# Priority Service Pricing with Heterogeneous Customers: Impact of Delay Cost Distribution

Ping Cao
School of Management, University of Science and Technology of China, Hefei, China 230026, pcao@ustc.edu.cn

Yaolei Wang
School of Management, University of Science and Technology of China, Hefei, China 230026, wangyaol@mail.ustc.edu.cn

Jingui Xie
School of Management, University of Science and Technology of China, Hefei, China 230026, xiej@ustc.edu.cn

This paper studies the priority pricing problem for a single-server queueing system with two priority classes in which customers have different sensitivities to delay. The system makes a fixed-delay announcement to inform arriving customers of the expected delay for each class, whereupon each customer must decide which class to join. Any customer who joins the priority class is charged a fixed priority price. Our examination of customers' joining behavior under any given priority prices reveals that there can be multiple equilibrium delays and that the number of those delays depends on the structure of customers' delay cost distribution. We characterize the stability of these equilibria and show that the system can reach the largest or smallest equilibrium by making a proper initial delay announcement. In addition, we consider two pricing problems to maximize the system's long-run average revenue and social welfare respectively. The results derived here establish that both the revenue-maximizing price and the social welfare–maximizing price are quite sensitive to the delay cost distribution. Finally, we investigate the influence of the number of priority classes by extending the two–priority-class model to a multiple–priority-class model.

*Key words*: priority service; pricing; heterogeneous delay cost rate; delay announcement; equilibrium delay
*History*:

## 1. Introduction

Delay is a frequent phenomenon in both service and manufacturing systems. The *delay cost rate*—defined as the delay cost per unit time—is a commonly used economic measure of delay (Reinertsen 2009). Customers in the real world are heterogeneous in their delay sensitivity, and a delay-sensitive customer has a high delay cost rate (Akşin et al. 2013). In a typical service system, such customers may be willing to pay extra in order to reduce their waiting time. Hence managers adopt "priority" service pricing as a valuable revenue management tool used to differentiate customers and improve profitability (Afèche 2013). The four examples listed next are taken from the service and manufacturing industries.

• SeaWorld, the water-based amusement park, offers a "Quick Queue" program that gives patrons (who are willing to pay some $30) front-of-the-line admission to its most popular rides (Gavirneni and Kulkarni 2016).

• When applying for a visa, the applicant can obtain priority service by paying an extra fee; doing so puts the applicant's paperwork ahead of all non-priority applications.

• Benjamin Franklin Plumbing, a nationally branded franchising organization in the United States, offers a "Ben Franklin Society" program that allows its members to move to the top of the appointment book if they want to get their plumbing service quickly.

• The Timken Company, a leading manufacturer of tapered roller bearings, offers a special program called "Bearing Express"; the customers enrolled in that program can pay an additional price to reduce the standard lead time by half. Timken implements this program by assigning different levels of priority to orders (Gilland and Warsing 2009).

The characterization of customers' delay cost is crucial to setting an effective pricing strategy. Gavirneni and Kulkarni (2016) point out that the heterogeneity of delay cost depends on the specific service offered. For amusement parks, customers' delay costs depend on individual income and wealth. In the case of consular services, the cost could reflect both ability to pay and the imminence of travel plans. For the manufacturing industry, delay costs most often depend on the urgency of the orders. In short, there is a wide distribution of customers' delay costs across different service types and their associated underlying environments. Yet the extant literature typically assumes that customers' delay costs are distributed in a specific way (see e.g. Naor 1969; Gilland and Warsing (2009); Afèche 2013). That assumption gives rise to the following fundamental questions.

1. How important is it to *generalize* the distribution assumption of customers' delay cost?
2. How does the customers' delay cost distribution affect the service system's pricing strategy and revenue?

We shall address these questions by considering a general distribution for customers' delay costs. Following the line of Naor (1969), we assume that a customer's delay cost is a linear function of the waiting time.[1] In order to assess the effect of a general delay cost distribution on customers' "joining" behavior and on the pricing strategy of system managers, we start by considering a stylized M/G/1 queueing system with two priority classes. In a typical setting, arriving customers decide which queue to join by comparing utilities of each option; that process requires such system-related information as the expected waiting time of each class and the current number of waiting customers. The system often provides this information by way of a delay announcement. We shall consider a specific delay announcement mechanism, the *fixed delay* (FD) announcement, whereby

[1] Other authors (e.g., Akan et al. 2012) have studied convex or convex–concave delay costs. However, we do not consider a nonlinear cost structure in this paper.

each arriving customer is informed of long-run average waiting time for each class. Next we analyze customers' joining equilibrium—in terms of its existence, uniqueness, and stability—when the priority price remains fixed. The results of that analysis are then used to characterize the revenue-maximizing price and the social welfare–maximizing price as well as to obtain closed-form versions of those optimal prices for some delay cost distributions. Finally, we extend our results based on two priority classes to multiple classes and investigate the implications of offering a small number of priority classes.

### 1.1. Main Results and Contributions

Our paper makes several contributions to the literature. First, we study customers' behavioral equilibrium under delay announcement in the context of a queueing game. Most research on queueing games has assumed that customers can obtain the waiting time's exact value, which they use to make joining decision; thus customers' behavioral equilibrium is presumed to be uniquely determined under a fixed decision variable (here, the priority price). In service systems, however, it is difficult for customers to know the exact waiting time. A more realistic setting is one in which the system will provide a delay announcement that merely guides customer decisions. It is therefore worthwhile to study how the equilibrium will be formed under a given delay announcement scheme. Our paper explores customers' behavioral equilibrium under the FD announcement in terms of two aspects: the equilibrium's existence and uniqueness; and the evolution of customer response to the system's delay announcement. We find that the equilibrium delay may not be unique and that the number of equilibria depends on the delay cost distribution structure. For uniform and exponential distributions, uniqueness is guaranteed. Yet multiple equilibria may exist for unimodal distributions, in which case the queueing game literature's typical assumption (i.e., exact knowledge of waiting times) is invalid. It follows that the distribution structure of customers' delay costs plays a critical role in customers' behavioral equilibrium and hence also in the system's pricing strategies. This connection demonstrates the importance of being able to generalize any assumption about the distribution of customers' delay costs. If there are multiple equilibria then we are naturally interested in identifying the most desirable and reachable one. Thus we are motivated to study these equilibria by analyzing the evolution of delay announcements under the iterative scheme proposed by Armony et al. (2009). More specifically, we derive an explicit characterization of the stability of these equilibria and demonstrate that the largest or smallest equilibrium delay can be reached if the initial delay announcement is appropriate.

Second, we analyze the priority prices that maximize the system's long-run average revenue and the ones that maximize social welfare. We focus in particular on examining the effect of delay cost distributions on these optimal prices. To build a consistent framework, we first analyze

those two optimization problems and characterize the structural properties of related problems. We then investigate these pricing strategies under some specific distribution families—namely, the uniform, exponential, Weibull, power, and log-normal distributions. Our results indicate that: (i) the optimal prices are sensitive to the specific delay cost distribution family considered; (ii) if the delay cost distribution is mis-specified then the relative revenue loss can be significant, especially when the system is congested; and (iii) some managerial insights reported in the literature are likewise sensitive to the underlying distributional assumptions. For example, Gilland and Warsing (2009) report that the social welfare–maximizing price is equal to the revenue-maximizing price. In contrast, our study reveals that this claim holds only under certain distributions, such as the uniform distribution (for some parameters) and the exponential distribution.

Finally, we investigate the influence of the number of priority classes by extending our model to multiple priority classes. Thus we employ a transformation of variables to establish the existence of a delay equilibrium and then analyze the corresponding two optimization problems. Our numerical studies show that the service system's revenue and social welfare both increase with the number of priority classes and that the most benefit is gained by offering a relatively small number of such classes.

### 1.2. Organization of the Paper

The rest of our paper proceeds as follows. Section 2 gives a detailed review of the related literature, and the two–priority-class queueing model is introduced in section 3. In section 4, we discuss customers' equilibrium behavior under any given priority price. Sections 5 and 6 describe the service system's optimal pricing strategies from the perspectives of maximizing social welfare and of maximizing its own long-run average revenue. Then, in section 7, we discuss these pricing strategies under four distribution families and conduct several numerical studies. Section 8 is devoted to analysis of the service system when multiple priority classes are offered. We conclude in section 9 with a summary of our findings and some suggestions for future research.

## 2. Literature Review

Our work is distinct from the literature on queueing games in two respects. First, instead of assuming a particular delay cost distribution, we consider a general distribution of customers' delay cost in a service system priority pricing problem. Second, we evaluate the effect of the system's delay announcement on customer's joining behavior upon arrival. We examine the properties of customers' behavioral equilibrium—especially its evolution and stability—as well as its effect on decision making, which has not received much attention in queueing game literature.

*Distributional assumptions on customers' delay cost.* In the literature on queueing games, customers' delay cost distribution is assumed to be either discrete or continuous. In the first category of research, a common assumption is that customers are *homogeneous* with respect to delay sensitivity—in other words, that their delay cost rates are identical (see e.g. Naor 1969; Mendelson 1985; Hassin 1986; Chen and Frank 2004; Debo et al. 2008; Hassin and Roet-Green 2017). Other studies view customers as being *heterogeneous* as regards delay cost. Plambeck (2004) and Afèche (2013) assume that there are two customer types, each with a different but constant delay cost rate. Hassin (1995), Ha (2001) and Ata and Peng (2018) consider multiple types of customers by partitioning them into a finite number of sets, where each set consists of customers who have the same (and constant) delay cost rate.

Unlike the works just cited, which study only the discrete types of customers' delay cost, some of the literature assumes that customers' delay cost rates follow a special *continuous* distribution. For instance, Zhang et al. (2007), Gilland and Warsing (2009), and Yu et al. (2013) presuppose a uniform distribution of those rates. Gavirneni and Kulkarni (2016) study an M/G/1 variation of the two-class non-preemptive priority model in which the customers' delay cost rate is modeled as a Burr distribution. For a comprehensive survey of the various heterogeneity assumptions made about customers' delay cost rates, see Hassin (2016). Our paper also considers a continuous distribution of customers' delay costs. Yet in contrast with the scholars just cited, we assume that the customers' delay cost follows a general continuous distribution rather than a specific one. Our reason for adopting this approach is the empirical evidence presented by Lu et al. (2013), who use supermarket data and find that customers' delay sensitivity is random and follows a continuous distribution; see also Akşin et al. (2013), who assume that customers' unit waiting costs follow a log-normal distribution whose parameters they estimate based on data from a call center. Our examining of the various distributions reveals that not all managerial implications described in literature hold true when the underlying distribution is mis-specified—a finding that underscores the importance of distributional structure in decision making. Thus robust managerial insights depend on the care taken to adopt reasonable assumptions about that structure.

*Priority service pricing.* The priority service pricing problem is a classical topic in queueing games; representative works include Mendelson and Whang (1990), Hassin (1995), and Kittsteiner and Moldovanu (2005) (to name just a few). Hassin and Haviv (2003) offer an excellent survey of the literature on this topic. Our paper segments a continuum of customers into a finite set of priority levels, a strategy that is closely related to that employed by Gilland and Warsing (2009), Gavirneni and Kulkarni (2016), and Nazerzadeh and Randhawa (2018).

Gilland and Warsing (2009) consider a priority pricing problem in an M/M/1 queueing system with two or more priority classes; assuming that customers' delay cost rates are drawn from a

uniform distribution and that all customers must be served. They conclude that the revenue-maximizing price is equal to the social welfare–maximizing price—a conclusion that may fail to hold more generally (i.e., under other delay cost distributions). Recall that Gavirneni and Kulkarni (2016) study an M/G/1 queueing system in which there are two priority classes and customers' delay costs follow a Burr distribution. Our paper considers a similar priority pricing problem but assumes that those costs follow a general distribution.

Nazerzadeh and Randhawa (2018) also consider a general delay cost distribution as they study the problem of offering a menu of price and delay options that maximizes an M/M/1 queueing system's revenue. One difference between that paper and ours is that, to realize the offered expected delays, customers will not be served under strict priority policy. These authors show that offering two service priorities is asymptotically optimal on the square-root scale.

All of the research just cited implicitly assumes that the true expected delay is known to arriving customers—from which it follows that customers' joining equilibrium is uniquely determined under a fixed priority price. In practice, however, the actual (expected) delay is often unknown to arriving customers and so the system typically makes a delay announcement to guide customers in their joining decision. Our paper investigates customers' joining equilibria under a prespecified delay announcement scheme, a topic that is understudied in previous research on priority service pricing.

*Delay announcement.* Service systems often provide waiting time information to arriving customers via delay announcements, which can effectively manage customers' behavior. Such announcements have been extensively studied in the queueing literature (see e.g. Whitt 1999; Armony and Maglaras 2004a, 2004b; Allon and Bassamboo 2011). Guo and Zipkin (2007) consider a queue with balking under three levels of delay information, and they identify conditions under which the system's performance is improved by providing more accurate delay information. Armony et al. (2009) propose two delay announcement schemes for an unobservable G/GI/$s$+GI queueing system: FD announcement and DLS announcement (announcing the delay of the *last* customer to enter service). They study the performance effect of delay announcements made to arriving customers and analyze, in a fluid model framework, the equilibrium behavior under each scheme. These authors find that multiple equilibria can arise if a monotonicity condition is violated. Ibrahim et al. (2017) consider a single-class M/M/N+M queueing system under DLS announcement and find that this announcement scheme may not always be accurate; they also provide conditions under which the DLS prediction is asymptotically accurate. Our study differs from that of Armony et al. in that we consider an M/G/1 queueing system with two priority classes, restrict our attention to FD announcements, and account for the system's pricing decision.

## 3. Model

We consider a queueing system with a single server and two classes. Customers arrive at the system according to a Poisson process with rate $\lambda$, and the service time $B$ follows a general distribution. Upon arrival, each customer decides whether to join class 1 by paying an additional fixed price $K$ after the system announces each class's long-run expected delay. Customers in class 1 are served with strict priority, and the customers within each class are served on a first-come-first-served (FCFS) basis. We assume that the service is non-preemptive: once served, a customer cannot be "preempted out" prior to service completion. We assume also that the system's traffic intensity $\rho := \lambda\mathbb{E}[B]$ is less than 1, a condition that guarantees the queueing system's stability.

Customers are heterogeneous in their evaluation of delay, or their delay cost rate, measured in dollars per unit time. We assume that delay cost rate, denoted by $C$, is a positive random variable with cumulative distribution function (CDF) $H(x)$ and probability density function (PDF) $h(x)$. The focus of this paper is on customers' delay cost distribution in Assumption 1 which is general enough to include many distribution families, such as log-normal distribution.

ASSUMPTION 1. *The PDF of delay cost rate $h(x)$ is continuous and quasi-concave.*[2]

Queue lengths are *not* observable to customers. A customer arriving in the system will receive a delay announcement $(w_1, w_2)$, where $w_1$ and $w_2$ are the average waiting times in (respectively) class 1 and class 2. Given this information, the customer decides which class to join by comparing her expected disutility from joining either class. For a customer whose delay cost rate is $c$, the disutility of joining class 1 is $u_1 = K + cw_1$;[3] that from joining class 2 is $u_2 = cw_2$. Hence the customer will join class 1 if $K + cw_1 < cw_2$ or will join class 2 otherwise. Note that, by Assumption 1, the customer delay cost distribution $H$ is continuous. Hence, the probability that the disutility of joining class 1 equals to that of joining class 2 is 0. Therefore, our analysis remains valid regardless of whether these customers join class 1 or class 2. However, if $H$ is discrete then a mixed equilibrium will arise (Anand et al. 2011).

[2] A function $h(x)$ is *quasi-concave* if it is increasing with $x \leq x^*$ and decreasing with $x \geq x^*$ for some $x^*$.

[3] The only customer heterogeneity we consider is that in their delay cost rates, although our model could capture unobserved heterogeneities or factors if it incorporated a random shock (cf. Ata and Peng 2018). However, doing so would unduly complicate our paper's entire analysis. Because we are interested in the delay cost rate distribution's effect on customers' joining equilibrium and system performance, we forgo all concern about other heterogeneities.

Remark 1 (Customers' Disutility). Our model employs the notion of customers' *disutility*—rather than the more commonly used *utility*—to characterize customers' joining behavior. The reason is that, if utility is used, then customers may balk (i.e., decline to enter the system) if they anticipate a negative surplus from joining one of the classes (a fortiori from joining either class). However, in some systems all customers may need the service; also, customer valuation of the service (e.g., consular processing) may be so high that balking is not a viable option.

In the next section we show that there may be multiple equilibrium delays for a given priority price $K$. One goal of the system manager is to make proper delay announcements—in other words, announcements likely to result in the equilibrium that is most favorable (from the system's perspective). That topic is the subject of section 4. The system manager is also tasked with setting the priority service price. We investigate this issue from the perspective of maximizing system revenue (in section 5) and also from the perspective of maximizing social welfare (in section 6).

## 4. Customers' Joining Equilibrium under a Fixed Priority Price

We now discuss customers' equilibrium behavior under a fixed priority price $K$. Because $K$ is fixed in this section, we omit it to simplify the notation whenever no confusion could result. Suppose the system manager makes a delay announcement $(w_1, w_2)$. We assume that $w_1 < w_2$, for all customers would otherwise choose class 2 upon arrival. A customer joins class 1 *if and only if* her delay cost rate $c$ satisfies the strict inequality $c > K/(w_2 - w_1)$. So from the system manager's perspective, the probability that an arriving customer joins class 1 is given by

$$p(w_1, w_2) := \mathbb{P}\left\{C > \frac{K}{w_2 - w_1}\right\} = 1 - H\left(\frac{K}{w_2 - w_1}\right). \tag{1}$$

Thus the customers joining class 1 and class 2 form two Poisson processes with respective arrival rates $\lambda p(w_1, w_2)$ and $\lambda(1 - p(w_1, w_2))$.

Let $W_i(w_1, w_2)$ be the steady-state delay in class $i$ $(i = 1, 2)$, and let $W(w_1, w_2)$ be the system's steady-state delay following delay announcement $(w_1, w_2)$. Previous research (Gautam 2012) has established that, if the service rate is the same across different classes, then the expectation $\mathbb{E}[W(w_1, w_2)]$ does not depend on the routing policy; we therefore simply write it as $\mathbb{E}[W]$. It follows from standard results for an M/G/1 non-preemptive priority queue that the average delay in each class is given by the following lemma due to Adan and Resing 2002.

Lemma 1. *Given a delay announcement* $(w_1, w_2)$*, the average system delay is*

$$\mathbb{E}[W] = p(w_1, w_2)\mathbb{E}[W_1(w_1, w_2)] + (1 - p(w_1, w_2))\mathbb{E}[W_2(w_1, w_2)] = \frac{\lambda \mathbb{E}[B^2]}{2(1 - \rho)}, \tag{2}$$

*where*

$$\mathbb{E}[W_1(w_1,w_2)] = \frac{\lambda \mathbb{E}[B^2]}{2(1-p(w_1,w_2)\rho)} = \frac{1-\rho}{1-p(w_1,w_2)\rho}\mathbb{E}[W] \quad \text{and} \tag{3}$$

$$\mathbb{E}[W_2(w_1,w_2)] = \frac{\lambda \mathbb{E}[B^2]}{2(1-p(w_1,w_2)\rho)(1-\rho)} = \frac{1}{1-p(w_1,w_2)\rho}\mathbb{E}[W]. \tag{4}$$

The service system's delay announcement will be viewed as trustworthy only if it is consistent with the *actual* average delay (Armony et al. 2009). That criterion leads to the following definition of equilibrium delay.

Definition 1 (Equilibrium Delay). An announced delay $(\tilde{w}_1, \tilde{w}_2) \in \mathbb{R}^2_+$ is an *equilibrium delay* of the system provided that $\mathbb{E}[W_i(\tilde{w}_1, \tilde{w}_2)] = \tilde{w}_i$ for $i = 1, 2$.

If there exists a number $\tilde{w}_2$ such that $\mathbb{E}[W_2(\tilde{w}_1, \tilde{w}_2)] = \tilde{w}_2$, then $\mathbb{E}[W_1(w_1,w_2)] = (1-\rho)\mathbb{E}[W_2(w_1,w_2)] = (1-\rho)\tilde{w}_2$. Hence it follows from Definition 1 that $((1-\rho)\tilde{w}_2, \tilde{w}_2)$ is an equilibrium delay. Conversely, if $(\tilde{w}_1, \tilde{w}_2)$ is an equilibrium delay then $\tilde{w}_1 = \mathbb{E}[W_1(\tilde{w}_1, \tilde{w}_2)] = (1-\rho)\mathbb{E}[W_2(\tilde{w}_1, \tilde{w}_2)] = (1-\rho)\tilde{w}_2$. Therefore, it suffices to focus on the delay in class 2. For any equilibrium delay $(\tilde{w}_1, \tilde{w}_2)$, combining (1) and (4) yields

$$\mathbb{E}[W] = \tilde{w}_2(1-\rho) + \rho H\left(\frac{K}{\rho \tilde{w}_2}\right)\tilde{w}_2. \tag{5}$$

If we define the right-hand side (RHS) of (5) as an auxiliary function,

$$F_e(x) := x(1-\rho) + \rho H\left(\frac{K}{\rho x}\right)x \quad \text{for } x > 0, \tag{6}$$

then finding equilibria is equivalent to finding positive real roots of $F_e(x) = \mathbb{E}[W]$.

### 4.1. Existence of Equilibrium Delay

The first important issue concerns the existence and uniqueness of equilibrium delay. We show in this section that there always exists an equilibrium delay, although it may not be unique. The number of equilibrium delays depends on the structure of customers' delay cost distribution. Most previous research has sought to ensure analytical tractability by assuming that the delay cost rate is uniformly (or exponentially) distributed. Indeed, the equilibrium delay is unique under that assumption. However, empirical studies show that the PDF may be quasi-concave. For example, Akşin et al. (2013) demonstrate that the log-normal distribution, whose density function is quasi-concave, fits actual delay cost data quite well. It is therefore necessary to investigate equilibrium delays under a general delay cost rate distribution. We obtain the following result.

THEOREM 1 **(Existence of Equilibrium)**. *There exists at least one equilibrium delay.*

The proofs of all results are relegated to section A of the Appendix; here we simply offer a sketch of the main idea. In light of the discussion after Definition 1, there is a one-to-one correspondence between equilibrium delays and the positive real roots of $F_e(x) = \mathbb{E}[W]$. Theorem 1 follows by showing that, for any $y > 0$, there always exists an $x > 0$ such that $F_e(x) = y$.[4] Further investigation into the structure of $F_e(x)$ yields our next theorem, as follows.

THEOREM 2 **(Multiple Equilibria)**. *If Assumption 1 holds, then there exist at most three equilibrium delays.*

REMARK 2. The proof of Theorem 2 shows that, if $h(x)$ is decreasing in $x$, then the equilibrium delay is unique. One can use the same argument to show that, if $h(x)$ has a more complicated structure than the one described in Assumption 1, then there could exist *more* than three equilibrium delays.

The following example illustrates that there may, indeed, be more than a single equilibrium delay.

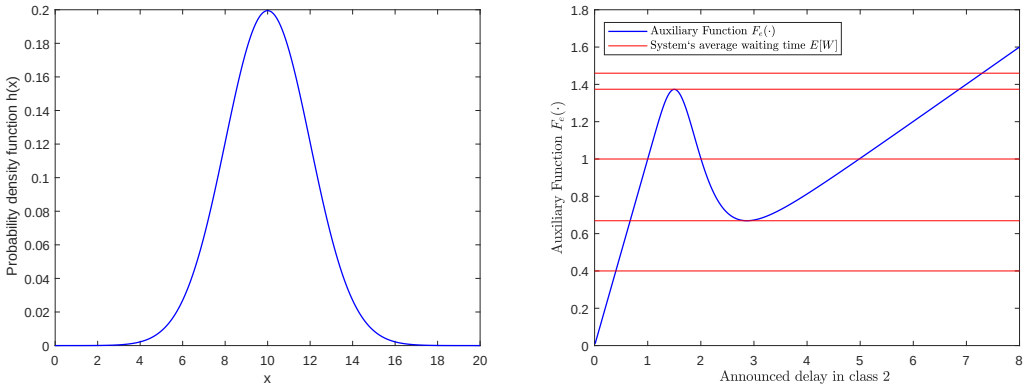
EXAMPLE 1. Suppose that $C := \max\{Z, 0\}$ for $Z \sim N(\mu, \sigma^2)$; that is, let $Z$ be normally distributed with mean $\mu$ and variance $\sigma^2$. The parameters are set as follows: $\mu = 10$, $\sigma = 2$, $\rho = 0.8$, and the priority price is $K = 15$. Figures 2(a) and 2(b) plot (respectively) the PDF of $C$ and the auxiliary function $F_e$ defined in (6). Figure 2(b)shows that the number of roots of $F_e(x) = \mathbb{E}[W]$ varies from 1 to 3 as $\mathbb{E}[W]$ varies from 0 to $\infty$. If we let $L_{\min}$ and $L_{\max}$ denote (respectively) the local minimum and local maximum of $F_e$, then the following more specific statements hold.

(i) If $L_{\min} < \mathbb{E}[W] < L_{\max}$, then there are three equilibrium delays.
(ii) If $\mathbb{E}[W] = L_{\min}$ or $\mathbb{E}[W] = L_{\max}$, then there are two equilibrium delays.
(iii) If $\mathbb{E}[W] > L_{\max}$ or $\mathbb{E}[W] < L_{\min}$, then there is a unique equilibrium delay.

The preceding results and example demonstrate that the system's equilibrium delays might not be unique, a conclusion at odds with previous research. In Zhou et al. (2014), for instance, there is a unique equilibrium delay whereas we show that there may be multiple equilibrium delays.

[4] We remark that the existence of a pure equilibrium is guaranteed under the non-preemptive service rule. If the service is preemptive, then a pure equilibrium delay may not exist and a mixed equilibrium delay will arise.

**Figure 1 (a) Probability Density Function of $C$ in Example 1; (b) Function $F_e(\cdot)$**




(a) (b)

The latter possibility raises two questions. First, which equilibria are stable? Real-world service systems can be perturbed by stochastic fluctuations, so a stable equilibrium is preferable. Second, which equilibrium delay is the most favorable? Different equilibrium delays may result in different system performance levels, and managers are naturally inclined to identify the equilibrium delay that optimizes performance. We shall provide detailed answers to these questions in sections 4.2 and 5, respectively.

Next we adopt some notation that will prove useful in the sequel. Let $W_e(K) := \{(\tilde{w}_{1,i}, \tilde{w}_{2,i}) : 1 \leq i \leq I(K)\}$ be the set of all equilibrium delays under price $K$ such that $\tilde{w}_{2,i}$ is strictly increasing in $i$, where $I(K)$ is the number of equilibrium delays under that price. It then follows from (3) that

$$\tilde{w}_{1,i} = (1-\rho)\tilde{w}_{2,i} \quad \text{for all } \; i = 1, 2, \ldots, I(K). \tag{7}$$

Our next lemma follows immediately from the expressions $F_e(\tilde{w}_2) = \mathbb{E}[W]$ and $x(1-\rho) < F_e(x) \leq x$ for $x > 0$, which give the lower and upper bounds of $\tilde{w}_2$.

LEMMA 2. *For all $i = 1, 2, \ldots, I(K)$, we have that $(1-\rho)\mathbb{E}[W] \leq \tilde{w}_{1,i} < \mathbb{E}[W] \leq \tilde{w}_{2,i} < \mathbb{E}[W]/(1-\rho)$.*

### 4.2. Stability of Equilibrium Delays

We now propose an iterative method for investigating the stability of these equilibria. Following Armony et al. (2009), we assume that the system makes an initial delay announcement $(w_1(0), w_2(0)) \in \mathbb{R}^2_+$, after which the system observes the actual expected steady-state delay $(w_1(1), w_2(1))$ of those customers who have already been served. The system then makes a second delay announcement and observes the actual expected steady-state delay $(w_1(2), w_2(2))$ of those

customers served under delay announcement $(w_1(1), w_2(1))$. This procedure is repeated ad infinitum. If the system makes a delay announcement $(w_1(t), w_2(t))$ at period $t \geq 0$, then the probability that an arriving customer joins class 1 can be written as

$$p(t) := 1 - H\left(\frac{K}{w_2(t) - w_1(t)}\right), \tag{8}$$

which will generate the actual average delay, $(w_1(t+1), w_2(t+1))$, as follows:

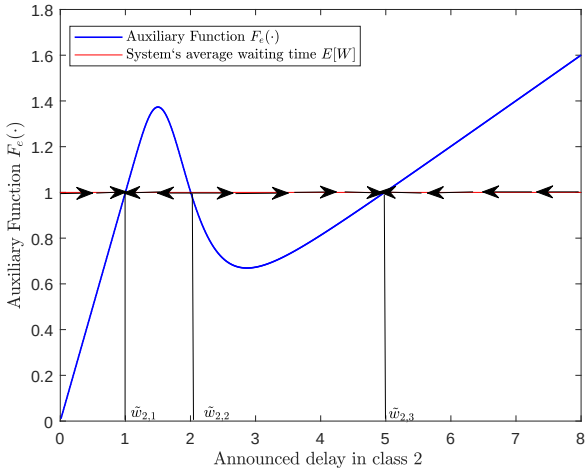
$$w_1(t+1) = \frac{\lambda \mathbb{E}[B^2]}{2(1-p(t)\rho)} = \frac{(1-\rho)\mathbb{E}[W]}{1-p(t)\rho}, \tag{9}$$

$$w_2(t+1) = \frac{\lambda \mathbb{E}[B^2]}{2(1-p(t)\rho)(1-\rho)} = \frac{\mathbb{E}[W]}{1-p(t)\rho}. \tag{10}$$

This delay will be announced at period $t+1$. We mention that this iterative algorithm is a natural way to compute the equilibrium delay, but it does not actually correspond to a natural evolution of the system—that is, unless there is substantial time between successive iteration steps. Otherwise, the system would not be able to reach a steady state before the delay announcement is changed.

Figure 2 illustrates a dynamic process of delay announcements for a case in which there are three equilibrium delays. We can see that: if the announced delay for joining class 2 at period 1 (i.e., $w_2(1)$) is less than $\tilde{w}_{2,1}$, then $w_2(t)$ will increase to $\tilde{w}_{2,1}$ as $t$ increases; if $\tilde{w}_{2,1} < w_2(1) < \tilde{w}_{2,2}$, then $w_2(t)$ will decrease to $\tilde{w}_{2,1}$ as $t$ increases; if $\tilde{w}_{2,2} < w_2(1) < \tilde{w}_{2,3}$, then $w_2(t)$ will increase to $\tilde{w}_{2,3}$ as $t$ increases; and if $w_2(1) > \tilde{w}_{2,3}$, then $w_2(t)$ will decrease to $\tilde{w}_{2,3}$ as $t$ increases.

**Figure 2 Evolution of the Delay Announcement when there are Multiple Equilibrium Delays**




Proposition 1 characterizes the dynamic behavior of delay announcements.

PROPOSITION 1. *For any fixed priority price $K$, the following statements hold.*

(i) *If $w_2(0) - w_1(0) > \rho\tilde{w}_{2,I(K)}$, then $(w_1(t), w_2(t)) \to (\tilde{w}_{1,I(K)}, \tilde{w}_{2,I(K)})$ as $t \to \infty$; in particular, if $w_2(0) - w_1(0) \geq \rho\mathbb{E}[W]/(1-\rho)$ then $(w_1(t), w_2(t)) \to (\tilde{w}_{1,I(K)}, \tilde{w}_{2,I(K)})$ as $t \to \infty$.*

(ii) *If $w_2(0) - w_1(0) \leq \rho\tilde{w}_{2,1}$, then $(w_1(t), w_2(t)) \to (\tilde{w}_{1,1}, \tilde{w}_{2,1})$ as $t \to \infty$; in particular, if $w_2(0) - w_1(0) \leq \rho\mathbb{E}[W]$ then $(w_1(t), w_2(t)) \to (\tilde{w}_{1,1}, \tilde{w}_{2,1})$ as $t \to \infty$.*

(iii) *If $\rho\tilde{w}_{2,i} < w_2(0) - w_1(0) \leq \rho\tilde{w}_{2,i+1}$ for some $1 \leq i \leq I(K) - 1$, then $(w_1(t), w_2(t)) \to (\tilde{w}_{1,i}, \tilde{w}_{2,i})$ as $t \to \infty$ provided that $\rho\mathbb{E}[W] < (w_2(0) - w_1(0))\big[1 - \rho + \rho H\big(K/(w_2(0) - w_1(0))\big)\big]$ and $(w_1(t), w_2(t)) \to (\tilde{w}_{1,i+1}, \tilde{w}_{2,i+1})$ as $t \to \infty$ provided that $\rho\mathbb{E}[W] \geq (w_2(0) - w_1(0))\big[1 - \rho + \rho H\big(K/(w_2(0) - w_1(0))\big)\big]$.*

REMARK 3. The validity of Proposition 1 does not rely on Assumption 1. If the initial announced delay gap $w_2(0) - w_1(0)$ between two classes exceeds $\rho\mathbb{E}[W]/(1-\rho)$, then the announced delay will converge to the largest equilibrium delay. If the initial delay announcement satisfies $w_2(0) - w_1(0) \leq \rho\mathbb{E}[W]$, then the announced delay will converge to the smallest equilibrium delay. The implication is that the system can always reach the largest or the smallest equilibrium delay by choosing the proper initial delay announcement—that is, even if the cost distribution $H$ is unknown. We say that the sequence of delay announcements $\{(w_1(t), w_2(t)); t \in \mathbb{N}\}$ *reaches* an equilibrium delay $(\tilde{w}_1, \tilde{w}_2)$ if $(w_1(t), w_2(t)) \to (\tilde{w}_1, \tilde{w}_2)$ as $t \to \infty$. In that case, we refer to the equilibrium delay $(\tilde{w}_1, \tilde{w}_2)$ as being *reachable*.

REMARK 4. If there exists a unique equilibrium delay—that is, if $I(K) = 1$—then the sequence of announced delays following any initial delay announcement will converge to the unique equilibrium delay, which must be (globally) stable.

Our next theorem, which is based on Proposition 1, characterizes the stability of equilibrium delays.

THEOREM 3 **(Stability of Equilibrium)**. *Consider any equilibrium delay $(\tilde{w}_{1,i}, \tilde{w}_{2,i})$, $i \in \{1, 2, \ldots, I(K)\}$. If $F_e(\cdot)$ is strictly increasing at point $\tilde{w}_{2,i}$, then the equilibrium delay is (locally) stable; otherwise, it is unstable.*

According to this theorem, an equilibrium $(\tilde{w}_{1,i}, \tilde{w}_{2,i})$ is stable if and only if $F_e'(\tilde{w}_{2,i}) > 0$. An unstable equilibrium is undesirable because even a small perturbation in the delay announcement may lead to another equilibrium. We can further illustrate Theorem 3 by referencing the conditions enumerated in Example 1. Here we will, accordingly, again have three cases as follow.

(i) If $L_{\min} < \mathbb{E}[W] < L_{\max}$, then both the largest and the smallest equilibrium delays are stable.

(ii) If $\mathbb{E}[W] = L_{\max}$ then the largest (resp. smallest) equilibrium delay is stable (resp. unstable), and if $\mathbb{E}[W] = L_{\min}$ then the largest (resp. smallest) equilibrium delay is unstable (resp. stable).

(iii) If $\mathbb{E}[W] > L_{\max}$ or $\mathbb{E}[W] < L_{\min}$, then any unique equilibrium delay is stable.

## 5. Maximizing Service System Revenue

In this section we study the optimal priority pricing problem with the objective of maximizing the service system's long-run average revenue. Quantities associated with the price $K$ will now be so marked in order to emphasize that dependence.

If the equilibrium delay is $(\tilde{w}_1, \tilde{w}_2)$, then the average revenue of the service system is

$$\lambda K p(\tilde{w}_1, \tilde{w}_2; K) = \lambda K \left(1 - H\left(\frac{K}{\rho \tilde{w}_2}\right)\right),$$

as follows from (1) and (7). The term on the RHS of this equation is strictly increasing in $\tilde{w}_2$. So given a price $K$, the largest equilibrium delay $(\tilde{w}_{1,I(K)}, \tilde{w}_{2,I(K)})$ is the most desirable for a manager seeking to maximize the system's average revenue. Recall from Remark 3 that the largest equilibrium delay can be reached if the initial delay announcement is properly provided.

Remark 5 (Stability of the Largest Equilibrium Delay). A system manager will certainly want to know whether the largest equilibrium delay (i.e., the most desirable delay) is stable for a given priority price $K$. If the PDF of delay cost $h(x)$ is decreasing in $x$, then the equilibrium delay is unique and thus stable (see Remark 4). Now we consider a quasi-concave PDF of delay cost, in which the function $F_e$ takes a structure pictured in Figure 2(b) in view of the proof of Theorem 2. Let $F_{\min}(K) = F_e(x_{\min}(K); K)$ and $F_{\max}(K) = F_e(x_{\max}(K); K)$, where $x_{\min}(K)$ is the local minimizer of $F_e(\cdot; K)$ and $x_{\max}(K)$ is its local maximizer. Since $F_e(x; K)$ is increasing in $K$ for any fixed $x$, it follows that both $F_{\min}(K)$ and $F_{\max}(K)$ are strictly increasing in $K$. We therefore conclude, based on Figure 2(b) and Theorem 3, that the largest equilibrium delay is unstable if and only if $F_{\min}(K) = \mathbb{E}[W]$. Hence there exists at most one price, denoted by $K_L$, such that the corresponding largest equilibrium delay is unstable.

In this section, we study the optimal priority price that maximizes the service system's average revenue. Let $p^*(K) := p(\tilde{w}_{1,I(K)}, \tilde{w}_{2,I(K)}; K)$ be the largest probability of joining class 1 under price $K$. Our next lemma states that the number of customers who join class 1 will be fewer when the price is higher, which is consistent with intuition.

LEMMA 3. *Under Assumption 1, $p^*(K)$ is both continuous and decreasing in $K$ for $K>0$, and it satisfies*

$$p^*(K) = 1 - H\left(\frac{K(1-p^*(K)\rho)}{\rho\mathbb{E}[W]}\right). \tag{11}$$

*Moreover, $p^*(K)$ is strictly decreasing in $K>0$ when $h$ has support $[0,\infty)$.*

REMARK 6 (FINITE SUPPORT). If $h$ has finite support $[a,b]$ and satisfies Assumption 1, then we can show that: (a) if $K \le (a\rho\mathbb{E}[W])/(1-\rho)$ then $p^*(K)=1$; and (b) if $K \ge b\rho\mathbb{E}[W]$ then $p^*(K)=0$. We can also show that, if $(a\rho\mathbb{E}[W])/(1-\rho) < K < b\rho\mathbb{E}[W]$, then $p^*(K)$ is strictly between 0 and 1 and is strictly decreasing in $K$. Hence the upper and lower bounds of $K$ are defined by, respectively, $(a\rho\mathbb{E}[W])/(1-\rho)$ and $b\rho\mathbb{E}[W]$.

Lemma 3 suggests that there is a trade-off between the price $K$ and the joining probability $p^*(K)$. If so, then there could be a price $K^*$ such that the average revenue $r(K) := \lambda K p^*(K)$ is maximized.

Before discussing the optimal price $K^*$, we digress somewhat by presenting an example in which the delay cost rate $C$ follows a two-point distribution. Our aim is to demonstrate that an optimal price may not exist if Assumption 1 fails to hold.

EXAMPLE 2. Let $C$ follow a two-point distribution; that is, suppose that $\mathbb{P}\{C=a\}=p$ and $\mathbb{P}\{C=b\}=1-p$ for some numbers $a$, $b$, and $p$ with $0 \le a < b$ and $0<p<1$. We consider three overlapping intervals: $I_1 = [0, a\rho\mathbb{E}[W]/(1-\rho))$; $I_2 = [a\rho\mathbb{E}[W]/(1-(1-p)\rho), b\rho\mathbb{E}[W]/(1-(1-p)\rho))$; and $I_3 = [b\rho\mathbb{E}[W], \infty)$. Now, for any $K \ge 0$, if $K$ is located within $m \in \{1,2,3\}$ of these three intervals then there are $m$ equilibrium delays. Furthermore, if we define $i(K) := \min\{j \in \{1,2,3\} : K \in I_j\}$ then

$$p^*(K) = \begin{cases} 1 & \text{if } \ i(K)=1, \\ 1-p & \text{if } \ i(K)=2, \\ 0 & \text{if } \ i(K)=3. \end{cases}$$

It is easy to show that: (a) if $\rho < ((1-p)b-a)/((1-p)(b-a))$, then $r(K)$ attains its supremum $(\lambda(1-p)b\rho\mathbb{E}[W])/(1-(1-p)\rho)$ as $K \uparrow (b\rho\mathbb{E}[W])/(1-(1-p)\rho)$; (b) if $\rho > ((1-p)b-a)/((1-p)(b-a))$, then $r(K)$ attains its supremum $\lambda a\rho\mathbb{E}[W]/(1-\rho)$ as $K \uparrow a\rho\mathbb{E}[W]/(1-\rho)$; and (c) if $\rho = ((1-p)b-a)/((1-p)(b-a))$, then $r(K)$ attains its supremum as $K \uparrow (b\rho\mathbb{E}[W])/(1-(1-p)\rho)$ or $K \uparrow a\rho\mathbb{E}[W]/(1-\rho)$. Yet for none of these cases is there an optimal price $K^*$ such that $r(K^*) = \sup_{K\ge 0}\{r(K)\}$.

The main reason for this pathological result is that $H$, the CDF of the delay cost rate $C$, is not continuous and is flat in some intervals. Example 2 shows that the differentiable property of $H$ plays a crucial role in guaranteeing the existence of the optimal price in the revenue maximization problem.

We next analyze the optimal price $K^*$. To ease the analysis, we transform the price $K$ to $p^*(K)$, the probability of joining class 1. Note that $p^*(K)$ can attain any value in $(0,1]$—that is, by Lemma 3 and since $p^*(0)=1$ and $p^*(\infty)=0$. Hence $K(p):=\max\{K\geq 0: p^*(K)=p\}$ is well-defined for any $p\in(0,1]$. In addition, for any $K$ such that $p^*(K)=p$ we have—by (11)—that

$$H\left(\frac{K(1-p\rho)}{\rho\mathbb{E}[W]}\right)=1-p.$$

This equality yields the following explicit expression for $K(p)$:

$$K(p)=\frac{H^{-1}(1-p)\rho\mathbb{E}[W]}{1-p\rho}. \tag{12}$$

Here $H^{-1}$ is the *inverse function* of $H$, defined as $H^{-1}(y):=\inf\{x\geq 0: H(x)\geq y\}$ for $y\in[0,1]$.

Let $M(p):=pH^{-1}(1-p)/(1-p\rho)$. Then we have the following result, which characterizes the relationship between the original optimization problem $\max_{K\geq 0} r(K)$ and the transformed optimization problem $\max_{0<p\leq 1} M(p)$.

THEOREM 4. *Suppose Assumption 1 holds; suppose also that the function $M(p)$ is strictly quasi-concave in $p\in(0,1]$ and thus has a unique maximizer, denoted by $p^*\in(0,1)$. Then $r(K)$ is quasi-concave in $K>0$ and has the unique maximizer $K^*=K(p^*)$.*

The conditions given in Theorem 4 may seem restrictive. However, our next proposition shows that they are satisfied for many distribution families.

PROPOSITION 2. *If the delay cost rate $C$ satisfies one of the following conditions, then $M(p)$ is strictly quasi-concave for $p\in(0,1]$.*

(i) *$C$ follows the Burr distribution, with $H(x)=1-(1+(x/a)^d)^{-k}$, $x\geq 0$, whose parameters satisfy $kd>1$.*

(ii) *$C$ follows the Weibull distribution.*

(iii) *$C$ follows the uniform distribution, with $H(x)=(x-a)/(b-a)$ for $x\in[a,b]$, whose parameters satisfy $\rho<(b-2a)/(b-a)$.*

(iv) *$C$ follows the log-normal distribution with $H(x)=\Phi((\ln x-\mu)/\sigma)$, $x>0$, whose parameters satisfy $\sigma\rho<\sqrt{\pi/2}$; here $\Phi$ is the CDF of the standard normal distribution.*

(v) *$m^2(x)+m'(x)>0$ for all $x\geq 0$, where $m(x):=h(x)/(1-H(x))$ is the hazard rate of the distribution function $H$.*

Note that the exponential distribution satisfies condition (v) in Proposition 2 because its hazard rate is a positive constant. As a matter of fact, any distribution with an increasing hazard rate satisfies condition (v). Note also that Theorem 4 and Proposition 2(i) jointly imply not only the existence and uniqueness of the revenue-maximizing price but also the quasi-concavity of the Burr distribution's revenue function when $kd > 1$, which extends Theorem 2 in Gavirneni and Kulkarni (2016). If the conditions in Proposition 2 fail then the function $M(p)$ might not be quasi-concave; see Appendix B.2 for additional details.

We conclude this section with an example to show that the price $K_L$ under which the largest equilibrium delay is unstable might be equal to the revenue-maximizing price $K^*$.

EXAMPLE 3. Let $C \sim \text{lognormal}(\mu, \sigma^2)$. Then $M(p) = p\exp\{\mu + \sigma\Phi^{-1}(1-p)\}/(1-p\cdot\rho)$ and so $p^*$ depends only on $\sigma$ and $\rho$. Substituting $p^*$ into (12) now yields $K^* = \rho\mathbb{E}[W]\exp\{\mu + \sigma\Phi^{-1}(1-p^*)\}/(1-\rho p^*)$. If the system manager sets price $K^*$, then

$$F_e(x;K^*) = x(1-\rho) + \rho H\left(\frac{K^*}{\rho x}\right)x = x(1-\rho) + \rho\Phi\left(\frac{\log(\mathbb{E}[W]/(1-\rho p^*)x)}{\sigma} + \Phi^{-1}(1-p^*)\right)x.$$

Any equilibrium $(\tilde{w}_1, \tilde{w}_2)$ satisfies $F_e(\tilde{w}_2; K^*) = \mathbb{E}[W]$. That is, $v := \tilde{w}_2/\mathbb{E}[W]$ satisfies $f_e(v) = 1$, where

$$f_e(x) := x(1-\rho) + \rho\Phi\left(\frac{\log(1/(1-\rho p^*)x)}{\sigma} + \Phi^{-1}(1-p^*)\right)x.$$

The largest equilibrium delay $(\tilde{w}_{1,I(K^*)}, \tilde{w}_{2,I(K^*)})$ is unstable if and only if $v$ is the local minimum of $f_e$. Testing different values of $\sigma$ and $\rho$ in MATLAB reveals that, if $\sigma = 0.24$ and $\rho = 0.85$, then $p^* \approx 0.98$ and the corresponding $v \approx 6.01$—values that correspond exactly to the local minimizer of $f_e$. The implication is that the corresponding largest equilibrium delay is unstable: $K^* = K_L$.

Example 3 states that the largest equilibrium delay under the revenue-maximizing price $K^*$ might be unstable. However, in view of Remark 5, the largest delay under $K^*$ will be stable in nearly all cases. If the corresponding largest equilibrium delay for price $K^*$ is unstable, then we can lower the priority price slightly to a point where that delay is stable. In other words, the system manager can obtain the desired stability by sacrificing just a small amount of revenue.

## 6. Maximizing Social Welfare

In this section we consider the optimal pricing problem from the social planner's perspective, in which a priority price is set such that social welfare (i.e., the sum of customer utilities and the service system's utility) is maximized. For this problem, the purpose of setting a priority price is not to maximize the system's revenue but rather to reduce the overall delay cost by distinguishing

among customers who are relatively less and more sensitive to delays and then giving priority to the latter.

To fix ideas, for now we assume that $v$, the value of receiving a service, is homogeneous across customers. Given priority price $K$ and a corresponding equilibrium delay $(\tilde{w}_1, \tilde{w}_2)$, a customer with delay cost rate $c \le K/(\tilde{w}_2 - \tilde{w}_1)$ will join class 2, which incurs average delay cost $c\tilde{w}_2$, while a customer with delay cost rate $c > K/(\tilde{w}_2 - \tilde{w}_1)$ will join class 1, which incurs average delay cost $c\tilde{w}_1$. Hence, the social welfare–maximizing price $K$ and the equilibrium delay $(\tilde{w}_1, \tilde{w}_2)$ are given by

$$\begin{aligned} S(K) &:= \int_0^{K/(\tilde{w}_2-\tilde{w}_1)} (v - c\tilde{w}_2)\, dH(c) + \int_{K/(\tilde{w}_2-\tilde{w}_1)}^{\infty} (v - c\tilde{w}_1)\, dH(c) \\ &= \int_0^{K/(\rho\tilde{w}_2)} (v - c\tilde{w}_2)\, dH(c) + \int_{K/(\rho\tilde{w}_2)}^{\infty} (v - c(1-\rho)\tilde{w}_2)\, dH(c) \\ &= v - \tilde{w}_2 \bigg( \int_0^{\infty} c\, dH(c) - \rho \int_{K/(\rho\tilde{w}_2)}^{\infty} c\, dH(c) \bigg), \end{aligned} \tag{13}$$

where the second equality follows from (7). Here the total customer delay cost under priority price $K$ and equilibrium delay $(\tilde{w}_1, \tilde{w}_2)$ is

$$\tilde{w}_2 \bigg( \int_0^{\infty} c\, dH(c) - \rho \int_{K/(\rho\tilde{w}_2)}^{\infty} c\, dH(c) \bigg).$$

Note that

$$p(\tilde{w}_1, \tilde{w}_2; K) = 1 - H\bigg(\frac{K}{\tilde{w}_2 - \tilde{w}_1}\bigg) = 1 - H\bigg(\frac{K}{\rho\tilde{w}_2}\bigg)$$

and

$$\tilde{w}_2 = \frac{\mathbb{E}[W]}{1 - p(\tilde{w}_1, \tilde{w}_2; K)\rho}.$$

So if we put $p = p(\tilde{w}_1, \tilde{w}_2; K)$, then

$$\frac{K}{\rho\tilde{w}_2} = H^{-1}(1-p), \quad \tilde{w}_2 = \frac{\mathbb{E}[W]}{1-p\rho}; \tag{14}$$

hence the expression for social welfare can be written as

$$S(K) = v - \frac{\mathbb{E}[W]}{1-p\rho} \bigg( \int_0^{\infty} c\, dH(c) - \rho \int_{H^{-1}(1-p)}^{\infty} c\, dH(c) \bigg). \tag{15}$$

We also have our next lemma, which states that $p(\tilde{w}_1, \tilde{w}_2; K)$ has range $[0, 1]$.

LEMMA 4. *As the price $K$ varies from 0 to $\infty$, $p(\tilde{w}_1, \tilde{w}_2; K)$ can attain any value in $[0, 1]$.*

This lemma and equation (15) together imply that the problem of maximizing social welfare is equivalent to the following optimization problem:

$$\min_{0 \le p \le 1} f_S(p) := \frac{1}{1-p\rho} \bigg( \int_0^{\infty} c\, dH(c) - \rho \int_{H^{-1}(1-p)}^{\infty} c\, dH(c) \bigg), \tag{16}$$

whose solution can be characterized as follows.

THEOREM 5. *Under Assumption 1, the optimization problem* (16) *admits a unique minimizer* $p_S^* \in (0,1)$, *where* $p_S^* = 1 - H(q_S^*)$ *and where* $q_S^*$ *is the unique positive root of*

$$g_S(q) := q(1-\rho(1-H(q))) - \int_0^\infty c \cdot h(c)\, dc + \rho \int_q^\infty c \cdot h(c)\, dc = 0. \tag{17}$$

Combining Theorem 5 with (14) allows us to characterize the social welfare–maximizing equilibrium delay $(\tilde{w}_{1,S}^*, \tilde{w}_{2,S}^*)$ and the optimal price $K_S^*$ as follows:

$$\tilde{w}_{2,S}^* = \frac{\mathbb{E}[W]}{1-p_S^*\rho}, \quad \tilde{w}_{1,S}^* = (1-\rho)\tilde{w}_{2,S}^*, \quad K_S^* = \frac{\rho \mathbb{E}[W] \cdot H^{-1}(1-p_S^*)}{1-p_S^*\rho}. \tag{18}$$

Note that $(\tilde{w}_{1,S}^*, \tilde{w}_{2,S}^*)$ may not be the largest equilibrium delay under price $K_S^*$. This outcome can be explained by inspecting (13). Given priority price $K$, social welfare might not be maximized at $\tilde{w}_2 = \tilde{w}_{2,I(K)}$ because the function

$$G(w;K) := w\left( \int_0^\infty c\, dH(c) - \rho \int_{K/(\rho w)}^\infty c\, dH(c) \right)$$

is not necessarily monotone in $w$.

Analogously to the case of maximizing the service system's revenue, the question arises of whether the social welfare–maximizing equilibrium delay $(\tilde{w}_{1,S}^*, \tilde{w}_{2,S}^*)$ is stable. According to Theorem 3, the equilibrium delay $(\tilde{w}_{1,S}^*, \tilde{w}_{2,S}^*)$ is stable if and only if $F_e(\cdot; K_S^*)$ is strictly increasing at point $\tilde{w}_{2,S}^*$. In the case of a power distribution with $\alpha = 2$ and $\rho = 0.9$ (see section 7.4), one can easily calculate that $F_e'(\tilde{w}_{2,S}^*; K_S^*) = -0.0659 < 0$. Hence the social welfare–maximizing equilibrium delay is not necessarily stable. Even so, we can identify a condition under which that delay is stable.

PROPOSITION 3. *If Assumption 1 holds and if the function* $c \cdot h(c)$ *is decreasing in* $c$, *then the social welfare–maximizing equilibrium delay* $(\tilde{w}_{1,S}^*, \tilde{w}_{2,S}^*)$ *is stable when the priority price is set to* $K_S^*$.

If $(\tilde{w}_{1,S}^*, \tilde{w}_{2,S}^*)$ is unstable then any small perturbation will drive the system to other stable equilibrium delays, which would degrade social welfare. It follows that the priority pricing queueing system examined here is likely not an efficient (much less the best) way to maximize social welfare. Alternatives such as a threshold policy or a dynamic priority service policy may be more effective than the static priority policy studied in this paper.

## 7. Special Distribution Families and Numerical Study

We now investigate how both the revenue-maximizing price and the social welfare–maximizing price are affected by different families of delay cost distributions. In particular, we derive these optimal prices for uniform, exponential, Weibull, and power distributions—after which we conclude this section with a numerical study.

### 7.1. Uniform Distribution

Previous research reflects wide adoption of the uniform distribution; see, for example, Zhang et al. (2007), Gilland and Warsing (2009), and Yu et al. (2013). Let the delay cost rate $C$ follow a uniform distribution with support $[a,b]$.

We assume that the system manager's goal is to maximize the system's long-run average revenue. It then follows from the first-order optimality condition and the proof of Proposition 2 that

$$p^* = \min\left\{\frac{1-\sqrt{(1-b\rho(b-a)^{-1})^+}}{\rho}, 1\right\}$$

for $p^*$ as defined in Theorem 4. Hence the revenue-maximizing price is given by

$$K^* = \max\left\{(b-a)\big(1-\sqrt{(1-b\rho(b-a)^{-1})^+}\big)\cdot\mathbb{E}[W], \frac{a\rho}{1-\rho}\cdot\mathbb{E}[W]\right\}. \tag{19}$$

Now suppose that the system manager wishes also to maximize social welfare. Solving $g_S(q_s^*)=0$ for $g_S$ as defined in (17), we obtain

$$q_S^* = \frac{b\rho+a-b+(b-a)\sqrt{1-\rho}}{\rho}$$

and so the social welfare–maximizing price is given by

$$K_S^* = \frac{b\rho-b+a+(b-a)\sqrt{1-\rho}}{\sqrt{1-\rho}}\cdot\mathbb{E}[W]. \tag{20}$$

Note that

$$\begin{aligned} F_e'(\tilde{w}_{2,S}^*; K_S^*) &= 1-\rho+\rho H(q_S^*) - \rho q_S^*\cdot h(q_S^*) \\ &= 1-\rho+\rho\frac{q_S^*-a}{b-a} - \rho q_S^*\frac{1}{b-a} \\ &= \frac{b-a-b\rho}{b-a}. \end{aligned}$$

We conclude that $(\tilde{w}_{1,S}^*, \tilde{w}_{2,S}^*)$ is stable if and only if $\rho < (b-a)/b$. Because $\rho<1$, this equilibrium delay must be stable when $a=0$.

A comparison between (19) and (20) immediately yields the following result.

PROPOSITION 4. *If $a=0$ in the case of a uniform distribution, then $K^*=K_S^*$.*

The condition that customers' delay costs follow a uniform distribution with $a=0$ appears often in the literature (Gilland and Warsing 2009; Yu et al. 2013). Proposition 4 states that, under this condition, the system manager can set a price at which both the system's revenue and social welfare are maximized—provided there are only two priority classes. The same result is obtained also by Gilland and Warsing (2009, Prop. 3). However, we must point out that this condition typically fails to hold in practical situations, which results in revenue-maximizing and social welfare–maximizing prices that differ.

### 7.2. Exponential Distribution

Now we assume that $H(x)=1-e^{-\kappa x}$ for $x\geq 0$, where $\kappa>0$ is the rate parameter. By the first-order optimality condition, $p^*$ is the unique root of $\ln p+1-\rho\cdot p=0$ in $(0,1)$ (see Lemma 8(i) in Appendix B) and the revenue-maximizing price is

$$K^*=\frac{\rho\mathbb{E}[W]}{\kappa}.$$

The corresponding optimal revenue is $R^*=\lambda\rho p^*\mathbb{E}[W]/\kappa$.

It follows from (17) that $q_S^*$ is the unique positive root[5] of $\kappa q-1+\rho e^{-\kappa q}=0$ and that the social welfare–maximizing price is

$$K_S^*=\frac{q_S^*\rho\mathbb{E}[W]}{1-\rho e^{-\kappa q_S^*}}=\frac{\rho\mathbb{E}[W]}{\kappa}=K^*.$$

That is: in the case of an exponential distribution and two priority classes, the social welfare–maximizing price *always* equals the revenue-maximizing price. Moreover, the equilibrium delay is unique under any price (by Remark 2), from which it follows that the social welfare–maximizing equilibrium delay must be stable (by Remark 4).

### 7.3. Weibull Distribution

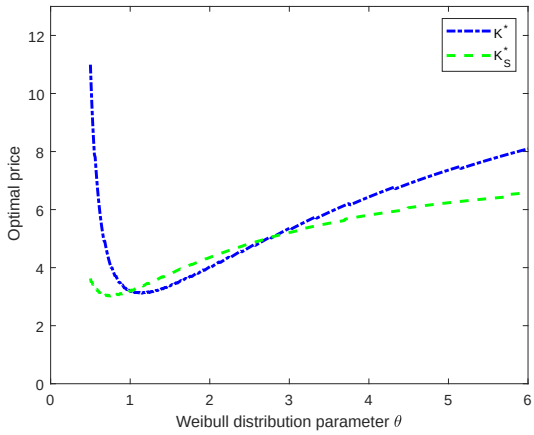
The distribution function for a Weibull distribution is $H(x;\tau,\theta)=1-e^{-(x/\tau)^\theta}$ for $x\geq 0$, where $\theta>0$ is the shape parameter and $\tau>0$ is the scale parameter. We shall use $K^*(\tau,\theta)$ and $K_S^*(\tau,\theta)$ to denote (respectively) the revenue-maximizing price and social welfare–maximizing price under a Weibull distribution. Our next lemma states that both $K^*(\tau,\theta)$ and $K_S^*(\tau,\theta)$ are radically homogeneous in $\tau$.

[5] Let $e^{-\kappa q}=p$; then $\kappa q-1+\rho e^{-\kappa q}=-\ln p-1+\rho p$. Then, by Lemma 8(ii), $\kappa q-1+\rho e^{-\kappa q}=0$ has only one positive root (which we denote by $q_S^*$). Also, $p^*=e^{-\kappa q_S^*}$.

LEMMA 5. *For each $\tau > 0$, we have $K^*(\tau,\theta) = \tau K^*(1,\theta)$ and $K_S^*(\tau,\theta) = \tau K_S^*(1,\theta)$.*

Next we investigate the relationship between $K^*(\tau,\theta)$ and $K_S^*(\tau,\theta)$. In view of Lemma 5, it suffices for this purpose to compare $K^*(1,\theta)$ and $K_S^*(1,\theta)$. Figure 3 shows that, as $\theta$ varies, no relationship between $K^*(1,\theta)$ and $K_S^*(1,\theta)$ dominates.[6] Note that $\theta = 1$ corresponds to the case of an exponential distribution.

**Figure 3 Comparison of $K^*$ and $K_S^*$ for Weibull Distribution**




### 7.4. Power Distribution

We follow Guo and Zipkin (2007) in considering the power distribution, whose PDF is $H(x) = x^\alpha$ for $\alpha > 0$ and $x \in [0,1]$. A distinct feature of the power distribution is that it can capture different kinds of distribution structures (concave, linear, and convex) as $\alpha$ varies. Using logic similar to that in section 7.1, we find that the revenue-maximizing price is

$$K^* = \frac{(2\rho)^{1-1/\alpha}\left(2\rho - 1 - \alpha + \sqrt{(1+\alpha)^2 - 4\rho\alpha}\right)^{1/\alpha}}{1 - \alpha + \sqrt{(1+\alpha)^2 - 4\rho\alpha}} \cdot \mathbb{E}[W]$$

and that the social welfare–maximizing price is

$$K_S^* = \frac{q_S^{*2}\rho}{(1-\rho)\alpha(1-q_S^*)} \cdot \mathbb{E}[W];$$

here $q_S^*$ is the unique root of $(1-\rho)(\alpha q + q - \alpha) + \rho q^{\alpha+1} = 0$ in $(0,1)$ (see Lemma 8(iii)). We remark that the social welfare–maximizing equilibrium delay may not be stable as $\alpha$ varies ( see the paragraph before Proposition 3).

Much as in the case of the Weibull distribution, for the power distribution there is no dominant relationship between $K^*$ and $K_S^*$; this claim is evidenced by Figure 4. In Figure 5 it is clear that

[6] The system parameters used here (and in Figure 4) are $\mathbb{E}[B] = 1$, $\mathbb{E}[B^2] = 2$, and $\rho = 0.8$.

both the revenue-maximizing price and the optimal revenue increase with and are also extremely sensitive to $\alpha$, or the structure of $H(\cdot)$. The uniform distribution, which is widely assumed in the literature, corresponds to the special case of $\alpha = 1$. Hence a manager who wrongly assumes a uniform distribution may incur a substantial loss of revenue.

**Figure 4 Comparison of $K^*$ and $K_S^*$ for Power Distribution**

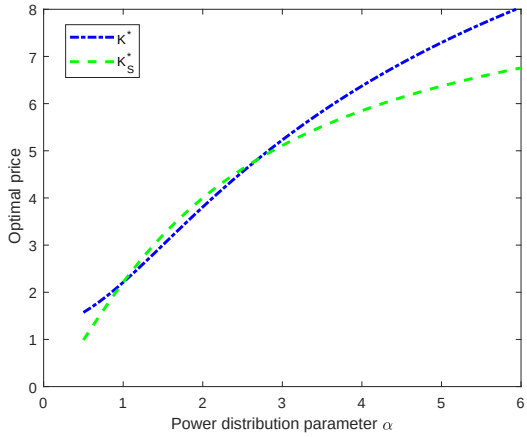
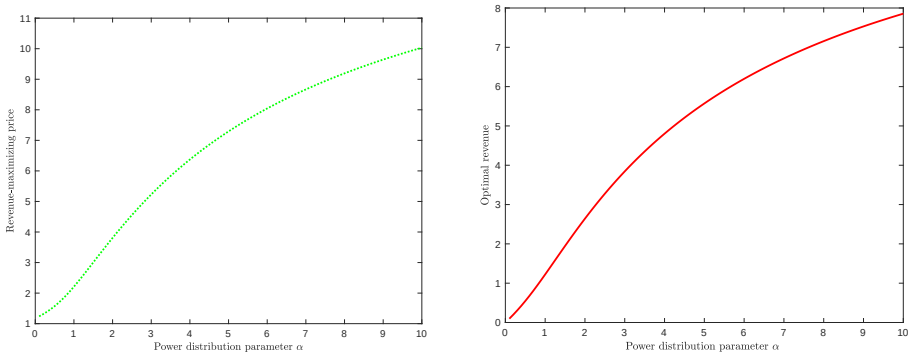



**Figure 5 Power Distribution**




(a) Revenue-maximizing price (b) Optimal revenue

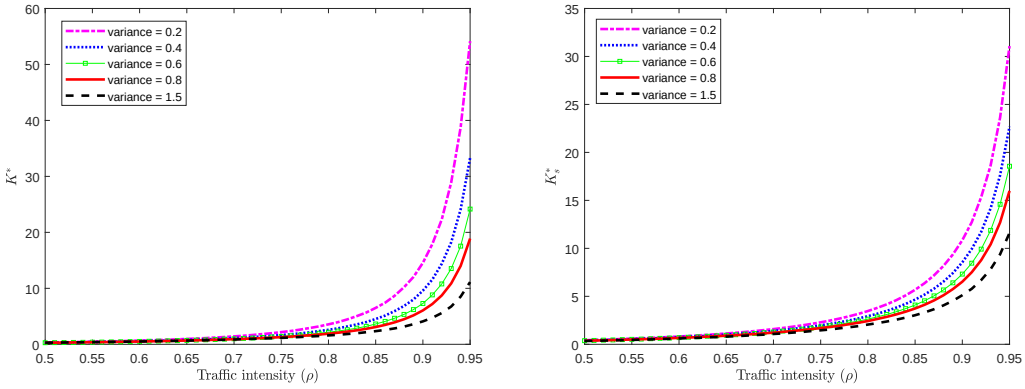
## 7.5. Numerical Study

We now conduct numerical studies to examine how various delay cost rate distributions affect the revenue-maximizing price and social welfare–maximizing price. Thus we investigate the effects of traffic intensity $\rho$ and of delay cost distributions on these two optimal prices. For that purpose, we assess these effects from two perspectives. First, we use a given distribution family (here, log-normal distributions) and consider distributional heterogeneity by changing the variance while keeping the mean fixed. Second, we compare different distribution families. In the following numerical experiments, we set $\mathbb{E}[B] = 1$ and $\mathbb{E}[B^2] = 2$.

**7.5.1. Identical Distribution Family.** In this numerical study we focus on the log-normal distribution: $C \sim \text{lognormal}(\mu, \sigma^2)$. We choose distribution parameters such that the mean is 0.75 and the variances are 0.2, 0.4, 0.6, 0.8, and 1.5 (see Appendix C.1 for detailed descriptions of these parameters).

*Effect of traffic intensity $\rho$ on optimal prices.* As Figure 6 illustrates, optimal prices are increasing in the system's traffic intensity $\rho$. Therefore, the system manager should raise the priority price in response to greater congestion.

**Figure 6 Optimal Prices for Log-Normal Distribution Family.**

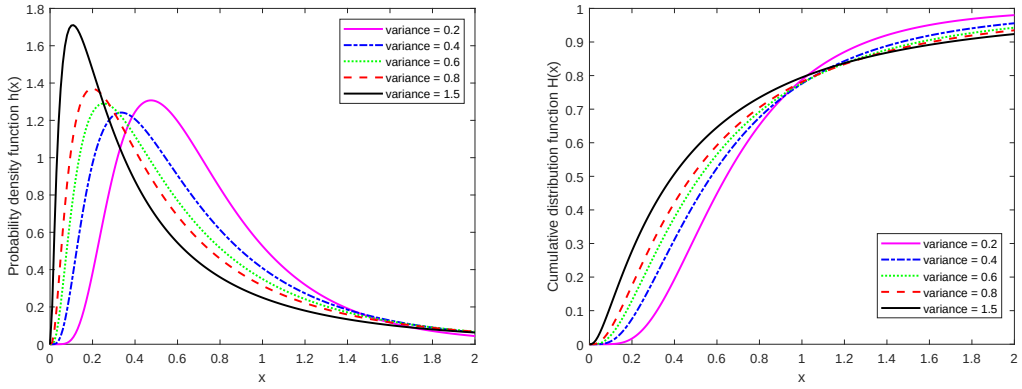



(a) Revenue maximization (b) Social welfare maximization

*Effect of delay cost distribution on optimal prices.* Figure 6 indicates also that the optimal prices are quite sensitive to delay cost distributions, especially when the traffic intensity $\rho$ is high. Furthermore, optimal prices decrease with increases in the delay cost distribution's variance. This can be explained by observing Figure 7, which plots the PDFs and CDFs for log-normal distributions with the parameters just described. The delay cost distribution becomes more left-skewed as variance increases, which means that a higher proportion of customers have a low delay cost rate and so are unwilling to pay additional fee to join the priority class. In this case, then, the system manager should reduce the price as a means of attracting more customers to join that class.

*Effect of traffic intensity $\rho$ and delay cost distribution on the probability of joining the priority class.* Another quantity of interest is the probability of joining the priority class, which is called the "participation level" in Gavirneni and Kulkarni (2016). Figure 8 plots the probabilities of joining the priority class for various traffic intensities in the cases of revenue maximization and social welfare maximization, each of which is sensitive to the distribution parameters. We find that $p_S^*$ is lower than $p^*$; that is, the social welfare–maximizing price results in a smaller fraction of priority customers than does the revenue-maximizing price—in line with the observations of Naor (1969)

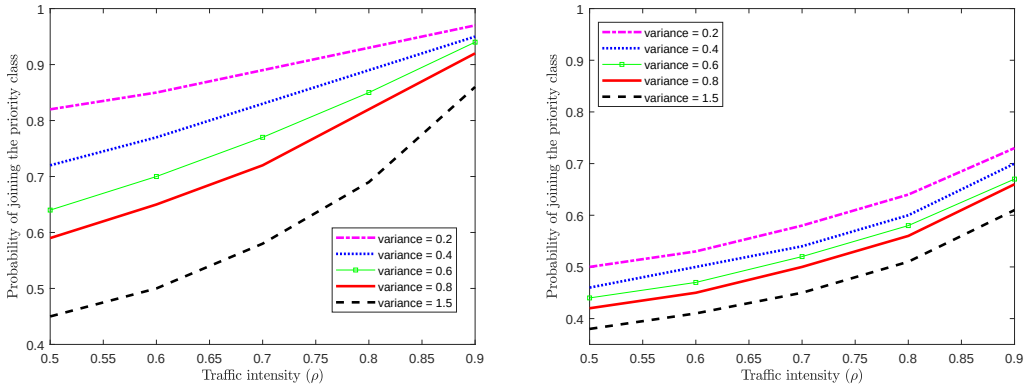
**Figure 7 PDFs and CDFs for Log-normal Distribution Family.**




(a) Probability density function (b) Cumulative distribution function

and Gavirneni and Kulkarni (2016). Another noteworthy finding is that higher traffic intensity induces a larger fraction of customers to join the priority class despite a higher priority service price being charged.

**Figure 8 Probability of Joining the Priority Class for Log-normal Distribution Family.**




(a) Revenue maximization (b) Social welfare maximization

**7.5.2. Different Distribution Families.** Next we investigate and compare the optimal prices and the probability of joining the priority class under the uniform, exponential, Weibull, power, and log-normal distribution families. Here we set the parameters such that all five of these distributions have the same mean (0.75) but different variances; see Table 1 for the specifics.

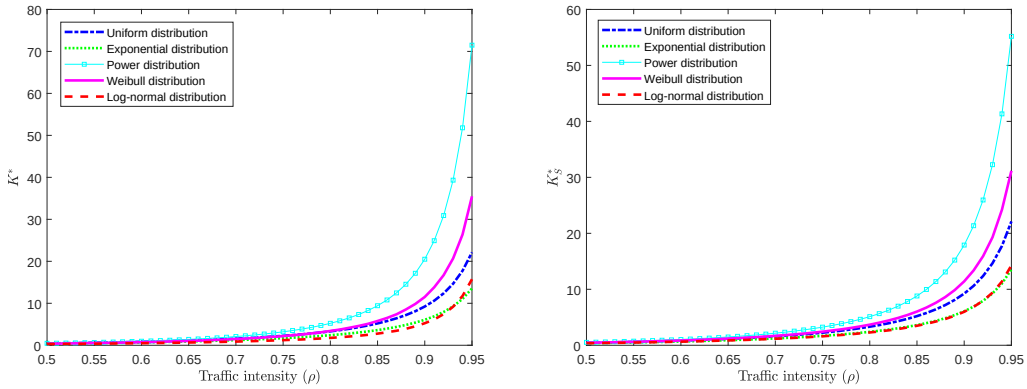
Figure 9 and Figure 10 plot (respectively) the optimal prices and the probabilities of joining the priority class for various levels of traffic intensity under these five distributions. We find that the conclusions drawn in section 7.5.1 are fairly robust in different distribution families. For the five

**Table 1 Distribution Parameters for the Second Numerical Study**

| Distribution | Uniform | Exponential | Weibull | Power | Log-normal |
|---|---|---|---|---|---|
| Parameters | $a=0, b=3/2$ | $\kappa=4/3$ | $\tau=0.8463, \theta=2$ | $\alpha=3$ | $\mu=-0.7985, \sigma=1.0107$ |

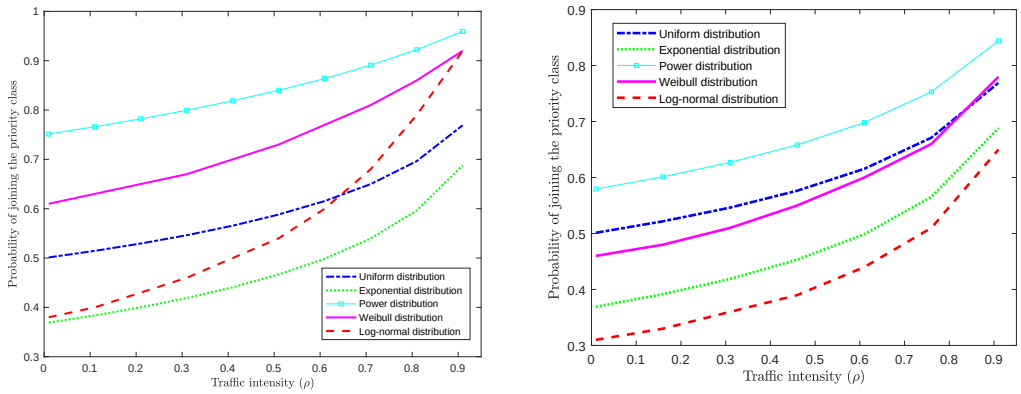
families we tested, optimal prices and the probability of a customer joins the priority class both increase with traffic intensity $\rho$ and—especially when $\rho$ is high—are extremely sensitive to delay cost distribution.

**Figure 9 Optimal Price For Different Distribution Families with the Same Mean.**




(a) Revenue maximization (b) Social welfare maximization

**Figure 10 Probability of Joining the Priority Class for Different Distribution Families with the Same Mean.**




(a) Revenue maximization (b) Social welfare maximization

In the parameter settings used for this numerical study, only the mean of the delay cost rate remains fixed. However, one might argue that optimal prices are actually more sensitive to variance

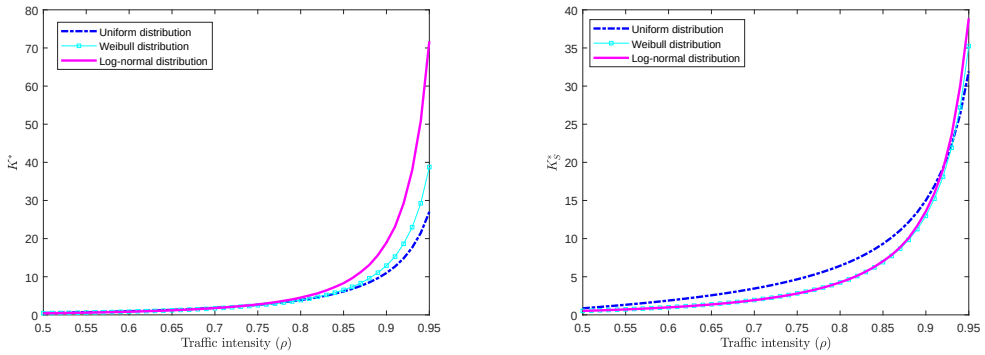
of the delay cost rate than to the distribution family. To test that possibility, our next numerical study focuses on two-parameter distributions—namely, the uniform, Weibull, and log-normal distributions—and sets the same mean (0.8874) and variance (0.2360) for both. The other distribution parameters for this study are listed in Table 2.

**Table 2 Distribution Parameters for the Third Numerical Study**

| Distribution | Uniform | Weibull | Log-normal |
|---|---|---|---|
| Parameters | $a = 0.0460, b = 1.7288$ | $\tau = 1, \theta = 1.9$ | $\mu = -0.2506, \sigma = 0.5120$ |

It is clear from Figure 11 that, when the system is relatively congested, the optimal prices are *not* primarily determined by the first two moments of the delay cost distribution. That finding confirms the value of carefully examining the distribution of customers' delay costs before any pricing decisions are made.

**Figure 11 Optimal Prices for Different Distributions with the Same First Two Moments.**

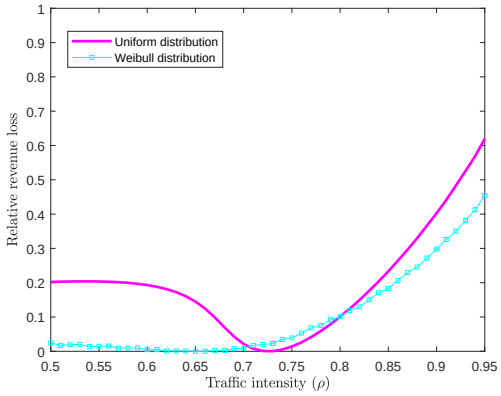



(a) Revenue maximization (b) Social welfare maximization

Finally, we evaluate how a mis-specified distribution affects system's revenue. For this purpose, we provisionally assume that customers' delay cost follows a log-normal distribution. Now suppose that the system manager (wrongly) believes this delay cost to be uniformly distributed. In that event, the manager will set the priority price $K^*_{\text{uniform}} = \arg\max_{K\geq 0} r_{\text{uniform}}(K)$ rather than the true optimal price $K^*_{\text{lognormal}} = \arg\max_{K\geq 0} r_{\text{lognormal}}(K)$, where $r_{\text{uniform}}(K)$ and $r_{\text{lognormal}}(K)$ are the system's revenue under price $K$ given that customers' delay cost follows uniform and log-normal distributions respectively. Then the *relative revenue gap*, defined as $[(r_{\text{lognormal}}(K^*_{\text{lognormal}}) - r_{\text{lognormal}}(K^*_{\text{uniform}}))/r_{\text{lognormal}}(K^*_{\text{lognormal}})] \times 100\%$, quantifies the relative revenue loss if the system manager mistakes a log-normal distribution for a uniform one. One can similarly consider the case in which the system manager wrongly regards customers' delay cost distribution as Weibull.

Figure 12 plots the relative revenue losses for these two distributions, where the parameters in Table 2 are used to make the mean and variance equal. These comparisons demonstrate that the revenue loss due to a misspecified distribution can be significant when traffic intensity is high.

**Figure 12 Relative Revenue Loss when Customers' Delay Cost Follows a Log-normal Distribution**




## 8. An Extension: Multiple Priority Classes

This section extends our previous two–priority-class queueing model to a model of multiple priority classes. We shall consider $N$ different priority classes, indexed via $\mathcal{N} := \{1, 2, \ldots, N\}$, such that class $i$ is strictly prioritized over class $j$ for any $i < j$. The system manager's objective is to set a priority price *menu* $\boldsymbol{K} := (K_i; i \in \mathcal{N})$, where $K_i$ is the price charged for joining class $i$, such that the resulting long-run revenue (or social welfare) is maximized. Just as in the two–priority-class queueing model, we let $K_N = 0$ and assume that all customers eventually join the service system.[7]

We continue to assume that the queues are not observable to customers. Hence the system will devise a delay announcement menu $\boldsymbol{w} := (w_i; i \in \mathcal{N})$ for arriving customers, where $w_i$ is the average waiting time for class-$i$ customers. The joining decision of each customer is based on minimizing her disutility from the corresponding average delay cost *plus* the charged priority price. Therefore, a customer whose delay cost rate is $c$ will join class $i$ provided that

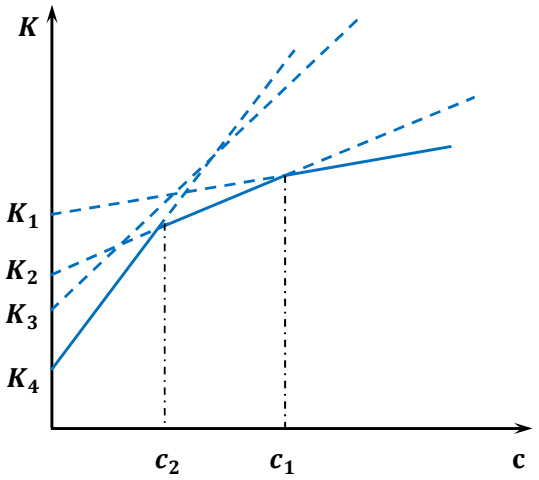
$$i = \arg\min_{j \in \mathcal{N}} \{K_j + cw_j\}. \tag{21}$$

According to the service principle, $w_1 < w_2 < \cdots < w_N$. It should therefore suffice to consider the case $K_1 > K_2 > \cdots > K_N$, since if $K_i \le K_{i+1}$ for some $i$ then all customers will prefer class $i$ to class $i+1$.

One could use (21) to make the same argument as that in section 4. Yet here we adopt the alternative approach of Nazerzadeh and Randhawa (2018) because it facilitates the characterization of arrival rates and leads to a simpler optimization problem.

[7] If $K_N \neq 0$ then, for all $i \in \mathcal{N}$, we let $K_i' := K_i - K_N$ and regard $K_i'$ as the priority price charged for joining class $i$.

**Figure 13 Effective Cost Function of a Four-Class System**




To illustrate this approach, we first look at Figure 13, which plots the four maps $c \mapsto K_i + cw_i$ $(i = 1, 2, \ldots, 4)$ for some price–delay menu $(\boldsymbol{K}, \boldsymbol{w})$. The solid line represents the effective cost of joining the system and thus also characterizes customers' optimal joining behavior. It is clear from this graph that, for a customer with delay cost rate $c$: if $c < c_2$ then it is optimal to join class 4; if $c_2 \le c < c_1$, it is optimal to join class 2; and if $c \ge c_1$ then it is optimal to join class 1. Because no customers will join class 3, $(K_3, w_3)$ can be removed from the price–delay pair and this four-class system will, in effect, have only three classes.

These considerations suggest that the system could present arriving customers with a *delay threshold menu* $(\boldsymbol{c}, \boldsymbol{w})$ rather than with a price–delay menu. We define the *threshold set* as $\boldsymbol{c} := (c_i; i \in \mathcal{N} \cup \{0\})$ for $c_0 = \infty \ge c_1 \ge \ldots \ge c_{N-1} \ge c_N = 0$. So for each $i = 1, \ldots, N$, a customer whose delay cost rate is in the interval $[c_i, c_{i-1})$ will join class $i$. Moreover, the threshold $c_i$, $i = 1, 2, \ldots, N-1$ can be determined by solving $K_i + c_i w_i = K_{i+1} + c_i w_{i+1}$. Therefore,

$$K_i = \sum_{j=i}^{N-1} c_j (w_{j+1} - w_j), \quad i = 1, 2, \ldots, N-1. \tag{22}$$

In the sequel, we replace $\boldsymbol{K}$ with $\boldsymbol{c}$. Then the probability of joining class $i$ can be written as

$$p_i(\boldsymbol{c}) := H(c_{i-1}) - H(c_i). \tag{23}$$

Hence it follows from standard results for the M/G/1 non-preemptive priority queue (see e.g. Adan and Resing 2002) that the average delay for class-$i$ customers is

$$\mathbb{E}[W_i] = \frac{\lambda \mathbb{E}[B^2]}{2\big(1 - \rho \sum_{j=1}^{i-1} p_j(\boldsymbol{c})\big)\big(1 - \rho \sum_{j=1}^{i} p_j(\boldsymbol{c})\big)} = \frac{\lambda \mathbb{E}[B^2]}{2\big(1 - \rho \bar{H}(c_{i-1})\big)\big(1 - \rho \bar{H}(c_i)\big)}, \tag{24}$$

where $\rho = \lambda \mathbb{E}[B]$.

### 8.1. Equilibrium Delay

Given a priority price menu $\boldsymbol{K}$, a delay announcement menu $\boldsymbol{w}$ is an *equilibrium delay* if $\mathbb{E}[\boldsymbol{W}] = \boldsymbol{w}$ for $\boldsymbol{W} = (W_i; i \in \mathcal{N})$. We can use (24) to determine the equilibrium delay $\boldsymbol{w}$ via the equilibrium threshold set $\boldsymbol{c}$. It follows from (22) and (24) that the equilibrium threshold set $\boldsymbol{c}$ satisfies

$$c_i \cdot \frac{\lambda\rho\mathbb{E}[B^2](\bar{H}(c_{i+1}) - \bar{H}(c_{i-1}))}{2(1-\rho\bar{H}(c_{i-1}))(1-\rho\bar{H}(c_i))(1-\rho\bar{H}(c_{i+1}))} = K_i - K_{i+1}, \quad i = 1, 2, \ldots, N-1. \tag{25}$$

We can now present the following result, which is an extension of Theorem 1 for the case of multiple priority classes. Note that—as in the two-class case covered by Theorem 2—the equilibrium delay may not be unique.

THEOREM 6. *Suppose that the CDF of delay cost rate $H$ is continuous. Then there exists at least one equilibrium delay.*

The main idea of proving Theorem 6 is as follows. Fix any $c_1 > 0$. Solving (25) with $i = 1$ will identify the corresponding $c_2$ (if it exists). Then solving (25) iteratively with $i = k$ ($k = 2, \ldots, N-1$) yields the corresponding $c_3, \ldots, c_N$. An equilibrium delay corresponds to a solution sequence $\{c_i; i \in \mathcal{N}\}$ with $c_N = 0$. Now, to prove the existence of an equilibrium delay, it suffices to show that the solution sequence is continuous with respect to $c_1$. In fact, one can find all the equilibrium delays by varying $c_1$.

### 8.2. Maximizing Service System Revenue under Multiple Priority Classes

Next we consider a priority pricing problem in which the manager seeks to maximize the system's long-term average revenue by offering a price–delay menu when there are $N$ priority classes. We denote that revenue by $\lambda\sum_{i=1}^{N-1} K_i p_i(\boldsymbol{K})$, where $p_i(\boldsymbol{K})$ is the probability of joining class $i$ under price set $\boldsymbol{K}$ with its corresponding equilibrium delay announcement. So given (22) and (23), the system manager wishes to solve the following optimization problem:

$$\max_{\boldsymbol{c}} \quad \lambda \sum_{i=1}^{N-1} \left[ (H(c_{i-1}) - H(c_i)) \sum_{j=i}^{N-1} c_j (w_{j+1} - w_j) \right] \tag{26}$$

$$\text{s.t.} \quad w_i = \frac{\lambda\mathbb{E}[B^2]}{2(1-\rho\bar{H}(c_{i-1}))(1-\rho\bar{H}(c_i))}, \quad i = 1, 2, \ldots, N, \tag{27}$$

$$c_0 = \infty \geq c_1 \geq \ldots \geq c_{N-1} \geq c_N = 0.$$

If we interchange the order of summation in (26), use (27), and put $d_i = \bar{H}(c_i)$, then this problem can be simplified to

$$\max_{\boldsymbol{d}=(d_i;i\in\mathcal{N}\cup\{0\})} \frac{\lambda^2\mathbb{E}[B^2]\rho}{2}\sum_{i=1}^{N-1} H^{-1}(1-d_i)\cdot d_i \frac{d_{i+1}-d_{i-1}}{(1-\rho d_{i-1})(1-\rho d_i)(1-\rho d_{i+1})} \tag{28}$$
$$\text{s.t.} \quad d_0 = 0 \le d_1 \le \ldots \le d_{N-1} \le d_N = 1.$$

However, (28) is *not*, in general, a convex programming problem—a fact that renders the analysis rather difficult. We shall therefore resort to the numerical study undertaken in section 8.4, a (local) optimum can be identified numerically with the aid of either a genetic or a gradient descent algorithm. Note that (28) can also be solved under some specific distributions, as shown by the following result (for the uniform distribution).

PROPOSITION 5. *Suppose that the delay cost rate $C \sim U(0,b)$. Then the optimal solution to* (28) *is given by*

$$d_i^* = \frac{1-(1-\rho)^{i/N}}{\rho}, \quad i = 0,1,\ldots,N,$$

*and so the revenue-maximizing prices are*

$$K_i = \sum_{j=i}^{N-1} \frac{b[(1-\rho)^{1/N}-(1-\rho)]}{\rho}\left[\frac{\lambda\mathbb{E}[B^2]}{2(1-\rho)^{(2j+1)/N}} - \frac{\lambda\mathbb{E}[B^2]}{2(1-\rho)^{(2j-1)/N}}\right], \quad i = 0,1,\ldots,N.$$

*In this case, the service system's optimal revenue is*

$$\frac{\lambda^2\mathbb{E}[B^2]b}{2\rho^2}\left[\frac{(2-\rho)\rho}{1-\rho} - \left[(1-\rho)^{-(N-1)/N} - (1-\rho)^{(N-1)/N} + (N-1)((1-\rho)^{-1/N} - (1-\rho)^{1/N})\right]\right].$$

It is worth noting that, if $C \sim U(a,b)$ with $a > 0$, then Proposition 5 no longer holds. For a detailed discussion, see Remark 8 in Appendix A.

### 8.3. Maximizing Social Welfare under Multiple Priority Classes

Here we consider the social welfare maximization problem described in section 6. When there are more than two priority classes, the system manager must solve a different optimization problem as follows:

$$\min_{\mathbf{c}} \sum_{i=1}^{N} w_i \int_{c_i}^{c_{i-1}} x\, dH(x)$$
$$\text{s.t.} \quad w_i = \frac{\lambda\mathbb{E}[B^2]}{2(1-\rho\bar{H}(c_{i-1}))(1-\rho\bar{H}(c_i))}, \quad i = 1,2,\ldots,N,$$
$$c_0 = \infty \ge c_1 \ge \cdots \ge c_{N-1} \ge c_N = 0.$$

If we put $d_i = \bar{H}(c_i)$ and replace the variable $x$ in the integral with $y = \bar{H}(x)$, then this problem can be reduced to

$$\min_{\boldsymbol{d}=(d_i;i\in\mathcal{N}\cup\{0\})} \frac{\lambda\mathbb{E}[B^2]}{2}\sum_{i=1}^{N}\frac{\int_{d_{i-1}}^{d_i} H^{-1}(1-y)\,dy}{(1-\rho d_{i-1})(1-\rho d_i)} \quad \text{s.t.} \quad d_0 = 0 \le d_1 \le \cdots \le d_{N-1} \le d_N = 1. \tag{29}$$

This problem, too, is (in general) not a convex programming; hence we shall examine it mainly by way of numerical examples. However, (29) can be simplified considerably for some special distributions, as our next proposition demonstrates for the case of a uniformly distributed delay cost rate.

PROPOSITION 6. *Suppose that the delay cost rate $H \sim U(a,b)$. Then the optimal solution to* (29) *is given by*

$$d_i^* = \frac{1-(1-\rho)^{k/N}}{\rho}, \quad i = 0,1,\ldots,N,$$

*in which case the social welfare–maximizing prices are*

$$K_i = \sum_{j=i}^{N-1}\frac{b(1-\rho)^{1/N} - b(1-\rho) + a}{\rho}\left[\frac{\lambda\mathbb{E}[B^2]}{2(1-\rho)^{(2j+1)/N}} - \frac{\lambda\mathbb{E}[B^2]}{2(1-\rho)^{(2j-1)/N}}\right], \quad i = 0,1,\ldots,N.$$

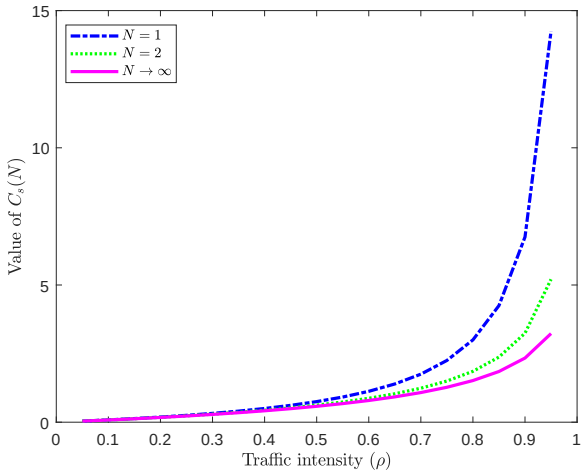
*We can also calculate the optimal total delay cost as*

$$\frac{\lambda\mathbb{E}[B^2]}{2\rho}\left[\frac{b(\rho-1)+a}{1-\rho} + \frac{b-a}{2\rho}N((1-\rho)^{-1/N} - (1-\rho)^{1/N})\right].$$

It follows from Propositions 5 and 6 that the revenue-maximizing prices are equal to the social welfare–maximizing prices when the delay cost follows a uniform distribution $U(a,b)$ with $a = 0$; this result is in line with that of Gilland and Warsing (2009, Prop. 3). By Remark 8, however, these optimal prices will *not* be equal if $a > 0$.

We conclude this discussion by using Proposition 6's results to assess how the *number* of priority classes affects the total customers' delay cost. The same parameter settings are used here as in section 7.5 (viz., $\mathbb{E}[B] = 1$ and $\mathbb{E}[B^2] = 2$). Since $\rho = \lambda\mathbb{E}[B]$, it follows that the optimal total customers' delay cost under $N$ priority classes can be written as $C_s(N) = (b(\rho-1)+a)/(1-\rho) + (b-a)/2\rho \cdot N((1-\rho)^{-1/N} - (1-\rho)^{1/N})$. We can then easily obtain the three equalities $C_s(1) = -b + (2b - \rho(b-a))/2(1-\rho)$, $C_s(2) = -b + a/(1-\rho) + (b-a)/\sqrt{1-\rho}$, and $\lim_{N\to\infty} C_s(N) = -b + a/(1-\rho) - (b-a)\ln(1-\rho)/\rho$. Moreover, $C_s(N)$ is strictly decreasing in $N$. Figure 14 plots the value of $C_s(N)$ $(N = 1, 2, \infty)$, as $\rho$ varies, while using the values given in Table 1 for $a$ and $b$. This graph shows that increasing the number of priority classes leads to less total delay cost, especially when $\rho$ is close to 1. Yet when $\rho < 0.7$, there is negligible benefit to introducing more than two priority classes.

**Figure 14** **Value of $C_s(N)$ for $N = 1, 2, \infty$ as a Function of $\rho$**




### 8.4. Effect of Number of Priority Classes

We now use numerical examples to examine how the number $N$ of priority classes affects the system's revenue and total delay cost. Toward that end, we put $\rho = 0.9$ and use the same parameters given for the distribution families (uniform, power, exponential, and log-normal) discussed in section 7.5. (Results for $\rho = 0.8$ are presented in Appendix C.2.)

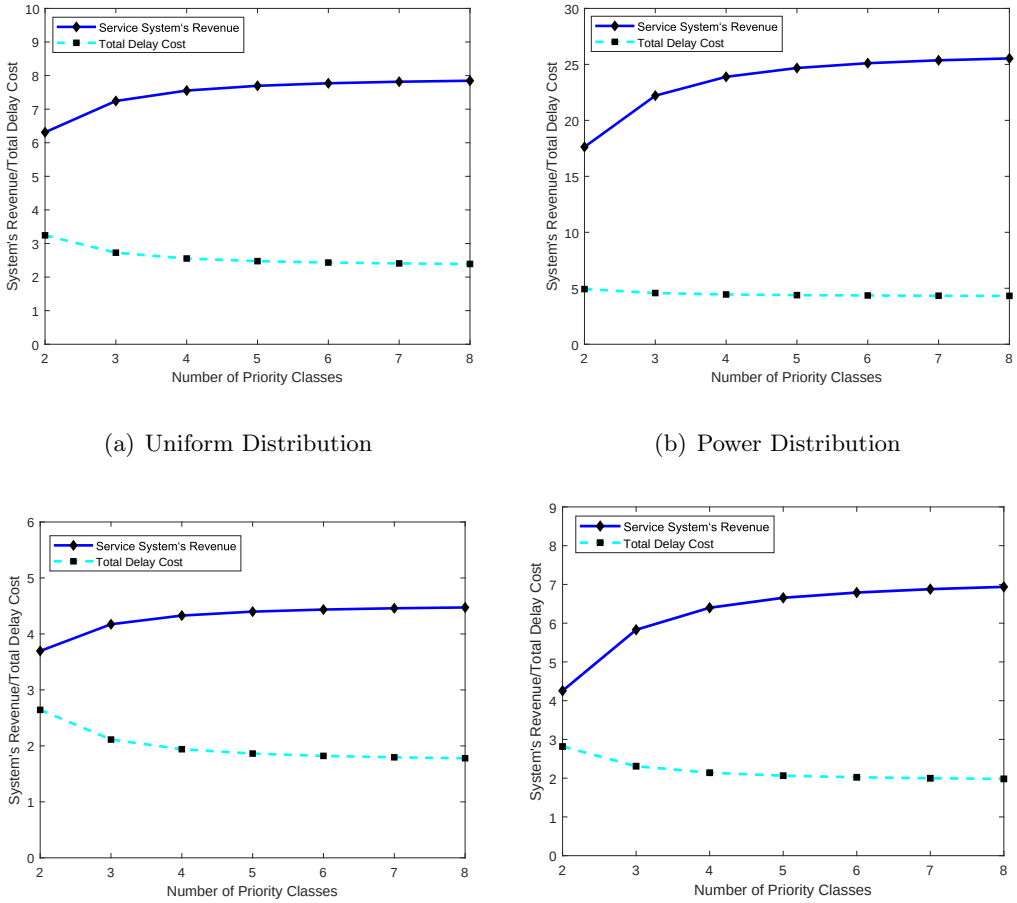
Figure 15 plots the service system's optimal revenue and total waiting cost with respect to the number of priority classes for each of the four distribution families we tested. By increasing the number of priority classes, the service provider can more effectively segment customers based on their delay cost rates—thereby exploiting greater price discrimination to increase revenue. At the same time, such finer segmentation allows each customer to be more closely targeted to their needs; hence the total delay cost declines (equivalently, social welfare improves) with a greater number of priority classes.

The curves in Figure 15 flatten out rather quickly, which indicates that most of the benefit the service system can gain from a complex pricing scheme (or an infinite number of priority classes in the extreme case) can be gained by offering a relatively small number of priority classes. In fact, it is seldom practical to offer either a continuum of prices or a large number of prices because doing so would unnecessarily complicate not only customer decisions but also system announcements of equilibrium delays.

## 9. Conclusion

The main takeaway from our study is that customers' delay cost distribution plays a key role in a service system's pricing strategies. Hence that distribution should be carefully estimated before a system manager makes any pricing decisions. Different delay cost distributions can lead to

**Figure 15 Maximal Revenue and Minimal Delay Cost with Respect to the Number of Priority Classes**




(a) Uniform Distribution

(b) Power Distribution

(c) Exponential Distribution

(d) Log-normal Distribution

completely different pricing strategies. In this paper, customers' choice behavior is discussed under a fixed price; we find that there might be multiple equilibria, the number of which depends on the structure of customer's delay cost distribution. In addition, we investigate how the customer response to a system's delay announcement evolves and then identify a condition under which these delay equilibria are stable. Our study of optimal pricing strategies—those that maximize the service system's long-run average revenue or social welfare—reveals that these optimal prices are highly sensitive to the delay cost distribution. Finally, we extend our results of two priority classes to multiple classes and investigate the implications of offering a small number of priority classes.

Because the distribution $H$ of customers' delay costs is seldom known in practice, an important research question remains: How can the optimal priority price be identified when that distribution is unknown? Given a delay announcement $(w_1, w_2)$ and a priority price $K$, one can use historical data (on customers' joining decisions) only to estimate the single quantile $H((K/(w_2 - w_1))$. Determining the entire distribution $H(\cdot)$ requires that the system manager vary $w_1$, $w_2$, and $K$—but while

bearing in mind two effects. First, announced delays $w_1$ and $w_2$ should be consistent with the actual delay. If the system manager varies $(w_1, w_2)$ frequently, customers will be aware of that and will no longer trust the system, which will hinder the system's efficient operations. Second, varying price of $K$ will incur a revenue loss. Therefore, the priority price and corresponding delay should be varied in a smart way so that the loss of revenue and goodwill is minimized. It is worth mentioning that in the absence of customers' choice equilibrium behavior, this is, in fact, an optimal pricing problem without knowing the demand function. This type of problem has been studied in the dynamic pricing literature, by using online learning method (see Besbes and Zeevi (2009, 2015) and the references therein). However, their results cannot directly apply to our problem due to a different problem setting. Hence, a dedicated analysis is required if $H$ is unknown.

We conclude by identifying several directions for future research. First, as just mentioned, an online learning method could be used to achieve better knowledge of delay cost distributions. Second, we have assumed that a delay's cost is (stochastically) linear in its delay. Yet the delay cost may, in practice, have a more complicated form—such as those exhibited by the convex–concave delay costs examined by Ata and Olsen (2009) and Akan et al. (2012). Finally, additional insights could follow from considering other delay announcement schemes, such as the state-dependent announcement studied by Armony and Maglaras (2004b).

## References


Adan, I. and J. Resing (2002). *Queueing Theory*. Eindhoven University of Technology.

Afèche, P. (2013). Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing & Service Operations Management 15*(3), 423–443.

Akan, M., B. Ata, and T. Olsen (2012). Congestion-based lead-time quotation for heterogenous customers with convex-concave delay costs: Optimality of a cost-balancing policy based on convex hull functions. *Operations Research 60*(6), 1505–1519.

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

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

Anand, K. S., M. F. Paç, and S. Veeraraghavan (2011). Quality-speed conundrum: Trade-offs in customer-intensive services. *Management Science 57*(1), 40–56.

Armony, M. and C. Maglaras (2004a). On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research 52*(2), 271–292.

Armony, M. and C. Maglaras (2004b). Contact centers with a call-back option and real-time delay information. *Operations Research 52*(4), 527–545.

Armony, M., N. Shimkin, and W. Whitt (2009). The impact of delay announcements in many-server queues with abandonment. *Operations Research 57*(1), 66–81.

Ata, B. and T. L. Olsen (2009). Near-optimal dynamic lead-time quotation and scheduling under convex-concave customer delay costs. *Operations Research 57*(3), 753–768.

Ata, B. and X. Peng (2018). An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. *Operations Research 66*(1), 163–183.

Besbes, O. and A. Zeevi (2009). Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research 57*(6), 1407–1420.

Besbes, O. and A. Zeevi (2015). On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science 61*(4), 723–739.

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

Debo, L. G., L. B. Toktay, and L. N. Van Wassenhove (2008). Queuing for expert services. *Management Science 54*(8), 1497–1512.

Gautam, N. (2012). *Analysis of Queues: Methods and Applications*. CRC Press.

Gavirneni, S. and V. G. Kulkarni (2016). Self-selecting priority queues with Burr distributed waiting costs. *Production and Operations Management 25*(6), 979–992.

Gilland, W. G. and D. P. Warsing (2009). The impact of revenue-maximizing priority pricing on customer delay costs. *Decision Sciences 40*(1), 89–120.

Guo, P. and P. Zipkin (2007). Analysis and comparison of queues with different levels of delay information. *Management Science 53*(6), 962–970.

Ha, A. Y. (2001). Optimal pricing that coordinates queues with customer-chosen service requirements. *Management Science 47*(7), 915–930.

Hassin, R. (1986). Consumer information in markets with random product quality: The case of queues and balking. *Econometrica*, 1185–1195.

Hassin, R. (1995). Decentralized regulation of a queue. *Management Science 41*(1), 163–173.

Hassin, R. (2016). *Rational Queueing*. CRC press.

Hassin, R. and M. Haviv (2003). *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, Volume 59. Springer.

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

Ibrahim, R., M. Armony, and A. Bassamboo (2017). Does the past predict the future? The case of delay announcement in service systems. *Management Science 63*(6), 1762–1780.

Kittsteiner, T. and B. Moldovanu (2005). Priority auctions and queue disciplines that depend on processing time. *Management Science 51*(2), 236–248.

Lu, Y., A. Musalem, M. Olivares, and A. Schilkrut (2013). Measuring the effect of queues on customer purchases. *Management Science 59*(8), 1743–1763.

Mendelson, H. (1985). Pricing computer services: Queueing effects. *Communications of the ACM 28*(3), 312–321.

Mendelson, H. and S. Whang (1990). Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research 38*(5), 870–883.

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

Nazerzadeh, H. and R. S. Randhawa (2018). Near-optimality of coarse service grades for customer differentiation in queueing systems. *Production and Operations Management 27*(3), 578–595.

Plambeck, E. L. (2004). Optimal leadtime differentiation via diffusion approximations. *Operations Research 52*(2), 213–228.

Reinertsen, D. G. (2009). *The Principles of Product Development Flow: Second Generation Lean Product Development*, Volume 1. Celeritas Publishing.

Whitt, W. (1999). Improving service by informing customers about anticipated delays. *Management Science 45*(2), 192–207.

Yu, D. Z., X. Zhao, and D. Sun (2013). Optimal pricing and capacity investment for delay-sensitive demand. *IEEE Transactions on Engineering Management 60*(1), 124–136.

Zhang, Z., D. Dey, and Y. Tan (2007). Pricing communication services with delay guarantee. *INFORMS Journal on Computing 19*(2), 248–260.

Zhou, W., X. Chao, and X. Gong (2014). Optimal uniform pricing strategy of a service firm when facing two classes of customers. *Production and Operations Management 23*(4), 676–688.

# APPENDIX

The appendix consists of three sections. Appendix A provides the proofs of main results, Appendix B gives some auxiliary results and Appendix C presents supporting materials for numerical studies.

## A. Proofs of Main Results

*Proof of Theorem 1.* Suppose that $(\tilde{w}_1, \tilde{w}_2)$ is an equilibrium delay. Then, it follows from Definition 1, (3) and (4) that

$$\tilde{w}_1 = \frac{\lambda \mathbb{E}[B^2]}{2(1-p(\tilde{w}_1,\tilde{w}_2)\rho)}, \ \tilde{w}_2 = \frac{\lambda \mathbb{E}[B^2]}{2(1-p(\tilde{w}_1,\tilde{w}_2)\rho)(1-\rho)}.$$

Hence, we have

$$\tilde{w}_1 = (1-\rho)\tilde{w}_2. \tag{30}$$

It follows from (1), (2), and (30) that

$$\begin{aligned}\tilde{w}_2 &= \frac{\lambda \mathbb{E}[B^2]}{2(1-p(\tilde{w}_1,\tilde{w}_2)\rho)(1-\rho)} = \frac{\mathbb{E}[W]}{1-p(\tilde{w}_1,\tilde{w}_2)\rho} \\ &= \frac{\mathbb{E}[W]}{1-\left(1-H\left(\frac{K}{\tilde{w}_2-\tilde{w}_1}\right)\right)\rho} = \frac{\mathbb{E}[W]}{1-\left(1-H\left(\frac{K}{\rho\tilde{w}_2}\right)\right)\rho} = \frac{\mathbb{E}[W]}{1-\rho+\rho H\left(\frac{K}{\rho\tilde{w}_2}\right)}.\end{aligned}$$

In view of (6), the equilibrium delay satisfies $F_e(\tilde{w}_2) = \mathbb{E}[W]$.

It follows from Lemma 6 in Appendix B that there exists a number $\tilde{w}_2' > 0$ such that $F_e(\tilde{w}_2') = \mathbb{E}[W]$. Let $\tilde{w}_1' = (1-\rho)\tilde{w}_2'$. It is straightforward to verify that $(\tilde{w}_1', \tilde{w}_2')$ is an equilibrium delay. Thus, there exists at least one equilibrium delay. □

*Proof of Theorem 2.* As discussed in section 4, equilibrium delays have a one-to-one map to the roots of $F_e(x) = \mathbb{E}[W]$. Using the chain rule, we have

$$\begin{aligned}F_e'(x) &= 1-\rho+\rho H\left(\frac{K}{\rho x}\right) + \rho x \cdot h\left(\frac{K}{\rho x}\right)\cdot\left(-\frac{K}{\rho x^2}\right) \\ &= 1-\rho+\rho H\left(\frac{K}{\rho x}\right) - \frac{K}{x}\cdot h\left(\frac{K}{\rho x}\right).\end{aligned} \tag{31}$$

We claim that

(*a*) If $h(x)$ is decreasing in $x$, then $F_e'(x)$ is decreasing in $x$ and the equilibrium delay is unique.

(*b*) If $h(x)$ is quasi-concave in $x$, then $F_e'(x)$ is quasi-convex in $x$. In this case, there might be one, two, or three equilibrium delays.

*Proof of Claim* (*a*). Let $x_1 < x_2$. It follows from (31) that

$$\begin{aligned}F_e'(x_2) - F_e'(x_1) &= \rho H\left(\frac{K}{\rho x_2}\right) - \frac{K}{x_2}\cdot h\left(\frac{K}{\rho x_2}\right) - \rho H\left(\frac{K}{\rho x_1}\right) + \frac{K}{x_1}\cdot h\left(\frac{K}{\rho x_1}\right) \\ &= -\rho\int_{\frac{K}{\rho x_2}}^{\frac{K}{\rho x_1}} h(y)dy - \frac{K}{x_2}\cdot h\left(\frac{K}{\rho x_2}\right) + \frac{K}{x_1}\cdot h\left(\frac{K}{\rho x_1}\right) \\ &\le -\rho\left(\frac{K}{\rho x_1} - \frac{K}{\rho x_2}\right) h\left(\frac{K}{\rho x_1}\right) - \frac{K}{x_2}\cdot h\left(\frac{K}{\rho x_1}\right) + \frac{K}{x_1}\cdot h\left(\frac{K}{\rho x_1}\right) \\ &= 0,\end{aligned}$$

where the second equality holds as $H(x)-H(y)=\int_y^x h(z)dz$ for any $x,y$, and the first inequality follows since $h(x)$ is decreasing in $x$. Hence, $F_e'(x)$ is decreasing in $x$.

It follows from (31) that $F_e'(x)\to 1-\rho>0$ as $x\to\infty$. Hence, $F_e'(x)>0$ for all $x>0$, i.e., $F_e(x)$ is increasing in $x$ and thus the equilibrium delay is unique.

*Proof of Claim* $(b)$. Since $h(x)$ is quasi-concave in $x$, there exists a changeover point $x_0\geq 0$ such that $h(x)$ is increasing in $x$ for $x\leq x_0$ and decreasing in $x$ for $x\geq x_0$. Using a similar argument as in the proof of Claim $(a)$, we can show that $F_e'(x)$ is decreasing in $x$ for $x\leq K/(\rho x_0)$, and increasing in $x$ for $x\geq K/(\rho x_0)$. Hence, $F_e'(x)$ is quasi-convex in $x$.

If $F_e'(K/(\rho x_0))>0$, then $F_e'(x)>0$ for all $x>0$. Hence, there exists a unique solution to $F_e(x)=\mathbb{E}[W]$, and thus the equilibrium delay is unique.

Next, we consider the case of $F_e'(K/(\rho x_0))<0$. Note that $F_e'(0^+):=\lim_{x\downarrow 0}F_e'(x)>0$ (see Lemma 7 in Appendix B), $F_e'(x)\to 1-\rho>0$ as $x\to\infty$ . Due to the quasi-convexity of $F_e'(x)$ in $x$, there exist two numbers $x_1\in(0,K/(\rho x_0)]$ and $x_2\in[K/(\rho x_0),\infty)$ such that $F_e'(x)>0$ if $x\in(0,x_1)$, $F_e'(x)\leq 0$ if $x\in[x_1,x_2]$ and $F_e'(x)>0$ if $x\in(x_2,\infty)$, i.e., $F_e(x)$ is increasing if $x\in(0,x_1)$, decreasing if $x\in[x_1,x_2]$, and increasing if $x\in(x_2,\infty)$. See Figure 2(b) for an illustration. Hence, if $\mathbb{E}[W]$ is larger than $F_e(x_1)$ or less than $F_e(x_2)$, there exists a unique equilibrium delay; if $\mathbb{E}[W]$ is equal to $F_e(x_1)$ or $F_e(x_2)$, there exist two equilibrium delays; and if $\mathbb{E}[W]$ is larger than $F_e(x_2)$ and less than $F_e(x_1)$, there exist three equilibrium delays. □

*Proof of Proposition 1.* For any $t\geq 1$, we have $w_1(t)=(1-\rho)w_2(t)$ by virtue of (9) and (10). Hence, $p(t)=\bar{H}(K/(\rho w_2(t)))$. It holds that

$$w_2(t+1)=\frac{\lambda\mathbb{E}[B^2]}{2(1-p(t)\rho)(1-\rho)}=\frac{\mathbb{E}[W]}{1-\rho+\rho H\left(\frac{K}{\rho w_2(t)}\right)}. \tag{32}$$

Define

$$d(w):=\frac{\mathbb{E}[W]}{1-\rho+\rho H\left(\frac{K}{\rho w}\right)}.$$

Obviously, $d(w)$ is increasing in $w$. It follows from (32) that

$$w_2(t+1)=d(w_2(t)),\ \text{for all } t\geq 1. \tag{33}$$

Moreover, $\tilde{w}_2=d(\tilde{w}_2)$ for any equilibrium delay $(\tilde{w}_1,\tilde{w}_2)$.

First, we claim that under any initial delay announcement $(w_1(0),w_2(0))\in\mathbb{R}_+^2$,

$$(w_1(t),w_2(t))\to(\tilde{w}_1,\tilde{w}_2)\ \text{as } t\to\infty \text{ for some equilibrium delay } (\tilde{w}_1,\tilde{w}_2). \tag{34}$$

If $w_2(0)\leq w_2(1)$, then $w_2(1)=d(w_2(0))\leq d(w_2(1))=w_2(2)$. By a straightforward induction on $t$, we know that $w_2(t)$ is increasing in $t$. Moreover, it follows from (32) that $w_2(t)\leq\mathbb{E}[W]/(1-\rho)$

for $t \geq 1$. Hence, as $t \to \infty$, $w_2(t) \to w_2^*$ for some $w_2^*$ and thus $w_1(t) = (1-\rho)w_2(t) \to (1-\rho)w_2^*$. Letting $t \to \infty$ in (33) yields $w_2^* = d(w_2^*)$. Hence, $((1-\rho)w_2^*, w_2^*)$ is an equilibrium delay. Similarly, if $w_2(1) < w_2(0)$, then $w_2(2) = d(w_2(1)) \leq d(w_2(0)) = w_2(1)$, which suggests that $w_2(t)$ is decreasing in $t$. Obviously, $w_2(t) \geq \mathbb{E}[W]$ in view of (32). Hence, $w_2(t) \to w_2^*$ for some $w_2^*$. Letting $t \to \infty$ in (33), we conclude that $((1-\rho)w_2^*, w_2^*)$ is also an equilibrium delay. Therefore, (34) holds in either case.

Next, we prove the claimed properties (i)–(iii) separately.

(i) If $w_2(0) - w_1(0) \geq \rho\tilde{w}_{2,I(K)}$, then it follows from (8) and (10) that

$$w_2(1) = \frac{\mathbb{E}[W]}{1-p(0)\rho} = \frac{\mathbb{E}[W]}{1-\rho+\rho H\left(\frac{K}{w_2(0)-w_1(0)}\right)} \geq \frac{\mathbb{E}[W]}{1-\rho+\rho H\left(\frac{K}{\rho\tilde{w}_{2,I(K)}}\right)} = \tilde{w}_{2,I(K)},$$

where the last equality holds since $F_e(\tilde{w}_{2,I(K)}) = \mathbb{E}[W]$. Then $w_2(2) = d(w_2(1)) \geq d(\tilde{w}_{2,I(K)}) = \tilde{w}_{2,I(K)}$, which implies $w_2(t) \geq \tilde{w}_{2,I(K)}$ for all $t \geq 1$. It follows from (34) that $(w_1(t), w_2(t))$ converges to an equilibrium delay. Since $w_2(t) \geq \tilde{w}_{2,I(K)}$ for all $t \geq 1$ and $(\tilde{w}_{1,I(K)}, \tilde{w}_{2,I(K)})$ is the largest equilibrium delay, it must hold that $(w_1(t), w_2(t)) \to (\tilde{w}_{1,I(K)}, \tilde{w}_{2,I(K)})$ as $t \to \infty$.

The second claim in (i) follows immediately from Lemma 2.

The proof of (ii) is quite similar to that of (i), and thus omitted for brevity.

(iii) If $\rho\tilde{w}_{2,i} < w_2(0) - w_1(0) \leq \rho\tilde{w}_{2,i+1}$, then we have $\tilde{w}_{2,i} < w_2(1) \leq \tilde{w}_{2,i+1}$ according to (8) and (10). With a similar argument as that of (i), we can show that $\tilde{w}_{2,i} < w_2(t) \leq \tilde{w}_{2,i+1}$.

Note that it follows from (8), (10) and (32) that

$$w_2(1) = \frac{\mathbb{E}[W]}{1-\rho+\rho H\left(\frac{K}{w_2(0)-w_1(0)}\right)},$$
$$w_2(2) = \frac{\mathbb{E}[W]}{1-\rho+\rho H\left(\frac{K}{\rho w_2(1)}\right)}.$$

Hence, $w_2(2) < w_2(1)$ is equivalent to $\rho w_2(1) < w_2(0) - w_1(0)$, i.e., $\rho\mathbb{E}[W] < [w_2(0) - w_1(0)][1-\rho+\rho H(\frac{K}{w_2(0)-w_1(0)})]$. Therefore, if $\rho\mathbb{E}[W] < [w_2(0) - w_1(0)][1-\rho+\rho H(\frac{K}{w_2(0)-w_1(0)})]$, then $w_2(2) < w_2(1)$. As in the proof of (34), $w_2(t)$ will be decreasing in $t$ and thus $(w_1(t), w_2(t)) \to (\tilde{w}_{1,i}, \tilde{w}_{2,i})$ as $t \to \infty$. If $\rho\mathbb{E}[W] \geq [w_2(0) - w_1(0)][1-\rho+\rho H(\frac{K}{w_2(0)-w_1(0)})]$, then $w_2(2) \geq w_2(1)$. In this case, $w_2(t)$ will be increasing in $t$ and thus $(w_1(t), w_2(t)) \to (\tilde{w}_{1,i+1}, \tilde{w}_{2,i+1})$ as $t \to \infty$. □

*Proof of Theorem 3.* We perform a perturbation analysis by considering the effect of a small deviation from the equilibrium delay $(\tilde{w}_{1,i}, \tilde{w}_{2,i})$. Suppose that the announced delay is now $(\tilde{w}_1', \tilde{w}_2')$ with $\tilde{w}_2' = \tilde{w}_{2,i} + \epsilon$, where $|\epsilon|$ is sufficiently small.

First, we consider the case in which $F_e(\cdot)$ is strictly increasing at $\tilde{w}_{2,i}$. If $\epsilon > 0$, then $F_e(\tilde{w}_2') > F_e(\tilde{w}_{2,i}) = \mathbb{E}[W]$. It follows from Proposition 1 (i) and (iii) that the announced delay will converge

to $(\tilde{w}_{1,i}, \tilde{w}_{2,i})$. If $\epsilon < 0$, then $F_e(\tilde{w}_2') < F_e(\tilde{w}_{2,i}) = \mathbb{E}[W]$. It follows from Proposition 1 (ii) and (iii) that the announced delay will converge to $(\tilde{w}_{1,i}, \tilde{w}_{2,i})$. Hence, in this case, the equilibrium delay $(\tilde{w}_{1,i}, \tilde{w}_{2,i})$ is stable.

Next, we consider the case in which $F_e(\cdot)$ is not strictly increasing at $\tilde{w}_{2,i}$. We distinguish three subcases: (i) $F_e(\cdot)$ is strictly decreasing at $\tilde{w}_{2,i}$; (ii) $F_e(\cdot)$ has a local minimizer at $\tilde{w}_{2,i}$; and (iii) $F_e(\cdot)$ has a local maximizer at $\tilde{w}_{2,i}$.

*Subcase 1: $F_e(\cdot)$ is strictly decreasing at $\tilde{w}_{2,i}$.* If $\epsilon > 0$, then $F_e(\tilde{w}_2') < \mathbb{E}[W]$. It follows from Proposition 1 (iii) that the announced delay will converge to $(\tilde{w}_{1,i+1}, \tilde{w}_{2,i+1}) \neq (\tilde{w}_{1,i}, \tilde{w}_{2,i})$. If $\epsilon < 0$, then $F_e(\tilde{w}_2') > \mathbb{E}[W]$. It follows from Proposition 1 (iii) that the announced delay will converge to $(\tilde{w}_{1,i-1}, \tilde{w}_{2,i-1})$.

*Subcase 2: $F_e(\cdot)$ has a local minimizer at $\tilde{w}_{2,i}$.* If $\epsilon > 0$, then $F_e(\tilde{w}_2') > \mathbb{E}[W]$. It follows from Proposition 1 (iii) that the announced delay will converge to $(\tilde{w}_{1,i}, \tilde{w}_{2,i})$. If $\epsilon < 0$, then $F_e(\tilde{w}_2') > \mathbb{E}[W]$. It follows from Proposition 1 (iii) that the announced delay will converge to $(\tilde{w}_{1,i-1}, \tilde{w}_{2,i-1})$.

*Subcase 3: $F_e(\cdot)$ has a local maximizer at $\tilde{w}_{2,i}$.* If $\epsilon > 0$, then $F_e(\tilde{w}_2') < \mathbb{E}[W]$. It follows from Proposition 1 (iii) that the announced delay will converge to $(\tilde{w}_{1,i+1}, \tilde{w}_{2,i+1})$. If $\epsilon < 0$, then $F_e(\tilde{w}_2') < \mathbb{E}[W]$. It follows from Proposition 1 (iii) that the announced delay will converge to $(\tilde{w}_{1,i}, \tilde{w}_{2,i})$.

In either of the above three cases, the equilibrium delay $(\tilde{w}_{1,i}, \tilde{w}_{2,i})$ is unstable. □

*Proof of Lemma 3.* Choose any initial delay announcement $(w_1(0), w_2(0)) \in \mathbb{R}_+^2$ such that $w_2(0) - w_1(0) \geq \rho\mathbb{E}[W]/(1-\rho)$. Then, it follows from Proposition 1 that $(w_1(t), w_2(t)) \to (\tilde{w}_{1,I(K)}, \tilde{w}_{2,I(K)})$ and thus $p(t;K) \to p^*(K)$ as $t \to \infty$ (here we write $p(t;K)$ instead of $p(t)$ to demonstrate the dependence of $p(t)$ on $K$). Moreover, we have

$$p(t+1;K) = 1 - H\left(\frac{K}{\rho w_2(t+1)}\right) = 1 - H\left(\frac{K(1-p(t;K)\rho)}{\rho\mathbb{E}[W]}\right), \tag{35}$$

where the second equality uses (10). Note that $p(0;K) = 1 - H(K/(w_2(0) - w_1(0)))$ is decreasing in $K$. By induction on $t$ and (35), it is straightforward to show that $\{p(t;K); t \in \mathbb{N}\}$ is a monotonically decreasing sequence such that $p(t;K)$ is continuous and decreasing in $K$ for all $t \geq 0$. Hence, it follows from Dini's theorem and uniform limit theorem that $p(t;K) \to p^*(K)$ uniformly as $t \to \infty$, and $p^*(K)$ is continuous and decreasing in $K$. Moreover, letting $t \to \infty$ in (8) yields (11).

We show that $p^*(K)$ cannot be a constant in any nonempty interval. If it fails to hold, there exist numbers $K_1 < K_2$ and $p \in [0,1]$ such that $p^*(K) = p$ for all $K \in [K_1, K_2]$. It follows from (11) that $p = 1 - H(K(1 - p \cdot \rho)/(\rho\mathbb{E}[W]))$ and thus $H(x) = 1 - p$ for all $x \in [K_1(1 - p \cdot \rho)/(\rho\mathbb{E}[W]), K_2(1 - p \cdot \rho)/(\rho\mathbb{E}[W])]$, which contradicts with the assumption that $H$ has support $[0, +\infty)$. Hence, $p^*(K)$ cannot be a constant in any nonempty interval and thus $p^*(K)$ is strictly decreasing in $K$. Using the same argument, one can show Remark 6. □

*Proof of Theorem 4.* Letting $p^*(K) = p$ and noting that $p^*(K)$ can take any value in $(0,1]$, the problem $\max_{K\geq 0} r(K)$ is equivalent to $\max_{p\in(0,1]} pK(p)$, i.e., $\max_{p\in(0,1]} M(p)\cdot\rho\mathbb{E}[W]$.

If $M(p)$ is quasi-concave in $p\in(0,1]$ with its maximizer $p^*$, then $r(K)$ attains its maximum at $K^* = K(p^*)$.

If $h$ has support $[0,\infty)$, then for each $K$, $K(p^*(K)) = K$. It follows from Lemma 3 that $r(K)$ is strictly quasi-concave in $K>0$ and has a unique maximizer $K^* = K(p^*)$.

Note that under Assumption 1, the support of $h$ takes the form $[a,b]$, where $a\geq 0$ and $b\leq\infty$. It follows from Lemma 3 and Remark 6 that $r^*(K)$ takes value $\lambda K$ if $K\leq(a\rho\mathbb{E}[W])/(1-\rho)$, and is strictly quasi-concave in $K$ if $(a\rho\mathbb{E}[W])/(1-\rho) < K < b\rho\mathbb{E}[W]$, and takes value 0 if $K\geq b\rho\mathbb{E}[W]$. Hence, $r(K)$ is quasi-concave in $K>0$, and has a unique maximizer. □

*Proof of Proposition 2.* We prove each item one at a time.

(i) By simple algebraic manipulation, we have $M(p) = pa(p^{-1/k}-1)^{1/d}/(1-p\cdot\rho)$ for $p\in(0,1)$. Hence, its first-order derivative is given by

$$\frac{a(p^{-1/k}-1)^{1/d-1}(1-p^{1/k}+p\rho/kd-1/kd)}{(1-p\rho)^2\cdot p^{1/k}}.$$

Let $g(p) := 1-p^{1/k}+p\rho/kd-1/kd$. Then we have $g(0) = 1-1/kd>0$ and $g(1) = 1/kd\cdot(\rho-1)<0$. Note that $g''(p) = (k-1)p^{1/k-2}/k^2$. Hence, if $k\geq 1$, then $g(p)$ is convex in $p\in(0,1)$ and if $k<1$, then $g(p)$ is concave in $p\in(0,1)$. By considering $k\geq 1$ and $0<k<1$ separately, one can see that $g(p)=0$ has exactly one root $p^*\in(0,1)$ such that $g(p)>0$ for $p\in(0,p^*)$ and $g(p)<0$ for $p\in(p^*,1)$, which implies that $M(p)$ is strictly quasi-concave in $p\in(0,1)$.

(ii) A Weibull distribution function takes the form $H(x;\tau,\theta) = 1-e^{-(x/\tau)^\theta}$, $x\geq 0$. In this case, $M(p) = p\tau(-\ln p)^{1/\theta}/(1-p\cdot\rho)$ for $p\in(0,1)$, and thus its first-order derivative is

$$\frac{(-\ln p)^{1/\theta-1}\tau(\rho p-\theta\ln p-1)}{\theta(1-p\rho)^2}. \tag{36}$$

Let $g(p) := \rho p-\theta\ln p-1$. Then, $g(0^+) := \lim_{x\downarrow 0} g(x) = \infty$ and $g(1) = \rho-1<0$. Moreover, $g'(p) = \rho-\theta/p$. Hence $g(p)$ is first strictly decreasing in $p$ from 0 to $\theta/\rho$, and then strictly increasing from $\theta/\rho$ to $\infty$. Hence, $g(p)=0$ has exactly one root in $(0,1]$, which implies that $M(p)$ is strictly quasi-concave in $p\in(0,1]$.

(iii) In this case, $M(p) = p(b-(b-a)p)/(1-p\cdot\rho)$ for $p\in(0,1)$. Its first-order derivative is

$$(b-2(b-a)p+\rho(b-a)p^2)/(1-p\cdot\rho)^2. \tag{37}$$

Let $g(p) := b-2(b-a)p+\rho(b-a)p^2$. Then $g'(p) = -2(b-a)(1-\rho p)<0$. Note that $g(0) = b>0$, $g(1) = 2a-b+\rho(b-a)$. Hence, $M(p)$ is strictly quasi-concave in $p\in(0,1]$ if and only if $\rho < (b-2a)/(b-a)$.

(iv) In this case, $M(p) = pH^{-1}(1-p)/(1-p\cdot\rho) = pe^{\mu+\sigma\Phi^{-1}(1-p)}/(1-p\cdot\rho)$, and thus its first-order derivative is

$$M'(p) = \frac{e^{\sigma\Phi^{-1}(1-p)+\mu}\left(1-\frac{p\sigma(1-p\rho)}{\phi(\Phi^{-1}(1-p))}\right)}{(1-p\cdot\rho)^2}, \tag{38}$$

where $\phi$ is the probability density function of the standard normal distribution.

Let $f(x) := \phi(x) - \sigma(1-\Phi(x))(1-\rho+\rho\Phi(x))$. We claim that

$$\text{There exists a unique number } x_0 \text{ such that } f(x)<0 \text{ for } x<x_0 \text{ and } f(x)>0 \text{ for } x>x_0. \tag{39}$$

If this is true, then it follows from (38) that $M'(p)<0$ for $p>1-\Phi(x_0)$ then and $M'(p)>0$ for $p<1-\Phi(x_0)$. Hence, $M(p)$ is quasi-concave with maximizer $1-\Phi(x_0)$.

It remains to prove (39). First, we have

$$f(-\infty) := \lim_{x\to-\infty} f(x) = -\sigma(1-\rho) < 0,\ f(\infty) := \lim_{x\to\infty} f(x) = 0. \tag{40}$$

Note that $f(x) = (1-\Phi(x))\cdot[\phi(x)/(1-\Phi(x)) - \sigma(1-\rho+\rho\Phi(x))] > (1-\Phi(x))\cdot[\phi(x)/(1-\Phi(x)) - \sigma]$. Hence, it follows from Lemma 10 that there exists a positive number $x_1$ such that

$$f(x) > 0 \text{ for } x \geq x_1. \tag{41}$$

It follows from the continuity of $f$ in $x$ that there exists a root $x_0$ to $f(x)=0$. If the root is unique, then (39) holds. Below we argue by contradictory to show the uniqueness of the root of $f$.

Suppose it fails to hold. It follows from the continuity of $f$, (40) and (41) that there exist numbers $x_2$, $x_3$, $x_4$ such that $x_2<x_3<x_4$, $f(x_2)=f(x_3)=f(x_4)=0$ and $f'(x_2)>0$, $f'(x_3)<0$ and $f'(x_4)>0$. By mean value theorem there exist numbers $y_1\in(x_2,x_3)$, $y_2\in(x_3,x_4)$ and $y_3\in(x_4,+\infty)$ such that $f'(y_1)=f'(y_2)=f'(y_3)=0$. Hence, $g(x) := x+\sigma(2\rho(1-\Phi(x))-1)=0$ has three roots $y_1$, $y_2$ and $y_3$. Since $g(-\infty)=-\infty$ and $g(\infty)=\infty$, there exist two numbers $z_1\in(y_1,y_2)$, $z_2\in(y_2,y_3)$ such that $g'(z_1)=g'(z_2)=0$. Note that $g'(x) = 1-2\sigma\rho\phi(x) \geq 1-2\sigma\rho/\sqrt{2\pi} > 0$. Hence, $g(x)$ is strictly increasing in $x$, which contradicts with the derived result that $g(x)=0$ has three district roots.

(v) It follows from the implicit function theorem that if $H(x)=p$, then $(H^{-1})'(p) = 1/H'(x) = 1/h(x) = 1/h(H^{-1}(p))$. Hence, we have

$$\begin{aligned} M'(p) &= \frac{H^{-1}(1-p) - p(1-p\cdot\rho)(H^{-1})'(1-p)}{(1-p\cdot\rho)^2} \\ &= \frac{H^{-1}(1-p) - p(1-p\cdot\rho)/h(H^{-1}(1-p))}{(1-p\cdot\rho)^2}. \end{aligned}$$

Letting $x = H^{-1}(1-p)$, we define $g(x) := x - (1-H(x)) \cdot (1-\rho+\rho H(x))/h(x)$. Then, $g(0) = -(1-\rho)/h(0) < 0$ and $g(x) \to \infty$ as $x \to \infty$. Moreover, it follows from the definition of $m(x)$ that $1 - H(x) = \exp(-\int_0^x m(y)dy)$. Hence, we have

$$g(x) = x - \frac{1-\rho e^{-\int_0^x m(y)dy}}{m(x)},\ g'(x) = (1-\rho e^{-\int_0^x m(y)dy}) \cdot \left(1 + \frac{m'(x)}{m^2(x)}\right) > 0.$$

Therefore, there exists a number $x^*$ such that $g(x) < 0$ for $x \in [0, x^*)$ and $g(x) > 0$ for $x \in (x^*, \infty)$. Let $p^* = 1 - H(x^*)$. Then, $M(p)$ is strictly increasing in $p \in [0, p^*)$ and strictly decreasing in $p \in (p^*, 1]$. □

*Proof of Lemma 4.* Note that

$$p(\tilde{w}_1, \tilde{w}_2; K) = 1 - H\left(\frac{K(1-p(\tilde{w}_1, \tilde{w}_2; K)\rho)}{\rho \mathbb{E}[W]}\right). \tag{42}$$

With a similar argument as in the proof of Lemma 3, we can show that $p(\tilde{w}_1, \tilde{w}_2; K)$ is also continuous in $K$. Since $1 \geq 1 - p(\tilde{w}_1, \tilde{w}_2; K)\rho \geq 1 - \rho$, we have $p(\tilde{w}_1, \tilde{w}_2; 0) = 1 - H(0) = 1$ and $\lim_{K\to\infty} p(\tilde{w}_1, \tilde{w}_2; K) = 0$ in view of (42). This implies that $p(\tilde{w}_1, \tilde{w}_2; K)$ can attain any value in $(0, 1]$ as $K$ varies from 0 to $\infty$. □

*Proof of Theorem 5.* Let $H^{-1}(1-p) = q$, then $p = 1 - H(q)$. Since $H$ is continuous, $q$ can take any value in $[0, \bar{q}]$, where $\bar{q} := H^{-1}(1)$ might be $\infty$. Hence, the optimization problem (16) is equivalent to

$$\min_{0\leq q\leq \bar{q}} F_S(q) := \frac{1}{1-\rho(1-H(q))}\left(\int_0^\infty c \cdot h(c)dc - \rho\int_q^\infty c \cdot h(c)dc\right).$$

We have

$$\begin{aligned} F_S'(q) &= \rho h(q) \cdot \frac{\left(q(1-\rho(1-H(q))) - \int_0^\infty c \cdot h(c)dc + \rho\int_q^\infty c \cdot h(c)dc\right)}{(1-\rho(1-H(q)))^2} \\ &= \rho h(q) \cdot \frac{g_s(q)}{(1-\rho(1-H(q)))^2}. \end{aligned} \tag{43}$$

Note that $g_S(0) = -(1-\rho)\int_0^\infty c \cdot h(c)dc < 0$ and

$$\begin{aligned} g_S(\bar{q}) &= \bar{q} - \int_0^\infty c \cdot h(c)dc + \rho\int_{\bar{q}}^\infty c \cdot h(c)dc = \bar{q} - \int_0^{\bar{q}} c \cdot h(c)dc \\ &= \bar{q}\int_0^{\bar{q}} h(c)dc - \int_0^{\bar{q}} c \cdot h(c)dc = \int_0^{\bar{q}} (\bar{q} - c)h(c)dc > 0, \end{aligned}$$

where the second equality follows from $\int_{\bar{q}}^\infty c \cdot h(c)dc = 0$. Moreover,

$$g_S'(q) = 1 - \rho(1-H(q)) \geq 1 - \rho > 0, \text{ for all } q \geq 0.$$

Hence, there exists a unique root $q_S^* \in (0, \bar{q})$ to $g_S(q) = 0$ in $[0, \infty)$, such that $g_S(q) < 0$ for $q \in (0, q_S^*)$, and $g_S(q) > 0$ for $q \in (q_S^*, \bar{q})$. In view of (43), we know that $F_S(q)$ is strictly decreasing in $q$ for $q \in (0, q_S^*)$, and strictly increasing in $q$ for $q \in (q_S^*, \bar{q})$, i.e., it has a unique minimizer $q_S^*$. Hence, (16) admits a unique minimizer $p_S^* = 1 - H(q_S^*) \in (0, 1)$. □

*Proof of Proposition 3.* It suffices to show that $F_e(\cdot)$ is strictly increasing at point $\tilde{w}_{2,S}^*$. We have

$$\begin{aligned} F_e'(\tilde{w}_{2,S}^*) &= 1 - \rho + \rho H\left(\frac{K_S^*}{\rho \tilde{w}_{2,S}^*}\right) - \frac{K_S^*}{\tilde{w}_{2,S}^*} \cdot h\left(\frac{K_S^*}{\rho \tilde{w}_{2,S}^*}\right) \\ &= 1 - \rho + \rho H(q_S^*) - \rho q_S^* \cdot h(q_S^*), \end{aligned}$$

where the second equality follows from $K_S^*/(\rho \tilde{w}_{2,S}^*) = q_S^*$ in view of (18). Note that

$$g_S(q_S^*) = q_S^*(1 - \rho(1 - H(q_S^*))) - \int_0^\infty c \cdot h(c)dc + \rho \int_{q_S^*}^\infty c \cdot h(c)dc = 0.$$

Hence, we obtain

$$\begin{aligned} F_e'(\tilde{w}_{2,S}^*) &= \frac{\int_0^\infty c \cdot h(c)dc - \rho \int_{q_S^*}^\infty c \cdot h(c)dc}{q_S^*} - \rho q_S^* \cdot h(q_S^*) \\ &> \frac{\rho \int_0^\infty c \cdot h(c)dc - \rho \int_{q_S^*}^\infty c \cdot h(c)dc - \rho (q_S^*)^2 \cdot h(q_S^*)}{q_S^*} \\ &= \frac{\rho \int_0^{q_S^*} c \cdot h(c)dc - \rho \int_0^{q_S^*} q_S^* \cdot h(q_S^*)dc}{q_S^*} \\ &= \frac{\rho \int_0^{q_S^*} (c \cdot h(c) - q_S^* \cdot h(q_S^*))dc}{q_S^*} \\ &\geq 0, \end{aligned}$$

where the last inequality follows from the condition that $c \cdot h(c)$ is decreasing in $c$. Hence, the equilibrium delay $(\tilde{w}_{1,S}^*, \tilde{w}_{2,S}^*)$ is stable. □

*Proof of Lemma 5.* We attach symbol $\tau$ to demonstrate the dependence of these functions on $\tau$. In this case, we have $M(p; \tau) = p\tau(-\ln p)^{1/\theta}/(1 - p \cdot \rho)$ for $p \in (0, 1)$. Hence, its first-order derivative with respect to $p$ is given by

$$\frac{(-\ln p)^{1/\theta - 1} \tau (\rho p - \theta \ln p - 1)}{\theta (1 - p\rho)^2}.$$

Therefore, $p^*(\tau)$ is the unique root in $(0, 1)$ to $\rho p - \theta \ln p - 1 = 0$ (see Lemma 8 (iii)), which is independent of $\tau^*$. The revenue-maximizing price is

$$K^*(\tau, \theta) = K(p^*; \tau) = \tau \cdot \frac{(-\ln p^*)^{1/\theta - 1} \rho \mathbb{E}[W]}{\theta},$$

where the second equality uses (12). Hence, $K^*(\tau,\theta)=\tau K^*(1,\theta)$.

It follows from (17) and Lemma 9 that $q_S^*(\tau)$ is the unique root in $(0,\infty)$ to

$$\int_0^\infty e^{-(\frac{c}{\tau})^\theta}dc-\rho\int_q^\infty e^{-(\frac{c}{\tau})^\theta}dc-q=0.$$

Letting $c/\tau=\tilde{c}$, the above equation can be written as

$$\int_0^\infty e^{-\tilde{c}^\theta}d\tilde{c}-\rho\int_{q/\tau}^\infty e^{-\tilde{c}^\theta}d\tilde{c}-\frac{q}{\tau}=0.$$

Hence, $q_S^*(\tau)/\tau$ is a positive root of

$$\int_0^\infty e^{-c^\theta}dc-\rho\int_q^\infty e^{-c^\theta}dc-q=0,$$

which is equal to $q_S^*(1)$ by Lemma 9. Hence, we have $q_S^*(\tau)=\tau q_S^*(1)$. The social-welfare-maximizing price is

$$K_S^*(\tau,\theta)=\frac{q_S^*(\tau)\rho\mathbb{E}[W]}{1-\rho e^{-(\frac{q_S^*(\tau)}{\tau})^\theta}}=\frac{\tau q_S^*(1)\rho\mathbb{E}[W]}{1-\rho e^{-(q_S^*(1))^\theta}}=\tau K_S^*(1,\theta).\quad\square$$

*Proof of Theorem 6.* Consider any equilibrium delay $\boldsymbol{c}$ (if it exists) with $c_1=x$ for some $x>0$. It follows from (25) with $i=1$ and $c_0=\infty$ that

$$\frac{\bar{H}(c_2)}{1-\rho\bar{H}(c_2)}=\frac{2(K_1-K_2)(1-\rho\bar{H}(x))}{\lambda\rho\mathbb{E}[B^2]x}.\tag{44}$$

Now we try to find a $c_2\in[0,x]$ to satisfy (44). Note that $\bar{H}(c_2)/(1-\rho\bar{H}(c_2))$ is decreasing in $c_2$ and $c_1=x\geq c_2\geq 0$. Hence, we have

$$\frac{\bar{H}(x)}{1-\rho\bar{H}(x)}\leq\frac{\bar{H}(c_2)}{1-\rho\bar{H}(c_2)}\leq\frac{1}{1-\rho}.$$

Define

$$\underline{x}_2:=\inf\left\{x>0:\frac{2(K_1-K_2)(1-\rho\bar{H}(x))}{\lambda\rho\mathbb{E}[B^2]x}<\frac{1}{1-\rho}\right\},$$
$$\bar{x}_2:=\inf\left\{x>0:\frac{2(K_1-K_2)(1-\rho\bar{H}(x))}{\lambda\rho\mathbb{E}[B^2]x}\leq\frac{\bar{H}(x)}{1-\rho\bar{H}(x)}\right\},$$

where we use the convention $\inf\emptyset=\infty$. Since $2(K_1-K_2)(1-\rho\bar{H}(x))/(\lambda\rho\mathbb{E}[B^2]x)\to\infty$ as $x\to 0^+$ and $2(K_1-K_2)(1-\rho\bar{H}(x))/(\lambda\rho\mathbb{E}[B^2]x)\to 0$ as $x\to\infty$, we know that $\underline{x}_2>0$ and is finite. By continuity of $H$, we have $\bar{x}_2>\underline{x}_2$. We mention that here $\bar{x}_2$ might be $\infty$. It follows from the continuity of $H$ that for any $x\in[\underline{x}_2,\bar{x}_2]$, there exists a solution, denoted by $s_2(x)$, to (44), i.e.,

$$\frac{\bar{H}(s_2(x))}{1-\rho\bar{H}(s_2(x))}=\frac{2(K_1-K_2)(1-\rho\bar{H}(x))}{\lambda\rho\mathbb{E}[B^2]x}.$$

Moreover, it follows from the continuity of $H$ and the definition of $\underline{x}_2$ and $\bar{x}_2$ that we can choose $s_2$ such that $s_2(x)$ is continuous in $x\in[\underline{x}_2,\bar{x}_2]$ with $s_2(\underline{x}_2)=0$ and $s_2(\bar{x}_2)=\bar{x}_2$.

Next, we find a $c_3$ such that (25) with $i=2$ is satisfied given that $c_1=x$ and $c_2=s_2(x)$. For $x\in[\underline{x}_2,\bar{x}_2]$, it follows from (25) with $i=2$ that

$$\frac{\bar{H}(c_3)-\bar{H}(x)}{1-\rho\bar{H}(c_3)}=\frac{2(K_2-K_3)(1-\rho\bar{H}(x))(1-\rho\bar{H}(s_2(x)))}{\lambda\rho\mathbb{E}[B^2]s_2(x)}. \tag{45}$$

We define

$$\underline{x}_3:=\inf\left\{x\in[\underline{x}_2,\bar{x}_2]:\frac{2(K_2-K_3)(1-\rho\bar{H}(x))(1-\rho\bar{H}(s_2(x)))}{\lambda\rho\mathbb{E}[B^2]s_2(x)}<\frac{1-\bar{H}(x)}{1-\rho}\right\},$$
$$\bar{x}_3:=\inf\left\{x\in[\underline{x}_2,\bar{x}_2]:\frac{2(K_2-K_3)(1-\rho\bar{H}(x))(1-\rho\bar{H}(s_2(x)))}{\lambda\rho\mathbb{E}[B^2]s_2(x)}\le\frac{\bar{H}(s_2(x))-\bar{H}(x)}{1-\rho\bar{H}(s_2(x))}\right\}\wedge\bar{x}_2,$$

where we use $a\wedge b$ to denote $\min(a,b)$.

Similarly, one can show that for any $x\in[\underline{x}_3,\bar{x}_3]$, there exists a solution, denoted by $s_3(x)$, to (45). Moreover, $s_3(x)$ is continuous in $x\in[\underline{x}_3,\bar{x}_3]$ with $s_3(\underline{x}_3)=0$. For the value of $s_3(\bar{x}_3)$, there are two separate cases: the set in the definition of $\bar{x}_3$ is nonempty or empty. If it is nonempty, we have $s_3(\bar{x}_3)=\bar{x}_3\le\bar{x}_2$. If it is empty, we have $\bar{x}_3=\bar{x}_2$, in which $s_3(\bar{x}_3)=s_3(\bar{x}_2)\le\bar{x}_2$.

Using the above procedure, we can show that for any $k=2,3,\ldots,N$, there exists a nonempty interval $[\underline{x}_k,\bar{x}_k]\subset[\underline{x}_{k-1},\bar{x}_{k-1}]$ such that if $x\in[\underline{x}_k,\bar{x}_k]$, there exists a solution, denoted by $s_k(x)$, to (25) with $i=k-1$. Moreover, $s_k(x)$ is continuous in $x\in[\underline{x}_k,\bar{x}_k]$ with $s_k(\underline{x}_k)=0$.

Since any equilibrium delay $\boldsymbol{c}$ satisfies $c_N=0$, $(\underline{x}_N,s_2(\underline{x}_N),\ldots,s_N(\underline{x}_N))$ is a desired equilibrium delay. □

Remark 7. Theorem 6 still holds even if $H$ is not continuous. Please refer to Theorem 1 for the case of two priority classes and the proof of it. However, the argument will be more involved. Hence, the assumption of continuous $H$ is made, for the sake of analytical simplicity.

*Proof of Proposition 5.* First, we consider a general case that $C\sim U(a,b)$, i.e., $H(x)=(x-a)/(b-a)$ for $x\in[a,b]$. The problem (28) becomes

$$\begin{aligned}\max_{\boldsymbol{d}}\quad&\frac{\lambda^2\mathbb{E}[B^2]\rho}{2}\sum_{i=1}^{N-1}[(b-a)(1-d_i)+a]d_i\frac{d_{i+1}-d_{i-1}}{(1-\rho d_{i-1})(1-\rho d_i)(1-\rho d_{i+1})}\\ \text{s.t.}\quad&d_0=0\le d_1\le\ldots\le d_{N-1}\le d_N=1.\end{aligned}$$

By letting $1-\rho d_i=e_i$, the above problem can be further simplified to

$$\begin{aligned}\max_{\boldsymbol{e}}\quad&\frac{\lambda^2\mathbb{E}[B^2]}{2\rho}\sum_{i=1}^{N-1}\left(\frac{2b-2a-b\rho}{\rho}+\frac{b\rho-b+a}{\rho e_i}-\frac{(b-a)e_i}{\rho}\right)\cdot\left(\frac{1}{e_{i+1}}-\frac{1}{e_{i-1}}\right)\\ \text{s.t.}\quad&e_0=1\ge e_1\ge\ldots\ge e_{N-1}\ge e_N=1-\rho.\end{aligned} \tag{46}$$

Fix $i\in\{1,2,\ldots,N-1\}$ and fix $e_j$ for all $j\neq i$. Then if $2\le i\le N-2$, then the problem (46) becomes

$$\begin{aligned}\min_{e_i}\quad&e_i\left(\frac{1}{e_{i+1}}-\frac{1}{e_{i-1}}\right)+\frac{e_{i-1}-e_{i+1}}{e_i}\\ \text{s.t.}\quad&e_{i-1}\ge e_i\ge e_{i+1},\end{aligned}$$

whose optimal solution is $\sqrt{e_{i-1}e_{i+1}}$.

For $i=1$, the problem (46) becomes

$$\begin{array}{ll}\min\limits_{e_1} & e_1\left(\dfrac{1}{e_2}-\dfrac{1}{e_0}\right)+\dfrac{\frac{b\rho-b+a}{(b-a)e_0}+\frac{2b-2a-b\rho}{b-a}-e_2}{e_1}\\ \text{s.t.} & e_0=1\geq e_1\geq e_2,\end{array}$$

whose optimal solution is $\sqrt{e_0e_2}$ by noting that $e_0=1$. Hence, the optimal solution $\boldsymbol{e}^*$ (if it exists) satisfies the following chain rule

$$\frac{e_{k+1}^*}{e_k^*}=\frac{e_k^*}{e_{k-1}^*},\quad k=1,2,\ldots,N-2.$$

Let $e_1^*=\xi$, then we have $e_k^*=\xi^k$ for $k=0,1,\ldots,N-1$.

For $i=N-1$, the problem (46) becomes

$$\begin{array}{ll}\min\limits_{e_{N-1}} & e_{N-1}\left(\dfrac{1}{e_N}-\dfrac{1}{e_{N-2}}\right)+\dfrac{e_{N-2}-\frac{b\rho-b+a}{(b-a)e_N}-\frac{2b-2a-b\rho}{b-a}}{e_{N-1}}\\ \text{s.t.} & e_{N-2}\geq e_{N-1}\geq e_N=1-\rho.\end{array}\tag{47}$$

Let $A(\rho):=(b\rho-b+a)/((b-a)(1-\rho))+(2b-2a-b\rho)/(b-a)$. It is straightforward to verify that $1-\rho\leq A(\rho)$.

Now we consider the case that $a=0$. We have $A(\rho)=1-\rho=e_N$, which implies $e_{N-1}^*=\sqrt{e_{N-2}^*e_N^*}$. Hence, we have $\xi=(1-\rho)^{1/N}$, $e_k^*=(1-\rho)^{k/N}$ and thus $d_k^*=(1-(1-\rho)^{k/N})/\rho$, for $k=0,1,\ldots,N$. Plugging these values into (46) yields

$$\frac{\lambda^2\mathbb{E}[B^2]b}{2\rho^2}\left[\frac{(2-\rho)\rho}{1-\rho}-\left[(1-\rho)^{-(N-1)/N}-(1-\rho)^{(N-1)/N}+(N-1)((1-\rho)^{-1/N}-(1-\rho)^{1/N})\right]\right],$$

which is the optimal revenue of the service system. The revenue-maximizing prices are obtained by using (22), (24), $w_i=\mathbb{E}[W_i]$ and $1-\rho\bar{H}(c_i)=e_i$ for $0\leq i\leq N$. □

Remark 8 (Solution to the Problem (28) when $a>0$). In view of (47), we need to consider the following two cases:

*Case 1:* $e_{N-2}^*<A(\rho)$. Then, $e_{N-1}^*=e_N^*=1-\rho$ and thus we have $e_k^*=(1-\rho)^{k/(N-1)}$ for $k=0,1,\cdots,N-2$. The condition $e_{N-2}^*<A(\rho)$ can be written as $(1-\rho)^{(N-2)/(N-1)}<A(\rho)$.

*Case 2:* $e_{N-2}^*\geq A(\rho)$. The unconstrained solution of (47) is given by

$$e_{N-1}^u=\sqrt{\frac{(e_{N-2}^*-A(\rho))(1-\rho)e_{N-2}^*}{e_{N-2}-(1-\rho)}}.$$

In this case, by considering the relationship between $e_{N-1}^u$, $e_{N-2}^*$ and $e_N^*$, we have the following three subcases.

*Subcase 1:* $e^u_{N-1} \le e^*_N$. Then we have $e^*_{N-1} = e^*_N = 1-\rho$ and thus $e^*_k = (1-\rho)^{k/(N-1)}$ for $k = 0, 1, \ldots, N-2$. The condition $e^u_{N-1} \le e^*_N$ can be written as $[(1-\rho)^{(N-2)/(N-1)} - A(\rho)](1-\rho)^{(N-2)/(N-1)} \le (1-\rho)[(1-\rho)^{(N-2)/(N-1)} - (1-\rho)]$.

*Subcase 2:* $e^u_{N-1} \ge e^*_{N-2}$. Then, we have $e^*_{N-1} = e^*_{N-2}$ and thus $\xi = 1$, which implies that $e^*_k = 1$ for $k = 0, 1, \ldots, N-1$. The condition $e^u_{N-1} \ge e^*_{N-2}$ then can be written as $(1 - A(\rho))(1-\rho) \ge \rho$, which can not hold as $A(\rho) \ge 1-\rho$ and $\rho > 0$.

*Subcase 3:* $e^u_{N-1} \in (e^*_N, e^*_{N-2})$. Then, we have $e^*_{N-1} = e^u_{N-1}$. Hence, we have $e^*_k = \xi^k$ for $k = 0, 1, \ldots, N-1$, where $\xi$ satisfies $(1-\rho)^{1/(N-1)} < \xi < 1$ and $\xi^{2N-2} - (1-\rho)\xi^N - (1-\rho)\xi^{N-2} + (1-\rho)A(\rho) = 0$.

By summarizing the above results, we have the following characterization of the optimal $\mathbf{e}^*$: (i) If $[(1-\rho)^{(N-2)/(N-1)} - A(\rho)](1-\rho)^{(N-2)/(N-1)} \le (1-\rho)[(1-\rho)^{(N-2)/(N-1)} - (1-\rho)]$, then $e^*_k = (1-\rho)^{k/(N-1)}$ for $k = 0, 1, \ldots, N-2$ and $e^*_{N-1} = 1-\rho$; (ii) Otherwise, $e^*_k = \xi^k$ for $k = 0, 1, \ldots, N-1$, where $\xi$ is the root to $x^{2N-2} - (1-\rho)x^N - (1-\rho)x^{N-2} + (1-\rho)A(\rho) = 0$ in $((1-\rho)^{1/(N-1)}, 1)$.

*Proof of Proposition 6.* The optimization problem (29) becomes

$$\begin{aligned} \min_{\boldsymbol{d}} \quad & \frac{\lambda \mathbb{E}[B^2]}{2\rho} \sum_{i=1}^{N} \left( \frac{1}{1-\rho d_i} - \frac{1}{1-\rho d_{i-1}} \right) \cdot \left( b - \frac{(b-a)}{2}(d_i + d_{i-1}) \right) \\ \text{s.t.} \quad & d_0 = 0 \le d_1 \le \ldots \le d_{N-1} \le d_N = 1. \end{aligned} \tag{48}$$

By letting $1 - \rho d_i = e_i$, the above problem can be further simplified to

$$\begin{aligned} \min_{\mathbf{e}} \quad & \frac{\lambda \mathbb{E}[B^2]}{2\rho} \sum_{i=1}^{N} \left( \frac{1}{e_i} - \frac{1}{e_{i-1}} \right) \cdot \left( \frac{b(\rho-1)+a}{\rho} + \frac{(b-a)(e_i + e_{i-1})}{2\rho} \right) \\ \text{s.t.} \quad & e_0 = 1 \ge e_1 \ge \ldots \ge e_{N-1} \ge e_N = 1-\rho. \end{aligned} \tag{49}$$

Fix $i \in \{1, 2, \ldots, N-1\}$ and fix $e_j$ for all $j \ne i$. Then the problem (49) becomes

$$\begin{aligned} \min_{e_i} \quad & e_i \left( \frac{1}{e_{i+1}} - \frac{1}{e_{i-1}} \right) + \frac{e_{i-1} - e_{i+1}}{e_i} \\ \text{s.t.} \quad & e_{i-1} \ge e_i \ge e_{i+1}, \end{aligned}$$

whose optimal solution is $\sqrt{e_{i-1} e_{i+1}}$. That is, the optimal $\boldsymbol{e}^*$ satisfies the following chain rule

$$\frac{e^*_{k+1}}{e^*_k} = \frac{e^*_k}{e^*_{k-1}}, \quad k = 1, 2, \ldots, N-1.$$

Since $e^*_0 = 1$ and $e^*_N = 1-\rho$, we have $e^*_k = (1-\rho)^{k/N}$ and thus

$$d^*_k = \frac{1 - (1-\rho)^{k/N}}{\rho}, \text{ for } k = 0, 1, \ldots, N.$$

Plugging these values into (48) yields

$$\frac{\lambda \mathbb{E}[B^2]}{2\rho} \left[ \frac{b(\rho-1)+a}{1-\rho} + \frac{b-a}{2\rho} N((1-\rho)^{-1/N} - (1-\rho)^{1/N}) \right].$$

The social-welfare-maximizing prices are obtained by using (22), (24), $w_i = \mathbb{E}[W_i]$ and $1-\rho\bar{H}(c_i)) = e_i$ for $0 \le i \le N$. □

# B. Auxiliary Results

## B.1. Auxiliary Lemmas

LEMMA 6. *For any $y>0$, there exists a number $x>0$ such that $F_e(x)=y$, where $F_e(x)$ is defined in* (6).

*Proof.* For any function $f:\mathbb{R}_+\to\mathbb{R}$, we use $\Delta^- f(x)$ to denote $f(x)-f(x^-)$ and $\Delta^+ f(x)$ to denote $f(x^+)-f(x)$, where $f(x^-) := \lim_{y\uparrow x} f(y)$ and $f(x^+) := \lim_{y\downarrow x} f(y)$. Below we show the following claims for any $x_0>0$:

$(a)$ If $H(x)$ is continuous in $x$ at $K/(\rho x_0)$, then $F_e(x)$ is continuous in $x$ at $x_0$;

$(b)$ If $\Delta^- H(K/(\rho x_0))>0$, then $\Delta^+ F_e(x_0) = -\rho x_0 \Delta^- H(K/(\rho x_0)) < 0$;

$(c)$ $\Delta^- F_e(x_0) = 0$;

$(d)$ $F_e(0^+) = 0$;

$(e)$ $F_e(x)\to\infty$ as $x\to\infty$.

Therefore, if there is a jump in $F_e$, then $(b)$ and $(c)$ jointly imply that it must be a downward jump. Hence, Lemma 6 follows immediately in view of $(d)$ and $(e)$.

First, we prove $(a)$. For arbitrary $\epsilon>0$, it follows from the continuity of $H$ at $K/(\rho x_0)$ that there exists a number $\delta_H>0$, such that

$$\left|H(x)-H\left(\frac{K}{\rho x_0}\right)\right| < \frac{\epsilon}{3\rho x_0}, \text{ whenever } \left|x-\frac{K}{\rho x_0}\right| < \delta_H.$$

Choose

$$\delta = \min\left(\frac{\epsilon}{3}, \frac{x_0}{2}, \frac{\rho x_0^2 \delta_H}{2K}, \frac{2\epsilon}{\rho\left(H\left(\frac{K}{\rho x_0}\right)+\frac{\epsilon}{3\rho x_0}\right)}\right).$$

Then for any $x$ with $|x-x_0|<\delta$, we have $x > x_0-\delta \ge x_0/2 > 0$. Moreover,

$$\left|\frac{K}{\rho x}-\frac{K}{\rho x_0}\right| = \frac{K|x-x_0|}{\rho\cdot x\cdot x_0} < \frac{2K|x-x_0|}{\rho x_0^2} < \frac{2K\delta}{\rho x_0^2} \le \delta_H.$$

Hence, we have

$$\begin{aligned}
|F_e(x)-F_e(x_0)| &\le (1-\rho)|x-x_0| + \rho\left|H\left(\frac{K}{\rho x}\right)x - H\left(\frac{K}{\rho x_0}\right)x_0\right| \\
&\le (1-\rho)|x-x_0| + \rho\left|H\left(\frac{K}{\rho x}\right)-H\left(\frac{K}{\rho x_0}\right)\right|x_0 + \rho H\left(\frac{K}{\rho x}\right)|x-x_0| \\
&\le |x-x_0| + \rho x_0\left|H\left(\frac{K}{\rho x}\right)-H\left(\frac{K}{\rho x_0}\right)\right| \\
&\quad +\rho\left(H\left(\frac{K}{\rho x_0}\right)+\left|H\left(\frac{K}{\rho x}\right)-H\left(\frac{K}{\rho x_0}\right)\right|\right)\cdot\frac{|x-x_0|}{2} \\
&< \frac{\epsilon}{3} + \rho x_0\cdot\frac{\epsilon}{3\rho x_0} + \rho\left(H\left(\frac{K}{\rho x_0}\right)+\frac{\epsilon}{3\rho x_0}\right)\cdot\frac{\delta}{2} = \epsilon.
\end{aligned}$$

Therefore, $F_e(x)$ is continuous in $x$ at $x_0$.

The proof of $(b)$ and $(c)$ is straightforward by the definition of $F_e$ and the fact that the distribution function $H$ is right-continuous with left limits by its definition.

$(d)$ follows by virtue of $0 \le H(\cdot) \le 1$, and $(e)$ follows since $0 < \rho < 1$. □

LEMMA 7. *Suppose that Assumption 1 holds. Then, $F_e'(x) \to 1$ as $x \to 0^+$.*

*Proof.* Obviously, $H\left(K/(\rho x)\right) \to 1$ as $x \downarrow 0$. By virtue of (31), we only need to prove that $h(K/(\rho x))/x \to 0$ as $x \downarrow 0$, or equivalently, $h(y)y \to 0$ as $y \to \infty$.

Note that the changeover point of $h$ , denoted by $x_0$, must be finite. We show that for any $\epsilon > 0$, we have $h(y)y < \epsilon$ for sufficiently large $y$. For any $\delta \in (0,1)$, there exists a number $x_2 \ge x_0$ such that $\bar{H}(x_2) < \epsilon/2$. For sufficiently large $y$ such that $y > x_2$ and $h(y) < \epsilon/(2x_2)$, we have

$$\bar{H}(x_2) = \int_{x_2}^{\infty} h(z)dz \ge h(y) \cdot (y - x_2),$$

and thus $yh(y) \le h(y)x_2 + \bar{H}(x_2) < \epsilon$. Hence, $h(y)y \to 0$ as $y \to \infty$. □

LEMMA 8. *For any $\rho \in (0,1)$, there exists a unique root to $f(x) = 0$ in $(0,1)$, where $f$ is one of the following functions:*

(i) $f(x) = \ln x + 1 - \rho x$;

(ii) $f(x) = (1-\rho)(\alpha x + x - \alpha) + \rho x^{\alpha+1}$ *for* $\alpha > 0$;

(iii) $f(x) = \rho x - \theta \ln x - 1$ *for* $\theta > 0$;

*Proof.* We prove each part of the lemma separately.

(i) Note that $f(0^+) = -\infty$ and $f(1) = 1 - \rho > 0$. Since $f'(x) = 1/x - \rho$, we know that $f(x)$ is first strictly increasing in $x$ from 0 to $1/\rho$, and then strictly decreasing in $x$ from $1/\rho$ to $\infty$. Hence, $f(x) = 0$ has only one root in $(0,1)$.

(ii) Note that $f(0) = -\alpha(1-\rho) < 0$ and $f(1) = 1$. Since $f'(x) = (\alpha+1)(1-\rho+\rho x^{\alpha}) > 0$, we know that $f(x)$ is strictly increasing in $x$ from 0 to 1. Hence, $f(x) = 0$ has only one root in $(0,1)$.

(iii) Note that $f(0^+) = +\infty$ and $f(1) = \rho - 1 < 0$. Since $f'(x) = \rho - \theta/x$, we know that $f(x)$ is first strictly decreasing in $x$ from 0 to $\theta/\rho$, and then strictly increasing from $\theta/\rho$ to $\infty$. Hence, $f(x) = 0$ has only one root in $(0,1)$. □

LEMMA 9. *For any $\rho \in (0,1)$, there exists a unique root to $\int_0^{\infty} e^{-(\frac{c}{\tau})^{\theta}} dc - \rho \int_x^{\infty} e^{-(\frac{c}{\tau})^{\theta}} dc - x = 0$ in $(0,\infty)$, for $\theta > 0$ and $\tau > 0$.*

*Proof.* Let $f(x) := \int_0^\infty e^{-(\frac{c}{\tau})^\theta} dc - \rho \int_x^\infty e^{-(\frac{c}{\tau})^\theta} dc - x$. Then $f(0) = (1-\rho)\int_0^\infty e^{-(\frac{c}{\tau})^\theta} dc > 0$ and $f(\infty) = -\infty$. Since $f'(x) = \rho e^{-(x/\tau)^\theta} - 1 < 0$, we know that $f(x)$ is strictly decreasing in $x$ from 0 to $\infty$. Hence, $f(x) = 0$ has only one root in $(0, \infty)$. □

LEMMA 10. *Let $\lambda(x) := \phi(x)/(1-\Phi(x))$ be the hazard rate of the standard normal distribution. Then, $\lambda(x) > x$.*

*Proof.* Define $f(x) = \phi(x) - x(1-\Phi(x))$. It suffices to show that $f(x) > 0$. Note that $f'(x) = \phi'(x) + x\phi(x) - (1-\Phi(x)) = -(1-\Phi(x)) < 0$. Hence, $f(x) > 0$ follows immediately from $f(\infty) = 0$. □

## B.2. Revenue-Maximizing Price when Proposition 2's Conditions Fail

To find the revenue-maximizing price in each case, we need to solve the optimization problem $\sup_{p\in[0,1]} M(p)$ and find the maximizer $p^*$ (if it exists). Then, the revenue-maximizing price $K^*$ will be $K(p^*)$, where $K$ is defined in (12).

*Uniform Distribution.* Let $C \sim U(a,b)$ with $(b-2a)/(b-a) \le \rho$. Following the proof of Proposition 2 (iii), we have

$$M(p) = \frac{p(b-(b-a)p)}{1-p\cdot\rho}, \quad M'(p) = \frac{b-2(b-a)p+\rho(b-a)p^2}{(1-p\cdot\rho)^2}.$$

Again, we let $g(p) := b - 2(b-a)p + \rho(b-a)p^2$. Then $g'(p) = -2(b-a)(1-\rho p) < 0$. Note that $g(0) = b > 0$, $g(1) = 2a - b + \rho(b-a) \ge 0$ in view of $\rho \ge (b-2a)/(b-a)$. Hence, $M(p)$ is increasing in $p \in (0,1]$, which attains its maximum at $p^* = 1$. Besides, it follows from (12) that the revenue-maximizing price is $K^* = K(p^*) = a\rho\mathbb{E}[W]/(1-\rho)$.

The above calculation implies at optimality all customers will join class 1. We provide an explanation here. Note that $\rho \ge (b-2a)/(b-a)$ holds only if $a > 0$ and $\rho$ is rather large, which means that each customer has a minimum positive delay cost rate and the system is rather congested. For any arriving customer, if she joins class 2, she will be delayed for a substantially large time (since $\rho$ is large and class 1 is given strict priority), and thus incur a large delay cost (since her delay cost rate is not small). Hence, she will prefer to join class 1.

*Burr distribution.* The CDF is $H(x) = 1 - (1+(x/a)^d)^{-k}$ with $k, d > 0$, $kd \le 1$ and thus

$$M(p) = \frac{pa(p^{-1/k}-1)^{1/d}}{1-p\cdot\rho}.$$

If $kd < 1$, then we have

$$\lim_{p\to 0^+} p(p^{-1/k}-1)^{1/d} = \lim_{r\to\infty} \frac{r}{(r^d+1)^k} = \infty,$$

where the first equality follows by letting $r=(p^{-1/k}-1)^{1/d}$. Hence, we have $M(0^+):=\lim_{p\downarrow 0}M(p)=\infty$, which implies at optimality almost all customers will join class 2. Note that

$$K(p)=\frac{a(p^{-1/k}-1)^{1/d}}{1-p\cdot\rho}\cdot\rho\mathbb{E}[W]\to\infty,\ \text{as } p\to 0^+.$$

We provide an explanation here. Note that $1-H(x)=(1+(x/a)^d)^{-k}$ is of order $1/x^{dk}$. Hence, the delay cost rate is distributed with "heavy tail". For a large priority price $K$, it follows from (12) that there will be a fraction of order $1/K^{dk}$ customers joining class 1. Hence, the revenue will be of order $K^{1-dk}$ for large $K$, which induces the system to set price as high as possible since $dk<1$.

If $kd=1$, we distinguish two cases: $k<1$ versus $k\geq 1$.

Case 1: $k<1$. Recall the proof of Proposition 2 (i). Now we have $g(0)=0$, $g(1)<0$. Since $g'(p)=(\rho-d\cdot p^{1/k-1})/kd$, $g'(0)>0$. Moreover, $g''(p)<0$ for $p\in(0,1)$. Hence, $g(p)=0$ has exactly one root $p^*\in(0,1)$ such that $g(p)>0$ for $p\in(0,p^*)$ and $g(p)<0$ for $p\in(p^*,1)$, which implies that $M(p)$ is strictly quasi-concave in $p\in(0,1)$.

Case 2: $k\geq 1$. Then we have $g'(0)<0$. Hence, $g'(p)<0$ for all $p\in(0,1)$, which implies that $M(p)$ is decreasing in $p\in(0,1)$. In this cases, $M(p)$ takes its supremum $M(0^+)=a$ when $\rho\to 0^+$.

# C. Supporting Materials for Numerical Studies

## C.1. Formulas of the Mean and Variance

For the five distribution families considered in the main paper, we have explicit formulas of their means and variances, which are summarized in Table 3.

**Table 3 Formulas for Mean and Variance of the Various Distributions**

| Distribution | Parameters | Mean | Variance |
|---|---|---|---|
| Uniform | $a,b$ | $(a+b)/2$ | $(b-a)^2/12$ |
| Exponential | $\kappa$ | $\kappa$ | $\kappa^2$ |
| Weibull | $\tau,\theta$ | $\tau\Gamma(1+1/\theta)$ | $\tau^2[\Gamma(1+2/\theta)-(\Gamma(1+1/\theta))^2]$ |
| Power | $\alpha$ | $1/(\alpha+1)$ | $1/(\alpha+2)-1/(\alpha+1)^2$ |
| Log-normal | $\mu,\sigma$ | $e^{\mu+\sigma^2/2}$ | $(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$ |

For single-parameter distributions (exponetial and power), the mean and variance can not be set arbitrarily. Hence, in the second numerical study of section 7.5.2, we only consider two-parameter distributions. Using the formulas presented in Table 3, we can determine the parameters for the two-parameter distributions given the mean $\mathbb{E}[C]$ and the variance $\text{var}[C]$:

*Uniform distribution.* We have $a=\mathbb{E}[C]-\sqrt{3\text{var}[C]}$, $b=\mathbb{E}[C]+\sqrt{3\text{var}[C]}$. Note that $a\geq 0$ only if $(\mathbb{E}[C])^2\geq 3\text{var}[C]$.

*Weibull distribution.* We have no close-form expression as in uniform or log-normal distribution. Hence, we will choose $\tau$ and $\theta$ first to set the value of mean and variance.

*Log-normal distribution.* We have

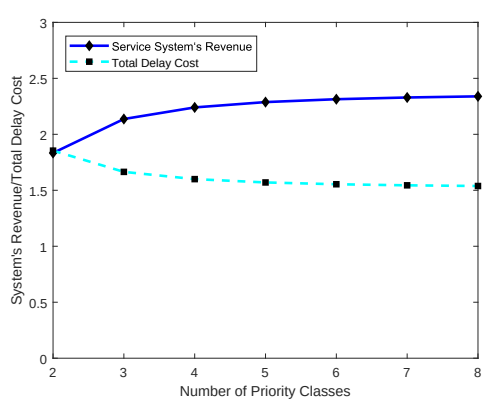
$$\mu = \ln(\mathbb{E}[C]) - \frac{1}{2}\ln\left(1 + \frac{\text{var}[C]}{\mathbb{E}[C]^2}\right), \quad \sigma^2 = \ln\left(1 + \frac{\text{var}[C]}{\mathbb{E}[C]^2}\right).$$

Recall that in section 7.5.1, we set the parameters such that the mean is 0.75 and the variances are $0.2, 0.4, 0.6, 0.8, 1.5$ respectively. Hence, the corresponding parameters $(\mu, \sigma)$ are (-0.4398, 0.5516), (-0.5563, 0.7329), (-0.6507, 0.8520), (-0.7300, 0.9406), and (-0.9373, 1.1399). The corresponding conditions in Proposition 2 are $\rho < 2.2721$, $\rho < 1.7107$, $\rho < 1.4710$, $\rho < 1.3325$ and $\rho < 1.0995$, which are all satisfied as $\rho < 1$.

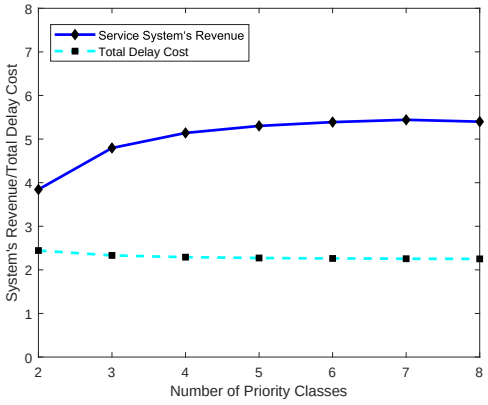
## C.2. Additional Numerical Examples

Figure 16 plots the optimal service system's revenue and total waiting cost with respect to the number of priority classes for four cases under $\rho = 0.8$. In comparison with Figure 15 under $\rho = 0.9$, we find that the value of providing more number of priority classes increases as the traffic intensity $\rho$ becomes larger. This is consistent with the finding in Figure 14, which only considers uniform distribution.

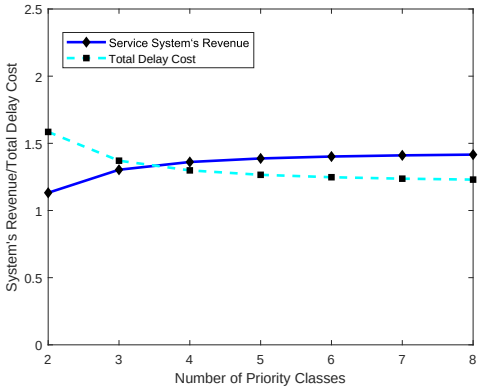
**Figure 16 Optimal Revenue and Delay Cost with Respect to the Number of Priority Classes When $\rho = 0.8$**



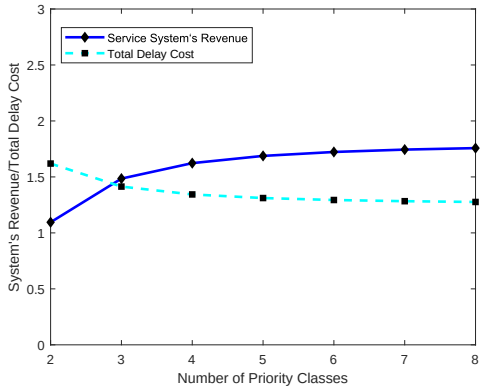

(a) Uniform Distribution




(b) Power Distribution




(c) Exponential Distribution




(d) Log-normal Distribution