Probability in the Engineering and Informational Sciences, **23**, 2009, 357–383. Printed in the U.S.A.
doi:10.1017/S0269964809000205

DYNAMIC PRICING TO CONTROL LOSS SYSTEMS WITH QUALITY OF SERVICE TARGETS

Robert C. Hampshire
Carnegie Mellon University
Pittsburgh, PA
E-mail: hamp@andrew.cmu.edu

William A. Massey
Princeton University
Princeton, NJ
E-mail: wmassey@princeton.edu

Qiong Wang
Bell Laboratories
Murray Hill, NJ
E-mail: qwang@research.bell-labs.com

Numerous examples of real-time services arise in the service industry that can be modeled as loss systems. These include agent staffing for call centers, provisioning bandwidth for private line services, making rooms available for hotel reservations, and congestion pricing for parking spaces. Given that arriving customers make their decision to join the system based on the current service price, the manager can use price as a mechanism to control the utilization of the system. A major objective for the manager is then to find a pricing policy that maximizes total revenue while meeting the quality of service targets desired by the customers. For systems with growing demand and service capacity, we provide a dynamic pricing algorithm. A key feature of our solution is congestion pricing. We use demand forecasts to anticipate future service congestion and set the present price accordingly.

1. INTRODUCTION

Many finite capacity service systems can be formulated as a *loss* system, where arriving customers are rejected at times when the system capacity is full. This is in contrast to a *delay* system, where the customers are put into a queue or a waiting line until service capacity becomes available. Examples of loss systems include private line communication services, call centers, hotels, and parking garages. A given system might also evolve from a delay system to a loss system as the nature of its offered services changes. Consider, for example, the underlying packet-switched data network of the Internet. It fits the description of a delay system when the network only offers the so-called "best-effort" service, where admission is open to all consumers and excess traffic (packets) is queued in the buffer. Nevertheless, for many recent real-time applications, such as voice over IP (VoIP) and Internet video, the network has to offer "guaranteed service," where consumers are rejected if there is not enough capacity to handle their demands. For such situations, a loss system formulation is more appropriate.

As is the case with many service facilities, customer arrivals at a loss system depend on the price of the service. Moreover, pricing serves the following three functions: a device for revenue collection, an instrument for admission control, and a control mechanism to attain a given quality of service (QoS). Note that admission control is the only way that the service manager can reduce the system load. Moreover, the QoS level is a measure of satisfaction for the customer.

For loss systems, the primary measure of the QoS is the *blocking rate*, or the probability that an arriving customer is denied service. For some cases, QoS is a contractual agreement between the customer and the service provider (e.g., leased lines and call centers). For other services, the QoS is a target that the provider aspires to meet (e.g., parking garages and hotels). Even in the latter case, missing a QoS target results in customer dissatisfaction and hurts the provider's business in the long run. Therefore, a sensible pricing policy should not only strive to maximize the provider's immediate revenue but also to satisfy the QoS target, which is a property that we require for the pricing policy developed in the article.

Early work on combining queuing and pricing, with applications to information systems, concentrate on the delay model (see Dewan and Mendelson [3], Mendelson [18], Mendelson and Whang [19], Stidham [21], and Westland [23]). This is also the case for Internet pricing (see Bailey and McKnight [1] and Mackie-Mason and Varian [15]). Pricing of a loss system that offers guaranteed services has been proposed in Wang, Peha, and Sirbu [22] and studied in great detail in Courcoubetis, Dimakis, and Reiman [2], Lanning, Massey, Rider, and Wang [14], and Maglaras and Zeevi [16].

State-dependent closed-loop pricing policies have been addressed by Fan-Orzechowski and Feinberg [8]. This results in a randomized pricing strategy. Similar pricing work with periodic arrival rates has been addressed by Yoon and Lewis [24] but without constraints on the blocking probabilities. Developing a revenue-maximizing pricing policy that accommodates time-varying demands and satisfies the required QoS constraint remains an open problem, even for the classical blocking model. The main difficulty is integrating the blocking rate calculation into the price optimization, especially in cases when customer arrivals are nonstationary. Our article addresses this issue and is based on work found in Hampshire [9].

We consider a growing service where the demand and capacity are large. In this setting, we develop a dynamic pricing solution that maximizes revenue while meeting the blocking rate target. The policy explicitly incorporates the nonstationarity of demand and is forward-looking by setting the price in anticipation of future congestion. We use a demand forecast to anticipate this future service congestion and set the present price accordingly.

There are several key contributions of our work:

- Our dynamic pricing policy is established using deterministic optimal control theory. Insight from Lagrangian mechanics is used to solve this problem. The notion of an opportunity cost per customer, which captures the monetary impact of admitting an additional customer on future congestion, emerges as an essential component of our policy.
- We approximate the uniform blocking rate constraint for the time-varying loss system by a simple threshold constraint for the mean of an infinite server queue with time-varying arrival rates. We can estimate this threshold by using a generic special function first introduced in Hampshire, Massey, Mitra, and Wang [10]. It is the inverse of the hazard rate function for a Normal distribution. Several new properties of this function are developed and used to further our analysis.
- We compare our algorithm with a static policy and a myopic policy. The static policy is the fixed price that maximizes revenue while meeting the blocking rate constraint. The myopic policy is a dynamic policy that uses only the present state of the system to determine the current price while meeting the blocking rate constraint. Our policy generates more revenue than both the static and myopic policies over a wide range of parameter values.
- Finally, we derive sensitivity results that provide managerial insights for quantifying the revenue trade-off between increasing the capacity and decreasing the quality of service.

In the section that follows, we introduce a carried load model of the service and formulate the pricing problem. The tools of calculus of variations and Lagrangian mechanics are employed in Section 3 to produce an approximate solution to the carried load pricing problem. We first reduce this problem to a constrained offered load problem that approximates the same blocking probabilities of the former. The latter problem can be solved using our Lagrangian dynamics. Our algorithm is then a numerical solution of this Lagrangian problem. In Section 4 the case of bounded elastic demand is treated with a numerical example. In Section 5 we can price our QoS metric by suggesting the percentage change in the blocking probability target that is needed to achieve a given percentage change in revenue. In a similar manner, we can also price the percentage change of capacity and show the trade-off between the two. Finally, in Section 6, we summarize our conclusions. This is followed by an Appendix on the properties of a special function that is relevant to the analysis in this article.

2. CARRIED LOAD MODEL AND PRICING PROBLEM

A traffic and offered load models are introduced for customer arrivals and resource demand processes. They form the foundation for the analytical tools needed to address our ultimate loss model: the carried load (blocking) system.

2.1. Traffic and Offered Load Models

We assume that customers arrive to the service system according to a *nonhomogeneous Poisson process* $\{ A(t) \mid t \geq 0 \}$ with rate function $\{ \lambda(t) \mid t \geq 0 \}$, where for all nonnegative integers n, we have

$$\mathsf{P}\{A(t) = n\} = \frac{1}{n!}\left(\int_0^t \lambda(s)\, ds\right)^n \exp\left(-\int_0^t \lambda(s)\, ds\right). \qquad \textbf{(2.1)}$$

The service system informs the arriving customers of their price of admission. We assume that each arriving customer assigns a utility value to the service. This value is private and hence unknown to the service manager. From the manager's perspective, each arriving customer has random, identically distributed utility values that are mutually independent. If the current service price is below an arriving customer's utility value, then the customer accepts that price and joins the system. If the current service price is above an arriving customer's utility value, then that customer rejects the price and does not join the system.

This procedure produces a thinned Poisson process with rate function $\lambda(t, \pi(t))$, where $\pi(t)$ is the service admission price at time t. We assume that the nonhomogeneous Poisson process with this rate function is our *traffic model* for customers requesting this service. We define a *dynamic policy* to be one that is deterministic but a function of our forecasted customer-demand rate function $\lambda(\cdot, \cdot)$ over a finite time horizon given by the interval $[0, T]$. We can now formally state our open-loop *traffic pricing* objective.

Optimization Problem 2.1 (*Traffic Pricing*): Find a dynamic pricing policy, denoted by $\{\pi(t) \mid 0 \leq t \leq T\}$, such that we

$$\text{maximize} \int_0^T \pi(t)\lambda(t, \pi(t))\, dt. \qquad \textbf{(2.2)}$$

This is a calculus of variation problem (see Ewing [7]) that reduces to a static optimization problem at each time t or

$$\pi(t)\lambda(t, \pi(t)) = \max_{z>0} z\, \lambda(t, z). \qquad \textbf{(2.3)}$$

We can restrict z to being positive since $z\lambda(t, z)$ is always a negative number when z is negative. When λ is smooth or differentiable, then π must solve the equation

$$\lambda(t, \pi(t)) + \pi(t)\, \frac{\partial \lambda}{\partial \pi}(t, \pi(t)) = 0 \quad \text{or} \quad \frac{\pi(t)}{\lambda(t, \pi(t))}\, \frac{\partial \lambda}{\partial \pi}(t, \pi(t)) = -1. \qquad \textbf{(2.4)}$$

The last statement says that the pricing policy solution to the traffic pricing problem has the property that the percentage increase in price is matched by the percentage decrease in demand. The traffic price corresponds to the optimal price assuming infinite capacity.

The *offered load model* represents the demand for the resources of a system with infinite capacity and corresponds to the infinite server queue, $\{Q_\infty(t) \mid t \geq 0\}$. Given our traffic model, the number of resources requested at any given time has a Poisson distribution with mean $q(t) \equiv E\left[Q_\infty(t)\right]$,

$$\mathsf{P}\{Q_\infty(t) = i\} = \frac{e^{-q(t)} q(t)^i}{i!}, \qquad \textbf{(2.5)}$$

whenever $Q_\infty(0)$ has a Poisson distribution (which includes $Q_\infty(0) = 0$). If the customer service times are exponential with rate μ, then

$$\frac{d}{dt} q(t) = \lambda(t, \pi(t)) - \mu q(t). \qquad \textbf{(2.6)}$$

The work of Eick, Massey, and Whitt [5] explored the dynamic properties of infinite server queues with nonhomogenous Poisson input. The solution of the offered load model provides a basis for the analysis of the loss system in the next subsection.

2.2. Carried Load Service Model and the Pricing Problem

The *carried load model* has finite capacity and arriving customers are denied service if all system resources are occupied. Throughout this article, the term *loss system* is used interchangeably with *carried load model*. Hampshire et al. [10] explored the connection between the offered load model and loss system in detail. We present a brief summary of this relationship.

Let $\{Q_C(t) | t \geq 0\}$ be the number of customers in a $M_t/M_t/C/C$ queue. The fundamental connection between our offered load and carried load models is given by the *modified offered load approximation* due to Jagerman [11]:

$$\mathsf{P}(Q_C(t) = C) \approx \beta_C\left(q(t)\right) = \mathsf{P}(Q_\infty(t) = C | Q_\infty(t) \leq C), \qquad \textbf{(2.7)}$$

where the formula given by $\beta_C(\cdot)$ is the Erlang blocking formula (see Erlang [6]). Observe that this approximation is *exact* when the Poisson arrival rate is constant and both systems are in steady state equilibrium. Estimates on the error of this approximation can be found in Massey and Whitt [17]. The use of this modified offered load approximation to the blocking rate improves upon the work of Lanning et al. [14]. They use the tail distribution of the time-varying infinite server queue to approximate the blocking rate of a loss system.

Due to the Poisson distribution of the one-dimensional distributions for the $M_t/G/\infty$ queue length process, it is natural to use the *square root staffing* representation found in Jennings, Mandelbaum, Massey, and Whitt [12] of the service capacity

$$C = \ulcorner q + x\sqrt{q}\, \urcorner, \tag{2.8}$$

where q in this section only refers to some generic offered load value and "$\lceil\cdot\rceil$" is the ceiling function. Now, the problem of determining the staffing level is transformed into finding an unknown continuous variable x. Since the mean of a Poisson random variable equals its variance, the term$\sqrt{q}$ can be thought of as the standard deviation of the offered load. Inspired by growing a business to match a corresponding growth in customer demand, we scale the arrival rate by $\eta > 0$ and, hence, the offered load q. Using the square root staffing rule and this scaling, Jagerman [11] showed that

$$\lim_{\eta\to\infty} \sqrt{\eta}\; \beta_{\ulcorner \eta q + x\sqrt{\eta q}\, \urcorner}(\eta q) = \frac{1}{\sqrt{q}}\, \frac{\phi(x)}{\Phi(x)}, \tag{2.9}$$

where

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{and} \quad \Phi(x) = \int_{-\infty}^{x} \phi(y)\, dy. \tag{2.10}$$

In other words, ϕ and Φ are, respectively, the density and cumulative distribution function for some standard Gaussian random variable (mean 0, variance 1). This asymptotic result states that if a manager wants to keep the blocking probability of the system below ϵ, this can be approximated by the relation

$$\frac{1}{\sqrt{q}}\frac{\phi(x)}{\Phi(x)} \leq \epsilon. \tag{2.11}$$

Now, we define $\psi(\cdot)$ to be the functional inverse of the ratio $\phi(\cdot)/\Phi(\cdot)$, or

$$\frac{\phi(\psi(y))}{\Phi(\psi(y))} = y \tag{2.12}$$

for all $y > 0$. Since $\phi(-x) = \phi(x)$ and $\Phi(-x) = 1 - \Phi(x)$, then the ratio $\phi(x)/\Phi(x)$ can be viewed as a hazard function. This function $\psi(\cdot)$ was first introduced in Hampshire et al. [10]. New properties of $\psi(\cdot)$ are critical for the analysis to follow and are presented in the Appendix. One essential property of $\psi(\cdot)$ is that it is a strictly decreasing function, so we have

$$\frac{1}{\sqrt{q}}\frac{\phi(x)}{\Phi(x)} \leq \epsilon \Rightarrow x \geq \psi\left(\epsilon\sqrt{q}\right). \tag{2.13}$$

This suggests that in terms of provisioning for a QoS level of ϵ, the smallest effective value for x is $x = \psi\left(\epsilon\sqrt{q}\right)$. Given C and ϵ, we can show that there exists a unique positive value θ such that

$$C = \ell_\epsilon(\theta), \quad \text{where } \ell_\epsilon(x) \equiv x + \psi\left(\epsilon\sqrt{x}\right)\sqrt{x}, \tag{2.14}$$

where we refer to this constant θ as the *critical offered load.* Asymptotically, θ is the largest offered load that has a blocking probability less than ϵ for C channels.

The critical offered load θ exists as a consequence of Corollary A.2 of the Appendix, which proves that $\ell_\epsilon(x)$ is an increasing function of x.

Now, we introduce the notion of *carried load revenue*. The problem faced by the service manager is to set a pricing policy that maximizes carried load revenue while satisfying the blocking probability constraints for the carried load system. The *carried load pricing problem* reduces to a search among all deterministic pricing policies $\{\pi(t) \mid 0 \le t \le T\}$ that yield a customer arrival rate function $\{\lambda(t, \pi(t)) \mid 0 \le t \le T\}$ that solves the following optimization problem.

Optimization Problem 2.2 (*Carried Load Pricing*): Find a dynamic pricing policy π such that we

$$\text{maximize} \int_0^T \pi(t)\lambda(t,\pi(t))\mathsf{P}\{Q_C(t) < C\}\,dt \quad \text{subject to} \max_{0\le t\le T} \mathsf{P}\{Q_C(t) = C\} \le \epsilon. \tag{2.15}$$

In what follows, we present the *constrained offered load pricing algorithm* that solves the carried load pricing problem under the blocking rate approximation.

3. ANALYZING THE PRICING PROBLEM

In this section we transform the optimization problem of *carried load* dynamic pricing optimization into an optimal control problem for the *offered load*. The key step that facilitates this transformation is approximating the blocking probability for a loss system by a nonlinear function of the offered load. Once the constrained optimal control problem is derived, we can use the tools of calculus of variations to derive necessary conditions of optimality.

3.1. Reduction to a Constrained Optimal Control Problem

Our set of feasible pricing policies are restricted to those for which

$$\max_{0\le t\le T} \mathsf{P}\{Q_C(t) = C\} \le \epsilon. \tag{3.1}$$

All such policies satisfy the following set of inequalities:

$$\begin{aligned}(1-\epsilon)\int_0^T \pi(t)\lambda(t,\pi(t))\,dt &\le \int_0^T \pi(t)\lambda(t,\pi(t))\mathsf{P}\{Q_C(t) < C\}\,dt \\ &\le \int_0^T \pi(t)\lambda(t,\pi(t))\,dt,\end{aligned} \tag{3.2}$$

which gives us the inequality

$$\left| \int_0^T \pi(t)\lambda(t,\pi(t))\,dt - \int_0^T \pi(t)\lambda(t,\pi(t))\mathsf{P}\{Q_C(t) < C\}\,dt \right| \times \left(\int_0^T \pi(t)\lambda(t,\pi(t))\,dt \right)^{-1} \leq \epsilon. \tag{3.3}$$

Thus, we see that for all pricing policies for which the probability of blocking is always less than ϵ, the relative error between the corresponding carried and offered load revenues is also less than ϵ.

This means that for small ϵ, we can approximate the pricing problem by maximizing the traffic revenue given by a pricing policy that satisfies the quality of service constraint or

$$\text{maximize} \int_0^T \pi(t)\lambda(t,\pi(t))\,dt \quad \text{subject to} \quad \max_{0\leq t\leq T} \mathsf{P}\{Q_C(t) = C\} \leq \epsilon. \tag{3.4}$$

Combining the modified offered load approximation with (2.9), the hazard rate limit result of Jagerman [11], we replace our quality of service constraint (3.1) with the offered load constraint of

$$\max_{0\leq t\leq T} q(t) \leq \theta, \quad \text{where } C = \ell_\epsilon(\theta). \tag{3.5}$$

This reduces our analysis of the pricing problem to the following approximation.

Optimization Problem 3.1 (*Constrained Offered Load Pricing*): Find a dynamic pricing policy such that we

$$\text{maximize} \int_0^T \pi(t)\lambda(t,\pi(t))\,dt \quad \text{subject to} \quad \max_{0\leq t\leq T} q(t) \leq \theta, \tag{3.6}$$

where

$$\frac{d}{dt}q(t) = \lambda(t,\pi(t)) - \mu q(t). \tag{3.7}$$

3.2. Lagrangian Dynamics

Next, we reformulate the constrained optimal control problem 3.1 as a classical physics problem of Lagrangian mechanics with the necessary auxiliary variables.

Optimization Problem 3.2 (*Constrained Offered Load Pricing*): Find a dynamic pricing policy, with auxiliary functions of time $p(\cdot)$, $q(\cdot)$, $x(\cdot)$, and $y(\cdot)$ such that

$$\text{maximize} \quad \int_0^T \mathcal{L}\Big(t, p(t), q(t), \overset{\bullet}{q}(t), \pi(t), x(t), y(t)\Big)\, dt, \tag{3.8}$$

where $\overset{\bullet}{q}(t) = dq(t)/dt$ and

$$\mathcal{L}\Big(t, p(t), q(t), \overset{\bullet}{q}(t), \pi(t), x(t), y(t)\Big)$$
$$= \pi(t)\lambda(t,\pi(t)) + p(t)\Big(\overset{\bullet}{q}(t) - \lambda(t,\pi(t)) + \mu q(t)\Big) + x(t)\big(C - \ell_\epsilon(q(t)) - y(t)^2\big). \tag{3.9}$$

Our revenue rate function $\mathcal{L}$, as given by (3.9), plays the role of the *Lagrangian* in physics, but we seek a *greatest* action principle here rather than a least one. The optimized integral for the total revenue as given by (3.8) is called the *action* for the system in classical mechanics. Using the *Euler–Lagrange equations* [13], the optimal $p(\cdot)$, $q(\cdot)$, $\pi(\cdot)$, $x(\cdot)$, and $y(\cdot)$ must satisfy the following set of Euler–Lagrange equations:

$$\frac{\partial \mathcal{L}}{\partial p} = 0 \Longrightarrow \overset{\bullet}{q}(t) = \lambda(t,\pi(t)) - \mu q(t), \tag{3.10}$$

$$\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \overset{\bullet}{q}} = \frac{\partial \mathcal{L}}{\partial q} \Longrightarrow \overset{\bullet}{p}(t) = \mu p(t) - x(t)\ell'_\epsilon(q(t)), \tag{3.11}$$

$$\frac{\partial \mathcal{L}}{\partial \pi} = 0 \Longrightarrow \lambda(t,\pi(t)) + \frac{\partial \lambda}{\partial \pi}(t,\pi(t)) = 0, \tag{3.12}$$

$$\frac{\partial \mathcal{L}}{\partial x} = 0 \Longrightarrow C - \ell_\epsilon(q(t)) = y(t)^2, \tag{3.13}$$

$$\frac{\partial \mathcal{L}}{\partial y} = 0 \Longrightarrow x(t)\cdot y(t) = 0. \tag{3.14}$$

These are the conditions satisfied by any extremal solution. For a local maximum, we must also have

$$(\pi(t) - p(t))\lambda(t,\pi(t)) = \max_{-\infty<z<\infty} (z - p(t))\lambda(t,z) \tag{3.15}$$

and $x(t) \geq 0$, since

$$-x(t)y(t)^2 = \max_{-\infty<z<\infty} -x(t)z^2 = 0. \tag{3.16}$$

These last two results are applications of the Pontryagin principle (see Pontryagin, Boltyanshii, Gamkredlidze, and Mishchenko [20]).

The offered load $q(t)$ and the price per customer $\pi(t)$ are the *state variables*. Their dynamics are constrained by the *Lagrange multiplier functions* $p(t)$ and $x(t)$.

The multiplier $p(t)$ is called the *dual variable* to $q(t)$. In classical mechanics $q(t)$ corresponds to the *position* variable and $p(t)$ corresponds to the *momentum* variable, where

$$p(t) = \frac{\partial \mathcal{L}}{\partial \overset{\bullet}{q}} \Big(t, p(t), q(t), \overset{\bullet}{q}(t), \pi(t), x(t), y(t)\Big) \qquad \textbf{(3.17)}$$

and the terminal condition $p(T) = 0$ holds. Moreover, $p(t)$ creates an *equality* constraint when we optimize the total revenue integral and the corresponding Euler–Lagrange equation (3.11) can be rewritten as

$$\overset{\bullet}{p}(t) = \frac{\partial \mathcal{L}}{\partial q} \Big(t, p(t), q(t), \overset{\bullet}{q}(t), \pi(t), x(t), y(t)\Big). \qquad \textbf{(3.18)}$$

Thus, when optimality is attained, $p(t)$ has the economic interpretation of being the *opportunity cost per customer*. This means that if at time t we introduce a new customer to the queuing system, then the resulting optimal profit derived is, up to first order, the original optimal value minus $p(t)$.

In a similar manner, $x(t)$ creates an *inequality* constraint when we optimize the total revenue integral. The Euler–Lagrange equation for $x(t)$ is equivalent to the desired inequality

$$C - \ell_\epsilon(q(t)) \geq 0. \qquad \textbf{(3.19)}$$

Moreover, $x(t)$ in this context has the economic interpretation of being the *marginal profit rate per channel* since we have by sensitivity analysis (see Section 5) that

$$\begin{aligned} \frac{d\mathcal{L}}{dC}\Big(t, p(t), q(t), \overset{\bullet}{q}(t), \pi(t), x(t), y(t)\Big) &= \frac{\partial \mathcal{L}}{\partial C}\Big(t, p(t), q(t), \overset{\bullet}{q}(t), \pi(t), x(t), y(t)\Big) \\ &= x(t). \qquad \textbf{(3.20)} \end{aligned}$$

Since $y(t)$ defines the inequality constraint, it is called the *slack variable*. The optimal solutions for both $x(t)$ and $y(t)$ make them *complementary variables* since the Euler–Lagrange equation for $y(t)$ yields $x(t)y(t) = 0$. Economically, this means that the marginal optimal revenue per channel is zero whenever there is *no* congestion or $y(t) > 0$.

Moreover, $\ell'_\epsilon(\theta)$ is the *marginal critical channel unit per customer* given the critical offered load of θ since it only makes a nonzero contribution during congestion when $x(t) > 0$, which is equivalent to $y(t) = 0$ or $q(t) = \theta$.

Finally, applying the Pontryagin principle in more detail yields the following optimization relations for $\pi(t)$:

$$(\pi(t) - p(t))\lambda(t, \pi(t)) = \begin{cases} \max_{z \geq 0} \, (z - p(t))\lambda(t, z) & \text{if } y(t) > 0 \\ \max_{z : \lambda(t,z) \leq \mu \cdot \theta} \, (z - p(t))\lambda(t, z) & \text{if } y(t) = 0. \end{cases} \qquad \textbf{(3.21)}$$

The last condition follows from the fact that $y(t) = 0$ if and only if $q(t)$ is at the constraint boundary θ, where its time derivative cannot be positive. Observe that if our demand function is a strictly decreasing function of the price for service, then it follows from (3.21) that $p(t) < \pi(t)$ always holds.

The following theorem provides a probabilistic formulation for $p(t)$ and $q(t)$ and gives us additional insight into their dynamical behavior. If we think of $q(t)$ as looking *backward* into the past as it moves *forward* in time, then $p(t)$ looks *forward* into the future as it moves *backward* in time. The opportunity cost per customer $p(t)$ is anticipating *future* levels of congestion in the system. This is precisely the mechanism that is needed for congestion pricing.

THEOREM 3.3: *Let Σ be a random service time that is exponentially distributed with mean $1/\mu$. We can rewrite the optimal solutions for p and q as*

$$q(t) = q(0)\mathsf{P}\{\Sigma > t\} + \mathsf{E}\left[\int_{t-\Sigma}^{t} \lambda(s,\pi(s))\,ds\right] \quad \text{and}$$

$$p(t) = \ell'_\epsilon(\theta)\mathsf{E}\left[\int_{t}^{t+\Sigma} x(s)\,ds\right], \tag{3.22}$$

where we use the convention that $\lambda(s,\cdot) = 0$ for all $s < 0$ and $x(s) = 0$ for all $s > T$. Moreover, p has the following three properties:

1. *We have $p(t) \geq 0$ for all $t \in [0,T]$.*
2. *If $p(t) = 0$ for some $t \in [0,T]$, then $p(s) = 0$ for all $s \in [t,T]$.*
3. *If $p(t) = 0$ for some $t \in [0,T)$, then $x(s) = 0$ for all $s \in [t,T]$.*

PROOF: By (3.10), the differential equation for q at time t is

$$\frac{d}{dt}q(t) = \lambda(t,\pi(t)) - \mu q(t). \tag{3.23}$$

Solving this inhomogeneous linear equation at time t, when we initialize at time 0, gives us

$$\begin{aligned} q(t) &= q(0)e^{\mu\cdot t} + \int_0^t \lambda(s,\pi(s))e^{-\mu\cdot(t-s)}\,ds \\ &= q(0)e^{\mu\cdot t} + \mathsf{E}\left[\int_0^t \lambda(s,\pi(s))1_{\{\Sigma > t-s\}}\,ds\right] \\ &= q(0)e^{\mu\cdot t} + \mathsf{E}\left[\int_{t-\Sigma}^t \lambda(s,\pi(s))\,ds\right]. \end{aligned}$$

Similarly, by (3.11), the differential equation for p at time t is

$$\frac{d}{dt}p(t) = \mu p(t) - \ell'_\epsilon(\theta)x(t). \tag{3.24}$$

Solving this inhomogeneous linear equation at time T, when we initialize at time t, gives us

$$p(T) = p(t)e^{\mu\cdot(T-t)} - \ell'_\epsilon(\theta)\int_t^T x(s)e^{\mu\cdot(T-s)}\,ds. \tag{3.25}$$

The terminal condition of $p(T) = 0$ and the convention for x gives us

$$\begin{aligned} p(t) &= \ell'_\epsilon(\theta)\int_0^t x(s)\cdot e^{-\mu\cdot(s-t)}\,ds \\ &= \ell'_\epsilon(\theta)\mathsf{E}\left[\int_t^T x(s)1_{\{\Sigma>s-t\}}\,ds\right] \\ &= \ell'_\epsilon(\theta)\mathsf{E}\left[\int_t^{t+\Sigma} x(s)\,ds\right]. \end{aligned}$$

The remaining properties for $p(t)$ follow from the fact that $\ell'_\epsilon(\theta) > 1-\epsilon$ (see Corollary A.2 of the Appendix) and $x(\cdot)$ is a nonnegative function of time. ■

3.3. Derivation of the Algorithm

Now we derive the constrained offered load pricing algorithm. Given the differential equations (3.10) and (3.11) for q and p, we can numerically integrate them as an autonomous system if we can compute π and x from a given pair of p and q.

First, we assume that $\lambda(t,\cdot)$ is a smooth, decreasing, invertible function of the price. Moreover, we assume that for all t, the equation

$$\lambda(t,z) + (z - p(t))\,\frac{\partial\lambda}{\partial\pi}(t,z) = 0 \tag{3.26}$$

has a unique solution in z that we call $\hat{\pi}(t)$, where we also have

$$\big(\hat{\pi}(t) - p(t)\big)\,\lambda(t,\hat{\pi}(t)) = \max_{z\geq 0}\,(z-p(t))\,\lambda(t,z). \tag{3.27}$$

Whenever $y(t) > 0$, it follows that the optimal price $\pi(t) = \hat{\pi}(t)$. We cannot use this solution for $\pi(t)$ whenever $y(t) = 0$ and $\mu\theta < \lambda(t,\hat{\pi}(t))$, since we now have $q(t) = \theta$. Since $\overset{\bullet}{q}(t) \leq 0$, we must have

$$\lambda(t,\pi(t)) \leq \mu\theta < \lambda\left(t,\hat{\pi}(t)\right) \quad\text{or}\quad \hat{\pi}(t) < \lambda(t,\cdot)^{-1}(\mu\theta) \leq \pi(t). \tag{3.28}$$

As a function of z, we are assuming that $(z-p(t))\,\lambda(t,z)$ has a unique critical point at $\hat{\pi}(t)$. Hence, this function must be a strictly decreasing function on the interval $[\hat{\pi}(t),\infty)$. This means that

$$\begin{aligned} \max_{z:\lambda(t,z)\leq\mu\cdot\theta}(z-p(t))\lambda(t,z) &= \max_{z:\lambda(t,z)=\mu\cdot\theta}(z-p(t))\lambda(t,z) \\ &= \big(\lambda(t,\cdot)^{-1}(\mu\theta) - p(t)\big)\,\mu\theta \end{aligned} \tag{3.29}$$

and so $\pi(t) = \lambda(t,\cdot)^{-1}(\mu\theta)$.

To obtain the nonzero expression for $x(t)$, first observe that the sensitivity relation (5.2) holds for any initial time t, where $0 \leq t < T$; we also have

$$\frac{d}{dC}\left(\pi(t)\lambda(t,\pi(t))\right) = x(t). \tag{3.30}$$

Now, observe that whenever we have $x(t) > 0$, it follows that $y(t) = 0$. This means that our system is congested or $\pi(t) = \lambda(t,\cdot)^{-1}(\mu\theta) = \lambda(t,\cdot)^{-1}\left(\mu\ell_{\epsilon}^{-1}(C)\right)$. From this formula, we then obtain

$$\begin{aligned}
x(t) &= \frac{d}{dC}\left\{\lambda(t,\cdot)^{-1}(\mu\theta)\mu\theta\right\} \\
&= \frac{d}{dC}\left\{\lambda(t,\cdot)^{-1}(\mu\ell_{\varepsilon}^{-1}(C))\mu\ell_{\epsilon}^{-1}(C)\right\} \\
&= \left\{\frac{d}{dC}\lambda(t,\cdot)^{-1}(\mu\ell_{\varepsilon}^{-1}(C))\right\}\mu\ell_{\epsilon}^{-1}(C) + \left\{\lambda(t,\cdot)^{-1}(\mu\ell_{\epsilon}^{-1}(C))\mu\,\frac{d}{dC}\ell_{\varepsilon}^{-1}(C)\right\} \\
&= \left\{\frac{d}{dC}\lambda(t,\cdot)^{-1}(\mu\ell_{\varepsilon}^{-1}(C))\right\}\mu\theta + \left\{\pi(t)\mu\frac{d}{dC}\ell_{\epsilon}^{-1}(C)\right\} \\
&= \left[\frac{\mu\theta}{\lambda'(t,\cdot)\circ\lambda(t,\cdot)^{-1}(\mu\ell_{\epsilon}^{-1}(C))} + \pi(t)\right]\mu\frac{d}{dC}\ell_{\epsilon}^{-1}(C) \\
&= \left[\frac{\mu\theta}{\lambda'(t,\pi(t))} + \pi(t)\right]\frac{\mu}{\ell_{\epsilon}'\circ\ell_{\epsilon}^{-1}(C)} \\
&= \left[\frac{\mu\theta}{\lambda'(t,\pi(t))} + \pi(t)\right]\frac{\mu}{\ell_{\epsilon}'(\theta)},
\end{aligned}$$

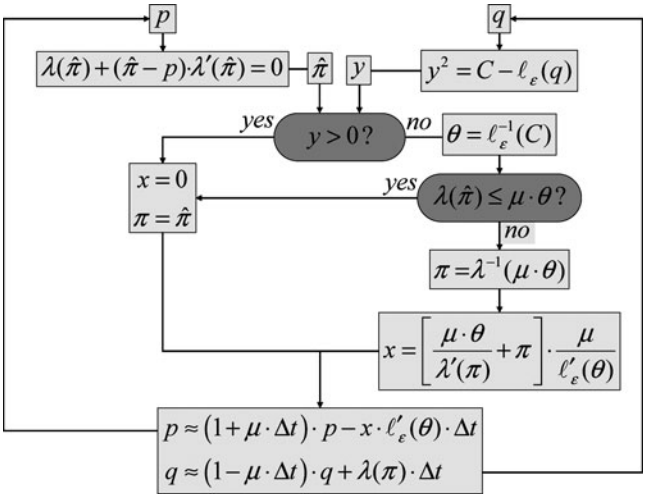
p
q
$\lambda(\hat{\pi})+(\hat{\pi}-p)\cdot\lambda'(\hat{\pi})=0$
$\hat{\pi}$
y
$y^2=C-\ell_{\varepsilon}(q)$
yes
$y>0?$
no
$\theta=\ell_{\varepsilon}^{-1}(C)$
$x=0$
$\pi=\hat{\pi}$
yes
$\lambda(\hat{\pi})\leq\mu\cdot\theta?$
no
$\pi=\lambda^{-1}(\mu\cdot\theta)$
$x=\left[\frac{\mu\cdot\theta}{\lambda'(\pi)}+\pi\right]\cdot\frac{\mu}{\ell_{\varepsilon}'(\theta)}$
$p\approx(1+\mu\cdot\Delta t)\cdot p-x\cdot\ell_{\varepsilon}'(\theta)\cdot\Delta t$
$q\approx(1-\mu\cdot\Delta t)\cdot q+\lambda(\pi)\cdot\Delta t$

FIGURE 1. The constrained offered load pricing algorithm.

where $\lambda'(t,\cdot) = \partial\lambda(t,\cdot)/\partial\pi$. Figure 1 presents a flowchart that summarizes our dynamic pricing algorithm.

4. NUMERICAL EXAMPLE

In this section, we demonstrate the application of our heuristic pricing approach through a numerical case study. The first step is to define the demand function. We assume that the user arrival rate has the following bounded elastic demand function

$$\lambda(t,\pi) = \frac{\gamma(t)}{(\alpha+\beta\pi)^{\sigma}}, \qquad \alpha > 0,\ \beta > 0, \text{and } \sigma > 1, \tag{4.1}$$

where π is some generic price value and

$$\gamma(t) = z[W - (2t/T - 1)^2], \qquad z > 0,\ W > 0. \tag{4.2}$$

This demand model differs from the standard *constant elasticity demand* (CED) function, $\lambda(t,\pi) = \beta\pi^{-\sigma}$, in two aspects. First, the CED function suffers from an anomaly that the revenue rate $\pi\lambda(t,\pi) = \beta\pi^{1-\sigma}$ can grow without bound as the price π approaches zero. With (4.1), we avoid this difficulty since

$$\max_{\pi\geq 0} \lambda(t,\pi) = \lambda(t,0) = \gamma(t)\alpha^{-\sigma}.$$

Note that our demand is *unimodal* for a fixed price π and it reaches its unique peak in the middle of the planning horizon ($t = T/2$).

From (4.1), we have for each t,

$$\frac{\pi(t)\lambda'(t,\pi(t))}{\lambda(t,\pi(t))} = -\sigma\frac{\beta\pi(t)}{\alpha+\beta\pi(t)}. \tag{4.3}$$

So, at any given price, a larger value of σ implies a larger percentage change of demand ($\triangle\lambda/\lambda$) with respect to the percentage change of price ($\triangle\pi/\pi$). Since the ratio on the right-hand side of (4.3) is a number between 0 and 1, so the price elasticity of demand for the percentage change in demand is bounded above in absolute value by σ. Also, in the limit as the price approaches infinity, the elasticity of demand actually becomes σ. With this demand function, the *optimal traffic price* (i.e., the price that maximizes the revenue in the absence of a capacity constraint, becomes)

$$\hat{\pi}_0 = \frac{\alpha}{\beta(\sigma-1)}, \tag{4.4}$$

which is constant despite demand variation.

Whenever $q(t) < \theta$, the optimal price $\pi(t)$ equals $\hat{\pi}(t)$ and

$$\hat{\pi}(t) = \hat{\pi}_0 + \frac{p(t)}{1-1/\sigma}, \tag{4.5}$$

where the corresponding arrival rate is

$$\lambda(t, \hat{\pi}(t)) = (1 - 1/\sigma)^{\sigma} \lambda(t, p(t)). \tag{4.6}$$

The opportunity cost per customer $p(t)$ can then be viewed as being proportional to a "future congestion tax" that raises the price $\hat{\pi}(t)$ above the optimal traffic price to keep the constrained offered load from exceeding the critical offered load θ. Our approach is forward-looking: It anticipates that users admitted during noncongested periods have a positive probability to contribute to future congestion.

Furthermore, when $q(t) < \theta$, we have $x(t) = 0$. The dynamics of $p(t)$ simply follow the equation $\overset{\bullet}{p} = \mu p$, and so

$$p(t) = p(0)e^{\mu t}. \tag{4.7}$$

Now, let t_1 be the first time s, where $q(s) = \theta$. The first step of our algorithm is to jointly find t_1 and $p(0)$.

Step 1: Solve for $p(0)$ and t_1, where

$$p(0) = \frac{e^{-\mu t_1}}{\beta} \left(\left(1 - \frac{1}{\sigma} \right) \delta(t_1) - \alpha \right) \tag{4.8}$$

and

$$\theta = q(0)e^{-\mu t_1} + \left(1 - \frac{1}{\sigma} \right)^{\sigma} \int_0^{t_1} \lambda \left(p(0)e^{\mu s}, s \right) e^{-\mu \cdot (t_1 - s)} \, ds, \tag{4.9}$$

where

$$\delta(t) = \left(\frac{\gamma(t)}{\mu \cdot \theta} \right)^{1/\sigma}.$$

Equation (4.8) is equivalent to $\lambda(t_1, \hat{\pi}) = \mu\theta$ and we rewrite this as

$$\left(1 - \frac{1}{\sigma} \right)^{\sigma} \lambda(t_1, p(0)e^{\mu t_1}) = \mu\theta. \tag{4.10}$$

The second equation follows from having $q(t_1) = \theta$.

The second step of our algorithm is to find the last time that our constrained offered load equals θ and call it t_2.

Step 2: Solve for t_2, where

$$\gamma(t_2) = \mu\theta \left(\frac{\alpha}{1 - 1/\sigma} \right)^{\sigma}. \tag{4.11}$$

Due to the unimodal structure of the demand, there is at most one congestion period. It follows that $p(t_2) = 0$, which implies that $\lambda(t_2, \hat{\pi}_0) = \mu\theta$. The third step is to compute the opportunity cost per customer.

Step 3: Solve $\overset{\bullet}{p} = \mu \cdot p - \ell'_\epsilon(\theta)x$ backward in time with $p(t_2) = 0$ and setting

$$x(t) = \begin{cases} 0 & \text{when } t < t_1 \\ (\delta(t)\,(1 - 1/\sigma) - \alpha)\,[\mu/(\beta \cdot \ell'_\epsilon(\theta))] & \text{when } t_1 \le t < t_2 \\ 0 & \text{when } t_2 \le t \le T. \end{cases} \tag{4.12}$$

The fourth and final step is to compute the constrained offered load.

Step 4: Solve $\overset{\bullet}{q} = \lambda(\pi) - \mu q$ forward in time given $q(0)$ and

$$\hat{\pi}(t) = \begin{cases} \hat{\pi}_0 + p(t)/(1 - 1/\sigma) & \text{when } t < t_1 \\ (\delta(t) - \alpha)\,/\beta & \text{when } t_1 \le t < t_2 \\ \hat{\pi}_0 & \text{when } t_2 \le t \le T. \end{cases} \tag{4.13}$$

4.1. Static Pricing Policy

We can compare our dynamic pricing policy to a *static* pricing policy. Let $\bar{\pi}$ be the fixed price that maximizes the total revenue subject to the uniform blocking constraint. Using the constrained offered load formulation, the optimization problem becomes the following.

Optimization Problem 4.1 (*Constrained Offered Load Static Pricing*): Find a price $\bar{\pi}$ such that we

$$\text{maximize } \bar{\pi} \int_0^T \lambda(t, \bar{\pi})\, dt \quad \text{subject to} \quad \max_{0 \le t \le T} q(t) \le \theta, \tag{4.14}$$

where

$$\frac{d}{dt} q(t) = \lambda(t, \bar{\pi}) - \mu q(t). \tag{4.15}$$

The solution to this optimization problem yields an optimal static price equal to traffic price $\bar{\pi} = \hat{\pi}_0$ whenever $\lambda\left(\bar{t}, \hat{\pi}_0\right) \le \mu\theta$, otherwise the optimal static price solves the equation $\lambda(\bar{t}, \bar{\pi}) = \mu\theta$, where $\bar{t}$ is the time of the peak offered load under the traffic price $\hat{\pi}_0$. The static price is set to ensure that the maximum of the offered load stays below the critical offered load θ.

4.2. Myopic Pricing Policy

We also can compare our dynamic pricing policy to a *myopic* pricing policy. This policy is purely reactive and does not anticipate any future congestion. Under the myopic policy, the traffic price is charged until the critical offered load is reached. When the QoS constraint becomes tight, the price is raised to maintain the constraint.

This occurs when the constrained offered load reaches its critical value θ. The myopic pricing policy is defined by

$$\pi^*(t) = \begin{cases} \hat{\pi}_0 & \text{when } t < t_1^* \\ (\delta(t) - \alpha)/\beta & \text{when } t_1^* \le t < t_2^* \\ \hat{\pi}_0 & \text{when } t_2^* \le t \le T, \end{cases} \qquad \textbf{(4.16)}$$

where t_1^* and t_2^* are the times of the start and end of congestion, respectively.

4.3. The Base Case

For our numerical base case we set the time horizon $T = 100$, capacity $C = 50$, and the target blocking rate $\epsilon = 0.01$ or 1%. The critical offered load for this case is $\theta = 37.98$. This means that

$$\ell_\epsilon(q(t)) = q(t) + \psi(\epsilon\sqrt{q(t)})\sqrt{q(t)} \le \ell_\epsilon(\theta) = C = 50 \quad \text{if and only if} \quad q(t) \le \theta. \qquad \textbf{(4.17)}$$

We assume that the average customer service time $1/\mu$ equals 30 time units. Finally, for the demand parameters, we let

$$\alpha = \beta = 0.05, \qquad \sigma = 2.0, \qquad z = 1.5, \quad \text{and} \quad W = 1.$$

Figure 2 plots the optimal pricing policies for the dynamic, myopic, and static cases. Each of these policies induce a corresponding arrival rate shown in the left-hand graph of Figure 3. Assuming the same fixed service rate, arrival rates are used to compute the constrained offered load via the ordinary differential equation (2.6). Under each pricing policy, the constrained offered load is displayed in the right-hand graph of Figure 3. In all cases, the time of the peak demand precedes the time of the peak offered load. This is a general feature of a time-varying queuing system.

The difference between the dynamic price and the traffic price is called the *future congestion tax*. It captures the amount an arriving customer pays in excess of the traffic price due to the blocking constraint.

Under the dynamic pricing policy, the arrival rate reaches its peak at time $t = 15$. Starting from $q(0) = 0$, the constrained offered load $q(t)$ grows until it reaches the critical offered load $\theta = 37.98$ at time $t_1 = 35.95$ (see the right-hand graph of Fig. 3). Once the offered load reaches the critical value, the price is set so that demand is constant and equals $\mu\theta$. The offered load remains at the critical offered load until $\lambda(t, \hat{\pi}(t)) < \mu\theta$ and the constrained offered load starts to fall off the boundary of the critical offered load value θ. When the QoS constraint becomes nonbinding and there is no congestion, the opportunity cost per customer equals zero and the price equals the traffic price.

Under the static pricing policy, the price $\bar{\pi} = 18.33$ is initially larger than the dynamic price. This leads to an arrival rate that is initially smaller than under the dynamic pricing policy. Hence, the offered load under the static price grows more

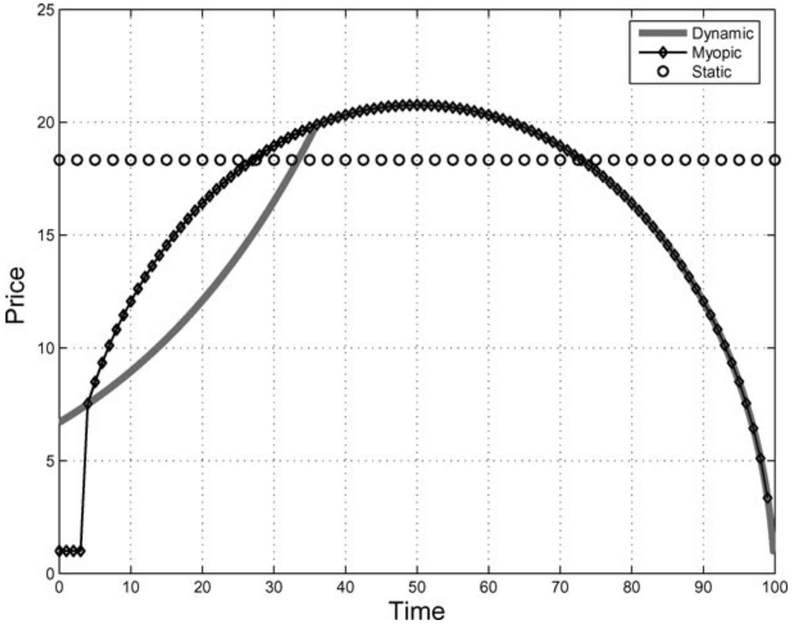

FIGURE 2. Dynamic, myopic, and static price.

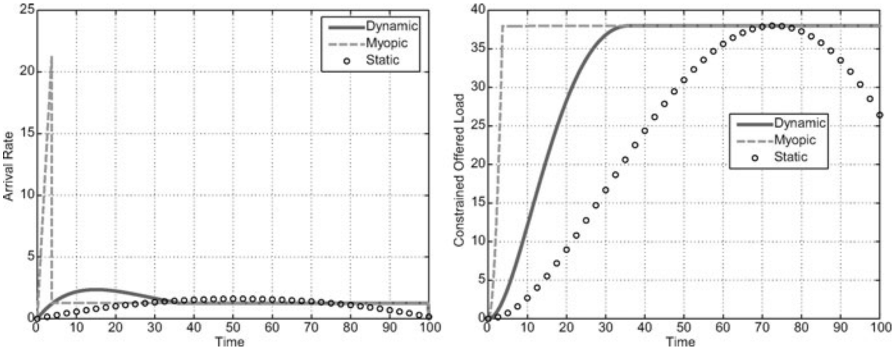

FIGURE 3. (*Left*) Arrival rate and (*right*) constrained offered load.

slowly than under the dynamic pricing policy as seen in the right-hand graph of Figure 3. The offered load under the static pricing policy reaches the critical threshold later ($\bar{t} = 72.9827$), as shown in Figure 3.

The price under the myopic policy is initially less than under the dynamic pricing policy. The low initial price leads to a large initial arrival rate. This induces excessive demand, causing the myopic constrained offered load to rise quickly. The critical threshold is reached sooner ($t_1^* = 3.6768$) than under the optimal dynamic pricing policy ($t_1 = 35.95$). Our dynamic policy efficiently smooths the demand during the precongestion period relative to the myopic policy.

We now return to the *original* loss system of Optimization Problem 2.2 and compare the revenue generated under each of the pricing policies. The total revenue under each pricing policy is computed for the *original* loss system with C channels. The transient state probability distribution is computed using the Kolmogorov forward equations. The resulting revenues under the dynamic, static, and myopic policies are 2147.7, 1955.3, and 2035, respectively. The percent revenue gains under the dynamic policy relative to the static and myopic are 8.95% and 5.35%, respectively.

In the next subsection we study the sensitivity of the percent revenue gain under the dynamic policy relative to the static and myopic policies as a function of the average service time and the price elasticity.

4.4. Sensitivity Analysis

We compare the percent revenue gain under the dynamic pricing policy relative to the static and myopic policies as a function of the average service time in the left-hand graph of Figure 4. In the right-hand graph of Figure 4, the beginning and end times of the congestion period are plotted as a function of the average service time.

As the average service time becomes smaller, the percent revenue gain of using the dynamic policy relative to static policy increases. During the precongestion period, the static price is larger than the dynamic price, so the static policy unduly depresses the arrival rate. If the average service time is small, then the system throughput is larger under the dynamic pricing policy, which leads to more revenue.

The onset of congestion under the myopic policy occurs earlier than under the dynamic policy. We observe that as the average service times increase, the percent revenue gain under the dynamic pricing policy increases. This is due to the fact that the opportunity cost of admitting an additional customer is greater when the average

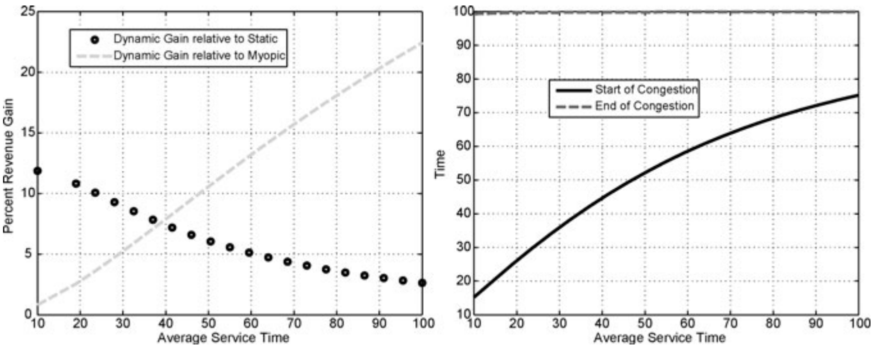

FIGURE 4. (*Left*) Percentage gain in revenue using dynamic pricing as a function of average service time and (*right*) congestion period as a function of average service time.

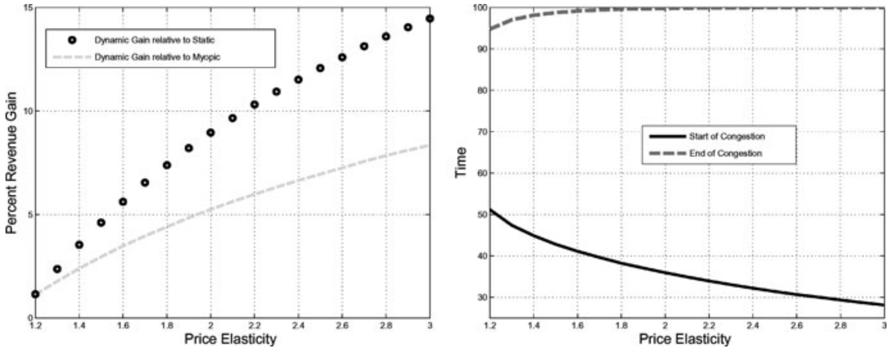

FIGURE 5. (*Left*) Percentage gain in revenue using dynamic pricing as a function of price elasticity (*right*) congestion period as a function of price elasticity.

service time is larger. Similarly, when the average service time is small, then the opportunity cost of admitting an additional customer is small. The myopic policy does not consider this opportunity cost, so when the opportunity cost is high, the dynamic policy yields a larger percentage revenue gain. The congestion period becomes smaller as the average service time increases. Thus, the time interval on which dynamic and myopic prices are equal decreases as the average service time increases.

In the left-hand graph of Figure 5, we compare the percent revenue gain under the dynamic pricing policy relative to the static and myopic policies as a function of price elasticity. In the right-hand graph of Figure 5, the beginning and end times of the congestion period under the dynamic and myopic policies are plotted as a function of price elasticity. As the customers become more price sensitive, the percentage revenue gain of the dynamic policy over both the static and myopic policies increases. The percentage revenue gain relative to the static policy grows faster, as a function of the elasticity, than its revenue gain relative to the myopic policy. In the left-hand graph of Figure 5, we observe that as the price elasticity σ grows, the congestion periods become longer, starting earlier and ending later.

5. SENSITIVITY RESULTS AND REVENUE TRADE-OFFS

The optimal constrained offered load revenue can be written as

$$
\begin{aligned}
R(T) &\equiv \int_0^T \mathcal{L}\Big(t, p(t), q(t), \dot{q}(t), \pi(t), x(t), y(t)\Big)\, dt \\
&= \int_0^T \pi(t)\lambda(t, \pi(t))\, dt. \qquad \textbf{(5.1)}
\end{aligned}
$$

Now, we quantify the sensitivity of the optimal revenue to various parameters of the carried load system.

THEOREM 5.1: *The optimal marginal revenues per channel and due to productivity gains are given respectively by the formulas*

$$\frac{dR}{dC}(T) = \int_0^T x(t)\,dt \quad \text{and} \quad \frac{dR}{d\mu}(T) = \int_0^T p(t)q(t)\,dt. \tag{5.2}$$

Moreover, the optimal marginal revenue per quality of service level ϵ is given by

$$\frac{dR}{d\epsilon}(T) = -\frac{\partial}{\partial\epsilon}\ell_\epsilon(\theta)\frac{dR}{dC}(T), \tag{5.3}$$

where in terms of $C = \ell_\epsilon(\theta)$, we have

$$-\frac{\partial}{\partial\epsilon}\ell_\epsilon(\theta) = \frac{\theta}{\epsilon\,(C-\theta(1-\epsilon))}. \tag{5.4}$$

PROOF: The envelope theorem (see Dixit [4]) leads to

$$\begin{aligned}\frac{dR}{dC}(T) &= \int_0^T \frac{\partial}{\partial C}\mathcal{L}\left(t, p(t), q(t), \overset{\bullet}{q}(t), \pi(t), x(t), y(t)\right)\,dt\\ &= \int_0^T x(t)\,dt, \end{aligned}\tag{5.5}$$

$$\begin{aligned}\frac{dR}{d\epsilon}(T) &= \int_0^T \frac{\partial}{\partial \epsilon}\mathcal{L}\left(t, p(t), q(t), \overset{\bullet}{q}(t), \pi(t), x(t), y(t)\right)\,dt\\ &= -\frac{\partial}{\partial\epsilon}\ell_\epsilon(\theta)\int_0^T x(t)\,dt, \end{aligned}\tag{5.6}$$

and

$$\begin{aligned}\frac{dR}{d\mu}(T) &= \int_0^T \frac{\partial}{\partial \mu}\mathcal{L}\Big(t, p(t), q(t), \overset{\bullet}{q}(t), \pi(t), x(t), y(t)\Big)\,dt\\ &= \int_0^T p(t)q(t)\,dt. \end{aligned}\tag{5.7}$$

Recalling that

$$C = \ell_\epsilon(\theta) = \theta + \psi\left(\epsilon\sqrt{\theta}\right)\sqrt{\theta},$$

we then have

$$
\begin{aligned}
-\frac{\partial}{\partial\epsilon}\ell_\epsilon(\theta) &= -\frac{\partial}{\partial\epsilon}\left[\theta + \psi\left(\epsilon\sqrt{\theta}\right)\sqrt{\theta}\right] \\
&= -\psi'\left(\epsilon\sqrt{\theta}\right)\theta \\
&= \frac{\theta}{\epsilon\sqrt{\theta}\left(\epsilon\sqrt{\theta} + \psi\left(\epsilon\sqrt{\theta}\right)\right)} \\
&= \frac{\theta}{\epsilon\left(\epsilon\theta + \psi\left(\epsilon\sqrt{\theta}\right)\sqrt{\theta}\right)} \\
&= \frac{\theta}{\epsilon\left(C - (1-\epsilon)\theta\right)}, \qquad \textbf{(5.8)}
\end{aligned}
$$

which completes the proof. ■

Since dR/dC equals the change in revenue due to an increase in the capacity and $dR/d\epsilon$ equals the change in revenue due to an incremental increase in the QoS target ϵ, Theorem 5.1 quantifies the trade-off between these two quantities. Now, we rewrite (5.3) in terms of percentage changes, which captures the elasticity of these parameters

$$
\frac{dR}{R}\left(\frac{d\epsilon}{\epsilon}\right)^{-1} = \frac{\theta}{C\left(C - \theta(1-\epsilon)\right)}\frac{dR}{R}\left(\frac{dC}{C}\right)^{-1}. \qquad \textbf{(5.9)}
$$

This suggests a trade-off for making a percentage change in the optimal revenue between the percentage change in ϵ and the percentage change in capacity C:

$$
\frac{\Delta\epsilon}{\epsilon} \approx \Gamma(C,\epsilon)\frac{\Delta C}{C}, \quad \text{where} \quad \Gamma(C,\epsilon) \equiv \frac{C\left(C - \theta(1-\epsilon)\right)}{\theta} \qquad \textbf{(5.10)}
$$

is called the *QoS–capacity efficiency ratio.*

This result implies that the percentage increase in optimal revenue obtained by a percentage increase in capacity can also be obtained by a percentage increase in the QoS level ϵ that is a factor of $\Gamma(C,\epsilon)$ times the capacity percentage increase. Moreover, this factor has the following asymptotic behavior for large C.

PROPOSITION 5.2: *For all* $0 < \epsilon < 1$, *we have*

$$
\lim_{C\to\infty}\Gamma(C,\epsilon) = \frac{1}{\epsilon} - 1. \qquad \textbf{(5.11)}
$$

Proof: By Corollary A.2 of the Appendix we know that $\ell_\epsilon(x)$ is an increasing function. Therefore, $C \to \infty$ implies that $\theta \to \infty$. Rewriting $\Gamma(C, \epsilon)$, we have

$$\frac{C\,(C - \theta(1-\epsilon))}{\theta} = \frac{\left(\theta + \psi\left(\epsilon\sqrt{\theta}\right)\sqrt{\theta}\right)\left(\theta\epsilon + \psi\left(\epsilon\sqrt{\theta}\right)\sqrt{\theta}\right)}{\theta} \tag{5.12}$$

$$= \left(\frac{1}{\epsilon} + \frac{\psi\left(\epsilon\sqrt{\theta}\right)}{\epsilon\sqrt{\theta}}\right)\epsilon\sqrt{\theta}\left(\epsilon\sqrt{\theta} + \psi\left(\epsilon\sqrt{\theta}\right)\right). \tag{5.13}$$

Using Theorem A.1, we have

$$\lim_{x\to\infty} \frac{\psi(x)}{x} = -1 \quad \text{and} \quad \lim_{x\to\infty} x \cdot (x + \psi(x)) = 1, \tag{5.14}$$

which completes the proof. ■

This gives us a simple rule of thumb for estimating the revenue trade-off between increasing the capacity and degrading the QoS. In order to increase revenue, a manager might consider increasing the relative amount of capacity C or increasing the relative amount of blocking ϵ. The *QoS–capacity efficiency ratio* for the base case of the numerical example is $\Gamma(C, \epsilon) = 16.32$. To induce the same percentage change in revenue, the percentage change in the blocking probability must be 16.32 times the percentage change in capacity. In the current example, a 10% percent increase in capacity produces the same percentage gain in revenue as a 163.2% percent increase in the probability of blocking ϵ. This implies that increasing C from 50 to 55 induces the same percent increase in revenue as changing the blocking probability ϵ from 1% to 2.163%.

Before making the decision to increase capacity or degrade the service quality, the manager must consider the costs of these two actions. The cost of increasing the relative capacity can be calculated directly, whereas the cost of degrading the service quality can be computed indirectly by measuring customer retention. The *QoS–capacity efficiency ratio* is a simple tool to facilitate this decision-making process.

6. SUMMARY AND CONCLUSIONS

We provide a dynamic pricing heuristic for time-varying loss systems that maximizes revenue while meeting QoS targets over a finite time horizon. The modified offered load (MOL) approximation and the ψ function, inverse of the normal hazard rate, are tools that convert the uniform QoS blocking constraint into a threshold constraint for the mean of the nonstationary infinite server queue. The resulting dynamic optimization problem is analyzed with calculus of variations, and we derive necessary conditions for the optimal pricing policy.

The opportunity cost per customer captures the “future congestion tax” levied on the system when admitting an additional customer. Using the constrained offered load and the opportunity cost per customer, our solution is able to anticipate future service congestion and set the present price accordingly.

Our dynamic pricing policy is shown numerically to produce more revenue than the static pricing policy. The static policy selects the fixed price that maximizes revenue while meeting the offered load constraint. As the average service time decreases and the price elasticity increases, so does the percent revenue gain generated by our dynamic policy.

Our dynamic pricing policy is also shown numerically to produce more revenue than a myopic pricing policy. The myopic policy does not consider future congestion. The myopic policy considers only the constrained offered load, not the opportunity cost per customer. As the average customer service time increases and the price elasticity increases, so does the percent revenue gain generated by our dynamic policy.

Finally, the revenue trade-off between increasing capacity and degrading the quality of service is considered. Increasing capacity allows more customers to be admitted for service, generating more revenue. Increasing the QoS target ϵ increases the critical offered load θ, allowing more customers to enter into service. The QoS-capacity efficiency ratio captures this revenue trade-off.

Acknowledgment

First and second authors supported by NSF grant DMI-0323668.

References

1. Bailey, J. & McKnight, L. (1997). *Internet economics*. Cambridge, MA: The MIT Press.
2. Courcoubetis, C.A., Dimakis, A., & Reiman, M.I. (2001). Providing band width guarantees over a best-effort network: call admission and pricing. *Proceedings of IEEE INFOCOM 2001*, pp. 459–467.
3. Dewan, S. & Mendelson, H. (1990). User delay cost and internal pricing for a service facility. *Management Science* 36(12): 1502–1517.
4. Dixit, A.K. (1990). *Optimization in economic theory*. Oxford: Oxford University Press.
5. Eick, S., Massey, W.A. & Whitt, W. (1993). The physics of the $M(t)/G/\infty$ queue. *Operations Research* 41(4): 400–408.
6. Erlang, A.K. (1918). Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges. *The Post Office Electrical Engineers' Journal* 10: 189–197.
7. Ewing, G.M. (1985). *Calculus of variations with applications*. New York: Dover Publications.
8. Fan-Orzechowski, X. & Feinberg, E.A. (2007). Optimality of randomized trunk reservation for a problem with multiple constraints. *Probability in the Engineering and Informational Sciences* 21(2): 189–200.
9. Hampshire, R.C. (2007). Dynamic queueing models for the operations management of communication services. *Ph.D. dissertation*, Princeton University, Princeton, NJ.
10. Hampshire R.C., Massey W.A., Mitra D. & Wang, Q. (2002). Provisioning of bandwidth sharing and exchange. In Telecommunications network design and economics and management: Selected proceedings of the 6th INFORMS telecommunications conferences. G. Anandalingam & S. Raghavan (eds.). Boston: Kluwer Academic, pp. 207–226.
11. Jagerman, D.L. (1975). Nonstationary blocking in telephone traffic. *Bell System Technical Journal* 54: 625–661.
12. Jennings, O.B., Mandelbaum, A., Massey, W.A. & Whitt, W. (1996). Server staffing to meet time-varying demand. Management Science 42(10): 1383–1394.

13. Lanczos, C. (1970). *The variational principles of mechanics*, 4th ed. New York: Dover Publications.
14. Lanning, S.G., Massey, W.A., Rider, B. & Wang, Q. (1999). Optimal pricing in queuing systems with quality of service constraints. In *Proceedings of the 16th International Teletraffic Congress*, Edinburgh, UK, pp. 747–756.
15. Mackie-Mason, J.F. & Varian, H. (1995). Pricing congestible network resources. *IEEE Journal on Selected Areas in Communications* 13(7): 1141–1149.
16. Maglaras, C. & Zeevi, A. (2005). Pricing and design of differentiated services: approximate analysis and structural insights. *Operations Research* 53: 242–262.
17. Massey, W.A. & Whitt, W. (1994). An analysis of the modified offered load approximation for the nonstationary Erlang loss model. *Annals of Applied Probability* 4(4): 1145–1160.
18. Mendelson, H. (1985). Pricing computer services: queuing effects. *Communications of the ACM* 28(3): 312–321.
19. Mendelson, H. & Whang, S. (1990). Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Operations Research* 38(5): 870–883.
20. Pontryagin, L.S., Boltyanshii, V.G., Gamkredlidze, R.V. & Mishchenko, E.F. (1962). *The mathematical theory of optimal processes*. New York: Wiley.
21. Stidham, S. (1992). Pricing and capacity decisions for a service facility: stability and multiple local optima. *Management Science* 38(8): 1121–1139.
22. Wang, Q., Peha, J.M. & Sirbu, M.A. (1997). Optimal pricing for integrated services networks. L. W. McKnight & J.P. Bailey (eds.), *Internet Economics*, Cambridge, MA: The MIT Press, pp. 352–376.
23. Westland, J.C. (1992). Congestion and network externalities in the short run pricing of information systems services. *Management Science* 38(6): 992–1099.
24. Yoon, S. & Lewis, M.E. (2004). Optimal pricing and admission control in a queueing system with periodically varying parameters. *Queueing Systems: Theory and Applications* 47(3): 177–199.

APPENDIX

Properties of the ψ function

The ψ function is introduced in Hampshire et al. [10], where many of its properties are derived. Below, we state and prove some additional properties that are used to further the analysis in this article.

THEOREM A.1: *For all $x > 0$, we have*

$$0 > \psi(x) + x - \frac{1}{x} > \frac{-1}{x^3 + x}. \tag{A.1}$$

PROOF OF THEOREM A.1: First, we show that $\psi(x) + x > 0$ for all $x > 0$ and converges to zero as $x \to +\infty$. Since $\phi'(y) = -y\phi(y)$, we have

$$\psi(x) + x = y + \frac{\phi(y)}{\Phi(y)} = \frac{y\Phi(y) + \phi(y)}{\Phi(y)} = \frac{\int_{-\infty}^{y} \Phi(z)\,dz}{\Phi(y)} > 0. \tag{A.2}$$

We now have $\psi(x) + x > 0$ for all positive x. If we set $\psi(x) = y$, then $x \to +\infty$ is equivalent to $y \to -\infty$. We then have $\lim_{x\to+\infty} \psi(x) + x = 0$, since by Hopital's rule,

$$\lim_{x\to\infty} \psi(x) + x = \lim_{y\to-\infty} \frac{\int_{-\infty}^{y} \Phi(z)\,dz}{\Phi(y)} = \lim_{y\to-\infty} \frac{\Phi(y)}{\phi(y)} = \lim_{y\to-\infty} \frac{\phi(y)}{-y\cdot\phi(y)} = \lim_{y\to-\infty} \frac{1}{-y} = 0. \tag{A.3}$$

Now, let

$$h(x) \equiv \psi(x) + x - \frac{1}{x}. \tag{A.4}$$

We have $\lim_{x\to+\infty} h(x) = 0$ and differentiating h gives us

$$h'(x) = \frac{-1}{x(x+\psi(x))} + 1 + \frac{1}{x^2} = \frac{h(x)}{x+\psi(x)} + \frac{1}{x^2}. \tag{A.5}$$

If $h(x) \geq 0$, then $h'(x) > 0$. This contradicts h converging to zero as $x \to +\infty$, so we must have $h(x) < 0$ for all $x > 0$.

Differentiating h a second time gives us

$$h''(x) = \frac{-h(x)}{(x+\psi(x))^2}\left(1 + \frac{-1}{x(x+\psi(x))}\right) + \frac{h'(x)}{x+\psi(x)} - \frac{2}{x^3} \tag{A.6}$$

$$= \frac{-h(x)^2}{(x+\psi(x))^3} + \frac{h'(x)}{x+\psi(x)} - \frac{2}{x^3}. \tag{A.7}$$

If we set $h'(x) = 0$, then, by (A.7), we must have $h''(x) < 0$. Hence, extreme points for h are always local maxima.

If $h'(x) \leq 0$ for some x, then h is locally decreasing. Since h is negative and convergent to zero for large x, then h must have a local minimum point, which is a contradiction. Therefore, $h'(x) > 0$ for all $x > 0$, and so

$$\frac{h(x)}{x+\psi(x)} + \frac{1}{x^2} > 0. \tag{A.8}$$

This gives us

$$h(x) > \frac{-1}{x^2}(x+\psi(x)) = \frac{-1}{x^2}\left(h(x) + \frac{1}{x}\right). \tag{A.9}$$

Finally, when resolve the inequality for h, we obtain

$$h(x) > \frac{-1}{x^3+x}, \tag{A.10}$$

which completes the proof. ■

Corollary A.2: *The following statements are true and equivalent:*

1. *The function $f(x) = \psi(x) + x$ is positive, decreasing, and convex.*
2. *The function $g(x) = x(\psi(x) + x)$ is positive, increasing, and less than* 1.

Moreover, if we let

$$\ell_\epsilon(x) \equiv x + \psi\left(\epsilon\sqrt{x}\right)\sqrt{x}, \tag{A.11}$$

then ℓ_ϵ is an increasing function with $\ell'_\epsilon(x) > 1 - \epsilon$.

Proof: The equivalence of the statements follows from the identity

$$f'(x) = \frac{-1}{g(x)} + 1. \tag{A.12}$$

Now, it remains only to prove the second statement.

Differentiating g gives us

$$g'(x) = \frac{g(x)}{x} - \frac{x}{g(x)} + x. \tag{A.13}$$

If we do it a second time, then

$$g''(x) = -\frac{g(x)}{x^2} - \frac{1}{g(x)} + \frac{g'(x)}{x} + \frac{x\,g'(x)}{g(x)^2} + 1. \tag{A.14}$$

If $g'(x) = 0$, then

$$g''(x) = -\frac{g(x)}{x^2} - \frac{1}{g(x)} + 1 < 0, \tag{A.15}$$

since (A.1) implies that $g(x) < 1$. Hence, all extreme points of g must be local maxima. If $g'(x) \leq 0$ for some x, then, by (A.15), it follows that g must be locally decreasing after x. However, g is positive and converges to 1 by (A.1), so g must have after x a local minimum. This is a contradiction, so $g'(x) > 0$ for all $x > 0$. Therefore, g is a strictly increasing function and this completes the proof. ■