Optimal Pricing for a Service Facility with Congestion Penalties

A Thesis
Presented to
The Academic Faculty

by

Idriss Maoui

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Industrial and Systems Engineering
Georgia Institute of Technology
May 2006

Optimal Pricing for a Service Facility with Congestion Penalties

Approved by:

Dr. Hayriye Ayhan, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Robert D. Foley, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Richard F. Serfozo
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yang Wang
School of Mathematics
Georgia Institute of Technology

Dr. Amy R. Ward
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: April 3, 2006

To Abderrahmane Abdelmoumene

ACKNOWLEDGEMENTS

I would like to thank all of my committee members. Thank you to Drs. Amy Ward, Richard Serfozo and Yang Wang for their teachings and their letters of recommendation as well as their comments and suggestions. I am greatly indebted to my advisors Drs. Hayriye Ayhan and Robert Foley for their commitment and understanding. Their advice and support were an invaluable help in my doctoral studies. Finally, I wish to thank my family for their unconditional love and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF NOTATION

(α, β) Support of willingness-to-pay distribution F, p. 10.

(α_i, β_i) Support of willingness-to-pay distribution F_i, p. 44.

c_s^2 Squared coefficient of variation of service time, p. 11.

$F(\cdot)$ Willingness-to-pay distribution, p. 10.

$f(\cdot)$ Probability density function of willingness-to-pay distribution, p. 10.

$F_i(\cdot)$ Willingness-to-pay distribution for class-i customers, p. 43.

$f_i(\cdot)$ Probability density function of willingness-to-pay distribution F_i, p. 43.

g_N^* Optimal revenue under dynamic pricing when capacity is N, p. 44.

h Holding cost per unit time incurred by each customer in the system, p. 12.

h_s Holding cost incurred per unit time when system is in state s, p. 45.

I Number of customer classes, p. 43.

$L(\rho, N)$ Expected number of customers in the system under traffic intensity ρ and capacity N, p. 11.

$L^q(\rho, N)$ Expected number of customers in the queue under traffic intensity ρ and capacity N, p. 40.

Λ Customers' arrival rate, p. 10.

$\lambda(y)$ Arrival rate of customers who are willing to pay a price of at most y, p. 11.

μ Service rate for one server, p. 11.

μ_s Service rate when system is in state s, p. 44.

N System capacity, p. 10.

N^* Optimal capacity level, p. 28.

$N(t)$ Counting process of customers' arrivals, p. 10.

$N(y,t)$ Counting process of customers who are willing to pay a price of at most y, p. 11.

$\pi_n(\rho, N)$ Stationary distribution of the queueing process under traffic intensity ρ and capacity N, p. 11.

$\pi_n(\mathbf{z})$ Stationary distribution of the queueing process under stationary pricing policy $\mathbf{z}$, p. 45.

p_s Probability that a customer is willing to enter a system with s customers, p. 29.

$r(\cdot)$ Hazard rate of willingness-to-pay distribution, p. 10.

$R(y,N)$ Long-run average profit per unit time for static price y and capacity N, p. 11.

$r_i(\cdot)$ Hazard rate of willingness-to-pay distribution F_i, p. 44.

$R(\mathbf{z})$ Long-run average profit per unit time for stationary pricing policy $\mathbf{z}$, p. 44.

$\rho(y)$ Traffic intensity when price is y, p. 11.

R_N^* Optimal revenue under static pricing when capacity is N, p. 12.

θ Reneging rate, p. 35.

$X(t)$ Number of customers in the system at time t, p. 44.

y Advertised price, p. 10.

y_N^* Optimal static price to be charged when capacity is N, p. 12.

$\hat{y}$ Maximum price under which we have a traffic intensity of 1, p. 11.

$\mathbf{z}$ Stationary pricing policy, where price $z_{i,s}$ is advertised to class-i customers when the system is in state s, p. 44.

$\mathbf{z}^*$ Optimal stationary pricing policy, p. 45.

SUMMARY

We consider the optimal pricing problem in a service facility in order to maximize its long-run average profit per unit time. We model the facility as a queueing process that may have finite or infinite capacity. Customers are admitted into the system if it is not full and if they are willing to pay the price posted by the service provider.

Moreover, the congestion level in the facility incurs penalties that greatly influence profit. We model congestion penalties in three different manners: holding costs, balking customers and impatient customers. First, we assume that congestion-dependent holding costs are incurred per unit of time. Second, we consider that each customer might be deterred by the system congestion level and might balk upon arrival. Third, customers are impatient and can leave the system with a full refund before being serviced.

We are interested in both static and dynamic pricing for all three types of congestion penalties. In the static case, we demonstrate that there is a unique optimal price that maximizes the long-run average profit per unit time. We also investigate how optimal prices vary as system parameters change. In the dynamic case, we show the existence of an optimal stationary policy in a continuous and unbounded action space that maximizes the long-run average profit per unit time. We provide explicit expressions for this policy under certain conditions. We also analyze the structure of this policy and investigate its relationship with our optimal static price.

CHAPTER I

INTRODUCTION

Determining the optimal price to be charged for a service facility is a critical decision for a manager. There is a trade-off between high prices and high demand that greatly influences revenue. Not all customers react the same way to advertised prices and the maximum amount that each customer is willing to pay is often random.

Moreover, the congestion level in the system affects the operating efficiency of the facility and incur penalties that reduce profit. In a congested service facility with scarce resources, entering customers might wait for service and might even be turned down by the service provider who can not accommodate them. There is a loss of service quality and customer goodwill associated with long waiting times. In addition, potential customers are often deterred by high congestion levels when this information is available to them through quoted lead times. Congestion plays a key role and affects the operating performance as well customers' behavior. A service provider who ignores congestion not only overestimates profit but also fails to realize that pricing can be used to control congestion costs. Setting high prices can be helpful to reduce congestion by dissuading customers from entering an overcrowded facility.

Although it has applications in other service industries, the work in this dissertation was originally motivated by the pricing problem of outsourced computer services. These services offer processing power, server time or bandwidth resources and are provided to businesses that do not have sufficient in-house capabilities. These solutions present an

inexpensive and flexible way to handle spikes in computing needs for businesses with limited resources. Sam Palmisano, IBM's CEO, forsees a near future in which "businesses [would] buy computing power on demand, similar to the way electricity is purchased" (see http://news.com.com/IBM+talks+up+computing+on+demand/2100-1001_3-963807.html). As businesses computing needs grow larger, these products give smaller companies access to supercomputing power that only very large corporations could afford (see http://www-03.ibm.com/press/us/en/pressrelease/7949.wss). The most prominent providers of such services include IBM, Hewlett-Packard, Cisco Systems, AT&T and Schlumberger. In the same fashion as utilities, the prices of these services should increase with congestion and usage. Our objective is to develop a better understanding of how congestion affects the optimal pricing decisions of the provider of such services.

We model the service facility as a queueing system with finite or infinite capacity (size). The queueing framework enables us to capture the variability of service times, customers' arrival times and customers' price sensitivity, as well as analytically tractable congestion penalties. We suppose that customers have independent identically distributed valuations of service and enter the system when it is not full and when their valuation is greater than the current advertised price. We will refer to the distribution of service valuation as *willingness-to-pay distribution* and we assume that the associated process is independent of arrival and service times and that prices are paid upon arrival.

We are interested in both static and dynamic pricing. The service provider is said to use static pricing, when prices are set at time zero and cannot be changed during the lifetime of the system. The service provider is said to use dynamic pricing, when prices can be adjusted in time. Note that this usually translates into having congestion-dependent prices; that is, prices that depend on the current congestion level of the system. When dynamic pricing

is in use, we refine our model by segmenting customers into a finite number of classes. Each class forms a homogeneous group where customers have the same willingness-to-pay distribution. However, the willingness-to-pay distribution might be different from one class to another. The service provider may advertise different prices to each customer class. This ability to have class-specific congestion-dependent prices is referred to as *dynamic precision pricing*. Moreover, when congestion penalties are incurred, dynamic pricing allows price adjustments to the current congestion level. Therefore, pricing can be dynamically adapted to the cost of congestion in each state.

The objective our work is to determine the optimal static and dynamic pricing policies that maximize the long-run average profit per unit time for a service facility subject to congestion penalties. Then, we seek to analyze structural properties of our results as well as their sensitivity to system parameters. This includes understanding the congestion control features of optimal prices and studying how optimal prices change as system parameters vary.

We develop our model by capturing congestion penalties in three different ways: holding costs, balking customers and impatient customers. In the holding cost model, we assume that state-dependent holding costs are incurred per unit of time. This is the case when each customer incurs a fixed cost per unit time spent in the system. One can think of this congestion penalty as a loss of customer goodwill that is proportional to the customer sojourn time. The longer the customer waits, the less likely the customer is to return. Although the system congestion is experienced by customers, the service provider indirectly bears its cost as she experiences its effects as a future loss of revenue.

In the model with balking customers, customers directly react to the current congestion level. In this case, we consider that each customer has a random congestion valuation

to make her decision whether to enter the system or not. Hence, only customers who are willing to pay the current price and who tolerate the current congestion level enter the system, otherwise they balk and do not pay. Note that this random congestion valuation can be considered as *willingness-to-wait* with a willingness-to-wait distribution. Such a behavior is experienced in service facilities where potential customers are quoted a service lead time and make their decision to purchase products based both on price and quoted waiting time. Internet commerce companies encounter this issue in periods of heavy demand, as they provide customers with expected shipping and delivery times.

In the model with impatient customers, customers who are willing to pay the current advertised price enter the system if it is not full and pay upon arrival. However, each customer waiting for service is impatient and reneges if he is not serviced prior to his maximum waiting time. We assume that customers' maximum waiting times are independent, identically distributed exponential random variables and that reneging customers are given a full refund. Note that in stable systems, this is equivalent to a model where customers pay upon service completion. This type of customer behavior is experienced in call centers, where customers have to wait for an operator to purchase a product. Impatient customers might renege and hang up and only customers who complete the phone transaction contribute to the bottom line.

The outline of this dissertation is as follows. In Chapter 2, we review the literature on the issue of pricing in queueing systems related to our work. Then, we decompose our analysis into two parts: static pricing in Chapter 3 and dynamic precision pricing in Chapter 4. In each of these two chapters, we describe our results for each congestion model: holding costs, balking customers and impatient customers. Chapter 5 contains the conclusion to our work as well directions for future research. In Appendices A and B, we show technical

results that support our work in Chapters 3 and 4, respectively.

CHAPTER II

LITERATURE REVIEW

The use of nontraditional pricing strategies in order to maximize profit in service facilities has generated much interest in the recent years. Although our work directly considers pricing in order to maximize profit, it is inspired by a series of papers on the more general topic of congestion control in queueing systems. We can group them in two categories, depending on whether the control is static or dynamic.

The paper by Naor [16] is the first one that combines the issues of pricing and congestion control in queues. Naor's work and many papers extending it (such as Knudsen [9] and Yechiali [21]) analyze systems where customers make a decision to enter a service facility based on its current queue length. Entering customers obtain a fixed reward and are charged a holding cost function of their time spent in the system. In order to maximize their utility, they decide to join or balk (join-balk rule). The service provider then imposes an entrance fee to induce an optimal customer admission rate. Larsen [10] and Hassin [6] evaluate the effect of releasing the expected queue length to potential customers as opposed to the current queue length.

Mendelson and Whang [15] consider customers who make their decision to enter the system based both on price and delay. Mendelson and Whang [15] also include different customer classes that have different demand functions and delay costs. Prices are then used by the decision maker as an incentive to induce optimal customer arrival rates and execution priorities.

Ittig [7] develops a model in which congestion is treated as a form of price. His objective is not optimal pricing but he determines the optimal number of servers for the service facility. He introduces a general demand function relating average waiting time and demand rate as well as a cost of service capacity. He sets up a nonlinear constrained optimization problem where the queueing link between demand rate and average waiting time is a constraint. Ittig [8] is also interested in estimating the optimal number of servers through transaction data when the relationship between demand and congestion is not explicitly known.

Ziya [22, 24] focuses on optimal static pricing for systems without holding costs in $M/G/1/\infty$ and $M/M/1/N$ queueing systems. Instead of using a congestion-based join-balk rule, he links the customers' arrival rate to the posted price through a random service valuation for each customer. He uses a willingness-to-pay distribution to capture the proportion of customers willing to pay the posted price and shows the existence of a unique optimal price that maximizes the long-run average profit. Ziya [22, 24] also exhibits how the optimal price changes as system parameters vary and addresses the issue of precision static pricing, where the service provider can advertise static class-specific prices in systems without holding costs.

In all the papers mentioned above, the system controls are static: that is, the controls are set by the decision maker once and remain unchanged throughout the life of the system. In the second group of papers, controls are allowed to depend on the state of the system (dynamic control). Stidham [19] develops a dynamic admission control model to optimize an infinite-horizon discounted reward with convex holding costs in single server queues. Stidham's decision variable is defined as whether to accept or reject an incoming job. Each accepted job yields a fixed deterministic reward. He shows the existence of a monotonic optimal stationary policy. He also extends his results to simple networks of queues.

On the other hand, George and Harrison [4] allow the service provider to dynamically control the service rate instead of the arrival rate. There is a penalty that depends on the chosen service rate and the objective is to minimize the long-run average cost in systems with holding costs.

Combining the problems of setting admission rates and service rates, Ata and Shneorson [2] consider a dynamic control model where the service provider sets state-dependent admission rates and service rates in an $M/M/1$ queue with holding costs. There is a reward associated with the chosen admission rates and a cost corresponding to the chosen service rates. After explicitly solving this problem, they analyze a decentralized model, where only service rates and prices are decision variables. The service provider must set them so that the optimal arrival rates are induced by customers maximizing their own utility.

Low [12], [13] is interested in dynamic pricing in $M/M/s$ queues with finite or infinite capacity but with a finite action space. Low does not use a willingness-to-pay distribution but each price in the action space corresponds to a given positive arrival rate. Low also considers state-dependent holding costs incurred as a lump sum as a customer arrives. He makes the extra assumption that holding costs are bounded and that the facility has multiple identical servers. He shows that optimal prices are nondecreasing as the system becomes congested and develops an algorithm to solve the Markov decision process formulation of the problem. Aktaran and Ayhan [1], as well as Çil, Karaesmen and Örmeci [3], further investigate the sensitivity of the optimal prices to system parameters. Paschalidis and Tsitsiklis [17] focus on models with multiple classes of customers that have different resource requirements without holding costs.

Differing from earlier work, we consider systems with holding costs, systems with balking customers and systems with impatient customers as alternative ways to capture congestion

penalties. We extend Ziya's work on static pricing by introducing these three types of congestion penalties and by considering capacity as a decision variable. We also extend Low's work on dynamic pricing with holding costs by introducing a general parameter structure and by considering a continuous unbounded action space for multiple customer classes in finite or infinite capacity systems. Unlike models with holding costs, the issue of pricing in queues with balking or impatient customers with refund has received little attention. Our objective is to analyze how congestion influences pricing in both static and dynamic pricing settings.

CHAPTER III

STATIC PRICING

3.1 Model Description

In this chapter, the service provider can only advertise one price at all times for all customers. Therefore, an optimal policy is characterized by a single advertised price. After describing the model, we determine the optimal prices for the service facility subject to each of the three congestion penalties. First, we focus on system with holding costs. Then, we analyze systems with balking customers, and finally, we consider systems with impatient customers.

We model the service facility as a single server system, where $N \leq \infty$ is the maximum number of customers allowed in the system at any time. Arriving customers enter if the system is not full and if they are willing to pay the price charged by the service provider.

Let y denote the mark-up charged for service. Note that the price to be charged is the sum of the mark-up and the variable cost of service. Without loss of generality, we assume that the variable cost of service is zero, so the mark-up is equal to the price. Let $N(t)$ be the number of arrivals in the time interval $(0, t]$. We assume that $\{N(t) : t \geq 0\}$ is a Poisson process with rate Λ . We call Λ the maximum arrival rate. For $y \geq 0$, let $F(y)$ be the proportion of customers willing to pay a price of at most y. We call $F(\cdot)$, the *willingness-to-pay distribution.* We assume that the cumulative distribution function $F(\cdot)$ is absolutely continuous with density $f(\cdot)$, support (α, β) and finite mean. Let $r(\cdot)$ denote the hazard rate function of $F(\cdot)$; that is, $r(y) = \frac{f(y)}{1-F(y)}$ for $\alpha < y < \beta$ and we define $r(y) = 0$ for $y \leq \alpha$ and $r(y) = \infty$ for $y \geq \beta$. In what follows, we assume that F has IGHR (Increasing

Generalized Hazard Rate); that is, $yr(y)$ is strictly increasing for all y in $[\alpha, \beta)$.

Assumption IGHR is equivalent to the demand function having decreasing price elasticity (see Proposition 2.1 in Ziya [23]). Many common distribution functions (such as exponential and uniform distributions) have this property that simply states that the demand becomes more elastic as prices decrease.

Let $N(y,t)$ be the number of customers who are willing to pay a price of at most y and arrive during $(0,t]$. Let $\lambda(y)$ denote the arrival rate of customers who are willing to pay a price of at most y, so that $\lambda(y) = \Lambda(1 - F(y)) = \lim_{t\to\infty} \frac{N(y,t)}{t}$.

Service times are independent, identically distributed random variables with distribution $G(\cdot)$, mean $\frac{1}{\mu}$ and squared coefficient of variation c_s^2. The service process, the arrival process and the process associated with the amounts successive customers are willing to pay are assumed to be independent.

When the price is y, the number of customers in the system forms a queueing process with Poisson arrival process $\{N(y,t) : t \geq 0\}$ and independent, identically distributed service times with c.d.f $G(\cdot)$. When an arriving customer is willing to pay the posted price, the customer enters the system if the system is not full; otherwise, the customer is lost.

Let $\rho(y) = \frac{\Lambda}{\mu}(1 - F(y))$ denote the traffic intensity when the price is y. Let $\hat{y}$ be the maximum price under which we have a traffic intensity of 1; that is, $\hat{y} = \sup\{y : \rho(y) = 1\}$ when $\frac{\Lambda}{\mu} \geq 1$. Note that when $\frac{\Lambda}{\mu} < 1$, $\hat{y} = -\infty$. We define the state of the system $X(t)$ as the number of customers in the system at time t. When they exist, $\{\pi_n(\rho, N)\}$ and $L(\rho, N)$ denote the stationary distribution and the expected number of customers in the system for traffic intensity ρ and capacity N.

Let $R(y, N)$ be the long-run average profit per unit time for a posted price y and capacity

N. The optimal static pricing problem can be formulated as:

$$\max_y R(y, N), \text{ subject to } y \geq 0.$$

When it exists and is unique, we let y_N^* denote the optimal price to be charged to maximize $R(y, N)$ and $R_N^* = R(y_N^*, N)$ denote the optimal objective value.

In the following, we consider the optimal pricing problem for the service facility to each of the three congestion penalties: first, we focus on system with holding costs. Then, we analyze systems with balking customers and finally, we consider systems with impatient customers.

3.2 Systems with Holding Costs

In this section, we capture congestion penalties through holding costs. More specifically, we assume that each entering customer pays the posted price at the time of arrival and incurs a cost of h per unit time while in the system as in Figure 1. To ensure that a positive long-run profit is attainable, we will assume that $\frac{h}{\mu} < \beta$.

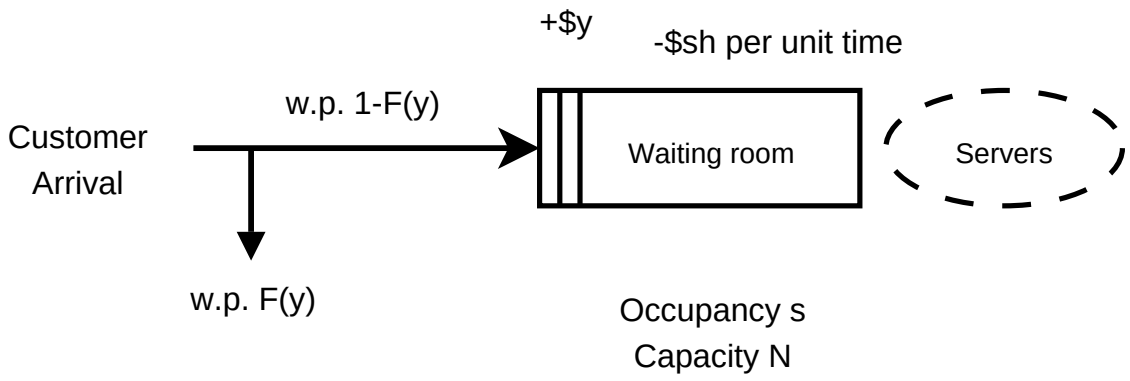

Figure 1: System with Holding Costs

3.2.1 Optimal Pricing for M/G/1/∞ Queues

In the following, we derive expressions for $R(y,\infty)$ and y^*_∞ when no further assumptions are made on the service time distribution. Only customers who are willing to pay the posted price y enter the system and they pay y immediately. Since they incur an additional cost of h per unit time that they spend in the system,

$$R(y,\infty) = \lim_{t\to\infty} \frac{yN(y,t) - h\sum_{k=1}^{N(y,t)} D_k}{t},$$

where $\{D_k : k = 1, 2, \ldots \infty\}$ is the sequence of the total waiting times for successive customers. Note that

$$\lim_{t\to\infty} \frac{\sum_{k=1}^{N(y,t)} D_k}{t} = \lim_{t\to\infty} \frac{\sum_{k=1}^{N(y,t)} D_k}{N(y,t)} \frac{N(y,t)}{t} = L(\rho(y),\infty).$$

Therefore, we can write the long-run average reward per unit time as

$$R(y,\infty) = y\lambda(y) - hL(\rho(y),\infty).$$

Clearly, if $\rho(y) \geq 1$ and $h > 0$, then $L(\rho(y),\infty) = \infty$ and $R(y,\infty) = -\infty$. From the Pollaczek-Khinchin formula [5], if $\rho(y) < 1$,

$$L(\rho(y),\infty) = \frac{\rho(y)(2-\rho(y)(1-c_s^2))}{2(1-\rho(y))}.$$

Therefore, if $\rho(y) < 1$,

$$R(y,\infty) = y\lambda(y) - h\frac{\rho(y)(2-\rho(y)(1-c_s^2))}{2(1-\rho(y))}.$$

Note that the long-run average reward function consists of two terms : the first describing the revenue through the arrival rate regardless of the service times, whereas the second accounts for the additional holding cost through the steady-state average number of customers in the system.

In the following result, we show the existence and the uniqueness of an optimal price.

Theorem 3.2.1 *There exists a unique optimal price given by :*

$$y^*_\infty = \inf\left\{y : r(y)\left(y - \frac{h}{\mu}\varphi(\rho(y))\right) \geq 1\right\},$$

where

$$\varphi(\rho) = \begin{cases} \frac{1+\rho(c_s^2-1)(1-\frac{\rho}{2})}{(1-\rho)^2} & \text{if } \rho < 1, h > 0, \\ \infty & \text{if } \rho \geq 1, h > 0, \\ 0 & \text{if } h = 0. \end{cases}$$

Proof First, assume $h > 0$ and let $\hat{\alpha} = \max(\alpha, \hat{y})$. Note that for all y less than or equal to $\hat{y}$, the reward function is equal to $-\infty$. Therefore, an optimal price, if it exists, has to be greater than $\hat{y}$. Since $F(\cdot)$ is absolutely continuous, for all y in $[\hat{\alpha}, \beta)$, $R(y, \infty)$ is continuous and a.e. differentiable on $[\hat{\alpha}, \beta)$. We can rewrite $R(y, \infty)$ as

$$R(y, \infty) = \rho(y)\left(\mu y - h\frac{(2 - \rho(y)(1 - c_s^2))}{2(1 - \rho(y))}\right).$$

Note that

$$\mu y - h\frac{(2 - \rho(y)(1 - c_s^2))}{2(1 - \rho(y))} \to \mu\beta - h > 0 \text{ as } y \text{ tends to } \beta\ .$$

Therefore, there exists y in $[\hat{\alpha}, \beta)$ such that $R(y, \infty) > 0$. Moreover, for all y in $[\hat{\alpha}, \beta)$, $R(y, \infty) < \infty$ and $R(y, \infty) \to 0$ as $y \to \beta$. So, there exists an optimal price in $[\hat{\alpha}, \beta)$.

If $R(y, \infty)$ is differentiable with respect to $y \in [\hat{\alpha}, \beta)$, then $\frac{\partial R(y,\infty)}{\partial y} > 0 (< 0)$ if and only if $r(y)(y - \varphi(y)) < 1 (> 1)$. Since there exists y in $[\hat{\alpha}, \beta)$ such that $R(y, \infty) > 0$ and $R(y, \infty) \to 0$ as $y \to \beta$, there exists y in $[\hat{\alpha}, \beta)$ such that $r(y)(y - \frac{h}{\mu}\varphi(\rho(y))) > 1$. Note that $\frac{\partial \varphi(\rho(y))}{\partial y} = -\frac{r(y)\rho(y)(c_s^2+1)}{(1-\rho(y))^3} \leq 0$, for all y in $[\hat{\alpha}, \beta)$, so $\varphi(\rho(y)) \geq \varphi(0) = 1$. Under Assumption IGHR, $r(y)(y - \frac{h}{\mu}\varphi(\rho(y)))$ is strictly increasing in the interval $[\inf\{y : r(y)(y - \frac{h}{\mu}\varphi(\rho(y))) \geq 1\}, \beta)$, so $R(y, \infty)$ is decreasing in the interval $(\inf\{y : r(y)(y - \frac{h}{\mu}\varphi(\rho(y))) \geq 1\}, \beta)$. In the

same fashion, $R(y,\infty)$ is increasing in the interval $(\hat{\alpha}, \inf\{y : r(y)(y - \frac{h}{\mu}\varphi(\rho(y))) \geq 1\})$. Eventually, we can conclude that $y^*_\infty = \inf\{y : r(y)(y - \frac{h}{\mu}\varphi(\rho(y))) \geq 1\}$. □

When $h = 0$, the result reduces to $y^*_\infty = \inf\{y : yr(y) \geq 1\}$, which agrees with the characterization of the optimal price in Proposition 3.3.1 of Ziya [22].

We can use Theorem 3.2.1 to derive the following result for M/M/1/∞ queueing systems.

Corollary 3.2.1 *If the service times are exponentially distributed and $h > 0$, then there exists a unique optimal price given by :*

$$y^*_\infty = \inf\left\{y : r(y)\left(y - \frac{h}{\mu}\varphi(\rho(y))\right) \geq 1\right\},$$

where

$$\varphi(\rho) = \begin{cases} \frac{1}{(1-\rho)^2} & \text{if } \rho < 1, h > 0, \\ \infty & \text{if } \rho \geq 1, h > 0, \\ 0 & \text{if } h = 0. \end{cases}$$

Depending on the willingness-to-pay distribution, it might be difficult to compute y^*_∞. The next result provides bounds (which are obtained by replacing $\varphi(\rho(y))$ by $\varphi(\frac{\Lambda}{\mu})$ in the expression of y^*_∞) on the optimal price. These bounds will also be used in the next section in order to compare the properties of systems with and without holding costs.

Proposition 3.2.1 *The unique optimal price y^*_∞ satisfies $\frac{h}{\mu} \leq y^*_\infty$. Moreover, if $\Lambda < \mu$, then*

$$y^*_\infty \leq \inf\left\{y : y\left(r(y) - \frac{h}{\mu}\varphi(\frac{\Lambda}{\mu})\right) \geq 1\right\}.$$

Proof First, suppose that $y^*_\infty < \frac{h}{\mu}$. If $h = 0$, then there is clearly a contradiction. Assume now that $h > 0$. Since $L(\rho(y),\infty) \geq \rho(y)$, we have $R(y^*_\infty,\infty) \leq y^*_\infty\lambda(y^*_\infty) - \frac{h}{\mu}\lambda(y^*_\infty) < 0$,

which is a contradiction since we proved in Theorem 3.2.1 that there exists y in $[\hat{\alpha}, \beta)$ such that $R(y, \infty) > 0$. Therefore, $\frac{h}{\mu} \leq y^*_\infty$.

Now suppose $\Lambda < \mu$. From Theorem 3.2.1, $y^*_\infty = \inf\{y : r(y)(y - \frac{h}{\mu}\varphi(\rho(y))) \geq 1\}$. Since $\varphi(\rho(\cdot))$ is nonincreasing, $\frac{1+\frac{\Lambda}{\mu}(c_s^2-1)(1-\frac{\Lambda}{2\mu})}{(1-\frac{\Lambda}{\mu})^2} = \varphi(\frac{\Lambda}{\mu}) \geq \varphi(\rho(y))$ for $y \geq 0$. Therefore, for all y in (α, β) such that $r(y)(y - \frac{h}{\mu}\varphi(\frac{\Lambda}{\mu})) \geq 1$, we have $r(y)(y - \frac{h}{\mu}\varphi(\rho(y)))$. This completes the proof. □

We now compare the optimal price and the optimal reward in two M/G/1/∞ systems (indexed by 1 and 2). These two systems differ by marginal holding cost, maximum arrival rates, service rates and squared coefficients of variation. Moreover, we also compare systems where the willingness-to-pay distributions are ordered in the stochastic ordering and hazard rate ordering. Recall that distribution F_1 is greater than or equal to distribution F_2 in the stochastic ordering ($F_1 \geq_{ST} F_2$) if and only if $F_1(y) \leq F_2(y), \forall y \geq 0$. Furthermore, distribution F_1 is greater than or equal to distribution F_2 in the hazard rate ordering ($F_1 \geq_{HR} F_2$) if and only if $r_1(y) \leq r_2(y), \forall y \geq 0$. Our objective is to compare the optimal prices $y^*_{\infty,1}$ and $y^*_{\infty,2}$ for these two systems.

In systems without holding cost, Ziya [22] shows that if $F_1 \geq_{HR} F_2$, then $y^*_{\infty,1} \geq y^*_{\infty,2}$. We show in the next proposition that this result still holds when holding costs are incurred. However, stochastic ordering of the willingness-to-pay distributions do not necessarily guarantee ordered optimal prices (see Section 3.4 in Ziya [22] for a counterexample). In the remainder of this section, parameters relative to system $i = 1, 2$ are indicated by subscript i.

Proposition 3.2.2 *Consider two systems 1 and 2 that satisfy all of the following :*

1. $\Lambda_1 \geq \Lambda_2$,

2. $F_1 \geq_{HR} F_2$,

3. $h_1 \geq h_2$,

4. $c_{s,1}^2 \geq c_{s,2}^2$,

5. $\mu_1 \leq \mu_2$.

Then, $y_{\infty,1}^* \geq y_{\infty,2}^*$.

Proof From Theorem 3.2.1, we have $y_{\infty,i}^* = \inf\{y : r_i(y)(y - \frac{h_i}{\mu_i}\varphi_i(\rho_i(y))) \geq 1\}$ for system $i = 1, 2$. Since hazard rate ordering implies stochastic ordering, $F_1 \geq_{ST} F_2$. In conjunction with conditions 1 and 5, this implies that $\rho_1(\cdot) \geq \rho_2(\cdot)$. Moreover, we showed in the proof of Theorem 3.2.1 that $\varphi(\cdot)$ is nondecreasing. Therefore, we have $\varphi_1(\rho_1(\cdot)) \geq \varphi_2(\rho_2(\cdot))$ from condition 4. Suppose that y is such that $r_1(y)(y - \frac{h_1}{\mu_1}\varphi_1(\rho_1(y))) \geq 1$. Then, $r_2(y)(y - \frac{h_2}{\mu_2}\varphi_2(\rho_2(y))) \geq 1$. Thus, $y_{\infty,2}^* \leq y_{\infty,1}^*$. □

As shown in the proof of Proposition 3.2.2, it is intuitive that systems with higher maximum arrival rate, smaller service rates, higher service variance and higher marginal holding cost yield higher long-run average holding cost. Therefore, it is not surprising that the optimal price should be higher when higher holding costs are incurred.

For systems with no holding cost, Ziya [22] shows that if $\frac{\Lambda_1}{\mu_1} \geq \frac{\Lambda_2}{\mu_2}$, then $y_1^* \geq y_2^*$. However, this result of Ziya does not extend to facilities with holding costs. Consider two M/M/1/∞ systems, with $\mu_1 = 1, \mu_2 = .1, \Lambda_1 = 0.5$, and $\Lambda_2 = 0.0001$. So, $\frac{\Lambda_2}{\mu_2} = 0.001 \leq \frac{\Lambda_1}{\mu_1} = 0.5$. For both systems, assume that the willingness-to-pay distribution is exponential with rate 1. According to Proposition 3.2.1 the optimal solution for system $i = 1, 2$ satisfies:

$$\frac{h}{\mu_i} \leq y_{\infty,i}^* \leq 1 + \frac{h\mu_i}{(\mu_i - \Lambda_i)^2}.$$

So, $y^*_{\infty,1} \leq 4h+1$ and $10h \leq y^*_{\infty,2}$. Consider $h = 1$. Therefore, $y^*_{\infty,1} < y^*_{\infty,2}$, although $\frac{\Lambda_1}{\mu_1} \geq \frac{\Lambda_2}{\mu_2}$.

In the next result, we analyze how the optimal reward varies as parameters change.

Proposition 3.2.3 *Consider two systems 1 and 2 that satisfy all of the following :*

1. $\Lambda_1 \geq \Lambda_2$,
2. $F_1 \geq_{ST} F_2$,
3. $h_1 \leq h_2$,
4. $c^2_{s,1} \leq c^2_{s,2}$,
5. $\mu_1 \geq \mu_2$.

Then, $R^*_{\infty,1} \geq R^*_{\infty,2}$.

Proof To prove this result, we split our proof into two parts. First, we show that the result holds when conditions 1 and 2 are changed to equalities. Second, we show that it holds when conditions 3,4 and 5 are changed to equalities. By composition, the result holds under all the conditions as well.

Suppose that conditions 3,4 and 5 hold and $\Lambda_1 = \Lambda_2$ and $F_1(\cdot) = F_2(\cdot)$. Recall that for all $y \geq 0$,

$$R_i(y, \infty) = y\lambda_i(y) - h_i L_i(\rho_i(y), \infty).$$

Conditions 3,4 and 5 imply that $h_1 L_1(\rho_1(y), \infty) \leq h_2 L_2(\rho_2(y), \infty)$. Therefore, $R_1(y, \infty) \geq R_2(y, \infty)$ and $R^*_{\infty,1} \geq R^*_{\infty,2}$.

Now suppose that 1 and 2 hold, whereas 3,4 and 5 are equalities. Since $F_1(\cdot)$ is absolutely continuous and $\lambda_1(\cdot) \geq \lambda_2(\cdot)$, there exists $\delta > 0$ such that $\lambda_1(y^*_2 + \delta) = \lambda_2(y^*_2)$. Therefore,

system 1 with price $y_2^* + \delta$ has the same arrival and service rates as system 2 with price y_2^*. Therefore, system 1 with price $y_2^* + \delta$ performs better than system 2 with optimal price. Hence, $R_{\infty,1}^* \geq R_{\infty,2}^*$ and the proof is complete. □

3.2.2 Optimal Pricing for M/M/1/N Queues

In this section, we study optimal pricing for capacitated queues. We focus on M/M/1/N queueing systems for which we can easily quantify the long-run average queue length and the long-run average reward function. We prove the existence of a unique optimal price under the IGHR assumption and derive ordering properties as system parameters change.

Only customers who are willing to pay the posted price y and find fewer than N customers in the system are allowed to enter. Therefore,

$$R(y, N) = \lim_{t\to\infty} \frac{yN_{in}(y,t) - h\sum_{k=1}^{N_{in}(y,t)} D_k}{t}$$

where $N_{in}(y,t)$ denotes the number of customers allowed in the system up to time t. From Little's result,

$$R(y, N) = y\lambda(y)(1 - \pi_N(\rho(y), N)) - hL(\rho(y), N).$$

Recall from Gross and Harris [5] that

$$L(\rho(y), N) = \frac{\rho(y)(1 - (N+1)\rho(y)^N + N\rho(y)^{N+1})}{(1-\rho(y))(1-\rho(y)^{N+1})} \text{ and}$$

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

We can also express the long-run average reward per unit time as

$$R(y, N) = y\mu(1 - \pi_0(\rho(y), N)) - h\rho(y)\frac{(1 - (N+1)\rho(y)^N + N\rho(y)^{N+1})}{(1-\rho(y))(1-\rho(y)^{N+1})}.$$

We demonstrate the existence and the uniqueness of an optimal price in Theorem 3.2.2.

Theorem 3.2.2 *There exists a unique optimal price given by :*

$$y_N^* = \inf\left\{y : r(y)\gamma_N^h(\rho(y))\left(y - \frac{h}{\mu}\varphi_N(\rho(y))\right) \geq 1\right\},$$

where

$$\varphi_N(\rho) = -\frac{\frac{\partial L(\rho,N)}{\partial \rho}}{\frac{\partial \pi_0(\rho,N)}{\partial \rho}} = \begin{cases} \frac{1-(N+1)^2\rho^N(1+\rho^2)+2N(N+2)\rho^{N+1}+\rho^{2N+2}}{(1-\rho)^2(1-(N+1)\rho^N+N\rho^{N+1})} & \text{if } \rho \neq 1, \\ \frac{1}{6}N^2 + \frac{1}{2}N + \frac{1}{3} & \text{if } \rho = 1, \end{cases}$$

and

$$\gamma_N^h(\rho) = \begin{cases} \frac{1+N\rho^{N+1}-(N+1)\rho^N}{(1-\rho^{N+1})(1-\rho^N)} & \text{if } \rho \neq 1, \\ \frac{1}{2} & \text{if } \rho = 1. \end{cases}$$

Proof First, we will prove that there exists an optimal solution. Note that for all y in $[\alpha, \beta)$, $L(\rho(y), N) \leq L(\rho(y), \infty) = \frac{\rho(y)}{1-\rho(y)}$. When y is in the neighborhood of β, $\rho(y) < 1$ and $L(\rho(y), \infty) < \infty$. So, for y in the neighborhood of β,

$$\begin{aligned} R(y, N) &\geq y\lambda(y)(1 - \pi_N(\rho(y), N)) - h\frac{\rho(y)}{1-\rho(y)} \\ &\geq \frac{\rho(y)}{1-\rho(y)}(\mu y(1 - \pi_N(\rho(y), N))(1 - \rho(y)) - h). \end{aligned}$$

Under the assumption that $\frac{h}{\mu} < \beta$,

$$\mu y(1 - \pi_N(\rho(y), N))(1 - \rho(y)) - h \to \mu\beta - h > 0, \text{as } y \to \beta.$$

Therefore, there exists y in $[\alpha, \beta)$ such that $R(y, N) > 0$. Note that $R(\cdot, N)$ is continuous and also bounded on $[\alpha, \beta)$ since $R(y, N) \to 0$ as $y \to \beta$. Hence, there exists an optimal price in $[\alpha, \beta)$.

If $R(y, N)$ is differentiable with respect to y in the interval $[\alpha, \beta)$, then $\frac{\partial R(y,N)}{\partial y} > 0 (< 0)$ if and only if $r(y)\gamma_N^h(\rho(y))(y - \frac{h}{\mu}\varphi_N(\rho(y))) < 1 (> 1)$. Since there exists y in $[\alpha, \beta)$ such that $R(y, N) > 0$ and $R(y, N) \to 0$ as $y \to \beta$, there exists y in $[\alpha, \beta)$ such that

$r(y)\gamma_N^h(\rho(y))(y-\frac{h}{\mu}\varphi_N(\rho(y)))\geq 1$. It remains to prove that $r(y)\gamma_N^h(\rho(y))(y-\frac{h}{\mu}\varphi_N(\rho(y)))$ is increasing in $[\inf\{y: r(y)\gamma_N^h(\rho(y))(y-\frac{h}{\mu}\varphi_N(\rho(y)))\geq 1\},\beta)$.

According to Lemma A.1 in Ziya [24], $\gamma_N^h(\rho(y))$ is nondecreasing for $y>0$. From Lemma A.0.1 $\varphi_N(\cdot)$ is nondecreasing, $\varphi(\rho(\cdot))$ is nonincreasing and nonnegative and $r(y)\gamma_N^h(\rho(y))(y-\frac{h}{\mu}\varphi_N(\rho(y)))$ is increasing in the interval $[\inf\{y: r(y)\gamma_N^h(\rho(y))(y-\frac{h}{\mu}\varphi_N(\rho(y)))\geq 1\},\beta)$. Hence, $R(y,N)$ is decreasing in $(\inf\{y: r(y)\gamma_N^h(\rho(y))(y-\frac{h}{\mu}\varphi_N(\rho(y)))\geq 1\},\beta)$. In the same fashion, $R(y,N)$ is increasing in $(\alpha,\inf\{y: r(y)\gamma_N^h(\rho(y))(y-\frac{h}{\mu}\varphi_N(\rho(y)))\geq 1\})$. Therefore,

$$y_N^*=\inf\left\{y: r(y)\gamma_N^h(\rho(y))\left(y-\frac{h}{\mu}\varphi_N(\rho(y))\right)\geq 1\right\}.$$

□

Since the optimal price is not always easy to compute, the next result provides some bounds on y_N^*.

Proposition 3.2.4 *The unique optimal price y_N^* satisfies $\frac{h}{\mu}\leq y_N^*$. Moreover,*

$$y_N^*\leq \inf\left\{y: r(y)\Gamma_N^h\left(y-\frac{h}{\mu}\Phi_N\right)\geq 1\right\},$$

where

$$\Gamma_N^h=\begin{cases}\frac{1+N(\frac{\Lambda}{\mu})^{N+1}-(N+1)(\frac{\Lambda}{\mu})^N}{(1-(\frac{\Lambda}{\mu})^{N+1})(1-(\frac{\Lambda}{\mu})^N)} & \text{if } \frac{\Lambda}{\mu}\neq 1,\\ \frac{1}{2} & \text{if } \frac{\Lambda}{\mu}=1,\end{cases}$$

and

$$\Phi_N=\begin{cases}\frac{1-(N+1)^2(\frac{\Lambda}{\mu})^N(1+(\frac{\Lambda}{\mu})^2)+2N(N+2)(\frac{\Lambda}{\mu})^{N+1}+(\frac{\Lambda}{\mu})^{2N+2}}{(1-\frac{\Lambda}{\mu})^2(1-(N+1)(\frac{\Lambda}{\mu})^N+N(\frac{\Lambda}{\mu})^{N+1})} & \text{if } \frac{\Lambda}{\mu}\neq 1,\\ \frac{1}{6}N^2+\frac{1}{2}N+\frac{1}{3} & \text{if } \frac{\Lambda}{\mu}=1.\end{cases}$$

Proof First, suppose that $y_N^*<\frac{h}{\mu}$. Since $L(\rho(y),N)\geq\rho(y)$, $R(y_N^*,N)\leq y_N^*\lambda(y_N^*)-\frac{h}{\mu}\lambda(y_N^*)<0$. We proved in Theorem 3.2.2 that there exists y in $[\alpha,\beta)$ such that $R(y,N)>0$. Thus, y_N^* can not be optimal. Therefore, $\frac{h}{\mu}\leq y_N^*$.

Recall from Theorem 3.2.2 that

$$y_N^* = \inf\left\{y : r(y)\gamma_N^h(\rho(y))\left(y - \frac{h}{\mu}\varphi_N(\rho(y))\right) \geq 1\right\}.$$

Moreover, we proved in Theorem A.0.1 that $\varphi_N(\rho(y))$ is nonincreasing with respect to y. Thus,

$$\varphi_N(\rho(y)) \leq \varphi_N(\frac{\Lambda}{\mu}) = \Phi_N.$$

In the same fashion, since $\gamma_N^h(\rho(y))$ is nondecreasing with respect to y,

$$\Gamma_N^h = \gamma_N^h(\frac{\Lambda}{\mu}) \leq \gamma_N^h(\rho(y)).$$

Let y be in $[\alpha, \beta)$ such that $r(y)\Gamma_N^h(y - \frac{h}{\mu}\Phi_N) \geq 1$. Using the previous orderings, we have

$$r(y)\gamma_N^h(\rho(y))\left(y - \frac{h}{\mu}\varphi_N(\rho(y))\right) \geq r(y)\Gamma_N^h\left(y - \frac{h}{\mu}\Phi_N\right) \geq 1.$$

Thus, $y_N^* \leq \inf\{y : r(y)\Gamma_N^h(y - \frac{h}{\mu}\Phi_N) \geq 1\}$. □

As in the $M/G/1/\infty$ case, we now compare the optimal prices of two systems with different parameters. In the remainder of this section, parameters relative to system $i = 1, 2$ are indexed by i.

First, we study how the optimal price y_N^* changes as capacity N increases. From Proposition 4.2 in Ziya [24], we know that when $h = 0$, the optimal price is increasing (decreasing) with respect to the capacity when $\frac{\Lambda}{\mu} > (<)\rho^c$, where ρ^c is called the critical traffic intensity ($\rho^c = (1 - F(\inf\{y : yr(y) \geq 2\}))^{-1}$). However, when $h > 0$, this is not always the case. Let $F(y) = 1 - e^{-\beta y}$, with $\beta = 0.1$ and $\Lambda = 8$, $\mu = 2$ and $h = 1$. In this case, when capacity is 5,6 and 7, the optimal prices y_5^*, y_6^* and y_7^* are 16.4204, 16.4064, 16.4245, respectively. Hence, the optimal price is not monotone in capacity. However, the next result shows that optimal prices are ordered with respect to other system parameters.

Proposition 3.2.5 *Consider two systems 1 and 2 that satisfy all of the following :*

1. $F_1 \geq_{HR} F_2$,
2. $\Lambda_1 \geq \Lambda_2$,
3. $\mu_1 \leq \mu_2$,
4. $h_1 \geq h_2$.

Then, $y^*_{N,1} \geq y^*_{N,2}$.

Proof Suppose conditions 1 through 4 hold. We have $\rho_1(y) \geq \rho_2(y), \forall y$ in $[\alpha, \beta)$ and $\frac{h_1}{\mu_1} \geq \frac{h_2}{\mu_2}$. Moreover, as shown in Theorem 3.2.2, $\varphi_N(\cdot)$ is nondecreasing and $\gamma^h_N(\cdot)$ is nonincreasing. Therefore, $\varphi_N(\rho_1(y)) \geq \varphi_N(\rho_2(y))$ and $\gamma^h_N(\rho_2(y)) \geq \gamma^h_N(\rho_1(y))$ for all y in $[\alpha, \beta)$.

Let $y \in [\alpha, \beta)$ be such that $r_1(y)\gamma^h_N(\rho_1(y))(y - \frac{h_1}{\mu_1}\varphi_N(\rho_1(y))) \geq 1$. Using the properties shown above, we have

$$r_2(y)\gamma^h_N(\rho_2(y))\left(y - \frac{h_2}{\mu_2}\varphi_N(\rho_2(y))\right) \geq r_1(y)\gamma^h_N(\rho_1(y))\left(y - \frac{h_1}{\mu_1}\varphi_N(\rho_1(y))\right) \geq 1.$$

Hence, $y^*_{N,2} \leq y^*_{N,1}$. □

Similar to the infinite capacity case, Proposition 3.5.1 in Ziya [22] shows that in systems with no holding cost, $\frac{\Lambda_1}{\mu_1} \geq \frac{\Lambda_2}{\mu_2}$ implies that $y^*_{N,1} \geq y^*_{N,2}$. This result cannot be extended to systems with holding costs. To see this, consider two M/M/1/2 systems, 1 and 2, where $\mu_1 = 1, \mu_2 = .1, \Lambda_1 = 0.5, \Lambda_2 = 0.0001$. So, $\frac{\Lambda_2}{\mu_2} = 0.001 \leq \frac{\Lambda_1}{\mu_1} = 0.5$. For both systems, assume that the willingness-to-pay distribution is exponential with rate 1. Therefore, we can use Proposition 3.2.4, which states that the optimal solution for system $i = 1, 2$ satisfies

$$\frac{h}{\mu_i} \leq y^*_{N,i} \leq \frac{1}{\Gamma^h_{N,i}} + \frac{h}{\mu}\Phi_{N,i}.$$

Note that $\Phi_{N,1} = 1.625$ and $(\Gamma^h_{N,1})^{-1} = 1.3125$. Therefore, $y^*_{N,1} \leq 1.3125 + 1.625h$ and $10h \leq y^*_{N,2}$. When $h = 1$, $y^*_{\infty,1} < y^*_{\infty,2}$. We can claim that the arrival rate, service rate and hazard rate orderings that hold when $h = 0$ in the $M/M/1/N$ case still hold when $h > 0$. However, as in the infinite capacity case, the traffic intensity ordering without holding costs cannot be extended when $h > 0$.

The following result shows that the optimal rewards are also ordered as the system parameters change.

Proposition 3.2.6 *Consider two systems 1 and 2 that satisfy all of the following :*

1. $\Lambda_1 \geq \Lambda_2$,
2. $F_1 \geq_{ST} F_2$,
3. $h_1 \leq h_2$,
4. $\mu_1 \geq \mu_2$.

Then, $R^*_{N,1} \geq R^*_{N,2}$.

Proof To prove this result, we split our proof into two parts as we did in the $M/G/1/\infty$ case. First, we show that the result holds when conditions 1 and 2 are replaced by equalities. Second, we show that it holds when conditions 3 and 4 are equalities. By composition, the result holds under all conditions as well.

Suppose that conditions 3 and 4 hold and $\Lambda_1 = \Lambda_2$ and $F_1(\cdot) = F_2(\cdot)$. Recall that for all $y \geq 0$,

$$R_i(y, N) = y\lambda_i(y)(1 - \pi_N(\rho_i(y), N)) - h_i L(\rho_i(y), N).$$

Conditions 3 and 4 imply that $h_1 L(\rho_1(y), N) \leq h_2 L(\rho_2(y), N)$ and that $\pi_N(\rho_1(y), N) \leq \pi_N(\rho_2(y), N)$. Therefore, $R_1(y, N) \geq R_2(y, N)$ and $R^*_{N,1} \geq R^*_{N,2}$.

When conditions 1 and 2 hold but conditions 3 and 4 are equalities, the proof is similar to the proof of Proposition 3.2.3 and is omitted. □

The following theorem shows that the infinite capacity model can be approximated by a finite capacity model of large size provided that it is stable for all prices. We show that both the optimal reward and optimal price of a finite capacity model converge to those of an infinite capacity system as the system size grows to infinity.

Theorem 3.2.3 *Under the stability condition* $\Lambda < \mu$, $R_N^* \to R_\infty^*$ *and* $y_N^* \to y_\infty^*$ *as* $N \to \infty$.

We need the following lemma before proving Theorem 3.2.3.

Lemma 3.2.1 *Under the stability condition* $\Lambda < \mu$, $R(y, N) \to R(y, \infty)$ *uniformly in* y *as* N *converges to infinity.*

Proof Let $y \geq 0$. Consider

$$R(y, \infty) = \lambda(y)y - h\sum_{n=0}^{\infty} n\pi_n(\rho(y), \infty).$$

Now consider a system with capacity N with posted price y and reward corresponding to this price

$$R(y, N) = \lambda(y)y(1 - \pi_N(\rho(y), N)) - h\pi_0(\rho(y), N)\sum_{n=0}^{N} n\frac{\lambda(y)^n}{\mu^n}.$$

We observe that $\pi_N(\rho(y), N) = \rho(y)^N \frac{1-\rho(y)}{1-\rho(y)^{N+1}} \to 0$ uniformly in y as N tends to infinity. In the same fashion, $\pi_0(\rho(y), N) = \frac{1-\rho(y)}{1-\rho(y)^{N+1}} \to (1 - \rho(y))$ uniformly in y as N goes to ∞. Finally, $\sum_{n=N}^{\infty} n\frac{\lambda(y)^n}{\mu^n} \leq \sum_{n=N}^{\infty} n\frac{\Lambda^n}{\mu^n} \to 0$ uniformly in y. Therefore, $R(y, N) \to R(y, \infty)$ uniformly in y as N goes to infinity, which proves the desired result. □

Proof of Theorem 3.2.3 According to the previous lemma, $R(y, N) \to R(y, \infty)$ uniformly in y. Therefore, R_N^* converges to R_∞^* as N goes to infinity.

Let $\{y^*_{N(m)}\}$ be a converging (to y) subsequence of $\{y^*_N\}$. Before we proceed, we need to show that such a subsequence exists and that $y < \infty$. Since $y^*_N \geq 0$ for all N, it suffices to show that y^*_N does not converge to infinity as N tends to infinity. Suppose that $\lim y^*_N = \infty$. Note that $R^*_N = R(y^*_N, N) \leq \lambda(y^*_N)y^*_N$. Since $\lambda(y^*_N)y^*_N \to 0$ as N tends infinity, $R^*_N \to 0$. But this is a contradiction since we showed that $R^*_N \to R^*_\infty > 0$. Therefore, $\{y^*_{N(m)}\}$ exists and has a finite limit.

To simplify the notation in the remainder of the proof, we use N instead of $N(m)$. Next, we show that $|R(y,\infty) - R^*_N| \to 0$ as N goes to infinity. Since $R^*_N \to R^*_\infty$, this implies that $R(y,\infty) = R^*_N$. Therefore, $y = y^*_\infty$ since y^*_∞ is unique as shown in Theorem 3.2.1. We have

$$\pi_0(\rho(y^*_N), N)^{-1} - \pi_0(\rho(y),\infty)^{-1} = \sum_{n=1}^{N} \rho(y^*_N)^n - \sum_{n=1}^{\infty} \rho(y)^n.$$

Therefore, for an arbitrary integer M between 1 and N,

$$\pi_0(\rho(y^*_N), N)^{-1} - \pi_0(\rho(y),\infty)^{-1} = \sum_{n=1}^{M} \rho(y^*_N)^n - \rho(y)^n + \sum_{n=M}^{N} \rho(y^*_N)^n - \sum_{n=M}^{\infty} \rho(y)^n.$$

Thus,

$$|\pi_0(\rho(y^*_N), N)^{-1} - \pi_0(\rho(y),\infty)^{-1}| \leq \sum_{n=1}^{M} |\rho(y^*_N)^n - \rho(y)^n| + 2\sum_{n=M}^{\infty} \frac{\Lambda^n}{\mu^n}.$$

First, let N go to infinity and then let M go to infinity. Consequently, $\pi_0(\rho(y^*_N), N) \to \pi_0(\rho(y),\infty)$.

We also have

$$\begin{aligned} R(y,\infty) - R^*_N =& \mu y(1 - \pi_0(\rho(y),\infty)) - \mu y^*_N(1 - \pi_0(\rho(y^*_N), N)) \\ & - h\pi_0(\rho(y),\infty)\sum_{n=1}^{\infty} n\rho(y)^n + h\pi_0(\rho(y^*_N), N)\sum_{n=1}^{N} n\rho(y^*_N)^n. \end{aligned}$$

So, for an arbitrary integer M between 1 and N,

$$\begin{aligned}|R(y,\infty) - R_N^*| \leq & \mu|y(1-\pi_0(\rho(y),\infty)) - y_N^*(1-\pi_0(\rho(y_N^*),N))| \\ & + h\sum_{n=1}^{M} n|\pi_0(\rho(y_N^*),N)\rho(y_N^*)^n - \pi_0(\rho(y),\infty)\rho(y)^n| \\ & + h\pi_0(\rho(y),\infty)\sum_{n=M}^{\infty} n\rho(y)^n + h\pi_0(\rho(y_N^*),N)\sum_{n=M}^{N} n\rho(y_N^*)^n.\end{aligned}$$

Then,

$$\begin{aligned}|R(y,\infty) - R_N^*| \leq & \mu|y(1-\pi_0(\rho(y),\infty)) - y_N^*(1-\pi_0(\rho(y_N^*),N))| \\ & + h\sum_{n=1}^{M} n|\pi_0(\rho(y_N^*),N)\rho(y_N^*)^n - \pi_0(\rho(y),\infty)\rho(y)^n| + 2h\sum_{n=M}^{\infty} n(\frac{\Lambda}{\mu})^n.\end{aligned}$$

First, let N go to infinity and then M go to infinity. We have $R_N^* \to R(y,\infty)$ for any converging subsequence. Therefore, $R(y,\infty) = R_\infty^*$, so $y = y_\infty^*$ is optimal for the infinite capacity system. Since the limit is unique, any converging subsequence y_N^* has the limit y_∞^*. Hence, $y = y_\infty^* = \lim_{N\to\infty} y_N^*$. □

3.2.3 Optimal Capacity in $M/M/1/N$ Queues

We showed that the infinite capacity model can be approximated by a finite capacity model of large size under the condition $\Lambda < \mu$. A natural question that stems from this result is whether there is a capacity level that maximizes the reward. Indeed, in our analysis so far, capacity is a given parameter. Now, we relax this constraint by allowing the service provider to set the capacity of the service facility in addition to the price. Note that the chosen capacity could be finite or infinite. Ziya [24] shows that systems with larger capacities always perform better when there is no holding cost. In this case, the service provider should have an infinite capacity system in order to maximize revenue. Thus, no customer is ever turned down due to capacity limitations. However, when $h > 0$, there is

a trade-off between large capacity and high holding costs. In the following, we show the existence of a capacity level $N^* < \infty$ that maximizes revenue when $\Lambda < \mu$ and $h > 0$.

Proposition 3.2.7 *If $\Lambda < \mu$ and $h > 0$, then there exists a capacity level $N^* < \infty$, such that $R^*_{N^*} = \sup_N R^*_N$. Consequently, there exists an optimal solution to $\sup_{y,N} R(y, N)$.*

We need the following lemma in order to prove Proposition 3.2.7.

Lemma 3.2.2 *If $\Lambda < \mu$ and $h > 0$, then, $\forall B \geq 0$, there exists $\overline{N} \in \mathbb{N}$ such that for all $y \leq B$, $R(y, N)$ is nonincreasing for $N \geq \overline{N}$.*

Proof Let $B \geq 0$. Instead of N being restricted to integer values, let N attain real values. Note that $R(y, N)$ is differentiable with respect to N. We will show that for all $y \leq B$ and N large enough, $\frac{\partial R(y,N)}{\partial N} \leq 0$. With some algebra,

$$\begin{aligned}
\frac{\partial R(y,N)}{\partial N} =& \frac{\rho(y)^{N+1}}{(1-\rho(y)^{N+1})^2}(-\mu y(1-\rho(y))\ln(\rho(y)) + h(1-\rho(y)^{N+1} + \ln(\rho(y)^{N+1}))) \\
\leq& \frac{\rho(y)^{N+1}}{(1-\rho(y)^{N+1})^2}\left(-\mu B \ln(\rho(B)) + h\left(1 + (N+1)\ln(\frac{\Lambda}{\mu})\right)\right) \\
\leq& 0, \text{ if } N \geq -\frac{\frac{\mu B}{h}\ln(\rho(B)) + 1}{\ln(\frac{\Lambda}{\mu})}.
\end{aligned}$$

Therefore, there exists $\overline{N} \in \mathbb{N}$ such that for all $y \leq B$ and $h > 0$, $R_N(y)$ is nonincreasing in N for $N \geq \overline{N}$. □

Proof of Proposition 3.2.7 We showed in Theorem 3.2.3 that y^*_N converges to y^*_∞. Therefore, let $\overline{y} = \sup_N\{y^*_N\} < \infty$. We use Lemma 3.2.2 to define

$$\overline{N} = 1 + \max\{N : \exists y \leq \overline{y}, R(y, N+1) > R(y, N)\}.$$

For $y \leq \overline{y}$ and $N \geq \overline{N}$, we have $R(y, N+1) \leq R(y, N)$. So, for $N \geq \overline{N}$,

$$R^*_{N+1} = \sup_{y \leq \overline{y}} R(y, N+1) \leq \sup_{y \leq \overline{y}} R(y, N) = R^*_N.$$

Therefore, R_N^* is nonincreasing for $N \geq \overline{N}$, which implies that $R_{N^*}^* = \sup_N R_N^*$ exists. $\square$

3.3 Systems with Balking Customers

In this section, we capture congestion penalties through balking customers.

3.3.1 Optimal Pricing in $M/M/1$ Queues

The model for the arrival, service and willingness-to-pay processes is the same as in the holding cost model with $h = 0$; that is, no holding cost is incurred. However, we suppose that potential customers make their decision to enter the system (if it is not full) based both on price and congestion. Therefore, we assume that each customer not only has a random *willingness-to-pay* but also a random *willingness-to-wait*. A customer's willingness-to-wait is the maximum current occupancy of the system so that the customer is willing to enter the facility. The customers' willingness-to-wait process forms a collection of independent, identically distributed discrete random variables. Hence, when $s < N$ customers are in system, an arriving customer who is willing to pay the advertised price accepts to join the system with probability p_s. We assume that $\{p_s : s \geq 0\}$ is nonincreasing in s, since customers are more likely to be deterred by high congestion levels. Without loss of generality, we suppose that $p_0 = 1$. To ensure the stability of the system when $N = \infty$, we suppose that $\Lambda \lim_s p_s < \mu$. To ease the notation in the following, we define $P_n = \prod_{s=0}^{n} p_s$, for $s = 1, \ldots, N-1$ and $P_{-1} = 1$.

As opposed to the holding cost model, we need not break down our work into the analysis of finite capacity queues and infinite capacity queues. We generally assume that $N \leq \infty$. However, we assume that the service times are Markovian with rate μ.

Hence, the customer admission process under static price y is a conditional (doubly stochastic) Poisson process with rate $1_{X(t)<N} p_{X(t)} \lambda(y)$. In the same fashion, the departure

process is a conditional Poisson process with rate $\mu 1_{X(t)>0}$. Under price y, the queueing system behaves as a Markovian birth-death process with birth rates $p_{X(t)}\lambda(y)$ and death rate μ as described in Figure 2.

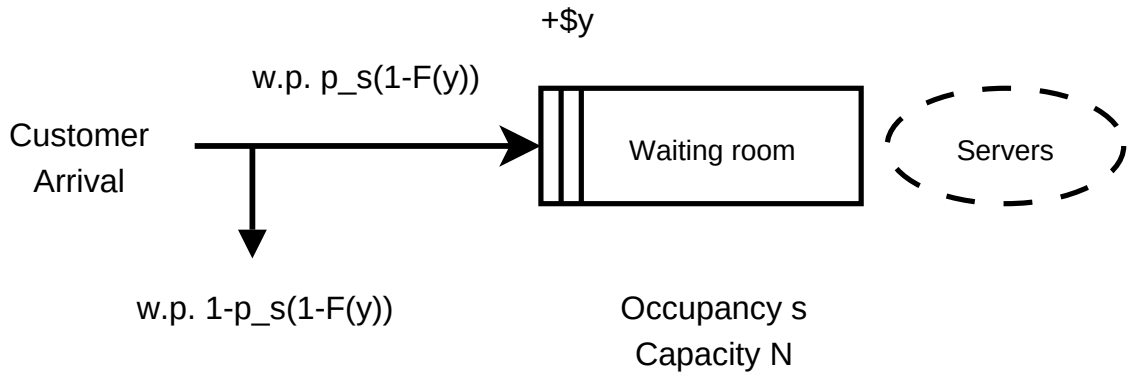

Figure 2: System with Balking Customers

Using the birth-death properties of the queueing process, the stationary distribution for price y is:

$$\pi_n(\rho(y), N) = \frac{\rho(y)^n P_{n-1}}{\sum_{n=0}^{N} \rho(y)^n P_{n-1}}, n = 0, \ldots, N.$$

The long-run average reward for price y can be expressed as

$$\begin{aligned} R(y, N) =& y\lambda(y) \sum_{n=0}^{N-1} p_n \pi_n(\rho(y), N) \\ =& y\lambda(y) \frac{\sum_{n=0}^{N-1} \rho(y)^n P_n}{\sum_{n=0}^{N} \rho(y)^n P_{n-1}}. \end{aligned}$$

Note that the reward differs from state to state as customers now react to both congestion and prices. Since the system is stable, all entering (paying) customers eventually get serviced and depart the system after a finite waiting time. Hence, we can also express the average

reward as

$$\begin{aligned} R(y,N) =& y\mu(1-\pi_0(\rho(y),N)) \\ =& y\mu\frac{\sum_{n=1}^{N}\rho(y)^n P_{n-1}}{\sum_{n=0}^{N}\rho(y)^n P_{n-1}}. \end{aligned}$$

In the following theorem, we show that exists a unique optimal price that maximizes the long-run average reward. We explicitly characterize this price in a similar fashion as in Theorems 3.2.1 and 3.2.2.

Theorem 3.3.1 *There exists a unique optimal price given by :*

$$y_N^* = \inf\{y : yr(y)\gamma_N^b(\rho(y)) \geq 1\},$$

where

$$\gamma_N^b(\rho) = \frac{\sum_{n=0}^{N-1}(n+1)\rho^n P_n}{\sum_{n=0}^{N-1}\rho^n P_n \sum_{n=0}^{N}\rho^n P_{n-1}},\ \rho \geq 0.$$

Proof The proof is similar to the proof of Theorem 3.2.2. First, we prove that there exists an optimal price. Since $F(\cdot)$ is absolutely continuous, $R(y,N)$ is continuous and a.e. differentiable on $[\alpha,\beta)$. Note that $R(y,N) \to 0$ as $y \to \beta$ and that $R(y,N) > 0$ in $[\alpha,\beta)$. Therefore, there exists an optimal price in $[\alpha,\beta)$.

Next, we prove the uniqueness of the optimal price. After some algebra, we show that for almost all y in $[\alpha,\beta)$, $\frac{\partial R(y,N)}{\partial y}$ satisfies:

$$\frac{\partial R(y,N)}{\partial y} = \lambda(y)\left(1-yr(y)\gamma_N^b(\rho(y))\right)\frac{\sum_{n=0}^{N-1}\rho(y)^n P_n \sum_{n=0}^{N}\rho(y)^n P_{n-1}}{\sum_{n=0}^{N}\rho(y)^n P_{n-1}^2}.$$

Note that, the above holds when $N=\infty$ as we interchange derivative and summation signs for power series of ρ within the radius of convergence.

If $R(y,N)$ is differentiable with respect to $y \in [\alpha,\beta)$, then $\frac{\partial R(y,N)}{\partial y} > 0 (< 0)$ if and only if $yr(y)\gamma_N^b(\rho(y)) < 1 (> 1)$.

Since there exists y in $[\alpha, \beta)$ such that $R(y, N) > 0$ and $R(y, N) \to 0$ as $y \to \beta$, there exists y in $[\alpha, \beta)$ such that $yr(y)\gamma_N^b(\rho(y)) > 1$. Under Assumption IGHR and using Lemma A.0.2, $yr(y)\gamma_N^b(\rho(y))$ is strictly increasing in $[\alpha, \beta)$, so $R(y, N)$ is decreasing in the interval $(\inf\{y : yr(y)\gamma_N^b(\rho(y)) \geq 1\}, \beta)$. In the same fashion, $R(y, N)$ is increasing in the interval $(\alpha, \inf\{y : yr(y)\gamma_N^b(\rho(y)) \geq 1\})$. Therefore, there exists a unique optimal price given by $y_N^* = \inf\{y : yr(y)\gamma_N^b(\rho(y)) \geq 1\}$. □

3.3.2 Properties of Optimal Prices and Optimal Profit

In the following, we investigate the sensitivity of optimal prices and optimal profits as parameters change. We compare two systems 1 and 2 that are identical except for some parameter whose ordering is known. Note that subscripts 1 and 2 refer to system 1 and 2, respectively.

Proposition 3.3.1 *Consider two systems 1 and 2 that satisfy all the following :*

1. $\frac{\Lambda_1}{\mu_1} \geq \frac{\Lambda_2}{\mu_2}$,
2. $F_1 \geq_{HR} F_2$.

Then, $y_{N,1}^* \geq y_{N,2}^*$.

Proof From Theorem 3.3.1, we have $y_{N,i}^* = \inf\{y : yr_i(y)\gamma_N^b(\rho_i(y)) \geq 1\}$ for system $i = 1, 2$. Since hazard rate ordering implies stochastic ordering, $F_1 \geq_{ST} F_2$. Using condition 1, this implies that $\rho_1(\cdot) \geq \rho_2(\cdot)$. Moreover, we showed in Lemma A.0.2 that $\gamma_N^b(\cdot)$ is nonincreasing. Therefore, we have $\gamma_N^b(\rho_1(\cdot)) \leq \gamma_N^b(\rho_2(\cdot))$. Suppose that y is such that $yr_1(y)\gamma_N^b(\rho_1(y)) \geq 1$. Then, $yr_2(y)\gamma_N^b(\rho_2(y)) \geq 1$. Thus, $y_{N,2}^* \leq y_{N,1}^*$. □

When congestion penalties are captured through balking customers, we note that optimal prices are ordered with the maximum traffic intensity $\frac{\Lambda}{\mu}$, whereas it does not hold in

the model with holding costs. We are also interested in the parameter sensitivity of the optimal reward. For instance, when Λ increases, more customers are admitted into the system. This generates more revenue on the one hand but also increases congestion and deters future potential customers on the other hand. In the next proposition, we show how the optimal reward varies as some parameters are increased or decreased.

Proposition 3.3.2 *Consider two systems 1 and 2 that satisfy all of the following :*

1. $\Lambda_1 \geq \Lambda_2$,
2. $F_1 \geq_{ST} F_2$,
3. $\mu_1 \geq \mu_2$.

Then, $R^*_{N,1} \geq R^*_{N,2}$.

Proof To prove this result, we split our proof into two parts. First, we show that the result holds when conditions 1 and 2 are changed to equalities. Second, we show that it holds when condition 3 is changed into an equality. By composition, the result holds under all the conditions as well.

Suppose that condition 3 holds and $\Lambda_1 = \Lambda_2$ and $F_1(\cdot) = F_2(\cdot)$. We will show that the generic expression $R(y, N) = y\mu(1 - \pi_0(\rho(y), N))$ is nondecreasing with respect to μ. We have

$$\frac{\partial R(y,N)}{\partial \mu} = y(1 - \pi_0(\rho(y), N)) \left(1 - \gamma^b_N(\rho(y))\right).$$

From Lemma A.0.2, $\gamma^b_N(\cdot)$ is nonincreasing, so $\gamma^b_N(\rho(y)) \leq \gamma^b_N(0) = 1$ for all y in $[\alpha, \beta)$. Therefore, $R(y, N)$ is nondecreasing in μ and $R^*_{N,1} \geq R^*_{N,2}$.

Now suppose that 1 and 2 hold, whereas 3 is an equality. Since $F_1(\cdot)$ is absolutely continuous and $\lambda_1(\cdot) \geq \lambda_2(\cdot)$, there exists $\delta > 0$ such that $\lambda_1(y^*_{N,2}+\delta) = \lambda_2(y^*_{N,2})$. Therefore,

system 1 with price $y^*_{N,2} + \delta$ has the same arrival and service rates as system 2 with price $y^*_{N,2}$. Therefore, system 1 with price $y^*_{N,2} + \delta$ performs better than system 2 with optimal price. Hence, $R^*_{N,1} \geq R^*_{N,2}$ and the proof is complete. □

In the holding cost model, we showed that system capacity was a critical parameter that could be adjusted to improve optimal profits. We now focus on the effect of the system capacity in the balking customer model. First, we show that optimal prices are not ordered in the system capacity N. We provide the following counterexample. Consider a system with $\Lambda = 30, \mu = 3, p_s = \frac{3}{3+.2s}$ for $s = 0, \ldots, N-1$. We suppose that customers have an exponentially distributed willingness-to-pay ($F(y) = 1 - e^{-y}$). In this case, we have $y^*_3 = 2.1964, y^*_4 = 2.1983$ and $y^*_5 = 2.1960$. Clearly, y^*_N is not monotone in N. However, we observe that R^*_N is monotone in N. We prove this claim in the next proposition.

Proposition 3.3.3 *Under the stability condition* $\Lambda \limsup_s p_s < \mu$, $R^*_N \uparrow R^*_\infty$ *and* $y^*_N \to y^*_\infty$ *as* $N \to \infty$,

Proof First, we show that for all $y \geq 0$, $R(y, N) \uparrow R(y, \infty)$ uniformly in y as N converges to infinity in order to show that $R^*_N \uparrow R^*_\infty$. Recall that for $K = N$ or $K = \infty$,

$$\begin{aligned} R(y, K) =& \mu y(1 - \pi_0(\rho(y), K)), \\ \pi_0(\rho(y), K)^{-1} =& \sum_{n=0}^{K} P_{n-1}\rho(y)^n. \end{aligned}$$

Hence, we only need to show that $\pi_0(\rho(y), N) \downarrow \pi_0(\rho(y), \infty)$ uniformly in y. Clearly, $\pi_0(\rho(y), N)$ is nonincreasing in N. We also have

$$0 \leq \pi_0(\rho(y), \infty)^{-1} - \pi_0(\rho(y), N)^{-1} = \sum_{n=N+1}^{\infty} P_{n-1}\rho^n(y) \leq \sum_{n=N+1}^{\infty} P_{n-1}\left(\frac{\Lambda}{\mu}\right)^n.$$

Since $\Lambda \limsup_s p_s < \mu$, $\pi_0(\rho(y), N) \downarrow \pi_0(\rho(y), \infty)$ uniformly in y implying that $R(y, N) \uparrow R(y, \infty)$ uniformly in y as well. Therefore, $R^*_N \uparrow R^*_\infty$.

To prove that $y_N^* \to y_\infty^*$, we use the same method as in the proof of Theorem 3.2.3. We repeat that proof, setting $h = 0$ and substituting $\{P_{n-1}\rho(y)^n\}$ for $\{\rho(y)^n\}$, $\{P_{n-1}\rho(y_N^*)^n\}$ for $\{\rho(y_N^*)^n\}$ and $\{P_{n-1}\frac{\Lambda^n}{\mu^n}\}$ for $\{\frac{\Lambda^n}{\mu^n}\}$. □

Therefore, the model with balking customers performs best when $N = \infty$ with no customer ever being turned down due to capacity limitations.

3.4 Systems with Impatient Customers

In this section, we model congestion penalties through impatient customers. As opposed the balking customer model, customers do not react to congestion upon arrival but during their waiting time. If a customer waits for too long before receiving service, the customer departs the system and receives a full refund.

3.4.1 Optimal Pricing in $M/M/1$ Queues

We assume that the arrival, service and willingness-to-pay processes are the same as described in the holding cost model with $h = 0$. Customers enter the system if it is not full and if they are willing to pay the price posted by the service provider. Payments are collected upon arrival. We suppose that each customer has a maximum waiting time in the queue that is exponentially distributed with rate θ . We will refer to θ as the *reneging rate*. We assume that the maximum waiting times for successive customers forms a collection of independent, identically distributed random variables that are independent of the arrival process and the service process. If a customer does not begin service prior to its maximum waiting time, the customer leaves (reneges) the system after receiving a full refund of the amount paid upon arrival. Note that customers in service are no longer subject to impatience. Congestion penalties are experienced the following way: as the system becomes congested, waiting times increase and customers are more likely to become impatient, leave

and claim a refund. Therefore, the service provider sustains losses when refunding impatient customers.

The model's features are represented in Figure 3.

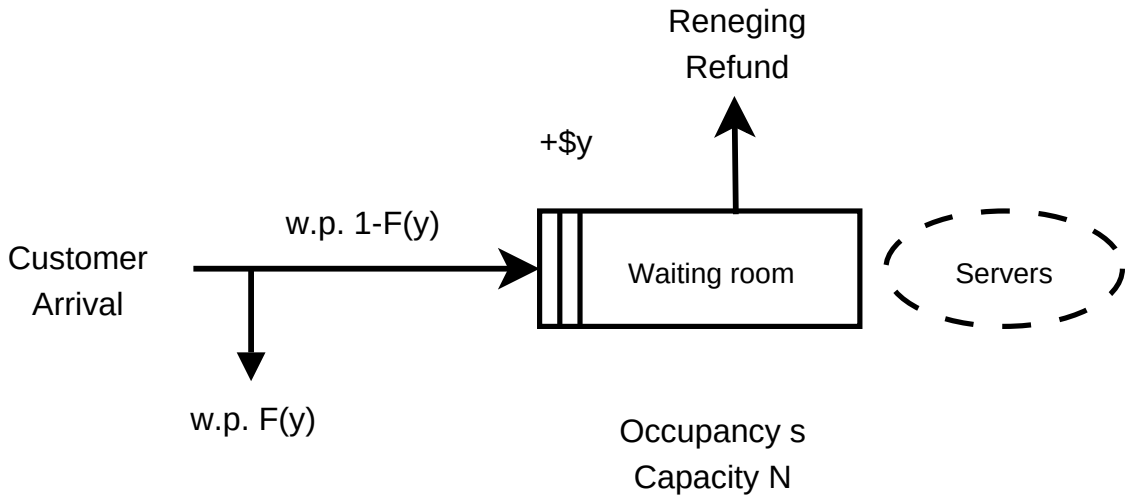

Figure 3: System with Impatient Customers

As in the previous section, we consider systems of finite or infinite capacity simultaneously ($N \leq \infty$) and exponentially distributed service times with rate μ. In the case where $N = \infty$, the stability of the system is ensured by impatient customers: all customers spend a finite expected time in the system.

Since the system is stable, only customers who depart the system after receiving service contribute to the long-run average profit. Impatient customers do not contribute to the long-run average profit although they make the system more congested. Note that we would observe the same long-run average profit, if payments were collected after each service completion instead of upon arrival.

In this chapter, we assume that the service provider can only advertise one price y at all

times. As in the holding cost model, the admission process is a conditional Poisson Process with rate $1_{X(t)<N}\lambda(y)$. However, the departure process differs significantly from the holding cost model. Customers may leave the system through service or due to impatience. Thus, the departure process (impatient departures and service departures) forms a conditional Poisson Process with rate $1_{X(t)>0}(\mu + (X(t) - 1)\theta)$. The resulting queueing process is a Markovian birth-death process with birth rates $\lambda(y)$ and death rates $\mu + (X(t) - 1)\theta$.

From Gross and Harris (p. 93) [5], the stationary distribution of the system is given by:

$$\pi_n(\rho(y), N) = \frac{\rho^n(y)Q_{n-1}}{\sum_{n=0}^{N} \rho^n(y)Q_{n-1}}, \text{ for } n = 0, \ldots, N$$

where $Q_n = \prod_{s=0}^{n} \frac{\mu}{\mu+s\theta}$, for $n = 0, \ldots, N-1$ and $Q_{-1} = 1$.

When price y is advertised, we can express the long-run average reward in two equivalent ways:

$$R(y, N) = \lambda(y)y \sum_{n=0}^{N-1} \pi_n(\rho(y), N) - \theta y \sum_{n=1}^{N} (s-1)\pi_n(\rho(y), N), \tag{1}$$

$$= \mu y(1 - \pi_0(\rho(y), N)) \tag{2}$$

In equation (1), we break down the long-run average profit into the revenue from customers' payments upon arrival and the refunds granted to impatient customers. Since the system is stable and all customers spend a finite amount of time in the system, only customers departing after service eventually contribute to the long-run average profit. Equation (2) uses this property and expresses the long-run average profit through the payments of serviced customers. Note that static pricing is key to deriving equation (2) since we need not know the payment history of all customers to compute the long-run average profit. We discuss this topic in further details in Section 4.2.3.

In the following theorem, we show that there exists a unique optimal price that maximizes the long-run average reward. We provide an explicit expression for the optimal price

to be charged.

Theorem 3.4.1 *There exists a unique optimal price given by :*

$$y_N^* = \inf\{y : yr(y)\gamma_N^r(\rho(y)) \geq 1\},$$

where

$$\gamma_N^r(\rho) = \frac{\sum_{n=0}^{N-1}(n+1)\rho^n Q_n}{\sum_{n=0}^{N-1}\rho^n Q_n \sum_{n=0}^{N}\rho^n Q_{n-1}},\ \rho \geq 0.$$

To prove this theorem, we notice analytical similarities with the model with balking customers. If the sequence $\{p_s\}$ is defined as $p_s = \frac{\mu}{\mu+s\theta}$, the maximization problem with balking customers becomes analytically equivalent to the one with impatient customers. Therefore, both models yield identical optimal prices and profits in the particular case when $p_s = \frac{\mu}{\mu+s\theta}, s = 0, \ldots, N-1$. We use this feature in the proof of Lemma A.0.3.

Proof of Theorem 3.4.1 First, we show in Lemma A.0.3 that $\gamma_N^r(\rho(y))$ is nonincreasing. Substituting $\gamma^r(\cdot)$ for $\gamma^b(\cdot)$, the remainder of the proof is identical to the proof of Theorem 3.3.1 and is omitted. □

We managed to show the existence of a unique optimal price that maximizes the long-run average reward per unit time. We noted that the optimization problem with impatient customers can be analytically treated as a particular case of the balking customers model with $p_s = \frac{\mu}{\mu+s\theta}, s = 0, \ldots, N-1$. This property becomes handy as we can refer to Section 3.3.2 when proving the next results.

3.4.2 Properties of Optimal Prices and Optimal Profit

We now investigate the sensitivity of optimal prices and optimal profits to system parameters. As previously, our objective is to show how optimal prices changes when system parameters increase or decrease. In the next two results, we compare two systems 1 and

2 that are identical except for some parameters that we specify. Parameters for system $i = 1, 2$ bear the additional subscript i.

Proposition 3.4.1 *Consider two systems 1 and 2 that satisfy all the following :*

1. $\Lambda_1 \geq \Lambda_2$,

2. $F_1 \geq_{HR} F_2$.

Then, $y^*_{N,1} \geq y^*_{N,2}$.

Proof From Lemma A.0.3, recall that $\gamma^r_N(\cdot)$ is nonincreasing. The rest of the proof is identical to the proof of Proposition 3.3.1 with $\mu_1 = \mu_2$, after subsituting $\gamma^r_N(\cdot)$ for $\gamma^b_N(\cdot)$.

□

As opposed to the balking customers model, there is no optimal price ordering with respect to $\frac{\Lambda}{\mu}$. We provide the following counterexample. Consider two systems 1 and 2 with $\Lambda_1 = 10, \Lambda_2 = 3, \mu_1 = 12, \mu_2 = 4$ and $\theta_1 = \theta_2 = 10$. Clearly, $\frac{\Lambda_1}{\mu_1} \geq \frac{\Lambda_2}{\mu_2}$. The optimal price to be charged when $N = 4$ is $y^*_{4,1} = 1.1362$ and $y^*_{4,2} = 1.1702$ for system 1 and 2, respectively. Note that $y^*_{4,1} < y^*_{4,2}$. We conclude that there is no price ordering in $\frac{\Lambda}{\mu}$.

Next, we focus on how the optimal long-run average profit varies when parameters change. For instance, it is intuitive that the higher reneging rate or the service rate, the more profitable the system. We formally prove this intuition in the next proposition.

Proposition 3.4.2 *Consider two systems 1 and 2 that satisfy all the following :*

1. $\Lambda_1 \geq \Lambda_2$,

2. $F_1 \geq_{ST} F_2$,

3. $\theta_1 \leq \theta_2$,

4. $\mu_1 \geq \mu_2$.

Then, $R^*_{N,1} \geq R^*_{N,2}$.

Proof We proceed by composition. We will show that the result holds when conditions $1, 2$ and 3 hold, while $\mu_1 = \mu_2$. Then, we show that the result holds when condition 4 holds and conditions 1,2 and 3 are modified into equalities. By composition, $R^*_{N,1} \geq R^*_{N,2}$ must then hold under all conditions.

Suppose that conditions $1, 2$ and 3 hold, while $\mu_1 = \mu_2 = \mu$. This clearly implies that $\pi_{0,1}(\rho_1(y), N)) \leq \pi_{0,2}(\rho_2(y), N))$. Recall that $R_i(y, N) = \mu y(1 - \pi_{0,i}(\rho_i(y), N))$. Therefore, $R_1(y, N) \leq R_2(y, N)$ and $R^*_{N,1} \geq R^*_{N,2}$. Now suppose that $\Lambda_1 = \Lambda_2 = \Lambda, F_1(\cdot) = F_2(\cdot) = F(\cdot)$ and $\theta_1 = \theta_2 = \theta$ while $\mu_1 \geq \mu_2$. Recall from (1) that for $y \geq 0$ and $i = 1, 2$,

$$\begin{aligned} R_i(y, N) &= \lambda(y)y \sum_{n=0}^{N-1} \pi_n(\rho_i(y), N) - \theta y \sum_{n=1}^{N} (s-1)\pi_n(\rho(y), N), \\ &= \lambda(y)y(1 - \pi_N(\rho_i(y), N)) - \theta y L^q(\rho_i(y), N), \end{aligned}$$

where $L^q(\rho_i(y), N)$ is the average queue size under traffic intensity $\rho_i(y)$. To make our notation consistent in the case when $N = \infty$, we suppose that $\pi_\infty(\cdot, \infty) = 0$. Since $\mu_1 \geq \mu_2$, we have $\rho_1(\cdot) \leq \rho_2(\cdot)$. Consequently, $\pi_N(\rho_1(y), N) \leq \pi_N(\rho_2(y), N)$ and $L^q(\rho_1(y), N) \leq L^q(\rho_2(y), N)$. Therefore, $R_1(y, N) \leq R_2(y, N)$ and $R^*_{N,1} \geq R^*_{N,2}$. By composition, the proof is complete. □

We now investigate how the optimal price and reward react to a change in system capacity. Not surprisingly, the results are similar to those of the balking customer model. First, we show that optimal prices are not ordered in the system capacity N. We provide the following counterexample inspired by the one given in Section 3.3.2. Consider a system with $\Lambda = 30, \mu = 3, \theta = .2$ and an exponentially distributed willingness-to-pay ($F(y) = 1 - e^{-y}$).

In this case, we have $y_3^* = 2.1964, y_4^* = 2.1983$ and $y_5^* = 2.1960$. Clearly, y_N^* is not monotone in N. However, we note that $R(y, N)$ is nondecreasing in N. This property is derived from (2) as $\pi_0(\rho(y), N)$ is nonincreasing in N. Hence, R_N^* is nondecreasing in N. We refine this result in the following proposition. We show that the optimal price and reward of finite capacity system of large size converges to those of an infinite capacity system. The proof of this proposition is identical to the proof of Proposition 3.3.3 using $\{Q_n\}$ in lieu of $\{P_n\}$ and is omitted.

Proposition 3.4.3 *As* $N \to \infty$, $R_N^* \uparrow R_\infty^*$ *and* $y_N^* \to y_\infty^*$.

Thus, systems with infinite capacity perform better that systems with finite capacity. Moreover, the optimal price and reward of a finite capacity system converge to those of an infinite capacity system when N grows to ∞.

3.5 Summary

In this chapter, we studied the optimal static pricing problem in a service facility modelled as single server queueing system subject to congestion penalties. Our objective has been to maximize the service provider's long-run average profit per unit time when only one price can be advertised at all times. Depending on whether there is a limit on the number of customers in the system, we considered systems of finite or infinite capacity. We successively analyzed three different ways in which the system congestion affects profit: holding costs, balking customers and impatient customers.

For each of the three congestion models, we showed the existence of unique optimal price to be charged. We derived expressions for the optimal price in each case. We also investigated how optimal prices and rewards vary as system parameters change. More specifically, we studied the influence of the system capacity N on optimal prices and rewards.

Although we noticed that optimal prices are not monotone with respect to capacity N in any of the three congestion models, system capacity is a critical parameter in order to improve optimal rewards. We showed that for all three congestion models, the optimal prices and rewards of a system with finite capacity converge to those of an unlimited capacity system. Moreover, in the holding cost model, we showed that there exists a finite capacity level that maximizes profit. Therefore, if the service provider has control over N, it does not make economic sense to have unlimited room for waiting customers when holding costs are incurred. It turns out that unlimited capacity incurs higher holding costs that could be avoided by limiting the number of customers in the system. Nevertheless, this does not hold for balking or impatient customers. For these two models, it is best to have unlimited system capacity as there is no direct penalty for having customers waiting in the queue.

CHAPTER IV

DYNAMIC PRECISION PRICING

4.1 Model Description

In this chapter, the service provider can dynamically adjust prices. Similarly to Chapter 3, we model the service facility as queueing system with finite or infinite capacity N. However, the queueing models we consider in this chapter have more general features than in the static pricing case. As in Chapter 3, we consider each of the three types of congestion penalties separately in three different models. This enables us to identify key properties and features of our solutions that are specific to the way congestion penalties are modelled.

In this chapter, we refine our framework by segmenting customers into I classes. Customers from class $i = 1, \ldots, I$ arrive according to a Poisson process with parameter $\Lambda_i > 0$. The arrival processes from customer classes are independent of each other. Note that this formulation is equivalent to having arriving customers randomly assigned to a specific class independently of everything else. The service provider can identify customers' classes upon arrival and can advertise class-specific prices. This ability is referred to as *precision pricing*. In this dissertation, we only consider dynamic precision pricing since Ziya [22] investigates precision pricing in the static case.

The maximum amounts that successive class $i = 1, \ldots, I$ customers are willing to pay are independent, identically distributed random variables with distribution F_i. The amount a class $i = 1, \ldots, I$ customer is willing to pay is independent of the amount a class $j = 1, \ldots, I$ customer is willing to pay for $i \neq j$. For all $i = 1, \ldots, I$, we assume that the cumulative distribution function $F_i(\cdot)$ is absolutely continuous with density $f_i(\cdot)$, support

(α_i, β_i) and finite mean. Let $r_i(\cdot)$ denote the hazard rate function of $F_i(\cdot)$; that is, $r_i(z) = \frac{f_i(z)}{1-F_i(z)}$ for $\alpha_i < z < \beta_i$. In all the following, we assume that F_i has IGHR (Increasing Generalized Hazard Rate); that is, $zr_i(z)$ is strictly increasing for all z in $[\alpha_i, \beta_i]$. The service provider can advertise different prices to different classes. Without loss of generality, only prices in $[\alpha_i, \beta_i]$ can be advertised to class-i customers.

We define the state of the system $X(t)$ as the number of customers in the system at time t. Let $\mathbf{z} \in [\alpha_1, \beta_1]^N \times \ldots \times [\alpha_I, \beta_I]^N$ be a pricing (decision) rule, where price $z_{i,s}$ is advertised to class-i customers when the system is in state s. Since there is a one-to-one relationship between decision rules and stationary policies, in an abuse of notation, we also denote by $\mathbf{z}$ the stationary pricing policy corresponding to the pricing rule $\mathbf{z}$; that is, $\mathbf{z}$ also denotes the policy of using pricing rule $\mathbf{z}$ at every decision epoch (see p. 20 of Puterman [18] for further details). Customers enter the system if it is not full and if they are willing to pay the price posted by the service provider upon arrival. In systems with balking customers, customers' decision to enter the system is also subject to the customers' willingness-to-wait as described in Section 3.3; the willingness-to-wait distribution is assumed to be same across customer classes. Hence, the admission process of customers under the stationary pricing policy $\mathbf{z}$ is a conditional (doubly stochastic) Poisson process with rate $1_{X(t)<N} \sum_{i=1}^{I} \lambda_i(z_{i,X(t)})$ (or $1_{X(t)<N} p_{X(t)} \sum_{i=1}^{I} \lambda_i(z_{i,X(t)})$ in systems with balking customers), where $\lambda_i(z) = \Lambda_i(1 - F_i(z))$. In the same fashion, the service process is a conditional Poisson process with rate $\mu_{X(t)} 1_{X(t)>0}$. Unless otherwise stated, $\{\mu_s\}$ are positive real numbers that are nondecreasing in s. Hence, the queueing system behaves as a Markovian birth-death process.

We let g_N^* denote the optimal dynamic average profit per unit time under capacity N over the set of all history-dependent randomized policies (see p.35-36 in Puterman [18]). Under stationary pricing policy $\mathbf{z}$, we denote the objective function by $R(\mathbf{z})$ and the stationary

probability distribution by $\{\pi_s(\mathbf{z})\}$. If there exists a unique optimal stationary pricing policy that maximizes the long-run average profit per unit time, we denote it by $\mathbf{z}^*$; that is, $R(\mathbf{z}^*) = g_N^*$.

We separate systems with finite capacity from systems with infinite capacity in our work. Indeed, we make extra assumptions and we use results from finite capacity systems in order to analyze systems of infinite capacity. In the following section, we focus on service facilities with finite capacity.

4.2 Queueing Systems with Finite Capacity

First, we consider the case of systems with finite capacity ($N < \infty$).

4.2.1 Systems with Holding Costs

In this section, the service facility is subject to holding costs. The service provider must pay a holding cost h_s per unit time spent in state s, where $0 = h_0 \leq h_1 \leq \ldots \leq h_N$ as it becomes more expensive to accommodate a larger number of customers. We assume that $\frac{h_1}{\mu_1} < \max \beta_i$ so that we have an attainable positive reward. Note that this structure is more general than having each customer incur a holding cost per unit time spent in the system where $h_s = hs$.

We use a Markov decision process (MDP) formulation to exhibit an optimal stationary policy. Note that the MDP associated with our system behaves as a birth-death process, with positive death rates, where the decision maker only controls the arrival rates. Therefore, the MDP is unichain for all stationary policies. We set up the system of average-cost

optimality equations (ACOE) as detailed in Theorem 5.2.2 of Lasserre and Hernández-Lerma [11]:

$$
\begin{aligned}
&l(-1) = 0, \\
&l(s) = \sup_{z_0,\ldots,z_I} \left\{ \frac{\sum_{i=1}^{I} \lambda_i(z_i)(z_i + l(s+1)) + \mu_s l(s-1) - g - h_s}{\sum_{i=1}^{I} \lambda_i(z_i) + \mu_s} \right\}, \text{ if } 0 \le s \le N-1, \\
&l(N) = l(N-1) - \frac{g + h_N}{\mu_N},
\end{aligned}
$$

where g is the gain and $l(\cdot)$ is the bias vector. Since the value of μ_0 does not matter as long as it is positive, we will consider $\mu_0 = \mu_1$ without loss of generality. In this system, we are solving for g and $l(\cdot)$.

We can transform these equations into a simpler equivalent form by letting $G(-1) = 0$ and $G(s) = l(s) - l(s+1),$ for $s = 0, \ldots, N-1$. Then,

$$G(-1) = 0, \tag{3}$$

$$g + h_s - \mu_s G(s-1) = \sum_{i=1}^{I} \sup_z \{(z - G(s))\lambda_i(z)\}, \text{ if } s = 0, \ldots, N-1, \tag{4}$$

$$G(N-1) = \frac{g + h_N}{\mu_N}. \tag{5}$$

If a solution $(g, G(\cdot), \mathbf{z})$ to the ACOE system exists, we call it a *canonical triplet*, where $\mathbf{z}$ are prices that achieve the suprema in (4). Precisely, for $s = 0, \ldots, N-1$ and $i = 1, \ldots, I$, the component $z_{i,s}$ of $\mathbf{z}$ satisfies $z_{i,s} = \arg\sup\{(z - G(s))\lambda_i(z)\}$.

In the following theorem, we explicitly characterize a unique optimal stationary policy.

Theorem 4.2.1 *There exists a canonical triplet $(g, G(\cdot), \mathbf{z})$ for the ACOE system (3)-(5). Moreover, the optimal long-run average reward is $g_N^* = g$ and $\mathbf{z}^* = \mathbf{z}$ is a unique optimal stationary policy, where, for $s = 0, \ldots, N-1$ and $i = 1, \ldots, I$,*

$$z_{i,s}^* = \inf \{z : r_i(z)(z - G(s)) \ge 1\}.$$

Before proving this theorem, we need the following two lemmas. Let $G(s, g)$ be the solution of (4) and (5) for $g \geq 0$.

Lemma 4.2.1 *For all $s = -1, \ldots, N-1$, $G(s, \cdot)$ is nondecreasing and continuous. Moreover, there exists $g \geq 0$ such that $G(-1, g) = 0$.*

Proof Note that $G(N-1, g) = \frac{g+h_N}{\mu_N}$ is continuous and nondecreasing in g. Suppose that $G(s, g)$ is nondecreasing and continuous in g for some state s between 0 and $N-1$. As $\sup\{\lambda_i(z)(z - G(s, g))\}$ is the supremum of a bounded continuous function of z, we can claim that $\mu_s G(s-1, g) = g - \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(s, g))\}$ is continuous and nondecreasing in g. By induction, for all $s = -1, \ldots, N-1$, $G(s, \cdot)$ is nondecreasing and continuous.

To complete the proof, we will show that $G(-1, 0) \leq 0$ and that there exists $g_b > 0$ such that $G(-1, g_b) \geq 0$. Hence, by continuity, there exists $g \in [0, g_b]$ such that $G(-1, g) = 0$. We know that $-\mu_0 G(-1, 0) = \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(0, 0))\} \geq 0$. Therefore, $G(-1, 0) \leq 0$. Now consider $g_b = \sum_{i=1}^{I} \sup\{\lambda_i(z) z\}$. Note that $G(N-1, g_b) = \frac{g_b + h_N}{\mu_N} \geq 0$. Suppose that $G(s, g_b) \geq 0$ for some $s = 0, \ldots, N-1$, then

$$g_b + h_s - \mu_s G(s-1, g_b) = \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(s, g_b))\} \leq \sum_{i=1}^{I} \sup\{\lambda_i(z) z\}.$$

Therefore, $\mu_s G(s-1, g_b) \geq g_b + h_s - \sum_{i=1}^{I} \sup\{\lambda_i(z) z\} \geq 0$. By induction, $G(-1, g_b) \geq 0$.

□

Lemma 4.2.2 *Let $(g, G(\cdot), \mathbf{z})$ be a canonical triplet. Then, for all $s = -1, \ldots, N-1$, $0 \leq G(s) \leq \frac{g+h_{s+1}}{\mu_{s+1}}$.*

Proof For all $s = 0, \ldots, N-1$, $\sup\{(z - G(s))\lambda_i(z)\} \geq 0$. Therefore, we have $G(s-1) \leq \frac{g+h_s}{\mu_s}$ from equation (4). Using equations (3) and (5) as well, $G(s) \leq \frac{g+h_{s+1}}{\mu_{s+1}}$ for all $s = -1, \ldots, N-1$.

Now suppose that there exists $s = 0, \ldots, N-1$ such that $G(s) < 0$. Since $G(-1) \geq 0$, there exists s such that $G(s) < 0$ and $G(s-1) \geq 0$. Hence, $\mu_{s+1}G(s) < \mu_s G(s-1)$. But we have

$$\begin{aligned}
\sum_{i=1}^{I} \lambda_i(z_{i,s+1})G(s+1) &= \sum_{i=1}^{I} \lambda_i(z_{i,s+1})z_{i,s+1} - g - h_{s+1} + \mu_{s+1}G(s) \\
&= \sum_{i=1}^{I} \lambda_i(z_{i,s+1})(z_{i,s+1} - G(s)) + \lambda_i(z_{i,s+1})G(s) + \mu_{s+1}G(s) - g - h_{s+1} \\
&< \sum_{i=1}^{I} \lambda_i(z_{i,s+1})(z_{i,s+1} - G(s)) + \lambda_i(z_{i,s+1})G(s) + \mu_s G(s-1) - g - h_s \\
&< \sum_{i=1}^{I} \lambda_i(z_{i,s+1})G(s).
\end{aligned}$$

If $\sum_{i=1}^{I} \lambda_i(z_{i,s+1}) = 0$, then $G(s) = \frac{g+h_{s+1}}{\mu_{s+1}} \geq 0$, which is impossible. Therefore, $G(s+1) < G(s) < 0$. Since $\mu_{s+2} \geq \mu_{s+1}$, we have $\mu_{s+2}G(s+1) < \mu_{s+1}G(s) < 0$. Consequently, we can repeat the argument above until we reach state $N-1$, for which $G(N-1) < 0$. But $G(N-1) = \frac{g+h_N}{\mu_N} \geq 0$, which yields a contradiction. Therefore, for all $s = -1, \ldots, N-1$, $0 \leq G(s) \leq \frac{g+h_{s+1}}{\mu_{s+1}}$. □

Proof of Theorem 4.2.1 The existence of a canonical triplet $(g, G(\cdot), \mathbf{z})$ to (3)-(5) is a direct consequence of Lemma 4.2.1. Since the state space is finite, we can refer to equation (5.2.12) in Lasserre and Hernández-Lerma [11] to prove that the canonical triplet $(g, G(\cdot), \mathbf{z})$ is an optimal solution. Therefore, $g_N^* = g$ and $\mathbf{z}^* = \mathbf{z}$.

It remains to show that $z_{i,s}^* = \inf\{z : r_i(z)(z - G(s)) \geq 1\}$ and that it is the unique optimal stationary policy. For $s = 0, \ldots, N-1$, and $i = 1, \ldots, I$, let

$$v_{i,s}(z) = \lambda_i(z)(z - G(s)),$$

$$v'_{i,s}(z) = (1 - F_i(z)) - f_i(z)(z - G(s)), \text{ a.e. on } [\alpha_i, \beta_i].$$

Note also that $v'_{i,s}(z) > (<) 0$ is equivalent to $r_i(z)(z - G(s)) < (>) 1$. The IGHR assumption implies that $r_i(z)(z - G(s)) \geq 1$ almost everywhere on $(\inf\{z : r_i(z)(z - G(s)) \geq 1\}, \beta_i)$.

Therefore, $v'_{i,s}(\cdot) > 0$ almost everywhere on $(\alpha_i, \inf\{z : r_i(z)(z - G(s)) \geq 1\})$ and $v'_{i,s}(\cdot) < 0$ almost everywhere on $(\inf\{z : r_i(z)(z - G(s)) \geq 1\}, \beta_i)$. Thus, $v_{i,s}(\cdot)$ is strictly unimodal and $z^*_{i,s} = \inf\{z : r_i(z)(z - G(s)) \geq 1\}$ is its unique maximizer on $[\alpha_i, \beta_i]$.

We still need to show that $\mathbf{z}^*$ is the unique optimal stationary policy. Under IGHR, $z^*_{i,s}$ is the unique maximizer of $\sup\{\lambda_i(z)(z - G(s))\}$. So, $g^*_N > \sum_{i=1}^{I} \lambda_i(z_i)(z_i - G(s)) + \mu_s G(s - 1) - h_s$ for all $z_i \neq z^*_{i,s}$. Since we have a unichain model, we can refer to Proposition 8.5.10 in Puterman [18] to prove the uniqueness of the optimal stationary policy $\mathbf{z}^*$. □

We are now able to characterize an optimal stationary policy explicitly. Note that it might be possible that $z^*_{i,s} = \beta_i$ for some state s. In this case, it is optimal for the service provider not to accept customers of class i when in state s. However, this can only happen if the class-i customers' willingness-to-pay distribution has finite support. Indeed, if F_i has infinite support, then for all $s = 0, \ldots, N - 1$, $\sup\{\lambda_i(z)(z - G(s))\} > 0$ and $z^*_{i,s} < \infty = \beta_i$. Moreover, note that $z^*_{i,s} = \inf\{z : (z - G(s))r_i(z) \geq 1\} \geq \inf\{z : zr_i(z) \geq 1\}$. Since $\inf\{z : zr_i(z) \geq 1\}$ is the optimal price to charge when demand function is $1 - F_i(z)$, we observe that holding costs and capacity limitations force the service provider to charge higher prices than she normally would if she had no constraints.

We will now exhibit structural properties of the derived optimal stationary policy. More specifically, we are in interested in the monotonicity of the optimal stationary policy. In the next Proposition, we demonstrate that the optimal prices to be charged are nondecreasing in the state index.

Proposition 4.2.1 *Suppose* $\{\mu_s\}_{s=0}^N$ *and* $\{h_s\}_{s=0}^N$ *are such that there exists an integer* q *between* 0 *and* N*, where* $\mu_0 \leq \mu_1 \leq \ldots \leq \mu_q = \mu_{q+1} = \ldots = \mu_N$ *and* $0 = h_0 = h_1 = \ldots = h_q \leq h_{q+1} \leq \ldots \leq h_N$*. Then,* $z^*_{i,s}$ *is nondecreasing in* s*.*

To prove this result, we need the following lemma.

Lemma 4.2.3 *Suppose* $\{\mu_s\}_{s=0}^N$ *and* $\{h_s\}_{s=0}^N$ *are such that there exists an integer* q *between* 0 *and* N*, where* $\mu_0 \leq \mu_1 \leq \ldots \leq \mu_q = \mu_{q+1} = \ldots = \mu_N$ *and* $0 = h_0 = h_1 = \ldots = h_q \leq h_{q+1} \leq \ldots \leq h_N$*. Then,* $G(\cdot)$ *is nondecreasing.*

Proof We decompose our proof into two parts. We will prove first by induction that $G(s)$ is nondecreasing for states $s = 0, \ldots, q-1$. Then, we will show the same for states $s = q-1, \ldots, N-1$.

Suppose $G(s-1) \leq G(s)$ for some state $s = 0, \ldots, q-2$, which is true when $s = 0$. Then, $\mu_s G(s-1) \leq \mu_{s+1} G(s)$ and

$$\begin{aligned}\sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(s+1))\} &= g_N^* - \mu_{s+1} G(s) \\ &\leq g_N^* - \mu_s G(s-1) \\ &\leq \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(s))\}.\end{aligned}$$

Hence, $\sum_{i=1}^I \lambda_i(z_{i,s}^*) G(s+1) \geq \sum_{i=1}^I \lambda_i(z_{i,s}^*) G(s)$. Therefore, either $\sum_{i=1}^I \lambda_i(z_{i,s}^*) = 0$ or $G(s) \leq G(s+1)$. We show that $\sum_{i=1}^I \lambda_i(z_{i,s}^*) = 0$ is impossible.

Assume that $\sum_{i=1}^I \lambda_i(z_{i,s}^*) = 0$. It implies that

$$G(s-1) = G(s) = \frac{g_N^*}{\mu_s} = \frac{g_N^*}{\mu_{s+1}} \text{ and}$$
$$\sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(s-1))\} = \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(s))\} = 0.$$

Therefore, $G(s-2) = \frac{g_N^*}{\mu_{s-1}} \geq \frac{g_N^*}{\mu_s} = G(s-1)$ and $\sum_{i=1}^I \sup\{\lambda_i(z)(z - G(s-2))\} \leq \sum_{i=1}^I \sup\{\lambda_i(z)(z - G(s-1))\} = 0$. We can repeat this argument until we reach the contradiction $G(-1) \geq \frac{g_N^*}{\mu_1}$. By induction, $G(s-1) \leq G(s)$ holds for $s = 0, \ldots, q-1$.

If $q = N$, the proof is complete. Otherwise, it remains to show by induction that $G(s-1) \leq G(s)$ for states $s = q, \ldots, N-1$. Recall that $G(N-1) = \frac{g_N^* + h_N}{\mu_N}$ and $G(N-2) \leq$

$\frac{g_N^*+h_{N-1}}{\mu_{N-1}}$. Therefore, $G(N-2) \leq \frac{g_N^*+h_{N-1}}{\mu_N} = G(N-1)$. Now suppose that $G(s) \geq G(s-1)$ for some state $s = q+1, \ldots, N-1$. Then,

$$\begin{aligned} g_N^* + h_s - \mu_N G(s-1) &= \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(s))\} \\ &\leq \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(s-1))\} \\ &\leq g_N^* - h_{s-1} - \mu_N G(s-2). \end{aligned}$$

Hence, $\mu_N(G(s-1) - G(s-2)) \geq h_s - h_{s-1} \geq 0$. By induction, $G(s) \geq G(s-1)$ for $s = q, \ldots, N-1$ and the proof is complete. □

Proof of Proposition 4.2.1 Recall that $z_{i,s}^* = \inf\{z : r_i(z)(z - G(s)) \geq 1\}$. Since $G(s)$ is nondecreasing in s, so is $z_{i,s}^*$. □

Therefore, in queues with the holding cost and service rate structure described above (such as multiple server systems), the service provider charges more as the system becomes congested. As a consequence, the admission rates are nonincreasing with respect to the number of people in the system. Hence, the optimal policy performs a congestion control that prevents high holding costs. Moreover, we showed in the proof of Lemma 4.2.3 that $\sum_{i=1}^{I} \lambda_i(z_{i,s}^*) > 0$ for all $s < q-1$. This property is quite intuitive since states 0 through q do not incur any holding cost, so that it is not profitable for the service provider to refuse entrance to customers in those states.

We are now interested in how the system reacts to an increase in capacity. A larger buffer size affects the holding costs as well as the revenue by welcoming more customers. As capacity increases, we show that the optimal prices decrease state by state whereas the optimal reward increases. In the following, subscripts 1 and 2 identify parameters for systems 1 and 2, respectively.

Proposition 4.2.2 *Consider two systems 1 and 2, where system 1 has capacity N and system 2 has capacity $N+1$. Then, $g^*_{N+1} \geq g^*_N$. If for all $s = 0, \ldots, N-1$, there exists $i = 1, \ldots, I$ such that $z^*_{i,s,1} < \beta_i$, then $z^*_{i,s,2} \leq z^*_{i,s,1}$.*

Proof It is straightforward to show that $g^*_{N+1} \geq g^*_N$ since the action space for system 2 includes the action space for system 1.

Suppose that $G_2(s) > G_1(s)$ for some state $s = 0, \ldots, N-1$. Therefore,

$$\sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G_2(s))\} < \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G_1(s))\} \text{ or}$$
$$\sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G_2(s))\} = \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G_1(s))\} = 0.$$

The latter case is impossible since it implies that $z^*_{i,s,1} = z^*_{i,s,2} = \beta_i$ for all i. So,

$$\begin{aligned}
\mu_s G_2(s-1) &= g^*_{N+1} - \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G_2(s))\} \\
&> g^*_N - \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G_1(s))\} \\
&> \mu_s G_1(s-1).
\end{aligned}$$

By induction, $0 = G_2(-1) > G_1(-1) = 0$, which yields a contradiction. Therefore, for all $s = 0, \ldots, N-1$, $G_2(s) \leq G_1(s)$ and consequently, $z^*_{i,s,2} \leq z^*_{i,s,1}$. □

We analyze now how the optimal reward varies as other parameters change. As earlier, subscripts 1 and 2 refer to system 1 and 2 respectively. For instance, if customers are willing to pay more, revenue increases but so do the system congestion and holding costs. In the next proposition, we characterize the sensitivity of the optimal reward to the willingness-to-pay distribution as well as other system parameters.

Proposition 4.2.3 *Consider two systems 1 and 2 that satisfy all of the following :*

1. $\Lambda_{i,1} \geq \Lambda_{i,2}$, *for* $i = 1, \ldots, I$,

2. $F_{i,1} \geq_{ST} F_{i,2}$, *for* $i = 1, \ldots, I$,

3. $h_{s,1} \leq h_{s,2}$, *for* $s = 0, \ldots, N$,

4. $\mu_{s,1} \geq \mu_{s,2}$, *for* $s = 1, \ldots, N$.

Then, $g^*_{N,1} \geq g^*_{N,2}$.

Proof We will prove the result by contradiction. Suppose that conditions 1,2,3 and 4 hold and $g^*_{N,1} < g^*_{N,2}$. Therefore, $G_1(N-1) = \frac{g^*_{N,1}+h_{N,1}}{\mu_{N,1}} < \frac{g^*_{N,2}+h_{N,2}}{\mu_{N,2}} = G_2(N-1)$. Suppose that $G_1(s) < G_2(s)$ for some $s = 0, \ldots, N-1$. Then,

$$\begin{aligned} g^*_{N,2} + h_{s,2} - \mu_{s,2}G_2(s-1) &= \sum_{i=1}^{I} \sup\{\lambda_{i,2}(z)(z - G_2(s)\} \\ &\leq \sum_{i=1}^{I} \sup\{\lambda_{i,1}(z)(z - G_1(s))\} \\ &\leq g^*_{N,1} + h_{s,1} - \mu_{s,1}G_1(s-1)). \end{aligned}$$

So, $G_1(s-1) < G_2(s-1)$. By induction, $0 = G_1(-1) < G_2(-1) = 0$, exhibiting a contradiction. Thus, $g^*_{N,1} \geq g^*_{N,2}$. □

We now compare our optimal policy with the optimal static price we derived in Theorem 3.2.2. Suppose now that $\mu_s = \mu$, $h_s = hs$, for all $s = 0, \ldots, N$ and that $I = 1$ as in Chapter 3. It is clear that dynamic pricing achieves a better optimal profit. Moreover, there is an ordering relationship between our optimal static price and our optimal stationary policy. As in Chapter 3, recall that $\pi_n(\rho, N)$ denote the stationary probability of n customers in the system under static pricing when traffic intensity is ρ. Recall also that $L(\rho, N)$ is the long-run expected number of customers in the system under static pricing and traffic intensity ρ.

Proposition 4.2.4 *Let $I = 1$ and let y_N^* denote the optimal static price for a system of capacity N. Then, for all $s = 1 \dots N$, $z_{1,0}^* \leq y_N^* \leq z_{1,N-1}^*$.*

Proof Since $I = 1$, we will omit the class subscript in this proof. For instance, we will write $r(\cdot), \lambda(y)$ and z_s^* instead of $r_1(\cdot), \lambda_1(y)$ and $z_{1,s}^*$. In this proof, we also use the quantity $\rho(y)$ that is defined as $\rho(y) = \frac{\lambda(y)}{\mu}$.

Recall that $z_{N-1}^* = \inf\{y : r(y)(y - \frac{g^*+Nh}{\mu}) \geq 1\}$ and $y_N^* = \inf\{y : r(y)\gamma_N^h(\rho(y))(y - \frac{h}{\mu}\varphi_N(\rho(y))) \geq 1\}$ from Theorem 3.2.2.

To prove that $y_N^* \leq z_{N-1}^*$, we will show that for all $y \geq 0$ such that $r(y)(y - \frac{g_N^*+Nh}{\mu}) \geq 1$, we have $r(y)\gamma_N^h(\rho(y))(y - \frac{h}{\mu}\varphi_N(\rho(y))) \geq 1$. Consider $y \geq 0$ such that $r(y)(y - \frac{g_N^*+Nh}{\mu}) \geq 1$. Since $g_N^* \geq \mu y(1 - \pi_0(\rho(y), N)) - hL(\rho(y), N)$, we can claim by Lemma A.0.4 that

$$\begin{aligned} y - \frac{g_N^* + Nh}{\mu} &\leq y\pi_0(\rho(y), N) - \frac{h}{\mu}(N - L(\rho(y), N)) \\ &\leq \pi_0(\rho(y), N)(y - \frac{h}{\mu}\varphi_N(\rho(y))) \\ &\leq \gamma_N^h(\rho(y))(y - \frac{h}{\mu}\varphi_N(\rho(y))). \end{aligned}$$

Hence, we have $r(y)\gamma_N^h(\rho(y))(y - \frac{h}{\mu}\varphi_N(\rho(y))) \geq 1$, proving that $y_N^* \leq z_{N-1}^*$.

To prove that $z_0^* \leq y_N^*$, we proceed by contradiction. Recall the alternate expression of z_0^* from Lemma B.0.6, which is $z_0^* = \sup\{y : g_N^* r(y) \leq \lambda(y)\}$. If $z_0^* = \beta$, then $g_N^* = 0$, which is impossible. Suppose that $\beta > z_0^* > y_N^*$, so that there exists y in (y_N^*, z_0^*) such that $g_N^* r(y) \leq \lambda(y)$ and $r(y)\gamma_N^h(\rho(y))(y - \frac{h}{\mu}\varphi_N(\rho(y))) > 1$. Using inequalities (2) and (4) from Lemma A.0.1, we have

$$\begin{aligned} \lambda(y)\gamma_N^h(\rho(y))(y - \frac{h}{\mu}\varphi_N(\rho(y))) &\leq \lambda(y)(1 - \pi_N(\rho(y), N))(y - \frac{h}{\mu}\varphi_N(\rho(y))) \\ &\leq \lambda(y)(1 - \pi_N(\rho(y), N))y - hL(\rho(y), N) \\ &\leq g_N^*. \end{aligned}$$

Therefore, $g_N^* r(y) \geq \lambda(y) r(y) \gamma_N^h(\rho(y))(y - \frac{h}{\mu}\varphi_N(\rho(y))) > \lambda(y)$, exhibiting a contradiction. Thus, we have proved that $z_0^* \leq y_N^*$. □

Therefore, y_N^* is convex combination of z_0^* and z_{N-1}^*. We can interpret the optimal static price as a "compromise" between z_0^* and z_{N-1}^*. On the one hand, when the system is empty, the service provider is willing to discount prices to attract customers. On the other hand, when the system is almost full, the service provider charges a premium for higher congestion costs. Under a static pricing scheme, the service provider does not have the possibility to differentiate states when pricing service. Hence, it is intuitive that the optimal price to be charged in this case lies in between the optimal dynamic prices charged in extremal states.

4.2.2 Systems with Balking Customers

We now consider systems with balking customers to capture congestion penalties. As in Section 3.3, customers entering the system are subject to their willingness-to-wait as well as their willingness-to-pay. A potential customer from any class arriving when the system is in state s decides to accept the current congestion level with probability p_s independently of everything else. We assume that $1 = p_0 \geq p_1 \geq \ldots \geq p_{N-1}$ as customers are more deterred by congested states.

As in the holding cost model, we use a Markov decision process (MDP) method. Under any pricing policy, the queue system is unichain as a birth-death process with positive death

rates. We have the following ACOE system:

$$G(-1) = 0, \tag{6}$$

$$g - \mu_s G(s-1) = p_s \sum_{i=1}^{I} \sup_z \{\lambda_i(z)(z - G(s))\}, \text{ if } s = 0, \ldots, N-1, \tag{7}$$

$$G(N-1) = \frac{g}{\mu_N}. \tag{8}$$

In the following theorem, we explicitly characterize a unique optimal stationary policy that maximize the long-run average profit. We give an explicit expression for the optimal prices to be charges in each state for each customer class.

Theorem 4.2.2 *There exists a canonical triplet $(g, G(\cdot), \mathbf{z})$ for the ACOE system (6)-(8). Moreover, the optimal long-run average reward is $g_N^* = g$ and $\mathbf{z}^* = \mathbf{z}$ is a unique optimal stationary policy, where, for $s = 0, \ldots, N-1$ and $i = 1, \ldots, I$,*

$$z_{i,s}^* = \inf\{z : r_i(z)(z - G(s)) \geq 1\}.$$

To show the existence of a canonical triplet, the proof is similar to the proof of Lemma 4.2.1 and is omitted. The remainder of Theorem 4.2.2 is proved similarly to Theorem 4.2.1 and we omit the proof as well. We state some useful properties of canonical triplets in the next lemma.

Lemma 4.2.4 *Let $(g, G(\cdot), \mathbf{z})$ a canonical triplet be for the ACOE system (6)-(8). Then,*

1. *for all $s = -1, \ldots, N-1$, $0 \leq G(s) \leq \frac{g}{\mu_{s+1}}$.*

2. *for all $s = 0, \ldots, N-1$, there exists $i = 1, \ldots, I$ such that $z_{i,s} < \beta_i$.*

Proof Let $(g, G(\cdot), \mathbf{z})$ a canonical triplet be for the ACOE system (6)-(8). First, we show that (1) holds. From equations (7) and (8), we immediately have $G(s) \leq \frac{g}{\mu_{s+1}}$ for $s = 0, \ldots, N-1$.

Suppose that there exists $s = 0, \ldots, N-1$ such that $G(s) < 0$. Since $G(-1) = 0$, there exists $s \geq 0$ such that $G(s) < 0$ and $G(s-1) \geq 0$. Hence, $\mu_{s+1}G(s) - \mu_s G(s-1) < 0$. But we have

$$\begin{aligned}
p_{s+1}\sum_{i=1}^{I}\lambda_i(z_{i,s+1})G(s+1) =& p_{s+1}\sum_{i=1}^{I}\lambda_i(z_{i,s+1})z_{i,s+1} - g + \mu_{s+1}G(s) \\
=& p_{s+1}\sum_{i=1}^{I}\lambda(z_{i,s+1})(z_{i,s+1} - G(s)) - g + \mu_{s+1}G(s) - \mu_s G(s-1) \\
&+ p_{s+1}\sum_{i=1}^{I}\lambda_i(z_{i,s+1})G(s) + \mu_s G(s-1) \\
<& p_s \sum_{i=1}^{I}\sup\{\lambda_i(z)(z - G(s))\} - g + \mu_s G(s-1) \\
&+ p_{s+1}\sum_{i=1}^{I}\lambda_i(z_{i,s+1})G(s) \\
<& p_{s+1}\sum_{i=1}^{I}\lambda_i(z_{i,s+1})G(s).
\end{aligned}$$

If $\sum_{i=1}^{I}\lambda_i(z_{i,s+1}) = 0$, then $G(s) = \frac{g}{\mu_{s+1}} \geq 0$, which is impossible. Therefore, $G(s+1) < G(s) < 0$. Since $\mu_{s+2} \geq \mu_{s+1}$, $\mu_{s+2}G(s+1) - \mu_{s+1}G(s) < 0$, so we can repeat the argument above until we reach state $N-1$, for which $G(N-1) < 0$. But $G(N-1) = \frac{g}{\mu_N} \geq 0$, which yields a contradiction. Therefore, for all $s = -1, \ldots, N-1$, $0 \leq G(s) \leq \frac{g}{\mu_{s+1}}$.

We now show that (2) holds. Now suppose that there exists $s = 0, \ldots, N-1$ such that $z_{i,s} = \beta_i$ for all $i = 1, \ldots, N-1$. This implies that $G(s) \geq \max_i \beta_i$. Therefore, $G(s-1) = \frac{g}{\mu_s} \geq \frac{g}{\mu_{s+1}} \geq G(s) \geq \max_i \beta_i$. Hence, $z_{i,s-1} = \beta_i$ for all $i = 0, \ldots, I$. By induction, we have $G(-1) = \frac{g}{\mu_0}$, exhibiting a contradiction. □

As opposed the holding cost model, the second property in Lemma 4.2.4 implies that it never optimal to set $z_{i,s} = \beta_i$ for all i in some state s. In other words, it is suboptimal for the service provider to refuse entry to all customers in some state $s \leq N-1$.

Using our solution characterization in Theorem 4.2.2, we are now able to derive structural and ordering properties of the optimal pricing solution and reward. In the next proposition, we demonstrate that the optimal prices to be charged are nondecreasing in the state index, under mild assumptions on p_s and μ_s.

Proposition 4.2.5 *Suppose* $\{\mu_s\}_{s=0}^N$ *and* $\{p_s\}_{s=0}^N$ *are such that there exists an integer* q *between* 0 *and* N*, where* $\mu_0 \leq \mu_1 \leq \ldots \leq \mu_q = \mu_{q+1} = \ldots = \mu_N$ *and* $1 = p_0 = p_1 = \ldots = p_{q-1} \geq p_q \geq \ldots \geq p_{N-1}$*. Then,* $G(s)$ *and* $z_{i,s}^*$ *are nondecreasing in* s*.*

Proof We split our proof into two parts. First, we prove by induction that $G(s-1) \leq G(s)$ for $s = 0, \ldots, q-1$. Then, we show that $G(s-1) \leq G(s)$ holds for $s = q, \ldots, N-1$ using an induction as well.

Suppose $G(s-1) \leq G(s)$ for some state $s = 0, \ldots, q-2$. It clearly holds for $s = 0$. Then, $\mu_s G(s-1) \leq \mu_{s+1} G(s)$ and

$$\begin{aligned}\sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(s+1))\} &= g_N^* - \mu_{s+1} G(s) \\ &\leq g_N^* - \mu_s G(s-1) \\ &\leq \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(s))\}.\end{aligned}$$

Hence, $\sum_{i=1}^I \lambda_i(z_{i,s}^*) G(s+1) \geq \sum_{i=1}^I \lambda_i(z_{i,s}^*) G(s)$. From the second result of Lemma 4.2.4, we have $\sum_{i=1}^I \lambda_i(z_{i,s}^*) > 0$. Hence, $G(s+1) \geq G(s)$. By induction, $G(s-1) \leq G(s)$ holds for $s = 0, \ldots, q-1$.

If $q = N$, the proof is complete. Otherwise, it remains to show by induction that $G(s-1) \leq G(s)$ for states $s = q, \ldots, N-1$. Now suppose that $G(s) \geq G(s-1)$ for some state $s = q+1, \ldots, N-1$. The first result of Lemma 4.2.4 shows that it holds when

$s = N - 1$. We have,

$$\begin{aligned} g_N^* - \mu_N G(s-1) &= p_s \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(s))\} \\ &\leq p_{s-1} \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G(s-1))\} \\ &\leq g_N^* - \mu_N G(s-2). \end{aligned}$$

Hence, $\mu_N(G(s-1) - G(s-2)) \geq 0$. By induction, $G(s) \geq G(s-1)$ for $s = q, \ldots, N-1$ and the proof is complete. □

As in the holding cost model, Proposition 4.2.5 shows that the optimal prices act as an indirect congestion control, deterring customers from entering congested states. The service provide must charge more in congested states in order to offset to future loss of potential customers deterred by the queue length. Let us analyze now how the optimal prices and reward change with respect to an increase in the system capacity N. Intuitively, the optimal reward g_N^* should increase as N grows. We demonstrate this in the next proposition. The proof of the following result is a minor change from the proof of Proposition 4.2.2 and can be found in the Appendix. Recall that we use the additional subscripts 1 and 2 when we compare two systems indexed by 1 and 2 respectively.

Proposition 4.2.6 *Consider two systems 1 and 2, where system 1 has capacity N and system 2 has capacity $N+1$. Then, $g_{N+1}^* \geq g_N^*$. Then $z_{i,s,2}^* \leq z_{i,s,1}^*$ for $s = 0, \ldots, N-1$ and $i = 1, \ldots, I$.*

In Section 3.3, we exhibited a unique optimal static price that maximizes the long-run average reward in the case of a unique customer class in a single server system. In the next Proposition, we compare this optimal static price with the optimal dynamic policy derived

in Theorem 4.2.2 in the case of a unique customer class ($I = 1$) in a single server system ($\mu_s = \mu$ for all $s = 1, \ldots, N$).

Proposition 4.2.7 *Let $I = 1$ and consider a single server system of capacity N. Let y_N^* denote the optimal static price to be charged. Then, for all $s = 1 \ldots N$, $z_{1,0}^* \leq y_N^* \leq z_{1,N-1}^*$.*

Proof Since $I = 1$, we omit the customer class subscript in this proof. Consider a single server system with service rate μ. Recall from Theorem 3.3.1 that $y_N^* = \inf\{y : yr(y)\gamma_N^b(\rho(y)) \geq 1\}$, where $\rho(y) = \frac{\lambda(y)}{\mu}$. First, we show that $y_N^* \leq z_{1,N-1}^*$. From Theorem 4.2.2, we have $z_{N-1}^* = \inf\{y : r(y)(y - \frac{g_N^*}{\mu}) \geq 1\}$

We will show that for all $y \geq 0$ such that $r(y)(y - \frac{g_N^*}{\mu}) \geq 1$, we have $r(y)\gamma_N^b(\rho(y)) \geq 1$. Consider $y \geq 0$ such that $r(y)(y - \frac{g_N^*}{\mu}) \geq 1$. Since $g_N^* \geq \mu y(1 - \pi_0(\rho(y), N))$, we have from Lemma A.0.5,

$$y - \frac{g_N^*}{\mu} \leq \pi_0(\rho(y), N)y \leq \gamma_N^b(\rho(y))y.$$

Therefore, we have $r(y)\gamma_N^b(\rho(y))y \geq 1$, implying that $y_N^* \leq z_{N-1}^*$.

We now prove that $z_0^* \leq y_N^*$. From Lemma B.0.6, we can express z_0^* as $z_0^* = \sup\{y : g_N^* r(y) \leq \lambda(y)\}$. Suppose that $z_0^* > y_N^*$, so that there exists y in (y_N^*, z_0^*) such that $g_N^* r(y) \leq \lambda(y)$ and $r(y)\gamma_N^b(\rho(y))y > 1$. From Lemma A.0.5, we have

$$\lambda(y)\gamma_N^b(\rho(y))y \leq \lambda(y)y\pi_0(\rho(y), N) \sum_{n=0}^{N-1} P_n \rho(y)^n \leq g_N^*.$$

Therefore, $g_N^* r(y) \geq \lambda(y) r(y)\gamma_N^b(\rho(y))y > \lambda(y)$ revealing a contradiction. Therefore, $z_0^* \leq y_N^*$. □

As in the model with holding cost, in single server queues with a unique customer class, the optimal static price lies in between the optimal dynamic prices to be charged in extremal states.

4.2.3 Systems with Impatient Customers

In this section, we model congestion penalties through impatient customers. The customer impatient behavior is the same as described in Section 3.4. Payments are collected upon arrival. Each customer entering the system has a random maximum waiting time distributed as an exponential random variable with parameter θ. The maximum waiting times of successive customers are assumed to be independent of each other, of the arrival process, of the service time process and of the customers' willingness-to-pay. If a customer does not begin service prior to his maximum waiting time, the customer reneges and receives a full refund from the service provider. We assume that the system has q identical servers with rate μ; that is, $\mu_s = (s \wedge q)\mu$. Only customers who are waiting for service are impatient. Therefore, we assume that $N > q$ without loss of generality.

As opposed to the static pricing scheme used in Section 3.4, the refunds received by impatient customers can now vary from one customer to another. Therefore, it seems that we need to include the prices paid by customers currently in the system in the system state description. In other words, defining the state of the system as the number of customers seems to be insufficient to use a Markovian analysis. We show that it is not the case when the queueing discipline is First-In First-Out (FIFO).

4.2.3.1 FIFO Queueing Discipline

Let $w_s, s = 0, \ldots, N-1$ denote the probability that a customer entering a system in state s will not renege and will leave upon service completion. In the remainder of this section, we assume that the system operates under a FIFO queueing discipline. Consequently, it is clear that w_s does not depend on the arrival process. We compute w_s in the next proposition.

Proposition 4.2.8 *Consider a system with impatient customers and q servers under FIFO queueing discipline. For $s = 0, \ldots, N-1$, we have $w_s = \frac{\mu q}{(s-q+1)^+\theta+\mu q}$.*

Proof If $s < q$, then $w_s = 1$. Otherwise, we can compute w_s the following way: $w_s = P(X_s \geq \sum_{n=q}^{s} Y_n)$ where $X_s \sim \text{Expo}(\theta)$ and $Y_n \sim \text{Expo}((n-q)\theta + \mu q)$. The random variable X_s represents the patience time of the customer of interest. The random variables $Y_q, \ldots, Y_s$ represent the interdeparture times of the customers waiting in line in front of the customer of interest. We have :

$$
\begin{aligned}
w_s &= P(X_s \geq \sum_{n=q}^{s} Y_n) \\
w_s &= \int_0^\infty \cdots \int_0^\infty P(X_s \geq y_q + \cdots + y_s) \prod_{n=q}^{s} (\mu q + (n-q)\theta) e^{-(\mu q + (n-q)\theta) q y_n} dy_n \\
w_s &= \int_0^\infty \cdots \int_0^\infty e^{-\theta(y_q + \cdots + y_s)} \prod_{n=q}^{s} (\mu q + (n-q)\theta) e^{-(\mu q + (n-q)\theta) q y_n} dy_n \\
w_s &= \prod_{n=q}^{s} \int_0^\infty (\mu q + (n-q)\theta) e^{-(\mu q + (n-q+1)\theta) y_n} dy_n \\
w_s &= \prod_{n=q}^{s} \frac{(n-q)\theta + \mu q}{(n-q+1)\theta + \mu q} \\
w_s &= \frac{\mu q}{(s-q)\theta + \mu q}
\end{aligned}
$$

□

As earlier, we use a Markov decision process formulation. However, we use the expected reward generated by a customer admission in state s instead of the actual reward. Under pricing policy $\mathbf{z}$, the expected reward generated by a class-i customer entering the system in state s is $z_{i,s} w_s$. Since w_s does not depend on the pricing policy and only depends on the current state s, using expected rewards spares us from keeping track of the prices paid by impatient customers (see p.20 in Puterman [18]).

We have the following system of ACOE:

$$G(-1) = 0, \tag{9}$$

$$g - (\mu_s + (s-q)^+\theta)G(s-1) = \sum_{i=1}^{I} \sup_z \{\lambda_i(z)(w_s z - G(s))\}, \text{ if } s = 0, \ldots, N-1, \tag{10}$$

$$G(N-1) = \frac{g}{\mu_N + (N-q)\theta}. \tag{11}$$

In the next theorem and lemma, we show the existence a canonical triplet $(g, G(\cdot), \mathbf{z})$ to (9)-(11) that corresponds to an optimal dynamic pricing solution. We also characterize the optimal prices to be charged.

Theorem 4.2.3 *There exists a canonical triplet $(g, G(\cdot), \mathbf{z})$ for the ACOE system (9)-(11). Moreover, the optimal long-run average reward is $g_N^* = g$ and $\mathbf{z}^* = \mathbf{z}$ is a unique optimal stationary policy, where, for $s = 0, \ldots, N-1$ and $i = 1, \ldots, I$,*

$$z_{i,s}^* = \inf\left\{z : r_i(z)\left(z - \frac{G(s)}{w_s}\right) \geq 1\right\}.$$

Since the state space is finite, the proof of Theorem 4.2.3 is similar to the proof of Theorem 4.2.1 and is omitted. The following lemma states useful properties of canonical triplets to (9)-(11).

Lemma 4.2.5 *Let $(g, G(\cdot), \mathbf{z})$ be a canonical triplet for the ACOE system (9)-(11). Then,*

1. *for all $s = -1, \ldots, N-1$, $0 \leq G(s) \leq \frac{g}{\mu_{s+1} + (s-q+1)^+\theta}$.*

2. *for all $s = 0, \ldots, N-1$, there exists $i = 1, \ldots, I$ such that $z_{i,s} < \beta_i$.*

Proof Let $(g, G(\cdot), \mathbf{z})$ be a canonical triplet for the ACOE system (9)-(11). First, we show that (1) holds. Equations (10) and (11), immediately imply that $G(s) \leq \frac{g}{\mu_{s+1} + (s-q+1)^+\theta}$ for $s = 0, \ldots, N-1$. Suppose that $G(s) < 0$ for some $s = 0, \ldots, N-1$. Since $G(-1) = 0$, there

exists $s \geq 0$ such that $G(s) < 0$ and $G(s-1) \geq 0$. Therefore, $(\mu_{s+1} + (s-q+1)^+\theta)G(s) - (\mu_s + (s-q)^+\theta)G(s-1) < 0$. We have

$$\begin{aligned}
\sum_{i=1}^{I} \lambda_i(z_{i,s+1})G(s+1) &= \sum_{i=1}^{I} \lambda_i(z_{i,s+1})w_{s+1}z_{i,s+1} - g + (\mu_{s+1} + (s-q+1)^+\theta)G(s) \\
&= \sum_{i=1}^{I} \lambda(z_{i,s+1})(w_{s+1}z_{i,s+1} - G(s)) - g + (\mu_{s+1} + (s-q+1)^+\theta)G(s) \\
&\quad + \sum_{i=1}^{I} \lambda_i(z_{i,s+1})G(s) \\
&< \sum_{i=1}^{I} \sup\{\lambda_i(z)(w_s z - G(s))\} - g + (\mu_s + (s-q)^+\theta)G(s-1) \\
&\quad + \sum_{i=1}^{I} \lambda_i(z_{i,s+1})G(s) \\
&< \sum_{i=1}^{I} \lambda_i(z_{i,s+1})G(s).
\end{aligned}$$

If $\sum_{i=1}^{I} \lambda_i(z_{i,s+1}) = 0$, then $G(s) = \frac{g}{\mu_{s+1}} \geq 0$, which is impossible. Hence, $G(s+1) < G(s) < 0$. Since $(\mu_{s+2} + (s+2-q)^+\theta) \geq (\mu_{s+1} + (s+1-q)^+\theta)$, $(\mu_{s+2} + (s+2-q)^+\theta)G(s+1) - (\mu_{s+1} + (s+1-q)^+\theta)G(s) < 0$, this argument can be repeated until we reach $G(N-1) < 0$. But $G(N-1) = \frac{g}{\mu_N + (N-q)^+\theta} \geq 0$, which yields a contradiction. Therefore, for all $s = -1, \ldots, N-1$, $0 \leq G(s) \leq \frac{g}{\mu_{s+1} + (s+1-q)^+\theta}$.

We now show that (2) holds. Now suppose that there exists $s = 0, \ldots, N-1$ such that $z_{i,s} = \beta_i$ for all $i = 1, \ldots, N-1$. This implies that $G(s) \geq \max_i \beta_i$. Therefore, $G(s-1) = \frac{g}{\mu_s + (s-q)^+\theta} \geq \frac{g}{\mu_{s+1} + (s+1-q)^+\theta} \geq G(s) \geq \max_i \beta_i$. Hence, $z_{i,s-1} = \beta_i$ for all $i = 0, \ldots, I$. By induction, we have $G(-1) = \frac{g}{\mu_0} = 0$, yielding a contradiction. □

We now focus on deriving structural and ordering properties of the optimal solution and reward. Similarly to the models with holding costs or balking customers, we show in the next proposition that the optimal prices to be charged are nondecreasing in the state index.

Proposition 4.2.9 *The sequences $\frac{G(s)}{w_s}$ and $z^*_{i,s}$ are nondecreasing in s.*

Proof First, we show that $\frac{G(s)}{w_s}$ is nondecreasing. As $z^*_{i,s} = \inf\left\{z : r_i(z)\left(z - \frac{G(s)}{w_s}\right) \geq 1\right\}$, this implies that $z^*_{i,s}$ is nondecreasing in s.

We split our proof into two parts. First, we prove by induction that $G(s-1) \leq G(s)$ for $s = 0, \ldots, q-1$ since $w_s = 1$ for $s < q$. This part of the proof is identical to the beginning of the proof of Proposition 4.2.5 and is omitted. Second, we show that $\frac{G(s-1)}{w_{s-1}} \leq \frac{G(s)}{w_s}$ holds for $s = q, \ldots, N-1$.

Suppose that $\frac{G(s-1)}{w_{s-1}} \leq \frac{G(s)}{w_s}$ for some state $s = q+1, \ldots, N-1$. The first part of Lemma 4.2.5 shows that it holds when $s = N-1$. Then,

$$\begin{aligned}
g^*_N - (\mu_N + (s-q)\theta)G(s-1) &= \sum_{i=1}^{I} \sup\{\lambda_i(z)(w_s z - G(s))\} \\
&= w_s \sum_{i=1}^{I} \sup\left\{\lambda_i(z)\left(z - \frac{G(s)}{w_s}\right)\right\} \\
&\leq w_{s-1} \sum_{i=1}^{I} \sup\left\{\lambda_i(z)\left(z - \frac{G(s-1)}{w_{s-1}}\right)\right\} \\
&\leq \sum_{i=1}^{I} \sup\{\lambda_i(z)(w_{s-1} z - G(s-1))\} \\
&\leq g^*_N - (\mu_N + (s-q-1)\theta)G(s-2).
\end{aligned}$$

Hence, $(\mu_N + (s-q)\theta)G(s-1) - (\mu_N + (s-q-1)\theta)G(s-2) \geq 0$ implying that $\frac{G(s-1)}{w_{s-1}} \geq \frac{G(s-2)}{w_{s-2}}$. By induction, $\frac{G(s)}{w_s} \geq \frac{G(s-1)}{w_{s-1}}$ for $s = q, \ldots, N-1$ and the proof is complete. □

In the next proposition, we investigate how the optimal prices and rewards change with an increase in capacity N. On the one hand, in systems of larger capacity, fewer customers are turned downed due to capacity limitations. On the other hand, the customers' waiting time and their likelihood of reneging are higher as N gets larger. We show that the optimal reward is nondecreasing in N and that the price to be charged to class-i customers in state s is nonincreasing in N. The proof of the next proposition given in the Appendix.

Proposition 4.2.10 *Consider two systems 1 and 2, where system 1 has capacity N and system 2 has capacity $N+1$. Then, $g^*_{N+1} \geq g^*_N$. Then $z^*_{i,s,2} \leq z^*_{i,s,1}$ for $s = 0, \ldots, N-1$ and $i = 1, \ldots, I$.*

In Section 3.4, we characterized a unique static optimal price y^*_N to be charged in the case of a single server $M/M/1$ queue with a unique customer class. In the following proposition, we investigate how the optimal dynamic prices derived in Theorem 4.2.3 compare with y^*_N, when there is a unique customer class ($I = 1$) in a single server queue ($\mu_s = \mu$ for all $s = 1, \ldots, N$).

Proposition 4.2.11 *Let $I = 1$ and consider a single server system of capacity N. Let y^*_N denote the optimal static price. Then, for all $s = 1 \ldots N$, $z^*_{1,0} \leq y^*_N \leq z^*_{1,N-1}$.*

Proof As $I = 1$, we omit the class subscript in the remainder of this proof. From Theorem 3.4.1, we have $y^*_N = \inf\{y : yr(y)\gamma^r_N(\rho(y)) \geq 1\}$, where $\rho(y) = \frac{\lambda(y)}{\mu}$. First, we show that $y^*_N \leq z^*_{1,N-1}$. From Theorem 4.2.3, recall that

$$z^*_{N-1} = \inf\left\{y : r(y)\left(y - \frac{g^*_N}{w_{N-1}(\mu + (N-1)\theta)}\right) \geq 1\right\} = \inf\left\{y : r(y)\left(y - \frac{g^*_N}{\mu}\right) \geq 1\right\}.$$

We prove that for all $y \geq 0$ such that $r(y)(y - \frac{g^*_N}{\mu}) \geq 1$, we have $r(y)\gamma^r_N(\rho(y)) \geq 1$. Consider $y \geq 0$ such that $r(y)(y - \frac{g^*_N}{\mu}) \geq 1$. Since dynamic pricing performs better than static pricing, we have $g^*_N \geq \mu y(1 - \pi_0(\rho(y), N))$. From Lemma A.0.5,

$$y - \frac{g^*_N}{\mu} \leq \pi_0(\rho(y), N)y \leq \gamma^r_N(\rho(y))y.$$

Hence, we conclude that $r(y)\gamma^r_N(\rho(y))y \geq 1$, implying that $y^*_N \leq z^*_{N-1}$.

We now show that $z^*_0 \leq y^*_N$. Recall from Lemma B.0.6 that z^*_0 satisfies $z^*_0 = \sup\{y : g^*_N r(y) \leq \lambda(y)\}$. Suppose that $z^*_0 > y^*_N$. Consequently, there must exist y in (y^*_N, z^*_0) such

that $g_N^* r(y) \leq \lambda(y)$ and $r(y)\gamma_N^r(\rho(y))y > 1$. From Lemma A.0.5, we have

$$\lambda(y)\gamma_N^r(\rho(y))y \leq \lambda(y)y\pi_0(\rho(y), N)\sum_{n=0}^{N-1} Q_n\rho(y)^n.$$

Note that

$$\begin{aligned}\lambda(y)y\pi_0(\rho(y), N)\sum_{n=0}^{N-1} Q_n\rho(y)^n =& \mu y\pi_0(\rho(y), N)\sum_{n=0}^{N-1}\frac{\lambda(y)^{n+1}}{\mu(\mu+\theta)\ldots(\mu+n\theta)}\\ =& \mu y\sum_{n=1}^{N}\pi_n(\rho(y), N)\\ =& \mu y(1-\pi_0(\rho(y), N)) \leq g_N^*.\end{aligned}$$

Consequently, $g_N^* r(y) \geq \lambda(y)r(y)\gamma_N^r(\rho(y))y > \lambda(y)$ and we reach a contradiction. We conclude that $z_0^* \leq y_N^*$. □

4.2.3.2 FIFO vs. LIFO

Under static pricing and with a unique customer class, it is clear that the optimal prices and reward do not depend on the queueing discipline. In this case, all customers pay the same price and can be considered as identical once in the system. However, we assumed that the service provider enforces a FIFO queueing discipline when analyzing the impatient customer model under dynamic precision pricing. In this section, we investigate how changing the queueing discipline to a preemptive Last-In First-Out (LIFO) discipline affects the optimal reward.

In the following, we assume that we only have one customer class ($I = 1$). Let $R^{LIFO}(\mathbf{z})$ denote the long-run average reward for the system under pricing policy $\mathbf{z}$ and preemptive LIFO queueing discipline. In the next theorem, we demonstrate that, a system under LIFO performs better than a system under FIFO provided that the advertised prices are nondecreasing in the state index.

Theorem 4.2.4 *If $I = 1$, then for any stationary pricing policy $\mathbf{z}$ such that z_s is nondecreasing in s, we have $R(\mathbf{z}) \leq R_{LIFO}(\mathbf{z})$.*

To prove this theorem, we need the following proposition and lemma.

Proposition 4.2.12 *Suppose that $I = 1$ and consider any queueing discipline and stationary pricing policy $\mathbf{z}$ such that $z_{1,s}$ is nondecreasing in s. Then, for all $s = 0, \ldots N-1$, when $s+1$ customers are in the system, the customer who was admitted into the system last paid at least $z_{1,s}$.*

Proof If $s+1$ is the current number of customers in the system, the last customer to enter the system encountered at least s customers in the system upon arrival. Therefore, the customer question paid at least $z_{1,s}$. □

Lemma 4.2.6 *Consider a system with a single customer class ($I = 1$), under FIFO discipline and stationary pricing policy $\mathbf{z}$. Then, the long-run average reward can also be expressed as*

$$R(\mathbf{z}) = \sum_{s=0}^{N-1} \mu_s z_{1,s} \pi_{s+1}(\mathbf{z}).$$

Proof Recall that

$$R(\mathbf{z}) = \sum_{s=0}^{N-1} \lambda_1(z_{1,s}) z_{1,s} w_s \pi_s(\mathbf{z}),$$

where $w_s = \frac{\mu q}{\mu q + (s-q+1)^+ \theta}$. Therefore,

$$\begin{aligned} R(\mathbf{z}) &= \sum_{s=0}^{N-1} \lambda_1(z_{1,s}) z_{1,s} \frac{\mu q}{\mu q + (s-q+1)^+ \theta} \pi_s(\mathbf{z}), \\ &= \sum_{s=0}^{N-1} \mu_{s+1} z_{1,s} \pi_{s+1}(\mathbf{z}). \end{aligned}$$

□

Lemma 4.2.6 shows that, under the FIFO queueing discipline, the system performs as if any exiting customer leaving s customers behind had paid price $z_{1,s}$ upon admission. We use this property to prove Theorem 4.2.4.

Proof of Theorem 4.2.4 Consider a system under preemptive LIFO queueing discipline. From Lemma 4.2.12, we can claim that any customer being serviced and leaving the system with s customers behind must have paid at least $z_{1,s}$ upon arrival. Therefore, we use Lemma 4.2.6 and claim that

$$R^{LIFO}(\mathbf{z}) \geq \sum_{s=0}^{N-1} \mu_{s+1} z_{1,s} \pi_{s+1}(\mathbf{z}) = R(\mathbf{z}).$$

□

A direct consequence of Theorem 4.2.4 is that a service provider using the optimal policy $\mathbf{z}^*$ as defined in Theorem 4.2.3 can improve profits by enforcing a LIFO queueing discipline. However, implementing LIFO incurs some hidden costs that are not captured in our model. Namely, under LIFO, customers who have been in the system for a long time may witness other customers with shorter waiting time being processed before them. There is a loss of customer goodwill that is associated with the customers' discontent. Therefore, the benefits from using a LIFO queueing policy might be offset by this hidden cost.

4.3 Queueing Systems with Infinite Capacity

In this section, we impose no limitation on the system capacity. This introduces some difficulties since the ACOE system now has infinitely many equations and solution triplets. Moreover, unlike in the finite capacity case, a canonical triplet does not always translate into an optimal stationary policy. Nevertheless, under certain parameter structures, we are able to find an optimal stationary policy that maximizes the long-term average profit per unit time.

4.3.1 Uniform Asymptotic Parameter Structure

First, we consider a particular parameter structure in the model with holding costs and the model with balking customers.

Definition 4.3.1 *The service system is said to have Uniform Asymptotic Parameter Structure (UAPS) if its parameters satisfy*

- *in the model with holding costs: there exists $N < \infty$, such that $h_s = h_N$, $\mu_s = \mu_N$ for $s \geq N$ and $\sum_{i=1}^{I} \Lambda_i < \mu_N$,*
- *in the model with balking customers: there exists $N < \infty$, such that $p_s = p_N$, $\mu_s = \mu_N$ for $s \geq N$ and $p_N \sum_{i=1}^{I} \Lambda_i < \mu_N$.*

Note that we do not have a UAPS in the model with impatient customers. Since each customer in the queue is impatient, the reneging rate cannot be constant past a certain state.

In order to show the existence of an optimal stationary policy, we use the mappings Ψ^h and Ψ^b defined as

$$\begin{aligned}
\Psi^h : R \times R^+ &\to R \\
(V, g) &\mapsto \frac{g + h_N - \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - V)\}}{\mu_N} \\
\Psi^b : R \times R^+ &\to R \\
(V, g) &\mapsto \frac{g - p_N \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - V)\}}{\mu_N}.
\end{aligned}$$

We show in Lemma B.0.7 that for all $g \geq 0$, $\Psi^h(\cdot, g)$ and $\Psi^b(\cdot, g)$ are nondecreasing and each have a unique fixed point. This fixed point is denoted by $FP^h(g)$ and $FP^b(g)$ for $\Psi^h(\cdot, g)$ and $\Psi^b(\cdot, g)$ respectively.

Instead of having all of the infinitely many equations of the ACOE system, we only consider a finite subset corresponding to states $s = 0, \ldots, N-1$. We use the functions FP^h and FP^b to define $G(s)$ for states $s \geq N-1$. Thus, we analyze the following systems of optimality equations.

- In the model with holding costs:

$$G(-1) = 0, \tag{12}$$

$$g + h_s - \mu_s G(s-1) = \sum_{i=1}^{I} \sup_z \{(z - G(s))\lambda_i(z)\} \qquad \text{if } s = 0, \ldots, N-1, \tag{13}$$

$$G(s) = FP^h(g) \qquad \text{if } s = N-1, \ldots, \infty. \tag{14}$$

- In the model with balking customers:

$$G(-1) = 0, \tag{15}$$

$$g - \mu_s G(s-1) = p_s \sum_{i=1}^{I} \sup_z \{(z - G(s))\lambda_i(z)\} \qquad \text{if } s = 0, \ldots, N-1, \tag{16}$$

$$G(s) = FP^b(g) \qquad \text{if } s = N-1, \ldots, \infty. \tag{17}$$

Note that if there exists a canonical triplet satisfying (12)-(14) or (15)-(17), it also satisfies the full system of optimality equations from the corresponding congestion penalty model. In the next theorem, we demonstrate that there exists canonical triplets to (12)-(14) and (15)-(17). Moreover, we prove the existence of an optimal stationary policy that we characterize in each UAPS model.

Theorem 4.3.1 *The following statements hold:*

1. *There exists a canonical triplet to the ACOE (12)-(14) and (15)-(17) respectively.*

2. *Let $(g, G(\cdot), \mathbf{z})$ is a canonical triplet to (12)-(14). Then, $\mathbf{z}^* = \mathbf{z}$ and $g^*_\infty = g$ in the UAPS model with holding costs. Moreover, $z^*_{i,s} = \inf\{z : r_i(z)(z - G(s)) \geq 1\}$ for all $s = 0, \ldots, \infty$ and $i = 1, \ldots, I$.*

3. *Let $(g, G(\cdot), \mathbf{z})$ is a canonical triplet to (15)-(17). Then, $\mathbf{z}^* = \mathbf{z}$ and $g^*_\infty = g$ in the UAPS model with balking customers. Moreover, $z^*_{i,s} = \inf\{z : r_i(z)(z - G(s)) \geq 1\}$ for all $s = 0, \ldots, \infty$ and $i = 1, \ldots, I$.*

Proof

1. First, we show that there exists a canonical triplet to (12)-(14). We only need to show the existence a canonical triplet to (12), (13) and $G(N-1) = FP^h(g)$. Then, by extending $G(\cdot)$ and $\mathbf{z}$ to $G(s) = G(N-1) = FP^h(g)$ and $z_{i,s} = z_{i,N-1}$ for $s \geq N-1$, $(g, G(\cdot), \mathbf{z})$ is also a solution of the full system of ACOE (12)-(14).

   To prove the existence of a canonical triplet to (12), (13) and $G(N-1) = FP^h(g)$, note that

   $$\Psi^h(0, \sum_{i=1}^{I} \sup\{\lambda_i(z)z\}) = \frac{h_N}{\mu_N} \geq 0,$$

   and consequently $FP^h(\sum_{i=1}^{I} \sup\{\lambda_i(z)z\}) \geq 0$. Therefore, the proof is exactly the same as in Lemma 4.2.1 except that $G(N-1, g) = FP^h(g)$ is now the starting point of the induction. In same fashion, we show the existence of a canonical triplet to (15)-(17).

2. Let $(g, G(\cdot), \mathbf{z})$ is a canonical triplet to (12)-(14). We now prove that it corresponds to an optimal solution in the UAPS model with holding costs. According to equation (5.2.12) of Lasserre and Hernández-Lerma [11], we only need to show that

   $$\lim_{t\to\infty} \inf_{d \in \Pi^{RH}} \frac{E^d_{s_0}[l(X(t))]}{t} = 0,$$

where Π^{RH} is the set of all history-dependent randomized policies and s_0 is the starting state at time $t = 0$. If it is the case, the canonical triplet corresponds to an optimal stationary policy. Recall that $l(s) - l(s+1) = G(s)$ for all $s = 0, \ldots, \infty$. Therefore, $l(s) = l(N) - (s - N)FP^h(g)$ for all $s \geq N$. Hence,

$$\begin{aligned} E^d_{s_0}[l(X(t))] =& E^d_{s_0}[l(X(t))|X(t) < N]P(X(t) < N) \\ &+ \Big(l(N) - (E^d_{s_0}[X(t)|X(t) \geq N] - N)FP^h(g)\Big) P(X(t) \geq N). \end{aligned}$$

Note that for all $d \in \Pi^{RH}$,

$$\Big|E^d_{s_0}[l(X(t))|X(t) < N]P(X(t) < N)\Big| \leq \max_{s \leq N-1} |l(s)|.$$

Therefore, we have

$$\lim_{t\to\infty} \inf_{d\in\Pi^{RH}} \left| \frac{E^d_{s_0}[l(X(t))|X(t) < N]P(X(t) < N)}{t} \right| = 0.$$

It remains to show that $\lim_{t\to\infty} \sup_{d\in\Pi^{RH}} \frac{E^d_{s_0}[X(t)|X(t)\geq N]P(X(t)\geq N)}{t} = 0$. First, note that

$$\sup_{d\in\Pi^{RH}} E^d_{s_0}[X(t)|X(t) \geq N]P(X(t) \geq N) \leq \sup_{d\in\Pi^{RH}} E^d_{s_0}[X(t)]$$

and that the supremum $\sup_{d\in\Pi^{RH}} E^d_{s_0}[X(t)]$ is attained for policy $\hat{\mathbf{z}}$, where $\hat{\mathbf{z}}$ denotes the stationary policy of charging price 0 to all customers in all states ($\hat{z}_{i,s} = 0$ for $i = 1, \ldots, I$ and $s = 0, \ldots, \infty$). We have

$$\lim_{t\to\infty} \sup_{d\in\Pi^{RH}} E^d_{s_0}[X(t)] \leq \lim_{t\to\infty} E^{\hat{\mathbf{z}}}_{s_0}[X(t)] < \infty.$$

Therefore,

$$\lim_{t\to\infty} \inf_{d\in\Pi^{RH}} \frac{E^d_{s_0}[l(X(t))]}{t} = 0,$$

and we can apply Theorem 5.2.4 from Lasserre and Hernández-Lerma [11] and claim that $\mathbf{z} = \mathbf{z}^*$ and $g = g^*_\infty$.

3. In the UAPS model with balking customers, the proof is identical to part 2 using $FP^b(g)$ in lieu of $FP^h(g)$ and is omitted.

□

This theorem enables us to explicitly characterize an optimal stationary policy in systems under UAPS. Note that the service provider charges the same price $z^*_{i,N-1}$ to class-i customers for all states $s \geq N-1$. This property is quite surprising since there is no apparent symmetry in the transition structure to justify it.

4.3.2 General Parameters: Systems with Holding Costs

Under Uniform Asymptotic Parameter Structure, we are able to explicitly derive an optimal stationary pricing policy. The assumption that $\mu_s = \mu_N$ for $s \geq N$ for some N is often encountered as servers become saturated with congestion. However, a linear holding cost structure does not allow us to use a UAPS. Neither do systems with impatient customers. In this section, we analyze infinite capacity systems with more general parameter structures.

We assume that the service system has q identical servers such that $\mu_s = \mu(s \wedge q)$. First, we focus on the model with holding costs with h_s increasing to infinity and integrable with respect to $\{\frac{(\sum_{i=1}^I \Lambda_i)^s}{\mu_1 \ldots \mu_s}\}$. We assume that $h_0 = h_1 = \ldots = h_q = 0$ and $\frac{h_1}{\mu_1} < \max \beta_i$ in order to have an attainable positive reward. We also suppose that $\sum_{i=1}^I \Lambda_i < \mu q$ so that the system is stable under any pricing policy.

First, let us consider willingness-to-pay distributions with finite support ($\beta_i < \infty$ for all $i = 1, \ldots, I$). In this case, we demonstrate in the next proposition that we can actually restrict our analysis to finite capacity systems; that is, it is optimal not to admit customers in the system past a certain finite congestion level.

Proposition 4.3.1 *If $\beta_i < \infty$ for all $i = 1, \ldots, I$, then $g^*_\infty = g^*_M$, where $M = \max\{s :$*

$h_s < \mu_s \max \beta_i\}$.

Proof To prove this proposition, we show that for any stationary policy of the infinite capacity system, one can find a stationary policy of the truncated M-capacity system that performs as well. Let $\mathbf{z}$ be a stationary pricing policy for the infinite capacity system such that $R(\mathbf{z}) > 0$. This policy exists since the assumption $\max \beta_i > \frac{h_1}{\mu_1}$ ensures the existence of a positive reward. Now consider the M-capacity stationary pricing policy $\mathbf{z}^{|\mathbf{M}}$, which is defined as the truncation of $\mathbf{z}$ up to state $M-1$ included. More precisely, $z_{i,s}^{|M} = z_{i,s}$ for all $s < M$ and $i = 1, \ldots, I$. We have

$$\begin{aligned} R(\mathbf{z}) &= \sum_{s=0}^{\infty} \left(\sum_{i=1}^{I} \lambda_i(z_{i,s}) z_{i,s} \right) \pi_s(\mathbf{z}) - h_{s+1}\pi_{s+1}(\mathbf{z}) \\ &= \sum_{s=0}^{\infty} \left(\sum_{i=1}^{I} \lambda_i(z_{i,s})(z_{i,s} - \frac{h_{s+1}}{\mu_{s+1}}) \right) \pi_s(\mathbf{z}) \\ &= \sum_{s=0}^{\infty} a_s \pi_s(\mathbf{z}) \text{ and} \\ R(\mathbf{z}^{|\mathbf{M}}) &= \sum_{s=0}^{M-1} a_s \pi_s(\mathbf{z}^{|\mathbf{M}}), \end{aligned}$$

where $a_s = \sum_{i=1}^{I} \lambda_i(z_{i,s})(z_{i,s} - \frac{h_{s+1}}{\mu_{s+1}})$. Clearly, the definition of M implies that $a_s \leq 0$ for $s \geq M$. It is straightforward to show that for all $s \leq M$,

$$\pi_s(\mathbf{z}^{|\mathbf{M}}) = \frac{\pi_s(\mathbf{z})}{\sum_{s=0}^{M} \pi_s(\mathbf{z})}.$$

Hence $R(\mathbf{z}^{|\mathbf{M}}) - R(\mathbf{z})$ has the same sign as $\sum_{s=0}^{M-1} a_s \pi_s(\mathbf{z}) - \sum_{s=0}^{\infty} a_s \pi_s(\mathbf{z}) \sum_{s=0}^{M} \pi_s(\mathbf{z})$ and we have

$$\begin{aligned} \sum_{s=0}^{M-1} a_s \pi_s(\mathbf{z}) - \sum_{s=0}^{\infty} a_s \pi_s(\mathbf{z}) \sum_{s=0}^{M} \pi_s(\mathbf{z}) &\geq \sum_{s=0}^{M-1} a_s \pi_s(\mathbf{z}) - \sum_{s=0}^{M-1} a_s \pi_s(\mathbf{z}) \sum_{s=0}^{M} \pi_s(\mathbf{z}) \\ &\geq \sum_{s=0}^{M-1} a_s \pi_s(\mathbf{z}) \left(1 - \sum_{s=0}^{M} \pi_s(\mathbf{z}) \right). \end{aligned}$$

Recall that $R(\mathbf{z}) > 0$ and

$$\sum_{s=0}^{M-1} a_s \pi_s(\mathbf{z}) \geq \sum_{s=0}^{M-1} a_s \pi_s(\mathbf{z}) + \sum_{s=M}^{\infty} a_s \pi_s(\mathbf{z}) \geq R(\mathbf{z}) > 0.$$

Therefore,

$$\sum_{s=0}^{M-1} a_s \pi_s(\mathbf{z}) - \sum_{s=0}^{\infty} a_s \pi_s(\mathbf{z}) \sum_{s=0}^{M} \pi_s(\mathbf{z}) \geq 0$$

and $R(\mathbf{z}^{|\mathbf{M}}) \geq R(\mathbf{z})$, proving the result. □

Proposition 4.3.1 shows that if all the willingness-to-pay distributions $F_1, \ldots, F_I$ have finite support, we can restrict our analysis to finite capacity queues and refer to Section 4.2.1. Therefore, without loss of generality, we now assume that at least one of the willingness-to-pay distributions $F_1, \ldots, F_I$ has infinite support in the rest of this section. To prove the existence of an optimal stationary policy, we approximate the infinite capacity system by a finite capacity model of large size. We validate this approximation through two limiting results in Proposition 4.3.2 and Theorem 4.3.2. Note that Weber and Stidham [20] provide a proof for the existence of an optimal stationary policy in the case of a compact action space.

Proposition 4.3.2 *If $\sum_{i=1}^{I} \Lambda_i < \mu q$, then $g_N^* \uparrow g_\infty^*$ as N goes to infinity.*

Theorem 4.3.2 *Let $(g_N^*, G(\cdot), \mathbf{z}^N)$ be the canonical triplet associated with the truncated system of capacity N. Then, under the stability condition $\sum_{i=1}^{I} \Lambda_i < \mu q$, there exists $\mathbf{z}$, such that $z_{i,s}^N \downarrow z_{i,s}$ $\forall i, s$ as $N \to \infty$. Moreover, $\mathbf{z} = \mathbf{z}^*$ is optimal for the infinite capacity model.*

We need the following lemma to prove Proposition 4.3.2 and Theorem 4.3.2.

Lemma 4.3.1 *Let $\mathbf{z}$ be a stationary pricing policy and $\mathbf{z}^{|\mathbf{N}}$ be the truncation of $\mathbf{z}$ up to state $N-1$. Under the stability condition $\sum_{i=1}^{I} \Lambda_i < \mu q$, $R(\mathbf{z}^{|\mathbf{N}}) \to R(\mathbf{z})$ as N goes to infinity.*

Proof Consider

$$\begin{aligned} R(\mathbf{z}) =& \sum_{s=0}^{\infty} \pi_s(\mathbf{z}) \sum_{i=1}^{I} \lambda_i(z_{i,s}) z_{i,s} - h_{s+1}\pi_{s+1}(\mathbf{z}) \\ =& \pi_0(\mathbf{z}) \sum_{s=0}^{\infty} \frac{\sum_i \lambda_i(z_{i,0}) \ldots \sum_i \lambda_i(z_{i,s-1})}{\mu_1 \ldots \mu_s} \left(\sum_{i=1}^{I} \lambda_i(z_{i,s})(z_{i,s} - \frac{h_{s+1}}{\mu_{s+1}}) \right). \end{aligned}$$

Moreover,

$$\begin{aligned} R(\mathbf{z}^{|\mathbf{N}}) =& \sum_{s=0}^{N-1} \pi_s(\mathbf{z}^{|\mathbf{N}}) \sum_{i=1}^{I} \lambda_i(z_{i,s}) z_{i,s} - h_{s+1}\pi_{s+1}(\mathbf{z}^{|\mathbf{N}}) \\ =& \pi_0(\mathbf{z}^{|\mathbf{N}}) \sum_{s=0}^{N-1} \frac{\sum_i \lambda_i(z_{i,0}) \ldots \sum_i \lambda_i(z_{i,s-1})}{\mu_1 \ldots \mu_s} \left(\sum_{i=1}^{I} \lambda_i(z_{i,s})(z_{i,s} - \frac{h_{s+1}}{\mu_{s+1}}) \right). \end{aligned}$$

Since $\pi_0(\mathbf{z}^{|\mathbf{N}})^{-1} = 1 + \sum_{s=0}^{N-1} \frac{\sum_i \lambda_i(z_{i,0}) \ldots \sum_i \lambda_i(z_{i,s})}{\mu_1 \ldots \mu_{s+1}}$, as N goes to infinity, $\pi_0(\mathbf{z}^{|\mathbf{N}}) \to \pi_0(\mathbf{z})$. Therefore, $R(\mathbf{z}^{|\mathbf{N}}) \to R(\mathbf{z})$, which proves the desired result. □

Proof of Proposition 4.3.2 We will prove this proposition by contradiction. From Proposition 4.2.2, we know that g_N^* is nondecreasing in N and that $\lim_N g_N^*$ exists and is less than or equal to g_∞^*. Now suppose that $\lim_N g_N^* < g_\infty^*$. Then, according to Lemma 4.3.1, there exists an N-capacity stationary policy $\mathbf{z}^N$, such that $\lim_N g_N^* < R(\mathbf{z}^N) < g_\infty^*$. As $R(\mathbf{z}^N) \leq g_N^*$, we have a contradiction and the proof is complete. □

Proof of Theorem 4.3.2 We proved in Proposition 4.3.2 that g_N^* converges to g_∞^*. Since we assumed that at least one of the willingness-to-pay distributions $F_1, \ldots, F_I$ has infinite support, we use Proposition 4.2.2 to claim that $z_{i,s}^N$ is a nonincreasing sequence in N. Therefore, $\lim_N z_{i,s}^N = z_{i,s}$ exists.

We now show that $|R(\mathbf{z}) - g_N^*| \to 0$. Since $g_N^* \to g_\infty^*$, it must imply that $R(\mathbf{z}) = g_\infty^*$ and that $\mathbf{z}$ is optimal.

To do so, we will prove first that for any s, $\pi_s(\mathbf{z}^N) \to \pi_s(\mathbf{z})$ as N goes to infinity. Since

$$\pi_s(\mathbf{z}^N) = \pi_0(\mathbf{z}^N) \frac{\sum_i \lambda_i(z_{i,0}^N) \ldots \sum_i \lambda_i(z_{i,s-1}^N)}{\mu_1 \ldots \mu_s}, \tag{18}$$

we only need to prove that $\pi_0(\mathbf{z}^N) \to \pi_0(\mathbf{z})$. We have

$$\pi_0(\mathbf{z}^N)^{-1} - \pi_0(\mathbf{z})^{-1} = \sum_{s=1}^{N} \frac{\sum_i \lambda_i(z_{i,0}^N) \dots \sum_i \lambda_i(z_{i,s-1}^N)}{\mu_1 \dots \mu_s} - \sum_{s=1}^{\infty} \frac{\sum_i \lambda_i(z_{i,0}) \dots \sum_i \lambda_i(z_{i,s-1})}{\mu_1 \dots \mu_s}. \tag{19}$$

Let M be an arbitrary integer smaller than N,

$$\begin{aligned} \pi_0(\mathbf{z}^N)^{-1} - \pi_0(\mathbf{z})^{-1} = & \sum_{s=1}^{M} \frac{\sum_i \lambda_i(z_{i,0}^N) \dots \sum_i \lambda_i(z_{i,s-1}^N)}{\mu_1 \dots \mu_s} - \frac{\sum_i \lambda_i(z_{i,0}) \dots \sum_i \lambda_i(z_{i,s-1})}{\mu_1 \dots \mu_s} \\ & + \sum_{s=M}^{N} \frac{\sum_i \lambda_i(z_{i,0}^N) \dots \sum_i \lambda_i(z_{i,s-1}^N)}{\mu_1 \dots \mu_s} \\ & - \sum_{s=M}^{\infty} \frac{\sum_i \lambda_i(z_{i,0}) \dots \sum_i \lambda_i(z_{i,s-1})}{\mu_1 \dots \mu_s}. \end{aligned} \tag{20}$$

Hence,

$$\begin{aligned} |\pi_0(\mathbf{z}^N)^{-1} - \pi_0(\mathbf{z})^{-1}| \leq & \sum_{s=1}^{M} \left| \frac{\sum_i \lambda_i(z_{i,0}^N) \dots \sum_i \lambda_i(z_{i,s-1}^N)}{\mu_1 \dots \mu_s} - \frac{\sum_i \lambda_i(z_{i,0}) \dots \sum_i \lambda_i(z_{i,s-1})}{\mu_1 \dots \mu_s} \right| \\ & + 2 \sum_{s=M}^{\infty} \frac{(\sum_i \Lambda_i)^s}{\mu_1 \dots \mu_s}. \end{aligned} \tag{21}$$

First, let N go to infinity and then let M go to infinity. We have $\pi_s(\mathbf{z}^N) \to \pi_s(\mathbf{z})$ for all $s \geq 0$ as N goes to infinity.

Now consider

$$\begin{aligned} R(\mathbf{z}) - g_N^* = & \sum_{s=0}^{\infty} \sum_{i=1}^{I} z_{i,s} \lambda_i(z_{i,s}) \pi_s(\mathbf{z}) - \sum_{s=0}^{N-1} \sum_{i=1}^{I} z_{i,s}^N \lambda_i(z_{i,s}^N) \pi_s(\mathbf{z}^N) \\ & - \sum_{s=1}^{\infty} h_s \pi_s(\mathbf{z}) + \sum_{s=1}^{N} h_s \pi_s(\mathbf{z}^N). \end{aligned}$$

So,

$$\begin{aligned} |R(\mathbf{z}) - g_N^*| \leq & \sum_{s=0}^{M} \sum_{i=1}^{I} |z_{i,s} \lambda_i(z_{i,s}) \pi_s(\mathbf{z}) - z_{i,s}^N \lambda_i(z_{i,s}^N) \pi_s(\mathbf{z}^N)| + \sum_{s=1}^{M} h_s |\pi_s(\mathbf{z}) - \pi_s(\mathbf{z}^N)| \\ & + 2I \sup_{i,z} \{z \lambda_i(z)\} \sum_{s=M}^{\infty} \frac{(\sum_{i=1}^{I} \Lambda_i)^s}{\mu_1 \dots \mu_s} + 2 \sum_{s=M}^{\infty} h_s \frac{(\sum_{i=1}^{I} \Lambda_i)^s}{\mu_1 \dots \mu_s}. \end{aligned}$$

Letting N go to ∞ yields

$$\lim_{N\to\infty} |R(\mathbf{z}) - g_N^*| \leq 2I \sup_{i,z}\{z\lambda_i(z)\} \sum_{s=M}^{\infty} \frac{(\sum_{i=1}^{I} \Lambda_i)^s}{\mu_1 \dots \mu_s} + 2 \sum_{s=M}^{\infty} h_s \frac{(\sum_{i=1}^{I} \Lambda_i)^s}{\mu_1 \dots \mu_s},$$

and letting M go to ∞ implies that $\lim_N g_N^* = R(\mathbf{z})$. Therefore, $\mathbf{z} = \mathbf{z}^*$ is an optimal stationary policy and the proof is complete. □

The key element of this result is that the stationary probability of being in highly congested states is negligible under any stationary policy. Thus, an infinite capacity system can be approximated by a system of large finite capacity. The optimal stationary policy exists and is the limit of finite-capacity optimal policies, which enables us to state the following proposition.

Proposition 4.3.3 *Under the stability condition* $\sum_{i=1}^{I} \Lambda_i < \mu q$, $z_{i,s}^*$ *is nondecreasing in* s.

Proof From Theorem 4.3.2, we know that $z_{i,s}^* = \lim_N z_{i,s}^N$, where $z_{i,s}^N$ is the optimal price at state s for the truncated N-capacity system. By Proposition 4.2.1, $z_{i,s}^N$ is nondecreasing in s. Hence, the same holds for $z_{i,s}^*$. □

Although we do not characterize the optimal stationary policy explicitly in this case, we can still derive some insights. Not surprisingly, the structures of the infinite and finite capacity optimal policies are the same. High prices are charged in congested states in order to minimize holding costs.

4.3.3 General Parameters: Systems with Balking Customers and Systems with Impatient Customers

In this section, we study systems with balking customers and systems with impatient customers when $N = \infty$ in cases when UAPS does not necessarily apply. We still assume that the service system has q identical servers such that $\mu_s = \mu(s \wedge q)$. Similarly to the case with

holding costs, we approximate systems of infinite capacity by systems of large finite capacity. We show the existence of an optimal pricing policy for both congestion models under a stability condition. In systems with balking customers, we assume that $\lim_s p_s \sum_{i=1}^{I} \Lambda_i < \mu q$ so that the system is stable under any stationary pricing policy. Moreover, we suppose that $1 = p_0 = p_1 = \ldots = p_{q-1} \geq p_q \geq \ldots \geq p_{N-1}$. In systems with impatient customers, we assume that a FIFO queueing discipline is enforced. Stability follows directly from the departures of impatient customers. As opposed the previous section, we need not consider separately willingness-to-pay distributions that have finite or infinite support (see Lemmas 4.2.4 and 4.2.5). In the following theorem, we show the existence of an optimal pricing solution in systems with balking customers and in systems with impatient customers. Similarly to Theorem 4.3.2, we use systems of large finite capacity to approximate infinite capacity systems.

Theorem 4.3.3 *The following holds for systems with balking customers if* $\lim_s p_s \sum_{i=1}^{I} \Lambda_i < \mu q$ *and for systems with impatient customers. Let* $(g_N^*, G(\cdot), \mathbf{z}^N)$ *be the canonical triplet associated with the truncated system of capacity* N*. Then,* $g_N^* \uparrow g_\infty^*$ *and there exists* $\mathbf{z}$*, such that* $z_{i,s}^N \downarrow z_{i,s}$ $\forall i, s$ *as* $N \to \infty$*. Moreover,* $\mathbf{z} = \mathbf{z}^*$ *is optimal for the infinite capacity model.*

Proof In systems with balking customers, the proof follows the same path as the proofs of Proposition 4.3.2, Lemma 4.3.1 and Theorem 4.3.2 by setting $h_s = 0$ and substituting $\{p_s\lambda_i(z_{i,s})\}$ for $\{\lambda_i(z_{i,s})\}$, $\{p_s\lambda_i(z_{i,s}^N)\}$ for $\{p_s\lambda_i(z_{i,s}^N)\}$ and $\{P_{s-1}(\sum_{i=1}^{I} \Lambda_i)^s\}$ for $\{(\sum_{i=1}^{I} \Lambda_i)^s\}$.

Consider now the case of impatient customers. We follow the same framework as in Proposition 4.3.2 and Theorem 4.3.2 using the long-run optimal rewards for stationary

pricing policy $\mathbf{z}$ and its truncation (up to state $N-1$ included) $\mathbf{z}^{|\mathbf{N}}$:

$$R(\mathbf{z}) = \pi_0(\mathbf{z}) \sum_{s=0}^{\infty} \frac{\sum_i \lambda_i(z_{i,0}) \dots \sum_i \lambda_i(z_{i,s-1})}{(\mu_1 + (1-q)^+\theta) \dots (\mu_s + (s-q)^+\theta)} \sum_{i=1}^{I} \lambda_i(z_{i,s}) w_s z_{i,s}$$

$$R(\mathbf{z}^{|\mathbf{N}}) = \pi_0(\mathbf{z}^{|\mathbf{N}}) \sum_{s=0}^{N-1} \frac{\sum_i \lambda_i(z_{i,0}) \dots \sum_i \lambda_i(z_{i,s-1})}{(\mu_1 + (1-q)^+\theta) \dots (\mu_s + (s-q)^+\theta)} \sum_{i=1}^{I} \lambda_i(z_{i,s}) w_s z_{i,s},$$

where

$$\pi_0(\mathbf{z})^{-1} = 1 + \sum_{s=0}^{\infty} \frac{\sum_i \lambda_i(z_{i,0}) \dots \sum_i \lambda_i(z_{i,s-1})}{(\mu_1 + (1-q)^+\theta) \dots (\mu_s + (s-q)^+\theta)} \text{ and}$$

$$\pi_0(\mathbf{z}^{|\mathbf{N}})^{-1} = 1 + \sum_{s=0}^{N-1} \frac{\sum_i \lambda_i(z_{i,0}) \dots \sum_i \lambda_i(z_{i,s-1})}{(\mu_1 + (1-q)^+\theta) \dots (\mu_s + (s-q)^+\theta)}.$$

It is straightforward to show that $R(\mathbf{z}^{|\mathbf{N}}) \to R(\mathbf{z})$ as N goes to infinity as in Lemma 4.3.1, implying that $g_N^* \uparrow g_\infty^*$ as in Proposition 4.3.2.

Let $(g_N^*, G(\cdot), \mathbf{z}^N)$ be the canonical triplet associated with the truncated system of capacity N. From Proposition 4.2.10, $z_{i,s}^N$ is nonincreasing in N for all $i = 1, \dots, I$ and $s < N$, therefore it has a limit $z_{i,s}$ corresponding to a stationary pricing policy $\mathbf{z}$. As in Theorem 4.3.2, we now show that $g_N^* \to R(\mathbf{z})$, which implies that $R(\mathbf{z}) = g_\infty^*$ and that $\mathbf{z}$ is an optimal pricing solution for the infinite capacity model.

We need to show first that for any s, $\pi_s(\mathbf{z}^N) \to \pi_s(\mathbf{z})$ as N goes to infinity. This is easily verified using equations (18) to (21) and substituting $(\mu_1 + (1-q)^+\theta) \dots (\mu_s + (s-q)^+\theta)$ for $\mu_1 \dots \mu_s$.

We have

$$R(\mathbf{z}) - g_N^* = \sum_{s=0}^{\infty} \sum_{i=1}^{I} w_s z_{i,s} \lambda_i(z_{i,s}) \pi_s(\mathbf{z}) - \sum_{s=0}^{N-1} \sum_{i=1}^{I} w_s z_{i,s}^N \lambda_i(z_{i,s}^N) \pi_s(\mathbf{z}^N).$$

So, for an arbitrary integer M smaller than N,

$$|R(\mathbf{z}) - g_N^*| \le \sum_{s=0}^{M} w_s \sum_{i=1}^{I} \left| z_{i,s} \lambda_i(z_{i,s}) \pi_s(\mathbf{z}) - z_{i,s}^N \lambda_i(z_{i,s}^N) \pi_s(\mathbf{z}^N) \right| + 2I \sup_{i,z} \{ z \lambda_i(z) \} \sum_{s=M}^{\infty} \frac{(\sum_{i=1}^{I} \Lambda_i)^s}{(\mu_1 + (1-q)^+\theta) \dots (\mu_s + (s-q)^+\theta)}.$$

Letting N go to infinity yields

$$\lim_{N\to\infty} |R(\mathbf{z}) - g_N^*| \leq 2I \sup_{i,z}\{z\lambda_i(z)\} \sum_{s=M}^{\infty} \frac{(\sum_{i=1}^{I} \Lambda_i)^s}{(\mu_1 + (1-q)^+\theta)\ldots(\mu_s + (s-q)^+\theta)}.$$

Letting M go to infinity implies that $g_N^* \to R(\mathbf{z})$ and the proof is complete. □

A direct consequence of Theorem 4.3.3 is that $z_{i,s}^*, i = 1, \ldots, I$ is nondecreasing in s in the model with balking customers as well as in the model with impatient customers. The congestion control performed by optimal prices in finite capacity systems extends to models with infinite capacity.

4.4 Summary

In this chapter, we allowed the service provider to adjust prices. We relaxed the model described in Chapter 3 by having multiple customer classes to whom the service provider can advertise specific prices. Our objective was to exhibit optimal dynamic pricing policies that maximize the long-run average reward in each of the three congestion models considered separately.

First, we focused on queues with finite capacity. For each congestion model, we characterized a unique optimal stationary policy using a Markov decision process formulation. In systems with impatient customers, we made the extra assumption of a FIFO queueing discipline to analyze the model without keeping track of payment history. With mild assumptions on the parameters' structure, we showed that the optimal prices to be charged are nondecreasing in the state index for each congestion model. Therefore, the service provider indirectly controls congestion penalties by deterring customers to enter congested states. Comparing our results with those of Chapter 3, we also demonstrated that, in each congestion model with a unique customer class, the optimal static price lies in between the optimal dynamic prices to be charged in extremal states. Moreover, in systems with

impatient customers, we investigated the impact of a queueing policy change on the optimal reward and showed that implementing a LIFO queueing policy improves the performance of the system.

Second, we considered systems with infinite capacity. We characterized an optimal stationary policy for systems with holding costs and systems with balking customers under Uniform Asymptotic Parameter Structure. Recall that UAPS describes systems whose parameters are the same for all states $s \geq N$ for some finite index N. In this case, we showed that the optimal solution has the same prices being charged to class-i customers across all states $s \geq N - 1$. In instances where UAPS does not necessarily hold, we showed the existence of an optimal dynamic solution by approximating systems of infinite capacity by systems of large finite capacity. For each of the three congestion models, we demonstrated that both the optimal prices and rewards of a finite capacity system converge to those of an infinite capacity system.

CHAPTER V

SUMMARY AND FUTURE RESEARCH

In this section, we summarize the main contributions of this dissertation and provide suggestions for future research.

5.1 Main Results

We analyzed the problem of optimal pricing in queueing systems where congestion plays a key role. We modelled congestion penalties in three different ways and analyzed both static pricing and dynamic pricing schemes in each case. We chose to analyze the three congestion penalties separately to isolate their respective effects on optimal rewards and pricing policies. Most of the literature in the area of pricing in queueing system only considers a single pricing scheme and a unique way of capturing congestion penalties (most often through holding costs) and does not provide a comprehensive study of how congestion affects profit. Moreover, the issue of optimal pricing in systems with balking customers and in systems with impatient customers with refund has received little attention. The same can be claimed about the comparison of optimal dynamic prices with static prices. We also noticed that much of literature imposes strong restrictions on the system parameters and the action space when analyzing dynamic pricing in infinite capacity queues.

The main result of our work is the determination of an optimal static price as well as an optimal stationary policy in each of the three congestion models. In each case, we compared the optimal static price with optimal dynamic prices. Although much of the literature considers the system capacity as a given parameter, we investigated its relationship

with optimal prices and rewards. For instance, we showed that, under static pricing, the service provider should restrict the system capacity when holding costs are incurred. In systems with balking customers and in systems with impatient customers, capacity should be unrestricted to maximize profit.

In systems with impatient customers under dynamic pricing, we showed that we need not keep track of payment history to determine an optimal stationary policy when a FIFO queueing discipline is enforced. This enabled us to keep a Markovian formulation with the system state describing the number of customers in the facility.

Another important contribution of our work is the analysis of dynamic pricing in infinite capacity queues. With mild assumptions on the action space, we demonstrated that the optimal dynamic prices and rewards of finite capacity systems converge to those of an infinite capacity system in each of the three congestion models. In cases with specific parameter structures, we explicitly characterized an optimal stationary policy when capacity is unlimited.

5.2 Future Research

In the following, we suggest some uninvestigated research topics related to our work.

5.2.1 Optimal Pricing in Systems with Multiple Congestion Penalties and Priorities

In this dissertation, we chose to analyze the three types of congestion penalties separately to isolate and compare the features and effects of each model. However, some service systems can very well experience a mix of congestion penalties in practice. For instance, combining impatient and balking customers is quite natural. Indeed, balking customers can be considered as "smart" impatient customers who are able to forecast their waiting time and might leave before even entering the system. Under dynamic pricing, it is straightforward to

combine any of the three models discussed in Chapter 4 of this dissertation. Nevertheless, the determination of optimal static prices with multiple types of penalties is more intricate. It is unclear whether the long-run average reward under static pricing is strictly unimodal when combining congestion penalties.

The other interesting feature of combining congestion penalties is the use of priorities. Consider a system where arriving customers can be impatient or not. A natural priority scheme would be to serve impatient customers first in order to minimize their waiting time and consequently curb the refunds paid to those who renege. The service provider may have the option to advertise different prices to customers that are impatient or not. Intuitively, impatient customers must be charged higher prices. We need a two-dimensional state space to describe the state of such a system as

$$X(t) = (\text{number of impatient customers at time } t, \text{ total number of customers at time } t).$$

The priority system and the transition structure disrupt the birth-death properties we have in this dissertation. Therefore, the determination of optimal dynamic pricing solutions requires the use of a policy iteration or value iteration method. One might also consider an asymptotic regime such as a fluid or diffusion approximation to investigate such a system where priorities significantly complicate the exact analysis.

5.2.2 Optimal Pricing with Adjustable Service Rate in Systems with Balking Customers and Systems with Impatient Customers

In our research, we considered service rates as given unadjustable parameters. By doing so, we assumed that service capacity was a sunk cost and that no action could be taken to expand or reduce the service offering. However, in some telecommunication service systems, the service rate could be dynamically adjusted. This is the case in wireless transmission applications where the power needed to transmit data packets can be tuned. There is a

clear tradeoff as lower transmission power consumes less energy but degrades the quality of service. Ata and Shneorson [2] investigate this issue and consider the dynamic pricing problem of maximizing the long-run average reward in $M/M/1$ queues with adjustable service rates. However, they only consider holding costs as a way to capture congestion penalties.

We can extend this model to systems with balking customers and systems with impatient customers. The model with impatient customers is particularly relevant in transmission networks with timeouts. If a data packet is not transmitted within a specific amount of time, the connection is dropped in the same fashion as an impatient customer would renege.

Let $c(\mu)$ denote the cost per unit time associated with offering service rate μ. We assume that $c(\cdot)$ is a continuous nondecreasing function and that $c(0) = 0$. Then, we can express the ACOE in $M/M/1/N$ systems with balking customers as:

$$
\begin{aligned}
&G(-1) = 0, \\
&\qquad g = p_s \sum_{i=1}^{I} \sup_{z} \{\lambda_i(z)(z - G(s))\} + \sup_{\mu>0} \{\mu G(s-1) - c(\mu)\}, \text{ if } s = 0, \ldots, N-1, \\
&\qquad g = \sup_{\mu>0} \{\mu G(N-1) - c(\mu)\}.
\end{aligned}
$$

The analysis of systems with impatient customers is more complex. In Section 4.2.3, we assumed that a FIFO queueing discipline was enforced. This implied that $w_s, s = 0, \ldots, N-1$ did not depend on the pricing policy (recall that w_s denotes the the probability that a customer entering a system in state s will not renege and will leave upon service completion). But w_s clearly depends on the service rate, which is a decision variable now. Therefore, the Markov decision process formulation we used in Section 4.2.3 no longer holds as the reward obtained in state s now depends on the service rates set for other states. Under a monotonicity assumption for the service rate policy, it may be possible to

bound the optimal reward. Otherwise, the analysis of this problem might require the use of an asymptotic regime approximation.

5.2.3 Optimal Pricing with Multiple Service Requirements

In this dissertation, we assumed that the customers from all classes require one unit of service; that is, every customer occupies exactly one slot in the system. Paschalidis and Tsitsiklis [17] investigate the case where the customers' service requirements vary from one class to the other and the service provider may advertise class-specific prices. They determine optimal static prices as well as an optimal stationary policy (class-specific congestion-dependent prices) that maximize the long-run average reward. They also demonstrate that the optimal dynamic reward can be approximately matched by suitably chosen static prices. However, they do not consider any congestion penalty in their model. Hence, they do not capture customers' aversion to congestion or the loss of goodwill incurred by keeping customers in the system.

We can extend the three congestion models discussed in this dissertation by considering the dynamic precision pricing problem with class-specific service requirements. This would enable us to further distinguish customer classes through their required service times. The birth-death structure no longer holds in this case and the MDP analysis would require a policy or value iteration method. As opposed to Paschalidis and Tsitsiklis [17], it is not likely that static pricing would closely match the performance of dynamic pricing when congestion penalties are incurred. Since higher penalties are incurred in congested states, it is intuitive that an optimal congestion-dependent pricing policy significantly improves profit compared to an optimal static pricing policy. For instance, as we demonstrated in Section 4.2.1, there are instances where holding costs make some customer classes unprofitable in highly congested states. Unlike static pricing, the use of dynamic pricing enables the service

provider to refuse entry to these classes, further improving profit.

APPENDIX A

COMPLEMENTARY RESULTS FOR CHAPTER 3

Lemma A.0.1 *The function $\varphi_N(\cdot)$ is nondecreasing on $[0, \infty)$.*

Proof In what follows, all derivatives are with respect to ρ. For simplicity, we omit the arguments of the functions. Since φ_N is differentiable ,

$$\varphi'_N = -\frac{L''\pi'_0 - L'\pi''_0}{\pi'^2_0},$$

$$L = \pi_0 F, \text{ where } F(\rho) = \sum_{n=1}^{N} n\rho^n,$$

$$L' = \pi'_0 F + F'\pi_0,$$

$$L'' = \pi_0 F'' + 2F'\pi'_0 + F\pi''_0 \text{ and}$$

$$L''\pi'_0 - L'\pi''_0 = 2F'\pi'^2_0 + F''\pi_0\pi'_0 - F'\pi_0\pi''_0.$$

But we also have $\pi'_0 = -\pi_0^2 \frac{F}{\rho}$ and

$$\pi''_0 = -\frac{F'\pi_0^2}{\rho} + \frac{F\pi_0^2}{\rho^2} - 2\frac{F\pi_0\pi'_0}{\rho} = -\frac{F'\pi_0^2}{\rho} + \frac{F\pi_0^2}{\rho^2} + 2\frac{\pi'^2_0}{\pi_0}.$$

Since $F'\pi_0\pi''_0 = -\frac{F'^2\pi_0^3}{\rho} + \frac{FF'\pi_0^3}{\rho^2} + 2F'\pi'^2_0$,

$$\begin{aligned} L''\pi'_0 - L'\pi''_0 =& F''\pi_0\pi'_0 + \frac{F'^2\pi_0^3}{\rho} - \frac{FF'\pi_0^3}{\rho^2} = -\frac{FF''\pi_0^3}{\rho} + \frac{F'^2\pi_0^3}{\rho} - \frac{FF'\pi_0^3}{\rho^2}, \\ =& -\frac{\pi_0^3}{\rho}\left(FF'' + F'\left(-F' + \frac{F}{\rho}\right)\right) \text{ and} \\ -F' + \frac{F}{\rho} =& \sum_{n=1}^{N} n\rho^{n-1} - \sum_{n=1}^{N} n^2\rho^{n-1} = -\sum_{n=2}^{N} n(n-1)\rho^{n-1}. \end{aligned}$$

Therefore,

$$\begin{aligned}
L''\pi_0' - L'\pi_0'' &= -\frac{\pi_0^3}{\rho}\left(\sum_{n=1}^{N} n\rho^{n-1}\sum_{n=1}^{N} n^2(n-1)\rho^{n-1} - \sum_{n=1}^{N} n(n-1)\rho^{n-1}\sum_{n=1}^{N} n^2\rho^{n-1}\right), \\
&= -\frac{\pi_0^3}{\rho^3}\left(\sum_{n=1}^{N} n\rho^{n}\sum_{n=1}^{N} n^2(n-1)\rho^{n} - \sum_{n=1}^{N} n(n-1)\rho^{n}\sum_{n=1}^{N} n^2\rho^{n}\right), \\
&= -\frac{\pi_0^3}{\rho^3}\left(\sum_{k=2}^{2N}\rho^k \sum_{n=\max(0,k-N)}^{\min(N,k)} n(k-n)(k-n-1)(k-2n)\right).
\end{aligned}$$

Let $a_k = \sum_{n=\max(0,k-N)}^{\min(N,k)} n(k-n)(k-n-1)(k-2n)$. We will prove that a_k is positive. Using the fact that $\max(0,k-N) = k - \min(N,k)$,

$$a_k = \sum_{n=\frac{k}{2}-\min(N,k)}^{\min(k,N)-\frac{k}{2}} \left(\frac{k}{2}-n\right)\left(\frac{k}{2}+n\right)\left(\frac{k}{2}+n-1\right)2n = \sum_{n=\frac{k}{2}-\min(N,k)}^{\min(k,N)-\frac{k}{2}} b_{n,k}.$$

Since for all $0 \le n \le \min(k,N) - \frac{k}{2}$,

$$b_{n,k} = \left(\frac{k}{2}-n\right)\left(\frac{k}{2}+n\right)\left(\frac{k}{2}+n-1\right)2n \ge -b_{-n,k} = \left(\frac{k}{2}-n\right)\left(\frac{k}{2}+n\right)\left(\frac{k}{2}-n-1\right)2n,$$

we have $a_k \ge 0$. Thus, $L''\pi_0' - L'\pi_0'' \le 0$ and ${\varphi_N}' \ge 0$. □

Lemma A.0.2 *For $N \le \infty$, the function $\gamma_N^b(\cdot)$ is nonincreasing on $[0,\infty)$.*

Proof We prove that $\gamma_N^b(\cdot)$ is nonincreasing for $\rho \ge 0$ by showing that its derivative is nonpositive. In what follows, all derivatives are with respect to ρ. For simplicity, we omit the arguments of the functions. Recall that

$$\gamma_N^b = \frac{\sum_{n=0}^{N-1}(n+1)\rho^n P_n}{\sum_{n=0}^{N-1}\rho^n P_n \sum_{n=0}^{N}\rho^n P_{n-1}} = \frac{L}{\sum_{n=0}^{N-1}\rho^{n+1}P_n} = \frac{L}{\pi_0^{-1}-1},$$

where

$$L = \frac{\sum_{n=0}^{N-1}(n+1)\rho^n P_n}{\sum_{n=0}^{N}\rho^n P_{n-1}}.$$

We define the quantity F as the following polynomial of ρ: $F = \frac{L}{\pi_0} = \sum_{n=1}^{N} n\rho^n P_{n-1}$. Now let us take the derivative of γ_N^b. Note that if $N = \infty$, we can still interchange derivative

and summation signs, since we consider power series of ρ within the radius of convergence.

$$\begin{aligned}
(\pi_0^{-1}-1)^2{\gamma_N^b}' &= L'\left(\frac{1}{\pi_0}-1\right)+L\frac{\pi_0'}{\pi_0^2}\\
&= \left(\pi_0'F+F'\pi_0\right)\left(\frac{1}{\pi_0}-1\right)+F\frac{\pi_0'}{\pi_0}\\
&= \left(-\pi_0^2\frac{F^2}{\rho}+F'\pi_0\right)\left(\frac{1}{\pi_0}-1\right)-\pi_0\frac{F^2}{\rho}\\
&= -2\pi_0\frac{F^2}{\rho}+\pi_0^2\frac{F^2}{\rho}+F'(1-\pi_0)\\
&= \frac{\pi_0^2}{\rho}\left(F'\rho(\pi_0^{-1}-1)-F^2(2\pi_0^{-1}-1)\right)\\
&= \frac{\pi_0^2}{\rho}\left(\sum_{n=1}^N n^2\rho^nP_{n-1}\sum_{n=1}^N\rho^nP_{n-1}-\left(\sum_{n=1}^N n\rho^nP_{n-1}\right)^2\left(2\sum_{n=1}^N\rho^nP_{n-1}+1\right)\right)
\end{aligned}$$

To analyze the sign of the quantity above, we break it down into two quantities:

$$\begin{aligned}
A &= \sum_{n=1}^N n^2\rho^nP_{n-1}\sum_{n=1}^N\rho^nP_{n-1}-(\sum_{n=1}^N n\rho^nP_{n-1})^2 \text{ and}\\
B &= 2\left(\sum_{n=1}^N n\rho^nP_{n-1}\right)^2\sum_{n=1}^N\rho^nP_{n-1}, \text{ so that}\\
{\gamma_N^b}' &= \frac{\pi_0^2}{\rho}\frac{A-B}{(\pi_0^{-1}-1)^2}.
\end{aligned}$$

It remains to show that $A\le B$.

Note that

$$\begin{aligned}
\sum_{n=1}^N n^2\rho^nP_{n-1}\sum_{n=1}^N\rho^nP_{n-1} &= \sum_{k=2}^{2N}\rho^k\sum_{n=(k-N)\vee1}^{N\wedge(k-1)} n^2P_{n-1}P_{k-n-1} \text{ and}\\
\left(\sum_{n=1}^N n\rho^nP_{n-1}\right)^2 &= \sum_{k=2}^{2N}\rho^k\sum_{n=(k-N)\vee1}^{N\wedge(k-1)} n(k-n)P_{n-1}P_{k-n-1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
A &= \sum_{k=2}^{2N}\rho^k\sum_{n=(k-N)\vee1}^{N\wedge(k-1)} n(2n-k)P_{n-1}P_{k-n-1} &(22)\\
&\le \sum_{k=3}^{2N}\rho^k\sum_{n=(k-N)\vee2}^{N\wedge(k-1)} n(2n-k)P_{n-1}P_{k-n-1}. &(23)
\end{aligned}$$

Inequality (23) holds since when $k = 2$, we have

$$\sum_{n=(k-N)\vee 1}^{N\wedge(k-1)} n(2n-k)P_{n-1}P_{k-n-1} = \sum_{n=1}^{1} n(2n-k)P_{n-1}P_{k-n-1} = 0.$$

Furthermore, when $n = 1$,

$$n(2n-k)P_{n-1}P_{k-n-1} = 1(2(1)-k)P_0P_{k-2} \leq 0.$$

On the other hand, we have

$$\begin{aligned}
B &= \left(\sum_{n=1}^{N} n\rho^n P_{n-1}\right)^2 2\sum_{n=1}^{N} \rho^n P_{n-1} && (24)\\
&= \sum_{k=3}^{3N} \rho^k \sum_{n=(k-N)\vee 2}^{2N\wedge(k-1)} \sum_{j=(n-N)\vee 1}^{N\wedge(n-1)} 2j(n-j)P_{j-1}P_{n-j-1}P_{k-n-1} && (25)\\
&\geq \sum_{k=3}^{3N} \rho^k \sum_{n=(k-N)\vee 2}^{N\wedge(k-1)} P_{n-1}P_{k-n-1} \sum_{j=1}^{n-1} 2j(n-j) && (26)\\
&\geq \sum_{k=3}^{2N} \rho^k \sum_{n=(k-N)\vee 2}^{N\wedge(k-1)} \frac{n(n+1)(n-1)}{3} P_{n-1}P_{k-n-1}. && (27)
\end{aligned}$$

Since $\{p_s\}$ is nonincreasing, $P_{j-1}P_{n-j-1} \geq P_{n-1}$, for all $j = (n-N)\vee 1, \ldots, N\wedge(n-1)$. Hence, inequality (26) holds.

To compare (23) and (27), it remains to show that for $2 \leq n \leq k-1$, $\frac{n(n+1)(n-1)}{3} \geq n(2n-k)$. Since $2 \leq n \leq k-1$, we have

$$n(2n-k) \leq n(n-1) \leq \frac{n(n+1)(n-1)}{3}.$$

Therefore, the right-hand side of (27) is greater than or equal to the right-hand side of (23) implying that $A \leq B$. Hence, ${\gamma_N^b}'(\cdot) \leq 0$ and $\gamma_N^b(\cdot)$ is nonincreasing □

Lemma A.0.3 *The function $\gamma_N^r(\cdot)$ is nonincreasing on $[0, \infty)$.*

Proof Recall that

$$\gamma_N^r(\rho) = \frac{\sum_{n=0}^{N-1}(n+1)\rho^n Q_n}{\sum_{n=0}^{N-1}\rho^n Q_n \sum_{n=0}^{N}\rho^n Q_{n-1}},$$

where $Q_n = \prod_{s=0}^{n} \frac{\mu}{\mu+s\theta}$ for $n = 0, \ldots, N-1$ and $Q_{-1} = 1$. Note that $\gamma_N^r(\cdot)$ has the same expression as $\gamma_N^r(\cdot)$, where $\{Q_n\}$ is substituted for $\{P_n\}$. To show that $\gamma_N^r(\cdot)$ is non increasing, we use Lemma A.0.2 after verifying that $\{Q_n\}$ and $\{P_n\}$ have the same properties. In Lemma A.0.2, we only use the fact that $\{P_n\}$ and $\{p_s\}$ are nonincreasing sequences. Similarly, $\{Q_n\}$ and $\{\frac{\mu}{\mu+s\theta}\}$ are nonincreasing sequences. Therefore, $\gamma_N^r(\cdot)$ is nonincreasing on $[0, \infty)$. □

Lemma A.0.4 *Consider an $M/M/1/N$ queue with holding costs. Then, for all $\rho \geq 0$,*

1. $\gamma_N^h(\rho) \geq \pi_0(\rho, N)$,
2. $1 - \pi_N(\rho, N) \geq \gamma_N^h(\rho)$,
3. $N - L(\rho, N) \geq \pi_0(\rho, N)\varphi_N(\rho)$,
4. $\rho\varphi_N(\rho)(1 - \pi_N(\rho, N)) \geq L(\rho, N)$,

where

$$\varphi_N(\rho) = -\frac{\frac{\partial L(\rho,N)}{\partial \rho}}{\frac{\partial \pi_0(\rho,N)}{\partial y}} = \begin{cases} \frac{1-(N+1)^2\rho^N(1+\rho^2)+2N(N+2)\rho^{N+1}+\rho^{2N+2}}{(1-\rho)^2(1-(N+1)\rho^N+N\rho^{N+1})} & \text{if } \rho \neq 1, \\ \frac{1}{6}N^2 + \frac{1}{2}N + \frac{1}{3} & \text{if } \rho = 1, \end{cases}$$

and

$$\gamma_N^h(\rho) = \begin{cases} \frac{1+N\rho^{N+1}-(N+1)\rho^N}{(1-\rho^{N+1})(1-\rho^N)} & \text{if } \rho \neq 1, \\ \frac{1}{2} & \text{if } \rho(y) = 1. \end{cases}$$

Proof For clarity, we will omit the arguments of the quantities we use in this proof. For instance, we will write π_0 instead of $\pi_0(\rho, N)$. All derivatives in this proof are with respect to ρ.

1. We prove first that $\gamma_N^h \geq \pi_0$. When $\rho = 1$, $\pi_0 = \frac{1}{N+1}$ and $\gamma_N^h = 1/2$, which agrees with our claim. Otherwise, recall that $\pi_0 = \frac{1-\rho}{1-\rho^{N+1}}$. Therefore,

$$\begin{aligned}
\frac{\pi_0}{\gamma_N^h} &= \frac{(1-\rho)(1-\rho^N)}{1+N\rho^{N+1}-(N+1)\rho^N} \\
&= \frac{(1-\rho)^2 \sum_{k=0}^{N-1} \rho^k}{(1-\rho)(\sum_{k=0}^{N-1} \rho^k - N\rho^N)} \\
&= \frac{(1-\rho) \sum_{k=0}^{N-1} \rho^k}{-\sum_{k=0}^{N-1}(1-\rho^k) + N(1-\rho^N)} \\
&= \frac{\sum_{k=0}^{N-1} \rho^k}{-\sum_{k=0}^{N-1} \sum_{s=0}^{k-1} \rho^s + N \sum_{s=0}^{N-1} \rho^s} \\
&= \frac{\sum_{k=0}^{N-1} \rho^k}{-\sum_{s=0}^{N-2} \rho^s (N-s-1) + N \sum_{s=0}^{N-1} \rho^s} \\
&= \frac{\sum_{k=0}^{N-1} \rho^k}{\sum_{s=0}^{N-2} \rho^s (s+1) + N\rho^{N-1}} \\
&= \frac{\sum_{k=0}^{N-1} \rho^k}{\sum_{s=0}^{N-1} \rho^s + \sum_{s=0}^{N-2} s\rho^s + (N-1)\rho^{N-1}} \\
&\leq 1.
\end{aligned}$$

2. Let us prove now that $1 - \pi_N \geq \gamma_N^h$, which is easily verified when $\rho = 1$. Now suppose $\rho \neq 1$. Since $1 - \pi_N = \frac{1-\rho^N}{1-\rho^{N+1}}$, we have

$$\begin{aligned}
\frac{1-\pi_N}{\gamma_N^h} - 1 &= \frac{(1-\rho^N)^2}{1+N\rho^{N+1}-(N+1)\rho^N} - 1 \\
&= \frac{1-2\rho^N+\rho^{2N}-1-N\rho^{N+1}+(N+1)\rho^N}{1+N\rho^{N+1}-(N+1)\rho^N} \\
&= \frac{\rho^N(\rho^N - N\rho + (N-1))}{1+N\rho^{N+1}-(N+1)\rho^N}.
\end{aligned}$$

We know from Ziya [24] that

$$1 + N\rho^{N+1} - (N+1)\rho^N \geq 0 \text{ and}$$

$$\begin{aligned}\rho^N - N\rho + (N-1) &= (\rho^N - 1) - N(\rho - 1) \\ &= (\rho - 1)\sum_{k=0}^{N-1}(\rho^k - 1) \\ &\geq 0.\end{aligned}$$

Therefore, $1 - \pi_N \geq \gamma_N^h$.

3. We need to show that $N - L \geq \pi_0\varphi_N$. When $\rho = 1$, we have $N - L = \frac{N}{2}$, $\pi_0 = \frac{1}{N+1}$ and $\varphi_N = \frac{1}{6}N^2 + \frac{1}{2}N + \frac{1}{3}$. So, $\frac{N-L}{\pi_0} = \frac{1}{2}N^2 + \frac{1}{2}N \geq \varphi_N$. Now suppose $\rho \neq 1$. We have $L = \pi_0 \sum_{n=0}^{N} n\rho^n = \pi_0 F$, where $F = \sum_{n=0}^{N} n\rho^n$. Recall from [14] that $\varphi_N = -\frac{L'}{\pi_0'}$. Using the fact that $\pi_0' = -\pi_0^2 \frac{F}{\rho}$,

$$\begin{aligned}\pi_0\varphi_N &= \pi_0 \frac{\pi_0' F + \pi_0 F'}{-\pi_0'} \\ &= -L - \frac{\pi_0^2}{\pi_0'} F' \\ &= -L + \frac{F'\rho}{F} \\ &= -L + \frac{\sum_{k=0}^{N} k^2\rho^k}{F} \\ &\leq N - L.\end{aligned}$$

4. It remains to prove that $\rho\varphi_N(1 - \pi_N) \geq L$. When $\rho = 1$, we have

$$\rho\varphi_N(1 - \pi_N) = \frac{N}{N+1}\left(\frac{1}{6}N^2 + \frac{1}{2}N + \frac{1}{3}\right) \geq \frac{N}{2} = L.$$

Now suppose $\rho \neq 1$. In the same fashion, $\varphi_N = -F - \frac{\pi_0}{\pi_0'}F'$. Moreover, note that

$\rho(1-\pi_N) = 1-\pi_0$. So,

$$\begin{aligned}
\rho\varphi_N(1-\pi_N) - L &= \varphi_N(1-\pi_0) - L, \\
&= \left(-F - \frac{\pi_0}{\pi_0'}F'\right)(1-\pi_0) - L \\
&= -F - \frac{\pi_0}{\pi_0'}F'(1-\pi_0) \\
&= -F + \rho^2\frac{F'}{F}\frac{1-\rho^N}{1-\rho} \\
&= -F + \rho^2\frac{F'}{F}\sum_{k=0}^{N-1}\rho^k.
\end{aligned}$$

Note that this quantity has the same sign as $-F^2 + \rho^2 F' \sum_{k=0}^{N-1}\rho^k$ and that

$$\begin{aligned}
-F^2 + \rho^2 F'\sum_{k=0}^{N-1}\rho^k &= -(\sum_{k=1}^{N} k\rho^k)^2 + \sum_{k=1}^{N} k^2\rho^k \sum_{k=1}^{N}\rho^k \\
&= \sum_{k=2}^{2N}\rho^k \sum_{n=\max(0,k-N)}^{\min(N,k)} n(n-k) + n^2 \\
&= \sum_{k=2}^{2N}\rho^k \sum_{n=\max(0,k-N)}^{\min(N,k)} n(2n-k) \\
&= \sum_{k=2}^{2N}\rho^k \sum_{n=\frac{k}{2}-\min(N,k)}^{\min(N,k)-\frac{k}{2}} (\frac{k}{2}+n)2n.
\end{aligned}$$

Let $b_{n,k} = (\frac{k}{2}+n)2n$ for $\frac{k}{2}-\min(N,k) \le n \le \min(N,k)-\frac{k}{2}$. For $0 \le n \le \min(N,k)-\frac{k}{2}$, we have

$$b_{n,k} = (\frac{k}{2}+n)2n \ge (\frac{k}{2}-n)2n = -b_{-n,k}.$$

Therefore,

$$\sum_{n=\frac{k}{2}-\min(N,k)}^{\min(N,k)-\frac{k}{2}} b_{n,k} = \sum_{n=\frac{k}{2}-\min(N,k)}^{\min(N,k)-\frac{k}{2}} (\frac{k}{2}+n)2n \ge 0.$$

Hence, we can claim that $\rho\varphi_N(1-\pi_N) \ge L$ and the proof is complete.

□

Lemma A.0.5 *Consider an $M/M/1/N$ queue under static price $y \ge 0$ in the model with balking customers. Then, the following holds:*

1. $\gamma_N^b(\rho(y)) \geq \pi_0(\rho(y), N)$

2. $\pi_0(\rho(y), N) \sum_{n=0}^{N-1} \rho(y)^n P_n \geq \gamma_N^b(\rho(y))$

Consider an $M/M/1/N$ queue under static price $y \geq 0$ in the model with impatient customers. Then, the following holds:

1. $\gamma_N^r(\rho(y)) \geq \pi_0(\rho(y), N)$

2. $\pi_0(\rho(y), N) \sum_{n=0}^{N-1} \rho(y)^n Q_n \geq \gamma_N^r(\rho(y))$

Proof We will only prove the result for the model with balking customers. In the model with impatient customers, the result is proved in the exact same fashion by substituting the sequence $\{Q_n\}$ for $\{P_n\}$.

Recall that

$$\gamma_N^b(\rho(y)) = \frac{\sum_{n=0}^{N-1}(n+1)\rho(y)^n P_n}{\sum_{n=0}^{N-1} \rho(y)^n P_n \sum_{n=0}^{N} \rho(y)^n P_{n-1}} = \pi_0(\rho(y), N) \frac{\sum_{n=0}^{N-1}(n+1)\rho(y)^n P_n}{\sum_{n=0}^{N-1} \rho(y)^n P_n}.$$

Therefore, $\gamma_N^r(\rho(y)) \geq \pi_0(\rho(y), N)$.

To prove that $\pi_0(\rho(y), N) \sum_{n=0}^{N-1} \rho(y)^n P_n \geq \gamma_N^b(\rho(y))$, we only need to show that

$$\left(\sum_{n=0}^{N-1} \rho(y)^n P_n \right)^2 \geq \sum_{n=0}^{N-1} (n+1)\rho(y)^n P_n.$$

We have

$$\begin{aligned}
\left(\sum_{n=0}^{N-1} \rho(y)^n P_n \right)^2 &= \sum_{k=0}^{2N-2} \rho(y)^k \sum_{n=(k-N+1)\vee 0}^{k\wedge(N-1)} P_n P_{k-n} \\
&\geq \sum_{k=0}^{N-1} \rho(y)^k \sum_{n=0}^{k} P_n P_{k-n} \\
&\geq \sum_{k=0}^{N-1} (k+1)\rho(y)^k P_k.
\end{aligned}$$

□

APPENDIX B

COMPLEMENTARY RESULTS FOR CHAPTER 4

Lemma B.0.6 *Let $I = 1$. Then, for $s = 0, \ldots, N-1$, we have:*

1. *in systems with holdings costs:*

$$z_{1,s}^* = \begin{cases} \sup\{z : (g_N^* + h_s - \mu_s G(s-1))r_1(z) \leq \lambda_1(z)\} & \text{if } g_N^* + h_s > \mu_s G(s-1) \\ \beta_1 & \text{otherwise} \end{cases}$$

2. *in systems with balking customers:*

$$z_{1,s}^* = \sup\{z : (g_N^* - \mu_s G(s-1))r_1(z) \leq \lambda_1(z)\},$$

3. *in systems with impatient customers:*

$$z_{1,s}^* = \sup\{z : (g_N^* - (\mu_s + (s-q)^+\theta)G(s-1))r_1(z) \leq w_s\lambda_1(z)\}.$$

Proof First, we prove the result for systems with holding costs. Since we only have one customer class, we omit the class subscript in the following. In Theorem 4.2.1, we proved that $z_s^* = \inf\{z : r(z)(z - G(s)) \geq 1\}$. Now suppose $g_N^* + h_s > \mu_s G(s-1)$. Therefore, $z_s^* < \beta$. It is straightforward to show that $\sup\{\lambda(z)(z - G(s)) - g_N^* - h_s + \mu_s G(s-1)\} = 0$ is equivalent to $\sup\{z - G(s) - \frac{g_N^* + h_s - \mu_s G(s-1)}{\lambda(z)}\} = 0$, where z_s^* is the unique price that attains the supremum. Let

$$t_s(z) = z - G(s) - \frac{g_N^* + h_s - \mu_s G(s-1)}{\lambda(z)},$$

$$t_s'(z) = 1 - (g_N^* + h_s - \mu_s G(s-1))\frac{r(z)}{\lambda(z)}, \text{ a.e. on } [\alpha, \beta].$$

Under IGHR, Proposition 5.1 of Ziya [23] shows that $t_s'(\cdot)$ is strictly decreasing almost everywhere on $(\inf\{z : zr(z) \geq 1\}, \beta)$, which includes (z_s^*, β). Hence, $t_s'(\cdot) < 0$ almost everywhere on (z_s^*, β). Therefore,

$$z_s^* = \sup\{z : t_s'(z) \geq 0\} = \sup\{z : (g_N^* + h_s - \mu_s G(s-1))r(z) \leq \lambda(z)\}.$$

Suppose that $g_N^* + h_s \geq \mu_s G(s-1)$. Since we always have $g_N^* + h_s \leq \mu_s G(s-1)$, we can claim that $g_N^* + h_s = \mu_s G(s-1)$ and $\sup\{(z - G(s))\lambda(z)\} = 0$. We have two possibilities: $z_s^* = \beta$ or $z_s^* = G(s) < \beta$. The latter is impossible since there must exist $\epsilon > 0$ such that $G(s) + \epsilon < \beta$ and $\epsilon\lambda(G(s) + \epsilon) > 0$. Therefore, $z_s^* = \beta$.

In systems with balking customers and systems with impatient customers, the proof only requires minor modifications and is omitted. □

Proof of Propositions 4.2.6 and 4.2.10:

We use the same method as in the proof of Proposition 4.2.2. We prove Proposition 4.2.6 first. It is clear that the optimal dynamic reward g_N^* is nondecreasing in N. To show that $z_{i,s,2}^* \leq z_{i,s,1}^*$, we will prove that $G_2(s) \leq G_1(s)$ for all $s = 0, \ldots, N-1$ in each case.

Suppose that $G_2(s) > G_1(s)$ for some state $s = 0, \ldots, N-1$. We have

$$p_s \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G_2(s))\} < p_s \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G_1(s))\} \text{ or} \tag{28}$$

$$\sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G_2(s))\} = \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G_1(s))\} = 0. \tag{29}$$

Equality (29) is impossible since it implies that $z_{i,s,1}^* = z_{i,s,2}^* = \beta_i$ for all i, which contradicts Lemma 4.2.4. Inequality (28) implies that

$$\mu_s G_2(s-1) = g_{N+1}^* - \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G_2(s))\} \tag{30}$$

$$> g_N^* - \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - G_1(s))\} > \mu_s G_1(s-1). \tag{31}$$

By induction , $0 = G_2(-1) > G_1(-1) = 0$, which yields a contradiction. Therefore, for all $s = 0, \ldots, N-1$, $G_2(s) \leq G_1(s)$ and consequently, $z^*_{i,s,2} \leq z^*_{i,s,1}$.

To prove Proposition 4.2.10, we repeat the proof from above substituting

$$\sum_{i=1}^{I} \sup\{\lambda_i(z)(w_s z - G_2(s))\} < \sum_{i=1}^{I} \sup\{\lambda_i(z)(w_s z - G_1(s))\} \text{ or} \tag{32}$$

$$\sum_{i=1}^{I} \sup\{\lambda_i(z)(w_s z - G_2(s))\} = \sum_{i=1}^{I} \sup\{\lambda_i(z)(w_s z - G_1(s))\} = 0 \tag{33}$$

for (28) and (29) and

$$\begin{aligned}(\mu_s + (s-q)^+\theta)G_2(s-1) &= g^*_{N+1} - \sum_{i=1}^{I} \sup\{\lambda_i(z)(w_s z - G_2(s))\} \\ &> g^*_N - \sum_{i=1}^{I} \sup\{\lambda_i(z)(w_s z - G_1(s))\} \\ &> (\mu_s + (s-q)^+\theta)G_1(s-1).\end{aligned}$$

for (30) and (31). Note that in this case, equality (33) is made impossible by Lemma 4.2.5.

□

Lemma B.0.7 *For all $g > 0$, $\Psi^h(\cdot, g)$ is a continuous nondecreasing contraction mapping from R to $(-\infty, \frac{g+h_N}{\mu_N}]$ and has a unique fixed point denoted by $FP^h(g)$. Moreover, $FP^h(\cdot)$ is increasing and continuous on $[0, \infty)$.*

For all $g > 0$, $\Psi^b(\cdot, g)$ is a continuous nondecreasing contraction mapping from R to $(-\infty, \frac{g}{\mu_N}]$ and has a unique fixed point denoted by $FP^b(g)$. Moreover, $FP^b(\cdot)$ is increasing and continuous on $[0, \infty)$.

Proof We will only prove the result for Ψ^h. The proof for Ψ^b is similar and is omitted. It is clear that $\Psi^h(V, g)$ is nondecreasing in V and continuity can be proven as $\Psi^h(V, g)$ depends on V through the supremum of a bounded continuous function of z.

We will now prove that $\Psi^h(\cdot, g)$ is a contraction. Suppose that $V_1 \leq V_2$. Then,

$$\begin{aligned}\mu_N(\Psi^h(V_2, g) - \Psi^h(V_1, g)) &= \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - V_1)\} - \sum_{i=1}^{I} \sup\{\lambda_i(z)(z - V_2)\} \\ &\leq \sum_{i=1}^{I} \lambda_i(z_i(V_1))(z_i(V_1) - V_1) - \lambda_i(z_i(V_1))(z_i(V_1) - V_2), \\ &\leq (V_2 - V_1) \sum_{i=1}^{I} \Lambda_i\end{aligned}$$

where $z_i(V)$ is the unique maximizer in $[\alpha_i, \beta_i]$ of $\lambda_i(z)(z - V)$. Therefore, $\Psi^h(V_2, g) - \Psi^h(V_1, g) \leq \frac{\sum_{i=1}^{I} \Lambda_i}{\mu_N}(V_2 - V_1)$, which proves that $\Psi^h(\cdot, g)$ is a contraction mapping and has a unique fixed point.

It remains to show that $FP^h(\cdot)$ is increasing and continuous. Let $0 \leq g_1 < g_2$. Therefore, $\Psi^h(FP^h(g_2), g_1) < \Psi(FP^h(g_2), g_2) = FP^h(g_2)$, so $FP^h(g_1) < FP^h(g_2)$, proving that $FP^h(\cdot)$ is increasing. As $\Psi^h(\cdot, g) \leq \frac{g + h_N}{\mu_N}$, we also have $FP^h(g) \leq \frac{g + h_N}{\mu_N}$.

To prove that $FP^h(\cdot)$ is continuous, we show by contradiction that it is both left-continuous and right-continuous. Let $g_n \uparrow g$ such that $g_n \geq 0$ for all n. Therefore, $FP^h(g_n)$ has a limit $\lim_n FP^h(g_n) \leq FP^h(g)$. Suppose that $\lim_n FP^h(g_n) < FP^h(g)$. Hence, $\lim_n FP^h(g_n) < \Psi(\lim_n FP^h(g_n), g)$. But $\lim_n FP^h(g_n) \geq FP^h(g_m), \forall m \geq 0$, so $\lim_n FP^h(g_n) > \Psi(\lim_n FP^h(g_n), g_m)$. As m goes to infinity, we have $\lim_n FP^h(g_n) \geq \Psi(\lim_n FP^h(g_n), g)$, yielding a contradiction. Hence, $FP^h(\cdot)$ is left-continuous.

In the same fashion, let $g_n \downarrow g$, such that $g_n \geq 0$ for all n. Therefore $FP^h(g_n)$ has a limit $\lim_n FP^h(g_n) \geq FP^h(g)$. Suppose that $\lim_n FP^h(g_n) > FP^h(g)$. Hence, $\lim_n FP^h(g_n) > \Psi^h(\lim_n FP^h(g_n), g)$. However, $\lim_n FP^h(g_n) \leq FP^h(g_m), \forall m \geq 0$, which implies that $\lim_n FP^h(g_n) < \Psi^h(\lim_n FP^h(g_n), g_m)$. Letting m go to infinity, we note that $\lim_n FP^h(g_n) \leq \Psi^h(\lim_n FP^h(g_n), g)$, yielding a contradiction. Therefore, $FP^h(\cdot)$ is continuous on $[0, \infty)$. □

REFERENCES

[1] Aktaran, T. and Ayhan, H., "Sensitivity of optimal prices to system parameters in a service facility, working paper," 2005.

[2] Ata, B. and Shneorson, S., "Dynamic control of an M/M/1 service system with adjustable arrival and service rates, under review," 2004.

[3] Çil, E. B., Karaesmen, F., and Örmeci, E. L., "Sensitivity analysis on a dynamic pricing problem of an m/m/c queuing system, under review,"

[4] George, J. M. and Harrison, J. M., "Dynamic control of a queue with adjustable service rate," *Operations Research*, vol. 49, pp. 720–731, 2001.

[5] Gross, D. and Harris, C., *Fundamentals of Queueing Theory*. John Wiley and Sons, New York, 1985.

[6] Hassin, R., "Consumer information in markets with random product quality : the case of queues and balking," *Econometrica*, vol. 54, pp. 1185–1195, 1986.

[7] Ittig, P. T., "Planning service capacity when demand is sensitive to delay," *Decision Sciences*, vol. 25, pp. 541–559, 1994.

[8] Ittig, P. T., "The real cost of making customers wait," *International Journal of Service Industry Management*, vol. 13, pp. 231–241, 2002.

[9] Knudsen, N. C., "Individual and social optimization in a multiserver queue with a general cost-benefit structure," *Econometrica*, vol. 40, pp. 515–528, 1972.

[10] Larsen, C., "Investigating sensitivity and the impact of information on pricing decisions in an M/M/1/∞ queueing model," *International Journal of Production Economics*, vol. 56-57, pp. 365–377, 1998.

[11] Lasserre, J.-B. and Hernández-Lerma, O., *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer-Verlag, 1996.

[12] Low, D. W., "Optimal dynamic pricing policies for an M/M/s queue," *Operations Research*, vol. 22, pp. 545–561, 1974.

[13] Low, D. W., "Optimal pricing for an unbounded queue," *IBM Journal of Research and Development*, vol. 18, pp. 290–302, 1974.

[14] Maoui, I., Ayhan, H., and Foley, R. D., "Optimal static pricing for a service facility with holding costs," under review.

[15] Mendelson, H. and Whang, S., "Optimal incentive-compatible priority pricing for the M/M/1 queue," *Operations Research*, vol. 38, pp. 870–883, 1990.

[16] Naor, P., "The regulation of queue size by levying tolls," *Econometrica*, vol. 37, pp. 15–24, 1969.

[17] Paschalidis, I. C. and Tsitsiklis, J. N., "Congestion-dependent pricing of network services," *IEEE/ACM Transactions on Networking*, vol. 8, pp. 171–184, 2000.

[18] Puterman, M., *Markov Decision Processes.* Wiley-Interscience, 1994.

[19] Stidham, S., "Optimal control of admission to a queueing system," *IEEE Transactions on Automatic Control*, vol. 30, pp. 705–713, 1985.

[20] Weber, R. R. and Stidham, S., "Optimal control of service rates in network queues," *Advances in Applied Probability*, vol. 19, pp. 202–218, 1987.

[21] Yechiali, U., "On optimal balking rules and toll charges in the G/M/1/s queue," *Operations Research*, vol. 19, pp. 349–370, 1971.

[22] Ziya, S., *Optimal pricing for a service facility.* PhD thesis, Georgia Institute of Technology, 2003.

[23] Ziya, S., Ayhan, H., and Foley, R. D., "Relationships among three assumptions in revenue management," *Operations Research*, vol. 52, pp. 804–809, 2004.

[24] Ziya, S., Ayhan, H., and Foley, R. D., "Optimal prices for finite capacity queueing systems," *Operations Research Letters*, to appear.