SECOND DEGREE PRICE DISCRIMINATION FOR AN *M/G/1* SYSTEM

Yong J. Kim, Dept. of MIS, Konkuk University, Seoul, Korea
yongjkim@kkucc.konkuk.ac.kr

1. Introduction

Consider an $M/G/1$ system with N classes having heterogeneous service time distributions. We can observe that each user's expected delay cost, $\theta ST(\underline{\lambda})$, is a random variable, that a user with a higher θ value is more impatient than users with lower θ values, and that he is willing to pay a higher premium (access charge) in exchange for a given reduction in waiting time. We define the system value as the sum of the *expected consumer's surplus* (CS) and *producer's surplus* (PS) of all service levels determined by the system manager. To avoid escalating complexity, we assume that service time distributions are homogeneous across the user population and that the total arrival rate to the system is Λ, but the rate of users actually entering the system is $\lambda < \Lambda$.

In this setting, prioritizing the user population according heterogeneous needs can be a policy improving system efficiency as well users' value: a typical user can select his priority at the time he submits a job to the system; he is charged a higher fixed cost in exchange for higher priority. Because users have heterogeneous goals and deadlines at different points in time, individual users' θ can be viewed as randomly distributed and so can aggregate user demand. Therefore, the system manager will try to maximize the net system value by utilizing the information on the distribution of users' valuation.

This type of price discrimination is called *second-degree price discrimination* in the theory of industrial organization. Offering a menu is also known as a *preference-revealing contract.* Because the contract does not require a user to select any particular service class, it is customary for the system manager to add incentive-compatible constraints (ICC) which induce the user to select the service class that is optimal from the system-wide net-value-maximization point. The system manager's problem is then to devise the optimal (p_i, ST_i)'s (service levels) in price *and* quality (sojourn time) domain.

In this proposal, we evaluate the critical assumptions used in past research in order to emphasize the soundness of our approach. In particular, we question the empirical availability of certain model parameters and argue that informational requirements on those parameters can be practically prohibitive and the number of priority classes may not be decided a priori. We consider a single class $M/G/1$ system in the pursuit of the optimal price; which is equal to the negative queuing delay externality. We show that the optimal access charge formula of our model is comparable to the one from Mendelson (1985). Finally, we discuss the welfare consequence of adopting priority systems.

2. System Manager's Informational Requirements

We briefly review the informational problems of implementing the priority queuing model of [MW90].

The net system value of [MW90] is defined as

$$\max_{\underline{\lambda}} \left\{ \sum_{j=1}^{N} V_j(\lambda_j) - TC(\underline{\lambda}) \right\} = \max_{\underline{\lambda}} \left\{ \sum_{j=1}^{N} V_j(\lambda_j) - \sum_{j=1}^{N} v_j \lambda_j ST_j(\underline{\lambda}) \right\}$$

The first order condition for optimal arrival rate is

$$V_i\,(\lambda_i) = v_i ST_i(\underline{\lambda}) + \sum_{j=1}^{N} v_j \lambda_j \frac{\partial ST_j(\underline{\lambda})}{\partial \lambda_j}. \qquad (1)$$

The delay cost term

$$TC(\underline{\lambda}) = \sum_{j=1}^{N} v_j \lambda_j ST_j(\underline{\lambda}) \qquad (2)$$

shows inherent negative externalities: if an additional job arrives, other jobs will experience longer queuing delays, and so the tagged job must be made to bear the cost of this negative externality. Otherwise, the system will be more congested than optimal from the system manager's point of view. Extending the Pigouvian tax for nonpreemptive $M/M/1$ with N classes, [MW90] proved that the optimal price for class-i users should be

$$p_i^* = \sum_{j=1}^{N} v_j \lambda_j^* \frac{\partial ST_j(\underline{\lambda}^*)}{\partial \lambda_i} \qquad (3)$$

where $\underline{\lambda}^* = (\lambda_1^*, \lambda_2^*, ..., \lambda_N^*)$ is the optimal traffic maximizing the net value of problem (1).

The derivation of (3) can be done intuitively. First, differentiating (1) with respect to λ_i, one can obtain the social marginal cost of having one more class-i job as

$$\frac{\partial TC(\underline{\lambda}^*)}{\partial \lambda_i} = \sum_{j=1}^{N} v_j \lambda_j^* \frac{\partial ST_j(\underline{\lambda}^*)}{\partial \lambda_i} + v_i ST_i(\underline{\lambda}^*), \qquad (4)$$

which is greater than the individual marginal cost, $v_i ST_i(\underline{\lambda}^*)$, by the amount equal to the right-hand-side of (4). When the system manager imposes (4) on class-i users, the class-i users' perception of marginal cost is equated to the system manager's view --the right-hand-side of (4) -- which leads the system to an optimal state.

However, p_i^* is not incentive-compatible if service time distributions are heterogeneous: if $p_i^* + v_i ST_i(\underline{\lambda}^*) > p_j^* + v_j ST_j(\underline{\lambda}^*)$, a class-$i$ user will select class-j priority. Thus [MW90] proposed an incentive-compatible pricing scheme, which is *Priority-and Time- Dependent* (PTD), and proved that the PTD pricing scheme is both optimal and incentive-compatible. In short, [MW90] is an improvement on the previous works in terms of its informational requirement: the system manager is assumed to know about the heterogeneous service time distributions, but does not need to differentiate individual users' class.

Having overcome the incentive-compatibility issue, the system manager still encounters at least two other fundamental information problems when he wants to implement the net-value-maximizing procedure of [MW90]. First, it is essential that he possess full information on the value function $V_i(\lambda_i)$ for solving the non-linear equations of (1).

An implicit assumption of [MW90] is that the system manager has full information on v_i and c_i for all $i = 1,...,N$. Because the v_i / c_i rule for nonpreemptive priority $M/G/1$ stipulates that higher priority should go to the higher v_i / c_i ratio, the success of implementing the model depends on the system manager's ability to correctly observe v_i and c_i for each class. The next example illustrates the potential complications associated with measuring v_i and c_i for $i = 1,...,N$.

EXAMPLE 1.

Suppose that an $M/M/1$ system has two user classes -- class 1 and class 2-- but the system manager can only observe the input and output from the system, and does not know a priori that there are two distinct user classes. What the system manager can observe from the server are $c_A^{(k)}$, the k-th moment ($k \geq 1$) of the service time distribution, and the aggregate arrival rate λ_g where

$$\lambda_g = \lambda_1 + \lambda_2 . \qquad (5)$$

The service process is characterized by the hyperexponential distribution: i.e.,

$$c_A^{(k)} = k!(\alpha c_1^k + (1-\alpha)c_2^k) \qquad (6)$$

where $\alpha = \lambda_1/(\lambda_1 + \lambda_2)$. If the system manager is fully aware that there are exactly two user classes in the system whose service time distributions are exponential with mean c_1 and c_2, then he can solve (5) and (6) simultaneously for c_1, c_2, λ_1, and λ_2. Because we have four variables to solve for, we need (9) and three equations from (6). In other words, λ_g, c_A, $c_A^{(2)}$, and $c_A^{(3)}$ need to be observed in order to solve for c_1, c_2, λ_1, and λ_2. With the correct values of c_1, c_2, λ_1, and λ_2, the system manager can implement a priority $M/M/1$ and, if necessary, priority- and time-dependent pricing schemes. However, the validity of this approach depends on the fact that there are exactly two user classes, and that the system manager knows this fact. All the system manager can measure is the aggregate arrival rate and the output data from the server.

The complexity of the problem increases if there are more than two classes in the system. For example, suppose that the manager knows that there are three classes in the system precisely. The aggregate arrival rate λ_g and the k-th moment of service time distribution observed by the system manager are

$$\lambda_g = \lambda_1 + \lambda_2 + \lambda_3 \qquad (7)$$

and $c_A^{(k)} = k!(\alpha_1 c_1^k + \alpha_2 c_2^k + \alpha_3 c_3^k) \qquad (8)$

where $\alpha_i = \lambda_i/(\lambda_1 + \lambda_2 + \lambda_3)$ for $i = 1,2,3$. Solving (7) and (8) as simultaneous equations for c_1, c_2, c_3, λ_1, λ_2, and λ_3 is not impossible, but it is necessary to observe c_A, $c_A^{(2)}$, $c_A^{(3)}$, $c_A^{(4)}$ and $c_A^{(5)}$ in addition to λ_g.

Worse yet, if the system manager does not know the exact number of user classes in the system, it is not possible to set up equations like (7) and (8). The fundamental questions are *why* there should be *N* user classes in the system and *how* the system manager can identify individual classes having heterogeneous service time distributions. Although one may argue that the basic premise of [MW90] is the existence of such information for classifying the users of heterogeneous service time distributions, it remains to ask whether such information is practically available for implementing a priority queue. The benefit of priority pricing, letting the user "put his money where his mouth is", can be diluted due to excessive informational requirements.

There are at least three salient features of the model presented in the following sections. First, it is another application of price-discrimination theory. Second, the system manager does not need to know delay cost per unit time parameters and service time distributions of separate user classes. As we showed in Example 1, such information may be hard to obtain. In contrast, our model starts from a simple $M/G/1$ non-priority queue and assumes that the manager can use the collected data such as service time distribution and system arrival rate in order to offer optimal prices for non-priority and priority queues. Third, the model actually complements the previous works, including [MW90]: it utilizes the noble incentive-compatibility pricing with reduced informational requirements and yet still maintains discrete priorities.

3. Optimal Pricing for Nonpreemptive Priority *M/G/1*

We assume that the system manager wants to maintain two classes in a nonpreemptive priority $M/G/1$ system and calculate the optimal access charges, when θ, the delay cost per unit time parameter, is a continuous variable. We assume that class-1 service has higher priority. Let $R(\theta)$, p_i, and $ST_i(\lambda_1,\lambda_2)$ denote the value of a job to a type θ user, the fixed access charge levied on a class-*i* job, and the expected sojourn time of non-preemptive priority $M/G/1/P$=2, respectively. Because $ST_1(\lambda_1,\lambda_2) < ST_2(\lambda_1,\lambda_2)$ and the service time distribution is homogeneous, the system manager should offer p_1 and p_2 such that $p_1 > p_2$. Given the choice between (p_1, ST_1) and (p_2, ST_2), a rational user will select the minimum of $p_1 + \theta ST_1$ and $p_2 + \theta ST_2$. Users with higher θ (i.e., more impatient) will select class-1 service while those with lower θ will select class-2 service.

Let θ_M be the crossover value, where

$$p_1 + \theta_M ST_1 = p_2 + \theta_M ST_2 .$$

We define the utility function $U_i(\theta)$ of type θ users who choose class-*i* service by

$$U_i(\theta) = R(\theta) - p_i - \theta ST_i(\lambda_1,\lambda_2) \quad (i = 1,2).$$

Individual rationality (IR) requires that a type θ user will join the system if

$$\max\{U_1(\theta), U_2(\theta)\} \geq 0 .$$

In other words, if he selects class-*i* service,

$$p_i + \theta ST_i < p_j + \theta ST_j \text{ and}$$

$$R(\theta) - p_i - \theta ST_i > 0 . \quad (9)$$

In order to make our derivation simple, we assume that the feasible range of θ satisfying (9) is continuous in the range $[\theta_L..\theta_H]$, where θ_L and θ_H denote the lowest and highest threshold values of θ. The continuous range of $\theta \in [\theta_L..\theta_H]$ implies that

$$R(\theta_H) - p_1 - \theta_H ST_1(\lambda_1,\lambda_2) = 0 \text{ and}$$

$$R(\theta_L) - p_2 - \theta_L ST_2(\lambda_1,\lambda_2) = 0 .$$

The steady state class-*i* arrival rates are then determined by θ_L, θ_M and θ_H:

$$\lambda_2 = \Lambda(F(\theta_M) - F(\theta_L)) \text{ and } \lambda_1 = \Lambda(F(\theta_H) - F(\theta_M)) .$$

where $F(\theta)$ is the distribution function for θ. The expected consumer surplus and producer surpluses are

$$CS = \lambda_2 \int_{\theta_L}^{\theta_M} \frac{(R(x) - p_2 - xST_2)f(x)}{F(\theta_M) - F(\theta_L)} dx$$

$$+ \lambda_1 \int_{\theta_M}^{\theta_H} \frac{(R(x) - p_1 - xST_1)f(x)}{F(\theta_H) - F(\theta_M)} dx$$

$$= \Lambda \int_{\theta_L}^{\theta_M} \left(R(x) - p_2 - xST_2\right) f(x) dx$$

$$+\Lambda\int_{\theta_M}^{\theta_H}\left(R(x)-p_1-xST_1\right)f(x)dx$$

$$PS=\lambda_2\int_{\theta_L}^{\theta_M}\frac{p_2 f(x)}{F(\theta_M)-F(\theta_L)}dx+\lambda_1\int_{\theta_M}^{\theta_H}\frac{p_1 f(x)}{F(\theta_H)-F(\theta_M)}dx$$

$$=\Lambda\int_{\theta_L}^{\theta_M}p_2 f(x)dx+\Lambda\int_{\theta_M}^{\theta_H}p_1 f(x)dx. \qquad (10)$$

Adding *PS* and *CS*, the net system value is defined as

$$SV=\Lambda\int_{\theta_L}^{\theta_M}(R(x)f(x)dx$$

$$-\Lambda\int_{\theta_L}^{\theta_M}xST_2 f(x)dx-\Lambda\int_{\theta_M}^{\theta_H}xST_1 f(x)dx. \qquad (11)$$

The optimization problem is then

$$\max_{p_1,p_2}\left\{\int_{\theta_L}^{\theta_M}(R(x)f(x)dx-\int_{\theta_L}^{\theta_M}xST_2 f(x)dx-\int_{\theta_M}^{\theta_H}xST_1 f(x)dx\right.$$

subject to

$$R(\theta_H(p_1,p_2))=p_1+\theta_H(p_1,p_2)ST_1(\lambda_1(p_1,p_2),\lambda_2(p_1,p_2))$$
$$R(\theta_L(p_1,p_2))=p_2+\theta_L(p_1,p_2)ST_2(\lambda_1(p_1,p_2),\lambda_2(p_1,p_2))$$
$$\lambda_2(p_1,p_2)=\Lambda(F(\theta_M(p_1,p_2))-F(\theta_L(p_1,p_2)))$$
$$\lambda_1(p_1,p_2)=\Lambda(F(\theta_H(p_1,p_2))-F(\theta_M(p_1,p_2)))$$
$$p_1+\theta_M(p_1,p_2)ST_1(p_1,p_2)=p_2+\theta_M ST_2(p_1,p_2).$$

We use $\lambda_1(p_1,p_2)$, $\lambda_2(p_1,p_2)$, $\theta_L(p_1,p_2)$, $\theta_M(p_1,p_2)$ and $\theta_H(p_1,p_2)$ interchangeably with λ_1, λ_2, θ_L, θ_M and θ_H, respectively.

Solving the first order conditions for (11), we obtain an optimal solution for p_1 and p_2:

$$p_1^*=\Lambda\left(\frac{\partial ST_2}{\partial\lambda_1}\int_{\theta_L}^{\theta_M}xf(x)dx+\frac{\partial ST_1}{\partial\lambda_1}\int_{\theta_M}^{\theta_H}xf(x)dx\right)$$

$$p_2^*=\Lambda\left(\frac{\partial ST_2}{\partial\lambda_2}\int_{\theta_L}^{\theta_M}xf(x)dx+\frac{\partial ST_1}{\partial\lambda_2}\int_{\theta_M}^{\theta_H}xf(x)dx\right)$$

The result is comparable to that of [MW90].

4. Concluding Remarks

The main motivation of this proposal arises from the strict informational requirement on v_i and $V_i(\cdot)$ without which the system manager cannot successfully implement the priority queuing model in previous chapters. Although our analysis here is limited to 2-class priority system, the results can be extended for *N*-class non-preemptive priority $M/G/1$.

One critical assumption of our analysis is the homogeneous service time distribution across all θ satisfying individual rationality. The assumption is based on the observation that the system manager can monitor only service times of finished jobs, not knowing which class each individual job belongs to. Thus the system manager partitions the users into an arbitrary number of classes, using the model of Section 4. However, after introducing non-preemptive priority $M/G/1$ service, the system manager can observe different (heterogeneous) service time distributions in each class, which expands his information on the empirical service time distributions of subgroups within the user population. After obtaining this information, he may be able to introduce a priority- and time-dependent (PTD) pricing scheme similar to the one of [MW90].

We note that there are fundamental differences between the two PTD pricing schemes. [MW90] assume that individual classes are disjoint, and a user's service time distribution is determined by which class he belongs to. Having to know about individual service time distributions causes stringent informational requirements as previously noted. The price discrimination model of this paper defines the users' service time distribution which is based on the class a user selects and the service time demanded by the user's job.

References

Mendelson, H., and S. Whang, "Optimal Incentive-compatible Priority Pricing for the $M/M/1$ Queue," Operations Research 38(5) (Sep.-Oct. 1990), 870-83

Whang, S., "Pricing Computer Systems: Incentive, Information and Queuing Effects," Ph.D. Dissertation, W.E. Simon Graduate School of Business Administration, University of Rochester, 1988

Wilson, R. B., "Economic Theories of Price Discrimination and Product Differentiation: A Survey," Technical notes, Stanford Business School, Stanford University, July 1991.

Wilson, R. B., "Efficient and Competitive Rationing," *Econometrica*, 57(1) (January 1989), 1-40