POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT POMS

Vol. 23, No. 4, April 2014, pp. 676–688
ISSN 1059-1478 | EISSN 1937-5956 | 14 | 2304 | 0676

DOI 10.1111/poms.12171

Optimal Uniform Pricing Strategy of a Service Firm When Facing Two Classes of Customers

Wenhui Zhou
School of Business Administration, South China University of Technology, Guangzhou 510640, China
whzhou@scut.edu.cn

Xiuli Chao
Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109, USA
xchao@umich.edu

Xiting Gong
Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T. Hong Kong
xtgong@se.cuhk.edu.hk

When facing heterogeneous customers, how should a service firm make its pricing decision to maximize revenue? If discrimination is allowed, then priority schemes and differentiated pricing are often used to achieve that. In many applications, however, the firm cannot or is not allowed to set discriminatory prices, for example, list price in retail stores, online shopping, and gas stations; thus a uniform price must be applied to all customers. This study addresses the optimal uniform pricing problem of a service firm using a queueing system with two classes of customers. Our result shows that the potential pool of customers plays a central role in the firm's optimal decision. Depending on the range of system parameters, which are determined explicitly by the primitive data, the firm's optimal strategy may choose to serve only one class of customers, a subset of a class of customers, or a combination of different classes of customers. In addition, the optimal price is in general not monotonic with respect to the potential market sizes because their changes may lead to a major shift in the firm's decision on which customer class to serve. However, unless such a shift occurs, the optimal price is weakly decreasing in the potential market sizes.

Key words: service system; queueing delays; delay-sensitive customers; optimal pricing
History: Received: April 2011; Accepted: August 2013 by Michael L. Pinedo, after 4 revisions.

1. Introduction

Delay is common in both manufacturing and service sectors. For example, in such industries as airplane manufacturing, ship building, textile mill products, steel, fabricated metals, and electric/nonelectric machinery, the average leadtimes/delays for orders can be in months or even years. For these industries, not only does the price of the product/service affect the customers' purchasing decision, but the delay is also an important factor that the customers consider. When discrimination is allowed, priority schemes have been proposed in the literature to maximize the firm's revenue. In many applications, however, the firm cannot or is not allowed to set discriminatory prices, for example, list price in retail stores, online shopping, and gas stations; thus a uniform price must apply to all customers. Similarly, the service cannot be discriminative either, and customers have to be served according to the first-come-first-served (FCFS) discipline. In such scenarios, even though the firm knows that there are different types of customers in the market, it has to offer a uniform price. Then, how should the firm make a uniform pricing decision based on its knowledge of the potential market structure and the characteristics of different classes of customers to maximize its expected profit/revenue?

This is the motivation for this study and the question we address. Throughout the article we refer to the system as a service system, even though it can also be a make-to-order manufacturing system. We assume that the firm is a monopoly in the sense that it provides an exclusive service to a designated market or there is little competition for the service provided. The firm faces two classes of customers, who may differ in their valuations of the service as well as their sensitivities to delays. The objective of the firm is to utilize its knowledge on the structure of its market base to find a uniform pricing decision that maximizes its expected revenue.

We model the problem by a queueing system with a Poisson arrival process of each class of customers and an arbitrary service time distribution and investigate the optimal pricing decision of the service firm. For any given service price by the firm, the customers make decisions on whether or not to join the service system. Thus, the resulting optimization problem of the service firm is a Stackelberg game. We solve this game problem and show that, depending on the range of system parameters, the firm's optimal pricing strategies can vary significantly. We obtain the optimal price for each and every range of the system parameters, and compare the optimal prices as well as revenues in the various parameter regions.

Roughly speaking, the literature on optimization of pricing and queues can be divided into two categories. In the first category, the queue is observable, and the resulting optimization problem is the dynamic control of queues, while in the second category, the queue is not observable and customers make decisions based on their perceptions on the average waiting times. Naor (1969) is the first to study the interactions between the price and queueing delays. He studies an observable M/M/1 queueing system with a single class of customers and customer balking and discusses the revenue-maximizing pricing decisions of the firm. Knudsen (1972) generalizes Naor's result to a multi-server queue with a nonlinear waiting cost function. Edelson and Hilderbrand (1975) investigate Naor's model with an unobservable queue and the customers make their decisions based on the expected waiting time. Other extensions of Naor's model include De Vany (1976), Lippman and Stidham (1977), Hassin (1986), and Chen and Frank (2004). All these studies assume that the customers in the market are homogeneous, and they study the relationship between socially and individually optimal decisions. For a comprehensive review on the design and control of queues, the reader is referred to Hassin and Haviv (2002) and Stidham (2009, 2011). There are also some studies on competition of multiple firms. In these studies, the firms set prices to compete for customers in the same pool and search for the pricing equilibrium. For example, Chen and Wan (2003) study the pricing equilibrium of two competitive firms with homogeneous customers under different market sizes.

Customers are usually heterogeneous; that is, they have different valuations for service as well as different sensitivities to delays. Most studies on heterogeneous customers focus on priority schemes (i.e., the priority of service and pricing decisions are jointly made by the firm with higher priority customers paying more for the same service) (e.g., Afanasyev and Mendelson 2010, Bradford 1996, Gilland and Warsing 2009, Hassin and Haviv 1997, Mendelson and Whang 1990, Rao and Petersen 1998, Van Mieghem 2000). When customers are heterogeneous and have private information, one approach in the operations literature is to design a menu of contracts with two attributes. The machinery in information economics such as adverse selection is used to design a menu, and by varying the values of the two attributes, the firm is able to extract the private information of customers and maximize its revenue. The key in such mechanism designs is the multi-dimensional nature of the contract, which enables the firm to coordinate different contract parameters to achieve a desired outcome. Mendelson and Whang (1990) are perhaps the first to present an incentive-compatible pricing scheme which induces the customers to reveal their true class information. There are only a few studies focusing on firms' non-discriminative decisions. Armony and Haviv (2003) study a uniform pricing problem of two competing firms under two classes of customers with the same valuation for service but different sensitivities to delays, and they identify various scenarios when symmetric, asymmetric, and continuum price equilibria exist and when a price equilibrium does not exist. Afèche and Mendelson (2004) investigate a uniform pricing and priority auction problem under continuum classes of customers with different valuations for service and delay costs.

In our model, the queue is not observable, and the two classes of customers have different valuations for the service and different sensitivities to their perceived delays. As a result, either class of customers may have a higher incentive to enter the service system, and the firm can only influence the customers' decisions on entering the system by its selling price. Once the customers enter the system, they are served on the FCFS basis. We will see later that in some scenarios, one class of customers always has a higher incentive than the other to join the service system regardless of the service price, while in other scenarios this is not true, and either class of customers may have a higher incentive to enter the service system, depending on the service price. To distinguish these two scenarios, in this study we refer to them as the *dominating customers case* and the *non-dominating customers case*, respectively. Note that if the two classes of customers differ in only one aspect, either in the valuation for service or in the sensitivity to delays, then the underlying scenario is always a dominating customers case. However, it is conceivable that a customer's value for service and its sensitivity to delay is correlated and in particular often positively correlated as they are both affected by some common factor of the individual customer. We refer to section 6.3.3 of Mandelbaum et al. (2001) for some empirical evidence on positive correlation between value of customer service and sensitivity to delays in a call center. To the best of our knowledge, similar studies all have focused on the dominating customers case. For example, Armony and Haviv (2003) consider a case where two classes of customers only differ in their sensitivities to delays, while Afanasyev and Mendelson (2010) study a problem where two classes of customers only differ in their valuations for service. On the other hand, both Armony and Haviv (2003) and Afanasyev and Mendelson (2010) study the pricing competition between two firms, which is more general than ours; we focus on a monopoly firm. As it turns out, the equilibrium customer behaviors and the optimal pricing of the firm for the non-dominating customers case are significantly more complicated than the dominating customers case. For both cases, we first analyze the customer equilibrium and, based on that, we then study the optimal pricing decisions for the firm. In contrast to the literature on comparing the effects of revenue-maximizing pricing and socially optimal pricing mentioned above, we are mainly concerned with the effects of the market structure on the firm's revenue-maximizing pricing decisions.

The first main insight offered in this study is that the market structure, which we refer to as the arrival rates of different classes of potential customers and their relative characteristics, plays a prominent role for the firm in making its pricing decisions. Indeed, as shown in this study, different values of the market sizes (arrival rates of different classes of potential customers) lead to completely different pricing strategies for the firm. Depending on the range of system parameters, which are explicitly determined by the system primitive data, the firm may choose to focus only on one class of customers, a subset of a class of customers, or a combination of different classes of customers, and it may decide to completely ignore some customer segments in setting its price. Therefore, to make an informed decision, it is crucial for the firm to understand its potential market base and the characteristics of the various customers. Second, our result reveals interesting non-monotonicity in the firm's optimal pricing with respect to potential market sizes. To maximize its revenue, the firm has to carefully set its price to shape up its market structure by deciding which and how many of customers to serve, and that leads to non-monotonicity in pricing with regard to potential market sizes. Specifically, the optimal price of the firm is in general not monotone with respect to the potential market sizes because the changes in the market sizes may lead to a major shift in the firm's decision on which customer classes to serve. However, we show that unless there is such a major change in the firm's optimal strategy, the optimal price is weakly decreasing in the potential market size of each class of customers.

Our results are closely related to some important phenomena often observed in service systems such as retail stores, restaurants, etc. It is common that some of the popular chain stores focus on different segments of the potential market; for example, some focus only on high end consumers, some focus only on low end consumers, while some others catch consumers from both ends. Our results show that this is due to the heterogeneity in the different classes of consumers: depending on the capacity of the service system and the potential market sizes of different classes of customers, the optimal strategy of the firm, including price, waiting times, quality of service, etc., leads to a corresponding optimal combination of different classes of customers to serve.

The rest of the article is organized as follows. The model formulation is introduced in section 2. In section 3, we discuss the equilibrium customer behavior for any given price set by the firm, and it is divided into two cases: dominating customers case and non-dominating customers case. In section 4, we study the firm's optimal pricing decision. Again, the results are presented for the two cases separately. We also present a numerical example to illustrate our results in section 4. Finally, concluding remarks are given in section 5. Throughout the article, we use increasing and decreasing in non-strict sense; that is, they represent "non-decreasing" and "non-increasing" respectively. All technical proofs and other supplementary materials are given in the online Supporting Information.

2. The Model

We consider a firm that provides a service to a market. There are two classes of potential customers, referred to as class-1 and class-2 customers, and they arrive according to independent Poisson processes with rates Λ_1 and Λ_2, respectively. The total arrival rate is denoted as $\Lambda = \Lambda_1 + \Lambda_2$. The service system is modeled as a queueing system. For simplicity, we focus on the case with a single server and will discuss the case with multiple servers in section 5. The service time for each customer is a random variable. Since we consider the scenario where the customers are indistinguishable by the firm, the service times for all customers are assumed to have the same distribution. Let S be a generic random service time with the mean value and the coefficient of variation denoted by μ^{-1} and c_v, respectively. The service discipline is FCFS. Each class of customers has a common perceived value of the service. Let v_i denote this perceived value for class-i customers, $i = 1,2$. Customers are also sensitive to delays. We assume the delay cost for a customer is in proportion to his sojourn time in the system. Let the delay cost per unit of time for class-i customers be denoted by d_i. In addition, we denote $w_i = v_i - d_i/\mu$. Then, w_i is the expected value of the service for a class-i customer when it receives service without any delay in queue. For convenience, we refer to w_i as the expected *net* value of the service for class-*i* customers. Without loss of generality, we assume that $w_1 \geq w_2 > 0$ and, when $w_1 = w_2$, $d_1 < d_2$. We also assume that all the above information/parameters are common knowledge to all potential customers as well as the service firm.

The firm posts a price p for its service, which cannot be differentiated for different customers. An arriving customer cannot observe the queue length and makes a decision on whether or not to enter the service system to maximize his expected utility. We refer to this decision as the customer's *joining decision*, and, rigorously speaking, it is to choose a probability to enter the system rather than to make an enter-it-or-leave-it choice. We assume that the utility for not joining the system is set as the zero reference for all customers. On the other hand, when a customer joins the system, he will not renege and will leave the system only after his service is completed. In this case, his expected utility equals his perceived value of the service minus the price of the service and the expected delay cost. Let $W_Q(\lambda)$ denote the expected waiting time in queue by an arbitrary customer, given that the total arrival rate at which customers join the system is λ. Then, the utility function of a class-*i* customer when joining the system is

$$\begin{aligned} u_i(\lambda, p) &= v_i - p - d_i(W_Q(\lambda) + \mu^{-1}) \\ &= w_i - p - d_i W_Q(\lambda), i = 1, 2. \end{aligned} \tag{1}$$

Thus, it is optimal for a class-*i* customer to enter (respectively, not enter) the service system when $u_i(\lambda,p)$ is positive (respectively, negative) and to enter the system with any probability when $u_i(\lambda,p)$ equals zero. Clearly, class-*i* customers will not join the system if $p > w_i$, $i = 1,2$.

It is well known that (see, e.g., Kleinrock 1975), for an M/G/1 queue, the expected waiting time in queue is

$$W_Q(\lambda) = \frac{\lambda(1 + c_v^2)}{2\mu(\mu - \lambda)}. \tag{2}$$

It is easy to see that $W_Q(\lambda)$ is strictly increasing and convex in λ in the stability region $0 \leq \lambda < \mu$. Hence the utility function $u_i(\lambda,p)$ is strictly decreasing and concave in λ.

Based on Equations (1) and (2), the following result compares the utility functions, $u_1(\lambda,p)$ and $u_2(\lambda,p)$, of the two classes of customers.

LEMMA 1. *For any price p of the service and* $0 \leq \lambda < \mu$, *the following results hold;*

(a) when $d_1 \leq d_2$, $u_1(\lambda,p) \geq u_2(\lambda,p)$ *for any* $\lambda \in [0,\mu)$;

(b) when $d_1 > d_2$, $u_1(\hat{\lambda}, p) = u_2(\hat{\lambda}, p)$, $u_1(\lambda,p) > u_2(\lambda, p)$ *when* $\lambda < \hat{\lambda}$, *and* $u_1(\lambda,p) < u_2(\lambda,p)$ *when* $\lambda > \hat{\lambda}$, *where*

$$\hat{\lambda} = \frac{2\mu^2(w_1 - w_2)}{2\mu(w_1 - w_2) + (1 + c_v^2)(d_1 - d_2)}. \tag{3}$$

Lemma 1 states that when $d_1 \leq d_2$, class-1 customers always have higher utility values than class-2 customers, regardless of the service price and the arrival rate. This is very intuitive, since class-1 customers have higher expected net valuation for the service and, meanwhile, are less sensitive to delays. In contrast, when $d_1 > d_2$, the utility value of class-1 customers can be either higher or lower than that of class-2 customers, depending on whether the arrival rate λ is lower or higher than the benchmark rate $\hat{\lambda}$. The reason is that since class-1 customers are more sensitive to delays, their utility decreases more quickly than class-2 customers' when the expected waiting time increases. Since the expected waiting time increases in the arrival rate λ, the utility value of class-1 customers is higher (respectively, lower) than that of class-2 customers when λ is low (respectively, high).

Class-1 customers always have higher utility values than class-2 customers when $d_1 \leq d_2$, but this dominating relationship is not true when $d_1 > d_2$. For convenience we shall refer to these two cases as the *dominating customers case* and the *non-dominating customers case*, respectively. As mentioned in the Introduction, the relevant studies on homogeneous customers have all focused on the dominating customers case by assuming either the same valuation for service or the same sensitivity for delays. In this study, we will consider both cases with the primary focus on the non-dominating customers case. We remark that the non-dominating customers case is often observed in applications since a customer's valuation for service may be positively correlated with his sensitivity to delays (i.e., w_i and d_i are positively correlated). This is because the valuation for service and the sensitivity to delays are often influenced by some common factors. As an example, consider the case of auto repair, then the utility of driving the car is such a common factor that impacts both the valuation for the repair service and the sensitivity to delays.

The objective of the service firm is to set a price p that maximizes its expected revenue per unit of time. To this end, the firm needs to understand how the two classes of potential customers react to its pricing decision. On the other hand, for any given service price, since customers are utility maximizers in deciding whether or not to enter the system, they will optimize for themselves and eventually reach an equilibrium. Therefore, the optimal pricing problem described above for the firm is a Stackelberg game. In the following sections, we will solve this game using the backward induction and discuss how the primitive data impact the firm's optimal pricing decision as well as the customers' equilibrium joining behaviors.

3. Equilibrium Customer Behaviors

In this section, we analyze the customers' equilibrium joining behaviors for a given service price p. The customers make decisions, based on the price p and the expected delay, on whether or not to join the system for service. Clearly, the joining decisions of customers determine the actual demand rate, which in turn affects the expected waiting time and utility of all customers. In other words, the customers play a game. In this section, we find the Nash equilibrium in the customers' joining strategies. Since customers within each class are homogeneous, we focus on equilibrium strategies that are symmetric among customers of the same class.

By definition, a Nash equilibrium is a profile of strategies where it is optimal for a class of customers to follow their prescribed strategy if the other class of customers follow their prescribed strategy.

REMARK 1. To prove a pair of joining strategies (θ_1, θ_2) is an equilibrium, we need to show that, when class-j customers follow strategy θ_j, it is optimal for class-i customers to follow strategy θ_i, $i \neq j = 1,2$. Note that, the equilibrium analysis of class-i customers, when the strategy of class-j customers is fixed, is quite similar to that of a system with only class-i customers. There is an extensive literature on customer joining equilibrium to queueing systems with one class of customers, and two types of equilibria have been studied, based on *social optimization* and *individual optimization* (e.g., Bell and Stidham 1983). The equilibrium analysis we utilize in characterizing the customer behaviors is, due to the nature of our problem, based on individual optimization.

We begin our analysis by introducing two important functions. For a given p (respectively, λ), denote $\lambda_i(p)$ (respectively, $p_i(\lambda)$) as the solution of the equation $u_i(\lambda,p) = 0$. Then, by Equations (1) and (2), $\lambda_i(p)$ and $p_i(\lambda)$ have the following explicit forms

$$\lambda_i(p) = \frac{2\mu^2(w_i - p)}{2\mu(w_i - p) + d_i(1 + c_v^2)}; \tag{4}$$

$$p_i(\lambda) = w_i - \frac{\lambda d_i(1 + c_v^2)}{2\mu(\mu - \lambda)}. \tag{5}$$

The functions $\lambda_i(p)$ and $p_i(\lambda)$ have very intuitive interpretations: for a given service price p, $\lambda_i(p)$ is the maximum total arrival rate at which a class-i customer would choose to enter the system, and for a given total arrival rate λ, $p_i(\lambda)$ is the maximum price a class-i customer is willing to pay for entering the system.

The following lemma summarizes some important properties of $\lambda_i(p)$ and $p_i(\lambda)$, $i = 1,2$, which can be directly verified from their expressions in Equations (4) and (5).

LEMMA 2. *For $i = 1,2$, suppose $p \leq w_i$ and $0 \leq \lambda < \mu$; then*

- *(a) $\lambda_i(p)$ is strictly decreasing and concave in p, and $p\lambda_i(p)$ is strictly concave in p;*
- *(b) $p_i(\lambda)$ is strictly decreasing and concave in λ, and $\lambda p_i(\lambda)$ is strictly concave in λ;*
- *(c) $\lambda_i(p)$ and $p_i(\lambda)$ are inverse functions of each other; that is, $\lambda_i(p_i(\lambda)) = \lambda$ and $p_i(\lambda_i(p)) = p$.*

From Lemma 1 and our previous discussions, it is conceivable that the customer joining behaviors for the non-dominating customers case are more complicated than those for the dominating customers case. Therefore, in what follows we study these two cases separately.

We use $(\theta_1^{EQ}(p), \theta_2^{EQ}(p))$ to denote the equilibrium joining probabilities of the two classes of customers when the firm posts a selling price p.

3.1. Dominating Customers Case

We first consider the case of dominating customers, that is, $d_1 \leq d_2$. In this case, class-1 customers have more incentive to enter the system than class-2 customers, and only if the system still has remaining capacity after serving all class-1 customers shall class-2 customers have the possibility of entering the service system.

The following theorem characterizes the equilibrium customer behaviors for the dominating customers case. Note that in this case, it follows from Equation (4) that $\lambda_1(p) \geq \lambda_2(p)$ for any $p \leq w_2$.

THEOREM 1. (*Customer equilibrium for dominating customers case*). *Suppose $d_1 \leq d_2$. For a given service price p, the customers' equilibrium joining behaviors can be characterized as follows:*

1. *if $\Lambda_1 \geq \lambda_2(p)$, then $(\theta_1^{EQ}(p), \theta_2^{EQ}(p)) = (\min\{1, \lambda_1(p)/\Lambda_1\}, 0)$;*
2. *if $\Lambda_1 < \lambda_2(p)$, then $(\theta_1^{EQ}(p), \theta_2^{EQ}(p)) = (1, \min\{1, (\lambda_2(p) - \Lambda_1)/\Lambda_2\})$.*

The result above shows that, in the case of dominating customers, for any given price p there exists a threshold $\lambda_2(p)$ such that no class-2 customers will enter the system for service if the arrival rate of class-1 customers is at or above this threshold. More specifically, if there are a large number of class-1 customers in the system (i.e., $\Lambda_1 \geq \lambda_1(p)$), then only some class-1 customers will enter the system for service (with the equilibrium entering rate being $\lambda_1(p)$), and no class-2 customers will enter the system for service. If the potential arrival rate of class-1 customers is lower than $\lambda_1(p)$ but higher than $\lambda_2(p)$, then all class-1 customers will enter the system but still no class-2 customers will enter the system for service in equilibrium. Only when there are relatively few class-1 customers in the system or the arrival rate of class-1 customers is below this threshold (i.e., $\Lambda_1 < \lambda_2(p)$) shall some or all class-2 customers enter the system for service. Moreover, the lower the service price, the higher this threshold is. Note that even when the arrival rate of potential class-1 customers is lower than the threshold $\lambda_2(p)$, some class-2 customers may still be turned away because it is possible that there exists a large number of class-2 customers in system, or $\Lambda_2 > \lambda_2(p) - \Lambda_1$.

3.2. Non-dominating Customers Case

We now turn to the non-dominating customers case, that is, $d_1 > d_2$. To characterize the equilibrium behaviors of the customers, we first define a reference price, which can be negative, by

$$\hat{p} = \frac{d_1 v_2 - d_2 v_1}{d_1 - d_2}. \tag{6}$$

It is easy to verify from Equation (4) that $\hat{p}$ is the price for the two arrival rate functions $\lambda_1(p)$ and $\lambda_2(p)$ to be equal. Other relationships among the reference price $\hat{p}$, the reference arrival rate $\hat{\lambda}$, $\lambda_i(p)$, and $p_i(\lambda)$ are presented in the following lemma.

LEMMA 3. *When $d_1 > d_2$, the following results hold:*

(a) $\hat{\lambda} < \lambda_1(p) < \lambda_2(p)$ *when* $p < \hat{p}$, $\hat{\lambda} = \lambda_1(\hat{p}) = \lambda_2(\hat{p})$, *and* $\hat{\lambda} > \lambda_1(p) > \lambda_2(p)$ *when* $p > \hat{p}$;
(b) $p_1(\lambda) > p_2(\lambda) > \hat{p}$ *when* $\lambda < \hat{\lambda}$, $\hat{p} = p_1(\hat{\lambda}) = p_2(\hat{\lambda})$, *and* $p_1(\lambda) < p_2(\lambda) < \hat{p}$ *when* $\lambda > \hat{\lambda}$.

Recall that $\lambda_i(p)$ is the maximum total arrival rate at which a class-*i* customer would choose to enter the system. Therefore, Lemma 3(a) states that, when the service price p is lower (respectively, higher) than the reference price $\hat{p}$, class-1 customers have less (respectively, more) incentive to enter the system than class-2 customers. The intuitions are as follows. If the service price p is smaller than $\hat{p}$, then $\lambda_i(p)$ being decreasing in p implies $\lambda_i(p) \geq \hat{\lambda}$, $i = 1,2$. According to Lemma 1, class-1 customers have a lower utility value than class-2 customers when $\lambda \geq \hat{\lambda}$. Thus, class-1 customers have less incentive to enter the system than class-2 customers with a low service price. Furthermore, Lemma 3(b) shows that class-1 customers are willing to pay more (respectively, less) for entering the system than class-2 customers when the total arrival rate is low (respectively, high). This result is also intuitive, since class-1 customers are more sensitive to delays under the non-dominating customers case.

With Lemma 3, the following theorem characterizes the equilibrium customer behaviors for the non-dominating customers case.

THEOREM 2 (*Customer equilibrium for non-dominating customers case*). *When $d_1 > d_2$, the equilibrium customer behaviors depend on whether $p \geq \hat{p}$ or $p < \hat{p}$. Specifically, if $p \geq \hat{p}$, then the customer equilibrium is given by cases* 1 *and* 2 *in exactly the same way as that in Theorem* 1. *If $p < \hat{p}$, then the customer equilibrium is given by*

3. *if* $\Lambda_2 \geq \lambda_1(p)$, *then* $(\theta_1^{EQ}(p), \theta_2^{EQ}(p)) = (0, \min\{\lambda_2(p)/\Lambda_2, 1\})$;
4. *if* $\Lambda_2 < \lambda_1(p)$, *then* $(\theta_1^{EQ}(p), \theta_2^{EQ}(p)) = (\min\{1, (\lambda_1(p) - \Lambda_2)/\Lambda_1\}, 1)$.

Theorem 2 shows that the roles of two classes of customers when $p \geq \hat{p}$ and $p < \hat{p}$ are reversed in equilibrium: in the first case, the equilibrium customer behaviors are the same as those for the dominating customers case, where class-1 customers have more incentive to enter the system for service, while in the second case, class-2 customers have more incentive to enter the service system and, indeed, only when there is not a sufficient number of class-2 customers shall some or all class-1 customers enter the system.

Having analyzed the equilibrium customer behaviors for any given price for service, p, in the next section we study the firm's optimal pricing problem. Recall that the objective of the service firm is to maximize its expected revenue.

4. Firm's Optimal Pricing Strategy

The firm sets the service price p to maximize its expected revenue, $\pi(p) = p\lambda(p)$, where $\lambda(p)$ is the equilibrium arrival rate at which the two classes of customers enter the system when the price of service is p. Note that $\lambda(p)$ is determined by Theorems 1 and 2, respectively, in the previous section for the dominating and non-dominating customers cases.

In the following, we explicitly solve the firm's optimization problem $\max_p \pi(p)$ for its optimal pricing strategy. We show that the market structure or the arrival rates of different classes of potential customers plays a critical role for the firm's optimal decisions. Depending on the ranges of system parameters, which are explicitly expressed in terms of the system primitive data, the firm may choose to focus only on one class of customers, a subset of a class of customers, or a combination of different classes of customers in setting its price. In addition, the optimal pricing of the firm is in general not monotone with respect to the potential market sizes or arrival rates of different classes of potential customers. More specifically, we show that the firm's optimal price is decreasing in both Λ_1 and Λ_2 when the customer base being served is unchanged, but this monotonicity result is not true in general. When there is a change in the customer base being served, the optimal price will encounter a jump or drop, destroying the monotonicity of the optimal price in Λ_1 or Λ_2.

Some technical preparation is needed to analyze the firm's problem. For $i = 1,2$, and on $p \le w_i$ and $0 \le \lambda < \mu$, we define

$$\pi_i(p) = p\lambda_i(p) \quad \text{and} \quad \tilde{\pi}_i(\lambda) = \lambda p_i(\lambda). \tag{7}$$

Thus, from the definition of $\lambda_i(p)$ and $p_i(\lambda)$, $\pi_i(p)$ and $\tilde{\pi}_i(\lambda)$ are the firm's maximum revenues when the service price is p and when the equilibrium arrival rate is λ, respectively, given there are only but sufficient class-i customers in the system. By Lemma 2, both $\pi_i(p)$ and $\tilde{\pi}_i(\lambda)$ are strictly concave functions. Furthermore, we denote

$$p_i^* = \arg\max_{p \le w_i} \pi_i(p) \quad \text{and} \quad \lambda_i^* = \arg\max_{0 \le \lambda < \mu} \tilde{\pi}_i(\lambda)$$

as the optimal price and the optimal equilibrium arrival rate, respectively, if there are only but sufficient class-i customers in the system. Since $\lambda_i(p)$ and $p_i(\lambda)$ are inverse functions of each other by Lemma 1, we have $\pi_i(p) = \tilde{\pi}_i(\lambda_i(p))$ and $\tilde{\pi}(\lambda) = \pi_i(p_i(\lambda))$ and $p_i^* = p_i(\lambda_i^*)$ and $\lambda_i^* = \lambda_i(p_i^*)$. In addition, it follows from Equations (5) and (7) that p_i^* is given explicitly as

$$p_i^* = w_i - \frac{\sqrt{d_i(1+c_v^2)(d_i(1+c_v^2)+2\mu w_i)} - d_i(1+c_v^2)}{2\mu}. \tag{8}$$

For $i = 1,2$, since $\tilde{\pi}_i(\lambda)$ is strictly concave in λ by Lemma 1 and achieves its maximum when $\lambda = \lambda_i^*$, it strictly increases from $\tilde{\pi}_i(0) = 0$ to $\tilde{\pi}_i(\lambda_i^*)$ when λ increases from 0 to λ_i^*. Thus, there exists a unique solution of λ to the equation $\tilde{\pi}_i(\lambda) = \min\{\tilde{\pi}_1(\lambda_1^*), \tilde{\pi}_2(\lambda_2^*)\}$ over $\lambda \in [0, \lambda_i^*]$, and we define it as $\underline{\lambda}_i^*$. Then, by definition, $0 \le \underline{\lambda}_i^* \le \lambda_i^*$ and $\tilde{\pi}_i(\underline{\lambda}_i^*) = \min\{\tilde{\pi}_1(\lambda_1^*), \tilde{\pi}_2(\lambda_2^*)\}$. That is, $\underline{\lambda}_i^*$ is the equilibrium arrival rate in the system with only class-i customers that yields the smaller one of the maximum revenues for the two systems with only class-1 customers and with only class-2 customers, respectively. Obviously, $\underline{\lambda}_i^* = \lambda_i^*$ if and only if $\tilde{\pi}_i(\lambda_i^*) \le \tilde{\pi}_j(\lambda_j^*)$, $j \ne i$.

Lastly, for $i \ne j = 1,2$, when $0 \le x \le \underline{\lambda}_j^*$, let $\underline{\lambda}_i(x)$ be the demand rate of class-i customers such that the revenue of the firm, when only serving class i customers, is the same as that when the firm only serves class-j customers of demand rate x; that is, $\underline{\lambda}_i(x)$ is the unique solution of λ to the equation $\tilde{\pi}_i(\lambda) = \tilde{\pi}_j(x)$ over $\lambda \in [0, \lambda_i^*]$. Thus, by definition, $0 \le \underline{\lambda}_i(x) \le \lambda_i^*$ and $\tilde{\pi}_i(\underline{\lambda}_i(x)) = \tilde{\pi}_j(x)$. Note that $\tilde{\pi}_j(x)$ strictly increases from $\tilde{\pi}_j(0) = 0$ to $\min\{\tilde{\pi}_1(\lambda_1^*), \tilde{\pi}_2(\lambda_2^*)\}$ when x increases from 0 to $\underline{\lambda}_j^*$. It follows that $\underline{\lambda}_i(x)$ strictly increases from 0 to $\underline{\lambda}_i(\underline{\lambda}_j^*) = \underline{\lambda}_i^*$ when x increases from 0 to $\underline{\lambda}_j^*$. From Equation (5), we obtain $\underline{\lambda}_i(x)$ as

$$\underline{\lambda}_i(x) = \frac{\tilde{\pi}_j(x) + \mu w_i - \sqrt{(\tilde{\pi}_j(x) + \mu w_i)^2 - 2(d_i(1+c_v^2) + 2\mu w_i)\tilde{\pi}_j(x)}}{d_i(1+c_v^2) + 2\mu w_i}\mu. \tag{9}$$

The following lemma summarizes some important relationships among λ_i^*, $\underline{\lambda}_i^*$, and $\underline{\lambda}_i(x)$, which will be used in characterizing the firm's optimal pricing strategy.

LEMMA 4. *For $i \ne j = 1,2$, the following results hold:*

(a) $0 \le \underline{\lambda}_i^* \le \lambda_i^*$ and $\underline{\lambda}_i^* = \lambda_i^*$ if and only if $\tilde{\pi}_i(\lambda_i^*) \le \tilde{\pi}_j(\lambda_j^*)$;
(b) $\underline{\lambda}_i(x)$ is strictly increasing in x when $0 \le x \le \underline{\lambda}_j^*$ with $\underline{\lambda}_i(0) = 0$ and $\underline{\lambda}_i(\underline{\lambda}_j^*) = \underline{\lambda}_i^*$;
(c) $\underline{\lambda}_1(x)$ and $\underline{\lambda}_2(x)$ are inverse functions of each other; that is, $\underline{\lambda}_i(\underline{\lambda}_j(x)) = x$ for any $0 \le x \le \underline{\lambda}_i^*$.

With the notation defined above, we are ready to study the optimal pricing decision of the service firm. To start with, the following result provides a lower bound for the firm's optimal price.

PROPOSITION 1. *The firm's optimal price is lower bounded by* $\min\{p_1^*, p_2^*\}$.

It will be seen that the firm's optimal price critically depends on the parameter settings, especially the potential arrival rates of two classes of customers, that is, Λ_1 and Λ_2. However, Proposition 1 shows that the firm's optimal price has a lower bound which is independent of Λ_1 and Λ_2. Also note that Proposition 1 holds under both the dominating customers case and the non-dominating customers case.

We now begin to analyze the firm's optimal price under each and every parameter setting. As can be expected, the optimal pricing decision for the dominating customers case is different and simpler than that for the non-dominating customers case; thus in the following we present the results for the two cases separately.

4.1. Dominating Customers Case

We first consider the case of dominating customers, that is, $d_1 \le d_2$. To completely characterize the optimal pricing decision in various scenarios, we need the following lemma.

LEMMA 5. *When* $d_1 \le d_2$, *then* $p_2^* < p_1^*$, $\underline{\lambda}_1^* < \underline{\lambda}_2^* = \lambda_2^* < \lambda_1^*$, *and* $\underline{\lambda}_2'(x) > 1$ *when* $0 < x < \underline{\lambda}_1^*$.

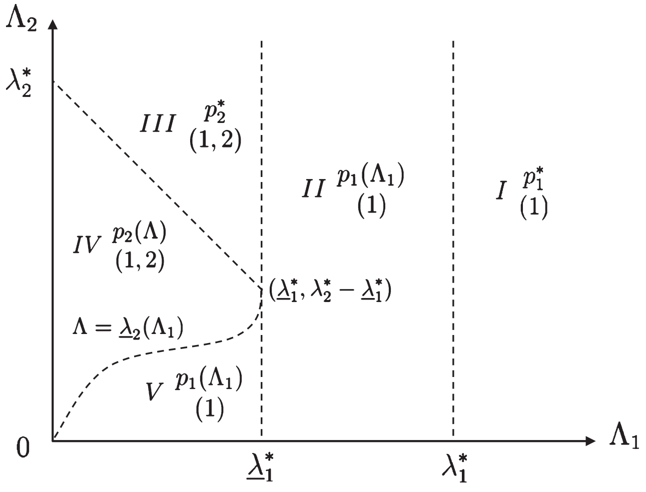
For convenience, we use p^{EQ} to represent the firm's optimal price and $(\theta_1^{EQ}, \theta_2^{EQ})$ to represent the customers' equilibrium probabilities of entering the service system. The following theorem fully characterizes the firm's optimal pricing strategy for the dominating customers case.

THEOREM 3 (*Optimal price for dominating customers case*). *When* $d_1 \le d_2$, *the optimal price of the firm depends on the market sizes of two classes of customers. Specifically,*

1. *if* $\Lambda_1 \ge \underline{\lambda}_1^*$, *then* $p^{EQ} = \max\{p_1^*, p_1(\Lambda_1)\}$ *and* $(\theta_1^{EQ}, \theta_2^{EQ}) = (\min\{1, \lambda_1^*/\Lambda_1\}, 0)$;
2. *if* $\Lambda_1 \le \underline{\lambda}_1^*$ *and* $\Lambda \ge \underline{\lambda}_2(\Lambda_1)$, *then* $p^{EQ} = \max\{p_2^*, p_2(\Lambda)\}$ *and* $(\theta_1^{EQ}, \theta_2^{EQ}) = (1, \min\{1, (\lambda_2^* - \Lambda_1)/\Lambda_2\})$;
3. *if* $\Lambda_1 \le \underline{\lambda}_1^*$ *and* $\Lambda \le \underline{\lambda}_2(\Lambda_1)$, *then* $p^{EQ} = p_1(\Lambda_1)$ *and* $(\theta_1^{EQ}, \theta_2^{EQ}) = (1, 0)$.

We illustrate the results in Theorem 3, including the optimal price and the customers being served, in Figure 1. In this figure, three dashed lines and one dashed curve divide the (Λ_1,Λ_2)-plane into five regions, where case 1 corresponds to regions (I) and (II), case 2 corresponds to regions (III) and (IV), and case 3 corresponds to region (V). There are two vertical lines in Figure 1, $\Lambda_1 = \underline{\lambda}_1^*$ and $\Lambda_1 = \lambda_1^*$, with $\underline{\lambda}_1^* < \lambda_1^*$ by Lemma 5. The third dashed line is $\Lambda_1 + \Lambda_2 = \lambda_2^*$, which intersects with $\Lambda_1 = \underline{\lambda}_1^*$ in the first quadrant because, by Lemma 5, $\lambda_2^* > \underline{\lambda}_1^*$. The curve, dividing regions (IV) and (V), is determined by the function $\Lambda_2 = \underline{\lambda}_2(\Lambda_1) - \Lambda_1$, which is strictly increasing in Λ_1 when $0 \le \Lambda_1 \le \underline{\lambda}_1^*$ by Lemma 5, and it starts from (0,0) and ends at $(\underline{\lambda}_1^*, \lambda_2^* - \underline{\lambda}_1^*)$ by Lemma 4. This shows that the curve intersects with the two straight lines at point $(\underline{\lambda}_1^*, \lambda_2^* - \underline{\lambda}_1^*)$.

Figure 1 Optimal Price for Dominating Customers Case


Λ_2
λ_2^*
III p_2^* $(1,2)$
II $p_1(\Lambda_1)$ (1)
I p_1^* (1)
IV $p_2(\Lambda)$ $(1,2)$
$(\underline{\lambda}_1^*, \lambda_2^* - \underline{\lambda}_1^*)$
$\Lambda = \underline{\lambda}_2(\Lambda_1)$
V $p_1(\Lambda_1)$ (1)
0
$\underline{\lambda}_1^*$
λ_1^*
Λ_1

The results in Theorem 3 can be interpreted as follows. When there are a large number of potential class-1 customers (i.e., $\Lambda_1 \ge \lambda_1^*$), the firm makes decision by completely ignoring the potential arrival rates of both classes of customers. The optimal price in this case is p_1^*, it is independent of the exact values of Λ_1 and Λ_2, and the firm only serves a subset of the class-1 customers; this is region (I). If Λ_1 is lower than λ_1^* but higher than $\underline{\lambda}_1^*$, the firm will still ignore class-2 customers, but will consider the potential arrival rate Λ_1 in its pricing decision. The optimal price in this case is $p_1(\Lambda_1)$, which is a function of Λ_1, and all class-1 customers are served; this is region (II). If Λ_1 is lower than $\underline{\lambda}_1^*$, but there are still a sufficient number of customers when adding all customers together, then the firm will consider both classes of customers in making its pricing decision and make sure that all class-1 customers are served. On the other hand, the firm does not want to lower its price by too much and will set the price to capture some of the class-2 customers; this is region (III). In regions (IV) and (V), the class-1 customers have a potential arrival rate below $\underline{\lambda}_1^*$, and there is not a sufficient customer base when all customers are combined. In these two regions, the firm does a trade-off analysis between serving only class-1 customers (by setting a sufficiently high price to turn away all class-2 customers) and serving all customers (by lowering the price). It turns out that in region (IV), the firm is better off to serve all customers with price $p_2(\Lambda)$, while in region (V), the firm is better off to turn away all the class-2 customers and only serve class-1 customers with price $p_1(\Lambda_1)$.

We now study the firm's optimal prices in different parameter regions in Figure 1. The optimal price in region (I) is p_1^*, which is independent of the arrival rates of both classes of customers. The optimal price for regions (II) and (V) is $p_1(\Lambda_1)$, which is strictly decreasing in the arrival rate of class-1 customers but independent of the arrival rate of class-2 customers. Since $p_1^* = p_1(\lambda_1^*)$, the firm's optimal price is decreasing in Λ_1 and Λ_2 in regions (I), (II), and (V). For region

(IV), the optimal price is $p_2(\Lambda_1 + \Lambda_2)$, which is strictly decreasing in $\Lambda_1 + \Lambda_2$, and the lowest price in this region is reached at $\Lambda_1 + \Lambda_2 = \lambda_2^*$, which is $p_2(\lambda_2^*) = p_2^*$. Thus, the firm's optimal price is decreasing in Λ_1 and Λ_2 in regions (III) and (IV). Note that the firm only serves class-1 customers in regions (I), (II), and (V), while it serves both classes of customers in regions (III) and (IV). Therefore, when the firm does not change the customer base being served, it will charge a lower optimal price when the potential arrival rate of either class of customers increases. This result is intuitive, since lowering the selling price is the only way for the firm to secure more demand from the potential demand pool.

From Figure 1, we can also easily observe that, for any fixed Λ_1, the firm's optimal price is decreasing in Λ_2, meaning that for the dominating customers case the firm will always charge a lower optimal selling price when there are more class-2 customers. However, this result is not true for class-1 customers. For any fixed positive Λ_2, the optimal price as a function of Λ_1 always decreases first, jumps up at one point, and then decreases until it finally drops to p_1^*. The intuitions are as follows. First, when Λ_1 is very small, it is not worthwhile for the firm to ignore class-2 customers; thus it will make the pricing decision to serve both classes of customers, and, as discussed above, the optimal price is decreasing in Λ_1. Second, since class-1 customers are more valuable to the firm for the dominating customers case, the firm will begin to ignore class-2 customers in setting its price when Λ_1 becomes large enough, as shown in Figure 1. Since at the switching value of Λ_1 the firm loses some class-2 customers when switching from serving both classes of customers to serving only class-1 customers, a price jump must be accompanied to maintain the firm's revenue. Finally, when Λ_1 becomes even larger, the firm will keep only serving class-1 customers and the optimal price will decrease in Λ_1 again.

We remark that there may exist multiple optimal prices for the boundary points of two adjacent regions in Figure 1. Note that the firm always has a unique optimal revenue under a given market structure (i.e., the relative values of Λ_1 and Λ_2). Then, the multiple optimal prices (if only) on the boundary points will inevitably result in multiple equilibrium customer markets that are served. For example, for any point (Λ_1,Λ_2) on the boundary curve between regions (IV) and (V), by Theorem 3, it is optimal for the firm to either set a price $p_1(\Lambda_1)$ and only serve all class-1 customers or set a price $p_2(\Lambda_1 + \Lambda_2)$ and serve all customers. Since the firm serves a smaller equilibrium market under the first price, it can be seen that $p_1(\Lambda_1)$ is higher than $p_2(\Lambda_1 + \Lambda_2)$. Similarly, for any point (Λ_1,Λ_2) on the boundary line between regions (II) and (III), it is optimal for the firm to either set a higher price $p_1(\Lambda_1)$ and only serve all class-1 customers or set a lower price p_2^* and serve all class-1 customers as well as some class-2 customers.

We now compare the firm's optimal revenues among different regions, and the results are reported in the following proposition. For comparison purposes we denote by $R^{(i)}$ the optimal revenue for an arbitrary point in region (i), which is clearly a function of both Λ_1 and Λ_2. When we write $R^{(i)} > R^{(j)}$, we mean that the firm's optimal revenue for any point in region (i) is larger than its optimal revenue for any point in region (j).

Proposition 2. *For the dominating customers case, the optimal revenues satisfy* (1) $R^{(I)} > R^{(II)} > R^{(III)} > R^{(IV)}$*; and* (2) $R^{(III)} > R^{(V)}$.

Note that the optimal revenues in some regions cannot always be compared in the above way. Nevertheless, we can show that the firm's optimal revenues in all of the five regions in Figure 1 are increasing in both Λ_1 and Λ_2; that is, the firm's optimal revenue is increasing in both Λ_1 and Λ_2. Thus, as the potential arrival rate of either class of customers increases, the firm will always be better off by obtaining a (weakly) higher optimal revenue. It is also interesting to note that for the dominating customers case, the firm's optimal revenue has an upper bound $\pi_1^*(p_1^*)$. Consequently, when the firm already has sufficient potential class-1 customers (i.e., $\Lambda_1 \geq \lambda_1^*$), adding more potential customers of any class will not improve the firm's optimal revenue.

4.2. Non-dominating Customers

In the case of non-dominating customers, that is, $d_1 > d_2$, the optimal pricing of the firm is more involved. In the following, the optimal pricing decisions are presented in four exclusive cases of system parameters according to the values of $\hat{p}, p_i^*$, and λ_i^*, i = 1,2, and they are sharply different from case to case. The results reveal that the firm should follow different pricing strategies in different regions of system parameters.

The following lemma will be used in characterizing the optimal prices in different regions of the system parameters.

Lemma 6. *When* $d_1 > d_2$*, if* $\hat{p} \geq p_1^*$*, then* $\hat{p} \geq p_2^*$*, and if* $\hat{p} \leq p_2^*$*, then* $\hat{p} \leq p_1^*$*; moreover,*

(a) if $\hat{p} \leq \min\{p_1^*, p_2^*\}$*, then* $p_1^*\lambda_1^* \geq p_2^*\lambda_2^*$*,* $\underline{\lambda}_1^* \leq \underline{\lambda}_2^* = \lambda_2^*$*, and* $\underline{\lambda}_2(x) \geq x$ *when* $0 \leq x \leq \underline{\lambda}_1^*$*;*

(b) if $\hat{p} \geq \max\{p_1^*, p_2^*\}$*, then* $p_1^*\lambda_1^* \leq p_2^*\lambda_2^*$*,* $\underline{\lambda}_1(\hat{\lambda}) = \underline{\lambda}_2(\hat{\lambda}) = \hat{\lambda} \leq \underline{\lambda}_2^* \leq \underline{\lambda}_1^* = \lambda_1^*$*,* $x \leq \underline{\lambda}_2(x) \leq \hat{\lambda}$ *when* $0 \leq x \leq \hat{\lambda}$ *and* $x \leq \underline{\lambda}_1(x) \leq \lambda_1^*$ *when* $\hat{\lambda} \leq x \leq \underline{\lambda}_2^*$*;*

(c) *if* $p_2^* < \hat{p} < p_1^*$, *then* $\hat{\lambda} < \underline{\lambda}_2^*$, $\underline{\lambda}_1(\hat{\lambda}) < \underline{\lambda}_1^*$, $x \le \underline{\lambda}_2(x) \le \lambda$ *when* $0 \le x \le \underline{\lambda}_1(\hat{\lambda})$, *and* $\hat{\lambda} \le \underline{\lambda}_2(x) \le \lambda_2^*$ *when* $\underline{\lambda}_1(\hat{\lambda}) \le x \le \underline{\lambda}_1^*$.

Note that the condition $p_2^* < \hat{p} < p_1^*$ in case (c) is equivalent to $\min\{p_1^*, p_2^*\} < \hat{p} < \max\{p_1^*, p_2^*\}$, since $p_1^* < \hat{p} < p_2^*$ is not a possible case by the first part of Lemma 6. Thus, the second part of Lemma 6 gives all possible cases on the relationship among $\hat{p}$, p_1^*, and p_2^*, and the corresponding relationships on the control parameters for each of the cases.

As it turns out, the structure of the optimal pricing strategy of the firm depends critically on the range of the system parameters. We are able to give the optimal pricing decision of the firm *in closed form* according to four distinct cases of the system parameters: (1) $\hat{p} \le \min\{p_1^*, p_2^*\}$; (2) $\hat{p} \ge \max\{p_1^*, p_2^*\}$; (3) $p_2^* < \hat{p} < p_1^*$ and $p_1^*\lambda_1^* \ge p_2^*\lambda_2^*$; and (4) $p_2^* < \hat{p} < p_1^*$ and $p_1^*\lambda_1^* < p_2^*\lambda_2^*$. The optimal pricing policy for case 1 is exactly the same as that of Theorem 3. That is, the firm's optimal pricing strategy for case 1 has the same structure as that for the dominating customers case. The reasons are as follows. From Proposition 1, the firm's optimal price is always at least $\min\{p_1^*, p_2^*\}$. Since $\hat{p} \le \min\{p_1^*, p_2^*\}$ under case 1, the firm's optimal price is always at least $\hat{p}$. Thus, it follows from Lemma 3 that class-1 customers always have more incentives to enter the system than do class-2 customers under the optimal price. Consequently, case 1 has the same structure on the optimal price as that for the dominating customers case. For all the other three cases, the firm's optimal price can be either higher or lower than $\hat{p}$, depending on the potential arrival rates of two classes of customers; thus either class of customers can have more incentive to enter the system than the other class.

In the following, we only present the results for case 2, and the results for the other two cases are given in the online Supporting Information for brevity.

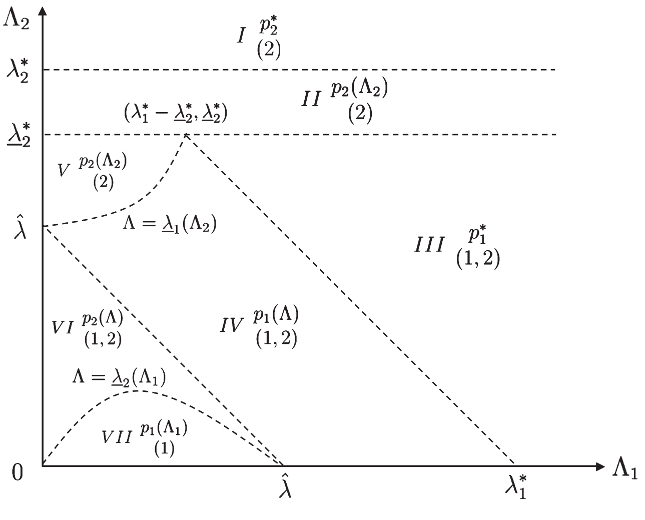
THEOREM 4 (*Optimal price for non-dominating customer case*). *When $d_1 > d_2$, the optimal price of the firm p^{EQ} and customers' equilibrium behavior $(\theta_1^{EQ}, \theta_2^{EQ})$ for case 2 is given according to the following subcases:*

1. *if* $\Lambda_2 \ge \underline{\lambda}_2^*$, *then* $p^{EQ} = \max\{p_2^*, p_2(\Lambda_2)\}$ *and* $(\theta_1^{EQ}, \theta_2^{EQ}) = (0, \min\{1, \lambda_2^*/\Lambda_2\})$;
2. *if* $\Lambda_2 \le \underline{\lambda}_2^*$ *and* $\Lambda_1 + \Lambda_2 \ge \max\{\hat{\lambda}, \underline{\lambda}_1(\Lambda_2)\}$, *then* $p^{EQ} = \max\{p_1^*, p_1(\Lambda_1 + \Lambda_2)\}$ *and* $(\theta_1^{EQ}, \theta_2^{EQ}) = (\min\{1, (\lambda_1^* - \Lambda_2)/\Lambda_1\}, 1)$;
3. *if* $\hat{\lambda} \le \Lambda_2 \le \underline{\lambda}_2^*$ *and* $\Lambda_1 + \Lambda_2 \le \underline{\lambda}_1(\Lambda_2)$, *then* $p^{EQ} = p_2(\Lambda_2)$ *and* $(\theta_1^{EQ}, \theta_2^{EQ}) = (0, 1)$;
4. *if* $\underline{\lambda}_2(\Lambda_1) \le \Lambda_1 + \Lambda_2 \le \hat{\lambda}$, *then* $p^{EQ} = p_2(\Lambda_1 + \Lambda_2)$ *and* $(\theta_1^{EQ}, \theta_2^{EQ}) = (1, 1)$;
5. *if* $\Lambda_1 \le \hat{\lambda}$ *and* $\Lambda_1 + \Lambda_2 \le \underline{\lambda}_2(\Lambda_1)$, *then* $p^{EQ} = p_1(\Lambda_1)$ *and* $(\theta_1^{EQ}, \theta_2^{EQ}) = (1, 0)$.

The results in Theorem 4 are illustrated in Figure 2, where four dashed lines and two dashed curves divide the (Λ_1,Λ_2)-plane into seven regions. The four dashed lines are $\Lambda_2 = \lambda_2^*, \Lambda_2 = \underline{\lambda}_2^*$, $\Lambda_1 + \Lambda_2 = \hat{\lambda}$, and $\Lambda_1 + \Lambda_2 = \lambda_1^*$, and the two curves are $\Lambda_1 + \Lambda_2 = \underline{\lambda}_1(\Lambda_2)$ and $\Lambda_1 + \Lambda_2 = \underline{\lambda}_2(\Lambda_1)$. The curve $\Lambda_1 + \Lambda_2 = \underline{\lambda}_2(\Lambda_2)$ is strictly decreasing in Λ_1 on $0 \le \Lambda_1 \le \lambda_1^* - \underline{\lambda}_2^*$; it starts at $(0, \hat{\lambda})$ and ends at $(\lambda_1^* - \underline{\lambda}_2^*)$, thus intersecting with the lines $\Lambda_2 = \underline{\lambda}_2^*$ and $\Lambda_1 + \Lambda_2 = \lambda_1^*$ at $(\lambda_1^* - \underline{\lambda}_2^*, \underline{\lambda}_2^*)$. The second curve, $\Lambda_1 + \Lambda_2 = \underline{\lambda}_2(\Lambda_1)$, stays in the first quadrant on $0 \le \Lambda_1 \le \hat{\lambda}$, and it starts at (0,0) and ends and intersects with the line $\Lambda_1 + \Lambda_2 = \hat{\lambda}$ at $(\hat{\lambda}, 0)$. Note that in Figure 2, we have displayed subcase 1 in regions (I) and (II) according to $\Lambda_2 \ge \lambda_2^*$ and $\underline{\lambda}_2^* \le \Lambda_2 \le \lambda_2^*$, respectively, and we have displayed subcase 2 in regions (III) and (IV) according to $\Lambda_1 + \Lambda_2 \ge \lambda_1^*$ and $\Lambda_1 + \Lambda_2 \le \lambda_1^*$, respectively.

Therefore, the optimal price of the firm is given in seven regions of the system parameters in Figure 2. In regions (I) to (V), the optimal price is lower than $\hat{p}$; thus class-2 customers have more incentive to enter the system than class-1 customers, while in regions (VI) and (VII), the optimal price is higher than $\hat{p}$, and thus class-1 customers have more incentive to enter the system than do class-2 customers. In region (I), there are sufficient potential class-2 customers; thus the firm, focusing on class-2 customers in this case for its revenue optimization, ignores class-1 customers as well as the exact arrival rate of class-2 customers and sets the price at p_2^*. In this subcase only a subset of class-2 customers are served and no class-1 customers are served; the firm sets the price as if there are infinitely many class-2 customers in the market. In region (II), there are no sufficient class-2 customers, and the

Figure 2 Optimal Policy when $\hat{p} \ge \max\{p_1^*, p_2^*\}$


Λ_2
λ_2^*
$\underline{\lambda}_2^*$
$\hat{\lambda}$
0
I p_2^* (2)
II $p_2(\Lambda_2)$ (2)
$(\lambda_1^* - \underline{\lambda}_2^*, \underline{\lambda}_2^*)$
V $p_2(\Lambda_2)$ (2)
$\Lambda = \underline{\lambda}_1(\Lambda_2)$
III p_1^* (1, 2)
VI $p_2(\Lambda)$ (1, 2)
IV $p_1(\Lambda)$ (1, 2)
$\Lambda = \underline{\lambda}_2(\Lambda_1)$
VII $p_1(\Lambda_1)$ (1)
$\hat{\lambda}$
λ_1^*
Λ_1

firm sets the price at $p_2(\Lambda_2)$, which depends on the potential arrival rate of class-2 customers Λ_2. In this region, the firm still completely ignores class-1 customers but sets a price to catch all class-2 customers. In region (III), the potential arrival rate of class-2 customers is so low that the firm has to reach out to class-1 customers, but, still, since there are a relatively large total number of customers in the market, the firm sets its price only to catch a subset of the class-1 customers. In regions (IV) and (V), there are not enough class-2 customers and neither is there a sufficient, though still quite some, total number of customers in the market. In this case, the firm does a trade-off analysis between serving only class-1 customers (by setting a high price to turn away all class-2 customers) and serving all customers (by lowering the price). It turns out that in region (IV) the firm is better off to serve only class-2 customers with price $p_2(\Lambda_2)$, while in region (V) it is optimal to serve all customers with price $p_1(\Lambda)$. Finally, in regions (VI) and (VII), the total number of customers in the market is very low, and in this case, the firm does a trade-off analysis between serving all customers (by setting a low price) and serving only class-1 customers (by setting a high price to turn away all class-2 customers). It turns out that in region (VI), the firm is better off to serve all customers with price $p_2(\Lambda)$, while in region (VII), it is better off to totally ignore the class-2 customers and serve all but only class-1 customers with price $p_1(\Lambda_1)$. It is interesting to note from Figure 2 that when Λ_2 is sufficiently small, the firm serves both classes of customers if Λ_1 is very small, only class-1 customers if Λ_1 becomes larger, and both classes of customers again when Λ_1 is large enough.

We illustrate Theorem 4 using the following numerical example, where the system parameters are given by

$$v_1 = 3, \quad v_2 = 2.9, \quad d_1 = 0.08, \quad d_2 = 0.03, \quad \mu = c_v = 1.$$

Since $w_1 = 2.92 > w_2 = 2.87$ and $d_1 > d_2$, this example belongs to the non-dominating customers case. In addition, it can be computed that

$$\hat{p} = 2.84, \quad \hat{\lambda} = 0.5, \quad p_2^* = 2.6914, \quad p_1^* = 2.6536,$$

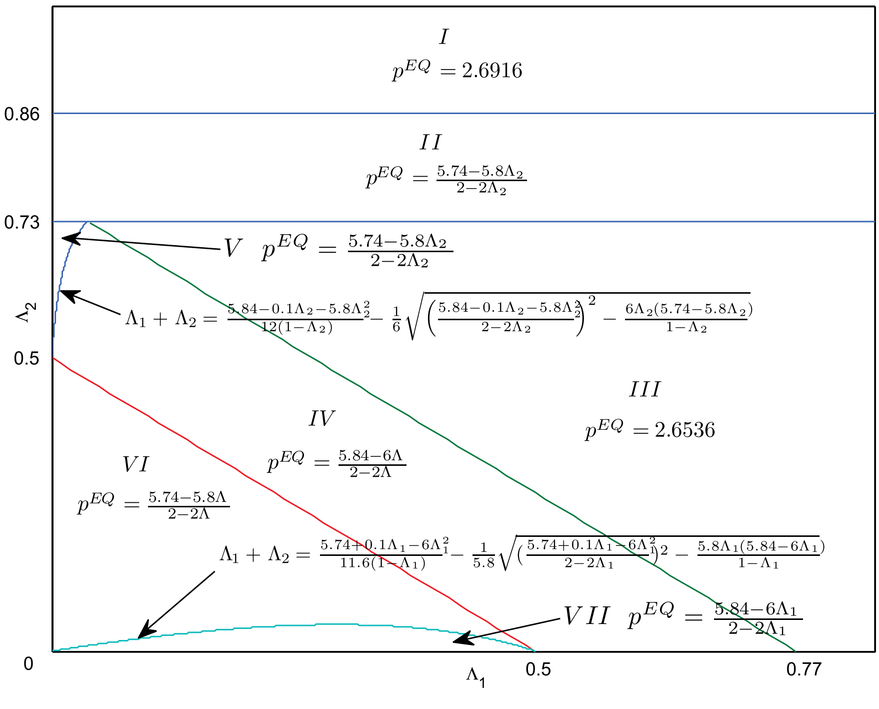
and

$$\lambda_2^* = 0.8562, \quad \lambda_1^* = \underline{\lambda}_1^* = 0.7691, \quad \underline{\lambda}_2^* = 0.7320.$$

Since $\hat{p} > \max\{p_1^*, p_2^*\}$, this example belongs to case 2. Thus, the firm's optimal prices p^* for this example are given by Theorem 4, and the values of p^* in different regions of (Λ_1,Λ_2) are displayed in Figure 3.

As seen from Figure 3, when Λ_2 is sufficiently small, then there are two values of Λ_1 that satisfy $\underline{\lambda}_1(\Lambda_1) = \Lambda_2$, call them $a_1^*(\Lambda_2)$ and $a_2^*(\Lambda_2)$, respectively, and suppose $a_1^*(\Lambda_2) < a_2^*(\Lambda_2)$. The firm serves both classes of customers when $\Lambda_1 < a_1^*(\Lambda_2)$; however, when $\Lambda_1 > a_1^*(\Lambda_2)$ but $\Lambda_1 < a_2^*(\Lambda_2)$, then the firm only serves class-1 customers, but it starts to serve both classes of customers again when $\Lambda_1 > a_2^*(\Lambda_2)$.

Figure 3 Numerical Example



REMARK 2. (1) $p_1^* = p_2^* = \hat{p}$ belongs to both case 1 and case 2, and the results are consistent; (2) if $p_2^* < \hat{p} < p_1^*$ and $p_1^*\lambda_1^* = p_2^*\lambda_2^*$, it belongs to both case 3 and case 4. In this scenario the results are consistent except when $\Lambda_1 \geq \lambda_1^*$ and $\Lambda_2 \geq \lambda_2^*$. In this region, there are two different optimal prices p_1^* and p_2^* for the firm. Both prices are optimal for the firm. When the price is p_1^*, only a subset of class-1 customers are served; while when the price is p_2^*, only a subset of class-2 customers are served. Both prices, however, give the same expected revenue for the firm. The same remark applies to several other scenarios when (Λ_1,Λ_2) falls at a boundary of two regions with different optimal prices.

We now study the firm's optimal prices under different system parameters for non-dominating customers case. First, we find that in all cases and within any range of system parameters, when the firm does not change the customer base being served, it will charge a lower optimal price when the potential arrival rate of either class of customers increases. This result is consistent with that of the dominating customers case. Second, the optimal price in Theorem 4 is not monotone in either Λ_1 or Λ_2. Similar comments apply to cases 3 and 4 presented in the online Supporting Information (and the optimal price is not monotone in Λ_1 but decreasing in Λ_2 for cases 3 and 4).

We finally compare the firm's optimal revenues in different regions. The comparison results of the revenues for case 1 is the same as that in Proposition 2. The results for case 2 is presented in the following, and the results for cases 3 and 4 are given in the online Supporting Information.

PROPOSITION 3. *The firm's optimal revenues in different regions in Figure* 2 *satisfy*

1. $R^{(I)} > R^{(II)} > R^{(III)} > R^{(k)} > R^{(VI)}$; *and*
2. $R^{(k)} > R^{(VII)}$, $k = IV,V$.

The same as in the dominating customers case, after checking all four cases 1–4, we find that the firm's optimal revenue in the non-dominating customers case is also increasing in both Λ_1 and Λ_2, with an upper bound of $\max\{\pi_1^*(p_1^*), \pi_2^*(p_2^*)\}$. Thus, combining the results for dominating and non-dominating customer cases, we conclude that the firm's optimal revenue function is increasing in the potential arrival rate of either class of customers, and the maximum revenue of the firm is always upper bounded by $\max\{\pi_1^*(p_1^*), \pi_2^*(p_2^*)\}$.

5. Conclusion

The message taken away from this study is that the potential market structure plays a key role for the firm; thus, it has to be fully investigated and understood before the firm makes the pricing decision. Different system parameters can lead to completely different pricing strategy for the firm. Depending on the range of system parameters, which are explicitly determined by the system primitive data, the firm may choose to focus only on one class of customers, a subset of a class of customers, or a combination of different classes of customers, and may decide to completely ignore some customer segments in setting its price. Another insight obtained from our analysis is that the optimal selling price of a firm is not always monotone in the potential market size or the arrival rates of potential customers. That is, when more customers are in the market it does not imply that the firm will increase its price to maximize its revenue; the firm may actually reduce its price. This is because the firm uses pricing as a strategy to shape up its market structure. Our findings are consistent with some interesting phenomena often observed in service systems such as restaurants, retail stores, etc., in which some focus only on high end consumers, some focus only on low end consumers, while some others catch consumers from both ends. Our results show that this is due to the heterogeneity in the different classes of consumers, and it is the firm's optimal strategy, on price, waiting time, quality of service, etc., that leads to this structure.

In this study, we focus on the case with two classes of customers. When there are more than two classes of customers, it is conceivable that the optimal pricing strategy of the firm will be more complicated. Nevertheless, some of the results can still be generalized. For example, Proposition 1 and the results discussed at the end of section 4 (i.e., the optimal price is lower bounded by the minimum of p_i^* over i, the optimal revenue function is increasing in the arrival rate of each class of customers, and the maximum revenue is upper bounded by the maximum of $\pi_i(p_i^*)$ over i) are not accidental, and they can be shown to be satisfied with an arbitrary number of classes of customers. In addition, as seen from our analysis, when customers have dominating relationships, then the model and results are extendable to either countable many or a continuum number of customer classes. In the model considered in this study, it would imply that w_i is increasing in i while d_i is decreasing in i.

Several other extensions are possible. The service system in this study is modeled as an $M/G/1$ queueing system. The results can be extended to the multiple-server case as long as we use any of the well-known queueing delay approximations for an $M/G/m$ queue. For example, we can use the delay approximation of Nozaki and Ross (1978) or the diffusion approximation of Whitt (1993) in the analysis of the customer utility functions. The structure of the optimal pricing for the multiple-server case turns out to be similar to the single-server case, though the detailed results vary. One possible extension is that the different classes of customers have different service time distributions, and another interesting problem is to relax the pricing scheme to allow differentiated prices for different classes of customers (but the service discipline remains FCFS), and investigate the improvement in the firm's optimum revenue when compared with the model of this study. These will be left as future research topics.

Acknowledgments

The authors are grateful to the department editor, the senior editor, and the anonymous reviewers for their constructive comments and suggestions, which have helped us to significantly improve both the quality and the exposition of the paper. In particular, Propositions 2 and 3 on the comparison of the firm's optimal revenues in different regions are added based on the reviewers' suggestions. The research of Wenhui Zhou is supported by NSFC under 71271089, 71131003, 71090403, and 70871040, and the research of Xiuli Chao and Xiting Gong is supported in part by the NSF under CMMI-0927631 and CMMI-1131249.

References

Afanasyev, M., H. Mendelson. 2010. Service provider competition: Delay cost structure, segmentation and cost advantage. *Manuf. Serv. Oper. Manag.* **12**(2): 213–235.

Afèche, P., H. Mendelson. 2004. Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Manage. Sci.* **50**(7): 869–882.

Armony, M., M. Haviv. 2003. Price and delay competition between two service providers. *Eur. J. Oper. Res.* **147**(1): 32–50.

Bell, C. E., S. Stidham, Jr. 1983. Individual versus social optimization in the allocation of customers to alternative servers. *Manage. Sci.* **29**(7): 831–839.

Bradford, R. 1996. Pricing, routing, and incentive compatibility in multiserver queues. *Eur. J. Oper. Res.* **89**(2): 226–236.

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

Chen, H., Y.-W. Wan. 2003. Price competition of make-to-order firms. *IIE Trans.* **35**(9): 817–832.

De Vany, A. 1976. Uncertainty, waiting time, and capacity utilization: A stochastic theory of product quality. *J. Polit. Econ.* **84**(3): 523–541.

Edelson, N. M., D. K. Hilderbrand. 1975. Congestion tolls for Poisson queueing processes. *Econometrica* **43**(1): 81–92.

Gilland, W. G., D. P. Warsing. 2009. The impact of revenue-maximizing priority pricing on customers delay costs. *Decis. Sci.* **40**(1): 89–120.

Hassin, R. 1986. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* **54**(5): 1185–1195.

Hassin, R., M. Haviv. 1997. Equilibrium threshold strategies: The case of queues with priorities. *Oper. Res.* **45**(6): 966–973.

Hassin, R., M. Haviv. 2002. *To Queue or Not to Queue*. International Series in Operations Research and Management Science, volume 59, Kluwer Publishing, Boston.

Kleinrock, L. 1975. *Queueing Systems, Volume I: Theory*. John Wiley & Sons, New York.

Knudsen, N. C. 1972. Individual and social optimization in a multi-server queue with a general cost-benefit structure. *Econometrica* **40**(3): 515–528.

Lippman, S., S. Stidham. 1977. Industrial versus social optimization in exponential congestion systems. *Oper. Res.* **25**(2): 233–247.

Mandelbaum, A., A. Sakov, S. Zeltyn. 2001. Empirical analysis of a call center. Technical Report, Technion–Israel Institute of Technology. Available at http://ie.technion.ac.il/serveng/References/ccdata.pdf (accessed date October 13, 2013).

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

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

Nozaki, S. A., S. M. Ross. 1978. Approximation in finite-capacity multi-server queues with Poisson arrivals. *J. Appl. Probability* **15**(4): 826–834.

Rao, S., E. R. Petersen. 1998. Optimal pricing of priority services. *Oper. Res.* **46**(1): 46–56.

Stidham, S. JR. 2009. *Optimal Design of Queueing Systems*. Chapman and Hall/CRC, Boca Raton, FL.

Stidham, S. JR. 2011. *Optimal Control of Queueing Systems*. Chapman and Hall/CRC, Boca Raton, FL.

Van Mieghem, J. A. 2000. Price and service discrimination in queueing systems: Incentive compatibility of Gcμ scheduling. *Manage. Sci.* **46**(9): 1249–1267.

Whitt, W. 1993. Approximations for the GI/G/m queue. *Prod. Oper. Manag.* **2**(2): 114–161.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1: Proofs.