big) < 0, \qquad \text{(A-14)}
\end{aligned}$$

that is, $X(\tilde{\mu})$ and $Y_0(\tilde{\mu})$ are negative correlated. Hence,

$$\left(V - c(n+1)\mathsf{E}_{\tilde{\mu}}\Big(\frac{1}{\tilde{\mu}}\Big) - \Big(V - c(n+1)\frac{\mathsf{E}_{\tilde{\mu}}\Big[\frac{1}{\tilde{\mu}}(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\Big]}\right)_{n=0} < 0. \tag{A-15}$$

However it is impossible to always have (A-15) for all n. For example, suppose $\tilde{\mu}$ follows a two-point distribution.

$$\tilde{\mu} = \begin{cases} 1 & \text{with probability } p, \\ 3 & \text{with probability } 1-p. \end{cases}$$

$\lambda = 1/2$. Then $\mathsf{cov}(X(\tilde{\mu}), Y_n(\tilde{\mu})) = p(1-p)(\frac{1}{3} - \frac{5}{3^{n+2}}) > 0$ for any $n \geq 1$. Hence, for this concrete example, (A-15) holds only for $n = 0$. □

Proof. **[of Lemma 3]:** We first show that $\mathcal{U}_o(s,n)$ is strictly decreasing in s. Taking the derivative with respect to s,

$$\frac{\partial \mathcal{U}_o(s,n)}{\partial s} = c(n-1) - cn\frac{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}s}\Big] \cdot \mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}^2 e^{-\tilde{\mu}s}\Big]}{\Big(\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}e^{-\tilde{\mu}s}\Big]\Big)^2}. \tag{A-16}$$

Using the Schwarz inequality, we have

$$\begin{aligned}
&\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}s}\Big] \cdot \mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}^2 e^{-\tilde{\mu}s}\Big] \\
&= \mathsf{E}_{\tilde{\mu}}\Big[\Big(\sqrt{(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}s}}\Big)^2\Big] \cdot \mathsf{E}_{\tilde{\mu}}\Big[\Big(\sqrt{(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}^2 e^{-\tilde{\mu}s}}\Big)^2\Big] \\
&\geq \Big(\mathsf{E}_{\tilde{\mu}}\Big[\sqrt{(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}s}} \cdot \sqrt{(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}^2 e^{-\tilde{\mu}s}}\Big]\Big)^2 \\
&= \Big(\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}e^{-\tilde{\mu}s}\Big]\Big)^2.
\end{aligned}$$

Thus

$$\begin{aligned}
\frac{\partial \mathcal{U}_o(s,n)}{\partial s} &= c(n-1) - cn\frac{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}s}\Big] \cdot \mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}^2 e^{-\tilde{\mu}s}\Big]}{\Big(\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}e^{-\tilde{\mu}s}\Big]\Big)^2} \\
&\leq c(n-1) - cn \\
&= -c < 0.
\end{aligned}$$

This implies the strict monotonicity of $\psi(s,n)$ with respect to s.

Next we show that $\mathcal{U}_o(s,n)$ is strictly decreasing in n. Because $\tilde{\mu} \geq \lambda$, it's sufficient to show that

$$\frac{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}s}\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}e^{-\tilde{\mu}s}\Big]}$$

is increasing in n, that is

$$\frac{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^{n+1}(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}s}\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^{n+1}(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}e^{-\tilde{\mu}s}\Big]} \geq \frac{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}s}\Big]}{\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^n(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}e^{-\tilde{\mu}s}\Big]}. \tag{A-17}$$

(A-17) is equivalent to

$$\begin{aligned}&\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^{n+1}(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}s}\Big]\cdot\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^{n}(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}e^{-\tilde{\mu}s}\Big]\\&\quad\geq\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^{n+1}(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}e^{-\tilde{\mu}s}\Big]\cdot\mathsf{E}_{\tilde{\mu}}\Big[(\frac{\lambda}{\tilde{\mu}})^{n}(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}s}\Big].\end{aligned}\tag{A-18}$$

Using the Schwarz inequality and following the line of from (A-7) to (A-9), (A-18) can be proved. We omit the details here. □

Proof. **[of Lemma 4]:** Note that

$$\begin{aligned}\frac{\partial\hat{\mathcal{U}}_o(s,n)}{\partial s}&=cn-cn\frac{\mathsf{E}\Big[(\frac{\lambda}{\tilde{\mu}})^{n}(1-\frac{\lambda}{\tilde{\mu}})e^{-\tilde{\mu}s}\Big]\cdot\mathsf{E}\Big[(\frac{\lambda}{\tilde{\mu}})^{n}(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}^2e^{-\tilde{\mu}s}\Big]}{\Big(\mathsf{E}\Big[(\frac{\lambda}{\tilde{\mu}})^{n}(1-\frac{\lambda}{\tilde{\mu}})\tilde{\mu}e^{-\tilde{\mu}s}\Big]\Big)^2}\\&\leq cn-cn\times 1=0.\end{aligned}$$

Using final steps similar to the proof of Lemma 3, we can show that the result holds. □

Proof. **[of Lemma 5]:** For any fixed τ^* and n_o^*, $P_{au}\geq P_{bq}$ is equivalent to

$$e^{-(\mu-\lambda)\tau^*}\geq\Big(\frac{\lambda}{\mu}\Big)^{n_o^*+1}.$$

Denote

$$l(\mu):=e^{-(\mu-\lambda)\tau^*},\quad r(\mu):=\Big(\frac{\lambda}{\mu}\Big)^{n_o^*+1}.$$

It can be shown that there are at most two intersection points on interval (λ,∞). The first intersection can be easily verified to be at λ. The second one, denoted by μ_0, is larger than λ. The later result directly follows from the identification of the monotonicity region of $l(\mu)/r(\mu)$ on $[\lambda,\infty)$ by taking its derivatives:

$$\mathsf{d}\Big(\frac{l(\mu)}{r(\mu)}\Big)\Big/\mathsf{d}\mu=e^{-(\mu-\lambda)\tau^*}\frac{1}{\lambda}\Big(\frac{\mu}{\lambda}\Big)^{n_o^*}\Big(n_0^*+1-\tau^*\mu\Big).$$

Thus, $P_{bq}\leq P_{au}$ if $\mu\leq\mu_0$ and $P_{bq}>P_{au}$ if $\mu>\mu_0$. □

Proof. **[of Corollary 1:** Proof directly follows from comparing the threshold properties of the conditional probabilities p_{bq} and p_{as} in equations (45)-(50). □