The optimal admission threshold in observable queues with state dependent pricing

Christian Borgs, Jennifer T. Chayes
Microsoft Research, {borgs, jchayes}@microsoft.com

Sherwin Doroudi
Tepper School of Business, Carnegie Mellon University, sdoroudi@andrew.cmu.edu

Mor Harchol-Balter
School of Computer Science, Carnegie Mellon University harchol@cs.cmu.edu

Kuang Xu
Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, kuangxu@mit.edu

We consider the social welfare model of Naor [20] and revenue-maximization model of Chen and Frank [7], where a single class of delay-sensitive customers seek service from a server with an observable queue, under state dependent pricing. It is known that in this setting both revenue and social welfare can be maximized by a threshold policy, whereby customers are barred from entry once the queue length reaches a certain threshold. However, no explicit expression for this threshold has been found. This paper presents the first derivation of the optimal threshold in closed form, and a surprisingly simple formula for the (maximum) revenue under this optimal threshold. Utilizing properties of the Lambert W function, we also provide explicit scaling results of the optimal threshold as the customer valuation grows. Finally, we present a generalization of our results, allowing for settings with multiple servers.

Key words: Queues, optimization, pricing, admission threshold, Lambert W function, multiserver systems

1. Introduction

We consider a setting similar to that in the work of Naor [20] and Chen and Frank [7], wherein a firm is selling a service to stochastically arriving customers. The customers gain utility from receiving the service, but this utility decreases as the delay experienced by customers increases. The goal of the firm is to price its service and choose an admission policy so as to maximize its earning rate, or, alternatively to maximize social welfare, that is, the rate at which the customers earn surplus and the firm earns revenue. As in the work of the aforementioned authors, we consider the *observable queue* setting, where customers can observe the current state of the queue when deciding whether to join the queue. Such settings arise naturally when the firm cannot hide customer orders or arise artificially because observable queues allow the firm greater flexibility in pricing.

Customers are assumed to be identical in the value they obtain from receiving service (their reservation value), V, and their delay cost per unit time, c. Customers arrive according to a Poisson process with rate λ and have independent and identically exponentially distributed service requirements with rate parameter μ. A customer is willing to join the queue and pay for service only if she expects that total costs—the sum of the price charged by the firm and the cost of waiting for service—will be no more than the value obtained from service.

Assume that the firm charges $p(n)$ when there are $n \geq 0$ customers in the system. When the queue is shorter, customers are willing to pay more to join the queue, so $p(n)$ should be nonincreasing in n. The revenue-maximizing prices also maximize social welfare, as these prices allow for the firm to extract all customer surplus. All other (lesser) prices—so long as they induce the same customer entry behavior as in the revenue-maximizing case—would also maximize social welfare, as this would simply be a direct transfer of surplus from the firm to the customers.

Prior work in this area began with Naor [20], who proposed an observable queueing model, where the firm charges a fixed price in order to maximize revenue. Naor was able to find the fixed price resulting in the optimal queue length. Chen and Frank [7] generalized Naor's model by allowing for state dependent pricing. They found that the earning rate is maximized by charging prices that are

as high as possible in each state while imposing an appropriate threshold state, such that arrivals may not join the queue beyond this threshold state; they do not explicitly derive this threshold. Naor also observed the importance of using an optimal threshold, this time, in the context of maximizing social welfare. Naor finds the optimal threshold numerically, but does not provide an explicit formula for the optimal threshold. A full review of the prior literature is presented in Section 2.

It is known that imposing a threshold will, in general, lead to a higher earning rate (or total surplus). While imposing a threshold causes some customers to be turned away (forgoing an opportunity to earn revenue), a threshold is nonetheless desirable as it forces the queue to stay short, maintaining high prices. Clearly, if the queue is sufficiently long, customers are unwilling to join the queue even if it is free, suggesting that there always exists a threshold that is *preferable to using no threshold at all*. However, the question remains as to what the *best* choice of threshold should be.

Although the advantages of thresholds in this setting have been understood for some time—and it has been possible to compute the optimal threshold numerically—to date, an explicit closed form solution has eluded researchers. Our primary contribution is deriving this threshold in closed form and expressing it in terms of the customer parameters V and c, and the queuing parameters λ and μ The formula we derive uses the well understood, but non-elementary, Lambert W (product logarithm) function.

Section 3 describes our model, while the full contributions of this paper are summarized in Section 4. In Section 5, we derive a closed form for the revenue-maximizing and socially optimal threshold. This is a result we have reported more briefly in [5]. In Section 6.1, we show examples where the choice of threshold is significant in that implementing the optimal threshold can lead to much higher gains in revenue than implementing a suboptimal threshold. In Section 6.2, we leverage this closed-form result in order to derive a strikingly simple approximation for the total surplus (earning rate) at the optimal threshold. In Section 6.3, we use these results to understand the asymptotic behavior of the optimal threshold and the earning rate as customer valuations

tend toward infinity. Finally, in Section 7, we consider an extension where the firm has multiple servers (i.e., an M/M/s queueing system) and examine the impact of the number of servers, s, on the optimal threshold and the optimal revenue. We again derive a closed-form expression for the optimal cutoff in this setting, along with an analogous approximation for the earning rate and asymptotic results for the optimal threshold.

2. Literature Review

This paper falls within the area of queueing with incentives. This research stream, described in detail in a book by Hassin and Haviv [15], divides into two broad groups: (i) models with observable queues and (ii) unobservable queues. Our work sits within the former category. Within both queueing paradigms, the primary objective is to either maximize social welfare or revenue (profit). In both settings, the system designer sets prices, admission policies, and/or scheduling policies.

2.1. Unobservable queues

Although we assume observable queues, we begin by describing the relevant literature for unobservable queues. A queue is said to be *unobservable* if an arrival cannot see the state of the queue (i.e., number of customers in the queue) at the time of arrival (and therefore, must make decisions based on average cases). Literature in the area of unobservable queues is primarily concerned with heterogenous user types, as opposed to our work, where customers are homogenous. In such models pricing is not only used as a means for earning revenue, but also as a means to distinguish customer type and prioritize them appropriately, in both the social welfare maximization and revenue maximization settings. Examples of papers in the area of social welfare welfare maximization with unobservable queues include the work of Ghanem [12] and Mendelson [19], while examples of papers in the area of revenue maximization with unobservable queues include the work of Plambeck [21] and Afèche [1]. In related work, rather than assuming having customers pay prices set by the system designer, Kittsteiner [16], assumes that customers bid for priority service.

2.2. Observable queues

A queue is said to be *observable* if an arrival can see how many customers are in the queue. Typically, papers with observable queueing models assume a single homogenous class of customers with known values and delay sensitivities. When maximizing revenue, observable queues may be preferable to their unobservable counterparts, as they allow the firm to charge higher prices when customers arrive to shorter (or empty) queues.

The earliest work fitting within this category, is the aforementioned paper by Naor [20], which considers both revenue maximization and social welfare optimization. In Naor's revenue-maximization model, prices are fixed. Knudsen extends Naor's model to one where the firm has multiple servers available [17]. Knudsen finds that an optimal threshold exists, but does not give this threshold in closed form; we also present closed-form expressions for the optimal cutoff in this setting. Other early work in this area, Balachandran [3] considers a model in which customers can purchase priority after observing the queue, while Alperstein [2] considers a similar model where a firm is able to extract all consumer surplus. These papers depart from Naor's fixed pricing framework, but price differentiation is introduced through priority classes chosen by customers, rather than depending directly on the state of the system.

Chen and Frank [7] introduce the notion of state dependent pricing for observable queues (i.e., settings where the firm can charge a different price depending on the number of customers in the system). They find that in order to maximize revenue, there exists an optimal threshold, k^*, such that customers should not be let in once the queue length reaches k^*. Although this threshold can be computed numerically without much difficulty, they do not give a closed form solution for the threshold. Yildirim and Hasenbein [22] study a model similar to that of Chen and Frank but with batch arrivals; they find thresholds for groups of different sizes implicitly.

Naor too derives an *implicit* expression for the optimal threshold, but this time for the case social welfare maximization [20]. The functional form of this expression, however, was understood only through numerical analysis, and unlike in our work, no asymptotic conclusions were drawn,

and no observations on the optimal consumer surplus (or optimal earning rate, in the case of revenue-maximization) were made.

Our method for finding the threshold involves using the forward difference operator, in a manner similar to the ordinary first order condition for differentiable functions. This approach has been used before by Economou and Kanta [11] for the analysis of thresholds in a similar model involving compartmented waiting space. They too do not find a closed form for the revenue-maximizing threshold.

Although in this paper we only consider observable queues, there exist other information structures that allow customers to decide whether to join a queue. For an overview of this area, one may consult the PhD thesis of Guo [13].

2.3. Customer balking strategy and admissions control literature

Threshold policies or strategies, arise in queueing contexts outside of pricing as well. In particular, our work is closely related to the literature on optimal and equilibrium balking strategies for observable queues. In these settings, there are often no prices, but delay sensitive customers must choose some threshold queue state, at which they will no longer join the queue. Qualitative results, and sometimes explicit thresholds are found by Burnetas and Economou [6], where the server must set up and by Economou and Kanta [10], where the server periodically breaks down and requires repair. Economou et.al. [9] consider general vacation times within a similar framework. While in our paper thresholds are imposed by the firm, in these papers, they are a result of customer behavior.

2.4. The Lambert W function

Finally, one of the key contributions of our work is the introduction of the Lambert W (product logarithm) function to this area of research. Corless, et.al. [8] provide an overview of the theory and applications of the Lambert W function. Although to our knowledge, this function has not appeared in the analysis of queueing models with pricing, it has been applied in related areas involving queueing: Libman and Orda [18] use the Lambert W function to compute the optimal

time to wait before retrying an action under M/M/1-induced delay, while Gupta and Weerawat [14] use the function to compute the optimal inventory level in a two stage queueing model.

3. The Model

In this section we introduce the queueing model, the customer parameters, and the choices available to the firm. In Section 3.4, we derive the firm's earning rate (revenue) function in closed form. Since maximizing the firm's earning rate simultaneously maximizes social welfare, as the firm extracts all consumer surplus, the model will be stated only as a **revenue-maximization problem**. Consumer surplus can be maximized by implementing the same admission threshold found in this paper and setting all prices to zero.

3.1. The queueing model

We consider an M/M/1 queueing system where customers arriving according to a Poisson process with rate λ are permitted to balk (i.e., decide not to join the queue). A single server serves customers in first come first serve order with service rate μ. We let $\rho \equiv \lambda/\mu$ and note that since not all arrivals will join, it is possible to have $\rho > 1$ (i.e., $\lambda > \mu$) and still maintain a stable queue.

3.2. Customers

Customers are homogenous and delay sensitive, each obtaining a value of V from service and experiencing a delay cost of c per unit time. It will be useful to define $\nu \equiv \mu V/c$. In order to ensure that not all customers will balk, we assume $\nu > 1$.

When a customer arrives, she observes the state of the system, n, which is the number of customers in the queue (including the customer being served, if any). Assuming that the firm sets the price to be $p(n)$ when there are n customers in the system, the customer will join the queue if and only if

$$V - c \cdot \mathbb{E}[T|n] - p(n) \geq 0, \tag{3.1}$$

where $\mathbb{E}[T|n]$ is the expected response time given that there are n customers in the system, not including the new arrival. Since the average service time is $1/\mu$ for all customers, $\mathbb{E}[T|n] = (n+1)/\mu$. We rewrite (3.1) to find that the customer joins the system if and only if

$$V - \frac{c}{\mu}(n+1) - p(n) \geq 0. \tag{3.2}$$

3.3. The firm

The firm makes two decisions: (i) a state dependent pricing scheme, $p(n)$, and (ii) a threshold $k \in \mathbb{Z}_+$, representing the maximum number of customers allowed to be in the system at any time. That is, a customer arriving when there are k customers in the system is refused service.

The optimal price in each state n is simply the maximum price that a newly arriving customer is willing to pay for service given that there are n customers in the queue. Charging any more would be introducing a *de facto* threshold, as new arrivals would simply refuse to enter the queue. Charging any less would be forgoing potential revenue without changing anything else in the system, as the customer would still enter, and thus the system would transition to state $n+1$, the only difference being the firm would earn less than it could in state n. From (3.2), we find that the firm should charge

$$\begin{aligned} p(n) &= \max\{p \in \mathbb{R} \colon V - c(n+1)/\mu - p \geq 0\} \\ &= V - \frac{c}{\mu}(n+1) \\ &= \frac{c}{\mu}(\nu - (n+1)) \end{aligned} \tag{3.3}$$

in state n, as in the model described by Chen and Frank [7]. The resulting system is an M/M/1/k system.

3.4. The M/M/1/k system and the earning rate

The Markov chain of the M/M/1/k system is shown in Figure 1. Let $\pi_n^{(k)}$ denote the time average probability of being in state n for the M/M/1/k. For all $k \geq 0$, and $n \in \{0, 1, \ldots, k\}$, the limiting probability of being in state n of the Markov chain is given by

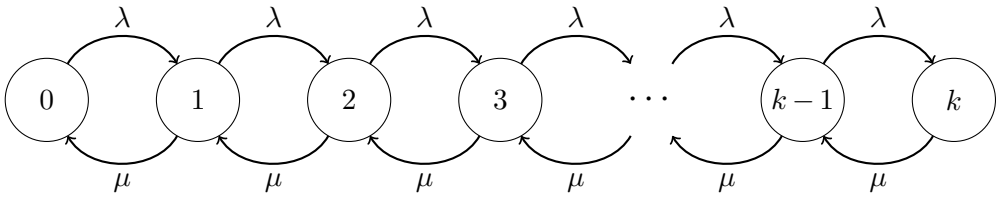

Figure 1 The Markov chain for threshold k, with the associated price written below each state. Note that there are no earnings associated with state k, as no arrivals are allowed to join the queue at this state.

$$\pi_n^{(k)} = \begin{cases} \dfrac{\rho^n(1-\rho)}{1-\rho^{k+1}} & \text{if } \rho \neq 1 \\ \\ \dfrac{1}{k+1} & \text{if } \rho = 1. \end{cases} \tag{3.4}$$

The earning rate for threshold k, $\mathcal{R}(k)$, is given by the expected price charged across all states (where the firm earns 0 at state k, since arrivals are rejected) scaled by the total arrival rate λ:

$$\mathcal{R}(k) = \lambda \sum_{n=0}^{k-1} p(n) \cdot \pi_n^{(k)}. \tag{3.5}$$

Assuming that $\rho \neq 1$, it follows from Eqs. (3.3), (3.4), and (3.5) that

$$\begin{aligned} \mathcal{R}(k) &= \frac{c\lambda}{\mu} \sum_{n=0}^{k-1} (\nu - (n+1)) \pi_n^{(k)} \\ &= \frac{c\rho}{(1-\rho)(1-\rho^{k+1})} \cdot \left[\nu(1-\rho^k)(1-\rho) + \rho^k(1+k-k\rho) - 1 \right]. \end{aligned} \tag{3.6}$$

Otherwise, $\rho = 1$, and we instead have

$$\mathcal{R}(k) = k \cdot \left(\frac{\nu}{k+1} - \frac{1}{2} \right). \tag{3.7}$$

4. Summary of results

The main problem in this paper is finding the optimal threshold, k^*, defined by any positive integer giving the solution to

$$\mathcal{R}(k^*) = \max_{k \in \mathbb{Z}^+} \mathcal{R}(k). \tag{4.1}$$

Our contributions are as follows:

- We find that the optimal threshold is given by the formula

$$k^* = \left\lceil (1-\rho)\nu + \frac{1}{1-\rho} - \frac{1}{\ln(\rho)} \cdot W\left(\frac{\ln(\rho)\cdot \rho^{(1-\rho)\nu+\frac{1}{1-\rho}}}{1-\rho}\right) - 2 \right\rceil, \tag{4.2}$$

where W is the greater branch of the Lambert W function for $\rho < 1$, and the lesser branch for $\rho > 1$ (see Theorem 1). Similar expressions exist in the multiserver setting where the firm has s servers.

- We find the following strikingly simple approximation for the optimal earning rate:

$$\mathcal{R}(k^*) \approx \mu\left(V - \frac{v}{\mu}(k^*+1)\right). \tag{4.3}$$

That is, the earning rate is approximately what we would earn if we could charge all customers the price we would charge if we admitted them at the threshold, and as soon as we served one customer there would always be another willing to pay for service. The nature of this approximation is made clear in the statement of Theorem 2.

- We find the following asymptotic characterizations for k^* as $\nu \to \infty$:

$$k^* \sim \begin{cases} (1-\rho)\nu & \text{if } \rho < 1 \\ \sqrt{2\nu} & \text{if } \rho = 1 \\ \log_\rho(\nu) & \text{if } \rho > 1 \end{cases} \tag{4.4}$$

and these asymptotic results can be used together with the aforementioned approximation to obtain the following asymptotic characterizations of $\mathcal{R}(k^*)$:

$$\mathcal{R}(k^*) \sim \begin{cases} \lambda V & \text{if } \rho < 1 \\ \mu V & \text{if } \rho \geq 1. \end{cases} \tag{4.5}$$

For further details, see Section 6.3.

- We generalize our results by finding the optimal threshold, k^* (also in closed form), for an M/M/s multiserver queueing system. We proceed to find analogues of the approximation for $\mathcal{R}(k^*)$ given by (4.3) and the asymptotic results for k^* given by (4.4). For further details, see Section 7.2.

5. The Analysis

This section contains the derivation of the closed form expression for the optimal threshold. We propose a technique involving the forward difference of $\mathcal{R}$, which we use to find the optimal threshold (Section 5.1). In deriving the threshold, we will discuss and make use of the Lambert W function and its two real-valued branches (Section 5.2).

5.1. The forward difference technique

Given λ, μ, V, and c, we seek to find the optimal threshold

$$k^* = \arg\max_{k \in \mathbb{Z}_+} \{\mathcal{R}(k)\}. \tag{5.1}$$

We know, due to a proof by Chen and Frank [7], that at least one such optimal threshold exists. In solving this maximization problem we extend $\mathcal{R}(\cdot)$ from the nonnegative integers to the reals and consider the function $\Delta\mathcal{R}(\cdot)$, defined by $\Delta\mathcal{R}(x) = \mathcal{R}(x+1) - \mathcal{R}(x)$, also known as the *forward difference* of $\mathcal{R}(\cdot)$. This function loosely captures the idea of a "discrete derivative," which we can use with the modified first order condition that all extrema of a function $f(\cdot)$ on the integers lie at points $\lceil x \rceil$ (or at boundaries), where x satisfies $\Delta f(x) \equiv f(x+1) - f(x) = 0$. In our case, either the optimal threshold, $k^* = \lceil x \rceil$, for x satisfying $\Delta\mathcal{R}(x) = 0$, or $k^* = 0$.

For the case where $\rho = 1$, applying this principle yeilds

$$k^* = \left\lceil \frac{1}{2}\left(\sqrt{1+8\nu} - 3\right) \right\rceil. \tag{5.2}$$

In order to use apply the forward difference technique to the real-valued extension of $\mathcal{R}$ given in (3.6) for the remaining case where $\rho \neq 1$, we solve the equation $\mathcal{R}(x) - \mathcal{R}(x+1) = 0$, for $x \in \mathbb{R}_+$. We call x the *unrounded optimal threshold*.

$$\begin{aligned} 0 &= \mathcal{R}(x) - \mathcal{R}(x+1) \\ &= \frac{c\rho}{(1-\rho)(1-\rho^{x+1})} \cdot \left[\nu(1-\rho^x)(1-\rho) + \rho^x(1+x-x\rho) - 1\right] \\ &\quad - \frac{c\rho}{(1-\rho)(1-\rho^{x+2})} \cdot \left[\nu(1-\rho^{x+1})(1-\rho) + \rho^{x+1}(1+(x+1)-(x+1)\rho) - 1\right]. \end{aligned} \tag{5.3}$$

Letting

$$G(\rho, \nu) \equiv \nu(1-\rho) + \frac{1}{1-\rho}, \tag{5.4}$$

and using the aid of Mathematica (for the sequence of algebraic steps, see [4]), we see that (5.3) is equivalent to

$$\ln(\rho) \cdot (G(\rho,\nu) + 2 - x)\, e^{\ln(\rho)\cdot(G(\rho,\nu)+2-x)} = \frac{\ln(\rho)\cdot\rho^{G(\rho,\nu)}}{1-\rho}. \tag{5.5}$$

Our goal is to solve this equation for x; however, elementary functions are insufficient to solve equation (5.5) in closed form, so we make use of the non-elementary Lambert W function.

DEFINITION 1. For all $z \geq -1/e$, the *Lambert W function* (also known as the product logarithm, or productlog function) is defined as either one of two real-valued functions (branches) giving the solution to

$$W(z)e^{W(z)} = z$$

We refer to the specific branches of this function as W_0 and W_{-1}, with $W_0(z) > W_{-1}(z)$ for all $z > -1/e$. The graphs of both branches are shown in Figure 2.

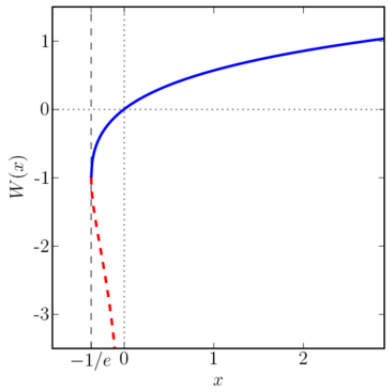

Figure 2 The real-valued branches of the Lambert W function. The solid curve gives the graph of the function W_0, and the dotted curve gives the graph of the function W_{-1}.

By the definition of W, the solutions to equations of the form $Xe^X = Y$ are $X = W(Y)$. Since Eq. (5.5) is of the form $Xe^X = Y$, applying this fact and rearranging terms, we have:

$$x = G(\rho,\nu) - \frac{1}{\ln(\rho)} \cdot W\left(\frac{\ln(\rho)\cdot\rho^{G(\rho,\nu)}}{1-\rho}\right) - 2 \tag{5.6}$$

Substituting back (5.4), we obtain a closed from expression for the unrounded optimal threshold:

$$x = (1-\rho)\nu + \frac{1}{1-\rho} - \frac{1}{\ln(\rho)} \cdot W\left(\frac{\ln(\rho)\cdot\rho^{(1-\rho)\nu+\frac{1}{1-\rho}}}{1-\rho}\right) - 2 \tag{5.7}$$

It follows that the optimal threshold $k^* = \lceil x \rceil$, where x is as given in (5.7) and can take on two values depending on which branch of W is used: W_0 or W_{-1}. In the next section, we discuss how to select the correct branch.

5.2. Selecting the correct branch of the Lambert W function

Theorem 1 states $k^* = \lceil x \rceil$ in closed form by indicating which branch should be selected when computing the optimal unrounded threshold, x. The proof relies on the observation that, when $\rho < 1$, only the W_0 branch yields a positive value for $\lceil x \rceil$, while, when $\rho > 1$, only the W_{-1} branch yields a positive value for $\lceil x \rceil$.

THEOREM 1. *For all positive $\lambda \neq 1$, the optimal threshold is*

$$k^* = \left\lceil (1-\rho)\nu + \frac{1}{1-\rho} - \frac{1}{\ln(\rho)} \cdot W_i\left(\frac{\ln(\rho) \cdot \rho^{(1-\rho)\nu + \frac{1}{1-\rho}}}{1-\rho} \right) - 2 \right\rceil \tag{5.8}$$

where $i = 0$ when $\rho < 1$ and $i = -1$ when $\rho > 1$.

In the remaining case where $\rho = 1$, the optimal threshold is $k^* = \left\lceil \left(\sqrt{1+8\nu} - 3\right)/2 \right\rceil$.

Proof. The case of $\rho = 1$ is a restatement of the result from (5.2), so we assume that $\rho \neq 1$. We consider the extension of $\mathcal{R}(k)$, as given in (3.6), from $\mathbb{Z}_+$ to $\mathbb{Z}$, and argue that $\mathcal{R}$ has two nontrivial extrema on $\mathbb{Z}$. One extremum, the optimal threshold, k^*, will be the positive local maximum (it is the global maximum on $\mathbb{Z}_+$), while the other will be the nonpositive local minimum; in particular the former is greater than the latter. It then follows that given ρ, the optimal threshold k^* is given by choosing whichever branch of the Lambert W function yields a higher value for (5.8).

Having found from (5.7) that $\Delta\mathcal{R}(x) = 0$ has at most two solutions for $x \in \mathbb{R}$ (one for each branch of W), it follows that $\mathcal{R}$ has at most two local extreme values. Hence, $\mathcal{R}$ has at most two nontrivial local extrema, where we consider a trivial local extremum to be any integer k such that $\mathcal{R}(k-1) = \mathcal{R}(k)$.

We know that there exists an optimal threshold $k^* > 0$. Moreover, since k^* must be a local maximum on $\mathbb{Z}$, it is one of at most two nontrivial local extrema. It follows that the optimal threshold k^* is given by

$$\left\lceil (1-\rho)\nu + \frac{1}{1-\rho} - \frac{1}{\ln(\rho)} \cdot W_i\left(\frac{\ln(\rho)\cdot\rho^{(1-\rho)\nu+\frac{1}{1-\rho}}}{1-\rho}\right) - 2 \right\rceil \tag{5.9}$$

with i set to the appropriate value (one of 0 or -1). Consequently, should $\mathcal{R}$ have another nontrivial local extremum, then it is also given by (5.9), except with i is set to the other possible value. Now observe that

$$\begin{aligned}\lim_{k\to-\infty}\mathcal{R}(k) &= \lim_{k\to-\infty}\frac{\rho}{1-\rho^{k+1}}\cdot\frac{1}{1-\rho}\cdot\left[\nu(1-\rho^k)(1-\rho)+\rho^k(1+k-k\rho)-1\right] \\ &= \begin{cases}+\infty & \text{if } \rho<1 \\ \rho/(1-\rho)\cdot((1-\rho)\nu-1) & \text{if } \rho>1\end{cases}\end{aligned} \tag{5.10}$$

In particular, we find that there exist values $k<0$ for which $\mathcal{R}(k)>\mathcal{R}(0)=0$, as both of these limits are positive. Since $\mathcal{R}(1)>\mathcal{R}(0)=0$ (as $\nu>1$), this is sufficient to establish that there exists some $k'\le 0$ such that k' is a local minimum for $\mathcal{R}$ on $\mathbb{Z}$. Hence, $\mathcal{R}$ does indeed have another local extremum, k', and this is the other value obtained from (5.9).

Therefore, one of the two nontrivial local extrema is $k'\le 0$, the local minimum of $\mathcal{R}$, while the other is $k^*>0$, the local maximum of $\mathcal{R}$. In particular $k'\le 0<k^*$. It follows that $k^*=\lceil x\rceil$, where x is the larger of the two solutions to the equation $\Delta\mathcal{R}(x)=0$, that is, the larger of the two possible values from (5.9).

When $\rho<1$ the coefficient of the Lambert W function in (5.9), $-1/\ln(\rho)$, is positive, and since $W_0(x)\ge W_{-1}(x)$, the value of (5.9) is greater when $i=0$. Similarly, when $\rho>1$, the coefficient is negative, so the value of (5.9) is greater when $i=-1$. Thus, k^* is given by $i=0$ for $\rho<1$ and $i=-1$ for $\rho>1$. □

6. Results: The optimal earning rate and the asymptotic characterization of k^*

While Theorem 1 gives a closed form expression for the optimal threshold, k^*, this section describes the impact of k^* on the earning rate, and gives an asymptotic characterization of k^*. In Section 6.1, we argue that implementing the optimal threshold can lead to significant gains in revenue over a suboptimal threshold. We find that the greatest gains, and therefore the most important settings for implementing the best threshold, are those where $\rho>1$. Next, in Section 6.2, we derive

a strikingly simple approximation for the optimal earning rate, $\mathcal{R}(k^*)$. Finally, in Section 6.3, we use our expression from Theorem 1 to derive asymptotic results on k^* and $\mathcal{R}(k^*)$ as $\nu \to \infty$.

6.1. Significance of the optimal threshold

In this section, we study the significance of the threshold on the earning rate. The numerical examples for the plots in this section are chosen for illustrative purposes.

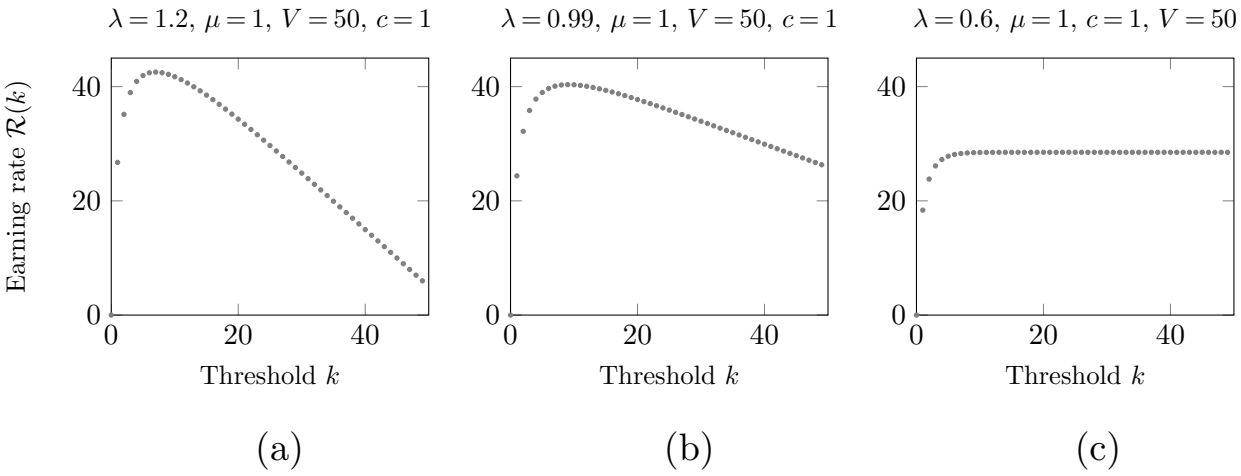

Figure 3 Plots of the earning rate $\mathcal{R}(k)$ vs. the threshold k when $\mu = 1$, $\nu = V = 50$, $c = 1$ and (a) $\rho = \lambda = 1.2$, (b) $\rho = \lambda = 0.99$, and (c) $\rho = \lambda = 0.6$. The optimal thresholds are (a) $k^* = 7$, (b) $k^* = 9$, and (c) $k^* = 21$.

Our findings suggest that in settings where $\rho > 1$, choosing a threshold that is too high can be arbitrarily detrimental, suggesting that it is crucial to implement the optimal (or at least some near-optimal) threshold. This is illustrated by Figure 3(a), where we have a "moderately high" value of $\rho = \lambda = 1.2$, and $\mu = 1$, $V = 50$, and $v = 1$. The earning rate at the optimal threshold $k^* = 7$ is over 600% higher than the earning rate attained by implementing the natural threshold $k = V - 1 = 49$. Implementing the thresholds $k = 6$ or $k = 8$, yield negligibly smaller earnings. Meanwhile, when compared to an intermediate threshold such as $k = 25$, the optimal threshold yields 43% higher revenue.

Similar findings emerge for values of ρ close to 1, although the potential earnings by implementing the optimal threshold are less than when $\rho > 1$. For the case illustrated in Figure 3(b) the earning

rate at the optimal threshold $k^* = 9$ is about 12% higher than the intermediate threshold of $k = 25$ and 53% higher than the extreme threshold of $k = 49$.

We find that the choice of threshold is relatively insignificant, however, when $\rho < 1$ by a considerable margin. In such settings, nearly all thresholds yield virtually the same earning rate. This is because the system will rarely enter higher states, as the likelihood of being in such a state decreases geometrically, and such decreases are considerably sharp when $\rho \ll 1$. We illustrate the "low ρ case" with Figure 3(c), where the optimal threshold is $k^* = 21$. However, employing any threshold with $k \geq 7$ yields revenues within 1% of optimal.

In summary, we observe that the choice of threshold is nearly inconsequential for low ρ, as in (c), but is much more significant for high ρ, as in (a); when ρ is even higher than 1.2, the "nearly linear" decline is steeper.

6.2. The earning rate at the optimal threshold

We can determine the earning rate at the optimal threshold by substituting our closed form expression for k^* from Theorem 1 into $\mathcal{R}(k)$, as given in (3.6). Since the expression for k^* involves the ceiling function, $\mathcal{R}(k^*)$ will generally not have a simple form. Recall that

$$x = (1-\rho)\nu + \frac{1}{1-\rho} - \frac{1}{\ln(\rho)} \cdot W\left(\frac{\ln(\rho) \cdot \rho^{(1-\rho)\nu + \frac{1}{1-\rho}}}{1-\rho}\right) - 2 \tag{6.1}$$

with $W = W_0$ for $\rho < 1$ and $W = W_{-1}$ for $\rho > 1$, is the *unrounded optimal threshold*. We compute $\mathcal{R}(x)$ and obtain the approximation $\mathcal{R}(x) \approx \mathcal{R}(k^*)$. The following strikingly simple result expresses $\mathcal{R}(x)$ in terms of x itself.

THEOREM 2. *Let x be the unrounded optimal threshold. Then*

$$\begin{aligned} \mathcal{R}(x) &= c(\nu - (x+1)) \\ &= \mu\left(V - \frac{c}{\mu}(x+1)\right). \end{aligned}$$

Proof. For both the cases of $\rho \neq 1$ and $\rho = 1$, the result follows from straightforward substitution via Mathematica. The algebraic steps are detailed in [4]. □

When $x \in \mathbb{Z}$, we have $x = k^*$, and so $\mathcal{R}(k^*) = V - (k^* + 1)$ *holds exactly.* In this integral case, where thresholds k^* and $k^* + 1$ are both optimal, there is a queueing theoretic justification for this result, providing some intuition. We start with

$$\mathcal{R}(k^*) = \mathcal{R}(k^* + 1) \tag{6.2}$$

with the left-hand (right-hand) side corresponding to the M/M/1/k^* (respectively, M/M/1/$k^* + 1$) Markov chain, with prices $p(n) = (c/\mu)(\nu - (n + 1))$ at each state where we allow customers to enter. In the "$k^* + 1$ chain" (the one corresponding to the right-hand side), with probability q we are in one of the states that are also in the "k^* chain," and with probability $1 - q$, we are in state $k^* + 1$. Since this chain is a birth-death process, we can decompose it and see that we earn at a rate of $\mathcal{R}(k^*)$ when transitioning to all but the final state (prices are the same in both settings), and we earn $(c/\mu)(\nu - (k^* + 1))$ when an arrival enters the queue at state k^*, and the system transitions to state $k^* + 1$. However, to decompose the earnings properly, we may equivalently consider that we earn nothing when transitioning to state $k^* + 1$ (just as would happen when there is an "dropped arrival" at state k^* in the "k^* chain"), and instead earn $(c/\mu)(\nu - (k^* + 1))$ when leaving state $k^* + 1$. This is a valid reformulation of the model as every time we enter state $k^* + 1$, we must eventually leave it, transitioning back to state k^*. Transitioning from $k^* + 1$ to k^* occurs at rate of μ, so we have an additional contribution of $c(1 - q)(\nu - (k^* + 1))$ to the earning rate from this part of the chain. Therefore, we have

$$\begin{aligned}
\mathcal{R}(k^*) &= \mathcal{R}(k^* + 1) \\
&= q \cdot \mathcal{R}(k^*) + c(1 - q)(\nu - (k^* + 1)) \\
&= c(\nu - (k^* + 1)) \\
&= \mu \left(V - \frac{c}{\mu}(k^* + 1) \right)
\end{aligned}$$

with the penultimate equality following from straightforward algebra. Whenever $x \notin \mathbb{Z}$, we instead have the approximation

$$\mathcal{R}(k^*) \approx \mu \left(V - \frac{c}{\mu}(k^* + 1) \right). \tag{6.3}$$

We interpret this result as follows: the earning rate is approximately what we would earn if we could charge all customers the price we would charge if we admitted them at the threshold, and as soon as we served one customer there would always be another willing to pay for service. Together with the asymptotic results in Section 6.3, this allows us to compute how the rate of earning changes as V grows.

6.3. Asymptotic analysis

In this section we give three asymptotic results, as $\nu \to \infty$. We can interpret the regime where $\nu \to \infty$, as the regime when customer valuations, $V \to \infty$ (with μ and c fixed), when service rate $\mu \to \infty$ (with V and c fixed and λ varying so that ρ remains fixed), or when $c \to 0^+$ (with V and μ fixed). Note that asymptotic results for ρ are less important, as it is easy to see that as $\rho \to \infty$ and $\nu > 1$, we have $k^* = 1$, while as $\rho \to 0$, the choice of threshold becomes irrelevant. This analysis would not be possible were it not for the expression for k^* that we derived in Section 5, and moreover, the derivation of these asymptotic results will necessarily make use of both branches of the Lambert W function. Our results are summarized in Table 1. We also briefly discuss the asymptotic properties of the earning rate, using the result in Section 6.2.

We use the notation $f \sim g$ as $x \to \infty$, to mean that

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1 \tag{6.4}$$

while $f \sim g$ as $x \to 0^-$ is defined analogously. We also write $f(x) = o(g(x))$ if

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0. \tag{6.5}$$

THEOREM 3. *For a fixed $\rho < 1$, as $\nu \to \infty$, $k^* \sim (1-\rho)\nu$.*

Proof. When $\rho < 1$, we let $W = W_0$. Noting that W is continuous around 0 and $W(0) = 0$, we have

$$\lim_{\nu\to\infty} k^* = \lim_{\nu\to\infty} \left\lceil (1-\rho)\nu + \frac{1}{1-\rho} - \frac{1}{\ln(\rho)} \cdot W\left(\frac{\ln(\rho) \cdot \rho^{(1-\rho)\nu + \frac{1}{1-\rho}}}{1-\rho} \right) - 2 \right\rceil$$

$$
\begin{aligned}
&= \lim_{\nu\to\infty}\left((1-\rho)\nu - \frac{1}{\ln(\rho)}\cdot W\left(\frac{\ln(\rho)\cdot\rho^{(1-\rho)\nu+\frac{1}{1-\rho}}}{1-\rho}\right)\right) + o(\nu)\\
&= \lim_{\nu\to\infty}\left((1-\rho)\nu\right) - \frac{1}{\ln(\rho)}\cdot\lim_{\nu\to\infty}\left(W\left(\frac{\ln(\rho)\cdot\rho^{(1-\rho)\nu+\frac{1}{1-\rho}}}{1-\rho}\right)\right) + o(\nu)\\
&= \lim_{\nu\to\infty}\left((1-\rho)\nu\right) - \frac{1}{\ln(\rho)}\cdot W(0) + o(\nu)\\
&= \lim_{\nu\to\infty}\left((1-\rho)\nu\right) + o(\nu).
\end{aligned}
$$

It then follows from the limit above that $k^* \sim (1-\rho)\nu$. □

THEOREM 4. *For $\rho = 1$, as $\nu\to\infty$, $k^* \sim \sqrt{2\nu}$.*

Proof. When $\rho = 1$, we have

$$
\begin{aligned}
\lim_{\nu\to\infty} k^* &= \lim_{\nu\to\infty}\left\lceil\frac{1}{2}\left(\sqrt{1+8\nu}-3\right)\right\rceil\\
&= \frac{1}{2}\lim_{\nu\to\infty}\left(\sqrt{8\nu}+o\left(\sqrt{\nu}\right)\right)\\
&= \lim_{\nu\to\infty}\left(\sqrt{2\nu}+o\left(\sqrt{\nu}\right)\right).
\end{aligned}
$$

It then follows from the limit above that that $k^* \sim \sqrt{2\nu}$. □

THEOREM 5. *For a fixed $\rho > 1$, as $\nu\to\infty$, $k^* \sim \log_\rho \nu$.*

Proof. When $\rho > 1$, we let $W = W_{-1}$ and note that we have the property that

$$
W(x) \sim -\ln(-1/x) - \ln(\ln(-1/x)) \tag{6.6}
$$

as $x\to 0^-$ (see [8]). Applying this property, and the fact that

$$
\frac{\ln(\rho)\cdot\rho^{(1-\rho)\nu+\frac{1}{1-\rho}}}{1-\rho} \to 0 \tag{6.7}
$$

(and is negative) as $\nu\to\infty$, we have

$$
\begin{aligned}
k^* &= \left\lceil(1-\rho)\nu + \frac{1}{1-\rho} - \frac{1}{\ln(\rho)}\cdot W\left(\frac{\ln(\rho)\cdot\rho^{(1-\rho)\nu+\frac{1}{1-\rho}}}{1-\rho}\right) - 2\right\rceil\\
&\sim (1-\rho)\nu + \frac{1}{\ln(\rho)}\cdot\left(\ln\left(\frac{(\rho-1)\cdot\rho^{(\rho-1)\nu-\frac{1}{1-\rho}}}{\ln(\rho)}\right) + \ln\left(\ln\left(\frac{(\rho-1)\cdot\rho^{(\rho-1)\nu-\frac{1}{1-\rho}}}{\ln(\rho)}\right)\right)\right)\\
&= (1-\rho)\nu + \log_\rho\left(\frac{(\rho-1)\cdot\rho^{(\rho-1)\nu-\frac{1}{1-\rho}}}{\ln(\rho)}\right) + \log_\rho\left(\ln\left(\frac{(\rho-1)\cdot\rho^{(\rho-1)\nu-\frac{1}{1-\rho}}}{\ln(\rho)}\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= (1-\rho)\nu + (\rho-1)\nu - \frac{1}{1-\rho} + \log_\rho\left(\frac{\rho-1}{\ln(\rho)}\right) + \log_\rho\left(\ln\left(\frac{(\rho-1)\cdot\rho^{(\rho-1)\nu-\frac{1}{1-\rho}}}{\ln(\rho)}\right)\right)\\
&\sim \log_\rho\left(\ln\left(\frac{(\rho-1)\cdot\rho^{(\rho-1)\nu-\frac{1}{1-\rho}}}{\ln(\rho)}\right)\right)\\
&= \log_\rho\left((\rho-1)\ln(\rho)\nu - \frac{\ln(\rho)}{1-\rho} + \ln\left(\frac{\rho-1}{\ln(\rho)}\right)\right)\\
&\sim \log_\rho(\nu),
\end{aligned}
$$

with the last step following because the argument of the logarithm is an affine function of ν. □

The asymptotic results and their impact on the earning rate at the optimal threshold, $\mathcal{R}(k^*)$, are summarized in the Table 1. The main observation is that for very large ν, k^* is a decreasing proportion of ν as ρ increases. Specifically when $\rho<1$, k^* is linearly proportional to ν; when $\rho=1$, k^* is proportional to the square root of ν; when $\rho>1$, k^* is logarithmic in ν. Hence, we see a very sharp difference, when we cross from $\rho<1$ to $\rho>1$, in accordance with the plots in Section 6.1. It is important to note that these asymptotic conditions provide good approximations only when ν is *very* large, and they are primarily illustrative in understanding the structure of k^* and providing intuition rather than being effective for actual implementation. This holds especially for the $\rho>1$ result, as the convergence of $W_{-1}(x)$ as $x\to 0^-$ is particularly slow.

Behavior of k^* as $\nu\to\infty$	Behavior of $\mathcal{R}(k^*)$ as $\nu\to\infty$	Condition on ρ
$k^*\sim(1-\rho)\nu$	$\mathcal{R}(k^*)\sim\lambda V$	$\rho<1$
$k^*\sim\sqrt{2\nu}$	$\mathcal{R}(k^*)\sim V$	$\rho=1$
$k^*\sim\log_\rho(\nu)$	$\mathcal{R}(k^*)\sim\mu V$	$\rho>1$

Table 1 Summary of asymptotic results as $\nu\to\infty$.

On the other hand, the asymptotic results for $\mathcal{R}$, which follow from the structure of $\mathcal{R}$ and the approximation $\mathcal{R}(k)\approx\mu(V-(k^*-1))$, are reasonably accurate even for low values of V. We are also able to derive asymptotic comparative statics results. In particular, as $\nu\to\infty$ (and in particular, as $V\to\infty$), we have

$$
\frac{\partial\mathcal{R}(k^*)}{\partial V}\sim\begin{cases}\lambda & \text{if } \rho\le 1\\ \mu & \text{if } \rho\ge 1.\end{cases} \tag{6.8}
$$

Intuitively, as V grows without bound, the earnings are proportional to service *throughput*.

7. Multiserver systems

Until now, we have assumed an M/M/1 queueing model. We now generalize to an M/M/s queueing model, where the firm has s identical servers. This allows us to study the impact of the number of servers, s, on the optimal threshold, k^*, and the optimal revenue, $\mathcal{R}(k^*)$.

7.1. The multiserver model

As in the single-server model, we let V, c, and λ be the customer valuation, delay cost per unit time, and arrival rate, respectively, while we assume that *each* of the s servers serves customers with rate μ. We redefine the load to be $\rho \equiv \lambda/(s\mu)$ and redefine $\nu \equiv s\mu V/c$; we again assume that $\nu > 1$. That is, we replace the service rate μ in the original definitions of ρ and ν with the *aggregated service rate* $s\mu$. As the single-server model corresponds to the case where $s = 1$, these definitions are consistent with the original definitions.

We observe that all customers arriving to a system where there are $n < s$ customers present (i.e., a system where there is at least one idle server) incur an expected waiting cost of c/μ. We also observe that all customers arriving to a system where there are $n \geq s$ customers present (i.e., a system where all servers are busy), incur an expected waiting cost of $c/\mu + c(n-s+1)/(s\mu) = c(n+1)/(s\mu)$: in expectation, these customer must wait $1/(s\mu)$ units of time for each of the $n-s+1$ customers in the queue (including itself), before receiving service for an additional $1/\mu$ units of time. Hence, the firm's optimal state-dependent pricing scheme is given by

$$\begin{aligned} p(n) &= \begin{cases} V - c/\mu & \text{if } n < s \\ V - c(n+1)/(s\mu) & \text{if } n \geq s \end{cases} \\ &= \begin{cases} c(\nu - s)/(s\mu) & \text{if } n < s \\ c(\nu - (n+1))/(s\mu) & \text{if } n \geq s. \end{cases} \end{aligned}$$

When imposing a threshold at state k, we have an M/M/s/k queueing system, as depicted in Figure 4 for the case where $s = 3$. As in the case of a single server, we use the corresponding limiting probabilities $\pi_i^{(s;k)}$ to give an expression for the earning rate,

$$\mathcal{R}(k) = c\rho \left((\nu - s) \sum_{n=0}^{s-1} \pi_n^{(s;k)} + \sum_{n=s}^{k-1} (\nu - (n+1)) \, \pi_n^{(s;k)} \right). \tag{7.1}$$

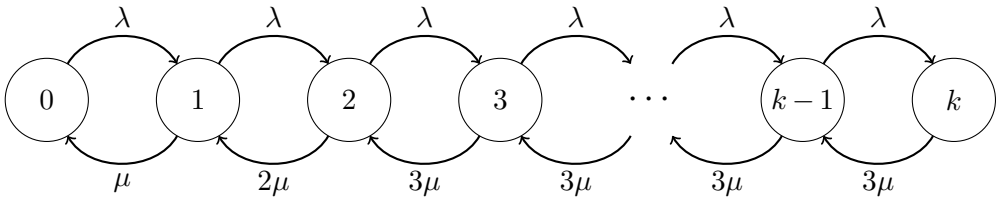

Figure 4 The Markov chain for a system with $s=3$ servers and threshold k, with the associated price written below each state.

7.2. The optimal threshold in multiserver systems

For a given value of s, we may again apply the methodology in Section 5 to obtain closed-form solutions for the optimal threshold, k^*, with the aid of a computer.

THEOREM 6. *Let*

$$Q(s,\rho) = se^{s\rho}\Gamma(s,s\rho) = \sum_{j=0}^{s-1} \frac{s!(s\rho)^j}{j!}$$

$$C_1 = s^s\rho^2$$

$$C_2 = (1-\rho)(s\rho)^s + (1-\rho)^2 Q(s,\rho)$$

$$C_3 = (s\rho)^s(1+\rho(s\rho-s-2)) + (1-\rho)^2(1-s\rho)Q(s,\rho),$$

where $\Gamma(\cdot,\cdot)$ is the incomplete gamma function.

For an s-server system, the optimal threshold, k^*, is given by

$$k^* = \max\{s, \lceil x \rceil\}, \text{ with } x = (1-\rho)\nu - \frac{C_3}{C_2} - \frac{1}{\ln(\rho)} \cdot W_i\left(\frac{C_1\ln(\rho)\cdot\rho^{-\frac{C_3}{C_2}+(1-\rho)\nu}}{C_2}\right), \tag{7.2}$$

where $i=0$ when $\rho<1$ and $i=-1$ when $\rho>1$. In the remaining case where $\rho=1$, we have

$$k^* = \max\{s, \lceil x \rceil\}, \text{ with } x = s - \frac{Q(s,1)}{s^s} + \sqrt{2\nu - 2s + \frac{Q(s,1)}{s^s}\cdot\left(1+\frac{Q(s,1)}{s^s}\right) + \frac{1}{4}} - \frac{3}{2},$$

where we define $k^* \equiv s$ when $x \notin \mathbb{R}$.

Proof. Let $\mathcal{R}$ be as given in (7.1). In the case where $\rho \neq 1$, using Mathematica, we can show that multiplying both sides of $\mathcal{R}(k+1) - \mathcal{R}(k) = 0$ by a well-chosen nonzero value yields the equivalent equation

$$C_1\rho^k + C_2(k-(1-\rho)\nu) + C_3 = 0,$$

where C_1, C_2, and C_3 are given in the statement of the theorem. Solving this equation using the Lambert W function gives x as in (7.2), when using the branch of the W function resulting in the larger value. By arguments similar to those presented in Theorem 1, if $\lceil x \rceil \geq s$, then $\lceil x \rceil$ is the optimal threshold. Otherwise, the optimal threshold is s, as since $\nu > 1$, it is always in the firm's best interest to use all available servers. A similar approach gives the result for $\rho = 1$. □

We note that $Q(s, \rho)$ is an $s-1$ degree polynomial in ρ that does not depend on ν or k, from which it follows that C_1, C_2, and C_3 are also polynomials in ρ that do not depend on ν or k. The closed forms results for k^* in the s-server setting are beneficial in that they allow us to obtain results analogous to those derived for the single-server setting.

COROLLARY 1. *For an s-server system, the results of Theorems 2, 3, 4, and 5 hold, where we replace μ with $s\mu$ and interpret ρ and ν according to the revised definitions presented in this section: $\rho \equiv \lambda/(s\mu)$ and $\nu \equiv s\mu V/c$.*

Proof. For Theorem 2, we verify with Mathematica that $\mathcal{R}(x+1) - \mathcal{R}(x) = 0$ is equivalent to $\mathcal{R}(x) - c(\nu - (x+1)) = 0$. The asymptotic results in Theorems 3, 4, and 5 are straightforward to prove by mimicking the proofs of the original theorems (replacing the expressions for k^* as appropriate), where we observe that C_1, C_2, and C_3 are constant in ν. □

We observe that the alternative representation of the result from Theorem 2 becomes

$$\mathcal{R}(x) = s\mu \left(V - \frac{c}{s\mu}(x+1) \right).$$

This expression has the same interpretation as in the single server setting, because in the multi-server setting, the firm uses its entire service capacity, $s\mu$, when serving customers at the optimal threshold, $k^* \geq s$. The asymptotic results from Theorems 3, 4, and 5 suggest that the number of servers, s, has no asymptotic effect on the optimal threshold, as long as the aggregated service rate, $s\mu$, remains constant.

7.3. The impact of the number of servers

We study the impact of the number of servers, s, on both the optimal threshold, k^*, and the resulting optimal earning rate, $\mathcal{R}(k^*)$. We examine an s-server system under high, intermediate,

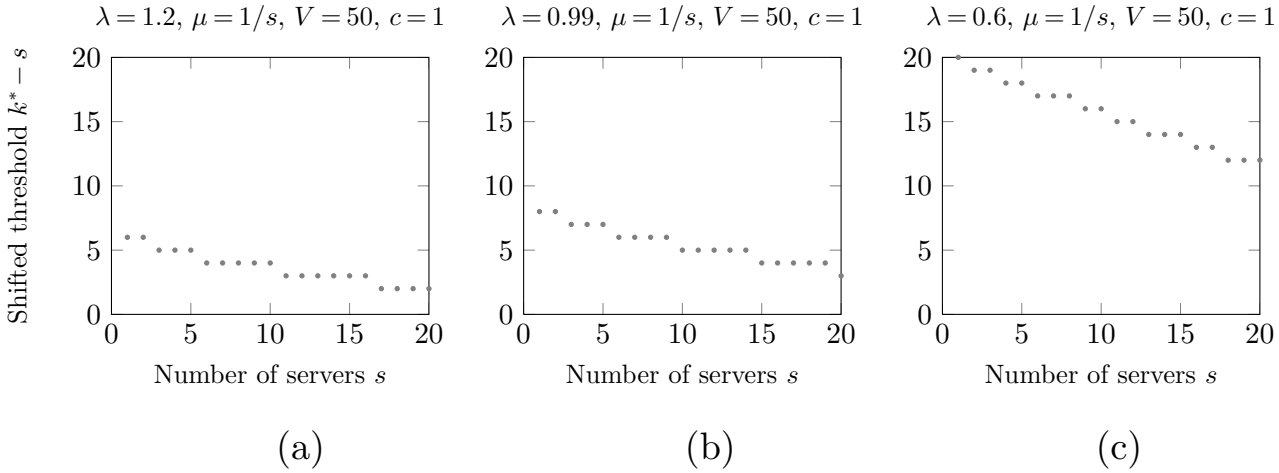

Figure 5 Plots of the shifted optimal threshold, $k^* - s$, versus the number of servers, s, when $\mu = 1/s$, $\nu = V = 50$, $c = 1$ and (a) $\rho = \lambda = 1.2$, (b) $\rho = \lambda = 0.99$, and (c) $\rho = \lambda = 0.6$.

and low load with $\mu = 1/s$, $V = 50$, and $c = 1$; we vary μ with the number of servers, s, in order to ensure that ρ and ν remain fixed. It will be more useful to examine $k^* - s$ rather than k^*: if we are willing to accept any customers, then we are willing to accept at least s customers, since we can charge the same price for the first s customers. We call $k^* - s$ the shifted optimal threshold; it can be interpreted as positioning the threshold after the $(k^* - s)$-th spot in the *queue* (rather than the *system*), and hence is the same as the optimal maximum queue length. We find that the shifted optimal threshold falls steadily as the number of servers grows. This is likely due to the fact that the firm does not need to maintain a long queue when there is also space for customers at the servers.

Figure 6 captures the impact of the number of servers, s, on the earning rate, $\mathcal{R}(k^*)$. Since the service requirement of all customers increases as s increases, due to our assumption that $\mu = 1/s$, the earning rate is decreasing in s. We note that as s increases, the optimal earning rate falls at a greater rate for higher ρ (Figure 6(a)) than for lower ρ (Figure 6(c)).

8. Further work

It would be interesting to extend the forward difference techniques in this paper to solve extensions of this problem. One extension would be to allow the customers to be heterogeneous in terms of their value for service, V, without allowing the firm to maintain multiple queues or practice

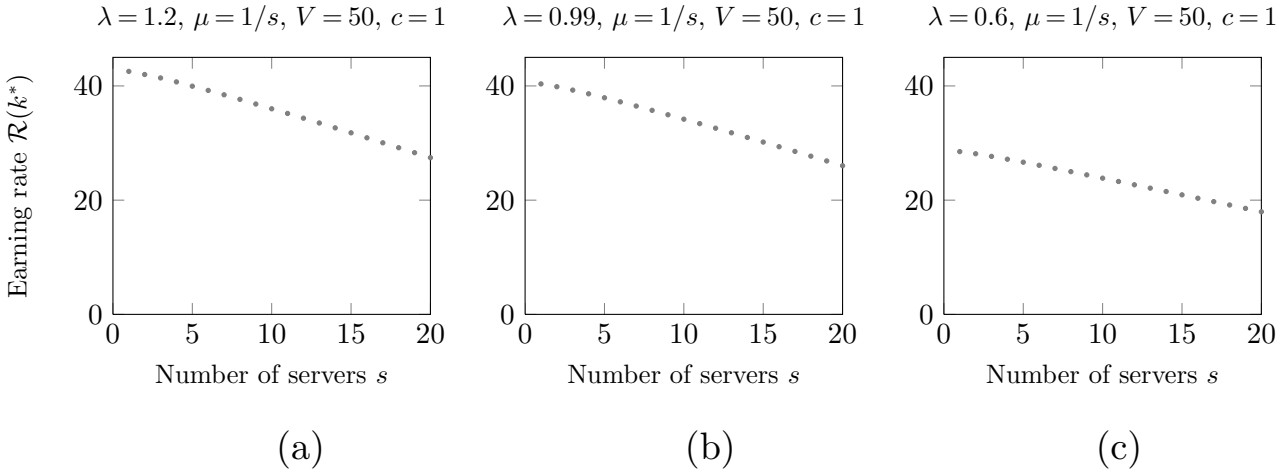

Figure 6 Plots of the optimal earning rate, $\mathcal{R}(k^*)$, versus the number of servers, s, when $\mu = 1/s$, $\nu = V = 50$, $c = 1$ and (a) $\rho = \lambda = 1.2$, (b) $\rho = \lambda = 0.99$, and (c) $\rho = \lambda = 0.6$.

price discrimination. Similarly, it would also be desirable to extend these results to finding the optimal threshold for heterogeneous waiting costs, c, but this would likely require the introduction of additional techniques, as such settings generally require multiple thresholds. Another difficult extension would be to include discount rates in the earning rate function. This may require using the techniques in this paper in conjunction with dynamic programming.

References

[1] P. Afèche, *Incentive-compatible revenue management in queueing systems: optimal strategic delay and capacity*, Working paper, University of Toronto (2010).

[2] H. Alperstein, *Optimal pricing policy for the service facility offering a set of priority prices*, Management Science **34** (1988), no. 5, 666–671.

[3] KR Balachandran, *Purchasing priorities in queues*, Management Science **18** (1972), no. 5, Part 1, 319–326.

[4] C. Borgs, J.T. Chayes, S. Doroudi, M. Harchol-Balter, and K. Xu, *The optimal admission threshold in observable queues with state dependent pricing*, Tech. report, CMU-CS-12-145, School of Computer Science, Carnegie Mellon University, 2012.

[5] ———, *Pricing and queueing*, ACM SIGMETRICS Performance Evaluation Review **40** (2012), no. 3, 71–73.

[6] A. Burnetas and A. Economou, *Equilibrium customer strategies in a single server Markovian queue with setup times*, Queueing Systems **56** (2007), no. 3, 213–228.

[7] H. Chen and M.Z. Frank, *State dependent pricing with a queue*, IIE Transactions **33** (2001), no. 10, 847–860.

[8] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, and D.E. Knuth, *On the Lambert W function*, Advances in Computational Mathematics **5** (1996), no. 1, 329–359.

[9] A. Economou, A. Gómez-Corral, and S. Kanta, *Optimal balking strategies in single-server queues with general service and vacation times*, Performance Evaluation **68** (2011), no. 10, 967–982.

[10] A. Economou and S. Kanta, *Equilibrium balking strategies in the observable single-server queue with breakdowns and repairs*, Operations Research Letters **36** (2008), no. 6, 696–699.

[11] ———, *Optimal balking strategies and pricing for the single server Markovian queue with compartmented waiting space*, Queueing Systems **59** (2008), no. 3, 237–269.

[12] S.B. Ghanem, *Computing center optimization by a pricing-priority policy*, IBM Systems Journal **14** (1975), no. 3, 272–291.

[13] P. Guo, *Analysis and comparison of queues with different levels of delay information*, Ph.D. thesis, Duke University, 2007.

[14] D. Gupta and W. Weerawat, *Supplier-manufacturer coordination in capacitated two-stage supply chains*, European Journal of Operational Research **175** (2006), no. 1, 67–89.

[15] R. Hassin and M. Haviv, *To queue or not to queue: Equilibrium behavior in queueing systems*, Kluwer, 2003.

[16] T. Kittsteiner and B. Moldovanu, *Priority auctions and queue disciplines that depend on processing time*, Management Science **51** (2005), no. 2, 236–248.

[17] N.C. Knudsen, *Individual and social optimization in a multiserver queue with a general cost-benefit structure*, Econometrica: Journal of the Econometric Society (1972), 515–528.

[18] L. Libman and A. Orda, *Optimal retrial and timeout strategies for accessing network resources*, Networking, IEEE/ACM Transactions on **10** (2002), no. 4, 551–564.

[19] H. Mendelson and S. Whang, *Optimal incentive-compatible priority pricing for the m/m/1 queue*, Operations Research **38** (1990), no. 5, 870–883.

[20] P. Naor, *The regulation of queue size by levying tolls*, Econometrica: journal of the Econometric Society **37** (1969), no. 1, 15–24.

[21] E.L. Plambeck, *Optimal leadtime differentiation via diffusion approximations*, Operations Research **52** (2004), no. 2, 213–228.

[22] U. Yildirim and J.J. Hasenbein, *Admission control and pricing in a queue with batch arrivals*, Operations Research Letters **38** (2010), no. 5, 427–431.