Admission Policies in Loss Queueing Models with Heterogeneous Arrivals

E. Carrizosa • E. Conde • M. Muñoz-Márquez
Facultad de Matemáticas, Universidad de Sevilla, Tarfia s/n, 41012 Sevilla, Spain
Facultad de Matemáticas, Universidad de Sevilla, Tarfia s/n, 41012 Sevilla, Spain
Departamento de Matemáticas, Universidad de Cádiz, Sacramento 82, Escuela Politécnica de Cádiz, 11002 Cádiz, Spain

In this paper we consider a loss system where the arrivals can be classified into different groups according to their arrival rate and expected service time. While the standard admission policy consists of rejecting only those customers who arrive when all servers are busy, we address the problem of finding the optimal static admission policy (with respect to a given reward structure) when customers can be discriminated according to the group they belong to, thus customers of some groups might be automatically rejected (even if some servers remain idle) in order to enhance the global efficiency of the system. The optimality of a $c\mu$-rule is shown, from which finite-time algorithms for the one- and two-server cases are derived.
(*Algorithms; Multichannel Queues; Nonlinear Programming; Optimization*)

1. Introduction

Motivated by real-world applications, there has been an increasing interest in the optimal design and control of queueing models (see, e.g., Crabill et al. 1977, Harel 1990, Hillier and Lieberman 1990, Mendelson and Whang 1990, Viscolani 1993, Walrand 1988, Zipkin 1986, and the references therein), where the set of parameters that optimize a certain performance measure is sought. In these problems, the controllable parameters are usually the number of servers, the arrival and service rate, and the capacity of the system.

Finding the optimal values of the controllable parameters is usually reduced to solving a mathematical program, where the controllable parameters play the role of decision variables, and the objective function to be optimized is the performance measure. The determination of such optimal parameters provides operation rules that optimize the performance of the system.

In this paper we address a design problem for a loss system with heterogeneous arrivals, where it is sought the optimal coverage with respect to a given cost structure: any accepted (respectively, rejected) customer induces a cost or disutility r (respectively, $\hat{r}$), with $r < \hat{r}$, and the performance measure is the expected disutility per unit time.

A direct application of this model appears when one has a queueing system with hierarchical service facilities: a primary and a back-up facility. Suppose that the nature of the service is such that no queue is allowed at the primary facility; hence, any customer finding the primary facility busy must be rerouted and served at a higher cost by the back-up system. Then, minimizing the overall cost does not mean allocating to the primary facility all the customers finding some of its servers available, but finding an optimal allocation rule. This situation appears, e.g., when one introduces mobile emergency units to serve a certain community, where one has different points, representing populated areas (towns, . . .); according to their own features, each area has its own arrival rate and service time, related with the distance from the home location of the servers to the area (Chiu and Larson 1985). Any call finding all the units busy is served by an exogenous system (at a higher cost). Since service times are dependent of travel times, it should be intuitively obvious that the optimization of the system performance—e.g., by maximizing

0025-1909/98/4403/0311$05.00

the expected number of calls served—would imply to serve just those customers close enough to the facility, condemning the outliers to be systematically served by the exogenous server.

Applications of this approach to Computer Science can be found in Xu et al. (1992). See also Miller (1969) for another rejection model in loss systems.

In spite of their practical interest, these rejections policies have not been extensively considered in the literature (some exceptions are Batta 1988, Lippman and Ross 1971), mainly due to the mathematical difficulties inherent to such models. Needless to say also that a much harder to analyze dynamic control of the system, taking into account the number of busy servers and their remaining processing time is to outperform a static policy, which decides whether to accept or not according to just the group the customer belongs to and the existence of idle servers. Nevertheless, there are situations where obtaining in real time the information needed by dynamic rules is so costly or difficult that one can use a static model as an approximation to the much less tractable dynamic models.

The rest of the paper is organized as follows. In §2 the model is formally introduced, and some properties of the corresponding mathematical program are discussed. Section 3 is devoted to the statement of optimality conditions and localization results on the optimal solution, which are used in §4 to design resolution procedures. The paper ends presenting some conclusions and extensions in §5.

2. The Model

Let $J = \{1, 2, \ldots, n\}$ represent n types of customers requesting service of a system consisting of a finite number c of identical servers. Customers of type i (hereafter called i-customers) arrive following a Poisson process with rate $\lambda_i > 0$, the arrival processes of the different groups being independent. The duration of each i-customer service is modelled as a random variable with mean s_i ($0 < s_i < \infty$). No queue is allowed, which means that an arrival that finds the c servers busy is referred to a back-up service system. In addition, any i-customer that arrives when at least one facility is idle is accepted (and its service starts immediately) with probability x_i, and is rejected (rerouted) with probability $1 - x_i$, the x_i's being controllable variables. Any accepted (respectively, lost) i-customer induces a cost to the system of r_i (respectively, $\hat{r}_i$) monetary units, where $r_i < \hat{r}_i$.

Under the assumptions above, the goal is to determine the value of $x = (x_1, \ldots, x_n)$ that minimizes the expected cost per unit time.

Obviously, given $x \in [0, 1]^n$, this system behaves as an $M/G/c/c$ system with arrival rate $\lambda(x)$ and mean service time $s(x)$ given by

$$\lambda(x) = \sum_{i=1}^{n} \lambda_i x_i,$$

$$s(x) = \sum_{i=1}^{n} \frac{\lambda_i x_i}{\lambda(x)} s_i,$$

with the convention that $s(x) = 0$ if $\lambda(x) = 0$.

Let $\rho = (\rho_1, \ldots, \rho_n)$ denote the vector of loads offered by the different groups of customers, i.e.,

$$\rho_i = \lambda_i s_i, \quad i = 1, 2, \ldots, n.$$

Then, the fraction of time that at least one out of the c servers is idle is given by $\Psi_c(\rho \cdot x)$, where $u \cdot v$ denotes the usual scalar product in $\mathbb{R}^n$ and $\Psi_c(t)$ is the fraction of time that an $M/G/c/c$ system with offered load t has at least one server idle, (see Kleinrock 1975), i.e.,

$$\Psi_c(t) = \frac{\sum_{k=0}^{c-1} t^k/k!}{\sum_{k=0}^{c} t^k/k!}.$$

The expected number of i-customers per unit time that enter into the system is given by $\lambda_i x_i \Psi_c(\rho \cdot x)$, and the expected number of rejected i-customers per unit time equals $\lambda_i(1 - x_i) + \lambda_i x_i(1 - \Psi_c(\rho \cdot x)) = \lambda_i(1 - x_i\Psi_c(\rho \cdot x))$. Hence, the expected total cost per unit time is given by

$$\Psi_c(\rho \cdot x) \sum_{i=1}^{n} \lambda_i r_i x_i + \sum_{i=1}^{n} \lambda_i \hat{r}_i (1 - x_i \Psi_c(\rho \cdot x)). \quad (2.1)$$

For each $i = 1, 2, \ldots, n$, define the *rejection surcharge* Δ_i as the difference between the individual cost of rejected and accepted i-customers, i.e.,

$$\Delta_i = \hat{r}_i - r_i.$$

In terms of these parameters, the main result of the paper—Theorem 3.4—states that the optimal policy is a $c\mu$-rule, since it discriminates groups according to the ratios $(\Delta_i)/s_i$, i.e., customers are sorted according to a

single measure: the rejection surcharge per unit service time.

Since (2.1) turns out to be

$$\sum_{i=1}^{n} \lambda_i \hat{r}_i - \Psi_c(\rho \cdot x) \sum_{i=1}^{n} \lambda_i \Delta_i x_i,$$

determining the vector x minimizing the expected cost per unit time is equivalent to solving the following maximization problem

$$\max \Psi_c(\rho \cdot x) \sum_{i=1}^{n} \lambda_i \Delta_i x_i,$$

$$\text{s.t. } x \in [0, 1]^n.$$

To simplify the notation, we first transform the cost structure into an equivalent one by making the two following assumptions.

ASSUMPTION A1. *We assume that*

$$\Delta_i = 1 \quad \forall i = 1, \ldots, n. \tag{A1}$$

Assumption A1 supposes no loss of generality. Indeed, if A1 did not hold, one could define for each $i = 1, 2, \ldots, n$

$$\begin{cases} \tilde{\lambda}_i = \lambda_i \Delta_i, \\ \tilde{s}_i = s_i / \Delta_i, \\ \tilde{\Delta}_i = 1 \end{cases} \tag{2.2}$$

Then, it is easily seen that the system with parameters $\tilde{\lambda}_i$, $\tilde{s}_i$, $\tilde{\Delta}_i$ gives for all $x \in [0, 1]^n$ the same value of the performance measure as the original system.

ASSUMPTION A2. *We assume that*

$$s_1 < s_2 < \cdots < s_n. \tag{A2}$$

Assumption A2 supposes no loss of generality. Indeed, if $s_i = s_j$ for some i, j, $i \neq j$, we can construct an equivalent system with $n - 1$ types of customers, where groups i and j are mixed in one group with arrival rate $\lambda_i + \lambda_j$ and mean service time $s_i = s_j$, and, after relabelling the groups, if necessary, the Assumption A2 is verified.

Hereafter, unless explicitely mentioned, we assume that Assumptions A1 and A2 hold, and stress that this is done just for notational convenience; if they do not hold, one can always transform the original parameters using (2.2) and sorting the service times.

Under such assumptions, the mathematical program of interest is

$$\max F(x) = (\lambda \cdot x)\Psi_c(\rho \cdot x),$$

$$\text{s.t. } x \in [0, 1]^n, \tag{P}$$

where $\lambda = (\lambda_1, \ldots, \lambda_n)$.

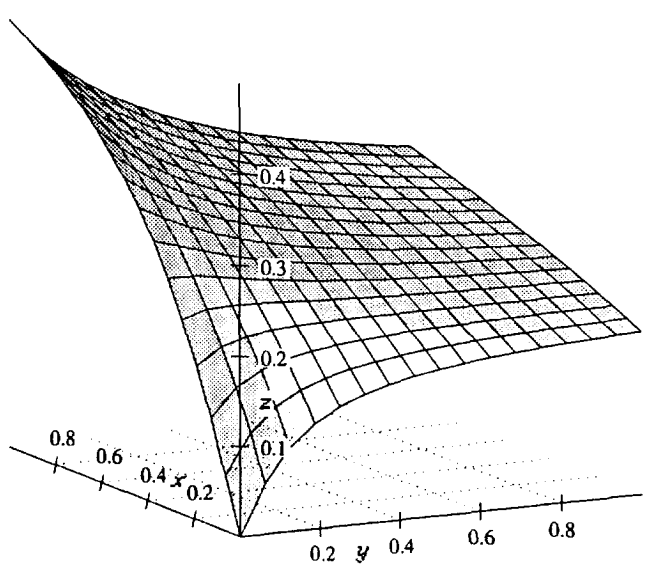
As F is continuous and the feasible set is compact, the maximum of (P) is attained at some $x^* \in [0, 1]^n$.

The case $n = 1$ (homogeneous arrivals) is straightforward: $F(x)$ gives then the throughput of an $M/G/c/c$ system with expected service time s_1 and arrival rate $\lambda_1 x_1$, which is concave (see Harel 1990) and obviously increasing in x_1. Hence, the optimal solution is $x_1^* = 1$, which corresponds to the policy of rejecting only those customers who find the c servers busy.

Unfortunately, these properties do not extend to the case $n > 1$ and contrary to most models encountered in the literature (Grassman 1983, Harel 1990, Harel and Zipkin 1987, Viscolani 1993, Yao and Shantikumar 1987, Zipkin 1986 among others), F is not concave, see Figure 1. This, at least at first glance, makes the optimization process more difficult.

In spite of its lack of concavity, F enjoys interesting mathematical properties (in fact some generalized con-

Figure 1 *F* **Is Not Concave**



cavity), as shown in Theorem 2.1). We recall that a differentiable function g on a convex set S is said to be *pseudoconcave* iff

$$\nabla g(x)\cdot(y - x) > 0 \quad \text{whenever } x, y \in S, g(y) > g(x).$$

See, e.g., Avriel et al. (1988) or Martos (1977) for further properties on pseudoconcave functions.

THEOREM 2.1. *F is pseudoconcave.*

The proof can be found in the appendix.

See also Barros and Frenk (1995) for another "tractable" queueing model with nonconcave objective function.

3. Optimality Conditions

The purpose of this section is to state optimality conditions for problem (P). Expressing F as in (7.1), we see that (P) is a *nonlinear fractional program* with convex feasible region (the n-cube $[0, 1]^n$) and pseudoconcave objective function (Theorem 2.1). The pseudoconcavity of F enables the statement of optimality conditions in terms of the gradient ∇F of F (see, e.g., Corollary 7.49 of Martos 1977). More precisely, if we denote by $D(x)$ the set of feasible directions at $x \in [0, 1]^n$, then

$$\begin{aligned} &x \in [0, 1]^n \text{ is an optimal solution to } (P) \\ &\quad \text{iff } \nabla F(x)\cdot d \leq 0 \quad \forall d \in D(x). \end{aligned} \tag{3.1}$$

The gradient of F is easily obtained,

$$\begin{aligned} \nabla F(x) = {} & \lambda\Psi_c(\rho\cdot x) \\ & + (\lambda\cdot x)\Psi_c'(\rho\cdot x)\rho \quad \forall x \in [0, 1]^n. \end{aligned} \tag{3.2}$$

We will combine (3.1) and (3.2) above to obtain optimality conditions. First we show that no interior point x can be optimal.

LEMMA 3.1. *Suppose that $x = (x_1, \ldots, x_n)$ is an optimal solution to (P), and $0 < x_k < 1$. Then,*

$$s_k = \frac{-\Psi_c(\rho\cdot x)}{(\lambda\cdot x)\Psi_c'(\rho\cdot x)}.$$

PROOF. Let e^k be the vector with 1 at the kth coordinate and zeros everywhere else. As $0 < x_k < 1$, both e^k and $-e^k$ are feasible directions, which, by (3.1), implies that

$$\nabla F(x)\cdot e^k = 0$$

By (3.2), the result follows. □

The next lemma shows that any optimal policy is deterministic except for at most one class of customers.

LEMMA 3.2. *If x is an optimal solution to (P), then x has at most one fractional component.*

The proof is straightforward by the lemma above and Assumption A2. □

This result leads to an interesting consequence about the geometrical structure of the set of optimal solutions, stated in the theorem below.

THEOREM 3.1. *The set of optimal solutions to (P) is a closed (possibly degenerate) segment contained in an edge of $[0, 1]^n$.*

PROOF. As, by Theorem 2.1, F is pseudoconcave, it is continuous and quasiconcave (see Theorem 7.28 of Martos 1977). Hence, the set S of optimal solutions to (P) is a closed convex set. Furthermore, S is contained in an edge of $[0, 1]^n$. Indeed, else, there would exist x^1, $x^2 \in S$ not contained in the same edge. By convexity of S and the quasiconcavity of F, the whole segment with endpoints x^1 and x^2 would consist of optimal solutions. Then there would exist an optimal solution with at least two fractional components, which, by Lemma 3.2, is a contradiction. Hence, the result holds. □

COROLLARY 3.1. *If $c > 1$, then the set of optimal solutions to (P) is a singleton.*

See the appendix for the proof.

REMARK 3.1. If $c > 1$, the corollary above shows that there exists a unique optimal solution. However, if $c = 1$, (P) might have multiple optimal solutions. As a simple example, take $n = 2$, $\lambda_1 = \lambda_2 = 1$, $s_1 = 1$, $s_2 = 2$. Then, it is easily checked that the set of optimal solutions to (P) is the closed segment with endpoints (1, 0) and (1, 1). In fact, it will be shown in §4 that, for $c = 1$, the set of optimal policies consists of either one nonrandomized policy or the set of mixtures of two nonrandomized policies.

By the theorem above, any optimal solution x to (P) must be either a vertex or a point on an edge of $[0, 1]^n$, thus the search of optimal solutions to (P) is reduced to the 0- and one-dimensional faces of the n-cube $[0, 1]^n$.

We state below the corresponding optimality conditions. Define, for each $x \in [0, 1]^n$ the sets $I(x)$ and $J(x)$ as

$$I(x) = \{i : x_i = 0\}, \quad J(x) = \{i : x_i = 1\}.$$

THEOREM 3.2. (optimality conditions at a vertex) *Let $x \neq 0$ be a vertex of $[0, 1]^n$. Then, x is an optimal solution to (P) iff*

$$\min_{i \in I(x)} s_i \geq \frac{-\Psi_c(\rho \cdot x)}{(\lambda \cdot x)\, \Psi_c'(\rho \cdot x)} \geq \max_{i \in J(x)} s_i. \tag{3.3}$$

The proof can be found in the appendix.

THEOREM 3.3. (optimality conditions on edges). *Let $x \in [0, 1]^n$ be such that $0 < x_k < 1$ for some k, and $x_i \in \{0, 1\}$ $\forall i \neq k$. Then, x is an optimal solution to (P) iff*

$$\min_{i \in I(x)} s_i > s_k = \frac{-\Psi_c(\rho \cdot x)}{(\lambda \cdot x)\, \Psi_c'(\rho \cdot x)} > \max_{i \in J(x)} s_i. \tag{3.4}$$

The proof of this theorem runs parallel to that of the theorem above, just taking into account Lemma 3.1 and Assumption A2. □

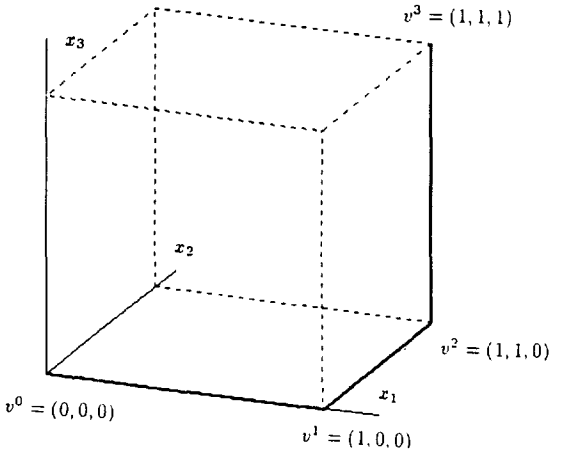
REMARK 3.2. Although the conditions (3.3) and (3.4) have been obtained under Assumptions A1 and A2, they lead to a meaningful interpretation in terms of the original setting: It is optimal to serve only those customers whose rejection surcharge per unit service time is high enough.

The results stated so far suggest a naïve procedure for solving (P): Evaluate the 2^n vertices of $[0, 1]^n$, and solve the $n2^{n-1}$ one-dimensional nonlinear fractional programs obtained by optimizing F on each edge. However, we can go much further; as shown below, neither *all* the vertices nor *all* the edges are true candidates to contain optimal solutions.

For this purpose, let v^i, $i = 0, \ldots, n$ be the vector with 1 at the first i coordinates and 0 everywhere else. Let $\alpha : [0, n] \to [0, 1]^n$ be the natural parametrization of the path through $v^0, v^1, \ldots, v^n$ (see Figure 2 for an illustration of α when $n = 3$).

The next theorem shows that any optimal solution must be contained in the path $\alpha([0, n])$, i.e., since, by Assumption A2, the service times are given in increasing order, the optimal policy belongs to the class of $c\mu$ rules.

Figure 2 **The Set $\alpha([0, n])$ when $n = 3$**



THEOREM 3.4. *If x is an optimal solution to (P), then there exists some t, $0 < t \leq n$ such that $x = \alpha(t)$.*

PROOF. By Lemma 3.2, x has at most one fractional component, thus either (3.3) or (3.4) apply.

Hence, A2 implies that

$$\left.\begin{array}{l} \text{if } x_i = 0, \quad \text{then } x_j = 0 \ \forall j > i \\ \text{if } x_i > 0, \quad \text{then } x_j = 1 \ \forall j < i \end{array}\right\}, \tag{3.5}$$

thus $x = \alpha(t)$ for some t. □

REMARK 3.3. The $c\mu$-rule optimality is no longer true when one restricts the set of policies to nonrandomized ones, i.e., when, instead of solving (P) one wants to solve its $\{0, 1\}$-version (P_{NR})

$$\max F(x),$$
$$\text{s.t. } x \in \{0, 1\}^n. \tag{P_{NR}}$$

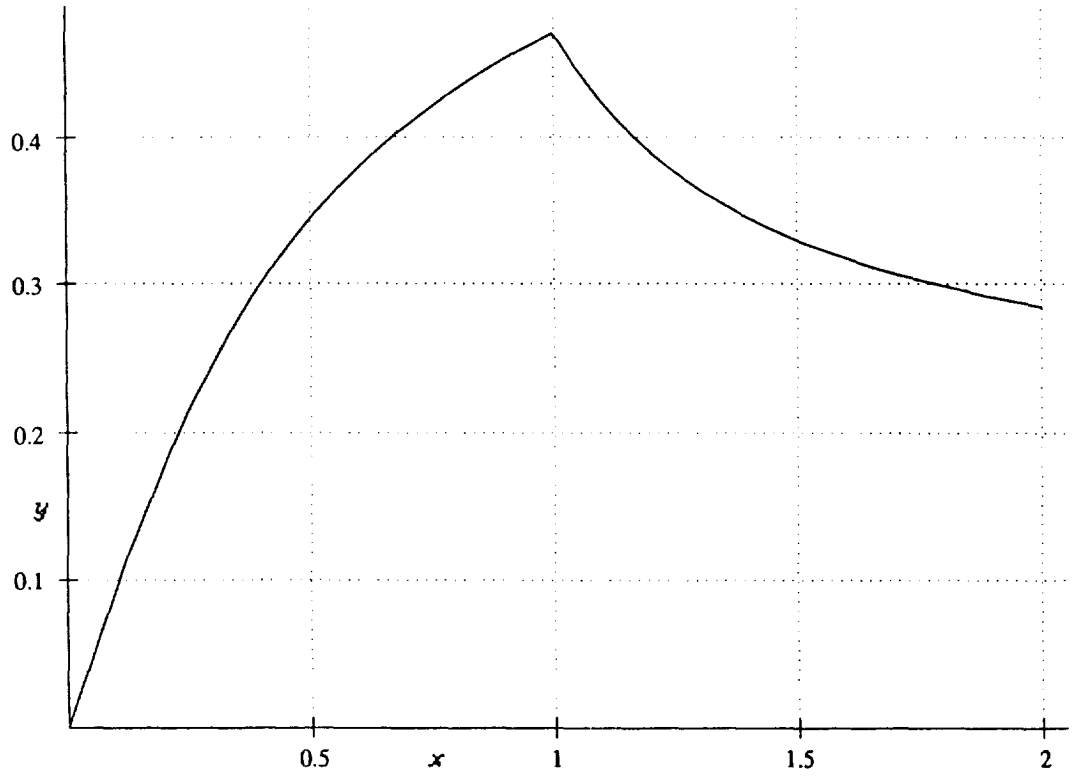
For instance, if we take $c = 2$, $n = 3$, $\lambda_1 = 1$, $s_1 = 1$, $\lambda_2 = 10$, $s_2 = 2.5$, $\lambda_3 = 0.5$, $s_3 = 0.75$, it is easily seen that the optimal solution to (P_{NR}) is the vector $x^* = (1, 0, 1)$, which is not in the path $\{\alpha(t) : t \in [0, 3]\}$.

The theorem above shows that (P) is equivalent to the one-dimensional problem $(\tilde{P})$

$$\max \tilde{F}(t) = F(\alpha(t)),$$
$$(\text{s.t. } t \in [0, n]. \tag{$\tilde{P}$}$$

As $\alpha(\cdot)$ is not differentiable at integer points, $\tilde{F}$ might not be differentiable, but it remains directionally differentiable.

Figure 3 The $\tilde{F}$ Corresponding to F of Figure 1



What does not seem so intuitive is the fact that $\tilde{F}(\cdot)$ is also unimodal. As an example, Figure 3 shows the $\tilde{F}$ corresponding to the function F of Figure 1.

In fact, as the next theorem shows, $\tilde{F}(\cdot)$ is a directionally differentiable (although non-differentiable) *semilocally pseudoconcave* function on $[0, n]$, i.e.,

$$\tilde{F}'(t; s-t) > 0 \quad \text{whenever } \tilde{F}(s) > \tilde{F}(t), \quad s, t \in [0, n]$$

where $\tilde{F}'(t; s-t)$ stands for the directional derivative of $\tilde{F}$ at t in the direction $s-t$. See Kaul and Kaur (1982) for further results on this concept.

THEOREM 3.5. *$\tilde{F}(\cdot)$ is semilocally pseudoconcave on $[0, n]$.*

For the proof, see the appendix.

4. Finding an Optimal Rejection Policy

In the section above we have shown that finding an optimal policy is equivalent to solving the one-dimensional problem $(\tilde{P})$,

$$\max \tilde{F}(t) = F(\alpha(t)),$$
$$\text{s.t. } t \in [0, n]. \qquad (\tilde{P})$$

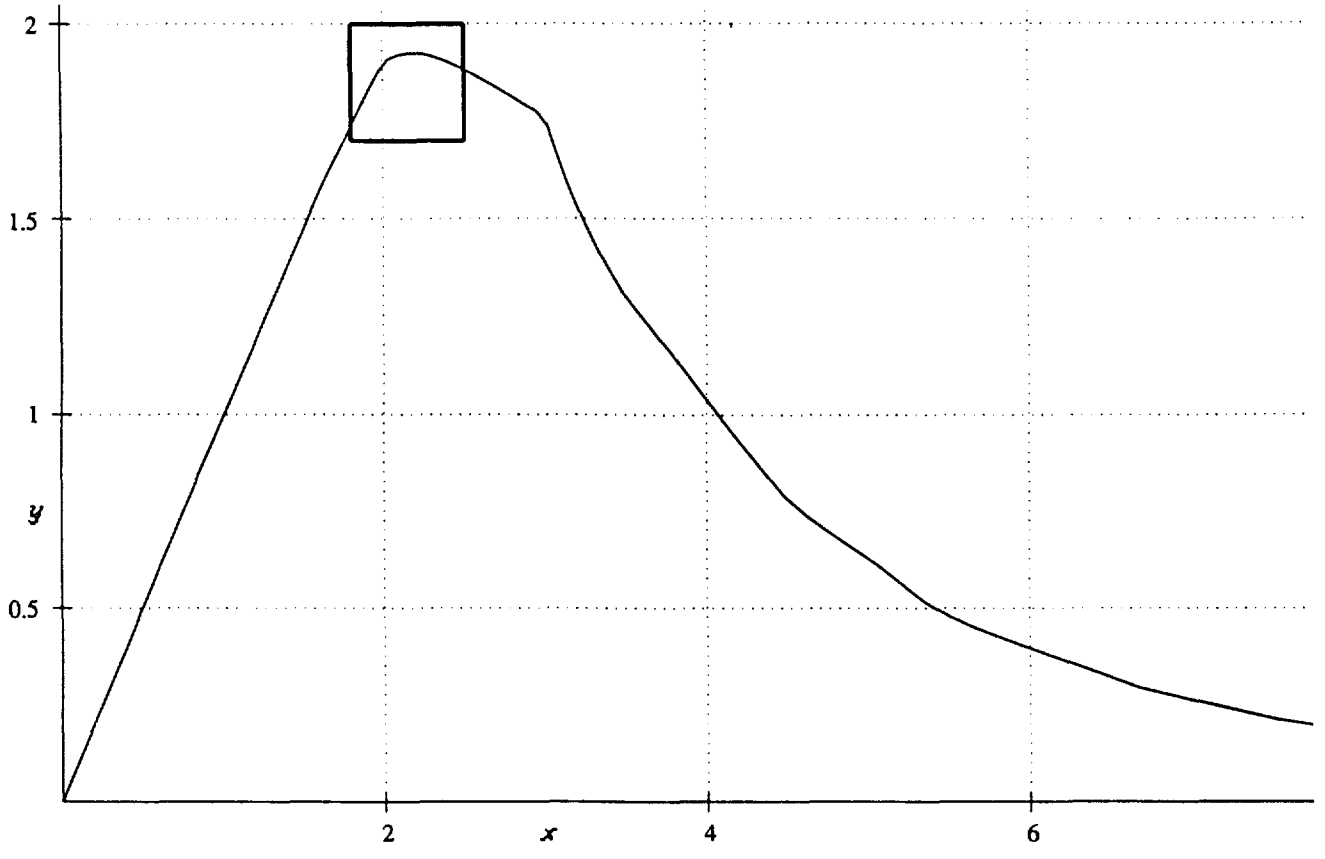
Furthermore, by Theorem 3.5 the objective function is semilocally pseudoconcave, which enables the resolution (up to a prespecified accuracy ϵ) of $(\tilde{P})$ by a variety of well-known methods (see, e.g., Chapter 8 of Bazaraa and Shetty 1979). An illustration is given in the next example.

EXAMPLE 4.1. Consider a system with $c = 3$ servers and $n = 8$ classes of customers. The i-customers arrive following a Poisson process with arrival rate $\lambda_i = 1$ and have expected service time $s_i = i^3/10$, $i = 1, \ldots, 8$.

The graph of $\tilde{F}$ is plotted in Figure 4, and Figure 5 represents the portion boxed in Figure 4. Note that $\tilde{F}$ attains its maximum at a noninteger point, i.e., optimality is attained at a randomized rejection policy.

In order to solve $(\tilde{P})$, we have chosen the golden section method (see, e.g., Bazaraa and Shetty 1979). For a prespecified accuracy of $\epsilon = 0.001$ we needed 18 iterations; the results are shown in Table 1.

Figure 4 **The Graph of $\tilde{F}$**



Hence, we take as solution the point $t^* = 2.18013$, which gives $\tilde{F}(t^*) = 1.92477$, and corresponds to the following policy: when an arriving i-customer finds at least one server idle, it is rejected by the system with probability 0 if $i = 1, 2$; with probability $3-t^* = 0.81987$ if $i = 3$ and with probability 1 if $i > 3$. □

The methodology above applies for any value of c. However, when $c \leq 2$, the functions F and $\tilde{F}$ have a much simpler shape, which enables us to solve *exactly* the problem $(\tilde{P})$ and also (P) in finite time.

We explore first the single-server case, i.e., $c = 1$. Then, (P) takes the form

$$\max F(x) = (\lambda \cdot x)/(1 + \rho \cdot x),$$
$$\text{s.t. } x \in [0, 1]^n. \qquad (P)$$

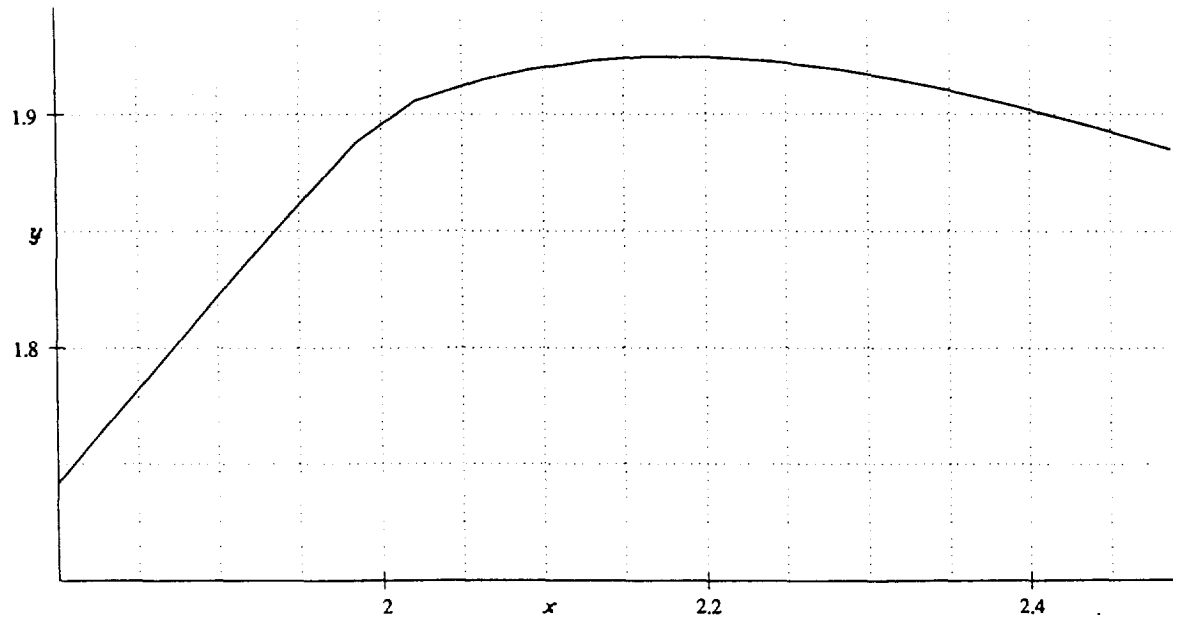
This implies that (P) becomes a linear fractional program, thus the objective function F is not only pseudoconcave but pseudomonotonic. See, e.g., Avriel et al. (1988) or Martos (1977) for further properties on this class of functions.

The single-server assumption enables to strengthen or ease some of the results obtained in previous sections. For instance, we know from §3 that the set S of optimal solutions to (P) might not be a singleton, but is contained in an edge of $[0, 1]^n$. The pseudomonotonicity of F implies that S must equal the convex hull of the optimal vertices. Hence, S must have one of the two following forms:

1. The singleton $\{\alpha(k)\}$ for some $k \in \{1, 2, \ldots, n\}$.
2. The closed segment with endpoints $\{\alpha(k)\}$ and $\{\alpha(k + 1)\}$ for some $k \in \{1, 2, \ldots, n - 1\}$.

In other words, there always exists some k such that the nonrandomized policy consisting of automatically rejecting *all* the i-customers $(i > k)$ and serving every i-customer $(i = 1, 2, \ldots, k)$ who finds idle servers is optimal. Hence, in

Figure 5 $\tilde{F}$ **Attains Its Maximum at a Noninteger Point**



order to find an optimal policy we can restrict our attention to the set $\{\alpha(1), \alpha(2), \ldots, \alpha(n)\}$ as candidate points (i.e., $\{\alpha(1), \alpha(2), \ldots, \alpha(n)\}$ is a *finite dominating set*, see Hooker et al. 1991), which suggests the following approach: Evaluate F at $\alpha(i)$ for all i, and take as optimal solution the $\alpha(i^*)$ giving the highest value of F.

This simple algorithm runs in $O(n^2)$ time, which is acceptable for moderate values of n. If n is so large that even an $O(n^2)$ computing time should be considered as prohibitive, more subtle procedures can be used. Indeed, as shown in Hansen et al. (1991), it is possible to use binary search techniques to obtain an overall complexity of $O(n)$ time and space.

Table 1 **Solving** $(\tilde{P})$ **by the Golden Section Method**

Iterat.	Accuracy	Optimal Interval
1	3.05573	(0.00000, 4.94427)
2	1.88854	(0.00000, 3.05573)
3	1.16718	(1.16718, 3.05573)
4	0.72136	(1.88854, 3.05573)
5	0.44582	(1.88854, 2.60990)
	...	...
16	0.00224	(2.17719, 2.18082)
17	0.00138	(2.17858, 2.18082)
18	0.00086	(2.17943, 2.18082)

We examine now the case $c = 2$. It is rather easy to find the integer k such that the optimal solution to $(\tilde{P})$ either is k or belongs to the open interval $(k, k + 1)$. Indeed, such conclusion can be obtained after checking the optimality conditions at vertices and performing some iterations of, for instance, a binary-search procedure. If k is optimal to $(\tilde{P})$, then $\alpha(k)$ is optimal to (P). On the contrary, if the optimal solution is $t^* \in (k, k + 1)$, then, by the semilocal pseudoconcavity of $\tilde{F}$, t^* is the *unique* root in $(k, k + 1)$ of the nonlinear equation

$$\tilde{F}'(t) = 0. \tag{4.1}$$

By (7.5), it is easily seen that when $c = 2$, (4.1) can be written in the form $Q(t) = 0$, where Q is a polynomium of degree not greater than 3. Hence, the optimal solution to $(\tilde{P})$ is the unique root of Q in $(k, k + 1)$, which, as is well-known, can be obtained *exactly*.

5. Conclusions

In this paper we have addressed a design problem associated with a loss model with heterogeneous arrivals, where the decision variables represent the probability

that a customer of each class is rejected by the system when he finds some servers idle.

After imposing a cost structure on the model, the search of a cost-optimal policy is reduced to solving a nonlinear mathematical program. The objective function of the problem addressed may not be concave; nevertheless, we state some properties of the problem (optimality of a $c\mu$-rule) that enable us to reduce the optimization to solving an equivalent unimodal one-dimensional problem for which several well-known resolution techniques can be applied.

Finally, we have shown that an optimal policy for the two-server case can be found as a root of a polynomial function of degree 3, whilst the one-server case leads to a linear fractional program with a simple structure, which can be solved in $O(n^2)$ time by straightforward techniques, and in linear time by binary search.

Extensions of these results to systems governed by dynamic state-dependent policies, or to $M/G/1/\infty$ systems are interesting questions which remain open.[1]

[1] The research of the authors is partially supported by Spanish DGICYT grant PB93-0927. This support is gratefully acknowledged.

Appendix

LEMMA 7.1. *The function* $1/\Psi_c(\cdot)$ *is convex on* $[0, \infty)$. *Moreover, if* $c > 1$ *then* $1/\Psi_c(\cdot)$ *is strictly convex on* $[0, \infty)$.

PROOF. The case $c = 1$ is straightforward ($1/\Psi_1(t) = 1 + t$), so we consider only the case $c > 1$. As

$$1/\Psi_c(t) = \frac{\sum_{k=0}^{c} t^k/k!}{\sum_{k=0}^{c-1} t^k/k!} = 1 + \frac{t}{c}(1 - \Psi_{c-1}(t)),$$

and the function $t \mapsto t\Psi_{c-1}(t)$ is concave (see, e.g., Corollary 1 of Harel 1990), it follows that $1/\Psi_c(\cdot)$ is convex. In order to check that it is also strictly convex, observe that, otherwise, $1/\Psi_c(\cdot)$ should be a polynomial of degree at most one in some nondegenerate interval I, i.e., there would exist α, β such that

$$\frac{\sum_{k=0}^{c} t^k/k!}{\sum_{k=0}^{c-1} t^k/k!} = \alpha t + \beta \quad \forall t \in I.$$

Equating coefficients, one would obtain $\alpha = 0 = 1/c$, which is a contradiction. □

PROOF OF THEOREM 2.1. Observe that

$$F(x) = \frac{\lambda \cdot x}{1/\Psi_c(\rho \cdot x)}. \tag{7.1}$$

By Lemma 7.1, the function $t \mapsto 1/\Psi_c(t)$ is convex. Hence, its composition with the linear function $x \mapsto \rho \cdot x$ is convex. Hence, F is the quotient of a nonnegative linear and a positive convex function, thus F is pseudoconcave, as asserted. □

PROOF OF COROLLARY 3.1. Suppose, on the contrary, that (P) has at least two different optimal solutions x, y. Then, by the theorem above, both x and y belong to the same edge of $[0, 1]^n$.

Within the interval I of endpoints x, y, one deduces from Theorem 5.17 of Avriel et al. (1988) and our Lemma 7.1 that F is strictly pseudoconcave on I, which is a contradiction with the simultaneous optimality of x and y. □

PROOF OF THEOREM 3.2. First, observe that $\Psi_c(t)$ is strictly decreasing in t (see Harel 1990), i.e.,

$$\Psi_c'(t) < 0 \quad \forall t. \tag{7.2}$$

Define, for each $i = 1, \ldots, n$, the vector e^i as in the proof of Lemma 3.1. Obviously

$$D(x) = \text{cone}(\{e^i : i \in I(x)\} \cup \{-e^i : i \in J(x)\}), \tag{7.3}$$

where cone (A) represents the cone generated by the elements of A. Hence, by (3.1), (3.2), and (7.3), it follows that

x is optimal iff

$$\begin{cases} \lambda_i \Psi_c(\rho \cdot x) + (\lambda \cdot x)\, \Psi_c'(\rho \cdot x)\, \rho_i \le 0 & \forall i \in I(x), \\ \lambda_i \Psi_c(\rho \cdot x) + (\lambda \cdot x)\, \Psi_c'(\rho \cdot x) \rho_i \ge 0 & \forall i \in J(x), \end{cases}$$

which, by (7.2), turns out to be equivalent to (3.3). □

PROOF OF THEOREM 3.5. For each $k = 1, \ldots, n$ let the vector e^k be defined as in the proof of Lemma 3.1. We will just show that $\tilde{F}(\cdot)$ is semilocally pseudoconcave at noninteger points $t \in [0, n]$, namely,

$$\begin{aligned} &s, t \in [0, n],\ s \ne t,\ \tilde{F}'(t; s - t) \le 0 \\ &\quad \text{implies } \tilde{F}(s) \le \tilde{F}(t). \end{aligned} \tag{7.4}$$

The proof for integer points t is completely analogous and will not be given here.

Hence, we assume that $t \in [0, n]$ is not integer, thus there exists $k \in \{0, 1, \ldots, n-1\}$ such that $k < t < k + 1$. Within the interval $[k, k + 1]$ the function $\tilde{F}$ takes the form

$$\tilde{F}(s) = \left(\sum_{i<k} \lambda_i + \lambda_k(s - k)\right) \cdot \Psi_c\left(\sum_{i<k} \rho_i + \rho_k(s - k)\right) \quad \forall s \in [k, k + 1].$$

Hence, $\tilde{F}$ is differentiable at t, and

$$\begin{aligned} \tilde{F}'(t) &= \nabla F(\alpha(t)) \cdot e^k \\ &= \lambda_k \Psi_c(\rho \cdot \alpha(t)) + (\lambda \cdot \alpha(t)) \Psi_c'(\rho \cdot \alpha(t)) \rho_k. \end{aligned} \tag{7.5}$$

Hence, one has

$$\tilde{F}'(t) \le 0 \quad \text{iff } \Psi_c(\rho \cdot \alpha(t)) + (\lambda \cdot \alpha(t)) \Psi_c'(\rho \cdot \alpha(t))\, s_k \le 0.$$

Since, by (7.2) and A2,

$$\Psi_c(\rho\cdot\alpha(t)) + (\lambda\cdot\alpha(t))\Psi_c'(\rho\cdot\alpha(t))s_j$$
$$\leq \Psi_c(\rho\cdot\alpha(t)) + (\lambda\cdot\alpha(t))\Psi_c'(\rho\cdot\alpha(t))s_k \ \forall j \geq k,$$
$$\Psi_c(\rho\cdot\alpha(t)) + (\lambda\cdot\alpha(t))\Psi_c'(\rho\cdot\alpha(t))s_j$$
$$\geq \Psi_c(\rho\cdot\alpha(t)) + (\lambda\cdot\alpha(t))\Psi_c'(\rho\cdot\alpha(t))s_k \ \forall j \leq k,$$

it follows that

$$\left.\begin{aligned}\tilde{F}'(t) \leq 0 \quad &\text{iff } \max_{j\leq k} \nabla F(\alpha(t))\cdot e^j \geq 0\\ \tilde{F}'(t) \geq 0 \quad &\text{iff } \min_{j\leq k} \nabla F(\alpha(t))\cdot e^j \leq 0\end{aligned}\right\}. \tag{7.6}$$

Now we are in position to show (7.4); let $s \neq t$ be such that $\tilde{F}'(t)(s-t) = \tilde{F}'(t; s-t) \leq 0$. We can distinguish the cases $s > t$ and $s < t$.

If $s > t$, then $\tilde{F}'(t) \leq 0$. Hence, by (7.6),

$$\nabla F(\alpha(t))\cdot e^j \leq 0 \quad \forall j \geq k,$$

thus

$$\nabla F(\alpha(t))\cdot d \leq 0 \quad \forall d \in \text{cone}(\{e^j : j \geq 0\}).$$

Hence, by Theorem 2.1,

$$F(\alpha(t)) \geq F(y) \quad \forall y \in [0,1]^n$$
$$\text{such that } y - \alpha(t) \in \text{cone}(\{e^j : j \geq k\}).$$

Since $s \geq t$, by construction of $\alpha(\cdot)$ one has that

$$\alpha(s) - \alpha(t) \in \text{cone}(\{e^j : j \geq k\}).$$

Hence,

$$\tilde{F}(t) = F(\alpha(t)) \geq F(\alpha(s)) = \tilde{F}(s), \tag{7.7}$$

as asserted.

Now we consider the case $s < t$. Then, $\tilde{F}(t) \geq 0$, thus, by (7.6),

$$\nabla F(\alpha(t))\cdot e^j \geq 0 \quad \forall j \leq k,$$

thus

$$\nabla F(\alpha(t))\cdot d \geq 0 \quad \forall d \in \text{cone}(\{e^j : j \leq 0\}),$$

which turns out to be equivalent to

$$\nabla F(\alpha(t))\cdot d \leq 0 \quad \forall d \in \text{cone}(\{-e^j : j \leq 0\}).$$

As for the case $s > t$, this leads to (7.7), and (7.4) follows. □

References

Avriel, M., W. E. Diewert, S. Schaible, and I. Zang, *Generalized Concavity*, Plenum Press, 1988.

Barros, A. I. and J. B. G. Frenk, "Generalized Fractional Programming and Cutting Plane Algorithms," *J. Optimization Theory and Applications*, 87 (1995), 103–120.

Batta, R., "Single-server Queueing-location Models with Rejection," *Transportation Sci.*, 22 (1988), 209–216.

Bazaraa, M. S. and C. M. Shetty, *Nonlinear Programming, Theory and Algorithms*, Wiley, New York, 1979.

Chiu, S. S. and R. C. Larson, Locating an n-server Facility in a Stochastic Environment, *Computers and Operations Res.*, 12 (1985), 509–516.

Crabill, T. B., D. Gross, and M. J. Magazine, "A Classified Bibliography of Research on Optimal Design and Control of Queues," *Operations Res.*, 25 (1977), 219–232.

Grassmann, W., "The Convexity of the Mean Queue Size of the $M/M/c$ Queue with Respect to the Traffic Intensity," *J. Applied Probability*, 20 (1983), 916–919.

Hansen, P., M. V. Poggi de Aragao, and C. C. Ribeiro, "Hyperbolic 0–1 Programming and Query Optimization in Information Retrieval," *Math. Programming*, 52 (1991), 255–263.

Harel, A., "Convexity Properties of the Erlang Loss Formula," *Oper. Res.*, 38 (1990), 499–505.

—— and P. H. Zipkin, "The Convexity of a General Performance Measure for the Multiserver Queues," *J. Applied Probability*, 24 (1987), 725–736.

Hillier, F. S. and G. J. Lieberman, *Introduction to Stochastic Models in Operations Research*, McGraw-Hill, New York, 1990.

Hooker, J. N., R. S. Garfinkel, and C. K. Chen, "Finite Dominating Sets for Network Location Problems," *Operations Res.*, 39 (1991), 100–118.

Kaul, R. N. and S. Kaur, "Generalizations of Convex and Related Functions," *EJOR*, 9 (1982), 369–377.

Kleinrock, L., *Queueing Systems*, Vol. I, Wiley, New York, 1975.

Lippman, S. A. and S. H. Ross, "The Streetwalker's Dilemma: A Job Shop Model," *SIAM J. Applied Mathematics*, 20 (1971), 336–342.

Martos, B., *Nonlinear Programming Theory and Methods*, North-Holland, Amsterdam, 1977.

Mendelson, H. and S. Whang, "Optimal Incentive-compatible Priority for the $M/M/1$ Queue," *Oper. Res.*, 38 (1990), 870–883.

Miller, B. L., "A Queueing Reward System with Several Customer Classes," *Management Sci.*, 16 (1969), 234–245.

Viscolani, B., "Optimal Design of a Multiservice System: The Line-penalty Problem," *EJOR*, 67 (1993), 242–247.

Walrand, J., *An Introduction to Queueing Networks*, Prentice-Hall, New York, 1988.

Xu, S. H., R. Righter, and J. G. Shantikumar, "Optimal Dynamic Assignment of Customers to Heterogeneous Servers in Parallel," *Oper. Res.*, 40 (1992), 1126–1138.

Yao, D. D. and J. G. Shanthikumar, "The Optimal Input Rate to a System of Manufacturing Cells," *INFOR*, 25 (1987), 57–65.

Zipkin, P. H., "Models for Design and Control of Stochastic Multi-item Batch Production Systems," *Oper. Res.*, 34 (1986), 91–104.

Accepted by Linda V. Green; received May 16, 1994. This paper has been with the authors 18 months for 2 revisions.