Signaling Quality via Queues

Laurens G. Debo*  Christine A. Parlour†
Booth School of Business  Haas School of Business
University of Chicago  UC Berkeley
Chicago, IL 60637  Berkeley, CA 94720

Uday Rajan‡
Ross School of Business
University of Michigan
Ann Arbor, MI 48109.

April 6, 2010

*Tel: (773) 702-7140, E-mail: laurens.debo@chicagobooth.edu
†Tel: (510) 643-9391, E-mail: parlour@haas.berkeley.edu
‡Tel: (734) 764-2310, E-mail: urajan@umich.edu
Signaling Quality via Queues

Abstract

We consider a firm’s choice of service rate in the following environment. The firm may have high or low quality, and sells a good to consumers who are heterogeneously informed. Consumers arrive according to a Poisson process and are serviced in a random period of time. If a consumer arrives when another consumer is being serviced, he must join a queue. Consumers observe the length of the queue before making their purchasing decision. The firm may choose a fast or slow service rate. A faster rate requires a costly investment in technology. We show that, in equilibrium, informed consumers join the queue if it is below a threshold. The threshold varies with the quality of the good, so an uninformed consumer updates her belief about quality on observing the length of the queue. The strategy of an uninformed consumer has a “hole”: she joins the queue at lengths both below and above the hole, but not at the hole itself. When all consumers are informed, the high-quality firm has a greater incentive to speed up than the low-quality firm. However, the high-quality firm selects a slower service rate than the low-quality firm if there are a lot of queue lengths between the hole in an uninformed consumer’s strategy and the threshold at which informed consumers balk from its queue. Strikingly, if the proportion of informed consumers is low, the high-quality firm may choose the slow service rate even if the technological cost of speeding up is zero. The queue can therefore be a valuable signaling device for a high-quality firm.

Key Words: Queuing Game, Quality Signaling, Service Rate Selection, Strategic Consumer Behavior.
1 Introduction

Consumers frequently have to wait before they can consume a product or service. Lines outside nightclubs, rides at amusement parks and waiting lists for new products are part of everyone’s experience. Toys and innovative products, whose value cannot easily be communicated, exhibit similar phenomena. For example, Cabbage Patch Kids in 1983 and Beanie Babies in the 1990s had significant waiting times, and queues formed in front of stores when these products came on the market. The prevalence of queuing in order to consume begs the question: Do firms have a strategic incentive to manipulate impatient customers’ waiting times to generate more demand?

We develop a model in which a firm sells a good that can be of either high or low quality. Impatient consumers arrive at the market according to a Poisson process. Purchasing the good entails joining a “first-come, first-served” queue. The firm controls the distribution of the queue length by choosing the rate at which it services customers. Some consumers are informed, and know the quality of the good. Others are uninformed, and must infer the quality from the length of the queue. An arriving consumer does not observe the entire history of the game and so does not know how many people arrived before her. She also cannot observe the firm’s service rate choice; she only sees the queue.

The firm chooses either a fast or a slow service rate. We analyze the equilibrium service rate strategies for each of the high and low-quality firms. In particular, we characterize conditions under which the high-quality firm would prefer to slow down (i.e., choose the slow service rate). If the feasible service rates are close to the arrival rate, and the proportion of informed consumers is small, the high-quality firm chooses the slow service rate, even if there is no technological cost to speeding up. Essentially, it uses the length of the queue to signal its quality to uninformed consumers, thereby increasing demand.

As is standard in most queuing models, informed consumers adhere to a “threshold” strategy, where the threshold depends on the quality of the good. That is, they join the queue if it is short enough, and balk at longer queues. In contrast, uninformed agents, who have to infer the quality of the firm from the queue, play a non-threshold strategy. Their Bayesian updating process leads to a “hole” in their joining decision. That is, there is exactly one queue length at which they balk, which lies between the thresholds at which informed consumers balk when faced with low- and high-quality firms. Importantly, they join the queue at both smaller and larger queue lengths. Since no uninformed consumer joins the queue at a hole, longer queues can only arise if an informed consumer who knows the firm has high quality were to join at that length. Hence, an uninformed consumer
arriving when the queue is above the hole infers that the firm has high quality. The hole thus serves as a natural filter of quality information for uninformed consumers.

The hole in the uninformed consumer’s strategy affects the incentives of each type of firm to speed up (i.e., choose the fast service rate) in different ways. We assume each type of firm has the same profit margin and maximizes the revenue rate per unit time. For the low-quality firm, the queue never rises above the hole, since informed consumers balk at a threshold weakly lower than the position of the hole. Thus, the low-quality firm prefers to keep the queue at lengths below the hole, and naturally has an incentive to choose the fast service rate. The high-quality firm, on the other hand, loses uninformed consumers at the hole, but wins them back at higher queue lengths. If the proportion of informed consumers is high, the high-quality also has an incentive to select a fast service process. However, if there is a large proportion of uninformed consumers, the high-quality firm wishes to avoid having the queue stall at the hole. Enabling the queue to cross the hole allows it to capture the uninformed demand that exists at longer queue lengths. The high-quality firm can do this by selecting the slow service process, even if increasing the service rate comes at a low (or no) cost.

Our paper makes the following contributions. First, we highlight the role of a queue in the process by which uninformed consumers learn about product quality. As we show, inferences from a queue are subtle. If the queue is sufficiently long, uninformed consumers rightly deduce the product has high quality. However, at shorter queue lengths, the posterior probability that a product has high quality may sometimes decrease in the queue length.

Second, we show that the informational role played by the queue provides a firm with an incentive to slow down service. More broadly, the queue in our model is simply a device via which a firm communicates information to uninformed consumers about the strength of demand, and hence about their own valuation for the product. With new products, the fraction of informed consumers is often low. There is then a hidden cost to capacity expansion: It reduces information transmission to uninformed consumers. As a result, it may be better to gradually build up sales of the product over time. Such phased roll-outs are common with, for example, small budget movies. In the latter case, capacity is determined by the number of screens the movie is showing on, and is often kept low in the first few weeks to allow information to disseminate to uninformed consumers.

It is natural to ask whether a firm might adjust its price to signal the quality of its product. Bagwell and Riordan (1991) provide a model in which a low-quality firm can always earn a positive profit (even when it is recognized to have low quality), and the demand for a high-quality firm decreases with price. In such a setting, they show that there
are equilibria in which the high-quality firm signals with a high price.

However, in many contexts, firms may have limited ability to provide other signals about quality. For example, Becker (1991) argues that, if a consumer’s utility is increasing in the total number of consumers buying a product, a producer such as a restaurant or Broadway show will prefer to have a low price and excess demand, rather than raise the price. When the price cannot be used as a signal and consumers have imperfect information, Banerjee (1992) and Bikhchandani, Hirshleifer and Welch (1992) show that herd behavior may result. That is, consumers who know the entire history of the game may prefer to ignore their own signals and imitate other consumers before them. In our model, such behavior cannot arise since consumers are either fully informed or fully uninformed. We postulate that, if the good is an experience good that each consumer only purchases once, it is difficult to use price as a signal. Further, if the proportion of informed consumers is low, it is not too costly for a low-quality firm to mimic a high-quality one on price.

Similarly, other signals about quality, such as directly reporting the service rate, may also be infeasible. For example, if a firm reports its service rate, the low-quality firm has an incentive to lie and make the same report as the high-quality firm, so the problem is substantially the same as it is in this paper. However, if the service rate is verifiable, a separating equilibrium may exist in which the quality is known to all consumers. Thus, our model applies primarily to the case in which the service rate is not verifiable.

In the queueing literature, equilibrium joining strategies are examined by Naor (1969) and the subsequent literature on economic aspects of queueing (see Hassin and Haviv, 2003, for an excellent overview). When there are positive waiting costs, agents play a threshold strategy: they join the queue as long as it is not too long. Beyond some threshold, there is a congestion effect: the waiting costs imply that joining the queue is not worthwhile. The service quality is typically assumed to be common knowledge. The literature on service rate decisions with observable queues is sparse (see Hassin and Haviv, Chapter 8). Hassin (1986) considers a firm choosing the profit-maximizing number of single-server facilities, and finds that forcing a firm to make its queue length unobservable reduces social welfare if an observable queue would increase the firm’s profit. Debo and Veeraraghavan (2009) show that, in a setting with only uninformed consumers, a high-quality firm that faces a greater cost to speeding up will choose a lower service rate than a low-quality firm.

We set up our model in Section 2. The consumer strategy is analyzed in Section 3, and Section 4 considers the selection of service rate by each type of firm. Section 5 provides some concluding remarks. All proofs are contained in the appendix.
2 Model

We consider the market for an experience good, which may be a physical product or a service. The good may have a high (\( h \)) or low (\( \ell \)) quality. In addition, firms choose the speed at which they provide the good to their customers (“the service rate”).

The sequence of events is as follows. First, nature chooses the type of the firm which is privately observed by the firm. The firm then chooses a service rate \( \mu \in \{\mu, \bar{\mu}\} \). The mean service time per consumer is \( \frac{1}{\mu} \). Without loss of generality, we normalize the cost of choosing rate \( \mu \) to zero; and let \( k \geq 0 \) be the technological cost per unit of time of choosing rate \( \bar{\mu} \). We assume that \( k \) does not depend on the quality of the firm and represents a non-verifiable expenditure of resources. Once chosen, the service rate is held constant for the rest of the game. For \( \theta \in \{h, \ell\} \), let \( \beta_\theta \) denote the probability that a firm of type \( \theta \) chooses service rate \( \bar{\mu} \). The strategy of the firm may then be represented as \( \beta = (\beta_h, \beta_\ell) \).

Risk-neutral consumers arrive at the market according to a Poisson process with parameter \( \Lambda \), after the firm has chosen its service rate. A proportion \( q \) of consumers is informed, and knows the quality of the good. The remaining proportion \( 1-q \) is uninformed. Uninformed consumers have a prior belief that the good is high quality with probability \( p \). The utility an agent obtains from purchasing and consuming a good of quality \( \theta \) is \( v_\theta \), with \( v_h > v_\ell \). This utility is net of the good’s price, which is not explicitly modeled.

The firm cannot communicate its service rate to the consumers. Hence, upon arrival a consumer does not observe the service rate of the firm but does observe the number of people waiting to be served by the firm. If agents arrive faster than they are serviced, they form a queue. The queue is served on a first-come first-served basis. A consumer suffers a disutility \( c > 0 \) per unit of time that she has to wait to obtain the good, starting from her initial arrival to the market.

Each consumer takes an action \( a \in \{\text{join}, \text{balk}\} \), where \( a = \text{join} \) is the decision to acquire the good, or to join the queue. Once she joins the queue, she may not renege; i.e., she cannot leave until she has been served. Joining the queue is therefore synonymous with consuming the good. If she chooses \( a = \text{balk} \) (i.e., not acquire the good), she obtains a reservation utility of zero. Thus, each consumer will join the queue only if the expected utility from joining exceeds zero.

We assume that an informed consumer obtains a positive utility from consuming the low-quality good whenever she finds no one else ahead of her, even if the firm has chosen the slow service rate. This ensures that the low-quality firm can earn a strictly positive profit. Otherwise, the low-quality firm may exit the market, in which case consumers would know
any good offered was of high quality.

**Assumption 1** \( v_\ell > \frac{c}{\mu} \).

In what follows, we characterize Markov perfect Bayesian equilibria of the game. In outlining the model, for generality we consider mixed strategies for both the firm and consumers. In the bulk of our analysis, however, we restrict attention to the case in which each firm type plays a pure strategy; that is, \( \beta_\theta \in \{0, 1\} \) for each \( \theta \). In the consumer continuation game (i.e., once the firm has chosen its service rate), we consider equilibria in which the strategy of a consumer depends only on whether she is informed and on the length of the queue when she arrives at the market, but not on the exact time at which she arrives or on the number of consumers that have preceded her. Thus, the equilibria are both symmetric across consumers and time-invariant.

### 2.1 Consumers’ Objective Functions

Consider an informed consumer who knows the firm has high quality. Suppose the high-quality firm chooses the fast service rate \( \bar{\mu} \). Suppose further that there are already \( n \) consumers in the queue (including the consumer currently being served) when the informed consumer arrives at the market. Suppose she joins the queue. Due to the memory-less property of the exponential distribution, the expected service time for each customer is \( \frac{1}{\mu} \). Therefore, the total expected time to service the \( n \) consumers already in the queue and the new arrival is \( \frac{n+1}{\mu} \). Hence, her expected utility from joining the queue is \( v_h - (n + 1)\frac{c}{\mu} \).

Define \( \bar{N} = \lfloor v_h \frac{\bar{\mu}}{c} \rfloor \), where \( \lfloor x \rfloor \) is the largest integer less than or equal to \( x \). Then, if there are already \( \bar{N} \) people waiting in the queue, an informed consumer who knows the firm is of high quality will not join the queue. Of course, if the high firm chooses the slow service rate \( \underline{\mu} \), or the firm has low quality, an informed consumer may balk at smaller queue lengths as well. Since no informed consumer will join the queue at \( \bar{N} \) (or higher), regardless of the quality of the firm or the service rate it chooses, no uninformed consumer will join at this length either (or at higher lengths). Hence, \( \bar{N} \) is the largest queue length that can be observed in any equilibrium of the game. Let \( \mathcal{N} = \{0, 1, \cdots, \bar{N}\} \).

In general, a consumer may play a mixed strategy, and join the queue with some probability between 0 and 1. Thus, a mixed strategy for an informed agent is a mapping \( \sigma_i : \{h, \ell\} \times \mathcal{N} \rightarrow [0, 1] \), and a mixed strategy for an uninformed agent is a mapping \( \sigma_u : \mathcal{N} \rightarrow [0, 1] \). The overall consumer strategy profile may then be represented as \( \sigma = (\sigma_i, \sigma_u) \).
By contrast, uninformed consumers do not know the firms’ quality choice. First, consider the expected utility of the informed consumers. Let $\theta \in \{h, \ell\}$, let $\hat{\mu}(\beta_\theta) = \frac{1}{\beta_\theta + (1-\beta_\theta)\frac{1}{\beta}}$ be the weighted harmonic mean of the service rate distribution chosen by the firm. Given the firm’s strategy, the expected time to service each consumer is $\frac{1}{\hat{\mu}(\beta_\theta)}$. The expected utility of an informed consumer if she joins a queue containing $n$ consumers is therefore

$$w_i(n, \theta, \beta) = v_\theta - (n + 1)c \frac{1}{\hat{\mu}(\beta_\theta)}.$$ (1)

An uninformed consumer who finds $n$ consumers waiting in line ascribes a probability $\gamma(n)$ to the firm being of high quality. Thus, she expects the waiting time per consumer to be $\gamma(n)\frac{1}{\hat{\mu}(\beta_h)} + (1-\gamma(n))\frac{1}{\hat{\mu}(\beta_\ell)}$. Her expected utility if she joins the queue when there are already $n$ consumers waiting in line is

$$w_u(n, \gamma, \beta) = v_\ell + \gamma(n)(v_h - v_\ell) - (n + 1)c \left\{ \frac{1}{\hat{\mu}(\beta_h)} + \gamma(n) \left( \frac{1}{\hat{\mu}(\beta_h)} - \frac{1}{\hat{\mu}(\beta_\ell)} \right) \right\}.$$ (2)

### 2.2 Firm’s Objective Function

Suppose a firm of type $\theta$ chooses a service rate $\mu$. Informed consumers directly observe the quality of the firm, and hence, given the firm’s strategy, know $\mu$. For uninformed consumers, their posterior beliefs over quality, $\gamma(\cdot)$, lead to an expectation over $\mu$. Thus, a firm’s choice of service rate affects the strategies of both informed and uninformed consumers.

Each type of firm chooses its own strategy $\beta_\theta$ to maximize its expected payoff, given $\sigma$. Recall that the price is not a choice variable for the firm; rather the price is implicit in the values $v_h, v_\ell$. We therefore assume that each type of firm earns a profit of $r$ per consumer that it serves. We assume that $r$ does not depend on the quality of the firm. Hence, the firm maximizes its throughput per unit time in the long-run; that is, given the stationary distribution over queue length induced by $\sigma$ and $\mu$.

Let $X_\theta(t \mid \mu, \sigma)$ be a random variable that denotes the number of consumers in the queue at some time $t \geq 0$, given that the firm has quality $\theta$, the service rate of the firm is $\mu$, and the strategy profile of consumers is denoted by $\sigma$. The arrival and service processes, together with the strategies of informed and uninformed consumers, imply that $X_\theta(t \mid \mu, \sigma)$ follows a Markov process. Since the queue length process is irreducible and aperiodic, it has a stationary distribution with support contained in $\mathcal{N}$. Let $\pi_\theta(n, \mu, \sigma)$ denote the stationary probability of observing $n$ consumers in the queue under the given conditions. That is, $\pi_\theta(n, \mu, \sigma) = \lim_{t \to \infty} \text{Prob}(X_\theta(t \mid \mu, \sigma) = n)$.

Let $\hat{\pi}_\theta(n, \beta, \sigma)$ for $\theta \in \{h, \ell\}$ be the probability of observing $n$ consumers in the system upon arrival when the firm’s strategy is $\beta$ and the consumers’ strategy is $\sigma$. By the PASTA
(Poisson Arrivals See Time Averages) property of queueing systems (see Wolff, 1982), in
the limit a newly-arriving agent will be faced with a distribution over queue length equal
to the stationary distribution. That is, \( \pi_\theta(n, \beta, \sigma) = \beta_\theta \pi_\theta(n, \bar{\mu}, \sigma) + (1 - \beta_\theta) \pi_\theta(n, \mu, \sigma) \) for \( \theta \in \{h, \ell\} \).

In general, the distribution defined by \( \pi_h(\cdot) \) will be different than the distribution defined by \( \pi_\ell(\cdot) \) for two reasons. First, even if both types of firm chose the same service rate, informed consumers follows a strategy that varies with firm type. Second, the two types of firm may adopt different service rate strategies. Thus, uninformed consumers can update their prior beliefs about firm quality on observing the length of the queue when they arrive.

We derive the stationary distribution over queue length as follows. Suppose a firm with type \( \theta \) has chosen a service rate \( \mu \). Let \( s_\theta(n, \sigma) \) be the ex ante probability that an agent who plays strategy \( \sigma \) joins a queue that already has \( n \) agents. Then,

\[
s_\theta(n, \sigma) = (1 - q) \sigma_u(n) + q \sigma_i(\theta, n) \quad \text{for } \theta \in \{h, \ell\}.
\]  

(3)

If there are \( n \) agents in the queue, the rate at which a new agent joins the queue is \( \Lambda s_\theta(n, \sigma) \). Although the firm may be playing a mixed strategy, each type chooses a single realized service rate, which remains constant throughout the consumer game. The rate at which agents leave the queue is equal to the service rate \( \mu \). Hence, for any consumer strategy profile \( \sigma \), the induced queuing system is a birth and death process. The stationary probability for each queue length \( n \), \( \pi_\theta(n, \mu, \sigma) \), is then derived in a standard manner from the resulting flow balance equations.

**Lemma 1** Suppose the firm has type \( \theta \in \{h, \ell\} \) and chooses service rate \( \mu \), and all agents follow the strategy profile \( \sigma \). Then, the stationary probabilities over different queue lengths are given by:

\[
\pi_\theta(0, \mu, \sigma) = \frac{1}{1 + \sum_{n=1}^{N} \left( \frac{1}{\mu} \right)^n \prod_{j=0}^{n-1} s_\theta(j, \sigma)}
\]

\[
\pi_\theta(n, \mu, \sigma) = \pi_\theta(0, \mu, \sigma) \left( \frac{1}{\mu} \right)^n \prod_{j=0}^{n-1} s_\theta(j, \sigma).
\]

(4)

(5)

Let \( R_\theta(\mu, \sigma) \) be the firm \( \theta \)'s revenue per unit of time when the consumer strategy profile is \( \sigma \) and the firm chooses service rate \( \mu \). Under these conditions, the firm is busy with stationary probability \( 1 - \pi_\theta(0, \mu, \sigma) \). Since the expected time to service a consumer is \( \frac{1}{\mu} \), the revenue per unit of time is \( r\mu \). Hence, the expected payoff of firm \( \theta \) when it chooses service rate \( \mu \) is \( R_\theta(\mu, \sigma) = r\mu(1 - \pi_\theta(0, \mu, \sigma)) \).
2.3 Definition of Equilibrium

We consider a Markov-perfect Bayesian equilibrium of the game: both types of firm and both types of consumer must maximize their respective expected payoffs, and, where possible, the beliefs of uninformed consumers must obey Bayes’ rule.

**Definition 1** A Markov-perfect Bayesian equilibrium is defined by a triple \((\beta, \sigma, \gamma)\) that satisfies the following properties.

(i) Each type of firm maximizes its expected payoff:

\[
\beta_\theta \in \arg\max_{\beta \in [0,1]} \bar{\beta}[R_\theta(\bar{\mu}, \sigma) - k] + (1 - \bar{\beta})R_\theta(\mu, \sigma) \text{ for } \theta \in \{h, \ell\}. \tag{6}
\]

(ii) Each type of consumer maximizes her expected utility:

\[
\sigma_i(j,n) \in \arg\max_{\tilde{\sigma} \in [0,1]} \tilde{\sigma}w_i(n, \theta, \beta) \text{ for } \theta \in \{h, \ell\} \text{ and each } n \in N, \tag{7}
\]

\[
\sigma_u(n) \in \arg\max_{\tilde{\sigma} \in [0,1]} \tilde{\sigma}w_u(n, \gamma, \beta) \text{ for each } n \in N. \tag{8}
\]

(iii) Where possible, the beliefs \(\gamma\) are derived from the strategies using Bayes’ rule. That is, whenever the denominator is strictly positive,

\[
\gamma(n) = \frac{p\tilde{\pi}_h(n, \beta, \sigma)}{p\tilde{\pi}_h(n, \beta, \sigma) + (1-p)\tilde{\pi}_l(n, \beta, \sigma)}. \tag{9}
\]

Throughout the paper, we focus on pure strategy equilibria. That is, a firm of type \(\theta\) chooses a service rate \(\mu_\theta\) (so that \(\beta_\theta \in \{0, 1\}\)) and, given the firm’s strategy, consumers play a pure strategy in the continuation game.

3 Continuation Equilibria in the Consumer Game

We first consider continuation equilibria in the consumer game that arises after each firm type \(\theta\) has chosen its service rate strategy, \(\beta_\theta\). The optimal strategy of an informed consumer is straightforwardly characterized as a threshold strategy, which is standard in the queueing literature (see, for example, Naor, 1969, and Hassin and Haviv, 2003). She knows the quality of the good, and hence joins the queue as long as it is not too long. Let \(\mu_\theta\) be the service rate chosen by the firm of type \(\theta\). Then, an informed consumer joins as long as the queue length \(n\) is less than \(n_\theta = \lfloor v_\theta \mu_\theta \rfloor \) (note that for each \(\theta\), \(n_\theta\) depends on \(\beta_\theta\); for notational brevity, we suppress this dependence going forward). Further, if \((n_\theta + 1)\frac{c}{\mu_\theta} > v_\theta\), it is a unique best response for an informed consumer to balk at \(n_\theta\).
Lemma 2 Suppose \( n_\theta \frac{c}{\mu_\theta} < v_\theta \) for each \( \theta = h, \ell \). Then, it is a unique best response for an informed consumer to play \( \sigma_i(\theta, n) = 1 \) for \( n < n_\theta \) and \( \sigma_i(\theta, n) = 0 \) for \( n \geq n_\theta \).

For the rest of the paper, we assume that the condition in Lemma 2 holds. For analytic convenience, we further assume that \( n_\ell < n_h \). That is, regardless of the service rate strategies chosen by the firms, the threshold at which informed consumers balk from the high-quality firm strictly exceeds the corresponding threshold for the low-quality firm. Assumption 2 essentially implies that the range in consumer valuations \( (v_h - v_\ell) \) has a greater impact on the consumer’s choice than the range in service rates \( \mu - \bar{\mu} \). If the assumption is satisfied, then, for any pair of strategies \( (\mu_h, \mu_\ell) \) chosen by the different types of the firm, it will be the case that \( n_h > n_\ell \).

Assumption 2 \( [v_\ell(\bar{\mu}/c)] < [v_h(\mu/c)] \).

Of course, uninformed consumers will always join queues that are shorter than \( n_\ell \), since they earn a positive payoff even if the firm has low quality. Consider queue lengths between \( n_\ell \) and \( n_h \). If the queue did not communicate any information about quality, uninformed consumers would join at any length \( n \) as long as \( pv_h + (1 - p)v_\ell \geq (n + 1)c \left( \frac{\mu_h}{\mu_h} + \frac{(1 - p)}{\mu_\ell} \right) \). However, as Lemma 1 shows, the two types of firm generate different probability distributions over queue length. Hence, an uninformed consumer updates her belief on seeing the length of the queue.

As we show below, the equilibrium strategy of an uninformed consumer is typically not a threshold strategy. Rather, it is characterized by a “hole.” That is, there exists exactly one queue length \( \hat{n} \) between \( n_\ell \) and \( n_h \) at which the uninformed consumer does not join the queue. At every other queue length less than \( n_h \), she follows the strategy of an informed consumer who knows the firm has high quality; that is, she joins the queue. The intuition underlying this strategy is as follows. At any queue length below \( n_\ell \), informed consumers also join the queue. Hence, regardless of beliefs, uninformed consumers find it optimal to join the queue. At a queue length of \( n_\ell \) or above, however, the beliefs start to matter. Suppose there is some queue length \( n > n_\ell \) at which uninformed consumers balk. If an uninformed consumer sees any queue length higher than \( n \), she must believe the firm has high quality. Thus, her strategy at \( n \) and all higher queue lengths mimics that of an informed agent who knows the firm has high quality.

Proposition 1 In any pure strategy equilibrium of the consumer game, there exists a \( \hat{n} \in \{n_\ell, n_\ell + 1, \ldots, n_h\} \) such that \( \sigma_u(\hat{n}) = 0 \) and \( \sigma_u(n) = \sigma_i(h, n) \) for all \( n \in N \setminus \hat{n} \).
Consider the Bayesian updating problem faced by an uninformed consumer. She observes the queue length \( n \). In equilibrium, high and low-quality firms generate different distributions over queue length. Suppose each type of firm plays a pure strategy. That is, a firm of type \( \theta \) chooses a service rate \( \mu_\theta \in \{\bar{\mu}, \mu\} \). The uninformed consumer’s posterior belief that the firm has high quality may be represented as

\[
\gamma(n) = \frac{p \pi_h(n, \mu_h, \sigma)}{p \pi_h(n, \mu_h, \sigma) + (1 - p) \pi_\ell(n, \mu_\ell, \sigma)},
\]

where \( \sigma \) is the equilibrium in the consumer game. If \( \sigma \) is a pure strategy, there is a sharp characterization of the posterior belief \( \gamma \). In particular, for any queue length \( n \), \( \gamma(n) \) can be represented in terms of the likelihood ratio at queue length zero, \( \frac{\pi_\ell(0, \mu_\ell, \sigma)}{\pi_h(0, \mu_h, \sigma)} \).

**Proposition 2** Suppose each type of firm plays a pure strategy, with type \( \theta \) choosing service rate \( \mu_\theta \in \{\bar{\mu}, \mu\} \). Suppose further that a pure strategy equilibrium \( \sigma \) results in the consumer game, in which informed consumers balk at \( n_\theta \) for a type \( \theta \) firm, and uninformed consumers join at every queue length \( n \leq n_\ell \) except at \( \hat{n} \). Then, the uninformed consumer’s posterior that the firm has high quality is given by

\[
\gamma(n) = \frac{p \pi_h(n, \mu_h, \sigma)}{p \pi_h(n, \mu_h, \sigma) + (1 - p) \pi_\ell(n, \mu_\ell, \sigma)},
\]

where

\[
\phi(n, \mu_h, \mu_\ell, \sigma) = \begin{cases} 
\phi(0, \mu_h, \mu_\ell, \sigma) \left( \frac{\mu_h}{\mu_\ell} \right)^n & \text{if } n \in \{0, \ldots, n_\ell\} \\
\phi(0, \mu_h, \mu_\ell, \sigma) \left( \frac{1 - q}{1 - q} \right) \left( \frac{\mu_h}{\mu_\ell} \right)^n & \text{if } n \in \{n_\ell + 1, \ldots, \hat{n}\} \\
0 & \text{if } n \in \{\hat{n} + 1, \ldots, n_h\}
\end{cases}
\]

with \( \phi(0, \mu_h, \mu_\ell, \sigma) = \frac{\pi_\ell(0, \mu_\ell, \sigma)}{\pi_h(0, \mu_h, \sigma)} \).

Consider a queue length less than or equal to \( n_\ell \), the threshold at which an informed consumer balks at the low-quality firm. Since informed consumers join at such a queue length, an uninformed consumer’s posterior probability that the firm has high quality decreases in the queue length if \( \mu_h > \mu_\ell \) and increases if \( \mu_h < \mu_\ell \). If the high-quality firm has the faster service rate, over this range a higher queue length is relatively more likely to come from a low-quality firm, so that the posterior belief places lower weight on the firm being of high quality. The converse argument applies if the high-quality firm has a slower service rate than the low-quality firm.

Once the threshold \( n_\ell \) is reached, only uninformed consumers join the low-quality firm, whereas all consumers join the high-quality firm. The probability that a newly-arrived consumer will join the queue is therefore \( (1 - q) \) for the low-quality firm and one for the high-quality firm. Now, the posterior belief that the firm has high quality decreases in the queue length only if \( (1 - q) \mu_h > \mu_\ell \), and increases if \( (1 - q) \mu_h < \mu_\ell \). At the threshold \( \hat{n} \),
uninformed consumers do not join the queue. Hence, the queue for the low-quality firm can never grow larger than \( \hat{n} \). It follows immediately that at any queue length \( \hat{n} + 1 \) or higher, even an uninformed consumer knows the firm has high quality.

**Example 1**

To illustrate an uninformed consumer’s belief updating process and the resulting consumer equilibrium, we consider the following numeric example. Let \( p = 0.5, \quad q = 0.1, \quad v_h = 10, \quad v_\ell = 2, \quad c = 0.4005, \quad \Lambda = 1, \quad \mu_h = 1.025, \quad \text{and} \quad \mu_\ell = 0.975. \) Then, \( n_h = 25 \) and \( n_\ell = 4. \) In Figure 1, we show an uninformed agent’s posterior expectations about consumption utility and the waiting cost if she joins at each queue length. Both depend on her posterior belief that the firm has high quality. For queue lengths zero to four (i.e., weakly below \( n_\ell \)), the posterior belief falls in the queue length, since \( \mu_h > \mu_\ell \). For queue lengths between 5 and 16, the posterior belief rises in the queue length, since the proportion of informed consumers is sufficiently high (in particular, \((1 - q)\mu_h < \mu_\ell\)). At 16, the expected consumption value is just below the expected waiting cost, so the uninformed consumer does not join the queue. Thus, \( \hat{n} = 16 \). At queue lengths 17 and beyond, the uninformed consumer knows the firm has high quality, so the posterior expectation jumps to \( v_h \), and the waiting cost becomes linear in queue length. Overall, the uninformed consumer joins at every queue length between 0 and 24 except at 16.

We next consider the implications of consumers playing a pure strategy for the payoff (or expected revenue per unit of time) of the firm. For any consumer strategy \( \sigma = (\sigma_i, \sigma_u) \), let \( \hat{n}(\sigma) = \min \{n \mid \sigma_u(n) = 0\} \). When \( \sigma_u \) is a pure strategy, it follows that \( \hat{n}(\sigma) \) is the queue length at which the uninformed consumer’s strategy has a hole. Thus, if the firm has low-quality, its queue never extends beyond \( \hat{n}(\sigma) \). At queue lengths below \( n_\ell \), both informed and uninformed agents join the queue, and at queue lengths between \( n_\ell \) and \( \hat{n}(\sigma) - 1 \), only uninformed agents join. In contrast, the high quality firm has all agents joining the queue at all queue lengths below \( n_h \), except for the queue length \( \hat{n}(\sigma) \), at which only informed agents join.

These properties allow us to simplify the expected revenue expressions for each type of firm. We fix a particular pure strategy for consumers, \( \sigma \), and determine the revenue of each type of firm if it picks an arbitrary service rate \( \mu \). The consumer strategy need not be a best response to the firm’s strategy. For each \( \theta = h, \ell, \) let \( \tilde{n}_\theta(\sigma) = \min \{n \mid \sigma_i(\theta, n) = 0\} \) be the first queue length at which informed consumers balk. We say a consumer pure strategy \( \sigma \)
This figure shows an uninformed consumer’s posterior expectations about value (+) and the total waiting cost (○) on the Y-axis, plotted against the queue length when she arrives on the X-axis. The parameters are: \( p = 0.5, q = 0.1, v_h = 10, v_\ell = 2, c = 0.4005, \Lambda = 1, \mu_h = 1.025, \) and \( \mu_\ell = 0.975. \)

Figure 1: Uninformed consumer’s posterior expectations of value and waiting cost

satisfies the necessary conditions of equilibrium in the continuation game if \( \tilde{n}_h(\sigma) \geq \tilde{n}_\ell(\sigma) \) and \( \hat{n}(\sigma) \in \{\tilde{n}_\ell(\sigma), \tilde{n}_\ell(\sigma) + 1, \cdots, \tilde{n}_h(\sigma)\}. \) Of course, in equilibrium, \( \tilde{n}_\theta(\sigma) \) will equal \( n_\theta, \) as defined before Lemma 2. For the purposes of the next Lemma, however, the consumer strategy is held fixed at \( \sigma, \) regardless of the firm’s choice of service rate. As we show, conditional on consumers playing a pure strategy, both types of firm would prefer to have the hole in the uninformed consumer’s strategy at as high a queue length as possible.

**Lemma 3** Fix a pure strategy for consumers, \( \sigma, \) that satisfies the necessary conditions for equilibrium in the continuation game. Suppose a firm of type \( \theta \in \{h, \ell\} \) chooses service rate
µ, and consumers play σ. Then, the expected revenue per unit of time for each type of firm is

\[
R_h(\mu, \sigma) = r\mu \left( 1 - \frac{1}{\sum_{j=0}^{\hat{n}(\sigma)} (\Lambda/\mu)^j + q \sum_{j=\hat{n}(\sigma)+1}^{\hat{n}_h(\sigma)} (\Lambda/\mu)^j} \right),
\]

\[
R_\ell(\mu, \sigma) = r\mu \left( 1 - \frac{1}{\sum_{j=0}^{\hat{n}_\ell(\sigma)} (\Lambda/\mu)^j + \sum_{j=\hat{n}_\ell(\sigma) + 1}^{\hat{n}_h(\sigma) + 1} (1 - q)^{j-\hat{n}_\ell(\sigma)} (\Lambda/\mu)^j} \right).
\]

Further, fixing \(\hat{n}_h(\sigma)\) and \(\hat{n}_\ell(\sigma)\), the expected revenue \(R_\theta\) increases in \(\hat{n}(\sigma)\) for \(\theta \in \{h, \ell\}\).

Therefore, all else equal, each type of firm will prefer that the hole in an uninformed consumer’s strategy, \(\hat{n}\), occur at a high rather than low queue length. Of course, the service rate choices of each type of firm may affect an informed consumer’s strategy as well. Nevertheless, to highlight the role of information contained in the queue length on the firm’s optimal strategy, from now on we restrict attention to parameter values under which there is a hole in an uninformed consumer’s strategy exactly at \(n_\ell\). Given the strategy of each type of firm, \(n_\ell\) represents the lowest queue length at which the hole can occur, and hence is also the queue length at which the hole has the greatest impact on the firm’s strategy.

The strategy of each type \(\theta\) is characterized by its choice of service rate \(\mu_\theta\). We show that, regardless of the service rate chosen by each type of firm, a continuation equilibrium in which the uninformed consumer’s strategy has a hole at \(n_\ell\) exists as long as the prior probability \(p\) (i.e., the probability the firm has high quality) is sufficiently low. The intuition here is straightforward. By definition, an informed consumer will not join the queue of a low-quality firm when the queue length is \(n_\ell\). Thus, if an uninformed consumer places sufficiently low probability on the firm having high quality, she will not join the queue at \(n_\ell\) either. The condition that her posterior at \(n_\ell\) place sufficiently low probability on the firm having high quality can then be translated into a condition on the prior probability \(p\).

Recall that \(n_\theta = \lfloor v_\theta \frac{\mu_\theta}{c} \rfloor\) is the threshold at which informed consumers balk when the firm has type \(\theta\). We assume that \(n_h\) and \(n_\ell\) are invariant to the service rate chosen by each type of firm.

**Assumption 3** \([v_\ell(\bar{\mu}/c)] = [v_\ell(\mu/c)]\) and \([v_h(\bar{\mu}/c)] = [v_h(\mu/c)]\).

Assumption 3 effectively implies that the fast and slow service rates (\(\bar{\mu}\) and \(\mu\)) are close to each other. In other words, if a firm speeds up the rate at which it services consumers, it does so through an incremental operational improvement. With this additional assumption, we show that if the prior probability the firm has high quality is low enough, there is a
continuation equilibrium in the consumers’ game with a hole in the uninformed consumer’s strategy at $n_\ell$.

**Proposition 3** Suppose a firm with type $\theta \in \{h, \ell\}$ chooses service rate $\mu_\theta \in \{\bar{\mu}, \bar{\mu}\}$. Then, there exists a $p_0 \in (0,1)$ such that if $p \leq p_0$, there is a pure strategy equilibrium $\sigma$ in the consumers’ continuation game, with $\hat{n}(\sigma) = n_\ell$. As $v_h \to \infty$, $p_0 \to 0$.

A case of interest in what follows will be when $v_h$ becomes large. As the proposition states, $p_0 \to 0$ as $v_h$ increases toward infinity. To maintain a hole in an uninformed consumer’s strategy at any fixed finite queue length as $v_h$ becomes large, the prior probability the firm has high quality must tend toward zero. Otherwise, the posterior probability the firm has high quality remains bounded away from zero, and if $v_h$ is sufficiently large, an uninformed consumer will join the queue at that length. The same intuition applies when the hole is specifically at $n_\ell$.

In each of the results in Section 4, we assume that the condition in Proposition 3 is satisfied. That is, the probability the firm has high quality is sufficiently low, given the other parameters of the model. We further assume that the equilibrium described in Proposition 3 results in the consumer game, so that the hole in the uninformed consumers’ strategy is exactly at $n_\ell$.

### 4 Equilibrium Choice of Service Rates

We now turn to the optimal choice of service rates by the firms. For any service rates chosen by the two types of firm, we focus on the pure strategy equilibrium in which the uninformed consumer’s strategy has a hole exactly at $n_\ell$. Since we consider service rates $\bar{\mu}$ and $\bar{\mu}$ that are close to each other, let $\mu_0 > 0$ be the mean of the fast and slow service rates. Then, for any $\epsilon \in (0, \mu_0)$, define $\bar{\mu} = \mu_0 + \epsilon$ and $\bar{\mu} = \mu_0 - \epsilon$.

Let $\hat{\sigma}$ denote the consumer strategy in which informed consumers join the queue of a firm of type $\theta$ at all queue lengths up to and including $n_\theta - 1$, and uninformed consumers join the queue at all queue lengths up to $n_h - 1$ except at the length $n_\ell$. For each $\theta$, $n_\theta = \lfloor v_\theta(\mu_0/c) \rfloor$, which is an immediate function of the parameters of the model. Hence, the consumer strategy $\hat{\sigma}$ is well-defined. Further, if $p \leq p_0$, regardless of the service rate strategy chosen by each type of firm, there is an equilibrium of the consumer continuation game in which the hole in an uninformed consumer’s strategy is exactly at $n_\ell$.

Now, consider a firm of type $\theta$. Let $\Delta_\theta$ denote the marginal revenue the firm gains from choosing the fast service rate $\bar{\mu} = \mu_0 + \epsilon$ rather than the slow service rate $\bar{\mu} = \mu_0 - \epsilon$, given
that consumers are playing the strategy \( \hat{\sigma} \). Then,

\[
\Delta_\theta(\mu_0, \epsilon) = R_\theta(\mu_0 + \epsilon, \hat{\sigma}) - R_\theta(\mu_0 - \epsilon, \hat{\sigma}).
\] (10)

Observe that, from Lemma 3, for each \( \theta \) the revenue rate \( R_\theta \) depends on \( q \), the proportion of informed consumers. Hence, \( \Delta_\theta \) also depends on \( q \). For any \( \mu_0 \), the firm will strictly prefer the fast service rate \( \mu_0 + \epsilon \) over the slow service rate \( \mu_0 - \epsilon \) if and only if \( \Delta_\theta(\mu_0, \epsilon) > k \), the marginal cost of installing added capacity (i.e., of choosing the fast service rate).

We first provide conditions under which both firms will choose the fast service rate in equilibrium. Of course, the technological cost of service, \( k \), must be sufficiently small. Further, for the high-quality firm to choose fast service, it must be the case that \( v_h \) is not too high. The condition in Proposition 4 below ensures that \( n_h \leq 2n_\ell + 1 \).

**Proposition 4** Suppose \( p \leq p_0 \). Then,

(i) There exists a range of \( k \) such that in equilibrium the low-quality firm chooses the fast service rate \( \mu_0 + \epsilon \).

(ii) If \( v_h \leq 2v_\ell + \frac{\mu_0}{c} \), there exists a range of \( k \) such that in equilibrium the high-quality firm chooses the fast service rate \( \mu_0 + \epsilon \).

In the remainder of the paper, we consider two cases in which the high-quality firm prefers to slow down in equilibrium. Given Proposition 4, such cases must inevitably involve a large \( v_h \). We first consider service rates \( \bar{\mu} \) and \( \underline{\mu} \) that surround the arrival rate, \( \Lambda \), and show that the incentives of the high-quality firm to speed up depend on the fraction of informed consumers, \( q \). We then assume both service rates are strictly greater than, but close to, \( \Lambda \), and show that the high-quality firm has a strict incentive to slow down, regardless of \( q \).

**4.1 Service Rates on Either Side of Arrival Rate**

Suppose \( \mu_0 = \Lambda \). Then, the slow service rate \( \underline{\mu} \) is strictly less than \( \Lambda \), and the fast service rate \( \bar{\mu} \) is strictly greater than \( \Lambda \). We consider the relative incentive each type of firm has to choose the fast service rate \( \bar{\mu} = \Lambda + \epsilon \) over the slow service rate \( \underline{\mu} = \Lambda - \epsilon \). We obtain clean analytic expressions for the derivative of revenue of type \( \theta \) with respect to \( \mu, \frac{\partial R_\theta}{\partial \mu}(\mu, \hat{\sigma}) \), in the limit as \( \epsilon \) goes to zero (i.e., \( \underline{\mu} \) and \( \bar{\mu} \) each approach \( \Lambda \)). Using these limiting analytic expressions as our starting point, we are then able to provide comparisons in a neighborhood of service rates around \( \Lambda \).
We first provide the limiting expressions for the rate of change of revenue with respect to $\mu$, when the service rate exactly equals the arrival rate.

**Lemma 4** Suppose $\mu = \Lambda$ and $p \leq p_0$. Then, the rate of change of revenue with respect to $\mu$ is:

\[
\frac{\partial R_h(\Lambda, \hat{\sigma})}{\partial \mu} = \frac{r [2(n_h - n_\ell)^2 q^2 - (n_h - 3n_\ell - 1)(n_h - n_\ell)q + n_\ell(n_\ell + 1)]}{2[n_\ell + 1 + q(n_h - n_\ell)]^2},
\]
\[
\frac{\partial R_\ell(\Lambda, \hat{\sigma})}{\partial \mu} = \frac{r n_\ell}{2(n_\ell + 1)}.
\]

Consider the rate of change of the revenue of the high-quality firm in the limit, when $\mu = \Lambda$. First, suppose that $q = 1$; i.e., all consumers are informed. Then, $\frac{\partial R_h}{\partial \mu}(\Lambda, \hat{\sigma}) = \frac{r n_h}{2(n_h + 1)}$, which exceeds $\frac{\partial R_\ell}{\partial \mu}(\Lambda, \hat{\sigma}) = \frac{r n_\ell}{2(n_\ell + 1)}$ since $n_h > n_\ell$. That is, when all consumers are informed, the high-quality firm has a greater incentive to speed up than the low-quality firm. Next, suppose that $q = 0$, so that all consumers are uninformed. Then, the two derivatives in Lemma 4 are each exactly equal to $\frac{r n_\ell}{2(n_\ell + 1)}$. This is intuitive: if there are no informed consumers, the two types of firm have exactly the same stationary distribution over queue length. Hence, each type of firm has the same incentive to speed up.

In equilibrium, a firm of type $\theta$ will strictly prefer the fast service rate $\Lambda + \epsilon$ whenever $\Delta_\theta(\epsilon) > k$; that is, the increase in revenue from speeding up the service rate exceeds the cost of the technology. We have a closed-form expression for $R_\theta$ only in the limit, $\bar{\mu}$ and $\mu$ are each equal to $\Lambda$ (so that $\epsilon = 0$). To the extent that a firm has a strict incentive to speed up (i.e., choose $\bar{\mu} = \Lambda + \epsilon$) or slow down (i.e., choose $\mu = \Lambda - \epsilon$), results obtained when $\mu = \Lambda$ will continue to hold for service rates in an $\epsilon$-neighborhood of $\Lambda$.

Define a threshold proportion of informed consumers, $\hat{q}$ as follows:

\[
\hat{q} = \left(1 - \frac{1}{n_h - n_\ell}\right) \frac{n_\ell + 1}{n_\ell + 2}.
\]

It is immediate that the threshold $\hat{q}$ is increasing in $n_h - n_\ell$.

Our main result is described as follows. Consider any $q > \hat{q}$. Then, if $\underline{\mu}$ and $\bar{\mu}$ are sufficiently close to $\Lambda$, the marginal revenue of the high-quality firm exceeds the marginal revenue of the low-quality firm; that is, $\Delta_h(\Lambda, \epsilon) > \Delta_\ell(\Lambda, \epsilon)$. The high-quality firm therefore has a greater incentive to speed up than the low-quality firm. Thus, there is a range of costs at which the high-quality firm chooses the fast service rate $\bar{\mu}$ and the low-quality firm chooses the slow service rate $\underline{\mu}$. Similarly, if $q < \hat{q}$ and $\underline{\mu}$ and $\bar{\mu}$ are sufficiently close to $\Lambda$,
the low-quality firm has a greater incentive to speed up than the high-quality firm; that is, \( \Delta_h(\Lambda, \epsilon) < \Delta_\ell(\Lambda, \epsilon) \). Thus, there is another range of costs at which the high-quality firm chooses the slow service rate \( \mu \) and the low-quality firm chooses the fast service rate \( \bar{\mu} \).

**Proposition 5** For every \( q \in (0, 1) \), there exists an \( \hat{\epsilon}(q) > 0 \) such that if \( \epsilon \in (0, \hat{\epsilon}(q)] \) and \( p \leq p_0 \):

(i) If \( q < \hat{q} \), then \( \Delta_h(\Lambda, \epsilon) < \Delta_\ell(\Lambda, \epsilon) \). Hence, there exists a range for \( k \) such that in equilibrium the high-quality firm chooses the slow service rate \( \Lambda - \epsilon \) and the low-quality firm chooses the fast service rate \( \Lambda + \epsilon \). However, the converse service rate strategies cannot be sustained in equilibrium.

(ii) If \( q > \hat{q} \), then \( \Delta_h(\Lambda, \epsilon) > \Delta_\ell(\Lambda, \epsilon) \). Hence, there exists a range for \( k \) such that in equilibrium the high-quality firm chooses the fast service rate \( \Lambda + \epsilon \) and the low-quality firm chooses the slow service rate \( \Lambda - \epsilon \). However, the converse service rate strategies cannot be sustained in equilibrium.

Consider the trade-off faced by the high-quality firm as it chooses its service rate. Assume that \( n_\ell < n_h \), i.e. informed consumers will balk at strictly higher queue length when the quality is high than when the quality is low. Slowing down (i.e., choosing the slow service rate) implies that the average waiting time for consumers is longer. From the firm’s viewpoint, slowing down implies that higher queue lengths occur more often. Uninformed consumers join the queue at all lengths except at the hole, i.e., at queue length \( n_\ell \). Suppose the queue length is exactly \( n_\ell \) when an informed consumer arrives. Recall that \( n_\ell < n_h \), hence, the informed consumer will join, and the queue length will become \( n_\ell + 1 \). Therefore, if the proportion of informed consumers is relatively high, the hole is irrelevant, since it will be crossed whenever an informed consumer arrives. However, if the proportion of informed consumers is low, the hole is unlikely to be crossed simply by an arriving consumer. In this case, choosing the slow service rate is potentially valuable, since a greater amount of time is then spent at queue lengths above the hole.

Intuitively, the benefit of slowing down for the high-quality firm will depend on the number of queue lengths above the hole; i.e., on \( n_h - n_\ell \). If \( n_h \) is substantially greater than \( n_\ell \), there is a greater incentive to slow down and reach the higher queue lengths. Conversely, if \( n_h \) is close to \( n_\ell \), the incentive to slow down is weaker. Indeed, that may be observed directly from the fact that the threshold proportion of informed consumers, \( \hat{q} \), is strictly increasing in \( (n_h - n_\ell) \). Thus, the range of \( q \) for which the high-quality firm has a stronger incentive to slow down, compared to the low-quality firm, increases in \( (n_h - n_\ell) \).
Further, note that when \( n_h = n_\ell + 1 \), \( \hat{q} = 0 \). Thus, at the minimal distance between \( n_h \) and \( n_\ell \), the high-quality firm always has a stronger incentive to speed up, compared to the low-quality firm. It is only when the distance between \( n_h \) and \( n_\ell \) increases to two or more that the high-quality firm may have a weaker incentive to speed up.

Next, as a special case, consider \( v_h \) becoming large, keeping \( v_\ell \) fixed. Observe that, as \( v_h \) (and hence \( n_h \)) becomes large, the threshold \( \hat{q} \) (as defined in equation 11) goes to \( \frac{n_\ell + 1}{n_\ell + 2} \). Thus, it follows from Proposition 5 that for large values of \( v_h \), if \( q > \frac{n_\ell + 1}{n_\ell + 2} \), there is a range of \( k \) for which the high-quality firm speeds up and the low-quality firm slows down. Conversely, if \( q < \frac{n_\ell + 1}{n_\ell + 2} \), there is a range of \( k \) for which the high-quality firm slows down and the low-quality firm speeds up.

However, as \( v_h \) becomes large, we obtain even stronger results for the high-quality firm. Strikingly in this case, if the proportion of informed consumers is sufficiently low (in particular, \( q < \frac{1}{2} \)), the high-quality firm actually obtains a lower revenue when it speeds up to the fast service rate, compared to staying at the slow service rate. Hence, even if there is no cost to increasing the service rate, the high-quality firm will prefer to provide slow service.

**Proposition 6** There exist an \( \bar{\epsilon}(q) \) and a threshold value for the high-quality good \( \bar{v}(q) \) with the following properties: Suppose \( v_h \geq \bar{v}(q), \epsilon \in (0, \bar{\epsilon}(q)] \) and \( p \leq p_0 \). Then, \( \Delta_h(\Lambda, \epsilon) < 0 \) if \( q < \frac{1}{2} \) and \( \Delta_h(\Lambda, \epsilon) > 0 \) if \( q > \frac{1}{2} \). Hence, if \( q < \frac{1}{2} \), there is no value of \( k \) (including \( k = 0 \)) at which the high-quality firm chooses the fast service rate in equilibrium.

The intuition behind our results as follows. Speeding up service has two potential benefits for the high-quality firm. First, it may increase the threshold \( n_h \) at which informed consumers balk. Assumption 3 rules out this effect since \( n_h \) is fixed at the same level at both service rates \( \bar{\mu} \) and \( \mu \). Second, it shifts the stationary distribution of consumers toward the lower parts of the queue, in particular to queue lengths below the length at which the uninformed consumers have a hole in their strategy. When the proportion of informed consumers is low, speeding up service is less valuable, since the queue is less likely to cross the length at which uninformed consumers have a hole.

When the proportion of informed consumers is relatively high, however, even at that queue length, the likelihood is that the next consumer is an informed consumer. Hence, queue lengths above the hole are more likely to be reached, making speeding up more valuable. This value is enhanced when \( v_h \) is large, so that there are many queue lengths above the length at which the uninformed consumers’ strategy has a hole.
4.2 Both Service Rates Greater than Arrival Rate

In the previous subsection, the slow service rate was below the consumer arrival rate $\Lambda$ and the fast service rate was above the arrival rate. In this subsection, we consider the case in which both service rates, slow and fast, are strictly greater than the arrival rate $\Lambda$. Suppose $\mu_0 > \Lambda$, and as before let $\underline{\mu} = \mu_0 - \epsilon$ with $\bar{\mu} = \mu_0 + \epsilon$. Here, we assume that $\epsilon \in (0, \mu_0 - \Lambda]$, to ensure that $\underline{\mu}$ is weakly greater than $\Lambda$. We continue to assume that, for any service rate strategy of each type of firm, there is a pure strategy equilibrium $\hat{\sigma}$ with a hole in the uninformed consumer’s strategy exactly at $n_\ell$.

We show that, if the mean of feasible service rates, $\mu_0$, is sufficiently close to $\Lambda$, and $v_h$ is sufficiently high, the high quality firm will never choose the fast service rate, even if there is no cost to speeding up. This result mirrors the result in Proposition 6. Importantly, the slowing down by the high-quality firm in Proposition 6 occurs only when $q < \frac{1}{2}$, whereas in Proposition 7 below, it occurs for any value of $q$.

For any $n \geq 1$, define $\rho(n)$ to be the unique solution to the equation

$$\sum_{j=0}^{n} \rho^j + q \sum_{j=n+1}^{\infty} \rho^j = n + 1. \tag{12}$$

Observe that the left-hand side equals 0 when $\rho = 0$, and is infinite when $\rho = 1$. Further, it is strictly increasing in $\rho$. Hence, for any fixed value of $n$ and any $q \in (0,1)$, $\rho(n)$ exists, is unique, and is strictly less than one. Further, note that $\lim_{n \to \infty} \rho(n) = 1$.

**Proposition 7** Consider any $n \in \{1, 2, \cdots \}$. Suppose $\mu_0 \in (\Lambda, \Lambda/\rho(n))$ and $c n / \mu_0 < v_\ell < c(n+1)/\mu_0$. Then, there exists an $\epsilon'(n)$ such that, if $\epsilon \in (0, \epsilon'(n)]$ and $p \leq p_0$, then for any choice of service rate by each type of firm, there is a pure strategy equilibrium in the consumer continuation game in which $n_\ell = n$ and there is a hole in the uninformed consumer’s strategy at exactly $n_\ell$. Further, there exists a $v_h(n) > 2v_\ell + \frac{\mu_0}{\epsilon}$ such that if $v_h \geq v_h(n)$, for any value of $k$ (including $k = 0$), the high-quality firm chooses the slow service rate $\underline{\mu} = \mu_0 - \epsilon$.

Recall from Proposition 4 that, if $v_h$ is low, there is a range of costs for which the high-quality firm chooses the fast service rate. Proposition 7 shows the other case: if $v_h$ is sufficiently high, the high-quality firm will never have an incentive to speed up, even if the superior service technology comes at no cost. The expression $\Lambda/\rho(n)$ tends toward $\Lambda$ as $n$ becomes large. Thus, the higher the position of the hole in the uninformed consumer’s strategy, the narrower the range of service rates for which the high-quality firm will not speed up.
The intuition underlying this proposition again relates to the number of feasible queue
lengths beyond the length at which the uninformed consumer's strategy has a hole. Recall
that we fix the hole to be exactly at \( n_\ell \). If \( v_h \) is very high, there are a lot of queue lengths
above the hole. Intuitively, speeding up increases the proportion of time the high-quality
firm finds itself facing a queue below the length at which the hole exists. Thus, slowing
down is valuable when there are many queue lengths above the hole at which uninformed
consumers join. When \( n_h \) is sufficiently high, the high-quality firm actually loses revenue
by speeding up, and thus prefers to choose the slow service rate.

Consider the limiting case in which \( v_h \rightarrow \infty \). In the limit, informed consumers join
at all queue lengths, and uninformed consumers at all queue lengths except exactly at \( n_\ell \).
Thus, for any value of \( \mu \) strictly less than \( \Lambda \) the queue is almost always non-empty, and the
revenue rate of the high-quality firm (\( R_h \)) is exactly equal to \( r\mu \). However, if \( \mu > \Lambda \), there
is a positive probability that the queue length dwindles down to the hole. Since uninformed
consumers do not join at the hole, the revenue rate is strictly less than will be less than
\( r\Lambda \). Therefore, the high-quality firm’s revenue rate exhibits a local maximum exactly at
\( \mu = \Lambda \). Suppose, for example, \( \mu = \Lambda \) and \( \bar{\mu} \) is strictly greater than, but close to, \( \Lambda \). Then,
the high-quality firm’s revenue rate is decreasing in the service rate, so that it will never
speed up, even when \( k = 0 \). We illustrate this feature in Example 2.

Example 2

Set \( p = 0.05, q = 0.2, v_h = 20, v_\ell = 1, c = 0.04005 \) and \( \Lambda = 1 \). Let \( \bar{\mu} = 1.1 \) and \( \bar{\mu} = 1 \n(\text{i.e., let } \mu_0 = 1.05 \text{ and } \epsilon = 0.05) \). Then, \( n_\ell = 2 \) regardless of the service rate chosen by
the low-quality firm. For the high-quality firm, \( n_h = 49 \) if it chooses the slow service rate
\( \underline{\mu} = 1 \), and \( n_h = 54 \) if it chooses the high service rate \( \bar{\mu} = 1.1 \). For each choice of service
rate by each type of firm, the consumer game has a pure strategy equilibrium in which the
strategy of an uninformed consumer has a hole at exactly \( n_\ell \).

Since \( v_h \) is high and \( q \) is low, expect that there will be an equilibrium in which the
high-quality firm chooses the slow service rate. Accordingly, let \( \hat{\sigma} \) denote the consumer
strategy with \( n_\ell = 2, \hat{n} = n_\ell = 2, \) and \( n_h = 49 \). Fixing \( \hat{\sigma} \), suppose we vary the service rate
of each type of firm, and consider its revenue. The revenue expressions are as defined in the
statement of Lemma 3. For \( \mu \) varying from 0 to 2, we plot \( R_h(\mu, \hat{\sigma}) \) and \( R_\ell(\mu, \hat{\sigma}) \) in Figure
2.

As seen from the figure, the revenue of the high-quality firm, \( R_h \), has a decreasing region
approximately between \( \mu = 0.95 \) and \( \mu = 1.15 \). Thus, the revenue at the fast service rate \( \bar{\mu} \)
This figure shows the revenue rate of each type of firm when the consumer strategy is $\hat{\sigma}$, with $n_\ell = 2$, $n_h = 49$, and $\hat{n} = n_\ell = 2$. The parameters are: $p = 0.1$, $q = 0.2$, $v_h = 20$, $v_\ell = 1$, $c = 0.4005$, $\Lambda = 1$, $\bar{\mu} = 1.1$ and $\mu = 1$. For each $\mu$, we keep the consumer strategy fixed at $\hat{\sigma}$ and determine the revenue rate of each type of firm.

### Figure 2: Revenue rate for each type of firm

is smaller than the revenue at the slow service rate $\mu$. Therefore, there is no value of $k$ at which the high-quality firm will choose the fast service rate in equilibrium. Indeed, in this case, the equilibrium always has the high-quality firm choosing the slow service rate $\mu = 1$. If $k$ is sufficiently low (specifically, $k < 0.0312$) the low-quality firm chooses the fast service rate $\mu = 1.1$, whereas if $k$ is sufficiently high ($k > 0.0312$), the low-quality firm also chooses the slow service rate.

To summarize: The incentive for the high-quality firm to slow down is greatest when $v_h$ is very high, $v_\ell$ is low, and the prior probability $p$ is low. Under these conditions, $n_h - n_\ell$ is large and by Proposition 3, the hole in the uninformed consumer’s strategy is at $n_\ell$. As
Lemma 3 shows, since \( n_\ell \) is low, the hole at this length has a large impact on the revenue of the high-quality firm. In addition, there are many queue lengths above the hole, at which uninformed consumers join. Consider service rates close to \( \Lambda \). Then, if \( \mu < \Lambda < \bar{\mu} \) and less than 50% of the consumers are informed, by Proposition 6, the high-quality firm will always choose the slow service rate. Similarly, if \( \mu > \Lambda \), by Proposition 7, the high-quality firm will choose the slow service rate regardless of the proportion of informed consumers.

4.3 Numeric Example of Service Rate Equilibria

We now consider a numeric example to illustrate the broader generality of the results obtained in the previous section.

Example 3

Set \( p = 0.1, v_h = 20, v_\ell = 1, c = 0.4005, \Lambda = 1, \bar{\mu} = 1.15 \) and \( \mu = 0.85 \). We consider two cases for \( q \), the proportion of informed consumers: \( q = 0.05 \) (Case I) and \( q = 0.9 \) (Case II).

Our choice of parameters thus departs from the assumptions underlying our analytic results in two important ways.

(i) The two feasible service rates are not very close to each other. In particular, for each firm type \( \theta \), the threshold at which informed consumers balk, \( n_\theta \), varies depending on whether the firm chooses the high or slow service rate.

(ii) While a pure strategy exists in the consumer continuation game for each choice of service rates, the hole in the uninformed consumer’s strategy is not always at \( n_\ell \). In particular, when \( q \) is high, the hole is often at \( n_h \).

There are four possible pure strategy equilibria in the model: each type of firm could choose either a fast or a slow service rate. Table 1 provides the range of \( k \) for which each type of equilibrium is sustained. For each possible equilibrium, we also show the details of the consumer strategy: the thresholds \( n_\ell \) and \( n_h \) for informed consumers, and the length \( \hat{n} \) at which the strategy of uninformed consumers has a hole.

First, consider the consumer strategies for each choice of service rates by the two types of firm. In Case I, when the high-quality firm is slow, an empty or short queue is bad news (in the sense that it is very likely to belong to a low-quality firm), and the hole in the uninformed consumers’ strategy is at \( n_\ell = 2 \). When both firms are fast, the hole creeps up to \( n = 9 \). In Case II, with a high proportion of informed consumers, the situation is different. Observe that for any choice of service rates \( \mu_h \) and \( \mu_\ell \), it will be that \((1-q)\mu_h < \mu_\ell \). Thus,
This table shows the range of $k$ that supports different kinds of service rate equilibria for two sets of parameters. In each case, a pure strategy is played in the consumer game, and the corresponding values of $n_h, n_\ell$ and $\hat{n}$ are shown. The length $\hat{n}$ is the queue length at which an uninformed consumer’s strategy has a hole.

Table 1: Equilibrium service rate choices

<table>
<thead>
<tr>
<th>Possible Equilibrium</th>
<th>Case I $q = 0.05$</th>
<th>Case II $q = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both types of firm slow</td>
<td>$k \geq 0.1007$</td>
<td>$k \geq 0.0245$</td>
</tr>
<tr>
<td>$n_\ell, \hat{n}, n_h$</td>
<td>2, 2, 42</td>
<td>4, 48, 48</td>
</tr>
<tr>
<td>Both types of firm fast</td>
<td>$k \leq 0.1042$</td>
<td>$k \leq 0.020$</td>
</tr>
<tr>
<td>$n_\ell, \hat{n}, n_h$</td>
<td>2, 9, 57</td>
<td>5, 51, 51</td>
</tr>
<tr>
<td>High-quality firm fast, low-quality firm slow</td>
<td>$k \leq 0.1007$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$n_\ell, \hat{n}, n_h$</td>
<td>2, 2, 42</td>
<td>5, 5, 48</td>
</tr>
<tr>
<td>High-quality firm slow, low-quality firm fast</td>
<td>$k \leq 0.1007$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$n_\ell, \hat{n}, n_h$</td>
<td>2, 2, 42</td>
<td>5, 5, 48</td>
</tr>
</tbody>
</table>

an uninformed consumer’s posterior expectation of value will increase in $n$ for $n > n_\ell$. Given the high proportion of informed consumers, the increase is rapid enough to push the hole to $n_h$ if the high-quality firm is fast or both firms are slow. However, if the high-quality firm is slow and the low-quality firm is fast, once again a short queue is bad news, to the extent that the hole falls back to $n_\ell$.

Next, consider the equilibrium service rate choices. For either value of $q$, it is intuitive that both firms choose the slow service rate if the cost of speeding up ($k$) is sufficiently high. Similarly, we expect that in equilibrium both firms will choose the fast service rate if the $k$ is sufficiently low.

The qualitative difference between the two cases occurs in considering equilibria in which one firm has fast service and the other has slow service. Based on our analysis in Section 4, given the low proportion of informed consumers in Case I, we expect that there will be a range of costs that support an equilibrium in which the high-quality firm is slow and the
low-quality firm is fast. Indeed, we find that for \( k \leq 0.1007 \), such an equilibrium exists. Conversely, there is no equilibrium in which the high-quality firm is fast and the low-quality firm is slow. The intuition is that, when the proportion of informed consumers is low, the high-quality firm uses the length of the queue to signal its quality, and hence prefers the slow service rate.

The intuition is reversed in Case II, with a high proportion of informed consumers. Signaling quality now has limited value, and the high-quality firm has a greater incentive to speed up than the low-quality firm. Thus, there is a range of costs \( k \) (in particular, \( k \) between 0.0199 and 0.0245) at which an equilibrium with a fast high-quality firm and slow low-quality firm can be sustained. However, the converse strategies (i.e., a slow high-quality firm and a fast low-quality firm) cannot be supported in equilibrium.

In Case I, notice that for low costs of speeding up (\( k \leq 0.1007 \)) two equilibria exist: both firms fast and only the low-quality firm fast (as shown in the second and fourth rows of Table 1). Since the proportion of informed consumers is low, a hole in the uninformed consumers’ strategy at a high queue length \( n = 9 \) offers the low-quality firm an incentive to speed up. On the other hand, when the hole is at a low-queue length \( n = 2 \), the low-quality firm prefers to slow down.

Overall, the numeric example demonstrates that the intuition of Proposition 5 goes through even when the strategy of both informed and uninformed consumers changes depending on whether the service rate \( \mu \) or \( \bar{\mu} \) is chosen by either type of firm. With a low number of informed consumers, the high-quality firm slows down to signal via the queue. With a high proportion of informed consumers, however, the high-quality firm chooses fast service.

5 Conclusion and Discussion

Our characterization of the conditions under which firms of differing quality choose either fast or slow service rates translates quite naturally into predictions on how different types of firms strategically manipulate queues. These conditions all depend on the queue length at which the equilibrium hole in an uninformed consumer’s strategy appears. For the uninformed, queues below the hole are short enough such that low waiting costs make joining rational. At the hole, they balk, and above the hole, uninformed consumers perfectly infer that the good has high quality.

Both firm types lose the uninformed consumers at the hole. The high-quality firm knows it will win back these customers at queue lengths above the hole. However the hole will
only be filled (or crossed) if an informed consumer arrives at the market. Therefore, a high-
quality firm is worst off facing a market with few informed consumers and in which the
uninformed ascribe a low prior probability to it being of high quality. In this case, the hole
is at low queue lengths, with little likelihood it will be filled by an informed customer. It is
therefore in the high-quality firm’s best interest to keep the queue above the hole, leading it
to choose a slow service rate. Quite naturally, the revenue-destroying effect of a fast service
rate in these circumstances is at its highest when the slow service rate is slightly above
the consumer arrival rate, and when there are few informed consumers that could fill the
hole. Intuitively, the high quality firm can substitute for the effect of informed consumers
by slowing down the service rate, ensuring a higher probability that the queue is above the
hole.

In our model, the queue serves to communicate information to uninformed consumers
about the strength of demand, and hence about their own valuation for the product. More
broadly, when informational externalities are high, we expect firms to engage in phased roll-
outs of a new product. The informational effect of excess demand will be strongest early in
the product life-cycle as there are likely to be few informed consumers in the market. Our
analysis suggests that it might be optimal for such firms to gradually increase the service
rate over time as the fraction of informed consumers in the market increases. In case such
firms cover the demand rate too early in the product life cycle, when the fraction of informed
consumers is low, further service rate expansion would stall until the fraction of informed
consumers is high enough. Indeed, if excess demand provides an informational externality
that generates even more demand, it is no longer true that supply creates its own demand.
Rather, the lack of supply may lead to an increase in demand.
A  Appendix: Proofs

Proof of Lemma 1

$\pi_{\theta}(n, \mu, \sigma)$ is the stationary probability of observing a queue of length $n$ when the firm has quality $\theta$ and chooses service rate $\mu$, and agents play the strategy profile $\sigma$. This is the long run probability of a birth-death process. The birth rate at which agents join the queue when the queue length is $n$ is $\lambda_{s(\theta)}(n, \mu, \sigma)$. Once in the queue, consumers leave at the rate $\mu$ (the death rate). Thus, given firm quality $\theta$ and agents’ strategy $\sigma$, the flow balance equations are

$$
\pi_{\theta}(n-1, \mu, \sigma) \lambda_{s(\theta)}(n-1, \mu, \sigma) + \pi_{\theta}(n+1, \mu, \sigma) \mu = \pi_{\theta}(n, \mu, \sigma) [\lambda_{s(\theta)}(n, \mu, \sigma) + \mu] \quad \text{for } n \geq 1.
$$

Further, it must be that $\sum_{n=0}^{\infty} \pi_{\theta}(n, \mu, \sigma) = 1$.

Recursively solving this system of equations yields the expressions in the statement of the Lemma.

Proof of Lemma 2

Fix $\theta \in \{h, \ell\}$. Recall that $n_{\theta} = \lfloor \frac{v_{\theta} \ell \mu}{c} \rfloor$. Therefore, it follows that $(n_{\theta} + 1) \frac{c}{\mu_{\theta}} > v_{\theta}$. By assumption, $n_{\theta} \frac{c}{\mu_{\theta}} < v_{\theta}$. Consider a newly-arrived informed consumer who finds $n$ agents already in the queue before her (including the agent being currently served). If she joins the queue, her expected waiting cost is $(n + 1) \frac{c}{\mu_{\theta}}$. Therefore, for $n < n_{\theta} - 1$, the expected waiting cost is strictly smaller than the consumption value of the good, $v_{\theta}$. Hence, it is strictly optimal to join the queue (i.e., set $\sigma_i(j, n) = 1$) for $n$ in this range. Conversely, for $n \geq n_{\theta}$, the expected waiting cost strictly exceeds the consumption value of the good, $v_{\theta}$. Hence, it is strictly optimal to balk for $n$ in this range.

Proof of Proposition 1

Suppose there is a pure strategy equilibrium $\sigma$ in the consumer game. Consider a newly-arrived uninformed consumer. Suppose first that $n < n_{\ell}$. Since an informed consumer joins the queue of the low-quality firm for all $n < n_{\ell}$, it must be that $v_{\ell} > (n + 1) \frac{c}{\mu_{\ell}}$. Now, by Assumption 2, it follows that $v_h > (n + 1) \frac{c}{\mu_{h}}$. Hence, for any value of $\gamma(n) \in [0, 1]$, $w_u(n, \gamma, \beta) > 0$. Thus, the uninformed consumer should join the queue whenever $n < n_{\ell}$, so that in equilibrium $\sigma_u(n) = 1$ for $n$ in this range.

Now, suppose $n \geq n_{\ell}$. In a pure strategy equilibrium, $\sigma_u(n) = 1$ or $\sigma_u(n) = 0$ for all $n$.  

26
Suppose first that $\sigma_u(n) = 1$ for all $n < n_h$. By definition of $n_h$, $v_h - (n_h + 1)\frac{c}{\mu_\mu_h} < 0$. Hence, regardless of her belief $\gamma(n_h)$, the uninformed consumer does not join at $n_h$, so that $\sigma_u(n_h) = 0$.

Next, suppose there exists an $n \in \{n_\ell, n_\ell + 1, \ldots, n_h - 2\}$ at which $\sigma_u(n) = 0$. Then, $s_\ell(n, \sigma) = 0$, since an informed agent does not join the queue of a low-quality firm when $n > n_\ell$. However $s_h(n, \sigma) = q > 0$. Now, consider queue length $n + 1$. Since $n + 1 < n_h$, it follows that $\sigma_i(h, n + 1) = 1$. Then, from Lemma 1, it follows that $\pi_\ell(n + 1, \mu, \sigma) = 0$ and $\pi_h(n + 1, \mu, \sigma) > 0$. Therefore, from Bayes’ rule, it must be that $\gamma(n + 1) = 1$; that is, the uninformed consumer believes the firm has high quality with probability 1. Since the best response of an informed consumer is strict for every $n$, it must be that $\sigma_u(n + 1) = \sigma_i(n + 1)$. The same reasoning applies to any queue length $\tilde{n} > n + 1$; at any such queue length, $\gamma(\tilde{n}) = 1$, so that $\sigma_u(\tilde{n}) = \sigma_i(h, \tilde{n})$.

Finally, suppose that $\sigma_u(n_h - 1) = 0$. The queue length $n_h$ is never observed in equilibrium, so an uninformed consumer’s beliefs are arbitrary at that queue length. However, since $\sigma_i(h, n_h) = 0$, regardless of beliefs, $w_u(n_h, \gamma, \beta) < 0$, so that in equilibrium it must be that $\sigma_u(n_h) = 0 = \sigma_i(h, n_h)$.

**Proof of Proposition 2**

Suppose a firm of type $\theta$ chooses a pure strategy $\mu_\theta$, and there is a pure strategy $\sigma$ in the consumer game. Then, the overall probability a consumer joins the queue of the low-quality firm when there are already $j$ consumers in the queue is given by

$$s_\ell(\theta, \sigma) = \begin{cases} 1 & \text{if } j \in \{0, \ldots, n_\ell - 1\} \\ (1 - q) & \text{if } j \in \{n_\ell, \ldots, \tilde{n} - 1\} \\ 0 & \text{otherwise} \end{cases}$$

Similarly, for the high-quality firm,

$$s_h(\theta, \sigma) = \begin{cases} 1 & \text{if } j \in \{0, \ldots, \tilde{n} - 1\} \text{ or } j \in \{\tilde{n} + 1, \ldots, n_h - 1\} \\ q & \text{if } j = \tilde{n} \end{cases}$$

Now, for an uninformed consumer, the posterior probability the firm has high quality may be written as

$$\gamma(n, \mu_h, \mu_\ell) = \frac{p\pi_h(n, \mu_h, \sigma)}{p\pi_h(n, \mu_h, \sigma) + (1 - p)\pi_\ell(n, \mu_\ell, \sigma)} = \frac{p}{p + (1 - p)\frac{\pi_\ell(n, \mu_\ell, \sigma)}{\pi_h(n, \mu_h, \sigma)}}.$$
Let \( \phi_0 = \frac{\pi(0, \mu, \sigma)}{\pi_h(0, \mu_h, \sigma)} \). Then, it follows that

\[
\frac{\pi_\ell(n, \mu, \sigma)}{\pi_h(n, \mu, \sigma)} = \begin{cases} 
\phi_0(\mu_h, \mu_\ell, \sigma)(\mu_h/\mu_\ell)^n & \text{if } n \in \{0, \ldots, n_\ell\} \\
\phi_0(\mu_h, \mu_\ell, \sigma)(\mu_h/\mu_\ell)^n(1 - q)^{n - n_\ell} & \text{if } n \in \{n_\ell + 1, \ldots, \hat{n}\} \\
0 & \text{otherwise.}
\end{cases}
\]

The statement of the proposition now follows.

\[\square\]

**Proof of Lemma 3**

Suppose a firm with type \( \theta \) chooses a service rate \( \mu \), and suppose consumers play \( \sigma \), where \( \sigma \) is a pure strategy that satisfies the necessary conditions for equilibrium in the continuation game. Then, \( s_h(j, \sigma) = 1 \) for \( j \leq \hat{n}(\sigma) - 1 \) and \( j \in \{\hat{n}(\sigma) + 1, \ldots, \tilde{n}_h(\sigma) - 1\} \), with \( s_h(\hat{n}(\sigma)) = q \) (since only informed consumers join at \( \hat{n}(\sigma) \)). Substituting for \( s_h(j, \sigma) \) in equation (4) in Lemma 1 yields

\[
\pi_h(0, \mu, \sigma) = \frac{1}{\sum_{j=0}^{\hat{n}(\sigma)} (\Lambda/\mu)^j + q \sum_{j=\hat{n}(\sigma)+1}^{\tilde{n}_h(\sigma)} (\Lambda/\mu)^j}.
\]

By inspection, we observe that \( \pi_h(0, \mu, \sigma) \) declines in \( \hat{n}(\sigma) \), keeping \( \tilde{n}_h(\sigma) \) fixed.

Further, \( s_\ell(j, \sigma) = 1 \) for \( j \leq \hat{n}_\ell(\sigma) - 1 \), with \( s_\ell(j, \sigma) = 1 - q \) for \( j \in \{\hat{n}_\ell, \ldots, \hat{n}(\sigma) = 1\} \) (since only uninformed consumers join at these queue lengths) and \( s_\ell(j, \sigma) = 0 \) for \( j \geq \hat{n}(\sigma) \). Substituting for \( s_\ell(j, \sigma) \) in equation (4) in Lemma 1 yields

\[
\pi_\ell(0, \mu, \sigma) = \frac{1}{\sum_{j=0}^{\hat{n}(\sigma)} (\Lambda/\mu)^j + \sum_{j=\hat{n}(\sigma)+1}^{\tilde{n}_\ell(\sigma)} (1 - q)^{j-\hat{n}(\sigma)}(\Lambda/\mu)^j}.
\]

By inspection, we observe that \( \pi_\ell(0, \mu, \sigma) \) declines in \( \hat{n}(\sigma) \), keeping \( \tilde{n}_h(\sigma) \) fixed.

The expected revenue of firm \( \theta \) is \( R_\theta(\mu, \sigma) = r\mu(1 - \pi_\theta(0, \mu, \sigma)) \), from which the expressions in the statement of the Lemma follow. Since \( \pi_\theta(0, \mu, \sigma) \) decreases as \( \hat{n}(\sigma) \) increases (keeping fixed \( \tilde{n}_h \) and \( \tilde{n}_\ell \)), it follows that for each \( \theta \), the expected revenue \( R_\theta(\mu, \sigma) \) increases as \( \hat{n}(\sigma) \) increases.

\[\square\]

**Proof of Proposition 3**

Suppose each type of firm \( \theta \in \{h, \ell\} \) chooses a service rate \( \mu_\theta \in \{\mu, \hat{\mu}\} \). Consider a newly-arrived uninformed consumer who faces queue length \( n \). Suppose the consumer believes that all other consumers are playing a pure strategy \( \sigma \) with \( \hat{n}(\sigma) = n_\ell \), \( \sigma_i \) as shown by Lemma 2, and \( \sigma_u(n) = \sigma_i(h, n) \) for all \( n \neq n_\ell \). We will show that if \( p \) is sufficiently low, \( \sigma \) is a best response for the newly-arrived consumer, and hence constitutes an equilibrium in the continuation game.
Recall that $\gamma(n)$ is the posterior probability an uninformed consumer assigns to the firm having high quality, once he has observed the queue length $n$. Then, an uninformed consumer is willing to balk from a queue which already has $n$ consumers if
\[ \gamma(n)[v_h - (n + 1)\frac{c}{\mu_h}] + (1 - \gamma(n))[v_\ell - (n + 1)\frac{c}{\mu_\ell}] \leq 0, \tag{15} \]

Or, \[ \frac{1 - \gamma(n)}{\gamma(n)} \geq \frac{v_h - (n + 1)\frac{c}{\mu_h}}{(n + 1)\frac{c}{\mu_\ell} - v_\ell}. \tag{16} \]

From condition (iii) of the definition of equilibrium (Definition 1), it follows that $\frac{1 - \gamma(n)}{\gamma(n)} = \frac{(1 - p)\pi_\ell(n, \mu, \sigma)}{p\pi_h(n, \mu, \sigma)}$. Hence, assuming $\pi_\ell(n, \mu, \sigma) > 0$, it is a best response for the uninformed consumer to not join the queue at $n$ if
\[ \frac{1 - p}{p} \geq \frac{\pi_h(n, \mu, \sigma)}{\pi_\ell(n, \mu, \sigma)} \times \frac{v_h - (n + 1)\frac{c}{\mu_h}}{(n + 1)\frac{c}{\mu_\ell} - v_\ell}. \tag{17} \]

We evaluate the right-hand side of condition (20) at $n = n_\ell$. Consider the equations (13) and (14) for $\pi_h(0, \mu, \sigma)$ and $\pi_\ell(0, \mu, \sigma)$ in the proof of Lemma 3. Substitute $\mu = \mu_h$ for the high-quality firm, $\mu = \mu_\ell$ for the low-quality firm, and $n(\sigma) = n_\ell$. Then,
\[ \pi_h(0, \mu, \sigma) = \frac{1}{\sum_{j=0}^{n_\ell}(\Lambda/\mu)^k + q \sum_{j=n_\ell+1}^{n_h}(\Lambda/\mu)^k}, \tag{18} \]
\[ \pi_\ell(0, \mu, \sigma) = \sum_{j=0}^{n_\ell}(\Lambda/\mu_\ell)^j. \tag{19} \]

Now, for each $\theta = h, \ell$, both types of consumers join the queue for all $n < n_\ell$. Therefore, from equation (5) in Lemma 1, it follows that for each $\theta$, $\pi_\theta(n_\ell, \mu, \sigma) = \pi_\theta(0, \mu, \sigma) \left(\frac{\Lambda}{\mu}\right)^{n_\ell}$. Substituting the expressions for $\pi_h(\cdot)$ and $\pi_\ell(\cdot)$ on the right-hand side of condition (20), we obtain
\[ \frac{1 - p}{p} \geq \left(\frac{\mu_\ell}{\mu_h}\right)^{n_\ell} \left(\frac{v_h - (n + 1)\frac{c}{\mu_h}}{(n + 1)\frac{c}{\mu_\ell} - v_\ell}\right) \left(\frac{\sum_{j=0}^{n_\ell}(\Lambda/\mu_h)^j + q \sum_{j=n_\ell+1}^{n_h}(\Lambda/\mu_h)^j}{\sum_{j=0}^{n_\ell}(\Lambda/\mu_\ell)^j}\right). \tag{20} \]

Now, let
\[ \psi = \max_{\mu, \mu_\ell \in \{\mu, \hat{\mu}\}} \left(\frac{\mu_\ell}{\mu_h}\right)^{n_\ell} \left(\frac{v_h - c\frac{n_\ell+1}{\mu_h}}{c\frac{n_\ell+1}{\mu_\ell} - v_\ell}\right) \left(\frac{\sum_{j=0}^{n_\ell}(\Lambda/\mu_h)^j + q \sum_{j=n_\ell+1}^{n_h}(\Lambda/\mu_h)^j}{\sum_{j=0}^{n_\ell}(\Lambda/\mu_\ell)^j}\right) \tag{21} \]
and define $p_0 = \frac{1}{1+\psi}$. Clearly, $p_0 \in (0, 1)$. Further, if $p \leq p_0$ and all other consumers play a pure strategy $\sigma$ with $n(\sigma) = n_\ell$, it is a best response for a newly-arrived uninformed consumer who sees a queue of length $n_\ell$ to balk.

Now, by definition of $n_\ell$, it follows that it is a best response for an uninformed consumer to join at $n < n_\ell$. Further, at any $n \in \{n_\ell + 1, \ldots, n_h - 1\}$, the uninformed consumer
believes with probability 1 that the firm has high quality, so that it is a best response to set \( \sigma_u(n) = \sigma_i(h, n) \). Finally, at \( n = n_h \), it is a best response to set \( \sigma_u(n_h) = 0 \).

Finally, suppose the newly-arrived consumer is informed. In any pure strategy equilibrium, \( \sigma_i \) must be as shown in Lemma 2.

Hence, if \( p \leq p_0 \), the strategy \( \sigma \) constitutes a best response to itself. That is, if every other consumer plays \( \sigma \), it is a best response for a newly-arrived consumer to also play \( \sigma \). Hence, \( \sigma \) defines a pure strategy equilibrium of the continuation game.

Finally, consider the case \( v_h \to \infty \). Consider the expression for \( \psi \) in equation (21). It is immediate to observe that the denominators of all three terms remain bounded, whereas the numerator of the second term goes to infinity. Hence, \( \psi \to \infty \) as \( v_h \to \infty \), so that \( p_0 = \frac{1}{1+\psi} \to 0 \).

Proof of Proposition 4

Substitute \( \hat{n} = n_\ell \) into the revenue rate expressions in Lemma 3. We obtain

\[
R_h(\mu, \hat{\sigma}) = r\mu \left( 1 - \frac{1}{\sum_{j=0}^{n_\ell} (\Lambda/\mu)^j + q\sum_{j=n_\ell+1}^{n_h} (\Lambda/\mu)^j} \right),
\]

\[
R_\ell(\mu, \hat{\sigma}) = r\mu \left( 1 - \frac{1}{\sum_{j=0}^{n_\ell} (\Lambda/\mu)^j} \right).
\]

(i) Using the expression for the revenue rate of the low-quality firm in equation (23),

\[
\frac{\partial R_\ell}{\partial \mu}(\mu, \hat{\sigma}) = r \left( 1 - \frac{1}{\sum_{j=0}^{n_\ell} (\Lambda/\mu)^j} \right) - r\mu \left( \frac{(1/\mu) \sum_{j=1}^{n_\ell} (j(\Lambda/\mu)^j)}{\left\{ \sum_{j=0}^{n_\ell} (\Lambda/\mu)^j \right\}^2} \right).
\]

Letting \( \rho = \Lambda/\mu \), we can therefore write

\[
\frac{\partial R_\ell}{\partial \mu}(\mu, \hat{\sigma}) = r \frac{\left\{ \sum_{j=0}^{n_\ell} \rho^j \right\}^2 - \sum_{j=0}^{n_\ell} (j+1)\rho^j}{\left\{ \sum_{j=0}^{n_\ell} \rho^j \right\}^2}.
\]

Therefore, the sign of \( \frac{\partial R_\ell}{\partial \mu}(\mu, \hat{\sigma}) \) is equal to the sign of \( \left\{ \sum_{j=0}^{n_\ell} \rho^j \right\}^2 - \sum_{j=0}^{n_\ell} (j+1)\rho^j \). For \( i \in \{0, \ldots, n_\ell\} \), consider the coefficient of \( \rho^i \) in the expansion of \( \left\{ \sum_{j=0}^{n_\ell} \rho^j \right\}^2 = (1 + \rho + \rho^2 + \cdots + \rho^{n_\ell})^2 \). First, suppose \( i \) is odd. Then, \( \rho^i \) is achieved by multiplying both elements of pairs such as \( (\rho^i, 1), (\rho^{i-1}, \rho), (\rho^{i-2}, \rho^2) \), and so on. There are \( \frac{i+1}{2} \) such pairs. Thus, the coefficient of \( \rho^i \) is \( \frac{i+1}{2} \times 2 = i + 1 \). Next, suppose \( i \) is even. Then, \( \rho^i \) is achieved by multiplying both elements of pairs such as \( (\rho^i, 1), (\rho^{i-1}, \rho), (\rho^{i-2}, \rho^2) \), and so on. There are \( \frac{i}{2} \) such pairs. In addition, an extra \( \rho^i \) term is obtained by multiplying \( \rho^i/2 \) by itself. Thus, again the coefficient of \( \rho^i \) is \( i + 1 \).
Therefore,

\[ \left\{ \sum_{j=0}^{n_\ell} \rho^j \right\}^2 = \sum_{j=0}^{n_\ell} (j+1)\rho + H, \]

where \( H \) is strictly positive. In particular, \( H > \rho^{2n_\ell} > 0 \). Hence,

\[ \left\{ \sum_{j=0}^{n_\ell} \rho^j \right\}^2 - \sum_{j=0}^{n_\ell} (j+1)\rho > \rho^{2n_\ell} > 0. \]

Therefore, \( \frac{\partial R_h}{\partial \mu} > 0 \) for all \( \mu \). It follows that for any pair of feasible service rates \( \mu_0 - \epsilon \) and \( \mu_0 + \epsilon \), \( \Delta_{\ell}(\mu_0, \epsilon) > 0 \), and there is a range of \( k \) small enough such that firm \( \ell \) chooses the fast service rate.

(ii) Given the expression for \( R_h \) in equation (22), the rate of change of \( R_h \) with respect to \( \mu \) may be expressed as:

\[
\frac{\partial R_h}{\partial \mu} (\mu, \sigma) = r\mu \left( \sum_{j=0}^{n_\ell} \frac{[\Lambda/\mu]^j + q \sum_{j=n_\ell+1}^{n_h} [\Lambda/\mu]^j}{\left( \sum_{j=0}^{n_\ell} [\Lambda/\mu]^j + q \sum_{j=n_\ell+1}^{n_h} [\Lambda/\mu]^j \right)^2} \right) + 1 - \frac{r}{\sum_{j=0}^{n_\ell} \rho^j + q \sum_{j=n_\ell+1}^{n_h} \rho^j} \cdot \left[ -\sum_{j=0}^{n_\ell} j\rho^j - q \sum_{j=n_\ell+1}^{n_h} j\rho^j + \left( \sum_{j=0}^{n_\ell} \rho^j + q \sum_{j=n_\ell+1}^{n_h} \rho^j \right)^2 \right]
\]

Let \( \rho = \frac{\Lambda}{\mu} \). The sign of \( \frac{\partial R_h}{\partial \mu} \) at any point \( (\mu, \sigma) \) is equal to the sign of \( \frac{1}{r} \frac{\partial R_h}{\partial \mu} \), and hence equal to the sign of

\[
-\sum_{j=0}^{n_\ell} j\rho^j - q \sum_{j=n_\ell+1}^{n_h} j\rho^j + \left( \sum_{j=0}^{n_\ell} \rho^j + q \sum_{j=n_\ell+1}^{n_h} \rho^j \right)^2
\]

Let \( \Phi = -\sum_{j=0}^{n_\ell} (j+1)\rho^j - q \sum_{j=n_\ell+1}^{n_h} (j+1)\rho^j + \left( \sum_{j=0}^{n_\ell} \rho^j + q \sum_{j=n_\ell+1}^{n_h} \rho^j \right)^2 \). Then, the sign of \( \frac{\partial R_h}{\partial \mu} \) at any point is equal to the sign of \( \Phi \).

Now, observe that \( \left( \sum_{j=0}^{n_\ell} \rho^j + q \sum_{j=n_\ell+1}^{n_h} \rho^j \right)^2 = \left( \sum_{j=0}^{n_\ell} \rho^j + q \rho^{n_{\ell+1}} \sum_{j=0}^{n_h-n_{\ell-1}} \rho^j \right)^2 = \left( \sum_{j=0}^{n_\ell} \rho^j \right)^2 + q^2 \rho^{2(n_{\ell+1})} \left( \sum_{j=0}^{n_h-(n_{\ell+1})} \rho^j \right)^2 + 2q \rho^{n_{\ell+1}} \left( \sum_{j=0}^{n_\ell} \rho^j \right) \left( \sum_{j=0}^{n_h-(n_{\ell+1})} \rho^j \right). \)
Consider \((\sum_{j=0}^{M} \rho^j \bigg( \sum_{j=0}^{N} \rho^j \bigg))^2\) for a generic \(M\) and \(N\) with \(N \leq M\). Expanding terms, multiplying, and regrouping, we can show that

\[
\left( \sum_{j=0}^{M} \rho^j \right) \left( \sum_{j=0}^{N} \rho^j \right) = \sum_{j=0}^{N} (j + 1)\rho^j + (N + 1)\rho^N \sum_{j=1}^{M-N} \rho^j + \rho^M \sum_{j=1}^{N} (N - j + 1)\rho^j, \tag{24}
\]

\[
\left( \sum_{j=0}^{M} \rho^j \right)^2 = \sum_{j=0}^{M} (j + 1)\rho^j + \rho^M \sum_{j=1}^{M} (M - j + 1)\rho^j. \tag{25}
\]

Now, suppose \(v_h \leq 2v_{\ell} + \frac{w_0}{c}\), so that \(n_h \leq 2n_\ell + 1\). Then, using the expressions in (24) and (25) to substitute out for each of the three terms in the expansion of \((\sum_{j=0}^{n_\ell} \rho^j + q \sum_{j=n_\ell+1}^{n_h} \rho^j)^2\), we obtain

\[
\Phi = -\sum_{j=0}^{n_\ell} (j + 1)\rho^j - q \sum_{j=n_\ell+1}^{n_h} (j + 1)\rho^j + \sum_{j=0}^{n_\ell} (j + 1)\rho^j + \rho^{n_\ell} \sum_{j=1}^{n_h} (n_\ell - j + 1)\rho^j
\]

\[
+ q^2 \rho^{2(n_\ell+1)} \left[ \sum_{j=0}^{n_h-n_\ell} (j + 1)\rho^j + \rho^{n_h-(n_\ell+1)} \sum_{j=1}^{n_h-n_\ell} (n_h - n_\ell - j)\rho^j \right]
\]

\[
+ 2q \rho^{n_\ell+1} \left[ \sum_{j=0}^{n_h-n_\ell-1} (j + 1)\rho^j + (n_h - n_\ell)\rho^{n_h-n_\ell-1} \sum_{j=1}^{n_h-n_\ell-1} \rho^j + \rho^{n_\ell} \sum_{j=1}^{n_h-n_\ell-1} (n_h - n_\ell - j)\rho^j \right]
\]

Now, \(q \sum_{j=0}^{n_h-n_\ell} (j + 1)\rho^j = q \sum_{j=1}^{n_h-n_\ell} (n_\ell - j + 1)\rho^{n_\ell+j}\). Further, since \(n_h < 2n_\ell + 1\), it follows that \(n_h \leq 2n_\ell\), or \(n_h - n_\ell \leq n_\ell\). Hence, we can write \(\rho^{n_\ell} \sum_{j=1}^{n_\ell} (n_\ell - j + 1)\rho^j = \sum_{j=1}^{n_h-n_\ell} (n_\ell - j + 1)\rho^{n_\ell+j} + \sum_{j=n_h-n_\ell+1}^{n_\ell} (n_\ell - j + 1)\rho^{n_\ell+j}\). Finally, \(2q \rho^{n_\ell+1} \sum_{j=0}^{n_h-n_\ell} (j + 1)\rho^j = 2q \sum_{j=1}^{n_h-n_\ell} j\rho^{n_\ell+j}\). Substituting these expressions into the appropriate terms in the last equation for \(\Phi\), we obtain

\[
\Phi = \sum_{j=1}^{n_h-n_\ell} \rho^{n_\ell+j} [-q(n_\ell+j+1) + n_\ell - j + 1 + 2qj] + \Psi,
\]

where \(\Psi\) is a collection of all the remaining terms in \(\Phi\). Each of these terms is non-negative, and since \(n_h \geq n_\ell + 1\), it follows that the term \(q^2 \rho^{2(n_\ell+1)} \sum_{j=0}^{n_h-n_\ell+1} (j + 1)\rho^j\) is strictly positive. Hence, \(\Psi > 0\).

Simplifying further,

\[
\Phi = \sum_{j=1}^{n_h-n_\ell} \rho^{n_\ell+j}(1 - q)(n_\ell - j + 1) + \Psi.
\]

Since \(n_h - n_\ell \leq n_\ell\), each term in the first summation is strictly positive, as is \(\Psi\). Hence, \(\Phi > 0\). Since the sign of \(\frac{\partial R_h}{\partial \mu}\) equals the sign of \(\Phi\), it follows that \(\frac{\partial R_h}{\partial \mu} > 0\) whenever \(n_h \leq 2n_\ell + 1\).
That is, if \( v_h \leq 2v_\ell + \frac{\mu_0}{c} \), \( R_h \) is strictly increasing in \( \mu \) at each value of \( \mu \). Hence, \( \Delta_h(\mu_0, \epsilon) > 0 \). It follows that there exists a range of \( k \) small enough at which firm \( h \) chooses the fast service rate.

**Proof of Lemma 4**

Observe that under Assumption 3 (iii), regardless of firm’s choices over service rate, there is a pure strategy equilibrium in which the strategy of the uninformed consumers has a hole at \( n_\ell \). Further, given Assumption 3 (ii), \( n_\ell \) and \( n_h \) are invariant to firms’ service rate choices.

Recall the expressions for the revenue rates of the high and low-quality firms, as given in equations (22) and (23) in the proof of Proposition 4. First, consider the high-quality firm. We have

\[
\frac{\partial R_h}{\partial \mu} = r \left[ 1 - \frac{1}{\sum_{j=0}^{n_\ell} (\Lambda/\mu)^j + q + \sum_{j=n_\ell+1}^{n_h} (\Lambda/\mu)^j} - \frac{\sum_{j=0}^{n_\ell} (\Lambda/\mu)^j - q + \sum_{j=n_\ell+1}^{n_h} j(\Lambda/\mu)^j}{[\sum_{j=0}^{n_\ell} (\Lambda/\mu)^j + q + \sum_{j=n_\ell+1}^{n_h} j(\Lambda/\mu)^j]^2} \right].
\]

Now, when \( \mu = \Lambda \),

\[
\frac{\partial R_h}{\partial \mu} = r \left[ 1 - \frac{1}{n_\ell + 1 + q(n_h - n_\ell - 1)} - \frac{n_\ell(n_\ell + 1)/2 - q(n_h(n_h + 1)/2 - (n_\ell + 1)(n_\ell + 2)/2)}{[n_\ell + 1 - q(n_h - n_\ell - 1)]^2} \right].
\]

Collecting terms and simplifying yields the expression in the statement of the Lemma.

Similarly, for the low-quality firm, we have

\[
\frac{\partial R_\ell}{\partial \mu} = r \left[ 1 - \frac{1}{\sum_{j=0}^{n_\ell} (\Lambda/\mu)^j} - \frac{\sum_{j=0}^{n_\ell} j(\Lambda/\mu)^j}{[\sum_{j=0}^{n_\ell} (\Lambda/\mu)^j]^2} \right].
\]

Hence, when \( \mu = \Lambda \),

\[
\frac{\partial R_\ell}{\partial \mu} = r \left[ 1 - \frac{1}{n_\ell + 1} - \frac{n_\ell(n_\ell + 1)/2}{(n_\ell + 1)^2} \right] = \frac{rn_\ell}{2(n_\ell + 1)}.
\]

**Proof of Proposition 5**

Consider the expressions for the rate of change of revenue (with respect to \( \mu \)) for each type of firm in Lemma 4, when \( \mu = \Lambda \). Write \( n_h = n_\ell + m \), where \( m \geq 1 \) is an integer. Then,

\[
\frac{\partial R_h(\Lambda, \sigma)}{\partial \mu} = \frac{r \left[ 2m^2 q^2 + m(2n_\ell + 1 - m)q + n_\ell(n_\ell + 1) \right]}{2(n_\ell + 1 + qm)^2}.
\]
Let $\Gamma = \frac{1}{r} \left[ \frac{\partial R_h}{\partial \mu}(\Lambda, \hat{\sigma}) - \frac{\partial R_{h}}{\partial \mu}(\Lambda, \hat{\sigma}) \right]$. Then,

$$\Gamma = \frac{2m^2q^2 + m(2n_\ell + 1 - m)q + n_\ell(n_\ell + 1)}{2(n_\ell + 1 + mq)^2} \frac{n_\ell}{2(n_\ell + 1)}$$

$$2\Gamma = \frac{(n_\ell + 1)[2m^2q^2 + m(2n_\ell + 1 - m)q + n_\ell(n_\ell + 1)] - (n_\ell + 1 + mq)^2n_\ell}{(n_\ell + 1 + mq)^2(n_\ell + 1)}$$

(27)

The sign of $\Gamma$ equals the sign of numerator in equation (27), since the denominator is strictly positive. The numerator may be evaluated as:

$$(n_\ell + 1)[2m^2q^2 + m(2n_\ell + 1 - m)q + n_\ell(n_\ell + 1)] - (n_\ell + 1 + mq)^2n_\ell$$

$$= 2(n_\ell + 1)m^2q^2 + (2n_\ell + 1)(n_\ell + 1)mq - m^2(n_\ell + 1)q + n_\ell(n_\ell + 1)^2$$

$$- n_\ell[(n_\ell + 1)^2 + 2(n_\ell + 1)mq + m^2q^2]$$

$$= (n_\ell + 2)m^2q^2 - (n_\ell + 1)mq(m - 1) = mq[(n_\ell + 2)mq - (n_\ell + 1)(m - 1)].$$

Hence, the sign of the numerator is equal to the sign of the expression $[(n_\ell + 2)mq - (n_\ell + 1)(m - 1)]$. 

Now, recall that $\hat{q} = \left(1 - \frac{1}{n_h - n_\ell} \right) \frac{n_\ell + 1}{n_\ell + 2} = \left(\frac{m - 1}{m} \right) \frac{n_\ell + 1}{n_\ell + 2}$. Hence, it follows that the numerator in equation (27) is positive if $q > \hat{q}$ and negative if $q < \hat{q}$. Therefore, if $q > \hat{q}$, $\frac{\partial R_h}{\partial \mu}(\Lambda, \hat{\sigma}) > \frac{\partial R_h}{\partial \mu}(\Lambda, \hat{\sigma})$, and if $q < \hat{q}$, $\frac{\partial R_h}{\partial \mu}(\Lambda, \hat{\sigma}) < \frac{\partial R_h}{\partial \mu}(\Lambda, \hat{\sigma})$.

We now turn to the two parts of the statement of the proposition.

(i) Suppose $q < \hat{q}$. Observe that for each $\theta = h, \ell$, \(\lim_{\epsilon \to 0} \Delta_{\theta}(\Lambda, \epsilon) = \frac{\partial R_h}{\partial \mu}(\Lambda, \hat{\sigma})\). Since $\frac{\partial R_h}{\partial \mu} < \frac{\partial R_h}{\partial \mu}$, it follows that there exists an $\hat{\epsilon}(q) > 0$ such that, for all $\epsilon \in (0, \hat{\epsilon}]$, $\Delta_{h}(\Lambda, \epsilon) < \Delta_{\ell}(\Lambda, \epsilon)$. Now, for any $k \in (\Delta_{h}(\Lambda, \epsilon), \Delta_{\ell}(\Lambda, \epsilon))$, the high-quality firm chooses the slow service rate $\Lambda - \epsilon$ and the low-quality firm chooses the fast service rate $\Lambda + \epsilon$. Further, since $\Delta_{h}(\Lambda, \epsilon) < \Delta_{\ell}(\Lambda, \epsilon)$, there cannot exist a cost $k$ such that the high-quality firm chooses the fast service rate $\Lambda + \epsilon$ and the low-quality firm chooses the slow service rate $\Lambda - \epsilon$.

(ii) The argument when $q > \hat{q}$ is exactly similar to the argument in (i) above.

Proof of Proposition 6

(i) Consider the expression for $\frac{\partial R_h}{\partial \mu}$ when $\mu = \Lambda$, as shown in equation (26) in the proof of Proposition 5. Recall that $n_h = n_\ell + m$ in this expression. Consider the limit as $n_h \to \infty$, or, equivalently, $m \to \infty$.

In the limit as $n_h \to \infty$, we obtain

$$\lim_{n_h \to \infty} \frac{\partial R_h}{\partial \mu}(\Lambda, \hat{\sigma}) = \left(\frac{r}{2q^2} - \frac{q}{2} \right) = r \left(1 - \frac{1}{2q^2} \right).$$

(28)
The last expression is strictly negative whenever \( q < \frac{1}{2} \). Thus, for any \( q < \frac{1}{2} \), there exists a threshold \( \bar{n}(q) \) such that if \( n_h \geq \bar{n}(q) \), \( \frac{\partial R_h}{\partial \mu}(\Lambda, \bar{\sigma}) < 0 \). Similarly, for any \( q > \frac{1}{2} \), there exists a threshold \( \bar{n}(q) \) such that if \( n_h \geq \bar{n}(q) \), \( \frac{\partial R_h}{\partial \mu}(\Lambda, \bar{\sigma}) > 0 \). Define \( \bar{v}(q) = \bar{n}(q) \frac{c}{\mu_0 - c} \). Then, the condition \( n_h \geq \bar{n}(q) \) is equivalent to \( v_h \geq \bar{v}(q) \).

Now, \( \lim_{\varepsilon \to 0} \Delta_h(\Lambda, \varepsilon) = \frac{\partial R_h}{\partial \mu}(\Lambda, \bar{\sigma}) \). Suppose \( q < \frac{1}{2} \) and \( v_h \geq \bar{v}(q) \). Then, if \( \varepsilon \) is sufficiently close to zero, if must be the case that \( \Delta_h(\Lambda, \varepsilon) < 0 \). Formally, there exists some \( \bar{v}(\varepsilon) \) such that if \( \varepsilon \in (0, \bar{v}(\varepsilon)] \), then \( \Delta_h(\Lambda, \varepsilon) < 0 \). Hence, for parameters in this range, at any value of \( k \), including \( k = 0 \), the high-quality firm will choose the slow service rate \( \Lambda - \varepsilon \).

Finally, suppose \( q > \frac{1}{2} \) and \( v_h \geq \bar{v}(q) \). Then, there exists some \( \bar{v}(q) \) such that if \( \varepsilon \in (0, \bar{v}(\varepsilon)] \), then \( \Delta_h(\Lambda, \varepsilon) > 0 \).

**Proof of Proposition 7**

Consider the expression for \( R_h \) in equation (22), in the proof of Lemma 4. Let \( n_h \to \infty \), and denote \( \rho = \Lambda / \mu \). Let \( \bar{R}_h = \frac{1}{n} \lim_{n_h \to \infty} R_h(\mu, \bar{\sigma}) \). Then, we obtain

\[
\bar{R}_h(\mu, \bar{\sigma}) = \mu \left( 1 \frac{1}{1 - \sum_{j=0}^{n_h} \rho^j + q \lim_{n_h \to \infty} \sum_{j=n_h+1}^{\infty} \rho^j} \right)
\]

Hence, if \( \rho > 1 \) (i.e., \( \mu < \Lambda \)), \( \bar{R}_h = \mu \). If \( \rho < 1 \),

\[
\bar{R}_h(\mu, \bar{\sigma}) = \mu \left( 1 \frac{1}{1 - \rho^{n_{\tau+1}} + q \rho^{n_{\tau+1}} \rho^\mu} \right) = \mu \left( 1 \frac{1 - \rho}{1 - (1 - q)\rho^{n_{\tau+1}}} \right)
\]

It follows that for \( \rho < 1 \),

\[
\frac{\partial \bar{R}_h}{\partial \rho} = \Lambda \frac{-n_\ell (1 - q)\rho^{n_{\tau-1}}(1 - (1 - q)\rho^{n_{\tau+1}}) + (n_\ell + 1)(1 - q)\rho^{n_\ell} (1 - (1 - q)\rho^{n_{\tau+1}})}{[1 - (1 - q)\rho^{n_{\tau+1}}]^2}
\]

Let \( \Psi(\rho) = \rho [1 - (1 - q)\rho^{n_{\tau+1}}] - (1 - \rho)n_\ell \). Then, the sign of \( \frac{\partial \bar{R}_h}{\partial \rho} \) is equal to the sign of \( \Psi(\rho) \).

Now, \( \Psi(0) = -n_\ell < 0 \) and \( \Psi(1) = q > 0 \). Further,

\[
\Psi'(0) = 1 - (1 - q)\rho^{n_\ell} - n_\ell (1 - q)\rho^{n_\ell} + n_\ell = (n_\ell + 1)[1 - (1 - q)\rho^{n_{\tau+1}}] > 0.
\]

Hence, there exists a unique \( \hat{\rho}(n) \) which solves the equation

\[
\rho [1 - (1 - q)\rho^{n}] - (1 - \rho)n = 0, \quad \text{or,} \quad \frac{\rho}{1 - \rho} [1 - (1 - q)\rho^{n}] = n.
\]
Further, for any \( n_\ell \), if \( \rho \in (\hat{\rho}(n_\ell), 1) \), it follows that \( \frac{\partial \bar{R}_h}{\partial \rho} \) is increasing in \( \rho \). Since \( \rho = \frac{\Lambda}{\mu} \), it follows immediately that if \( \rho \in (\hat{\rho}(n_\ell), 1) \), then \( \frac{\partial \bar{R}_h}{\partial \mu} < 0 \). That is, \( \bar{R}_h \) is decreasing in \( \mu \).

That is, if \( \mu \in (1, \frac{\Lambda}{\hat{\rho}(n_\ell)}) \), it follows that \( \frac{\partial \bar{R}_h}{\partial \mu} < 0 \). Now, observe that for any finite \( v_h \),

\[
\frac{\partial \bar{R}_h}{\partial \mu}(\mu_0, \hat{\sigma}) = \lim_{\epsilon \to 0} \Delta(\mu_0, \epsilon) .
\]

Hence, \( \frac{\partial \bar{R}_h}{\partial \mu}(\mu_0, \hat{\sigma}) = \lim_{n_h \to \infty} \lim_{\epsilon \to 0} \Delta(\mu_0, \epsilon) \). Hence, there exists an \( n_h \) large enough (alternatively a \( v_h \) large enough) and an \( \epsilon' \) small enough such that the statement of the proposition follows. From Proposition 4, it follows that \( v_h(n) \) must exceed \( 2v_\ell + \frac{\mu_0}{\epsilon} \). \( \square \)
References


