https://dx.doi.org/10.1016/j.orl.2020.07.012

Equilibrium Analysis of Unobservable $M/M/n$ Priority Queues with Balking and Homogeneous Customers

Qingying Li

Glorious Sun School of Business and Management, Donghua University, Shanghai, China, liqingying@dhu.edu.cn

Pengfei Guo*

Faculty of Business, the Hong Kong Polytechnic University, Hong Kong, pengfei.guo@polyu.edu.hk

Yulan Wang

Faculty of Business, the Hong Kong Polytechnic University, Hong Kong, yulan.wang@polyu.edu.hk

Abstract: We investigate equilibrium queueing strategies in an unobservable $M/M/n$ non-preemptive priority queue with homogeneous customers. Arriving customers make decisions by choosing one of the three options: balking, joining the regular queue, or joining the priority queue with an additional fee. We adopt a sequential approach to analyze this two-dimensional queueing game. We first obtain the equilibrium choice between joining the priority or the regular queue given a fixed total joining rate, and then we determine the equilibrium total joining rate.

Keywords: Equilibrium Analysis, Unobservable Queue, Strategic Queueing, Priority Queue.

*Corresponding author.

1 Introduction

Many studies have been conducted to examine the effect of priority pricing on the customer queueing behavior and system performance. These studies mainly focus on customers' priority-buying decisions, i.e., either buying priority or not buying priority. Given a fixed potential demand rate, customers' decision is one dimensional. In practice, service systems commonly observe balking customers. In unobservable priority queues, when balking is allowed, arriving customers need to make their joining-or-balking decisions by choosing one of the following three options: balking, joining a regular queue, or joining a priority queue with an additional fee. When customers are homogeneous, their queueing strategy can be captured by a two-dimensional vector with elements to be probabilities of choosing two of the above three available options. Such kind of two-dimensional equilibrium analysis receives little attention in the literature, and is our research topic.

Extensive strategic queueing behavior have been conducted regarding various queueing settings; see [7] and [9] for comprehensive reviews of related work. Regarding customers' strategic queueing behavior in priority queues, earlier studies mainly consider observable queues [1,4,8]. In our work, customers make decisions according to their expected waiting time under an unobservable setting. We then review the studies on customers' strategic queueing decision in unobservable priority queues. Hassin [6] discusses the regulation of an unobservable $M/M/1$ queue by using bidding mechanisms to determine priority. The author first discusses customers' joining-or-balking decisions by considering a baseline model with homogeneous customers and no priority choice. Then, he extends the study to the case where customers can pay certain priority price, which is unknown to others, to enjoy a priority service over those paying a lower price. He also extends the study to the situation where customers are heterogeneous. Hassin and Haviv [9] (pp. 83–85) study customers' equilibrium queueing strategy in an $M/M/1$ queue with a priority option. They assume balking is not allowed, and hence customers can only either join the regular queue or join the priority queue. Given the total customer joining rate, our customer joining equilibrium can be considered as an extension of [9] from a one-server setting to a multi-server setting. However, in our setting, allowing customers to balk makes the decision space for customers becomes two dimensional. Afèche and Mendelson [2] consider both welfare maximization and profit maximization for an unobservable single-server queue under a general delay-cost function, where customers' service value is a random variable. Three pricing strategies, namely, uniform pricing, preemptive auction, and non-preemptive auction, are considered. Cao et al. [3] study a single-server service system with two queues, a regular queue and a priority queue, and assume balking is not allowed. Upon arrivals, heterogeneous customers are informed with the delay information (i.e., the average delay time) of both queues, and

make individual decisions on whether joining the regular queue or joining the priority queue with an additional fee. They then discuss the equilibria on the delay and investigate the priority pricing. Gavirneni and Kulkarni [5] consider an unobservable $M/G/1$ non-preemptive queue with heterogeneous customers, where customers' waiting cost is modeled by a Burr distribution. They examine customers' equilibrium queueing decisions among balking, joining the priority queue, and joining the regular queue. Given customers' joining probabilities to both types of queues, [5] derive the priority price that makes the given joining probabilities achievable under a concierge system. Unlike our work, they focus on investigating the waiting cost structure such that the service provider is profitable by offering the priority option. Note that all the aforementioned studies, except for the baseline model of [6] and [9] (pp. 83–85), consider heterogeneous customers while we consider homogeneous customers, whose equilibrium queueing strategies are much harder to derive.

Regarding the two-dimensional equilibrium analysis, to the best of our knowledge, only two recent studies, [10] and [13], consider the two-dimensional equilibrium analysis. Hassin and Roet-Green [10] investigate an $M/M/1$ service system where the queue length information is available to customers at a reasonable cost. Then, customers can choose among three options: joining the queue, inspecting the queue length, or balking directly. They establish the existence and uniqueness of the equilibrium by analyzing geometric properties of the customers' expected utility set via constructing a two-dimensional mapping from customers' choice probabilities to their utilities. Wang et al. [13] investigate the two-dimensional equilibrium analysis in an $M/M/1$ non-preemptive priority queue. They consider both observable and unobservable queues and derive customers' corresponding equilibrium joining strategies. For the unobservable queue, they solve the indifference equations that make customers indifferent between joining and balking to derive the unique equilibrium joining strategy. In our work, we extend the unobservable model to a more general $M/M/n$ multi-server priority queue, and investigate customers' two-dimensional equilibrium joining-or-balking decisions.

A typical approach to solve the equilibrium for this game is to consider customers' indifference choice among the three options, which boils down to solving two equations. This approach, however, is cumbersome and difficult to yield insights on the properties of the equilibrium such as uniqueness and stability. Here we take a sequential approach to solve the equilibrium: first, given the total joining rate, we investigate customers' joining choice between the regular queue and the priority queue; then, anticipating customers' equilibrium joining behavior between the two queues, we derive the equilibrium total joining rate. By employing such sequential approach, we are able to identify some properties of customers' utility. We show that joining customers' utility is discontinuous (with a downward jump) at this critical total joining rate. By further analyzing the property of the utility curves, we

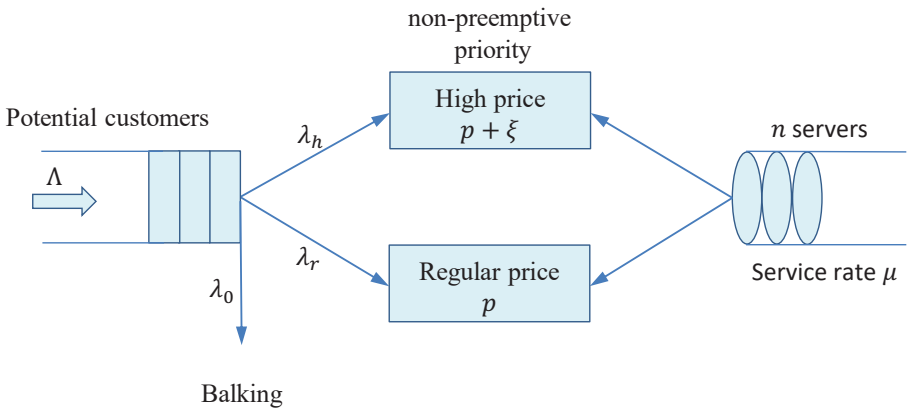


Figure 1: The model.

then derive customers' equilibrium joining rate. We show that there exist either a unique or multiple customer queueing equilibria, a result highly hinging on the magnitude of the potential customer demand.

2 Model Setup

Consider a service system with n identical servers. Queues are unobservable. Customers arrive according to a Poisson process with rate Λ. The service time of each server is exponentially distributed with rate μ. The service system provides two types of service, a regular queue (labeled r) and a priority queue (labeled h). An incoming customer needs to make her joining-or-balking decision by choosing from the following three options: balking, joining the regular queue or joining the priority queue with an additional fee. Customers who choose the regular queue need to pay the regular price p while customers who choose the priority queue need to pay a higher price $p+\xi$, where ξ is the additional fee for the priority option. Customers in the priority queue will be served first by the system following a *non-preemptive priority* discipline. Customers who join the same type of service, either the priority service or the regular service, are served based on the first-come first-serve principle. Once served, a customer receives a service reward R. All the above-mentioned parameters are strictly positive.

A customer's queueing strategy can be characterized by a two-dimensional vector (α_h, α_r) where $\alpha_h \in [0,1]$ and $\alpha_r \in [0,1]$ represent the probabilities of choosing the priority and regular queues, respectively, and $0 \leq \alpha_h + \alpha_r \leq 1$. Let $\lambda_i = \alpha_i \Lambda$, $i = h, r$, be the corresponding effective arrival rate into the respective queue. The total customer joining rate to the system

can then be denoted as $\lambda := \lambda_h + \lambda_r$. Thus, the customers' balking rate can be stated as $\lambda_0 := \Lambda - \lambda$. See Figure 1 for the illustration of the queueing system.

Customers are delay-sensitive and incur a waiting cost that is proportional to their waiting time in the system with a unit-time cost θ, where $\theta > 0$. Thus, their joining incentives are affected by the expected waiting times of each option. Regarding an $M/M/n$ non-preemptive priority queue, the expected waiting times in each queue have been analyzed and derived by [11]. Given the total customer joining rate to the system λ and the joining rate to the priority queue λ_h, let $W_h(\lambda, \lambda_h)$ and $W_r(\lambda, \lambda_h)$ denote the expected waiting times (including the service time) for a customer to join a priority and a regular queue, respectively. Then, according to [11], we have

$$W_h(\lambda, \lambda_h) = \frac{\phi(\lambda)}{n\mu - \lambda_h} + \frac{1}{\mu}, \tag{1}$$

and

$$W_r(\lambda, \lambda_h) = \frac{\phi(\lambda)}{n\mu - \lambda_h} \cdot \frac{n\mu}{n\mu - \lambda} + \frac{1}{\mu}, \tag{2}$$

where

$$\phi(\lambda) = \frac{(\frac{\lambda}{\mu})^n}{n!\left(1 - \frac{\lambda}{n\mu}\right)} \left[\sum_{k=0}^{n-1} \frac{(\frac{\lambda}{\mu})^k}{k!} + \frac{(\frac{\lambda}{\mu})^n}{n!\left(1 - \frac{\lambda}{n\mu}\right)}\right]^{-1}. \tag{3}$$

Note that $\phi(\lambda)$ is the Erlang-delay formula in an $M/M/n$ queue. It is the probability that all of the n servers are busy such that an arriving customer is blocked and has to wait in a queue.

Denote a customer's utility from joining the priority and regular queues as U_h and U_r, respectively. From (1) and (2), we then have

$$U_h(\lambda, \lambda_h) = R - (p + \xi) - \theta W_h(\lambda, \lambda_h) = R - (p + \xi) - \frac{\theta\phi(\lambda)}{n\mu - \lambda_h} - \frac{\theta}{\mu}, \tag{4}$$

and

$$U_r(\lambda, \lambda_h) = R - p - \theta W_r(\lambda, \lambda_h) = R - p - \frac{\theta\phi(\lambda)}{n\mu - \lambda_h} \cdot \frac{n\mu}{n\mu - \lambda} - \frac{\theta}{\mu}. \tag{5}$$

To avoid trivial cases, we assume that $R - \theta/\mu - p > 0$ to ensure that a customer served without waiting can obtain a positive utility. Customers' balking utility is normalized to be zero. Customers then make their strategic joining-or-balking decisions by comparing their utilities under the above three options.

3 Sequential Equilibrium Analysis

The typical way to derive customers' two-dimensional equilibrium strategy is to solve two indifferent equations, representing customers' indifference among three options [10,13]. Here, we apply a sequential approach to derive the equilibrium results. First, given a total customer joining rate, we investigate the customer's equilibrium joining strategy between the regular and priority queues; second, anticipating the customers' equilibrium joining strategy between these two types of queues, we then determine the equilibrium total customer joining rate. We use (λ, λ_h) to represent customers' two-dimensional strategy, where λ is the total joining rate, and λ_h is the joining rate to the high-priority queue. Note that to sustain a stable queueing system, we restrict our attention to the total joining rate λ such that $\lambda < n\mu$. Denote $\lambda_h^e(\lambda)$ as the equilibrium joining rate to the priority queue for a given total customer joining rate λ, and let λ^e be the equilibrium total customer joining rate.

3.1 Customers' joining equilibrium given total joining rate λ

We now analyze a customer's equilibrium joining behavior between a regular queue and a priority queue given the total customer joining rate λ. To derive customers' equilibrium joining strategy between the priority and regular queues, we first compare the customer' utilities from joining these two types of queues. From (4) and (5), we have

$$U_h(\lambda, \lambda_h) - U_r(\lambda, \lambda_h) = \frac{\theta\lambda\phi(\lambda)}{n\mu - \lambda} \cdot \frac{1}{n\mu - \lambda_h} - \xi. \tag{6}$$

Then, we can obtain the following result:

Lemma 1 *Consider the total joining rate λ such that $\lambda < n\mu$. (i) Both $U_r(\lambda, \lambda_h)$ and $U_h(\lambda, \lambda_h)$ decrease in λ and λ_h; (ii) the difference between these two joining utilities, $U_h(\lambda, \lambda_h) - U_r(\lambda, \lambda_h)$, increases in both λ and λ_h.*

Lemma 1 shows that as more customers join the system, the relative advantage of choosing the priority queue over the regular queue becomes larger. It implies that as more customers choose to buy priority, it becomes more beneficial for a tagged customer to buy as well. Such customer behavior is called the *follow-the-crowd (FTC)* behavior. Note that in a FTC setting, there often exist multiple equilibria [9]. Note also that Lemma 1 is an extension of the result for the $M/M/1$ queue, as stated in Lemmas 1 and 2 in [13].

Lemma 1 further implies that the lower bound of the joining utility difference, $\min_{\lambda_h} U_h(\lambda, \lambda_h) - U_r(\lambda, \lambda_h)$ is attained at $\lambda_h = 0$, that is, when no customer buys priority; while its upper bound, $\max_{\lambda_h} U_h(\lambda, \lambda_h) - U_r(\lambda, \lambda_h)$ is attained at $\lambda_h = \lambda$, that is, when all the joining customers buy priority and no customer joins the regular queue. Denote $\bar{\lambda}$ as the total

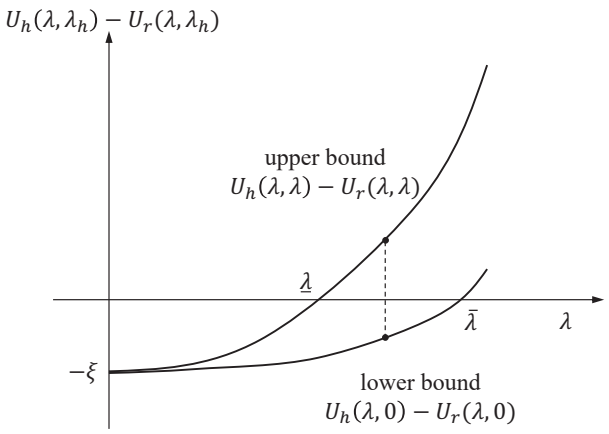


Figure 2: Customer Joining Utility Difference between the Priority and Regular Queues

customer joining rate that makes the customer indifferent between the two types of queues when no joining customers buy priority; that is, $\bar{\lambda}$ is the solution of the indifference equation "$U_h(\lambda, 0) - U_r(\lambda, 0) = 0$". Also, denote $\underline{\lambda}$ as the total customer joining rate that makes the customer indifferent between the two types of queues when all the joining customers buy priority; that is, $\bar{\lambda}$ is the solution of the indifference equation "$U_h(\lambda, \lambda) - U_r(\lambda, \lambda) = 0$". Then, according to Lemma 1, $\underline{\lambda} < \bar{\lambda}$; see Figure 2 for the illustration.

If there exists a joining rate λ_h that satisfies the indifference equation, $U_h(\lambda, \lambda_h) - U_r(\lambda, \lambda_h) = 0$ and $0 < \lambda_h < \lambda$, then we must have a mixed strategy in which a joining customer is indifferent between joining the regular or the priority queue and thus joins each queue with certain probability. The existence of such a λ_h requires that for a given total customer joining rate, the lower bound of the joining utility difference is negative (i.e., $U_h(\lambda, \lambda_h) - U_r(\lambda, \lambda_h)\big|_{\lambda_h=0} < 0$) while its upper bound is positive (i.e, $U_h(\lambda, \lambda_h) - U_r(\lambda, \lambda_h)\big|_{\lambda_h=\lambda} > 0$). Correspondingly, this requires that the total customer joining rate $\lambda \in (\underline{\lambda}, \bar{\lambda})$.

Recall that $U_h(\lambda, \lambda_h) - U_r(\lambda, \lambda_h)$ increases in the customer's joining rate to the priority queue λ_h. Hence, for a total customer joining rate $\lambda \in (\underline{\lambda}, \bar{\lambda})$, there exists a unique solution of the equation "$U_h(\lambda, \lambda_h) - U_r(\lambda, \lambda_h) = 0$", denoted by $\lambda_h^s(\lambda)$. From (6), we can derive that

$$\lambda_h^s(\lambda) = n\mu - \frac{\theta\lambda\phi(\lambda)}{\xi(n\mu - \lambda)}.$$

Therefore, when $\lambda \in (\underline{\lambda}, \bar{\lambda})$, a joining customer adopts a mixed strategy: joining the priority queue with probability $\delta_s := \lambda_h^s(\lambda)/\lambda$ and joining the regular queue with probability $1 - \delta_s$. It can be shown that $\lambda_h^s(\lambda)$ is decreasing in λ.

Besides the mixed-strategy equilibrium, there also exists a pure-strategy equilibrium in

which the joining customers either "all join" the regular queue or "all join" the priority queue, i.e., $\lambda_h = 0$ or $\lambda_h = \lambda$. A pure strategy $\lambda_h = 0$ is an equilibrium if, when all of the joining customers join the regular queue, a tagged customer has no incentive to deviate from this strategy. This is equivalent to the requirement that the utility of joining the regular queue is no less than that of joining the priority queue, i.e.,

$$U_r(\lambda, \lambda_h) - U_h(\lambda, \lambda_h)\big|_{\lambda_h \to 0^+} \geq 0, \tag{7}$$

or equivalently, $U_r(\lambda, 0) - U_h(\lambda, 0) \geq 0$. This requires that $\lambda \leq \bar{\lambda}$ as shown in Figure 2. Similarly, a pure strategy $\lambda_h = \lambda$ is an equilibrium if, when all of the joining customers join the priority queue, no customer wants to deviate, which requires

$$U_h(\lambda, \lambda_h) - U_r(\lambda, \lambda_h)\big|_{\lambda_h \to \lambda^-} \geq 0, \tag{8}$$

or equivalently, $U_h(\lambda, \lambda) - U_r(\lambda, \lambda) \geq 0$. This requires that $\lambda \geq \underline{\lambda}$ as depicted in Figure 2.

All the equilibria for the subgame are summarized in the following proposition.

Proposition 1 *Given a total customer joining rate λ, we have the following equilibria regarding the customers' joining behavior:*

(i.) if $\lambda \leq \underline{\lambda}$, in equilibrium, all the joining customers join the regular queue, i.e., $\lambda_h^e(\lambda) = 0$;

(ii.) if $\underline{\lambda} < \lambda < \bar{\lambda}$, there exist three equilibria, two pure ones with $\lambda_h^e(\lambda) = 0$ or $\lambda_h^e(\lambda) = \lambda$ and a mixed one with $\lambda_h^e(\lambda) = \lambda_h^s(\lambda)$;

(iii.) if $\lambda \geq \bar{\lambda}$, in equilibrium, all the joining customers join the priority queue, i.e., $\lambda_h^e(\lambda) = \lambda$.

Figure 3 illustrates all the equilibrium queueing strategies to the priority queue and the regular queue for the joining customers. Note that the results stated in Proposition 1 are similar to those obtained in a priority $M/M/1$ queue when balking is not allowed; see [9], pp. 83–85. It is interesting and worth noting that when the total customer joining rate $\lambda \in (\underline{\lambda}, \bar{\lambda})$, there exist three subgame customer joining equilibria: 'all buy priority', 'nobody buys priority', and 'some buy priority and others do not buy'. We now consider whether the equilibrium strategy is an *evolutionarily stable strategy* (ESS). According to [9], an equilibrium strategy y is said to be an ESS if for any $z \neq y$ which is a best response against y, y is better than z as a response to z itself. If an equilibrium strategy y is the unique best response against itself, it is necessarily an ESS. Among the foregoing three equilibrium strategies, the first two are pure strategies and the unique best responses against themselves

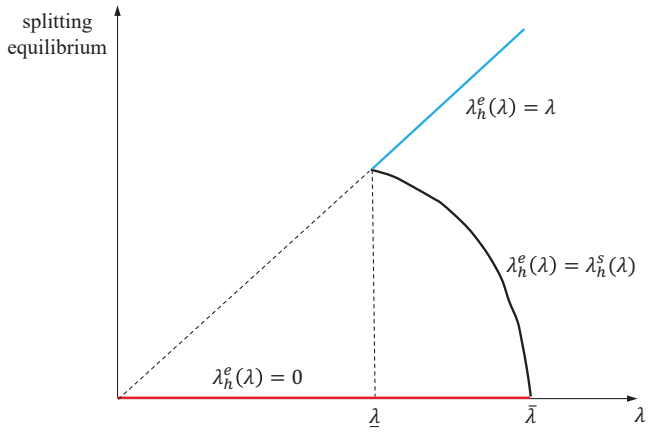


Figure 3: Joining Customers' Subgame Equilibrium Queueing Strategy

and thus are the ESS. The third equilibrium is a mixed-strategy equilibrium, where customers are indifferent between buying and not buying the priority. Hence, a strategy 'never buys priority' is also a best response against the mixed equilibrium strategy. Furthermore, given that all other customers adopt 'never buy priority', it is better for a tagged customer not to buy it either. Therefore, the mixed equilibrium is not an ESS.

3.2 Equilibrium total customer joining rate

In this section, we derive the customer's joining utility as a function of the total customer joining rate λ and then compare this utility with customer's balking utility to derive the equilibrium total customer joining rate λ^e. As we assume that customers are rational, their choice between joining and balking must result in a total joining rate λ less than $n\mu$; otherwise, the queue would become too long and customers would obtain negative utilities.

Recall that depending on the magnitude of the total customer joining rate λ, there exist at most three possible queueing strategies for those joining customers; see Proposition 1. If $\lambda \geq \underline{\lambda}$, $\lambda_h = \lambda$ is an equilibrium in which all joining customers buy priority. The customer's joining utility under this pure strategy, denoted by $u_1(\lambda)$, can be derived as

$$u_1(\lambda) = R - (p + \xi) - \theta W_h(\lambda, \lambda) = R - (p + \xi) - \frac{\theta\phi(\lambda)}{n\mu - \lambda} - \frac{\theta}{\mu}. \tag{9}$$

When $\underline{\lambda} < \lambda < \bar{\lambda}$, there exists a mixed-strategy equilibrium in which the joining customer joins the priority queue with certain probability and the regular queue with complementary probability; the equilibrium joining rate to the priority queue is $\lambda_h = \lambda_h^s(\lambda)$. Then, the

customer' joining utility under this mixed strategy, denoted as $u_2(\lambda)$, can be expressed as

$$u_2(\lambda) = R - (p+\xi) - \theta W_h(\lambda, \lambda_h^s(\lambda)) = R - (p+\xi) - \frac{\xi(n\mu - \lambda)}{\lambda} - \frac{\theta}{\mu}. \tag{10}$$

When $\lambda \leq \bar{\lambda}$, $\lambda_h = 0$ is an equilibrium in which nobody buys priority. Denote the customer' joining utility under this pure strategy as $u_3(\lambda)$. Then, we have

$$u_3(\lambda) = R - p - \theta W_r(\lambda, 0) = R - p - \frac{\theta\phi(\lambda)}{n\mu - \lambda} - \frac{\theta}{\mu}. \tag{11}$$

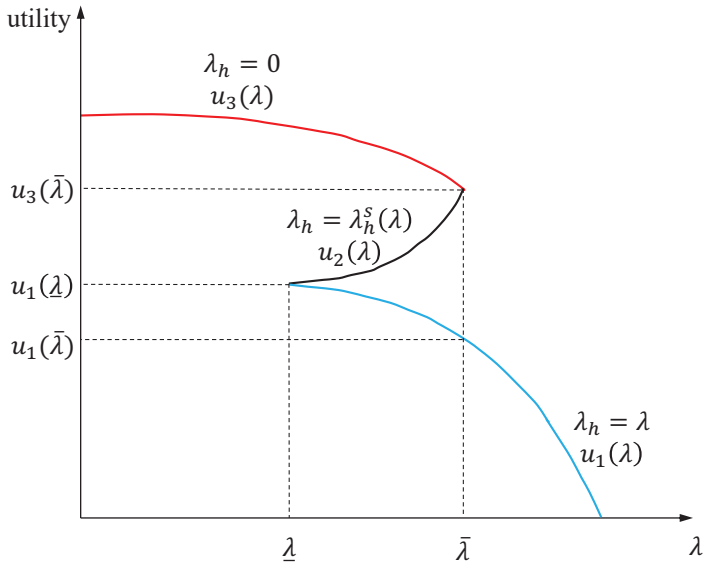


Figure 4: Joining Customer' Equilibrium Joining Utility versus the Total Customer Joining Rate λ

From [12], we know that the Erlang-delay formula $\phi(\lambda)$ increases in λ, the total customer joining rate. Then, it can be easily shown that the joining utilities under the two pure-strategy equilibria, $u_1(\lambda)$ and $u_3(\lambda)$, decrease in λ while the joining utility under the mixed-strategy equilibrium, $u_2(\lambda)$, increases in λ. For ease of presentation, we extend the domain of $u_2(\lambda)$ by including the two boundary points $\lambda = \underline{\lambda}$ and $\lambda = \bar{\lambda}$ according to equation (10). We can show that $u_1(\underline{\lambda}) = u_2(\underline{\lambda})$ and $u_3(\bar{\lambda}) = u_2(\bar{\lambda})$. This implies that the curves of these three equilibrium joining utility functions are connected at these two points and form a continuous grand curve over the whole range $\lambda \geq 0$. Moreover, when $\underline{\lambda} \leq \lambda \leq \bar{\lambda}$, $u_1(\lambda) \leq u_2(\lambda) \leq u_3(\lambda)$; see Figure 4 for the illustration. Figure 4 indicates that among these three utility function curves, only one of them shall intersect the x-axis (that is, the line representing zero utility) once. To check which utility function curve intersects the x-axis (i.e, whether the intersection point is located on the curve $u_1(\lambda)$, or $u_2(\lambda)$, or $u_3(\lambda)$), we shall first examine how the grand

curve— the whole composition of the three utility function curves— moves when the priority price ξ increases. It turns out that to capture its moving direction, we just need to check how the two connecting points on the grand curve, $(\underline{\lambda}, u_1(\underline{\lambda}))$ and $(\bar{\lambda}, u_3(\bar{\lambda}))$ move when ξ increases. The result is summarized in the following lemma.

Lemma 2 *(i) Both $\underline{\lambda}$ and $\bar{\lambda}$ increase in the priority price ξ, and (ii) both $u_1(\underline{\lambda})$ and $u_3(\bar{\lambda})$ decrease in ξ.*

Lemma 2 shows that as the priority price ξ increases, the two connecting points on the grand curve, $(\underline{\lambda}, u_1(\underline{\lambda}))$ and $(\bar{\lambda}, u_3(\bar{\lambda}))$ depicted in Figure 4, both move down and to the right on the $x-y$ plane. Thus, as ξ increases, the first utility function curve intersecting the x-axis is $u_1(\lambda)$, the joining utility function when all joining customers buy priority, then $u_2(\lambda)$, the joining utility function when joining customers join the priority/regular queue with certain probability, and last $u_3(\lambda)$, the joining utility function when no joining customers buy priority. Denote the total customer joining rate at the intersection point as λ^e.

It is common in the strategic queueing literature that the equilibrium total customer joining rate can happen at λ^e where only some customers join and they are indifferent between joining and balking, or at the boundary level Λ where all customers join the system and joining is still a preferred strategy. That is, the equilibrium is the minimum between the indifferent joining rate λ^e and the potential demand rate Λ. However, in our case, simply doing so may miss some equilibrium. We investigate the detailed equilibria according to the following cases.

Case (i): the equilibrium utility function $u_1(\lambda)$ intersects the x-axis and $\lambda^e \geq \bar{\lambda}$.

Figure 5(a) depicts this case. It can be shown that the existence of this case requires the priority price ξ satisfies $\xi \leq \xi_0$, where ξ_0 is the unique solution of

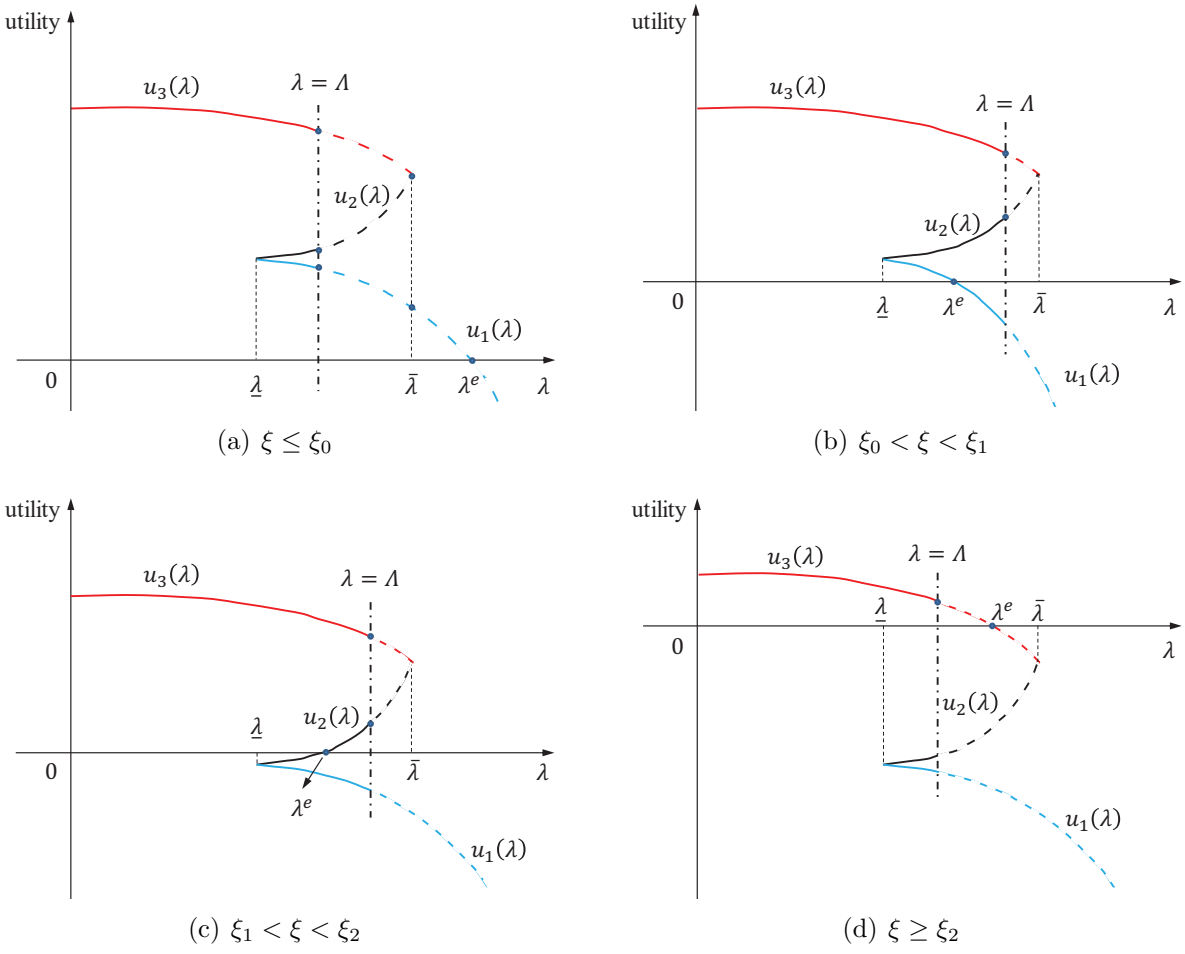
$$u_1(\bar{\lambda})|_{\xi=\xi_0} = 0.$$

As the equilibrium utility function $u_1(\lambda)$ intersects the x-axis, the intersection point $\lambda = \lambda^e$ shall solve the equation

$$R - (p+\xi) - \frac{\theta\phi(\lambda)}{n\mu - \lambda} - \frac{\theta}{\mu} = 0.$$

We then have the equilibrium outcomes regarding the customers' equilibrium queueing behavior, which hinges on the magnitude of the potential arrival rate Λ.

(a) If $\Lambda \geq \lambda^e$, the equilibrium total joining rate is λ^e and the equilibrium for the whole game is (λ^e, λ^e). In this equilibrium, every incoming customer joins the system with probability λ^e/Λ and all the joining customers buy the priority. In this situation, a total customer joining rate less than λ^e will result in a positive joining utility for



(a) $\xi \leq \xi_0$

(b) $\xi_0 < \xi < \xi_1$

(c) $\xi_1 < \xi < \xi_2$

(d) $\xi \geq \xi_2$

Figure 5: The Impact of Priority Pricing on Joining Customers' Joining Utility.

customers, motivating more customers to join the system until the joining utility is reduced to zero, while a total customer joining rate greater than λ^e will result in a negative joining utility, inducing more customers to balk until the number of joining customers is reduced to λ^e. Clearly, the equilibrium (λ^e, λ^e) is an ESS: first, the equilibrium that all the joining customers choose priority queue, i.e, $\lambda_h^s(\lambda^e) = \lambda^e$ is an ESS as discussed in §3.1; second, if other customers adopt a joining probability larger (respectively, smaller) than λ^e/Λ, their utility is negative (respectively, positive) and thus it is better off for a tagged customer to join with probability λ^e/Λ.

(b) If $\bar{\lambda} \leq \Lambda < \lambda^e$, the equilibrium total joining rate reaches the upper bound of the potential demand Λ and the equilibrium for the whole game is (Λ, Λ). In this equilibrium, all the customers join the priority queue and obtain a positive utility. This equilibrium strategy can be shown to be an ESS.

(c) If $\underline{\lambda} \leq \Lambda < \bar{\lambda}$, as shown in Figure 5(a), the vertical line $\lambda = \Lambda$ intersects the utility curves $u_1(\lambda)$, $u_2(\lambda)$, and $u_3(\lambda)$, respectively, and the utility at each of the three intersection points (note that the number of intersection points reduces to two when $\Lambda = \underline{\lambda}$) is positive. Since Λ is the maximum total demand rate, the corresponding joining strategies at these three intersection points, $(\Lambda, 0)$, $(\Lambda, \lambda_s^e(\Lambda))$, and (Λ, Λ), are all equilibria. Among the three equilibria, the equilibria $(\Lambda, 0)$ and (Λ, Λ) are the ESS according to the discussion in Case i(a)(see Figure 5(a)), while $(\Lambda, \lambda_s^e(\Lambda))$ is not an ESS because the mixed strategy, $\lambda_h^e(\Lambda) = \lambda_s^e(\Lambda))$ is not an ESS. Among the three equilibria, the equilibrium $(\Lambda, 0)$ yields the highest utility for customers; that is, it is a *Pareto-dominant* equilibrium.

(d) If $0 < \Lambda < \underline{\lambda}$, the equilibrium total joining rate is Λ and the two-dimensional equilibrium is $(\Lambda, 0)$; that is, in equilibrium, all customers join the regular queue. Again, this equilibrium is an ESS.

Case (ii): the equilibrium utility function $u_1(\lambda)$ intersects the x-axis and $\lambda^e < \bar{\lambda}$.

Figure 5(b) illustrates this case. It can be shown that the existence of this case requires the priority price ξ satisfies $\xi_0 < \xi \leq \xi_1$, where ξ_1 is the unique solution of

$$u_1(\underline{\lambda})|_{\xi=\xi_1} = 0.$$

Again, in this case, the intersection point $\lambda = \lambda^e$ shall solve the equation

$$R - (p + \xi) - \frac{\theta\phi(\lambda)}{n\mu - \lambda} - \frac{\theta}{\mu} = 0.$$

Note that in this case we have $\underline{\lambda} \leq \lambda^e < \bar{\lambda}$. Then, depending on the magnitude of the potential arrival rate Λ, we have the following equilibrium outcomes:

(a) If $\Lambda \geq \bar{\lambda}$, the equilibrium total joining rate is λ^e and the two-dimensional equilibrium is (λ^e, λ^e), in which every incoming customer joins the system with probability λ^e/Λ and all the joining customers buy priority. This equilibrium is an ESS.

(b) If $\lambda^e \leq \Lambda < \bar{\lambda}$, as shown in Figure 5(b), (λ^e, λ^e) again is the two-dimensional equilibrium. The vertical line $\lambda = \Lambda$ intersects the utility curves, and the joining utilities at the two intersection points, $u_1(\Lambda)$ and $u_2(\Lambda)$ are both positive. Since Λ is the maximum demand rate, the corresponding joining strategies at these two points, $(\Lambda, 0)$ and $(\Lambda, \lambda_h^s(\Lambda))$, are equilibria. Similar to that in Case(i), the customer queueing equilibrium $(\Lambda, 0)$ is an ESS while the other equilibrium $(\Lambda, \lambda_h^s(\Lambda))$ is not. Among the three equilibria, the equilibrium $(\Lambda, 0)$ yields the highest utility for customers and is the Pareto-dominant equilibrium.

(c) If $\underline{\lambda} \leq \Lambda < \lambda^e$, the vertical line $\lambda = \Lambda$ intersects all the three utility curves $u_1(\lambda)$, $u_2(\lambda)$, and $u_3(\lambda)$, and the joining utility at each intersection point is positive. Since Λ is the maximum potential demand, the corresponding joining strategies at all three intersection points, $(\Lambda, 0)$, $(\Lambda, \lambda_h^s(\Lambda))$, and (Λ, Λ) are equilibria. Similarly, the equilibrium $(\Lambda, \lambda_h^s(\Lambda))$ is not an ESS and the other two equilibria are the ESS. Among the three equilibria, $(\Lambda, 0)$ is the Pareto-dominant one.

(d) If $0 < \Lambda < \underline{\lambda}$, the equilibrium total joining rate is Λ and the two-dimensional equilibrium is $(\Lambda, 0)$. That is, in equilibrium, all customers join the regular queue. This equilibrium strategy is an ESS.

Case (iii): the equilibrium utility function $u_2(\lambda)$ intersects the x-axis.

Figure 5(c) depicts this case. It can be easily shown that the existence of this case requires the priority price ξ satisfies $\xi_1 < \xi < \xi_2$, where ξ_2 is the unique solution of

$$u_3(\bar{\lambda})|_{\xi=\xi_2} = 0.$$

As the utility curve $u_2(\lambda)$ intersects the x-axis at λ^e, by solving $u_2(\lambda) = 0$, we get

$$\lambda^e = \frac{\xi n \mu^2}{(R-p)\mu - \theta}.$$

Depending on the magnitude of the potential arrival rate Λ, we then have the following results regarding the equilibrium outcome.

(a) If $\Lambda \geq \bar{\lambda}$, the equilibrium total joining rate is λ^e and the two-dimensional equilibrium is $(\lambda^e, \lambda_h^s(\lambda^e))$. In this equilibrium, every incoming customer joins the system with probability λ^e/Λ and the joining customers buy priority with probability $\lambda_h^s(\lambda^e)/\lambda^e$. This strategy is not an ESS because the mixed strategy $\lambda_h^e(\lambda^e) = \lambda_h^s(\lambda^e)$ is not an ESS.

(b) If $\lambda^e \leq \Lambda < \bar{\lambda}$, as shown in Figure 5(c), $(\lambda^e, \lambda_h^s(\lambda^e))$ is the two-dimensional equilibrium. Besides, a close look at the intersection of the vertical line $\lambda = \Lambda$ with the grand utility curve shows that the joining utilities at the two intersection points, $u_1(\Lambda)$ and $u_2(\Lambda)$ are both positive. Since Λ is the maximum demand rate, the corresponding joining strategies at these two points, $(\Lambda, 0)$ and $(\Lambda, \lambda_h^s(\Lambda))$, are equilibria. Among the three equilibria, $(\lambda^e, \lambda_h^s(\lambda^e))$ and $(\Lambda, \lambda_h^s(\Lambda))$ are not the ESS, and the quilibrium $(\Lambda, 0)$ is an ESS and also Pareto-dominating.

(c) If $\Lambda < \lambda^e$, a close look at the intersection of the vertical line $\lambda = \Lambda$ with the grand utility curve shows that only the joining utility at the intersection point $u_1(\Lambda)$ is positive. Thus, the corresponding customer queueing strategy, $(\Lambda, 0)$, is the unique equilibrium.

Case (iv): the equilibrium utility function $u_3(\lambda)$ intersects the x-axis.

Figure 5(d) depicts this case. It can be shown that the existence of this case requires the priority price $\xi \geq \xi_2$. As the utility curve $u_3(\lambda)$ intersects the x-axis, the intersection point $\lambda = \lambda^e$ shall solve the following equation:

$$R - p - \frac{\theta\phi(\lambda)}{n\mu - \lambda} - \frac{\theta}{\mu} = 0.$$

We then have the following customer queueing equilibrium outcomes, which hinges on the magnitude of the potential arrival rate Λ.

(a) If $\Lambda \geq \lambda^e$, the equilibrium total joining rate is λ^e and the two-dimensional equilibrium is $(\lambda^e, 0)$. That is, in equilibrium, every incoming customer joins the system with probability λ^e/Λ and all the joining customers choose the regular queue. This equilibrium is an ESS.

(b) If $\Lambda < \lambda^e$, as shown in Figure 5(d), among all the intersection points of the vertical line $\lambda = \Lambda$ with the grand utility curve, only the one at $u_1(\Lambda)$ has a positive utility. The corresponding customer queueing strategy, $(\Lambda, 0)$ is the unique equilibrium and an ESS.

Note that the assumption $R - \theta/\mu - p > 0$ implies that $u_3(0) > 0$ and guarantees the existence of the intersection point of the grand utility curve with the x-axis (see Figure 5). A close look at the equilibrium outcomes under the above four exclusive and exhaustive cases reveals that when the potential demand is sufficiently large, the customer queueing equilibrium is unique. However, when the potential demand is low, there may exist multiple customer queueing equilibria. The following proposition summarizes the customers' equilibrium queueing strategies.

Proposition 2 *In a non-preemptive priority queue with potential demand Λ, given the priority price ξ, the customers' equilibrium queueing strategies are summarized in Table 1.*

Table 1: Customer Queueing Equilibrium Given the Priority Price ξ

ξ \ Λ	$\Lambda \in [0, \underline{\lambda}]$	$\Lambda \in (\underline{\lambda}, \frac{\xi n\mu^2}{(R-p)\mu-\theta})$	$\Lambda \in [\frac{\xi n\mu^2}{(R-p)\mu-\theta}, \bar{\lambda})$	$\Lambda \in [\bar{\lambda}, +\infty)$
$\xi \in (0, \xi_1]$	$(\Lambda, 0)$	$(\Lambda, 0)$ $(\Lambda, \lambda_h^s(\Lambda))^*$ $(\min\{\lambda^e, \Lambda\}, \min\{\lambda^e, \Lambda\})$		$(\min\{\lambda^e, \Lambda\}, \min\{\lambda^e, \Lambda\})$
$\xi \in (\xi_1, \xi_2]$	$(\Lambda, 0)$	$(\Lambda, 0)$	$(\Lambda, 0)$ $(\Lambda, \lambda_h^s(\Lambda))^*$ $(\lambda^e, \lambda_h^s(\lambda^e))^*$	$(\lambda^e, \lambda_h^s(\lambda^e))^*$
$\xi \in [\xi_2, +\infty)$	$(\min\{\lambda^e, \Lambda\}, 0)$	$(\min\{\lambda^e, \Lambda\}, 0)$		$(\lambda^e, 0)$

Note1: The equilibria marked with * are not the ESS.

Proposition 2 shows that the customer's equilibrium joining behavior highly hinges upon the magnitude of the potential demand Λ and the priority price ξ. Furthermore, there exists an unstable equilibrium when the priority price falls into an intermediate range (i.e., $\xi \in (\xi_1, \xi_2]$) and the potential arrival rate is sufficiently large (i.e., $\Lambda \geq \frac{\xi n\mu^2}{(R-p)\mu-\theta}$). A close look at Figure 5 indicates that the customers' joining utility under the queueing strategy $(\Lambda, 0)$ is the highest if there exist multiple equilibria, and thus this joining strategy Pareto dominates the others.

Acknowledgements

The first author Qingying Li was supported in part by the National Natural Science Foundation of China under the grant no. 71871052, 71832001, the Fundamental Research Funds for the Central Universities, and DHU Distinguished Young Professor Program. The second author Pengfei Guo was supported in part by the Research Grants Council of Hong Kong under grant no. 15506417. The third author Yulan Wang acknowledges the financial supports from the Research Grants Council of Hong Kong (RGC Reference Number: 15505318).

References

[1] Adiri, I., Yechiali, U. (1974). Optimal priority-purchasing and pricing decisions in nonmonopoly and monopoly queues. Operations Research, 22(5), 1051–1066.

[2] Afèche, P., Mendelson, H. (2004). Pricing and priority auctions in queueing systems with a generalized delay cost structure. Management Science, 50(7), 869-882.

[3] Cao, P., Wang, Y., Xie, J. (2019). Priority Service Pricing with Heterogeneous Customers: Impact of Delay Cost Distribution. Production and Operations Management, forthcoming.

[4] Erlichman, J., Hassin, R. (2015). Strategic overtaking in a monopolistic $M/M/1$ queue. IEEE Transactions on Automatic Control, 60(8), 2189-2194.

[5] Gavirneni, S., Kulkarni, V. G. (2016). Self-Selecting Priority Queues with Burr Distributed Waiting Costs. Production and Operations Management, 25(6), 979-992.

[6] Hassin, R. (1995). Decentralized regulation of a queue. Management Science, 41(1), 163-173.

[7] Hassin, R. (2016). Rational queueing. CRC Press, Taylor & Francis Group.

[8] Hassin, R., Haviv, M. (1997). Equilibrium threshold strategies: The case of queues with priorities. Operations Research, 45(6), 966-973.

[9] Hassin, R., Haviv, M. (2003). To queue or not to queue: Equilibrium behavior in queueing systems (Vol. 59). Springer Science & Business Media.

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

[11] Kella, O., Yechiali, U. (1985). Waiting times in the non-preemptive priority $M/M/c$ queue. Stochastic models, 1(2), 257-262.

[12] Lee, H.L., M.A. Cohen. 1983. A note on the convexity of performance Measures of $M/M/c$ queueing systems. Journal of Applied Probability, 20(4), 920–923.

[13] Wang, J., Cui, S., Wang, Z. (2019). Equilibrium Strategies in $M/M/1$ Priority Queues with Balking. Production and Operations Management, 28, 43-62.

Appendix: Proofs

Proof of Lemma 1: From (4), we can show that

$$\frac{dU_h(\lambda,\lambda_h)}{d\lambda} = -\frac{\theta\phi'(\lambda)}{n\mu-\lambda_h} \quad \text{and} \quad \frac{dU_h(\lambda,\lambda_h)}{d\lambda_h} = -\frac{\theta\phi(\lambda)}{(n\mu-\lambda_h)^2}.$$

Based on (5), we then have

$$\frac{dU_r(\lambda,\lambda_h)}{d\lambda} = -\frac{\theta n\mu}{n\mu-\lambda_h}\cdot\frac{\phi'(\lambda)(n\mu-\lambda)+\phi(\lambda)}{(n\mu-\lambda)^2} \quad \text{and} \quad \frac{dU_r(\lambda,\lambda_h)}{d\lambda_h} = -\frac{\theta n\mu\phi(\lambda)}{n\mu-\lambda}\cdot\frac{1}{(n\mu-\lambda_h)^2}.$$

As shown in [12], the Erlang-delay formula $\phi(\lambda)$ is increasing and convex in λ; that is, $\phi'(\lambda) > 0$ and $\phi''(\lambda) > 0$. This together with the above derivatives implies the first result.

By using $\phi(\lambda) > 0$ and $\phi'(\lambda) > 0$, from (6), we can show that

$$\begin{aligned}\frac{d}{d\lambda}\Big(U_h(\lambda,\lambda_h)-U_r(\lambda,\lambda_h)\Big) =& \frac{\theta}{n\mu-\lambda_h}\cdot\frac{(\lambda\phi'(\lambda)+\phi(\lambda)](n\mu-\lambda)+\lambda\phi(\lambda)}{(n\mu-\lambda)^2}\\ =& \frac{\theta}{n\mu-\lambda_h}\cdot\frac{\lambda\phi'(\lambda)(n\mu-\lambda)+n\mu\phi(\lambda)}{(n\mu-\lambda)^2} > 0,\\ \frac{d}{d\lambda_h}\Big(U_h(\lambda,\lambda_h)-U_r(\lambda,\lambda_h)\Big) =& \frac{\theta\lambda\phi(\lambda)}{n\mu-\lambda}\cdot\frac{1}{(n\mu-\lambda_h)^2} > 0.\end{aligned}$$

Proof of Lemma 2: From the definitions of $\bar{\lambda}$, we have

$$\frac{\theta\lambda\phi(\lambda)}{n\mu(n\mu-\lambda)}\big|_{\lambda=\bar{\lambda}} = \xi,$$

where the left-hand side term increases in λ. Thus, if ξ increases, $\bar{\lambda}$ increases. From the definition of $\underline{\lambda}$, we have

$$\frac{\theta\lambda\phi(\lambda)}{(n\mu-\lambda)^2}\big|_{\lambda=\underline{\lambda}} = \xi,$$

where

$$\frac{d}{d\lambda}\Big(\frac{\theta\lambda\phi(\lambda)}{(n\mu-\lambda)^2}\Big) = \frac{\theta}{(n\mu-\lambda)^4}\cdot\Big[(\lambda\phi'(\lambda)+\phi(\lambda))(n\mu-\lambda)^2+2(n\mu-\lambda)\lambda\phi(\lambda)\Big] > 0.$$

Thus, $\underline{\lambda}$ also increases in ξ.

We consider check the utility functions $u_1(\lambda)$ and $u_2(\lambda)$. By putting back the priority price parameter ξ, we then have

$$\begin{aligned}\frac{\partial u_1(\lambda,\xi)}{\partial\lambda} &= -\theta\frac{\phi'(\lambda)(n\mu-\lambda)+\phi(\lambda)}{(n\mu-\lambda)^2} < 0;\\ \frac{\partial u_2(\lambda,\xi)}{\partial\lambda} &= -\theta\frac{\phi'(\lambda)(n\mu-\lambda)+\phi(\lambda)}{(n\mu-\lambda)^2} < 0.\end{aligned}$$

Note that $\bar{\lambda}$ and $\underline{\lambda}$ are functions of ξ. We can show that

$$\frac{du_1(\bar{\lambda}(\xi),\xi)}{d\xi} = \frac{\partial u_1(\lambda,\xi)}{\partial\lambda}\Big|_{\lambda=\bar{\lambda}}\cdot\frac{d\bar{\lambda}}{d\xi}+\frac{\partial u_1(\lambda,\xi)}{\partial\xi} < 0,$$

where the last inequality holds because $\frac{\partial u_1(\lambda,\xi)}{\partial\lambda} < 0$, $\frac{d\bar{\lambda}}{d\xi} > 0$, and $\frac{\partial u_1(\lambda,\xi)}{\partial\xi} = -1 < 0$. We also have

$$\frac{du_2(\bar{\lambda}(\xi),\xi)}{d\xi} = \frac{\partial u_2(\lambda,\xi)}{\partial\lambda}\Big|_{\lambda=\underline{\lambda}}\cdot\frac{d\underline{\lambda}}{d\xi}+\frac{\partial u_2(\lambda,\xi)}{\partial\xi}\Big|_{\lambda=\underline{\lambda}} < 0,$$

where the last inequality holds because $\frac{\partial u_2(\lambda,\xi)}{\partial\lambda} < 0$, $\frac{d\underline{\lambda}}{d\xi} > 0$, and $\frac{\partial u_2(\lambda,\xi)}{\partial\xi} = -1-\frac{n\mu-\underline{\lambda}}{\underline{\lambda}} < 0$.