RESEARCH REPORT 37
23 OCTOBER 1962
I.E.R. 172-46

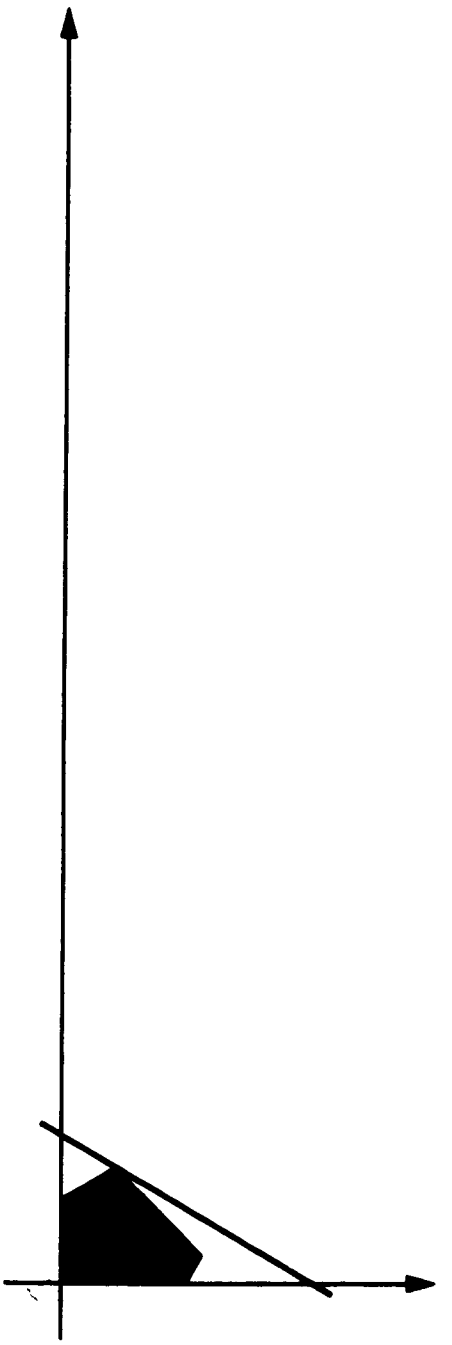


# MARKOV-RENEWAL PROGRAMMING

by

William S. Jewell





OPERATIONS RESEARCH CENTER

INSTITUTE OF ENGINEERING RESEARCH



UNIVERSITY OF CALIFORNIA - BERKELEY

# MARKOV-RENEWAL PROGRAMMING

by

William S. Jewell
Operations Research Center and
Department of Industrial Engineering
University of California, Berkeley

23 October 1962

Research Report 37

This research has been partially supported by the Office of Naval Research under Contract Nonr-222(83) with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

## ABSTRACT

A special structure in dynamic programming which has been studied by Bellman,[2] Blackwell,[6] D'Épenoux,[7] Derman,[8] Howard,[9] Manne,[13] Oliver,[14] Wolfe and Dantzig,[20] and others is the problem of programming over a Markov chain. This paper extends their results and solution algorithms to programming over a Markov-renewal process -- in which times between transitions of the system from state i to state j are independent samples from an inter-transition distribution which may depend on both i and j. For these processes, a general reward structure and a decision mechanism are postulated; the problem is to make decisions at each transition to maximize the total expected reward at the end of the planning horizon.

For finite-horizon problems, or infinite-horizon problems with discounting, there is no difficulty; the results are similar to previous work, expect for a new dependency upon the transition-time distributions being generally present. In the cases where the horizon extends towards infinity, or when discounting vanishes, however, a fundamental dichotomy in the optimal solutions may occur. It then becomes important to specify whether the limiting experiment is: (i) undiscounted, with the number of transitions $n \to \infty$, (ii) undiscounted, with a time horizon $t \to \infty$, or (iii) infinite n or t, with discount factor $\alpha \to 0$. In each case, a limiting form for the total expected reward is shown, and an algorithm developed to maximize the rate of return. The problem of finding the optimal or near-optimal policies in the case of ties in rate of return is still computationally unresolved.

Extensions to non-ergodic processes are indicated, and special results for the two-state process are presented. Finally, an example of machine maintenance and repair is used to illustrate the generality of the approach and the special problems which may arise.

# MARKOV-RENEWAL PROGRAMMING

## I. Introduction

An important special structure of dynamic programming occurs in the Markov decision processes first formulated by Bellman,[2][3] developed extensively by Howard,[9] and further analyzed by Oliver,[14] Manne,[13] D'Épenoux,[7] Blackwell,[6] Wolfe and Dantzig,[20] Derman,[8] and others. In this model, the system makes Markovian transitions from one to another of a finite set of states, accumulating a reward at each transition. A decision is made at each step from among a finite number of alternatives; this decision affects both the transition probabilities and the rewards obtained upon leaving the present state. The problem is to specify the policy of decisions to be made in each state which will maximize the total expected return at the end of the experiment.

The following cases are formulated by Howard:[9]

I. Discrete-parameter Markov chain
   - A. Finite number of transitions
   - B. Infinite number of transitions
   - C. Repeat of both cases with discounting

II. Continuous-parameter Markov chain
   - A. Finite planning horizon
   - B. Infinite planning horizon
   - C. Repeat of both cases with discounting

The models with finite horizons are all expressed in terms of the usual recursive relationships of dynamic programming,[2][4] whose solution techniques are well-known. Howard's contribution was the development of simple, finite, iterative techniques to find the optimal stationary policies to be followed in the infinite cases; since the total reward is unbounded in the undiscounted, infinite horizon cases, the rate of return becomes the

system objective. Blackwell[6] shows a similar algorithm for maximizing the return for a vanishing discount factor, and proves that among the optimal policies, there is one which is stationary.

The purpose of this paper is to generalize all of the above models and algorithms of Markov decision processes to a larger class of dynamic models in which the Markov-renewal process is used to describe system behavior. The important generalization provided by these processes is that time spent by the system between transitions may be a random variable. The resulting Markov-renewal decision processes will be seen to embrace a wider range of important operational problems, without seriously complicating the calculation of optimal policies.

The first three sections of the paper describe the properties of the Markov-renewal process, the reward structure assumed, and the decision process. The first cases analyzed are the finite-step and finite-time problems, both discounted and undiscounted. Next, discounted problems with an infinite horizon are examined, followed by a discussion of the difficulties of undiscounted, infinite-horizon models. Three distinct "infinite" cases are presented, with some remarks on the problems of ties and near-optimal policies. Extensions to nonergodic structures are described in the next section, followed by a comparison with some previous results. The paper closes with explicit formulae for the two-state process, a machine maintenance-repair example, and suggestions for further research.

## II. Markov-Renewal Processes

The Markov-renewal processes and the related semi-Markov processes were first studied by Levy, Smith, and Takács independently in 1954. In References [15] and [16] Pyke has summarized current results in this area, together with independent contributions and an exhaustive list of references. Other specific results are in References [17], [18], and [1].

Loosely speaking, Markov-renewal processes are generalizations of both the discrete- and continuous-parameter Markov chains in which the time between transitions of the system from state $i$ to state $j$ is a random variable obtained from a distribution which depends upon both $i$ and $j$. We shall only be concerned with Markov-renewal processes with a finite number of states,[16] labelled by some integer $i$, $i = 1, 2, \ldots, N$.

A particular realization of a Markov-renewal process consists of an initial integer $i_0$, followed by pairs of random variables, one of which is an integer, and the other a non-negative variable $\tau$, viz:

$$i_0 \; ; \; i_1 \, , \, \tau(i_0, i_1) \; ; \; i_2 \, , \, \tau(i_1, i_2) \; ; \; i_3 \, , \, \tau(i_2, i_3) \; ; \; \ldots$$

The integer $i_0$ represents the initial state of the system at time zero; it may be given uniquely, or determined from some initial distribution. The sequence of integers $i_1, i_2, i_3, \ldots$ represents the successive states of the system as it makes transitions between its allowed states at steps $1, 2, 3, \ldots$ These integers are generated by a Markov process, so that the conditional probability distribution

$$p_{ij} = \Pr\{i_{k+1} = j \,|\, i_k = i\} \qquad \begin{matrix} k = 0, 1, 2, \ldots \\ i, j = 1, 2, \ldots, N \end{matrix} \tag{1}$$

contains all of the information necessary to generate the successive states of the system, once $i_0$ is known.[11]

The sequence of non-negative variables represents the transition intervals between successive states. Thus $\tau(i_k, i_{k+1})$ is the time between the instants the system entered state $i_k$ at step $k$ and the time it entered state $i_{k+1}$ at the next transition. It is not necessary to describe the state of the system between these transition instants, in general; however, for convenience, one may speak of the system as being in state $i$, headed toward state $j$. Notice that it is necessary to select the next state immediately upon entering a given state so that the transition interval can be determined; in Markov-renewal processes this transition interval is determined from the stationary distribution functions:

$$F_{ij}(t) = \Pr\{\tau(i,j) \leq t\} \qquad \begin{matrix} t \geq 0 \\ i,j = 1,2,\ldots,N \end{matrix} \quad . \tag{2}$$

The moments of this distribution will be denoted by $\nu_{ij}^{(n)} = E\{[\tau(i,j)]^n\}$, $n = 0, 1, 2, \ldots$ ; the superscript (1) is suppressed for the mean transition interval. It is assumed that $F_{ij}(0) = 0$ for all $i, j$, so that $0 < \nu_{ij}^{(n)} \leq \infty$, for all $n$.

For convenience, several results of Pyke which will be used in the sequel are presented in Appendix A; full details and additional explicit formulae may be found in References [15], [16], [17], [18], and [1].

## III. The Reward Structure

We next describe the reward structure which will be assumed for the system. When the system makes a transition into state $i$, heading towards $j$, we assume that a fixed return of $R_{ij}$ dollars is received. Also, a variable running return of $r_{ij}$ dollars per unit time is assumed to be generated during the transition time, so that for a time $t$ since the last transition, a cumulative reward $R_{ij} + r_{ij}t$ $(0 \leq t \leq \tau(i,j))$ is generated. The expected return from state $i$, heading towards $j$ is:

$$\rho_{ij} = \int_0^\infty (R_{ij} + r_{ij}t)\, dF_{ij}(t) = R_{ij} + r_{ij}\nu_{ij} \quad . \tag{3}$$

If a discount factor $\alpha$ is used,

$$\rho_{ij}(\alpha) = \int_0^\infty \left( R_{ij} + r_{ij} \int_0^t e^{-\alpha x}\, dx \right) dF_{ij}(t)$$

$$= R_{ij} + \frac{r_{ij}}{\alpha} \left[ 1 - \int_0^\infty e^{-\alpha t}\, dF_{ij}(t) \right] \quad . \tag{4}$$

for all $i,j$. More general reward structures may also be used.

## IV. The Decision Process

It remains to describe the procedure by which the system behavior will be governed. Let us assume that there are a finite number of alternatives, $z = 1, 2, \ldots, Z$, available in each state of the system; selection of a certain alternative then influences the transition times and transition probabilities to the next state, as well as the rewards to be obtained during the interval until that transition. To put it another way, there are families of $Q_{ij}^z(t) = p_{ij}^z F_{ij}^z(t)$ functions, as well as decision-dependent rewards $R_{ij}^z$,

and reward rates, $r_{ij}^{z}$, for each $z = 1, 2, \ldots, Z$. The operating policy of the system is a selection of a $z$ to be used in each state of the system, possibly depending also on the remaining length of the experiment.

To summarize the system behavior under the influence of a given policy:

1. System enters state $i$.
2. Alternative, $z(i)$, is selected from among the available alternatives; it is a function only of the current system state $i$, and (possibly) the remaining length of the experiment.
3. Based upon $z(i)$, a next state $j$ is selected as a sample from the conditional probabilities, $p_{ij}^{z(i)}$; the sojurn time until that next state is entered, $\tau(i,j;z(i))$, is selected as a sample from the distribution $F_{ij}^{z(i)}(t)$.
4. For a clock time $t$ since state $i$ was entered, a cumulative reward

$$R_{ij}^{z(i)} + r_{ij}^{z(i)} t \qquad 0 \leq t \leq \tau(i,j;z(i))$$

is generated.

5. The system enters state $j$, and the process is repeated until the experiment is terminated.

The fundamental problem which we shall consider in this paper is the selection of the alternatives for each state, $z(i)$, which will maximize total expected return over the length of the experiment.

As we shall see, determination of this optimal policy will depend critically upon whether discounting is or is not used, or in the way in which certain limiting experiments are defined. In some cases, general results as to the optimal policy will not be obtainable.

## V. Finite Step, Discounted Case

The first experiment is the operation of the system for a fixed number of transition steps, $n$. Following the usual procedure of dynamic programming[2][4] define for all $i$, $n = 0, 1, 2, \ldots$

$V_i(n)$ = expected return obtained from an n-step process starting in state $i$, and using an optimal policy, $z^*(i, n)$.

Continuous discounting with parameter $\alpha$ per unit time will be used, so that $V_i(n)$ also depends upon $\alpha$.

Setting for convenience, $V_i(0) = 0$ $(i = 1, 2, \ldots, N)$, the optimal expected returns for a one-step process can be obtained for all $i$ through the relation:

$$V_i(1) = \max_z \rho_i^z(\alpha) \tag{5}$$

where

$$\rho_i^z(\alpha) = \sum_{j=1}^{N} p_{ij}^z \rho_{ij}^z(\alpha) = \sum_{j=1}^{N} p_{ij}^z \{R_{ij}^z + r_{ij}^z [1 - \tilde{f}_{ij}^z(\alpha)]/\alpha\} \tag{6}$$

is the expected discounted, one-step return, starting in state $i$ and following policy $z$. In (6) a tilde is used to indicate the Laplace-Stieltjes transform of $F_{ij}(t)$, or the Laplace transform of its derivative, $f_{ij}(t)$, if it exists:

$$\tilde{f}_{ij}(s) = \int_0^\infty e^{-st} dF_{ij}(t) = \int_0^\infty e^{-st} f_{ij}(t)\, dt \quad . \tag{7}$$

Similar notation for other transforms will be used in the sequel. Using the principle of optimality[2][4] the recurrence relations for $n = 2, 3, \ldots$

and all $i$, are:

$$V_i(n) = \max_z \left[ \rho_i^z(\alpha) + \sum_{j=1}^{N} p_{ij}^z \tilde{f}_{ij}^z(\alpha) V_j(n-1) \right] . \qquad (8)$$

The appearance of the factor $\tilde{f}_{ij}^z(\alpha)$ is due to the necessity for discounting the return with $n-1$ steps left by an amount which depends upon $\tau(i,j)$. The expected discounted return is then

$$\int_0^\infty e^{-\alpha t} V_j(n-1)\, dF_{ij}(t) = \tilde{f}_{ij}^z(\alpha) V_j(n-1) .$$

Equation (8) suggests a simple technique for computing the optimal policies, $z^*(i,n)$. One begins with $n = 1, 2, \ldots$, building up the optimal policies, and returns for successively larger problems. As Bellman and Dreyfus have pointed out[4] the computation is not complicated so much by the requirements for storing the sequences $\{V_i(n)\}$ and $\{z^*(i,n)\}$ as it is by the necessity of storing the $4Z$ matrices $(R_{ij}^z)$, $(r_{ij}^z)$, $(p_{ij}^z)$, and $(\tilde{f}_{ij}^z(s))$, each of which is of dimension $N^2$. The last matrix must of course be recalculated for each change in discount factor.

In the special case where transition intervals are all of fixed length $\tau$, then $\tilde{f}_{ij}(\alpha) = e^{-\alpha\tau} = \beta$, and (8) may be written as

$$V_i(n) = \max_z \left\{ \sum_{j=1}^{N} p_{ij}^z \left[ R_{ij}^z + r_{ij}^z \frac{(1-\beta)}{(-\ln\beta)} \tau + \beta V_j(n-1) \right] \right\} \qquad (9)$$

for $i = 1, 2, \ldots, N$ ; $n = 2, 3, \ldots$ . Upon redefinition of the expected reward per transition, (9) is seen to be equivalent to the discounted, finite-step Markov decision processes studied by Howard[9] and Blackwell.[6]

In the more general case of the Markov-renewal process, we obtain a complicated dependence on the discount factor because of the influence of the entire shape of the transition time distributions, expressed through the $\tilde{f}_{ij}^{z}(\alpha)$ . If the boundary rewards or penalties, $V_i(0)$ , are not zero, (5) should be replaced by:

$$V_i(1) = \max_z \left[ \rho_i^z(\alpha) + \sum_{j=1}^{N} p_{ij}^z \tilde{f}_{ij}^z(\alpha) V_j(0) \right] \qquad (5')$$

or, equivalently, the range of (8) and (9) can be extended to $n = 1, 2, 3, \ldots$ .

Because of the discounting, the sequence of expected returns $\{V_i(n)\}$ approaches a finite limit as $n$ approaches infinity, for all $\alpha > 0$ . It is not apparent what happens to the sequence of optimal decisions $\{z^*(i, n)\}$ ; we shall return to this point in a later section.

## VI. Finite Time, Discounted Case

Because the transitions in a Markov-renewal process occur stochastically over time, another possible experiment suggests itself — operation of the system for a fixed period of time, $t$ . For example, in certain operational problems it may be more realistic to think of a fixed horizon in time, rather than a horizon of fixed number of steps. The optimal policy, $z^*(i, t)$ , now depends upon the length of time the experiment has yet to run. Alternatives are still only selected at the transition instants; however, it is now possible to break off the experiment in the middle of some transition interval. Define for $t \geq 0$ , and all $i$ :

$V_i(t)$ = expected return obtained from a process which continues for $t$ units of time, starting in state $i$ , and using an optimal policy, $z^*(i, t)$ .

Let $V_{ij}(0)$ be the return (or penalty) imposed when system operation is terminated in state $i$, headed towards $j$.

Suppressing momentarily the policy to be followed, there are two possibilities: either the system has not transferred out of the starting state during the observation interval $(0, t]$, or it has made a transition to state $j$ at time $x$, $0 < x \leq t$. In the first case, the total discounted reward if the system was headed towards state $j$ is:

$$R_{ij} + r_{ij}[1 - e^{-\alpha t}]/\alpha + V_{ij}(0)\, e^{-\alpha t} \quad .$$

In the second case the reward is:

$$R_{ij} + r_{ij}[1 - e^{-\alpha x}]/\alpha + V_j(t - x)\, e^{-\alpha x} \quad .$$

The average discounted return for $t \geq 0$, and all $i$, is:

$$V_i(t) = \sum_{j=1}^{N} p_{ij} \left[ V_{ij}(0)\, e^{-\alpha t} F_{ij}^c(t) + R_{ij} + r_{ij} \int_0^t e^{-\alpha x} F_{ij}^c(x)\, dx \right.$$

$$\left. + \int_0^t e^{-\alpha x} V_j(t - x)\, dF_{ij}(x) \right] \qquad (10)$$

where $F_{ij}^c(t) = 1 - F_{ij}(t)$.

Because of the impossibility of zero intervals between transitions, the right hand side of (10) contains only the past history of $V_j(x)$ $(0 \leq x < t)$. The principle of optimality can thus be used to write the recurrence relation for expected total return when following an optimal policy as:

$$V_i(t) = \max_z \left\{ \sigma_i^z(\alpha, t) + \sum_{j=1}^{N} p_{ij}^z \left[ \int_0^t e^{-\alpha x} V_j(t-x) dF_{ij}^z(x) \right] \right\} \quad (11)$$

for $t \geq 0$ and all $i$, with:

$$\sigma_i^z(\alpha, t) = \sum_{j=1}^{N} p_{ij}^z \left[ V_{ij}(0)\, e^{-\alpha t} F_{ij}^{c,z}(t) + R_{ij}^z + r_{ij}^z \int_0^t e^{-\alpha x} F_{ij}^{c,z}(x)\, dx \right] . \quad (12)$$

With these definitions:

$$V_i(0) = \max_z \sum_{j=1}^{N} p_{ij}^z (V_{ij}(0) + R_{ij}^z) \quad , \quad (13)$$

i.e., it is possible to have a reward with a zero-length experiment. If desired, this anomaly can be removed by making the terminal reward a function of state $i$ only, i.e., $V_{ij}(0) = V_i(0)$ for all $i, j$, and by collecting the return $R_{ij}$ at some time slightly after $t = 0$. Or, $R_{ij}$ can be collected at the end of the transition period, in which case $R_{ij}$ should be replaced by by $R_{ij} \int_0^t e^{-\alpha x} dF_{ij}(x)$ in (4), (6) and (12).

There is, unfortunately, no general way in which (11) can be completely resolved. There are various approximation techniques in return- or policy-space[2][4] which may converge on an answer; however, if in fact digital computation is to be used, then one may proceed directly to discrete approximations of the continuous time variable in (11). Letting $t = k\Delta$ ($k = 1, 2, 3, \ldots$), one obtains for all $i$, the approximations:

$$V_i(k\Delta) = \max_z \left\{ q_i^z(\alpha, k\Delta) + \sum_{j=1}^{N} p_{ij}^z \sum_{\ell=1}^{k} e^{-\alpha\Delta\ell} V_j((k-\ell)\Delta) \cdot \right.$$

$$\left. \cdot \left[ F_{ij}^z(\ell\Delta) - F_{ij}^z(\ell\Delta - \Delta) \right] \right\} , \qquad (14)$$

which can be built up in the usual recursive manner. Round-off and truncation error limit the choice of small values of $\Delta$, but the use of more sophisticated quadrature techniques wil usually lead to results of sufficient accuracy.[4]

It should be obvious that the optimal strategies deduced for a time-horizon problem need not bear any resemblance to the optimal strategies for a problem with a fixed number of transitions, except possibly in the limit, if a stationary optimal policy exists. It is this point which we examine in the next section.

## VII. Discounted Cases with Infinite Step or Time Horizons

It is a simple matter to verify that expected returns $V_i(n)$ or $V_i(t)$ remain finite as $n$ or $t$ approach infinity in (8) and (11) for all $\alpha > 0$, since all elements of the matrix $\tilde{q} = (p_{ij}\tilde{f}_{ij}(s))$ lie in the interval $[0, 1)$ for all $s > 0$, and for all policies.

The limiting form of (8) as $n \to \infty$ is:

$$V_i = \lim_{n\to\infty} V_i(n) = \max_z \left\{ \rho_i^z(\alpha) + \sum_{j=1}^{N} p_{ij}^z \tilde{f}_{ij}^z(\alpha) V_j \right\} \qquad (15)$$

for all $i$.

From the Laplace transform of (11) and the well-known limit theorem of transform calculus, $\lim_{t\to\infty} V(t) \rightleftharpoons \lim_{s\to 0^+} [s\tilde{V}(s)]$, the limiting form for an infinite time horizon is:

$$V_i = \lim_{t\to\infty} V_i(t) = \max_z \left\{ \rho_i^z(\alpha) + \sum_{j=1}^{N} p_{ij}^z \tilde{f}_{ij}^z(\alpha) V_j \right\} \qquad (15')$$

since $\lim_{t\to\infty} \sigma_i(\alpha, t) = \rho_i(\alpha)$ .

Consider for a moment only the stationary policies; that is, ones which do not depend upon the number of steps (or time) since the beginning of the experiment, nor until the end of problem. It follows that the return from every stationary policy in (15) or (15') must satisfy the simultaneous equations $(i = 1, 2, \ldots, N)$

$$V_i = \rho_i(\alpha) + \sum_{j=1}^{N} p_{ij} \tilde{f}_{ij}(\alpha) V_j \qquad (16)$$

where the dependence upon the policy has been suppressed for clarity. In other words, by following the optimal stationary policies in either the n-step or t-horizon formuation, one obtains the same stationary policies, and the same limiting values for the expected total discounted return starting in state i !

We now present an algorithm related to the policy-space iterative technique of Howard[4][6][9] to find the optimal stationary policy for an infinite, discounted Markov renewal program. The flow chart for the algorithm is shown in Figure 1. Basically, the algorithm uses (16) to solve for a set of expected returns following some policy; then, those returns are used to select a better alternative in each state. When two successive

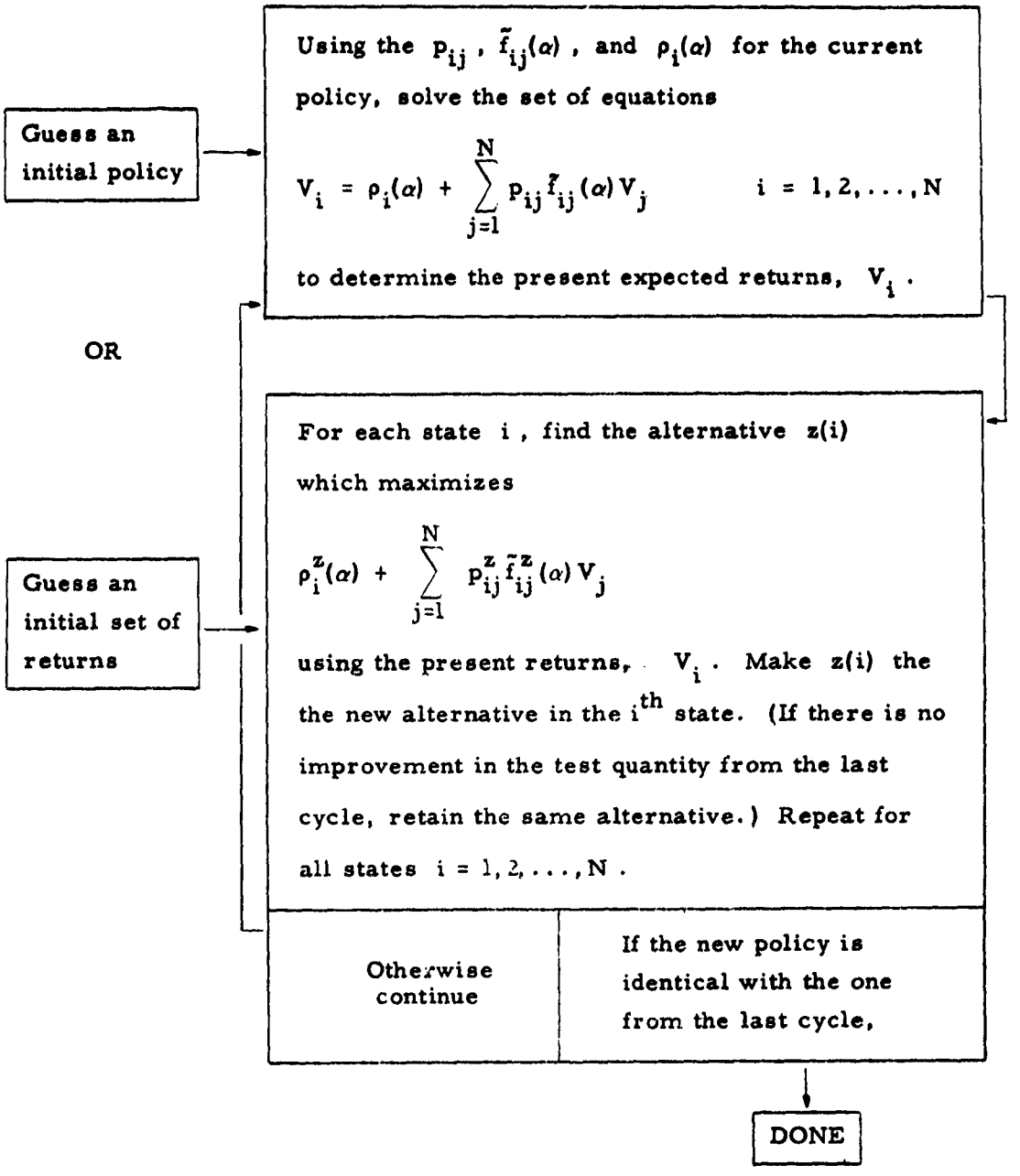



**Figure 1 – Flow chart of algorithm for optimal policy for infinite-horizon, discounted Markov-renewal program.**

policies are found to be identical, the algorithm terminates with an optimal stationary policy, and the maximum expected, discounted returns.

To prove that this algorithm converges, one must show that:

(1) It is always possible to solve the set of simultaneous equations.

(2) The policy-determining step strictly increases the expected return of at least one state in each cycle of the algorithm, if there was an improvement in the test quantity which led to the change in policy.

(3) If two successive policies are identical, then the algorithm has converged on the optimal policy, in the sense that no other policy can lead to higher expected returns for any state $i$.

(4) The algorithm terminates in a finite number of cycles.

The proof of (1) follows from the fact that all elements of the matrix $\tilde{q}(\alpha)$ lie in the interval $[0, 1)$ for all $\alpha > 0$, while the proof of (2) requires the additional observation that the diagonal elements of the matrix $[I - \tilde{q}(\alpha)]^{-1}$ are at least as great as one. The complete proof of this algorithm closely parallels that of Howard for the Markov decision process, and the reader is referred to Chapter 7 of Reference [9] for further details. The fact that there are only a finite number of policies guarantees convergence in a finite number of cycles.

There are also available some special-purpose linear programming algorithms to find the optimal stationary policy in a Markov decision process[7][20][14][1][8] which could equally well be applied to this problem.

It is important to notice that direct enumeration of all stationary policies usually not possible, since there are $Z^N$ of them.

## VIII. The Optimal Policy with Discounting

It is possible to produce a slightly stronger result than that of the last section; namely: among all the optimal policies for the infinite-step or infinite-time discounted case, there exists an optimal, stationary policy. This is important operationally since it may be difficult to follow a non-stationary policy. Our proof closely parallels that of Blackwell for the discounted Markov decision process,[6] and we shall only sketch in the major steps.

For the purposes of this section only, number the policy to be followed in the $i^{th}$ state in terms of the number of steps from beginning of the experiment; thus $z_n(i)$ is the policy to be followed if state $i$ is entered at the $n^{th}$ step from the beginning of the process. Letting $z_n$ be the N-dimensional vector of policies at the $n^{th}$ step, call $\gamma = \{z_1, z_2, z_3, \dots\}$ the sequence of policies to be followed, starting at step one. Finally a sequence of policies $\gamma^*$ is said to be optimal if $V_i(\gamma^*) \geq V_i(\gamma)$ for all $i = 1, 2, \dots, N$, and for all possible sequences of policies $\gamma$.

The essential step in Blackwell's proof lies in the observation that the transformation which maps the returns at the $(n-1)^{st}$ step into the returns at the $n^{th}$ step is monotone. In other words, if $V_j' \geq V_j$ for all $j$, then:

$$\rho_i^z + \sum_{j=1}^{N} \tilde{q}_{ij}^z V_j' \geq \rho_i^z + \sum_{j=1}^{N} \tilde{q}_{ij}^z V_j$$

for all $i$. But this fact follows from the observation that all of the $\tilde{q}_{ij}^z$ are non-negative.

The proof then proceeds in the following steps:

1. If the expected returns following the policy sequence $\{z^*\} = \{z_1^*, z_2^*, z_3^*, \dots\}$ are all greater than or equal to the corresponding returns following the policy sequence $\{f, z_1^*, z_2^*, \dots\}$, for all possible policy vectors $f$, then the sequence of policies $\{z^*\}$ is optimal.

2. If the expected returns from the policy sequence $\{f, z_1, z_2, \dots\}$ are all not less, but at least one is greater than, the corresponding returns from the policy sequence $\{z_1, z_2, z_3, \dots\}$; then, the stationary policy sequence $\{f, f, f, \dots\}$ stands in the same relationship to the policy sequence $\{z_1, z_2, z_3, \dots\}$.

3. Taking any stationary policy, either it is optimal, or there is an improvement possible with another stationary policy. Since there are only finitely many stationary policies, there must be one over which no improvement can be made; hence this one, which is found by the algorithm of Figure 1, must be optimal.

For details, the reader is referred to Section 3 of Reference [6].

A similar result is expected to obtain in the case of the infinite-time, discounted process because of the monotonicity of the transformations in (11), or the quantized equivalent (14), and because it is already been seen that the stationary optimal policies for the infinite-time case give the same expected return as for the infinite-step case.

## IX. The Problem of No Discounting

When the discount factor $\alpha$ approaches zero in either the finite-time or finite-step process, it is seen from (8), and (11) or (14) that no particular difficulties are encountered.

For the finite-step process, for $n = 1, 2, 3, \ldots$, and all $i$:

$$V_i(n) = \max_z \rho_i^z + \sum_{j=1}^{N} p_{ij}^z V_j(n-1) \tag{17}$$

with

$$\rho_i^z = \sum_{j=1}^{N} p_{ij}^z \left( R_{ij}^z + r_{ij}^z \nu_{ij}^z \right) \tag{18}$$

which can be solved through iteration on $n$.

In the finite-time process, for $t \geq 0$, and all $i$:

$$V_i(t) = \max_z \left\{ \sigma_i^z(t) + \sum_{j=1}^{N} p_{ij}^z \int_0^t V_j(t-x)\, dF_{ij}(x) \right\} \tag{19}$$

with

$$\sigma_i^z(t) = \sum_{j=1}^{N} p_{ij}^z \left[ V_{ij}(0) F_{ij}^{c,z}(t) + R_{ij}^z + r_{ij}^z \int_0^t F_{ij}^{c,z}(x)dx \right] \tag{20}$$

which can be resolved through a quantized approximation such as (21).

As the planning horizons approach infinity, the expected total returns without discounting also become infinite, and it is not clear what objective should be set for system optimization. There are three distinct "infinite" objectives which might be posed:

(i) Attempt to find policies in (17) which are optimal for all $n$ sufficiently large.

(ii) Attempt to find policies in (19) which are optimal for all $t$ sufficiently large.

or

(iii) Attempt to find policies in (15) which are optimal for all $\alpha$ sufficiently close to zero.

Unfortunately, there is no a priori reason to assume that the policies which might be found from these three approaches would in any way resemble each other. From Blackwell's investigation of case (iii) for the Markov decision process,[6] it is known that there may be both optimal and near-optimal policies as the discounting vanishes. He also shows that it may be very difficult, computationally speaking, to find all of the optimal or near-optimal policies. Finally, when discounting vanishes, the structure of the underlying Markov chain becomes more important in determining the nature of the limiting results, and this must be taken into consideration.

In order to partially circumvent some of these difficulties, the investigations of the limiting cases in the following sections will have the following additional restrictions:

[i] Only stationary policies will be investigated.

[ii] The Markov-renewal process will be assumed to have a single, finite, underlying Markov chain which is ergodic (irreducible and positive recurrent) for every policy.

[iii] All of the $\nu_{ij}$ are assumed finite.

These assumptions are not too unreasonable for real problem solutions, since a stationary policy is usually desirable for long-term planning -- primarily because of the stability it introduces, but also because of the ease of modification if the input data changes. The elimination of transient and absorbing states, or of multiple-chain structures, also presents no problems, since special extensions can be developed for these cases. If any of the $\nu_{ij}$ are

infinite, the process will tend to get stuck, on the average, in that state; thus, this state behaves as if it were absorbing, and should be separately handled.

The assumption of a finite number of states, and a finite number of alternatives, is very important and cannot be easily eliminated.

## X. Infinite Step, Discounted Case

The approach to be used is to show that (17) gives a limiting form* $V_i(n) \approx Gn + W_i$ for all $i$, and for a certain stationary policy $z(i)$. An algorithm will then be produced which finds the optimal stationary policy, in the sense that it produces a scalar $G$ which is at least as large as that obtained for any other policy.

Assume that we are following some stationary policy, and let $V(n)$ denote the column vector of expected returns at the $n^{th}$ step, $\rho$ the column vector of one-step returns, and $p$ the matrix $(p_{ij})$ of transition probabilities. Equation (17) can be written as:

$$V(n) = \rho + pV(n-1) = [I + p + p^2 + \ldots + p^{n-1}]\rho + V(0) \qquad n = 1, 2, \ldots \qquad (21)$$

where $I$ is the unit matrix. Then,

$$V(n) - V(n-1) = p^{n-1}\rho \quad .$$

But, if the Markov chain is ergodic, $p^{n-1}$ converges or is Cesàro-summable to a probability matrix $\Pi$, each row of which is the same (row) vector $\pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$, whose elements are all positive. In addition $\pi$ is the unique probability vector which satisfies $\pi p = \pi$; that is, it is

*The scalar $G$ should not be confused with the functions $G_{ij}(t)$.

the stationary vector for the ergodic chain.[11] Thus,

$$\lim_{n\to\infty} V(n) - V(n-1) = \Pi\rho \tag{22}$$

or,

$$\lim_{n\to\infty} V_i(n) - V_i(n-1) = \sum_{i=1}^{N} \pi_i \rho_i = G \quad .$$

for all $i$. The scalar $G$ is called the system gain by Howard; notice that for a single ergodic chain, it is independent of the state $i$.

Let $W_i(n) = V_i(n) - Gn$ ; then:

$$W(n) = [I + \rho + \rho^2 + \ldots + \rho^{n-1} - n\Pi]\rho + V(0)$$

$$= \left[ I + \sum_{j=1}^{n-1} (\rho^j - \Pi) - \Pi \right] \rho + V(0) \qquad n = 1, 2, \ldots \tag{23}$$

But, if the underlying chain is ergodic, $\lim_{n\to\infty} [I + \sum_{j=1}^{n-1} (\rho^j - \Pi)]$ converges or is Cesàro-summable to what Kemeny and Snell call the fundamental matrix[11] $Z = (z_{ij})$. By simple manipulations, it is seen that $Z$ is the well-defined inverse $(I - (\rho - \Pi))^{-1}$, and satisfies the relations $\rho Z = Z\rho$, $\pi Z = \pi$, and $I - Z = \Pi - \rho Z$. Thus,

$$W = \lim_{n\to\infty} W(n) = [Z - \Pi]\rho + V(0) \tag{24}$$

or,

$$V_i(n) \approx Gn - G + \sum_{j=1}^{N} z_{ij}\rho_j + V_i(0) \tag{25}$$

for all $i$, where

$$G = \sum_{i=1}^{N} \pi_i \rho_i \quad ; \quad Z = (I - P + \Pi)^{-1} \quad . \tag{26}$$

There is another interpretation of (25) which is of interest. Let $M_{ij,n}$ be the mean number of times the system enters state $j$ in $n$ transitions, if the system was started in state $i$. It is clear that for all $i, j$, and $n = 1, 2, \ldots$

$$M_{ij,n} = \sum_{k=0}^{n-1} p_{ij}^{(k)} = \delta_{ij} + \sum_{k=1}^{N} p_{ik} M_{kj,n-1} \tag{27}$$

so that, if $M_n$ is the matrix $(M_{ij,n})$,

$$V(n) = M_n \rho + V(0) \quad ; \qquad n = 1, 2, \ldots \tag{28}$$

that is, the expected value of a certain state above and beyond its terminal value is equal to the mean number of times that other states of the chain are visited in $n$ steps times the expected one-step rewards in those states, summed over all states which are visited.

Furthermore, it is known that for large $n$,[11] and all $i, j$

$$M_n \approx n\Pi + [Z - \Pi] \tag{29}$$

or

$$M_{ij,n} \approx n\pi_j + [z_{ij} - \pi_j] \quad ,$$

which is an alternate way of deducing (25). Thus the fact that the gain is the same for all initial states is a consequence of the limiting properties of the $M_{ij,n}$ which depends upon convergence to a stationary set of probabilities. The constant term represents the bias due to the initial starting state.

It now is possible to produce an algorithm to find the optimal stationary policy; this algorithm will parallel Figure 1, and the corresponding algorithm for Markov decision processes in Chapter 4, Reference [ 9].

There are some computational simplifications which can be made. At each step in the algorithm, we shall be testing to see which alternative will give the greatest possible increase in return; substituting the asymptotic expression (25) into (21), we obtain after clearing terms,

$$\sum_{j=1}^{N} z_{ij}\rho_j + G = \rho_i + \sum_{j=1}^{N}\sum_{k=1}^{N} p_{ij} z_{jk} \rho_k \qquad (30)$$

for all $i$. There are $N$ simultaneous equations in the $N + 1$ unknowns: $G$, and the $\sum_{j=1}^{N} z_{ij}\rho_j$. Thus, as Howard has pointed out, the unknowns cannot be resolved uniquely, but one can only find the relative values of the $\sum_{j=1}^{N} z_{ij}\rho_j$. These numbers are called the relative values of the policy, and are denoted by $V_i$; it is usual practice to set one of them, say $V_N$, equal to zero, and then solve the equations (30), which are now well-defined. For $i = 1, 2, \ldots, N-1$:

$$V_i + G = \rho_i + \sum_{j=1}^{N-1} p_{ij} V_j \qquad (31)$$

and

$$V_N = 0 \quad ; \quad G = \rho_N + \sum_{j=1}^{N-1} p_{Nj} V_j \quad .$$

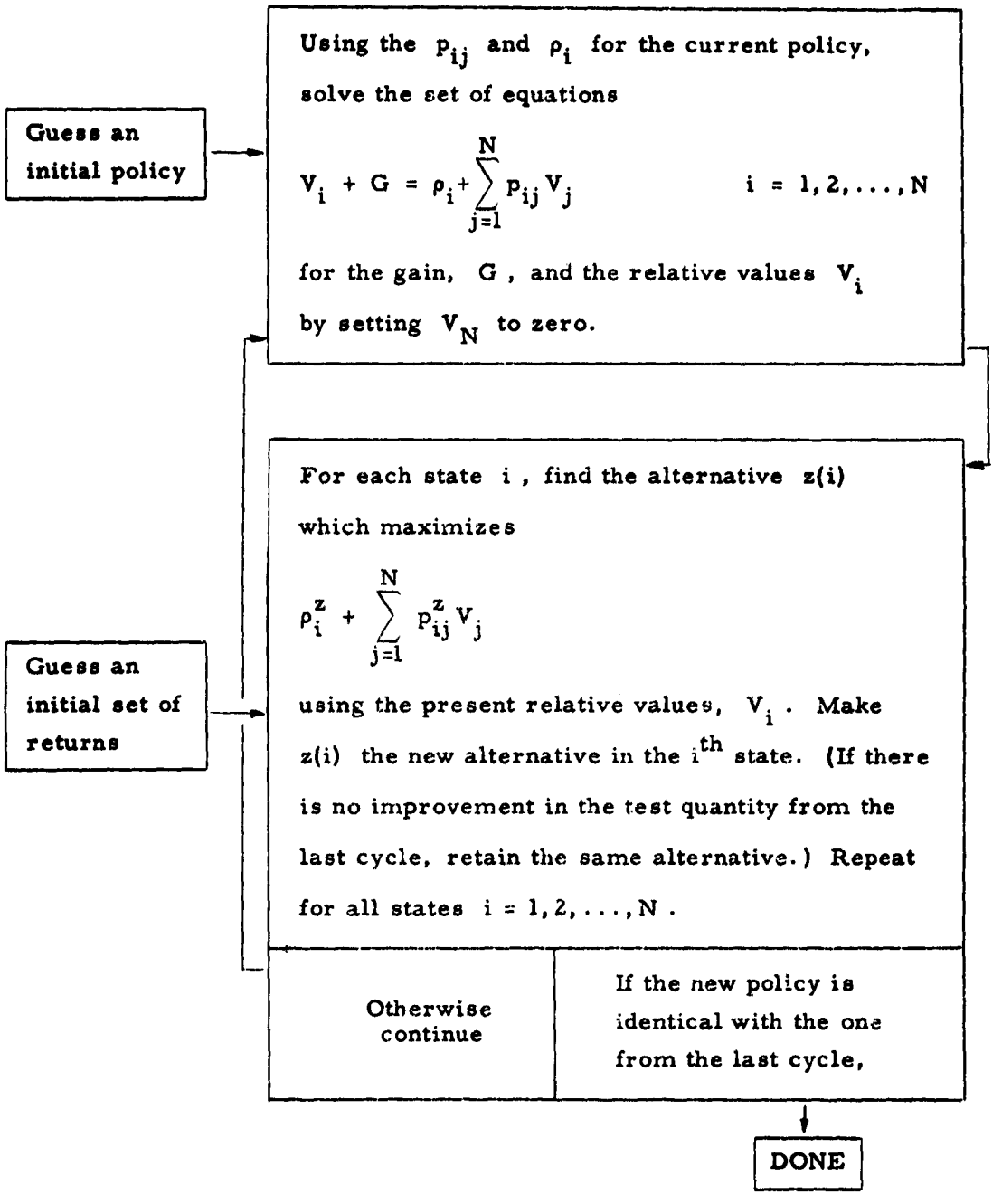



**Figure 2 – Flow chart of algorithm for optimal stationary policy for infinite-step, undiscounted Markov-renewal program.**

The advantage of the equations (31) is that they are much easier to compute at each step of the algorithm, rather than, say finding first $\Pi$ , then $Z$ , and then the test quantities.

The right-hand side of (31) is used as the test quantity for policy improvement in the algorithm shown in Figure 2; the procedure parallels that of Chapter 4, Reference [9], expect for the immediate expected reward, $\rho_i$ , given by (18), which is different, due to the nature of the Markov-renewal process. The same remarks which were made in conjunction with Figure 1 still obtain: the policy-determining step strictly increases the gain $G$ ; the algorithm converges on the optimal policy, when two succeeding policies are identical, and so on. The proof is elementary, and may be found in the above reference.

When the algorithm is terminated, an optimal stationary policy $z^*(i)$ will have been found, together with the maximal gain $G^*$ and relative values $V_i^*$ . At this point, one may calculate the stationary probabilities and the fundamental matrix, and thence find the optimal total return as:

$$V_i^*(n) \approx G^* n - G^* + V_i^* + V_N(0) + \sum_{j=1}^{N} z_{Nj}^* \rho_j \qquad (32)$$

for all $i$ .

## XI. Infinite Time, Undiscounted Case

Because of the close relationship between the infinite-step and infinite-time solutions in the discounted case, it might be expected that this section would be a repeat of the last one; however, one of our main results is that this is not the case. First, it is shown that (19) gives a limiting

form* $V_i(t) \approx gt + w_i$ as $t$ approaches infinity, for all $i$, and for a certain stationary policy, $z(i)$. We shall then produce an algorithm to find the stationary policy which maximizes the rate of increase of the expected value, $g$.

Assume some stationary policy and take the Laplace transform of (19). For all $i$, and for $s \geq 0$:

$$\tilde{V}_i(s) = \tilde{\sigma}_i(s) + \sum_{j=1}^{N} \tilde{q}_{ij}(s)\tilde{V}_j(s) \quad , \tag{33}$$

with

$$\tilde{\sigma}_i(s) = \sum_{i=1}^{N} p_{ij}\left[V_{ij}(0)\,[\,1 - \tilde{f}_{ij}(s)\,]/s + R_{ij}/s + r_{ij}[\,1 - \tilde{f}_{ij}(s)\,]/s^2\right] \tag{34}$$

Denoting the various column vectors by dropping the subscript, and the matrix of $(\tilde{q}_{ij}(s))$ by $\tilde{q}(s)$, Equation (33) becomes

$$\tilde{V}(s) = [\,I - \tilde{q}(s)\,]^{-1}\tilde{\sigma}(s) \quad . \tag{35}$$

The matrix $I - \tilde{q}(s)$ has an inverse for $s > 0$, but as $s$ approaches zero both the inverse and $\tilde{\sigma}(s)$ become ill-defined. This difficulty can be resolved through the use of first-passage time distributions, $G_{ij}(t)$, and the mean-entry counting functions, $M_{ij}(t)$, which are discussed in Appendix A.

*The scalar $g$ should not be confused with the functions $g_{ij}(t)$.

Combining (35) and the transform of (A. 6),

$$\tilde{V}(s) = \tilde{\sigma}(s) + \tilde{m}(s)\tilde{\sigma}(s) \tag{36}$$

or, for $t \geq 0$ , $i = 1, 2, \ldots, N$ :

$$V_i(t) = \sigma_i(t) + \sum_{j=1}^{N} \int_0^t \sigma_j(t - x)\, dM_{ij}(x) \quad . \tag{37}$$

The relationship between the expected return and the mean entry-counting functions in (37) should be compared with corresponding relationship (28) for the infinite-step process.

The $M_{ij}(t)$ are, of course, <u>not</u> the same as their discrete counterparts $M_{ij,n}$ , but are related to the first-passage time distributions $G_{ij}(t)$ through the transform of (A. 5), for $s \geq 0$ , and all $i, j$ :

$$\tilde{m}_{ij}(s) = \tilde{g}_{ij}(s) + \tilde{m}_{jj}(s)\tilde{g}_{ij}(s) \quad . \tag{38}$$

Because of the assumption of finite $r_{ij}$ , it turns out that all of the mean first-passage times, $\mu_{ij}$ , are also finite for an ergodic chain. At this point, it is convenient also to assume that the diagonal second moments, $\mu_{jj}^{(2)}$ are also finite. Finally, it is important to distinguish between two cases; either a $G_{jj}(t)$ is a lattice distribution, or it is not. In the first case, our results (39) and (40) hold only when averaged over the lattice period in question. It can be seen that a sufficient condition for a $G_{jj}(t)$ to be non-lattice distribution is that at least one nonzero $Q_{ij}(t)$ also be a non-lattice distribution, for $i = 1, 2, \ldots, N$ .

With these conditions, one has the interesting result for large $t$, and all $i, j$:[1][18]*

$$M_{ij}(t) \approx \frac{t}{\mu_{jj}} + \frac{\mu_{jj}^{(2)}}{2(\mu_{jj})^2} - \frac{\mu_{ij}}{\mu_{jj}} \quad . \tag{39}$$

This relationship may be found directly from (38), using well-known limit theorems of transform calculus; or, it follows directly from the observation that $M_{jj}(t)$ is the mean renewal (counting) function for the renewal process with inter-event distribution $G_{jj}(t)$, and the use of a theorem by Smith.[19] The limiting form of the mean entry-counting function is thus related only to the first and second moments of the first-passage time distributions.

By using the key renewal theorem[19] or through direct transform arguments, we then argue that (36) has the limit for large $t$, and all $i$:

$$V_i(t) \approx \left( \sum_{j=1}^{N} \frac{\rho_j}{\mu_{jj}} \right) t + \rho_i + \sum_{j=1}^{N} \left\{ \rho_j \left[ \frac{\mu_{jj}^{(2)}}{2(\mu_{jj})^2} - \frac{\mu_{ij}}{\mu_{jj}} \right] - \frac{\eta_j}{\mu_{jj}} \right\} \tag{40}$$

where $\eta_j$ is the area under $\rho_j - \sigma_j(t)$. From (20), this is:

$$\eta_j = \sum_{k=1}^{N} p_{jk} \left\{ -\nu_{jk} V_{jk}(0) + \frac{1}{2} \nu_{jk}^{(2)} r_{jk} \right\} \quad . \tag{41}$$

If the fixed rewards $R_{ij}$ are paid at the end of the interval before transition, instead of at the beginning, then a term $+\nu_{jk} R_{jk}$ must be added under the summation sign in (41).

*The slight discrepancy between (39) and Equation (A.9) of Reference [1] is due to their convention of smoothing out a lattice distribution function into the following period, while we use a symmetric smoothing.

Thus, the limiting form of the expected total return is of the form $V_i(t) \approx gt + w_i$, with the gain rate $g = \Sigma \rho_j/\mu_{jj}$ being the same for all initial states – a consequence of the assumption of ergodicity of the underlying Markov chain.

Appendix A presents formulae suitable for finding the $\mu_{jj}$ and the $\mu_{jj}^{(2)}$; calculation of these moments is not essential, however, to find the optimal stationary policy. Substituting the limiting form in (19)

$$w_i = \rho_i - g \sum_{j=1}^{N} p_{ij}[\nu_{ij} + W_j] + \left\{ \sigma_i(t) - \rho_i + \right.$$

$$\left. + \sum_{j=1}^{N} p_{ij} \left[ g \int_t^{\infty} F_{ij}^c(x) + w_j F_{ij}^c(t) \right] \right\} \qquad (42)$$

for all i. For large t, the terms in braces vanish because of the assumption about finite $\nu_{ij}$, and one obtains the following equations in the N + 1 unknowns, g and the $w_i$:

$$w_i + g\nu_i = \rho_i + \sum_{j=1}^{N} p_{ij} w_j \qquad (43)$$

for $i = 1, 2, \ldots, N$. Comparison with the corresponding infinite-step relations (31) reveals that they are identical except for the coefficient in front of the gain rate, g. Thus, unless all of the $\nu_i$ are identical, and equal, say $\nu_1 = \nu_2 = \ldots \nu_N = \nu$, then the solutions to the Equations (43) will not, in general, be equal to the solutions of (31). We shall return to a discussion of this point in a later section.

Furthermore, from our previous comments about the form of the Equations (43), we know that they cannot be used to solve for the complete $w_i$ given by (40), but, will only find the relative values of $v_i = w_i - w_N$, after setting the $N^{th}$ variable to zero. The following modified equations are used to solve for the relative values of the policy, $v_i$. For $i = 1, 2, \ldots, N-1$:

$$v_i + g\nu_i = \rho_i + \sum_{j=1}^{N-1} p_{ij}v_j \tag{44}$$

and

$$v_N = 0 \quad ; \quad g\nu_N = \rho_N + \sum_{j=1}^{N-1} p_{ij}v_j$$

Our algorithm for finding the optimal stationaly policy is shown in Figure 3. In basic structure, it is identical to the policy approximation algorithms of the previous sections; relative values and a gain rate obtained in a previous cycle are used in a test quantity to find a new policy with increased gain rate; the new policy is used to solve (44) for the new values and gain rate; and so on. The algorithm terminates when no change in policy can be made.

The one new feature is the form of the test quantity; Equations (44) must be divided through by the $\nu_i$ (which are nonzero and finite) to form a test quantity which has the dimensions of reward rate; this seems logical in view of the fact that the algorithm increases $g$ at each step. An alternative test quantity has been proposed by P. Schweitzer (Appendix B).

The proof is still elementary, and parallels those of the previous algorithms; basically one shows that $g^{z_2} > g^{z_1}$, if at some cycle the test

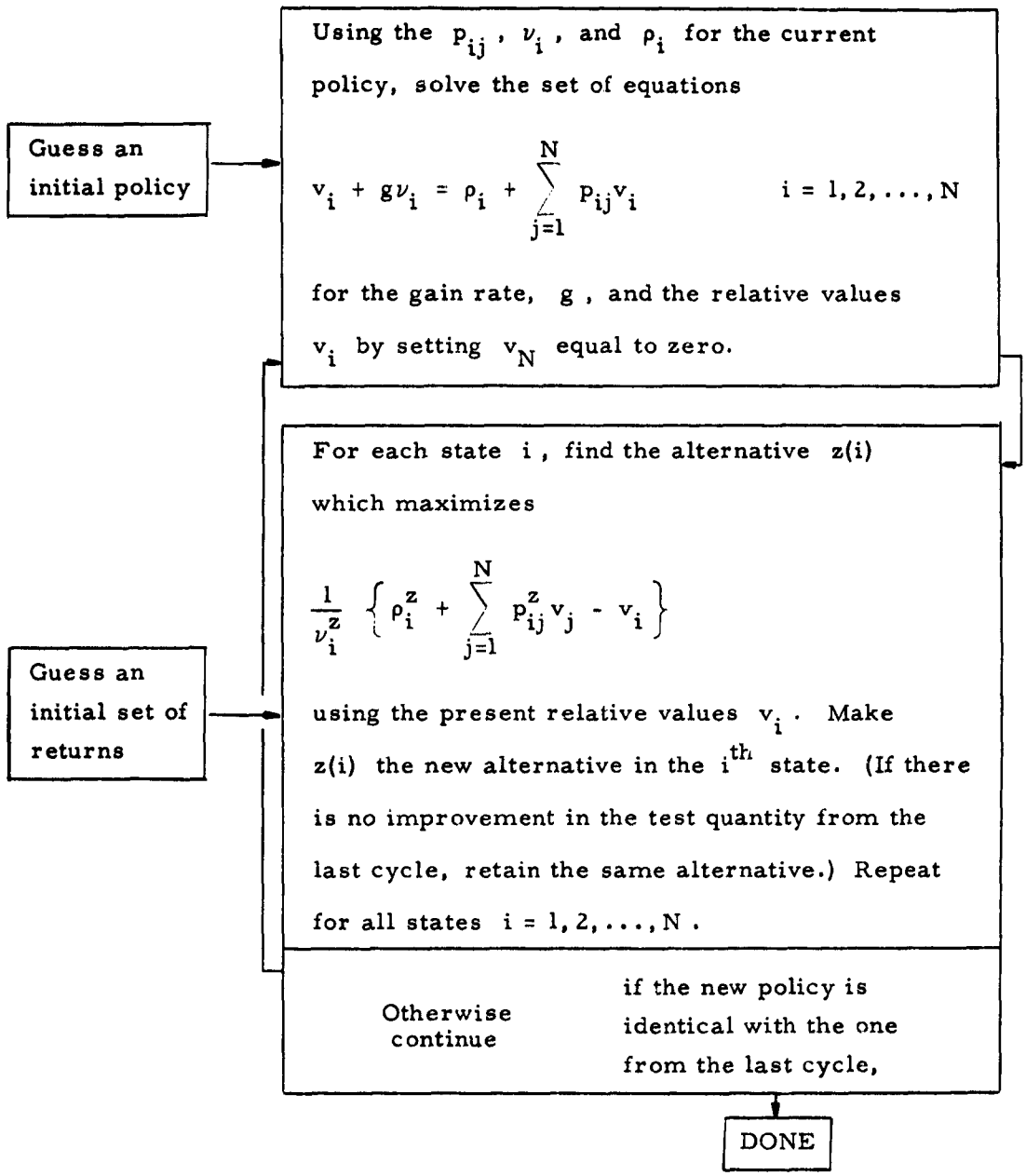



Figure 3 – Flow chart of algorithm for optimal stationary policy for the infinite-time, undiscounted Markov-renewal program, and for the infinite-time or -step, vanishing discount Markov-renewal program.

quantity indicated a change to policy $z_2$ using the relative values and gain rate of policy $z_1$. When the algorithm is terminated, an optimal stationary policy $z^*(i)$ will have been found, together with the maximal gain rate $g^*$, and relative values $v_i^*$. At this point, one may calculate the stationary probabilities, $\pi_i$ in the usual manner, and the $\mu_{ij}$ and $\mu_{jj}^{(2)}$ from Equations (A.8), (A.9) or (A.11). The optimal expected return is:

$$V_i(t) \approx g^* t + v_i^* + \rho_N + \sum_{j=1}^{N} \left\{ \rho_j \left[ \frac{\mu_{jj}^{(2)}}{2(\mu_{jj})^2} - \frac{\mu_{Nj}}{\mu_{jj}} \right] - \frac{\eta_j}{\mu_{jj}} \right\} \qquad (45)$$

for all $i$.

## XII. The Vanishing Discount Case

The final limiting case to be investigated is that of the infinite-step or infinite-time process whose discount factor, $\alpha$, vanishes. Thus we seek the limiting form for all $i$, with $\alpha$ approaching zero, of:

$$V_i = \rho_i(\alpha) + \sum_{j=1}^{N} p_{ij} \bar{f}_{ij}(\alpha) V_j \qquad (16)$$

with

$$\rho_i(\alpha) = \sum_{j=1}^{N} p_{ij} \left\{ R_{ij} + r_{ij} [1 - \bar{f}_{ij}(\alpha)] / \alpha \right\} \qquad (6)$$

A review of the steps encountered in finding the asymptotic form of the transform of (40) in the last section will indicate the necessary parallel between (33) and (16). It is easy to show that as $\alpha$ approaches zero, for all $i$:

$$V_i(\alpha) = \gamma/\alpha + \rho_i + \sum_{j=1}^{N} \left\{ \rho_j \left[ \frac{\mu_{jj}^{(2)}}{2(\mu_{jj})^2} - \frac{\mu_{ij}}{\mu_{jj}} \right] - \frac{\eta_j}{\mu_{jj}} \right\} + 0(\alpha) \quad (46)$$

with

$$\gamma = \sum_{j=1}^{N} \rho_j/\mu_{jj} \quad ; \quad \eta_j = \sum_{j=1}^{N} p_{jk} \left[ + \frac{1}{2} r_{jk} \nu_{jk}^{(2)} \right] \quad (47)$$

If the fixed rewards $R_{ij}$ are paid at the end of the transition interval, then a term $+R_{jk}\nu_{jk}$ must be added under the summation sign defining the $\eta_j$.

Thus, rather suprisingly, the criterion for optimization turns out again to be the gain rate, $\gamma = g$. The algorithm to be used is a repeat of that shown in Figure 3; even the same relative values $v_i$ being obtained. The one slight difference in this case is that the terminal rewards no longer enter into the calculation of the $w_i$; however, this does not affect the optimization, but merely the final expected return.

## XIII. Ties and Near-Optimal Policies

In each of the three limiting cases just discussed, the criterion for optimization has been the dominant term in the limiting form of the expected total return, either Gn, gt, or $\gamma/\alpha$. However, it may happen that when the algorithms of Figures 2 and 3 are carried out, there will be more than one optimal stationary policy – each with the same gain, or gain rate!

Blackwell[6] has considered this problem in detail for vanishing discount in Markov decision processes. It is shown that when the algorithm terminates with a single policy $z^*$, and when the test quantity is strictly less than $V_i^{z^*} + G^{z^*}$ <u>for all other alternatives, for each i</u>, then $z^*$ is optimal, in the sense that no other policy leads to a higher value for $V_i^z(\alpha)$,

for all $\alpha$ sufficiently close to zero. If a truely optimal policy cannot be found from this algorithm, Blackwell points out that there may be nearly-optimal policies, i.e., policies whose return converges to the return from an optimal policy, for $\alpha \to 0$. The near-optimal policies are just those for which both the gains (or gain rates) and the constant terms, $W_i$ or $w_i$, are comparable. The determination of all the near-optimal policies appears to be an arduous task, in general, in the event that there are ties, since the relative values, $V_i$ and $v_i$, are not sufficient for absolute comparisons.

Direct evaluation of all of the nearly- optimal policies may be feasible for small problems but is probably prohibitive for a "reasonable" real problem. On the other hand, one might claim that a real problem which gave two policies with the same gain rate had insufficient or inaccurate data! Nevertheless, the problem of ties is still an interesting and unresolved question, computationally.

## XIV. The Difference Between Limiting Cases

A disturbing feature of the three cases of limiting programs:

1. $\alpha = 0 \; ; \; n \to \infty$
2. $\alpha = 0 \; ; \; t \to \infty$
3. $n, t = \infty \; ; \; \alpha \to 0$

is the fact that they may give different optimal stationary policies. As we have seen, in the first case the algorithm maximizes the per-transition gain, $G = \sum_{i=1}^{N} \pi_i \rho_i$ ; while in the second and third cases the algorithm maximizes the gain rate, $g = \mathscr{g} = \sum_{i=1}^{N} \rho_i / \mu_{ii}$ .

A rationale for the equivalence of the second and third cases is that discounting may be interpreted as an experiment in which in each dt there is a probability $\alpha dt$ of entering an absorbing state, i.e., discontinuing the experiment. Thus, the behavior will reflect that of the time-horizon process,

rather than a transition – horizon process.

The difference between the first two cases is due to both the reward structure, and to the possible different sojurn times between transitions. For instance, consider a one-state periodic process of period $\nu$, in which a reward $R$ is given at every "transition" from the state back into itself. In order to select the maximum reward per transition, one would select the policy with largest $R$, no matter how large $n$ was. But in order to maximize reward rate for a large interval of time, one would select the policy with largest $R/\nu$!

The two criteria can be contrasted in more generality through (A.10).

$$G = \sum_{i=1}^{N} \pi_i \rho_i \quad \text{versus} \quad g = \frac{\sum_{i=1}^{N} \pi_i \rho_i}{\sum_{i=1}^{N} \pi_k \nu_k} \tag{48}$$

The gain rate is influenced by the mean sojurn times in the states of the system, weighted by the stationary probabilities of making a transition to those states – and this may change with a change in policy, even when $G$ does not.

In fact, one must make a distinction between <u>two</u> sets of stationary probabilities in Markov-renewal processes. The $\pi_i$ are the stationary probabilities which are the limiting values of being in state $i$ after $n$ transitions, as $n \to \infty$. There are also stationary probabilities, $P_i$, which are the limiting values of the probability of being in state $i$ at time $t$, as $t \to \infty$.[16] For an underlying ergodic chain. the two sets of probabilities are related by:

$$P_i = \frac{\nu_i}{\sum_{k=1}^{N} \pi_k \nu_k} \pi_i \qquad (49)$$

for $i = 1, 2, \ldots, N$. Thus, within the basic theory of Markov-renewal processes, there is a fundamental distinction between behavior of state probabilities from transition to transition, and behavior over time. This distinction is also well-known in the study of queueing problems in which the method of regeneration points is used.

We are thus forced to the conclusion that when considering Markov-renewal programs with an infinite horizon, one must decide whether the system will be operated for an infinite number of transitions, or for an infinite period of time!

A pertinent question is, when will these two experiments converge on the same stationary policy? From (48), a sufficient condition is that $\Sigma \pi_k \nu_k$ be independent of the possible policies; this would certainly be true, if for every policy, and for every pair of states, the $\nu_{ij}$ were all equal to $\nu$. Thus, for a Markov decision process, $G = g\nu$, independent of the policy under consideration. Of course, for some Markov-renewal programs, the policies may be identical because of the data of the problem.

## XV. Multiple Chain and Transient-State Problems

It is possible to extend the analyses of the previous sections to problems where the underlying Markov chain has several recurrent classes or where some of the states are transient or absorbing.

The problem of multiple chains has been discussed extensively by Howard.[9] The primary change in the algorithms of Figures 2 and 3 are the determination of a separate gain (or gain rate) for each class of recurrent

states. The policy determination routine selects alternatives in order to maximize the average gain "reached" from state $i$ ; if there are ties, a test quantity including the relative values is used. It should be emphasized that the procedure does not break ties between the optimal policies which maximize gain within each chain, but merely break ties within the algorithm which leads to one of these policies. For more details, the reader is referred to Reference [9]; the necessary changes for Markov-renewal programs can easily be deduced.

If there are transient states in the underlying Markov chain, then it is a simple matter to determine the expected number of steps until a given state in one of the recurrent chains is enteres.[11] The average return accumulated en route to absorption in the recurrent chain is then added to the return of the recurrent state (s) entered, in the obvious manner; further details are left to the reader. Absorbing states, for our purposes, may be treated like a one-state recurrent chain.

Once again it is emphasized that these special considerations relative to the underlying Markov chain are necessary only in limiting programs; when discounting is present, or when the process has a finite horizon, there is no difficulty.

## XVI. Limiting Results

The analyses presented here may be easily specialized to the results of Markov-decision processes. As an example, we specialize the results of the section on discounted programs to the continuous-parameter Markov process analyzed by Howard in Chapter 8, Reference [9].

For a continuous-parameter Markov process, $p_{ii} = 0$ , and $F_{ij}(t) = 1 - e^{-\omega_i t}$ , for all $i, j$ , and appropriate finite $\omega_i > 0$ .

Thus, $\tilde{f}_{ij}(\alpha) = \omega_i/(\alpha + \omega_i)$, and from (6):

$$\rho_i(\alpha) = \sum_{\substack{j=1 \\ j\neq i}}^{N} p_{ij} \left\{ \frac{\omega_i}{\alpha + \omega_i} R_{ij} + \frac{1}{\alpha + \omega_i} r_{ij} \right\} \tag{50}$$

for all $i$, when the rewards $R_{ij}$ are earned at the end of the transition interval. Making the running rewards only dependent upon $i$, $r_{ij} = r_i$, the expected return (16) must satisfy:

$$(\alpha + \omega_i)V_i = r_i + \omega_i \sum_{\substack{j=1 \\ j\neq i}}^{N} p_{ij} R_{ij} + \omega_i \sum_{\substack{j=1 \\ j\neq i}}^{N} p_{ij} V_j \tag{51}$$

for all $i$. Setting:

$$a_{ij} = \omega_i p_{ij} \qquad i \neq j$$

$$a_{ii} = -\omega_i$$

we finally obtain, for $i = 1, 2, \ldots, N$:

$$\alpha V_i = r_i + \sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij} R_{ij} + \sum_{j=1}^{N} a_{ij} V_j \tag{52}$$

which is essentially Equation (8.47) of Reference [7].

## XVII. The Two-State Processes

As a specific case of interest, we present explicit results for Markov-renewal programs with only two states. First, express the transition probabilities in terms of the off-diagonal elements:

$$P = \begin{bmatrix} 1 - p_{12} & p_{12} \\ p_{21} & 1 - p_{21} \end{bmatrix} , \tag{53}$$

which represents an ergodic chain for $p_{12}$ and $p_{21}$ greater than zero. The stationary transition probabilities for the Markov chain are:

$$\pi = (p_{12} + p_{21})^{-1} \{p_{21} , p_{12}\} \tag{54}$$

which, when used to calculate the fundamental matrix, gives:

$$Z - \Pi = (p_{12} + p_{21})^{-2} \begin{bmatrix} p_{12} & -p_{12} \\ -p_{21} & p_{21} \end{bmatrix} . \tag{55}$$

Similarly, the stationary probabilities over time for the Markov-renewal process are:

$$P = (\nu_1 p_{21} + \nu_2 p_{12})^{-1} \{\nu_1 p_{21} , \nu_2 p_{12}\} . \tag{56}$$

The mean first-passage times are:

$$(\mu_{ij}) = \begin{bmatrix} \dfrac{\nu_1 p_{21} + \nu_2 p_{12}}{p_{21}} & \dfrac{\nu_1}{p_{12}} \\ \dfrac{\nu_2}{p_{21}} & \dfrac{\nu_1 p_{21} + \nu_2 p_{12}}{p_{12}} \end{bmatrix} \tag{57}$$

while the diagonal terms of the second moment of first-passage times are:

$$\mu_{11}^{(2)} = \nu_1^{(2)} + 2p_{12}\nu_{12}\mu_{21} + \frac{p_{12}}{p_{21}}\left(\nu_2^{(2)} + 2p_{22}\nu_{22}\mu_{21}\right)$$
$$\mu_{22}^{(2)} = \nu_2^{(2)} + 2p_{21}\nu_{21}\mu_{12} + \frac{p_{21}}{p_{12}}\left(\nu_1^{(2)} + 2p_{11}\nu_{11}\mu_{12}\right) \qquad (58)$$

The limiting value of the per-transition gain is then

$$G = \frac{p_{21}\rho_1 + p_{12}\rho_2}{p_{21} + p_{12}} \quad ; \qquad (59)$$

while the infinite-step relative values from (31) are just

$$v_1 = \frac{\rho_1 - \rho_2}{p_{21} + p_{12}} \quad ; \quad v_2 = 0 \quad . \qquad (60)$$

The limiting value of the gain rate is:

$$g = \mathcal{g} = \frac{p_{21}\rho_1 + p_{12}\rho_2}{\nu_1 p_{21} + \nu_2 p_{12}} \qquad (61)$$

with the infinite-time or vanishing-discount relative values from (44):

$$v_1 = \frac{\nu_2\rho_1 - \nu_1\rho_2}{\nu_1 p_{21} + \nu_2 p_{12}} \quad ; \quad v_2 = 0 \qquad (62)$$

From these explicit formulae, the optimal policy can be found by direct evaluation, if the number of alternatives is not too large. The exact constant terms in each case can be then found from the relative values and (32), or (45), or (46)(47).

## XVIII. An Example

As an illustration, consider the two-state problem of a machine that is running (state 1) or has broken down (state 2). If the machine is running there are two maintenance alternatives:

Alternative A: $R_{12} = 0$ ; $r_{12} = \$100/\text{day}$ ; $\nu_{12} = 4$ days ; $p_{12} = 1$ .

Alternative B: $R_{12} = 0$ ; $r_{12} = \$\ 84/\text{day}$ ; $\nu_{12} = 5$ days ; $p_{12} = 1$ .

If the machine has broken down, there are two repair alternatives:

Alternative A: $R_{21} = 0$ ; $r_{21} = -\$65/\text{day}$ ; $\nu_{21} = 4$ days ; $p_{21} = 1$ .

Alternative B: $R_{21} = -\$100$ ; $r_{21} = -\$200/\text{day}$ ; $\nu_{21} = 1$ day ; $p_{12} = 1$ .

Alternative B may be thought of as an outside repairman whose expensive fixed and running charges are compensated for by his quick service time.

The finite- and infinite-step processes without discounting are independent of the complete transition time distributions, and depend only on the means, $\nu_{ij}$ . By direct evaluation:

$$G^{AA} = \$70 \quad ; \quad G^{AB} = \$50 \quad ; \quad G^{BA} = \$80 \quad ; \quad G^{BB} = \$60 \quad ,$$

so that (B, A) (expensive maintenance, cheap repair) is the optimal stationary policy for a large number of transitions. It is also the optimal policy for all values of n . Figure 4 shows total expected return as a function of n , starting in either state 1 or 2. The exact expressions are:

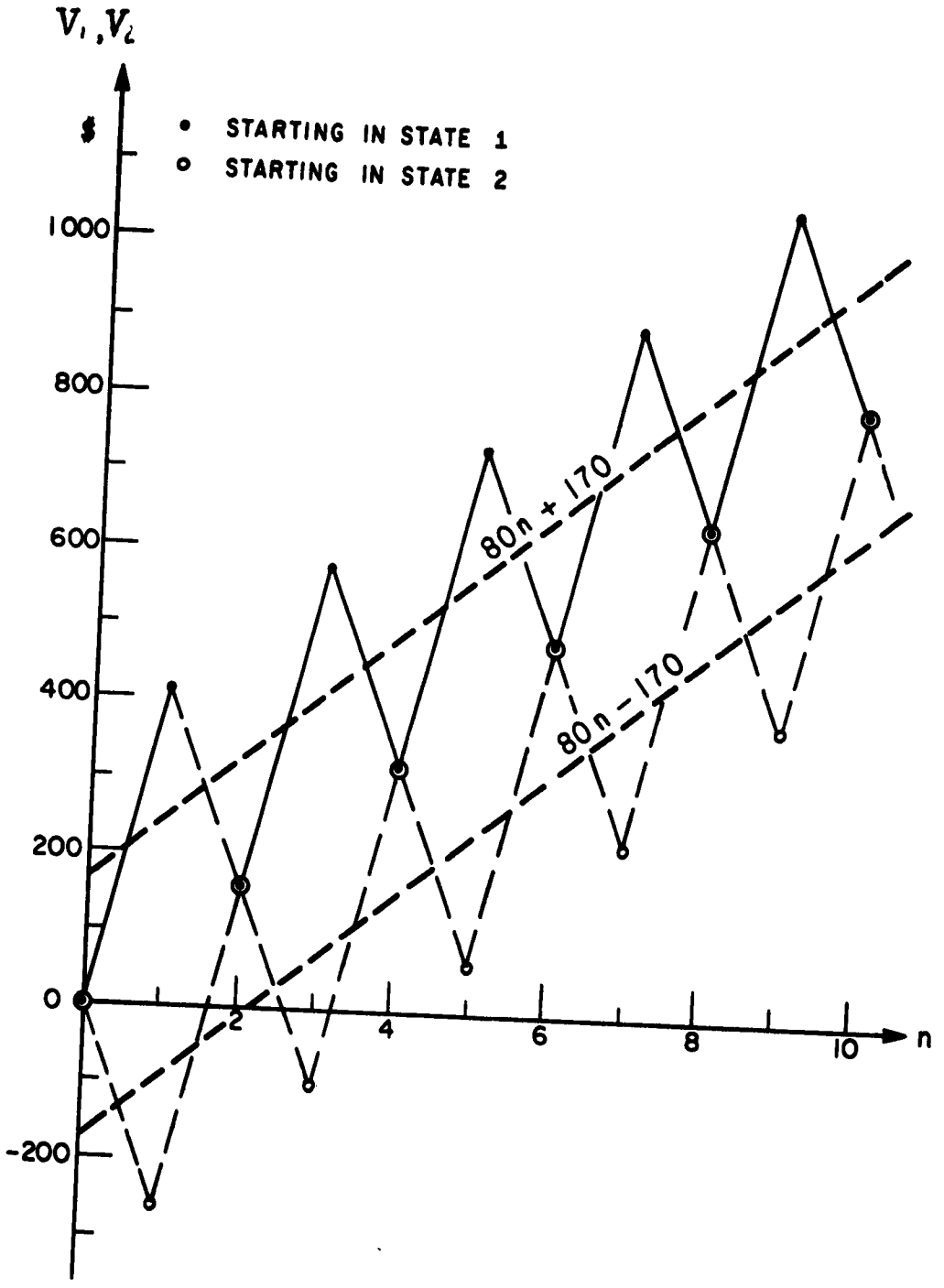



Figure 4 -- Optimal Expected Return versus Number of Transitions Remaining for Finite-Step Machine Maintenance and Repair Example.

$$V_1(n) = 80n + 170 + (-)^{n+1}170$$

$$V_2(n) = 80n - 170 - (-)^{n+1}170 \quad .$$

When these fluctuations are averaged out, the last term vanishes in each equation, so that the limiting form has $G = \$80$, and $W_1 = \$170$, $W_2 = -\$170$. These values may be checked from (32); (31) will, of course, only produce $V_1 = \$340$, $V_2 = 0$.

In the finite- and infinite-time processes without discounting, one must specify the form of the transition-time distributions; let us suppose that they are all degenerate, with the given means. Figure 5 shows, in heavy lines, the return obtained when following the optimal (nonstationary) policy for all $t$. The optimal policy itself is indicated by means of solid bars above and below the return curves; notice that by $t = 19\ 44/149$, the optimal policy in state 2 has stabilized to policy B, but that the return curves and the policy in state 1 do not stabilize until $t = 30$. At this point, the optimal policy in state 1 is to use either A or B, and the return curves have the form:

$$V_1(t) = 20t + 301\frac{111}{149} + u_1(t)$$

$$V_2(t) = 20t - 18\frac{38}{149} + u_2(t)$$

for $t \geq 30$. $u_1(t)$ and $u_2(t)$ are sawtooth curves of period one, whose time-average value is zero.

By direct evaluation,

$$g^{AA} = 17.50 \quad ; \quad g^{AB} = 20.0 \quad ; \quad g^{BA} = 17.77 \quad ; \quad g^{BB} = 20.0$$

in dollars per day,

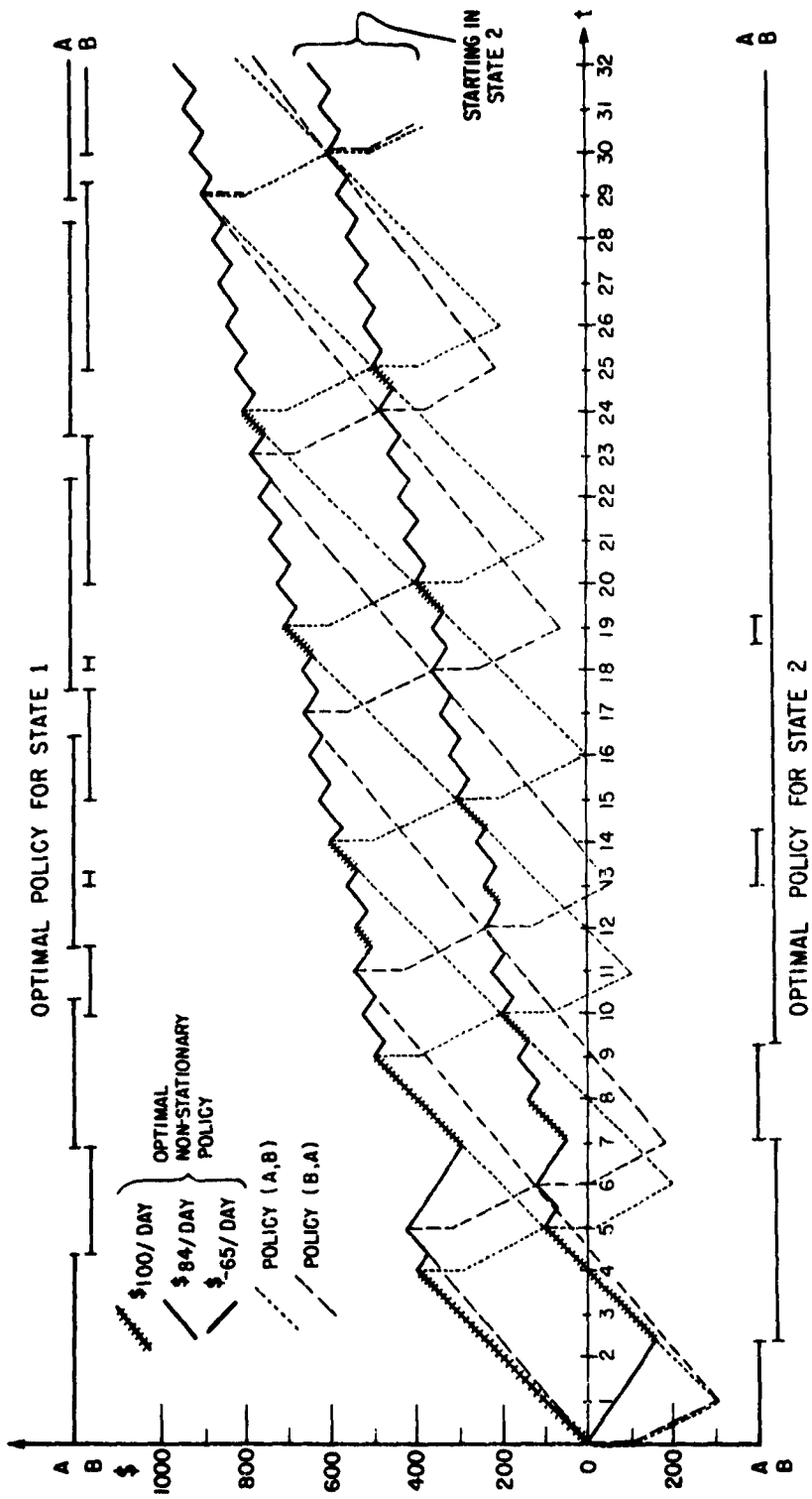



**Figure 5 — Total returns versus time remaining following the optimal non-stationary policy and two equivalent-gain-rate stationary policies for finite-time machine maintenance and repair example.**

so that either (A, B) or (B, B) are the maximal gain rate stationary policies, indicating that expensive repair should always be used! For comparison, Figure 5 also shown the total returns obtained when the stationary policies are followed for all time. For policy (A, B) (dotted line):

$$V_1(t) = 20t + 150 + u_3(t)$$
$$V_2(t) = 20t - 170 + u_4(t)$$

and for policy (B, B) (dashed line):

$$V_1(t) = 20t + 151\frac{2}{3} + u_5(t)$$
$$V_2(t) = 20t - 168\frac{1}{3} + u_6(t)$$

for all $t \geq 0$, where $u_3(t)$, $u_4(t)$, $u_5(t)$, and $u_6(t)$ are all sawtooth curves of period five (A, B), or six (B, B), whose time-average values are zero. The results can be obtained graphically, or from (40) and (41). A surprising result is that while both (A, B) and (B, B) are the limiting stationary policies of the optimal nonstationary policy, only (B, B) is the optimal, completely stationary policy, in the sense of maximizing the $w_i$, when there is a tie in the gain rates, $g$. This resolution of the tie cannot be found from (44), which gives relative values $v_1 = 320$; $v_2 = 0$, for both policies.

To illustrate the effect of the distribution shape upon the optimal policy, consider the infinite-horizon problem with discounting for the following distributions:

I. All distributions degenerate
II. $F_{12}^A$ and $F_{21}^A$ exponential, $F_{12}^B$ and $F_{21}^B$ degenerate.
III. $F_{12}^A$, $F_{21}^A$, $F_{12}^B$ exponential, $F_{21}^B$ degenerate.

In each case, the means are maintained at the values previously given. The resulting normalized optimal discounted returns, $\alpha V_1$ and $\alpha V_2$ are shown in Figure 6 versus $\alpha$ ; changes in policy for different regions of $\alpha$ are indicated by a vertical bar. Notice that any of the four possible policies may be optimal, depending upon the discount factor and the distribution assumed. As $\alpha \to 0$ , either AB or BB is selected as the optimal policy, but there are no near-optimal policies in the cases shown, even though they all tie in gain rate at the limit.

## XIX. Summary

Besides the extensions of the model to different classes of underlying Markov chains which have already been mentioned, there are other modifications which can be made:

1. A general reward structure, $R_{ij}(t, \tau_{ij})$ $0 \leq t \leq \tau_{ij}$
2. Termination rewards which depend upon the time until the next transition.
3. "Mixed" horizons, in which the process is terminated at the next transition after time $T$ , or at min $(n, T)$, etc.

These modifications do not change the solution algorithms in any substantial way.

A more difficult problem is the resolution of ties in infinite-horizon problems. Since both the gains and the relative values must be used in the algorithm to find a new policy which improves the gain, it would appear that resolution of ties would require knowledge about the transient part of the expected total reward. Another approach would be to use a secondary criterion, such as minimum variance, to resolve the ties.

In summary, we have considered an extension of previous work in Markov-decision processes into models which have Markov-renewal

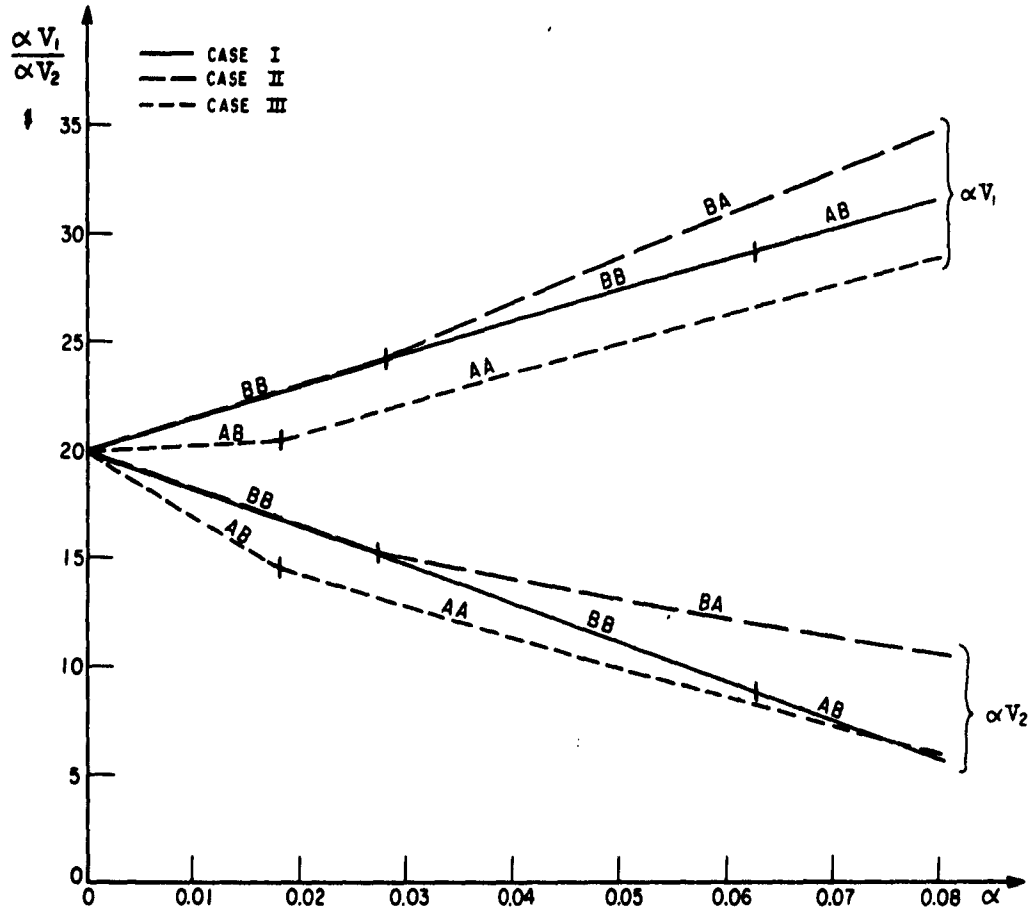



**Figure 6 — Normalized expected discount return versus discount factor for machine maintenance and repair example, showing the effect of different transition-time distributions.**

structures. It is felt that these models embrace a wider class of important operational problems, since in a Markov-renewal process the times between transitions can follow a random clock which depends on both the previous and the next state of the system. The policy-space algorithm remains much the same as in the Markov models; but now a fundamental distinction appears in infinite programs: are they infinite in time, or in number of transitions? Clarification of this distinction appear to be a fundamental part of Markov-renewal programs, and it will be of interest to see how this distinction will be reflected in application.

## APPENDIX A

Here are presented some results on Markov-renewal processes due to Pyke[15][16][17][18] and Barlow[1] which are used in the text.

The basic function is the joint conditional probability distribution:

$$Q_{ij}(t) = p_{ij} F_{ij}(t) = \Pr\{i_{k+1} = j \ , \ \tau(i,j) \leq t \,|\, i_k = i\} \qquad (A.1)$$

which is defined for all $i$ and $j$, $t \geq 0$, and $k = 0, 1, 2, \ldots$

Let $u$ represent the first passage time to state $j$, starting at state $i$; from the definition of the transition times:

$$\begin{aligned} u_{ij} &= \tau(i,j) \\ &\quad \text{or } \tau(i,k) + \tau(k,j) && (k \neq j) \\ &\quad \text{or } \tau(i,k) + \tau(k,\ell) + \tau(\ell,j) && (k, \ell \neq j) \end{aligned}$$

where each sequence is determined by the underlying $p_{ij}$. With this definition the first passage time from $i$ back to itself does not require the system to enter another state first. If the distribution function of first-passage times is denoted by:

$$G_{ij}(t) = \Pr\{u_{ij} \leq t\} \qquad \begin{matrix}(t \geq 0) \\ (i, j = 1, 2, \ldots, N)\end{matrix} \qquad (A.2)$$

then a simple renewal argument will give the following relationship between the $G_{ij}$ and the $Q_{ij}$:

$$G_{ij}(t) = Q_{ij}(t) + \sum_{\substack{k=1 \\ k \neq j}}^{N} \int_0^t G_{kj}(t - x)\, dQ_{ij}(x) \quad \begin{matrix}(t \geq 0) \\ (i, j = 1, 2 \ldots, N)\end{matrix} \qquad (A.3)$$

$G_{ij}(0) = 0$ because of the corresponding restriction on the $F_{ij}(0)$. The first and second moments of the first-passage times will be denoted by $\mu_{ij}$ and $\mu_{ij}^{(2)}$, respectively; these need not be finite, in general.

A variable of interest in Markov-renewal processes in $N_j(t)$, the number of times the system enters state $j$ in the interval $(0, t]$. In particular, the mean number of entries into state $j$ in the interval is defined as:

$$M_{ij}(t) = E\{N_j(t) \mid i_0 = i\} \qquad \begin{matrix}(t \geq 0)\\(i, j = 1, 2, \ldots, N)\end{matrix} \qquad (A.4)$$

From the definition of first-passage times:

$$M_{ij}(t) = G_{ij}(t) + \int_0^t G_{ij}(t - x)\, dM_{jj}(x) \,. \qquad \begin{matrix}(t \geq 0)\\(i, j = 1, 2, \ldots, N)\end{matrix} \qquad (A.5)$$

Finally, there is an interesting relationship between the $M_{ij}$ and the $Q_{ij}$:

$$M_{ij}(t) = Q_{ij}(t) + \sum_{k=1}^{N} \int_0^t Q_{ik}(t - x) dM_{kj}(x) \qquad \begin{matrix}(t \geq 0)\\(i, j = 1, 2, \ldots, N)\end{matrix} \qquad (A.6)$$

In particular, note that for a one-state process, $G(t) = Q(t) = F(t)$, and (A.5) and (A.6) reduce to the well-known equation from renewal theory[10][19]

$$M(t) = F(t) + \int_0^t F(t - x)\, dM(x) \,. \qquad (t \geq 0) \qquad (A.7)$$

It is this intimate relation with both Markov processes and renewal processes which led Pyke to define these processes as Markov-renewal. Asymptotic properties of the $M_{ij}(t)$ used in Equation (39) may be found in References [16] and [18].

Methods for the solution of the first-passage moments are needed in Equations (40) and (45) of the main text. From (A.3), or from Reference [1]:*

$$\mu_{ij} = \sum_{\substack{k=1 \\ k \neq j}}^{N} p_{ik}\mu_{kj} + \nu_i \quad ; \quad \nu_i = \sum_{k=1}^{N} p_{ik}\nu_{ik} \tag{A.8}$$

for all $i, j$. Also,

$$\mu_{ij}^{(2)} = \sum_{\substack{k=1 \\ k \neq j}}^{N} p_{ik}\{\mu_{kj}^{(2)} + 2\nu_{ik}\mu_{kj}\} + \nu_i^{(2)} \quad ; \quad \nu_i^{(2)} = \sum_{k=1}^{N} p_{ik}\nu_{ik}^{(2)} \tag{A.9}$$

for all $i, j$.

Both means are finite if the first and second means of the $F_{ij}(t)$ are finite, and the underlying chain is ergodic. In this case, the equations above are always well-defined.

If Equations (A.8) are multiplied by the stationary probabilities, $\pi_i$, and summed, there results the interesting relationship for all $j$:

$$\mu_{jj} = \frac{1}{\pi_j} \sum_{k=1}^{N} \pi_k \nu_k \quad . \tag{A.10}$$

*There are several typographic errors in [1] on pp. 53, 54. The equation before their (A.3) should have the indices of $\nu_{jk}^{(2)}$ reversed; the condition on the second summation in (A.4) should read $k \neq i$; and the equation after (A.4) should have $\pi_i$ as denominator of both terms on the right-hand side.

In a similar manner, from (A.9), for all $j$ :

$$\mu_{jj}^{(2)} = \frac{1}{\pi_j}\left\{\sum_{k=1}^{N} \pi_k \nu_k^{(2)} + 2 \sum_{\substack{k=1 \\ k\neq j}}^{N} \sum_{i=1}^{N} \pi_i p_{ik} \nu_{ik} \mu_{kj}\right\} . \qquad (A.11)$$

When $\nu_i = \nu_i^{(2)} = 1$ , these formulae reduce to well-known relationships for Markov chains.[11]

## APPENDIX B

In an unpublished report by P. Schweitzer of M.I.T. (private communication; March 13, 1963), an alternative test criterion has been proposed for a quantized version of the infinite time, undiscounted case. His criterion is to select, for each state $i$ , the alternative $z(i)$ which maximizes

$$\rho_i^z + \sum_{j=1}^{N} p_{ij}^z v_j - g\nu_i^z \qquad (B.1)$$

which may be contrasted with out test criterion of Figure 3:

$$\frac{1}{\nu_i^z}\left\{\rho_i^z + \sum_{j=1}^{N} p_{ij}^z v_j - v_i\right\} \quad .* \qquad (B.2)$$

Suppose one had a policy (vector) A which led to an improved policy B , and let $\Gamma_i \geq 0$ be the improvement in Schweitzer's test criterion for state $i$ and let $\gamma_i \geq 0$ be the improvement in our test criterion. Then if $\Delta g = g^B - g^A$ is the improvement in gain rate between policies A and B , it can be shown that:

$$\Delta g = \sum_{j=1}^{N} \frac{\Gamma_j}{\nu_j^B} P_j^B \quad \text{and} \quad \Delta g = \sum_{j=1}^{N} \gamma_j P_j^B \qquad (B.3)$$

*In both cases the relative values $v_i$ and the gain rate $g$ of the present policy are used.

Since the underlying Markov chain is ergodic, and all of the $\nu_j$ are assumed finite, both criteria lead to an improved policy. It is not known if one of them is computationaly more efficient than the other.

## REFERENCES


[1]. R.E. Barlow, "Applications of Semi-Markov Processes to Counter Problems," Studies in Applied Probability and Management Science, Chapter 3, (K.J. Arrow, S. Karlin, and H. Scarf, Eds.), Stanford University Press, Stanford, California, 1962.

[2]. R. Bellman, Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1957.

[3]. _____, "A Markovian Decision Process," J. Math. and Mech., Vol. 6, (1957), pp. 679-684.

[4]. _____, and S.E. Dreyfus, Applied Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1962.

[5]. D. Blackwell, "On the Functional Equation of Dynamic Programming," J. Math. Analysis and Appl., Vol. 2, (1961).

[6]. _____, "Discrete Dynamic Programming," Ann. Math. Stat., Vol. 33, (1962), pp. 719-726.

[7]. F. D'Épenoux, "Sur un problème de Production et de Stockage dans l'Aléatoire," Revue Francaise de Recherche Opérationelle, No. 14, (1960), pp. 3-16.

[8]. C. Derman, "On Sequential Decisions and Markov Chains," Manag. Sci., Vol. 9, (1962), pp. 16-24.

[9]. R.A. Howard, Dynamic Programming and Markov Processes, Technology Press and Wiley Press, New York, 1960.

[10]. W.S. Jewell, "The Properties of Recurrent-Event Processes," Operations Research, Vol. 8, (1960), pp. 446-472.

[11]. J.G. Kemeny and J.L. Snell, Finite Markov Chains, D. Van Nostrand, Princeton, New Jersey, 1960.

[12]. M. Klein, "Inspection-Maintenance-Replacement Schedules under Markovian Deterioration," Manag. Sci., Vol. 9 (1962), pp. 25-32.

[13]. A.S. Manne, "Linear Programming and Sequential Decisions," Manag. Sci., Vol. 6, (1960), pp. 259-267.

[14]. R.M. Oliver, "A Linear Programming Formulation of Some Markov Decision Processes," presented at a meeting of the Institute of Management Sciences and Operations Research Society of America, Monterey, California, April, 1960.

[15]. R. Pyke, "Markov Renewal Processes: Definitions and Preliminary Properties," Ann. Math. Stat., Vol. 33 (1961), pp. 1231-1242.

[16]. _____, "Markov Renewal Processes with Finitely Many States," Ann. Math. Stat., Vol. 33, (1961), pp. 1243-1259.

[17]. _____, "Markov Renewal Processes of Zero Order and Their Applications to Counter Theory," Studies in Applied Probability and Management Science, Chapter 11, (K. J. Arrow, S. Karlin, and H. Scarf, Eds.), Stanford University Press, Stanford, California, 1962.

[18]. R. Pyke, "Limit Theorems for Markov Renewal Processes," Technical Report No. 24, Columbia University (1961).

[19]. W. L. Smith, "Asymptotic Renewal Theorems," Proc. Roy. Soc., Edinburgh, Series A, Vol. 64, (1954), pp. 9-48.

[20]. P. Wolfe and G. B. Dantzig, "Linear Programming in a Markov Chain," Operations Research, Vol. 10, (1962). pp. 702-710.

# BASIC DISTRIBUTION LIST
# FOR UNCLASSIFIED TECHNICAL REPORTS

Head, Logistics and Mathematical
Statistics Branch
Office of Naval Research
Washington 25, D.C.

C.O., ONR Branch Office
Navy No. 100, Box 39, F.P.O.
New York City, New York

ASTIA Document Service Center
Arlington Hall Station
Arlington 12, Virginia

Institute for Defense Analyses
Communications Research Div.
von Neumann Hall
Princeton, New Jersey

Technical Information Officer
Naval Research Laboratory
Washington 25, D.C.

C.O., ONR Branch Office
346 Broadway, New York 13, NY
Attn: J. Laderman

C.O., ONR Branch Office
1030 East Green Street
Pasadena 1, California
Attn: Dr. A.R. Laufer

Bureau of Supplies and Accounts
Code OW, Dept. of the Navy
Washington 25, D.C.

Professor Russell Ackoff
Operations Research Group
Case Institute of Technology
Cleveland 6, Ohio

Professor Kenneth J. Arrow
Serra House, Stanford University
Stanford, California

Professor G.L. Bach
Carnegie Institute of Technology
Planning and Control of Industrial
Operations, Schenley Park
Pittsburgh 13, Pennsylvania

Professor A. Charnes
The Technological Institute
Northwestern University
Evanston, Illinois

Professor L. W. Cohen
Math. Dept., University of Maryland
College Park, Maryland

Professor Donald Eckman
Director, Systems Research Center
Case Institute of Technology
Cleveland, Ohio

Professor Lawrence E. Fouraker
Department of Economics
The Pennsylvania State University
State College, Pennsylvania

Professor David Gale
Dept. of Math., Brown University
Providence 12, Rhode Island

Dr. Murray Geisler
The RAND Corporation
1700 Main Street
Santa Monica, California

Professor L. Hurwicz
School of Business Administration
University of Minnesota
Minneapolis 14, Minnesota

Professor James R. Jackson
Management Sciences Research
Project, Univ. of California
Los Angeles 24, California

Professor Samuel Karlin
Math. Dept., Stanford University
Stanford, California

Professor C.E. Lemke
Dept. of Mathematics
Rensselaer Polytechnic Institute
Troy, New York

Professor W.H. Marlow
Logistics Research Project
The George Washington University
707 - 22nd Street, N.W.
Washington 7, D.C.

Professor Oskar Morgenstern
Economics Research Project
Princeton University
92 A Nassau Street
Princeton, New Jersey

## BASIC DISTRIBUTION LIST
## FOR UNCLASSIFIED TECHNICAL REPORTS

Professor R. Radner
Department of Economics
University of California
Berkeley, California

Professor Stanley Reiter
Department of Economics
Purdue University
Lafayette, Indiana

Professor Murray Rosenblatt
Department of Mathematics
Brown University
Providence 12, Rhode Island

Mr. J.R. Simpson
Bureau of Supplies and Accounts
Navy Department (Code W31)
Washington 25, D.C.

Professor A.W. Tucker
Department of Mathematics
Princeton University
Princeton, New Jersey

Professor J. Wolfowitz
Department of Mathematics
Lincoln Hall, Cornell University
Ithaca 1, New York