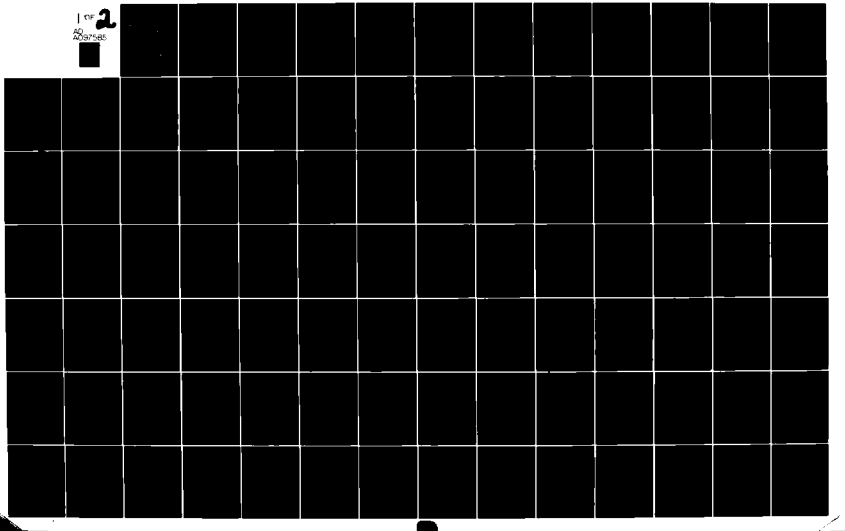
AD-A097 585 OFFICE OF NAVAL RESEARCH ARLINGTON VA F/G 5/1
NAVAL RESEARCH LOGISTICS QUARTERLY, VOLUME 28, NUMBER 1,(U)
MAR 81 S M SELIG

UNCLASSIFIED NL



NAVAL RESEARCH LOGISTICS QUARTERLY

Volume 28. Number 1,

Seymour M. Selig

195

MAR 1981
VOL. 28, NO. 1

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $11.15 a year in the U.S. and Canada, $13.95 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, P-35 (Revised 1-74).

A LINEAR PROGRAMMING MODEL FOR DESIGN OF COMMUNICATIONS NETWORKS WITH TIME VARYING PROBABILISTIC DEMANDS

K. O. Kortanek, D. N. Lee, G. G. Polak

Department of Mathematics
Carnegie-Mellon University
Pittsburgh, Pennsylvania

ABSTRACT

In this paper marginal investment costs are assumed known for two kinds of equipment stocks employed to supply telecommunications services: trunks and switching facilities. A network hierarchy is defined which includes important cases occurring in the field and also appearing in the literature. A different use of the classical concept of the marginal capacity of an additional trunk at prescribed blocking probability leads to a linear programming supply model which can be used to compute the sizes of all the high usage trunk groups. The sizes of the remaining trunk groups are approximated by the linear programming models, but can be determined more accurately by alternate methods once all high usage group sizes are computed. The approach applies to larger scale networks than previously reported in the literature and permits direct application of the duality theory of linear programming and its sensitivity analyses to the study and design of switched probabilistic communications networks with multiple busy hours during the day. Numerical results are presented for two examples based on field data, one of which having been designed by the multi-hour engineering method.

1. INTRODUCTION: A DESIGN SYNTHESIS PROBLEM

In this paper we treat telecommunications networks where customer demands for service are specified probabilistically between pairs of junctions according to different hours of the day. Telephone traffic may flow over a direct route which joins two distinct junctions or over an alternate route which is defined in prespecified network routing hierarchy. Networks which permit alternate routing of traffic are termed *switched* because switching operations are required to alternately route a call. The network routing hierarchy permits traffic which is blocked on a direct route to be switched through other junctions. The switching process tends to smooth out the peaks of traffic loads which occur throughout the network at different times of the day. Consequently, less equipment may be required to service the overall traffic load on the network than for a similar network without alternate routings.

An example of a network routing hierarchy is given below in Figure 1. It consists of junctions A through H and two different kinds of links joining certain pairs of junctions. A *link* is merely a dimensionless entity whose existence indicates that telephone calls, collectively termed *traffic*, may flow in either direction between the two junctions which it joins, without involving any other junction than these two. A dashed line designates a *direct link* while a solid line designates a *final link*. If there is a direct link between a call-origination junction and a

call-destination junction, then a telephone connection is first attempted on this link, the first choice route. Should the first choice connection fail, then an attempt is made to alternately route the call by way of final links, and in this case the traffic is referred to as *overflow traffic*. Arrows in Figure 1 indicate the overflow routing scheme. In case no direct link exists between a call origin and destination, then the call is also routed along the final links. Should a connection on final routes fail, we say that the call is "lost," and the caller must try to place the call again.

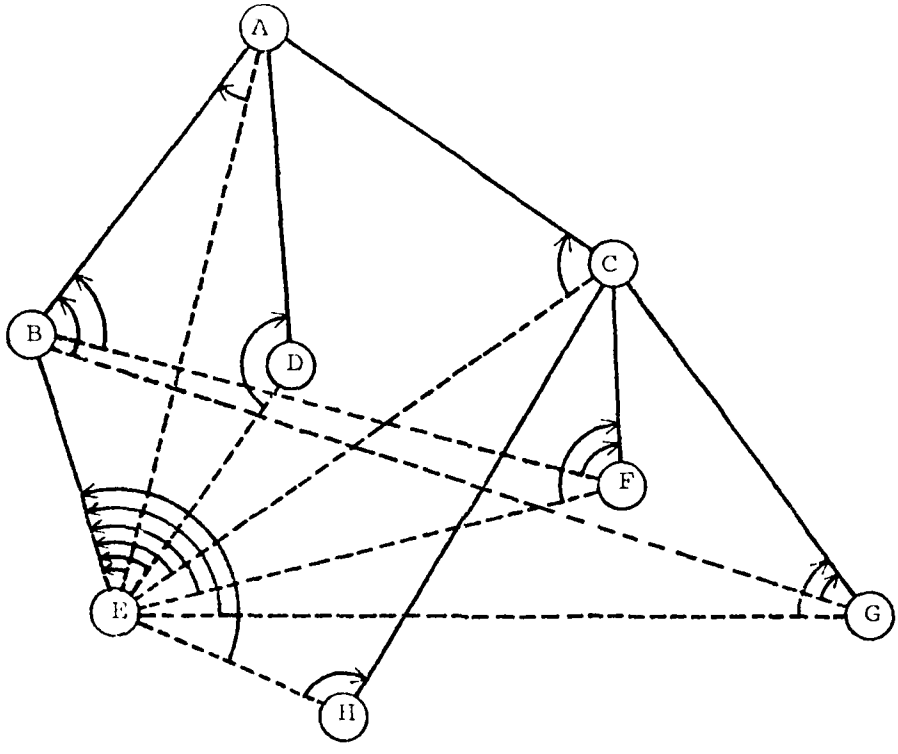


FIGURE 1. A network hierachy with direct (— — — —) and final (————) links where overflow from a direct link onto a final link is indicated by an arrow

The basic problem attacked in this paper is one of *design synthesis:* solve for least-cost equipment changes in a given network routing hierarchy which are sufficient to meet altered point-to-point customer demands for service during different times of day to within a prescribed blocking probability. The emphasis is on the provision of a telecommunications service by an optimal use of available equipment. The model we develop includes a probabilistic specification of customer demand by time of day and includes alternate routings, where each direct link has a uniquely specified alternate route in the hierarchy. It is a nonlinear integer program P, which takes as a basic "unit" of equipment the concept of a "trunk." The terminology requires elucidation.

In this paper a *trunk* shall merely refer to a channel which is required in order for a telephone call to transpire between two junctions in either direction. As such, it is a dimensionless quantity. The call carrying capacity of a trunk depends on the probabilistic mechanism underlying customer calling patterns. For example, during a fixed hour a trunk could carry 60 one-minute serially placed telephone calls. Under this discipline the total carried load during the

hour is 3600 call-seconds, denoted 36 CCS. Expressed another way, we observe that the probability of a call being blocked is zero. On the other hand should a demand for 60 one-minute calls occur simultaneously, then the offered load is still 36 CCS, but only 0.6 CCS is actually carried. The blocking probability is now 59/60.

A collection of trunks joining two distinct junctions is merely referred to as a *trunk group*. It is convenient to view a link as a trunk group. According to network engineering principles, it has been customary to assume that customer originated calls are generated by Poisson process and are assigned sequentially to a trunk group. These assumptions yield an important property which is fundamental to our development of a good linear programming approximation to the nonlinear integer program P, namely, that the carried load on the last trunk is monotonically decreasing with the number of trunks, see Messerli [13]. The necessary results upon which the linear programming construction is based are proved in the Appendix.

The nonlinear and linear supply models of this paper employ certain concepts of unit costs with respect to both trunking and switching. The definition of "cost" shall be limited to the incremental investment cost of providing a trunk on the direct route between two junctions and the incremental investment cost of providing a trunk along the uniquely specified alternate route connecting these two junctions. In addition, we shall include unit switching investment costs per CCS as a crude approximation for switching investments stemming from switching calls from one trunk group to another.

Finally, we present linear programming solutions for two network hierarchies occurring in the field, one of which has been designed using nonlinear steepest descent methods, see Eisenberg [5] and Elsner [6]. This section also contains a user's guide for implementation of the model.

2. APPROACHES TO DETERMINE TRUNKING AND SWITCHING REQUIREMENTS TO MEET DEMAND FOR SERVICE

Over the past 30 years it appears that there have been at least two basic approaches to the design synthesis problem discussed in the previous section.

The basic thrust of our paper proceeds according to what we term the *first approach* to the design problem. It incorporates specific probability distributions for each parcel of traffic, where a parcel is merely that portion of traffic which follows specific routes in the network. Different parcels experience different blocking probabilities, even on the very same trunk group. For example, a given trunk group may accommodate customer originated traffic governed by the Poisson probability distribution, and the group may also accomodate overflow traffic which is "peaked," in the sense that the mean of the distribution is less than its variance. Investigations of the blocking probabilities of individual parcels have been made by Wilkinson [20], Katz [12], and more recently by Deschamps [4].

The pioneering work representing a probabilistic approach which has had widespread use throughout the telecommunications industry is the 1954 paper by Truitt [19]. The generally accepted name of the method reflects the fact that economic considerations are also part of the analysis. The method is termed the "ECCS-method," where the letter "E" stands for "economic." The method was introduced by Truitt for the simplest of routing hierarchies consisting of a triad of junctions with one overflow possibility, and one specific time of day (single hour). The solved-for variables are the specific sizes of all trunk groups.

Further important extensions of the ECCS-method occurred in three directions. First, more accurate refinements of the overflow distributions themselves were made following the

"equivalent random method" of Wilkinson [20]. Second, more complicated network hierarchies were introduced, see for example Rapp [15]. The third advance involved incorporating traffic overflows and constraints on blocking probabilities for more than one time of day in the same cost minimization model, see Rapp [15] and Eisenberg [5]. It appears that it is necessary to consider overflow traffic for multiple times of day in order to determine trunk group sizes which meet stated blocking probability constraints. In addition, networks based on field data have been reported in Eisenberg [5] and Elsner [6] where potential cost savings may be realized by incorporating multiple times of day.

The *second major approach* to determine levels of telecommunications equipment appeared in the 1956 paper of Kalaba and Juncosa [11]. Their approach is based on a linear programming model for a classical routing problem having variable link capacities, and as such is a large scale one. Several contrasts to the first approach (embodied in the ECCS method) are apparent.

First, the parcels of traffic in the Kalaba-Juncosa model are deterministic. Traffic originating at junction i and terminating at junction j is a given constant, a_{ij}. Second, demands are specified for each year (or other relevant time period), in contrast to a specification for multiple "hours" within a fixed time period. Consequently, link capacities may be specified for ensuing future periods, but the impact of multiple busy periods within a given period has not been modeled.

In spite of severe deterministic assumptions, the pioneering linear programming model of Juncosa and Kalaba can theoretically accommodate all conceivable routing possibilities, for their traffic variables are indexed by an origin-destination point pair and also an intermediate switching point, over all possible triads.

About 5 years after the Juncosa-Kalaba paper, a series of papers written by Gomory and Hu on communication network flows appeared in the SIAM Journal [8], [9], [10]. Their work occurred over a 4-year period and expanded significantly the size of the linear programming network models that could be treated computationally. They were able to combine features of generalized linear programming decomposition techniques with efficient Ford-Fulkerson methods for solving network subproblems. Gomory and Hu also stressed the importance of including communications demands indexed by time, such as time of day, t. They proceeded under the reasonable assumption that the time variable assumes only a finite number of values. Alternatively, one could employ a continuous load curve with time-of-day varying demand.

Gomory and Hu illustrated their computational approach on a 10-node, 20-arc network with demands for two different time periods, and a given set of unit capacity (expansion) costs.

Based on discussions with engineers in the field, principally from the Long Lines Company of AT & T, we have found that both approaches have had significant impact in the actual design of telecommunications networks. The completely deterministic approach (the second approach) has been particularly important in delineating first choice, second choice, etc. alternate routes between pairs of junctions to be used in defining a network hierarchy. Once a network hierarchy is established, economies of scale are then achievable according to optimal use of the underlying probability distributions of originating and alternately routed customer traffic.

A convenient characterization of a network hierarchy, very useful to our approach, is introduced in the next section.

3. CHARACTERIZATION OF A NETWORK HIERARCHY

3.1. The Hierarchy Matrix

Given a network hierarchy such as Figure 1, let us list the junctions, termed *points*, as $p_1, p_2, \ldots, p_q$ where q is a positive integer. By a *calling pair* we shall mean an ordered pair of distinct points (a,b), "a" being referred to as *origin* and "b" as *destination*.

In Section 1 we defined what is meant by direct and final links. Any two distinct points may or may not be joined by a link, but no two points can be joined by both a direct and a final link. Each link may carry traffic for each of its two calling pairs, since traffic may flow in either direction between the two points it joins. For any calling pair (a,b) we assume that a call can be routed via a unique sequence of final links, which we shall term *the final routing of the calling pair* (a,b).

Let us list the set of final links by the positive integers, $J = 1, 2, \ldots, K$. We list the set of calling pairs also by positive integers, $i = 1, 2, \ldots, N_0$, where $N_0 = q(q-1)$ is the total number of calling pairs in the network.

For purposes of algebraic representation we display final routing as a matrix which has a row for each calling pair i and a column corresponding to each final link, J. We term this matrix the *hierarchy matrix*, denoted $[\pi_{iJ}]$, and specify the entry in the i-th row and J-th column to be a nonnegative integer defined as follows:

$$\pi_{iJ} = \begin{cases} \text{the integer-valued position of the } J\text{-th final link in the final routing of} \\ \text{calling pair } i, \text{ if final } J \text{ belongs to this sequence} \\ 0, \text{ otherwise.} \end{cases} \tag{1}$$

Observe that the row indices of the nonzero entries in the J-th column represent all the calling pairs which utilize final J in the final routing of calls. We denote the set of these nonzero indices Π_J.

A certain subset of the calling pairs may also be served by direct links, such as the ones drawn as dashed lines in Figure 1. These calling pairs are known as *high usage* calling pairs, and the direct links as high usage links. The case where there are no high usage links may be treated without loss of generality as one with high usage links having 0 number of trunks. Each high usage link provides a direct, first choice route exclusively for traffic between its endpoints, in both directions, while the remaining nonhigh usage calling pairs rely solely on final routing. Overflow traffic from a high usage calling pair shall merely follow the uniquely specified final routing for the pair.

Each high usage link is associated with two high usage calling pairs, each with the same points, but oppositely ordered. Thus, if there are M number of high usage links, there are $2M$ number of high usage calling pairs, and $2M$ is an even integer. Observe also that N_0, as the product of an odd and an even integer, is itself an even integer, and so for some integer N, $N_0 = 2N$.

This discussion suggests relabelling the calling pairs using the integers $-N, \ldots, -2, -1, 1, 2, \ldots, N$. For instance, 1 and -1 represent pairs of the same two points, but with opposite ordering, meaning the opposite direction for traffic. Let us further specify that the integers $-M, \ldots, -1, 1, \ldots, M$ are reserved for high usage pairs. Since existence of final

links is assumed and no calling pair is joined by both a direct and a final link, it follows that $M \leq N$. Moreover, i is a nonhigh usage calling pair if and only if $|i| > M$.

Consider Figure 1 for the purposes of illustration. There are 8 nodes, so there are $8(8-1) = 56$ calling pairs. Thus $N = 56/2 = 28$. There are 7 final links, and 8 high usage links, hence 16 high usage pairs, labelled $-8, -7, \ldots, -1, 1, 2, \ldots, 8$. The remaining calling pairs, labelled $-28, \ldots, -9, 9, \ldots, 28$ are serviced only by final routing. A portion of the hierarchy matrix is given in Table I. The full matrix has 56 rows, and 7 columns.

TABLE I. *A Portion of the Hierarchy Matrix of Figure 1*

Calling Pair and Its Integer Index		Final Link and Its Integer Index: AB	AC	AD	BE	CF	CG	CH
		1	2	3	4	5	6	7
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
(G,B)	−3	3	2	0	0	0	1	0
(F,B)	−2	3	2	0	0	1	0	0
(E,A)	−1	2	0	0	1	0	0	0
(A,E)	1	1	0	0	2	0	0	0
(B,F)	2	1	2	0	0	3	0	0
(B,G)	3	1	2	0	0	0	3	0
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
(A,F)	9	0	1	0	0	2	0	0

By labelling the high usage links by the integers $1, \ldots, M$, we can have link I correspond to the high usage pairs $-I$ and $+I$. We then relabel the final links as $M+1, \ldots, M+K$, and relabel the rows and columns of the hierarchy matrix in the same manner as we did the calling pairs and final links, respectively. Thus, $-N$ refers to the first row in the matrix, and $(M+1)$ the first column, but this departure from orthodox notation is compensated for by added convenience. In practice it is only a matter of defining two label vectors.

To summarize the listings, when we write "link (or trunk group) L," "high usage link I", and "final link J" it shall be understood that $L \in \{1, \ldots, M+K\}$, $I \in \{1, \ldots, M\}$, and $J \in \{M+1, \ldots, M+K\}$ respectively. Similarly for "calling pair j" and "high usage calling pair i," $j \in \{-N, \ldots, -1, 1, \ldots, N\}$ and $i \in \{-M, \ldots, -1, 1, \ldots, M\}$ respectively.

3.2. Classifying Point-to-Point Offered Loads

For each calling pair j there is a nonnegative demand for traffic denoted a_j termed *originating traffic*. Traffic is usually stated in units of erlangs, or in hundred call-seconds per hour (CCS) as discussed in Section 1.

Let J be a fixed final link. Traffic parcels offered to J consist of two types: overflow traffic from high usage calling pairs, and final-routed traffic from nonhigh usage calling pairs. The types are distinguished because of their different probability distributions, as seen in the next section.

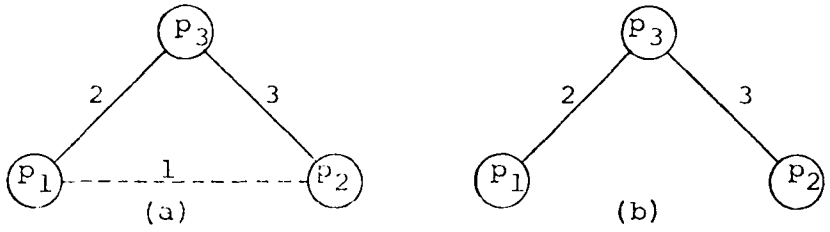
Because of this distinction, it is useful to separate Π_J into two subsets: $\Pi_J^0 = \{j \in \Pi_J : |j| \leq M\}$, i.e., the high usage calling pairs overflowing onto final link J, and $\Pi_J^1 = \{j \in \Pi_J : |j| > M\}$, i.e., those nonhigh usage calling pairs utilizing J in their final routing. Clearly,

$$\Pi_J = \Pi_J^0 \cup \Pi_J^1$$

and

$$\Pi_J^0 \cap \Pi_J^1 = \emptyset.$$

Two simple examples from Figure 2 below illustrate this classification, where in both networks (a) and (b) final links are designated by 2 and 3. Network (b) has no high usage calling pairs, while in (a) the single high usage link is denoted by 1.



Numbering Scheme for pairs in (a) and (b):

Calling pair	*Index Number*
(p_3,p_2)	-3
(p_3,p_1)	-2
(p_2,p_1)	-1
(p_1,p_2)	1
(p_1,p_3)	2
(p_2,p_3)	3

FIGURE 2. Two triads

In (a), $\Pi_2^0 = \Pi_3^0 = \{-1, 1\}$, while $\Pi_2^1 = \{-2, 2\}$, and $\Pi_3^1 = \{-3, 3\}$. In (b), $\Pi_2^0 = \Pi_3^0 = \emptyset$, while $\Pi_2^1 = \{-2, -1, 1, 2\}$ and $\Pi_3^1 = \{-3, -1, 1, 3\}$.

4. THE FORMULATION OF A NONLINEAR SUPPLY MODEL

4.1. Blocking Probabilities and Overflow Traffic

The call discipline is one of the factors in determining the relationship between the offered load to a trunk group and its carried load. Another key factor in determining carried loads is the assumption that customer originated traffic is Poisson distributed with arrival rate denoted by λ, see Messerli [13]. Fortunately, there is strong evidence to suggest that the number of calls occurring in a fixed, small time interval can be adequately modeled as a Poisson probability distribution. With these assumptions the distinction between a trunk group's offered load and carried load can now be made precise.

Assume that calls are assigned sequentially to a trunk group consisting of n trunks. Let λ denote the average customer arrival rate according to the Poisson distribution. The only assumption required on customer calling time is that it has finite mean μ. Otherwise, it may be arbitrarily distributed. Under these conditions the probability that all of the n trunks in the group are busy is given by the classical Erlang B-formula:

$$B(n,a) = (a^n/n!)/\sum_{k=0}^{n} (a^k/k!), \tag{2}$$

for $n = 0, 1, \ldots$, *where* $a = \lambda\mu$ *with its units termed erlangs*. The history of the original Erlang formula and its important generalizations may be found in Gnedenko-Kovalenko [7] and Syski [18].

An erlang is thus a measure of the flow of traffic per unit time. In the traffic engineering literature an erlang is one call-hour per hour, or equivalently 36 CCS per hour. The "hour" as the unit of time is so standard, it is usually dropped, and one says an erlang is 36 CCS. The value "a" in the Erlang formula is termed the mean of the offered load to the given trunk group. The expected overflow traffic is then $aB(n,a)$. With traffic flowing in both directions, similar formulas apply.

Suppose for some integer i, $-M \leqslant i \leqslant M$, a traffic intensity a_i from high usage calling pair i is offered to high usage link I consisting of x_I number of trunks, where $I = |i|$. (Through the paper we shall always take $I = |i|$ in that context.) According to (2) above, the probability that all x_I trunks are busy is $B(x_I, a_i + a_{-i})$, recalling traffic intensity a_{-i} running in the reverse direction shares the trunks on I. Hence, $a_i B(x_I, a_i + a_{-i})$ is the expected amount of traffic overflowing to the first link, J, in the final route sequence of i. Final link J, however, carries other parcels of traffic as well, as seen in Section 3.2: overflows from the other high usage calling pairs represented in Π_J^0, and traffic from the nonhigh usage pairs represented in Π_J^1. We next formalize the idea of the quality of service of the network and introduce a useful assumption on the marginal capacity of a trunk group.

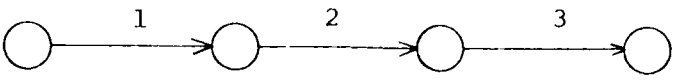
4.2. Network Quality of Service and an Assumption on Marginal Capacities

The important benefits of being able to compute changes in equipment stock to meet changes in demand were recognized much earlier by Kalaba and Juncosa [11], Gomory and Hu [8], [9], [10] and others. Fortunately, incremental studies on the network hierarchy introduced in Section 3 permit certain simplifying assumptions that make computations attractive. These assumptions relate to the concept of the marginal capacity of an additional trunk at a prescribed blocking probability. The resulting supply model is an optimization which is much simpler than would be possible when constructing a network *ab initio*. The assumptions and model are now presented.

DEFINITION: Assume Q_L^t, the traffic offered to a given link L at time t, has a fixed probability distribution, and let $m(Q_L^t)$ denote its mean. Given that L consists of x_L number of trunks, let $\mu_I(x_I, Q_I^t)$ denote the mean of the overflow distribution. The *carried load* is that portion of the offered load which does not overflow. For the case that L is a final link J, the *quality of service* ρ_J of final link J is defined by $\rho_J = 1 - \rho_J'$, where

$$\rho_J' = \max_t \frac{\mu_I(x_I, Q_I^t)}{m(Q_I^t)}.$$

Since the overflow is less than the offered load, ρ'_J lies between 0 and 1. The mean of the carried load is $m(Q_J^f) = \mu_J(x_J, Q_J^f)$. According to the network hierarchy, overflow from any final link is lost. Let us illustrate how these definitions are employed in calculating carried loads on serially connected final links, in a simple example consisting of final links 1, 2, and 3 as shown:



Assume that the only offered loads on 2 and 3 stem from carried loads on 1 and 2, respectively. Assume that there is only one time of day t_0 and one quality of service ρ. Thus,

$$\frac{\mu_J(x_J, Q_J)}{m(Q_J)} = \rho',$$

and $m(Q_{J+1}) = m(Q_J) - \mu_J(x_J, Q_J)$ for $J = 1, 2$. Hence, $m(Q_{J+1}) = m(Q_J)\rho$, $J = 1, 2$ and so $m(Q_3) = m(Q_1)\rho^2$.

A formal extension of this argument shows that for any final link J the mean of the overflow from any high usage calling pair $i \in \Pi_J^0$ is at least

$$a_i' B(x_J, a_i' + a'_{\cdot J})\rho^{(\pi_{iJ}-1)} \tag{3}$$

providing Π_J^0 is nonempty and where $\rho = \max\{\rho_J | J = M+1, \ldots, M+K\}$.

Marginal Capacity Assumption

Let ρ be fixed. For each J we assume that there exist two positive constants γ_J and b_J such that if τ^+ and τ^- are two offered loads having probability distributions, and $m(\tau^+) > 0$, then

$$\max_J \frac{\mu_J\left(x_J + \left\langle \frac{m(\tau^+)}{\gamma_J} \right\rangle, Q_J^f + \tau^+\right)}{m(Q_J^f + \tau^+)} \leqslant \rho' \ (= 1 - \rho), \tag{4a}$$

and if $0 < m(\tau^-) \leqslant b_J$ then

$$\max_J \frac{\mu_J\left(x_J - \left[\frac{m(\tau^-)}{\gamma_J}\right], Q_J^f - \tau^-\right)}{m(Q_J^f - \tau^-)} \leqslant \rho', \tag{4b}$$

where $\langle x \rangle$ is the smallest integer greater than or equal to x, termed the integer roundup of x, and where $[x]$ is the largest integer less than or equal to x termed the integer part of x. γ_J is termed the *marginal capacity of an additional trunk at quality of service* ρ.

Inequality (4a) states that when $\langle m(\tau^+)/\gamma_J \rangle$ number of trunks are added to final link J, then at least an additional amount of traffic $\rho\tau^+$ is carried. Inequality (4b) states that when $[m(\tau^-)/\gamma_J]$ number of trunks are removed from the trunk group, then the decrease in carried traffic is at most $\rho\tau^-$.

We assume throughout that each high usage group I consists of x_I (integer) number of trunks, and that each final group J consists of x_J number of trunks, establishing what we term the *existing network*. It is further assumed that the existing network can supply all service demanded a_i' for all pairs i and all times of day t with the provision of a quality of service ρ.

4.3. A Nonlinear Integer Programming Formulation for the Network Hierarchy of Section 3

The first task is to develop an expression for the sum of the traffic parcels of Section 3.2 offered to a final link J of the *existing network*.

4.3.1. Sum of Overflow Parcels Offered to Final J

Summing the mean overflows in (3) yields a lower bound for the mean total overflow traffic parcels offered to trunk group J. Let this sum be denoted by $L_J^{\text{II}}(t)$, i.e.,

$$L_J^{\text{II}}(t) = \sum_{M}^{M} a^i B(x_J, a_J^i + a^i{}_J)\rho^{(\pi_{iJ}-1)}\left(\frac{\pi_{iJ}}{\pi_{iJ}}\right) \tag{5}$$

for each final trunk group J, where we define $\left(\frac{\pi_{iJ}}{\pi_{iJ}}\right) = 0$ if $\pi_{iJ} = 0$. (This convention shall be used throughout the paper.) Thus, a term in the summation stemming from calling pair i is automatically set to 0 if final link J does not belong to the final routing of i. For the case that Π_J^{II} is empty, (5) automatically reduces to zero, a case, for example, which does not occur in Figure 1. An upper bound on the total overflow traffic to J is obtained by deleting the ρ-term in expression (5).

4.3.2. Sum of Parcels Offered to Final J from Nonhigh Usage Calling Pairs

For any $k \in \Pi_J^{\text{I}}$ it follows analogously to (4) that the expected portion of originating traffic parcel a_k^i offered to trunk group J is $a_k^i\rho^{(\pi_{kJ}-1)}$, provided that Π_J^{I} is nonempty. Summing all these parcels of traffic yields a sum which we denote $L_J^{\text{I}}(t)$:

$$L_J^{\text{I}}(t) = \sum_{M+K+N} a_k^i\rho^{(\pi_{kJ}-1)}\left(\frac{\pi_{kJ}}{\pi_{kJ}}\right). \tag{6}$$

4.3.3. A Constraint on the Sum of All Traffic Offered to Final J

The maximum total expected offered load E_J, which final group J of the existing network can service at blocking probability $1 - \rho$ is the maximum, over all times of day t, of the sums of both types of expected offered load parcels. Accordingly,

$$E_J = \max\{L_J^{\text{II}}(t) + L_J^{\text{I}}(t)\}. \tag{7}$$

Our modeling approach is basically an incremental one involving (i) modified offered loads $\tilde{a}_J^i$ for all pairs i, (ii) modifications of the number of trunks $\tilde{x}_J$, $L = 1, \ldots, M + K$, and (iii) a modification in the network service quality $\tilde{\rho}$. Under these three kinds of modifications, we may define quite analogously to (5) and (6) the expressions

$$\tilde{L}_J^{\text{II}}(t) \text{ and } \tilde{L}_J^{\text{I}}(t),$$

and analogous to (7) write

$$\tilde{E}_J = \max\{\tilde{L}_J^{\text{II}}(t) + \tilde{L}_J^{\text{I}}(t)\}. \tag{8}$$

The difference $\tilde{E}_J - E_J$ is the mean of the additional traffic distribution on final link J and so our assumption on capacities applies. Therefore, if $\tilde{E}_J - E_J > 0$, then by case (4a), only $\langle(\tilde{E}_J - E_J)/\gamma_J\rangle$ number of trunks need be added to final group J, where γ_J is the marginal

capacity of an additional trunk at blocking probability $1 - \tilde{\rho}$. Let γ_J denote the integer number of trunks required in group J in order to service initial demand E_J at the new service quality $\tilde{\rho}$. Hence, we obtain a feasibility requirement on the modified J trunk group size $\tilde{x}_J$,

$$\tilde{E}_J - E_J \leq \gamma_J(\tilde{x}_J - Y_J) \tag{9}$$

where $\tilde{x}_J$ is integer.

If $\tilde{E}_J - E_J < 0$, then we invoke a stronger version of the marginal capacity assumption regarding case (4b). We require that $\tau^- \equiv |\tilde{E}_J - E_J|$, a quantity which depends on the $\tilde{x}_J$ and certain $\tilde{x}_I$ (high usage size) variables, lie within the 0 to b_J range required in order for (4b) to hold. In other words, when $\lfloor |\tilde{E}_J - E_J|/\gamma_J \rfloor$ number of trunks are removed from Y_J, the resulting number of trunks,

$$\tilde{x}_J = Y_J - \lfloor |\tilde{E}_J - E_J|/\gamma_J \rfloor$$

may be offered the modified load at blocking probability $(1 - \tilde{\rho})$. It follows that the same feasibility requirement as (9) holds for this case too.

The system of inequalities (9), one inequality for each final group J, shall determine a set of constraints for the nonlinear supply model, and we shall write these constraints in greater detail when actually specifying the model. But, first we need to take account of the total switched traffic in the network.

4.3.4. Accounting for Total Switched Traffic

Let us work with the modified loads $\tilde{a}_i^I$, modified number of trunks $\tilde{x}_J$, and modified service quality $\tilde{\rho}$.

Let $\tilde{S}_I$ denote the total switched traffic throughout the network at time I. We shall now show that

$$\tilde{S}_I = \sum_{J=M+1}^{M+K} \sum_{\substack{i=1 \\ i \neq 0}}^{M} \{\tilde{a}_i^I B(\tilde{x}_J, \tilde{a}_i^I + \tilde{a}^I{}_J)\, \tilde{\rho}^{(\pi_{iJ}-1)}\} \left(\frac{\pi_{iJ}}{\pi_{iJ}}\right) + \sum_{J=M+1}^{M+K} \sum_{M<k\leq N} \{\tilde{a}_k^I\, \tilde{\rho}^{(\pi_{kJ}-1)}\} \left(\frac{\pi_{kJ}}{\pi_{kJ}}\right). \tag{10}$$

The amount of overflow traffic from high usage calling pair i destined for final J is $\tilde{a}_i^I B(\tilde{x}_J, \tilde{a}_i^I + \tilde{a}^I{}_J)$. However, before this particular parcel reaches J it must be consecutively switched at the point of origin, and the $(\pi_{iJ} - 1)$ points along the alternate route. Therefore, in this case the amount of switched traffic is:

$$\tilde{a}_i^I B(\tilde{x}_J, \tilde{a}_i^I + \tilde{a}^I{}_J)\, \tilde{\rho}^{(\pi_{iJ}-1)} \tag{11}$$

The same analysis applies to traffic from calling pairs served only by final routing. The traffic switched due to the originating load $\tilde{a}_k^I$, $k \in \Pi_J^I$ requiring final J for completion is

$$\tilde{a}_k^I\, \tilde{\rho}^{(\pi_{kJ}-1)}. \tag{12}$$

We now sum (11) over all high usage pairs and then over all finals. Similarly, (12) is summed over all remaining pairs and then over all finals. Finally, summing these two yields (10).

4.3.5. Cost Assumptions and the Nonlinear Model

Analogous to Eisenberg [5] and Elsner [6] we shall invoke simplifying cost assumptions for trunks and switching. We shall employ unit marginal investment costs per trunk and shall use the same cost for augmenting a trunk group as for diminishing a trunk group.* We shall denote the marginal cost per trunk for trunk group L by $c_L > 0$, $L = 1, \ldots, M + K$.

Changes in switching investment costs shall be approximated by using a marginal switching investment cost c per CCS of switched traffic, as for example in Eisenberg [5].

In the absence of real data and analogous to Eisenberg [5] we can merely set $c_L = \$1000$ for each trunk in group L, final or high usage, and also set $c = \$62$ (per CCS).

We are now ready to state the basic nonlinear programming supply model.

PROGRAM P: Assume an existing network (Section 3) has demands a_i^t for all pairs $i = -N, \ldots, -1, 1, \ldots, N$ integer group sizes x_L for high usage and final groups: $L = 1, \ldots, M + K$, and an overall network service probability ρ with marginal capacities γ_J, $J = M + 1, \ldots, M + K$. Let modified positive demands be denoted by $\tilde{a}_i^t$, and let $\tilde{\rho}$ denote a modified service probability with marginal capacity $\tilde{\gamma}_J$, $J = M + 1, \ldots, M + K$. Assume that c_L is the cost per trunk on trunk group L, $L = 1, \ldots, M + K$, and that c denotes the switching cost per CCS. Let E_J be defined according to (7). Compute

$$M_P = \min \sum_{L=1}^{M+K} c_L \tilde{x}_L + c\tilde{S} \tag{13a}$$

from among nonnegative integers $\tilde{x}_L$ for $L = 1, \ldots, M + K$, and real $\tilde{S}$ which satisfy:

$$\sum_{\substack{i=-M \\ i\neq 0}}^{M} \tilde{a}_i^t B(\tilde{x}_J, \tilde{a}_i^t + \tilde{a}^t{}_i)\tilde{\rho}^{(\pi_{iJ}-1)}\left(\frac{\pi_{iJ}}{\pi_{iJ}}\right) + \sum_{M<k\leq N}^{M} \tilde{a}_k^t \tilde{\rho}^{(\pi_{kJ}-1)}\left(\frac{\pi_{kJ}}{\pi_{kJ}}\right) - E_J \leqslant \tilde{\gamma}_J(\tilde{x}_J - Y_J) \tag{13b}$$

for each final J and each t, where Y_J is the required number of trunks in J for a $\tilde{\rho}$ service probability, the B-function given in (2), and

$$\sum_{J=M+1}^{M+K} \sum_{\substack{i=-M \\ i\neq 0}} \{\tilde{a}_i^t B(\tilde{x}_J, \tilde{a}_i^t + \tilde{a}^t{}_i)\,\tilde{\rho}^{(\pi_{iJ}-1)}\}\left(\frac{\pi_{iJ}}{\pi_{iJ}}\right) + \sum_{J=M+1}^{M+K} \sum_{M<|k|\leq N} \{\tilde{a}_k^t\,\tilde{\rho}^{(\pi_{iJ}-1)}\}\left(\frac{\pi_{kJ}}{\pi_{kJ}}\right) \leqslant \tilde{S} \tag{13c}$$

for each t.

*In practice, one rarely takes away existing equipment, but merely waits until the normal growth in message volume takes up the current slack.

Observe that the system of inequalities (13b) is merely (9) with full detail of the terms $\tilde{E}_l$ showing the $\tilde{x}_l$ as variables. On the other hand (13c) merely defines the maximum switched traffic in the network according to (10).

It is obvious that Program P is consistent because the $\tilde{x}_l$ variables may be taken arbitrarily large as well as the $\tilde{S}$ variable. P must have a finite minimum. Otherwise, some $\tilde{x}_l$ necessarily become arbitrarily large and since all cost coefficients are positive, the objective function would arbitrarily increase which is a contradiction.

Program P is a nonlinear integer programming problem which can be well approximated for practical purposes by a continuous convex program. In fact, even more can be done. Program P can be approximated by a finite linear program based on the special convexity property and monotonicity property of the Erlang *B*-function, see Messerli [13]. We focus now on how the linear programming approximation is constructed.

5. A LINEAR PROGRAMMING APPROXIMATION TO THE NONLINEAR PROGRAM P

5.1. The Convexity Properties of the Blocking Probabilities

In engineering practice, the definition of the "load on last trunk" with respect to a trunk group of size $n + 1$ which is offered the load "a" is defined by

$$D(n,a) = B(n,a) - B(n+1,a) \tag{14}$$

where the Erlang *B*-function is defined in (2), for $n = 0, 1, \ldots$, with $B(0,a) = 1$. Observe that $D(n,a) > 0$ for each nonnegative integer n. Messerli [13] gives a proof that for any fixed $a > 0$, $D(n,a)$ is strictly decreasing in the nonnegative integer variable n,

$$D(n+1,a) < D(n,a) \tag{15}$$

for $n = 0, 1, \ldots$

For "a" fixed define the polygonal function $\hat{B}(\cdot,a)$ from the non-negative reals to the nonnegative reals by

$$\hat{B}(x,a) = -D(n,a)x + (n+1)B(n,a) - nB(n+1,a), \tag{16}$$

where n is the integer part, $[x]$, of x. Note that $\hat{B}(r,a) = B(r,a)$ for each nonnegative integer r

The graph of the polygonal function $\hat{B}(\cdot,a)$ reveals its convexity and monotonicity properties, which are basic for the construction of the linear program.

For each nonnegative integer n the left-hand side of (16) defines an affine function on the nonnegative reals. The following cumulative-type expression for this affine function follows from Charnes-Cooper [1], pages 352-353.

For a fixed nonnegative integer n

$$-D(n,a)x + (n+1)B(n,a) - nB(n+1,a) = 1 + \sum_{r=0}^{n}(c_r - c_{r-1})(x - r) \tag{17}$$

for every real nonnegative x, where $c_{-1} = 0$ *and* $c_r = -D(r,a)$ *for* $r = 0, 1, \ldots$.

As strongly suggested by Figure 3, the following proposition yields a uniquely determined system of supporting hyperplanes for the epigraph K of the function $\hat{B}(\cdot,a)$. The proposition and its three corollaries shall be proved in the Appendix.

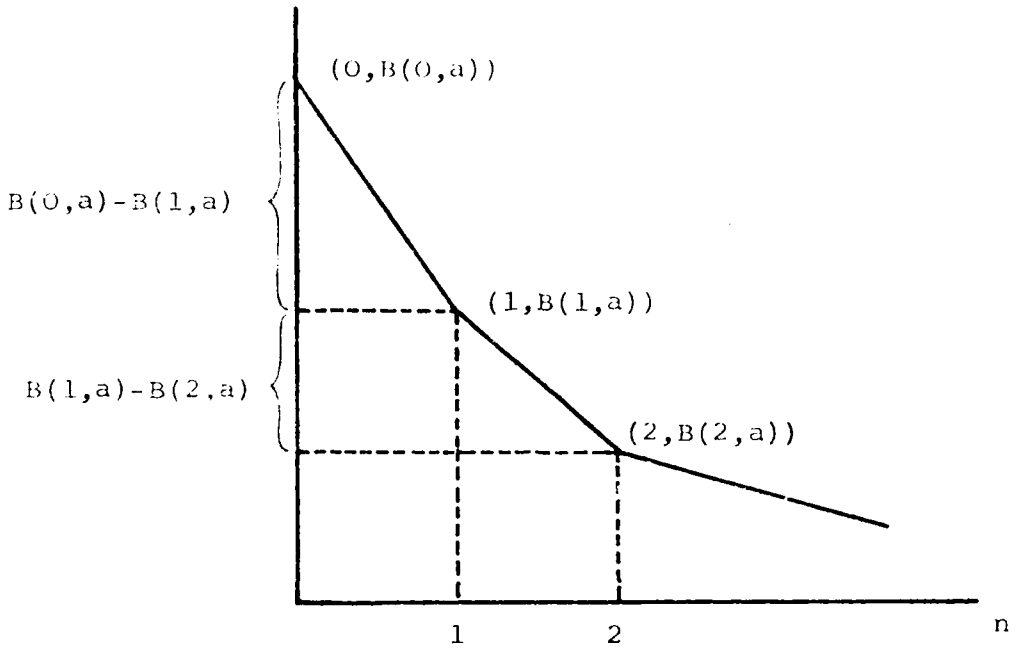


FIGURE 3 The polygonal function determined by the Erlang B-function on nonnegative integers

PROPOSITION 1: Let K be the epigraph of $\hat{B}(\cdot,a)$, $K = \{(z,x) \in \mathbb{R}^2 | x \geqslant 0$ and $z \geqslant \hat{B}(x,a)\}$. Let L be the set of all (z,x) in $\mathbb{R}^2$ which satisfy the semi-infinite system of linear inequalities

$$z - 1 \geqslant \sum_{r=0}^{n} (c_r - c_{r-1})(x - r) \tag{18}$$

for $x \geqslant 0$ and $n = 0, 1, 2, \ldots$.

Then $K = L$ and K is nonempty.

COROLLARY 1: Let x be nonnegative real. The $(\hat{B}(x,a),x)$ satisfies each inequality of (18) strictly except for (i), the inequality *indexed by* $\lfloor x \rfloor$ i.e., the inequality

$$z - 1 \geqslant \sum_{r=0}^{\lfloor x \rfloor} (c_r - c_{r-1})(x - r),$$

which it satisfies as an equality, and (ii) possibly the inequality indexed by $\lfloor x \rfloor - 1$ when $x \geqslant 1$. The latter inequality is satisfied as an equality if and only if x is a positive integer.

COROLLARY 2: Let U be a positive integer and set $K' = K \cap \{(z,x) | 0 \leqslant x \leqslant U\}$. Let L' be the set of all (z,x) which satisfy

$$z - 1 \geqslant \sum_{i=0}^{n} (c_i - c_{i-1})(x - i), \quad x \geqslant 0$$

for $n = 0, 1, \ldots, V - 1$. Then $K' = L'$.

COROLLARY 3: $(z,x) \in K'$ is an extreme point of K' if and only if x is a nonnegative integer and $z = B(x,a)$.

In view of Figure 3, which reflects the basic integer convexity property (15), these results are intuitively clear. They are formally proved in the Appendix.

5.2. The Key Approximation and the Linear Program

We now replace in Program P the B-function by the polygonal $\hat{B}$-function, and the integrality conditions on the $\tilde{x}_J$ variables are removed. Finally, upper bounding constraints $\tilde{x}_J \leqslant V_J$ are imposed, where the V_J are large positive integers.

The next step replaces each term $\tilde{a}_i^t B(x_J, \tilde{a}_i^t + \tilde{a}^t{}_{-1})$ in (13b) and (13c) with the new variable z_i^t and requires that

$$\tilde{a}_i^t \hat{B}(x_J, \tilde{a}_i^t + \tilde{a}^t{}_{-1}) \leqslant z_i^t.$$

The new approximation program so obtained, denoted P', is the following.

PROGRAM P': Same assumptions set as in P. Let V_J be large positive integers for high usage links. Compute

(19a) $$M_{P'} = \min \sum_{J=1}^{M+K} c_J \tilde{x}_J + c\tilde{S}$$

from among reals $\tilde{x}_J$, z_i^t, and $\tilde{S}$ which satisfy:

(19b) $X_J(t) \leqslant \tilde{x}_J$, where $X_J(t) =$

$$\left\{ \sum_{\substack{i=-M \\ i \neq 0}}^{M} z_i^t \rho^{(\pi_{iJ} - 1)} \left(\frac{\pi_{iJ}}{\pi_{iJ}} \right) + \sum_{M < |k| \leqslant N} \tilde{a}_k^t \rho_k^{(\pi_{kJ} - 1)} \left(\frac{\pi_{kJ}}{\pi_{kJ}} \right) - E_J + \tilde{\gamma}_J v_J \right\} / \tilde{\gamma}_J$$

for each final J and each t, and

(19c) $S(t) \leqslant \tilde{S}$, where $S(t) =$

$$\sum_{J=M+1}^{M+K} \sum_{\substack{i=-M \\ i \neq 0}}^{M} \{ z_i^t \rho^{(\pi_{iJ} - 1)} \} \left(\frac{\pi_{iJ}}{\pi_{iJ}} \right) + \sum_{J=M+1}^{M+K} \sum_{M < |k| \leqslant N} \{ \tilde{a}_k^t \rho^{(\pi_{kJ} - 1)} \} \left(\frac{\pi_{kJ}}{\pi_{kJ}} \right)$$

for each t, and

(19d) $$\tilde{a}_i^t \hat{B}(x_J, \tilde{a}_i^t + \tilde{a}^t{}_{-1}) \leqslant z_i^t$$

for each high usage calling pair i and time t, and

(19e) $$0 \leqslant \tilde{x}_J \leqslant V_J$$

for each high usage link J.

It is obvious now in view of Corollary 2 that P′ is equivalent to the finite linear program denoted LP′, obtained by replacing (19d) with the finite system of linear inequalities

$$\tilde{a}_i^l D(\nu, \tilde{a}_i^l + \tilde{a}^l{}_J)\tilde{x}_l + z_i^l \geq \tilde{a}_i^l(\nu + 1)B(\nu, \tilde{a}_i^l + \tilde{a}^l{}_J) - \tilde{a}_i^l \nu B(\nu + 1, \tilde{a}_i^l + \tilde{a}^l{}_J) \tag{20}$$

for $\nu = 0, 1, \ldots, V_l - 1$, and each high usage pair i, and each l. It is equally obvious that Program LP′ is consistent and has a finite minimum since the $\tilde{x}_l$ variables are bounded and all cost coefficients are positive. Hence P′ itself has optimal solutions.

We now use Corollary 1 of Proposition 1 to discuss the cost effects due to using an optimal solution of P′ as a solution to the integer program P. If high usage size $\tilde{x}_l^\bullet$ is not an integer, then $(B(<\tilde{x}_l^\bullet>, \tilde{a}^l + \tilde{a}^l{}_J), <\tilde{x}_K^\bullet>)$ is in the epigraph of $\hat{B}(\cdot, \tilde{a}_i^l + \tilde{a}^l{}_J)$ for each i, where $<\tilde{x}_l^\bullet>$ is the integer roundup. The roundup introduces an increase in the total cost associated with high usage group l, $(<\tilde{x}_l^\bullet> - \tilde{x}_l^\bullet)c_l$, where $0 \leqslant <\tilde{x}_l^\bullet> - \tilde{x}_l^\bullet < 1$. An offsetting cost effect from final groups J and switching $\tilde{S}$ occurs because from the monotonicity of $\hat{B}(\cdot, \tilde{a}^l + \tilde{a}^l{}_J)$, each z_i^l does not increase.

Finally, in order to insure quality of service $\tilde{p}$, noninteger final group sizes $\tilde{x}_l$ should be rounded up, thereby increasing total costs. Numerical estimates of these various offsetting cost effects due to round up of trunk group sizes determined by Program P′ have not been obtained. It appears to us that such estimates must stem from numerical experiments on field data. Certainly, as strongly suggested by Figure 3 and Proposition 1 and its Corollaries, integer programming pathologies from straightforward rounding processes do not occur.

Because of the linear inequality system (20), Program LP′ may be quite large and for practical purposes it would be useful to obtain an equivalent smaller problem in place of LP′. The monotonicity of the $\hat{B}$-function, essentially Corollary 1 of Proposition 1, suggests a useful procedure.

5.3. Solving the linear Program LP′ Through Bounded Variable Reductions

Let LP'_{BD} be the bounded variable version of LP′ obtained by replacing (19e) with

$$\alpha_l \leqslant \tilde{x}_l \leqslant \beta_l \tag{19e′}$$

for each high usage group, and in (20) restrict ν to: $\nu = \alpha_l, \ldots, \beta_l - 1$ where α_l and β_l are nonnegative integers such that $\beta_l - 1 - \alpha_l \geqslant 2$. The following is proved in the Appendix.

PROPOSITION 2: Under the above bounded variable assumptions:

(i) any optimal solution $\{(\tilde{x}_l^\bullet), \tilde{S}^*, (z_i^l)\}$ of LP'_{BD} is feasible for LP′, and

(ii) if for each high usage group l

$$\alpha_l < \tilde{x}_l^\bullet < \beta_l, \tag{21}$$

then this optimal solution is also optimal for Program LP′. Moreover, there exist α_l, β_l and an optimal solution of LP′ such that with respect to $\tilde{x}_l^\bullet$ of that solution, (21) holds.

6. COMPUTER PROGRAM AND RESULTS

6.1. Implementation of the Model

For all but the simplest network hierarchies, the large size of LP′ of Section 5.2 warrants the use of the bounded variable reduction program LP'_{BD}. Thus, according to Proposition 2 in Section 5.3, we may, in general, restrict $\tilde{x}_l$ in (19e′) to a range of 4 integers, that is, $\beta_l - \alpha_l = 3$ for each high usage link l. This in turn restricts ν to a range of 3 integers in (20). We shall also specify that there are a finite number, T, of periods (hours) during the calling day. For LP'_{BD}, the variables and constraints which occur are accounted for in Table 2.

TABLE 2. *The Variables and Constraints of Program* LP'_{BD}

Name			Number	Total
Variable	High Usage	$\tilde{x}_l$	M	$2M \times T + K + M + 1$
	Final	$\tilde{x}_J$	K	
	Switch	$\tilde{S}$	1	
	Overflow	z'_j	$2m \times T$	
Constraint	(19b)		$K \times T$	$T \times (K + 6M + 1) + 2M$
	(19c)		T	
	(20)		$2M \times 3 \times T$	
	(19e′)		$2M$	

For example, in Figure 1, $K = 7$, $M = 8$, and taking $T = 3$ we have a total of 64 variables and 184 constraints.

Let $A\xi \leqslant b$ denote the constraints system of LP'_{BD}, where $\xi = [(\tilde{x}_L), \tilde{S}, (z'_j)]^T$. It is easy to see that A is a sparse matrix; indeed, roughly 98% of its entries are zeros. Thus it requires some attention to enter each of these into its proper row and column. Figure 4 is a "blueprint" for the matrix A.

Calculating the entries of A requires computation of the erlang B-formula, (2) at integer values. However, the factorial terms involved quickly become too large for direct computation. Given some positive offered load a and positive number of trunks n, the following recursion is used:

$$B(n,a) = \frac{aB(n-1,a)}{n + aB(n-1,a)}; \quad B(0,a) = 1.$$

The "load on the last trunk" $D(n,a)$ which also appears in (20) merely requires computation of $B(n,a)$ and $B(n+1,a)$.

We need data on both the existing network and the modified network. As a simplifying assumption let us take $\rho = \tilde{\rho}$, meaning the quality of service is to be maintained at the same level. The necessary data then consist of a'_j and $\tilde{a}'_j$ for each calling pair j and for each time t, ρ, γ_J and $\tilde{\gamma}$ for each final J, and the sizes of the links on the existing network, x_L for each link L. Observe that since $\tilde{\rho} = \rho$, $Y_J = x_J$ for each final J, see Section 4.3.3. The hierarchy matrix introduced in Section 3.1 contains all the necessary information about final routing.

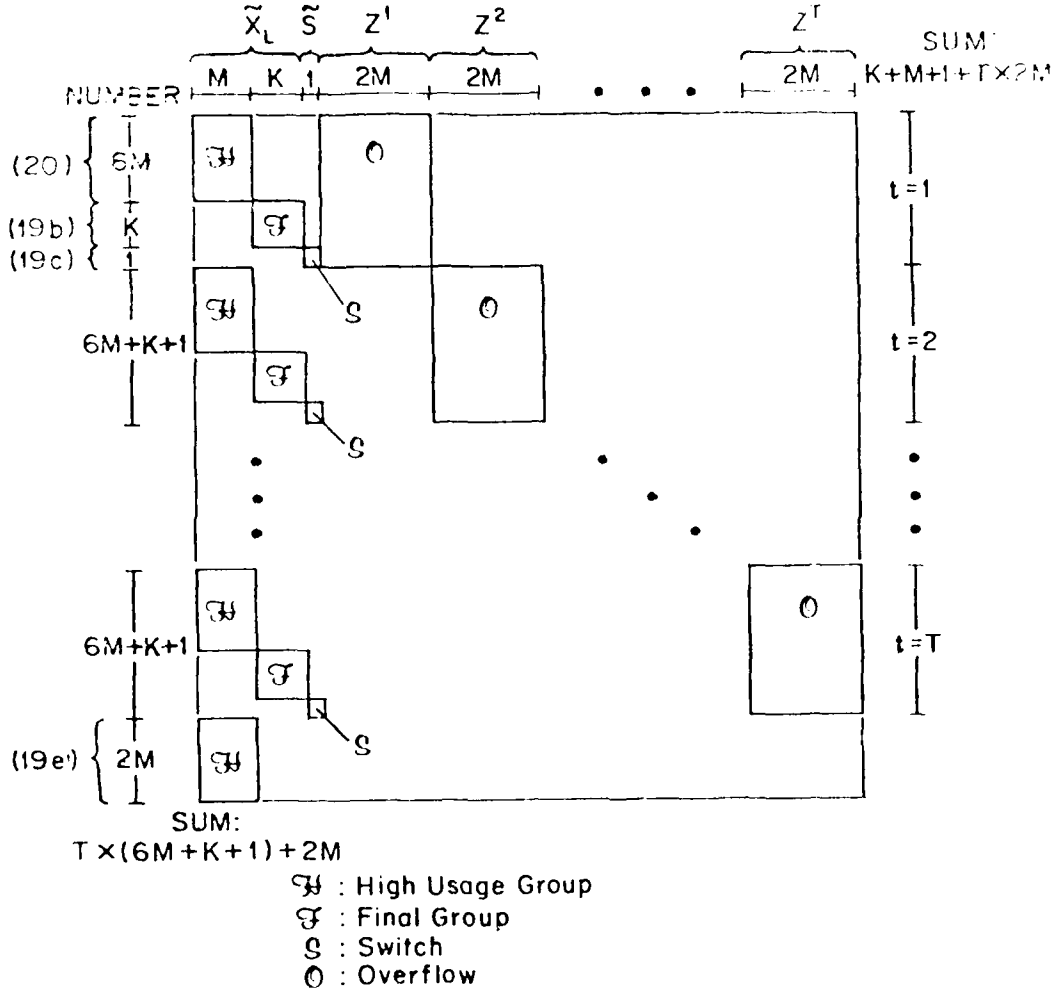


FIGURE 4. Structure of constraint A-matrix

With matrix A arranged as in Figure 4, the entries can be managed easily. Note that the block of the first $M + K + 1$ columns repeats itself for each time period while the overflow block shifts $2M$ columns for each consecutive time period. For $t = 1$, the first $6M + K + 1$ rows of A can be filled by the following piece of computer program. (Refer to Figure 5.)

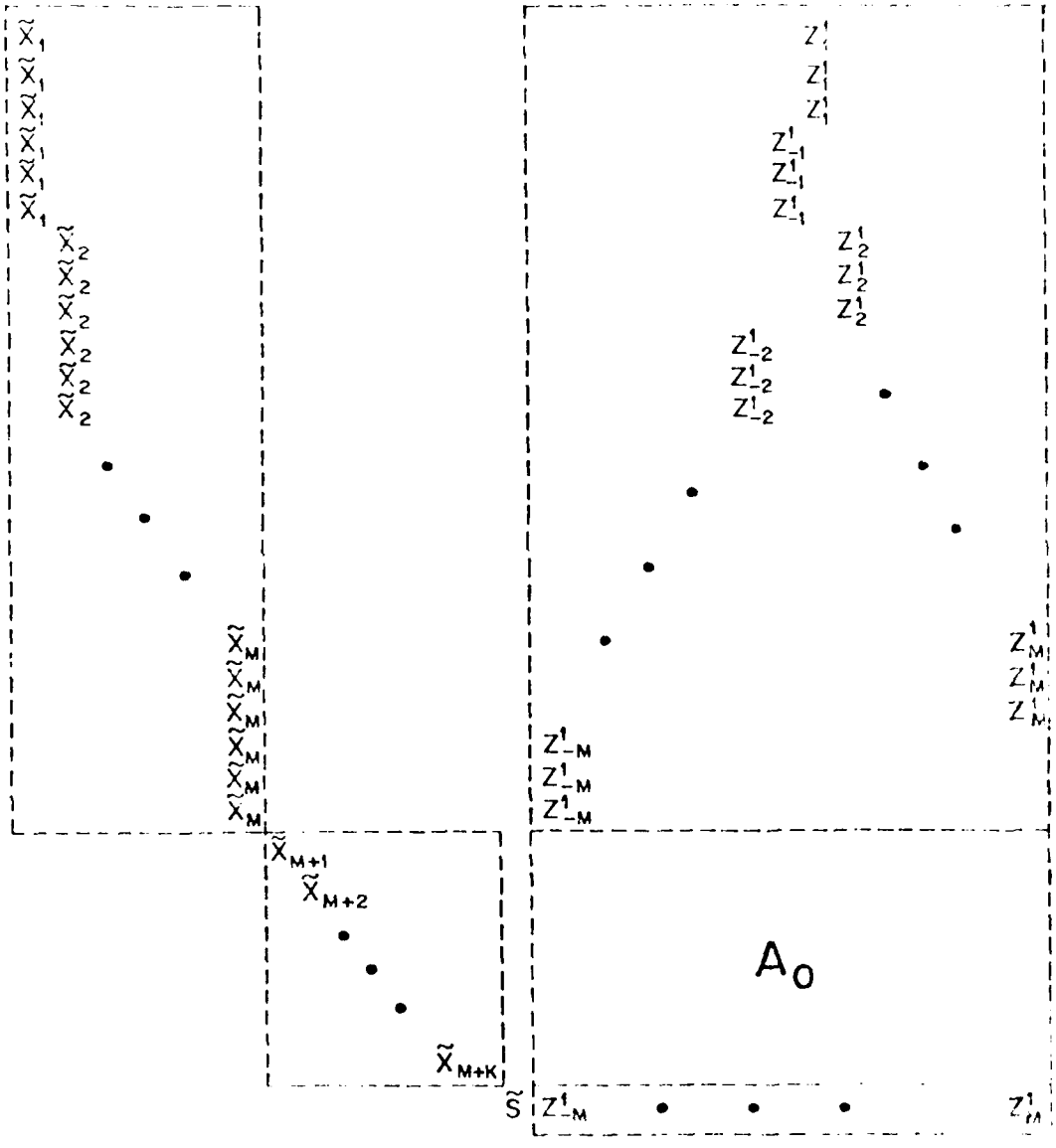


Fig. 4. Block of A matrix corresponding to $t = 1$ with variable in position of entries. Block A_0 consists of Z variables with nonzero coefficients determined by permissible overflows according to network metric.

Rows corresponding to (20):

```
FOR I:= 1 STEP 1 UNTIL M DO
FOR J:= 1 STEP 1 UNTIL 3 DO
BEGIN
        A[6*(I - 1) + J,I]:= -ã_I^1 D(α_I + J - 1,ã_I^1 + ã^1_I);
        A[6*(I - 1) + J + 3,I]:= -ã^1_I D(α_I + J - 1,ã_I^1 + ã^1_I);
        A[6*(I - 1) + J, 2M + K + 1 + I]:=
                A[6*(I - 1) + J + 3, 2M + K + 2 - I]:= -1
END;
```

Rows corresponding to (19b) and block A_0 of Figure 5:

```
FOR J:= M + 1 STEP 1 UNTIL M + K DO A[5M + I,I]:= -γ_I;
FOR I:= 1 STEP 1 UNTIL M DO
FOR J:= M + 1 STEP 1 UNTIL M + K DO
BEGIN
        A[5M + J,2M + K + 2 - I]:= IF π_IJ = 0 THEN 0
                ELSE ρ^(π_IJ - 1);
        A[5M + J,2M + K + 1 + I]:= IF π_IJ = 0 THEN 0
                ELSE ρ^(π_IJ - 1)
END;
```

The row corresponding to (19c):

```
A[6M + K + 1, M + K + 1]:= -1;
FOR I:= 1 STEP 1 UNTIL M DO
FOR J:= M + 1 STEP 1 UNTIL M + K DO
BEGIN
        A[6M + K + 1,2M + K + 2 - I]:= IF π_IJ = 0
                THEN A[6M + K + 1,2M + K + 2 - I]
        ELSE A[6M + K + 1,2M + K + 2 - I] + ρ^(π_IJ - 1);
        A[6M + K + 1,2M + K + 1 + I]:= IF π_IJ = 0
                THEN A[6M + K + 1,2M + K + 1 + I]
        ELSE A[6M + K + 1,2M + K + 1 + I] + ρ^(π_IJ - 1)
END;
```

Computing the right-hand side of the constraint system is straightforward, although (19b) involves E_J, the maximum load offered to final group J during the calling day, and it is the sum of many terms.

6.2. Numerical Examples

The model was implemented on two test examples, and the resulting bounded variable linear programs were solved at the Carnegie-Mellon University Computation Center on a DEC-20 machine using single-precision arithmetic.

6.2.1. First Example: A Network Based on California Field Data

We apply Program P of Section 4.3.5 to the network given in Eisenberg [5] and Elsner [6], which in turn is based on Gardena, California field data. The hierarchical structure of the network is given in Figure 6 below.

In this network there is only one originating office labelled p_0, and 43 terminating offices, labelled p_1 through p_{43}. This means that there is a demand for traffic associated with each of the 43 calling pairs (p_0,p_i), $i = 1, \ldots, 43$. All other ordered pairs of points are ignored. Every calling pair is also a high usage calling pair. The high usage links are labelled by the integers 1 through 43, and the finals by 44 through 87. Finals 45 through 87 which connect office p_{44}, the tandem switch, with each terminating office, are referred to as *tandem completing* groups. The overflow hierarchy is indicated in Figure 6.

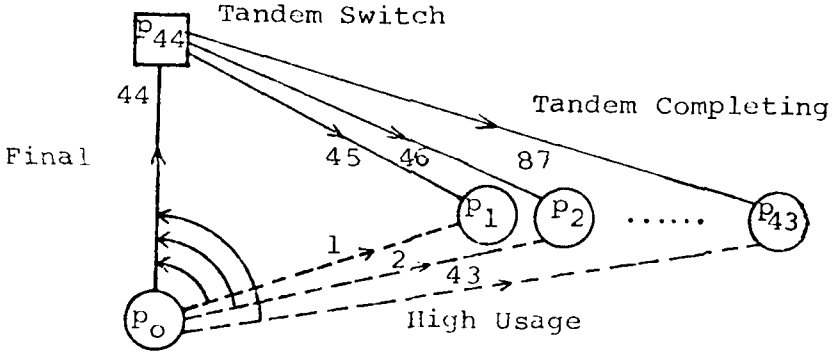


FIGURE 6 A network hierarchy based on Gardena, CA data, Eisenberg [5]

The hierarchy matrix, which has 43 rows (one for each calling pair) and 44 columns (one for each final), consists only of a single column of 1's (for final 44) next to a square (43 × 43) block with 2's along the diagonal, and 0's elsewhere. In fact, this hierarchy is so simple that the matrix itself need not be stored, since several statements written in a computer code can determine the entries of (19b) and (19c).

Base Demand

We assume that the network is constructed *ab initio*, namely all the initial demands between pairs of offices are zero and all initial trunk sizes are zero. According to (7), then, it follows that $E_J = 0$ for $J = 44, \ldots, 87$.

Incremental Demand

Positive incremental demands $\tilde{a}_i^t$ in CCS for each calling pair i and t are given in columns 2 and 3 of Table 3 below. As in [5] and [6], we take a marginal capacity of 30 CCS for all final groups and neglect blocking probability on the final link 44. Similarly, unit costs are $1000 per trunk and $62 switching cost per CCS incurred only at the tandem switch. With these specifications Program P of Section 4.3.5 becomes the following one.

$$\text{Find } M = \min 1000\left[\sum_{t=1}^{87} \tilde{x}_t\right] + 62\tilde{S}$$

subject to

$$\sum_{t=1}^{43} \tilde{a}_t^i B(\tilde{x}_t, \tilde{a}^i) \leqslant 30\tilde{x}_{44} \text{ for } i = 1, 2$$

$$\tilde{a}_t^i B(\tilde{x}_t, \tilde{a}^i) \leqslant 30\tilde{x}_{44+i} \text{ for } i = 1, \ldots, 43 \text{ for } t = 1, 2$$

and

$$\sum_{t=1}^{43} \tilde{a}_t^i B(\tilde{x}_t, \tilde{a}^i) \leqslant \tilde{S} \text{ for } t = 1, 2$$

where the $\tilde{x}_t$ are all nonnegative integers.

The above nonlinear integer program was approximated by the linear program derived by the methods of Section 5.2, which was then solved using suitable bounded variable reductions based on Section 5.3. The bounds of the high usage group sizes were chosen by our prior knowledge of Eisenberg's [5] and Elsner's [6] solutions. An optimal linear programming solution so obtained is termed the *incremented network*. Table 3 presents an incremented network and includes the overflows from the high usage trunk groups to the final trunk group 44.

Table 4 compares the sizes of the high usage trunk groups occurring in our incremented network with those computed in Eisenberg [5] and those computed in Elsner [6]. Finally, Table 5 gives some overall comparisons among the three solutions.

Remarks on Tables 3, 4, and 5

In Tables 3 and 4 each linear programming-determined high usage group size x, except #43, satisfies either (a), $\langle x \rangle - x < 10^{-6}$ or (b), $x - [x] < 10^{-6}$, and hence an integer is reported. High usage group #43 is truncated to 3 decimal places as are all overflows, the final group size, and tandem completing group sizes.

Eisenberg's multihour noninteger solution is not given in [5], and consequently the costs in Table 5 may be higher than for the noninteger solution.

Elsner's descent algorithm obtains a solution with a lower total cost than an integerized solution. The use of an approximation to the Erlang B-function (2) applicable to noninteger high usage trunk group sizes may account for this difference.

6.2.2. The Second Example: Figure 1's Network Hierarchy

We solve Program LP' of Section 5.2 applied to the network hierarchy of Figure 1 of Section 1 with the following specification of input data.

Base Demand

Traffic demand is assigned to all 56 pairs of points of Figure 1 by daytime, evening, and nighttime according to three basic kinds of pairs:

TABLE 3—*Specification of Incremented Offered Load Demands for Example 1 and an Optimal Linear Programming Solution with all Overflows from High Usage Groups*

Trunk Group	Offered Loads (CCS)		Overflow (CCS)		High Usage Trunks	Tandem-Completing Trunks
	Hour 1	Hour 2	Hour 1	Hour 2		
1	60	140	3.746	41.978	4	1.399
2	119	9	16.271	0.000	5	0.542
3	82	20	10.260	0.045	4	0.342
4	305	76	20.002	0.000	12	0.666
5	30	0	13.636	0.000	1	0.454
6	59	7	9.179	0.007	3	0.305
7	102	56	9.795	0.901	5	0.326
8	256	161	21.305	1.632	10	0.710
9	366	230	22.406	0.838	14	0.746
10	469	310	20.256	0.598	18	0.675
11	115	115	14.595	14.595	5	0.486
12	144	34	16.871	0.013	6	0.562
13	206	335	3.691	44.757	11	1.491
14	310	650	0.270	89.490	19	2.983
15	284	319	13.718	24.987	12	0.832
16	93	152	7.072	33.258	5	1.108
17	17	24	5.452	9.599	1	0.319
18	74	325	0.017	73.351	9	2.445
19	102	158	4.424	23.041	6	0.768
20	137	322	1.414	71.323	9	2.377
21	222	247	10.744	18.096	10	0.603
22	252	390	3.621	43.919	13	1.463
23	445	194	21.335	0.006	17	0.711
24	176	86	19.991	0.697	7	0.666
25	83	29	10.640	0.227	4	0.354
26	98	21	17.146	0.056	4	0.571
27	158	74	13.236	0.291	7	0.441
28	124	36	18.491	0.110	5	0.616
29	54	25	7.253	0.700	3	0.241
30	38	1	8.102	0.000	2	0.270
31	31	17	5.149	1.196	2	0.171
32	140	46	15.286	0.077	6	0.509
33	96	30	16.195	0.262	4	0.539
34	122	62	17.587	1.410	5	0.586
35	163	57	14.962	0.057	7	0.498
36	163	72	14.962	0.247	7	0.498
37	296	238	17.134	4.745	12	0.571
38	33	28	5.933	4.071	2	0.197
39	240	3	15.806	0.000	10	0.526
40	136	7	13.783	0.000	6	0.459
41	54	4	7.253	0.000	3	0.241
42	52	35	6.546	2.063	3	0.218
43	206	9	13.108	0.000	8.997	0.436
Totals of Columns	6712	5154	508.643	508.643	306.997	30.921

TABLE 4. *Comparison of Optimal High Usage Trunk Group Sizes Computed by the Multihour Method, A Descent Method, and Linear Programming for the Gardena Network*

Trunk Group	High Usage Group Sizes		
	From Multihour Method [5]	From Descent Method [6]	From Linear Programming
1	4	4.42	4
2	3	5.25	5
3	4	3.78	4
4	6	11.97	12
5	0	1.47	1
6	1	2.81	3
7	4	4.64	5
8	8	10.32	10
9	12	14.10	14
10	18	17.57	18
11	5	5.37	5
12	7	6.20	6
13	10	10.81	11
14	16	18.58	19
15	12	12.14	12
16	5	5.43	5
17	1	1.08	1
18	6	8.92	9
19	5	5.74	6
20	8	9.36	9
21	10	9.78	10
22	12	12.58	13
23	17	16.73	17
24	8	7.41	7
25	4	3.83	4
26	5	4.43	4
27	7	6.75	7
28	6	5.44	5
29	3	2.64	3
30	2	1.86	2
31	2	1.60	2
32	6	6.06	6
33	5	4.35	4
34	6	5.40	5
35	7	6.92	7
36	7	6.93	7
37	12	11.84	12
38	2	1.77	2
39	10	9.70	10
40	6	5.90	6
41	3	2.59	3
42	3	2.61	3
43	9	8.48	8.997
Totals	287	305.56	306.997

TABLE 5. *Comparisons of Total Number of Trunks, Switching Costs, and Total Costs for the Multihour, Descent, and Linear Programming Solutions of the Gardena Network*

Network Characteristics	Multihour [5]	Descent [6]	Linerar Programming
# High Usage Trunks	287	305.56	306.997
# Final Trunks	39	NA*	16.955
# Tandem Compl.	NA	NA	30.921
Switching Cost	$44,640	NA	$31,537
Total Cost	$405,315	$385,500	$386,410

*NA = not available

(1) each of the pairs (A,C) and (C,A) receive 500 CCS during daytime and 0 during the other two periods,

(2) each pair which includes exactly one of the pairs A or C receives 100 CCS during daytime and 0 during the other two periods, *and*

(3) each pair which excludes both points A and C receives 75 CCS during daytime, 200 CCS during evening, and 100 CCS during nighttime.

These choices were imagined upon viewing points A and C as "commercial" points and viewing all other points as "residential." They represent particular choices of the inputs $\bar{a}_j^t$, $j = 1, \ldots, 56$, of Program LP'. Analogous to the first example we assume that the cost per trunk is $1000, that the switching cost is $62 per CCS, the quality of service is 0.99, and that the marginal capacity of a trunk in a final group is 30 CCS. However, we did not neglect blocking on the final links. Using these inputs and the hierarchy of Figure 1, an optimal solution to LP' was obtained termed the *base network*.

Incremented Demand

Assume that an increase in demand of 20% occurs uniformly among all of the 56 calling pairs. With all other inputs to LP' remaining unchanged an optimal solution was obtained, termed (as before) the *incremented network*.

Moreover, Program LP' was solved under three additional restrictions on the time t, namely, all high usage links be sized according to: (a) daytime loads, (b) evening loads, and (c) nighttime loads, respectively. These restricted solutions result from the requirement that the network be "engineered" according to a fixed single hour, respectively. This is in contrast to the multihour solutions of the base and incremented networks, and provides a test of reasonableness of the multihour solutions.

For purposes of computer usage, the size of LP' was reduced by the bounded variable restrictions of Proposition 2 of Section 5.3. For example, setting the V_l bounds in (19e) at 25 for each high usage group yields a 64 variable linear program with 1240 constraints. This program was solved by solving a finite sequence of much smaller bounded variable programs (64 variables, 184 constraints). The results are given in Table 6 below.

TABLE 6. *Computer Results of Four Solutions of Program* LP', *Section 5.2: Base and Incremented Networks, and Network Single Hour Designs. Base Demand Incremented 20% Uniformly, $1000 Cost/Trunk, $62 Switching Cost/CCS, and 0.99 Quality of Service*

Final Links and Integer Index		Base Network	Incremented Network	Single Hour Designs		
				Daytime	Evening	Nighttime
AB	9	29.830	35.005	67.925	43.703	55.414
AC	10	87.149	104.643	104.314	111.299	111.400
AD	11	52.760	63.467	69.622	63.339	67.084
BE	12	14.224	17.041	39.273	21.936	31.838
CF	13	40.049	48.000	59.514	47.871	55.343
CG	14	40.049	48.000	59.514	47.871	55.343
CH	15	52.824	63.416	68.775	63.416	67.143
High Usage Links and Integer Index						
AE	1	1	2	11	0	0
BF	2	17	20	9	21	12
BG	3	17	20	9	21	12
CE	4	7	8	12	0	0
DF	5	18	20	9	21	12
EF	6	18	21	10	21	12
EG	7	18	21	10	21	12
EH	8	18	21	10	21	12
Total Switched Traffic (CCS)		8492.02	10173.11	13764.04	10138.47	12491.15
Total Cost (000)		$957.4	$1143.3	$1402.3	$1154.0	$1290.0

Observe that the multihour (incremented network) solution has a total cost which is less than each of the single hour design total costs, although the single evening hour solution is only 94% larger than the multihour solution. Apparently, the opportunity of engineering final groups *AB* and *AC* at another time, namely daytime, permits a slight saving in total cost.

7. CONCLUSIONS

In this paper it is recommended that linear programming be used to solve for changes in trunk group and switching equipment requirements necessary to provide for altered demands for telecommunications services and altered demands for service qualities. Obtaining solutions to this basic problem is a major goal of our supply model which seeks to minimize total incremental investments in both trunking and switching subject to these constraints.

The linear programming model distinguishes high usage trunk groups from final trunk groups according to the role each plays in the network hierarchy. The important subset of high usage group variables may be solved for by linear programming, and, in general, the costs due to straightforward integer rounding of these groups tend to be offsetting and integer round-off procedures easily maintain overall network quality of service.

Caution must be exercised, however, in the selection of the sizes of the final trunk groups because of the use of the marginal capacity assumption in the linear programming model. In practice, the actual values of the final group variables can be determine by methods which do not depend on the marginal capacity assumption, principally Wilkinson's Equivalent Random Method [2], [20]. This method is needed because of the various peakedness effects that occur in the probability distributions of alternately routed traffic, see also Deschamps [4].

The question of whether the linear program P' provides optimal solutions giving integral numbers of high usage trunk group sizes is still an open one. A related class of nonlinear integer programs which are solvable as linear programs is treated in Meyer [14], where various unimodularity assumptions are made. These assumptions do not apply in general to the class of network problems treated in this paper. The results of our linear programming experiments on two simple networks in the field may stimulate research on this question.

We shall leave the linear programming duality developments for a later paper. It appears that sensitivity and postoptimality analyses will be indeed useful for network design synthesis. Fortunately, by Proposition 1 and its corollaries it appears that a much smaller list of active dual variables will be required than the total number of constraints in program LP'.

Future work should also incorporate more than one alternate route in the network hierarchy, even though for many networks in the field the first and second choice routes are preeminent. Many networks given in the literature are included within the linear programming models of this paper. Large scale network optimizations made available through the modeling approach of this paper should enhance an effective integration of the supply model with a disaggregated econometric demand model for telecommunications services.

We conclude with an observation shared by Edward A. Silver and Stephen A. Smith, expressed in personal correspondence, that there is an interesting equivalence between telephone engineering and replenishment inventory systems, see [16] and [17]. Perhaps the design of more complex telecommunications network hierarchies may have application to the design of more complex replenishment inventory systems.

ACKNOWLEDGMENTS

We are grateful for helpful comments from a referee on earlier versions of this manuscript. We also wish to thank Dr. Ilker Baybars, School of Urban and Public Affairs, Carnegie-Mellon University, for obtaining an alternate characterization of the network hierarchy involving only graph theoretic terms. The research of the authors was partially supported by the National Science Foundation Grants NSF ENG76-05191 and ENG78-25488. In addition, G. G. Polak was supported by Research Grant DAAG29-77-G-0024, U.S. Army Research Office. The paper is a revision of earlier reports of December, 1977 and February, 1979 and has benefited greatly from earlier discussions with members of the Service, Rates, and Costs Department of the Long Lines Company of *AT&T*. Any errors or misinterpretations in the paper, however, remain the sole responsibility of the authors.

Appendix

PROOFS OF PROPOSITION 1, THREE COROLLARIES OF PROPOSITION 1, AND PROPOSITION 2

PROPOSITION 1. Let K be defined as,

$$K = \{(z,x) \in \mathbb{R}^2 \mid x \geq 0 \text{ and } z \geq \hat{B}(x,a)\}.$$

Let L be the set of all (z,x) in $\mathbb{R}^2$ which satisfy the semi-infinite system of linear inequalities

$$(1) \qquad z - 1 \geq \sum_{r=0}^{n} (c_r - c_{r-1})(x - r) \text{ and } x \geq 0$$

for $n = 0, 1, \ldots$.

Then $K = L$, and K is nonempty.

PROOF. Nonemptiness of K is most easily seen by observing that $(1,0) \in K$ since $\hat{B}(0,a) = B(0,a) = 1$.

Let (z,x) be an arbitrary point in K. Assume throughout that $\bar{n} = [\bar{x}]$, the integer part of $\bar{x}$. Applying (16) of Section 5.1 gives

$$z \geq D(n,a)x + (n+1)B(n,a) - \bar{n}B(\bar{n}+1,a),$$

and hence from (17) we have

$$(2) \qquad z - 1 \geq \sum_{r=0}^{n} (c_r - c_{r-1})(x - r).$$

Thus, (z,x) satisfies the particular inequality of (1) indexed by the nonnegative integer $\bar{n}$.

Consider now any integer $n, n \geq \bar{n} + 1$ and write

$$\hat{B}(\bar{x},a) - 1 + \Delta_1'' = \sum_{r=0}^{n} (c_r - c_{r-1})(\bar{x} - r)$$

where $\Delta_1'' = \sum_{r=\bar{n}+1}^{n} (c_r - c_{r-1})(\bar{x} - r)$. Now for any integer r, $\bar{n} + 1 \leq r \leq n$, it follows that $\bar{x} - r < 0$ because $\bar{n} \leq \bar{x} < \bar{n} + 1 \leq r$. In addition, $c_r - c_{r-1} > 0$ for each nonnegative integer r, and therefore $\Delta_1'' < 0$ for each integer n, $n \geq \bar{n} + 1$. Hence,

$$(3) \qquad \bar{z} - 1 \geq \hat{B}(\bar{x},a) - 1 > \hat{B}(\bar{x},a) - 1 + \Delta_1'' = \sum_{r=0}^{n} (c_r - c_{r-1})(\bar{x} - r),$$

for each integer n, $n \geq \bar{n} + 1$.

(2) and (3) together show that $(\bar{z},\bar{x})$ satisfies all those inequalities of (1) indexed by n, $n \geq n$. We now check that $(\bar{z},\bar{x})$ also satisfies those inequalities indexed by nonnegative integers n, $n \leq n - 1$.

If $\bar{n} = 0$, there is nothing to check for there are no such n. For $\bar{n} \geq 1$, let n satisfy $0 \leq n \leq \bar{n} - 1$ and write

$$\hat{B}(\bar{x},a) - 1 = \sum_{r=0}^{n} (c_r - c_{r-1})(\bar{x} - r) + \Delta_2''$$

where $\Delta y^n = \sum_{r=n+1}^{x} (c_r - c_{r-1})(x-r)$. For each integer r, $n+1 \leqslant r \leqslant n$, it follows that $x - r \geqslant 0$ and $c_r - c_{r-1} > 0$ as before. Hence, $\Delta y^n \geqslant 0$, and hence,

$$z - 1 \geqslant \hat{B}(x,a) - 1 \geqslant \sum_{r=0}^{n} (c_r - c_{r-1})(x-r) \tag{4}$$

for each integer n, $0 \leqslant n \leqslant \bar{n} - 1$. The latter finite system of inequalities (10) together with (2) and (3) show that $(z,\bar{x})$ satisfies (1), implying $K \subseteq L$ and in particular L is nonempty.

The other inclusion $L \subseteq K$ is trivial because any (z,x) in L satisfies in particular

$$z - 1 \geqslant \sum_{r=0}^{n} (c_r - c_{r-1})(\bar{x} - r).$$

Using (16) and (17) again shows $z \geqslant \hat{B}(\bar{x},a)$, i.e., $(\bar{z},\bar{x}) \in K$.

COROLLARY 1. Let $\bar{x}$ be nonnegative real. Then $(\hat{B}(\bar{x},a), \bar{x})$ satisfies each inequality of (1) strictly except for (i), the inequality indexed by $[\bar{x}]$, which it satisfies as an equality, and (ii) possibly the inequality indexed by $[\bar{x}] - 1$ when $[\bar{x}] \geqslant 1$. The inequality $[\bar{x}] - 1$ is satisfied as an equality if and only if $\bar{x}$ is a positive integer.

PROOF: Let $z = \hat{B}(x,a)$. Application of (3) shows that $(\bar{z},\bar{x})$ satisfies each inequality indexed by n, $n \geqslant \bar{n} + 1$, strictly, where $\bar{n} = [\bar{x}]$. By (16) and (17) of Section 5.1, it follows that (z,x) satisfies the inequality determined by $\bar{n}$ as an equality.

It only remains to prove that the inequalities indexed by nonnegative integers n, $n \leqslant \bar{n} - 2$ are satisfied strictly. There is nothing to check if $\bar{n} \leqslant 1$. For $\bar{n} \geqslant 2$, let n be any integer $0 \leqslant n \leqslant \bar{n} - 2$. Then

$$\hat{B}(\bar{x},a) - 1 = \sum_{r=0}^{n} (c_r - c_{r-1})(\bar{x} - r) + [A + (c_n - c_{n-1})(\bar{x} - \bar{n})]$$

where $A = \sum_{r=n+1}^{\bar{n}-1} (c_r - c_{r-1})(\bar{x} - r)$. Since $\bar{n} \leqslant \bar{x} \leqslant \bar{n} + 1$, it follows that $(c_n - c_{n-1})(\bar{x} - \bar{n}) \geqslant 0$ and $A > 0$. Hence,

$$\hat{B}(x,a) - 1 > \sum_{r=0}^{n} (c_r - c_{r-1})(\bar{x} - r)$$

for each integer n, $0 \leqslant n \leqslant \bar{n} - 2$.

The last assertion follows from examining

$$\hat{B}(\bar{x},a) - 1 = \sum_{r=0}^{n-1} (c_r - c_{r-1})(\bar{x} - r) + (c_n - c_{n-1})(\bar{x} - \bar{n})$$

where $\bar{n} = [\bar{x}] \geqslant 1$, for the inequality indexed by $\bar{n} - 1$ is satisfied as an equality if and only if $\bar{x} - n = 0$.

It will be useful later to include upper bounds on the x-variables in the set K. The following corollary states that in this case one only needs a finite number of the inequalities of (1).

COROLLARY 2. Let U be a positive integer and set $K' = K \cap \{(z,x) | 0 \leqslant x \leqslant U\}$. Let L' be the set of all (z,x) which satisfy

$$z \geq 1 \geq \sum_{i=0}^{x} (c_i - c_{i-1})(x - i), \quad x \geq 0$$

for $x = 0, 1, \ldots, u - 1$. Then $K = I$.

PROOF. Let $I = \{(z,x) : 0 \leq x \leq u - 1\}$. Then by Proposition 1, $K = I$. Since I incorporates the semi-infinite system (1), it follows immediately that $I = I$. On the other hand, let (z,x) be arbitrary in I. Then, $0 \leq x \leq u - 1$ and $x \leq u - 1$, where $u = \{c_i : B(i,a) < 1\}$. But membership in I implies

$$z \geq 1 \geq \sum_{i=0}^{x} (c_i - c_{i-1})(x - i).$$

Using (17) followed by (16) we find that $z \geq B(x,a)$ implying $(z,x) \in K$.

On the other hand, if $x = u$, then necessarily $x = U$. Moreover,

$$z \geq 1 \geq \sum_{i=0}^{U-1} (c_i - c_{i-1})(U - i).$$

But the right hand sum equals by (17),

$$1 - D(U - 1, a)[1 + (B(U - 1, a) - (U - 1)B(U, a)],$$

which is merely $1 - B(U,a)$. Therefore, in this case

$$z \geq B(U,a)$$

and $(z,x) \in K$ also. Thus, in either case $(z,x) \in K$ which *implies* (z,x) satisfies the entire inequality system (1) by Proposition 1. Hence, $(z,x) \in I$, and hence $I \subseteq I$. Therefore, $I = I$ which yields $K = I$.

COROLLARY 3. (z,x) is an extreme point of K if and only if x is a nonnegative integer and $z = B(x,a)$.

PROOF. There are only two variables z and x in the linear inequality system (1). Hence extreme points can only occur on the boundary of K at the intersection of a pair of linearly independent equations. By Corollary 1 a pair of linearly independent equations arise if and only if x is a nonnegative integer, and moreover, each nonnegative integer does satisfy two (adjacent) linearly independent equations. This includes the special cases of the endpoints where for $x = 0$, the additional inequality $x \geq 0$ is used and at $(B(U,a), U)$ the inequality $x \leq U$ is used.

PROPOSITION 2. Under the bounded variable assumptions made in Section 3.3

(i) any optimal solution $\{(\hat{x}_i^*), S^*, (z_i'^*)\}$ of LP'_{BD} is feasible for LP', and

(ii) if for each high usage group I

$$\alpha_I \leq x^* \leq \beta_I \tag{21}$$

then this optimal solution is also optimal for program LP'. Moreover, there exist α_I, β_I and an optimal solution of LP' such that with respect to $\hat{x}_i^*$ of that solution, (21) holds.

PROOF. If (19d) is strictly satisfied for any i and high usage calling pair i then z_i may be decreased to its lower bound without affecting feasibility. Thus $\{(\hat{x}_i^*), S^*, (z_i'^*)\}$ optimal for LP'_{BD} $\ldots$ $z_i = B(x_i, a_i + a_{i'})$ for each high usage calling pair i. By Corollary 1 for each

high usage calling pair i, $(z_i', \tilde{x}_i^*)$ satisfies (20) for every nonnegative integer z_i'. Since $z_i' \leq z_i^*$ and (19b) and (19c) are already satisfied, it follows that $\{(\tilde{x}_i^*), \tilde{S}^*, (z_i^*)\}$ satisfies all the constraints of LP'. This proves (i).

The first part of (ii) follows from linear programming duality theory. Because of (21) the two dual variables stemming respectively from the two bounding constraints on $\tilde{x}_i$ are both zero. Hence, one may delete these constraints in $LP_{(i)}$ and the same dual optimal solution prevails. Therefore, by duality $\{(\tilde{x}_i^*), \tilde{S}^*, (z_i^*)\}$ is optimal for the relaxed-variable constrained program LP'. The remaining statement of part (ii) follows from Corollary 1 and the fact that the nonnegative integers α_i, β_i satisfy $\beta_i - 1 - \alpha_i \geq 2$.

REFERENCES

[1] Charnes, A. and W.W. Cooper, *Management Models and Industrial Applications of Linear Programming*, Vols. I, II, (J. Wiley and Sons, Inc., New York 1961).

[2] Cooper, R.B., *Introduction to Queuing Theory*, (The Macmillan Company, New York 1972).

[3] Dantzig, G.B., *Linear Programming and Extensions*, (Princeton University Press, Princeton, NJ 1963).

[4] Deschamps, P.J., "Analytic Approximation of Blocking Probabilities in Circuit Switched Communication Networks," IEEE Transactions on Communications *COM-27*, 603-606 (1979).

[5] Eisenberg, M., "Engineering Traffic Networks for More Than One Busy Hour," Bell System Technical Journal *56*, 1-20 (1977).

[6] Elsner, W.B., "A Descent Algorithm for the Multihour Sizing of Traffic Networks," Bell System Technical Journal *56*, 1405-1430 (1977).

[7] Gnedenko, B.V. and I.N. Kovalenko, *Introduction to Queuing Theory*, translated by Eng. R. Condor, edited by D. Louvish, (Israel Program for Scientific Translations, Ltd., Jerusalem, 1968).

[8] Gomory, R.E. and T.C. Hu, "Multi-Terminal Network Flows," Journal of the Society for Industrial and Applied Mathematics *9*, 551-570 (1961).

[9] Gomory, R.E. and T.C. Hu, "An Application of Generalized Linear Programming to Network Flows," Journal of the Society for Industrial and Applied Mathematics *10*, 260-283 (1962).

[10] Gomory, R.E. and T.C. Hu, "Synthesis of a Communication Network," Journal of the Society for Indastrial and Applied Mathematics *12*, 348-369 (1964).

[11] Kalaba, R.E. and M.L. Juncosa, "Optimal Design and Utilization of Communications Networks," Management Science *3*, 33-44 (1956).

[12] Katz, S., "Statistical Performance Analysis of a Switched Communications Network," *Proceedings of the 5th International Teletraffic Congress*, 566-575 (1967).

[13] Messerli, E.J., "Proof of a Convexity Property of the Erlang B-Formula," Bell System Technical Journal *51*, 951-953 (1972).

[14] Meyer, R.R., "A Class of Nonlinear Integer Programs Solvable by a Single Linear Program," SIAM Journal of Control and Optimization *15*, 935-946 (1977).

[15] Rapp, Y.A., "Planning of Junction Networks with Non-Coincident Busy-Hours," Ericsson Technics *27*, 3-23 (1971).

[16] Silver, E.A. and S.A. Smith, "A Graphical Aid for Determining Optimal Inventories in a Unit Replenishment Inventory System," *Management Science 24*, 388-389 (1977).

[17] Smith, S.A., "Optimal Inventories for an (S-1, S) System with No Backorders," *Management Science 23*, 522-528 (1977).

[18] Syski, R., *Introduction to Congestion Theory in Telephone Systems*, (Oliver and Boyd, Ltd., London, 1960).

[19] Truitt, C.J., "Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks," Bell System Technical Journal *33*, 277-302 (1954).

[20] Wilkinson, R.I., "Theories for Toll Traffic Engineering in the U.S.A.," Bell System Technical Journal *35*, 421-514 (1956).

PREVENTIVE MAINTENANCE AND REPLACEMENT UNDER ADDITIVE DAMAGE

S. D. Chikte

Department of Electrical Engineering
University of Rochester
Rochester, New York

S. D. Deshmukh

Department of Managerial Economics and Decision Sciences
J.L. Kellogg Graduate School of Management
Northwestern University
Evanston, Illinois

ABSTRACT

A system deteriorates due to shocks received at random times, each shock causing a random amount of damage which accumulates over time and may result in a system failure. Replacement of a failed system is mandatory, while an operable one may also be replaced. In addition, the shock process causing system deterioration may be controlled by continuous preventive maintenance expenditures. The joint problem of optimal maintenance and replacement is analyzed and it is shown that, under reasonable conditions, optimal maintenance rate is decreasing in the cumulative damage level and that beyond a certain critical level the system should be replaced. Meaningful bounds are established on the optimal policies and an illustrative example is provided.

1. INTRODUCTION

In this section, we first introduce the reliability problem treated in this paper, provide a background in terms of the relevant literature, and summarize our assumptions and results.

A. Problem Statement

Consider a system that receives shocks at random points in time, each shock causing a random amount of damage which accumulates over time. As the cumulative damage level increases, the rate at which the system generates revenue declines and the probability of its failure increases. Replacement of a failed system is considered mandatory, while an operable one may also be replaced, a forced replacement being costlier than a voluntary one. In addition to replacement, the damage process can also be influenced by preventive maintenance expenditures; higher expenditure rates buffer the system more effectively, and hence decrease probabilistically the frequency of occurrence of shocks as well as their severity. Our problem is to determine an optimal policy that specifies a sequence of replacement and maintenance expenditure schedules so as to maximize the expected discounted net profit generated by the system over an infinite planning horizon.

B. Background Literature

The extensive literature on control of stochastically deteriorating systems has been surveyed by McCall [10] and Pierskalla and Voelker [11]. Of particular relevance here is the work by Taylor [15], Feldman [6,7] and Abdel-Hameed and Shimi [1] on optimal replacement of a system that is subject to shocks and failure, based on the theory of optimal stopping in Markov processes. On the other hand, Thompson [16] and Kamien and Schwartz [8] employ optimal control theory to characterize the time pattern of optimal maintenance expenditures that retard the system failure rate. In the former class of models the only decisions available are whether to replace the system or not, while in the latter class the state of the system at any time is described as being either working or failed. Our model incorporates the essential features of both of these two classes in that it allows for varying degrees of preventive maintenance (in addition to the replacement action) as well as a more detailed description of system deterioration (in addition to its description as working or failed). Our analysis is based on the methodology of stochastic dynamic programming, as in Derman [5], Ross [13] and others. Some preliminary work along these lines may be found in Chikte [3] and Chikte and Kozin [4].

C. Overview of Assumptions and Results

In Section 2, we define the state of the system in terms of its cumulative damage level, which increases randomly due to occurrence of shocks and is influenced by continuous maintenance expenditure and instantaneous replacement actions. The probabilistic rate at which damage accumulates is assumed to be decreasing in the maintenance expenditure rate (Assumption $P(i)$). Upon receiving a shock, the system may fail instantaneously with a probability that is assumed to be increasing in the resulting damage level but at a diminishing rate (Assumption $P(ii)$). If the system does fail, it must be replaced instantaneously by a new one at a fixed cost, while, even if it does not fail, it may still be replaced voluntarily at a lower cost (Assumption $I(ii)$). An operating system continuously generates revenue at a rate which decreases, but at a diminishing rate, as the cumulative damage level builds up (Assumption $E(i)$). Finally, we also introduce a condition (Assumption R) which ensures a profitable system operation (as in Theorem 4).

In Section 3, we first show that the maximum infinite horizon expected discounted net profit from system operation decreases at a diminishing rate as the cumulative damage level increases (Theorem 1). We then show that it is optimal to replace the system voluntarily as soon as its cumulative damage level exceeds a critical threshold (Theorem 2). As to the optimal preventive maintenance policy, we show that the maintenance expenditure rate should be reduced as the cumulative damage level builds up to the critical value (Theorem 3). Finally, we derive (in Theorem 5) meaningful bounds on the optimal policy. In particular, we show that postponement of voluntary replacement cannot be optimal if the extra profit from system operation until the next shock cannot justify the extra cost due to a possible failure and replacement at that shock. We also show that the optimal maintenance expenditure rate is always strictly less than the rate at which the system currently generates revenue.

Section 4 provides an example that illustrates the model and the results and Section 5 concludes the paper with some remarks on the type of information required for implementation.

2. MODEL FORMULATION

In this section, we first establish the notation and define the basic components of our model, then we present the assumptions made and finally we describe the overall model dynamics.

A. Notation and Definitions

Let a nonnegative random variable X_t denote the cumulative damage level of the system in operation at time $t \geq 0$; it is the sum total of the damages suffered due to shocks received by the system by time t.

The damage process affecting X_t is controlled by means of a continuous preventive maintenance expenditure rate $m \in [0, M]$, where (the budget) $M < \infty$. Maintenance is aimed at protecting the system from the undesirable environment so as to retard the rate at which shocks are received and to dampen the magnitudes of damages inflicted by them. Let $\lambda(m) > 0$ be the probabilistic rate at which shocks occur if the maintenance rate is m. Thus, if the maintenance rate is a constant m through time, the time interval between successive shocks is exponentially distributed with the mean $1/\lambda(m)$. Let a nonnegative random variable Y denote the magnitude of damage caused by a shock and let $G(\cdot|m)$ be the cumulative distribution function of Y, parametrized by the maintenance rate m. Thus, $\lambda(m)[1 - G(y|m)]$ is the probabilistic rate at which shocks causing damage in excess of y occur if the maintenance expenditure rate is m.

If $X_{t-} = x$ is the damage level just prior to time t and if the system receives a shock at time t, causing an additional damage of magnitude y (so that $X_t = z \equiv x + y$), then the system may fail instantaneously with a probability denoted by $p(z)$, depending on the new cumulative damage level z, while with probability $[1 - p(z)]$ it endures the shock and continues to operate. If the system fails at time t, it must be replaced immediately by a new one at the forced replacement cost $C_2 > 0$. Even if the system survives the shock, it may still be replaced instantaneously at a voluntary replacement cost $C_1 > 0$. In either case, the replacement decision at time t will be denoted as $d_t = 1$, while $d_t = 0$ corresponds to the nonreplacement decision.

If $X_t = x \geq 0$, let $r(x) \geq 0$ denote the instantaneous rate at which the system generates revenue from its operation. Suppose that future revenues and costs are discounted continuously at rate $\alpha > 0$, so that $e^{-\alpha t}$ is the present value of one dollar earned t time units from now.

By a (replacement and maintenance) policy δ we mean a pair (δ_1, δ_2) of functions of the system state x, denoted as $\delta_1: [0, \infty) \to \{0, 1\}$ and $\delta_2: [0, \infty) \to [0, M]$. Here, the replacement rule δ_1 specifies replacement of the system in state x (which is mandatory if the system is down) if $\delta_1(x) = 1$, while $\delta_1(x) = 0$ specifies the nonreplacement decision. Similarly, if the system is in state x, the maintenance rule δ_2 specifies a maintenance expenditure rate $\delta_2(x) \in [0, M]$. In light of the results by Stone [14] and Pliska [12] on controlled jump processes, it is reasonable to stipulate that δ revises the replacement and maintenance decisions only at shock times, depending on the state of the system then.

Finally, let $V_\delta(x)$ denote the net expected discounted return from employing the policy δ over an infinite planning horizon, starting with a system in state x at time 0. Let $V(x) = \sup_\delta V_\delta(x)$ be the maximum possible return obtainable. A policy δ^* is said to be optimal if $V_{\delta^*}(x) = V(x)$ for all $x \geq 0$. In order to characterize the optimal return function V and the optimal policy δ^*, we need to make certain assumptions on the model parameters.

B. Assumptions

Regarding the effectiveness of preventive maintenance expenditures in dampening the shock process, we assume that higher expenditure rates m protect the system better and as such

result in lower probabilistic rates $\lambda(m)[1 - G(y|m)]$ at which additional damages in excess of any given quantity y occur. (This is analogous to the stochastic monotonicity assumption, as, for example, in Derman [5].) As to the system failure process, it is reasonable to suppose that the system may fail only at shock times and that the probability $p(z)$ of its failure increases in the resulting cumulative damage level z but only at a decreasing rate. We state these probabilistic assumptions as

Assumption P

(i) For any fixed $y \geq 0$, $\lambda(m)[1 - G(y|m)]$ is continuous and nonincreasing in $m \in [0, M]$. In particular, taking $y = 0$, $\lambda(m)$ is continuous and nonincreasing in $m \in [0, M]$.

(ii) The failure probability $p(z)$ is nondecreasing and concave in the cumulative damage level $z \geq 0$*.

With respect to the economics of the system operation, we assume that an operating system in state $x \geq 0$ generates revenue at rate $r(x)$ which is nonincreasing and convex in the cumulative damage level. This reflects a degradation in the system performance as the damage accumulates but at a diminishing marginal rate. On the replacement cost side, we assume that the cost C_2 of replacing a failed system is higher than the cost C_1 of a voluntary replacement of a working system (possibly due to the salvage value differential), thereby providing an incentive to replace the system before failure. Also, to make this system operation and replacement a worthwhile undertaking, it is essential that the cost C_1 of a voluntary replacement be compensated for by the present value $r(0)/\alpha$ of the infinite horizon revenue that a system maintained in mint condition would generate. We summarize these economic conditions as

Assumption E.

(i) The revenue rate $r(x)$ is nonnegative, bounded, nonincreasing and convex in the damage level $x \geq 0$ and $r(0)/\alpha > C_1$.

(ii) The replacement costs C_1 and C_2 satisfy $C_2 > C_1 > 0$.

The above assumptions, *P* and *E*, will be used to characterize properties of the optimal value function V and the maintenance and replacement rules (in Theorems 1,2 and 3), while to show that V is positive (in Theorem 4) and to provide bounds on optimal policies (in Theorem 5) we impose the following simple and easily verifiable condition on the problem parameters, which ensures that the overall operation of the system is a profitable one.

Assumption R. There exists an $m^* \in [0, M]$ such that $m^* \leq \lim_{x \to \infty} r(x)$ and $[r(0) - m^*]/[\alpha + \lambda(m^*)] \geq C_2$.

It says that the net expected discounted profit generated by a new system that is maintained at a small enough expenditure rate m^* until the next shock makes up for the cost of a failure replacement that might be necessary. Generally speaking, if the revenues generated by operating the system are "high" enough in relation to the replacement costs, if the shock

*It is possible to relax Assumption $P(ii)$ by requiring that $p(\cdot)$ be concave only on the region of values of z on which $p(z)$ is not equal to one.

process is sufficiently "slow" and "mild" and if the failure probability is "small" enough, then it is possible to make the system operation a profitable one; Assumption R constitutes one particular set of such conditions.

C. The Model Dynamics

Under a policy $\delta = (\delta_1, \delta_2)$, the cumulative damage process $\{X_t; t \geq 0\}$ evolves as a nonterminating pure jump process as follows. If the state of the system in operation at time t is $X_t = x$ and if the replacement rule specifies $\delta_1(x) = 1$ then the system is renewed instantaneously, yielding $X_{t+} = 0$ at a voluntary replacement cost C_1, while $\delta_1(x) = 0$ leaves the system state unchanged until the next shock. Given $X_{t+} = x \geq 0$, the maintenance policy specifies a continuous maintenance expenditure rate $\delta_2(x) \in [0, M]$. Then the sojourn time S in state x is exponentially distributed with parameter $\lambda(\delta_2(x))$. During this interval, the system generates revenue at rate $r(x)$ and is maintained at an expenditure rate $\delta_2(x)$ and thus yields

$$(1) \qquad \int_0^\infty \left\{ \int_0^s e^{-\alpha u} [r(x) - m] du \right\} \lambda(m) e^{-\lambda(m)s} ds = [r(x) - \delta_2(x)]/[\alpha + \lambda(\delta_2(x))]$$

as the expected discounted profit until the next shock. Similarly, the net return from the next shock onwards will be discounted by the expected discount factor

$$(2) \qquad \int_0^\infty e^{-\alpha s} \lambda(\delta_2(x))\, e^{-\lambda(\delta_2(x))s} ds = \lambda(\delta_2(x))/[\alpha + \lambda(\delta_2(x))].$$

The next shock causes damage of magnitude y according to the distribution $G(dy|m)$, so that the postshock state is $X_{t+} = x + y$. At that instant the system fails with probability $p(x + y)$, in which case it must be replaced (i.e., $\delta_1(x + y) = 1$) at cost C_2, so that $X_{(t+s)+} = 0$. If the system does not fail, which happens with probability $[1 - p(x + y)]$, and if $\delta_1(x + y) = 1$, the system is replaced at cost C_1 and $X_{(t+s)+} = 0$, while if $\delta_1(x + y) = 0$ then the system continues to operate in state $X_{(t+s)+} = x + y$. In any case, $\delta_2(X_{(t+s)+})$ is the expenditure rate at which the system is maintained until the following shock, and the process repeats.

Our objective is to investigate an optimal decision rule $\delta^* = (\delta_1^*, \delta_2^*)$ which specifies the replacement and maintenance decisions $\delta_1^*(x)$ and $\delta_2^*(x)$ as functions of the cumulative damage level x at each shock instant, so as to yield the maximum expected discounted net return $V_\delta^*(x) = V(x)$ for each $x \geq 0$. In the next section, we analyze this problem in the stochastic dynamic programming framework.

3. OPTIMAL RETURN, REPLACEMENT AND MAINTENANCE

In subsection A below, we first provide the dynamic programming functional equation satisfied by the optimal return function $V(x)$, which is then shown to possess, under Assumptions P and E, certain "nice" properties. In subsection B, we make use of these properties of V to characterize the structure of optimal rules δ_1^* and δ_2^*, while in subsection C, Assumption R is employed to derive interesting bounds on δ_1^* and δ_2^*.

A. The Optimal Return Function

In order to analyze the optimal return $V(x)$, we first define $V_n(x)$ as the maximum expected discounted profit over an *infinite* time horizon, starting with a system in state x and given that exactly n more shocks will eventually occur. This is analogous to the approach in Lippman [8] and enables us to interpret n as the time index, yielding the Bellman dynamic programming recursion in a discrete time format as follows.

For all $n = 1, 2, \ldots$, and $x \geq 0$,

$$V_n(x) = \text{Max}\{U_n(0) - C_1,\ U_n(x)\}, \tag{3}$$

where

$$U_n(x) = \max_{m \in [0, M]} T_m V_{n-1}(x), \tag{4}$$

and the operator T_m is defined by

$$T_m V_{n-1}(x) = \{r(x) - m + \int_0^\infty \{V_{n-1}(x+y)[1 - p(x+y)] + [V_{n-1}(0) - C_2]p(x+y)\}\lambda(m)G(dy|m)\}/[\alpha + \lambda(m)] \tag{5}$$

and

$$V_0(x) = \text{Max}[r(0)/\alpha - C_1,\ r(x)/\alpha]. \tag{6}$$

These equations may be interpreted as follows. If the system in state x facing n more shocks is replaced voluntarily, the net optimal return would be $V_n(0) - C_1$, since n more shocks still remain. On the other hand, maintaining it at rate m yields $[r(x) - m]/[\alpha + \lambda(m)]$ as the expected discounted profit until the next shock, according to (1). If the next shock is of magnitude y (determined according to $G(dy|m)$), the optimal return from then on is $V_{n-1}(x+y)$ provided the system survives the shock (i.e., with probability $[1 - p(x+y)]$) and $V_{n-1}(0) - C_2$, otherwise, discounted by the expected discount factor $\lambda(m)/[\alpha + \lambda(m)]$, as in (2). Finally, with no more threat of future shocks (i.e., $n = 0$), maintenance expenditures are unnecessary and we may or may not replace the system, which will be operated from then onwards without further deterioration.

LEMMA 1: Under Assumptions P and E, for each $n = 0, 1, 2, \ldots$, the functions $V_n(x)$ and $U_n(x)$ are bounded, nonincreasing and convex in $x \geq 0$.

PROOF: Boundedness follows from

$$r(0)/\alpha \geq V_n(x) \geq -\{M + C_2[\alpha + \lambda(M)]\}/\alpha, \tag{7}$$

since $r(0)$ is the highest rate of return obtainable, while, in the worst case, infinitely many shocks occur and each requires a failure replacement in spite of employing the maximum possible maintenance rate. To prove monotonicity and convexity of V_n by induction on n, note from (6) and Assumption E (i) that V_0 has these properties. Suppose that V_{n-1} is nonincreasing and convex. From (3) and Assumption E (ii) we have $V_{n-1}(x+y) \geq V_{n-1}(0) - C_2$. Using this, together with the induction hypothesis and Assumption P, it can be checked that, for each y, the integrand in (5) is nonincreasing and convex in x. This, together with Assumption E (i) yields monotonicity and convexity of $T_m V_{n-1}$ for each m. Since these properties are preserved under the maximization operation, we have U_n and hence V_n nonincreasing and convex.

Q.E.D.

From the definition of V_n, it is easy to see that $V_n \leq V_{n-1}$ for all $n = 1, 2, \ldots$, i.e., permitting more shocks can not improve the total return obtainable. Thus, the sequence of functions $\{V_n,\ n = 0, 1, 2, \ldots\}$ is bounded as in (7) and nonincreasing, so that $V = \lim_{n \to \infty} V_n$ exists and is the maximum net expected discounted return over an infinite horizon, given that an unlimited number of shocks will eventually occur. By standard contraction mapping arguments, V is the unique solution to the following functional equation, which is similar to (3) and (4).

$$V(x) = \text{Max}\{V(0) - C_1,\ U(x)\},\ x \geq 0 \tag{8}$$

where

$$(9) \qquad U(x) = \max_{m \in [0,M]} T_m V(x)$$

and T_m is the operator defined in (5).

Since the properties of U_n and V_n is Lemma 1 are preserved upon taking limits as $n \to \infty$, we have the following.

THEOREM 1: Under Assumptions P and E, the optimal value functions $U(x)$ and $V(x)$ are bounded, nonincreasing and convex in $x \geq 0$.

B. Optimal Replacement and Maintenance Policy

From (8), it is clear that the optimal replacement rule δ_1^* specifies the replacement decision $\delta_1^*(x) = 1$ in state x if and only if $V(x) = V(0) - C_1$. Similarly, the optimal maintenance rule δ_2^* specifies in state x, the smallest expenditure rate $\delta_2^*(x)$ which attains the maximum of $T_m V(x)$ in (9) over $m \in [0,M]$; our continuity and compactness assumptions assure the existence of this maximizer.

We first show that, in our model, the optimal replacement rule δ_1^* has the well known control limit form (as in the models of Derman [5], Ross [13], Taylor [15], Feldman [6,7] and others).

THEOREM 2: Under Assumptions P and E, there exists an $x^* \in [0,\infty]$ such that $\delta_1^*(x) = 1$ if and only if $x \geq x^*$.

PROOF: By monotonicity of U and definition of δ_1^*, we may define

$$(10) \qquad x^* = \inf\{x \geq 0 : V(0) - C_1 \geq U(x)\}.$$

Q.E.D.

Next, we show that the optimal preventive maintenance expenditure rate is nonincreasing in the damage level of the system. This may be viewed as a stochastic analog of the result of Kamien and Schwartz [8] and Thompson [14], wherein the optimal maintenance rate is shown to be decreasing in the chronological age of the system. Indeed, it is reasonable to expect a reduction in continuous maintenance as instantaneous replacement becomes more imminent.

THEOREM 3: Under Assumptions P and E, the optimal maintenance rate $\delta_2^*(x)$ is nonincreasing in $x \in [0,x^*)$, where x^* is given by (10).

PROOF: If $x < x^*$, then from (8) and Theorem 1 we have $V(x) = U(x)$, which can be seen to be equivalent to

$$(11) \qquad \alpha V(x) = \max_{m \in [0,M]} [r(x) - m - f(x,m)]$$

where

$$(12) \qquad f(x,m) = \int_0^\infty \{[V(x) - V(x+y)] + [V(x+y) - V(0) + C_2]\, p(x+y)\}\, \lambda(m) G(dy|m).$$

Take $x_1 \leq x_2 < x^*$, so that we need to show that $\delta_2^*(x_1) \geq \delta_2^*(x_2)$. We first show that $[f(x_2,m) - f(x_1,m)]$ is nondecreasing in $m \in [0,M]$. Now

$$(13) \qquad [f(x_2,m) - f(x_1,m)] = \int_0^\infty g(x_1,x_2,y)\lambda(m)\, G(dy|m)$$

where

$$
\begin{aligned}
(14)\quad g(x_1,x_2,y) &= [V(x_2) - V(x_2+y)] - [V(x_1) - V(x_1+y)] \\
&\quad + [V(x_2+y) - V(0) + C_2]\, p(x_2+y) \\
&\quad - [V(x_1+y) - V(0) + C_2]\, p(x_1+y) \\
&= [V(x_2) - V(x_1)] + [V(x_1+y) - V(x_2+y)]\,[1 - p(x_1+y)] \\
&\quad + [V(x_2+y) - V(0) + C_2]\,[p(x_2+y) - p(x_1+y)].
\end{aligned}
$$

By monotonicity and convexity of V (Theorem 1) and monotonicity of p (Assumption P (ii)), the second term in the above expression is nonincreasing in y. Also monotonicity of V, concavity of p and the inequality $V(x) - V(0) + C_2 \geqslant V(x) - V(0) + C_1 \geqslant 0$ (since $C_2 \geqslant C_1$ and V satisfies (8)) imply that the third term is nonincreasing in y. Thus, $g(x_1,x_2,y)$ is nonincreasing in y. This, coupled with Assumption $P(i)$ now implies that $[f(x_2,m) - f(x_1,m)]$ is nondecreasing in $m \in [0,M]$. Since $\delta_2^*(x)$ attains the maximum on the right hand side of (11), the above implies that $\delta_2^*(x_1) \geqslant \delta_2^*(x_2)$, whenever $x_2 \geqslant x_1$, because otherwise we would have

$$
\begin{aligned}
&[r(x_2) - \delta_2^*(x_2) - f(x_2,\delta_2^*(x_2))] - [r(x_1) - \delta_2^*(x_2) - f(x_1,\delta_2^*(x_2))] \\
&< [r(x_2) - \delta_2^*(x_1) - f(x_2,\delta_2^*(x_1))] - [r(x_1) - \delta_2^*(x_1) - f(x_1,\delta_2^*(x_1))]
\end{aligned}
$$

i.e.

$$
\begin{aligned}
&[r(x_2) - \delta_2^*(x_2) - f(x_2,\delta_2^*(x_2))] + [r(x_1) - \delta_2^*(x_1)\, f(x_1,\delta_2^*(x_1))] \\
&< [r(x_2) - \delta_2^*(x_1) - f(x_2,\delta_2^*(x_1))] + [r(x_1) - \delta_2^*(x_2 - f(x_1,\delta_2^*(x_2))],
\end{aligned}
$$

contradicting optimality of $\delta_2^*(x)$ when in state x.

Q.E.D.

Thus, by Theorems 2 and 3, the higher the state of deterioration of the system the less should be the maintenance effort to prevent further deterioration and, as soon as the deterioration level exceeds a critical value, the system should be replaced by a new one.

C. Bounds on Optimal Policy and Return

So far, with Assumptions P and E, there is no guarantee that even the optimal policy will result in a profitable system operation over the longrun. This is precisely the purpose of Assumption R, as the following Theorem 4, shows and this fact will also be needed to establish bounds on x^* and $\delta_2^*(x)$ in Theorem 5 below.

THEOREM 4: With Assumptions P, E and R, the optimal return $V(x)$ is positive for all $x \geqslant 0$.

PROOF: Consider a policy $\delta = (\delta_1,\delta_2)$, where $\delta_1(x) = 1$ and $\delta_2(x) = m^*$ for all $x \geqslant 0$, where m^* is as in Assumption R; thus δ replaces the system at every shock and always specifies the constant maintenance rate m^*. The total expected discounted return, starting in state x and following this policy δ, is, therefore

$$
\begin{aligned}
V_\delta(x) &= [r(x) - m^*]/[\alpha + \lambda(m^*)] + \lambda(m^*)\,[r(0) - m^*]/[\alpha[\alpha + \lambda(m^*)]] \\
&\quad - \lambda(m^*)[K(x) + \lambda(m^*)K(0)/\alpha]/[\alpha + \lambda(m^*)],
\end{aligned}
$$

where

$$
K(x) = C_1 + (C_2 - C_1) \int_0^x p(x+y)\, G(dy \mid m^*)
$$

is the expected replacement cost upon receiving a shock. Since $C_2 \geq K(x) \geq K(0)$, we have

$$V_\delta(x) \geq [r(x)-m^*]/[\alpha + \lambda(m^*)] + \lambda(m^*)/\alpha\{[r(0)-m^*]/[\alpha + \lambda(m^*)]-C_2\} > 0,$$

by Assumption R. Since $V(x) \geq V_\delta(x)$ for all x, the proof is completed.

Q.E.D.

Our final objective is to derive bounds on the optimal policy δ^*.

THEOREM 5: Under Assumptions P, E and R, we have

(15) $x^* \leq b$,

where $b = \inf B$,

$$B = \{x \geq 0; \max_{m \in [0,M]} \{[r(x)-m]/\lambda(m) - (C_2 - C_1)\int_0^\infty p(x+y)G(dy|m)\} \leq 0\}$$

and

(16) $\delta_2^*(x) < r(x)$, $x \in [0,x^*)$.

PROOF: To prove (15), in view of (8), it suffices to show that $V(x) > U(x)$ whenever $x \in B$. Suppose $x \in B$ and $V(x) = U(x)$. Now

$$\begin{aligned}
U(x) &\leq \max_{m \in [0,M]} \{ r(x)-m + \int_0^\infty \{V(x)[1-p(x+y)] \\
&\quad + [V(0) - C_2]p(x+y)\}\lambda(m)G(dy|m)\}/[\alpha + \lambda(m)]\} \\
&\leq V(x) \max_{m \in [0,M]} \{\lambda(m)/[\alpha + \lambda(m)]\} \\
&\quad + \max_{m \in [0,M]} \{r(x) - m + \int_0^\infty [-V(x) \\
&\quad + V(0) - C_2]p(x+y)\lambda(m)G(dy|m)\}/[\alpha + \lambda(m)]\} \\
&\leq V(x) \max_{m \in [0,M]} \{\lambda(m)/[\alpha + \lambda(m)]\} \\
&\quad + \max_{m \in [0,M]} \{r(x) - m - \int_0^\infty (C_2 - C_1)p(x+y)\lambda(m)G(dy|m)\}/[\alpha + \lambda(m)]\} \\
&\leq V(x) \max_{m \in [0,M]} \lambda(m)/[\alpha + \lambda(m)] \\
&< V(x),
\end{aligned}$$

yielding a contradiction. In the above argument, the first inequality follows from $V(x+y) \leq V(x)$ (Theorem 1), the third one from $V(x) \geq V(0) - C_1$, the fourth one from the fact that $x \in B$ and the last one from $V(x) > 0$ (Theorem 4). To prove (16) by contradiction, suppose that $\delta_2^*(x) \geq r(x)$ for some $x \in [0,x^*)$. Then

$$\begin{aligned}
V(x) &= U(x) \\
&= \{r(x) - \delta_2^*(x) + \int_0^\infty \{V(x+y)[1 - p(x+y)] \\
&\quad + [V(0) - C_2]p(x+y)\}\lambda(\delta_2^*(x))G(dy|\delta_2^*(x))\}/[\alpha + \lambda(\delta_2^*(x))] \\
&\leq \lambda(\delta_2^*(x))/[\alpha + \lambda(\delta_2^*(x))] \int_0^\infty \{V(x+y)[1-p(x+y)] \\
&\quad + [V(0) - C_2]p(x+y)\}G(dy|\delta_2^*(x))
\end{aligned}$$

$$\leqslant \lambda(\delta_2^*(x))/[\alpha + \lambda(\delta_2^*(x))]\{V(x)$$
$$+ \int_0^\infty [V(0) - C_2 - V(x)]p(x+y)G(dy|\delta_2^*(x))\}$$
$$\leqslant V(x)\lambda(\delta_2^*(x))/[\alpha + \lambda(\delta_2^*(x))]$$
$$< V(x),$$

again yielding a contradiction. Here the first two equalities follow from the definitions of x^* and δ_2^*, respectively. The first inequality follows from the hypothesis that $\delta_2^*(x) \geqslant r(x)$, the second inequality from monotonicity of V, the third one from $x < x^*$ and the last one follows by positivity of V.

Q.E.D.

The bound in (15) may be interpreted in terms of a "one-stage-look-ahead" stopping policy, as, for example, in Ross [13, p. 183]. Suppose we postpone the voluntary replacement of an operating system until the next shock in the hope of "squeezing" additional revenue out of it. However, such a postponement would involve the risk of a higher forced replacement cost due to possible failure the next shock might cause. The first part of Theorem 5 in essence juxtaposes these two conflicting factors in specifying an optimal replacement strategy. It asserts that if the net expected revenue until the next shock, $[r(x) - m]/\lambda(m)$, cannot at the least overcome the expected extra cost $(C_2 - C_1)\int_0^\infty p(x+y)G(dy|m)$ due to possible failure replacement at the next shock, for any choice of maintenance rate m, then it is best to replace the system right away instead of waiting. The second part of the theorem says that "living beyond one's means" cannot be the best maintenance strategy, even in a favorable environment, i.e., that the optimal maintenance rate is always strictly less than the rate at which the machine generates revenues, as given in (16).

4. AN EXAMPLE

In this section, we illustrate the model and results by providing explicit solutions for a specific example. Consider a system which fails when the cumulative damage first exceeds a prespecified threshold d (see, e.g., Buckland [2], Section 1-10), so that the failure probability function $p(\cdot)$ is given by

$$p(z) = \begin{cases} 0 & \text{if } 0 \leqslant z < d \\ 1 & \text{if } z \geqslant d \end{cases} \tag{17}$$

which is trivially nondecreasing and concave on [0,d) as per (the footnote of) Assumption P(ii). Suppose that the shock rate $\lambda(m) \equiv \lambda > 0$, independent of the maintenance rate $m \in [0,M]$, and that each shock causes either zero damage (so that the system survives) with probability m/M or damage of magnitude d (resulting in a system failure) with probability $1 - m/M$, i.e., the distribution of damage caused by a shock is

$$G(y|m) = \begin{cases} m/M & \text{if } 0 \leqslant y < d \\ 1 & \text{if } y \geqslant d \end{cases} \tag{18}$$

Then, $\lambda(m)\,[1 - G(y|m)]$ is (linearly) decreasing in m, as required in Assumption P(i). We may take the economic parameters r, C_1 and C_2 to be arbitrary ones satisfying Assumption E, although for expositional simplicity we take the reward function $r(x)$ to be strictly decreasing and convex in x (e.g., $r(x) = ke^{-x}$ with $k > 0$) and, to rule out trivial solutions, we suppose that

(19) $[r(0) - r(d)]/(\alpha + \lambda) \geq C_1$.

The optimality equations (8) and (9) now become

(20) $V(x) = \text{Max}\ \{V(0) - C_1, U(x)\}$

where

(21) $(\alpha + \lambda)\ U(x) = \underset{m \in [0,M]}{\text{Max}}\ \{r(x) - m + \lambda [V(x)m/M + (V(0) - C_2)\ (1 - m/M)]\}$.

The optimal solutions V, δ_1^* and δ_2^* depend upon relative magnitudes of certain problem parameters, as given in the following three disjoint and exhaustive cases. In each case, it can be verified in a straightforward manner that the given solutions satisfy (20) and (21).

CASE (i): $C_2 - C_1 > M/\lambda$.

In this case, the optimal return is the convex nonincreasing function given by

(22) $$V(x) = \begin{cases} [r(x) - M]/\alpha & \text{if } x \leq x^* \\ [r(0) - M]/\alpha - C_1 & \text{if } x > x^* \end{cases}$$

where the critical replacement level x^* satisfies

(23) $r(x^*) = r(0) - \alpha\ C_1$.

In light of (19) and the strict monotonicity of r, we have $x^* < d$ and that x^* is unique. As for the maintenance rule, we have $\delta_2^*(x) = M$ for all $x \in [0, x^*)$, specifying the maximum maintenance rate until replacement, since in this case the replacement cost differential is higher than the maximum maintenance cost until failure.

CASE (ii): $C_2 - C_1 \leq M/\lambda < C_2$.

In this case the solution turns out to be

(24) $$V(x) = \begin{cases} [r(x) - M]/\alpha, & 0 \leq x \leq \bar{x} \\ [r(x) + \lambda [(r(0) - M)/\alpha - C_2]]/(\alpha + \lambda), & \bar{x} < x \leq x^* \\ [r(0) - M]/\alpha - C_1 & x > x^* \end{cases}$$

where x^* satisfies

(25) $r(x^*) = r(0) - \alpha\ C_1 - M + \lambda (C_2 - C_1)$

and $\bar{x}$ satisfies

(26) $r(\bar{x}) = r(0) - \alpha [C_2 - M/\lambda]$.

Again, δ_1^* specifies replacement whenever $x \geq x^*$. The optimal maintenance policy δ_2^* is of the "bang-bang" type and is specified in terms of the switchpoint $\bar{x}$ (which is less than x^* since $(C_2 - C_1) \leq M/\lambda$) as follows:

(27) $$\delta_2^*(x) = \begin{cases} M & \text{if } 0 \leq x < \bar{x} \\ 0 & \text{if } \bar{x} \leq x < x^* \end{cases}$$

From (26), note that $\bar{x}$ is increasing in C_2.

CASE (iii) $C_2 - C_1 < C_2 \leq M/\lambda$.

In this final case, we get

$$(28)\qquad V(x)=\begin{cases} r(x)/(\alpha+\lambda)+\lambda[r(0)/(\alpha+\lambda)-C_2]/\alpha & x<x^* \\ r(0)/\alpha-C_1-\lambda C_2/\alpha & x\geqslant x^* \end{cases}$$

where the control limit x^* satisfies

$$(29)\qquad r(x^*)=r(0)-(\alpha+\lambda)C_1$$

and identically zero maintenance rate (i.e., $\delta_2^*(x)=0$ for all $x\in[0,x^*)$) is optimal.

In all three cases, from (23), (25) and (29), we observe that the optimal control limit x^* is decreasing in C_2 (or (C_2-C_1)) and increasing in C_1. Similarly, the switch-point $\bar{x}$ at which the optimal maintenance rate switches from M to 0 is increasing in C_2 (or (C_2-C_1)) and decreasing in C_1. Thus, the higher the replacement cost differential (C_2-C_1), the greater should be the intensity of preventive maintenance and replacement effort. For a concrete example, consider $d=1$, $r(x)=1-x$ and $C_1>0$ fixed. Then Figure 1 displays the parametric behavior of the optimal critical value x^* (shown by the solid line) and the optimal switch point $\bar{x}$ (shown by the broken line) as the forced replacement cost C_2 is varied.

Finally, notice that under optimal policy, in cases (i) and (ii) at most one replacement ever takes place, while in case (iii) the system is replaced at every shock. In short, our analysis has delineated conditions under which it will be optimal to actually utilize the maintenance capability available for buffering the system completely from shocks.

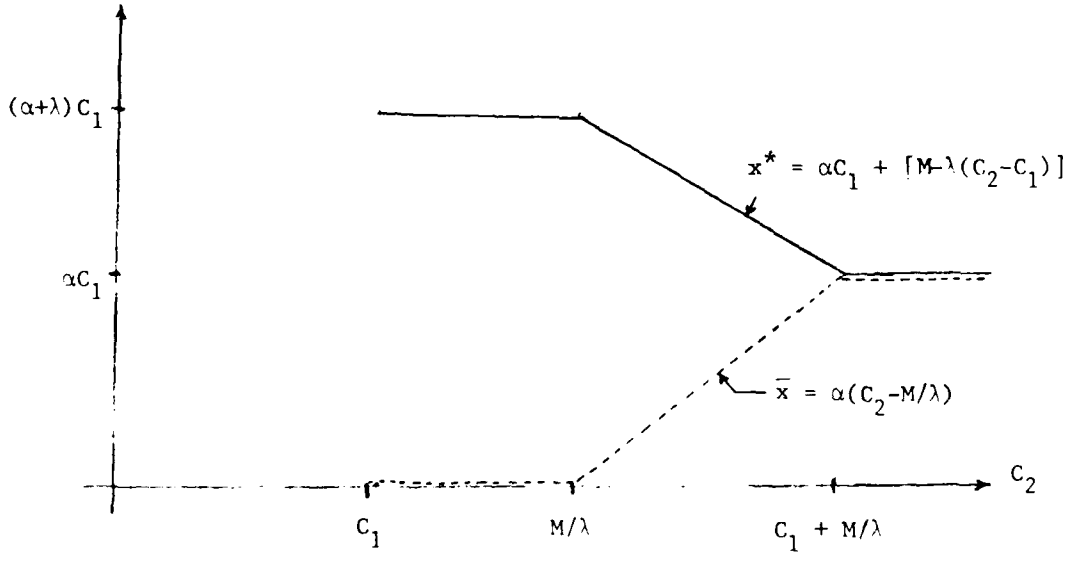


FIGURE 1

5. CONCLUDING REMARKS

In this paper, we have integrated the problems of determining optimal preventive maintenance and replacement schedules for a system that is subject to stochastic deterioration and failure induced by a shock process. Under reasonable assumptions, we have proved that the maximum obtainable return and the optimal policies have appealing features and we have illustrated these by means of an example. We conclude the paper by discussing some implementational aspects of the model.

In practice, the state of the system x may be observed in terms of some convenient surrogate measure of system efficiency, accuracy or wear such as the production (or revenue) rate, fraction defective produced, energy consumption rate, etc. Accounting and financial information may be used to estimate the discount rate and replacement costs C_1 and C_2, which depend upon such economic factors as wage levels, prices and opportunity costs of lost production during replacement delays. Statistical estimation of the parameters λ and G of the shock process and the failure probability p would require observations on the system performance together with simulation experiments. The numerical computation of optimal policy itself would require discretization of the performance space and maintenance rates m, so that standard algorithms such as the policy improvement routine (see, e.g., Ross [13]) can be employed. Given the simple structure of the optimal policy, its implementation may be based on a control chart type procedure by establishing control limits $\{x_1^*, x_2^*, \ldots x_k^*\}$ on the system deterioration level x, so that, if $x \geq x_k^*$ the system should be replaced; otherwise if $x_i^*{}_1 \leq x \leq x_i^*$, it should be maintained at an expenditure rate m_i, greater deterioration corresponding to smaller expenditures. The selection of control limits may also be based on simulation studies.

ACKNOWLEDGMENTS

We wish to thank the referee for suggesting several organizational improvements in the paper. S. D. Deshmukh's research was supported by a Kellogg Research Chair and the J.L. Kellogg Center for Advanced Study in Managerial Economics and Decision Sciences.

REFERENCES

[1] Abdel-Hameed, M.S. and I.N. Shimi, "Optimal Replacement of Damaged Devices," Journal of Applied Probability, 15, 153-161 (1978).

[2] Buckland, W.R., *Statistical Assessment of Life Characteristics*, (Charles Griffin and Co. Ltd, London, 1964).

[3] Chikte, S.D., *Markovian Decision Models for Optimal Dynamic Resource Allocation Problems*, unpublished Ph.D. dissertation, Polytechnic Institute of New York, Brooklyn, New York (1977).

[4] Chikte, S.D. and F. Kozin, "Optimal Preventive Maintenance and Replacement Strategies Under Markovian Deteriorations," *Proceedings of the 8th Annual Modeling and Simulation Conference*, University of Pittsburgh, Pittsburgh, Pennsylvania 359-363 (1977).

[5] Derman, C., "On Optimal Replacement Rules when Changes of State are Markovian," *Mathematical Optimization Techniques*, Chapter 9, R. Bellman, editor (University of California Press, Berkeley and Los Angeles, California, 1963).

[6] Feldman, R.M., "Optimal Replacement with Semi-Markov Shock Models," Journal of Applied Probability, *13*, 108-117 (1976).

[7] Feldman, R.M. "Optimal Replacement with Semi-Markov Shock Models Using Discounted Costs," Mathematics of Operations Research, *2*, 78-90 (1977).

[8] Kamien, M.I. and N.L. Schwartz, "Optimal Maintenance and Sale Age for a Machine Subject to Failure," Management Science, *17*, B495-B504 (1971).

[9] Lippman, S.A., "Optimal Pricing to Retard Entry," Review of Economic Studies, *47*, 723-731 (1980).

[10] McCall, J.J., "Maintenance Policies for Stochastically Failing Equipment: A Survey," Management Science, *11*, 493-524 (1965).

[11] Pierskalla, W.P. and J.A. Voelker, "A Survey of Maintenance Models: The Control and Surveillance of Deteriorating Systems," Naval Research Logistics Quarterly, *23*, 353-388 (1976).

[12] Pliska, S.R., "Controlled Jump Processes," Stochastic Processes and Their Applications, *3*, 259-282 (1975).

[13] Ross, S.M., *Applied Probability Models with Optimization Applications*, (Holden-Day, San Francisco, California, 1970).

[14] Stone, L.D., "Necessary and Sufficient Conditions for Optimal Control of Semi-Markov Jump Processes," SIAM Journal on Control, *11*, 187-201 (1973).

[15] Taylor, H.M., "Optimal Replacement Under Additive Damage and Other Failure Models," Naval Research Logistics Quarterly, *22*, 1-18 (1975).

[16] Thompson, G.L., "Optimal Maintenance Policy and Sale Date of a Machine," Management Science, *14*, 543-550 (1968).

OPTIMAL MAINTENANCE MODELS FOR SYSTEMS SUBJECT TO FAILURE—A REVIEW

Y.S. Sherif

Department of Industrial and Systems Engineering
The University of Alabama in Huntsville
Huntsville, Alabama

M.L. Smith

Department of Industrial Engineering
Texas Technological University
Lubbock, Texas

ABSTRACT

This paper is a state-of-the-art review of the literature related to optimal maintenance models of systems subject to failure. The emphasis is on work appearing since the 1976 survey, "A Survey of Maintenance Models: The Control and Surveillance of Deteriorating Systems," by W.P. Pierskalla and J.A. Voelker, published in this journal.

1. INTRODUCTION

Maintenance involves planned and unplanned actions carried out to retain a system in or restore it to an acceptable condition. Optimal maintenance policies aim to minimize downtime while providing for the most effective use of systems in order to secure the desired results at the lowest possible costs. Proper maintenance techniques have been emphasized over the past two decades due to increased complexity of systems, increased quality requirements and rising costs of material and labor. The two old concepts of maintenance: loving care (the reliability of the equipment is directly proportional to the frequency of maintenance), and emergency replacement (operate equipment until it is inoperable) may not be optimal. Both methods lead to improper maintenance, excessive breakdowns, and high costs. Since the 1965 and 1967 surveys on maintenance by McCall and Barzilovich [315,34], a great deal of research has been done in the field of optimal maintenance modeling, involving the aspects of optimal preventive and preparedness maintenance policies. Tables 1-3 give the references in various classifications. Some references appear more than once in Table 1 because these papers consider two or more topics related to maintenance models. Also, some papers are not referred to in Table 2 because the topics of these papers were not concerned with any specific model type.

2. OPTIMAL MAINTENANCE MODELS

The literature related to optimal maintenance models is classified as follows:

Optimal Maintenance Models

1. Deterministic Models
2. Stochastic Models
   - A. Under Risk
   - B. Under Uncertainty
     1. Simple System
     2. Complex System
        - a. Preventive Maintenance (periodic, sequential) or
        - b. Preparedness Maintenance (periodic, sequential, opportunistic).

Optimization techniques employed for obtaining optimal maintenance policies include the following:

Linear programming
Nonlinear programming
Dynamic programming
Pontryagin maximum principle
Mixed-integer programming
Decision theory
Search techniques
Heuristic approaches

The characteristics of each optimal maintenance model considered in this survey will be explained briefly.

2.1 Deterministic Models

These models incorporate the following assumptions:

- The outcome of every maintenance action is nonrandom.
- Maintenance action restores the system to its original state.
- The purchase price and salvage value of the system are taken as given functions of its age.
- Aging (were and tear) increases the costs of operating the system.
- Aging failure is not necessarily operational failure.
- All failures are new, and can be observed instantaneously.
- By prolonging the operating life of the system through maintenance, costs are incurred and benefits may increase.

The optimal maintenance policy for deterministic models is periodic and the times between successive maintenance actions must be equal.

2.2 Stochastic Models Under Risk

Risk is a time-dependent property that is measured by probability. For stochastically failing equipment under risk, it is impossible to predict the exact time of failure; but the distributions of the time to failure of each component of the system are known.

2.2.1 Simple System Preventive Maintenance Model (periodic, sequential)

This model utilizes the following assumptions:

- The system time to failure is a random variable with known distribution
- The system is either operating or failed.
- Failure is an absorbing state.
- Maintenance action regenerates the system immediately upon completion.
- The intervals between successive regeneration points are independent random variables.
- The maintenance cost is generally higher if undertaken after an operational failure than before.

The optimal policy for various assumptions is as follows:

- For systems with unlimited lifetime, the optimal preventive maintenance policy is the strictly periodic one—i.e., maintain system at failure or at an age t_i, whichever occurs first.
- For systems with constant failure rate (exponential); maintain at failure.
- For systems with increasing failure rate (Weibull, gamma, ..., etc., for some parameter values), maintain on progressive schedule.
- For systems with limited lifetime (process with a relatively short lifetime, or equipment subject to rapid technological change), the best preventive policy is the sequential one. This sequential policy recalculates the maintenance age t after each overhaul. It actually attempts to minimize the expected cost of system operation over the remaining life of the process.

2.2.2 Simple System Preparedness Maintenance Model (periodic, sequential)

This model utilizes the following assumptions:

- The system time to failure is a random variable with known distribution.
- The actual state of the system is known with certainty only at the time of inspection or maintenance.
- Failure is an absorbing state.

2.2.3 Complex System Preventive Maintenance Model (periodic, sequential, opportunistic)

This model is an extension of 2.2.1 for complex systems. The optimal policy for various assumptions is as follows:

- If the parts constituting the complex system are interconnected in such a way that they can be considered as stochastically and economically independent, then the optimal maintenance policy for this complex system reduces to that of the simple system, i.e., employ a periodic or sequential preventive maintenance policy for each separate part.

- If individual parts cannot be considered as stochastically and economically independent, then a policy called the opportunistic maintenance policy will be more effective. Under this policy, the maintenance of a single uninspected part depends on the state of one or more continuously inspected parts. The opportunistic maintenance policy is advantageous when the cost of a joint maintenance action is less than the sum of the cost of the separate maintenance actions.
- If a complex system is composed of a large collection of identical units of equipment, then a block maintenance policy may be advantageous. Under this policy, each unit is replaced on failure, and all units are replaced at periodic intervals, T, $2T$, $3T$, ..., without regard to individual unit age. Scheduled and unscheduled maintenance can be combined. Consequently, this policy is easier to implement, and results in lower administrative and maintenance costs.

2.2.4 Complex System Preparedness Maintenance Model (periodic, sequential, opportunistic)

This model is an extension of 2.2.2 for complex systems. The optimal policy for various assumptions is as follows:

- If the complex system is under continuous surveillance, then this model reduces to the preventive maintenance model described under 2.2.3.
- If the complex system is not inspected, then the only maintenance policy to secure the highest level of preparedness is replacement.

2.3 Stochastic Models Under Uncertainty

For stochastically failing equipment under uncertainty, the exact time of failure and the distribution of the time to failure are not known.

2.3.1 Preventive Maintenance Model for Simple and Complex Systems

The optimal policy for various assumptions is obtained as follows:

- When the system is new or failure data are not known, the minimax techniques are applied.
- When information about the system (failure rate, ... etc.) is partially known, Chebyshev-type bounds are applied.
- When subjective beliefs about the system failure are known, Bayesian adaptive techniques are applied.

2.3.2 Simple (complex) System Preparedness Maintenance Model

The techniques of minimax strategies, Chebyshev-type bounds and Bayesian adaptive policies can be applied to this model as explained under item 2.3.1.

TABLE 1. *General Classification*

Type	References
Inspection	24, 29, 30, 81, 104, 106, 241,242, 251, 253, 254, 259, 263, 280, 297, 337, 434, 435, 475, 482, 498, 518
Maintenance	1, 5-10, 13, 15, 16, 24, 31-37, 41, 49-58, 60-70, 74-82, 89-97, 101-106, 115-120, 121-131, 146, 154-161, 161-169, 172, 177-183, 193, 202, 205 211, 224, 233, 238, 242-252, 255-257, 266-277, 283-290, 295-301, 304, 315-329, 334-348, 359-382, 384-403, 408-421, 433, 438, 444-452, 470-475, 482-499, 502-513, 516
Reliability	2-8, 11-16, 21-29, 38, 39, 49-71, 77-92, 108-119, 123-133, 147-156, 162-189, 197-213, 221-238, 241-245, 258-271, 277-294, 303-318, 314-317, 326-347, 352-369, 372-376, 385-392, 398, 408, 419, 431-438, 450-464, 494-500, 506, 511, 515-522
Optimization Techniques	5, 13, 18, 42, 103, 140, 141, 188, 195, 197, 219, 244, 248, 252-256, 357-367, 383, 418, 442, 480, 487, 497
Decision Theory	15-20, 26-33, 34-51, 73-90, 99-117, 134-152, 158-162, 190-201, 217-228, 230-237, 243-251, 255-260, 273-276, 294-297, 307-322, 336-340, 355-360, 404-413, 422-430, 439-443, 455-469, 475-489, 512-516, 523, 524

TABLE 2. *Classification of Maintenance Models by Type*

Type	References
Deterministic	7, 26, 41, 42, 120, 128, 234, 264, 335, 365, 366, 380, 384, 394, 395, 401, 442, 487
Stochastic, Under Risk	
Preventive	4, 6, 13, 15, 17, 21, 22-29, 44-51, 69-72, 73-80, 114-116, 119-124, 138-152, 175-184, 225-240, 251, 261-280, 323-352, 364-400, 441-486, 513-516
Preparedness	10, 32, 43, 56, 73, 81, 92, 100, 104, 106, 170, 179, 211, 213, 237, 241, 243, 253, 259, 297-300, 324, 338, 342, 343, 369, 403, 411, 435, 469, 485
Opportunistic	19, 47, 199, 316, 353-355, 384, 392, 402, 437, 483, 493, 511, 512
Stochastic, Under Uncertainty	10, 18, 34, 43, 75, 86, 90, 100, 104, 105, 117, 118, 125, 126, 132, 157, 167, 173, 185, 195, 209, 236, 252, 255, 373, 417, 418, 422, 433, 436

TABLE 3. *Classification by Type of Applicable Optimization Techniques*

Technique	References
Decision Theory	1-4, 6-11, 14-39, 43-139, 142-190, 192-218, 220-226, 228-243, 245, 249-251, 259-337, 359, 361-370, 372-376, 385-387, 390-396, 400-441, 443-460, 463-479, 481, 483-496, 500-524
Dynamic Programming	40-42, 191, 199, 219, 244, 246, 252, 255-258, 360, 442, 462, 480, 488
Linear Programming	103, 271, 497
Mixed-Integer Programming	384-497
Nonlinear Programming	191, 227, 247, 248, 257
Pontryagin's Maximum Principle	5, 13, 140, 141, 389, 447, 482
Search Techniques	339, 371, 377
Simulation Techniques	374, 397, 499

ACKNOWLEDGMENT

Concerning the reference list, we have tried to be reasonably complete, however those papers which were not included were either considered not to bear directly on the topics of this survey or inadvertently overlooked. We apologize to both the researchers and readers if we have omitted any relevant papers.

We would like to thank the editor of the *NRLQ* and the referee for their excellent and exhaustive review.

REFERENCES

[1] Adam, E.L., and M.E. Pohlen, "A Scoring Methodology for Equipment Replacement Model Evaluation," AIIE Transactions, *6*, 338-343 (1976).

[2] Aggarwal, K.K., "Minimum Cost Systems with Specified Reliability," IEEE Transactions on Reliability, R-26, *3*, 166-167 (1977).

[3] Aggarwal, K.K., K.B. Misra and J.S. Gupta, "Reliability Evaluation, A Comparative Study of Different Techniques," Microelectronics and Reliability, *14*, 49-56 (1975).

[3] A-Hammed, M.S., and F. Proschan, "Shock Models with Underlying Birth Process," Journal of Applied Probability, *12*, 18-28 (1975).

[5] Ahmed, N.U., and K.E. Schenk, "Optimal Availability of Maintainable Systems," IEEE Transactions on Reliability, R-27, *1*, 41-45 (1978).

[6] Alam, M., and V.V.S. Sarma, "Optimum Maintenance Policy for an Equipment Subject to Deterioration and Random Failure," IEEE Transactions on Systems, Man, and Cybernetics, SMC-4, *4*, 172-175 (1974).

[7] Alchian, A., "Economic Replacement Policy," in *Discounted Cash Flow Analysis: Stochastic Extensions*, A. Reismen and A. Rao, AIIE, Economy Division, Monograph Series, 16-19 (1973).

[8] Alexanian, L.T., and D.E. Brodie, "A Method for Estimating the Reliability of ICS," IEEE Transactions on Reliability, R-26, *5*, 359-361 (1977).

[9] Allen, S.G., and D.A. D'Esopo, "An Ordinary Policy for Repairable Stock Items," Operations Research, *16*, 3, 669-674 (1968).

[10] Ansell, J., A. Bendell, S. Humble, and C.S. Mudhar, "3-State and 5-State Reliability Models," IEEE Transactions on Reliability, R-29, *2*, 176-177 (1980).

[11] Antelman, G.R., and I.R. Savage, "Characteristic Functions of Stochastic Integrals and Reliability Theory," Naval Research Logistics Quarterly, *12*, 3, 199-222 (1965).

[12] Arinc Research Corporation, *Reliability Engineering*, Prentice Hall, Englewood Cliffs, N.J., 1964).

[13] Arora, S.R., and P.T. Lele, "A Note on Optimal Maintenance Policy and Sale Date of A Machine," Management Science, *13*, 3, 170-173 (1970).

[14] Arrow, K., D. Levhari, and E. Sheshinski, "A Production Function for the Replacement Problem," The Review of Economic Studies, *39*, 241-249 (1972).

[15] Ascher, H., "Evaluation of Repairable System Reliability Using the Bad as Old Concepts," IEEE Transactions on Reliability, R-17, *2*, 105-110 (1968).

[16] Ascher, H., and H. Feingold, "Is There Repair After Failure?" IEEE Annual Reliability and Maintainability Symposium, 190-197 (1978).

[17] Ashar, K.G., "Probabilistic Model of Systems Operation with a Varying Degree of Spares and Services Facilities," Operations Research, *8*, 5, 707-718 (1960).

[18] Aumann, R.J., and M. Maschler, "Some Thoughts on the Minimax Principle," Management Science, *18*, 5, P54-P63 (1972).

[19] Bagdonavius, V.B., "A Statistical Test of a Model of Additive Accumulation of Damage," SIAM Theory of Probability and Its Applications, *23*, 2, 385-389 (1978).

[20] Bansard, J.P., J.L. Descamps, G. Maarek, and G. Morihain, "Study of Stochastic Renewal Process Applied to a Trigger-Off Policy for Replacing Whole Sets of Components," *Proceedings of the International Conference on Operations Research*, 235-264 (1970).

[21] Barlow, R.E., "Analysis of Retrospective Failure Data Using Computer Graphics," *IEEE Annual Reliability and Maintenance Symposium* (1978).

[22] Barlow, R.E., and A.W. Marshall, "Bounds on Distributions with Monotone Hazard Rates," Anals of Mathematical Statistics, *35*, 3, 1234-1274 (1964).

[23] Barlow, R.E., and L.C. Hunter, "Mathematical Models for System Reliability," Sylvania Technologist, *13*, 1, 16-31 (1960).

[24] Barlow, R.E., and L.C. Hunter, "Optimum Preventive Maintenance Policies," Operations Research *8*, 1, 90-100 (1960).

[25] Barlow, R.E., and L.C. Hunter, "Reliability Analysis of a One-Unit System," Operations Research 9, 2, 200-208 (1961).

[26] Barlow, R.E., and F. Proschan, "Comparison of Replacement Policies and Renewal Theory Implications," Annal Mathematical Statistics, *35*, 2, 577-589 (1964).

[27] Barlow, R.E., and F. Proschan, "Planned Replacement," in *Studies in Applied Probability and Management Science*, K.J. Arrow, S. Karlin, and H. Scarf, (Stanford University Press, Stanford, California, 1962).

[28] Barlow, R.E., and F. Proschan, *Statistical Theory of Reliability and Life Testing*, (Holt, Rinehart and Winston, N.Y. 1975).

[29] Barlow, R.E., F. Proschan, and L.C. Hunter, *Mathematical Theory of Reliability*, (John Wiley, N.Y. 1965).

[30] Barlow, R.E., L.C. Hunter, and F. Proschan, "Optimum Checking Procedures," Journal of the Society for Industrial and Applied Mathematics, *11*, 4, 1078-1095 (1963).

[31] Barr, J.L., and K.E. Knight, "Technological Change and Learning in the Computer Industry," Management Science, *14*, 11, 661-681 (1968).

[32] Bar-Shalom, Y., and A.I. Cohen, "Optimal Resource Allocation for an Environmental Surveillance System," IEEE Transactions on Systems, Man, and Cybernetics, SMC-6, *6*, 391-400 (1976).

[33] Bar-Shalom, Y., R.E. Larson, and M. Grossbert, "Allocation of Stochastic Control Theory to Response Allocation under Uncertainty," IEEE Transactions on Automatic Control, AC-19 *1*, 1-7 (1974).

[34] Barzilovich, Y.Y., "Maintenance of Complex Technical Systems, II (Survey)," Engineering Cybernetics, *1*, 63-75 (1967).

[35] Barzilovich, Y.Y., "Servicing of Complex Technical Systems, I," Engineering Cybernetics, *5*, 67-81 (1966).

[36] Basker, B.A., and T.M. Husband, "Determination of Optimal Overhaul Intervals and Inspection Frequencies—A Case Study," Microelectronics and Reliability, *17*, 2, 313-315 (1978).

[37] Basker, B.A., and T.M. Husband, "Simulating Maintenance Work in an Engineering Firm: A Case Study," Microelectronics and Reliability, *16*, 5, 571-581 (1977).

[38] Basu, A.P., "Estimates of Reliability for Some Distribution Useful in Life Testing," Technometrics, *6*, 2, 215-219 (1964).

[39] Bazovsky, I., Reliability Theory and Practice, (Prentice Hall, Englewood Cliffs, N.J., 1961).

[40] Bellman, R., *Dynamic Programming*, (Princeton Univ. Press Princeton, N.J., 1957).

[41] Bellman, R., "Equipment Replacement Policy," Journal of the Society for Industrial and Applied Mathematics, *3*, 3, 133-136 (1955).

[42] Bellman, R., and S.E. Dreyfus, *Applied Dynamic Programming*, (Princeton Univ. Press *Princeton*, N.J., 1962).

[43] Beja, A., "Probability Bounds in Replacement Policies for Markov Systems," Management Science, *16*, 3, 257-264 (1969).

[44] Berg, M., "A Proof of Optimality for Age Replacement Policies," *Journal of Applied Probability*, *13*, 4, 751-759 (1976).

[45] Berg, M., "General Trigger-Off Replacement Procedures for Two-Unit Systems," Naval Research Logistics Quarterly, *25*, 1, 15-29 (1978).

[46] Berg, M., "Optimal Replacement Policies for Tow-Unit Machines with Increasing Running Costs—I," Stochastic Processes and Their Applications, *4*, 80-106 (1976).

[47] Berg, M., and B. Epstein, "Comparison of Age, Block, and Failure Replacement Policies," IEEE Transactions on Reliability, R-27 *1*, 25-29 (1978).

[48] Bergman, B., "Optimal Replacement Under a General Failure Model," Advances in Applied Probability, *10*, 2, 431-451 (1978).

[49] Bigel, G., and J. Winston, "Reliability and Maintainability Growth of a Modern High Performance Aircraft, the F14A," Microelectronics and Reliability, *19*, 1, 31-38 (1979).

[50] Birnbaum, Z.W., J.D. Esary, and A.W. Marshall, "A Stochastic Characterization of Wear-Out for Components and Systems," Annals of Mathematical Statistics, *37*, 1, 816-825 (1966).

[51] Blumenthal, S., J.A. Greenwood, and L. Herbach, "A Comparison of the Bad as Old and Superimposed Renewal Models," Management Science, *23*, 3, 280-285 (1976).

[52] Boness, A.J., and A.N. Schwartz, "A Cost-Benefit Analysis of Military Aircraft Replacement Policies," Naval Research Logistics Quarterly, *16*, 2, 237-257 (1969).

[53] Borgerson, B.R., and R.L. Freitas, "A Reliability Model for Gracefully Degrading and Standby Sparing Systems," IEEE Transactions on Computers, C-24, 5, 517-525 (1975).

[54] Bosch, G., "Model for Failure Rate Curves," Microelectronics and Reliability, *19*, 4, 371-376 (1979).

[55] Bovaird, R.L., "Characteristics of Optimal Maintenance Policies," Management Science, *7*, 3, 238-253 (1961).

[56] Bracken, J., and K.W. Simmon, "Minimizing Reductions in Readiness Caused by Time-Phased Decreases in Aircraft Overhaul and Repair Activities," Naval Research Logistics Quarterly, *13*, 2, 159-165 (1966).

[57] Brandt, E., and D.R. Limaye, "Mathematical Analysis of Downtime," Naval Research Logistics Quarterly, *17*, 2, 525-534 (1970).

[58] Branson, M.H., and B. Shah, "Reliability Analysis of Systems Comprised of Units with Arbitrary Repair-Time Distributions," IEEE Transactions on Reliability, R-20, *4*, 217-223 (1971).

[59] Britney, R.R., "The Reliability of Complex Systems with Dependent Sub-System Failures: An Absorbing Markov Chain Model," Technometrics, *16*, 2, 245-250 (1974).

[60] Brown, D.B., and H.F. Martz, Jr., "A Two Phase Algorithm for the Maintenance of a Deteriorating Component System," IEEE Transactions on Reliability, R-20, *1*, 28-32 (1971).

[61] Bryson, A.E., and Y.C. Ho, *Applied Optimal Control*, (Blaisdell Publishing Co., Waltham, Mass. 1969).

[62] Bryson, M.C., and M.M. Siddiqui, "Some Criteria for Aging," Journal of the American Statistical Association, *64*, 328, 1472-1483 (1969).

[63] Bury, K.V., "On Product Reliability under Random Field Loads," IEEE Transactions on Reliability, R-27, *4*, 258-260 (1978).

[64] Butler, D.A., "A Hazardous-Inspection Model," Management Science, *25*, 1, 79-89 (1979).

[65] Buzacott, J.A., The Effect of Non-stationary Breakdown and Random Processing Times on the Capacity of Flow Lines with In-Process Inventory," AIIE Transactions, *4*, 4, 308-312 (1972).

[66] Calabro, S.R., Reliability Principles and Practices, (McGraw-Hill, New York, 1962).

[67] Campbell, R.L., "General Maintenance Techniques for Large Digital Systems," Annals of Reliability and Maintainability, AIAA, 5, 247-255 (1966).

[68] Canfield, R.V., and L.E. Borgman, "Some Distributions of Time to Failure for Reliability Applications," Technometrics, 17, 2, 263-268 (1975).

[69] Chan, P.K.W., and T. Downs, "Optimization of Maintained Systems," IEEE Transaction on Reliability, R-29, 1, 42-44 (1980).

[70] Chandhuri, D., and K.C. Sahu, "Preventive Maintenance Interval for Optimal Reliability of Deteriorating System," IEEE Transactions on Reliability, R-26, 5, 371-371 (1977).

[71] Chandra, S. and D.B. Owen, "On Estimating the Reliability of a Component Subject to Several Different Stresses," Naval Research Logistics Quarterly, 23, 1, 31-39 (1975).

[72] Chang, F.Y., and W.T. Thompson, "Bayes Analysis of Reliability for Complex Systems," Operations Research, 24, 1, 156-168 (1976).

[73] Chitgopekar, S.S., "A Note on the Costly Surveillance of a Stochastic System," Naval Research Logistics Quarterly, 21, 3, 365-371 (1974).

[74] Cho, H.H., "On the Proper Preventive Maintenance," *9th National Symposium on Reliability and Quality Control*, 431-438 (1963).

[75] Chu, W.W., "Adaptive Diagnosis of Faulty Systems," Operations Research, 16, 4, 915-927 (1968).

[76] Cinlar, E., *Introduction to Stochastic Processes*, (Prentice-Hall, Englewood Cliffs, N.J. 1975).

[77] Clarke, J.M., "No-Growth Reliability Curves," The Journal of Environmental Sciences, 22, 2, 35-38 (1979).

[78] Clement, R.R., "Reliability of Multichannel Systems," SIAM Review, 22, 1, 88-95 (1980).

[79] Cliff, R.A., "Acceptable Testing of VLSI Components which Contain Error Corrections," IEEE Transactions on Computers, C-29, 2, 125-134 (1980).

[80] Coleman, J.J., "Reliability of Aircraft Structures in Resisting Chance Failure," Operations Research 7, 5, 639-645 (1959).

[81] Coleman, J.J., and T.J. Abrams, "Mathematical Model for Operational Readiness," Operations Research, 10, 1, 126-138 (1962).

[82] Corder, A.S., *Maintenance Management Techniques*, (McGraw-Hill, New York, N.Y., 1976).

[83] Costes, A., C. Landrault, and J.C. Laprie, "Reliability and Availability Models for Maintained Systems Featuring Hardware Failure and Design Faults," IEEE Transactions on Computers, C-27, 6, 548-560 (1978).

[84] Cox, D.R., Renewal Theory (Methuen and Co. London, England: 1971).

[85] Cox, D.R., and P.A.W. Lewis, *The Statistical Analysis of Series of Events* (Methuen and Co. London, England: 1966).

[86] Cox, D.R., and W.L. Smith, "In the Superposition of Renewal Processes," Biometrika, 41, 1, 91-99 (1954).

[87] Cozzolino, J.M., "Probabilistic Models of Decreasing Failure Rate Processes," Naval Research Logistics Quarterly, 15, 3, 351-374 (1968).

[88] Cozzolino, J.M., "The Optimal Burn-In Testing of Repairable Equipment," Naval Research Logistics Quarterly, 17, 2, 167-181 (1970).

[89] Crabill, T.P., "Optimal Control of a Maintenance System with Variable Service Rates," Operations Research 22, 4, 736-745 (1974).

[90] Crelin, G.L., "The Philosophy and Mathematics of Bayes Equation," IEEE Transactions on Reliability, R-21, 3, 131-135 (1972).

[91] Crow, L.H., "Reliability Analysis for Complex Repairable System," in *Reliability and Biometry*, F. Proschan and R. Serfling, Editors (SIAM, Philadelphia, Pa., 1974).

[92] Crow, L.H., and I.N. Shimi, "Maximum Likelihood Estimation of Life-Distributions from Renewal Testing," Annals of Mathematical Statistics, *43*, 6, 1827-1838 (1972).

[93] D'aversa, J.S., and J.F. Shapiro, "Machine Maintenance and Replacement by Linear Programming and Enumeration," Journal of the Operational Research Society, *29*, 8, 759-768 (1978).

[94] Davis, D.J., "An Analysis of Some Failure Data," Journal of the American Statistical Association, *47*, 258, 113-150 (1952).

[95] Dean, M.M., *Optimization by Variational Methods*, (McGraw-Hill N.Y., 1969).

[96] Dean, B.V., "Replacement Theory," in *Progress in Operations Research*, *1*, R.L. Ackoff, Editor (John Wiley and Sons, New York, N.Y. 1963).

[97] Deavers, K.L., and J.J. McCall, "An Analysis of Procurement and Product Improvement Decisions," The RAND Corp., RM 3859-PR (1963).

[98] Denardo, E., and B. Fox, "Nonoptimality of Planned Replacement Intervals of Decreasing Failure Rate," Operations Research, *15*, 2, 358-359 (1967).

[99] Derman, C., *Finite State Markovian Decision Processes*, (Academic Press, New York, N.Y. 1970).

[100] Derman, C., "On Minimax Surveillance Schedules," Naval Research Logistics Quarterly, *8*, 4, 415-419 (1961).

[101] Derman, C., "On Optimal Replacement Rules When Changes of State are Markovian," in *Mathematical Optimization Techniques*, R. Bellman, Editor, (University of California Press, Berkeley, Calif. 1963).

[102] Derman, C., "Optimal Replacement and Maintenance under Markovian Determination with Probability Bounds on Failure," Management Science, *9*, 3, 478-481 (1963).

[103] Derman, C., "On Sequential Decisions and Markov Chains," Management Science, *9*, 1, 16-24 (1962).

[104] Derman, C., and M. Klein, "Surveillance of Multi-Component Systems: A Stochastic Traveling Salesman Problem," Naval Research Logistics Quarterly, *13*, 2, 103-11 (1966).

[105] Derman, C., G.J. Lieberman, and S.M. Ross, "A Renewal Decision Problem," Management Science, *24*, 5, 554-561 (1978).

[106] Derman, C., and J. Sacks, "Replacement of Periodically Inspected Equipment," Naval Research Logistics Quarterly, *7*, 4, 597-607 (1960).

[107] Dhillon, B.S., "A Common Cause Failure Availability Model," Microelectronics and Reliability, *17*, 6, 583-584 (1978).

[108] Dhillon, B.S., "A Two Failure Modes Systems with Cold Stand-by Units," Microelectronics and Reliability, *18*, 3, 251-252 (1978).

[109] Dhillon, B.S., and C. Singh, "Bibliography of Literature on Fault Trees," Microelectronics and Reliability, *17*, 501-503 (1978).

[110] Dhillon, B.S., and C. Singh, "On Fault Trees and Their Reliability Evaluation Methods," Microelectronics and Reliability, *19*, 1, 57-63 (1979).

[111] Diaconis, Persi, and Donald Ylvisaker, "Conjugate Priors for Exponential Families," The Annals of Statistics, *7*, 2, 269-281 (1979).

[112] Dick, R.S., "The Reliability of Repairable Complex Systems—Part B: The Dissimilar Machine Case," IEEE Transactions on Reliability, R-12, *1*, 1-8 (1963).

[113] Dickinson, W.E., and R.M. Walker, "Reliability Improvement by the Use of Multiple-Element Switching Circuits," I.B.M. Journal, *2*, 2, 142-147 (1958).

[114] Diston, M., and G. Weiss, "Burn-in Programs for Repairable Systems," IEEE Transactions on Reliability, R-25, *5*, 265-267 (1973).

[115] DiVeroli, J.C., "Optimal Continuous Policies for Repair and Replacement," Operational Research Quarterly, *25*, 1, 89-98 (1974).

[116] Donelson, J., "Cost Model for Testing Program Based on Nonhomogeneous Poisson Failure Model," IEEE Transactions on Reliability, R-26, *2*, 189-195 (1977).

[117] Doob, J.L., *Stochastic Processes*, (John Wiley and Sons, New York, N.Y., 1953).

[118] Drake, A.W., "Bayesian Statistics for the Reliability Engineer," *Proceedings of the Annual Symposium on Reliability*, 315-320 (1966).
[119] Drenick, R.F., "The Failure Law of Complex Equipment," Journal of the Society for Industrial and Applied Mathematics, *8*, 4, 680-690 (1960).
[120] Dreyfus, S., "A Generalized Equipment Study," Journal of the Society for Industrial and Applied Mathematics, *8*, 3, 425-435 (1960).
[121] Drinkwater, R.W., and N.A.J. Hastings, "An Economic Replacement Model," Operational Research Quarterly, *18*, 2, 121-138 (1967).
[122] Dryden, J.A., and J.P. Large, "A Critique of Spacecraft Cost Models," The RAND Corporation, R-2196-1-AF (1977).
[123] Dubey, S.D., "Asymptotic Properties of General Estimation of the Weibull Parameters," Technometrics, *7*, 3, 423-434 (1965).
[124] Eble, F.A., Maintenance Strategies for Ambiguous Faults," *IEEE Proceedings, Reliability and Maintainability Symposium, Philadelphia, Pennsylvania*, 238-245 (1972).
[125] Eckles, J.E., "Optimum Maintenance with Incomplete Information," Operations Research, *16*, 5, 1058-1067 (1968).
[126] Edwards, W., "Cognitive Processes and Assessment of Subjective Probability Distributions," Journal of the American Statistical Association, *70*, 350, 291-293 (1975).
[127] Eisen, M., and M. Leibowitz, "Replacement of Randomly Deteriorating Equipment," Management Science, *9*, 2, 268-276 (1963).
[128] Eisenhut, P.S., "New Insights into the Life Cycle Approach," AIIE Transactions, *5*, 2, 150-155 (1973).
[129] Elandt-Johnson, R.C., "Some Prior and Posterior Distributions in Survival Analysis: A Critical Insight on Relationships Derived from Cross-Sectional Data," Journal of the Royal Statistical Society, *42*, 1, 96-106 (1980).
[130] Elsayed, E.A., and B.S. Dhillon, "Repairable Systems with One Standby Unit," Microelectronics and Reliability, *19*, 3, 243-246 (1979).
[131] Emoto, S.E., and R.E. Schafer, "On the Specification of Repair Time Requirements," IEEE Transactions on Reliability, R-29, *1*, 13-16 (1980).
[132] Ermer, D.S., "A Bayesian Model of Machining Economics for Optimization by Adaptive Control," ASME Journal of Engineering for Industry, *92*, B, 3, 628-632 (1970).
[133] Esary, J.D., and A.W. Marshall, "Multivariate Distributions with Increasing Hazard Rate Average," The Annals of Probability, *7*, 2, 359-370 (1979).
[134] Esary, J.D., A.W. Marshall, and F. Proschan, "Shock Models and Wear Processes," Annals of Probability, *1*, 4, 627-649 (1973).
[135] Esary, J.D., and F. Proschan, "Relationship Between System Failure Rate and Component Failure Rates," Technometrics, *5*, 2, 183-189 (1963).
[136] Everett, H., "Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," Operations Research, *11*, 3, 399-417 (1963).
[137] Falkner, C.H., "Jointly Optimal Deterministic Inventory and Replacement Policies," Management Science, *16*, 9, 622-635 (1970).
[138] Falkner, C.H., "Jointly Optimal Inventory and Maintenance Policies for Stochastically Failing Equipment," Operations Research, *16*, 3, 587-601 (1968).
[139] Falkner, C.H., "Optimal Spares for Stochastically Failing Equipment," Naval Research Logistics Quarterly, *16*, 3, 287-295 (1969).
[140] Fan, L.T., *The Continuous Maximum Principle—A Study of Complex Systems Optimization*, (John Wiley, New York, N.Y., 1966).
[141] Fan, L.T., C.S. Wang, *The Discrete Maximum Principle—A Study of Multistage Systems Optimization*, (John Wiley, New York, N.Y. 1964).
[142] Feeney, G.J., and C.C. Sherbrooke, "The (S-1,S) Inventory Policy Under Compound Poisson Demand," Management Science, *12*, 5, 391-411 (1966).

[143] Feldman, R.M., "Optimal Replacement for Systems Governed by Markov Additive Shock Processes," Annals of Probability, *5*, 3, 413-429 (1977).

[144] Feldman, R.M., "Optimal Replacement With Semi-Markov Shock Model," Journal of Applied Probability, *13*, 1, 108-117 (1976).

[145] Feller, W., *An Introduction to Probability Theory and Its Applications*, John Wiley and Sons New York, N.Y. 1971).

[146] Feyerherm, M.P., and H.W. Kennedy, "Practical Maintainability Numbers," *6th National Symposium on Reliability and Quality Control*, 343-346 (1960).

[147] Flehinger, B.J., "A General Model for the Reliability Analysis of Systems under Various Preventive Maintenance Policies," Annals of Mathematical Statistics, *33*, 1, 137-156 (1962).

[148] Flehinger, B.J., "Reliability Improvement Through Redundancy at Various System Levels," I.B.M. Journal, *2*, 2, 148158 (1958).

[149] Flehinger, B.J., "System Reliability as a Function of System Age: Effects of Intermittent Component Usage and Periodic of Intermittent Component Usage and Periodic Maintenance," Operations Research, *8*, 1, 30-44 (1960).

[150] Flehinger, B.J. and P.A. Lewis, "Two-Parameter Lifetime Distributions for Reliability Studies of Renewal Processes," I.B.M. Journal *3*, 1, 58-74 (1959).

[151] Folkman, J., and S. Port, "Optimal Procedures for Stochastically Failing Equipment," Journal of Applied Probability, *3*, 521-537 (1966).

[152] Folks, J.L., and R.H. Browne, "On the Interpretation of the Observed Confidence," Technometrics, *17*, 3, 287-290 (1975).

[153] Ford, R., Jr. and D.R. Fulkerson, *Flows in Networks*, (Princeton University Press, Princeton, N.J. 1962).

[154] Forman, E.H., and N.D. Singpurwalla, "An Empirical Stopping Rule for Debugging and Testing Computer Software," Journal of the American Statistical Association, *72*, 360, 750-757 (1977).

[155] Foster, F.G., and I. Elce, "A Simulation Programme for Machine Maintenance, Telephone Traffic and Stock Models," Operational Research Quarterly, *14*, 3, 333-342 (1963).

[156] Fox, B., "Age Replacement with Discounting," Operations Research, *14*, 3, 533-537 (1966).

[157] Fox, B., "Adaptive Age Replacement," Journal of Mathematical Analysis and Application, *18*, 2, 365-376 (1967).

[158] Fox, B., "Markov Renewal Programming by Linear Fractional Programming," SIAM Journal of Applied Mathematics, *14*, 6, 1418-1432 (1966).

[159] Frair, L.C., P.M. Ghare, and K.L. Frair, "Optimization of System Reliability via Redundancy and/or Design considerations," IEEE Transactions on Reliability, R-29, *1*, 33-35 (1980).

[160] Friedman, L. and I.B. Bertsbakh, "Maximum Likelihood Estimation in a Minimum-Type Model with Exponential and Weibull Failure Modes," Journal of the American Statistical Association, *75*, 370, 460-465 (1980).

[161] Gabrielson, I.N., "Frequency of Maintenance," *Annals of Reliability and Maintainability AIAA*, *5*, 428-433 (1966).

[162] Gandara, A., and M.D. Rich, "Reliability Improvement Warranties for Military Procurement," The RAND Corporation, R-2264-AF (1977).

[163] Gangadharan, A.C., and S.J. Brown, *Failure Data and Failure Analysis: In Power and Processing Industries*, American Society for Mechanical Engineers, Inc., New York, N.Y. (1977).

[164] Gani, J., "On the Age Distribution of Replacement Ranked Elements," Journal of Mathematical Analysis and Applications, *10*, 3, 589-597 (1965).

[165] Garriba, S., G. Reina, and A. Volta, "Availability of Repairable Units when Failure and Restoration Rates Age in Real Time," IEEE Transactions on Reliability, R-25, *2*, 88-94 (1976).

[166] Gaver, D.P., Jr., "Time to Failure and Availability of Paralleled Systems with Repair," IEEE Transactions on Reliability, R-12, 2, 30-38 (1963).

[167] Gaver, D.P., Jr., and M. Mazumdar, "Some Bayes Estimates of Long-Run Availability in a Two-State System," IEEE Transactions on Reliability, R-18, 4, 184-189 (1969).

[168] Gen, M., H. Okuno and S. Shinofuji, "An Optimizing Method in System Reliability with Failure Modes by Implicit Enumeration Algorithm," Journal of the Operations Research Society of Japan, *19*, 2, 99-116 (1976).

[169] Gertsbakh, I.B., *Models of Preventive Maintenance*, (Elsevier-North Holland Co. New York, N.Y., 1977).

[170] Gertsbakh, I.B., "Optimum Use of Reserve Elements," Engineering Cybernetics, 5, 73-78 (1966).

[171] Gertsbakh, I.B., "Reliability Characteristics of Alternative Check Up Schedules for Detecting Hidden Failures," Journal of the Operational Research Society, *29*, 12, 1219-1229 (1978).

[172] Gertsbakh, I.B., and K.B. Kordonskiy *Models of Failure*, (Springer-Verlag New York, N.Y., 1969).

[173] Girshick, M., and H. Rubin, "A Bayes Approach to a Quality Control Model," Annals of Mathematical Statistics, *23*, 1, 116-125 (1952).

[174] Glasser, G.J., "The Age Replacement Problem," Technometrics, 9, 1, 83-91 (1967).

[175] Glass, B., "An Optimum Policy for Detecting a Fault in a Complex System," Operations Research, 7, 4, 468-477 (1959).

[176] Goheen, T.C., "On the Optimal Operating Policy for the Machine Repair Problem when Failure and Repair Times have Erlang Distribution," Operations Research, *25*, 3, 484-492 (1977).

[177] Goldman, A.S., and T.B. Slattery, *Maintainability: A Major Element of System Effectiveness*, (John Wiley and Sons, New York, N.Y., 1967).

[178] Golodnikov, A.N., and L.S. Stoikova, "Determination of the Optimal Preventive Replacement Period on the Basis of Information on the Mathematical Expectation and Variance of the Trouble-Free System Operation Time," Cybernetics, *11*, 3, 431-440 (1979).

[179] Goodman, L., "Methods of Measuring Useful Life of Equipment Under Operational Conditions," Journal of the American Statistical Association, *48*, 263, 503-551 (1953).

[180] Gopal, K., K.K. Aggarwal and J.S. Gupta, "A New Approach to Reliability Optimization in General Modular Redundant Systems," Microelectronics and Reliability, *18*, 5, 419-422 (1978).

[181] Gopal, K., K.K. Aggarwal, and J.S. Gupta, "Reliability Optimization in Systems with Many Failure Modes," Microelectronics and Reliability, *18*, 5, 423-425 (1978).

[182] Gopalan, M.N., and K.Y. Marathe, "Availability Analysis of 1-Server *n*-Unit System with Slow Switch Subject to Maintenance," IEEE Transactions on Reliability, R-29, 2, 189 (1980).

[183] Gradon, F., *Maintenance Engineering*, (John Wiley and Sons, New York, N.Y., 1973).

[184] Green, A.E. and A.J. Bourne, *Reliability Technology* (John Wiley and Sons, London, England, 1972).

[185] Grohowski, G., W.C. Hausman, and L.R. Lamberson, "A Bayesian Statistical Inference Approach to Automobile Reliability Estimation," Journal of Quality Technology, *8*, 4, 197-208 (1976).

[186] Gronchko, D., *Operations Research and Reliability*, (Gordon and Breach, Service Publishers, New York, N.Y., 1971).

[187] Gross, Donald and John F. Ince, "Spares Provisioning for Repairable Items: Cyclic Queues in Light Traffic," AIIE Transactions, *10*, 3, 307-314 (1978).

[188] Guild, R.D., and J.D. Chipps, "High Reliability Systems by Multiplexing," Journal of Quality Technology, 9, 2, 62-69 (1977).

[189] Gupta, P.P., "Complex System Reliability with High General Repair Time Distribution under Preemptive Resume Repair Discipline," Microelectronics and Reliability, *12*, 351-356 (1973).

[190] Haber, S., and R. Sitgreaves, "A Methodology for Estimating Expected Usage of Repair Parts with Application to parts with no Usage History," Naval Research Logistics Quarterly, *17*, 2, 535-546 (1970).

[191] Hadley, G., *Nonlinear and Dynamic Programming* (Addison-Wesley, Reading, Mass., 1966).

[192] Hadley, G., and T.M. Whittin, *Analysis of Inventory Systems* (Prentice Hall, Englewood Cliffs, N.J., 1963).

[193] Hall, K.M., "System Maintainability," *8th National Symposium on Reliability and Quality Control*, 310-321 (1962).

[194] Hallburg, O., "Failure Rate as a Function of Time Due to Lon-Normal Life Distributions of Weak Parts," Microelectronics and Reliability, *16*, 2, 155-158 (1977).

[195] Hammond, J.S., "Simplifying the Choice Between Uncertain Prospects where Preference is Nonlinear," Management Science, *20*, 7, 1047-1072 (1974).

[196] Hanscom, M. and R. Cleroux, "The Block Replacement Problem," Journal of Statistical Computation and Simulation, *3*, 233-248 (1975).

[197] Harris, C.M., and N.D. Singpurwalla, "Life Distributions Derived from Stochastic Hazard Functions," IEEE Transactions on Reliability, R-17, 2, 70-79 (1968).

[198] Hart, S., E.E. Daniel Jr., and T.J. Hodgson, "An Efficient Computational Alternative to Using Linear Programming to Design Oil Pollution Detection Schedules, AIIE Transactions, *10*, 1, 48-51 (1978).

[199] Hastings, N.A.J., *Dynamic Programming With Management Applications* (The Butterworth Group, London, England (1973).

[200] Hastings, N.A.J., "The Repair Limit Replacement Method," Operations Research Quarterly, *20*, 3, 337-350 (1969).

[201] Hastings, N.A.J., and A.K.S. Jardine, "Component Replacement and the Use of Relcode," Microelectronics and Reliability, *19*, 1, 56-69 (1979).

[202] Hatoyama, Y., "On Optimal Policies for Multi-Repair Type Markov Maintenance Models," Journal of the Operations Research Society of Japan, *22*, 2, 106-121 (1979).

[203] Haviland, R.P., *Engineering Reliability and Long Life Design* (D. Van Nostrand Co. Princeton, N.J., 1964)

[204] Hayes, J.P., "A Graph Model for Fault-Tolerant Computing," IEEE Transactions on Computers, C-25, 9, 875-884 (1976).

[205] Heggerston, H.E., and T.E. Brazier, "A Decision Table for Determining a Base-Depot Maintenance Policy for Aircraft Engines," *Annals of Reliability and Maintainability AIAA*, *5*, 84-91 (1966).

[206] Henin, C., "Optimal Allocation of Unreliable Components for Maximizing Expected Profit Over Time," Naval Research Logistics Quarterly, *20*, 3, 395-403 (1973).

[207] Henin, C., "Optimal Replacement Policies for a Single Loaded Sliding Standby," Management Science, *18*, 11, 706-715 (1972).

[208] Hevesh, A.H., and D.J. Harrahy, "Effects of Failure on Phased-Array Radar System," IEEE Transactions on Reliability, R-15, *1*, 22-32 (1969).

[209] Higgins, J.J., and C.P. Tsokos, "Sensitivity of Bayes Estimates of Reciprocal MTBF and Reliability to an Incorrect Failure Model," IEEE Transactions on Reliability, R-26, *4*, 286-289 (1977).

[210] Hildebrand, J.K., *Maintenance Turns to the Computer*, Mass. Cahners Books International, Boston, Mass. (1972)

[211] Hilliard, J.E., "An Approach to Cost Analysis of Maintenance Float Systms," AIEE Transactions, *8*, 1, 128-133 (1976).

[212] Hillier, F.S., "Surveillance Programs for Lots in Storage," Technometrics, *4*, 4 (1962).

[213] Hitomi, K., N. Nakamura and S. Inoue, "Reliability Analysis of Cutting Tools," ASME Journal of Engineering for Industry, *101*, 2, 185-190 (1979).

[214] Hjorth, U., "A Reliability Distribution with Increasing, Decreasing, Constant and Bathtub-Shaped Failure Rates," Technometrics, *22*, 1, 99-107 (1980).
[215] Hodgson, V., and T.L. Hebble, "Nonpre-emptive Priorities in Machine Interference," Operations Research, *15*, 2, 245-253 (1967).
[216] Holland, C.W., and R.A. McLean, "Applications of Replacement Theory," AIEE Transactions, *7*, 1, 42-47 (1975).
[217] Hosford, J.E., "Measures of Dependability," Operations Research *8*, 1, 53-64 (1960).
[218] Hopkins, D., "Infinite-Horizon Optimality in an Equipment Replacement and Capacity Expansion Model," Management Science, *18*, 3, 145-156 (1971).
[219] Howard, R.A., Dynamic Programming and Markov Processes (M.I.T. Press, Cambridge, Mass., 1960).
[220] Howard, R.A., "The Foundation of Decision Analysis," IEEE Transactions on Systems, Science and Cybernetics, SSC4, *3*, 212-221 (1968).
[221] Howard, R.R., W.J. Howard, and F.A. Hadden, "Study of Downtime in Military Equipment," 5th National Symposium on Reliability and Quality Control, 402-408 (1959).
[222] Howard, W., "Chain Reliability: A Simple Failure Model for Complex Mechanisms," The RAND Corp. RM1058 (1953).
[223] Hotelling, H., "A General Mathematical Theory of Depreciation," Journal of the American Statistical Association, *20*, 151, 340-353 (1925).
[224] Hsu, J.I.S., "An Empirical Study of Computer Maintenance Policies," Management Science, *15*, 4, B180-B195 (1968).
[225] Hunter, L.C., and F. Proschan, "Replacement when Constant Failure Rate Precedes Wearout," Naval Research Logistics Quarterly, *8*, 2, 127-136 (1961).
[226] Inagaki, T., K. Inoue and H. Akashi, "Improvement of Supervision Schedules for Protective Systems," IEEE Transactions on Reliability, R-28, 2, 141-144 (1979).
[227] Inagaki, T., K. Inoue, and H. Akashi, "Optimal Reliability Allocation under Preventive Maintenance Schedule," IEEE Transactions on Reliability, R-27, *1*, 39-40 (1978).
[228] Ingram, C.R., and R.L. Scheaffer, "On Consistent Estimation of Age Replacement Intervals," Technometrics, *18*, 2, 213-219 (1976).
[229] Intehard, D., and S. Karlin, "Optimal Policy for Dynamic Inventory Process with Non-Stationary Stochastic Demands," in *Studies in Applied Probability and Management Science*, K.J. Arrow, S. Karlin, and H. Scarf, Editors (Stanford Univ. Press, Stanford, Calif., 1962).
[230] Intriligator, M.D., *Mathematical Optimization and Economic Theory*, Prentice Hall (Englewood Cliffs, N.J., 1971).
[231] Jacobson, S.E., and S. Arunkumar, "Investment in Series and Parallel Systems to Maximize Expected Life," Management Science, *19*, 9, 1023-1028 (1973).
[232] Jain, A., and K.P.K. Nair, "Comparison of Replacement Strategies for Items that Fail," IEEE Transactions on Reliability, R-23, *4*, 247-251 (1974).
[233] Jardine, A.K.S., "Determination of Optimal Maintenance Times," The Plant Engineer, *13*, 6, 109-114 (1973).
[234] Jardine, A.K.S., *Maintenance, Replacement and Reliability*, (John Wiley and Sons, New York, N.Y., 1973).
[235] Jardine, A.K.S., "Maintenance and Replacement," in *Handbook of Operations Research*, 2, J.J. Moder and S.E. Elmaghraby, editors, Van Nostrand Reinhold New York, N.Y. (1978).
[236] Jarris, J.E., "A Method for Automating the Visual Inspection of Printed Wiring Boards," IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-2, *1*, 77-82 (1980).
[237] Johnson, E., "Computation and Structure of Optimal Reset Policies," Journal of the American Statistical Association, *62*, 320, 1462-1487 (1967).
[238] Johnson, N., "On Optimizing Maintainability," Microelectronics and Reliability, *17*, 1, 41-46 (1978).

[239] Jorgensen, D.W., and J.J. McCall, "Optimal Replacement Policies for a Ballistic Missile," Management Science, 9, 3, 358-379 (1963).
[240] Jorgenson, D.W., J.J. McCall, and R. Radner, *Optimal Replacement Policy* (The RAND Corp., Santa Monica, Calif., 1967).
[241] Kaio, N., and S. Osaki, "Optimum Inspection-Ordering Policies with Salvage Cost," Microelectronics and Reliability, *18*, 3, 253-257 (1978).
[242] Kaio, N., and S. Osaki, "Optimum Ordering Policies with Lead Time for an Operating Unit in Preventive Maintenance," IEEE Transactions on Reliability, R-27, *4*, 270-271 (1978).
[243] Kaio, N., and S. Osaki, "Optimum Planned Maintenance with Salvage Cost," International Journal of Production Research, *16*, 3, 249-257 (1978).
[244] Kalaba, R., "Optimum Preventive Sampling Via Dynamic Programming," Operations Research, 6, 3, 439-440 (1958).
[245] Kalback, J.F., "Effect of Preventive Maintenance on Reliability," 6th National Symposium on Reliability and Quality Control, 484-488 (1960).
[246] Kalman, P.J., "A Dynamic Nonlinear Constrained Optimal Replacement Model," Naval Research Logistics Quarterly, *23*, 4, 639-651 (1976).
[247] Kalman, P.J., "A Stochastic Constrained Optimal Replacement Model," Naval Research Logistics Quarterly, *17*, 2, 547-553 (1970).
[248] Kalman, P.J., "A Stochastic Constrained Optimal Replacement Model: The Case of Ship Replacement," Operations Research, *20*, 2, 327-334 (1972).
[249] Kalyman, B.A., "Machine Replacement with Stochastic Costs," Management Science, *18*, 5, 288-298 (1972).
[250] Kameyawa, M., and T. Higuchi, "Design of Dependent Failure Tolerant Microcomputer System Using Triple-Modular Redundancy," IEEE Transactions on Computers, C-29, 2, 202-205 (1980).
[251] Kamins, M., "Determining Checkout Intervals for Systems Subject to Random Failure," The RAND Corporation, RM 2578 (1960).
[252] Kamien, M.I., and N.L. Schwartz, "Optimal Maintenance and Sale Age for a Machine Subject to Failure," Management Science, *17*, 8, B495-B504 (1971).
[253] Kander, Z., "Inspection Policies for Deteriorating Equipment Characterized by *N* Quality Levels," Naval Research Logistics Quarterly, *25*, 2, 243-255 (1978).
[254] Kander, Z., and P. Naor, "Optimization of Inspection Policies by Classical Methods," *Proceedings of the 3rd Israel Conference on Operations Research*, 1969 (Gordon and Breach, New York, N.Y., 1970).
[255] Kander, Z., and A. Raviv, "Maintenance Policies when Failure Distribution of Equipment is Only Partially Known," Naval Research Logistics Quarterly, *21*, 3, 419-429 (1974).
[256] Kao, E.P., "Optimal Replacement Rules when Changes of States are Semi-Markovian," Operations Research, *21*, 6, 1231-1249 (1973).
[257] Kaplan, S., "A Note on a Constrained Replacement Model for Ships Subject to Degradation in Utility," Naval Research Logistics Quarterly, *21*, 3, 563-568 (1974).
[258] Kapur, K.C., and L.R. Lamberson, *Reliability in Engineering Design*, (John Wiley and Sons, New York, N.Y., 1977).
[259] Katz, I., "A New Concept of Planned Inspections," Annals of Reliability and Maintainability AIAA, 5, 416-420 (1966).
[260] Kaufmann, A., D. Grouchko, and R. Cruon, Mathematical Models for the Study of the Reliability of Systems (Academic Press, New York, N.Y., 1977).
[261] Kaz, E., "The Effectiveness of Preventive Maintenance," International Journal of Production Research, *14*, 3, 329-344 (1976).
[262] Keller, A.Z., *Uncertainty in Risk and Reliability in Management*, (Adam Hilger Co., London, England 1975).

[263] Keller, J.B., "Optimum Checking Schedules for Systems Subject to Random Failure," Management Science, *21*, 3, 256-260 (1974).

[264] Kent, A., "The Effect of Discounted Cash Flow in Replacement Analysis," Operational Research Quarterly, *21*, 1, 113-118 (1960).

[265] Khalifa, D., M. Hottenstein, and S. Aggarwal, "Cannibalization Policies for Multistage Systems," Operations Research, *25*, 6, 1032-1039 (1977).

[266] Khandelwal, D.N., J. Sharma, and L.M. Ray, "Optimal Periodic Maintenance Policy for Machines Subject to Deterioration and Random Breakdown," IEEE Transactions on Reliability, R-28, *4*, 328-330 (1979).

[267] Khatib, H., "Maintenance Scheduling of Generating Facilities," IEEE Transactions on Power Apparatus and Systems, PAS-98, *5*, 1604-1608 (1979).

[268] Kirkman, R.A., "Failure Concepts in Reliability Theory," IEEE Transactions on Reliability, R-12, *4*, 1-10 (1963).

[269] Kirkman, R.A., "Methods of Predicting Electronic Failures," Annals of Reliability and Maintainability AIAA, *5*, 75-83 (1966).

[270] Kivenson, G., *Durability and Reliability in Engineering Design*, (Hayden Book Co., New York, N.Y., 1971).

[271] Klein, M., "Inspection, Maintenance, Replacement Schedules under Markovian Deterioration," Management Science, *9*, 1, 25-32 (1962).

[272] Kleyle, R., "Maintainability Test Plans Based on Trichotomous Classification of Repair Times," Journal of Quality Technology, *9*, 1, 21-27 (1977).

[273] Kolesar, P., "Minimum Cost Replacement under Markovian Deterioration," Management Science, *12*, 9, 694-706 (1966).

[274] Kolesar, P., "Randomized Replacement Parts which Maximize the Expected Cycle Length of Equipment Subject to Markovian Deterioration," Management Science, *13*, 11, 867-876 (1967).

[275] Kontolem, J.M., "Availability of Networks Subject to Scheduled Maintenance," IEEE Transactions on Reliability, R-28, *1*, 90 (1979).

[276] Kredentser, B.P., "Distribution of the Time to the First Failure of a class of Complex Systems," Cybernetics, *14*, 6, 899-902 (1979).

[277] Krishna, G., K.K. Aggarwal and J.S. Gupta, "A New Approach to Reliability Optimization in General Modular Redundant Systems," Microelectronics and Reliability, *18*, 5, 419-422 (1978).

[278] Krishna, G., K.K. Aggarwal and J.S. Gupta, "Reliability Optimization in Systems with Many Failure Modes," Microelectronics and Reliability, *18*, 5, 423-425 (1978).

[279] Krohn, C.A., "Hazard Versus Renewed Rate of Electronic Items," IEEE Transactions on Reliability, R-18, *2*, 64-73 (1969).

[280] Kulshrestha, D.K., "Reliability with Preventive Maintenance," Metrika, *19*, 2, 216-226 (1972).

[281] Kulshrstha, D.K., "Reliability of a Repairable Multicomponent System with Redundancy in Parallel," IEEE Transactions on Reliability, R-19, *2*, 50-53 (1970).

[282] Kumamato, H., K. Tanaka, K. Inoue and E.J. Henley, "Dagger-Sampling Monte Carlo for System Unavailability Evaluation," IEEE Transactions on Reliability, R-29, *2*, 122-125 (1980).

[283] Lagokos, S.W., "A Covariate Model for Partially Censored Data Subject to Competing Causes of Failure," Applied Statistics, *27*, 3, 235-241 (1978).

[284] Lambe, T.A., "The Decision to Repair or Scrap a Machine," Operational Research Quarterly, *25*, 1, 99-110 (1974).

[285] Lambert, B.K., A.G. Walvekar, and J.P. Hirmas, "Optimal Redundancy and Availability Allocation in Multistage Systems," IEEE Transactions on Reliability, R-20, *3*, 182-185 (1971).

[286] Landers, R.R., *Reliability and Product Assurance*, (Prentice Hall, Englewood Cliffs, N.J., 1963).

[287] Lanzenauer, C.H., and D.D. Wright, "Developing an Optimal Repair Replacement Strategy for Pallets," Naval Research Logistics Quarterly, *25*, 1, 169-178 (1978).

[288] Lewis, P.A., "A Branching Poisson Process Model for the Analysis of Computer Failure Patterns," Journal of the Royal Statistical Society, B26, *3*, 398-456 (1964).

[289] Lewis, P.A., "Implications of a Failure Model for Use and Maintenance of Computers," Journal of Applied Probability, *1*, 2, 347-368 (1964).

[290] Lincoln, T.L., and G.H. Weiss, "A Statistical Evaluation of Recurrent Medical Examination," Operations Research, *12*, 2, 187-205 (1964).

[291] Lloyd, D.K., and M. Lipow, *Reliability: Management, Methods and Mathematics*, (Prentice Hall, Englewood Cliffs, N.J., 1962).

[292] Locks, M.O., "System Reliability Analysis," Microelectronics and Reliability, *18*, 4, 335-345 (1978).

[293] Lomax, K.S., "Business Failures: Another Example of the Analysis of Failure Data," Journal of the American Statistical Association, *49*, 2 to 8, 847-852 (1954).

[294] Lotka, A.J., "A Contribution to the Theory of Self-Renewing Aggregates with Special Reference to Industrial Replacement," Annals of Mathematical Statistics, *10*, 1, 1-25 (1939).

[295] Lu, K.S., and R. Saeks, "Failure Prediction for an On-Line Maintenance System in a Poisson Shock Environment," IEEE Transactions on Systems, Man and Cybernetics, SMC-9, *6*, 356-362 (1979).

[296] Luss, H., "Maintenance Policies when Deterioration Can be Observed by Inspections," Operations Research, *22*, 1, 117-128 (1974).

[297] Luss, H., and Z. Kander, "Inspection Policies when Duration of Checking is Non-Negligible," Operational Research Quarterly, *25*, 2, 299-309 (1974).

[298] Lwin, T., and N. Singh, "Prediction of Future Failures of a System," Microelectronics and Reliability, *15*, 5, 485-488 (1976).

[299] Makabe, H., and H. Morimura, "A New Policy for Preventive Maintenance," Journal of the Operations Research Society of Japan, *4*, 3, 110-124 (1963).

[300] Makabe, H., and H. Morimura, "On Some Preventive Maintenance Policies," Journal of the Operations Research Society of Japan, *6*, 1, 17-47 (1963).

[301] Makabe, H., and H. Morimura, "On Some Preventive Maintenance Policies," Journal of the Operations Research Society of Japan, *6*, 1, 17-47 (1963).

[302] Malik, M.A.K., "Average Life of Equipment Subject to Compound Failures," ASME Journal of Engineering for Industry, *99*, B, 2, 631-633 (1977).

[303] Malik, M.A.K., "Reliable Preventive Maintenance Scheduling," AIIE Transactions, *11*, 3, 221-228 (1979).

[304] Mann, L., *Maintenance Management*, (Lexington Books, D.C. Heath and Co., Lexington, Mass. 1976).

[305] Mann, N., R. Schaefer, and N. Singpurwalla, *Methods for Statistical Analysis of Reliability and Life Data*, (John Wiley and Sons, New York, N.Y. 1974).

[306] Manne, A.S., "Capacity Expansion and Probabilistic Growth," Econometrica, *29*, 4, 631-649 (1961).

[307] Manne, A.S., and A. Veinott, "Optimal Plant Size with Arbitrary Increasing Time Paths of Demand," in *Investment for Capacity Expansion, Size, Location, and Time-Phasing*, A.S. Manne, Editor (The M.I.T. Press Cambridge, Mass. 1967).

[308] Marshall, A.W., and F. Proschan, "Classes of Distributions Applicable in Replacement with Renewal Theory Implications," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 395-415 (1972).

[309] Marshall, A.W., and M. Shaked, Multivariate Shock Models for Distributions with Increasing Hazard Rate Average," The Annals of Probability, *7* 2, 343-358 (1979).

[310] Martz, H.F., Jr., "On Single Cycle Availability," IEEE Trans. Reliability, R-20, *1*, 21-23 (1971).

[311] Masse, P., *Optimal Investment Decisions*, (Prentice-Hall Englewood Cliffs, N.J. 1962).
[312] Masso, J., and M.L. Smith, "Interstage Storages for Three Stage Lines Subject to Stochastic Failures," AIIE Transactions, *6*, 4, 354-358 (1974).
[313] Matveyev, A.V., "Estimation of the Probability of Parametric Failure," Engineering Cybernetics, *16*, 4, 76-82 (1978).
[314] Mazunder, M., "Some Estimates of Reliability Using Interference Theory," Naval Research Logistics Quarterly, *17*, 2, 159-165 (1970).
[315] McCall, J.J., "Maintenance Policies for Stochastically Failing Equipment: A Survey," Management Science, *11*, 5, 493-624 (1965).
[316] McCall J.J., "Operating Characteristics of Opportunistic Replacement and Inspection Policies," Management Science, *10*, 1, 85-97 (1963).
[317] McGregor, M.A., "Approximation Formulas for Reliability with Repair," IEEE Transactions on Reliability, R-12, *4*, 64-92 (1963).
[318] Meisel, W.S., "On-Line Optimization of Maintenance and Verification Schedules," IEEE Transactions on Reliability, R-18, *4*, 200-201 (1969).
[319] Menipaz, E., "Cost Optimization of Some Stochastic Maintenance Policies," IEEE Transaction on Reliability, R-28 *2*, 133-136 (1979).
[320] Mercer, A., "Some Simple Wear-Dependent Renewal Processes," Journal of the Royal Statistical Society, Ser. B, *23*, 2, 268-276 (1961).
[321] Meyer, P.L., *Introductory Probability and Statistical Applications*, (Addison-Wesley, Reading, Mass. 1966).
[322] Meyer, R.A., "Equipment Replacement under Uncertainty," Management Science, *17*, 11, 750-758 (1971).
[323] Meyers, R., and R.S. Dick, "Some Considerations of Scheduled Maintenance," *8th National Symposium on Reliability and Quality Control* 343-356 (I.R.E., New York, N.Y. 1962).
[324] Mine, H., and H. Kawai, "An Optimal Inspection and Replacement Policy," IEEE Transactions on Reliability, R-24 *3*, 3-5-309 (1975).
[325] Mine, H. and H. Kawai, "An Optimal Maintenance Policy for a 2-Unit Parallel System with Degraded State," IEEE Transactions on Reliability, R-23, *2*, 81-86 (1976).
[326] Mine, H., and H. Kawai, "Preventive Maintenance of a 1-Unit System with a Wearout State," IEEE Transactions on Reliability, R-23, *1*, 24-29 (1976a).
[327] Mine, H., and T. Nakagawa, "A Summary of Optimum Preventive Maintenance Policies Maximizing Interval Reliability," Journal of the Operations Research Society of Japan, 21, *2*, 205-216 (1978).
[328] Mine, H., and T. Nakagawa, "Age Replacement Model with Mixed Failure Times," IEEE Transactions on Reliability, R-27, *2*, 173 (1978).
[329] Mine, H., and T. Nakagawa, "Interval Reliability and Optimum Preventive Maintenance Policy," IEEE Transactions Reliability, R-26, *2*, 131-133 (1977).
[330] Mine, H., and S. Osaki, "On Failure Time Distributions for Systems of Dissimilar Units," IEEE Transactions on Reliability, R-18, *4*, 165-168 (1969).
[331] Moore, J.R., Jr., "Forecasting and Scheduling for Past-Model Replacement Parts," Management Science, *18*, B200-B213 (1971).
[332] Morey, R.C., "Some Stochastic Properties of a Compound Renewal Damage Model," Operations Research, *14*, 5, 902-908 (1966).
[333] Morgan, D.E., and D.J. Taylor "A Survey of Methods of Achieving Reliable Software," Computer, *10*, 2, 44-53 (1977).
[334] Morimura, H. "On Some Preventive Maintenance Policies for IFR," Journal of the Operations Research Society of Japan, *12*, 3, 94-124 (1970).
[335] Morse, P.C., *Queues, Inventories and Maintenance*, (John Wiley and Sons, New York, N.Y. 1958).
[336] Moses, M., "Dispatching and Allocating Servers to Stochastically Failing Networks," Management Science, *18*, 6, B289-B300 (1972).

[337] Munford, A.G., and A.K. Shahani, "A Nearly Optimal Inspection Policy," Operational Research Quarterly, *23*, 3, 373-379 (1972).

[338] Munford, A.G., and A.K. Shahani, "An Inspection Policy, for the Weibull Case," Operational Research Quarterly, *24*, 3, 453-488 (1973).

[339] Murphy, R.A., "The Effect of Surge on System Availability," AIIE Transactions, *7*, 4, 439-443 (1975).

[340] Musa, J.D., "A Theory of Software Reliability and Its Applications," IEEE Transactions on Software Engineering, SE-1, *3*, 312-327 (1975).

[341] Muth, E.J., "A Method for Predicting System Downtime," IEEE Transactions on Reliability, R-17 *2*, 97-102 (1968).

[342] Myhre, J.M., A.M. Rosenfield, and S.C. Saunders, "Determining Confidence Bounds for Highly Reliable Coherent Systems Based on a Paucity of Compound Failure," Naval Research Logistics Quarterly, *25*, 2, 213-227 (1978).

[343] Nakagawa, T., "Reliability Analysis of Standby Repairable Systems when an Emergency Occurs," Microelectronics and Reliability, *17*, 4, 461-464 (1978).

[344] Nakagawa, T., "Optimum Preventive Maintenance Policies for Repairable Systems," IEEE Transactions on Reliability, R-26, *3*, 168-173 (1977).

[345] Nakagawa, T., "Optimum Policies When Preventive Maintenance is Imperfect," IEEE Transactions on Reliability, R-28, *4*, 331-332 (1979).

[346] Nakagawa, T., "The Expected Number of Visits to State K before a Total System Failure of a Complex System with Repair Maintenance," Operations Research, *22*, 1, 108-111 (1976).

[347] Nakagawa, T., and S. Osaki, "Markov Renewal Process with Some Non-Regeneration Points and Their Applications to Reliability Theory," Microelectronics and Reliability, *15*, 6, 633-636 (1976).

[348] Nakagawa, T., and S. Osaki, "Optimum Preventive Maintenance Policies for a 2-Unit Redundant System," IEEE Transactions on Reliability, R-23, 86-91, (1974).

[349] Nakagawa, T., and S. Osaki, "Optimum Preventive Maintenance Policies Maximizing the Mean Time to the First System Failure for a Two-Unit Standby Redundant System," Journal of Optimization Theory and Applications, *14*, 115-129 (July 1974).

[350] Nakagawa, T., and S. Osaki, "Some Aspects of Damage Models," Microelectronics and Reliability *13*, 4, 253-257 (1974).

[351] Nakagawa, T., and S. Osaki, "The Optimal Repair Limit Replacement Policies," Operational Research Quarterly, *25*, 2, 311-317 (1974).

[352] Nakagawa, T., Y. Sawa, and Y. Suzuki, "Reliability Analysis of Intermittently Used Systems when Failures are Detected Only During a Usage Period," Microelectronics and Reliability, *15*, 1, 35-38 (1976).

[353] Nakagawa, T., and K. Yasui, "Approximate Calculation of Block Replacement with Weibull Failure Times," IEEE Transactions on Reliability, R-27, *4*, 268-269 (1978).

[354] Nakamichi, H., J. Fukata, S. Takamatsu, and M. Kodoma, "Reliability Consideration on a Repairable Multicomponent System with Redundancy in Parallel," Journal of the Operations Research Society of Japan, *17*, 1, 39-48 (1974).

[355] Nair, K.P.K., and V.P. Marathe, "Multistage and Replacement Strategies," Operations Research, *14*, 5, 874-887 (1966b).

[356] Nair, K.P.K, and V.P. Marathe, "On Multistage Replacement Strategies," Operations Research *14*, 3, 537-539 (1966a).

[357] Nair, K.P. K., and M.D. Naik, "Multistage Replacement Strategies," Operations Research, *13*, 2, 279-290 (1965a).

[358] Nair, K.P.K., and M.D. Naik, "Multistage Replacement Strategies with Finite Duration of Transfer," Operations Research, *13*, 5, 828-835 (1965b).

[359] Nathan, I., "Modified Long Term System Reliabilty Models for Maintained Systems Considering Imperfect Failure Detection, Repair and Sparing," *Annual Symposium on Reliability*, 654-669 (1966).

[360] Nemhauser, G.L., *Introduction to Dynamic Programming*, (John Wiley and Sons, New York, N.Y. 1967).

[361] Newman, C.P., and N.M. Bonhomme, "Evaluation of Maintenance Policies Using Markov Chains and Fault Tree Analysis," IEEE Transactions on Reliability, R-24, *1*, 37-45 (1975).

[362] Newbrough, E.T., *Effective Maintenance Management*, (McGraw-Hill New York, N.Y. 1967).

[363] Nikitin, N.N. and V.D. Razevig, "Evaluation of the Transfer Coefficients of a Markov Process Using the Stochastic Equation of a Dynamic System," Automation and Remote Control, *37*, 4, 512-520 (1976).

[364] Noonan, G.C., and C.G. Fain, "Optimal Preventive Maintenance Policies when Immediate Detection of Failure is Uncertain," Operations Research, *10*, 3, 407-410 (1962).

[365] Ohashi, M. and T. Nishida, "Optimum Preventive Maintenance Policy for a 1-Unit System," IEEE Transactions on Reliability, R-29, *2*, 174-175 (1980).

[366] Okumoto, K., and S. Osaki, "Optimum Policies for a Standby System with Preventive Maintenance," Journal of the Operational Research Society, *28*, 415-523 (1977).

[367] Onaga, K., "Maintenance and Operating Characteristics of Communications Networks," Operations Research, *17*, 2, 311-336 (1969).

[368] Osaki, S., "An Intermittently Used System with Preventive Maintenance," Journal of the Operations Research Society of Japan, *15*, 2, 102-111 (1972).

[369] Osaki, S., and T. Nakagawa, "A Note on Age Replacement," IEEE Transactions on Reliability, R-24, *1*, 92-94 (1975).

[370] Osaki, S., and A. Sakura, "A Two-Unit Standby Redundant System with Repair and Preventive Maintenance," Journal of Applied Probability, *7*, 3, 641-648 (1970).

[371] Ozan, T.M., C.T. Ng, and O. Saateioglu, "*Application of Search Theory to Maintenance* and Inspection Problems," International Journal of Production Research, *14*, 1, 85-90 (1976).

[372] Pandit, S.M., "Data Dependent Systems Approach to Stochastic Tool Life Reliability, ASME Journal of Engineering for Industry, *100*, 3, 318-322 (1978).

[373] Papadopoulos, A.S., "The Burr Distribution as a Failure Model from a Bayesian Approach," IEEE Transactions on Reliability, R-27, *5*, 369-371 (1978).

[374] Park, K.S., "Gamma Approximation for Preventive Maintenance Scheduling," AIIE Transactions, *7*, 4, 393-397 (1975).

[375] Park, K.S., "Optimal Number of Minimal Repairs before Replacement," IEEE Transactions on Reliability, R-28, *2*, 137-140 (1979).

[376] Pashkovskiy, G.S., "Hierarchical Models of Experiments in Bayesian Theory of Decision Making Generalization of Problems of Search for Defects," Engineering Cybernetics, *15*, 6, 72-79 (1977).

[377] Pashkovskiy, G.S., "Methods of Optimizing Programs of Successive Search for Defects," Engineering Cybernetics, *9*, 2 (1971).

[378] Peterson, E.L., "Maintainability Application to System Effectiveness Quantification," IEEE Transactions on Reliability, R-20, *1*, 3-7 (1971).

[379] Peterson, E.L., and H.B. Loo, "Maintainability Derivations Using the Analytical Maintenance Model," IEEE Transactions on Reliability, R-17, *2*, 111-114 (1968).

[380] Peterson, C.L., and J.H. Milligan, "Economic Life Analysis for Machinery Replacement Decisions," Transactions of the ASAE, *19*, 5, 819-826 (1976).

[381] Phoenix, S.L., "The Asymptotic Distribution for the Time to Failure of a Fiber Bundle," Advances in Applied Probability, *11*, 1, 153-187 (1979).

[382] Phoenix, S.L., "The Asymptotic Time to Failure of A Mechanical System of Parallel Members," SIAM Journal on Applied Mathematics, *34*, 227-246 (1978).

[383] Pierre, D.A., and M.J. Lowe, *Mathematical Programming via Augmented Lagrangians: An Introduction with Computer Programs*, (Addison-Wesley, Reading, Mass., 1975).

[384] Pierskalla, W.P., and J.A. Voelker, "A Survey of Maintenance Models: The Control and Surveillance of Deteriorating Systems," Naval Research Logistics Quarterly, *23*, 3, 353-388 (1976).

[385] Pieruschka, E., *Principles of Reliability*, (Prentice Hall Englewood Cliffs, N.J. 1963).

[386] Platz, O., "Availability of a Renewable Checked System," IEEE Transactions on Reliability, R-25, *1*, 56-58 (1976).

[387] Plesser, K., and J. Field, "Cost Optimized Burn-in for Repairable Electronic Systems," IEEE Transactions on Reliability, R-26, *3*, 195-197 (1977).

[388] Polovko, A.M., *Fundamentals of Reliability Theory*, (Academic Press, New York, N.Y. 1968).

[389] Pontryagin, L.S., V.G. Boltyanskii, R.V. Gamkrelidze, and E.F. Mischenko, *The Mathematical Theory of Optimal Processes*, (Interscience New York, N.Y. 1962).

[390] Port, S.C., "Optimal Procedure for the Installation of a Unit Subject to Stochastic Failures," Journal of Mathematical Analysis and Applications, 9, 3, 491-497 (1964).

[391] Port, S.C., and J. Falkman, "Optimal Procedures for Stochastically Failing Equipment," Journal of Applied Probability, *3*, 2, 521-537 (1966).

[392] Porteus, E.L., and Z.F. Lansdowne, "Optimal Design of a Multi-Item, Multi-Location Multi-Repair Type Repair and Supply System," Naval Research Logistics Quarterly, *21*, 2, 213-237 (1974).

[393] Prakesh, C., "Analysis of Non-Catastrophic Failures in Electronic Devices Due to Random Noise," Microelectronics and Reliability, *16*, 5, 581-588 (1977).

[394] Preinreich, G.A.D., "The Economic Life of Industrial Equipment," Econometrica, *8*, 1, 12-44 (1940).

[395] Priel, V.A., *Systematic Maintenance Organization*, (Macdonald and Evans London, England 1974).

[396] Pritsker, A.B., "The Setting of Maintenance Tolerance Limits," Journal of Industrial Engineering, *14*, 2, 80-86 (1963).

[397] Pritsker, A.B. "The Monte-Carlo Approach to Setting Maintenance Tolerance Limits," Journal of Industrial Engineering, *14*, 3, 115-118 (1963).

[398] Proctor, C.L., and Y.T. Wang, "Optimal Maintenance Policy for Systems that Experience State Degradations," Microelectronics and Reliability, *14*, 2, 199-202 (1975).

[399] Proschan, F., "Optimal System Supply," Naval Research Logistics Quarterly, *7*, 2, 609-646 (1960).

[400] Proschan, F., "Theoretical Explanation of Observed Decreasing Failure Rate," Technometrics, *5*, 3, 375-383 (1963).

[401] Quayle, N.J.T., "Damaged Vehicles—Repair or Replace," Operational Research Quarterly, *23*, 1, 83-87 (1972).

[402] Radner, R., and D.W. Jorgenson, "Opportunistic Replacement of a Single Part in the Presence of Several Monitored Parts," Management Science, *10*, 1, 70-83 (1963).

[403] Radner, R., and D.W. Jorgenson, "Optimal Replacement and Inspection of Stochastically Failing Equipment," in K.J. Arrow, S. Karlin, and H. Scard (eds.), *Studies in Applied Probability and Management Science*, (Stanford University Press, Stanford, Calif., 1962).

[404] Raiffa,f H., and R. Schlaifer, *Applied Statistical Decision Theory*, (Harvard University Cambridge Mass. 1961).

[405] Ramalingam, S., "Tool-Life Distributions, Part 2: Multiple-Injury Tool-Life Model," ASME Journal of Engineering for Industry, *99*, B, 3, 523-531 (1977).

[406] Ramalingam, S., and J.D. Watson, "Tool-Life Distributions, Part 1: Single-Injury Tool-Life Model," ASME Journal of Engineering for Industry, *99*, B, 3, 519-522 (1977).

[407] Ramalingam, S., and J.D. Watson, "Tool Life Distributions, Part 4: Minor Phases in Work Material and Multiple-Injury Tool Failure," ASME Journal of Engineering for Industry, *100*, 2, 201-209 (1978).

[408] Ramalingam, S., Y.L. Peng, and J.D. Watson, "Tool Life Distributors, Part 3: Mechanism of Single Injury Tool Failure and Tool Life Distribution in Interrupted Cutting," ASME Journal of Engineering for Industry, *100*, 2, 193-200 (1978).

[409] Ran, A., and S.I. Rosenland, "Age Replacement with Discounting for a Continuous Maintenance Cost Model," Technometrics, *18*, 4, 459-465 (1976).

[410] Rao, S.S., and C.P. Reddy, "Reliability Analysis of Machine Tool Structures," ASME Journal of Engineering for Industry, *99*, B, 4, 882-888 (1977).

[411] Raykin, A.L., A.F. Rubtsov, and V.S. Penin, "Reliability of Systems with Regularly Renewed Replacements," Engineering Cybernetics, *4*, 9 (1964).

[412] Reinitz, R.C., and L. Karasyk, "A Stochastic Model for Planning Maintenance of Multi-Part Systems," *Proceedings of the Fifth International Conference on Operations Research*, Tavistock Publications, London, England, 703-713 (1969).

[413] Reynolds, G.H., "A Shuttle Car Assignment Problem in the Mining Industry," Management Science, *17*, 9, 652-655 (1971).

[414] Riddick, R., "The Effect of Scheduled Repair Cycle on Marine Equipment Reliability," Annual Reliability and Maintainability Symposium IEEE, (1967).

[415] Roberts, N.H., *Mathematical Methods in Reliability Engineering*, (McGraw Hill New York, N.Y. 1964).

[416] Robertson, J.L., "A Method of Computing Survival Probabilities of Several Targets Versus Several Weapons," Operations Research, *4*, 5, 546-557 (1956).

[417] Roeloffs, R., "Minimax Surveillance Schedules for Replacement Units," Naval Research Logistics Quarterly, *14*, 4, 461-471 (1967).

[418] Roeloffs, R., "Minimax Surveillance Schedules with Partial Information," Naval Research Logistics Quarterly, *10*, 4, 307-322 (1963).

[419] Rohal-Ilkiv, B., "Availability of A Redundant System with Replacement and Repair," IEEE Transactions Reliability, R-28, *1*, 83 (1979).

[420] Rolfe, A.J., "Markov Chain Analysis of a Situation where Cannibalization is the Only Repair Activity," Naval Research Logistics Quarterly, *17*, 2, 151-158 (1970).

[421] Roll, Y., and P. Noar, "Preventive Maintenance of Equipment Subject to Continuous Deterioration and Stochastic Failure," Operational Research Quarterly, *19*, 1, 61-73 (1968).

[422] Rose, M., "Determination of the Optimal Investment in End Products and Repair Resources," Naval Research Logistics Quarterly, *20*, 1, 147-159 (1973).

[423] *Rosenfield*, D., "Markovian Deterioration with Uncertain Information," Operations Research, *24*, 1, 141-155 (1976).

[424] Ross, S.M., *Applied Probability Models with Optimization Applications*, (Holden-Day, San Francisco, Calif. 1970).

[425] Ross, S.M., "A Markovian Replacement Model with A Generalization to Include Stocking," Management Science, *15*, 11, 702-715 (1969a).

[426] Ross, S.M., "Average Cost Semi-Markov Decision Processes," Journal of Applied Probability, *7*, 3, 649-656 (1970).

[427] Rozhdestvenskiy, D.V., and G.N. Fanarzhi, "Reliability of a Duplicated System with Renewal and Preventive Maintenance," Engineering Cybernetics, *8*, 3, 475-479 (1970).

[428] Sackrowitz, H., and E. Samel-Cahn, "Inspection Procedures for Markov Chains," Management Science, *21*, 3, 261-270 (1974).

[429] Saisieni, M.W., "A Markov Chain Process in Industrial Replacement," Operational Research Quarterly, *7*, 4, 148-154 (1956).

[430] Salter, R.G., "Geriometry: A Cumulative Stress-Damage Measurement of Useful Life of Mechanical Systems," The RAND Corporation, P-5962 (1978).

[431] Sandler, G.H., *System Reliability Engineering*, (Prentice-Hall Englewood Cliffs, N.J. 1963).

[432] Sarma, V.V.S., and M. Alam, "Optimal Maintenance Policies for Machines Subject to Deterioration and Intermittent Breakdowns," IEEE Transactions on Systems, Man and Cybernetics, SMC-5, *3*, 396-398 (1975).

[433] Sathi, P.T., and W.M. Hancock, "A Bayesian Approach to the Scheduling of Preventive Maintenance," AIIE Transactions, *5*, 172-179 (1973).

[434] Savage, I.R., "Cycling," Naval Research Logistics Quarterly, *3*, 3, 163-175 (1956).

[435] Savage, I.R., "Surveillance Problem," Naval Research Logistics Quarterly, *9*, 3, 187-209 (1962).

[436] Savage, I.R., "Surveillance Problems: Poisson Models with Noise," Navel Research Logistics Quarterly, *11*, 1, 1-13 (1964).

[437] Scheaffer, R.L., "Optimum Age Replacement Policies with an Increasing Cost Factor," Technometrics, *13*, 1, 139-144 (1971).

[438] Schiff, A.J., R.E. Torres-Cabrejos and J.T.P. Yao, "Evaluating The Seismic Reliability of Electrical Equipment Containing Ceramic Structural Members," Earthquake Engineering and Structural Dynamics, *7*, 1, 85-98 (1979).

[439] Schlaifer, R., *Probability and Statistics for Business Decisions*, (McGraw-Hill New York, 1959).

[440] Schrady, D.A., "A Deterministic Inventory Model for Repairable Items," Naval Research Logistics Quarterly, *14*, 3, 391-398 (1967).

[441] Scott, M., "Distributions of the Number of Tasks by A Repairable Machine," Operations Research, *20*, 851-859 (1972).

[442] Schwartz, A.N., J.A. Sheler and C.R. Cooper, "Dynamic Programming Approach to the Optimization of Naval Aircraft Rework and Replacement Policies," Naval Research Logistics Quarterly, *18*, 3, 395-414 (1971).

[443] Schweitzer, P., "Optimal Replacement Policies for Hyper-Exponentially and Uniformly Distributed Lifetimes," Operations Research, *15*, 2, 360-362 (1967).

[444] Sculli, D., and A.W. Suraweera, "Tramcar Maintenance," Journal of the Operational Research Society, *30*, 9, 809-814 (1979).

[445] Senju, S., "A Probabilistic Approach to Preventive Maintenance," Journal of the Operations Research Society of Japan, *1*, 2, 49-58 (1957).

[446] Serfozo, R., "A Replacement Problem Using a World Identity for Discounted Variables," Management Science, *20*, 9, 1314-1315 (1974).

[447] Sethi, S.P., "Simultaneous Optimization of Preventive Maintenance and Replacement Policy for Machines: A Modern Control Theory Approach," AIIE Transactions, *5*, 2, 156-163 (1973).

[448] Shedletsky, J.J., and E.J. McCluskey, "The Error Latency of a Fault in a Sequential Digital Circuit," IEEE Transactions on Computers, C-25, *6*, 635-659 (1976).

[449] Shelley, B.F. "Maintenance Manhour Distributions," Annual Symposium on Reliability, 704-711 (1966).

[450] Sherwin, D.J., "Hyper Exponentially Distributed Failures to Process Plant," *Proceedings of 5th Symposium on Reliability Technology*, University of Bradford, U.K. (1978).

[451] Sherwin, D.J., "Inspection Intervals for Condition-Maintained Items which Fail in An Obvious Manner," IEEE Transactions on Reliability, R-28, *1*, 85-89 (1979).

[452] Singh, C., "A Matrix Approach to Calculate The Failure Frequency And Related Indices," Microelectronics and Reliability, *19*, 4, 395-398 (1979).

[453] Shooman, M.L., *Probabilistic Reliability: An Engineering Approach*, (McGraw-Hill, New York, N.Y., 1968).

[454] Shultis, J.K., and N.D. Eckhoff, "Selection of Beta Prior Distribution Parameters from on Component Failure Data," IEEE Transaction on Power Apparatus and Systems, PAS-98, *2*, 400-407 (1979).

[455] Silver, E.A., "Inventory Allocation Among an Assembly and its Repairable Subassemblies," Naval Research Logistics Quarterly, *19*, 2, 261-280 (1972).

[456] Silver, E.A., "Inventory Control under a Probabilistic Time-Varying Demand Pattern," AIIE Transactions, *4*, 371-379 (1978).

[457] Simpson, V.P., "Optimum Solution Structure for a Repairable Inventory Problem," Operations Research, *26*, 2, 271-281 (1978).

[458] Singh, C., "On the Behavior of Failure Frequency Bounds," IEEE Transactions on Reliability, R-26, *1*, 63-66 (1977).

[459] Singh, C., "Tie Set Approach to Determine the Frequency of System Failure," Microelectronics and Reliability, *14*, 3, 292-294 (1975).

[460] Singh, C., and R. Billington, "A New Method to Determine the Failure Frequency of a Complex System," IEEE Transactions on Reliability, R-23, *4*, 231-234 (1974).

[461] Singpurwalla, N.D., "Estimating Reliability Growth Using Time Series Analysis," Naval Research Logistics Quarterly, *25*, 1, 1-14 (1978).

[462] Sivazlian, B.D., "On a Discounted Replacement Problem with Arbitrary Repair Time Distribution," Management Science, *19*, 11, 1301-1309 (1973).

[463] Sivazlian, B.D., and J.F. Mahoney, "Group Replacement of a Multi Component System which is Subject to Deterioration Only," Advances in Applied Probability, *10*, 4, 867-885 (1978).

[464] Skakela, J., and b. Robal-Ilkiv, "2-Unit Redundant Systems with Replacement and Repair," IEEE Transactions on Reliability, R-28, *4*, 294-296 (1977).

[465] Smallwood, R., and E. Sandik, "The Optimal Control of Partially Observable Markov Processes over a Finite Horizon," Operations Research *21*, 4, 1071-1088 (1973).

[466] Soland, R.M., "A Renewal Theoretic Approach to the Estimation of Future Demand for Replacement Parts," Operations Research, *16*, 1, 36-51 (1968).

[467] Soland, R.M., "Bayesian Analysis of the Weibull Process with Unknown Scale Parameter and its Application to Acceptance Sampling," IEEE Transactions on Reliability, R-17, *2*, 84-90 (1968).

[468] Srinivasan, V.S., "The Effect of Standby Redundancy in System's Failure with Repair Maintenance," Operations Research, *14*, 6, 1024-1036 (1966).

[469] Srivastara, A.K., and G.E. Rehkugler, "Strain Rate Effects in Similitude Modelling of Plastic Deformation of Structures Subject to Impact Loading," Transactions of the ASAE, *19*, 4, 617-621 (1976).

[470] Stevenson, T.E., and R.J. McNichols, "Maintenance Float for Small Fleet Sizes," *IEEE Proceedings Reliability and Maintainability Symposium Philadelphia, Penn.*, 246-249 (1973).

[471] Staller, D.S., "A Failure Model for Equipments Undergoing Complex Operation," Operations Research, *6*, 5, 723-728 (1958).

[472] Sule, R. and B. Harmon, "Determination of Coordinated Maintenance Scheduling Frequencies for a Group of Machines," AIIE Transactions, *11*, 1, 48-53 (1979).

[473] Swell, B.H., "Reliability Growth Management," The Journal of Environmental Sciences, *22*, 1, 9-12 (1979).

[474] Subramanian, R., and M.R. Wolff, "Age Replacement in Simple Systems with Increasing Loss Function," IEEE Transactions on Reliability, R-25, *1*, 32-34 (1976).

[475] Tadikamalla, P.R., "An Inspection Policy for the Gamma Failure Distribution," Journal of Operational Research Society, *30*, 1, 77 80 (1979).

[476] Tahara, A., and T. Nishida, "Optimal Replacement Policy for Minimal Repair Model," Journal of the Operations Research Society of Japan, *18*, 3, 113-124 (1975).

[477] Taylor, H.M., "Optimal Replacement under Additive Damage and Other Failure Models," Naval Research Logistics Quarterly, *22*, 1, 1-18 (1975).

[478] Taylor, H.M., "Optimal Stopping in a Markov Process," Annals of Mathematical Statistics, *39*, 4, 1333-1344 (1968).

[479] Taylor, J., and R.R.P. Jackson, "Application of Birth and Death Processes to Provisions to Spare Machines," Operational Research Quarterly, *5*, 4, 95-108 (1954).

[480] Terborgh, G., *Dynamic Replacement Policy*, (McGraw-Hill New York, N.Y., 1949).

[481] Thomas, M.U., "On Minimizing the Mean Detection Time to Failures Subject to Detection Error," Journal of Quality Technology, *7*, 2, 59-61.

[482] Thompson, G.L., "Optimal Maintenance Policies and Sale Date of a Machine," Management Science, *14*, 9, 543-550 (1968).

[483] Thorburn, D., "An Inventory Depletion with Random and Age-Dependent Lifetimes," Navel Research Logistics Quarterly, *25*, 3, 395-404 (1978).

[484] Tilquin, C., and R. Cleroux, "Block Replacement Policies with General Cost Structures," Technometrics, *17*, 3, 291-298 (1975).

[485] Tilquin, C., and R. Cleroux, "Periodic Replacement with Minimal Repair at Failure and Adjustment Costs," Naval Research Logistics Quarterly, *22*, 2, 243-254 (1975).

[486] Tosch, T.J., and P.T. Holmes, "A Bivariate Failure Model," Journal of the American Statistical Association, *75*, 370, 415-417 (1980).

[487] Turban, E., "The Use of Mathematical Models in Plant Maintenance Decision Making," Management Science, *13*, 6, B342-B358 (1967).

[488] Veinott, A.F., Jr., "Optimal Policy for a Multi-Product, Dynamic, Non-stationary Inventory Problem," Management Science, *12*, 3, 206-222 (1965).

[489] Veinott, A.F., "Discrete Dynamic Programming with Sensitive Discount Optimality Criteria," Annals of Mathematical Statistics, *40*, 1635-1660 (1969).

[490] Veinott, A.F., Jr., and H.M. Wagner, "Optimal Capacity Scheduling—I, II," Operations Research, *10*, 4, 518-532 (1962).

[491] Vered, G., and U. Yechiali, "Optimal Structures and Maintenance Policies for PABX Power Systems," Operations Research, *27*, 1, 37-47 (1979).

[492] Vergin, R.C., "Optimal Renewal Policies for Complex Systems," Naval Research Logistics Quarterly, *15*, 4, 523-534 (1968).

[493] Vergin, R.C., "Scheduling Maintenance and Determining Crew Size for Stochastically Failing Equipment," Management Science, *13*, 2, B52-B65 (1966).

[494] Vesely, W.E., "The Evaluation of Failure and Failure Related Data," Proceedings, 1973 Annual Reliability and Maintainability Symposium, IEEE, New York, N.Y., 500-506 (1973).

[495] Wang, R.C., "Replacement Policy with Unobservable States," Journal of Applied Probability, *14*, 2, 340-348 (1977).

[496] Wagner, H.M., and G. Glaser, "Preventive Maintenance," Tech. Report Number 1, NSF 24064, Applied Mathematics and Statistics Lab., Stanford University, Stanford, Calif. (1963).

[497] Wagner, H.M., R.J. Giblio, and R.G. Glaser, "Preventive Maintenance Scheduling by Mathematical Programming," Management Science, *10*, 2, 315-334 (1964).

[498] Wattanapanom, N. and L. Shaw, "Optimal Inspection Schedules for Failure Detection in a Model Where Tests Hasten Failure," Operations Research, *27*, 2, 303-317 (1979).

[499] Weir, K., "Analysis of Maintenance Man Loading via Simulation," IEEE Transactions on Reliability, R-20, *3*, 164-169 (1971).

[500] Weiss, H.K., "Estimation of Reliability Growth in a Complex System with a Poisson Type Failure," Operations Research, *4*, 5, 532-545 (1956).

[501] Weiss, G.H., "On the Theory of Replacement of Machinery with Random Failure Time," Naval Research Logistics Quarterly, *3*, 4, 279-293 (1956).

[502] Welker, E.L., "Relationship between Equipment Reliability, Preventive Maintenance Policy, and Operating Costs," *Proc. of the Fifth Annual Symp. on Reliability and Quality in Electronics, IRE New York, N.Y.*, 270-280 (1959).

[503] White, C., III, "A Markov Quality Control Process Subject to Partial Observation," Management Science, *23*, 8, 843-852 (1977).

[504] White, C.C., III, "Optimal Inspection and Repair of a Production Process Subject to Deterioration," Journal of the Operations Research Society, *29*, 3, 235-243 (1978).

[505] White, E.N., *Maintenance Planning, Control and Documentation*, (Gower Press, Essex England: 1973).

[506] Widawsky, W.H., "Reliability and Maintainability Parameters Evaluated with Simulation," IEEE Transactions on Reliability, R-20, *3*, 158-164 (1971).

[507] Williams, P.B., "Taking Another Look at Electrical System Reliability," Public Utilities Fortnightly, *99*, 6, 23-26 (1977).

[508] Wilson, M.A., "The Learning Curve in Maintenance Analysis," Annals. of Reliability and Maintainability AIAA, *5*, 434-439 (1966).

[509] Winston, W., "Optimal Control of Discrete and Continuous Time Maintenance Systems with Variable Service Rates," Operations Research, *25*, 2, 259-268 (1977).

[510] Winter, B.B., "Introduction to Cyclic Replacement Systems," IEEE Transactions on Reliability, R-12, 36-49 (1963).

[511] Wohl, J.G., "System Operational Readiness and Equipment Dependability," IEEE Transactions on Reliability, R-15, *1*, 1-6 (1966).

[512] Wolf, M.R. and R. Subramanian, "Optimal Re-adjustment Intervals," Operations Research, *22*, 1, 191-197 (1974).

[513] Woodman, R.C., "Replacement Policies for Components that Deteriorate," Operational Research Quarterly, *18*, 3, 267-281 (1967).

[514] Woodman, R.C., "Replacement Rules for Single and Multi-Component Equipment," Journal of the Royal Statistical Society, Series C, *18*, 1, 31-40 (1969).

[515] Wright, L.W., "An Overview of Electronic Part Failure Analysis Experience," IEEE Transactions on Reliability, R-17, *1*, 5-9 (1968).

[516] Yaakov, R., and A. Sachish, "Combined Overhaul and Replacement Policies for Deteriorating Equipment," Journal of the Operations Research Society of Japan, *21*, *2*, 274-286 (1978).

[517] Yadav, R.P.S., "A Reliability Model for Stress Versus Strength Problem," Microelectronics and Reliability, *12*, 119-123 (1973).

[518] Yamada, S., "Optimum Number of Checks in Checking Policy," Microelectronics and Reliability, *16*, 4, 589-591 (1977).

[519] Zachs, S., and W. Fenske, "Sequential Determination of Inspection Epochs for Reliability Systems with General Lifetime Distributions," Naval Research Logistics Quarterly, *20*, 3, 377-386 (1973).

[520] Zelen, M., (ed.) *Statistical Theory of Reliability*, (University of Wisconsin Press, Madison, Wisconsin, 1964).

[521] Zhuk, P.I., and E.I. Litrak, "Reliability Parameters for a System with Mobile Reserves and Restricted Repair," Cybernetics, *14*, 6, 894-898 (1979).

[522] Ziehms, H., "Approximations to the Reliability of Phased Missions," Naval Research Logistics Quarterly, *25*, 2, 229-242 (1978).

[523] Zampi, A., R. Levi, and G.L. Ravignani, "Multi-Tool Machining Analysis, Part 1: Tool Failure Patterns and Implications," ASME Journal of Engineering for Industry, *101*, 2, 230-236 (1979).

[524] Zuckerman, D., "Replacement Models under Additive Damage," Naval Research Logistics Quarterly, *24*, 4, 549-588 (1977).

BOUNDS FOR STRENGTH-STRESS INTERFERENCE VIA MATHEMATICAL PROGRAMMING*

Geung-Ho Kim

State University of New York
Buffalo, New York

ABSTRACT

Problems of bounding $\Pr\{X > Y\}$, when the distribution of X is subject to certain moment conditions and the distribution of Y is known to be of convex-concave type, are treated in the framework of mathematical programming. Juxtaposed are two programming methods: one is based on the notion of weak duality and the other on the geometry of a certain moment space.

INTRODUCTION

Let X and its cumulative distribution function $F(\cdot)$ represent the strength variation of a certain system, and let Y and its c.d.f. $G(\cdot)$ represent the variation of the stress to which the system is subject. When X and Y are statistically independent, the quantity $R \equiv \Pr\{X > Y\} = \int G(t)\, dF(t)$ is commonly referred to as the reliability of the system. The problem of estimating R has been addressed in the literature in a variety of different contexts. Among such contributions are *Birnbaum and McCarty's [2] nonparametric procedures for confidence* intervals, Govindarajulu's [7] improvement on Birnbaum and McCarty's work via asymptotic normality of R, Church and Harris' [4] parametric procedures for UMVU estimation and confidence intervals, Enis and Geisser's [6] Bayesian inferences on R, and Bhattacharyya and Johnson's [1] generalization to multicomponent systems. In this study, we are interested, rather, in bounding R with respect to F for fixed G, and in determining the corresponding extremal distributions F^*.

We note that this problem is a slight modification of the classical variational problem underlying the Tchebycheff inequality; we need only replace G in the expression for R by a fixed symmetric set characteristic function, and then maximize R subject to given values of the first two moments of X. We note as well that both problems are special cases of what in Karlin and Studden [9] (Ch. XII), are called "generalized Tchebycheff problems," which are treated there, essentially, by the duality theory of linear programming.

The fact that problems of the Tchebycheff type can be solved effectively in the framework of linear programming theory has also been documented, for example, in Isii [8], Whittle [17], and Pyne [13]. In this paper, we treat the optimization of R through programming approaches, which, though kindred in spirit to the above, do seem to be especially well tailored to our problem, when the further assumption is made that G is "strictly unimodal, i.e., is strictly convex to the left of some point, and strictly concave to the right.

*This research was partially supported by the Air Force Office of Scientific Research, through Grant #78-3518 to Iowa State University.

The first approach, essentially a linear specialization of the weak duality argument of David and Kim [5], recommends itself for its simplicity, but fails when extremal distributions do not exist. The second approach, based on the geometry of a certain moment space, does succeed in such situations, but is less direct. We note in passing that Brook's [3] bounding of moment generating functions offers still a third programming alternative for the optimization of R. In Sec. 2, we outline the first method in conjunction with a certain pair of linear programs (P_I, D_I), and the second method in conjunction with a certain geometrically motivated program P_{II}. Sec. 3, devoted to the first method, illustrates the construction of extremal c.d.f.'s, in the context of two simple examples. Sec. 4 illustrates the second method, using the examples of Sec. 3.

2. LINEAR PROGRAMMING FORMULATIONS

Define the following classes of functions:

$\mathcal{F}$ The class of "generalized c.d.f.'s" F of the form cF, where $c \geqslant 0$ and F a c.d.f. on the line.

$\mathcal{F}_n$ The class of discrete c.d.f.'s F on the line with at most $n+1$ jumps.

$\mathcal{G}_1$ The class of c.d.f.'s G on the line that are strictly convex to the left of 0, and strictly concave to the right.

$\mathcal{G}_2$ The class of c.d.f.'s G on the line that are strictly concave on $[0, \infty]$ and identically 0 otherwise.

$\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$. Further, we assume that a $G \in \mathcal{G}$ possesses probability density function $g(\cdot)$.

For a given n-tuple $\underline{h}(t) \equiv (h_1(t), \ldots, h_n(t))$, where each $h_k(t)$ is a piecewise continuous function on the line, let $CH[\underline{h}(E_1)]$ denote the convex hull generated by $\underline{h}(t)$ when we vary t over the line. For a given point, $b = (b_1, \ldots, b_n) \in CH[\underline{h}(E_1)]$, we formulate a linear program:

$$P_I: \text{ maximize } \int G(t)\, dF(t)$$

$$\text{subject to } \int dF(t) = 1 \tag{2.1a}$$

$$\int h_k(t)\, dF(t) = b_k, \quad 1 \leqslant k \leqslant n \tag{2.1b}$$

$$\text{and } F \in \mathcal{F}_n \tag{2.2}$$

where G is a fixed c.d.f. in $\mathcal{G}$.

We note that the underlying space $\mathcal{F}$ in (2.2), being free from normalization, is a convex cone. Exploiting the cone structure of both (2.1) and (2.2) in conjunction with standard dual cone theory (Luenberger [10], (p. 157), and Sposito [15], (p. 261)), we may write down a linear program formally dual to P_I:

$$D_I: \text{ minimize } \lambda \beta'$$

$$\text{subject to } \lambda \eta(t)' \geqslant G(t), \ \forall\, t \in E_1, \tag{2.3}$$

$$\text{and } \lambda = (\lambda_0, \lambda_1, \ldots, \lambda_n) \in E_{n+1}, \tag{2.4}$$

where $\beta \equiv (1, b)$ and $\underline{\eta}(t) \equiv (1, \underline{h}(t))$.

Note first, by (2.1) and (2.3), that, if F^0 is feasible for P_I and λ^0 is feasible for D_I, then

$$\int G(t)\, dF^0(t) \leqslant \int \lambda^0 \eta^0(t)^T\, dF^0(t) = \lambda^0 \beta^T. \tag{2.5}$$

Therefore, (as suggested in David and Kim [5], and Pukelsheim [12]), if we find a feasible solution pair (F^*, λ^*) that satisfies

$$\int [\lambda^* \eta(t)^T - G(t)]\, dF^*(t) = 0, \tag{2.6}$$

then (F^*, λ^*) is in fact an optimal pair for (P_I, D_I), and, certainly, F^* satisfying (2.6) will need to concentrate its mass on the "set of osculation"

$$T(\lambda^*) = \{\tau \in E_1 \mid \lambda^* \eta(\tau)^T - G(\tau) = 0\}, \tag{2.7}$$

whose cardinality is bounded usually by $n + 2$, when the functions $\{G(t), \eta(t)\}$ are linearly independent on the line. (See Karlin & Studden [9].) Sec. 3 contains detailed demonstrations on how to construct an extremal c.d.f. F^* (which turns out to be supported at only n points in one example, and $(n-1)$ points in another example).

Now, with special reference to the second approach, consider the following program P_I', an essentially finite dimensional version P_I:

$$\begin{aligned} P_I'\text{: maximize } & \int G(t)\, dF(t) \\ \text{subject to } & \int h_k(t)\, dF(t) = b_k,\ 1 \leqslant k \leqslant n \\ \text{and } & F \in \mathcal{F}_a. \end{aligned}$$

The fact that P_I and P_I' yield the same optimal value follows from the general considerations in Rogosinski [14] and Mulholland and Rogers [11]. The reduction of P_I to P_I' provides a useful geometric version of our problem, in that the class $\mathcal{F}_a$ of P_I' generates the convex hull $CH[\Gamma]$ of the trace

$$\Gamma = \{\underset{\sim}{x} = (x_1, \ldots, x_n, x_{n+1}) \mid x_k = h_k(t),\ 1 \leqslant k \leqslant n, \text{ and}$$
$$x_{n+1} = G(t), \text{ for some } t \in E\}.$$

Hence, we are led to the equivalent program

$$P_{II}\text{: } \sup_{\underset{\sim}{x} \in C} x_{n+1}$$

where $C = CH[\Gamma]$, and $\mathcal{L}_b = \{\underset{\sim}{x} \mid x_k = b_k,\ 1 \leqslant k \leqslant n,\ x_{n+1} \in E_1\}$.

See Van Slyke and Wets [16], and Pyne [13] for similar constructions.

Since the set C in E_{n+1} is convex, the optimal value x_{n+1}^* of P_{II} may be obtained by associating this value with a suitable supporting hyperplane H_b of C at the boundary point $(b_1, \ldots, b_n, x_{n+1}^*)$. Finding the equation for H_b is not easy, however, since C is known only through Γ. In Sec. 4 the problem of finding H_b is attacked by considering H_b as a certain limit of all hyperplanes in E_{n+1} that cut or touch the set Γ.

3. ILLUSTRATION OF THE FIRST APPROACH

EXAMPLE 1. We wish to find the maximum reliability R^* of a system whose strength distribution F is known to have mean 0 and variance $b > 0$, when the distribution of the stress

to which the system is subject is given by a known continuous c.d.f. G in $\mathcal{G}$. We compute R^* via constructing of an extremal c.d.f. F^*. (Remark: There is no loss of generality in fixing the common value of the mean of F and the mode of G at zero. If their common value is in fact, say, a positive value M, then the corresponding F^{**} is obtained by shifting F^* obtained below to right by M.) Specializing the program pair (P_I, D_I) of Sec. 2 to this problem, we find the program pair

$$P_{II}: \text{ maximize } \int G(t)\, dF(t)$$

$$\text{subject to } \int dF(t) = 1$$

$$\int t\, dF(t) = 0 \tag{3.1}$$

$$\int t^2\, dF(t) = b, \tag{3.2}$$

$$\text{and } F \in \mathcal{F}.$$

$$D_{II}: \text{ minimize } \lambda_0 + \lambda_2 b$$

$$\text{subject to } \lambda_0 + \lambda_1 t + \lambda_2 t^2 \geqslant G(t),\ \forall t \in E_1, \tag{3.3}$$

$$\text{and } \underset{\sim}{\lambda} = (\lambda_0, \lambda_1, \lambda_2) \in E_3.$$

The osculating set $T(\underset{\sim}{\lambda}^*)$ of Sec. 2 now is the set of τ's where the parabola $P(\tau) \equiv \lambda_0^* + \lambda_1^* \tau + \lambda_2^* \tau^2$ lying above the "convex-concave" function $G(\tau)$ touches $G(\tau)$. At such τ, the derivatives $P'(\tau)$ and $G'(\tau)$ must coincide, and, since $P'(\tau)$ is linear and $G'(\tau)$ is either "increasing-decreasing" if $G \in \mathcal{G}_1$ or "identically zero-decreasing" if $G \in \mathcal{G}_2$, there can be at most two such τ's. We recall from Sec. 2 that the spectrum of F^* must be contained in $T(\underset{\sim}{\lambda}^*)$. Hence, in view of restriction (3.2), the spectrum of F^* consists of exactly two points s and t (with respective weights p and $(1-p)$), which, in addition, must be of opposite sign in view of restriction (3.1), say, $s < 0 < t$.

Pooling all our findings and restrictions, we write down the following nonlinear relations in the six unknowns s, t, p, λ_0^*, λ_1^*, and λ_2^*:

$$s\,p + t(1-p) = 0 \tag{3.5}$$

$$s^2 p + t^2(1-p) = b \tag{3.6}$$

$$\lambda_0^* + \lambda_1^* \tau + \lambda_2^* \tau^2 = G(\tau), \quad \tau = s, t \tag{3.7}$$

$$\lambda_1^* + 2\lambda_2^* \tau = g(\tau), \quad \tau = s, t \tag{3.8}$$

$$\lambda_2^* > 0. \tag{3.9}$$

Moreover, the optimality condition (2.6) adds the further requirement

$$G(s)\,p + G(t) \cdot (1-p) = \lambda_0^* + \lambda_2^* b. \tag{3.10}$$

Solving (3.5)-(3.10) for the six unknowns reduces to finding a positive t (and negative $s = -b/t$) satisfying

$$1/2[g(t) + g(-b/t)] \cdot [t + b/t] = G(t) - G(-b/t). \tag{3.11}$$

For $G \in \mathcal{G}_1$, relation (3.11) implies that t should be chosen such that the area under the density $g(\cdot)$ between $-b/t$ and t equals the area of the trapezoid formed by the four points $\{(-b/t, 0), (-b/t, g(-b/2)), (t, g(t)), (t, 0)\}$. For $G \in \mathcal{G}_2$, it turns out as well that we are to equate the area under $g(\cdot)$ between 0 and t with the area of the triangle formed by points $\{(-b/t, 0), (t, g(t)), (t, 0)\}$.

EXAMPLE 2. Here we consider a slight modification of Example 1; i.e., the restrictions (3.1) and (3.2) are replaced respectively by

$$\int t\,dF(t) = b_1 \tag{3.12}$$

$$\int |t|\,dF(t) = b_2 \tag{3.13}$$

The osculating set $T(\underline{\lambda}^*)$ of Sec. 2 now is the set of τ's where the wedge $W(\tau) \equiv \lambda_0^* + \lambda_1^*\tau + \lambda_2^*|t|$ lying above $G(\tau)$ touches $G(\tau)$. In view of the strict concavity of $G(\tau)$ for $t \geq 0$ and the specification of $T(\underline{\lambda}^*)$ above, the set $T(\underline{\lambda}^*)$ is reduced to a certain nonnegative singleton, which fact, in turn, implies that we should confine our search for an extremal F^* to the class of degenerate c.d.f.'s. However, since those F's that satisfy (3.11) and (3.12) with $b_1 \neq b_2$ cannot be degenerate, the weak duality method requiring, as it does, the existence of an extremal c.d.f. F^* is not applicable in this case. Only in the trivial case $b_1 = b_2 > 0$, can we fix F^* through $T(\underline{\lambda}^*)$; i.e., by weak duality.

4. ILLUSTRATION OF THE SECOND APPROACH

EXAMPLE a. (Example 2 of Sec. 3) Specializing of the formulation of P_{II} in Sec. 2 yields

$$P_{IIa}: \sup_{x \in C} x_3$$

where

$$C = CH[\Gamma],\ \Gamma = \{x \mid x_1 = t,\ x_2 = |t|,\ \text{and } x_3 = G(t);$$

$$\text{some } t \in E_1\} \tag{4.1}$$

and

$$\mathcal{L}_b = \{x \mid x_1 = b_1,\ x_2 = b_2,\ \text{and } x_3 \in E_1\}. \tag{4.2}$$

Since no triple of distinct points of Γ can be colinear, any such triple determines as hyperplane that cuts or touches Γ. The idea of our second approach is then to find a "best triple" that yields the highest hyperplane at $b = (b_1, b_2)$ among the collection $\mathcal{W}$ of all "qualified" triples w. In what follows, using the unimodality of G, we are able to reduce $\mathcal{W}$ to the collection $\mathcal{V}$ of "qualified" pairs v of distinct points in Γ.

To see this in detail, we first introduce the notation $\Pi(S)$ for the projection of the set S into the plane $x_3 = 0$. Now partition Γ into Γ^- and Γ^+, where Γ^- is the left side of Γ, corresponding to $t < 0$, and Γ^+ is the right side of Γ, corresponding to $t \geq 0$, and define

$$\mathcal{W} = \{w \mid w \text{ satisfies the condition that } b \in CH[\Pi(w)]\},$$

$$\mathcal{W}^+ = \{w \mid w \in \mathcal{W},\ \text{and any two of the three points of } w \text{ in}$$

$$\Gamma^+,\ \text{with the remaining point in } \Gamma^-\},\ \text{and } \mathcal{W}^- = \mathcal{W} - \mathcal{W}^+.$$

Also, define $h(w;b)$ = height at b of the hyperplane determined by a triple w. (Note that P_{IIa} is equivalent to find $\sup_{w \in \mathcal{W}} h(w;b)$.)

Using essentially the convexity of $G(t)$ for $t < 0$, it can be demonstrated that

LEMMA 1. For any triple $w \in \mathcal{W}^-$, there is a triple $w' \in \mathcal{W}^+$ such that $h(w';b) \geq h(w;b)$. Hence, $\sup_{w \in \mathcal{W}} h(w;b) = \sup_{w \in \mathcal{W}^+} h(w;b)$. Next, we define the collection of pairs

$\Gamma = \{v | \text{one point of } v \text{ in } \Gamma^* \text{ and the other point in } v$
is Γ^*; and also $b = CH[\Pi(v)]\}$.

Using the strict concavity of $G(t)$ for $t \geq 0$, we find

LEMMA 2. $\sup_{v \in \Gamma^*} h(v;b) = \sup_{v \in \Gamma} h(v;b)$.

To evaluate the right hand side of the above, we need an explicit expression for $h(v;b)$: Parametrizing by the (acute) angle α between $\Pi(CH[v])$ and $\Pi(\Gamma^*)$, and redefining h accordingly, we write

$$h(\alpha;b) = G(-\delta - \sigma \tan\alpha) \cdot \delta/(\delta + \sigma \tan\alpha)$$
$$G(\sigma + \delta \cot\alpha) \cdot \sigma \tan\alpha/(\delta + \sigma \tan\alpha),$$

where

$$\delta = 1.2(b_2 - b_1),\ \sigma = 1.2(b_2 + b_1), \text{ and } \alpha \in (\alpha_*, \pi/2).$$

A further reparametrization by $p = (\delta + \sigma \tan\alpha)$, (and redefining h accordingly), yields

$$h(p;b) = pG(-\delta/p) + (1-p)\, G(\sigma/(1-p)),\ p \in (0,1). \tag{4.3}$$

With the expression (4.3) for h, the monotonicity of G allows the conclusion that

$$\sup_{v \in \Gamma} h(p;b) = G(\sigma).$$

Moreover, if $b_1 \neq b_2$, then there does not exist an extremal c.d.f. achieving the optimal value $G(\sigma)$ of $P_{\Pi(\Gamma)}$, corresponding to the fact that Γ is not bounded. If $b_1 = b_2 > 0$, however, there is an extremal c.d.f. (degenerate at b_1) achieving $G(\sigma)$.

EXAMPLE b. (Example 1 of Sec. 3) In this case, (4.1) and (4.2) of Example a are replaced by

$$\Gamma = \{x : x_1 = t,\ x_2 = t^2, \text{ and } x_3 = G(t);\ \text{some } t \in E_1\}. \tag{4.4}$$

$$\mathcal{L} = \{x : x_1 = 0,\ x_2 = b,\ \text{and } x_3 = \in E_1\}. \tag{4.5}$$

Following the analogous argument, we reparametrize $v \in \Gamma^*$, the modification of Γ pertinent to (4.4) and (4.5), by $\theta \in [0,\pi]$, where the angle θ is between the line extending $\Pi(CH[v])$ and the x_1-axis.

For given $b = (0,b)$, by letting $p = 1/2 \tan\theta$, we find

$$h(p,b) = G(p + r(p)) \cdot \{(r(p) - p)/2r(p)\}$$
$$+ G(p - r(p)) \cdot \{(r(p) + p)/2r(p)\}, \tag{4.6}$$
where $r(p) = (p^2 + b)^{1/2}$.

It is easy to check that $h(p;b)$ is concave on $(-\infty, 0]$ and convex on $(0,\infty)$ in view of the fact that $G(\cdot) \in \mathcal{G}$. Differentiating (4.6) with respect to p and setting it equal to 0 yields

$$g(p + r(p)) + g(p - r(p)) = \{G(p + r(p)) - G(p - r(p))\}/r(p), \tag{4.7}$$

which is to be solved for the p^* in $(-\infty, 0]$ that maximized (4.6). We note that (4.7) does reduce to (3.11) with $t = p + r(p)$.

6. ACKNOWLEDGMENTS

I would like to thank C. R. Mischke for suggesting the problem considered here. I am also indebted to H. T. David for numerous helpful suggestions and comments offered in the course of this study.

REFERENCES

[1] Bhattacharyya, G.K. and R.A. Johnson, "Strength-Stress Models for System Reliability," *Reliability and Fault Tree Analysis* (R.E. Barlow, Editor), Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania (1975).

[2] Birnbaum, Z.W. and R.C. McCarty, "A Distribution-Free Upper Confidence Bound for Pr $\{Y < X\}$ Based on Independent Samples of X and Y" Annals of Mathematical Statistics *29*, 558-562 (1958).

[3] Brook, D., "Bounds for Moment Generating Functions and for Extinction Probabilities," Journal of Applied Probability, *3*, 171-178 (1966).

[4] Church, J.D. and B. Harris, "The Estimation of Reliability from Stress-Strength Relationships," Technometrics, *12*, 49-54 (1970).

[5] David, H.T. and Geung-Ho Kim, "Pragmatic Optimization of Information Functionals," *Optimizing Methods in Statistics—Proceedings of the International Conference on Optimization in Statistics*, Bombay, India, 1977. Rustagi, J.S., Editor, (Academic Press, New York, N.Y. 1979).

[6] Enis, P. and Geisser, S. "Estimation of the Probability that $Y < X$," Journal of the American Statistical Association, *66*, 162-168 (1971).

[7] Govindarajulu, Z. "Distribution-Free Confidence Bounds for $P(X < Y)$," Annals of the Institute of Statistical Mathematics, *20*, 229-238 (1968).

[8] Isii, K., "Inequalities of the Types of Chebychev and Cramér-Rao and Mathematical Programming," Annals of the Institute of Statistical Mathematics, *16*, 277-293 (1964).

[9] Karlin, S. and W.J. Studden, *Tchebycheff Systems: with Applications in Analysis and Statistics*, (Wiley-Interscience, New York, N.Y. 1966).

[10] Luenberger, D.G., *Optimization by Vector Space Methods*. (John Wiley and Sons, Inc., New York, N.Y., 1969).

[11] Mulholland, H.P. and C.A. Rogers, "Representation Theorems for Distribution Functions," Proceedings of the London Society, Series 3, 177-223 (1958).

[12] Pukelsheim, "A Quick Introduction to Mathematical Programming with Applications to Most Powerful Tests, Nonnegative Variance Estimation, and Optimal Design Theory," Technical Report No. 128, Stanford University, Stanford, California (1978).

[13] Pyne, D.A., "Duality in Abstract Mathematical Programming with Applications to Statistical Problems," Unpublished Ph.D. Thesis, Iowa State University, Ames, Iowa (1972).

[14] Rogosinski, W.W. "Moments of Nonnegative Mass," Proceedings of the Royal Society of London, Series A, *245*, 1-27 (1958).

[15] Sposito, V.A., Linear and Nonlinear Programming, (Iowa State University Press, Ames, Iowa, 1975).

[16] Van Slyke, R.M. and R.J.B. Wets, "A Duality Theorem for Abstract Mathematical Programmings with Applications to Optimal Control Theory," Journal of Mathematical Analysis and Application *22*, 679-706 (1968).

[17] Whittle, P., *Optimization under Constraints: Theory and Applications of Nonlinear Programming* (Wiley-Interscience, New York 1971).

BOUNDS AND ELIMINATION IN GENERALIZED MARKOV DECISIONS

Gary J. Koehler

Micro Data Base Systems, Inc. *
Lafayette, Indiana

ABSTRACT

In discounted Markov decision processes bounds on the optimal value function can be computed and used to eliminate suboptimal actions. In this paper we extend these procedures to the generalized Markov decision process. In so doing we forfeit the contraction property and must base our analysis on other procedures. Duality theory and the Perron-Frobenius theorem are the main tools.

1. INTRODUCTION

In this paper a finite state and action, infinite horizon, generalized Markov decision process consists of a finite set of s states denoted by S; a finite set of actions A_i for each $i \in S$; an immediate reward c_i^a for each $i \in S$ and $a \in A_i$ and a weighted "generalized probability" p_{ij}^a for each $i, j \in S$ and $a \in A_i$. Let $\Delta = \underset{i=1}{\overset{s}{X}} A_i$ denote the set of decisions. For $\delta \in \Delta$, c^δ refers to the $s \times 1$ reward vector where $c_i^{\delta(i)}$ is the immediate reward for using action $\delta(i)$ in state i and P_δ is the $s \times s$ generalized probability matrix associated with using decision δ. A generalized Markov decision process requires that

1. $P_\delta \geqq 0$ for each $\delta \in \Delta$
2. $\rho(P_\delta) < 1$ for at least one $\delta \in \Delta$
3. $D = \{v : v \geqq P_\delta v + c^\delta, \ \delta \in \Delta\} \neq \varnothing$

where $\rho(P)$ is the spectral radius of the square matrix P.

Let $\mathscr{L}_\delta(\cdot)$ and $\mathscr{L}(\cdot)$ be defined over R^s where

$$\mathscr{L}_\delta(v) = P_\delta v + c^\delta$$

and

$$\mathscr{L}(v) = \underset{\delta \in \Delta}{V\text{-Max}} \mathscr{L}_\delta(v)$$

where V-Max means vector maximization. Since each P_δ is isotone (i.e., $x \geqq y$ implies $P_\delta x \geqq P_\delta y$), $\mathscr{L}_\delta$ and $\mathscr{L}$ are accordingly isotone. Notice that $\mathscr{L}$ may not be a contraction mapping or even an N-stage contraction mapping and thus may not possess a unique fixed point [2]. Since $D \neq \varnothing$, it is easy to show that $\mathscr{L}$ has at least one fixed point [7,8]. Let $F = \{v : v = \mathscr{L}(v)\}$ be the set of fixed points of $\mathscr{L}$. We wish to solve

*This research was performed when the author was affiliated with the School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

$$v^* = \underset{v \in I}{V\text{-Min}}\ v.$$

This problem is well defined and is motivated in [7,8].

Such problems were studied in [7] as a generalization of [8] and encompass traditional discounted Markov decisions [6], the discounted processes investigated by Veinott [17] and the more general processes resulting from the duals to linear programs with (hidden) Leontief Substitution Systems and (hidden) essentially Leontief Substitution Systems. The latter two cases include such applications as completely-ergodic nondiscounted Markov decision processes [9], shortest path problems (with or without cycles), and the stopping model of Denardo and Rothblum [3].

It has long been known in the context of the traditional discounted Markov decision process [10,12,13,15] and more recently in the discounted processes of Veinott [12] that bounds of the form $l \leqq v^* \leqq u$ can be constructed and used to eliminate inferior actions from further consideration as potential candidates of an optimal stationary policy [4,5,11,12,13,15].

In this paper we extend the development and usage of bounds on v^* to the generalized Markov decision setting. Since most results in the literature were developed using a contraction argument and the generalized process does not usually possess this property, we must utilize a slightly different set of machinery. We will rely heavily on duality theory and the Perron-Frobenius theorem (see Varga [16] or Seneta [14]).

2. NOTATION AND PRELIMINARY RESULTS

Let x and y be two vectors. Write $x \geqq y$ (respectively, $x > y$) if $x_i \geqq$ (respectively, $x_i >$) y_i for every i. Also write $x \geqslant y$ if $x \geqq y$ but $x \neq y$. Let $L(x) = \{z : z \leqq x\}$ and if T is a set, let $L(T) = \underset{x \in I}{U} L(x)$. If P is a square matrix, $p(P)$ will denote the spectral radius of P. If $P \geqq 0$ and square then the Perron-Frobenius theorem gives us that $Px = p(P)x$ for some $x \geqslant 0$ and $p(P) \geqq 0$. $(I-P)^{-1}$ exists and is nonnegative if $p(P) < 1$.

From [1,18] we have that v^* is given by some $\delta^* \in \Delta$ where $p(P_{\delta^*}) < 1$ and $v^* = (I-P_{\delta^*})^{-1} c^{\delta^*}$. In this paper we are interested in finding v^* by successively iterating $\mathscr{L}$. That is, $v^n = \mathscr{L}^n(v^0)$ where v^0 is an initial guess of v^*. Let $C = \{v : v^* = \lim_{n \to \infty} \mathscr{L}^n(v)\}$ be the set of all starting points leading to v^* under the successive application of $\mathscr{L}$. $C \neq \varnothing$ since $v^* \in C$. In both the discounted Markov decision process and Veinott's discounted process $C = R^S$. In general, however, $C \neq R^S$ [7,8]. A useful result obtained by Koehler [7,8] is that $L(C) \subseteq C$ so, since $v^* \in C$, $L(v^*) \subseteq C$. The following short example gives a case where $L(v^*) = C$. The problem is:

State	Action	P_{ij}^a		c_i^a
1	1	0	0	0
	2	0	1	0
2	1	1	0	0

It is readily determined that $v^* = 0$, $D = F = \{\lambda e : \lambda \geqq 0\}$ and $\mathscr{L}(v^*) = \{x : x \leqq 0\}$ where e is a vector of ones. For any starting vector v^0 we get for $n \geqq 2$

$$\mathscr{L}^n(v^0) = \begin{pmatrix} \text{Max } (0, v_1^0) \\ \text{Max } (0, v_2^0) \end{pmatrix}$$

where $h = 1$ and $l = 2$ if n is even and $h = 2$ and $l = 1$ if n is odd. Hence, $\mathscr{L}^n(v^0)$ converges if and only if $v^0 \leqq 0$, i.e., $v^0 \in L(v^*)$.

From a practical point of view, it is easy to pick a point of $L(v^*)$. For example, let $v^0 = Md$ where $M << 0$ and $d > 0$. Thus, since $L(v^*)$ may be C and picking points of $L(v^*)$ is relatively easy, when we restrict attention to the case where $v^0 \in L(v^*)$ we do so without much practical loss of generality. In the previous section we defined $D = \{v: v \geqq P_\delta v + c^\delta, \delta \in \Delta\} = \{v: v \geqq \mathscr{L}(v)\}$. We wish to express this set in one further way. Define the vector f_i^a by

$$f_{ij}^a = \begin{cases} 1 - P_{ij}^a & i = j \\ -P_{ij}^a & i \neq j \end{cases}$$

where $i, j \in S$ and $a \in A_i$. Let F be a matrix having each f_i^a as a row where $a \in A_i$, $i \in S$. Corresponding to F, let c be a vector of the c_i^a values. Then we can write D as $D = \{v: Fv \geqq c\}$. The matrix F is essentially Leontief [7] and since $\rho(P) < 1$ for some $\delta \in \Delta$, the set $\{x: Fx > 0, x \geqq 0\}$ is nonempty [18].

3. ELIMINATION OF SUBOPTIMAL ACTIONS

Suppose one has bounds l and u such that $l \leqq v^* \leqq u$. If action $a \in A_i$ is part of an optimal policy, then the inequality $v_i \geqq \Sigma P_{ij}^a v_j + c_i^a$ must be tight at v^*. Clearly then, if the above inequality is never tight in the polytope $B = \{x: l \leqq x \leqq u\}$ it cannot be tight at v^* and should be eliminated from further consideration. A typical test for checking this condition is if

$$\Sigma P_{ij}^a u_j + c_i^a < l_i \tag{1}$$

then a is suboptimal (see [5,11,12,13,15] for such examples).

A tighter test results directly from duality theory. The inequality $v_i \geqq \Sigma P_{ij}^a v_j + c_i^a$ is not tight in B if and only if

$$\bar{c}_i^a > 0 \text{ and } (1 - P_{ii}^a)(u_i - l_i) < \bar{c}_i^a \tag{2}$$

or

$$\bar{c}_i^a < 0 \text{ and } \sum_{f_{ij}^a < 0} f_{ij}^a (u_j - l_j) > \bar{c}_i^a$$

where $\bar{c}_i^a = c_i^a - \Sigma f_{ij}^a l_j$.

Notice that the test in (1) can never eliminate an action when $\bar{c}_i^a > 0$ but that (2) allows this condition. Anything eliminated by test (1) is removed by (2).

4. BOUNDS FOR THE GENERALIZED PROBLEM

We begin our development of bounds on v^* by considering the restricted case where $\Delta = \{\delta\}$ and $\rho(P_\delta) < 1$. That is, we wish to determine bounds on $v^* = (I - P_\delta)^{-1} c^\delta$. Both Porteus [12] and, indirectly, Veinott [17] have investigated this case. Porteus first transforms the process into an equivalent one where the new transition matrix $\tilde{P}_\delta$ has all equal row sums (which are necessarily less than 1.0). Once this has been accomplished, bounds such as [10,12,13,15] can be computed. Here we do not transform the data.

For the time being, let us suppress δ. Let $d > 0$ but otherwise arbitrary. Let a and b satisfy

$$ad \leqq v^{n+1} - v^n \leqq bd. \tag{3}$$

We wish to develop bounds of the form

$$l = v^{n+k} + \alpha d \leqq v^* \leqq v^{n+k} + \beta d = u \tag{4}$$

where k is a nonnegative integer. Here $v^n = \mathcal{L}^n(v^0)$, $v^0 \in C = R^S$ and $\lim v^n = v^*$.

Since P^k is isotone, from (3) we have

$$aP^k d \leqq P^k v^{n+1} - P^k v^n \leqq bP^k d. \tag{5}$$

Multiply (5) by P to get

$$aP^{k+1} d \leqq P^{k+1} v^{n+1} - P^{k+1} v^n \leqq bP^{k+1} d$$

and add this to (5). We get

$$a(I + P)P^k d \leqq P^{k+1} v^{n+1} - P^k v^n \leqq b(I + P)P^k d.$$

Repeating this procedure and taking limits gives

$$a(I - P)^{-1} P^k d \leqq P^k v^* - P^k v^n \leqq b(I - P)^{-1} P^k d. \tag{6}$$

PROPOSITION 1: Let $v^* = Pv^* + c$ where $P \geqq 0$ and $\rho(P) < 1$. Let $v^{n+1} = Pv^n + c$ and $d > 0$. a and b are such that

$$ad \leqq v^{n+1} - v^n \leqq bd$$

and $k \geqq 0$ *and integral, then*

$$\alpha d + v^{n+k} \leqq v^* \leqq \beta d + v^{n+k}$$

where

$$\begin{aligned} &\beta \geqq 0 \quad \text{if } b = 0 \\ &\beta \geqq b\bar{\gamma} \quad \text{if } b > 0 \\ &\beta \geqq b\chi \quad \text{if } b < 0 \end{aligned}$$

and

$$\begin{aligned} &\alpha \leqq 0 \quad \text{if } a = 0 \\ &\alpha \leqq a\chi \quad \text{if } a > 0 \\ &\alpha \leqq a\bar{\gamma} \quad \text{if } a < 0 \end{aligned}$$

where

$$\bar{\gamma}(\chi) = \text{Max (Min) } x'P^k d$$

subject to

$$x'd - x'Pd = 1$$

$$x'(I - P) \geqq 0.$$

PROOF: We will prove the result for β and note that the proof for α follows in a similar manner. Let $y \equiv b(I - P)^{-1} P^k d \geqq v^* - v^{n+k}$ as given in (6). Then, by duality, $y \leqq \beta d$ if and only if $x'(I - P) \geqq 0$ implies $\beta x'(I - P)d \geqq bx'P^k d$. Notice that $x'(I - P) \geqq 0$ implies $x \geqq 0$ since $(I - P)^{-1} \geqq 0$. Also, since $\rho(P) < 1$, using the Perron-Frobenius theorem we

get that $x'(I - P) = 0$ if and only if $x = 0$ and $x'(I - P) \geqslant 0$ whenever $x \geqslant 0$ gives $x'(I - P) \geqq 0$. There is such an $x \geqslant 0$ (use $x' = d'(I - P)^{-1}$). Hence, β must satisfy

$$\beta \geqq \frac{bx'P^k d}{x'(I - P)d}$$

whenever $x'(I - P) \geqq 0$ with $x \geqslant 0$. Enumerating the cases where $b = 0$, $b < 0$ and $b > 0$ gives the results of the theorem. We need only show that the objective function of the linear programs is bounded for all feasible points. Suppose this is not true. Then there is a ray $z \geqslant 0$ such that $z'd - z'Pd = 0$ or $z'(I - P)d = 0$. Since $d > 0$ and $z'(I - P) \geqq 0$, $z'(I - P) = 0$. This gives that $z = 0$, a contradiction.

Some useful cases follow.

COROLLARY 1: When $k = 0$,

$$\bar{\gamma} = \frac{1}{1 - q} \quad \chi = \frac{1}{1 - r}$$

and when $k = 1$

$$\bar{\gamma} = \frac{q}{1 - q} \quad \chi = \frac{r}{1 - r}$$

where

$$q(r) = \text{Max (Min) } x'Pd$$

s.t.

$$x'd = 1$$

$$x'(I - P) \geqq 0.$$

Note that q and r are both strictly less than 1.0.

COROLLARY 2: If d is an eigenvector of P with $Pd = \lambda d$, then

$$\bar{\gamma} = \chi = \frac{\lambda^k}{1 - \lambda}.$$

Most of the bounds reported for discounted Markov decisions fall into one of the two cases given above. Usually d is a vector of ones.

While determining $\bar{\gamma}$ or χ is, in general, a nontrivial task, one can usually obtain useful bounds on $\bar{\gamma}$ and χ and use these. For example, the Perron-Frobenius eigenvector is a feasible solution so

$$\bar{\gamma} \geqq \frac{p^k(P)}{1 - p(P)} \geqq \chi \geqq 0.$$

Also, as is commonly known,

$$\operatorname{Min}_i \frac{(Pd)_i}{d_i} \leqq p(P) \leqq \operatorname{Max}_i \frac{(Pd)_i}{d_i}.$$

The dual problems also provide bounds although one must obtain tight enough upper bounds to be meaningful.

We now return our attention to determining bounds for the generalized Markov decision process. In the following we assume $v^0 \in L(v^*)$ so that

$$l^0 \equiv v^0$$

and

$$l_i^{n+1} = \text{Max}\ (l_i^n,\ v_i^{n+1}) \tag{7}$$

provides us with a lower bound to v^* at each iteration. An upper bound is not as easy to derive.

In the unlikely event that δ^* is known, one can use the upper bound developed in Proposition 1 since

$$P_{\delta^*}^k(v^{n+1} - v^n) \leqq bP_{\delta^*}^k d$$

plus

$$P_{\delta^*}^{k+1}(v^{n+1} - v^n) \leqq bP_{\delta^*}^{k+1} d$$

gives

$$P_{\delta^*}^k \mathscr{L}_{\delta^*}(v^{n+1}) - P_{\delta^*}^k v^n \leqq b(I + P_{\delta^*})P_{\delta^*}^k d$$

or, in the limit,

$$P_{\delta^*}^k v^* - P_{\delta^*}^k v^n \leqq b(I - P_{\delta^*})^{-1} P_{\delta^*}^k d.$$

That is,

$$v^* - \mathscr{L}_{\delta^*}^k(v^n) \leqq b(I - P_{\delta^*})^{-1} P_{\delta^*}^k d.$$

The resulting bound is

$$v^* \leqq \mathscr{L}_{\delta^*}^k(v^n) + \beta d \leqq v^{n+k} + \beta d \tag{8}$$

where β is given in Proposition 1 and $\bar{\gamma}$ and χ correspond to P_{δ^*}. For $k = 0$ we get $v^* \leqq v^n + \beta d$ and for $k = 1$ we get $v^* \leqq \mathscr{L}_{\delta^*}(v^n) + \beta d \leqq v^{n+1} + \beta d$.

We realize, of course, that if δ^* is known, one would ignore all other $\delta \in \Delta$ and work only with δ^*. A more reasonable case is if δ^* is unknown but $\bar{\gamma}_{\delta^*}$ is known. Since $v^0 \leqq v^*$, $v^{n+1} \geqslant v^n$ for all n unless $v^n = v^*$. Hence, $v^{n+1} - v^n \leqslant bd$ implies $b > 0$. Thus, knowledge of $\bar{\gamma}_{\delta^*}$ is sufficient for determining an upper bound on v^*.

As an illustration of (8) and the elimination procedure of (2) consider the following example:

Example 1

State	Action	P_{ij}^a		c_i^a
1	1	0	2	2
	2	0	1	3
2	1	2	0	−6
	2	1	0	−3
	3	0	0	−1

Note that no P_δ has all its rows less than one. Here

$$v^* = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

$$\delta^* = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$p(P_{\delta^*}) = 0$$

$$d = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \bar{\gamma}_{\delta^\infty} = 1 \text{ for } k = 1.$$

Let $v^0 = -10e$. Then

$$v^1 = \begin{bmatrix} -7 \\ -1 \end{bmatrix}$$

so

$$v^1 - v^0 = \begin{bmatrix} 3 \\ 9 \end{bmatrix}.$$

Let $b = 9$ and $\beta = 9$. Then.

$$\begin{bmatrix} -7 \\ -1 \end{bmatrix} \leqq v^* \leqq \begin{bmatrix} -7 \\ -1 \end{bmatrix} + \begin{bmatrix} 9 \\ 9 \end{bmatrix} = \begin{bmatrix} 2 \\ 8 \end{bmatrix}.$$

Using the elimination procedure of (2) we get

State	Action	$\bar{c}_i^a$	Test Value	
1	1	7	9	
	2	9	9	
2	1	−19	−18	Eliminate
	2	−9	−9	
	3	0	0	

Suppose neither δ^* nor $\bar{\gamma}_{\delta^\infty}$ is known. Consider the following. Let $\Delta' \subseteq \Delta$ such that $\delta^* \in \Delta'$, $\delta \in \Delta'$ implies $p(P_\delta) < 1$ and if $\delta(i) = \bar{\delta}(i)$ for some $\bar{\delta} \in \Delta'$ for each $i = 1, \ldots, s$ then $\delta \in \Delta'$. A special case is $\Delta' = \{\delta^*\}$. After appropriate permutations we could write $F = (F_1, F_2)$ where F_1 corresponds to Δ'. The matrix F_1 is totally Leontief and has several desirable features, one of which is that the set $\{x : F_1 x \leqq 0, x \geqq 0\}$ is empty [18]. Let $F_1 = B - Q$ where each column of Q looks like $e_i - f_i^a$ where $a = \delta(i)$ for some $\delta \in \Delta'$. B then has unit vector columns and each row has at least one $+1$.

In a manner analogous to the procedures leading to (6) and (8) we can determine conditions on β such that

$$bP_\delta^k d \leqq (I - P_\delta)u \tag{9}$$
$$u \leqq \beta d$$

for all $\delta \in \Delta'$ and thus obtain an upper bound to $\bar{\gamma}_{\delta^\infty}$. System (9) can be written as

$$bd \leqq F_1' u \qquad k = 0$$
$$u \leqq \beta d$$

and

$$bQ'd \leqq F_1' u \qquad k = 1$$
$$u \leqq \beta d.$$

The following result follows:

PROPOSITION 2.

Let $d > 0$ and b satisfy

$$v^{n+1} - v^n \leqq bd$$

where $v^0 \in L(v^*)$ and $v^n = \mathscr{L}^n(v^0)$. Let F_1 be constructed as given above. Then

$$v^* \leqq v^{n+k} + \beta d$$

if $k = 0$ and

$$\beta \geqq b \text{ Max } d'x$$

s.t.

$$\begin{aligned} d'F_1x &= 1 \\ F_1x &\geqq 0 \\ x &\geqq 0 \end{aligned}$$

or if $k = 1$ and

$$\beta \geqq b \text{ Max } d'Qx$$

s.t.

$$\begin{aligned} d'F_1x &= 1 \\ F_1x &\geqq 0 \\ x &\geqq 0 \end{aligned}$$

PROOF: Let $g = bd$ if $k = 0$ and $g = bQ'd$ if $k = 1$. Then $g \leqq F_1'u$, $u \leqq \beta d$ has a solution if and only if $x \geqq 0$, $F_1x \geqq 0$ implies $\beta d'F_1x \geqq g'x$. The rest follows as in Proposition 1 except here we note that the constraint set is bounded since $\{x: F_1x = 0, x \geqq 0\}$ is empty.

The dual linear programs provide upper bounds to the solutions of the problems in Proposition 2 and these in turn are upper bounds to $\bar{\gamma}_{\delta^*}$. The bounds of Proposition 2 are used as

$$v^* \leqq v^{n+k} + \beta d.$$

The final case we consider is when no Δ' can be determined due, perhaps, to the necessity of knowing that $\delta^* \in \Delta'$. In such a case one is faced with the unpleasant task of determining a $\bar{\gamma}_\delta$ for each $\delta \in \Delta$ where $p(P_\delta) < 1$ and then using the largest such value in determining β. This would involve solving

(10) $$\text{Max } x'P_\delta^k d$$

s.t.

$$\begin{aligned} x'd - x'P_\delta d &= 1 \\ x'(I - P) &\geqq 0 \\ x &\geqq 0 \end{aligned}$$

for each $\delta \in \Delta$. Unbounded or infeasible problems can be ignored. While this procedure would be a considerable task, if a decision problem is to be solved a large number of times with only the c_i^a elements changing, then it may be of value to determine a bound for β in this fashion.

As an example, the optimal solution values to (10) for each $\delta \in \Delta$ of the problem in Example 1 are:

δ	Value ($k = 1$)
(1,1)	No Solution
(1,2)	No Solution
(1,3)	2
(2,1)	No Solution
(2,2)	No Solution
(2,3)	1

Thus, without knowledge of δ^* one would have to use $\beta \geqq 2b$. Note also that $\Delta' = \left\{\begin{bmatrix}1\\3\end{bmatrix}, \begin{bmatrix}2\\3\end{bmatrix}\right\}$ and the procedure of Proposition 2 would have led to $\beta \geqq 2b$ also.

As a final note, it is not always possible to abstract a $\Delta' \subseteq \Delta$ containing all $\delta \in \Delta$ having $p(P_\delta) < 1$ with no $\delta \in \Delta'$ having $p(P_\delta) \geqq 1$. For example,

State	Action	P_{ij}^a
1	1	0 0
	2	0 1
2	1	1 0
	2	0 0

we find that $p(P_\delta) < 1$ only for $\delta \in \left\{\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}2\\2\end{bmatrix}\right\}$. This set does not qualify for a Δ' set since $\delta = \begin{bmatrix}2\\1\end{bmatrix}$ has $p(P_\delta) = 1$ yet $\delta(1)$ and $\delta(2)$ are represented in the set. Hence, one may *have* to use (10) instead of Proposition 2.

REFERENCES

[1] Cottle, R.W. and A.F. Veinott, Jr., "Polyhedral Sets Having a Least Element," Mathematical Programming, *3*, 238-249 (1972).

[2] Denardo, E.V. "Contraction Mapping in the Theory Underlying Dynamic Programming," SIAM Review, *9*, 165-177 (1967).

[3] Denardo, E.V. and U.G. Rothblum, "Optimal Stopping, Exponential Utility, and Linear Programming," School of Organization and Management, Yale University, (October 1977).

[4] Grinold, R.C. "Elimination of Suboptimal Actions in Markov Decision Problems," Operations Research, *21*, 848-851 (1973).

[5] Hastings, N. and J. Mello, "Tests for Suboptimal Actions in Discounted Markov Programming," Management Science, *19*, 1019-1022 (1973).

[6] Howard, R.A., *Dynamic Programming and Markov Processes*, (MIT Press, Cambridge, Massachusetts, 1960)

[7] Koehler, G.J. "Generalized Markov Decision Processes," forthcoming RAIRO Operations Research.

[8] Koehler, G.J. "Value Convergence in a Generalized Markov Decision Process," SIAM Journal of Optimization and Control, *17*, 180-186 (1979).

[9] Koehler, G.J., A.B. Whinston and G.P. Wright, "An Iterative Procedure for Non-Discounted Discrete-Time Markov Decisions," Naval Research Logistics Quarterly, *21*, 719-723 (1974).

[10] MacQueen, J.B. "A Modified Dynamic Programming Method for Markovian Decision Problems," Journal of Mathematical Analysis and Application *14*, 38-43 (1966).

[11] MacQueen, J.B. "A Test for Suboptimal Actions in Markovian Decision Problems," Operations Research, *5*, 559-561 (1967).
[12] Porteus, E.L. "Bounds and Transformations for Discounted Finite Markov Decision Chains," Operations Research, *23*, 761-784 (1975).
[13] Porteus, E.L. "Some Bounds for Discounted Sequential Decision Processes," Management Science, *18*, 7-11 (1971).
[14] Seneta, E., *Non-negative Matrices*, (John Wiley and Sons, New York, 1973).
[15] Totten, J.C., "Computational Methods for Finite State Finite Valued Markovian Decision Problems," Report 71-9 Operations Research Center, University of California, Berkeley, California (1971).
[16] Varga, R.S., *Matrix Iterative Analysis*, (Prentice Hall, Englewood Cliffs, New Jersey, 1962).
[17] Veinott, A.F., Jr., "Discrete Dynamic Programming with Sensitive Discount Optimality Criteria," Annals of Mathematical Statistics, *40*, 1635-1660 (1969).
[18] Veinott, A.F., Jr., "Extreme Points of Leontief Substitution Systems," Linear Algebra and Its Applications, *1*, 181-194 (1968).

SURROGATE DUALITY IN A BRANCH-AND-BOUND PROCEDURE

Mark H. Karwan

Department of Industrial Engineering
State University of New York at Buffalo
Buffalo, New York

Ronald L. Rardin

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia

ABSTRACT

Recent research has led to several surrogate multiplier search procedures for use in a primal branch and bound procedure. As single constrained integer programming problems, the surrogate subproblems are also solved via branch-and-bound. This paper develops the inner play between the surrogate subproblem and the primal branch-and bound trees which can be exploited to produce a number of computational efficiencies. Most important is a restarting procedure which precludes the need to solve numerous surrogate subproblems at each node of a primal branch-and bound tree. Empirical evidence suggests that this procedure greatly reduces total computation time.

1. INTRODUCTION

Consider the general integer linear programming problem:

$$(P) \quad Min\ cx \text{ subject to } Ax \underset{(\leq)}{\leqslant} b$$

where $S = \{x \geqslant 0:\ Gx \leqslant h,\ x \text{ satisfies some discrete constraints}\}$. Here, A and G are $m \times n$ and $q \times n$ matrices respectively, with all vectors having the appropriate dimension.

The *surrogate relaxation* of the problem (P) associated with any $v \geqslant 0$ is

$$(P^v) \quad Min\ cx \text{ subject to } v(Ax-b) \underset{(\leq)}{\leqslant} 0.$$

If we define the function

$v(\cdot)$ = The value of an optimal solution to problem $(\cdot)$ if one exists and $+\infty$ if the problem is infeasible

then clearly $v(P^v)$ provides a lower bound on $v(P)$ for any $v \geqslant 0$. The best such bound is achieved by the *surrogate dual*.

$$(D_v) \quad \underset{v \geqslant 0.}{Max} \{v(P^v)\}$$

Only in rare integer programs would one expect such a dual problem to directly produce a solution to (P). Thus, the importance of duals in integer programming centers on their ability to produce bounds for a branch-and-bound procedure. By careful partitioning of the constraints of a problem into those which are relaxed $Ax \leqslant b$, and those which are enforced $x \in S$, problems, (P^v), can be created which are easier to solve than (P). Thus the bound $v(P^v)$ is easier to obtain, and searches over $v \geqslant 0$ will produce improved bounds. The successful application of duality in a branch-and-bound scheme can be seen to depend on the quality of these bounds and the ease of computing the bounds, since one must repeat the procedure over and over with different candidate sets.

Recent research (see Karwan and Rardin [6]) has produced a number of surrogate multiplier search procedures. Empirical results [5] suggest surrogate duals may close a significant fraction of the gap between the values of the lagrangian dual and the primal problems.

In this paper, the intent is to more fully develop the inner play between the surrogate dual and the primal in a branch-and-bound procedure. When the two are considered conjunctively a number of advantages are gained beyond the providing of a bound by the surrogate dual. A number of general observations will first be made with respect to the surrogate dual. Then specific issues or parts of the general branch-and-bound procedure will be developed in their relationship with the surrogate dual.

2. SURROGATE SUBPROBLEMS

Consider the surrogate relaxation of (P) for any $v \geqslant 0$. Note that (P^v) is itself an integer linear programming problem with a single main constraint $v(Ax-b) \leqslant 0$. Thus, it is a knapsack problem with a set of side constraints, $x \in S$. A number of solution techniques have appeared in the literature for the case of $S = \{x: x \geqslant 0,\ x \text{ bounded above}\}$. Basically these can be divided into two categories, dynamic programming procedures and branch-and-bound or implicit enumeration procedures. For a good review of the dynamic programming procedures, see Garfinkel and Nemhauser [3]. It will soon become evident that a branch-and-bound procedure will be more convenient in solving (P^v), because the relation between the primal and knapsack branch-and-bounds can be exploited. Moreover, Cabot [1], Kolesar [7], Fayard and Plateau [2], and Greenberg and Hegerich [4], among others, have developed branch-and-bound procedures which proved computationally more efficient than the dynamic programming approaches. Finally, Karwan and Rardin [6] have shown that each surrogate relaxation need not be solved optimally. Only a feasible solution with value less than or equal to the incumbent solution value of the surrogate dual is necessary for terminating the solution of (P^v). By solving (P^v) via a branch-and-bound procedure such solutions are readily available, require no extra computations, and lead to fewer iterations (choices of v) in solving (D_v). In a dynamic programming procedure, however, a feasible solution is generally not available until optimality is obtained so that (P^v) must be solved completely. For these reasons, and more to become apparent upon seeing the inner play with (P), the remainder of this paper assumes surrogate relaxation subproblems are best solved via a branch-and-bound procedure.

Role of the Primal Incumbent in $(P^{\nu}(T))$

In branch-and-bound procedure, the set of feasible solutions to (P) is partitioned into independent subsets by an enumeration which places additional constraints on integer variables. The unenumerated portion of (P) is represented by a list of *candidate problems*, each of which is simply (P) with certain additional constraints $x \in T$ appended. To facilitate the discussion, we define $P(T)$ to be the same as (P) except that x is restricted to $x \in T$. We also define $v^*(P)$ to be the value of the best currently known feasible solution to (P), i.e. the value of the incumbent solution used to provide an upper bound on the optimal solution value.

Note that $v(D_{\nu}(T))$ is being employed as a bound for some candidate problem $P(T)$ in the primal branch-and-bound procedure. However, $v(P^{\nu}(T))$ is a valid bound in $P(T)$ for all $\nu \geqslant 0$, not just the ν which maximizes $v(P^{\nu}(T))$. Thus, $(D_{\nu}(T))$ need not be solved optimally if $v(P^{\nu}(T))$, for some ν used on the way to solving $(D_{\nu}(T))$, is sufficient to fathom $P(T)$, i.e., $v(P^{\nu}(T)) \geqslant v^*(P)$.

Conversely, the value of the incumbent in the primal, $v^*(P)$, may be used as an upper bound in solving any (P^{ν}). That is, if no completion of a candidate problem in (P^{ν}) can produce a solution with value less than $v^*(P)$, that candidate problem in (P^{ν}) may be fathomed. If all candidate problems in the knapsack $(P^{\nu}(T))$ fail to produce a solution with value less than $v^*(P)$, then it can be concluded that $v(P^{\nu}(T)) \geqslant v^*(P)$ so that the candidate problem $P(T)$ may be fathomed in the primal.

3. CONDITIONAL BOUNDS AND BRANCHING VARIABLES

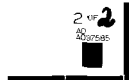
The rationale for the interaction between the two branch-and-bound procedures with respect to conditional bounds and branching rules can perhaps best be understood via a 0-1 integer programming example. Later a procedure for the general case will be presented. Consider Figure 1 which presents a branch-and-bound tree for the problem $(P^{\nu}(T))$ where $P(T)$ is a given candidate problem from the primal tree. This tree may result from the application of any branch-and-bound procedure for solving $(P^{\nu}(T))$. The solution is found at node 8 with value v^8. Since the full tree is shown and an optimal solution has been found, $v^1, v^2, \ldots, v^7$ must all be $\geqslant v^8$.

Now a number of important observations may be made. If v^8 is accepted as the optimal solution value for $(D_{\nu}(T))$ and the candidate problem $P(T)$ is not able to be fathomed ($v^8 < v^*(P)$) then a branching variable must be chosen and a conditional bound computed for each of the two new nodes created in the primal tree. Note that if x_1 is chosen as the branching variable, then a valid bound on any solution to $(P(T \cap \{x: x_1 = 1\}))$ is given by $v = \text{Min}(v^2, v^3)$. Also, since v^8 was the optimal value of $(P^{\nu}(T))$, $v \geqslant v^8$. So even though $v^8 < v^*(P)$, it is possible that the bound $v \geqslant v^*(P)$ so that no completion of $P(T \cap \{x: x_1 = 1\})$ will ever need be considered. It follows that x_1 is a good candidate for a branching variable in the primal tree. Note that a conditional bound for branching on x_2 may be taken as Min (v^5, v^7, v^3) for $x_2 = 0$ and Min (v^5, v^8, v^2) for $x_2 = 1$. One problem is that all of the end nodes for which x_2 is a free variable must be included (hence v^5) in calculating both bounds. x_1 is the only variable for which no free end nodes may exist, and we will choose it as the branching variable.

What is required to implement the branching procedure suggested above is the saving of the minimum value or bound on the end nodes for each of the two sides of the tree defined by the first branching variable. An end node may be recognized as one from which a fathoming occurs. Thus, before fathoming it is necessary to determine which side of the tree one is on, check to see if the bound on that node is less than the saved bound for that side of the tree

AD-A097 585 OFFICE OF NAVAL RESEARCH ARLINGTON VA F/G 5/1
NAVAL RESEARCH LOGISTICS QUARTERLY. VOLUME 28. NUMBER 1.(U)
MAR 81 S M SELIG

UNCLASSIFIED NL



and if necessary, replace that saved bound. Then after solving $(P^{\nu}(T))$ one will have $\nu(P^{\nu}(T))$ as the bound on one side (ν^8 in Figure 1), and a bound saved for branching on the nonoptimal side of the tree ($\text{Min}(\nu^2, \nu^3)$ in Figure 1).

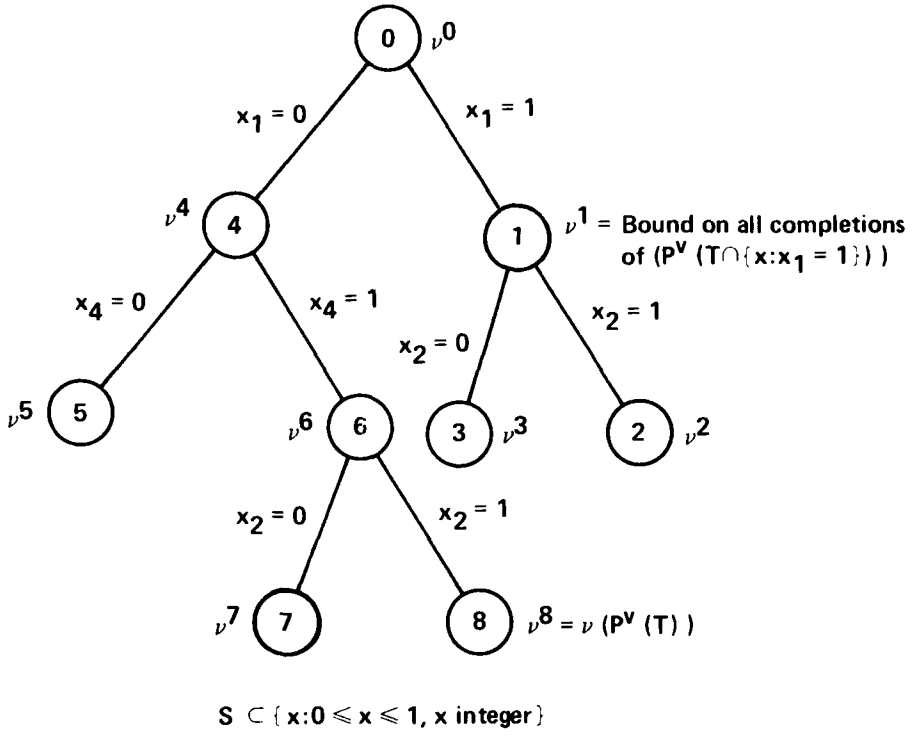


FIGURE 1. Example of a Branch-and-Bound Tree for $(P^{\nu}(T))$

4. INTERACTION OF THE SURROGATE SEARCH MASTER PROBLEMS

The two surrogate dual algorithms which appear most promising as discussed in Karwan and Rardin [6] both keep a list of the x's generated by each surrogate relaxation and solve a master problem involving these x's to obtain a new surrogate multiplier v. These master problems, one for each candidate problem in a primal branch-and-bound procedure, may be seen to interact in such a way as to save a great deal of time in solving (D_S) at any proceeding node in a primal tree.

Consider the primal branch-and-bound tree shown in Figure 2 for a 0-1 integer linear programming problem. Assume that a master problem, or at least a list of the x^i generated in solving $(D_S(\phi))$ at node 0, has been kept and it is now time to branch on x_1. Scan the master problem at node 0 and place all x^i, $i=1, 2, \ldots, k$ which satisfy $x_1^i = 0$ in a new master problem for solving $(D_S(T))$ at node 1 of the primal tree. All solutions $x \in S$, $x_1 = 1$ such that $cx < \nu(D_S(\phi))$ have been made infeasible by the optimal surrogate multiplier at node 0. If one is to use $\nu(D_S(\phi))$ as a bound after branching on x_1, then all of these x's must be included in the master problem at node 1. This is valid since the candidate problem at node 1 is a more constrained version of (P), and all the x's put in the master problem satisfy this extra constraint.

This procedure may be continued as follows. In solving $(D_S(T))$ at node 1, possibly more x's are generated. When branching to node 2, all x's in the master problem at node 1 with $x_4 = 0$ may be put in the master problem to begin solving the surrogate dual at node 2.

Any candidate problem may be chosen to be explored next in a branch-and-bound procedure and a number of strategies have been suggested. The "last-in first-out" or LIFO procedure always chooses the most recently added member of the candidate list to explore. Referring to Figure 2, the nodes have been numbered in the order in which a LIFO procedure might explore them. Hence, the order of branching is from node 0 to node 1 to node 2 and to node 3 at which time node 3 is fathomed, either because the incumbent solution to (P) was exceeded, a feasible solution was obtained, or it was determined that $x_1 = 0$, $x_4 = 0$ and $x_3 = 1$ precluded any feasible solution to (P). Thus "back-tracking" goes to node 4 which is also fathomed, leading back to node 5. In a LIFO procedure note that there are never more than two nodes at any given level of the tree, a level being defined by the number of fixed variables or extra constraints on (P). For instance in Figure 2, the fathoming of nodes 2 and 5 must occur before node 6 is chosen as the node from which to branch. In large integer programming problems, where many x's from previous surrogate master problems are to be stored, storage can be a main concern and it is minimized by using the LIFO branching procedure.

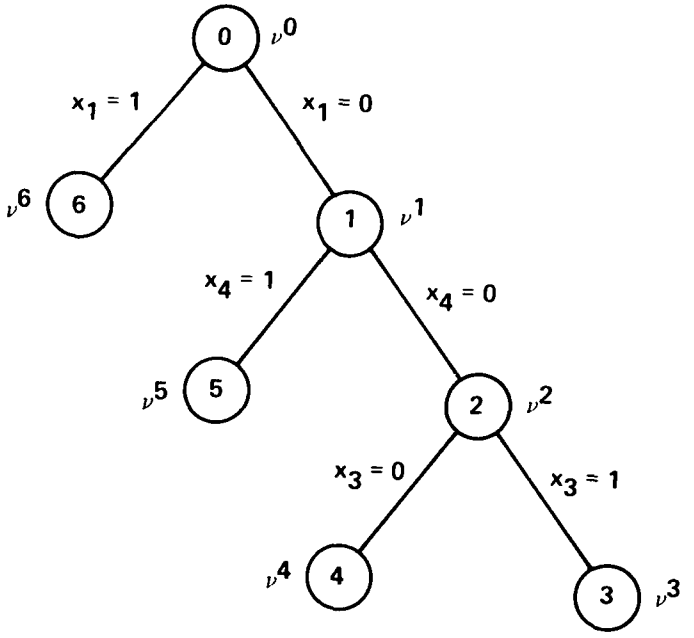


FIGURE 2 Example of a Primal Branch-and-Bound Tree

The master problem interactions can be shown to be very efficient in terms of a LIFO branching procedure for (P). Again consider Figure 2 and the following use of a "current table" and a "save table." At node 0, the master problem consists of the following x's, say for n = dimension of x = 5.

x^1 0 0 1 1 1
x^2 1 0 1 0 1
x^3 0 1 1 0 0
x^4 1 1 0 1 0

Branching takes place to node 1. Those x's which have $x_1 = 0$ (x^1 and x^3) are placed in the "current table" for the "current" or next-to-be-explored candidate problem. The other x's (x^2 and x^4) are placed in the "save table" and it is noted that at level 1 of the tree, the next open slot in the save table is in row 3. Node 1 is now explored and some new x's are generated and put in the current table which becomes

x^1 0 0 1 1 1
x^3 0 1 1 0 0
x^5 0 0 1 0 1
x^6 0 1 0 1 1

Now it is time to branch to node 2, so those x's which have $x_4 = 0$ remain in the current table, i.e., x^3 and x^5. x^1 and x^6 are placed in the save table and it is noted that the next open slot in the save table at level 2 of the tree is 5.

The current table is now

x^3 0 1 1 0 0
x^5 0 0 1 0 1

and the save table is

x^2 1 0 1 0 1 $L = 1$
x^4 1 1 0 1 0
- - - - - - - - - - - - - - -
x^1 0 0 1 1 1 $L = 2$
x^6 0 1 0 1 1

Assume that, in contrast to Figure 2, fathoming occurs at node 2, possibly after generating some more x's. Now the current table can be cleared since it is no longer necessary to explore any candidate problem with $x_1 = 0$ and $x_4 = 0$. In fact, these x's will never be generated or needed again, since either x_1 or x_4 or both will always be fixed at 1 in any future candidate problems. Now the LIFO branching procedure goes to node 5 with $x_1 = 0$ and $x_4 = 1$. But some of these x's are stored in the save table from the last slot in the save table ($5 - 1 = 4$) back to the next available slot stored after the previous level, level 1, which is slot number 3. These are put in the current table which is now

x^1 0 0 1 1 1
x^6 0 1 0 1 1

and the save table is now

x^2 1 0 1 0 1
x^4 1 1 0 1 0

Possibly more x's are generated at node 5 and placed in the current table. A fathoming then occurs at node 5 and a "backtracking" takes place to node 6. The "other side" of level 2 has been explored so the backtracking must be to level 1. The current table is again cleared and the elements in the save table from the last slot to the first slot for level 1 savings (slot 1) are placed in the current table. The procedure continues, with only two lists being necessary to

easily store, update, and use all of the x's generated by solving surrogate relaxations throughout the primal branch-and-bound procedure. Note that no x's will be regenerated using this procedure, and again that once the current candidate problem is fathomed those x's may be taken out of storage.

The following is a formal outline for branching and fathoming while employing the current and save tables in a LIFO branching procedure for a general integer linear programming problem. Let

L	= current level in primal branch-and-bound tree
T_{tI}^1, T_{tI}^2	= two new candidate problems created at level L, T_{tI}^1 is candidate problem chosen to explore next
SAVBND (L)	= bound saved for candidate problem at level L which is not being explored next
NXSV	= next available slot of the save table
NXCR	= next available slot of the current table
NSV(L)	= next available slot of the save table at level L in the primal tree
$v^*(P)$	= incumbent solution value to (P)

Branching:

If SAVBND (L) $< v^*(P)$, place all x's from the current table satisfying $x \in T_{tI}^2$ into the save table, updating NXSV. In any case, let NSV (L) = NXSV and remove all x's satisfying $x \in T_{tI}^2$ from the current table, closing up the current table and updating NXCR. Determining if $x \in T_{tI}^2$ is done simply by checking the single component of x upon which the branching occurred.

Fathoming:

Clear the current table by setting NXCR = 1. (If T_{tI}^2 has already been explored, SAVBND(L) = $+\infty$.) If SAVBND(L) $\geq v^*(P)$ replace NXSV by NSV($L-1$) and L by $L-1$ until a candidate problem is found to explore. Place rows NSV($L-1$) to NSV(L)-1 from the save table into the current table. Update NXCR.

After branching or fathoming more x's are generated while solving ($D_\zeta(T)$) and placed in the current table until it is time to branch or fathom again.

Although formally developed here for a LIFO branching procedure, the current and save table concept can be used for any primal branching procedure (e.g., least lower bound) by scanning a single save table for x's which satisfy the constraints on the present candidate problem. As seen above, this scanning is done very efficiently in the LIFO procedure by simply keeping an indicator (NSV(L)) for each level (L) of the primal tree. In any case, when a candidate problem is fathomed, the appropriate x's may be taken out of storage and will never be needed or regenerated again.

5. COMPUTATIONAL ANALYSIS

A set of randomly generated 0-1 integer programming test problems (see Karwan [5]) was used to demonstrate the developments discussed in this paper. A LIFO branching procedure was employed in the primal branch-and-bound tree and the LRMP procedure, (see Karwan and Rardin [6]) was the surrogate dual multiplier search procedure employed.

Table 1 presents the results of employing the above techniques on three problem sizes with a low and a high density and five replications per cell. One of the principal causes for interest in surrogate duals is improvement in bounds. The percent of the LP to IP gap closed by the surrogate dual, i.e.,

$$(v(D_S) - v(LP))/(v(P) - v(LP))$$

appears substantial. The large range for a given cell is perhaps to be expected with such unstructured randomly generated problems.

Some measure of the efficiency of the interaction between the primal and the subproblem branch-and-bound procedures is provided by the remaining columns of Table 1. As expected, the principal part of all time spent on candidate problems is consumed in knapsack subproblems. Values in column 8 range from 71%-82%. However, the number of knapsack subproblems solved at any particular node is quite small (column 6). The small numbers are a consequence of the save table—current table scheme developed in Section 4. Another indication of the efficiency of the save table approach is the relation between the mean time to solve the first surrogate dual (column 4) and the mean time to solve all surrogate duals (column 7). For larger problem sizes the average surrogate dual—which begins with many x^k saved from previous knapsacks—solves in 5-10% of the time for the first dual.

TABLE 1. *Primal Branch-and-Bound Empirical Results*
(Five Replications per cell)

(1) Problem Size: Density	(2) % of LP to IP Gap closed by D_S Average [Range]	(3) Total Time of Branch-and-Bound Procedure (Sec.)	(4) Time for First D_S (i.e., $D_S(0)$) (Sec.)	(5) Total Nodes Explored	(6) Knapsacks Solved Per Node	(7) D_S Time/ Node (Sec.) (% of 1st Node)	(8) Percent in Knapsack Time/Node
5 × 10	74.8	.034	.024	2.0	2.67	.013	72.3
10	[8.3, 100]					(54.2)	
10 × 20	15.2	3.20	.525	18.0	5.31	.122	71.3
10	[3.7, 48.8]					(3.8)	
15 × 30	5.9	35.93	3.10	109.4	5.69	.337	73.0
10	[1.7, 14.0]					(10.9)	
5 × 10	51.7	.140	.057	4.8	3.37	.028	82.1
25	[11.4, 100]					(20.0)	
10 × 20	13.4	6.18	1.26	44.0	3.88	.139	79.1
25	[10.8, 28.1]					(11.0)	
15 × 30	6.6	107.3	4.92	308.2	4.18	.325	79.7
25	[3.3, 8.7]					(6.6)	

REFERENCES

[1] Cabot, V.A., "An Enumeration Algorithm for Knapsack Problems," Operations Research, *18*, 306-311 (1970).

[2] Fayard, D. and G. Plateau, "Resolution of the 0-1 Knapsack Problem: Comparison of Methods," *Mathematical Programming*, *8*, 272-307 (1975).

[3] Garfinkel, R.S. and G.L. Nemhauser, *Integer Programming*, (John Wiley and Sons, New York, N.Y., 1972).

[4] Greenberg, H. and R.L. Hegerick, "A Branch Search Algorithm for the Knapsack Problem," Management Science, *16*, 327-332 (1970).

[5] Karwan, M.H., "Surrogate Constraint Duality and Extensions in Integer Programming,"

Ph.D. dissertation, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia (1976).

[6] Karwan, M.H., and R.L. Rardin, "Surrogate Dual Multiplier Search Procedures in Integer Programming," School of Industrial and Systems Engineering, Report Series No. J-77-13, Georgia Institute of Technology, Atlanta, Georgia (1976).

[7] Kolesar, P.J., "A Branch and Bound Algorithm for the Knapsack Problem," Management Science, *13*, 723-735 (1967).

EXTREME SOLUTIONS OF THE TWO MACHINE FLOW-SHOP PROBLEM

Wlodzimierz Szwarc

School of Business Administration
University of Wisconsin-Milwaukee
Milwaukee, Wisconsin

ABSTRACT

The paper provides a new theoretical framework to identify extreme solutions of the two machine flow-shop problem. Some remarkable properties of these solutions have been developed. As a result the problem of generating minimal solutions can be decomposed into a number of smaller subproblems.

1. INTRODUCTION

The well known two machine flow-shop problem can be formalized as follows. Find a permutation $P = p_1, p_2, \ldots, p_n$ of numbers $1, 2, \ldots, n$ that minimizes

$$T(P) = \max_{1 \leq u \leq n} \left\{ \sum_{r=1}^{u} A_{p_r} + \sum_{r=u}^{n} B_{p_r} \right\} \tag{1}$$

where A_r, B_r, $r = 1, 2, \ldots, n$ are given positive numbers. According to the flow-shop terminology $T(P)$ is the completion time of n items processed in a sequence $p_1, p_2, \ldots, p_n$ while A_r and B_r are operation times of item r on machines A and B. Each item is to be processed first on A, then on B.

Let $I = (1, 2, \ldots, n)$ be the set of all items, and i and j two arbitrary items of I. Introduce the following relation

$$R(i,j) \overset{df}{=} [\min(A_i, B_j) \leqslant \min(A_j, B_i)]. \tag{2}$$

Notice that $R(i,j)$ or $R(j,i)$ holds for every pair $i,j \in I$. We say that $P = p_1, p_2, \ldots, p_n$ is an $R-$sequence if

$$i < j \Rightarrow R(p_i, p_j), \ \forall\, i, j \in I. \tag{3}$$

As shown in Section 2 every R-sequence minimizes (1). The set of R-sequences is usually a small portion of the set of all minimal solutions.

This paper examines the properties of extreme (minimal, maximal) solutions of the flow-shop problem. It provides necessary and sufficient minimality conditions (Section 3) simpler than those of [4] along with sufficient maximality conditions (Section 6). It introduces a critical item concept (Section 4) that leads to several remarkable properties of extreme solutions. As a result the problem of generating minimal solutions can be decomposed into several smaller several smaller subproblems (Section 5).

2. *R*-SEQUENCES

Let Π be the set of all n-element permutations of $1, 2, \ldots, n$. We will use symbols P, Q, P', Q', P_k to indicate those permutations. Let $Q = \sigma ij\pi$, and $Q' = \sigma ji\pi$ be two elements of Π.

LEMMA 1: $(2) \Rightarrow [T(Q) \leqslant T(Q')], \forall \sigma ij\pi, \sigma ji\pi \in \Pi$.

PROOF: Due to (1)

$$T(Q) = \max\left[\sum_{r \in \sigma ij\pi} A_r + T(\pi), \sum_{r \in \sigma i} A_r + T(ij) + \sum_{r \in \pi} B_r, T(\sigma) + \sum_{r \in ij\pi} B_r\right]^*$$

$$T(Q') = \max\left[\sum_{r \in \sigma ji\pi} A_r + T(\pi), \sum_{r \in \sigma j} A_r + T(ji) + \sum_{r \in \pi} B_r, T(\sigma) + \sum_{r \in j\pi} B_r\right].$$

Consider inequality

$$T(ij) \leqslant T(ji) \tag{4}$$

which is equivalent to (2).† The theorem holds since (4) implies $T(Q) \leqslant T(Q')$.

Let $P = p_1, p_2, \ldots, p_n$ be an *R*-sequence

THEOREM 1: P minimizes (1).

PROOF: Consider an arbitrary sequence $P' \in \Pi$, $P' \neq P$. Then $P' = \sigma p_i p_j \pi$ for some $i < j$. According to (3) and Lemma 1, $T(\sigma p_i p_j \pi) \leqslant T(P')$. Hence, P' along with every permutation other than P can be eliminated from Π as nonoptimal. The well known Johnson's Algorithm [2] of constructing sequence $P = p_1, p_2, \ldots, p_n$ can be defined in the following manner:

STEP 1: Find $\min(\min_r A_r, \min_r B_r)$.

STEP 2:

(a) If the minimum is at $A_q \leqslant B_r$, define $J = \{p_1\}$ where p_1 is the element with the smallest subscript among the elements of the set $\{r \mid A_r = A_q\}$.

(b) If the minimum is at $B_q < A_r$, define $J = \{p_n\}$ where p_n is the element with the largest subscript among the elements of the set $\{r \mid B_r = B_q\}$.

STEP 3: Replace I by $I - J$ and repeat Steps 1 and 2 until all elements of P are determined.

COROLLARY 1: Johnson's Algorithm produces an *R*-sequence.

---

*$T(\pi), T(\sigma), T(ij), T(ji)$ are defined by (1) for sequences π, σ, ij, ji. Hence,

$T(ij) = \max(A_i + B_j + B_\pi, A_i + A_j + B_\pi)$

$T(ji) = \max(A_j + B_i + B_\pi, A_j + A_i + B_\pi)$

†To see it subtract $A_i + B_j + A_j + B_\pi$ from both sides of (4)

PROOF: Let $P = 1, 2, \ldots, n$. Assume that $R(i,j)$ does not hold for some $i < j$. Then $\min(A_i, B_j) > \min(A_j, B_i)$. Consequently, Johnson's Algorithm will place element j in front of element i contrary to our assumption, Q.E.D.

Introduce the following notations:

$$E_r = A_r - B_r,\quad I = \{r \mid r \in I,\ E_r < 0\},$$

$$I^+ = \{r \mid r \in I,\ E_r > 0\},\quad I^0 = \{r \mid r \in I,\ E_r = 0\}.$$

Let $P = p_1, p_2, \ldots, p_n$ be an arbitrary R-sequence. Then the following obvious properties hold:

PROPERTY 1: The elements of I^- are arranged in a nondecreasing order of the A_r and precede the elements of I^+ that are arranged in a nonincreasing order of the B_r.

PROPERTY 2: The elements of I^0 can be placed in any order as long as they do not precede (follow) an item with a smaller $A_r(B_r)$.

PROPERTY 3: Any subsequence of P is an R-sequence.* Consider a sequence $\sigma, \sigma \subset I$, and an R-sequence π, where $\pi \subset I - \sigma$.

PROPERTY 4: $T(\sigma\pi) \leqslant T(\sigma\pi')$, $T(\pi\sigma) \leqslant T(\pi'\sigma)$ for all possible permutations π' of the elements of π.

PROOF: According to (1)

$$T(\sigma\pi) = \max\left[\sum_{\sigma} A_r + T(\pi), T(\sigma) + \sum_{\pi} B_r\right],$$

$$T(\sigma\pi') = \max\left[\sum_{\sigma} A_r + T(\pi'),\ T(\sigma) + \sum_{\pi'} B_r\right].$$

Hence, $T(\pi) \leqslant T(\pi') \Rightarrow T(\sigma\pi) \leqslant T(\sigma\pi')$, Q.E.D.

One can similarly prove $T(\pi\sigma) \leqslant T(\pi'\sigma)$. According to Property 4, to find a sequence that minimizes (1), provided σ is fixed, arrange the items that follow (precede) σ in an R-sequence.

This rule may not be valid if σ occupies a middle position. Consider the following example (Figure 1):

	A_r	B_r
1	2	3
2	6	8
3	9	10
4	3	1

FIGURE 1

*This does not mean that a subsequence of every optimal sequence is optimal (see Remark 1, Section 3).

Assume that we are to find a sequence that minimizes (1) where $\sigma = 3$ occupies the second place. Although 124 is the only R-sequence of $I - \sigma = (1, 2, 4)$, 1324 is not the best sequence since $T(2314) = 29 < T(1324) = 30$.

3. NECESSARY AND SUFFICIENT OPTIMALITY CONDITIONS

Define $W(\pi) = T(\pi) - \sum_{\pi} B_r$, for $\pi \subset I$. Then

$$W(P) = T(P) - \sum_{r=1}^{n} B_r = \max_{1 \le u \le n} \left[A_{p_u} + \sum_{r=1}^{u-1} E_{p_r} \right]. \tag{5}$$

Consequently, the minimization of (1) is equivalent to that of (5). As known $W(P)$ is the idle time of machine B while processing sequence P. Let σ and π be two sequences, $\pi \subset I - \sigma$.

PROPERTY 5: $W(\sigma\pi) \geqslant W(\sigma)$

PROOF: According to (5)

$$W(\sigma\pi) = \max\left[W(\sigma), \sum_{\sigma} E_r + W(\pi) \right], \text{Q.E.D.}$$

Observe that $W(\sigma\pi) \geqslant W(\pi)$ may not hold. Consider two sequences $P = \sigma\gamma i\pi$ and $Q = \sigma i\gamma\pi$, define the following conditions:

$$A_i \leqslant W(P) - \sum_{\sigma} E_r, \tag{6}$$

$$A_i - B_i \leqslant W(P) - \sum_{\sigma} E_r - W(\gamma). \tag{7}$$

For $P = \sigma ji\pi$ and $Q = \sigma ij\pi$ formula (7) becomes

$$A_i - B_i \leqslant W(P) - \sum_{\sigma} E_r - A_j. \tag{7a}$$

We will show

PROPERTY 6:

$$E_i \leqslant 0 \Rightarrow [(6) \Leftrightarrow [W(Q) \leqslant W(P)]],$$

$$E_i > 0 \Rightarrow [[(6) \text{ and } (7)] \Leftrightarrow [W(Q) \leqslant W(P)]].$$

PROOF: Due to (5)

$$W(P) = \max\left[W(\sigma), W(\gamma) + \sum_{\sigma} E_r, A_i + \sum_{\sigma\gamma} E_r, W(\pi) + \sum_{\sigma\gamma i} E_r \right],$$

$$W(Q) = \max\left[W(\sigma), A_i + \sum_{\sigma} E_r, W(\gamma) + \sum_{\sigma} E_r + E_i, W(\pi) + \sum_{\sigma i\gamma} E_r \right].$$

If $E_i \leqslant 0$ then $W(Q) \leqslant W(P)$ whenever $A_i + \sum_{\sigma} E_r \leqslant W(P)$. On the other hand if $A_i + \sum_{\sigma} E_r > W(P)$ then $W(Q) \geqslant A_i + \sum_{\sigma} E_r > W(P)$, Q.E.D.

One can similarly prove case $E_i > 0$.

Assume that P is an optimal sequence, which means that P minimizes (1) and (5).

COROLLARY 2: Q is optimal *if and only if* one of the following conditions hold:

1. (6) if $E_i \leqslant 0$, or

2. (6), (7) if $E_i > 0$.

Consider the following example (Figure 2):

	A_i	B_i
1	1	2
2	3	5
3	6	4
4	6	3

FIGURE 2

$P = 1234$ optimal and $W(P) = 5$. Let $\sigma = \phi$, $i = 2, 3$. Permutation 2134 is optimal due to $E_2 < 0$, and (6) $(3 < 5 - 0)$, while sequence 3124 is not optimal $(6 < 5 - 0)$.

Assume $\sigma = (1)$, $\gamma = (2)$, $i = 3$. Both conditions (6) and (7) are met $(6 \leqslant 5 - (-1)$, $6 - 4 \leqslant 5 - (-1) - 3)$. Consequently, 1324 is optimal. Observe that neither 2134 nor 1324 is an R-sequence.

REMARK 1: Although 1324 is optimal its subsequence 132 is not, since $T(132) = 16 > T(123) = 14$.

Usually the number of optimal solutions far exceeds the number of R-sequences. Consider the following case:

$$E_i > 0, \forall i, \max_i B_i < \min_j A_j, B_i \neq B_j, \forall i \neq j.$$

While there is only one R-sequence the number of *all* optimal solutions (where the last element is an item with the smallest B_i) is $(n-1)!$.

4. CRITICAL ITEMS

Element u is a *critical* item of an optimal sequence $P = \sigma u \pi$ if

$$W(P) = A_u + \sum_{\sigma} E_r \left(\text{or } T(P) = \sum_{\sigma u} A_r + \sum_{u\pi} B_r\right).$$

Assume in this section $E_r \neq 0$, $r \in I$. Let $P = \sigma_1 \sigma_2 iuj \pi_1 \pi_2$ be an R-sequence, and u its critical item. Suppose we move u upward in front of $\sigma_2 i$, or downward behind $j\pi_1$. Will the resulting sequence be optimal? The following theorems resolve this issue.

THEOREM 2: $Q = \sigma_1 u \sigma_2 ij \pi_1 \pi_2$ is optimal if and only if

(8) $$E_r > 0, \; B_r = B_u, \; r \in \sigma_2 iu.$$

PROOF:

$\Rightarrow$: 1. If $E_u < 0$, and $E_r < 0$, $r \in \sigma_1\sigma_2 i$. Hence,

$$W(Q) \geqslant W(\sigma_1 u) \geqslant A_u + \sum_{\sigma_1} E_r > A_u + \sum_{\sigma_1\sigma_2 i} E_r = W(P),$$

contrary to the assumption, Q.E.D.

2. If $E_i < 0$ one can show as before that $W(Q) > W(P)$, Q.E.D.

Since $E_u > 0$ and $E_i > 0$ then $B_i \geqslant B_u$ (Property 1).

3. If $B_i > B_u$ then

$$W(Q) \geqslant W(\sigma_1 u \sigma_2 i) \geqslant A_i + \sum_{\sigma_1 u \sigma_2} E_r = A_i + \sum_{\sigma_1\sigma_2} E_r + A_u - B_u >$$

$$> A_i + \sum_{\sigma_1\sigma_2} E_r + A_u - B_i = W(P), \text{Q.E.D.}$$

$E_i > 0$, $E_u > 0$ and $B_i = B_u$ imply (8) due to Property 1.

$\Leftarrow$: Condition (8) along with Property 1 imply that Q is an R-sequence. One can similarly prove the following.

THEOREM 3: $Q' = \sigma_1\sigma_2 ij\pi_1 u\pi_2$ is optimal if and only if

(9) $$E_r < 0, \quad A_r = A_u, \quad r \in uj\pi_1.$$

Consider sequences Q and Q' of Theorems 2 and 3.

PROPERTY 7: 1. If Q is optimal then i is its critical item,

2. If Q' is optimal then j is its critical item.

PROOF: The optimality of Q implies (8). Hence,

$$W(Q) \geqslant W(\sigma_1 u \sigma_2 i) \geqslant A_i + \sum_{\sigma_1 u \sigma_2} E_r = A_u + \sum_{\sigma_1\sigma_2 i} E_r = W(P) = W(Q), \text{Q.E.D.}$$

The proof of the second part is symmetrical.

Due to (8) and (9)

$$W(Q) > W(\sigma_1 u), \quad W(Q') > W(\sigma_1\sigma_2 ij\pi_1 u).$$

Hence, u is no longer a critical item of Q or Q'.

Suppose that we move element i of an R-sequence $P = \sigma_1\sigma_2 u\pi_1 i\pi_2$ ahead of its critical item u. The following theorem resolves the optimality issue of the resulting sequence.

THEOREM 4: $Q = \sigma_1 i\sigma_2 u\pi_1\pi_2$ is optimal if and only if

(10) $$E_u < 0, \quad E_i < 0, \quad \sigma_2 = \phi, \quad A_i = A_u.$$

PROOF:

$\Rightarrow$: If $E_r > 0$, then (Property 5) $W(Q) \geq W(\sigma_1 i \sigma_2 u) \geq A_u + \sum_{\sigma_1 i \sigma_2} E_r > A_u + \sum_{\sigma_1 \sigma_2} E_r = W(P)$, which is in contradiction with the optimality of Q. Hence, $E_i < 0$. This implies $E_r < 0$, $r \in \sigma_1 \sigma_2 u \pi_1$ and $A_i \geq A_u$ (Property 1).

$$W(Q) \geq W(\sigma_1 i) \geq A_i + \sum_{\sigma_1} E_r \geq A_u + \sum_{\sigma_1 \sigma_2} E_r = W(P).$$

Thus, $W(Q) > W(P)$ if $\sigma_2 \neq \phi$ or $A_i > A_u$, Q.E.D.

$\Leftarrow$: Due to Property 1 (P is an R-sequence) condition (10) implies $E_r < 0$, $A_i = A_u$, $r \in u\pi_1 i$. Hence, Q is an R-sequence, Q.E.D.

Optimal Presequences:

Given an R-sequence $P = \alpha\pi \in \Pi$, then π is also an R-sequence (Property 3). Consider a permutation $Q = \sigma\pi \in \Pi$.

We say that σ is an optimal presequence when $Q = \sigma\pi$ is optimal. Q is *uniquely* determined for each σ, once P is given. Hence, to find all optimal sequences it is sufficient to generate all optimal s-element presequences for each $s \leq n-1$, given an R-sequence P.

REMARK 2: According to Property 5 presequence σi may be optimal only if σ is optimal.

REMARK 3: Formulas (6) and (7) allow to determine the optimality of presequence σi provided

1. σ is already known to be an optimal presequence.
2. P is a known R-sequence.

Let $P = \alpha u \beta$ be an R-sequence and u its critical item. Consider another sequence Q.

THEOREM 5: If Q is optimal then

1. The elements of αu precede those of β whenever $E_u > 0$, or
2. The elements of $u\beta$ follow those of α whenever $E_u < 0$.

PROOF:

CASE $E_u > 0$: Let $Q = \sigma\pi$ where σ is an optimal presequence. Assume that αu is a k-element sequence ($k \leq n-1$). For each $s \leq k$ consider sets of s-element optimal presequences σ. According to Theorem 4 no element of β belongs to an optimal σ if $s = 1$. Due to the same Theorem and Remark 2 this is also true for $s = 2, 3, \ldots, k$, Q.E.D.

The proof of second case is symmetrical.

5. GENERATING OPTIMAL SEQUENCES

Consider an R-sequence $P = \sigma iuj\pi$ where the critical item u is the s-th element of P. Theorems 2, 3 and Property 7 imply:

COROLLARY 3: If none of the conditions (8) and ( ) holds then u remains the s-th element of *every* optimal sequence.

Element u is also a critical item of *every* optimal sequence. To see it assume that $Q = \alpha u \beta$ where α and β are permutations of elements of σi and $j\pi$, respectively. Then,

$$W(Q) \geqslant W(\alpha u) \geqslant A_u + \sum_{\alpha} E_r = A_u + \sum_{\sigma i} E_r = W(P) = W(Q), \text{Q.E.D.}$$

5.1 Let $P = \alpha_0 u_1 \alpha_1 u_2 \ldots \alpha_{q-1} u_q \alpha_q$ be an R-sequence where none of (8) and (9) hold for critical items $u_t, 1 \leqslant t \leqslant q$. We will show that the problem of generating optimal presequences $\sigma \subset I$ can be decomposed into $q + 1$ separate subproblems. Consider an optimal presequence $\sigma = \alpha_0' u_1 \alpha_1' u_2 \ldots \alpha_{t-1}' u_t \sigma_t$, $t \leqslant q$ where α_s' is a permutation of the elements of α_s while $\sigma_t \subset \alpha_t$, for each $0 \leqslant s \leqslant t$. Then,

$$W(P) = A_{u_t} + \sum_{\sigma} E_r - \sum_{u_t \sigma_t} E_r, \quad 1 \leqslant t \leqslant q. \tag{11}$$

Formulas (6) and (7) remain in their original form for $t = 0$ while for $t \geqslant 1$ they become

$$A_t \leqslant B_{u_t} - \sum_{\sigma_t} E_r, \tag{6'}$$

$$A_t - B_t \leqslant B_{u_t} - \sum_{\sigma_t} E_r - W(\gamma), \tag{7'}$$

where $\sigma_t \subset \alpha_t, \gamma_t \subset \alpha_t$. To illustrate the decomposition technique along with the generating procedure consider the example of reference [3] (Figure 3).

	A_r	B_r
1	2	3
2	4	5
3	6	30
4	30	4
5	4	1

FIGURE 3

$P = 12345$ is an R-sequence, $W(P) = 4$ and $u_1 = 3$, $u_2 = 4$, $u_3 = 5$. Consequently, $\alpha_0 = (1, 2)$, $\alpha_1 = \alpha_2 = \alpha_3 = \phi$. Since the assumptions of Corollary 4 are met for all u_t (they automatically hold for u_2 and u_3 since $\alpha_2 = \alpha_3 = \phi$) every optimal solution $Q = \ldots 345$. It only remains to find optimal one element presequences of α_0 since α_0 is a two element set. Due to $E_r < 0$, $r \in \alpha_0$ it is sufficient to check (6). Presequence 2 is optimal since (6) holds for $i = 2$, $\sigma = \phi$ (see Remark 3) in addition to the known optimal presequence 1. Consequently, 12 and 21 are optimal arrangements of α_0. There are only two optimal sequences 12345 and 21345.*

5.2. Consider some critical item u of an R-sequence P where

1. (8) holds for some $\sigma_2 i = i_1 i_2 \ldots i_s i$, or
2. (9) holds for some $j\pi_1 = j j_1 \ldots j_s$.

According to Theorems 2 and 3 we can generate R-sequences, say, P_k by arranging the elements of $\sigma_2 iu$ or $uj\pi_2$ of P in the following manner:

$$i_1 i_2 \ldots i_s iu, \quad u i_1 i_2 \ldots i_s i, \quad \ldots, \quad i_2 i_3 \ldots i_s iu i_1 \tag{12}$$

*The authors of [3] using a lexicographic search procedure examined (in this example) lower bounds for 9 presequences with the number of elements ranging from 3 to 5 (in 7 presequences).

(13) $uj j_1 \cdots j_\lambda,\ j j_1 \cdots j_\lambda u,\ \ldots,\ j_\lambda u j j_1 j_2 \cdots j_{\lambda-1}$

In view of Property 7 the critical items are the last elements of the sequences of (12) and the first elements of the sequences of (13)*.

To find the optimal permutations we apply the procedure of Section 5.1 to *each* P_λ assuming that none of its critical items can be moved.

To illustrate this case consider the following example (Figure 4):

	1	2	3	4	5	6	7
A_r	3	4	8	5	6	3	2
B_r	7	7	5	4	4	2	1

FIGURE 4

$P = \bar{1}234\bar{5}67$ is an R-sequence, $W(P) = 3$, $u_1 = 1$, $u_2 = 5$.

The dashes indicate the critical items of P.

Since (8) and (9) hold for $\sigma_2 i = (4)$ and $j\pi_1 = (2)$ four R-sequences are generated (see (12) and (13)).

$$P_1 = \bar{1}234\bar{5}67,\quad P_2 = \bar{2}134\bar{5}67,\quad P_3 = \bar{1}235\bar{4}67,\quad P_4 = \bar{2}135\bar{4}67.$$

To generate optimal sequences out of P_1 observe that $P_1 = \alpha_0 u_1 \alpha_1 u_2 \alpha_2$ where $\alpha_0 = \phi$ $\alpha_1 = (2,3,4)$ $\alpha_2 = (6,7)$. According to (6') and (7') the list of optimal arrangements of α_1 and α_2 is 234, 243, 423, and 67, 76 respectively. Consequently, P_1 generates six sequences $\bar{1}234\bar{5}67$, $\bar{1}243\bar{5}67$, $\bar{1}234\bar{5}67$, $\bar{1}423\bar{5}76$, $\bar{1}243\bar{5}76$, $\bar{1}423\bar{5}76$.

To find the remaining optimal sequences we have to verify (6') and (7') for $\alpha_1 = (1,3,4)$, $(2,3,5)$ and $(1,3,5)$ since $\alpha_2 = (6,7)$ is the *same* for all four sequences P_λ. The total number of optimal solutions is 24 while the number of R-sequences is 4.

5.3. Consider the case when $E_r = 0$ for $r \in I^0 \neq \phi$. Let P be an R-sequence. We can assume (Property 1) that $P = \alpha\beta\gamma$ where

$$E_r < 0,\ r \in \alpha,\quad E_r = 0,\ r \in \beta,\quad E_r > 0,\ r \in \gamma.$$

Let $\max_{r \in \beta} A_r = A_u$. Consider sequence $\alpha\gamma$.

THEOREM 6: $W(P) = \max[W(\alpha\gamma), A_u + \sum_{\alpha} E_r]$

PROOF: Let u be a critical item of P. Examine three cases:

1. $u \in \gamma$. Then $P = \alpha\beta\gamma_1 u\gamma_2$ and

$$W(P) = A_u + \sum_{\alpha\beta\gamma_1} E_r = A_u + \sum_{\alpha\gamma_1} E_r = W(\alpha\gamma).$$

2. $u \in \alpha$. The proof is similar to that of the previous case.
3. $u \in \beta$. $P = \alpha\beta_1 u\beta_2\gamma$, and

$$W(P) = A_u + \sum_{\alpha\beta_1} E_r = A_u + \sum_{\alpha} E_r.$$

Expression $A_u + \sum_{\sigma} E_r$ is maximal for $u = v$, Q.E.D.

Theorems 2, 3, and 4 remain valid even for $E_u = 0$ as long as $E_r \neq 0$, for $r \in I - u$.

We offer the following procedure of generating optimal sequences:

STEP 1: Delete set I^0 from I and find an R-sequence $\alpha\gamma$.

STEP 2: Apply the generating procedure of Section 5.2 to sequence π where

$$\pi = \begin{cases} \alpha\gamma \text{ if } A_v + \sum_{\sigma} E_r < W(\alpha\gamma), \\ \alpha v\gamma \text{ if } A_v + \sum_{\sigma} E_r \geq W(\alpha\gamma). \end{cases}$$

STEP 3: For each sequence π generate n-element optimal sequences by placing the remaining items of β in the appropriate places using formula (6).

To illustrate the procedure expand the example of Figure 2 by adding two new elements 5 and 6 where $A_5 = B_5 = 8$, $A_6 = B_6 = 5$.

STEP 1: We already know that $\alpha\gamma = 1234$ is an R-sequence.

STEP 2: $\pi = 12534$ since $A_5 + \sum_{\sigma} E_r = W(\alpha\gamma) = 5$.

Observe that $\alpha_3 = (3)$, and elements 5, 3, 4 cannot be moved. Handling set $\alpha_0 = (1, 2)$ we obtain two optimal sequences $12\bar{5}\bar{3}\bar{4}$ and $21\bar{5}\bar{3}\bar{4}$.

STEP 3: Condition (6) for $i = 6$, $W(P) = W(\alpha\gamma) = 5$ becomes

$$5 \leq 5 - \sum_{\sigma} E_r. \tag{14}$$

Consider Figures 5 and 6 where the $\sum_{\sigma} E_r$ are written on the margins of the tables (except $\sum_{\sigma} E_r = 0$ for $\sigma = \phi$).

1	1	2	-1
2	3	5	-3
5	8	8	-3
3	6	4	-1
4	6	3	2

FIGURE 5

2	3	5	-2
1	1	2	-3
5	8	8	-3
3	6	4	-1
4	6	3	2

FIGURE 6

According to (14) presequence $\sigma 6$ is optimal if and only if $\sum_{\sigma} E_r \leq 0$. Hence, element 6 can be placed everywhere as long as it precedes element 4. Consequently, there are ten optimal sequences.

6. MAXIMAL SOLUTIONS

Let P be an R-sequence produced by Johnson's Algorithm. It is easy to see that a reversed sequence $P' = p_1, p_2, \ldots, p_n$ maximizes (1). Without loss of generality we can assume $p = 1, 2, \ldots, n$, and

$$\left.\begin{array}{c} E_r > 0,\ r \leqslant t,\ E_r \leqslant 0,\ r \geqslant t+1, \\ B_1 \leqslant B_2 \leqslant \ldots \leqslant B_t,\ A_{t+1} \geqslant A_{t+2} \geqslant \ldots \geqslant A_n \end{array}\right\} \tag{15}$$

for some $0 \leqslant t \leqslant n$.*

THEOREM 7: Element t or $t+1$ is a critical item of P'.

PROOF: Define

$$K_i = \sum_{r=1}^{i} A_r - \sum_{r=1}^{i-1} B_r$$

then,

$$W(P') = \max_{1 \leqslant i \leqslant n} K_i$$

CASE 1: $1 \leqslant i \leqslant t$. Then $B_i \leqslant B_{i+1}$ and $B_{i+1} < A_{i+1}$ imply $B_i < A_{i+1}$. Consequently, $K_i < K_{i+1}$.

CASE 2: $t+1 \leqslant i \leqslant n$. Then, $A_i \leqslant A_{i+1}$ and $A_{i+1} \leqslant B_{i+1}$ imply $A_i \leqslant B_{i+1}$. Hence, $K_i \leqslant K_{i+1}$. Combining both cases we have

$$K_1 < K_2 < \ldots < K_t \text{ and } K_{t+1} \geqslant K_{t+2} \geqslant \ldots \geqslant K_n.$$

Thus, $W(P') = \max_{t \leqslant i \leqslant t+1} K_i$, Q.E.D.

Let u be the critical item of a reversed Johnson sequence $P' = 1, 2, \ldots, n$. It is easy to see that any sequence $Q = \alpha u \pi$ maximizes (1) as long as α is a permutation of $1, 2, \ldots, u-1$ while β is a permutation of $u+1, \ldots, n$.

COROLLARY 4: The minimum number of maximal sequences is $(u-1)!(n-u)!$, where u=t or t+1

Reversing a minimal (non R) sequence does not *necessarily* produce a maximizing sequence. Consider the example on Figure 2. Although 1324 is minimal the reversed permutation 4231 does not maximize (1) since $T(4231) = 21 < T(4321) = 23$. Let $P' = 1, 2, \ldots, n$ be a reversed Johnson's sequence. Consider a set of $n-1$ — element permutations π of numbers $r \in I-(i)$. It is obvious that $\pi' = 1, 2, \ldots, i-1, i+1, \ldots, n$ maximizes $W(\pi)$.

Due to Theorem 7

$$W(P') = K_u,\ u = t, t+1$$

where

$$\left.\begin{array}{c} A_{t-1} \leqslant B_t,\ A_t \geqslant B_{t+1}, \text{ if } u = t \\ A_{t-1} \geqslant B_t,\ A_{t-2} \leqslant B_{t+1}, \text{ if } u = t+1 \end{array}\right\} \tag{16}$$

*For $t = 0$ no $I_r < 0$ while for $t = n$ no $I_r > 0$.

while

$$B(\pi') = \bar{K}_{v}^{dl} K_{v} \quad E_{t}$$

where

$$v = \begin{cases} u, & \text{if } t \neq u, \\ u-1 \text{ or } u+1, & \text{if } t = u. \end{cases}$$

THEOREM 8. $B(\pi) \leqslant B(P)$.

PROOF. There are two cases:

$t \neq u$: 1. If $t < u$ then $t \leqslant t_v$ and $E_t > 0$. Hence, $B(P) - B(\pi') = E_t > 0$.
2. If $t > u$ then $B(P) = B(\pi')$. Q.E.D.

$t = u$: Four sub cases are to be considered:

1. $u = t$, $v = t-1$; 2. $u = t$, $v = t+1$; 3. $u = t+1$, $v = t$; 4. $u = t+1$, $v = t+2$.
$B(P) - B(\pi') = B - A$ (case 1), $A_{t+1} - A$ (case 2), $B - A_{t+1}$ (case 3), $A_{t+2} - A_{t+1}$ (case 4).

The nonnegativity of $B(P) - B(\pi')$ follows directly from
1. (16) for cases 1 and 3.
2. (15) and (18) for cases 2 and 4 ($A_{t+1} \leqslant B_t$, $B_t < A_t$ – case 2).

Let σ be a permutation of numbers $i \in J$ where J is a proper subset of I. Theorem 8 implies

PROPERTY 8: $\max_{\sigma} W(\sigma) \leqslant \max_{P} W(P)$.

Property 8 does not imply $W(\sigma) \leqslant W(P)$. To see it consider the following example

	A_i	B_i
1	9	5
2	4	6
3	6	8

FIGURE 7

Let $P = 123$ and $\sigma = 13$. Still $W(P) = 9 < W(\sigma) = 10$.

BIBLIOGRAPHY

[1] Bagga, P.C., "The Two Stage Production Schedules with All Optimals," *Advancing Frontiers in Operational Research*, (Hindustan Publishing Corporation, New Delhi, India, 145-150 1969).

[2] Johnson, S.H., "Optimal Two and Three Stage Production Schedules with Set-up Times Included," *Naval Research Logistics Quarterly*, *1*, 61-68 (1954).

[3] Pandit, S.N.N. and Y.V. Subrahmanyam, "Enumeration of All Optimal Job Sequences," Opsearch, *12*, 33-39 (1975).

[4] Potts, C.N., "The Two Machine Permutation Scheduling Problem," submitted for publication.

A THEORETICAL AND COMPUTATIONAL COMPARISON OF "EQUIVALENT" MIXED-INTEGER FORMULATIONS*

R. R. Meyer

University of Wisconsin
Madison, Wisconsin

ABSTRACT

This paper provides a theoretical and computational comparison of alternative mixed integer programming formulations for optimization problems involving certain types of economy-of-scale functions. Such functions arise in a broad range of applications from such diverse areas as vendor selection and communications network design. A "nonstandard" problem formulation is shown to be superior in several respects to the traditional formulation of problems in this class.

1. FORMULATIONS: EQUIVALENT AND OPTIMAL

This paper provides a theoretical and computational comparison of alternative mixed integer programming formulations for optimization problems involving certain types of economy-of-scale functions. Such functions arise in a broad range of applications from such diverse areas as vendor selection and communications network design. A "*non*standard" problem formulation is shown to be superior in several respects to the traditional formulation of problems in this class.

This first section describes a rigorous approach to formulating certain optimization problems through the use of "minimization models" [4,5,6]. The minimization model concept is then used as the basis for defining a family of "equivalent" formulations as well as a means of defining an "optimal" formulation. Sections 2 and 3 establish the optimality of a very compact formulation for functions belonging to a class of economy-of-scale functions. Computational results for a communications network problem are then given to illustrate the superiority of this formulation as compared to a "standard" formulation of the problem.

The *economy-of-scale property* that we will consider is encountered in a broad variety of cost functions for goods ranging from doughnuts to telecommunications links. Roughly speaking, a function is said to have an economy-of-scale property if the cost (per unit) of a commodity decreases if certain "large" quantities of the commodity are purchased. A simple example of such a cost function, but one which serves to illustrate some of the properties that we wish to consider, is a "cheaper-by-the-dozen" function defined as follows: let x_1 denote the number of *single* units of a commodity with the cost per *single* unit being a *positive* constant c_1; let x_{12} denote the (nonnegative, integer) number of *dozens* (groups of 12) purchased, the price per dozen being a positive constant $c_{12} < 12c_1$ (so that it is *cheaper* to purchase a dozen than it is to

*This research was supported by the National Science Foundation under grant MCS 7605841 A01.

purchase 12 single units), and let $k(x)$ (see Figure 1 for a typical $k(x)$) denote the "cheaper-by-the-dozen" function that represents the *minimum* cost of purchasing *at least* x units. For simplicity in this example, x and x_1 will be assumed to be *continuous* variables. It is easily seen that $k(x)$ can be compactly represented as

$$k(x) = \min \quad c_1 x_1 + c_2 x_2 \qquad (1.1)$$

$$\text{s.t.} \quad x_1 + 12 x_2 \geq x$$

$$x_1, x_2 \geq 0, \quad x_2 \text{ integer.}$$

That is, substituting any constant x for x in the right hand side of (1.1) yields an optimization problem (in the variables x_1 and x_2) whose optimal value is precisely $k(x)$. Of course, the piecewise-linear function $k(x)$ can be represented in many other ways, but, as will be seen, the representation (1.1) is not only compact but also is useful in formulating optimization problems involving $k(x)$.

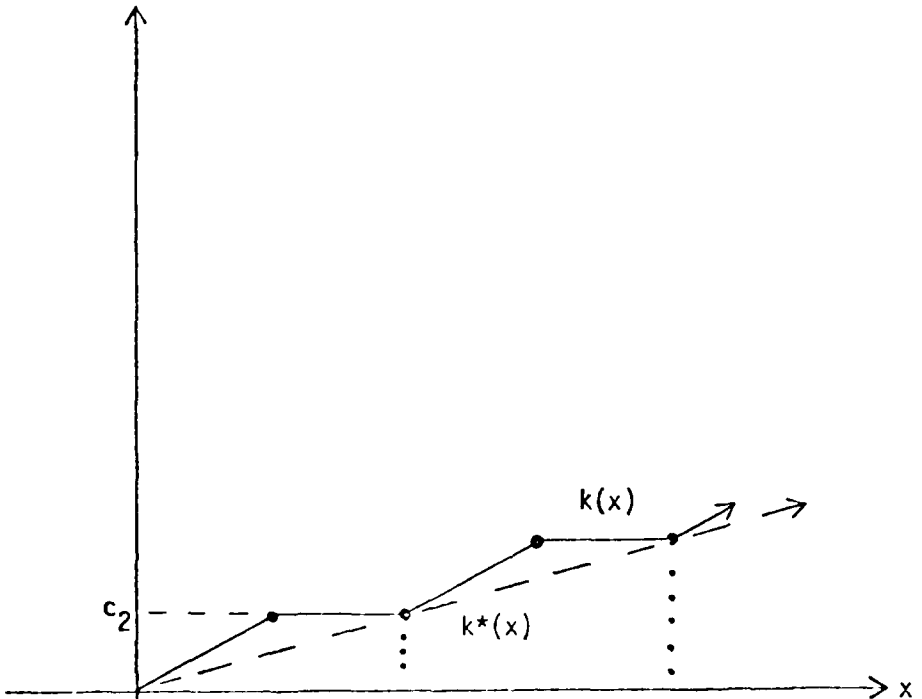


FIGURE 1. Functions $k(x)$ and $k^*(x)$

The RHS of (1.1) is an example of a *mixed-integer minimization model* (MIMM), a concept that was described in [4,5,6]. To define this concept, suppose f is a function from $\mathbb{R}^1$ into $(-\infty, +\infty]$, and that the following equation holds for all x belonging to a subset S of $\mathbb{R}^1$:

$$f(x) = \min \quad cy \tag{1.2}$$

$$\text{subject to} \quad Ay = b - \hat{A}x,$$

$$y \geq 0 \text{ and } y_i \text{ integer for } i \in I,$$

where I is a subset of $\{1, \ldots, n\}$, b is an element of $\mathbb{R}^m$, and c, A, and $\hat{A}$ are of dimensions $1 \times n$, $m \times n$ and $m \times 1$ respectively. (We will assume that the optimization problem on the rhs on (1.2) has an optimal solution if its feasible set is nonempty, and that the "optimal value" is defined to be $+\infty$ if the feasible set is empty.) The expression on the rhs of (1.2) is said to be a *MIMM for f on S* (for our purposes, it is convenient to assume that S is *convex*, although for the general theoretical development of MIMM's given in [4], this is not necessary). As noted in [4], the utility of MIMM's arises in part from the fact that for any set $\tilde{S} \subseteq S$ the following two problems are *equivalent*:

$$\min f(x) + \hat{f}(x,z) \tag{1.3}$$

$$\text{s.t. } x \in \tilde{S}, \ (x,z) \in T,$$

and

$$\min cy + \hat{f}(x,z) \tag{1.4}$$

$$\text{s.t. } x \in \tilde{S}, \ (x,z) \in T,$$

$$Ay = b - \hat{A}x,$$

$$y \geq 0 \text{ and } y_i \text{ integer for } i \in I.$$

The problems (1.3) and (1.4) are *equivalent* in the sense that (1.3) has a feasible solution if and only if (1.4) has a feasible solution, and (x^*,z^*) is an optimal solution of (1.3) if and only if there exists a y^* such that (x^*,y^*,z^*) is an optimal solution of (1.4). From a computational point of view the transformation from (1.3) to (1.4) may allow the replacement of a *piecewise-linear* objective function term $f(x)$ by a *linear* objective function term cy. Thus, if an optimization problem has only linear constraints and objective function terms for which MIMM's exist, then this conversion procedure may be carried out *term-by-term* until the original problem has been transformed into a mixed integer program (MIP). Note, however, that although this MIP will be equivalent to the original problem, equivalence may be destroyed if the integrality constraints on the newly added variables are deleted, a *relaxation* which is usually the first step of an algorithm for the solution of an MIP. In particular, the relaxation of the integrality constraints of a MIMM will yield a parametrically defined family of problems (a *linear programming minimization model* (LPMM)) whose optimal value must be (see [4]) a *convex* function on all of $\mathbb{R}^1$. Thus, this relaxation will mean that a nonconvex objective function term of the original formulation is replaced by a convex approximation. In algebraic terms, defining

$$f^*(x) \equiv \min cy \tag{1.5}$$

$$\text{s.t. } Ay = b - \hat{A}x, \ y \geq 0,$$

it follows that f^* is *convex* on $\mathbb{R}^1$, so that if f (as defined in (1.2)) is *nonconvex* on S (in the sense that there exist points $x_1, x_2, x \in S$ and a $\lambda \in (0,1)$ such that $x = \lambda x_1 + (1-\lambda)x_2$ and $f(x) > \lambda f(x_1) + (1-\lambda) x_2$), then f and f^* *cannot* coincide over all of S (in particular, they would not agree at x). The difference $f(x) - f^*(x)$ (which is always nonnegative because of the relaxation of the constraints) will be termed the *relaxation error* of the LPMM at x.

In the case of the MIMM (1.1), for example, the optimal value function (for $x \geq 0$) for the LPMM obtained by relaxing the integrality constraints of (1.1) is easily seen to be the *linear*

function $k^*(x) = |x - 1/2|$. The relaxation error in this particular case is thus the difference between the values $k(x)$ and $k^*(x)$ (see Figure 1). Note that this difference is *positive* unless x is an integer multiple of 1/2. (For $x = 0$, $k(x) = k^*(x) = 0$, but we are concerned here only with nonnegative values of x.)

In comparing *alternative* MIMM formulations, a comparison of the behavior of the relaxation errors establishes the relative accuracy of the approximations used in the first step of the solution of the respective MIP's. Thus, if f^{**} is the optimal value function of the continuous relaxation of a *different* MIMM for f, and $f^*(x) \geq f^{**}(x)$ for all $x \in S$ (which we write as $f^* \geq f^{**}$), then the MIMM (1.2) may be considered to be at least as good (with respect to the relaxation error criterion) as the MIMM from which f^{**} was derived. Moreover, if it can be established that the inequality $f^* \geq f^{**}$ holds for *all* convex functions f^{**} satisfying $f \geq f^{**}$, then the MIMM giving rise to f^* will be *optimal* from the standpoint of error in a relaxation solution strategy, and will therefore be said to be *relaxation-optimal* on S. (As will be seen, a function may have more than one relaxation-optimal MIMM, so additional MIMM criteria also will be considered.) In order to more easily describe results of this type, it is convenient to introduce some additional terminology. If h is a function mapping a convex set T into $[-\infty, +\infty]$, the *convex envelope* of h on T (which may be thought of geometrically as the *largest* convex function *below* h on T), denoted by $c^*(h, x, T)$, is the function satisfying the relations:

(1.6) $\quad c^*(h, x, T) \leq h(x)$ for all $x \in T$,

(1.7) $\quad c^*(h, x, T)$ is convex on T,

(1.8) $\quad$ if $g(x) \leq h(x)$ for all $x \in T$ and g is convex on T,

$\quad$ then $g(x) \leq c^*(h, x, T)$ for all $x \in T$.

(In places where reference to the variable is not needed, we will write $c^*(h, T)$ in place of $c^*(h, x, T)$.) Existence and uniqueness of $c^*(h, T)$ easily follow from the fact that the pointwise supremum of a family of convex functions is convex. Defining on T the set of functions

$$C(h, T) \equiv \{g \mid g \text{ is convex on } T,\ g \leq h\},$$

$c^*(h, T)$ is simply the supremum of $C(h, T)$. It might be noted that the domain T plays a very significant role in determining the convex envelope. That is, the value of the convex envelope at a particular point may be different for different choices of T. This aspect of the convex envelope will be taken up in Section 2.

The optimal value function of a LPMM, in addition to being convex, is also piecewise-linear (PL), and it is also convenient to introduce some terminology for piecewise-linear functions of a single variable, which are our principal concern in this paper.

We will say that a real-valued function h defined on a closed interval $[\alpha_0, \alpha_p] \subset \mathbb{R}^1$ is a *piecewise-linear function on* $[\alpha_0, \alpha_p]$ *with breakpoints* $\alpha_0 < \alpha_1 < \cdots < \alpha_p$ if h is *affine* on each subinterval $[\alpha_{i-1}, \alpha_i]$ and $[h(\alpha_{i-1}) - h(\alpha_{i-2})]/(\alpha_{i-1} - \alpha_{i-2}) \neq [h(\alpha_i) - h(\alpha_{i-1})]/(\alpha_i - \alpha_{i-1})$ for $i = 1, \ldots, p-1$ (that is, the slope to the left of α_i differs from the slope to the right of α_i).

The basic result that will be used to establish that certain formulations yield convex envelopes is the sufficiency part of the following theorem:

THEOREM 1: Let g be a lower semi-continuous (l.s.c.) function mapping $[\alpha_0, \alpha_p]$ into $(-\infty, +\infty]$ with $g(\alpha_i) < +\infty$ for $i = 0, \ldots, p$.

Let g^* be a convex piecewise-linear function on $[\alpha_0, \alpha_p]$ with breakpoints $\alpha_0 < \alpha_1 < \cdots \alpha_p$, and let $g^*(x) \leq g(x)$ for $x \in [\alpha_0, \alpha_p]$. A necessary and sufficient condition for g^* to be the convex envelope of g on $[\alpha_0, \alpha_p]$ is that $g^*(\alpha_i) = g(\alpha_i)$ for $i = 0, \ldots, p$.

Proof. To establish *sufficiency*, suppose that $\tilde{g} \in C(g, [\alpha_0, \alpha_p])$. Then for any $x \in [\alpha_0, \alpha_p]$ there exists at least one pair $\alpha_{i-1}, \alpha_{i+1}$ of breakpoints such that $x \in [\alpha_{i-1}, \alpha_{i+1}]$. Choosing $\lambda \in [0, 1]$ such that $x = \lambda\alpha_i + (1 - \lambda)\alpha_{i-1}$, we have (using the convexity of $\tilde{g}$): $\tilde{g}(x) \leq \lambda\tilde{g}(\alpha_i) + (1 - \lambda)\tilde{g}(\alpha_{i-1}) \leq \lambda g(\alpha_i) + (1 - \lambda) g(\alpha_{i-1}) = \lambda g^*(\alpha_i) + (1 - \lambda) g^*(\alpha_{i-1}) = g^*(x)$. Thus, $\tilde{g}(x) \leq g^*(x)$ for any $x \in [\alpha_0, \alpha_p]$ establishing that $g^* = c^*(g, [\alpha_0, \alpha_p])$.

To show necessity, suppose that $g(\alpha_0) - g^*(\alpha_0) \equiv \epsilon_0 > 0$. Since g is l.s.c. and g^* is continuous, there exists a $\delta_0 \in (0, \alpha_0)$ such that $\alpha_0 \leq x \leq \alpha_0 + \delta_0$ implies $g(x) \geq g(\alpha_0) - \epsilon_0/2$ and $g^*(x) \leq g^*(\alpha_0) + \epsilon_0/2 = g(\alpha_0) - \epsilon_0/2$.

Now consider the PL function $\tilde{g}$ (see Figure 2) with breakpoints at $\delta\alpha_0, \alpha_0 + \delta_0, \alpha_1, \ldots, \alpha_p$ and function values $\tilde{g}(\alpha_0) = g(\alpha_0) - \epsilon_0/2$, $\tilde{g}(\alpha_0 + \alpha_0) = g^*(\alpha_0 + \delta_0)$, $\tilde{g}(\alpha_i) = g^*(\alpha_i)$ $(i = 1, \ldots, p)$. Note that $\tilde{g}(\alpha_0) > g^*(\alpha_0)$, but that $\tilde{g}(x) = g^*(x)$ for $x \in [\alpha_0 + \delta_0, \alpha_p]$ and that $\tilde{g}$ is a convex function on $[\alpha_0, \alpha_p]$. Finally, the relations $\tilde{g}(\alpha_0) = g(\alpha_0) - \epsilon_0/2$ and $\tilde{g}(\alpha_0 + \delta_0) = g^*(\alpha_0 + \delta_0) \leq g^*(\alpha_0) + \epsilon_0/2 = g(\alpha_0) - \epsilon_0/2$ imply $\tilde{g}(x) \leq g(\alpha_0) - \epsilon_0/2$ for $x \in [\alpha_0, \alpha_0 + \delta]$, so that $\tilde{g}(x) \leq g(x)$ for $x \in [\alpha_0, \alpha_p]$. Thus, $\tilde{g}(x) \in C(g(x), [\alpha_0, \alpha_p])$ and $\tilde{g}(\alpha_0) > g^*(\alpha_0)$, contradicting the hypothesis that $g^*(x) = c^*(g, x, [\alpha_0, \alpha_p])$. A contradiction may be similarly obtained if $g(\alpha_p) > g^*(\alpha_p)$. For an interior breakpoint α_i, the construction of a suitable $\tilde{g}$ is similar (see Figure 3), except that the breakpoints of $\tilde{g}$ (where it coincides with g^*) are taken to be $\alpha_0, \ldots, \alpha_{i-1}, \alpha_i - \delta_i, \alpha_i + \delta_i, \ldots, \alpha_p$, where $0 < \delta_i < \min\{\alpha_i - \alpha_{i-1}, \alpha_{i+1} - \alpha_i\}$ is chosen so that, defining $\epsilon_i \equiv g(\alpha_i) - g^*(\alpha_i) > 0$, we have, for $x \in [\alpha_i - \delta_i, \alpha_i + \delta_i]$, the inequalities $g(x) \geq g(\alpha_i) - \epsilon_i/2$, $g^*(x) \leq g^*(\alpha_i) + \epsilon_i/2$. Because of the change in slope at breakpoints, it may be verified that $g^*(\alpha_i) < \tilde{g}(\alpha_i)$, and thus a contradiction may be obtained.

Note that for sufficiency, lower semi-continuity of g is *not* required. In this paper we are primarily concerned with the sufficiency part of this theorem, but it should be noted that in [4] the lower semi-continuity of optimal value functions of MIMM's was established under rationality assumptions on the coefficients of the MIMM.

It might also be noted that the argument used in the proof can be used to show that g does *not* have a PL convex envelope if $g(\alpha_0) = +\infty$ or $g(\alpha_p) = +\infty$, since this would mean that $g^*(\alpha_0) < g(\alpha_0)$ or $g^*(\alpha_p) < g(\alpha_p)$ for any PL function g^*. On the other hand, a PL convex envelope may exist if there are *interior* points x of $[\alpha_0, \alpha_p]$ with the property that $g(x) = +\infty$. This allows the domain of g to have "gaps" on which g may be thought of as being $+\infty$. Such gaps often occur in optimal value functions of MIMM's.

From Figure 1, one might conjecture that k^* is the convex envelope of k on R^1. This is indeed true, and in Section 2 we will use the approach of Theorem 1 to establish a more general result from which this follows as a special case.

2. THE UNBOUNDED CASE

In this section we will consider MIMM's for a broad class of economy-of-scale functions that includes the economy-of-scale function $k(x)$ of the previous section. Specifically, we will develop relaxation-optimal MIMM's for the class of functions whose elements may be represented as optimal value functions of the following type.

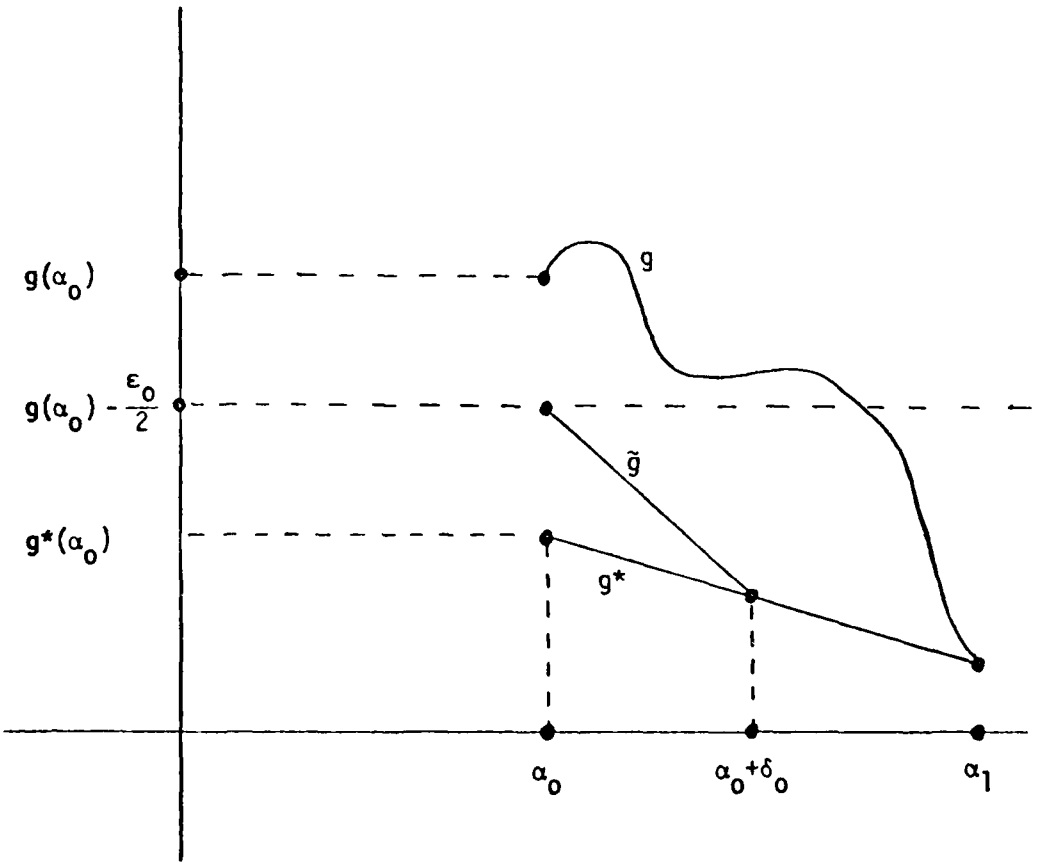


FIGURE 2. Proof that $g(\alpha_0) > g(\alpha_1)$

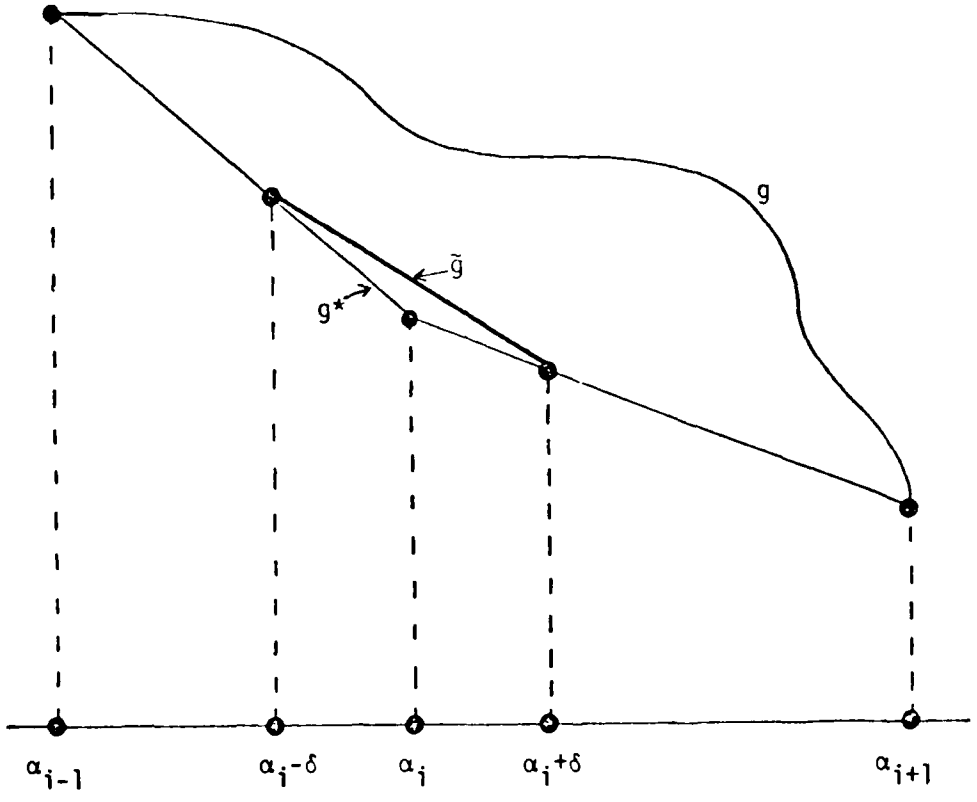


FIGURE 3 The case $g^*(\alpha_i) < \tilde{g}(\alpha_i)$ where $0 < \delta < p$

$$f_1(x) \equiv \min cy \tag{2.1}$$

$$\text{s.t. } ay \geqslant x$$

$$y \geqslant 0,\ y_i \text{ integer for } i \in I,$$

where $c = (c_1, \ldots, c_n) \geqslant 0$, $a = (a_1, \ldots, a_n) > 0$,

and I is a subset of $\{1, \ldots, n\}$.

(The case in which there are $c_i = 0$ is not of economic interest, but is included for mathematical completeness. The sign restrictions on c and a do serve to guarantee the existence of an optimal solution for all x, but, as shown in Appendix A, could be replaced by this hypothesis. In the next section, where *bounds* on the y_i are assumed, it will be seen that these sign restrictions have greater significance.) Note that the class of functions representable in the form (2.1) includes fixed-charge functions and economy-of-scale functions allowing several *different* volume discounts (as opposed to only one in the case of $k(x)$). (The computational results in Section 5 deal with an example in which $n = 3$.) For notational convenience we will assume that the variables have been ordered so that

$$c_1/a_1 \equiv r_1 \leqslant c_2/a_2 \equiv r_2 \leqslant \ldots \leqslant c_n/a_n \equiv r_n. \tag{2.2}$$

From a cost viewpoint, this means that, on as *per unit* basis, the most "economical" purchase quantity is a_1, the next most economical is a_2, etc., and the right-hand side x represents the minimum amount to be purchased

Consider the continuous relaxation of the MIMM in (2.1), which has the optimal value function defined by

$$f_1^*(x) = \min cv \quad \text{s.t. } av \geq x,\ v \geq 0. \tag{2.3}$$

The following lemma states that f_1^* is linear on R_+^1, and provides the basis for a proof of the relaxation-optimality of the MIMM on the RHS of (2.1).

LEMMA 1. For $x \in \mathbb{R}_+^1$, $f_1^*(x) = r_1 x$.

PROOF. Note that, for any $x \geq 0$, the dual of (2.3) may be written as

$$\max vx \quad \text{s.t. } va \leq c,\ v \geq 0. \tag{2.4}$$

By setting $v_1^* = r_1/a_1$ and $v_2^* = v_3^* = \cdots = v_n^* = 0$ and $v^* = r_1$, we obtain primal and dual feasible solutions with common objective function value $r_1 x$. This is thus the optimal value, $f_1^*(x)$.

Having obtained a closed form representation of $f_1^*(x)$, the relationship between f_1 and f_1^* is easily established.

THEOREM 2. The following relations hold between f_1 and f_1^*:

$$f_1(x) = f_1^*(x) \text{ for } x = k \cdot a_1 \ (k = 0, 1, \ldots) \tag{2.5}$$

$$f_1^* = c^*(f_1, \mathbb{R}_+^1). \tag{2.6}$$

PROOF. Since $f_1^*(x) \leq f_1(x)$ for $x \in \mathbb{R}_+^1$, (2.5) may be established by showing that, for $x = k \cdot a_1$ $(k = 0, 1, 2, \ldots)$, (2.1) has a *feasible* solution with objective function value $f_1^*(ka_1) = r_1 \cdot ka_1 = kc_1$. Such a feasible solution is obtained by setting $v_1 = k$ and $v_2 = v_3 = \cdots = v_n = 0$.

To prove (2.6), it suffices to show that for any $x' \in \mathbb{R}_+^1$, there exist $x_1', x_2' \in \mathbb{R}_+^1$ such that, for some $\lambda' \in [0, 1]$, we have $\lambda' x_1' + (0 - \lambda')\ x_2' = x'$ and $f_1^*(x') = \lambda' f_1(x_1') + (1 - \lambda') f_1(x_2')$, since any $f \in c^*(f_1, \mathbb{R}_+^1)$ must satisfy $f(x') \leq \lambda' f_1(x_1') + (1 - \lambda') f_1(x_2')$. These quantities are obtained by taking $x_1' = 0$, $x_2' = k \cdot a_1$, where k is an integer chosen such that $ka_1 \geq x'$, and λ' such that $(1 - \lambda')ka_1 = x'$. Then $\lambda' f_1(x_1') + (1 - \lambda')\ f_1(x_2') = 0 + (1 - \lambda') r_1 ka_1 = r_1 x' = f_1^*(x')$.

It is of some mathematical interest to note that the constraint $av \geq x$ in (2.3) is satisfied as an equality by an optimal solution of (2.3). The observation may be used to establish that f_1^* is also the convex envelope on $\mathbb{R}_+^1$ of the optimal value function in the corresponding *equality-constrained* case:

$$f_1(x) \equiv \min cv \quad \text{s.t. } av = x \quad v \geq 0,\ v_i \text{ integer for } i \in I. \tag{2.7}$$

This result follows since $f_1(x) = f_1^*(x)$ for $x = k \cdot a_1$ $(k = 0, 1, \ldots)$. Since f_1^* may be written in the form (2.3) with the constraint $av \geq x$ replaced by $av = x$, it follows by the analog of Theorem 2 that the modified MIMM is relaxation-optimal in the equality-constrained case (2.7) as well.

On the other hand, it is not always possible to establish relaxation-optimality if a positive *constant* is added to the RHS of the constraint with RHS x in (2.1) (negative constants pose no difficulty, as we will show in Section 3). An example illustrating the difficulties that may arise in this case is given in Appendix B. However, it is possible to extend the results of this section to the case in which nonnegative *bounds* are imposed on the variables. This case is taken up in Section 3.

Finally, in the case that the a_i are all *rational*, Theorem 1 is a special case of a result of Blair and Jeroslow [2], who considered a system of constraints and showed that the convex envelope of the optimal value function of the MIMM (for $x \in \mathbb{R}^m$)

$$\min cy \tag{2.8}$$

$$\text{s.t. } Ay \geqslant x,\ y \geqslant 0,\ y_i \text{ integer for } i \in I,$$

coincides with the optimal value function of the continuous relaxation of the MIMM. The thrust of the next section can thus be viewed as an extension of this result to certain cases in which *nonzero constants* are allowed in the constraints of (2.8). (In general the Blair-Jeroslow result does *not* extend to the nonhomogeneous case, as may be ascertained from the examples in Section B.)

3. BOUNDS ON y

For most integer programming codes, it is necessary to have bounds on the integer variables. If the range of the y variables in (2.3) is restricted by the imposition of bounds, then the corresponding optimal value function on $\mathbb{R}^1$ is piecewise-linear (where it is finite), but the relaxation-optimality property of Section 2 may nonetheless be extended to this case. We first consider the case of upper bounds, and then the case of upper and lower bounds. As in Section 2 we assume that $c \geqslant 0$ and $a > 0$. (By making some obvious extensions, the constraint $a > 0$ may be removed, but as may be seen from an example in Appendix B, sign restrictions on c are needed in the bounded case to guarantee relaxation-optimality.)

Specifically, instead of the MIMM in (2.1) we first consider

$$f_2(x) = \min cy \tag{3.1}$$

$$\text{s.t. } ay \geqslant x$$

$$0 \leqslant y \leqslant u$$

$$y_i \text{ integer}, i \in I$$

where $c \geqslant 0$, $a > 0$, the ordering assumption (2.2) is assumed to be satisfied, and the u_i are nonnegative constants with u_i integer for $i \in I$. To prove relaxation-optimality we will show that the convex envelope of f_2 on $D \equiv [0, au]$, denoted by $c^*(f_2, D)$, is given by the optimal value function of the continuous relaxation:

$$f_2^*(x) \equiv \min cy \tag{3.2}$$

$$\text{s.t. } ay \geqslant x$$

$$0 \leqslant y \leqslant u.$$

(We are not concerned with $x > au$ since $f_2(x) = f_2^*(x) = +\infty$ for such x.)

For notational convenience in stating a closed form expression for $f^*_2(x)$, we make the following definitions:

$$b_j \equiv \sum_{i=1}^{j} a_i u_i, \quad d_j \equiv \sum_{i=1}^{j} c_i u_i \quad (j = 0, \ldots, n),$$

where it is understood that $b_0 = 0$ and $d_0 = 0$.

The following is the analog of Lemma 1:

LEMMA 2: $f^*_2(x) = r_j(x - b_j) + d_j$ for $b_j \leq x \leq b_{j+1}$
$(j = 0, \ldots, n-1)$

PROOF: The proof is analogous to that of Lemma 1. For any x, the dual of (2.7) is given by

$$\max_{v,w} vx - wu$$
$$\text{s.t. } va - w \leq c, \; v \geq 0, \; w \geq 0.$$

In addition, for any $x \in D$, the optimal solutions of the primal and dual problems are as follows: if $b_j \leq x \leq b_{j+1}$, set $v^* = u_i$ for $i \leq j$, set $y^*_i = 0$ for $i > j+1$, and choose y^*_{j+1} such that $a_j y^* = x$; set $v^* = r_{j+1}$, $w^*_i = r_{j+1}a_i - c_i$ for $i \leq j$, and $w^*_i = 0$ for $i > j$.

Note from Lemma 2 that the breakpoints of f^*_2 are contained in the set $\{b_0, \ldots, b_n\}$. By applying Theorem 1, we can obtain the following analog of Theorem 2:

THEOREM 3: The following relationships hold between f_2 and f^*_2:

$$f_2(\bar{x}) = f^*_2(\bar{x}) \text{ if } \bar{x} = b_j \; (j = 0, \ldots, n), \tag{3.3}$$

$$f^*_2 = c^*(f_2, D). \tag{3.4}$$

PROOF: The relation (3.3) follows from considering the feasible solution with $y^*_i = u_i$ for $i \leq j$ and $y^* = 0$ for $i > j$. The relation (3.4) then follows directly from (3.3) and Theorem 1.

In a branch-and-bound algorithm in which the y_i are used as the branching variables, the formulation (3.1) has the additional very nice property of yielding a relaxation-optimal formulation *at each node* in the tree, since relaxation-optimality is *not* affected by the imposition of additional integer upper and lower bounds on the y_i in (3.1). This is because introduction of nonnegative *lower* bounds is equivalent to the addition of a *negative* constant to the RHS of the constraint $ay \geq x$. Since a constraint of the form $ay \geq x - \gamma$, where $\gamma \geq 0$ implies an optimal value of 0 for $x \in [0, \gamma]$ in both the corresponding MIMN and its relaxation, it is easily shown that a translation of variables leads to the following result (see Appendix C for details):

COROLLARY 1: For $x \geq 0$, let

$$f_1(x) \equiv \min_{y} cy$$
$$\text{s.t. } ay \geq x$$
$$l \leq y \leq u$$
$$y_i \text{ integer}, \; i \in I.$$

where $l_i \geq 0$ and l_i and u_i are integer for $i \in I$, then the MIMM is *relaxation-optimal* on any interval $[\alpha, au]$, where $\alpha \in [0, a]$.

In the next two sections we will compare these results to a "standard" approach to formulation that yields relaxation-optimal MIMM's for quite general piecewise-linear functions.

4. AN ALTERNATE APPROACH

A standard and quite general approach to modelling continuous piecewise-linear nonconvex functions is to employ the so-called "λ formulation" of separable programming with the additional restrictions that at most two λ_i are allowed to be positive and that these must be "consecutive." We will see that, while this approach also yields relaxation-optimal models, it can, in contrast to the approach of Section 3, lead to computational difficulties in the absence of special provisions for handling the variables.

Assume that $\hat{f}$ is a piece wise-linear function on $[\alpha_0, \alpha_p]$ with breakpoints $\alpha_0 < \alpha_1 < \ldots < \alpha_p$. (It is possible to deal with l.s.c. "piecewise-linear" functions by a slightly different formulation technique (see [4]), but, aside from the need for more complex notation, the results are essentially the same.) Consider the following MIMM for $\hat{f}$:

$$
\begin{aligned}
(4.1) \qquad \hat{f}(x) = \min_{\lambda,\,\delta} \; & \sum_{i=0}^{p} f(\alpha_i)\lambda_i \\
\text{s.t.} \; & \sum_{i=0}^{p} \alpha_i \lambda_i = x \\
& \sum_{i=0}^{p} \lambda_i = 1, \; \lambda_i \geq 0 \; (i = 0, \ldots, p) \\
& \lambda_0 \leq \delta_0 \\
& \lambda_1 \leq \delta_0 + \delta_1 \\
& \vdots \\
& \lambda_{p-1} \leq \delta_{p-2} + \delta_{p-1} \\
& \lambda_p \leq \delta_{p-1} \\
& \sum_{i=0}^{p-1} \delta_i = 1, \; \delta_i \geq 0 \text{ and integer } (i = 0, \ldots, p-1)
\end{aligned}
$$

and let $\hat{f}^*$ denote the optimal value function corresponding to the continuous relaxation of the RHS of (4.1). Note that $\hat{f}^* \in C(\hat{f}, [\alpha_0, \alpha_p])$.

THEOREM 4: The MIMM on the RHS of (4.1) is relaxation-optimal on $[\alpha_0, \alpha_p]$.

PROOF: Let $\bar{x} \in [\alpha_0, \alpha_p]$ and let $\bar{\lambda}$ be chosen so that $\hat{f}^*(\bar{x})$ is obtained by setting $\lambda_i = \bar{\lambda}_i$ in the corresponding LPMM, so that $\hat{f}^*(\bar{x}) = \sum f(\alpha_i) \cdot \bar{\lambda}_i \geq \sum c^*(\hat{f}, \alpha_i, [\alpha_0, \alpha_p]) \cdot \bar{\lambda}_i \geq c^*(\hat{f}, \bar{x}, [\alpha_0, \alpha_p])$. Since $\hat{f}^* \in C(\hat{f}, [\alpha_0, \alpha_p])$, this implies that $\hat{f}^*(\bar{x}) = c^*(\hat{f}, \bar{x}, [\alpha_0, \alpha_p])$ and the conclusion follows. ∎

While Theorem 4 implies that the standard MIMM will also be relaxation-optimal for a *continuous* economy-of-scale function in the class considered in Section 3, the MIMM (4.1) has

several computational disadvantages. One obvious disadvantage is its sheer size, since the number of constraints and variables in (4.1) is determined by the number of breakpoints of f, whereas this is not the case for the formulations of Sections 2 and 3. A more subtle disadvantage is the failure of the integer variables δ_i of (4.1) to directly reflect physical quantities. In particular, the δ_i all have cost coefficients of 0 and, moreover, a 0 "branch" on a δ_i has no effect on the allowable range of x values unless it has the largest or smallest index of any δ_i not yet fixed. While these disadvantages may be alleviated via the use of "Special Ordered Set" (SOS) strategies for branching (see [1]), such strategies are often not available in MIP codes (see [3]). In particular, SOS strategies are not fully implemented on the Univac EMPS-MIP code in use at the Madison Academic Computing Center, and in the next section we compare results obtained with EMPS and the formulation approaches of Section 3 and 4. (It should be noted that the use of an SOS strategy has the advantage of imposing disjoint upper and lower bounds on the range of the variable x in (4.1) when SOS branching is performed. Branching on the y_i in (3.1) imposes upper bounds on x, but does not directly impose lower bounds. Lower bounds on the range of x may be directly imposed by adding to (3.1) constraints of the form

$$x \geq ay - \hat{a}z,$$

plus additional constraints of the form $z_i \leq y_i$. By selecting the coefficients a to reflect maximum "surpluses" so that for any $\hat{x} \in [0,au]$, a $\bar{y}$ yielding an optimal solution to (3.1) for $x = \hat{x}$ will satisfy $\hat{x} \geq a\bar{y} - \hat{a}\bar{z}$ for some feasible $\bar{z}$, relaxation optimality will be preserved. This follows easily from the fact that, by assumption, the optimal value function of the MIMM remains $f(x)$, while the optimal value of the continuous relaxation, which cannot increase beyond $c^*(I_+,[0,au])$ (in spite of the added constraint) must also remain the same. Some theoretical and computational aspects of such lower bound constraints as well as some other modelling refinements to deal with upper bounds on x are currently under investigation.)

5. A COMPUTATIONAL COMPARISON

In this section we consider a comparison of solution times for different formulations of the following communications network problems: determine the minimum cost network (see Table 1) that meet specified demands (see Table 2) between six distinct pairs of cities (A,B), (A,C), (A,D), (B,C), (B,D), and (C,D), where the communication traffic between the elements of a city-pair may be routed via any acyclic path between the cities (there are 5 such routes between each city-pair).

TABLE 1. *Costs*

Arc	Single Channel	12 Channels	60 Channels
A-B	789.75	7028.77	17690.40
B-C	878.25	7992.07	21341.47
C-D	1407.70	13232.38	42512.54
D-A	654.90	5697.63	13098.00
D-B	1045.60	9619.52	28022.08
C-A	1236.57	11500.10	35860.53

TABLE 2 *Two Sets of Communications Demands*

City-Pair	Demand Set I	Demand Set II
A-B	2	4
B-C	10	10
C-D	46	64
D-A	5	5
D-B	2	10
C-A	4	14

Algebraically, this problem has the form:

$$\min \sum_{i=1}^{6} h_i(x_i)$$

$$\text{s.t.} \sum_{j=1}^{5} z_{j,k} = d_k \quad (k = 1, \ldots, 6)$$

$$\sum_{(j,k)\in A_i} z_{j,k} = x_i \quad (i = 1, \ldots, 6)$$

$$x_i, z_{j,k} \geq 0,$$

where $z_{j,k}$ represents the number of channels on the j^{th} path between the k^{th} city-pair, d_k is the total number of channels needed by the k^{th} city-pair, A_i is the set of pairs (j,k) such that the corresponding path uses arc i, x_i is the total number of channels on arc i, and $h_i(x_i)$ is the minimum cost of leasing at least x_i channels on arc i. (Note that the h_i are economy-of-scale functions of the type considered in Sections 2 and 3 with $n = 3$. For computational convenience the variables associated with single channels on arcs were assumed continuous. Because of the fixed demands, bounds could be imposed on all variables. General integer variables were decomposed into 0–1 variables, since the FMPS-MIP code requires this.)

The computational results of Table 3 illustrate the dramatic difference in solution behavior and times between the formulation approaches of Sections 3 and 4. The MIP code used was the Univac FMPS-MIP code (level 7R1) and the problems were run on the Madison Academic Computing Center Univac 1110. For demand set I, the Section 3 formulation requires only about 1/4 the computer time of the Section 4 formulation. For demand set II, the solution time for the Section 3 formulation is 15 seconds, whereas the FMPS system was unable to solve the Section 4 formulation. Similar behavior was observed in runs using a locally developed MIP code, IPMIXD, which successfully solved both I-S and II-S, but failed to solve either I-L or II-L because of storage overflows.

TABLE 3. *Problem Sizes and Solution Times*

Problem	Rows	Columns	0-1 Variables	Solution Time (Sec.)
I-S*	12	54	18	4
I-L†	76	122	40	15
II-S	12	60	24	15
II-L	116	202	80	‡

*I denotes demand set I, S denotes "short" formulation

†L denotes "long" standard formulation

‡ FMPS system forced termination of run with message "numerical errors"

A number of other versions of the problems were run in which some of the cost function terms were modelled via the Section 3 approach and the remainder via the Section 4 approach. In all cases the results were worse then those obtained via the Section 3 approach.

6. CONCLUSION

For piecewise-linear functions belonging to a broad class of economy-of-scale functions, a compact mixed-integer programming formulation has been described. This formulation was then shown to behave at least as well as any other mixed-integer formulation of the function in

terms of the approximation error resulting from the relaxation of integrality constraints. Moreover, a computational comparison (using a communications network problem as a test problem) showed the superiority of the compact formulation over a standard mixed-integer formulation of the same problem.

ACKNOWLEDGMENT

Jay M. Flesher of the Madison Academic Computing Center assisted in the development of the test problems and obtained the computational results of Section 5.

REFERENCES

[1] Beale, E.M.L. and J.J.H. Forrest, "Global Optimization Using Special Ordered Sets," *10*, (1976) 52-69 Mathematical Programming.

[2] Blair, C.E. and R.G. Jeroslow, "The Value Function of a Mixed Integer Program: II," Management Science Research Report *377*, GSIA, Carnegie-Mellon University, (December 1976).

[3] Land, A. and S. Powell, "Computer Codes for Problems of Integer Programming," to appear in the proceedings of the conference, Discrete Optimization, (1977), held at the University of British Columbia, Vancouver, August, (1977).

[4] Meyer, R.R., "Integer and Mixed-Integer Programming Models: General Properties," Journal of Optimization Theory and Applications *16* (1975), 191-206 (1975).

[5] Meyer, R.R., "Mixed Integer Minimization Models for Piecewise-Linear Functions of a Single Variable," Discrete Math, *16* 163-171 (1976).

[6] Meyer, R.R. and M.V. Thakkar, "Rational Mixed-Integer Minimization Models," Mathematics Research Center Report #1552, University of Wisconsin-Madison, (1976).

APPENDIX A

To justify the statement in Section 2 that the restrictions $a_i > 0$ and $c_i \geq 0$ can be replaced by assuming that (2.1) has an optimal solution for $x \geq 0$, we consider the remaining cases: (1) i such that $c_i \geq 0$ and $a_i \leq 0$ (2) i such that $c_i < 0$ and $a_i > 0$, and (3) i such that $c_i < 0$ and $a_i < 0$.

CASE 1: For those i such that $c_i \geq 0$ and $a_i \leq 0$, one may obtain an equivalent problem by deleting the corresponding variables y_i from the problem, since, for any $x_i \geq 0$, an optimal solution may be obtained in which such $y_i = 0$.

CASE 2: If there are i such that $c_i < 0$ and $a_i > 0$, then clearly the objective function of (2.1) must be unbounded from below, so this case is ruled out by the existence of an optimal solution.

CASE 3: If, for some i, $c_i < 0$ and $a_i < 0$, then either all $a_i \leq 0$, in which case (2.1) is infeasible for $x > 0$, or there exists at least one j such that $a_j > 0$. In the latter case let $r' \equiv \min\{c_k/a_k | a_k > 0\}$ and $r'' \equiv \max\{c_k/a_k | c_k < 0, a_k < 0\}$. If $r'' \leq r'$, then, assuming that the variables are ordered so that $a_1 > 0$ and $c_1/a_1 = r'$, it may be seen from obvious extensions of the proofs of Lemma 1 and Theorem 2 that the desired result holds. On the other hand, if $r'' > r'$, then the objective function of (2.1) is unbounded from below for all x. This follows by letting $r' = c_1/a_1$ and $r'' = c_p/a_p$, noting that $c_1/-c_p < a_1/-a_p$, and choosing a rational $\theta > 0$ such that $c_1/-c_p < \theta < a_1/-a_p$, from which it follows that $a_1 + a_p\theta > 0$ and $c_1 + c_p\theta < 0$. Now choose an integer $M > 0$ such that $M\theta$ is integer and note that the relations $a_1 \cdot M + a_p \cdot M\theta > 0$ and $c_1 M + c_p \cdot M\theta < 0$ imply unboundedness.

APPENDIX B

Here we consider several examples to illustrate the difficulties that can arise when one attempts to extend the results of Sections 2 and 3 by either (1) inserting a positive constant on the RHS of the constraint involving x, or (2) relaxing sign restrictions in the bounded case, or (3) allowing more than one constraint involving x in the bounded case.

The following illustrates the difficulties that may arise when a *positive constant* appears in the RHS of a MIMM (see Figure 4):

$$k_1(x) \equiv \min\ y_1 + 10y_2$$
$$\text{s.t. } y_1 + 12y_2 \geq x + 10$$
$$y_1,\ y_2 \geq 0,\ y_2 \text{ integer.}$$

In this case, the convex envelope of $k_1(x)$ on R^1 is easily seen to have a value of 10 on $[0,2]$, so that it does not coincide at $x = 0$ with the optimal value function of the continuous relaxation of the MIMM as given by:

$$k_1^*(x) \equiv \min\ y_1 + 10y_2$$
$$\text{s.t. } y_1 + 12y_2 \geq x + 10$$
$$y_1,\ y_2 \geq 0,$$

since $k^*_1(0) = 10 \cdot \frac{10}{12} < 10$.

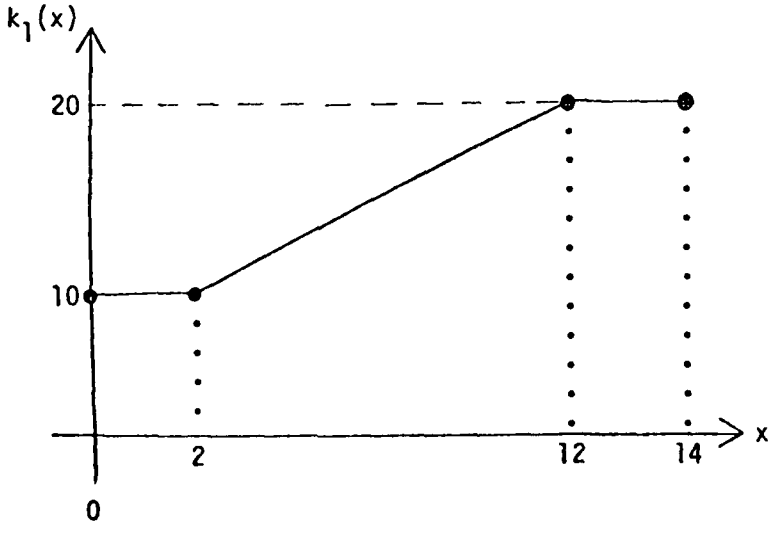


FIGURE 4. $k_1(x)$ on $[0,14]$

Note also that the addition of *bounds* does not help, since defining

$$\begin{aligned} k_2(x) \equiv \min_{y} \; & y_1 + 10y_2 \\ \text{s.t. } & y_1 + 12y_2 \geq x + 10 \\ & 0 \leq y_1 \leq 10 \\ & 0 \leq y_2 \leq 1 \\ & y_2 \text{ integer} \end{aligned}$$

yields $k_1(x) = k_2(x)$ for $x \in [0, 12]$, and $k_1(x)$ coincides with its convex envelope on $[0,12]$, whereas the optimal value function of the continuous relaxation is again strictly less than $k_1(x)$ at $x = 0$.

Now consider the following example in which a RHS constant is not present in the constraint involving x, but there are *negative* coefficients:

$$\begin{aligned} k_3(x) \equiv \min_{y_1', y_2} \; & -y_1' + 10y_2 \\ \text{s.t. } & -y_1' + 12y_2 \geq x \\ & 0 \leq y_1' \leq 10 \\ & 0 \leq y_2 \leq 1 \\ & y_2 \text{ integer.} \end{aligned}$$

Making the change of variables $y_1' = 10 - y_1$ we have

$$\begin{aligned} k_3(x) = -10 + \min_{y_1, y_2} \; & y_1 + 10y_2 \\ \text{s.t. } & y_1 + 12y_2 \geq x + 10 \\ & 0 \leq y_1 \leq 10 \\ & 0 \leq y_2 \leq 1 \\ & y_2 \text{ integer,} \end{aligned}$$

so that $k_3(x) = -10 + k_2(x)$. It is easily seen that while k_3 coincides with its convex envelope on $[0,12]$, it differs from the optimal value function of the corresponding continuous relaxation at $x = 0$.

In our last example, we consider the case of *two* constraints with *positive* coefficients and rhs x:

$$\begin{aligned} k_4(x) \equiv \min \; & y_1 + y_2 \\ \text{s.t. } & 2y_1 + 4y_2 \geq x \\ & 4y_1 + 3y_2 \geq x \\ & 0 \leq y_1, y_2 \leq 1 \\ & y_1, y_2 \text{ integer.} \end{aligned}$$

In this case the optimal value function is finite for $x \leqslant 6$, and is easily seen to have the values

$$k_4(x) = \begin{cases} 0 \text{ for } x = 0 \\ 1 \text{ for } 0 < x \leqslant 3 \\ 2 \text{ for } 3 < x \leqslant 6. \end{cases}$$

Thus, the convex envelope of k_4 on $[0,6]$ is simply $x/3$. On the other hand, for $x = 5$ the continuous relaxation of the above MIMM for k_4 is easily seen to have optimal value $3/2$ for $x = 5$ (choose $y_1 = \frac{1}{2}$, $y_2 = 1$), and therefore it does *not* coincide with the convex envelope, which has value $5/3$ at $x = 5$.

APPENDIX C

We wish to establish relaxation-optimality in the case of both upper and lower bounds as considered in Corollary 1. Define

(C.1) $$f_3(x) \equiv \min_y cy$$
$$\text{s.t. } ay \geqslant x,$$
$$l \leqslant y \leqslant u'$$
$$y_i \text{ integer}, \ i \in I$$

and

(C.2) $$f_3^*(x) \equiv \min_y cy$$
$$\text{s.t. } ay \geqslant x$$
$$l \leqslant y \leqslant u',$$

where $l \geqslant 0$ and l_i and u_i' are integer for $i \in I$. By making the substitutions $y = z + l$, $x = t + al$, and $\tilde{u} = u' - l$, we have

$$f_3(x) = cl + \min_z cz$$
$$\text{s.t. } az \geqslant t, \ 0 \leqslant z \leqslant \tilde{u}, \ z_i \text{ integer}, \ i \in I$$
$$= cl + \tilde{f}_2(t) = cl + \tilde{f}_2(x - al),$$

where

$$\tilde{f}_2(t) \equiv \min_z cz$$
$$\text{s.t. } az \geqslant t, \ 0 \leqslant z \leqslant \tilde{u}, \ z_i \text{ integer}, \ i \in I.$$

Similarly, $f_3^*(x) = cl + \tilde{f}^*_2(x - al)$ where

$$\tilde{f}^*_2(t) \equiv \min_z cz$$
$$\text{s.t. } az \geqslant t, \ 0 \leqslant z \leqslant \tilde{u}.$$

STOCHASTIC MODELS FOR SPREAD OF MOTIVATING INFORMATION

Menachem Berg

*University of Haifa**
Haifa, Israel

ABSTRACT

In this work we consider spread of information which motivates the hearer to perform some specified action. The time to completion of an action is assumed to be a random variable and the main focus is on the number of completed actions by time t, $Y(t)$. Some models, which reflect different degree of centralization in the spread process, are analyzed and the distribution of $Y(t)$, as well as that of some other stochastic processes of interest, are obtained. The relevance to propagation of epidemics is pointed out.

All models are solved by employing two interrelated concepts, namely, the order statistics property of stochastic processes and the binomial closure property of collections of distributions. In this respect, the work also serves as an illustration of the application of these useful concepts.

1. INTRODUCTION

In this work we shall consider several spread of information models. While the term information is meant in a broad sense we are particularly referring to messages which motivate the hearers to perform some specified action. This could be a marketing leaflet which stimulates the reader to buy some commodity or a military call up order which requires the report of its recepient at some predetermined place. The spreading itself could be carried out by a single spreader (possibly a source), by means of a hierarchy of spreaders or by anyone who has heard the information. The models which will be discussed in this work corresponds to this varying degree of centralization in the spread process.

All models start with a single initial spreader — having more than one would merely require convoluting the results — and the spread rate is always of a homogeneous Poisson type. The time to completion of the specified action is assumed to be a random variable, independent from hearer to hearer, with a general cumulative distribution function $H(\cdot)$. It should be noted that an action need not involve physical efforts and may even be instantaneous so that $H(\cdot)$ is indeed, the c.d.f. of the period of time elapsed between the receipt of the information and the completion of the action.

The quantity we are mainly interested in is the number of hearers who have completed the action by time t or alternatively, the number of completed actions by time t. Besides computing the distribution of this stochastic process we shall also obtain the distribution of associated stochastic processes of interest such as the number of hearers up to time t or the number

*Presently visiting the Department of Quantitative Methods, University of Illinois at Chicago Circle

of responsive spreaders or hearers up to time t (when the possibility of "defection" is taken into account).

It is instructive to note that the above models bear relevance to propagation of epidemics. The vocabulary should then be translated as follows: Information-Disease, Spreader-Carrier, Hearer-Infectious, Source-Virus. The specified action could be interpreted as any event of interest such as recovery or the less fortunate outcome.

For literature on spread of rumors see Dietz [3] and Bartholomew [2]. A comprehensive treatise on spread of epidemics can be found in Bailey [1].

2. SOME PRELIMINARY RESULTS

Let us first present two concepts, which we shall use extensively in the sequel.

DEFINITION 1: A stochastic process with unit jumps, $Y(t)$, is said to have the order statistics (abbreviated: OS) property if conditioned on $Y(t) = n$, the unordered times of jumps are distributed as a random sample of size n from a c.d.f. $F_t(\cdot)$ which we shall call the kernal c.d.f.

NOTE: In this work we shall consider only processes with continuously distributed "inter-jump" intervals so that $F_t(\cdot)$ will always be a continuous function.

DEFINITION 2: A collection of discrete nonnegative distributions $\mathscr{P}$ is said to be binomially closed (abbreviated: BC) if for every $P \in \mathscr{P}$ and any $0 \leqslant \gamma \leqslant 1$ there exists a $\tilde{P} \in \mathscr{P}$ such that

$$N \sim P;\ X\Big|_{N=n} \sim \text{Binomial } (n, \gamma) \rightarrow X \sim \tilde{P}$$

or, restated, if N is distributed according to a member of $\mathscr{P}$ and the conditional distribution of X given $N = n$, is Binomial with parameters (n, γ), then the unconditional distribution of X is also a member of $\mathscr{P}$.

Of particular interest are collections which are parametric families of distributions depending on some parameter θ. In this case the above definition can be reworded as follows:

DEFINITION 2': A parametric family of distributions $\mathscr{P} = \{P_\theta, \theta \in \Theta\}$ is said to be BC if for every $\theta \in \Theta$ and any $0 \leqslant \gamma \leqslant 1$ there exists a $\tilde{\theta} \in \Theta$ such that

$$N \sim P_\theta;\ X\Big|_{N=n} \sim \text{Binomial } (n, \gamma) \rightarrow X \sim P_{\tilde{\theta}}$$

where $\tilde{\theta} = \tilde{\theta}(\theta, \gamma)$.

The function $\tilde{\theta}(\theta, \gamma)$ will be called the transformation function.

Examples of uniparametric BC families of distribution are:

1. The Poisson family of distributions

$$P_\theta(x) = e^{-\theta}\frac{\theta^x}{x!},\ x = 0, 1, 2, \ldots;\ \theta \in \Theta = [0, \infty).$$

In this case the transformation function is

(1) $$\tilde{\theta}(\theta, \gamma) = \theta\gamma.$$

2. The Binomial family of distributions

$$P_\theta(x) = \binom{K}{x}\theta^x(1-\theta)^{K-x}, \quad x = 0, \ldots, K; \quad \theta \in \Theta = [0, 1].$$

Here again

(2) $$\tilde{\theta}(\theta, \gamma) = \theta\gamma.$$

3. The Geometric family of distributions

$$P_\theta(x) = \theta(1-\theta)^x, \quad x = 0, 1, 2, \ldots; \quad \theta \in \Theta = [0, 1].$$

Here

(3) $$\tilde{\theta}(\theta, \gamma) = [1 + \gamma(\theta^{-1} - 1)]^{-1}.$$

A useful tool for verifying whether a particular collection of distribution is BC is provided by the following characterization theorem.

PROPOSITION 1: Let $\mathscr{P}$ be a collection of nonnegative discrete distributions and let $\mathscr{G}$ be the corresponding collection of moment generating functions where the m.g.f. associated with a distribution P is given by $G(z) = \sum_{x=0}^{\infty} z^x P(x)$. Then $\mathscr{P}$ is BC if and only if $\mathscr{G}$ is closed under a linear transformation of its independent variable, i.e., for every $P \in \mathscr{P}$ and any $0 \leqslant \gamma \leqslant 1$ there exists a $\tilde{P} \in \mathscr{P}$ such that

$$G(\gamma z + 1 - \gamma) = \tilde{G}(z)$$

where $G(\tilde{G})$ is the m.g.f. associated with $P(\tilde{P})$.

The proof of this Proposition is straightforward. When dealing with parametric families of distributions we have the equivalent:

PROPOSITION 1': A family of nonnegative discrete distribution $\mathscr{P} = \{P_\theta, \theta \in \Theta\}$ is BC if and only if for every $\theta \in \Theta$ and any $0 \leqslant \gamma \leqslant 1$ there exists a $\tilde{\theta} \in \Theta$ such that

(4) $$G_\theta(\gamma z + 1 - \gamma) = G_{\tilde{\theta}}(z)$$

where G_θ is the m.g.f. associated with P_θ.

Due to the one to one correspondence between distributions and m.g.fs, the transformation function $\tilde{\theta}(\theta, \gamma)$ is the same function in both collections.

COROLLARY: If the collection $\mathscr{P}$ is BC then the collection $\mathscr{P}^{(x)}$, formed by taking the x-th convolution of each member of $\mathscr{P}$, is BC too. The assertion is valid not only for positive integers x but for any positive real x for which there exists a corresponding collection $\mathscr{G}^{(x)}$ of proper m.g.fs. In the parametric family context we state that if $\mathscr{P} = \{P_\theta, \theta \in \Theta\}$ is BC then so is $\mathscr{P}^{(x)} = \{P_\theta^{(x)}, \theta \in \Theta\}$ where $P_\theta^{(x)}$ is the x-th convolution of P_θ with itself. In this case the transformation function $\tilde{\theta}(\theta, \gamma)$ remains invariant under the operation, i.e., it is independent of x.

The corollary follows immediately from Proposition 1 (or 1') due to the multiplicative property of m.g.fs.

The following proposition relates the two concepts of OS property of stochastic processes and BC property of collections of distributions.

PROPOSITION 2: If a stochastic process $N(t)$, with $N(0) = 0$, processes the OS property, then the collection of distributions of $N(t)$, $t \geq 0$:

$$\mathscr{P} = \{P_{N(t)},\ t \geq 0\}$$

is BC.

PROOF: From the OS property of $N(t)$, we can conclude that $P(N(s) = j/N(t) = n) = \binom{n}{j} F_t^j(s)\ (1 - F_t(s))^{n-j}$, for all $0 \leq s \leq t$ and all integers $0 \leq j \leq n$ (where $0^0 \equiv 1$). Hence,

$$P(N(s) = j) = \sum_{n=j}^{\infty} \binom{n}{j} F_t^j(s)\,(1 - F_t^{n-j}(s))\,P(N(t) = n),\quad j = 0, 1, 2, \ldots.$$

Multiplying both sides by z^j and summing over j from 0 to ∞ we obtain after some manipulations

$$G_{N(s)}(z) = G_{N(t)}(z\,F_t(s) + 1 - F_t(s))$$

where $G_{N(t)}(z) = \sum_{n=0}^{\infty} z^n\,P(N(t) = n)$, is the m.g.f. of the distribution of $N(t)$.

Hence, for every $t \geq 0$ and for any $0 \leq \gamma \leq 1$ there exists an $s (0 \leq s \leq t)$, such that

$$G_{N(s)}(z) = G_{N(t)}(z\gamma + 1 - \gamma),$$

which is the solution of equation

$$F_t(s) = \gamma. \tag{5}$$

Such a unique solution does exist since $F_t(s)$ is continuously increasing from 0 to 1 in the interval [0,t]. Proposition 2 now follows from Proposition 1.

We are now in a position to state the main theorem.

PROPOSITION 3: In an information spread process (of the type described in the Introduction) let $Y(t)$ be the number of hearers who initiated an action up to time t and let $X(t)$ be the number of completed actions by time t. Then, if the stochastic process $Y(t)$ possesses the OS property, the distribution of $X(t)$ belongs, for all $t \geq 0$, to the collection $\mathscr{P} = \{P_{X(t)},\ t \geq 0\}$.

PROOF: Assume that $Y(t) = n$. Then, since $Y(t)$ possesses the OS property, the unordered points of time at which the n hearers received the information are distributed as a random sample of size n from a c.d.f. $F_t(u)$. Moreover, the probability that a hearer who got the message at time $u\,(u \leq t)$ will complete the action by time t is $H(t - u)$. Combining these two facts we have

$$X(t)\Big|_{Y(t)=n} \sim \text{Binomial } (n,p)$$

where

(6) $$p = \int_0^t H(t-u)dF_t(u).$$

Now, by Proposition 2, the collection $\mathscr{P} = \{P_{Y(t)},\ t \geq 0\}$ is BC and hence, by the very definition of this property, the distribution of $X(t)$ belongs to $\mathscr{P}$ as well.

3. HIERARCHICAL SPREADING

We begin with a simple model in which a single spreader circulates a piece of information according to a Poisson process with parameter λ i.e., the "interhearing" times are exponentially distributed with parameter λ. Upon receiving the information, any hearer initiates an action whose time to completion is distributed according to a general c.d.f. $H(\cdot)$. It is assumed that an action can be initiated only when the information (which could be a leaflet or a form) has been received directly from the initial spreader.

By assumption, $N(t)$ is a Poisson process, viz.,

$$N(t) \sim \text{Poisson } (\lambda t).$$

It is well known that a Poisson process possesses the OS property with a kernal c.d.f.,

$$F_t(u) = \frac{u}{t}, \quad 0 \leq u \leq t,$$

so that by Proposition 3 the distribution of $X(t)$ belongs, for any $t \geq 0$, to the collection $\mathscr{P} = \{P_{X(t)},\ t \geq 0\}$. This collection, however, is identical with the Poisson family of distributions and therefore, by (1),

(7) $$X(t) \sim \text{Poisson } \left[\lambda \int_0^t H(u)du\right]$$

since here

$$\theta = \lambda t \text{ and, by (6), } \gamma = p = \frac{1}{t}\int_0^t H(u)du.$$

Thus,

(8) $$E[X(t)] = \lambda \int_0^t H(u)du$$

and

(9) $$G_{X(t)}(z) = \exp\left[-\lambda(1-z)\int_0^t H(u)du\right].$$

Let us now drop the assumption that all hearers do act and introduce a probability α for a hearer to be responsive and perform the action. The number of responsive hearers up to time t $Y(t)$, is again a Poisson process with parameter $\lambda\alpha$ which enables us to repeat the above arguments with $\lambda\alpha$ instead of λ. Therefore, by (7),

$$X(t) \sim \text{Poisson } \left[\lambda\alpha \int_0^t H(u)du\right].$$

A natural extension of the above single spreader model is achieved by designating some of the hearers as spreaders. These spreaders, however, do not perform the action. Specifically, we have an initial spreader who begins at time 0 to circulate the information among, what we shall call, second generation spreaders. These spreaders pass on the information to hearers who perform the action. All spreading is done according to a Poisson process with parameter λ. The total number of completed actions by time t, can be expressed as

$$X_2(t) = \sum_{i=1}^{N(t)} X_{i,1}(t)$$

where

$S(t)$ is the number of second generation spreaders who have received the information by time t,

$X_{(i)}(t)$ is the number of completed actions up to time t by hearers of the i-th second generation spreader. By assumption,

$$S(t) \sim \text{Poisson } (\lambda t)$$

and hence, by the OS property of the Poisson process,

$$G_{X_2(t)}(z) = \sum_{n=0}^{\infty} \left[\frac{1}{t} \int_0^t G_{X_{(i)}(t-y)}(z)\,dy \right]^n e^{-\lambda t} \frac{(\lambda t)^n}{n!} \tag{10}$$

$$= \exp\left[-\lambda t + \lambda \int_0^t G_{X_{(i)}(t)}(z)\,dy \right].$$

Substituting (9) into (10) yields,

$$G_{X_2(t)}(z) = \exp\left[-\lambda t + \lambda \int_0^t e^{\lambda(1-z)\int_0^y H(u)\,du}\,dy \right] \tag{11}$$

with

$$E[X_2(t)] = \left. \frac{dG_{X_2(t)}(z)}{dz} \right|_{z=1} = \lambda^2 \int_0^t (t-u) H(u)\,du. \tag{12}$$

The total number of people who know the information by time t (including second generation spreaders) can be represented as

$$N_2(t) = \sum_{i=1}^{S(t)} (N_{(i)}(t) + 1)$$

where,

$N_{(i)}(t)$ is the number of hearers of the i-th second generation spreader, up to time t.

Noting that the m.g.f. of $N_{(i)}(t)$ is obtainable from the m.g.f. of $X_{(i)}(t)$, by setting $H(u) = 1$ $(u \geq 0)$, and since $G_{N_{(i)}(t)+1}(z) = zG_{N_{(i)}(t)}(z)$, the m.g.f. of $N_2(t)$ can be shown to be

$$G_{N_2(t)}(z) = \exp\left[-\lambda t + \frac{z}{1-z}(1 - e^{-\lambda t(1-z)}) \right] \tag{13}$$

with $N_2(t)$

$$E[N_2(t)] = \lambda t + \frac{(\lambda t)^2}{2}.$$

If the possibility of "defection" is taken into account and we let $1 - \beta$ be the probability that a second generation spreader does not spread and $1 - \alpha$ be the probability that a hearer does not perform the action, then, repeating the above arguments, we obtain

$$G_{X_2(t)}(z) = \exp\left[-\lambda\beta t + \lambda\beta \int_0^t e^{\lambda\alpha(1-z)\int_0^y H(u)\,du}\,dy \right]$$

with

$$E[X_2(t)] = \lambda^2\alpha\beta \int_0^t (t-u)H(u)du.$$

We now proceed to consider a general spreading hierarchy. Thus, in a structure of order k, the process starts at time 0 with an initial spreader who circulates the information among the second generation spreaders, who pass it on to third generation spreaders and so on until the k-th generation spreaders spread the message through the rest of the population who perform the action. All spreading is assumed to be according to a Poisson process with parameter λ and the time to completion of the action has a c.d.f. $H(\cdot)$. Spreaders do not perform the action.

In order to obtain the distribution of $X_k(t)$ (the index k denotes the order of the spreading hierarchy), we first make the observation that a second generation spreader replicates, with regard to his branch, the role of the initial spreader for a structure of order $k-1$. Hence, using once more the OS property of the Poisson process, we obtain the recursive equation

$$G_{X_{k+1}(t)}(z) = \exp\left[-\lambda t + \lambda \int_0^t G_{X_k(y)}(z)dy\right], \quad k = 2, 3, 4, \ldots \tag{14}$$

where $G_{X_2(t)}(z)$ is given by (11).

Taking the derivative of (14) with respect to z and setting $z = 1$, yields a set of recursive equations for the expectations of $X_k(t)$ $(k = 2, 3, \ldots)$. Solving these equations, while recalling the initial value $E(X_2(t))$ in (12), we find

$$E[X_k(t)] = \frac{\lambda^k}{(k-1)!}\int_0^t (t-y)^{k-1}H(y)dy, \quad k = 2, 3, \ldots. \tag{15}$$

(In fact, both (14) and (15) also hold for $k = 1$ which represents a single spreader model.)

For small t, a higher order of the spreading hierarchy would not necessarily increase the expected number of completed actions—since spreaders do not perform the action—but for larger t this will be the case. When t tends to ∞ it can be shown, using an Abelian argument on Laplace transforms, that

$$E[X_{k+1}(t)] - E[X_k(t)] \underset{t\to\infty}{\to} \infty, \text{ for any finite } k.$$

Similar arguments with respect to $N_k(t)$—the total number of people who know the information by time t (including spreaders), yield the recursive equation

$$G_{N_{k+1}(t)}(z) = \exp\left[-\lambda t + \lambda z \int_0^t G_{N_k(y)}(z)dy\right], \quad k = 2, 3, \ldots \tag{16}$$

where $G_{N_2(t)}(z)$ is given by (13). It can be shown from (16) and (13) that

$$E[N_k(t)] = \sum_{i=1}^{k} \frac{(\lambda t)^i}{i!} \tag{17}$$

which indicates, as one would have intuitively expected, that the higher the order of the spreading hierarchy, the faster the spread of the information.

It is interesting to investigate the behavior of $X_k(t)$ and $N_k(t)$ when $k \to \infty$. For $X_k(t)$ we can get from (14) that

$$G_{X_\infty(t)}(z) = 1 \iff P[X_\infty(t) = 0] = 1, \tag{18}$$

which is not surprising since if everybody spreads there is no one to carry out the action. From (16) we can obtain an integral equation for the m.g.f. of $N_\infty(t)$, the solution of which is,

$$G_{N_\infty(t)}(z) = e^{-\lambda t}[1 - z(1 - e^{-\lambda t})]^{-1}. \tag{19}$$

The m.g.f. in (19) corresponds to the distribution

$$P[N_\infty(t) = n] = e^{-\lambda t}(1 - e^{-\lambda t})^n, \quad n = 0, 1, 2, \ldots$$

i.e.,

$$N_\infty(t) \sim \text{Geometric } (e^{-\lambda t}) \tag{20}$$

with

$$E[N_\infty(t)] = e^{\lambda t} - 1.$$

This result could have been obtained directly from (17).

An important generalization of the hierarchical spreading model arises when the spreading rate of the initial spreader, which could be a source, is different from those of the subsequent spreaders. Repeating the arguments in the above model, when the initial spreader circulates the information according to a Poisson process with rate μ, yields for $\tilde{X}_k(t)$ and $\tilde{N}_k(t)$ (which correspond to $X_k(t)$ and $N_k(t)$, respectively, in the ordinary case)

$$G_{\tilde{X}_{k+1}(t)}(z) = \exp\left[-\mu t + \mu \int_0^t G_{X_k(y)}(z)\,dy\right] \tag{21}$$

$$G_{\tilde{N}_{k+1}(t)}(z) = \exp\left[-\mu t + \mu z \int_0^t G_{N_k(y)}(z)\,dy\right] \quad k = 2, 3, \ldots. \tag{22}$$

Differentiating (21) and (22) with respect to z and setting $z = 1$, yields

$$E[\tilde{X}_k(t)] = \mu \frac{\lambda^{k-1}}{(k-1)!} \int_0^t (t-y)^{k-1} H(y)\,dy = \frac{\mu}{\lambda} E[X_k(t)]$$

and

$$E[\tilde{N}_k(t)] = \frac{\mu}{\lambda} \sum_{i=1}^{k} \frac{(\lambda t)^i}{i!} = \frac{\mu}{\lambda} E[N_k(t)].$$

when k tends to ∞, $\tilde{X}_\infty(t)$ behaves as $X_\infty(t)$, (see Equation (18)). For $\tilde{N}_\infty(t)$ we have, recalling (19),

$$G_{\tilde{N}_\infty(t)}(z) = e^{-\lambda t}[1 - z(1 - e^{-\lambda t})]^{-\mu/\lambda},$$

which corresponds to the distribution,

$$P(\tilde{N}_\infty(t) = n) = \binom{\mu/\lambda + n - 1}{n} e^{-\lambda t}(1 - e^{-\lambda t})^n, \quad n = 0, 1, \ldots.$$

That is,

$$\tilde{N}_\infty(t) \sim \text{Negative Binomial}\left(\frac{\mu}{\lambda}, e^{-\lambda t}\right). \tag{23}$$

4. FREE SPREAD OF INFORMATION

In this model we make no prior designation of spreaders and assume that every hearer may pass on the information in addition to performing the action. At first glance, it may look contradictory that a person can do both simultaneously, but one should bear in mind our introductory remark that an action need not involve physical efforts. In fact, an action could even be instantaneous in which case $H(\cdot)$ is the c.d.f. of the time until the action is taken.

As usual the process starts at time 0 with an initial spreader who circulates the information according to a Poisson process with parameter λ. Any hearer of the information initiates an action, whose time to completion is distributed according to a c.d.f. $H(\cdot)$, and, simultaneously, goes on spreading the information at the same rate (Poisson with parameter λ). The number of hearers up to time t, $N(t)$, should have the same distribution as $N_\infty(t)$ in the previous model, so that

$$N(t) \sim \text{Geometric } (e^{-\lambda t}). \tag{24}$$

This result is also obtainable by the following argument. The time at which the n-th person received the message T_n, can be expressed as the sum of the successive "interhearing" periods of the first n hearers. It can now be observed that these periods correspond, in reverse order, to the "interfailure" periods of a system which is composed of n units in parallel each having an exponential lifetime distribution. T_n is therefore distributed as the lifetime of this system, i.e.,

$$P(T_n \leqslant t) = (1 - e^{-\lambda t})^n, \quad t \geqslant 0,$$

which, recalling the relation $P(T_n \leqslant t) = P(N(t) \geqslant n)$, yields (24).

The process $N(t)$ possesses the OS property [4] with a kernel c.d.f.

$$F(u) = \frac{e^{\lambda u} - 1}{e^{\lambda t} - 1} \qquad 0 \leqslant u \leqslant t. \tag{25}$$

Hence, the distribution of $X(t)$ belongs, for any $t \geqslant 0$, to the collection $\mathcal{P} = \{P_{X(t)}, t \geqslant 0\}$ which coincides with the Geometric family of distributions. Therefore, using (3), we have

$$X(t) \sim \text{Geometric } \left\{\left[1 + e^{\lambda t}\int_0^t \lambda e^{-\lambda u} H(u)\,du\right]^{-1}\right\} \tag{26}$$

with

$$E[X(t)] = e^{\lambda t}\int_0^t \lambda e^{-\lambda u} H(u)\,du$$

since here

$$\theta = e^{-\lambda t}$$

and, by (6),

$$\gamma = p = (1 - e^{-\lambda t})^{-1}\int_0^t \lambda e^{-\lambda u} H(u)\,du. \tag{27}$$

Let us now generalize the model by making the response of the hearers to both spreading and acting probabilistic. More precisely, we assume that every hearer is either interested or uninterested, with probabilities β and $1 - \beta$ respectively, where uninterested hearers neither spread nor act while those interested do spread but still may not perform the action with probability $1 - \alpha$.

Letting $S(t)$ be the number of interested hearers up to time t, it can be verified that $S(t)$ is the same type of birth process as $N(t)$, only with $\lambda\beta$ instead of λ. Thus,

$$S(t) \sim \text{Geometric } (e^{-\lambda\beta t}). \tag{28}$$

Using the OS property of $S(t)$ and the BC property of the Binomial family of distributions, with transformation function $\tilde{\theta}(\theta, \gamma) = \theta\gamma$, it can be verified that

$$X(t)\Big|_{S(t)=n} \sim \text{Binomial } (n, \alpha\tilde{p}), \tag{29}$$

where

$$\tilde{p} = (1 - e^{-\lambda\beta t})^{-1}\int_0^t \lambda\beta e^{-\lambda\beta u} H(u)\,du.$$

Applying Proposition 3 and using (3) with $\theta = e^{-\lambda\beta t}$ and $\gamma = \alpha\bar{p}$ we obtain

$$X(t) \sim \text{Geometric}\left\{\left[1 + \alpha e^{\lambda\beta t}\int_0^t \lambda\beta e^{-\lambda\beta u}H(u)du\right]^{-1}\right\} \tag{30}$$

with

$$E[X(t)] = \alpha e^{\lambda\beta t}\int_0^t \lambda\beta e^{-\lambda\beta u}H(u)du.$$

Like in the hierarchical spreading model we can now generalize this model by allowing the rate of the initial spreader (which could be a source) to be different from those of the other spreaders. Thus, if the initial spreader circulates the information according to a Poisson process with parameter μ, the distribution of $N(t)$ should be identical to that of $\tilde{N}_\alpha(t)$ in the hierarchical spreading model (Equation (23)), i.e.,

$$N(t) \sim \text{Negative Binomial}\left(\frac{\mu}{\lambda},\ e^{-\lambda t}\right).$$

It can be shown that the process $N(t)$ possesses the OS property with the kernel probability distribution function in (25). Furthermore, the m.g.f. of a Negative Binomial distribution with parameters (x,θ) is the m.g.f. of a Geometric distribution with parameter θ, taken to the power x $(x > 0)$. Hence, using the corollary of Proposition 1', we can conclude that the Negative Binomial family of distributions with parameter $\theta \in [0,1]$ is BC, for any $x > 0$ with $\tilde{\theta}$ given by (3). Therefore, by Proposition 3,

$$X(t) \sim \text{Negative Binomial}\left(\frac{\mu}{\lambda},\ \left[1 + e^{\lambda t}\int_0^t \lambda e^{-\lambda u}H(u)du\right]^{-1}\right)$$

with

$$E[X(t)] = \frac{\mu}{\lambda}e^{\lambda t}\int_0^t \lambda e^{-\lambda u}H(u)du.$$

5. SPREAD BY A SOURCE

In this model, we have a source (some media) which, form time 0 on, transmits a piece of information to a population of size N. Any member of the population may hear the information in any interval $(t,t+\Delta t)$, independently of other members, with probability $\lambda\Delta t + 0(\Delta t)$, at which moment he initiates an action whose time to completion has a c.d.f. $H(\cdot)$. The distribution of the number of hearers up to time t, $N(t)$, is

$$N(t) \sim \text{Binomial}\ (N, 1 - e^{-\lambda t}),$$

since the c.d.f. of the time until any one of them will hear the information is given by

$$L(u) = 1 - e^{-\lambda u}, \quad u \geqslant 0.$$

Moreover, the stochastic process $N(t)$ possesses the OS property with a kernel c.d.f.

$$F_t(u) = \frac{1 - e^{-\lambda u}}{1 - e^{-\lambda t}}, \quad 0 \leqslant u \leqslant t.$$

Applying now Proposition 3 and using (2) we obtain

$$X(t) \sim \text{Binomial}\left(N, e^{-\lambda t}\int_0^t \lambda e^{\lambda u}H(u)du\right)$$

with

$$E[X(t)] = Ne^{-\lambda t}\int_0^t \lambda\ e^{\lambda u}H(u)du.$$

since here $\theta = 1 - e^{-\lambda t}$ and, by (6),

$$\gamma = p = (e^{\lambda t} - 1)^{-1} \int_0^t \lambda e^{\lambda u} H(u)\,du. \tag{31}$$

Let us now relax the assumption that all the population is exposed to the source and introduce a probability β that a member of the population will hear the information at all. We further make the response of the hearers to the stimulus probabilistic and let α be the probability that a hearer does initiate an action.

Denote by Z the number of people who are exposed to the source. Then,

$$Z \sim \text{Binomial } (N,\beta).$$

Given $Z = m$ the conditional distribution of $N(t)$, the number of hearers (out of the m exposed) up to time t, is

$$N(t)\Big|_{Z=m} \sim \text{Binomial } (m, 1 - e^{-\lambda t}).$$

Using the OS property of $N(t)$ we can show, (like in the previous model—see Equation (29)), that

$$X(t)\Big|_{N(t)=n,\, Z=m} \sim \text{Binomial } (n,\alpha p), \quad 0 \leqslant n \leqslant m \leqslant N$$

where p is given by (31).

Applying Proposition 3 and then unconditioning with respect to Z (which amounts to one more use of the BC property of the Binomial family of distribution) we finally obtain

$$X(t) \sim \text{Binomial } \left(N,\alpha\beta\, e^{-\lambda t} \int_0^t \lambda\, e^{\lambda u} H(u)\,du\right)$$

with

$$E[X(t)] = N\alpha\beta\, e^{-\lambda t} \int_0^t \lambda\, e^{\lambda u} H(u)\,du.$$

6. MORE GENERAL SPREAD PROCESSES

Throughout this work we have assumed that the spread rate is of a homogeneous Poisson type. In this section we shall employ our procedure to solve the nonhomogeneous case.

Specifically, assume that the spread rate of any active spreader at time t is a function of t: $\lambda(t)$. Beginning with the single spreader model we have the well known result

$$N(t) \sim \text{Poisson } (\Lambda(t))$$

where

$$\Lambda(t) = \int_0^t \lambda(u)\,du.$$

The nonhomogeneous Poisson process also possesses the OS property with a kernel c.d.f.

$$F_t(u) = \frac{\Lambda(u)}{\Lambda(t)} \quad 0 \leqslant u \leqslant t$$

so that by following the arguments in the homogeneous case we can show that

$$X(t) \sim \text{Poisson } \left(\int_0^t \lambda(u) H(t-u)\,du\right)$$

with

$$E[X(t)] = \int_0^t \lambda(u) H(t-u)\,du$$

and

$$G_{X(t)}(z) = \exp\left[-(1- \quad)\int_0^t \lambda(u) H(t-u)\,du\right].$$

Continuing to a hierarchical spread structure of order 2 we have

$$G_{X(t)}(z) = \sum_{n=0}^{\infty} e^{-\Lambda(t)} \frac{\Lambda^n(t)}{n!} \left[\int_0^t \frac{\lambda(y)}{\Lambda(t)} G_{X(t,y)}(z)\,dy\right]^n$$

$$= \exp\left[-\Lambda(t) + \int_0^t \lambda(y)\, G_{X(t,y)}(z)\,dy\right]$$

where $X(y,t)$ is the number of completed actions by time t generated by a single spreader who operates in the time interval $[y,t]$. The m.g.f. of this r.v. is given by

$$G_{X(\cdot)}(z) = \exp\left[-(1-z)\int_y^t \lambda(u) H(t-u)\,du\right].$$

In a similar way we can obtain results for higher orders of spreading structures.

Proceeding to the free spread model, the solution of the Kolmogorov backward equations for the probabilities $P(N(t) = n)$ $n = 0, 1, 2, \ldots$, yields

$$N(t) \sim \text{Geometric } (e^{-\Lambda(t)}).$$

It can also be directly verified that $N(t)$ possesses the OS property with a kernel c.d.f.

$$F_t(u) = \frac{e^{\Lambda(u)} - 1}{e^{\Lambda(t)} - 1}, \quad 0 \leqslant u \leqslant t.$$

Repeating the arguments in the homogeneous case we finally obtain here

$$X(t) \sim \text{Geometric } \left(\left[1 + \int_0^t \lambda(u) e^{\Lambda(u)} H(t-u)\,du\right]^{-1}\right)$$

with

$$E[X(t)] = \int_0^t \lambda(u) e^{\Lambda(u)} H(t-u)\,du.$$

Turning to the last model, in which the information is spread by a source, we now assume that each member of the population may hear the information in the interval $(t, t+\Delta t)$ with probability $\lambda(t)\Delta t + o(\Delta t)$. We then have

$$N(t) \sim \text{Binomial } (N,\ 1 - e^{-\Lambda(t)})$$

and moreover, the process $N(t)$ possesses the OS property with a kernel c.d.f.

$$F_t(u) = \frac{1 - e^{-\Lambda(u)}}{1 - e^{-\Lambda(t)}}, \quad 0 \leqslant u \leqslant t.$$

Applying Proposition 3 yields here

$$X(t) \sim \text{Binomial } \left(N, \int_0^t \lambda(u)\, e^{-\Lambda(u)} H(t-u)\,du\right)$$

with

$$E[X(t)] = N\int_0^t \lambda(u)\, e^{-\Lambda(u)} H(t-u)\,du.$$

As a matter of fact, our procedure can be used in this model to obtain a complete general solution.

Recalling our definition of $L(\cdot)$ as the c.d.f. of the time, since the beginning of the transmission of the information, until a member of the population hears it, we have

$$N(t) \sim \text{Binomial } (N, L(t)).$$

It is not difficult to observe that by its very nature, the process $N(t)$, possesses the OS property with a kernel c.d.f.

$$F_t(u) = \frac{L(u)}{L(t)}, \quad 0 \leqslant u \leqslant t.$$

Using our procedure, we finally obtain

$$X(t) \sim \text{Binomial } \left[N, \int_0^t H(t-u)dL(u)\right]$$

with

$$E[X(t)] = N \int_0^t H(t-u)dL(u).$$

Note that in this case the distribution of $X(t)$ could also have been obtained directly by defining a "success," for any member of the population, as the event of having accomplished the action by time t.

REFERENCES

[1] Bailey, N.T.J., *The Mathematical Theory of Infectious Diseases and Its Applications*, (Griffin, London, 1975).

[2] Bartholomew, D.J., *Stochastic Models for Social Processes*, (2nd edition, John Wiley and Sons, New York, N.Y., 1973).

[3] Dietz, K., "Epidemics and Rumours: A Survey," Journal of the Royal Statistical Society Series, A *130*, 505-528 (1967).

[4] Neuts, M. and S.I. Resnick, "On the Times of Births in a Linear Birth Process," Journal of the Australian Mathematical Society, *12*, 473-475 (1971).

MAXIMAL NASH SUBSETS FOR BIMATRIX GAMES

M. J. M. Jansen

Department of Mathematics
Catholic University
Nijmegen, The Netherlands

ABSTRACT

In this work maximal Nash subsets are studied in order to show that the set of equilibrium points of a bimatrix game is the finite union of all such subsets. In addition, the extreme points of maximal Nash subsets are characterized in terms of square submatrices of the payoff matrices and dimension relations are derived.

1. INTRODUCTION

A *bimatrix game* is defined by a pair (A,B) of real $m \times n$-matrices. A strategy for player I (II) is an element of $S^m (S^n)$, where $S^n := \{p \in \mathbb{R}^n;\ p \geq 0,\ \sum_{i=1}^{n} p_i = 1\}$. Corresponding to the strategy pair $(p,q) \in S^m \times S^n$ the payoffs are pAq^t and pBq^t, respectively.

A pair $(p,q) \in S^m \times S^n$ is called an *equilibrium point* of the $m \times n$-bimatrix game (A,B) if $pAq^t = \max_{p \in S^m} pAq^t$ and $pBq^t = \max_{q \in S^n} pBq^t$. The set of all equilibrium points of (A,B), which is nonempty by a theorem of J. F. Nash [9,10], will be denoted by $E(A,B)$.

NOTATION: For a natural number m, let $\mathbb{N}_m := \{1, \ldots, m\}$. The elements of the basis of unit vectors of $\mathbb{R}^m$ are denoted by $e_1, \ldots, e_m$. For a finite set S, $|S|$ is the number of elements of S. The convex hull of a set $S \subset \mathbb{R}^m$ is denoted by $\mathrm{conv}(S)$. If $C \subset \mathbb{R}^m$ is a convex set, then we write $\mathrm{ext}(C)$, $\dim(C)$ and $\mathrm{relint}(C)$ for the set of extreme points of C, the dimension of (the affine hull of) C and the relative interior of C, respectively.

Let (A,B) be an $m \times n$-bimatrix game and let $(p,q) \in S^m \times S^n$. It is well-known (Cf. [7], theorem 4) that $(p,q) \in E(A,B)$ iff $C(p) \subset M(A;q)$ and $C(q) \subset M(p;B)$, where $C(p)$ (the *carrier* of p) $:= \{i \in \mathbb{N}_m;\ p_i > 0\}$, $C(q) := \{j \in \mathbb{N}_n;\ q_j > 0\}$, $M(A;q) := \{i \in \mathbb{N}_m;\ e_i Aq^t = \max_{k \in \mathbb{N}_m} e_k Aq^t\}$ and $M(p;B) := \{j \in \mathbb{N}_n;\ pBe_j^t = \max_{k \in \mathbb{N}_n} pBe_k^t\}$.

The organization of the paper is as follows. In Section 2 we show that the set of equilibrium points of a bimatrix game is the union of convex polytopes. The equilibrium point set can therefore be constructed if we know the extreme points of these convex polytopes. These so-called extreme equilibrium points are studied in the third section. As a by-product we find that the set of equilibria is in fact a finite union. Finally, dimension relations are given for the convex polytopes mentioned before.

2. THE STRUCTURE OF MAXIMAL NASH SUBSETS

DEFINITIONS. Let (A,B) be a bimatrix game and let $S \subset E(A,B)$. We call two equilibrium points (p,q), $(p',q') \in S$ *S-interchangeable* if $(p,q') \in S$ and $(p',q) \in S$. We call two equilibrium points *interchangeable* if they are $E(A,B)$-interchangeable. We call S a *Nash subset* for the game (A,B) if every pair of equilibrium points in S is S-interchangeable. A Nash subset S is called a *maximal Nash subset* for the game (A,B) if there exists no Nash subset $T \subset E(A,B)$ such that S is properly contained in T.

The term maximal Nash subset was first introduced by G. A. Heuer and C. B. Millham in [4]. J. F. Nash, who already considered such sets in 1951 [10], called them *sub-solutions*. These authors showed that a maximal Nash subset for an $m \times n$-bimatrix game is a closed and convex subset of $S^m \times S^n$. The following theorem implies that a maximal Nash subset is in fact the Cartesian product of two convex polytopes.

THEOREM 1. Let (A,B) be an $m \times n$-bimatrix game and let S be a maximal Nash subset for the game (A,B). Suppose that $(\mathring{p},\mathring{q}) \in \text{relint}(S)$. Then $S = K(\mathring{q}) \times L(\mathring{p})$, where $K(\mathring{q}) = \{p \in S^m; (p,\mathring{q}) \in E(A,B)\}$ and $L(\mathring{p}) = \{q \in S^n; (\mathring{p},q) \in E(A,B)\}$ are convex polytopes.

PROOF. Let $\pi_1(S) = \{p \in S^m;$ there exists a $q \in S^n$ with $(p,q) \in S\}$ and $\pi_2(S) = \{q \in S^n,$ there exists a $p \in S^m$ with $(p,q) \in S\}$. Since it is clear that $S = \pi_1(S) \times \pi_2(S)$, the theorem is proved if we can show that $\pi_1(S) = K(\mathring{q})$ and $\pi_2(S) = L(\mathring{p})$. The inclusions $\pi_1(S) \subset K(\mathring{q})$ and $\pi_2(S) \subset L(\mathring{p})$ are immediate. Suppose that $p \in K(\mathring{q})$ and $q \in \pi_2(S)$. Since $q \in \text{relint}\,\pi_2(S)$, Theorem 6.4 of [11] implies that there is a $q' \in \pi_2(S)$ and a $\lambda \in (0,1)$ such that $q = \lambda q + (1-\lambda)q'$. From $q \in \pi_2(S) \subset L(\mathring{p})$ it follows that $\mathring{p} \in K(q)$. Also, $\mathring{p} \in K(q')$. Hence, $K(q) \cap K(q') \neq \phi$ and Lemma 3.5 of [4] implies that $K(\mathring{q}) = K(q) \cap K(q')$. So $p \in K(q)$ and $\{p\} \times \pi_2(S) \subset E(A,B)$. If $p \notin \pi_1(S)$, then $\text{conv}(\pi_1(S) \cup \{p\}) \times \pi_2(S)$ is a Nash subset properly containing the maximal Nash subset S. This leads to a contradiction. So $p \in \pi_1(S)$ and we have proved that $K(\mathring{q}) \subset \pi_1(S)$. In a similar manner, one can show that $L(\mathring{p}) \subset \pi_2(S)$. Finally, it is well-known that $K(\mathring{q})$ and $L(\mathring{p})$ are convex polytopes. ||

The following Lemma can be proved in the same way as Theorem 1 in [2].

LEMMA 1. Let (A,B) be a bimatrix game. If C is a convex subset of $E(A,B)$, then every pair of equilibrium points in C is interchangeable.

It is well-known that a maximal Nash subset is a convex set not properly contained in any other convex subset of the set of equilibrium points. This property is characteristic for maximal Nash subsets as we will prove now.

THEOREM 2. Let (A,B) be a bimatrix game and let C be a convex subset of $E(A,B)$ not properly contained in any other convex subset of $E(A,B)$. Then C is a maximal Nash subset for the game (A,B).

PROOF. (a) First we prove that $\tilde{C} = \{(p,q) \in E(A,B)$; there exists an $(x,y) \in E(A,B)$ with (x,q), $(p,y) \in C\}$ is a convex set. If (p,q), $(\bar{p},\bar{q}) \in \tilde{C}$, then there exist (x,y), $(\bar{x},\bar{y}) \in E(A,B)$ such that (x,q), $(p,y) \in C$ and $(\bar{x},\bar{q})$, $(\bar{p},\bar{y}) \in C$. But then $(\lambda x + (1-\lambda)\bar{x}, \lambda q + (1-\lambda)q)$, $(\lambda p + (1-\lambda)\bar{p}, \lambda y + (1-\lambda)\bar{y}) \in C$, for all $\lambda \in (0,1)$. In view of the foregoing Lemma, we may conclude that $(\lambda x + (1-\lambda)\bar{x}, \lambda y + (1-\lambda)\bar{y}) \in E(A,B)$ and $(\lambda p + (1-\lambda)\bar{p},$

$\lambda q + (1-\lambda)q) \in E(A,B)$, for all $\lambda \in (0,1)$. Consequently, $\lambda(p,q) + (1-\lambda)(\bar{p},\bar{q}) \in \tilde{C}$, for all $\lambda \in (0,1)$. (b) Also, $\tilde{C}$ is a Nash subset. If (p,q), $(p,\bar{q}) \in \tilde{C}$, then there exist (x,y), $(\bar{x},\bar{y})$ $\in E(A,B)$ as in (a). Note that (x,y), $(\bar{x},\bar{y}) \in \tilde{C}$. So Lemma 1 implies (with $\tilde{C}$ in the role of C) that $(x,y) \in E(A,B)$. Similarly, $(p,\bar{q}) \in E(A,B)$. Since $(p,\bar{q})$, $(\bar{x},y) \in E(A,B)$ and (p,y), $(\bar{x},\bar{q}) \in \tilde{C}$, it follows that $(p,\bar{q}) \in \tilde{C}$. Similarly, $(\bar{p},q) \in \tilde{C}$, and (p,q) and $(\bar{p},q)$ are $\tilde{C}$-interchangeable. (c) Because $\tilde{C}$ is convex and $C \subset \tilde{C}$, it follows that $C = \tilde{C}$. So, in view of (b), C is a Nash subset. It is obvious that, in addition, C is a maximal Nash subset. ||

COROLLARY 1 (Cf. [2], Theorem 1): If (A,B) is a bimatrix game, then $E(A,B)$ is convex if and only if $E(A,B)$ is a Nash subset.

REMARK 1. Let (A,B) be a bimatrix game and let $(p,q) \in E(A,B)$. Since $\{(p,q)\}$ is a Nash subset for the game (A,B), we can, applying Zorn's lemma, find a maximal Nash subset containing (p,q). Consequently, every equilibrium point of the game (A,B) is contained in a maximal Nash subset and $E(A,B)$ is the union of such subsets.

3. EXTREME POINTS OF MAXIMAL NASH SUBSETS

For a matrix game L. S. Shapley and R. N. Snow [12] characterized all pairs of extreme optimal strategies of the players. We want to describe for the case of bimatrix games, the extreme points of the maximal Nash subsets. Our approach incorporates the work of H. W. Kuhn [5] and O. L. Mangasarian [6].

DEFINITION: An equilibrium point of a bimatrix game (A,B) is called an *extreme equilibrium point* if it is an extreme point of some maximal Nash subset for the game (A,B).

In [6], O.L. Mangasarian introduced, for an $m \times n$-bimatrix game (A,B), the convex polyhedral sets $P_B := \{(p,\beta) \in S^m \times \mathbb{R};\ pBe_j^t \leqslant \beta \text{ for all } j \in \mathbb{N}_n\}$ and $Q_1 := \{(q,\alpha) \in S^n \times \mathbb{R};\ e_i A q^t \leqslant \alpha \text{ for all } i \in \mathbb{N}_m\}$. These sets play also a role in the proof of the following

THEOREM 3: The set of equilibrium points of a bimatrix game is a (not necessarily disjunct) union of a finite number of maximal Nash subsets.

PROOF: Let S be a maximal Nash subset for the game (A,B) and suppose that $(p,q) \in \text{ext}(S)$ and that $(\mathring{p},\mathring{q}) \in \text{relint}(S)$. Then, by Theorem 1, we have $p \in \text{ext}(K(\mathring{q}))$ and $q \in \text{ext}(L(\mathring{p}))$. The reader can easily prove that this implies that $(p,pB\mathring{q}^t) \in \text{ext}(P_B)$ and that $(q,\mathring{p}Aq^t) \in \text{ext}(Q_1)$ (Cf. [5], Lemma 1). Hence, if (p,q) is an extreme equilibrium point of the game (A,B), then $(p,pBq^t,q,pAq^t) \in \text{ext}(P_B) \times \text{ext}(Q_1)$. Since $\text{ext}(P_B)$ and $\text{ext}(Q_1)$ are finite sets, the number of extreme equilibrium points of the game (A,B) is also finite. Hence, the number of maximal Nash subsets is finite. ||

REMARK 2: In [6], O. L. Mangasarian called an element $(p,q,\alpha,\beta) \in S^m \times S^n \times \mathbb{R} \times \mathbb{R}$ an extreme equilibrium point of the $m \times n$-bimatrix game (A,B), if $(p,\beta) \in \text{ext}(P_B)$, $(q,\alpha) \in \text{ext}$ (Q_1) and $p(A+B)q^t = \alpha + \beta$. It is easy to show that a point (p,q,α,β) is an extreme equilibrium point in the sense of O. L. Mangasarian if and only if (p,q) is an extreme equilibrium point in the sense of definition 1 and if furthermore $\alpha = pAq^t$ and $\beta = pBq^t$. Therefore, Theorem 3 implies the Lemma on page 779 of [6].

REMARK 3: The extension of Theorem 3 to the case of more than two players does not necessarily hold. On page 3 of [2], H. H. Chin, T. Parthasarathy and T. E. S. Raghavan give an example of noncooperative 3-person game, where all the players have the set S^2 as strategy space and where the set of equilibrium points is equal to the (convex) set $\{((\lambda, 1-\lambda), (\lambda, 1-\lambda), (\lambda, 1-\lambda)) \in S^2 \times S^2 \times S^2; \lambda \in [0,1]\}$. This set of equilibrium points is the union of an uncountable number of maximal Nash subsets.

For a proof of the following theorem, see Lemma 2 of H. W. Kuhn [5].

THEOREM 4: Let (A,B) be an $m \times n$-bimatrix game. If (p,q) is an extreme equilibrium point of the game (A,B) and γ is the number of elements of the carrier of q, then there exists a $\gamma \times \gamma$-submatrix K of A such that [renumber, if necessary, the rows and columns of A in such a way that K is in the upper left corner of A]

(1) the $(\gamma+1) \times (\gamma+1)$-matrix $\tilde{K} := \begin{bmatrix} K & -1 \\ 1 & 0 \end{bmatrix}$ is nonsingular,

(2) $q_j = (\det(\tilde{K}))^{-1} \sum_i K_{ij}$ if $j \in C(q)$ and

[K_{ij} is the cofactor of the element k_{ij}]

(3) $pAq^t = \det(K)/\det(\tilde{K})$.

An analogous statement can be formulated with respect to the connection of the vector p and the number pBq^t with a certain square submatrix of B.

REMARK 4: Let (A,B) be a bimatrix game. Without loss of generality we may suppose that $A > 0$ and $B < 0$. Let S be a maximal Nash subset for the game (A,B). Suppose that $(p,q) \in \text{relint}(S)$ and that $L(p) = \{q\}$. Note that the proof of Theorem 4 is based on the fact that the rank of the matrix $A(S) := [a_{ij}]_{i \in M(A,q), j \in C(q)}$ equals $|C(q)|$. Using the fact that $A > 0$, Theorem 4 (3) implies that $\dim L(p) = |C(q)| - \text{rank } A(S)$. We shall see in Theorem 5 that a similar statement holds for sets $L(p)$ with more than one element. If $K(q) = \{p\}$, then $\dim K(q) = |C(p)| - \text{rank } B(S)$, where $B(S) := [b_{ij}]_{i \in C(p), j \in M(p,B)}$.

4. A DIMENSION RELATION FOR MAXIMAL NASH SUBSETS

The purpose of this section is to extend the dimension relations as given by C. B. Millham in [8]. The relations derived below include, in contrast to the results in Millham's paper, those for the zero-sum case (Cf. [1], [3]).

LEMMA 2: Let (A,B) be a bimatrix game and let S be a maximal Nash subset for (A,B). Suppose that $(\mathring{p},\mathring{q}) \in \text{relint}(S)$. Then, for all $(p,q) \in S$, $C(p) \subset C(\mathring{p})$, $C(q) \subset C(\mathring{q})$, $M(A;q) \supset M(A;\mathring{q})$ and $M(p;B) \supset M(\mathring{p};B)$.

PROOF: Suppose that $p \in K(\mathring{q})$, $p \neq \mathring{p}$. Because $\mathring{p} \in \text{relint } K(\mathring{q})$, there exist a $\tilde{p} \in K(\mathring{q})$ and a $\lambda \in (0,1)$ such that $\mathring{p} = \lambda p + (1-\lambda)\tilde{p}$. This implies that $C(p) \subset C(\mathring{p})$. Now, for $j \in M(\mathring{p};B)$,

$$\mathring{p}B\mathring{q}^t = \mathring{p}Be_j^t = \lambda pBe_j^t + (1-\lambda)\tilde{p}Be_j^t \leq \lambda pB\mathring{q}^t + (1-\lambda)\tilde{p}B\mathring{q}^t = \mathring{p}B\mathring{q}^t.$$

This is possible only if $pBe_j^t = pB\mathring{q}^t$. So $j \in M(p;B)$ and we have proved that $M(\mathring{p};B) \subset M(p;B)$. The other assertions are proved in a similar way. ||

DEFINITION: Let (A,B) be a bimatrix game and let S be a maximal Nash subset for the game (A,B). In view of Lemma 2, the matrices

$$A(S) := [a_{ij}]_{i \in M(A;q),\, j \in C(q)} \text{ and } B(S) := [b_{ij}]_{i \in C(p),\, j \in M(p,B)}$$

do not depend on the choice of the point $(p,q) \in \text{relint}(S)$. We call $A(S)$ and $B(S)$ the S-submatrices of A and B, respectively.

THEOREM 5: Let (A,B) be an $m \times n$-bimatrix game with $A > 0$ and $B < 0$. Let S be a maximal Nash subset for the game (A,B). If $(p,q) \in \text{relint}(S)$,

then (1) $\dim L(p) = |C(q)| - \text{rank } A(S)$

and (2) $\dim K(q) = |C(p)| - \text{rank } B(S)$.

PROOF: We only prove (1). If $L(p)$ has only one element, we are finished (Remark 4). Suppose now that $L(p)$ contains more than one element. There is no loss of generality in supposing that $C(q) = \{1, \ldots, \gamma\}$, where $\gamma = |C(q)|$. Let $d := \gamma - \text{rank } A(S)$. Choose a basis $\hat{x}(1), \ldots, \hat{x}(d)$ of $\text{Ker } A(S) := \{x \in \mathbb{R}^\gamma;\ A(S)x^t = 0\}$ in such a way that, for each $k \in \mathbb{N}_d$, $\hat{q} + \hat{x}(k) > 0$, where $\hat{q} := (q_1, \ldots, q_\gamma)$, and $pAq^t - e_iAq^t > e_iAx(k)^t$ for each $i \notin M(A;q)$, where $x(k) := (\hat{x}(k), 0, \ldots, 0) \in \mathbb{R}^n$. We normalize the vectors $q + x(k)$ in such a way that the normalized result $y(k)$ is an element of S^n. We leave it to the reader to show that the vectors $q, y(1), \ldots, y(d)$ are linearly independent vectors in $L(p)$. Hence, $\dim L(p) \geqslant d$. Suppose now that there exists a vector $y(d+1) \in \text{relint } L(p)$ such that the vectors $y(1) - q, \ldots, y(d+1) - q$ are linearly independent. Then, in view of Lemma 2, $C(y(d+1)) = C(q)$ and $M(A;y(d+1)) = M(A;q)$. So if $\hat{y}(k) := (y(k)_1, \ldots, y(k)_\gamma)$, for each $k \in \mathbb{N}_{d+1}$, then $A(S)\{\hat{y}(k)/pAy(k)^t - \hat{q}/pAq^t\} = 0$, for each $k \in \mathbb{N}_{d+1}$. This is impossible since $\dim \text{Ker } A(S) = d$. So $\dim L(p) = d$. ||

It is easy to prove that Theorem 1 in [8] is implied by Theorem 5.

For a matrix game A, the only maximal Nash subset is the set S of all pairs of optimal strategies for both players. In this case, the S-submatrix of A equals the essential submatrix of A (Cf. [3], page 44) and the dimension relation for matrix games follows from Theorem 5.

ACKNOWLEDGMENT

The author is indebted to Dr. S. H. Tijs for many helpful comments and discussions. Also the suggestions of the referee are gratefully acknowledged.

REFERENCES

[1] Bohnenblust, H.F., S. Karlin and L.S. Shapley, "Solutions of Discrete, Two Person Games," Annals of Mathematics Studies, *24*, 51-72 (1950).

[2] Chin, H., T. Parthasarathy and T.E.S. Raghavan, "Structure of Equilibria in N-person Non-Cooperative Games," International Journal of Game Theory, *3*, 1-19 (1974).

[3] Gale, D. and S. Sherman, "Solutions of Finite Two-person Games," Annals of Mathematics Studies, *24*, 37-49 (1950).

[4] Heuer, G.A. and C.B. Millham, "On Nash Subsets and Mobility Chains in Bimatrix Games," Naval Research Logistics Quarterly, *23*, 311-319 (1976).

[5] Kuhn, H.W., "An Algorithm for Equilibrium Points in Bimatrix Games," Proceedings, National Academy of Science U.S.A., *47*, 1656-1662 (1961).

[6] Mangasarian, O.L., "Equilibrium Points of Bimatrix Games," Journal of the Society of Industrial and Applied Mathematics, 12, 778-780 (1964).
[7] Millham, C.B., "On Nash Subsets of Bimatrix Games with Special Properties," Naval Research Logistics Quarterly, 19, 709-714 (1972).
[8] Millham, C.B., "On Nash Subsets of Bimatrix Games," Naval Research Logistics Quarterly, 21, 307-317 (1974).
[9] Nash, J.F., "Equilibrium Points in n-person Games," Proceedings National Academy of Science U.S.A., 36, 48-49 (1950).
[10] Nash, J.F., "Non-Cooperative Games," Annals of Mathematics, 54, 286-295 (1951).
[11] Rockafellar, R.T., Convex Analysis, Princeton, (Princeton, Princeton University Press, 1970).
[12] Shapley, L.S. and R.N. Snow, "Basic Solutions of Discrete Games," Annals of Mathematics Studies, 24, 27-35 (1950).
[13] Vorobev, N.N., "Equilibrium Points in Bimatrix Games," Theory of Probability and Its Applications, 3, 297-309 (1958).

A CHARACTERIZATION OF THE VALUE OF ZERO-SUM TWO-PERSON GAMES

S. H. Tijs

Department of Mathematics
Catholic University
Nijmegen, The Netherlands

ABSTRACT

For the family D consisting of those zero sum two-person games which have a value, the value-function on D is characterized by four properties called objectivity, monotony, symmetry and sufficiency.

INTRODUCTION

In a beautiful paper, E. I. Vilkas gave a characterization of the value-function, defined on the class of all finite matrix games [2]. In [1], pp. 60-65 this result was extended to the class of all finite and semi-infinite matrix games.

The purpose of this paper is to deduce characterizing properties for the value-function on the set of all determined two-person games. The organization of the paper is as follows: the necessary notation and definitions are given in sections 1 and 2; in section 3, properties for the value-function are presented, which are shown in section 4 to be characteristic of this function.

1. A *(zero-sum) two-person game* is an ordered triple $< X, Y, K >$, in which X and Y are nonempty sets (called the *pure strategy spaces* of player I and player II, respectively) and $K : X \times Y \to \mathbb{R}$ is a real-valued function on the Cartesian product of X and Y (called the *pay off function* of player I).

2. Let $< X, Y, K >$ be a two-person game. For each $x \in X$ ($x \in Y$) let us denote the probability measure on X (Y) with mass 1 in x (y) by e_x (e_y). Let P_X be the set of all convex combinations of elements of $\{e_x : x \in X\}$, likewise let P_Y be the convex hull of $\{e_y : y \in Y\}$. Then the two-person game $< P_X, P_Y, E_K >$ with

$$E_K(\mu, \nu) := \int\int K(x,y)\, d\mu(x)\, d\nu(y) \text{ for each } (\mu, \nu) \in P_X \times P_Y$$

is called the *e-mixed extension of the game* $< X, Y, K >$. The *lower value* $\sup_{\mu \in P_X} \inf_{\nu \in P_Y} E_K(\mu, \nu)$ of the game $< P_X, P_Y, E_K >$ is denoted by $\underline{v}(X, Y, K)$ and the *upper value* $\inf_{\nu \in P_Y} \sup_{\mu \in P_X} E_K(\mu, \nu)$ is denoted by $\bar{v}(X, Y, K)$. Note that

$$-\infty \leq \underline{v}(X, Y, K) \leq \bar{v}(X, Y, K) \leq \infty.$$

If $\underline{v}(X, Y, K) = \bar{v}(X, Y, K)$ for a game $< X, Y, K >$, then we say that the game is a *determined game*. In that case, the common value is denoted by $v(X, Y, K)$ and called the *value* of (the e-mixed extension of) $< X, Y, K >$. The family of determined games is denoted by D.

3. In this section we want to look at some distinguished properties of the value-function $v : D \to [-\infty, \infty]$. For this purpose we need some definitions.

DEFINITION 1: The *transpose* of a two-person game $<X,Y,K>$ is the two-person game $<Y,X,-K>$ where

$$K'(y,x) = K(x,y) \text{ for each } (y,x) \in Y \times X.$$

DEFINITION 2: Let $<X,Y,K>$ be a two-person game and let S be a nonempty subset of X. Then we say that S is *sufficient for player I in the game* $<X,Y,K>$ if for each $x \in X - S$ there exists a $\mu \in P_S$ such that

$$E_K(\mu, e_y) \geq K(x,y) \text{ for each } y \in Y.$$

DEFINITION 3: Let $<X,Y,K>$ be a two-person game and let T be a nonempty subset of Y. We say that T *is sufficient for player II in the game* $<X,Y,K>$ if T is sufficient for player I in the game $<Y,X,-K'>$.

THEOREM 1:

(P.1) ["Objectivity"] Let $<X,Y,K>$ be a two-person game and suppose that $X = \{a\}$, $Y = \{b\}$. Then $<X,Y,K> \in D$ and $v(X,Y,K) = K(a,b)$.

(P.2) ["Monotonicity"] Let $<X,Y,K> \in D$ and $<X,Y,L> \in D$ and suppose that $L \geq K$ (i.e. $L(x,y) \geq K(x,y)$ for each $(x,y) \in X \times Y$). Then $v(X,Y,L) \geq v(X,Y,K)$.

(P.3) ["Symmetry"] Let $<X,Y,K> \in D$. Then $<Y,X,-K'> \in D$ and $v(Y,X,-K') = -v(X,Y,K)$.

(P.4) ["Sufficiency"] Let $<X,Y,K>$ be a two-person game, and $\phi \neq S \subset X$ and let $K' : S \times Y \to \mathbb{R}$ be the restriction of K to $S \times Y$. Suppose that S is sufficient for player I in the game $<X,Y,K>$. Then $<S,Y,K'> \in D$ iff $<X,Y,K> \in D$ and

$$v(X,Y,K) = v(S,Y,K') \text{ if } <S,Y,K> \in D.$$

PROOF: (P.1) and (P.2) are obvious. (P.3) follows from the fact that

$$-E_K(\mu,\nu) = E_{-K}(\nu,\mu) \text{ for each } (\mu,\nu) \in P_X \times P_Y.$$

Now let us prove (P.4). First we note that P_S can be seen (in an obvious manner) as a subset of P_X, and that E_K is the restriction of E_K to $P_S \times P_Y$.

Take $\alpha \in P_X$. Then there exist $n \in \mathbb{N}$, $x^1, x^2, \ldots, x^n \in X$ and $p_1, p_2, \ldots, p_n \in [0,\infty)$ such that $\sum_{i=1}^{n} p_i = 1$ and $\alpha = \sum_{i=1}^{n} p_i e_{x^i}$. Since S is sufficient for player I in the game $<X,Y,K>$ for each $i \in \{1, \ldots, n\}$, there exists an $\alpha_i \in P_S$ such that

$$E_K(\alpha_i, e_y) \geq K(x^i, y) \text{ for each } y \in Y.$$

[If $x^i \in S$, then we can take $\alpha_i = e_{x^i}$.] Let $\bar{\alpha} := \sum_i p_i \alpha_i$. Then $\bar{\alpha} \in P_S$ and

$$E_K(\bar{\alpha}, e_y) = \sum_i p_i E_K(\alpha_i, e_y) \geq \sum_i p_i K(x^i, y) = E_K(\alpha, e_y) \text{ for each } y \in Y.$$

But then,

(i) $E_K(\bar{\alpha}, \nu) \geq E_K(\alpha, \nu)$ for each $\alpha \in P_1$ and each $\nu \in P_2$.

This implies that

$$\sup_{\mu \in P_1} E_K(\mu, \nu) = \sup_{\alpha \in P_1} E_K(\alpha, \nu) \text{ for each } \nu \in P_2$$

and thus

(ii) $v(S, Y, K') = v(X, Y, K)$.

From (i) we may also conclude that

$$\inf_{\nu \in P_2} E_K(\bar{\alpha}, \nu) \geq \inf_{\nu \in P_2} E_K(\alpha, \nu) \text{ for each } \alpha \in P_1$$

and then

(iii) $\underline{v}(S, Y, K') = \underline{v}(X, Y, K)$.

Now (P.4) follows from (ii) and (iii). ||

4. The following theorem shows that the properties (P.1)-(P.4) characterize the value-function $v : D \to [-\infty, \infty]$.

THEOREM 2: Let $f : D \to [-\infty, \infty]$ be a function with the following four properties:

(Q.1) If $X = \{a\}$, $Y = \{b\}$ and if K is a real-valued function on $X \times Y$, then $f(X, Y, K) = K(a, b)$.

(Q.2) For each $<X, Y, K> \in D$, $<X, Y, L> \in D$ with $L \geq K$: $f(X, Y, L) \geq f(X, Y, K)$.

(Q.3) For each $<X, Y, K> \in D$: $f(Y, X, -K') = -f(X, Y, K)$.

(Q.4) For each $<X, Y, K> \in D$ and $<S, Y, K'> \in D$, where $S \subset X$, K' is the restriction of K to $S \times Y$ and where S is sufficient for player I in the game $<X, Y, K>$, we have $f(S, Y, K') = f(X, Y, K)$.

Then $f(X, Y, K) = v(X, Y, K)$ for each $<X, Y, K> \in D$.

PROOF: First we note that (Q.3) and (Q.4) imply

(Q.5) For each $<X, Y, K> \in D$ and $<X, T, K''> \in D$, where $T \subset Y$, K'' is the restriction of K to $X \times T$ and where T is sufficient for player II in the game $<X, Y, K>$, we have $f(X, T, K'') = f(X, Y, K)$.

Now take an $<X, Y, K> \in D$ with $v(X, Y, K) \in (-\infty, \infty]$ and take a real number t such that $v(X, Y, K) > t$. We want to prove that $f(X, Y, K) \geq t$. For this purpose we introduce the following five two-person games:

(1) $<X \cup \{a\}, Y, L>$ where $a \notin X$ and where $L(x, y) := K(x, y)$ for each $(x, y) \in X \times Y$ and $L(a, y) := t$ for each $y \in Y$.

(2) $<X \cup \{a\}, Y, M>$ where $M(x, y) :=$ minimum $\{K(x, y), t\}$ for each $(x, y) \in (X \cup \{a\}) \times Y$.

(3) $\langle X \cup \{a\}, Y \cup \{b\}, N \rangle$ where $b \notin Y$ and where $N(x,y) = M(x,y)$ for each $(x,y) \in (X \cup \{a\}) \times Y$ and $N(x,b) = t$ for each $x \in X \cup \{a\}$

(4) $\langle \{a\}, Y \cup \{b\}, N' \rangle$ where N' is the restriction of N to $\{a\} \times Y \cup \{b\})$.

(5) $\langle \{a\}, \{b\}, N'' \rangle$ where N'' is the restriction of N to $\{a\} \times \{b\}$

Since $v(X,Y,K) > t$, there exists a $\mu \in P_X$ such that

$$E_f(\mu, e_j) \geq E_K(\mu, e_j) \geq t = f(a,y) \text{ for each } y \in Y$$

Hence, X is sufficient for player I in the game $\langle X \cup \{a\}, Y, L \rangle$. By (P.4) and (Q.4) we may conclude that

(Q.6) $\langle X \cup \{a\}, Y, L \rangle \in D$ and $f(X \cup \{a\}, Y, L) = f(X, Y, K)$

It follows from (Q.1) that

(Q.7) $f(\{a\}, \{b\}, N) = t$.

In the game $\langle \{a\}, Y \cup \{b\}, N \rangle$ the set $\{b\}$ is sufficient for player II because

$$E_N(e_i, e_b) = N(a,y) \leq t \text{ for each } y \in Y$$

Then $\langle \{a\}, Y \cup \{b\}, N \rangle \in D$ in view of (P.3) and (P.4); now (Q.5) implies

(Q.8) $f(\{a\}, Y \cup \{b\}, N) = f(\{a\}, \{b\}, N)$.

In the game $\langle X \cup \{a\}, Y \cup \{b\}, N \rangle$ the set $\{a\}$ is sufficient for player I because for each $x \in X$

$$t = E_N(e_x, e_b) \geq N(x,y) \text{ for each } y \in Y \cup \{b\}.$$

By (P.4) and (Q.4) we obtain: $\langle X \cup \{a\}, Y \cup \{b\}, N \rangle \in D$ and

(Q.9) $f(X \cup \{a\}, Y \cup \{b\}, N) = f(\{a\}, Y \cup \{b\}, N)$.

It is easy to see that Y is sufficient for player II in the game $\langle X \cup \{a\}, Y \cup \{b\}, N \rangle$. Hence, by (P.3) and (P.4): $\langle X \cup \{a\}, Y, M \rangle \in D$, and then by (Q.5):

(Q.10) $f(X \cup \{a\}, Y, M) = f(X \cup \{a\}, Y \cup \{b\}, N)$.

Now $L \geq M$ and then by (Q.2) we have

(Q.11) $f(X \cup \{a\}, Y, L) \geq f(X \cup \{a\}, Y, M)$.

Combining (Q.6)-(Q.11) we obtain: $f(X,Y,K) \geq t$. Thus, we have proved that $f(X,Y,K) \geq t$ for each $\langle X,Y,K \rangle \in D$ with $v(X,Y,K) \in (-\infty, \infty]$ and each $t < v(X,Y,K)$. But then

(Q.12) $f(X,Y,K) \geq v(X,Y,K)$ for each $\langle X,Y,K \rangle \in D$.

It follows from (Q.3), (Q.12) and (P.3) that

(Q.13) $f(X,Y,K) = -f(Y,X,K) \leq -v(Y,X,K) = v(X,Y,K)$ for each $\langle X,Y,K \rangle \in D$.

Properties (Q.12) and (Q13) imply the conclusion of the theorem. □

REFERENCES

[1] Tijs, S.H., *Semi-Infinite and Infinite Matrix Games and Bimatrix Games*, Ph.D. dissertation, Department of Mathematics, Catholic University, Nijmegen, The Netherlands (1975).

[2] Vilkas, E.I., "Axiomatic Definition of the Value of a Matrix Game," Theory of Probability and its Applications *8*, 304-307 (1963).

MANPOWER MODELING IN COST EFFECTIVENESS STUDIES OF USAF PROGRAM TO REDUCE THE INCIDENCE OF HEART DISEASE*

Clifford C. Petersen

Purdue University
West Lafayette, Indiana

ABSTRACT

Planning for a cardiovascular disease reduction program, soon to be initiated by the United States Air Force, has required an evaluation of its expected cost effectiveness. During the course of this evaluation, it was necessary to consider manpower flows and their expected changes in response to the disease reduction program. This paper describes several manpower models that were applied: a simple expected value equilibrium model, a cross-sectional model that considered the length of service of personnel, and a staffing model used to optimize the allocation of paramedics to the many Air Force bases of various sizes. The relevance of these models to the cost effectiveness evaluation is shown but the detailed cost effectiveness results are not presented.

Analyses are being performed to evaluate the cost effectiveness of a U.S. Air Force health program that is soon to be initiated. The "Health Evaluation and Risk Tabulation" (HEART) program will be directed toward cardiovascular disease that strikes several hundred Air Force personnel annually and results in a considerable loss of personnel through death and disability.

THE HEART PROGRAM

In very general terms, the HEART program will involve processing all military personnel in the Air Force to establish each individual's risk of future heart disease, followed by treatment of those found to be at high risk. This will be done by measurement of systolic blood pressure, serum cholesterol, glucose intolerance, and determining heart abnormality (left ventricular hypertrophy) by means of an electrocardiogram. Also, it will be determined whether the individual smokes cigarettes regularly. These data and age are used with the risk coefficients developed through the Framingham Study [2] to calculate for the individual the probability of occurrence of a cardiovascular incident within eight years. The coefficients are based on over 20 years of followup on a large civilian population, and have succeeded in clustering about 25 percent of the heart incidents into the top decile of risk. The possibility of coefficient modification and the inclusion of other risk indicators is being anticipated in the USAF program.

The calculated risks will serve to identify the most susceptible fraction of the USAF for treatment, and recalculation after treatment will serve, in some measure, to show the

*Based on part of the research performed for the USAF School of Aerospace Medicine by Purdue University under Contract F33615-77-C-0624.

improvement that was obtained. Obviously, the ultimate benefit will become apparent only in the long term when the actual incidence of heart disease can be observed. In addition to the treatment of high risk personnel, all personnel will be reached through a general education program to encourage improved dietary habits and cessation of smoking.

Various analyses are being directed toward therapy effectiveness, threshold selection policies, the effect of measurement error, and operational procedures, as well as toward the evaluation of cost effectiveness. Statistical and probability models and extensive computer simulation are being used. This paper, however, will describe only the application of manpower planning models to the determination of the cost effectiveness of the HEART program. The population numbers and dollar costs that will be used herein are altered and somewhat incomplete but serve for illustrative purposes; the actual analysis used the complete and most recent information on population, turnover of personnel, pay scales, and policies. The complete cost effectiveness analysis will not be presented as it is only intended here to show the applicability of several manpower planning models to that analysis.

THE COST REDUCTION PROBLEM

Only the costs to the U.S. government that will be affected by the HEART program need be considered. The major present costs that will be changed are those associated with USAF personnel departing from service and their subsequent replacement. It is necessary to identify and associate costs with the various ways in which personnel leave the Air Force. These costs are different for enlisted personnel and officers because of pay scales, and different for flyers (pilots and navigators) and nonflyers because of the considerable cost of training a replacement flyer. An additional cost, estimated at $1,000,000 per year, is that due to loss of aircraft because of heart attacks suffered by the pilots.

The various types of departure will now be described briefly. Voluntary and involuntary separation (or simply separation) includes resignation, failure to reenlist, and reduction-in-force terminations. Except in the case of flyers, these types of departure are considered to incur negligible costs. Voluntary retirement (or simply retirement) occurs when an individual retires with from 20 to 30 years of service. The departure cost is substantial, including payment of 50 to 75 percent of the individual's salary to the individual or his spouse for a period usually in excess of 30 years. Disability retirement, disability separation, and assignment to the temporary disability retirement list (TDRL) are forms of departure for reason of 30% or more disability, and must be considered separately for cardiovascular (CV) disabilities and other (non-CV) disabilities. Because cardiovascular related separations and TDRL's practically always become permanent, they are lumped with CV disability-retirements in this analysis. The departure cost is substantial, including hopitalization and continuing payments to the individual and to the spouse over a period usually in excess of 30 years. Departure by death is self-explanatory and its cost is analogous to that for disability retirements.

In determining the cost per retirement, disability retirement, or death, it seems reasonable that the long series of benefits paid to the individual or spouse should be discounted. It is perhaps not surprising that it was difficult to determine what rate to use, and that the agreed upon approach was to use two rates, 5 and 10 percent, for separate analyses. The cost of CV departures decreases by approximately 20% when changing from 5% to 10% discounting.

Table 1 summarizes the approximate costs for each type of departure:

TABLE 1. *Cost of Each Departure from U.S. Air Force (Thousands of Dollars)*

	Officers		Enlisted Personnel
	Flyers	Nonflyers	All Categories
Separation	280.0	0	0
Retirement	474.1	194.1	109.9
CV Disability-Retirement	531.9	251.0	153.7
CV Death	431.7	151.3	85.3
Non-CV Disability-Retirement	449.8	155.8	72.0
Non-CV Death	353.6	81.1	32.4
All future obligations brought to present worth using a 5 percent discount rate.			

Given the cost for each departure, and knowing the present average number of CV disability-retirement and CV death departures over the past several years, it is simple to calculate the annual departure cost due to cardiovascular disease. The anticipated effectiveness of therapy in reducing CV incidence, through the HEART program, can then be assumed (we've used 20 percent here). A naive approach to determining the cost reduction is to claim 20 percent of the annual CV departure cost (from which the operating cost of the HEART program would be subtracted to obtain the net annual savings). This approach, however, neglects the effect of the reduction in CV departures upon the other types of departure. It neglects, for example, the possibility that a person saved from CV death may be killed in another way, or that he must ultimately leave in some manner, typically incurring a departure cost. Also ignored is the beneficial effect of the HEART program in delaying the occurrence of heart attacks in the individuals who will still suffer them.

To deal with the interaction between the various types of departure, two different models were formulated. Both are based on the assumption of a steady state manpower system.

THE STEADY STATE SYSTEM

Although the U.S. Air Force will probability never be in a true steady state condition, it is as reasonable to use such a condition for the analysis as to hypothesize any other unknown future state. The strategy is to model a steady state force having the same size, distribution of personnel, and departure rates as the present force, and then to hypothesize a 20 percent reduction in the cardiovascular departure rates and determine what the new steady state condition would be. The difference in annual departure costs associated with the two systems would be attributable to the 20 percent reduction in CV incidence. Proportional cost changes would result from any other assumed reduction in CV incidence.

There are two primary requirements that must be satisfied in order to maintain steady state. Obviously, the annual number of new entries must equal the annual number of departures for each class of personnel. In addition, for each class, the total length of service in years of all persons departing in one year, must equal the number of persons in the system (a consequence of N man-years of service being accumulated each year by a force of size N).

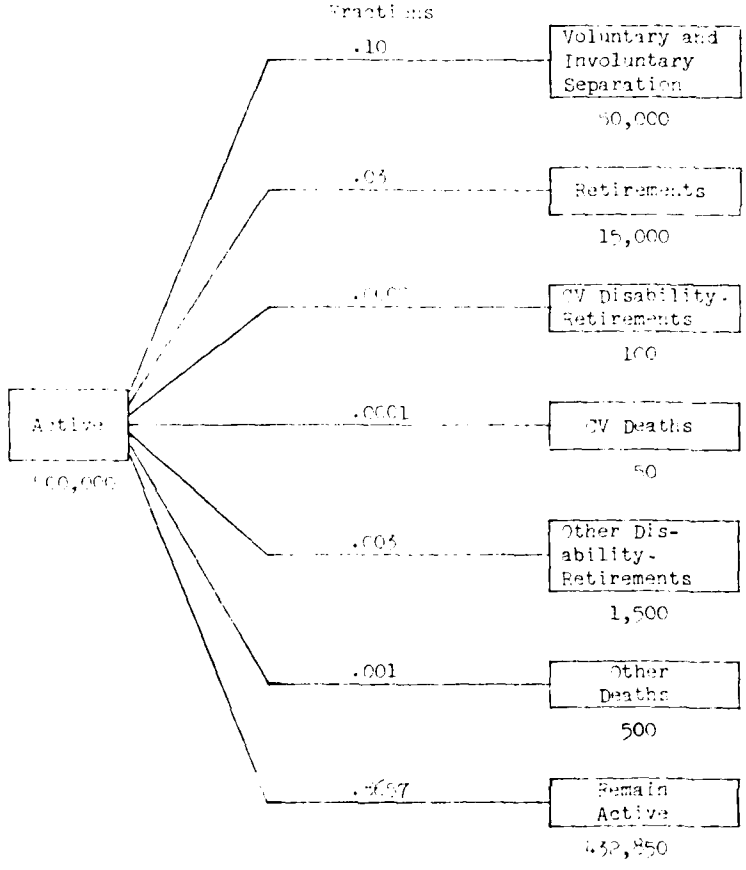
In the present U.S. Air Force, the number of new entrants is less than the number of departures for enlisted men. Also, the total length of service for enlisted men departing per

year exceeds the number of this class of personnel, and confirms a shrinking force with a large fraction of the population having many years of service. The steady state turnover rate indicated by length of service of those departing (reciprocal of average length of service) is somewhat below the actual turnover rate being experienced. In contrast, the situation is reversed for enlisted women. Nevertheless, with certain assumptions as to future recruitment and incentives it seems reasonable to conceive of a model of the USAF at steady state.

EXPECTED VALUE EQUILIBRIUM MODEL

This model for adjusting the other departure rates as the CV departure rates decrease is simple and requires little data. It ignores the length of service requirement for steady state, simply assuming that it will be met.

The initial steady state flows is illustrated for enlisted men in Figure 1 and requires that we know only the steady state total population count which will remain constant, the fractions for the various types of departure, and consequently the fraction remaining active each year.



No. of departures = 67,150

Turnover rate = 13.430 percent

FIGURE 1 Annual flow rates, enlisted men, before HEART program

The CV departure rates are then considered as being reduced by 20 percent. It is assumed that the manpower flow that would have departed due to CV disease is diverted to the other types of departure, and to continuing service, in proportion to their respective rates. A rationalization of this assumption is possible by viewing the departure fractions or rates as probabilities. Individuals who win a reprieve from CV disease, and will have to be routed to other types of departure or to continuing service, are distributed according to the appropriate probabilities. At steady state, with reduced CV departure rates, we recalculate new rates for other types of departure and extend them to numbers of departures as shown in Figure 2.

Flows are adjusted for the other classes of personnel in a similar manner. For pilots and navigators, however, there is a large replacement training cost for the increased voluntary and involuntary separations.

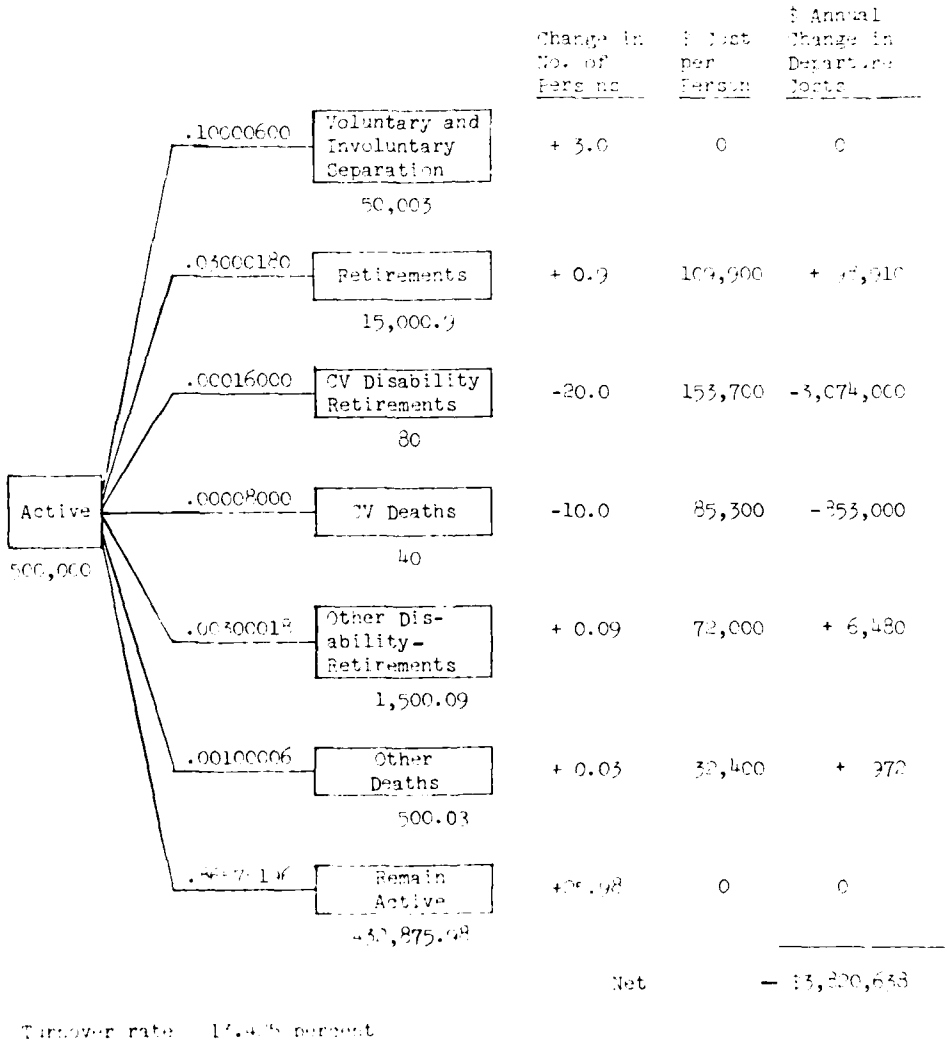


FIGURE 2. Annual flow rates, enlisted men, after HEART program

TWO-CHARACTERISTIC CROSS-SECTIONAL MODEL

This model [1] was chosen to impose the effect of length of service on the analysis. This effect is important because it is presumed that the effect of implementing the HEART program will be not only to reduce the rate of CV departures but to postpone the time of departure of the fraction who will still depart because of cardiovascular reasons. The data requirements are reasonable, not requiring detailed tracking of cohorts, but primarily adding data regarding the average length of service at departure.

In this model we define a matrix, P, of one-step transition probabilities, where each state is described by two characteristics, a status and a length of service. The model has more capability than will be used, as it serves our need by defining only one status, namely "active," rather than various ranks, for example. The analysis will be performed separately for each class of personnel and we will assume no flow of personnel between classes, such as from enlisted to officer or vice versa.

Given the matrix P, completely defined by knowing the average length of service at time of departure and the fraction of total departures for each type of departure, we note that the limit P^n will be the steady state transition matrix. This matrix will have identical rows, Π, where the j th element, π_j, is the proportion of the population in state j at equilibrium. The vector Π is determined by solving

$$\Pi = \Pi P$$

$$\sum_{\text{All } j} \pi_j = 1.$$

The strategy will be to find the steady state departures for our initial data, then to reduce the CV departure probabilities by 20 percent and increase the length of service to CV departure by an estimated two years, and again find the steady state condition. Considering the cost of each type of departure, the annual savings in departure costs due to the effect of the HEART program will then be calculated.

EXAMPLE: Use of the P matrix will be demonstrated with a very small example. Following this the enlisted personnel will be analyzed to show some of the adjustment that had to be made in our assumptions, and to give results for comparison with those of the expected value equilibrium model.

Suppose there is an organization with one class of personnel and three types of departure. Each year from now on five persons will resign after two years of service, five will be disabled after three years of service, and fifteen will retire after five years of service. Since the total length of service of the departing personnel is 100 man-years, the size of the organization must be 100 at steady state. All departing personnel are immediately replaced. We wish to find the steady state distribution of personnel by length of service.

Each year 25 persons enter the system and each year's group behaves as follows:

End of Year	Fraction Remaining For the Next Year
1	1.0
2	0.8
3	0.6
4	0.6
5	0

Defining active duty states (A, n), where n is the number of years of completed service, the one-step transition matrix, P, is as shown below. The probabilities of changing in one year from state A, i to state $A, i + 1$ for $i = 0, 1, 2, 3, 4$ are obtained from the table of fraction remaining, shown above. The probabilities of changing from state A, i to state $A, 0$ are probabilities of leaving the system, in which event a new person enters the system with 0 years of service.

		To:				
		$A, 0$	$A, 1$	$A, 2$	$A, 3$	$A, 4$
From:	$A, 0$		1.			
	$A, 1$	0.2		0.8		
	$A, 2$	0.25			0.75*	
	$A, 3$					1.
	$A, 4$	1.				

*Fraction going from A,2 to A,3 = 0.6/0.8 = .75

Solving $\Pi = \Pi P$ and $\sum_{\text{All}_i} \pi_i = 1$, we obtain $\Pi = [0.25, 0.25, 0.20, 0.15, 0.15]$.

The interpretation of this solution is that for this organization, there will be 25 percent new personnel, 25 percent who have completed one year of service, 20 percent who have completed two years, 15 percent who have completed three years, and 15 percent who have completed four years of service and who will retire at year end.

If the P matrix and its resulting equilibrium distribution Π are judged applicable, that is, if the mechanisms underlying the departures are such that the numbers of departures of the different types stay in the same relative proportions, then for any size organization the numbers of departures at steady state can easily be derived.

To conclude this example, assume that the desired size of the organization is 120, and that $\Pi = [0.25, 0.25, 0.20, 0.15, 0.15]$ still applies as the distribution of length of service. From π_0 we know that 25% of the organization (30 persons) will depart each year. These departures are then prorated over the types of departure as

$$\text{Resignations} = 30 \times 5/25 = 6$$

$$\text{Disabilities} = 30 \times 5/25 = 6$$

$$\text{Retirements} = 30 \times 15/25 = 18$$

Enlisted Personnel—Initial

The force size for enlisted personnel is 500,000. However, the presently observed numbers of departures from the Air Force as shown in Table 2(a) are not consistent with a steady state model with 500,000 population. In fact, they imply an equilibrium population of 545,450. One way to retain our observed departure information in a steady state model having a population of 500,000 is to decrease the number of each type of departure to 500,000/545,000 of its observed value. This is based on the equilibrium requirement that:

$$\sum_{\text{All}_i} n_i s_i = N$$

TABLE 2. *Calculation of Enlisted Steady State Departures Before HEART Program*

(a)	Average Length of Service, (Years)	Observed Number Departing
Separation	4	50,000
Retirement	22	15,000
CV Disability-Retirement	20	100
CV Deaths	19	50
Non-CV Disability-Retirement	6	1,500
Non-CV Deaths	7	500
		67,150
(Implied steady state population: 545,450)		
(b) The annual number of departures for a force of 500,000 at steady state, with proportional scaling, are:		
	Separation	45,833.71
	Retirement	13,750.12
	CV Disability-Retirement	91.67
	CV Death	45.83
	Non-CV Disability-Retirement	1,375.01
	Non-CV Death	458.34
		61,554.68
(c) The annual number of departures for a force of 500,000 at steady state, with selective scaling, are:		
	Separation	47,000.
	Retirement	13,480.
	CV Disability-Retirement	100.
	CV Death	50.
	Non-CV Disability-Retirement	1,500.
	Non-CV Death	500.
		62,630.

where n_i is the number of departures of the i th type, s_i is their average length of service, and N is the total population. Because all s_i remain the same, a change in N can be accommodated by changing all n_i proportionally as shown in Table 2(b). Proportional scaling, as just described, seems valid in making small adjustments, but for large adjustments such as this the reasonableness of the effect on each type of departure deserves examination.

Another way to construct the steady state model is to decrease the numbers of departures selectively. The data on the number of disability-retirements and deaths, whether from CV disease or other causes, as observed in a present force of 500,000 enlisted personnel should not be treated cavalierly. They should not be adjusted appreciably in the initial steady state model because there is no logical basis for reducing the incidence in contradiction to the medical records. The types of departures that can logically be reduced to achieve a hypothesized steady state Air Force, are separations and retirements, assuming that Air Force inducements and policies were modified to effect such reductions. A reasonable assumption is that the annual number of separations can be reduced about 6 percent (from 50,000 to 47,000). Leaving the number of departures for disability-retirement and death unaltered, we derive the required annual number of retirements, n_r, from

$$\sum_{\text{All } i} n_i s_i = N.$$

$47,000(4) + n_r(22) + 109(20) + 50(19) + 1,500(6) + 500(7) = 500,000$ obtaining 13,480 retirements annually. The steady state results are shown in Table 2(c). These results will be used to represent the initial departure distribution of enlisted personnel, before installation of the HEART program.

Enlisted Personnel—After

It was shown, in the example presented earlier, that if given an initial set of data comprising the annual number of each type of departure and the average age at departure, the steady state size of the population can be calculated. Also, an equilibrium distribution, H, can be calculated to describe the distribution by length of service (as well as the annual number of each type of departure).

If the initial set of data is perturbed, a new population size and new numbers of each type of departure can be calculated for steady state. The desired population size can be restored by proportional or selective scaling.

The perturbation applied to the initial steady state data for enlisted personnel is the assumed effect of the HEART program, that is, a reduction of CV departures by 20 percent and an increase of two years in the average age of those departing because of CV disease. Table 3(a) shows the steady state result for these assumptions, and an implied population of 499,660. The population was restored to 500,000 by proportional scaling. Table 3(b) shows the result, after scaling, and the change from the "before HEART program" result of Table 2(c).

TABLE 3. *Calculation of Enlisted Steady State Departures After HEART Program*

(a)	Average Length of Service (Years)	Assumed Number Departing
Separation	4	47,000.00
Retirement	22	13,480.00
CV Disability-Retirement	22	80.00
CV Death	21	40.00
Non-CV Disability-Retirement	6	1,500.00
Non-CV Death	7	500.00
		62,600.00
(Implied steady state population: 499,660)		

(b) The annual number of and changes in departures for a force of 500,000 at steady state, after proportional scaling, are:

	Number Departing	Change in Enlisted Personnel Departing Due to HEART Program
Separation	47,031.96	+31.96
Retirement	13,489.17	+9.17
CV Disability-Retirement	80.06	−19.94
CV Death	40.03	−9.97
Non-CV Disability-Retirement	1,501.02	+1.02
Non-CV Death	500.35	+0.35
	62,642.59	

Using the departure changes shown in Table 3(b) and the departure costs of Table 1, the annual change in departure costs of enlisted personnel due to the HEART program is a decrease of $2,822,656 per year. This model realistically yields a larger increase in retirements than shown by the expected value equilibrium model (Figure 2), thereby accounting for most of the reduction in savings (vs. $3,820,638).

Annual reductions for other classes of personnel, and for other assumptions of effectiveness of the HEART program, are generated in a similar way.

ALLOCATIONS OF PARAMEDICS

One of the primary *increased* costs of the HEART program is that of additional personnel needed to operate the program.* A problem arises in the efficient allocation of numbers of paramedics to the various USAF bases while recognizing that the bases are of different sizes.

Knowing the number of military personnel of each base, and assuming a risk threshold that is consistently used at all bases and that will place an identical fraction of each base's population under treatment, we can define the following:

X = the specified fraction of base population that is to be treated in the therapy group.

P_i = the number of paramedics required at the i th base.
B_i = the population of the i th base.

Knowing the details of the proposed treatment and screening tasks, and that there is one nurse available part-time at each base, we determined x_i, the capacity, or maximum fraction of the base population that can be treated, as a function of base size and number of paramedics allocated. This involved careful analysis of the time required for each task as well as consideration of allowances for rest breaks and vacations. The resulting capacity for the i th base is determined as:

$$\text{Hours for Screening} + \text{Hours for Therapy} + \text{Hours for Group Sessions} = \text{Available Hours of Nurse} + \text{Paramedics}$$

$$.19803(B_i) + 7.875(B_i)(x_i) + 1050 = 1800 + 1800(P)$$

$$\text{or} \qquad x_i = \max\left[\frac{1800(P_i) - .19803(B_i) + 750}{7.875(B_i)}, 0\right].$$

Nonnegativity must be enforced explicitly. Some of the effort of the nurse and paramedics involves screening of all base personnel and it would be possible to obtain a negative value for x_i (the fraction that can be treated) if the screening effort exceeded the available manpower.

The objective is to determine the minimal set of P_i such that $x_i \geq X$ for all i. We will first formulate a simple mathematical programming approach for determining P_i and then use it to justify an even simpler computerized allocation scheme.

Mathematical Programming Model

All USAF bases may be grouped into 26 size ranges, and we will define N_i as the number of bases of the i th size, $i = 1, 2, \ldots, 26$. We wish to determine P_i, the number of paramedics to assign to all bases of size i, for any value of X that is chosen. The integer linear program is:

$$\text{Min} \sum_{i=1}^{26} N_i P_i \quad \text{subject to twenty-six}$$

constraints, one for each base size, of the form

$$P_i \geq \frac{7.875(B_i)(X) + .19803(B_i) - 750}{1800}$$

where the variables P_i are nonnegative integers.

Each constraint is a function of only one variable, P_i, since X is fixed and B_i is known. Solution of such a program to determine each P_i and to minimize the total required number of paramedics would be possible but very time consuming. As an alternate method, note that each constraint may be satisfied by merely fixing P_i as the smallest feasible nonnegative integer. This will obviously minimize the objective function because minimizing each term of a sum minimizes the sum.

An optimal assignment will not necessarily produce full utilization of all paramedics, but there will be no assignment using fewer paramedics which will permit treatment of the stated fraction (X) of the population.

Simple Allocation Algorithm

The allocation algorithm starts with a specified value of X, and considers only the nurse assigned to each base. The maximum possible therapy group size, x_i, with full utilization of this allocation is then calculated and updated for each base. If $x_i \geq X$ for all i, this allocation is optimal for the stated X. Otherwise the base (or bases) with the smallest fraction of personnel in therapy (x_i) is then "given" one paramedic, and the calculations are updated. This procedure is continued, assigning additional paramedics, until the desired therapy group fraction is attained for all bases. In summary, the procedure initializes the P_i vector at 0 and determines X, which, because of a uniform threshold policy at all bases, will be the smallest fraction among all bases. Then the P_i vector is increased in the most efficient manner until the specified value of X is attained.

The procedure is easily continued to obtain solutions for an entire range of X values. A typical set of solutions shows the total number of paramedics required to range from 254 for a 7 percent therapy group to 567 for a 19 percent group. For the 7 percent therapy group, the individual base requirements range from 0 to 5 paramedics and for the 19 percent group, from 0 to 11. Overall utilization for the two cases is .75 and .88, respectively.

TOTAL COST EFFECTIVENESS

The total cost effectiveness was expressed as a net annual savings and was a function of the risk threshold selected (which, in turn, governed how many people would be treated) and the assumptions made regarding the effectiveness of therapy.

Net annual savings = Departure cost reduction + lost aircraft cost reduction + cost reduction in CV nondepartures* − paramedic costs − operating, drug, and test costs.

It has not been the intent of this paper to present the results of the cost effectiveness analysis but only to describe several manpower planning models used in performing it. The models permit estimation of changes of some of the complex cost elements. Computer experimentation was then possible to aid in certain decisions such as determination of therapy group size and treatment intensity [3].

REFERENCES

[1] Grinold, R.C. and K.T. Marshall, *Manpower Planning Models*, (Elsevier North-Holland, Inc., 1977).

[2] Kannel, W.B., D. McGee, and T. Gordon, "A General Cardiovascular Risk Profile: The Framingham Study," The American Journal of Cardiology, (July 1976).

[3] Petersen, C.C., A. Ravindran, A. Sweet, and L. Cote, "Analyses in Support of the HEART Program," Project Report, School of Industrial Engineering, Purdue University, West Lafayette, Indiana (September 25, 1978).

*Hospitalization and noneffectiveness costs of the personnel suffering mild CV disease which does not result in disability-retirement or death will also be reduced by an effective HEART program

AN EMPIRICAL EVALUATION OF FURTHER APPROXIMATIONS TO AN APPROXIMATE CONTINUOUS REVIEW INVENTORY MODEL

Donald L. Byrkett

Systems Analysis Department
Miami University
Oxford, Ohio

ABSTRACT

This paper describes an empirical evaluation of several approximations to Hadley and Whitin's approximate continuous review inventory model with backorders. It is assumed that lead time demand is normally distributed and various exponential functions are used to approximate the upper tail of this distribution. These approximations offer two important advantages in computing reorder points and reorder quantities. One advantage is that normal tables are no longer required to obtain solutions, and a second advantage is that solutions may be obtained directly rather than iteratively. These approximations are evaluated on two distinct inventory systems. It is shown that an increase in average annual cost of less that 1% is expected as a result of using these approximations. The only exception to this statement is with inventory systems in which a high shortage cost is specified and ordering costs are unusually low.

INTRODUCTION

This paper is concerned with Hadley and Whitin's [4] approximate continuous review inventory model in which a fixed quantity of an individual item is ordered each item the inventory position (units on hand plus units on order minus backorders) reaches the reorder point. After a lead time has elapsed, the entire order is received. It is assumed that reorder quantities and reorder points are established independently for each item and that the distribution of lead time demand can be approximated by a normal distribution.

The optimal reorder point and reorder quantity for this model are determined by minimizing a cost function including the expected number of orders placed per unit time, the expected number of backorders per unit time, and an approximation to the expected holding cost per unit time. The solution which minimizes this approximate cost function is found by an iterative algorithm that converges quite rapidly.

To find the optimal solution, it is necessary to calculate the expected number of backorders per period for a given policy. If lead time demand is assumed to be normally distributed, then this requires the evaluation of the standardized normal loss integral. Several authors [5,7,8] have developed exponential functions to approximate the expected number of backorders per period. This not only alleviates the iterative solution but saves the table look-up required to evaluate the normal loss integral. This paper evaluates these approximations.

MODEL DERIVATION

The following notation, from Hadley and Whitin [4] is used.

Q	=	order quantity (units)
r	=	reorder point (units)
λ	=	demand rate (units/year)
A	=	ordering cost ($/order)
C	=	cost of item ($/unit)
I	=	carrying cost ($/$ value of stock/year)
π	=	backordering cost ($/unit backordered)
α	=	probability of a stockout occurring during a lead time
β	=	probability that any unit demanded cannot be filled from stock
$h(t)$	=	probability density function of lead time demand
μ	=	mean lead time demand (units)
σ	=	standard deviation of lead time demand
$H(r)$	=	probability lead time demand exceeds the reorder point (complementary cumulative distribution)
$\bar{n}(r)$	=	expected number of backorders during a lead time when the reorder point is r (units)

The expected anual cost (AC) of operating the inventory system is represented in the equation below, assuming that shortages are backordered.

$$AC = \frac{A\lambda}{Q} + IC\left(\frac{Q}{2} + r - \mu\right) + \frac{\lambda\pi}{Q}\,\bar{n}(r). \tag{1}$$

The first term represents the expected ordering cost, the second term the expected carrying cost, and the third term the expected number of backorders. It is the second term in this equation that is an approximation, since the average inventory level is estimated as though there are no backorders. If the expected number of backorders is small, the approximation is very good (see Gross and Ince [3]). The third term may also be considered an approximation since the lead time demand is approximated by the normal distribution.

The values of Q and r that minimize the above annual cost function can be found by the simultaneous solution of the two equations below.

$$Q = \left\{\frac{2\lambda}{IC}\,(A + \pi\bar{n}(r))\right\}^{1/2}. \tag{2}$$

$$H(r) = \frac{QIC}{\pi\lambda}. \tag{3}$$

An iterative solution is suggested by Hadley and Whitin [4] that will work as long as $\frac{QIC}{\pi\lambda} < 1$. This is fine, since as $H(r)$ approaches 1, the approximation to carrying cost becomes rather poor and, thus, the model is not appropriate.

In practice, it is often difficult to estimate the backordering cost, π. To avoid this problem, one may instead specify a desired service level. One approach (α service policy) is to specify α, the probability of a stockout occurring during a lead time. A second approach (β service policy) is to specify β, the probability that any unit demanded cannot be filled from stock. Since π is not specified, it can be eliminated from (2) and (3) above (see Nahmias [6]), yielding the following:

$$Q = \frac{\bar{n}(r)}{H(r)} + \left\{\left[\frac{\bar{n}(r)}{H(r)}\right]^2 + \frac{2\lambda A}{IC}\right\}^{1/2}. \tag{4}$$

In addition, the α service policy requires

$$H(r) = \alpha. \tag{5}$$

And, the β service policy requires

$$\frac{n(r)}{Q} = \beta. \tag{6}$$

Values of Q and r may be found directly from Equations (4) and (5) for the α service policy and values of Q and r may be found iteratively from Equations (4) and (6) for the β service policy. See Nahmias [6] for an appropriate algorithm for finding optimal values of Q and r.

FURTHER APPROXIMATIONS

When applying the above model to an inventory system with many parts, it is typically assumed as an approximation that the lead time demand follows a particular distributional form for all parts. A very convenient approximation and the one assumed in this paper is the normal distribution. That is, it is assumed that $h(t)$ is the probability density function of the normal distribution with the mean u and standard deviation σ. Therefore, $H(r)$ and $\bar{n}(r)$ may be calculated from the equations below, where $Z(t)$ is the probability density function of the unit normal.

$$H(r) = \int_{\frac{r-u}{\sigma}}^{\infty} Z(t)\, dt. \tag{7}$$

$$\bar{n}(r) = \sigma \int_{\frac{r-u}{\sigma}}^{\infty} \left(t - \frac{r-u}{\sigma}\right) Z(t)\, dt. \tag{8}$$

The integral in (7) is the complementary cumulative distribution of the unit normal and is tabulated in any standard statistics book. The integral in (8) is referred to as the standardized normal loss integral and is tabulated in Brown [2]. The tabulated integrals in (7) and (8) are required to solve Equations (2) and (3), Equations (4) and (5), and Equations (4) and (6).

Two approximations have been suggested to avoid the table look-up required by (7) and (8). One approximation, suggested by Schroeder [8] and Herron [5], is to use an exponential function, of the form ae^{-bt}, to approximate the integral in Equation (7). Using this approximation,

$$H(r) = ae^{-b\left(\frac{r-u}{\sigma}\right)} \text{ and } \bar{n}(r) = \frac{\sigma a}{b} e^{-b\left(\frac{r-u}{\sigma}\right)}.$$

The exponential approximation not only avoids the table look-up required to calculate $H(r)$ and $\bar{n}(r)$ but also avoids the iterative solution procedures required to find optimal values of Q and r. If the expression for $\bar{n}(r)$ is substituted into the annual cost Equation (1) and partial derivatives are set equal to zero, the optimal value of Q is as follows regardless of whether π, α, or β is specified.

$$Q = \frac{\sigma}{b} + \left\{\left[\frac{\sigma}{b}\right]^2 + \frac{2A\lambda}{IC}\right\}^{1/2}. \tag{9}$$

The optimal value of r is presented in Equation (10a), (10b), and (10c) for π specified, α specified, and β specified, respectively.

(10a) $$r = u - \frac{\sigma}{b} \ln\left[\frac{QIC}{\pi\lambda a}\right].$$

(10b) $$r = u - \frac{\sigma}{b} \ln(\alpha/a).$$

(10c) $$r = u - \frac{\sigma}{b} \ln\left[\frac{Q\beta b}{\sigma a}\right].$$

Note that the optimal values of Q and r do not require an iterative solution. These values represent approximate solutions when $h(t)$ is assumed to be the normal probability density function. They are approximate, since the complementary cumulative distribution function of the unit normal is approximated using an exponential function.

A second approximation, suggested by Herron [5] and Parker [7] is to use an exponential function of the same form to approximate the integral in Equation (8). Using this approximation, $\bar{n}(r) = \sigma a e^{b\frac{(r-\mu)}{\sigma}}$ and $H(r) = abe^{-b\frac{(r-\mu)}{\sigma}}$.

Likewise, this approximation avoids the table look-up required to find $H(r)$ and $\bar{n}(r)$ and avoids the iterative solution procedure required to find optimal values of Q and r. The optimal value of Q is the same as that specified in Equation (9) and the optimal value of r is presented in Equations (11a), (11b), and (11c) for π specified, α specified, and β specified respectively.

(11a) $$r = \mu - \frac{\sigma}{b} \ln\left[\frac{QIC}{\pi\lambda ab}\right].$$

(11b) $$r = \mu - \frac{\sigma}{b} \ln(\alpha/ab).$$

(11c) $$r = \mu - \frac{\sigma}{b} \ln\left[\frac{Q\beta}{\sigma a}\right].$$

Thus, both approximations allow optimal values of Q and r to be calculated directly and avoid the problems of looking-up data in the normal tables. The purpose of this paper is to evaluate the accuracy of these approximations.

PARAMETER ESTIMATION

Figure 1 contains a plot of the log of the complementary cumulative unit normal distribution and the log of the standardized normal loss integral versus K, the number of standard deviations above zero. For the exponential functions to be a good fit, these plots should be straight lines. Obviously, there is a rather slow gradual curvature to both lines but a straight line does not appear to be a bad approximation.

Table 1 contains the parameter estimates obtained by the various authors and the range of K that was used to obtain these estimates. Herron used two straight lines to obtain a better fit of the curvature.

This author developed his own parameter estimates for the standardized normal loss integral by fitting a least square regression line to twenty-one points in the range $1.0 \leqslant K \leqslant 3.0$. This was done in order to evaluate a method analogous to that used by Schroeder [8] for the second type of approximation.

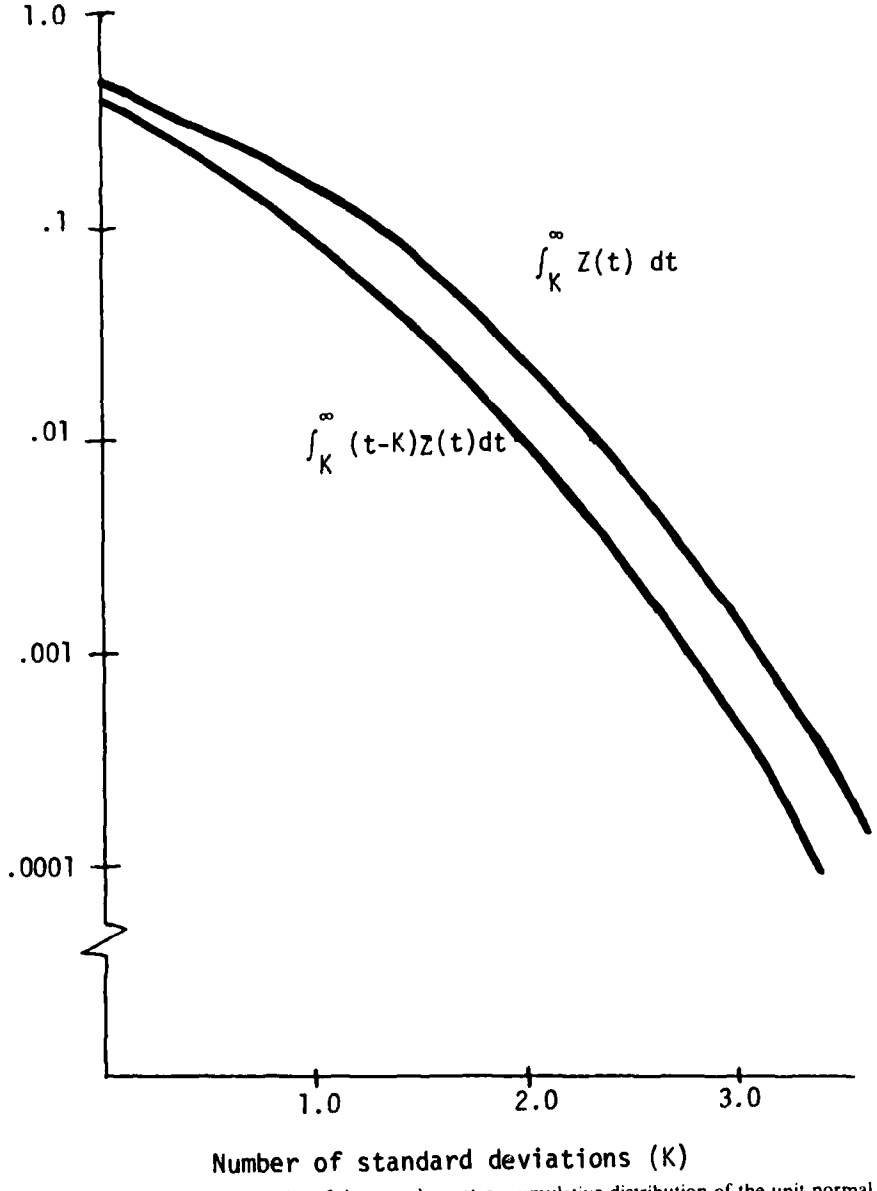


FIGURE 1. Semilogarithmic plot of the complementary cumulative distribution of the unit normal and the standardized normal loss integral versus K standard deviations above zero.

TABLE 1. *Parameter Estimates*

Author [reference]	Estimates		Range of fit
	a	b	
A. Approximations to complementary cumulative distribution of unit normal			
1. Schroeder [8]	2.8800	2.4900	$1.0 \leqslant K \leqslant 3.0$
2. Herron [5]	.5500	.7530	$0 \leqslant K \leqslant 1.5$
	3.6000	.3870	$1.5 \leqslant K \leqslant 3.0$
B. Approximation to standardized normal loss integral			
1. Byrkett	1.5792	2.6879	$1.0 \leqslant K \leqslant 3.0$
2. Herron [5]	.4400	.5760	$0 \leqslant K \leqslant 1.5$
	2.4900	.3460	$1.5 \leqslant K \leqslant 3.0$
3. Parker [7]	.4500	1.6949	$0 \leqslant K \leqslant 1.4$

The reader should notice that Parker [7] developed his approximation in the range $0 \leqslant K \leqslant 1.4$. It is felt that for most inventory systems, including those evaluated in this paper, it is preferable to use an approximation of the upper tail of the distribution, for example $1.0 \leqslant K \leqslant 3.0$. For this reason, Parker's approximation will not be given further evaluation.

COMPARISON WITH TABLED VALUES

One approach to measuring the accuracy of the approximates outlined in the previous section is to compare the values obtained using the approximations with the corresponding tabled values. Table 2 displays these results. The approximations to the complementary cumulative distribution were used to compute tabled values of the complementary cumulative distribution at intervals of .05 in the range $1.0 \leqslant K \leqslant 3.0$. Likewise, the approximations to the standardized normal loss integral were used to compute tabled values of the standardized normal loss integral. Three criteria are used to compare the approximations to the tabled values; the mean absolute deviation, the mean squared deviation, and the mean percentage deviation.

TABLE 2. *Comparison of Approximations to Tabled Values**

Author	Absolute deviation		Squared deviation		Percentage deviation	
	Mean	Max.	Mean	Max.	Mean	Max.
A. Approximations to complementary cumulative distribution of unit normal						
1. Schroeder [8]	.0090	.0801	.000405	.006420	13	50
2. Herron [5]	.0025	.0129	.000015	.000166	7	115
B. Approximations to standaridized normal loss integral						
1. Byrkett	.0025	.0241	.000032	.000581	11	30
2. Herron [5]	.0010	.0058	.000003	.000033	7	12

*Entries in table calculated by comparing approximate value with tabled value at intervals of .05 between $K = 1.0$ and $K = 3.0$.

Two observations may be made from this table. First, the approximations of the standardized normal loss integral are closer to the tabled values than those of the complementary cumulative distribution of the unit normal according to all criteria. This indicates that the lower curve in Figure 1 is closer to linear than the upper curve and that approximating this curve may produce a smaller error. Second, Herron's two line approximation is preferable to Schroeder's and Byrkett's one line approximation, according to all criteria.

Though these results tend to favor the estimates developed by Herron [5], they are by no means conclusive with respect to their economic effects in controlling inventory.

COMPARISON OF OPERATING POLICIES

The major concern in using these approximations is how much influence they will have on the cost of operating an inventory system. It is possible to simply use an iterative algorithm to find optimal values of Q and r and to look-up values of $H(r)$ and $\bar{n}(r)$ from normal tables. This, however, requires significantly more computer time than the one or two line approximations discussed above. For example, the CPU time required to execute the iterative algorithms and to look-up values of $H(r)$ and $\bar{n}(r)$ from normal tables for all cases discussed below was 113.44 seconds. This compares with 2.57 seconds for the one line approximations and 4.42 seconds for the two line approximations. If you consider an inventory system with many thousand items and frequent updating, the savings in computer time can be substantial.

With the exponential approximations, it is possible to calculate the optimal reorder points and reorder quantities directly without any table look-ups. To determine how much this computational advantage costs, the average annual cost of the solution of Equations (2) and (3), Equations (4) and (5), and Equations (4) and (6) are compared with the annual cost of the approximations given by Equations (9) and (10a,b,c) and Equations (9) and (11a,b,c). Equation (1), using a table look-up, will be used to compare the resulting operating policies.

It is not valid to compare the approximations on a single item from the inventory, since all approximations may do equally well on a given item. Rather, the approximations must be compared on the entire inventory, or on at least a representative cross section of the entire inventory. In this study, cross sections of two different inventories are used to compare the approximations. One inventory, called the Maintenance Inventory, contains equipment and parts for maintaining a large fleet of maintenance vehicles, including cars, trucks, graders, and so forth. Forty items were selected from this inventory and estimates made of λ, C, μ, and σ. A second inventory, called the Warmdot Inventory, contains spare parts for heating and air conditioning equipment. Brown [1] contains λ and C for sixteen items from this inventory. Estimates were made of μ and σ by assuming a three month lead time and using the derived relationship in Brown [1], $\sigma = .21\mu^{.8} C^{.2}$.

In addition to comparing the approximations on two different inventories, all three methods of specifying shortages costs (π, α, and β) are considered. The shortage costs, ordering costs, and carrying costs are run at three levels each to determine the impact of these costs. The costs used are as follows:

Ordering cost	(A)	— \$1, \$10, \$100
Carrying cost	(I)	— .1, .2, .3
Shortage cost		— Low, Medium, High
	π	— \$10, \$100, \$1000
	α	— .15, .10, .05
	β	— .10, .01, .001.

Table 3 summarizes the results of using the single line approximation of Schroeder [8] and the two line approximation of Herron [5] to the complementary cumulative distribution of the unit normal. The numbers reported in these tables are percentage increases in annual costs for all sample parts in the Maintenance Inventory for the various cases. Similarly, Table 4 summarizes the results of using the single line approximation by Byrkett and the two line approximation by Herron [5] to the standardized normal loss integral. Again, these results are for the Maintenance Inventory. Similar results are developed for the Warmdot Inventory though these have not been included in order to conserve space.

TABLE 3. *Percentage Increase in Annual Costs Using Approximations to the Complementary Cumulative Distribution of the Unit Normal (Maintenance Inventory)*

A	I	Approximations*	Shortage Cost Specified								
			π			α			β		
			10	100	1000	.15	.10	.05	.10	.01	.001
1	.1	1	**	1.0	4.6	.6	.3	.4	1.3	.4	.7
		2	**	1.4	3.6	2.0	3.4	.0	1.3	.4	.5
	.2	1	**	1.1	3.5	.6	.3	.4	1.3	.4	.8
		2	**	.9	2.9	2.3	2.8	.0	1.5	.4	.5
	.3	1	**	1.6	3.1	.6	.3	.4	1.3	.4	.9
		2	**	.5	2.6	2.4	4.1	.0	1.6	.4	.6
10	.1	1	**	.4	2.5	.4	.1	.3	1.3	.0	.3
		2	**	.6	1.9	.8	1.5	.0	.4	.1	.2
	.2	1	**	.7	2.2	.5	.2	.4	1.2	.0	.4
		2	**	.4	1.6	1.0	1.8	.0	.5	.1	.2
	.3	1	**	1.3	1.8	.4	.1	.3	1.1	.1	.7
		2	**	.1	1.2	1.4	2.4	.0	.8	.0	.4
100	.1	1	**	.2	.6	.2	.0	.2	1.4	.0	.0
		2	**	.1	.4	.2	.3	.0	.1	.1	.0
	.2	1	**	.7	.6	.3	.1	.2	1.6	.0	.1
		2	**	.1	.3	.3	.4	.0	.2	.2	.0
	.3	1	**	1.7	.5	.3	.1	.3	1.5	.0	.2
		2	**	.2	.3	.4	.7	.0	.2	.1	.1

*1 — Schroeder [8], 2 — Herron [5]

** Indicates that the assumption $\frac{QIC}{\pi\lambda} < 1$ was violated for one or more items in the inventory.

It is difficult to draw conclusions from the raw data provided in Tables 3 and 4. Thus, Table 5 is developed which summarizes the results in Tables 3 and 4 and the corresponding results for the Warmdot Inventory by averaging these percentage increases across all tables by group. A regression model was developed using the percentage increase in annual costs as the dependent variable. The independent variables were all 0-1 variables used to represent the seven groups listed in Table 5 and all of the two factor interactions. One by one the independent variables used to represent the seven factors were removed from the model to test the

statistical relationship of the given variable with the percentage increase in cost. An F test was used with significance level set of 99%. This will insure that the family of seven tests has a joint significance level of at least 93%. Based on these seven tests, the variables found to be significant are the inventory system under study, the level of the ordering cost, and the level of the shortage cost. The approximation approach and the method of approximation were nearly significant but not at the confidence level specified.

TABLE 4. *Percentage Increase in Annual Costs Using Approximations to the Standardized Normal Loss Integral (Maintenance Inventory)*

A	I	Approximation*	Shortage Cost Specified								
			π			α			β		
			10	100	1000	.15	.10	.05	.10	.01	.001
1	.1	1	**	.4	2.9	.2	.5	.1	.7	.5	.2
		2	**	.4	2.4	.8	.4	.2	.1	.4	.2
	.2	1	**	.9	2.2	.8	.5	.1	1.7	.5	.3
		2	**	.2	1.2	.1	.4	.2	.1	.1	.2
	.3	1	**	1.8	1.7	.8	.5	.1	1.7	.5	.3
		2	**	.2	.6	.1	.4	.2	.1	.1	.2
10	.1	1	**	.2	1.2	.6	.2	.0	1.8	.0	.1
		2	**	.3	.4	.1	.3	.1	.3	.1	.1
	.2	1	**	.8	.8	.8	.4	.1	1.8	.1	.2
		2	**	.1	.4	.2	.4	.1	.1	.1	.2
	.3	1	**	1.4	.7	.7	.2	.0	1.8	.1	.3
		2	**	.1	.3	.2	.3	.3	.1	.2	.4
100	.1	1	**	.4	.2	.4	.1	.0	2.0	.1	.2
		2	**	.1	.1	.1	.2	.1	.2	.0	.3
	.2	1	**	1.2	.1	.4	.2	.0	2.1	.1	.0
		2	**	.2	.1	.1	.3	.1	.3	.0	.0
	.3	1	**	2.0	.1	.5	.2	.0	2.0	.1	.1
		2	**	.4	.0	.1	.3	.1	.2	.0	.1

*1 — Byrkett, 2 — Herron [5]

** Indicates that the assumption $\frac{QIC}{\pi\lambda} < 1$ was violated for one or more items in the inventory.

Several observations may be made from Table 5. First, the average percentage increase over all groups studied is only .71%. This indicates that the approximations are quite effective. However, maximum percentage increase of 12.9% indicates that under some conditions the approximations are not so effective. Second, the approximations are much more effective for service level type policies (α and β) than for shortage cost type policies (π). Third, the approximations are more effective for inventory systems in which the ordering costs are relatively high ($100) than for inventory systems with relatively low ordering cost $1 and $10). Fourth, the approximations are more effective for inventory systems in which the shortage costs are relatively low than those with high shortage costs. Fifth, though the differences are not great, it appears that the two line approximations to the standardized normal loss integral produces the best results.

It was noted in the previous paragraph that the inventory system under study was found to be a significant factor. Though the difference in mean percentage increase is not great, there is significant interaction between the inventory system and the specification of shortage cost.

the level of ordering cost, and the level of shortage cost. The interaction between the inventory system and the method of specifying shortage cost is illustrated in Table 6A. Notice that the approximations are more effective for the maintenance inventory system when π is specified and vice versa when a service level is specified (α and β).

TABLE 5. *Percentage Increase in Annual Costs by Group*

Group	Mean	Standard deviation	Maximum
Over all groups	.71	1.24	12.9
Approximation approach			
$H(r)$	.82	1.51	10.7
$\bar{n}(r)$	.59	1.37	12.9
Inventory system*			
Maintenance inventory	.62	.80	4.6
Warmdot inventory	.79	1.84	12.9
Specification of shortage cost*			
π	1.55	2.30	12.9
α	.30	.57	4.1
β	.41	.59	2.3
Approximation			
Single line ($1 \leqslant K \leqslant 3$)	.81	1.45	10.7
Two lines ($0 \leqslant K \leqslant 3$)	.60	1.43	12.9
Ordering cost*			
1	1.26	2.11	12.9
10	.63	1.07	5.5
100	.24	.41	2.1
Carrying cost			
.1	.66	1.23	7.1
.2	.69	1.35	8.4
.3	.77	1.71	12.9
Shortage cost*			
Low	.57	.65	2.4
Medium	.50	.90	5.3
High	1.03	2.14	12.9

*Indicates this variable is significant using F-test with $\alpha = .01$

Other selected interactions are also illustrated in Table 6. Table 6B indicates that the approximations are least effective for inventory systems with low ordering costs in which a shortage cost is specified. Table 6C indicates that the approximations are also least effective for inventory systems with a high specified shortage cost. Table 6D combines the results of Tables and 6B and 6C and indicates that the approximations are least effective for approximations with a combination of a low ordering cost and a high shortage cost.

TABLE 6. *Percentage Increase in Annual Costs for Selected Two Factor Interactions*

A. Specification of shortage cost versus inventory system

Inventory system	Specifications of shortage cost		
	π	α	β
Maintenance	1.02	.48	.49
Warmdot	1.90	.11	.34

B. Specification of shortage cost versus ordering cost

Ordering cost	Specification of shortage cost		
	π	α	β
1	3.01	.51	.54
10	1.31	.27	.41
100	.33	.11	.28

C. Specification of shortage cost versus shortage cost

Shortage cost	Specification of shortage cost		
	π	α	β
Low	.16	.41	.92
Medium	.93	.42	.15
High	2.86	.07	.17

D. Ordering cost versus shortage cost

Shortage cost	Ordering cost		
	1	10	100
Low	.78	.56	.36
Medium	.98	.34	.18
High	1.93	.98	.19

SUMMARY AND CONCLUSIONS

Hadley and Whitin's [4] approximate continuous review inventory model has received frequent analysis in the literature [recently, 3 and 6], though little has been reported of actual use. Perhaps the reason for this apparent lack of use is the requirement that a probability distribution be specified for lead time demand, and the requirement that a backordering cost or service level be specified. Moreover, even if one is willing to specify the normal distribution for lead time demand and an appropriate backordering cost or service level, one still may be hesitant about using the iterative solution algorithm that requires the use of normal tables. The purpose of this paper is to evaluate some approximations that relieve the latter two deterrents to using this model.

This evaluation produced the following results:

1. The exponential approximations result in operating policies very near those of iterative algorithms. The average increase in annual costs as a result of using these approximations is .71%, depending on certain characteristics of the inventory system.

2. It is preferable to approximate the standardized normal loss integral with an exponential function than to approximate the complementary cumulative distribution function.

3. The approximations are closer for α or β specified policies, than for π specified policies.

4. A single line approximation in the range $K = 1.0$ to $K = 3.0$ is nearly as effective as a two line approximation.

5. The approximations are least effective for inventory systems with low ordering costs and high specified shortage costs.

REFERENCES

[1] Brown, R.G., "Estimating Aggregate Inventory Standards," Naval Research Logistics Quarterly *10*, 55-71 (1963).

[2] Brown, R.G., *Decision Rules for Inventory Management*, (Holt, Rinehart, and Winston, New York, N.Y., (1967).

[3] Gross, D. and J.F. Ince, "A Comparison and Evaluation of Approximate Continuous Review Inventory Models," International Journal of Production Research, *13*, 9-23 (1975).

[4] Hadley, G. and T.M. Whitin, *Analysis of Inventory Systems*, (Prentice-Hall, Englewood Cliffs, New Jersey 1963).

[5] Herron, D.P., "Inventory Management for Minimum Cost," Management Science *14*, B219-B235 (1967).

[6] Nahmias, S., "On the Equivalence of Three Approximate Continuous Review Inventory Models," Naval Research Logistics Quarterly, *23*, 31-36 (1976).

[7] Parker, L.L., "Economical Reorder Quantities and Reorder Points With Uncertain Demand," Naval Research Logistics Quarterly, *11*, 351-358 (1964).

[8] Schroeder, R.G., "Managerial Inventory Formulations With Stockout Objective and Fiscal Constraints," Naval Research Logistics Quarterly, *21*, 375-388 (1974).

A NOTE ON THE MIXTURE OF NEW WORSE THAN USED IN EXPECTATION

Kishan G. Mehrotra

Syracuse University,
Syracuse, New York

1. INTRODUCTION

The class of distributions which are new worse than used in expectation (NWUE) was first introduced by Marshall and Proschan [2]. These classes play an important role in the theory of reliability and in particular arise quite naturally in considering replacement policies. A nonnegative distribution F with survival function $\bar{F}$ and expected value μ is said to be NWUE if

$$\mu \bar{F}(t) \leqslant \int_t^\infty \bar{F}(x)\, dx \quad \text{for all } t \geqslant 0.$$

In this note we are interested in the following question: Is the class of NWUE preserved under arbitrary mixture? Barlow and Proschan [1] conjectured that NWUE is not preserved under arbitrary mixtures. In section 1 of this note we present examples which verify this conjecture and in Section 2 we give some other elementary properties of distribution of this class.

2. NWUE IS NOT PRESERVED

The examples considered below are obtained in view of Lemmas 1 and 2 of the next section. That is, we take two specific NWUE distributions $\bar{F}_1$ and $\bar{F}_2$ with respective expectations μ_1 and μ_2 such that $\mu_1 > \mu_2$ and $\bar{F}_1$ crosses $\bar{F}_2$ from above. Then, at a point t beyond the point of intersection of $\bar{F}_1$ and $\bar{F}_2$ the defining equation of NWUE is not satisfied.

Consider the mixture

$$\bar{F}(x) = \frac{1}{2}\,(\bar{F}_1(x) + \bar{F}_2(x)) \quad \text{for all } x \geqslant 0$$

where

$$\bar{F}_1(x) = e^{-(x/0.8)} \quad \text{for } x \geqslant 0$$

and

$$\bar{F}_2(x) = e^{-\delta k} \quad \text{for } (k-1)\delta \leqslant x < k\delta, \quad k = 1, 2, \ldots .$$

Thus, F_1 is the exponential distribution function with expected value $\mu_1 = .8$ and clearly NWUE. $F_2(x)$ is a slight modification of distributions considered by Barlow and Proschan [1] in Section 5.9 of Chapter 6. Since $\bar{F}_2(x)$ is easily seen to be a NWU, by (2.4) of Chapter 6 of Barlow and Proschan [1], it is clearly NWUE. The expected value of the random variable with distribution function F_2 is

$$\mu_2(\delta) = \frac{\delta e^{-\delta}}{1-e^{-\delta}}$$

and

$$\int_t^\infty \bar{F}_2(x)\,dx = (k\delta - t)\, e^{-k\delta} + \frac{\delta e^{-(k+1)\delta}}{1-e^{-\delta}}$$

where k is an integer such that $(k-1)\delta \leqslant t < k\delta$. For a given δ set

$$L(\delta,t) = \frac{1}{2}(\mu_1 + \mu_2(\delta))\frac{1}{2}(\bar{F}_1(t) + \bar{F}_2(t))$$

and

$$R(\delta,t) = \frac{1}{2}\int_t^\infty \{\bar{F}_1(x) + \bar{F}_2(x)\}\,dx.$$

Then, for $\delta = .5$, $\mu_2(.5) = .77074$. For $t = .5-\epsilon$, where ϵ is positive and very small, for instance, $\epsilon = .001$, $L(.5, .5-\epsilon) \doteq .44836$ and $R(.5, .5-\epsilon) \doteq .44784$. Clearly, $L(.5, .5-\epsilon) > R(.5, .5-\epsilon)$. Thus, the mixture $\{\frac{1}{2}\bar{F}_1(x) + \bar{F}_2(x)\}$ is not NWUE. This inequality holds for values of t slightly less than 1 and 2.

The above examples clearly show that NWUE is not preserved under the mixture, as conjectured by Barlow and Proschan [1].

3. SOME ADDITIONAL PROPERTIES OF NWUE DISTRIBUTION

LEMMA 1: Let F be the class of NWUE distributions with equal mean μ. Then any arbitrary mixture of $F_\alpha \in F$ is NWUE.

PROOF: Let $F = \int F_\alpha\, dG(\alpha)$ for arbitrary distribution function G.

Then,

$$\mu_f = \int_0^\infty \bar{F}(x)dx = \int_0^\infty \int \bar{F}_\alpha(x)dG(\alpha)dx = \int [\int_0^\infty \bar{F}_\alpha(x)dx]dG(\alpha)$$

$$= \int \mu\, dG(\alpha) = \mu$$

where the second inequality holds by Fubini's Theorem. Next, for arbitrary $t \geqslant 0$,

$$\int_t^\infty \bar{F}(x)dx = \int_t^\infty [\int \bar{F}_\alpha(x)dG(\alpha)]\,dx = \int [\int_t^\infty \bar{F}(x)dx]dG(\alpha)$$

$$\geqslant \int \mu \bar{F}_\alpha(t)\,dG(\alpha) = \mu\bar{F}(t).$$

Thus, F is NWUE.

LEMMA 2: Let F_1 and F_2 be two NWUE distribution functions such that $\bar{F}_1$ crosses $\bar{F}_2$ once from below. Let $\mu_1 > \mu_2$ where μ_i is the mean associated with F_i. Then for any $p, 0 \leqslant p \leqslant 1, \bar{F}(x) = p\bar{F}_1(x) + q\bar{F}_2(x)$ is NWUE: $q = 1-p$.

PROOF:

$$\left\{\int_t^\infty \bar{F}(x)dx - \mu\bar{F}(t)\right\} = (p+q)\int_t^\infty \{p\bar{F}_1(x) + q\bar{F}_2(x)\}dx - (p\mu_1 + q\mu_2)(p\bar{F}_1(t) + q\bar{F}_2(x))$$

$$= p^2\left[\int_t^\infty \bar{F}_1(x)dx - \mu_1\bar{F}_1(t)\right] + q^2\left[\int_t^\infty \bar{F}_2(x)dx - \mu_2\bar{F}_2(t)\right]$$

$$+ pq\left[\int_t^\infty \{\bar{F}_1(x) + \bar{F}_2(x)\}dx - \mu_1\bar{F}_2(t) - \mu_2\bar{F}_1(t)\right].$$

To show that this expression is positive for all t, it is sufficient to show that the third term is positive, because the first two terms are positive by assumption F_i is NWUE, $i = 1, 2$,.

Let t_0 be the point where $\bar{F}_1$ crosses $\bar{F}_2$ from below. Then for $t > t_0, \bar{F}_1(t) > \bar{F}_2(t)$. Thus,

$$\int_t^\infty \{\bar{F}_1(x) + \bar{F}_2(x)\}dx - \mu_1\bar{F}_2(t) - \mu_2\bar{F}_1(t) \geq (\mu_1 - \mu_2)(\bar{F}_1(t) - \bar{F}_2(t)) \geq 0.$$

For $t < t_0$.

$$\int_t^\infty \{\bar{F}_1(x) + \bar{F}_2(x)\}dx = \{\mu_1\bar{F}_2(t) + \mu_2\bar{F}_1(t)\} \geq (\mu_1 - \mu_2)\int_t^\infty \left\{\frac{\bar{F}_1(x)}{\mu_1} - \frac{\bar{F}_2(x)}{\mu_2}\right\}dx$$

$$= (\mu_1 - \mu_2)\int_0^t \left\{\frac{\bar{F}_2(x)}{\mu_2} - \frac{\bar{F}_1(x)}{\mu_1}\right\}dx.$$

But for $t < t_0, \dfrac{\bar{F}_2(x)}{\mu_2} \geq \dfrac{\bar{F}_1(x)}{\mu_2} \geq \dfrac{\bar{F}_1(x)}{\mu_1}$. Therefore, the above integral is positive.

The following result provides a lower bound for the distribution function for any member of NWUE class in terms of its expectation.

LEMMA 3: If F is a NWUE and μ is its expectation, then $F(t) \geq \dfrac{t}{t+\mu}$ for all $t \geq 0$.

PROOF: $\mu\bar{F}(t) \leq \displaystyle\int_t^\infty \bar{F}(x)dx = \mu - \int_0^t \bar{F}(x)dx \leq \mu - t\bar{F}(t)$ because $\bar{F}(x) \geq \bar{F}(t)$ for all $x \leq t$. Thus, $\bar{F}(t) \leq \dfrac{\mu}{\mu+t}$ or $F(t) \geq \dfrac{t}{\mu+t}$, $t \geq 0$.

To show that the above bound is sharp we consider the following example. Let X be a nonnegative random variable such that

$$P[X = 0] = \alpha$$

$$P[0 < X < a] = 0$$

and for $X > a$, the density is given by

$$f(x) = (1-\alpha)\frac{1}{\lambda}\, e^{-(x-a)/\lambda}$$

where $\lambda = a(\alpha - (1-\alpha)^2)/(1-\alpha)^2$, and $0<\alpha<1$ is chosen so that $\alpha - (1-\alpha)^2 > 0$.

The expected value, μ_a, of the random variable X is given by

$$\mu_a = (1-\alpha)\,[a + \lambda]$$

and the distribution function F is such that

$$\bar{F}_a(x) = \begin{cases} (1-\alpha) & \text{for } 0 \leqslant x < a \\ (1-\alpha)\, e^{-(x-a)/\lambda} & \text{for } x \geqslant a, \end{cases}$$

Clearly, F is NWUE. Moreover, for $x = a$

$$\frac{x}{x + \mu_a} = \frac{a}{a + \mu_a} = (1-a) = \bar{F}_a(a),$$

implying that for any give $\mu > 0$, and for each $a > 0$, there exists a NWUE F_a such that $F_a(a) = \dfrac{a}{a + \mu}$.

REFERENCES

[1] Barlow, Richard E. and F. Proschan, *Statistical Theory of Reliability and Life Testing: Probability Models.* (Holt, Rinehart and Winston, New York, N.Y. 1975.)

[2] Marshall, A.W. and F. Proschan, "Classes of Distributions Applicable in Replacement with Renewal Theory Implications." *Sixth Berkeley Symposium* I, 395-416, (1970).

A NOTE ON OPTIMAL SWITCHING BETWEEN TWO ACTIVITIES

Steven E. Shreve

Department of Mathematics
Carnegie-Mellon University
Pittsburgh, Pennsylvania

ABSTRACT

Let f_1 and f_2 map $[0,T]$ into the real numbers. A system is following either f_1 or f_2 and earning the associated reward $\int f_1$ or $\int f_2$, respectively. It is possible at any time to switch from f_i to f_j by paying a switching cost $b > 0$. We determine a switching policy which maximizes the total reward. Conditions which guarantee a planning horizon are established.

INTRODUCTION

In many endeavors one must choose one of two activities, each of which has a time-varying reward. There is usually a cost associated with switching from one activity to the other. Such is the case in fisheries, where a fisherman chooses each day to fit his boat for deep or shallow fishing, and this paper stems from a model of such behavior. We model this situation in the following way. Let f_1 and f_2 map $[0,T]$ into the real numbers. A system is following either f_1 or f_2 and earning the associated reward $\int f_1$ or $\int f_2$, respectively. It is possible at any time to switch from f_i to f_j by paying a switching cost $b > 0$. For example, if the system begins following f_1, switches at time $t_1 \geqslant 0$ to f_2, and then switches back to f_1 at time $t_2 \geqslant t_1$, the total reward is

$$\int_0^{t_1} f_1 + \int_{t_1}^{t_2} f_2 + \int_{t_2}^{T} f_1 - 2b.$$

The problem of optimal switching between two activities has been studied by Pekelman [2], who required switching to occur in a continuous fashion with a bounded rate. In our case, switching occurs instantaneously. Pekelman derived the nature of an optimal policy using Lagrange multipliers, and showed the existence of planning and forecast horizons. Our problem is simpler, and our analysis relies on dynamic programming. We also characterize planning horizons.

REDUCTION AND ASSUMPTIONS

A function f is said to change sign at t if f takes both negative and positive values in any neighborhood of t. We assume

(A1) The set of points in $[0,T]$ where $f_1 - f_2$ changes sign is nonempty and finite. Let $0 < t_1 \leqslant t_2 \leqslant \ldots \leqslant t_n < T$ be an enumeration of this set.

(A2) There is no interval in $[0,T]$ on which $f_1 - f_2$ is identically zero. A model which does not satisfy this assumption can be reduced to one which does.

We note that the performance of any policy is dominated by the performance of a policy which switches only on the set $\{0, t_1, t_2, \ldots, t_n\}$. If a policy mandates a switch at $s \in (t_k, t_{k+1})$, then the switch can be relocated from s to t_k, t_{k+1}, or to coincide with some other switch in (t_k, t_{k+1}), so that the reward is not decreased. If two switches coincide, they can both be eliminated with no loss of reward. It is clear then that the switch at s can be moved to t_k, t_{k+1}, or eliminated altogether.

This observation restricts our attention to policies which switch only at $\{0, t_1, \ldots, t_n\}$. Since there are only finitely many such policies, an optimal policy exists. Let $t_0 = 0$ and define

$$\alpha_k = \int_{t_k}^{t_{k+1}} (f_2 - f_1).$$

A policy which mandates following f_2 on the intervals $[t_{k_j}, t_{k_j+1}]$, $0 \leqslant t_{k_1} < t_{k_2} < \ldots < t_{k_m} \leqslant t_n$ earns reward

$$\sum_{j=1}^{m} \alpha_{k_j} + \int_0^T f_1 - cb,$$

where c is the number of switches incurred by the policy. We have thus reduced our problem to the following sequential optimization model.

M: At each stage k, a system is in either state 0 or state 1. A policy $\pi = (\mu_0, \mu_1, \ldots, \mu_n)$ is a sequence such that each μ_k maps $\{0,1\}$ into {Hold, Switch} or simply {H,S}. If the k-th state is x_k, then the k-th control is $u_k = \mu_k(x_k)$, the $(k+1)$-*st* state is

$$x_{(k+1)} = f(x_k, u_k) = \begin{cases} x_k & \text{if} \quad u_k = H, \\ 1 - x_k & \text{if} \quad u_k = S, \end{cases} \tag{1}$$

and the reward associated with (x_k, u_k) is

$$g(x_k, u_k) = \begin{cases} 0 & \text{if} \quad x_k = 0, \ u_k = H, \\ -b + \alpha_k & \text{if} \quad x_k = 0, \ u_k = S, \\ \alpha_k & \text{if} \quad x_k = 1, \ u_k = H, \\ -b & \text{if} \quad x_k = 1, \ u_k = S. \end{cases} \tag{2}$$

We wish to find a policy π which maximizes

$$J_\pi(x_0) = \sum_{k=0}^{n} g(x_k, u_k). \tag{3}$$

This is a finite stage, deterministic, dynamic programming model with two states and two actions. The dynamic programming algorithm for this model is simple and computationally feasible. This model has, however, a special feature which leads to a more efficient algorithm. It is apparent that whenever $0 \leqslant \alpha_k \leqslant 2b$ $(-2b \leqslant \alpha_k \leqslant 0)$, there is nothing to be gained by switching from 0 to 1 (1 to 0) at stage k and back to 0(1) at stage $k+1$. To build on this observation, we define a model more general than M.

DEFINITION: We say a dynamic programming model N is *alternating* if it has two states 0 and 1, two actions H and S, system equation (1), one-stage reward (2) and objective functional (3). We require that $b \geqslant 0$ and $A_N = (\alpha_0, \alpha_1, \ldots, \alpha_n)$ is an ordered set of real numbers such that $\alpha_0 \neq 0$, and the nonzero members of the set have alternating signs. A policy for an alternating model is a sequence $\pi = (\mu_0, \mu_1, \ldots, \mu_n)$ such that μ_k maps $\{0,1\}$ into

$\{H,S\}$. We say π is *feasible* if $\alpha_k = 0$ implies $\mu_k(x_k) = H$, regardless of the choice of x_0. We say π is optimal if π maximizes $J_\pi(x_0)$ over all feasible policies (independent of x_0), and the reward $J_\pi(x_0)$ corresponding to an optimal π is called the *value function*.

The model M is an alternating model with every α_k different from zero. Given an alternating model N, we can construct a related alternating model $\phi(N)$ by the following procedure:

(P) Let $m \leqslant n$ be the largest integer for which $\alpha_m \neq 0$. If α_m and α_0 are the only nonzero members of A_N, set $\phi(\alpha_k) = \alpha_k$, $k = 0, \ldots, n$. Otherwise, determine the index $\bar{k}$, $1 \leqslant \bar{k} \leqslant m-1$, of the smallest nonzero $|\alpha_k|$. If more than one such $|\alpha_k|$ exists, choose the smallest index. If $|\alpha_{\bar{k}}| > 2b$, set $\phi(\alpha_k) = \alpha_k$, $k = 0, \ldots, n$. If $|\alpha_{\bar{k}}| \leqslant 2b$, let $\bar{p} = \max\{p \mid 0 \leqslant p \leqslant \bar{k}-1, \alpha_p \neq 0\}$, $\bar{q} = \min\{q \mid \bar{k}+1 \leqslant q \leqslant n, \alpha_q \neq 0\}$, and set

$$\phi(\alpha_k) = \alpha_k, \quad k \neq \bar{p}, \ k \neq \bar{k}, \ k \neq \bar{q},$$

$$\phi(\alpha_{\bar{p}}) = \alpha_{\bar{p}} + \alpha_{\bar{k}} + \alpha_{\bar{q}},$$

$$\phi(\alpha_{\bar{k}}) = \phi(\alpha_{\bar{q}}) = 0.$$

The model $\phi(N)$ is the model N with each α_k replaced by $\phi(\alpha_k)$. It is easily verified that $\phi(N)$ is alternating.

Either the models $\phi(N)$ and N are the same, or else $\phi(N)$ is simpler than N is the sense that $A_{\phi(N)}$ contains more zeroes than A_N. For example, if $A_N = (-1, 3, 0, 0, -2, 5)$ and $b = 1$, then $A_{\phi(N)} = (-1, 6, 0, 0, 0, 0)$.

LEMMA: Let N be an alternating model and let $\phi(N)$ be derived from N by procedure (P). Then every optimal policy in $\phi(N)$ is also optimal in N.

PROOF: Since every feasible policy in $\phi(N)$ is feasible in N and leads to the same reward in both models, it suffices to show that both models have the same value function. We will show this by producing a policy which is optimal in N and feasible in $\phi(N)$. There is nothing to prove when $\phi(N) = N$, so we assume the contrary, i.e., $|\alpha_k| \leqslant 2b$.

The dynamic programming algorithm for N takes the following form. For $k = 0, 1, \ldots, n$, if $\alpha_k \neq 0$,

$$J_k(0) = \max\{J_{k+1}(0), -b + \alpha_k + J_{k+1}(1)\} \tag{4}$$

$$J_k(1) = \max\{\alpha_k + J_{k+1}(1), -b + J_{k+1}(0)\}, \tag{5}$$

while if $\alpha_k = 0$,

$$J_k(0) = J_{k+1}(0), \tag{6}$$

$$J_k(1) = J_{k+1}(1), \tag{7}$$

where $J_{n+1}(0) = J_{n+1}(1) = 0$. Define $\pi = (\mu_0, \mu_1, \ldots, \mu_n)$ by

$$\mu_k(0) = \begin{cases} H & \text{if} \quad J_k(0) = J_{k+1}(0), \\ S & \text{if} \quad J_k(0) > J_{k+1}(0), \end{cases} \tag{8}$$

$$\mu_k(1) = \begin{cases} H & \text{if} \quad J_k(1) = \alpha_k + J_{k+1}(1), \\ S & \text{if} \quad J_k(1) > \alpha_k + J_{k+1}(1). \end{cases} \tag{9}$$

The policy π is optimal for (N) [1, p. 50]. We show it is feasible for $\phi(N)$, i.e., for any initial state x_0,

(10) $\mu_k(x_k) = H.$

(11) $\mu_q(x_q) = H.$

Observe first that (4)-(7) imply

(12) $J_k(0) \leqslant b + J_k(1), \quad k = 0, 1, \ldots, n+1.$

(13) $J_k(1) \leqslant b + J_k(0), \quad k = 0, 1, \ldots, n+1.$

Recall that $\alpha_k \neq 0$ and $|\alpha_k| \leqslant 2b$. Since $J_{p+1} = J_k$ and $J_{k+1} = J_q$, we can and do assume for simplicity of notation that $\bar{p} = \bar{k} - 1$, $\bar{q} = \bar{k} + 1$. Thus, we have $|\alpha_{\bar{k}-1}| > |\alpha_{\bar{k}}|$, $|\alpha_{\bar{k}+1}| \geqslant |\alpha_{\bar{k}}|$.

We assume $\alpha_{\bar{k}} < 0$. The other case is treated similarly. We have

$\alpha_{\bar{k}-1} > 0$, $\alpha_{\bar{k}+1} > 0$. From (12) we have

$$-b + J_{\bar{k}+2}(0) \leqslant J_{\bar{k}+2}(1) < \alpha_{\bar{k}+1} + J_{\bar{k}+2}(1),$$

so (5) and (9) imply

(14) $J_{\bar{k}+1}(1) = \alpha_{\bar{k}+1} + J_{\bar{k}+2}(1), \quad \mu_{\bar{k}+1}(1) = H.$

Since $\alpha_{\bar{k}} < 0$, (13) implies

$$-b + \alpha_{\bar{k}} + J_{\bar{k}+1}(1) < -b + J_{\bar{k}+1}(1) \leqslant J_{\bar{k}+1}(0),$$

so from (4) and (8) we have

(15) $J_{\bar{k}}(0) = J_{\bar{k}+1}(0), \quad \mu_{\bar{k}}(0) = H.$

Since $|\alpha_{\bar{k}}| \leqslant |\alpha_{\bar{k}+1}|$, we have from (12),

$$-b + J_{\bar{k}+2}(0) \leqslant J_{\bar{k}+2}(1) \leqslant \alpha_{\bar{k}} + \alpha_{\bar{k}+1} + J_{\bar{k}+2}(1).$$

Since $|\alpha_{\bar{k}}| \leqslant 2b$, we also have

$$-2b + \alpha_{\bar{k}+1} + J_{\bar{k}+2}(1) \leqslant \alpha_{\bar{k}} + \alpha_{\bar{k}+1} + J_{\bar{k}+2}(1).$$

Together with (4) and (14), these inequalities yield

$$-b + J_{\bar{k}+1}(0) \leqslant \alpha_{\bar{k}} + \alpha_{\bar{k}+1} + J_{\bar{k}+2}(1) = \alpha_{\bar{k}} + J_{\bar{k}+1}(1).$$

From (5) and (9) we now have

(16) $J_{\bar{k}}(1) = \alpha_{\bar{k}} + J_{\bar{k}+1}(1), \quad \mu_{\bar{k}}(1) = H.$

Equations (15) and (16) imply (10). It remains to establish (11).

Since $\alpha_{\bar{k}+1} > 0$, (5), (9) and (12) imply $\mu_{\bar{k}+1}(1) = H$. If $J_{\bar{k}+2}(0) \geqslant -b + \alpha_{\bar{k}+1} + J_{\bar{k}+2}(1)$, then (4) and (8) imply

$$J_{\bar{k}+1}(0) = J_{\bar{k}+2}(0), \quad \mu_{\bar{k}+1}(0) = H,$$

and (11) follows. On the other hand, if $J_{\bar{k}+2}(0) < -b + \alpha_{\bar{k}+1} + J_{\bar{k}+2}(1)$, then

(17) $J_{\bar{k}+1}(0) = -b + \alpha_{\bar{k}+1} + J_{\bar{k}+2}(1), \quad \mu_{\bar{k}+1}(0) = S,$

and (11) will hold if and only if $x_{\bar{k}+1} = 1$ (independent of the choice of x_0). Since $\alpha_{\bar{k}-1} + \alpha_{\bar{k}} > 0$, (15), (17), (14) and (16) imply

$$\begin{aligned} J_k(0) = J_{k+1}(0) &= -b + \alpha_{k+1} + J_{k+2}(1) \\ &< -b + \alpha_{k-1} + \alpha_k + \alpha_{k+1} + J_{k+2}(1) \\ &= -b + \alpha_{k-1} + \alpha_k + J_{k+1}(1) \\ &= -b + \alpha_{k-1} + J_k(1). \end{aligned}$$

From (4) and (8) we see that

$$J_{k-1}(0) = -b + \alpha_{k-1} + J_k(1), \quad \mu_{k-1}(0) = S. \tag{18}$$

Since $\alpha_{k-1} > 0$, (12) implies

$$-b + J_k(0) \leqslant J_k(1) \leqslant \alpha_{k-1} + J_k(1),$$

and (5) and (9) yield

$$J_{k-1}(1) = \alpha_{k-1} + J_k(1), \quad \mu_{k-1}(1) = H. \tag{19}$$

Equations (1), (18) and (19) imply $x_k = 1$. Equations (1) and (16) imply $x_{k+1} = 1$, as was to be proved. Q.E.D.

We state now a theorem which gives a simple construction of an optimal policy for an alternating model N for which $\phi(N) = N$. We show also that any alternating model can be reduced to this case.

THEOREM: Let N be alternating model for which $\phi(N) = N$. If A_N has only two nonzero members α_0 and α_m, then an optimal policy $\pi = (\mu_0, \mu_1, \ldots, \mu_m)$ for (N) is given by

$$\mu_k(x_k) = H, \quad k \neq 0, \quad k \neq m, \tag{20}$$

$$\mu_m(x_m) = \begin{cases} S \text{ is } x_m = 0, \ \alpha_m > b \\ \quad \text{or } x_m = 1, \ \alpha_m < -b, \\ H \text{ otherwise.} \end{cases} \tag{21}$$

$$\mu_0(x_0) = \begin{cases} S \text{ if } x_0 = 0, \ -b + \alpha_0 + J_m(1) > J_m(0), \\ \quad \text{or } x_0 = 1, \ -b + J_m(0) > \alpha_0 + J_m(1), \\ H \text{ otherwise.} \end{cases} \tag{22}$$

where

$$J_m(0) = \max\{0, -b + \alpha_m\},$$
$$J_m(1) = \max\{\alpha_m, -b\}.$$

If A_N has more than two nonzero members, then the policy defined by

$$\mu_k(x_k) = \begin{cases} S \text{ if } x_k = 0, \ \alpha_k > b, \\ \quad \text{or } x_k = 1, \ \alpha_k < -b, \\ H \text{ otherwise.} \end{cases} \tag{23}$$

is optimal for N. If $\phi(N) \neq N$, then there exists some positive integer i such that $\phi^{i+1}(N) = \phi^i(N)$, and any optimal policy for $\phi^i(N)$ is also optimal for N.

PROOF: For the trivial case where A_N has only two nonzero members, the optimality of the policy given by (20)-(22) follows directly from the dynamic programming algorithm (4)-(9). Suppose now A_N has three or more nonzero members and m is the largest index with $\alpha_m \neq 0$. Since $\phi(N) = N$, $\alpha_k \neq 0$ implies $|\alpha_k| > 2b$ for $1 \leqslant k \leqslant m-1$. The optimality of (23) follows from (4)-(9), (12) and (13). Finally, if $\phi(N) \neq N$, then $A_{\phi(N)}$ contains fewer nonzero elements than A_N. After finitely many iterations of ϕ, we must obtain $\phi^l(N)$ such that $\phi^{l+1}(N) = \phi^l(N)$. Q.E.D.

EXISTENCE OF PLANNING HORIZONS.

Suppose in an alternating model N we have $\alpha_k > 2b$ for some k. Then, in the notation of procedure (P), either $\phi(\alpha_k) = \alpha_k$, $\phi(\alpha_k) = \alpha_k + \alpha_k + \alpha_q \geqslant \alpha_k$, or $\phi(\alpha_k) = 0$. The last case occurs if $k = \bar{q}$, in which case $\bar{p} < k$, $\phi(\alpha_p) = \alpha_p + \alpha_k + \alpha_k \geqslant \alpha_k$, and $\phi(\alpha_{p+1}) = \ldots = \phi(\alpha_k) = 0$. For any i, we will have either $\phi^l(\alpha_k) \geqslant \alpha_k > 2b$, or else $\phi^l(\alpha_l) > 2b$, where $l < k$ and $\phi^l(\alpha_{l+1}) = \ldots = \phi^l(\alpha_k) = 0$. If $\phi^{l+1}(N) = \phi^l(N)$, then the optimal policy of the Theorem guarantees that $x_{k+1} = 1$. Thus, we can disregard α_j for $j \geqslant k+1$ in determining an optimal policy for stages 0 through k. If $\alpha_k < -2b$, a similar argument holds, where now we have $x_{k+1} = 0$.

In conclusion, if $|\alpha_k| > 2b$, or for any $i \geqslant 1$, $|\phi^l(\alpha_k)| > 2b$, then we can solve the smaller problem of optimal switching between any stage $h < k$ and stage k independent of the values of α_j, where j does not satisfy $h \leqslant j \leqslant k$, and the policy thereby obtained will be part of an optimal policy for the full problem.

REFERENCES

[1] Bertsekas, D.P. *Dynamic Programming and Stochastic Control*, (Academic Press, New York, N.Y. 1976.)

[2] Pekelman, D. "On Optimal Utilization of Production Processes," *Operations Research 27*, 260-278 (1979).

NEWS AND MEMORANDA

THE 1980 LANCHESTER PRIZE

Call for Nominations

Each year since 1954 the Council of the Operations Research Society of America has offered the Lanchester Prize for the best English-language published contribution in operations research. The Prize for 1980 consists of $2000 and a commemorative medallion.

The screening of books and papers for the 1980 Prize will be carried out by a committee appointed by the Council of the Society. To be eligible for consideration, the book or paper must be nominated to the Committee. Nominations may be made by anyone; this notice constitutes a call for nominations.

To be eligible for the Lanchester Prize, a book, a paper or a group of books or papers must meet the following requirements:

(1) It must be on an operations research subject,

(2) It must carry a current award year publication date, or, if a group, or at least one member of the group must carry a current award year publication date,

(3) It must be written in the English language, and

(4) It must have appeared in the open literature.

The books(s) or papers(s) may be a case history, a report of research representing new results, or primarily expository.

For any nominated set (e.g., article and/or book) covering more than the most recent year, it is expected that each element in the set represents work from one continuous effort, such as a multi-year project or a continuously written, multi-volume book.

Judgments will be made by the Committee using the following criteria:

(1) The magnitude of the contribution to the advancement of the state of the art of operations research,

(2) The originality of the ideas or methods,

(3) New vistas of application opened up,

(4) The degree to which unification or simplification of existing theory or method is achieved, and

(5) Expository clarity and excellence.

Nominations should be sent to:

Linus E. Schrage
Graduate School of Business
University of Chicago
1101 East 58th Street
Chicago, Illinois 60637

Nominations may be in any form, but must include as a minimum the title(s) of the paper(s) or book, author(s), place and date of publication, and six copies of the material. Supporting statements bearing on the worthwhileness of the publication in terms of the five criteria will be helpful, but are not required. Each nomination will be carefully screened by the Committee; nominations must be received by May 30, 1981, to allow time for adequate review.

Announcement of the results of the Committee and ORSA Council action, as well as award of any prize(s) approved, will be made at the National Meeting of the Society, October 19-21, 1981 in Houston, Texas.

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

MARCH 1981
VOL. 28, NO. 1
NAVSO P-1278

CONTENTS

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217