The Annals of Mathematical Statistics
1971, Vol. 42, No. 2, 477–496

MARKOV RENEWAL PROGRAMS WITH SMALL INTEREST RATES[1]

BY ERIC V. DENARDO

Yale University

1. Introduction. When decisions and rewards are introduced into a Markov renewal process, the resulting optimization problem is called a *Markov renewal program.* After the appearance of Pyke's [32] basic papers on Markov renewal processes, Markov renewal programming was initially developed by de Cani [7], Howard [22], Jewell [24] and Schweitzer [34], [35], with many subsequent contributions. The Markov renewal programming model allows random transitions, thereby generalizing a discrete-time (fixed transition time) Markov decision problem treated by Shapley [36], Howard [21], Blackwell [2] and several others.

Typically, the expectation $V(t)$ of the total income received until time t approaches infinity as $t \to \infty$. (The symbol $V(t)$, along with others used here, is defined precisely in Section 3; actually, $V(t)$ is a vector having one component per state.) When a positive interest rate s is introduced into the model, income earned at time t is discounted by the factor e^{-st}. This renders finite the expectation $v(s)$ of the total discounted income over an infinite planning horizon whenever $V(t)$ grows slower than exponentially.

This paper is concerned primarily with the limiting behavior of $v(s)$ as the interest rate s vanishes. Fortunately, discounting corresponds precisely to taking a Laplace–Stieltjes transform, so that a substantial theory can be invoked. In particular, the behavior of $v(s)$ as $s \to 0+$ is related through Abelian and Tauberian theorems to the behavior of $V(t)$ as $t \to \infty$.

This paper's basic tool is a partial (Laurent) expansion of $v(s)$ in powers of s, which is obtained in Theorem 1. The hypothesis of this theorem is that the reward structure have finite $(n+1)$st moment and that the transition structure have finite $(n+2)$nd moment. Under these conditions, Theorem 1 verifies that

$$v(s) = s^{-1}V_{-1} + V_0 + \cdots + s^n V_n + o(s^n)$$

and evaluates the coefficients $V_{-1}, V_0, \cdots, V_n$. This expansion generalizes results obtained by Jewell [21] for the Markov renewal program and by Blackwell [2] and Miller and Veinott [31] for the discrete-time model. The technology used to obtain this expansion applies equally well to two features of the Markov renewal process itself, namely the last-observed-state function $P(t)$ and the expected-number-of-transitions function $M(t)$. Corollary 1 applies the technology of Theorem 1 to justify and evaluate comparable series expansions of the Laplace–Stieltjes transforms of $P(t)$ and $M(t)$. In so doing, the corollary generalizes results of Pyke [32], Jewell [24], [25], Kshrisagar and Gupta [29], and Hunter [23].

Received May 5, 1969; revised September 21, 1970.

[1] This research was partially supported by the National Science Foundation under grant GK-13757.

Different interest rates reflect different trade-offs between immediate income and future income, so that one can expect the optimal policy to vary with the interest rate. However, Blackwell [2] analyzed the discrete-time model to show that a particular policy is optimal, simultaneously, for all sufficiently small interest rates. He called such a policy "optimal." Theorem 3 treats the Markov renewal program and gives conditions under which a policy exists that is optimal in Blackwell's sense. Example 1 shows constructively that when these conditions are violated there may not exist a policy that is optimal in this sense.

The equation displayed above suggests that when the interest rate is small the decision-maker maximize the "gain" V_{-1}, break ties by maximizing the "bias" V_0, etc. This observation gives rise to a family of successively more refined criteria of which the first two are called gain-optimality and bias-optimality. This family has a natural limit, and one is lead to suspect that a policy that is optimal with respect to the limiting criteria is optimal in Blackwell,s sense. Theorem 4 gives conditions under which this is the case, and Example 2 shows that the result need not hold when the conditions are violated.

Theorem 2 and Corollary 2 study the equations encountered when linear programming or policy iteration is used to compute a gain-maximizing policy. They interpret the ambiguity in the solution of these equations as the gain of a Markov renewal program having altered reward structure. The utility of such alterations was first observed by Veinott [37] in an analysis of the discrete-time model. The altered Markov decision problem is used here to compute a bias-optimal policy, and then to compute a policy that is optimal with respect to the limiting criteria suggested by the partial power series expansion of $v(s)$. Barring the pathologies illustrated by Examples 1 and 2, this policy will be optimal in Blackwell's sense.

Section 5 specializes these results to two simpler decision problems, the discrete-time model and an exponential model in which rewards are linear and transition times are exponentially distributed. Only a fraction of the material in that section is new, but it is presented within a unified framework.

2. The model. This paper treats a Markov renewal programming model having finitely many *states* that are numbered 1 through N. We speak of a state as being observed for only an instant. The time interval between observation of successive states is a random variable taking values in $[0, \infty)$. Each state i has associated with it a finite set D_i of decisions. The instant state i is observed, some decision k in D_i must be selected and cannot be changed until the instant the next state is observed.[2]

It suffices for our purposes to focus attention upon stationary non-randomized decision procedures. For this reason we define a *policy* as a function δ whose domain is the set of states and that for each state i assumes some value $\delta(i)$ in D_i.

[2] Chitgopekar [3] has analyzed a model having linear costs but allowing more general policies involving "hesitation".

Using policy δ means that the decision-maker selects decision $\delta(i)$ each instant at which state i is observed. The collection Δ of all such policies is called the *policy space*.

The probabilistic structure of the model is most easily described in terms of a sequence $\{S_n, t_n \mid n = 0, 1, \cdots\}$ of pairs of random variables. The index n indicates the nth observed state. State S_n is observed at time t_n, with $t_n \geqq t_{n-1}$ and $t_0 = 0$. The transition structure is governed by a set of functions $Q_{ij}^k(x)$ defined for $x \geqq 0$, $i, j = 1, \cdots, N$ and k in D_i. Let

$$Q_{ij}^k(x) = \Pr\{S_{n+1} = j,\, t_{n+1} \leqq t + x \mid S_n = i,\, t_n = t,\, \delta(i) = k\}.$$

Thus $Q_{ij}^k(x)$ is the joint probability that the next observed state is j and that the transition time is no greater than x, given observation of state i and given decision k. The notation indicates that $Q_{ij}^k(x)$ is independent of n and t_n. For a fixed policy δ, the set $\{Q_{ij}^{\delta(i)}(\cdot) \mid i, j = 1, \cdots, N\}$ is the family of transition probabilities associated with a *semi-Markov process* or, equivalently, a *Markov renewal process*; cf. Pyke [32]. Transitions are assumed to occur with probability 1, so that $\sum_{j=1}^N Q_{ij}^k(\infty) = 1$ for each state i and decision k. Certain transitions can be instantaneous with probability one, allowing $\sum_{j=1}^N Q_{ij}^k(0) = 1$ for particular i and k. However, we exclude ergodic chains all of whose transitions have this form, so that $\mathscr{E}\{t_N \mid S_0\} > 0$, independent of S_0. Zero-time transitions seldom occur in real-world models, but they often occur when these models are reorganized for efficient computation; cf. [9].

The income structure of the Markov renewal programming model is specified by a collection of functions of the form $R_i^k(x)$ defined for $x \geqq 0$, k in D_i and $i = 1, \cdots, N$. $R_i^k(x)$ is the expectation of the income earned during the time interval $[0, \min(x, t_1)]$ given the conditions $S_0 = i$ and $\delta(i) = k$. Like the transition structure, the income structure is assumed to be time invariant. So, roughly speaking,[3] $R_i^k(x)$ is also the expectation of the income earned during the interval $[t, \min(t+x, t_{n+1})]$ given the conditions $t_n = t$, $S_n = i$ and $\delta(i) = k$. It is reasonable to assume, as we do, that $R_i^k(x)$ has bounded variation on $[0, \infty)$, though weaker conditions would suffice. Note that for a given policy the sequence $\{S_n, t_n \mid n = 0, 1, \cdots\}$ is Markov in the sense that observing state S_n at time t_n regenerates both the transition and the income structure of the process.

Except for the selection of a performance criterion, the model is now completely specified. When a positive interest rate is included, the natural criterion is to maximize the expectation of the discounted income stream. As the interest rate approaches zero, a family of alternative criteria emerges, of which the simplest is to maximize the rate of income per unit time.,Definition and discussion of this family of criteria is postponed until Section 4, at which point the requisite notation will be on hand.

The model and its analysis simplify markedly in two familiar cases. In the *discrete-time* case treated by Shapley [36], Howard [21] and many others, each

[3] The intent of this interpretation of $R_i^k(x)$ is probably quite clear, even though it is imprecise when $t_{n-1} = t_n$. $R_i^k(x)$ could be properly defined in terms of the triplet (n, S_n, t_n).

transition takes exactly one unit of time, so that $t_n = n$. Income r_i^k is earned each epoch in which state i is observed and decision k is selected. It simplifies certain formulas (see (42)–(43) below) to assume that r_i^k is earned at the end of the epoch; i.e., $R_i^k(x) = 0$ for $x < 1$ and $R_i^k(x) = r_i^k$ for $x \geqq 1$. Howard [21], Zachrisson [41] and others have treated a decision problem whose transition law is determined by the differential equation of a stationary continuous-time Markov process. Miller [30] and Rykov [33] showed that attention can be confined to stationary policies, which simplifies this model to the *exponential* case with linear returns and exponential transition times; that is,

$$Q_{ij}^k(x) = p_{ij}^k[1-\exp(-\lambda_i^k x)] \quad \text{and} \quad R_i^k(x) = r_i^k[1-\exp(-\lambda_i^k x)].$$

3. Policy Evaluation. In preparation for the optimization problem, we now fix upon a particular policy and develop some of its probabilistic and income generating properties. To simplify the notation, dependence on the decision k and policy δ is dropped temporarily. This development is organized so as to obtain equations (1)–(7) in an intuitively appealing manner that omits justification of the required integrations. The needed justification follows (7) and uses arguments similar to Pyke's [32].

Let $V_i(t)$ denote the expectation of the undiscounted income earned during the interval $[0, t]$ if state i is observed at time 0. The behavior of $V_i(t)$ as $t \to \infty$ is intimately related to the behavior as $s \to 0^+$ of its Laplace–Stieltjes transform

$$v_i(s) = \int_{0-}^{\infty} e^{-st}\, dV_i(t). \tag{1}$$

Interpret s as an interest rate that is compounded instantaneously, so that a dollar received at time t has present value e^{-st} at time 0. Then $v_i(s)$ has a second interpretation as the present value at time 0 of the entire income stream, given initial state i and interest rate s.

The technique we shall use to analyze $V_i(t)$ applies equally well to the analysis of two related features of the semi-Markov process identified by $Q(t)$. Let $M(t)$ be the $N \times N$ matrix whose ijth element $M_{ij}(t)$ is the expectation of the number of occurrence of state j during the interval $[0, t]$, given that state i is observed initially. Similarly, let $P(t)$ be the $N \times N$ matrix whose ijth element $P_{ij}(t)$ is the probability that state j was the last observed state during the interval $[0, t]$, given that state i was observed initially. Conditioning $M_{ij}(t)$, $V_i(t)$ and $P_{ij}(t)$ on the first transition produces the renewal equations

$$M_{ij}(t) = \delta_{ij} + \textstyle\sum_{k=1}^N \int_{0-}^{t+} dQ_{ik}(x) M_{kj}(t-x) \tag{2}$$

$$V_i(t) = R_i(t) + \textstyle\sum_{j=1}^N \int_{0-}^{t+} dQ_{ij}(x) V_j(t-x) \tag{3}$$

$$P_{ij}(t) = H_{ij}(t) + \textstyle\sum_{k=1}^N \int_{0-}^{t+} dQ_{ik}(x) P_{kj}(t-x), \tag{4}$$

where δ_{ij} is the Kroeneker symbol and $H_{ij}(t) = \delta_{ij}[1 - \sum_{l=1}^N Q_{il}(t)]$. As in (1), we use small letters for Laplace–Stieltjes transforms, so that the transforms of $M_{ij}(t)$, $Q(t)$ and $P(t)$ become $m_{ij}(s)$, $q(s)$, and $p(s)$. Take Laplace–Stieltjes transforms of (2)–(4) (noting that the transform of the convolution is the product of the transforms), write them in matrix notation, and rearrange them slightly to produce the simple expressions

$$[I - q(s)]\, m(s) = I \tag{5}$$

$$[I - q(s)]\, v(s) = r(s) \tag{6}$$

$$[I - q(s)]\, p(s) = h(s). \tag{7}$$

An argument akin to Pyke's [32] is now used to justify the integrations in (1)–(7). Let $Q^{*n}(t)$ denote the n-fold convolution of $Q(t)$ with itself, so that

$$Q_{ij}^{*2}(t) = \textstyle\sum_{k=1}^{N} \int_{0-}^{t+} dQ_{ik}(x) Q_{kj}(t-x).$$

Observe that $\sum_{n=0}^{m} Q_{ij}^{*n}(t)$ is the expectation of the number of occurrences of state j in the interval $[0, \min(t_m, t)]$ given $S_0 = i$. Let ε be the expectation of the smallest non-zero transition time. The prohibition of ergodic chains of zero-time transitions, along with a simple computation, produces the result $P(t_m < m\varepsilon/2N) \to 0$ as $m \to \infty$. This suffices for both facts in

$$M_{ij}(t) = \textstyle\sum_{n=0}^{\infty} Q_{ij}^{*n}(t) = O(t),$$

the latter meaning that $M_{ij}(t)/t$ is bounded as $t \to \infty$. Isolating the $n = 0$ term in the above produces (2). Taking transforms produces (8), when one notes that $\{[q(s)]^n\}$ is a geometric series.

$$m(s) = \textstyle\sum_{n=0}^{\infty} [q(s)]^n = [I - q(s)]^{-1} = O(1/s), \qquad \text{as} \quad s \to 0. \tag{8}$$

Equation (8) also implies that $m(s)$ and $q(s)$ commute. Similar arguments lead to (3) and (4). To obtain (3), compute the expected income until the earlier of time t and the time of the nth transition. Then let $n \to \infty$ and isolate the 0th term in series.

Expansions of $v(s)$, $m(s)$ and $p(s)$. Our primary concern here is to use (6) to find a partial power series (Laurent) representation of $v(s)$ in terms of the moments of $Q(t)$ and $R(t)$. Due to the similarities between (5), (6) and (7), our technique for expanding $v(s)$ also yields partial power series representations for $m(s)$ and $p(s)$. Define the normalized moments Q_n and R_n by

$$Q_n = \textstyle\int_{0-}^{\infty} t^n \, dQ(t)/n!, \qquad R_n = \int_{0-}^{\infty} t^n \, dR(t)/n!. \tag{9}$$

Of course, Q_n is an $N \times N$ matrix, and R_n is an $N \times 1$ vector.

Note that Q_0 is the transition matrix of the embedded Markov chain. We now review some well-known (cf. Doob [16] page 175) Markov chain theory. As $n \to \infty$, $[Q_0]^n$ converges $(C, 1)$ to a stochastic matrix P^* that is characterized by the equations $P^*(I - Q_0) = 0$ and $P^*\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is the $N \times 1$ vector of 1's. In addition, $P^* = Q_0 P^* = P^* P^*$. In general, Q_0 can have several ergodic chains and some transient states. Number the ergodic chains from 1 through n, and let S_m denote the set of states in ergodic chain m, for $m = 1, \cdots, n$. Also, let T denote the (possibly empty) set of transient states. With $P_i{}^*$ denoting the ith row of P^*, the ergodic states have

$$(10) \qquad P_i{}^* = \pi^m \qquad \text{for all } i \in S_m, \quad m = 1, \cdots, n,$$

where π^m contains the stationary probability distribution for chain m and zeros; π^m is characterized by the equations $\pi^m(I - Q_0) = 0$, $\pi^m\mathbf{1} = 1$ and $\pi_j{}^m = 0$ for $j \notin S_m$. For transient states,

$$(11) \qquad P_i{}^* = \sum_{m=1}^n t_{im}\pi^m \qquad \text{for all} \quad i \in T,$$

where $t_{im} \geqq 0$, $\sum_{m=1}^n t_{im} = 1$ and t_{im} is the probability of eventual absorption into ergodic chain m if one starts in (transient) state i.

Kemeny and Snell [28] showed that the matrix $(I - Q_0 + P^*)$ is invertible and called its inverse Z the *fundamental matrix*. (Actually, ([28] page 100) proves this only for the ergodic case, but the general case is a simple extension.) Since $P^*(I - Q_0 + P^*) = P^* = (I - Q_0 + P^*)P^*$, the matrix Z satisfies $P^* = P^*Z = ZP^*$. In dynamic programming, it is slightly more convenient to work with the matrix $H = Z - P^*$. The preceding facts readily combine to verify that $0 = P^*H = HP^*$. Routine manipulations now yield $(I - Q_0)H = H(I - Q_0) = I - P^*$; for instance, $(I - Q_0)H = (I - Q_0 + P^*)H = (I - Q_0 + P^*)(Z - P^*) = I - P^*$.

Lemma 1 studies the equation $(I - Q_0)x = b$ and in doing so uses the definition $v = Q_1\mathbf{1}$, so that v_i is the mean time to transition given state i.

LEMMA 1. *Let b be an $N \times 1$ vector satisfying $P^*b = 0$ and let c be any $N \times 1$ vector.*

(i) *Then an $N \times 1$ vector x satisfies $(I - Q_0)x = b$* iff *$x = Hb + y$ for an $N \times 1$ vector y satisfying $y = P^*y$.*

(ii) *Suppose $(I - Q_0)x = b$ and $P^*Q_1x = P^*c$. Then $x = Hb + y$ with*

$$(12) \qquad y_i = C^m = \pi^m[c - Q_1Hb]/\pi^m v \qquad \textit{for } i \in S_m, \quad m = 1, \cdots, n$$

$$(13) \qquad y_i = \sum_{m=1}^n t_{im}C^m \qquad \textit{for} \quad i \in T.$$

PROOF. For (i), first take $x = Hb$. Then $(I - Q_0)x = (I - Q_0)Hb = (I - P^*)b = b$. Now suppose $x = Hb + y$ with $y = P^*y$. Then $(I - Q_0)x = b + (I - Q_0)y = b + (I - Q_0)P^*y = b$. To complete (i), suppose $(I - Q_0)x = b$. Then $(I - Q_0)(x - Hb) = 0$; premultiplying by H produces $(I - P^*)(x - Hb) = 0$ and hence $y = P^*y$ with $y = x - Hb$.

For (ii), consider an ergodic state $i \in S_m$. From (i), $y = P^*y$, so that (10) implies $y_i = \pi^m y = C^m$, a constant independent of i; thus the transient case follows from the ergodic case and (11). Next substitute $x = Hb + y$ into $P^*Q_1x = P^*c$, yielding $P^*Q_1y = P^*[c - Q_1Hb]$. From (12), (ii) will be proved when it is shown that $P_i{}^*Q_1y = \pi^m vC^m$. For $j \in S_m$, one has $(Q_1y)_j = \sum_k (Q_1)_{jk}y_k = v_jC^m$, implying $P_i{}^*Q_1y = \pi^m vC^m$ as desired. □

To calculate the $N \times 1$ vector x satisfying (ii) of Lemma 1, we must determine P^* and H, which can be done with one matrix inversion; cf. Schweitzer [35]. Suppose Q_i is finite, which implies Q_j is finite for $j < i$. A standard Abelian relationship (cf. Chung [5] page 156) between the moments of a function and the behavior of its transform for small s yields

$$(14) \qquad q(s) = Q_0 - sQ_1 + \cdots + (-s)^i Q_i + o(s^i).$$

Similarly, if R_j is defined and finite,

$$(15) \qquad r(s) = R_0 - sR_1 + \cdots + (-s)^j R_j + o(s^j).$$

LEMMA 2. *Suppose the $N \times 1$ vector b satisfies $P^*b = 0$. Then, $m(s)b = o(1/s)$.*

PROOF. Note in (8) that $m(s) = O(1/s)$. First, we show that $m(s)(I - Q_0) = o(1/s)$. From (8), $m(s)$ and $q(s)$ commute, so that $m(s)[I - q(s)] = I$. (14) with $i = 0$ gives $q(s) = Q_0 + o(1)$. Combining the two gives

$$(16) \qquad m(s)(I - Q_0) = I + m(s)o(1) = I + O(1/s)o(1) = I + o(1/s) = o(1/s).$$

Since $P^*b = 0$, Lemma 1 shows that $(I - Q_0)Hb = b$ and therefore that $m(s)b = m(s)(I - Q_0)Hb = o(1/s)Hb = o(1/s)$, as desired. □

Lemma 1 and Lemma 2 are the principle ingredients in our proof of Theorem 1, which analyzes the equation

$$(17) \qquad v(s) = s^{-1}V_{-1} + V_0 + \cdots + s^n V_n + e_n(s).$$

Theorem 1 gives conditions under which $e_n(s) = o(s^n)$ and evaluates the coefficients $V_{-1}, \cdots, V_n$ under these conditions. To interpret V_{-1}, we apply the standard Tauberian theorem ([17] page 420) for non-decreasing functions. Since $V(t) = (M * R)(t)$ and $M(t)$ is non-decreasing, (17) suffices for $V(t) \sim tV_{-1}$ as $t \to \infty$. Consider the difference $V(t) - tV_{-1}$ and related quantities. If $V(t) - tV_{-1}$ converges $(C, 0)$ or $(C, 1)$, the standard Abelian theorem ([40] page 182) guarantees that the limit is V_0. In the discrete time case, Kemeny and Snell [28] show that $M(t) - tP^* \to H$ as t approaches infinity though the integers, where the convergence is (C, 1) when P is cyclic. Similarly, one can show that

$$(18) \qquad V(t) - tV_{-1} \to V_0\,(C, 1) \qquad \text{as} \quad t \to \infty,$$

where the convergence in (18) is $(C, 0)$ whenever $dQ^n(t)/dt$ exists and is strictly positive for some n and all t in some interval.

Equations (17) and (18) relate to the optimization problem. When the interest rate s is small, (17) prompts the decision maker to select a policy that maximizes V_{-1} and break ties by maximizing V_0. (18) plays the same role when $s = 0$.

THEOREM 1. *Suppose Q_{n+2} is finite and R_{n+1} is defined and finite. Then,* (17) *holds with $e_n(s) = o(s^n)$. Moreover, for $i = -1, 0, \cdots, n$, the vector V_i is the unique solution of equations*

$$(19a) \qquad (I - Q_0)V_i = b_i,$$

$$(19b) \qquad P^*Q_1 V_i = P^*c_i$$

where $c_{-1} = R_0, b_{-1} = 0, b_i = c_{i-1} - Q_1 V_{i-1}$ *and, for* $i \geqq 0$,

$$c_i = (-1)^{i+1} R_{i+1} + \textstyle\sum_{j=2}^{i+2} (-1)^j Q_j V_{i+1-j}. \tag{20}$$

PROOF. First, observe that c_i and b_i depend only on variables $V_{-1}, V_0, \cdots, V_{i-1}$ and moments Q_0 through Q_{i+2} and R_0 through R_{i+1}. Applying this observation recursively shows that V_i is defined in terms of moments Q_0 through Q_{i+2} and R_0 through R_{i+1}. In particular, the expression for V_n involves only the moments of $Q(t)$ and $R(t)$ that are hypothesized as defined and finite.

Lemma 1 shows that (19)–(20) have unique solutions for $i = -1, 0, \cdots, n$. To complete the proof, we need only show that these solutions give $e_n(s) = o(s^n)$ in (17). For this purpose, substitute (14) with $i = n+2$, (15) with $j = n+1$ and (17) into (6). With $V_{-1}, \cdots, V_n$ as specified by (19)–(20), gather coefficients (the first $n+1$ of which are zero) to obtain

$$[I - q(s)]e_n(s) = [c_n - Q_1 V_n]s^{n+1} + o(s^{n+1}). \tag{21}$$

Since $P^*[c_n - Q_1 V_n] = 0$, Lemma 2 and (8) apply to (21), giving

$$\begin{aligned} e_n(s) &= m(s)[c_n - Q_1 V_n]s^{n+1} + m(s)o(s^{n+1}) \\ &= o(s^{-1})s^{n+1} + O(s^{-1})o(s^{n+1}) = o(s^n). \quad \square \end{aligned} \tag{22}$$

Lemma 1 and Theorem 1 provide an explicit, systematic method for computing V_{-1} through V_n, once P^*, H and the requisite moments of $R(t)$ and $Q(t)$ have been calculated. The following corollary exploits the similarity of (5) and (7) to (6). In it, Theorem 1 is adapted to obtain all desired coefficients of $m(s)$ and $p(s)$. Let $H_n = \int_{0-}^{\infty} t^n \, dH(t)/n!$ and observe that $H_0 = 0$. Note that H_n is finite whenever Q_n is finite.

COROLLARY 1. *Suppose* Q_{n+2} *is finite. Then,*

$$m(s) = \textstyle\sum_{i=-1}^{n} s^i M_i + o(s^n) \tag{23}$$

$$p(s) = \textstyle\sum_{i=0}^{n} s^i P_i + o(s^n) \tag{24}$$

where P_i *and* M_i *are the unique solutions of the equations*

$$(I - Q_0)M_i = B_i, \tag{25a}$$

$$P^* Q_1 M_i = P^* C_i \tag{25b}$$

$$(I - Q_0)P_i = D_i, \tag{26a}$$

$$P^* Q_1 P_{i,} = P^* E_i \tag{26b}$$

with

$C_{-1} = I, B_{-1} = D_{-1} = E_{-1} = 0, B_i = C_{i-1} - Q_1 M_{i-1}, D_i = E_{i-1} - Q_1 P_{i-1}$, *and*

$$C_i = \textstyle\sum_{j=2}^{i+2} (-1)^j Q_j M_{i+1-j}, \tag{27}$$

$$E_i = (-1)^{i+1} H_{i+1} + \textstyle\sum_{j=2}^{i+2} (-1)^j Q_j P_{i+1-j}. \tag{28}$$

PROOF. First consider the variants of Lemmas 1 and 2 in which the $N \times 1$ vectors b and c are replaced by $N \times N$ matrices B and C with $P^*B = 0$. Each column of B and C may be treated separately, so these variants stand. Then, to prove the corollary for $m(s)$ and $p(s)$, repeat the proof of Theorem 1 using these variants of Lemmas 1 and 2 and substituting, respectively, into (5) and (7), rather than (6). It remains only to show that $P_{-1} = 0$, which follows trivially from $H_0 = 0$. □

Of course, if the expansions of $v(s)$, $m(s)$ and $p(s)$ were all required, it would be most efficient to deal first with $m(s)$ and then use the equations $v(s) = m(s)r(s)$ and $p(s) = m(s)h(s)$ for the others. For any of these expansions, the constants like c_{i-1} and b_i are so closely interrelated as to suggest simpler expressions than indicated by Theorem 1 and its corollary. Unfortunately, none of the expressions seems to simplify in the general case, even though they simplify markedly in the discrete-time and exponential cases treated in Section V of this paper.

Several authors have shown how to compute the moments in semi-Markov processes. Pyke [32] computes P_0 and M_{-1}. He also related $M(t)$ to the first passage time distribution. Barlow and Proschan [1] provide a systematic method that can readily be extended to obtain all desired moments of the first passage time distribution. Jewell [25] first obtained formulas for M_0 and V_0. Kshirsagar and Gupta [29] and, more recently, Hunter [23] also obtained formulas for M_{-1} and M_0, the latter by a route that recognises Z (and hence H) as a generalized inverse of $I - P$. Readable introductions to semi-Markov processes are found in Fox [19] and Çinlar [6].

In this context, our method is integrated and obtains all desired moments of $m(s)$, $p(s)$ and $v(s)$. It has the added advantage of integrating effectively with dynamic programming, as Theorem 2 will attest.

Lemma 1 and Theorem 1 provide explicit expressions for V_{-1}, V_0, etc. For notational convenience, set $V_{-1} = g$ and $V_0 = w$, so that

$$g_i = C^m = \pi^m R_0 / \pi^m v \qquad \text{for} \quad i \in S_m, \quad m = 1, \cdots, n$$

$$g_j = \textstyle\sum_{m=1}^{n} t_{jm} C^m \qquad \text{for} \quad j \in T.$$

The explicit formula for w is rather complex and is provided below for the case in which P has one ergodic chain.

$$w = P^*(-R_1 + Q_2 g / \pi v + (I - P^* Q_1 / \pi v)\, H\, (R_0 - Q_1 g).$$

An altered semi-Markov process. Equations (17) and (18) motivate us to call V_{-1} and V_0 the *gain* and *bias*, respectively. Note from (19a), (19b) that the gain depends only on the *triplet* (Q_0, Q_1, R_0), not on higher moments of $Q(t)$ or $R(t)$. A policy that maximizes the gain may be computed efficiently by any of several linear programming and policy iteration schemes that are partially described in Section 4 and 5 and more completely specified in [10], [12] and the references cited there. Linear programming and policy iteration approaches that find a gain-maximizing policy both lead to consideration of equations

(29a) $$(I-Q_0)g = 0,$$

(29b) $$(I-Q_0)x = R_0 - Q_1 g,$$

where g and x are $N \times 1$ vectors. These equations are treated by the corollary that follows

THEOREM 2. *Suppose Q_{n+2} is finite and R_{n+1} is defined and finite. Consider a solution z of the equation $(I-Q_0)z = b_n$. Then $V_n - z$ is the gain rate for any semi-Markov process with rewards having triplet $(Q_0, Q_1, c_n - Q_1 z)$.*

PROOF. To obtain the left-most of the two equations below, subtract $(I-Q_0)z = b_n$ from (19a) with $i = n$. For the right-hand equation, subtract the identity $P^*Q_1 z = P^*Q_1 z$ from (19b) with $i = n$.

$$(I-Q_0)(V_n - z) = 0, \qquad P^*Q_1(V_n - z) = P^*[c_n - Q_1 z]$$

Since $b_{-1} = 0$ and $c_{-1} = R_0$, the above pair of equations play the same role as do (19a), (19b) with $i = -1$. This identifies $V_n - z$ as the gain rate for the process with rewards $C_n - Q_1 z$ until transition, completing the proof. □

Theorem 2 introduces a secondary semi-Markov process that has altered rewards but need not have altered transition structure. This altered process forms the foundation for the optimization procedures introduced in Section 4. Veinott [37] first observed the utility of such alterations in an analysis of the discrete-time model. Equations, (29a), (29b) are now treated by

COROLLARY 2. *Suppose Q_2 is finite and R_1 is defined and finite. Then every solution (g, x) of (29a), (29b) has $g = V_{-1}$. Moreover, $V_0 - x$ is the gain rate for any semi-Markov process with rewards having triplet (Q_0, Q_1, R') with*

$$R' = -R_1 + Q_2 V_{-1} - Q_1 x.$$

PROOF. First note that (29a) and (29b) are identical to (19a) with $i = -1$ and with $i = 0$, respectively. So, the pair (V_{-1}, V_0) satisfies (29a), (29b). Premultiply (29b) by P^* to obtain (19b) with $i = -1$. This and Theorem 1 imply $g = V_{-1}$. For the remainder of Corollary 2, apply Theorem 2 with $n = 0$. □

4. Optimization. Attention is now returned to the optimization model having several alternative decisions per state. As before, policy δ specifies for each state i a decision $\delta(i)$ in D_i. Superscripts are introduced to denote dependence upon δ, so that $v^\delta(s)$ is the discounted income function for policy δ. As in equation (1), the ith component of the $N \times 1$ vector $v^\delta(s)$ is the expectation of the total discounted income under the following circumstances: state i is observed initially, policy δ is used, the planning horizon is infinite, and the interest rate is s.

Several optimality criteria are now introduced and interrelated. It is then shown how to compute policies that are optimal with respect to most of these criteria. The first portion of this discussion is virtually independent of Section 3.

s-Optimal and optimal policies. For $s > 0$, we call policy δ *s-optimal* if $v^{\delta}(s) \geqq v^{\eta}(s)$ for every policy η in the policy space Δ. Since $v^{\delta}(s)$ has N components, an s-optimal policy must attain N maxima simultaneously. Jewell [24] showed that an s-optimal policy exists. Moreover, it is known ([8] Example 4) that each s-optimal policy in Δ maximizes the discounted income stream over the broader class of all randomized history-remembering decision procedures that exclude switching decisions in mid-transition. So no advantage obtains in discounted dynamic programming from the introduction of this broader class of decision procedures.

Following Blackwell [2], we call a policy *optimal* if it is s-optimal for all sufficiently small positive s. Blackwell first demonstrated, nonconstructively, that an optimal policy exists in the discrete-time case. Let Δ_{opt} denote the (possibly empty) set of optimal policies. We will shortly give a condition that suffices for Δ_{opt} to be non-empty, followed by an example having $\Delta_{\text{opt}} = \emptyset$. Consider any sequence $t_n \to 0+$. Since there are only finitely many policies, one policy λ is t_n-optimal for infinitely many n. Define Δ_{seq} as the set consisting of each policy λ that is s_n-optimal for every n in at least one sequence $s_n \to 0+$. The preceding observation assures Δ_{seq} be non-empty. By definition, Δ_{seq} contains Δ_{opt}.

In preparation for Theorem 3, note that since $v^{\delta}(s)$ is a Laplace–Stieltjes transform it is analytic ([40] page 57) in the open half-plane $Re(s) > 0$. Theorem 3 turns on the behavior of $v(s)$ at the origin. Since $sv(s) \to V_{-1}$ as $s \to 0$, $v(s)$ is not analytic at the origin whenever $V_{-1} \neq 0$. The question is whether the origin is an isolated singularity; i.e., whether $sv(s)$ is analytic at $s = 0$.

THEOREM 3. *Suppose for every δ that v^{δ} has an isolated singularity or is analytic at $s = 0$. Then Δ_{opt} is non-empty.*

PROOF. Since $\Delta_{\text{opt}} \subset \Delta_{\text{seq}} \neq \emptyset$, it suffices to show that $\Delta_{\text{opt}} = \Delta_{\text{seq}}$. Suppose Δ_{seq} contains exactly one policy. Then this policy must be optimal for all sufficiently small interest rates, and $\Delta_{\text{seq}} = \Delta_{\text{opt}}$. Suppose Δ_{seq} contains multiple policies. Pick any state i and any distinct policies δ and η in Δ_{seq}. Necessarily, there exists a sequence $s_n \to 0+$ such that $v_i^{\delta}(s_n) = v_i^{\eta}(s_n)$ for each n. Set $g(s) = s[v_i^{\delta}(s) - v_i^{\eta}(s)]$. The hypothesis and Theorem 1 with $n = -1$ imply that $g(s)$ is analytic in a domain containing the origin and the open half-plane $Re(s) > 0$. Since $g(s_n) = 0$ for each n, the identity theorem ([20] page 199) for analytic functions yields $g(s) \equiv 0$ and hence $v^{\delta}(s) \equiv v^{\eta}(s)$, in which sense policies in Δ_{seq} are indistinguishable. Hence, $\Delta_{\text{seq}} = \Delta_{\text{opt}}$. □

Equations (41) and (43) in Section 5 indicate that the hypothesis of Theorem 3 is satisfied in the discrete-time and exponential cases. An example of a Markov renewal program for which no optimal policy exists is built upon the function-transform pair given in Doetsch ([15] page 241) as

$$F(t) = \frac{\sin(2t)^{\frac{1}{2}} \sinh(2t)^{\frac{1}{2}}}{(\pi t)^{\frac{1}{2}}}, \quad f(s) = s^{\frac{1}{2}} \sin(1/s).$$

Note that the Laplace–Stieltjes transform $f(s)$ oscillates as s decreases to zero. We equate return function $R_1{}^{b}(t)$ with $F(t)$ even though $F(t)$ is not of bounded variation on $[0, \infty]$. The assumption that each $R_i{}^k(t)$ be of bounded variation was only used to obtain (15) with $j = 0$, and $f(s) = o(1)$ without this assumption.

EXAMPLE 1. (Δ_{opt} is empty). State 1 is the only state. There are two decisions, b and c, with $Q_{11}^b(t) = Q_{11}^c(t) = 1 - e^{-t}$. Decision c generates no income, so that $0 = R_1{}^c(t) = V_1{}^c(t)$. Decision b has income $R_1{}^b(t) = F(t)$ so that $V_1{}^b(t)$ has transform $v_1{}^b(s) = [\sin(1/s)]s^{\frac{1}{2}}/(1+s)$. Since $v_1{}^b(s)$ oscillates about zero as s decreases to zero, no policy is optimal, and Δ_{opt} is empty. Note also that neither $v_1{}^b(s)$ nor $f(s)$ is analytic at the origin. □

Again using superscripts to denote dependence upon policies, Theorem 1 assures that

$$v^\delta(s) = s^{-1}V_{-1}^\delta + V_0{}^\delta + \cdots + s^n V_n{}^\delta + o(s^n) \tag{30}$$

whenever Q_{n+2}^δ is finite and R_{n+1}^δ is defined and finite. Equation (30) seems related to Theorem 3. However, the proof of Theorem 3 exploited Theorem 1 only for $n = -1$, which only requires finiteness of $Q_1{}^\delta$ and $R_0{}^\delta$. Moreover, the hypothesis of Theorem 3 holds in some cases when higher moments of $Q^\delta(t)$ and/or $R^\delta(t)$ diverge. When this happens, $v^\delta(s)$ has the Laurent expansion about the origin $v^\delta(s) = \sum_{n=-1}^\infty s^n X_n$; but for $n \geqq 0$ the coefficient X_n is not specified by (19a), (19b), and so (30) is meaningless.

Related optimality criteria. Theorem 3 fails to indicate how to compute Δ_{opt} when this set is non-empty. However, (18) and (30) suggest a second line of approach that admits of computation when appropriate moments of $Q^\delta(t)$ and $R^\delta(t)$ exist. When the interest rate is small, these equations motivate the decision-maker to select policy δ so as to maximize V_{-1}^δ, break ties by maximizing $V_0{}^\delta$, etc. So, with this in mind and with $\Delta_{-2} = \Delta$, define the (possibly empty) set Δ_k recursively by

$$\Delta_k = \{\delta \in \Delta_{k-1} \mid V_k{}^\delta \geqq V_k{}^n \quad \text{for all} \quad \eta \in \Delta_{k-1}\}.$$

As suggested by (18), we call policies in Δ_{-1} and Δ_0 *gain-optimal* and *bias-optimal*, respectively.

Since a policy in Δ_k must attain N maxima simultaneously, it is not clear *a priori* that this set is non-empty for any k greater than -2. But (30) with $n = -1$ indicates that each policy λ in Δ_{seq} is gain-optimal. In fact, (30) shows that $\Delta_{\text{seq}} \subset \Delta_n$ whenever each policy $\eta \in \Delta_{n-1}$ has $V_n{}^\eta$ defined. By Theorem 1, this occurs when each policy $\eta \in \Delta_{n-1}$ has finite $(n+2)$nd moments of $Q_t(\cdot)$ and finite $(n+1)$st moments of $R^\eta(\cdot)$. Since $\{\Delta_k\}$ is non-increasing in k, we can define a "limit" set Δ_∞ by the following rule: if Δ_n contains a single policy for some n, set $\Delta_\infty = \Delta_n$; otherwise set $\Delta_\infty = \lim_{n\to\infty} \Delta_k$. Clearly, $\Delta_{\text{seq}} \subset \Delta_\infty$ whenever $\Delta_\infty \neq \emptyset$. We shall show shortly how to compute policies in Δ_k and Δ_∞—policies whose merits are indicated by

THEOREM 4. $\Delta_\infty = \Delta_{\text{opt}}$ *in either of the two cases*:

(i) Δ_∞ *contains a single policy*,

(ii) Δ_∞ *contains several policies, each of which has a Laurent expansion about the origin.*

PROOF. It has been shown that $\Delta_{\text{opt}} \subset \Delta_{\text{seq}} \neq \varnothing$ and that $\Delta_{\text{seq}} \subset \Delta_\infty$ whenever $\Delta_\infty \neq \varnothing$. In (i), Δ_∞ consists of a single policy, which necessitates $\Delta_{\text{opt}} = \Delta_{\text{seq}} = \Delta_\infty$. In (ii), any two policies δ and η in Δ_∞ have $v^\delta(s) \equiv v^\eta(s)$, so that $\Delta_{\text{opt}} = \Delta_{\text{seq}} = \Delta_\infty$. □

It seems reasonable that Δ_∞ contain a single policy in most applications. However, examples can be constructed in which Δ_∞ contains multiple policies having different income functions, and the qualification in (ii) of Theorem 4 is not superfluous. The difficulty lies in the fact that functions are not characterized by their moments, as Stieltjes ([40] page 126) illustrated with the function

$$G(x) = \int_0^x e^{-t^{1/4}} \sin(t^{\frac{1}{4}})\, dt$$

having $\int_0^\infty x^n \, dG(x) = 0$ for $n = 0, 1, \cdots$.

EXAMPLE 2. ($\Delta_{\text{opt}} \neq \Delta_\infty$, with $\Delta_\infty \neq \varnothing$). There are two states, 1 and 2. State 2 is absorbing and yields no income. State 1 has two decisions, d and e. Choosing either of these yields 1 dollar at the moment of transition, with

$$dR_1{}^d(t) = dQ_{12}^d(t) = e^{-t^{1/4}}\, dt/24$$
$$dR_1{}^e(t) = dQ_{12}^e(t) = e^{-t^{1/4}}[1 + \sin(t^{\frac{1}{4}})]\, dt/24.$$

Thus corresponding moments of $V_1{}^d(t)$ and $V_1{}^e(t)$ are finite and identical, which implies $\Delta_\infty = \{d, e\}$. Since different functions cannot have the same transform, Δ_{opt} can contain at most one element of $\{d, e\}$. □

In the discrete-time and exponential cases, one can show (see Theorem 5) that $\Delta_\infty = \Delta_{N-1}$, where N is the number of states. This allows us to terminate the computation procedure with the determination of Δ_{N-1}. One has no such assurance in the general Markov renewal program, since for any k one can construct functions $R^\delta(t)$ and $R^\eta(t)$ that are identical in their first k moments and differ in their $(k+1)$st moment. (In fact, Boas [40] page 139, showed that any sequence of moments is realiable in a function $R^\delta(t)$ of bounded variation.)

Two other optimality criteria have been suggested for the discrete-time and continuous-time models. In our notation, Veinott [38] would call policy δ *n-discount optimal* if

$$\text{(31)} \qquad \liminf_{s \to 0+} s^{-n}[v^\delta(s) - v^\eta(s)] \geqq 0$$

for every policy η, stationary or non-stationary. One sees from (30) that whenever Δ_n is defined, a stationary policy is n-discount optimal; in fact, every policy in Δ_n is n-discount optimal.

Veinott [39] also observed a time-domain analogue of n-discount optimality. Multiplying the transform by s amounts in this case to averaging the function; so Veinott calls policy δ(−1)-*average optimal* if

$$\text{(32)} \qquad \liminf_{t \to \infty} t^{-1}[V^\delta(t) - V^\eta(t)] \geqq 0$$

where η is any policy, stationary or not. Derman [14] and Brown [3] provided the first proofs of the equivalence between (−1)-discount and (−1)-average optimality in the discrete-time case. The time-domain correspondent to dividing by s is to integrate once. So, let $I_0{}^\delta(t) = V^\delta(t)$ and $I^\delta_{n+1}(t) = \int_{0-}^{t} I_n{}^\delta(x)\,dx$ for $n \geqq 0$. For $n \geqq 0$, Veinott calls policy δ *n-average optimal* if, in our notation,

$$\text{(33)} \qquad \liminf_{t\to\infty}\,[I_n{}^\delta(t) - I_n{}^\eta(t)] \geqq 0 \qquad (C, 1)$$

for every policy η, stationary or not. For the discrete-time case, Veinott demonstrated equivalence between n-discount optimality and n-average optimality, extending a result Miller and the author [13] obtained for the case $n = 0$. This equivalence has not yet been extended to the general Markov renewal program.

Computation. Assume from now on that $\Delta_{\text{opt}} = \Delta_\infty \neq \varnothing$, so that the pathologies indicated by Example 1 and Example 2 do not arise. The computation of bias-optimal and optimal policies can be parsed into a sequence of simpler Markov renewal programs, each of which may be solved by linear programming or policy iteration.

Toward this end, some notation is now introduced. Components of the triplet $(Q_0{}^\delta, Q_1{}^\delta, R_0{}^\delta)$ are defined, as before, by

$$p^k_{ij} = Q^k_{ij}(\infty), \qquad v^k_{ij} = \int_{0-}^{\infty} t\, dQ^k_{ij}(t), \qquad v_i{}^k = \sum_{j=1}^N v^k_{ij}, \quad r_i{}^k = R_i{}^k(\infty).$$

It simplifies the notation to set $g^\delta = V^\delta_{-1}$ and $w^\delta = V_0{}^\delta$. Also, for a bias-optimal policy λ, set $g^* = g^\lambda$ and $w^* = w^\lambda$.

The policy iteration formulations are described first, and the procedures are then adapted for linear programming. The basic tool in policy iteration is a routine that finds a gain-optimal policy; its termination conditions are described in terms of a policy δ, a pair (g, x) of N-vectors, and the inequalities

$$\text{(34)} \qquad \sum_{j=1}^N p^k_{ij}\, g_j \leqq g_i$$

$$\text{(35)} \qquad r_i{}^k + \sum_{j=1}^N p^k_{ij} x_j \leqq x_i + \sum_{j=1}^N v^k_{ij} g_j.$$

In the general *multi-chain* case, the *termination conditions* for policy iteration are

(a) each pair (i, k) satisfies (34)

(b) each pair (i, k) that satisfies (34) as an equality satisfies (35)

(c) the pair $[i, \delta(i)]$ satisfies (34) and (35) as equalities, for $i = 1, \cdots, n$.

Theorem 1 and condition (c) imply $g = g^\delta$, and (cf. [12]) conditions (a) and (b) imply $g^\delta = g^*$. In the *single-chain* case, which often arises in practice, each gain-optimal policy as a single ergodic chain and perhaps some transient states. In this case, all components of g are identical', so that each pair (i, k) satisfies (34) as an equality and therefore (35).

Most policy iteration routines (cf. [12], [24]) replace $\sum_{j=1}^N v^k_{ij} g_j$ by $v_i{}^k g_i$ in (35), which simplifies the computation. By Theorem 1, this simplification affects only the x_j for transient states. To save time, one can use this simpler variant of (35) to find g^* and a gain-optimal policy and then restart the policy iteration by using (35) to evaluate x_j for the transient states.

Finding a bias-optimal policy by policy iteration. The problem of finding a bias-optimal policy is now parsed into a sequence of at most three simpler Markov renewal programs. Step one is to use policy iteration to find a gain-optimal policy δ and a pair (g, x) satisfying the aforementioned termination conditions.

For step two, define Δ' as the set of all policies η such that $(i, \eta(i))$ satisfies (34) and (35) as equalities for $i = 1, \cdots, N$. Note that δ is in Δ'. Theorems 1 and 2 show that each policy η in Δ' has $g^\eta = g^*$ and $w^\eta = x + \bar{g}^\eta$, where $\bar{g}^\eta$ is the gain rate for a semi-Markov process with rewards having triplet $(Q_0{}^\eta, Q_1{}^\eta, \bar{R}^\eta)$ with $\bar{R}^\eta = c_0{}^\eta - Q_1{}^\eta x = -R_1{}^\eta + Q_2{}^\eta g^* - Q_1{}^\eta x$. Hence, maximizing $\bar{g}^\eta$ over Δ' maximizes w^η over this set. This maximization can be accomplished efficiently by policy iteration; to do so, replace Δ by Δ' and $r_i{}^{\eta(i)}$ by $\bar{R}_i{}^{\eta(i)}$ and the use the original policy iteration routine. If Δ' contains a bias-optimal policy, the altered policy iteration routine will find it.

It sometimes occurs that Δ' contains no bias-optimal policies. The difficulties that can arise involve only the transient states and are identical in the Markov renewal and discrete-time models, for which reason the reader is referred to [11] (particularly Lemma 4 and Lemma 6) for verification of the following three facts. In the single-chain case, every bias-optimal policy is contained in Δ', so that the difficulty cannot arise. In the multichain case, the policy δ that maximizes $\bar{g}^\delta$ over Δ' may have $w_i{}^\delta < w_i{}^*$, but only for a state (or states) i that is transient under every bias-optimal policy. In this case, restart the original policy iteration routine with the attempt to improve on the pair (g^δ, w^δ). This routine will terminate, perhaps immediately, with a bias-optimal policy.

Finding an optimal policy by policy iteration. The procedure just described for finding a policy in Δ_0 readily adapts to find a policy in Δ_1. As Theorem 1 and Theorem 2 suggest, one simply repeats the procedure after making the following four changes. First, replace Δ by the set of policies η that satisfy $(I - Q_0{}^\eta)g^* = b_{-1} = 0$ and $(I - Q_0{}^\eta)w^* = b_0$. Second, with δ as the bias-optimal policy determined above, find any solution x of the equation $(I - Q_0{}^\delta)x = b_1$. Third, set Δ' equal to the set of all policies ξ in (the refined set) Δ such that $(I - Q_0{}^\xi)x = b_1$. Finally, replace $R_0{}^\eta$ by $c_0{}^\eta - Q_1{}^\eta w^*$ and replace $\bar{R}_0{}^\eta$ by $c_1{}^\eta - Q_1{}^\eta x$.

Having made these changes, repeat steps two and three of the procedure for computing a bias-optimal policy. Theorem 1 and Theorem 2 verify that step two maximizes $V_1{}^\eta$ over those η in Δ', and step three plays the same role as before. So, the end result is a policy δ in Δ_1.

The adaptation procedure just described is systematic and can be reapplied as many times as are required, provided $R_i{}^k(\cdot)$ and $Q_{ij}^k(\cdot)$ have the requisite moments. Successive applications of this procedure refine Δ to the set of all optimal policies. If Δ is refined to a single policy, the procedure terminates. Should there be multiple optimal policies, this procedure will find them all, but it will fail to indicate when it has done so.

Adaptation for linear programming. Linear programming can be used to find a gain-optimal policy; see [10], [12], and the references cited there. The problem of finding a bias-optimal policy by linear programming can be parsed into a sequence of three linear programs that are roughly analogous to the three policy iteration routines used to find a bias-optimal policy. In the most general multichain case, the terminal-conditions for linear programming differ from those given earlier for policy iteration. For this reason, definition of Δ' is slightly different. For the same reason, one does not return to the original linear program after completing the second step by linear programming. Rather, one solves a special linear program designed specifically to treat the transient states.

These three linear programs are precisely analogous to the three used in [11] to find a bias-optimal policy in the discrete time-case. The only differences are that g_i is replaced by $\sum_{j=1}^{N} v_{ij}^{k} g_j$ where appropriate and $\bar{R}^{\eta}$ is used in the step two. The details are not reproduced here. As with policy iteration, linear programming can be applied repeatedly in search of an optimal policy.

A lexicographic policy iteration routine. The policy iteration and linear programming formulations just described are organized to strive for a policy in Δ_{-1}, then a policy in Δ_0, then in Δ_1, etc. Miller and Veinott [31] (also, see Veinott [37], [38]) treat the discrete-time case with a policy iteration routine organized along different coordinates. Roughly their routine has the effect of comparing two policies' coefficients lexicographically—first s^{-1}, then s^0, etc. We shall briefly summarize an adaptation of these procedure to Markov renewal programming. For simplicity, preselect an integer p as a truncation constant.

Policy iteration routines alternate two steps. The *policy evaluation* step evaluates policy δ by calculating $V^{\delta}_{-1}, \cdots, V_p{}^{\delta}$, which Lemma 1 and Theorem 1 show how to do. The *policy improvement* step involves test quantities $z(i, k, n)$ defined in terms of ${}_nR_i{}^k$, the nth normalized moment of $R_i{}^k(t)$, and ${}_nQ_i{}^k$, the $1 \times N$ vector whose jth component is the nth normalized moment of $Q_{ij}^{k}(t)$. With ${}_{-1}R_i{}^k = 0$, define $z(i, k, n)$ in a manner related to Theorem 1 as

$$z(i, k, n) = (-1)^n{}_nR_i{}^k + \textstyle\sum_{m=0}^{n+1} (-1)^m{}_mQ_i{}^k V^{\delta}_{n-m} - (V_n{}^{\delta})_i.$$

Note from Theorem 1 that $z(i, k, n) = 0$ whenever $k = \delta(i)$. With $\xi \neq \delta$, policy ξ is considered an *improvement* over δ if for $i = 1, \cdots, N$ either

(a) $\xi(i) = \delta(i)$ or

(b) some n satisfies $z(i, k, n) > 0$ and $z(i, k, m) = 0$ for $-1 < m < n$.

Rule (b) is lexicographic. When $p = 0$, the improvement step is identical to that in Howard's [21] multichain policy iteration routine. Using techniques in [31] or [12], one can show that every ξ that is an improvement over δ has $v^{\xi}(s) > v^{\delta}(s)$ for all sufficiently small s.

If an improvement can be found, the cycle is repeated by evaluating the improved policy. If not, it can be shown that the terminal policy δ is contained in Δ_{p-1}.

5. Discrete-time and exponential cases. The results developed in Sections 3 and 4 simplify considerably for the discrete-time and exponential cases of Markov renewal programming. In both of these models, Theorems 1 and 2 allow computation of the moments of $V^{\delta}(t)$ and of optimal policies. However, the method given in this section is more direct and revealing. Except for part of Theorem 5, the results are not new; cf. [30], [31], [37].

Policy evaluation. The Laurent expansion of $v^{\delta}(s)$ is now obtained, first for the exponential case and then for the discrete time case. As in Section 3, dependence on δ is dropped temporarily. For both cases, let R denote the $N \times 1$ vector of rewards until transition and P denote the transition matrix of the embedded Markov chain. For the exponential case, direct computation yields $r_i(s) = R_i/(1+v_i s)$ and $q_{ij}(s) = P_{ij}/(1+v_i s)$. Define the $N \times N$ matrix D by $D_{ij} = \delta_{ij} v_i$, which allows $r(s)$ and $q(s)$ to be written in matrix notation as the power series

$$(36) \qquad r(s) = \textstyle\sum_{i=0}^{\infty}(-Ds)^i R, \qquad q(s) = \sum_{i=0}^{\infty}(-Ds)^i P\ .$$

To evaluate the coefficients in the Laurent expansion $v(s) = \sum_{i=-1}^{\infty} s^i V_i$, substitute this expression and (36) into (6) and equate coefficients of s. There results

$$(37) \qquad (I-P)V_{-1} = 0$$

$$(38) \qquad (I-P)V_i - \textstyle\sum_{j=1}^{i+1}(-D)^j P V_{i-j} = (-D)^i R, \qquad \text{for} \quad i \geqq 0.$$

Next, we manipulate (37) and (38) into a form akin to (19a), (19b). Note that the equations in (37)–(38) starting with $(I-P)V_{i-1}$ and $(I-P)V_i$ differ mainly by a factor of $-D$. Premultiply the former by D and add the result to the latter; most of the terms cancel, leaving

$$(39) \qquad (I-P)V_i + DV_{i-1} = \delta_{i0} R \qquad \text{for} \quad i \geqq 0.$$

To proceed further, we must introduce the continuous-time analogues of P^* and H. Set $B = D^{-1}(I-P)$. Doob [16] and Kemeny and Snell [27] verify that $P(t) \to A$ as $t \to \infty$, where the $N \times N$ matrix A satisfies $BA = AB = 0$ and $A = A^2$. Moreover, $B+A$ is invertible and Kemeny and Snell call $Z = (B+A)^{-1}$ the *fundamental matrix* for continuous-time Markov chains. With $H = Z-A$, one also has $HB = I-A$.

To evaluate V_i, first premultiply (39) by AD^{-1}, leaving

$$(40) \qquad AV_{i-1} = \delta_{i0} AD^{-1} R, \qquad \text{for} \quad i \geqq 0.$$

To evaluate V_{-1}, premultiply (37) by HD^{-1} to obtain $V_{-1} = AV_{-1} = AD^{-1}R$, the last by (40). For $i \geqq 0$, premultiply (39) by HD^{-1} to obtain $V_i = AV_i - HV_{i-1} + \delta_{i0} HD^{-1} R$. Equation (40) gives $AV_i = 0$, and recursive substitution yields $V_i = (-H)^i HD^{-1} R$. This evaluates all coefficients of the Laurent expansion

$$(41) \qquad v(s) = s^{-1} AD^{-1} R + \textstyle\sum_{i=0}^{\infty}(-sH)^i HD^{-1} R$$

obtained by Veinott [38] by adapting a method of Miller and Veinott [31].

In the discrete-time case, the delay of one time unit in transition and reward produces the transforms $r(s) = R\,e^{-s}$ and $q(s) = P\,e^{-s}$. It proves convenient to introduce the one-period interest rate β equivalent to instantaneous interest rate s.

So, set $e^{-s} = 1/(1+\beta) = 1-\beta+\beta^2- \cdots$, which allows us to express $r(s)$ and $q(s)$ in terms of β as

$$(42) \qquad r[\log(1+\beta)] = \textstyle\sum_{i=0}^{\infty}(-\beta)^i R, \qquad q[\log(1+\beta)] = \textstyle\sum_{i=0}^{\infty}(-\beta)^i P$$

The only differences between (36) and (42) are that D is replaced by I and s by $\log(1+\beta)$. When $D = I$, there is no difference between P^* and A or between the two definitions of H. So, the development for the exponential case repeats exactly, giving

$$(43) \qquad v[\log(1+\beta)] = \beta^{-1}P^*R+\textstyle\sum_{i=0}^{\infty}(-\beta H)^i HR$$

for the discrete time case. Miller and Veinott [31] obtained this expansion by a related method. Subsequently, Veinott [38] observed that the resolvent theory in Kato ([26] pages 36 ff.) also serves this purpose.

Optimization. The optimization procedures also simplify for the discrete and exponential cases. Note that $H\mathbf{1} = 0$ in the discrete-time and exponential cases, and consider

LEMMA 3. *Let X and Y be $N \times N$ matrices such that* $0 = X\mathbf{1} = Y\mathbf{1}$. *If $X^n b = Y^n b$ for $n = 0, \cdots, N-1$, then $X^n b = Y^n b$ for every $n \geqq 0$.*

PROOF. Let $a_n = X^n b$ and $b_n = Y^n b$ for $n = 0, 1, \cdots$. Assume the inductive hypothesis that $a_i = b_i$ for $i = 0, \cdots, n$ and that $a_n = \sum_{i=0}^{n-1} c_i a_i$ for some scalars $c_0, \cdots, c_{n-1}$. Premultiplying this expression by X and Y yields

$$a_{n+1} = \textstyle\sum_{i=1}^{n} c_{i-1}a_i = b_{n+1},$$

which completes the inductive step. The inductive hypothesis is established trivially if the set $\{a_0, \cdots, a_{N-1}\}$ is dependent. So, suppose this set is linearly independent. Since it contains N elements, $\mathbf{1} = \sum_{i=0}^{N-1} d_i a_i$ for some scalars $d_0, \cdots, d_{N-1}$. Since $X\mathbf{1} = Y\mathbf{1} = 0$, premultiplying the preceding expression by X and Y gives

$$0 = \textstyle\sum_{i=1}^{N} d_{i-1}a_i = \sum_{i=1}^{N} d_{i-1}b_i.$$

We cannot have $d_{N-1} = 0$, for this would imply that $\{a_0, \cdots, a_{N-1}\}$ were dependent. So we have $b_N = a_N = \sum_{i=1}^{N-1} (d_{i-1}/d_{N-1})a_i$, which establishes the inductive hypothesis. □

The data needed to compute the Laurent expansion (41) or (43) is all contained in the triplet (P, v, R), where P is the transition matrix of the embedded Markov chain, v is the vector of mean transition times and R is the vector of expected rewards until transition. The discrete-time case has $v = \mathbf{1}$ and $D = I$. Theorem 2 and Corollary 3 simplify in the discrete-time and exponential models to

THEOREM 5. *In the discrete-time (exponential) model $\Delta_{N-1} = \Delta_{\text{opt}} \neq \varnothing$. With $i \geqq 0$ consider an $N \times 1$ vector z satisfying $(I-P)z = -DV_{i-1}+\delta_{i0}R$. Then $V_i - z$ is the gain of the discrete-time (exponential) process with triplet $(P, v, -Dz)$.*

PROOF. Since every policy has a Laurent expansion, Δ_{N-1} and Δ_{opt} are nonempty. Lemma 3 indicates that all policies in Δ_{N-1} must have identical Laurent expansions. So, all are optimal. Subtract the equation $(I-P)z = -DV_{i-1}+\delta_{i0}R$ from (39) to obtain $(I-P)(V_i-z) = 0$. Premultiply this expression by HD^{-1} to obtain $V_i-z = A(V_i-z) = A(-z)$, since (40) gives $AV_i = 0$ for $i \geqq 0$. We then have $V_i-z = AD^{-1}(-Dz)$, so that comparison with (41) and (43) completes the proof. □

The information in Theorem 5 was observed by Miller and Veinott [31], except for our replacement of Δ_N with Δ_{N-1}. We close by noting several differences between the general development in Theorem 2 and this particularization. In the discrete-time and exponential cases, the altered Markov decision process has reward vector $-z$ *per unit time*, rather than $c_n{}^{\delta}-Q_1{}^{\delta}z$ until transition. Only the latter is policy-dependent. Theorem 5 holds for $i \geqq 0$, but Theorem 2 holds for $i = -1$ in addition. This occurs because the interpretation of $-z$ as a reward vector rests upon $AV_i = 0$, which is only true when $i \geqq 0$. In the discrete-time model, the expressions for V_n in Theorems 2 and 5 differ, since V_n is the coefficient of s^n in one and of β^n in the other.

Acknowledgment. While the final responsibility for rigor and clarity of exposition rests with the author, both aspects of this paper benefited from Ward Whitt's insightful reading of an earlier draft.

REFERENCES

[1] Barlow, R. E. and Proschan, F. (1965). *Mathematical Theory of Reliability*. Wiley, New York.

[2] Blackwell, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719–726.

[3] Brown, B. W. (1965). On the iterative method of dynamic programming on a finite space discrete time Markov process. *Ann. Math. Statist.* **36** 1279–1285.

[4] Chitgopekar, S. S. (1969). Continuous time Markovian sequential control processes. *SIAM J. Control* **7** 367–389.

[5] Chung, K. L. (1968). *A Course in Probability Theory*. Harcourt, Brace & World, New York.

[6] Çinlar, E. (1969). Markov renewal theory. *Adv. Appl. Probability* **1** 123–187.

[7] de Cani, J. S. (1964). A dynamic programming algorithm for embedded Markov chains when the planning horizon is at infinity. *Management Sci.* **10** 716–733.

[8] Denardo, E. V. (1967). Contraction mappings in the theory underlying dynamic programming. *SIAM Rev.* **9** 165–177.

[9] Denardo, E. V. (1968). Separable Markov decision problems. *Management Sci.* **14** 451–462.

[10] Denardo, E. V. (1970a). On linear programming in a Markov decision problem. *Management Sci.* **16** 281–288.

[11] Denardo, E. V. (1970b). Computing bias-optimal policies in discrete and continuous Markov decision problems. *Operations Res.* **18** 279–289.

[12] Denardo, E. V. and Fox, B. L. (1968). Multichain Markov renewal programs. *SIAM J. Appl. Math.* **16** 468–487.

[13] Denardo, E. V. and Miller, B. L. (1968). An optimality condition for discrete dynamic programming with no discounting. *Ann. Math. Statist.* **39** 1220–1227.

[14] Derman, C. (1964). On sequential control processes. *Ann. Math. Statist.* **35** 341–349.

[15] Doetsch, G. (1961). *Guide to the Application to Laplace Transforms*. Van Nostrand, Princeton.

[16] Doob, J. (1953). *Stochastic Processes*. Wiley, New York.

[17] Feller, W. (1966). *An Introduction to Probability Theory and its Applications* **2**. Wiley, New York.

[18] Fox, B. L. (1968). (g, w)-optimal in Markov renewal programs. *Management Sci.* **15** 210–212.

[19] Fox, B. L. (1968). Semi-Markov processes: A primer. RM-5803-PR, The RAND Corp., Santa Monica.

[20] Hille, E. (1959). *Analytic Function Theory* **1**. Ginn, New York.

[21] Howard, R. A. (1960). *Dynamic Programming and Markov Processes.* Technology Press of M.I.T., Cambridge.

[22] Howard, R. A. (1963). Semi-Markovian decision processes. *Proc. Internat. Statist. Inst.* Ottawa, Canada.

[23] Hunter, J. J. (1969). On the moments of Markov renewal processes. *Adv. Appl. Probability* **1** 188–210.

[24] Jewell, W. S. (1963). Markov-renewal programming I: formulation, finite return models; Markov-renewal programming II: infinite return models, example. *Operations Res.* **11** 938–971.

[25] Jewell, W. S. (1964). Limiting covariance in Markov renewal processes. ORC 64–16, Operations Research Center, Univ. of California, Berkeley.

[26] Kato, T. (1966). *Perturbation Theory for Linear Operators.* Springer-Verlag, New York.

[27] Kemeny, J. G. and Snell, J. L. (1960). Finite continuous-time Markov chains. *Theor. Probability Appl.* **6** 101–105.

[28] Kemeny, J. G. and Snell, J. L. (1961). *Finite Markov Chains.* Van Nostrand, Princeton.

[29] Kshirsagar, A. M. and Gupta, Y. P. (1967). Asymptotic values of the first two moments in Markov renewal processes. *Biometrica* **54** 597–603.

[30] Miller, B. L. (1968). Finite state continuous time Markov decision processes with an infinite planning horizon. *J. Math. Anal. Appl.* **22** 552–569.

[31] Miller, B. L. and Veinott, A. F. Jr. (1969). Discrete dynamic programming with a small interest rate. *Ann. Math. Statist.* **40** 366–370.

[32] Pyke, R. (1961). Markov renewal processes: definitions and preliminary properties; and Markov renewal processes with finitely many states. *Ann. Math. Statist.* **32** 1231–1259.

[33] Rykov, V. V. (1966). Markov decision processes with finitely many states. *Theor. Probability Appl.* **11** 302–311.

[34] Schweitzer, P. (1963). Private communication to W. S. Jewell [24].

[35] Schweitzer, P. (1965). *Perturbation theory and Markovian decision processes.* Ph.D. dissertation, Massachusetts Inst. of Technology.

[36] Shapley, L. S. (1953). Stochastic games. *Proc. Nat. Acad. Sci. USA* **39** 1095–1100.

[37] Veinott, A. F., Jr. (1966). On finding optimal policies in discrete dynamic programming with no discounting. *Ann. Math. Statist.* **37** 1284–1294.

[38] Veinott, A. F., Jr. (1969a). Discrete dynamic programming with sensitive discount optimality criteria. *Ann. Math. Statist.* **40** 1635–1660.

[39] Veinott, A. F., Jr. (1969b). Discrete dynamic programming with sensitive average optimality criteria. Operations Research Society of America, 36th National Meeting.

[40] Widder, D. F. (1946). *The Laplace Transform.* Princeton Univ. Press.

[41] Zachrisson, L. E. (1964). Markov games. In *Advances in Game Theory*, ed. M. Dresher, L. S. Shapley and A. W. Tucker. Princeton Univ. Press.