

# Optimal Pricing for a Service Facility

Serhan Ziya\*, Hayriye Ayhan\*\*, Robert D. Foley\*\*

\*Department of Statistics and Operations Research  
University of North Carolina  
CB# 3180, 213 Smith Building, Chapel Hill, NC 27599-3180  
E-mail: [ziya@email.unc.edu](mailto:ziya@email.unc.edu)

\*\*School of Industrial and Systems Engineering  
Georgia Institute of Technology  
765 Ferst Drive, Atlanta, GA 30332-0205  
E-mail: [hayhan@isye.gatech.edu](mailto:hayhan@isye.gatech.edu), [rfoley@isye.gatech.edu](mailto:rfoley@isye.gatech.edu)

January, 2004

### **Abstract**

This paper investigates optimal pricing policies for a service facility modeled as a queueing system. Arriving customers are accepted if they are willing to pay the price charged by the service provider and if there is room in the waiting area. The capacity of the waiting area can be either finite or infinite. We determine expressions for the optimal prices that maximize the service provider's long-run average profit, and we prove some structural results on the optimal policies exploring their relationships with the customers' willingness to pay and system parameters such as service speed and waiting room capacity. We show that the optimal price is not necessarily higher for systems where customers are willing to pay more, and the relationship between the optimal price and waiting room capacity depends on customer demand and other system parameters. Under certain assumptions, we give necessary and sufficient conditions for the optimal price to be an increasing or decreasing function of the waiting room capacity.

# 1 Introduction

The goal of this paper is to determine the best price to charge customers in a service facility and to investigate how the best price changes as assumptions about the system change. By “best,” we mean the price that maximizes long run average profit per unit time. Our “system” is assumed to be a single server queue, and each customer has her own cut-off point for how much she is willing to pay for service. If the price charged is more than the amount the customer is willing to pay, the customer will not purchase service. If the advertised price is less than or equal to a customer’s cut-off point and there is room in the waiting area when the customer arrives, the customer joins the queue.

We restrict attention to the simplest pricing strategy in which there is one fixed, static price for all customers. Clearly, there is a trade-off. A higher price yields more revenue per customer, but fewer customers purchase service. Surprisingly, even if all customers are willing to pay more, it may be optimal to charge less; see the example in Section 6. We do prove a result that agrees with the conventional wisdom of charging more to customers who are willing to pay more, but the result requires that an additional condition holds.

The primary contribution of this paper is the analysis of optimal pricing decisions for finite waiting room capacity systems. In the literature, there is significant amount of work on pricing in queueing systems; however, pricing for blocking systems (where the waiting room capacity is finite) has not received much attention. We review some of this literature below.

In most of the earlier work, although there is no external restriction, the number of customers waiting is bounded due to delay sensitivity of the customers. In contrast, in our model, the service provider puts a bound on the number of customers waiting in the system (or there is a physical constraint that limits the number of customers). Customers are not delay sensitive; however, the restriction on the number of waiting customers guarantees a certain service level for the customers.

Our interests are also different from the work in the literature. The earlier work mostly concentrates on congestion control, using prices to achieve social optimality, the effects of waiting costs on the optimal policies etc. On the other hand, our objective is to investigate the effects of changes in the system parameters (e.g. waiting room capacity) and customers’ willingness-to-pay on the optimal prices that maximize the service provider’s profits. One of the important results of this paper is on the relationship between the optimal price and waiting room capacity. Under certain conditions, we show that optimal price is increasing or decreasing in the waiting room capacity depending on the customer demand for the system.

Even though we are mainly interested in finite capacity systems, as it is shown in the following sections, pricing decision for infinite capacity systems and pricing decision for finite capacity systems are closely related. More precisely, optimal prices for infinite capacity systems provide bounds on the optimal prices for finite capacity systems. Therefore, we also include infinite capacity systems in our analysis.

To our knowledge, Naor [22] is the first to analyze the use of price as a regulatory tool with quantitative arguments. There is significant amount of work that follows Naor’s paper extending and generalizing the basic principles and ideas which are first pointed out quantitatively in his paper. The fundamental conclusion/idea of this body of research is that in the presence of waiting costs, if the arriving customers are interested in maximizing

their own benefits, the customer arrival rate is suboptimal for the organization as a whole; however, prices can be used to induce the self-optimizing customers to act in a socially optimal way. Some examples of this avenue of work are Yechiali [30], Knudsen [12], Edelson and Hildebrand [6] and Lippman and Stidham [16].

Another important paper is by Mendelson [19]. Mendelson analyzed a queueing system within a microeconomic framework. In Mendelson's model, capacity (the term capacity is used to refer to service capacity) is a long term decision variable. The author studies the effects of queueing costs on the pricing and capacity policies. See also Dewan and Mendelson [5], Stidham [27], Ha [9], and Mendelson and Whang [20] for work that follows Mendelson [19].

There is also significant amount of work on dynamic pricing policies for queueing systems. Some examples are Low [17], [18], Johansen [10], Lippman [15], and Paschalidis and Tsitsiklis [23]. The paper by Paschalidis and Tsitsiklis is particularly relevant. The authors show that in some cases static pricing policies are asymptotically optimal and provide computational results showing that optimal static prices can perform almost as well as optimal dynamic policies.

Almost all the work mentioned so far (except [17] and [23]) considers models where there is no external restriction on the number of customers waiting. Courcoubetis and Reiman [4] analyze optimal pricing decisions for a loss system in an asymptotic regime where the capacity and the potential load of the system go to infinity. Their results are in agreement with some of the results given in our paper. Miller and Buckman [21] are interested in optimal transfer prices for a service department modelled as an M/M/s/s queueing system. Sumita, Masuda, and Yamakawa [28] analyze loss externalities and find optimal transfer prices for a finite buffer system. There is also some work that considers optimal pricing policies for loss networks with QoS (Quality of Service) guarantees. For examples, see Kelly [11], Wang, Peha, and Sirbu [29], and Lanning, Massey, Rider, and Wang [14].

The remainder of the paper is organized as follows. Section 2 describes the model in more detail and introduces notation. In Section 3, we derive expressions for the expected profit per unit time. In Section 4, we discuss the existence of the optimal prices and compare the optimal prices for infinite capacity systems with the optimal prices for finite capacity systems. In Section 5, we give expressions for optimal prices under certain assumptions. Section 6 discusses the relationships between the optimal prices and customers' willingness-to-pay. In Section 7, we investigate how the optimal price changes as the customer demand rate or the service rate changes. Section 8 contains our results on the relationships between optimal prices and waiting room capacity. Section 9 contains concluding remarks. Proofs of our results are given in the Appendix.

## 2 Model Description

We consider a single server service system. The queue may have either a finite or infinite capacity waiting area, and we let  $m \leq \infty$  denote the maximum number of customers in the system at any time. Customers who arrive when the system is full are lost.

The arrival process is allowed to be quite general. We simply let  $N(t)$  be the random number of customer arrivals during  $(0, t]$  where  $0 \leq N(t) < \infty$ , and assume that  $N(t)/t$

converges to a strictly positive finite number  $\Lambda$ , almost surely. We call  $\Lambda$ , the maximum arrival rate. (If the reader would like a particular example to keep in mind, a Poisson process with rate  $\Lambda$  is a good choice.)

Since the objective is to maximize long run average profit, we need some assumptions on the cost to serve customers. We assume that the total cost to serve a group of customers is simply a linear function of the number of customers served, and we let  $c$  denote the variable cost per customer. The amounts that customers are willing to pay for are assumed to be independent, identically distributed non-negative random variables. Let  $y$  denote the mark-up over the variable cost  $c$  per customer; hence, the price would be the sum of the mark-up and the variable cost  $c$ . Let  $F(y)$  denote the proportion of customers willing to pay a mark-up of at most  $y$ . The cumulative distribution function  $F(\cdot)$  will be referred to as the willingness-to-pay distribution. Also, let  $\bar{F}(y)$  be defined as

$$\bar{F}(y) = \lim_{\varepsilon \downarrow 0} 1 - F(y - \varepsilon).$$

Then, the probability that an arriving customer is willing to join the system is  $\bar{F}(y)$  where  $y$  is the mark-up. Note that when  $F(\cdot)$  is a continuous function,  $\bar{F}(y) = 1 - F(y)$  for all  $y$ .

The distribution  $F(\cdot)$  is assumed to be absolutely continuous with density  $f(\cdot)$  and support  $(\alpha, \beta)$  with  $0 \leq \alpha < \beta \leq \infty$  and  $F(y) < 1$  for all  $y < \beta$ . The hazard rate function for  $F(\cdot)$  will be denoted by  $r(y) = f(y)/\bar{F}(y)$  for  $y \in [\alpha, \beta)$ . Unless otherwise stated, the mean of  $F(\cdot)$  is assumed to be finite. Henceforth, we only consider mark-ups  $y$  that satisfy

$$\alpha \leq y < \beta. \tag{1}$$

Let  $N(y, t)$  be the number of customers who are willing to pay (a mark-up of) at least  $y$  and arrive during  $(0, t]$ . If we let  $\lambda(y)$  denote the arrival rate of customers willing to pay  $y$  or more, then

$$\lambda(y) = \Lambda \bar{F}(y). \tag{2}$$

Note that  $\lambda(0) = \Lambda$ , and we are assuming that all arrivals are willing to pay at least the variable cost  $c$  (or, equivalently, that we ignore all customers who would not pay a price of at least  $c$ ). In this paper, without loss of generality, we assume that  $c = 0$  so that the mark-up  $y$  is also the price. However, all the results given in the following sections can easily be generalized to the case with strictly positive service costs ( $c > 0$ ).

Service times are assumed to be i.i.d. random variables with c.d.f.  $G(\cdot)$  and mean  $\mu^{-1}$ . Thus, the service rate is  $\mu$ , and we assume that  $0 < \mu < \infty$ . Service protocol is assumed to be First-Come-First-Served (FCFS) although most of our results hold for more general service disciplines. The service process, arrival process, and amounts that the customers are willing to pay are assumed to be mutually independent.

To avoid complications in the construction of the queue length process, we make a certain assumption on the order of events occurring at some time  $t$ . We assume that at any time point  $t$ , first all scheduled departures are allowed to occur. Then, new arrivals are allowed to enter the system. If any of these new arrivals have zero service time and are next to be serviced, they are allowed to depart allowing for more arrivals. Continuing in this way, all events that can occur at time  $t$  are allowed to occur. Note that this ‘‘event order’’ assumption we make is the same as in Sonderman [26].

For each price  $y \geq 0$ , the number of customers in the system forms a G/GI/1/m queue with arrival rate  $\lambda(y)$ , service distribution  $G(\cdot)$  with mean  $\mu^{-1}$ , capacity  $m \leq \infty$ , and traffic intensity  $\rho(y) = \lambda(y)/\mu$ . Furthermore, when  $m < \infty$  and the price is  $y$ , we assume that the long-run fraction of customers blocked - which we refer to as the blocking probability - is a constant  $B_N(\lambda(y), m)$ . More precisely, let  $N^B(y, t)$  be the number of blocked customers during  $(0, t]$  among the customers who are willing to pay at least the price  $y$ . We assume that

$$\lim_{t \rightarrow \infty} N_m^B(y, t)/N(y, t) = B_N(\lambda(y), m) \text{ a.s.} \quad (3)$$

Note that the subscript  $N$  is to indicate that the blocking probability is a function of the arrival process. The blocking probability depends upon the price  $y$  through the arrival rate  $\lambda(y)$ . Under our assumptions, setting a price  $y$  corresponds to randomly deleting each potential customer from the arrival process  $N$  with probability  $1 - \lambda(y)/\Lambda$ . When the arrival process is stationary and ergodic, a sufficient condition for the blocking probability ( $B_N(\lambda(y), m)$ ) to exist is that with probability 1 at most one customer departs at any time and departure time of a served customer does not coincide with an arrival; see Franken et. al. [8].

We consider two different alternatives for collecting payments:

- All customers pay upon joining the system (payment at acceptance points),
- All customers pay upon service completion (payment at departure points).

Fix the arrival process  $N$ , willingness-to-pay distribution  $F(\cdot)$ , service time distribution  $G(\cdot)$ , and let  $R_a(y, m)$  be the long-run average revenue per unit time with price  $y$  and system capacity  $m \leq \infty$  assuming customers pay upon acceptance into the system. Similarly, define  $R_d(y, m)$  except that customers pay upon completing service. For  $m < \infty$ , it turns out that  $R_d(y, m) = R_a(y, m)$  so we will simply write  $R(y, m)$ . The objective is to determine the price  $y$  that maximizes  $R_d(y, m)$  for each  $m$  and also the price  $y$  that maximizes  $R_a(y, m)$  for each  $m$ .

We let  $Y_m^*$  be the set of optimal prices when the system capacity is  $m < \infty$ ,  $Y_a^*$  be the set of optimal prices when the capacity is infinite and customers pay at acceptance points, and  $Y_d^*$  be the set of optimal prices when the capacity is infinite and customers pay at departure points. When the optimal prices are unique, we use  $y_m^*$ ,  $y_a^*$  and  $y_d^*$  to denote the respective unique optimal prices.

Finally, we let  $y^\circ = \sup\{y : \rho(y) = 1\}$ . In other words,  $y^\circ$  is the largest price for which the arrival rate is equal to the service rate when such a price exists. Note that if  $\Lambda/\mu < 1$  then no such price exists and  $y^\circ = -\infty$ . On the other hand, if  $\Lambda/\mu \geq 1$ , then  $y^\circ$  exists and lies in the support of  $F(\cdot)$ .

### 3 Objective Functions

In this section, we derive expressions for  $R_a(y, \infty)$ ,  $R_d(y, \infty)$ , and  $R(y, m)$ , the objective functions for the three different models we are interested in.

### 3.1 When $m = \infty$ and payments at acceptance points— $R_a(y, \infty)$

For the infinite capacity system and under the assumption that customers pay at acceptance points, the objective function is easy to derive. Since every customer who is willing to pay at least  $y$  is accepted to the system and each customer pays  $y$ , we have:

$$R_a(y, \infty) = \lim_{t \rightarrow \infty} \frac{yN(y, t)}{t} = y\lambda(y). \quad (4)$$

Obviously, in this setting, assumptions on the service times are irrelevant.

### 3.2 When $m = \infty$ and payments at departure points— $R_d(y, \infty)$

Under the assumption that customers pay at departures and  $m = \infty$ , the objective function can be expressed in two parts depending on the value of  $\rho(y)$ . Since the customers pay as they leave the system, we are interested in the departure rate from the system. Let  $N^d(y, t)$  denote the total number of customers who departed (and therefore paid for their service) from the system until time  $t$ . Then, it can be shown that (see Sigman [25])

$$\lim_{t \rightarrow \infty} \frac{N^d(y, t)}{t} = \begin{cases} \mu & \text{if } \rho(y) \geq 1 \\ \lambda(y) & \text{if } \rho(y) < 1 \end{cases}.$$

Thus, the objective function can be defined as

$$R_d(y, \infty) = \begin{cases} y\mu & \text{if } \rho(y) \geq 1 \\ y\lambda(y) & \text{if } \rho(y) < 1 \end{cases}. \quad (5)$$

### 3.3 When $m < \infty$ and payments at either acceptance or departure points— $R(y, m)$

If the capacity  $m$  is finite, the long run average reward per unit time is the same regardless of whether payments occur at acceptance points or at departure points since the difference in the amount of revenue under the two schemes at any time is smaller than  $my$ . This means that while deriving the objective function, we can assume either payment alternative. We derive two different expressions for  $R(y, m)$  by considering the two different payment schemes, and we will use both expressions. First, by considering payments at acceptance points,

$$\begin{aligned} R(y, m) &= \lim_{t \rightarrow \infty} \frac{y(N(y, t) - N_m^B(y, t))}{t} \\ &= y \lim_{t \rightarrow \infty} \frac{N(y, t)}{t} \frac{N(y, t) - N^B(y, t)}{N(y, t)}. \end{aligned}$$

Thus, using (3), we have

$$R(y, m) = y\lambda(y)[1 - B_N(\lambda(y), m)]. \quad (6)$$

We now derive a second expression for  $R(y, m)$  by assuming that customers pay as they leave the system. Let  $S(y, t)$  be the total time the server is busy until time  $t$  and let  $S_n$  denote the service time for the  $n$ th customer. We use  $N_m^d(y, t)$  to denote the total number

of customers departed (completed service) until time  $t$  given that the capacity is  $m$  and price is  $y$ . Note that  $\lim_{t \rightarrow \infty} N_m^d(y, t)/t$  exists since  $N(y, t) - N_m^B(y, t) - m \leq N_m^d(y, t) \leq N(y, t) - N_m^B(y, t)$ . Thus, we have

$$\begin{aligned} \sum_{n=1}^{N_m^d(y, t)} S_n &\leq S(y, t) \leq \sum_{n=1}^{N_m^d(y, t)+1} S_n \\ \frac{\sum_{n=1}^{N_m^d(y, t)} S_n}{N_m^d(y, t)} &\leq \frac{S(y, t)}{N_m^d(y, t)} \leq \frac{\sum_{n=1}^{N_m^d(y, t)+1} S_n}{N_m^d(y, t)+1} \frac{N_m^d(y, t)+1}{N_m^d(y, t)}. \end{aligned}$$

From the strong law of large numbers,

$$\lim_{t \rightarrow \infty} \frac{S(y, t)}{N_m^d(y, t)} = \mu^{-1}. \quad (7)$$

Then, we can derive  $R(y, m)$  as follows:

$$R(y, m) = \lim_{t \rightarrow \infty} \frac{y N_m^d(y, t)}{t} = y \lim_{t \rightarrow \infty} \frac{\mu S(y, t)}{t}.$$

Since we know the last limit exists, define  $\pi_0(y, m) = 1 - (\lim_{t \rightarrow \infty} S(y, t)/t)$ , which is the long run proportion of time that the system is empty; hence  $1 - \pi_0(y, m)$  is the long run proportion of time that the server is working, and we have our second expression

$$R(y, m) = y\mu[1 - \pi_0(y, m)]. \quad (8)$$

## 4 Existence of Optimal Prices and Infinite Capacity vs. Finite Capacity Systems

In this section, we will be concerned with the existence of optimal prices and we will give an ordering result for the optimal prices for finite and infinite capacity systems. As it will be shown in the following, optimal prices may not exist in some cases. However, the assumptions on  $F(\cdot)$  given in Section 2 ensure that optimal prices exist for the infinite capacity systems. For the finite capacity systems, the optimal prices exist if  $B_N(\lambda(y), m)$  (or  $\pi_0(y, m)$ ) is a continuous function of  $y$ .

Proposition 4.1 states that for the infinite capacity systems, the optimal prices exist, and if there is a unique optimal price for the infinite capacity system with customers paying at acceptance points, then this optimal price is a lower bound on any optimal price for any finite capacity system.

**Proposition 4.1** *For the infinite capacity systems, optimal prices exist ( $Y_a^* \neq \emptyset$  and  $Y_d^* \neq \emptyset$ ). Furthermore, if an optimal price also exists for a system with capacity  $m < \infty$  ( $Y_m^* \neq \emptyset$ ) and  $y_a^* \in Y_a^*$  is the unique optimal price for the infinite capacity system under payments at acceptance points, then any optimal price for the  $m$  capacity system is at least as large as  $y_a^*$  ( $y_a^* \leq \inf\{y : y \in Y_m^*\}$  for any  $m < \infty$ ).*

The key to the proof of Proposition 4.1 is the fact  $B_N(\lambda(y), m)$  is non-increasing in  $y$ , which follows from the main result of Ziya, Ayhan and Foley [32]. If the blocking term  $B_N(\lambda(y), m)$

is strictly decreasing in  $y$ , it can be shown that Proposition 4.1 holds even when there is more than one optimal price for the infinite capacity system. In that case, the result holds for any optimal price of  $R_a(y, \infty)$ .

The blocking term  $B_N(\lambda(y), m)$  is not known in general, which makes it impossible to find the optimal prices for finite capacity systems in most cases. However, Proposition 4.1 suggests that even if it is not possible to find an optimal price for the finite capacity system, the facility manager can at least find a lower bound on any of the optimal prices by solving the infinite capacity problem.

The proof of Proposition 4.1 uses the assumption that  $F(\cdot)$  has a finite mean to prove the existence of the optimal price for the infinite capacity systems. If  $F(\cdot)$  has an infinite mean, then the optimal price either exists and is finite, or it does not exist. As an example, consider the distribution function  $F(y) = 1 - \frac{1}{y+\varepsilon}$  for  $0 < 1 - \varepsilon \leq y < \infty$ , which has an infinite mean. It can be shown that if  $\varepsilon < 0$ , then  $y_a^* = 1 - \varepsilon$ . On the other hand, if  $\varepsilon > 0$ , then the optimal price does not exist, since for any price  $y$ , there is always a better price which is greater than  $y$  and  $y = \infty$  is not optimal. Note that for  $\varepsilon = 0$ , any  $y \in [1, \infty)$  is optimal.

## 5 Optimal Price Expressions

In this section, we are interested in deriving conditions and expressions for unique optimal prices. To ensure uniqueness, we will make an assumption on the function

$$e(y) = yr(y) \text{ for } y \in [\alpha, \beta).$$

**Assumption IPE (Increasing Price Elasticity):**  $e(y)$  is strictly increasing for  $y \in [\inf\{y : e(y) \geq 1\}, \beta)$ .

It turns out that the function  $e(y) = yr(y)$  is the price elasticity of the demand  $\lambda(y) = \Lambda\bar{F}(y)$ . To be more precise,

$$e(y) = - \lim_{\Delta y \rightarrow 0} \frac{\frac{\lambda(y+\Delta y) - \lambda(y)}{\lambda(y)}}{\frac{\Delta y}{y}}.$$

Hence, having  $e(y)$  increasing is equivalent to having the price elasticity function increasing. Then, assumption IPE is equivalent to the assumption that price elasticity is increasing over the range of prices for which demand is elastic. (We say that demand is elastic for a certain price if the price elasticity for that price is at least 1.) Under IPE then, demand is inelastic for low prices and elastic for higher prices.

According to the terminology of Lariviere and Porteus [13], IPE is the same as the distribution function  $F(\cdot)$  having an IGFR (increasing generalized failure rate). As Lariviere and Porteus indicate this assumption is satisfied by many widely used continuous distribution functions. It is obviously satisfied for distributions with a non-decreasing failure rate  $r(\cdot)$  but also for many other distributions due to the factor  $y$  in  $yr(y)$  (e.g. uniform and Weibull distributions).

Under IPE, we can show that there are unique optimal prices for infinite capacity systems and also for finite capacity systems under exponentially distributed interarrival and service times.

**Proposition 5.1**

- (i) Under IPE,  $R_a(y, \infty)$  and  $R_d(y, \infty)$  both have unique optimal prices. That is  $Y_a^* = \{y_a^*\}$  and  $Y_d^* = \{y_d^*\}$ , where

$$\begin{aligned} y_a^* &= \inf[y : e(y) \geq 1], \\ y_d^* &= \begin{cases} y_a^* & \text{if } \Lambda/\mu < 1 \\ \max\{y_a^*, y^o\} & \text{if } \Lambda/\mu \geq 1 \end{cases} \end{aligned} .$$

For  $\Lambda/\mu \geq 1$ ,  $\max\{y_a^*, y^o\} = y^o$  if  $e(y^o) \geq 1$  (or if  $\rho(y_a^*) \geq 1$ ); otherwise,  $\max\{y_a^*, y^o\} = y_a^*$ .

- (ii) Suppose that  $R(y, m)$  has a unique local maximum and  $\pi_0(y, m)$  is differentiable in  $y$ . Then, the optimal price,  $y_m^*$  can be defined as

$$y_m^* = \inf[y : \frac{y\pi_0'(y, m)}{1 - \pi_0(y, m)} \geq 1]$$

where  $\pi_0'(y, m)$  is the derivative of  $\pi_0(y, m)$  w.r.t.  $y$ .

- (iii) Suppose that interarrival and service times have exponential distributions. Then, under IPE,  $R(y, m)$  has a unique optimal price; that is,  $Y_m^* = \{y_m^*\}$ . Furthermore,

$$y_m^* = \inf[y : e(y)\gamma_m(y) \geq 1]$$

where

$$\gamma_m(y) = \begin{cases} \frac{1+m(\rho(y))^{m+1}-(m+1)(\rho(y))^m}{(1-(\rho(y))^{m+1})(1-(\rho(y))^m)} & \text{if } \rho(y) \neq 1 \\ \frac{1}{2} & \text{if } \rho(y) = 1 \end{cases} .$$

We have used Assumption IPE for proving uniqueness. This assumption ensures that the revenue rate function  $y\lambda(y)$  is quasi-concave over the interval where it is strictly positive. In the pricing and revenue management literature, several other assumptions are made on this “revenue function” in order to have analytically tractable models. A common property of these assumptions is that each ensures that the “revenue function” is quasi-concave. These various assumptions are compared in Ziya, Ayhan and Foley [31].

## 6 Optimal Prices and Stochastic Ordering

In this section, we try to answer the following question: Should customers who are willing to pay more be charged more? Although the answer to this question might appear to be yes, we will show that it is not always true. This section contains an example of two identical systems A and B except that the  $n$ th arrival to B is always willing to pay more than the corresponding arrival to A. Even though it might appear that higher prices should be set in System B, we will give an example where A should have the higher price.

First, if  $F_B(y) \leq F_A(y)$  for all  $y \in (-\infty, \infty)$  then  $F_A$  is smaller than  $F_B$  with respect to stochastic dominance, i.e.  $F_A \leq_{st} F_B$  (see Shaked and Shanthikumar [24]).

**Example:** Let  $F_A(y) = \frac{y}{2.2}$  for  $y \in [0, 2.2]$  and  $F_B(y) = \ln(y)$  for  $y \in [1, e]$ . One can check to see that  $F_A \leq_{st} F_B$ , and the corresponding functions  $e_A(y)$  and  $e_B(y)$  are strictly increasing, which implies that IPE holds. Let  $U_1, U_2, \dots$  be iid uniform  $(0, 1)$  random variables, and suppose that the  $n$ th customer to A is willing to pay  $F_A^{-1}(U_n)$  while the  $n$ th customer to B is willing to pay  $F_B^{-1}(U_n)$ . Since  $F_B^{-1}(U_n) > F_A^{-1}(U_n)$ , the  $n$ th customer to B is willing to pay more than the  $n$ th customer to A. However, using Proposition 5.1, it can be shown that in System A  $y_a^* = 1.1$  whereas in System B  $y_a^* = 1$ .

If we turn our attention to the optimal price when customers pay at departure, the relationship between  $y_d^*$  for the two systems depends on the value of  $\Lambda/\mu$ . There exists some  $\kappa$  such that if  $\Lambda/\mu > (<)\kappa$ , then  $y_d^*$  for System B is strictly larger (smaller) than in A. It can be shown in this example that  $\kappa$  is roughly 1.10535. Similarly, the relationship between the optimal prices for the finite capacity system also depends upon the value of  $\Lambda/\mu$ . As an example, let  $m = 1$ . Then, it can be shown that if  $\Lambda/\mu > (<)\theta$ , then  $y_1^*$  in System B is strictly greater (smaller) than the one in System A. For this example,  $\theta$  is approximately equal to 0.18294.

This counter-intuitive result shows that it is not always better to charge higher prices to customers who are willing to pay more when by “pay more” we mean stochastic dominance. However, if by “pay more”, we mean that willingness-to-pay distribution functions are stochastically ordered with respect to hazard rate, then it is possible to prove that the optimal prices will also be ordered in the same direction. This result is given in Proposition 6.1 but first, we define hazard rate ordering.

Consider two distribution functions  $F_A$  and  $F_B$ . If  $F_A$  and  $F_B$  are both defined over non-negative sets, absolutely continuous, with hazard rate functions  $r_A$  and  $r_B$ , respectively, and  $r_A(y) \geq r_B(y)$  for  $y \geq 0$  then  $F_A$  is smaller than  $F_B$  with respect to the hazard rate ordering, i.e.  $F_A \leq_{hr} F_B$  (see Shaked and Shanthikumar [24]).

**Proposition 6.1** *Consider two systems with willingness-to-pay distribution functions  $F_A$  and  $F_B$ , respectively where  $F_A \leq_{hr} F_B$ . Then, under IPE,  $y_a^*$  ( $y_d^*$ ) in System A is less than or equal to  $y_a^*$  ( $y_d^*$ ) in System B. Furthermore, if interarrival and service times are exponentially distributed, the same holds for  $y_m^*$  for  $m < \infty$ .*

From Proposition 6.1, we know that having hazard rate order between the willingness-to-pay distributions suffices to ensure that optimal prices are ordered in the same direction for infinite capacity systems and also for finite capacity systems if interarrival and service times are exponentially distributed. The counterexample we provided at the beginning of this section shows that usual stochastic order is not sufficient. However, one might wonder whether some additional conditions would imply such an order. For example, suppose that all the customers are willing to pay an additional  $s$  dollars. Would that imply that optimal price also increases by  $s$  dollars, or at least increases? The following proposition answers these questions.

**Proposition 6.2** *Consider two systems with willingness-to-pay distribution functions  $F_A$  and  $F_B$ , respectively where  $F_A(y) = F_B(y + s)$  for  $\alpha_A \leq y \leq \beta_A$ . Then, under IPE,  $y_a^*$  ( $y_d^*$ ) in System B is greater than or equal to  $y_a^*$  ( $y_d^*$ ) and less than or equal to  $y_a^* + s$  ( $y_d^* + s$ ) in System A. Furthermore, if interarrival and service times are exponentially distributed, the same holds for  $y_m^*$  for  $m < \infty$ .*

According to Proposition 6.2, if the customers are willing to pay an additional  $s$  dollars, then the optimal price is also higher. Furthermore, the new optimal price is at most  $s$  dollars higher. To show that the optimal price may increase by something other than  $s$  dollars, consider the following example. Suppose that we have infinite capacity, customers pay at acceptance times and the willingness-to-pay distribution is Uniform  $(0, 100)$ . Then, the optimal price is \$50. If each customer were willing to pay \$10 more, then the willingness-to-pay distribution would be Uniform  $(10, 110)$ , and the optimal price would be \$55. Note that even though each customer is willing to pay \$10 more, the optimal price only increases by \$5.

## 7 Optimal Prices, Customer Arrival Rate and Service Rate

In this section, we are interested in how the optimal prices change as the customer demand rate  $\Lambda$  and service rate  $\mu$  change. Having a stochastically larger willingness-to-pay distribution function implies that customer demand for each price is also higher. However, as shown in the previous section, this does not imply that the optimal price is also higher. Another way to have a demand function larger for each price is to have a larger customer arrival rate  $\Lambda$ . One might wonder if a larger  $\Lambda$  implies a larger optimal price. Similarly, suppose that the current server is replaced by a faster one. How are the optimal prices influenced by such a change? One would expect  $\Lambda$  and  $\mu$  to have opposite effects on the optimal price, and this is indeed the case.

We know from Proposition 5.1 (i) that  $y_a^*$  is not a function of  $\Lambda$  and  $\mu$ . Therefore, changes in these rates have no effect on  $y_a^*$ . However,  $y_d^*$  and  $y_m^*$  (for exponential interarrival and service times) are functions of both  $\Lambda$  and  $\mu$ . To be more precise, they are functions of  $\Lambda/\mu$  and as stated in Proposition 7.1, they are non-decreasing in  $\Lambda/\mu$ .

**Proposition 7.1** *Consider two systems with demand rates  $\Lambda_A$  and  $\Lambda_B$ , service rates  $\mu_A$  and  $\mu_B$ , respectively. Suppose that  $\Lambda_B/\mu_B > \Lambda_A/\mu_A$  and IPE holds. Then,  $y_d^*$  in System A is less than or equal to  $y_d^*$  in System B. Furthermore, if interarrival and service times are exponentially distributed, the same holds for  $y_m^*$  for  $m < \infty$ .*

Assuming that the service speed remains the same, the scenario considered in Proposition 7.1 corresponds to a case where the customer population increases in such a way that the proportion of customers at different valuation levels remain the same (i.e., willingness-to-pay distribution does not change). Hence, if customers' heterogeneity in their valuations do not change as a result of new customers, then the optimal prices either remain the same or they are higher.

## 8 Optimal Prices and Waiting Room Capacity

In this section, we first consider the optimal price  $y_m^*$  as  $m$  goes to infinity. Then, we will investigate more closely how the optimal price changes as the capacity  $m$  increases or decreases.

## 8.1 Optimal Price as the Capacity Converges to Infinity

For finite capacity systems with general service and interarrival times, it is not possible to find expressions for optimal prices since we do not have expressions for  $B_N(\cdot)$  or  $\pi_0(\cdot)$  for such general systems. Therefore, one would be interested in approximations to the optimal price. In this section, we show that as the capacity converges to infinity, the optimal price for the finite capacity system converges to the optimal price for the infinite capacity system (when payments are made at departures) under certain conditions. This means that for sufficiently large capacity, the optimal price can be approximated with  $y_d^*$ , which has the simple expression given in Proposition 5.1.

In the proof of this limit result, a key assumption is the following:

$$\lim_{m \rightarrow \infty} \pi_0(y, m) = 0 \text{ for } \rho(y) \geq 1.$$

We believe that this assumption holds under quite general conditions if not always. One might be more skeptical whether it holds for  $\rho(y) = 1$ . However, using the two equivalent expressions of  $R(y, m)$ , one can show that for  $\rho(y) = 1$ ,  $\lim_{m \rightarrow \infty} \pi_0(y, m) = \lim_{m \rightarrow \infty} B_N(\lambda(y), m)$ . Then, if one limit is strictly positive so is the other and this is a good reason to believe that the assumption holds under quite general conditions. Note that it can be easily shown that it holds if interarrival and service times are exponentially distributed.

We now give two lemmas. These lemmas are used in the proof of Proposition 8.1, but they are given here since they have some independent interest.

**Lemma 8.1** *For  $m \geq 1$ ,  $B_N(\lambda(y), m + 1) \leq B_N(\lambda(y), m)$  and  $\pi_0(y, m + 1) \leq \pi_0(y, m)$ . Therefore,  $R(y, m + 1) \geq R(y, m)$  for any  $y$ .*

**Lemma 8.2** *Suppose that  $\lim_{m \rightarrow \infty} \pi_0(y, m) = 0$  for  $\rho(y) \geq 1$ . Then,  $R(y, m)$  converges to  $R_d(y, \infty)$  as  $m$  converges to infinity.*

Lemma 8.1 states the intuitive result that for any price  $y$ , it is always better to have higher capacity, and Lemma 8.2 states that the objective function for the finite capacity system converges to the objective function for the infinite capacity system under the assumption that customers pay at departure times. Using these lemmas, we prove the following proposition.

**Proposition 8.1** *Let  $\{y_m^*\}$  for  $m = 1, 2, \dots$  be a sequence of optimal prices for  $R(y, m)$ , and suppose that  $\lim_{m \rightarrow \infty} \pi_0(y, m) = 0$  for  $\rho(y) \geq 1$ . Then, under IPE,  $y_m^*$  converges to  $y_d^*$  as  $m$  converges to  $\infty$ .*

If  $\rho(y_a^*) \leq 1$ , then  $y_m^*$  converges to  $y_a^*$  by Propositions 5.1 and 8.1. On the other hand if  $\rho(y_a^*) \geq 1$ , then  $y_m^*$  converges to  $y^o$ . In other words, as the capacity converges to infinity, the optimal price always converges to a price under which the arrival rate of the customers is at most equal to the service rate. Note that Courcoubetis and Reiman [4] prove a similar result under a different formulation. They conclude that as the capacity converges to infinity, under the asymptotically optimal prices, the system is either unloaded (i.e., arrival rate is less than service rate) or critically loaded (i.e., arrival rate is equal to service rate).

## 8.2 Monotonicity of the Optimal Price in the Waiting Room Capacity

In this section, we look more closely at how the optimal price for the finite capacity system  $y_m^*$  changes with the capacity  $m$ . An initial guess would be that the optimal price is monotone increasing or monotone decreasing in the capacity. However, it turns out that in general optimal price is not monotone as the following example shows.

**Example:** Suppose that arrivals are Poisson and service times are deterministic. Let the willingness-to-pay distribution  $F(\cdot)$  be Uniform(0,10), maximum arrival rate  $\Lambda = 2.9$  and service rate  $\mu = 1$ . Then, it can be shown that  $y_1^* \simeq 6.638$ ,  $y_2^* \simeq 6.522$  and  $y_3^* \simeq 6.546$ . Hence, the optimal price decreases as the capacity is increased from 1 to 2, but increases as the capacity is increased from 2 to 3.

However, if we assume that interarrival and service times are exponentially distributed, then we can show that the optimal price is monotone in the capacity. Interestingly, whether it is monotone non-decreasing or monotone non-increasing depends on the value of  $\Lambda/\mu$ . This monotonicity result is given in Proposition 8.2 along with other ordering relations.

**Proposition 8.2** *Let  $\rho^c = 1/\bar{F}(y^c)$  where*

$$y^c = \begin{cases} \inf\{y : e(y) \geq 2\} & \text{if there exists } y < \infty \text{ s.t. } e(y) \geq 2 \\ \infty & \text{otherwise,} \end{cases}$$

*and suppose that interarrival and service times are exponentially distributed. Then, under IPE:*

- (i) *If  $\Lambda/\mu = \rho^c$  then  $y_a^* \leq y_m^* = y^c = y^o = y_d^*$  for  $m < \infty$ .*
- (ii) *If  $\Lambda/\mu > \rho^c$  then  $y_a^* \leq y^c \leq y_1^* \leq y_2^* \leq \dots \leq y_m^* \leq y_{m+1}^* \leq \dots \leq y_d^* = y^o$ .*
- (iii) *If  $\Lambda/\mu < \rho^c$  then  $y_a^* \leq y_d^* \leq \dots \leq y_{m+1}^* \leq y_m^* \leq \dots \leq y_2^* \leq y_1^* \leq y^c$ . Furthermore, if  $1 \leq \Lambda/\mu < \rho^c$ , then  $y^o \leq y_m^*$  for  $m < \infty$ .*

As it can be seen from Proposition 8.2, the ordering relations depend on whether the value of  $\Lambda/\mu$  is equal to, greater than or smaller than  $\rho^c$ . Because of this property we call  $\rho^c$  the “critical traffic intensity” and corresponding price  $y^c$  the “critical price”. The expression  $\Lambda/\mu$  gives the maximum traffic intensity the service facility can observe. Hence, it is a measure of the level of “relative demand” for the service (“relative” in the sense that demand is normalized by the service capacity of the system) and we call it the “offered load”. If the offered load is at some critical level  $\rho^c$ , the optimal price is the same whatever the capacity is. If the load is high, the optimal price increases with the waiting room capacity; on the other hand, if the load is low, the optimal price decreases with the waiting room capacity. Note that since  $\rho^c$  and  $y^c$  only depend on the willingness-to-pay distribution function, whether the offered load is low or high is also determined by the customers’ willingness-to-pay distribution.

The expression  $\Lambda/\mu$  is the offered load for the system; however, it is not the load observed by the server. There are two mechanisms that control the arrival rate seen by the server. One is the waiting room capacity and the other is the price. For fixed price, systems with

smaller capacities see lower arrival rates since they lose more customers. As we increase the waiting room capacity, the effect of the waiting room capacity on the arrival rate decreases, and pricing becomes a more effective decision. If the offered load is high, the service provider can set relatively high prices while keeping the server utilization high. In this case, as the waiting room capacity increases, the service provider can increase the price to take advantage of the increased “effective” demand caused by the additional waiting space. However, if the offered load is low, the service provider is not as powerful. Charging high prices would result in having low utilization levels. In this case, the service provider does not view an increase in the waiting room capacity as an opportunity to increase the prices, but rather as an opportunity to induce more demand by lowering prices. When the capacity is small, since many customers are lost even if they are willing to pay the price, the service provider finds it more profitable to charge a higher price.

A more technical explanation for this monotonicity is as follows: First, note that if  $\Lambda/\mu > \rho^c$ , we have  $\rho(y_m^*) \geq 1$  for all  $m < \infty$ . This means that even though the optimal price increases as  $m$  increases (hence  $\rho(y_m^*)$  decreases), still for any  $m$ , the customer arrival rate for price  $y_m^*$  is more than the service rate. Therefore, as the capacity increases, the service provider sets the price so that each served customer pays more, but not so much as to cause the server to become idle frequently due to the smaller arrival rate induced by the increase in price. Note that, for small capacity, there is a higher chance for the server to be idle. Therefore, lower prices (hence higher arrival rates) are set in such cases. On the other hand, if  $\Lambda/\mu < \rho^c$ , we have  $\rho(y_m^*) \leq 1$  for all  $m < \infty$ . In other words, the server will be idle quite often whatever the capacity is. Arguing in the same way as in the previous case, we might be tempted to believe that for small capacity, the price should be decreased (hence arrival rate is increased) so as to decrease the idleness. However, in this case that is a quite costly alternative, since that would cause each customer to pay extremely low prices due to the low demand. In general, what we have here is a trade-off between the idleness and the revenue obtained from each served customer, and the relationship between the two depends on the offered demand for the service ( $\Lambda/\mu$ ).

## 9 Conclusions

The majority of the work on the optimal pricing for queueing systems considers models with infinite waiting room capacity, which are in general easier to analyze compared to systems with finite waiting room capacity. The difficulty in the analysis of the finite capacity systems arises due to the presence of the blocking term (in our model, the term  $B_N(\lambda(y), m)$ ) in the objective functions. In fact, in many cases, it is impossible to find expressions for the optimal price since we do not have an expression for  $B_N(\lambda(y), m)$  under general arrival and service processes. However, optimal prices for infinite capacity systems can be used to make more informed decisions for finite capacity systems as well. Under the assumption that customers pay at acceptance times, the optimal price for the infinite capacity model is a lower bound on any optimal price for any finite capacity system. Furthermore, for large capacity, the optimal price is close to the optimal price for the infinite capacity model under the assumption that customers pay as they leave.

As stated in Proposition 8.2, the relationship between the price and the waiting room

capacity depends on the offered load. The service provider has two conflicting objectives: billing many customers, which is equivalent to keeping the server busy, and billing a large amount for each customer. As the waiting room size decreases, the service provider might also decrease the price so that the server will have less idle time. This is exactly what happens when the offered load is high. On the other hand, when the offered load is low, the service provider is not as powerful, and the relationship between the optimal price and the capacity is reversed. As the capacity gets smaller, charging lower prices to raise the arrival rate is a costly strategy. Instead, the service provider chooses to get paid more for each accepted customer. The trade-off between the price and arrival rate works in two different ways depending on the offered load being low or high.

Changes in customers' willingness-to-pay might have some counter-intuitive effects on the optimal prices. There are some cases where increases in each customer's willingness-to-pay decrease the optimal price. In other words, if there is a shift in the customers' willingness-to-pay distribution function and this shift results in having a larger distribution in the usual stochastic sense, the new optimal price is not necessarily higher. However, if the shift causes the new distribution to be larger in the hazard rate order, then the new optimal price is higher.

## A Appendix

In this section, we provide the proofs of the results given in the previous sections and also give the lemmas which are needed to prove some of our results.

The proof of the following lemma is omitted since it is straightforward but requires tedious algebra.

**Lemma A.1** *Let  $m \geq 1$  be a finite integer. Then:*

- (i)  $\frac{1+mz^{m+1}-(m+1)z^m}{(1-z^m)(1-z^{m+1})}$  is a decreasing function of  $z$  for  $z > 0$ .
- (ii)  $\frac{1+(m+1)z^{m+2}-(m+2)z^{m+1}}{(1-z^{m+1})(1-z^{m+2})} - \frac{1+mz^{m+1}-(m+1)z^m}{(1-z^m)(1-z^{m+1})} < 0$  for  $z > 1$ .
- (iii)  $\frac{1+(m+1)z^{m+2}-(m+2)z^{m+1}}{(1-z^{m+1})(1-z^{m+2})} - \frac{1+mz^{m+1}-(m+1)z^m}{(1-z^m)(1-z^{m+1})} > 0$  for  $0 < z < 1$ .

The following corollary follows from Lemma A.1. Part (i) follows from Lemma A.1(i) and part (ii) follows from Lemma A.1(ii) and from Lemma A.1(iii).

**Corollary A.1** *Let  $\gamma_m(y)$  be defined as in Proposition 5.1 (iii). Then, we have*

- (i)  $\gamma_m(y)$  is strictly decreasing in  $\rho(y)$  and non-decreasing in  $y$ ,
- (ii) If  $\rho(y) > (<)(=)1$ , then  $\gamma_{m+1}(y) < (>)(=)\gamma_m(y)$ .

### Proof of Proposition 4.1:

We first prove that  $Y_a^* \neq \emptyset$ . Since  $F(\cdot)$  is assumed to be absolutely continuous and  $\Lambda > 0$ ,  $y\lambda(y)$  is a continuous non-negative function of  $y$ . Since  $\alpha < \beta$ , there exists  $\alpha < y < \beta$  such

that  $y\lambda(y) > 0$ . If  $\beta < \infty$ , we have  $\beta\lambda(\beta) = 0$ . If  $\beta = \infty$ , it follows from the finite mean assumption on  $F(\cdot)$  that  $\lim_{y \rightarrow \infty} y\lambda(y) = 0$  (see Chung [3]). Since the function  $y\lambda(y)$  is continuous, we conclude that there exists a finite value of  $y$  for which the function  $y\lambda(y)$  is maximized. Hence,  $Y_a^* \neq \emptyset$ .

If  $\Lambda/\mu < 1$ , then it follows immediately from (5) that  $Y_d^* \neq \emptyset$ . If  $\Lambda/\mu \geq 1$ , then we have the same result from (5) and the fact that  $y^o < \infty$ .

Suppose that  $Y_m^* \neq \emptyset$ . Let  $y_a^*$  be the unique element of  $Y_a^*$  and let  $y_m^* \in Y_m^*$ . It is sufficient to show that if  $y_a^* \neq y_m^*$ , then  $y_a^* < y_m^*$ . Since  $y_m^*$  is optimal for the m-capacity system,

$$R(y_m^*, m) \geq R(y_a^*, m).$$

Using (6), we get

$$\frac{y_m^* \lambda(y_m^*)}{y_a^* \lambda(y_a^*)} \geq \frac{1 - B_N(\lambda(y_a^*), m)}{1 - B_N(\lambda(y_m^*), m)}.$$

Note that the numerator and denominator on the left hand side are the expected long run average revenue expressions for the infinite capacity system (under ‘‘payments at acceptance points’’) with prices  $y_m^*$  and  $y_a^*$ , respectively. Then, since  $y_a^*$  is the unique optimal price for the infinite capacity system,

$$\frac{y_m^* \lambda(y_m^*)}{y_a^* \lambda(y_a^*)} < 1.$$

This implies

$$\frac{1 - B_N(\lambda(y_a^*), m)}{1 - B_N(\lambda(y_m^*), m)} < 1.$$

Therefore,

$$B_N(\lambda(y_a^*), m) > B_N(\lambda(y_m^*), m).$$

Finally, the result follows from the main result of Ziya, Ayhan and Foley [32].  $\square$

### Proof of Proposition 5.1:

(i) Note that if  $R_a(y, \infty)$  is differentiable at  $y \in [\alpha, \beta)$ , then  $\frac{dR_a(y, \infty)}{dy} > (<)0$  if and only if  $e(y) = yr(y) < (>)1$ . We know from the proof of Proposition 4.1 that there exists  $\alpha < y < \beta$  for which  $R_a(y, \infty)$  is strictly positive. We also know that if  $\beta < \infty$  then  $R_a(\beta, \infty) = 0$ ; if  $\beta = \infty$ , then  $\lim_{y \rightarrow \beta} R_a(y, \infty) = 0$ . It follows that, since  $R_a(y, \infty)$  is continuous, there exists  $y \in (\alpha, \beta)$  such that  $e(y) \geq 1$ . Then, since  $e(y)$  is increasing for  $y > \inf[y : e(y) \geq 1]$ ,  $R_a(y, \infty)$  is decreasing for any  $y > \inf[y : e(y) \geq 1]$  for which  $R_a(y, \infty)$  is differentiable. Similarly,  $R_a(y, \infty)$  is increasing for  $y < \inf[y : e(y) \geq 1]$ . Finally, since  $F(\cdot)$  is absolutely continuous,  $R_a(y, \infty)$  is differentiable a.e., continuous, and we conclude that  $y_a^* = \inf[y : e(y) \geq 1]$ .

If  $\Lambda/\mu < 1$ , then  $\rho(y) < 1$  for all  $y$  and by (5)  $R_d(y, \infty) = R_a(y, \infty)$ . Thus,  $y_d^* = y_a^*$ .

Suppose that  $\Lambda/\mu \geq 1$ . Then, there exists  $y$  such that  $\rho(y) = 1$  and we can rewrite (5) as

$$R_d(y, \infty) = \begin{cases} y\mu & \text{if } y \leq y^o \\ y\lambda(y) & \text{if } y > y^o \end{cases} \quad (9)$$

From the definition of  $y^o$  and continuity of  $F(y)$ , we have  $\lambda(y^o) = \mu$ , and thus  $R_d(y, \infty)$  is continuous. Note that it follows immediately from (9) that the best price in the set  $\{y : y \leq$

$y^o\}$  is  $y^o$ . If there exists  $\varepsilon > 0$  such that  $e(y^o + \varepsilon) < 1$ , then using the arguments of part (i), we conclude that the best price in the set  $\{y : y > y^o\}$  is  $y_a^*$  and  $y_a^* \lambda(y_a^*) \geq y^o \lambda(y^o) = y^o \mu$ . Thus,  $y_a^* = y_a^* = \max\{y_a^*, y^o\}$ . If there exists no  $\varepsilon > 0$  such that  $e(y^o + \varepsilon) < 1$ , then once again using the arguments of part (i), we conclude that the best price in the set  $\{y : y > y^o\}$  is no better than  $y^o$  and  $y_a^* \leq y^o$ . Hence,  $y_a^* = y^o = \max\{y_a^*, y^o\}$ .

Using similar arguments, we can also show that if  $y^o r(y^o) \geq (<)1$ , then  $y_a^* \leq (\geq)y^o$ . Similarly, if  $\rho(y_a^*) \geq (<)1$ , then  $y_a^* \leq (>)y^o$  where we also make use of the fact that  $\rho(y^o) = 1$ .

(ii) Using the expression of  $R(y, m)$  given in (8) and taking its derivative, it can be shown that  $\frac{dR(y, m)}{dy} > (<)0$  if

$$y \frac{\pi'_0(y, m)}{1 - \pi_0(y, m)} < (>)1. \quad (10)$$

Since  $R(y, m)$  is assumed to have a unique local maximum, it follows that  $y_m^* = \inf[y : \frac{y \pi'_0(y, m)}{1 - \pi_0(y, m)} \geq 1]$ .

(iii) Since the system is an M/M/1/m queueing system, we have an explicit expression for  $\pi_0(y, m)$ ,

$$\pi_0(y, m) = \begin{cases} \frac{1 - \rho(y)}{1 - \rho(y)^{m+1}} & \text{if } \rho(y) \neq 1 \\ \frac{1}{m+1} & \text{if } \rho(y) = 1 \end{cases}.$$

Suppose that  $R(y, m)$  is differentiable at  $y$ . Then, we know from the proof of part (ii) that  $\frac{dR(y, m)}{dy} > (<)0$  if  $y \frac{\pi'_0(y, m)}{1 - \pi_0(y, m)} < (>)1$ . It can be shown that  $y \frac{\pi'_0(y, m)}{1 - \pi_0(y, m)} = yr(y)\gamma_m(y) = e(y)\gamma_m(y)$ . It can also be shown that  $\pi_0(y, m)$  is differentiable a.e. and continuous. Thus,  $R(y, m)$  is also differentiable a.e. and continuous. We know from Proposition 4.1 that  $y_m^* \in [\inf\{y : e(y) \geq 1\}, \beta)$ , and since we are assuming IPE, we know that  $e(y)$  is strictly increasing in the same interval. Finally, from Corollary A.1 (i), we conclude that  $e(y)\gamma_m(y)$  is also strictly increasing over  $[\inf\{y : e(y) \geq 1\}, \beta)$ . Hence,  $R(y, m)$  has a unique local maximum (and therefore a unique global maximum) and we conclude that

$$y_m^* = \inf[y : e(y)\gamma_m(y) \geq 1].$$

□

### Proof of Proposition 6.1:

Note that in this proof, we add A and B as subscripts to our original notation to indicate systems A and B.

Now, there are three different settings.

(i) Suppose that the capacity is infinite and customers pay at arrivals. Then,  $y_{aB}^* = \inf\{y : e_B(y) \geq 1\}$  and therefore, from assumption IPE,  $e_B(y_{aB}^* + \varepsilon) \geq 1$  for any  $\varepsilon > 0$  such that  $(y_{aB}^* + \varepsilon) < \beta$ . Since  $F_B \geq_{hr} F_A$ ,  $r_B(y) \leq r_A(y)$ . This implies that  $e_A(y_{aB}^* + \varepsilon) \geq 1$ . Finally, since  $y_{aA}^* = \inf\{y : e_A(y) \geq 1\}$  we conclude that  $y_{aA}^* \leq y_{aB}^*$ .

(ii) Suppose that the capacity is infinite and customers pay at departures. If  $\Lambda/\mu < 1$ , then the result immediately follows from part (i) and Proposition 5.1 (i). If  $\Lambda/\mu \geq 1$ , using Proposition 5.1 (i), it is sufficient to show that  $y_B^o \geq y_A^o$ . However, this immediately follows from the definitions of  $y_B^o$  and  $y_A^o$ , and using the fact that  $F_B \geq_{hr} F_A$  implies that  $F_B \geq_{st} F_A$ .

(iii) Suppose that interarrival and service times are exponentially distributed and there is a finite capacity,  $m$ . Then,  $y_{mB}^* = \inf[y : e_B(y)\gamma_{mB}(y) \geq 1]$ . Let  $\varepsilon > 0$  be such that

$y_{mB}^* + \varepsilon < \beta$ . Then, from assumption IPE and Corollary A.1 (i) (note that interval  $[\inf\{y : e_A(y) \geq 1\}, \beta)$  contains  $y_{mA}^*$  by Proposition 4.1),  $e_B(y_{mB}^* + \varepsilon)\gamma_{mB}(y_{mB}^* + \varepsilon) \geq 1$ . Since  $F_B \geq_{hr} F_A$ ,  $r_B(y) \leq r_A(y)$ . Moreover,  $\rho_B(y) \geq \rho_A(y)$  since  $F_B \geq_{hr} F_A$  implies  $F_B \geq_{st} F_A$ . Then, from Corollary A.1 (i), it follows that  $e_A(y_{mB}^* + \varepsilon)\gamma_{mA}(y_{mB}^* + \varepsilon) \geq 1$ . Finally, since  $y_{mA}^* = \inf[y : e_A(y)\gamma_{mA}(y) \geq 1]$ , we conclude that  $y_{mA}^* \leq y_{mB}^*$ .  $\square$

**Proof of Proposition 6.2:**

The proof is similar to the proof of Proposition 6.1. We add A and B as subscripts to our original notation to indicate systems A and B. It can easily be shown that  $yr_B(y) = yr_A(y-s)$  for  $\alpha_B \leq y \leq \beta_B$ . Suppose that the capacity is infinite. Then, the result immediately follows if the payments are made at acceptance times. If the payments are made at departure times, then the result follows from  $yr_B(y) = yr_A(y-s)$  in addition to the fact that  $y_B^o = y_A^o + s$ . If the capacity is finite, the result follows from  $yr_B(y) = yr_A(y-s)$  and  $\rho_B(y) = \rho_A(y-s)$  for  $\alpha_A \leq y \leq \beta_A$  in addition to Corollary A.1 (i).  $\square$

**Proof of Proposition 7.1:**

Note that in this proof, we add A and B as subscripts to our original notation to indicate systems A and B.

Suppose that there is infinite capacity and customers pay at departures. Since  $\Lambda_B/\mu_B > \Lambda_A/\mu_A$ , we have  $y_B^o > y_A^o$ . We know from Proposition 5.1 (i) that  $y_{aB}^* = y_{aA}^*$ . Then, it again follows from Proposition 5.1 (i) that  $y_{dA}^* \leq y_{dB}^*$ .

Suppose that interarrival and service times are exponentially distributed and there is a finite capacity,  $m$ . Then,  $y_{mB}^* = \inf[y : e(y)\gamma_{mB}(y) \geq 1]$ . Let  $\varepsilon > 0$  be such that  $y_{mB}^* + \varepsilon < \beta$ . Then, from IPE and Corollary A.1 (i) (note that interval  $[\inf\{y : e(y) \geq 1\}, \beta)$  contains  $y_{mA}^*$  by Proposition 4.1),  $e(y_{mB}^* + \varepsilon)\gamma_{mB}(y_{mB}^* + \varepsilon) \geq 1$ . Since  $\rho_B(y) > \rho_A(y)$ , from Corollary A.1 (i), it follows that  $e(y_{mB}^* + \varepsilon)\gamma_{mA}(y_{mB}^* + \varepsilon) \geq 1$ . Since  $y_{mA}^* = \inf[y : e(y)\gamma_{mA}(y) \geq 1]$ , we conclude that  $y_{mA}^* \leq y_{mB}^*$ .  $\square$

**Proof of Lemma 8.1:**

Proof is based on a coupling argument. Consider two systems, System 1 and System 2. Suppose that System 1 has a capacity of  $m$  and System 2 has a capacity of  $m + 1$ . Let  $T^n$  denote the arrival time of the  $n$ th customer who is willing to pay  $y$  but who may not be able to enter the system (both System 1 and System 2) due to waiting room capacity and  $S^n$  denote the service time for the  $n$ th customer serviced (in both System 1 and System 2). We further define the following:

$A_i(t)$ : Number of customers accepted to System  $i$  until (and including) time  $t$ .

$D_i(t)$ : Number of customers departed from System  $i$  until (and including) time  $t$ .

$I_i(t)$ : Idle time for server (System)  $i$  until time  $t$ .

We will show that for  $t \geq 0$ ,

$$\begin{aligned} A_1(t) &\leq A_2(t), \\ D_1(t) &\leq D_2(t), \\ I_1(t) &\geq I_2(t). \end{aligned} \tag{11}$$

Let  $\tau$  be defined as

$$\tau = \inf\{t : A_1(t) > A_2(t)\}$$

and suppose for contradiction that  $\tau < \infty$ . Then, we know that

$$A_1(t) \leq A_2(t) \text{ for } t \in [0, \tau). \quad (12)$$

Since service time for some  $n$ th customer is the same for both systems, this also implies that

$$D_1(t) \leq D_2(t) \text{ for } t \in [0, \tau). \quad (13)$$

Let  $A_2(\tau) = k - 1$ . Then, from the definition of  $\tau$ , we conclude that after all the scheduled events occur at time  $\tau$ , System 2 is full and customer at the end of the queue is customer  $k - 1$ . Customer  $k$  joins System 1 but not System 2. Then,  $D_2(\tau) = k - 2 - m$ . On the other hand, since customer  $k$  joins System 1 at time  $\tau$ , we have  $k - m \leq D_1(\tau)$ .

Then, we conclude that  $D_1(\tau) - D_2(\tau) \geq 2$ . This together with (13) imply that at time  $\tau$ , there has been at least one departure from System 1 such that service time for that customer started earlier than the corresponding customer of System 1. Since service time for the  $n$ th customer is the same in System 1 and System 2 for all  $n$ , this is a contradiction to (12) and (13).

Thus, we conclude that (11) holds for all  $t \geq 0$ . The fact that  $A_1(t) \leq A_2(t)$  implies that  $B_N(\lambda(y), m + 1) \leq B_N(\lambda(y), m)$  and the fact that  $I_1(t) \geq I_2(t)$  implies that  $\pi_0(y, m + 1) \leq \pi_0(y, m)$  for  $m \geq 1$ . Finally,  $R(y, m + 1) \geq R(y, m)$  follows from (6) and (8).  $\square$

**Proof of Lemma 8.2:**

Let  $\bar{y}$  be such that  $\rho(\bar{y}) = 1$ . Using (6) and (8), it can easily be shown that

$$B_N(\lambda(\bar{y}), m) = \pi_0(\bar{y}, m).$$

From the limit assumption, we know that  $\lim_{m \rightarrow \infty} \pi_0(\bar{y}, m) = 0$ . Then, we also have

$$\lim_{m \rightarrow \infty} B_N(\lambda(\bar{y}), m) = 0.$$

From the main result of Ziya, Ayhan and Foley [32], we know that,  $B_N(\lambda(y), m) \leq B_N(\lambda(\bar{y}), m)$  for  $\rho(y) < 1$ . This implies that  $\lim_{m \rightarrow \infty} B_N(\lambda(y), m) = 0$  for  $\rho(y) \leq 1$ . Then, using (6), we have  $\lim_{m \rightarrow \infty} R(y, m) = y\lambda(y)$  for  $\rho(y) \leq 1$ . Also, using (8) and the limit assumption again, we conclude that  $\lim_{m \rightarrow \infty} R(y, m) = y\mu$  for  $\rho(y) \geq 1$ . Hence,  $\lim_{m \rightarrow \infty} R(y, m) = R_d(y, \infty)$ .

$\square$

**Proof of Proposition 8.1:**

It can be shown that  $R(y, m)$  is bounded using the fact that  $R(y, m) \leq R_a(y, \infty)$  and the finite mean assumption for  $F(\cdot)$  implying that

$$\limsup_{y \rightarrow \infty} R_a(y, \infty) = 0.$$

From Lemma 8.1, we also know that  $R(y, m)$  is non-decreasing in  $m$ . Then, we have (see Bartle and Sherbert [1] and Fischer [7]):

$$\lim_{m \rightarrow \infty} (\sup_y R(y, m)) = \sup_y (\lim_{m \rightarrow \infty} R(y, m)).$$

We know from Lemma 8.2 that

$$\lim_{m \rightarrow \infty} R(y, m) = R_d(y, \infty).$$

Then, we have

$$\lim_{m \rightarrow \infty} (\sup_y R(y, m)) = \sup_y (R_d(y, \infty)),$$

which can also be written as (since we assume the existence of optimal prices)

$$\lim_{m \rightarrow \infty} R(y_m^*, m) = R_d(y_d^*, \infty). \quad (14)$$

Now, suppose for contradiction that  $y_m^*$  does not converge to  $y_d^*$ . Then there exists a neighborhood of  $y_d^*$ ,  $V$ , such that if  $n$  is any natural number, then there is a natural number  $k = k(n) \geq n$  such that  $y_k^* \notin V$  (see Bartle [2]). Let  $V = \{y : y_d^* - \delta_1 < y < y_d^* + \delta_2\}$  where  $\delta_1 > 0$  and  $\delta_2 > 0$ . Then,

$$R(y_k^*, k) = \sup_{\{y: y \notin V\}} R(y, k) \leq \sup_{\{y: y \notin V\}} R_d(y, \infty) = \max(R_d(y_d^* - \delta_1), R_d(y_d^* + \delta_2))$$

where the last equality follows from the fact that under IPE,  $R_d(y, \infty)$  has a unique local maximum (it is either unimodal or decreasing) and  $y_d^*$  is the unique optimal solution.

Let  $\psi > 0$  be defined such that  $R_d(y_d^*, \infty) - \max(R_d(y_d^* - \delta_1), R_d(y_d^* + \delta_2)) = \psi$ . Then,

$$R_d(y_d^*, \infty) - R(y_k^*, k) \geq R_d(y_d^*, \infty) - \max(R_d(y_d^* - \delta_1), R_d(y_d^* + \delta_2)) = \psi > 0.$$

Let  $0 < \varepsilon < \psi$ , then we know from (14) that there exists a natural number  $N(\varepsilon)$  such that for  $m \geq N(\varepsilon)$  we have

$$0 \leq R_d(y_d^*, \infty) - R(y_m^*, m) < \varepsilon.$$

However, we also know that there exists  $k \geq N(\varepsilon)$  such that

$$R_d(y_d^*, \infty) - R(y_k^*, k) \geq \psi$$

which is a contradiction. Hence,  $y_m^*$  converges to  $y_d^*$ .  $\square$

### Proof of Proposition 8.2:

The proof follows immediately from Lemmas A.2, A.3 and A.4. In part (i),  $y^c = y^o$  follows from the fact that IPE implies that  $F(y)$  is strictly increasing for  $y \in [y_a^*, \beta)$  which in turn implies that  $y^o$  is the only price for which  $\rho(y) = 1$ . The fact that  $y_d^*$  is a lower or an upper bound follows from Proposition 8.1 and the fact that  $y_a^*$  is always a lower bound follows from Proposition 4.1.  $\square$

**Lemma A.2** *Let*

$$y^c = \begin{cases} \inf\{y : e(y) \geq 2\} & \text{if there exists } y < \infty \text{ s.t. } e(y) \geq 2 \\ \infty & \text{otherwise,} \end{cases}$$

$\rho^c = 1/\bar{F}(y^c)$ , and  $m < \infty$ . Suppose that interarrival and service times are exponentially distributed. Then, under IPE, we have:

- (i) If  $\Lambda/\mu > \rho^c$ , then  $y^c \leq y_m^*$ .
- (ii) If  $\Lambda/\mu < \rho^c$ , then  $y_m^* \leq y^c$ .
- (iii) If  $\Lambda/\mu = \rho^c$ , then  $y_m^* = y^c$ .

*Proof* First, suppose that there exists no  $y$  such that  $e(y) \geq 2$ . In such a case,  $y^c = \infty$  and  $\rho^c = \infty$ . Then, since  $\Lambda$  is finite and  $\mu$  is strictly positive  $\Lambda/\mu < \rho^c$ . Hence, we conclude that if there is no  $y$  such that  $e(y) \geq 2$ , then  $\Lambda/\mu < \rho^c$ ; or, if  $\Lambda/\mu \geq \rho^c$ , then there exists  $y$  such that  $e(y) \geq 2$ . This means that in parts (i) and (iii) below, there exists  $y$  such that  $e(y) \geq 2$ .

(i) If  $\Lambda/\mu \geq \rho^c$ , we have  $\rho(y^c) \geq 1$ . Then, from Corollary A.1 (i),  $\gamma_m(y^c) \leq \frac{1}{2}$ . We can also write  $\gamma_m(y^c - \varepsilon) \leq \frac{1}{2}$  for any  $\varepsilon > 0$  such that  $y^c - \varepsilon \geq \alpha$ . From the definition of  $y^c$ , we have  $e(y^c - \varepsilon) < 2$ . Thus,  $e(y^c - \varepsilon)\gamma_m(y^c - \varepsilon) < 1$ . Hence, it follows from Proposition 5.1 (iii) that  $y^c \leq y_m^*$ .

(ii) Suppose that  $\Lambda/\mu \leq \rho^c$ . If there exists no  $y$  such that  $e(y) \geq 2$ , then  $y^c = \infty$  and the result immediately follows. Now, suppose that there exists such  $y$  and  $y^c < \infty$ . Since  $\Lambda/\mu \leq \rho^c$ , we have  $\rho(y^c) \leq 1$ . Then, from Corollary A.1 (i),  $\gamma_m(y^c) \geq \frac{1}{2}$ . We can also write  $\gamma_m(y^c + \varepsilon) \geq \frac{1}{2}$  for any  $\varepsilon > 0$  such that  $y^c + \varepsilon < \beta$ . Since there exists  $y$  such that  $e(y) \geq 2$ , from the definition of  $y^c$  and the assumption IPE, we have  $e(y^c + \varepsilon) \geq 2$ . Then,  $e(y^c + \varepsilon)\gamma_m(y^c + \varepsilon) \geq 1$ , and finally Proposition 5.1 (iii) implies that  $y_m^* \leq y^c$ .

(iii) It follows from parts (i) and (ii) above.  $\square$

**Lemma A.3** *Let  $m, n < \infty$  and  $n > m$ . Suppose that interarrival and service times are exponentially distributed and IPE holds. Then:*

- (i) If  $\rho(y_m^*) \geq 1$ , then  $y_n^* \geq y_m^*$ .
- (ii) If  $\rho(y_m^*) \leq 1$ , then  $y_n^* \leq y_m^*$ .

*Proof* (i) Suppose for contradiction that  $y_n^* < y_m^*$ . Then there exists  $\delta > 0$  such that  $y_n^* + \delta < y_m^*$ . By Proposition 5.1 (iii), we can write  $e(y_n^* + \delta)\gamma_m(y_n^* + \delta) < 1$ . Since  $\rho(y_m^*) \geq 1$ , it is also true that  $\rho(y_n^* + \delta) \geq 1$  and from Corollary A.1 (ii), we have

$$e(y_n^* + \delta)\gamma_n(y_n^* + \delta) \leq e(y_n^* + \delta)\gamma_m(y_n^* + \delta) < 1.$$

This is a contradiction to the optimality of  $y_n^*$ , since  $y_n^*$  satisfies

$$y_n^* = \inf\{y : e(y)\gamma_n(y) \geq 1\}.$$

Hence,  $y_n^* \geq y_m^*$ .

(ii) Let  $\varepsilon > 0$  be such that  $y_m^* + \varepsilon < \beta$ . Since  $\rho(y_m^*) \leq 1$ , we also have  $\rho(y_m^* + \varepsilon) \leq 1$ . Then from Corollary A.1 (ii),  $\gamma_n(y_m^* + \varepsilon) \geq \gamma_m(y_m^* + \varepsilon)$ . Hence, we have

$$e(y_m^* + \varepsilon)\gamma_n(y_m^* + \varepsilon) \geq e(y_m^* + \varepsilon)\gamma_m(y_m^* + \varepsilon).$$

From Proposition 5.1 (iii) and Assumption IPE, we have  $e(y_m^* + \varepsilon)\gamma_m(y_m^* + \varepsilon) \geq 1$ . This implies that  $e(y_m^* + \varepsilon)\gamma_n(y_m^* + \varepsilon) \geq 1$  and we conclude that  $y_n^* \leq y_m^* + \varepsilon$ . Taking the limit as  $\varepsilon$  approaches zero, we find  $y_n^* \leq y_m^*$  for all  $m < \infty$ .  $\square$

**Lemma A.4** *Suppose that interarrival and service times are exponentially distributed and IPE holds. If  $e(y^o) \geq (<)2$ , then  $y_m^* \leq (\geq)y^o$ .*

*Proof* If  $e(y^o) \geq 2$ , then since  $\gamma_m(y^o) = \frac{1}{2}$  we have  $e(y^o)\gamma_m(y^o) \geq 1$ . Then, from Proposition 5.1 (iii),  $y_m^* \leq y^o$ . Similarly, if  $e(y^o) < 2$ , then  $e(y^o)\gamma_m(y^o) < 1$  and by Proposition 5.1 (iii), we conclude that  $y_m^* \geq y^o$ .  $\square$

## References

- [1] R. G. Bartle and D. R. Sherbert, Introduction to Real Analysis, John Wiley & Sons, 1992, pp. 44.
- [2] R. G. Bartle, The Elements of Real Analysis, John Wiley & Sons, 1976, pp. 99.
- [3] K. L. Chung, A Course in Probability Theory, Academic Press, 1974, pp. 49.
- [4] C. A. Courcoubetis and M. I. Reiman, “Pricing in a Large Single Link Loss System,” Teletraffic Engineering in a Competitive World, P. Key and D. Smith (Editors), Elsevier, 1999, pp. 737–746.
- [5] S. Dewan and H. Mendelson, User Delay Costs and Internal Pricing for a Service Facility, Management Science 36 (1990), 1502–1517.
- [6] N. M. Edelson and D. K. Hildebrand, Congestion Tolls for Poisson Queueing Processes, Econometrica 43 (1975), 81–92.
- [7] E. Fischer, Intermediate Real Analysis, Springer-Verlag, 1983, pp. 102.
- [8] P. Franken, D. Konig, U. Arndt, V. Schmidt, Queues and Point Processes, Akademie - Verlag, Berlin, 1981, pp. 112.
- [9] A. Y. Ha, Incentive-Compatible Pricing for a Service Facility with Joint Production and Congestion Externalities, Management Science 44 (1998), 1623–1636.
- [10] S. G. Johansen, Optimal Prices of an M/G/1 Jobshop, Operations Research 42 (1994), 765–774.
- [11] F. P. Kelly, “Charging and Accounting for Bursty Connections,” Internet Economics, Lee McKnight and Joseph Bailey (Editors), The MIT Press, 1997, pp. 253–278.
- [12] N. C. Knudsen, Individual and Social Optimization in a Multiserver Queue with a General Cost-Benefit Structure, Econometrica 40 (1972), 515–528.
- [13] M. A. Lariviere and E. L. Porteus, Selling to the Newsvendor: An Analysis of Price-Only Contracts, Manufacturing and Service Operations Management 3 (2001), 293–305.
- [14] S. Lanning, W. A. Massey, B. Rider, and Q. Wang, Optimal Pricing in Queueing Systems with Quality of Service Constraints, Proceedings of the 16th International Teletraffic Congress - ITC, 1999, pp. 747–756.

- [15] S. A. Lippman, Applying a New Device in the Optimization of Exponential Queueing Systems, *Operations Research* 23 (1975), 687–708.
- [16] S. A. Lippman and S. Stidham, Jr., Individual versus Social Optimization in Exponential Congestion Systems, *Operations Research* 25 (1977), 233–247.
- [17] D. W. Low, Optimal Dynamic Pricing Policies for an M/M/s Queue, *Operations Research* 22 (1974), 545–561.
- [18] D. W. Low, Optimal Pricing for an Unbounded Queue, *IBM Journal of Research and Development* 18 (1974), 290–302.
- [19] H. Mendelson, Pricing Computer Services: Queueing Effects, *Communications of the ACM* 28 (1985), 312–321.
- [20] H. Mendelson and S. Whang, Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue, *Operations Research* 38 (1990), 870–883.
- [21] B. L. Miller and A. G. Buckman, Cost Allocation and Opportunity Costs, *Management Science* 33 (1987), 626–639.
- [22] P. Naor, The Regulation of Queue Size by Levying Tolls, *Econometrica* 37 (1969), 15–24.
- [23] I. C. Paschalidis and J. N. Tsitsiklis, Congestion-Dependent Pricing of Network Services, *IEEE/ACM Transactions on Networking* 8 (2000), 171–184.
- [24] M. Shaked and J. G. Shanthikumar, *Stochastic Orders and Their Applications*, Academic Press, Inc., 1994.
- [25] K. Sigman, *Stationary Marked Point Processes - An Intuitive Approach*, Chapman & Hall, 1995, pp. 134.
- [26] D. Sonderman, Comparing Multi-Server Queues with Finite Waiting Rooms, I: Same Number of Servers, *Advances in Applied Probability* 11 (1979), 439–447.
- [27] S. Stidham, Jr., Pricing and Capacity Decisions for a Service Facility: Stability and Multiple Local Optima, *Management Science* 38 (1992), 1121–1139.
- [28] U. Sumita, Y. Masuda, and S. Yamakawa, Optimal Internal Pricing and Capacity Planning for Service Facility with Finite Buffer, *European Journal of Operational Research* 128 (2001), 192–205.
- [29] Q. Wang, J. M. Peha, and M. A. Sirbu, “Optimal Pricing for Integrated Services Networks,” *Internet Economics*, Lee McKnight and Joseph Bailey (Editors), The MIT Press, 1997, pp. 353–376.
- [30] U. Yechiali, On Optimal Balking Rules and Toll Charges in the G/M/1 Queueing Process, *Operations Research* 19 (1971), 349–370.

- [31] S. Ziya, H. Ayhan and R. D. Foley, Relationships Among Three Assumptions in Revenue Management, to appear in Operations Research, available at [www.unc.edu/~ziya/rev-mgmt-assumptions.pdf](http://www.unc.edu/~ziya/rev-mgmt-assumptions.pdf).
- [32] S. Ziya, H. Ayhan and R. D. Foley, A Monotonicity Result for the Blocking Probability in a G/GI/c/m Queueing System, under review, available at [www.unc.edu/~ziya/monotone-blocking.pdf](http://www.unc.edu/~ziya/monotone-blocking.pdf).