The Annals of Mathematical Statistics
1969, Vol. 40, No. 5, 1635–1660

DISCRETE DYNAMIC PROGRAMMING WITH SENSITIVE DISCOUNT OPTIMALITY CRITERIA[1]

By Arthur F. Veinott, Jr.

Stanford University

1. Introduction. This paper is concerned with stationary finite state and action Markovian decision processes where future rewards are discounted and the transition matrices are substochastic. The discrete time parameter case is treated in Sections 2–4 with analogous results being obtained for the continuous time parameter case in Section 5.

In Section 2 we generalize some results of Shapley [32], Bellman [4], Howard [20], Blackwell [5], Eaton and Zadeh [15], Derman [8], and Denardo [7] which are concerned with the use of the methods of successive approximation and policy improvement in finding a policy π that maximizes the expected infinite horizon discounted reward $V_\rho(\pi)$ where ρ is the rate of interest. Our contribution is to the case where $-1 < \rho \leqq 0$, e.g., corresponding to inflation. One main new result in this section, Corollary 4, is that the series defining $V_\rho(\pi)$ converges absolutely for every (some) π for all one period rewards if and only if the same assertion is true when π ranges over the stationary policies. This result is proved by dynamic programming methods.

Up to now we have supposed the interest rate to be fixed. A stronger concept of optimality would be to find a policy that maximizes $V_\rho(\cdot)$ for a sufficiently small interval of interest rates. Blackwell [5] has shown that there is a stationary policy maximizing $V_\rho(\cdot)$ for all $\rho > 0$ close enough to zero. Recently Miller and the author [30] discovered a policy improvement algorithm for finding such a policy. The algorithm exploits the fact that for stationary policies the Laurent expansion about the origin of the return $V_\rho(\cdot)$ has a simple form. Analogous results hold in the transient case where one seeks to maximize $V_\rho(\cdot)$ for all $\rho < 0$ close enough to zero. Section 3 consists of an expository development of the Laurent expansion mentioned above together with some new results on properties of its coefficients.

In Section 4 we introduce the following new optimality criteria. For each $n = -1, 0, 1, \cdots$, a policy π^* is called $n^\pm$ discount optimal if (for n^-, we consider only the transient case)

$$\liminf_{\rho \to 0\pm} |\rho|^{-n}[V_\rho(\pi^*) - V_\rho(\pi)] \geqq 0, \quad \text{for all} \quad \pi.$$

The sensitivity of this criterion increases with n. And when n is as large as the number of states, the criterion is shown to be equivalent to Blackwell's criterion

Received 3 September 1968; revised February 1969.

[1] This research was supported by the National Science Foundation under Grant GK-1420, the Office of Naval Research under Contract Nonr-225(77) (NR-347-010), and the IBM Research Center.

cited above. We also characterize the class of stationary $n^{\pm}$ discount optimal policies and show that the algorithm given in [30] can be terminated more rapidly when only $n^{\pm}$ discount optimal policies are sought. This leads to efficient ways of implementing that algorithm. In particular, Blackwell's [5] and the author's [34] algorithms for respectively finding -1^{+} and 0^{+} discount optimal policies emerge as special cases.

In [35] we investigate a family of sensitive averaging criteria and obtain the algorithm given in [30] in a quite different way. We also establish the equivalence of $n^{\pm}$ discount optimality with the corresponding averaging criterion.

2. Maximal expected reward: discrete parameter.

Preliminaries. Consider a system which is observed at each of a sequence of points in time labeled 1, 2, $\cdots$. At each of those points the system is found to be in one of S states labeled 1, 2, $\cdots$, S or to have "stopped". Each time the system is observed in state s, an action a is chosen from a finite set A_s of possible actions and a reward $r(s, a)$ is received. The conditional probability that the system is observed in state t at time $N + 1$ given that it is found in state s at time N, that action a is taken at that time, and given the observed states and actions taken at times 1, 2, $\cdots$, $N - 1$, is assumed to be a nonnegative function $p(t \mid s, a)$ depending only on t, s, and a. The corresponding conditional probability that the system is observed to have stopped at time $N + 1$ is $1 - \Sigma_t p(t \mid s, a)$. Once the system is observed to have stopped, it remains stopped and earns no rewards.

Let $F = \mathbf{X}_{s=1}^{S} A_s$. A *policy* is a sequence $\pi = (f_1, f_2, \cdots)$ of elements f_N of F. Using the policy π means that if the system is observed in state s at time N, the action chosen at that time is $f_N(s)$, the sth coordinate of f_N. Let $f^{\infty} \equiv (f, f, \cdots)$ and call f^{∞} a *stationary policy*. If π is a policy, let π^N denote the first N components of π and call ${}^N\pi \equiv (\pi^N, \pi^N, \cdots)$ a *periodic policy*. Let f^N denote the first N components of f^{∞}.

For any $f \varepsilon F$, let $r(f)$ be the $S \times 1$ column vector whose sth component is $r(s, f(s))$, and let $P(f)$ be the $S \times S$ matrix whose stth element is $p(t \mid s, f(s))$. If $\pi = (f_1, f_2, \cdots)$, let $P^N(\pi) = P(f_1) \cdots P(f_N)$. Thus $P^N(f^{\infty}) = P(f)^N$ and $P^0(\pi) = I$. Call π *transient* if $\sum_{N=0}^{\infty} P^N(\pi)$ converges. If $\pi = (f_i)$ is transient, the S-vector $V(\pi)$ of expected total returns starting from each state, given by

$$V(\pi) \equiv \sum_{N=0}^{\infty} P^N(\pi) r(f_{N+1}), \tag{1}$$

converges absolutely.

In the remainder of this section we shall *drop* the assumption that the row sums of $P(\cdot)$ be one or less; however, we *retain* the hypothesis that $P(\cdot)$ is nonnegative. The definition of a transient policy given above is equally valid in this case. Of course, we can then no longer think of the elements of $P(f)$ as probabilities. But there are many alternate interpretations of the generalized model which, for brevity, we do not mention here.

If $B = (b_{ij})$ is a complex matrix, let $\|B\| \equiv \max_i \Sigma_j |b_{ij}|$ be its norm. Assume B is $S \times S$ and denote by $\sigma(B)$ its spectrum and $|\sigma(B)|$ its spectral radius. In

the sequel, we often use the well-known fact [22, page 30] that the following are equivalent: (i) $|\sigma(B)| < 1$, (ii) $\|B^N\| < 1$ for some $N \geqq 1$, (iii) $B^N \to 0$ as $N \to \infty$, and (iv) $\sum_{N=0}^{\infty} B^N$ converges absolutely. These conditions imply that $[I - B]$ is nonsingular and its inverse has the Neumann series expansion given in (iv) above. If also B is real and non-negative, the above four conditions are equivalent to: (v) $[I - B]$ is nonsingular and has a non-negative inverse. If further B is *substochastic*, i.e., $\|B\| \leqq 1$, then the five aforementioned conditions are equivalent to: (vi) $\|B^S\| < 1$ and (vii) $[I - B]$ is nonsingular. Thus [if $\|P(f)\| \leqq 1$], f^∞ is transient if and only if $P(f)$ satisfies any one of the conditions [(i)–(vii)] (i)–(v) given above.

Maximal Expected Reward. Under the hypothesis $\|P(f)\| < 1$ for $f \varepsilon F$, Shapley [32] showed that there is a stationary policy maximizing $V(\cdot)$ over all stationary policies by the method of successive approximations. Howard [20, pages 83 ff] devised a finite policy improvement method for finding such a policy. Blackwell [5] showed that $V(\cdot)$ assumes its maximum over all policies among the stationary policies.

Our aim here is to establish these same results under the weaker hypothesis that each stationary policy is transient. Partial results in this direction have been obtained by Eaton and Zadeh [15], Derman [8], and especially Denardo [7].

The following lemma is basic to most of what follows. It generalizes a result of Howard [20, page 87] from stationary to nonstationary policies.

LEMMA 1. *If* $\pi = (g_i)$ *and* $\pi^* \equiv (f_i)$ *are transient, then*

$$V(\pi) - V(\pi^*) = \sum_{N=0}^{\infty} P^N(\pi) v(g_{N+1}, \pi^*) \tag{2}$$

where $v(g, \pi^*) \equiv V(g, \pi^*) - V(\pi^*)$. *If also* $\pi = g^\infty$, *then*

$$V(g^\infty) - V(\pi^*) = [I - P(g)]^{-1} v(g, \pi^*). \tag{3}$$

PROOF.

$$V(\pi) - V(\pi^*) = \sum_{N=0}^{\infty} [V(\pi^{N+1}, \pi^*) - V(\pi^N, \pi^*)] = \sum_{N=0}^{\infty} P^N(\pi) v(g_{N+1}, \pi^*).$$

One immediate consequence of this result is the following.

LEMMA 2. *Suppose* $f \varepsilon F$ *and every stationary policy is transient. Then either* $v(g, f^\infty) > 0$ *for some* $g \varepsilon F$ *or* $v(g, f^\infty) \leqq 0$ *for all* $g \varepsilon F$. *In the former case* $V(g^\infty) > V(f^\infty)$. *The latter case occurs if and only if* $V(g^\infty) \leqq V(f^\infty)$ *for all* $g \varepsilon F$.

The next two corollaries come from Denardo [7]. They generalize results of Shapley [32].

COROLLARY 1. *If every stationary policy is transient, there is a stationary policy that maximizes* $V(\cdot)$ *over the class of stationary policies.*

PROOF. Apply the policy improvement method implied by Lemma 2.

Because of Corollary 1 we may define $V^* \equiv \max_{f \varepsilon F} V(f^\infty)$ provided every stationary policy is transient. Also define the operator $\mathfrak{R}$ mapping E^S into itself by

$$\mathfrak{R}V \equiv \max_{g \varepsilon F} [r(g) + P(g)V], \qquad V \varepsilon E^S. \tag{4}$$

Notice that $\mathfrak{R}$, and hence $\mathfrak{R}^n$, is monotone (nondecreasing). The next result is an easy consequence of Corollary 1 and Lemma 2.

COROLLARY 2. *If every stationary policy is transient, V^* is the unique fixed point of $\mathfrak{R}$.*

We say that the dynamic program $(\tilde{r}(\cdot), \tilde{P}(\cdot))$ is *positively similar* to $(r(\cdot), P(\cdot))$ if both are defined on F and there is an $S \times S$ diagonal matrix B having positive diagonal elements for which $\tilde{r}(f) = Br(f)$ and $\tilde{P}(f) = BP(f)B^{-1}$ for all $f \in F$. Positive similarity is evidently an equivalence relation on the class of dynamic programs on F. On using some obvious definitions, we have $\tilde{P}^N(\pi) = BP^N(\pi)B^{-1}$, $\tilde{\mathfrak{R}}^n = B\mathfrak{R}^nB^{-1}$, and when π is transient, $\tilde{V}(\pi) = BV(\pi)$.

A property of a dynamic program is said to be *invariant* if it holds in all positively similar dynamic programs or holds in none of them. The following examples of invariant properties of dynamic programs will be used frequently in the sequel: (i) a policy is transient, (ii) a policy maximizes the total expected reward, (iii) the iterates of the "optimal return" operator (4) converge geometrically to the unique fixed point of the operator. A property that is not invariant is the magnitude of the norm of a transition matrix.

LEMMA 3. (Hoffman) *For every $\epsilon > 0$ and every dynamic program $(r(\cdot), P(\cdot))$, there is a positively similar dynamic program $(\tilde{r}(\cdot), \tilde{P}(\cdot))$ for which* $\max_g \|\tilde{P}(g)\| < \max_g |\sigma(P(g))| + \epsilon$.

PROOF. It suffices to show that $\max_g |\sigma(P(g))| < 1$ implies $\max_g \|\tilde{P}(g)\| < 1$. By Corollary 2, there is a unique S-vector v, $v \geqq 1$, satisfying

$$v = \max_{g \in F} [1 + P(g)v]. \tag{5}$$

Let B be the diagonal matrix whose sth diagonal element is v_s^{-1}. Pre-multiplying (5) by B and subtracting the positive vector $B1$ gives $1 - B1 = \max_g \tilde{P}(g)1$, which completes the proof.

The conclusion of Lemma 3 cannot be strengthened to $\max_g \|\tilde{P}(g)\| = \max_g |\sigma(P(g))|$ as the following example shows.

EXAMPLE. If P is the matrix with rows (1, 1) and (0, 1), then $|\sigma(P)| = 1$. Also the rows of every positively similar $\tilde{P}$ are of the form $(1, b)$ and (0, 1) with $b > 0$ so $\|\tilde{P}\| = 1 + b > 1$.

The next corollary sharpens and generalizes results of Shapley [32] and Denardo [7, pages 169–170].

COROLLARY 3. *Suppose every stationary policy is transient. Then $\mathfrak{R}^nV \to V^*$ as $n \to \infty$. Moreover, for each α for which $\max_g |\sigma(P(g))| < \alpha < 1$ and each V, there is a constant K such that $\|\mathfrak{R}^nV - V^*\| \leqq K\alpha^n$, $n = 0, 1, \cdots$.*

PROOF. By Lemma 3 there is a positively similar dynamic program $(\tilde{r}(\cdot), \tilde{P}(\cdot))$ with $\max_g \|\tilde{P}(g)\| < \alpha$. Then as Shapley [32] has shown, $\tilde{\mathfrak{R}}$ is a contraction with modulus α under $\|\cdot\|$. Hence, by the Banach fixed point theorem, the conclusions of the corollary hold for $(\tilde{r}(\cdot), \tilde{P}(\cdot))$. Thus, by invariance they hold for $(r(\cdot), P(\cdot))$, which completes the proof.

The equivalence of 1^0 and 3^0 of the next corollary reading with "some" was shown by Derman [8, page 19] under the hypothesis that each transition matrix

is substochastic. This last hypothesis appears to be needed for his proof, although as we show here, it is not essential to the result.

COROLLARY 4. *The following four statements are equivalent.*

1^0. *Every (some) stationary policy is transient.*

2^0. *Every (some) periodic policy is transient.*

3^0. *Every (some) policy is transient.*

4^0. *For some* $N \geqq 1$, $\|P^N(\pi)\| < 1$ *for every (some)* π.

If also $\|P(g)\| \leqq 1$ *for all* $g \varepsilon F$, *the above are equivalent to*

5^0. $\|P^S(\pi)\| < 1$ *for every (some)* π.

PROOF. We establish the equivalence of 1^0–4^0 first. On reading with "every", notice that $4^0 \Rightarrow 3^0 \Rightarrow 2^0 \Rightarrow 1^0$. Also $1^0 \Rightarrow 4^0$ by Lemma 3 and the invariance of these properties.

Turning now to the results reading with "some", it is clear that $1^0 \Rightarrow 2^0 \Rightarrow 3^0 \Rightarrow 4^0$. Therefore, suppose 4^0 holds. Let $r(g) = -1$ for all g. Then by 4^0, ${}^N\pi$ is transient so $V({}^N\pi)$ is finite. Now, as in Strauch [33], $0 \geqq \mathfrak{R}^n 0 \geqq V({}^N\pi)$ and $\mathfrak{R}^n 0 \downarrow -V(\geqq V({}^N\pi))$, say, as $n \to \infty$. Moreover, because $\mathfrak{R}$ is continuous, $-V$ is a fixed point of $\mathfrak{R}$. Thus, for some f, $V = 1 + P(f)V = \cdots = \sum_{N=0}^{n} P(f)^N 1 + P(f)^{n+1}V \geqq \sum_{N=0}^{n} P(f)^N 1$, which implies f^∞ is transient and proves 1^0.

We now give an independent proof of the equivalence of 1^0–5^0 for the case $\|P(g)\| \leqq 1$ for all $g \varepsilon F$. Reading first with "every", we have $5^0 \Rightarrow 4^0 \Rightarrow 3^0 \Rightarrow 2^0 \Rightarrow 1^0$. Therefore, suppose 1^0 holds.

It will be convenient to say the system is in state $S + 1$ when it is stopped. Let $\pi = (f_i)$ be any policy and s, $1 \leqq s \leqq S$, any state. Let $T^1 = \{s\}$ and let T^N be the set of states at time $N (= 2, 3, \cdots)$ that are accessible from state s at time 1 when using the policy π.

It suffices to show state $S + 1$ is accessible from (each) state s in at most S steps under π, for then $\|P^S(\pi)\| < 1$. This will be so if $N \geqq 1$ and $(S + 1) \varepsilon \bigcup_{i=1}^{N} T^i$ imply $T^{N+1} \not\subset \bigcup_{i=1}^{N} T^i$. We prove this last assertion by contradiction. Thus suppose $T^{N+1} \subset \bigcup_{i=1}^{N} T^i$ for some $N \geqq 1$. Then define $g \varepsilon F$ by the rule:

$$g(t) = f_j(t) \quad \text{for} \quad t \varepsilon T^j - \bigcup_{i=1}^{j-1} T^i, \quad j = 1, 2, \cdots, N,$$

and arbitrarily otherwise. Under g^∞, the only states accessible from states in $\bigcup_{i=1}^{N} T^i$ are those in $\bigcup_{i=1}^{N} T^i$. In particular, state $S + 1$ is not accessible from any state in $\bigcup_{i=1}^{N} T^i$ under g^∞. Thus g^∞ is not transient, which is a contradiction and completes the proof.

It remains to consider the result reading with "some". Clearly, $1^0 \Rightarrow 5^0 \Rightarrow 2^0 \Rightarrow 3^0 \Rightarrow 4^0$. Thus, suppose 4^0 holds. Then, there is a smallest integer K for which $\|P^K(\pi)\| < 1$. Let T^N be the set of states at time N from which state $S + 1$ is accessible by time $K + 1$. By definition of K, $T^1 = \{1, \cdots, S\}$. Write $\pi = (f_i)$ and define $g \varepsilon F$ by the rule:

$$g(t) = f_j(t) \quad \text{for} \quad t \varepsilon T^j - \bigcup_{i=j+1}^{K} T^i, \quad j = 1, 2, \cdots, K.$$

Under g^∞, state $S + 1$ is accessible from every state so g^∞ is transient, completing the proof.

Since ${}^N\pi$ is transient if and only if $|\sigma(P^N(\pi))| < 1$, it is clear that the equivalence of 1^0 and 2^0 of Corollary 4 can be restated in the following interesting form. (c.f., Bellman [4, pages 328–332]).

COROLLARY 5.

$$\min_g |\sigma(P(g))| \leqq |\sigma(P^N(\pi))|^{1/N} \leqq \max_g |\sigma(P(g))|$$

for every $N \geqq 1$ *and* π. *Moreover, the upper (lower) bound is attained for every* $N \geqq 1$ *by any stationary policy* $\pi = g^\infty$ *for which* g *achieves the maximum (minimum) on the right (left).*

With the aid of Corollary 4, we see from Lemma 1 that Lemma 2 generalizes immediately as follows.

THEOREM 1. *Suppose every stationary policy is transient and* π^* *is a policy. If* $v(g, \pi^*) > 0$ *for some* $g \varepsilon F$, *then* $V(g^\infty) > V(\pi^*)$. *Also* $v(g, \pi^*) \leqq 0$ *for all* $g \varepsilon F$ *if and only if* $V(\pi) \leqq V(\pi^*)$ *for all* π.

By combining this theorem with Corollary 1 and Lemma 2, we get the following immediate generalization of Corollary 1.

COROLLARY 6. *If every stationary policy is transient, there is a stationary policy that maximizes* $V(\cdot)$ *over the class of all policies.*

Theorem 1 and Corollary 6 were established by Blackwell [5] for the case $\|P(g)\| < 1$ for all $g \varepsilon F$ and by Denardo [7] under a hypothesis something like the "every" part of 4^0 of Corollary 4.

3. Laurent expansion of the resolvent: discrete parameter. If Ω is a set of complex numbers, denote its complement by Ω^c. If B is an $S \times S$ complex matrix, then the matrix function $R_\lambda(B) \equiv [\lambda I - B]^{-1}$, defined for $\lambda \varepsilon \sigma(B)^c$, is called the *resolvent* of B. An excellent account of the properties of resolvents of finite matrices relevant for this paper is given in Kato [22, pages 36 ff.]. (However, the reader should note that Kato calls $[B - \lambda I]^{-1}$ the resolvent of B. Our terminology follows that used in Dunford and Schwarz [12, page 566].) This section is largely expository with probably only (9), (25), and Lemma 7 being new.

The principal purpose of this section is to develop and interpret the coefficients of the Laurent expansion [26, page 117] about the origin of the resolvent of $Q \equiv P - I$ where P is an $S \times S$ substochastic matrix (Theorem 2 below). This expansion is obtained in [22, page 39] by means of contour integration and in [30] by elementary matrix calculations. In this section we pursue the latter method, obtaining here more detailed results. The expansion plays a key role in the development of algorithms for finding $n^{\pm}$ discount optimal policies in Section 4.

The resolvent. The resolvent $R_\lambda(B) \equiv R_\lambda$ of B is easily seen [22, page 36] to satisfy the *resolvent equation*

$$R_\lambda - R_\mu = (\mu - \lambda) R_\lambda R_\mu \tag{6}$$

for $\lambda, \mu \varepsilon \sigma(B)^c$. Moreover, if $\lambda \varepsilon \sigma(B)^c$, then [22, page 37]

$$R_\lambda^{n+1} = (-1)^n (n!)^{-1} \frac{d^n}{d\lambda^n} R_\lambda, \qquad n = 1, 2, \cdots. \tag{7}$$

If $|\sigma(\beta B)| < 1$ where $\beta \equiv \lambda^{-1}$, then $\lambda \varepsilon \sigma(B)^c$ and one has the Neumann series expansion $R_\lambda = \sum_{i=0}^{\infty} \beta^{i+1} B^i$. Thus on differentiating with respect to λ and using (7) we get

$$R_\lambda{}^{n+1} = \sum_{i=0}^{\infty} \binom{i+n}{n} \beta^{i+n+1} B^i, \qquad n = 0, 1, \cdots, \tag{8}$$

provided $|\sigma(\beta B)| < 1$. This last result must be refined for our applications.

It is observed in [23, page 23] that $1 \varepsilon \sigma(B)^c$ provided

$$\lim_{N\to\infty} (N+1)^{-1} \sum_{i=0}^{N} B^i = 0.$$

In this event the series on the right of (8) may be divergent when $\lambda = \beta = 1$. However, since R_λ is continuous in $\lambda \varepsilon \sigma(B)^c$ by Cramér's rule, it is clear upon letting $\beta \uparrow 1$ in (8) that the series on the right of (8) is summable (A) (Abel) to $R_1{}^{n+1}$ when $\lambda = \beta = 1$. (The summability terminology used here is that of Hardy [18, pages 7, 96].) Actually we shall show that an even stronger statement is true, viz., the indicated series is summable $(C, n+1)$ (Cesaro) to $R_1{}^{n+1}$ (see (9) below). This result is known where $S = 1$ (the scalar case) and $n = 0, 1, \cdots$ [25, page 481], and where $S > 1$ (the matrix case) and $n = 0$ [23, page 23]. See [25, pages 479, 489] for examples of such series in the scalar case that are not (C, n) summable. Thus the result is sharp.

LEMMA 4. *The following are equivalent*

1^0. $\lim_{n\to\infty} (N+1)^{-1} \sum_{i=0}^{N} B^i = 0$.

2^0. $1 \varepsilon \sigma(B)^c$ *and* $\lim_{N\to\infty} N^{-1} B^N = 0$.

If either 1^0 *or* 2^0 *holds, then*

$$R_1{}^{n+1} = \sum_{i=0}^{\infty} \binom{i+n}{n} B^i \qquad (C, n+1), \qquad n = 0, 1, \cdots. \tag{9}$$

PROOF. Evidently

$$[I - B] \sum_{i=0}^{N} B^i = I - B^{N+1}, \qquad N = 0, 1, \cdots. \tag{10}$$

That 2^0 implies 1^0 follows by premultiplying (10) by R_1, dividing by $N+1$, and letting $N \to \infty$. It remains to show that 1^0 implies 2^0 and (9). On dividing (10) by $N+1$ and letting $N \to \infty$, we see that 1^0 implies the second assertion in 2^0. Averaging (10) we get

$$[I - B][(N+1)^{-1} \sum_{j=0}^{N} \sum_{i=0}^{j} B^i] = I - B[(N+1)^{-1} \sum_{i=0}^{N} B^i]. \tag{11}$$

Since the right side of (11) converges to I as $N \to \infty$, $[I - B]$ must be nonsingular so $1 \varepsilon \sigma(B)^c$ and 2^0 holds. Thus on premultiplying (11) by R_1 and letting $N \to \infty$ we see that (9) holds for $n = 0$.

In order to complete the proof it will be convenient to use the $(t+1)$-fold summation notation $\Sigma^{N,t}$ defined by

$$\sum_{i=0}^{N,t} (\cdot)_i \equiv \sum_{i_t=0}^{N} \cdots \sum_{i_2=0}^{i_3} \sum_{i_1=0}^{i_2} (\cdot)_i.$$

It will suffice to show by induction on n that

$$\lim_{N\to\infty} \binom{N+n+1}{n+1}^{-1} \sum_{i=0}^{N,t} \binom{i+n}{n} B^i = 0, \qquad t = 0, 1, \cdots, n; \tag{12}$$

$$= R_1{}^{n+1}, \qquad t = n+1;$$

for $n = 0, 1, \cdots$. We have just shown (12) holds for $n = 0$ and $t = 1$. Also (12) holds by hypothesis for $n = t = 0$. Suppose now that (12) holds for the integer $n - 1 \geqq 0$ and consider n. We establish (12) for $0 \leqq t \leqq n + 1$ by induction on t. One has the following generalization of (10),

$$[I - B] \sum_{i=0}^{N} \binom{i+n}{n} B^i = \sum_{i=0}^{N} \binom{i+n-1}{n-1} B^i - B\binom{N+n}{n} B^N, \tag{13}$$

for $N = 0, 1, \cdots$. Premultiplying (13) by $K_N \equiv \binom{N+n+1}{n+1}^{-1} R_1$ and letting $N \to \infty$, it is clear from the induction hypothesis and 2^0 that (12) holds for $t = 0$. Suppose now (12) holds for some $t - 1$, $0 \leqq t - 1 < n + 1$, and consider t. Summing (13) yields

$$[I - B] \sum_{i=0}^{N,t} \binom{i+n}{n} B^i = \sum_{i=0}^{N,t} \binom{i+n-1}{n-1} B^i - B \sum_{i=0}^{N,t-1} \binom{i+n}{n} B^i, \tag{14}$$

for $N = 0, 1, \cdots$. On premultiplying (14) by K_N and letting $N \to \infty$, we see that (12) holds for t by the induction hypothesis and the fact [18, page 100] that a series that is (C, n) summable is also $(C, n + 1)$ summable to the same sum. This completes the proof.

The reduced resolvent. In the remainder of this section we shall be concerned with an $S \times S$ substochastic matrix P and the Cesaro limit

$$P^* \equiv \lim_{N \to \infty} (N + 1)^{-1} \sum_{i=0}^{N} P^i$$

which is known [11, page 175] to exist. Moreover, P^* uniquely satisfies

$$PP^* = P^*P = P^*P^* = P^*. \tag{15}$$

A useful implication of (15) is

$$(P - P^*)^n = P^n - P^*, \qquad n = 1, 2, \cdots. \tag{16}$$

It follows from (16) and Lemma 4 on setting $B = P - P^*$ that $1 \varepsilon \sigma(P - P^*)^c$, or equivalently $0 \varepsilon \sigma(Q - P^*)^c$. Also from (15), $P^*Q = 0$. Combining these facts, we see that $P^* = 0$ if and only if $0 \varepsilon \sigma(Q)^c$. These results will be used frequently in the sequel.

For any square matrix B and scalar δ, let $\sigma_\delta(B) = \sigma(B) - \{\delta\}$.

Although Q is singular when $P^* \neq 0$, we show below there is a matrix function H_ρ defined for $\rho \varepsilon \sigma_0(Q)^c$ and having a rational representation on its domain, for which $R_\rho(Q) = \rho^{-1}P^* + H_\rho$ for $\rho \varepsilon \sigma(Q)^c$. The matrix function H_ρ is called the *reduced resolvent* of Q [22, page 40]. In the remainder of this section we shall develop a number of its properties, many of which appear in [22, pages 36 ff].

LEMMA 5. (*a*) *There is a unique* (*common*) *solution* H_ρ *of the two systems*

$$\begin{bmatrix} \rho I - Q \\ P^* \end{bmatrix} H_\rho = \begin{bmatrix} I - P^* \\ 0 \end{bmatrix} = H_\rho \begin{bmatrix} \rho I - Q \\ P^* \end{bmatrix} \tag{17}$$

if and only if $\rho \varepsilon \sigma_0(Q)^c$. *Moreover,* H_ρ *has a rational representation in* ρ *on* $\sigma_0(Q)^c$.

(b) *If* $\rho \varepsilon_\sigma (Q)^c$, *then* $\rho^{-1}P^*$ *is finite and* $R_\rho \equiv R_\rho(Q)$ *satisfies*

$$H_\rho = [I - P^*]R_\rho = R_\rho - \rho^{-1}P^* = R_\rho[I - P^*]. \tag{18}$$

PROOF. For brevity we note only that the unique solution to (17) is given by $H_\rho \equiv R_\rho(Q)(I - P^*)$ for $\rho \varepsilon \sigma(Q)^c - \{0\}$ and (recalling $0 \varepsilon \sigma(Q - P^*)^c$) $H_0 \equiv R_0(Q - P^*)(I - P^*)$ for $\rho = 0$.

LEMMA 6. *If* $\rho, \eta \varepsilon \sigma_0(Q)^c$, *then*

$$H_\rho - H_\eta = (\eta - \rho)H_\rho H_\eta \qquad \textit{(resolvent equation)} \tag{19}$$

and

$$H_\rho H_\eta = H_\eta H_\rho . \tag{20}$$

If $\eta \varepsilon \sigma_0(Q)^c$ *and* $1 \varepsilon \sigma((\eta - \rho)H_\eta)^c$, *then* $\rho \varepsilon \sigma_0(Q)^c$ *and*

$$H_\rho = [I - (\eta - \rho)H_\eta]^{-1}H_\eta = H_\eta[I - (\eta - \rho)H_\eta]^{-1}. \tag{21}$$

If $\eta \varepsilon \sigma_0(Q)^c$ *and* $|\sigma((\eta - \rho)H_\eta)| < 1$, *then* $\rho \varepsilon \sigma_0(Q)^c$ *and*

$$H_\rho = \sum_{n=0}^{\infty} (\eta - \rho)^n H_\eta^{n+1}. \tag{22}$$

If $\rho \varepsilon \sigma_0(Q)^c$, *then*

$$H_\rho^{n+1} = (-1)^n(n!)^{-1}\frac{d^n}{d\rho^n}H_\rho, \qquad n = 1, 2, \cdots . \tag{23}$$

If $\rho \neq -1$ *and* $|\sigma(\beta(P - P^*))| < 1$ *where* $\beta \equiv (1 + \rho)^{-1}$, *then* $\rho \varepsilon \sigma_0(Q)^c$ *and*

$$H_\rho^{n+1} = \sum_{i=0}^{\infty} \binom{i+n}{n}\beta^{i+n+1}(P^i - P^*), \qquad n = 0, 1, \cdots . \tag{24}$$

On letting $H \equiv H_0$,

$$H^{n+1} = \sum_{i=0}^{\infty} \binom{i+n}{n}(P^i - P^*) \qquad (C, n+1), \qquad n = 0, 1, \cdots . \tag{25}$$

PROOF. The formulas (19)–(23) follow readily from Lemma 5.

The hypothesis of (24) implies $1 \varepsilon \sigma(\beta(P - P^*))^c$ which, by Lemma 5, is equivalent to $\rho \varepsilon \sigma(Q - P^*)^c \subset \sigma_0(Q)^c$. Also from (16)

$$\begin{aligned} H_\rho &= R_\rho(Q - P^*) - \beta P^* = R_{1+\rho}(P - P^*) - \beta P^* \\ &= \sum_{i=0}^{\infty}\beta^{i+1}(P - P^*)^i - \beta P^* = \sum_{i=0}^{\infty}\beta^{i+1}(P^i - P^*), \end{aligned}$$

which verifies (24) for $n = 0$. Differentiating both sides of (24) (for $n = 0$) with respect to ρ and using (23) establishes (24).

To prove (25) recall on setting $B = P - P^*$, that 1^0 of Lemma 4 holds. Thus $R_1(P - P^*)^{n+1} = \sum_{i=0}^{\infty}\binom{i+n}{n}(P - P^*)^i \qquad (C, n+1)$, which, in view of (16) and $H^{n+1} = R_1(P - P^*)^{n+1} - P^*$, establishes (25) and completes the proof.

By combining (18) and (22) we obtain the promised Laurent expansion of $R_\rho(Q)$. The reader should note the probabilistic interpretation (24), (25) of the powers of H appearing in the expansion.

THEOREM 2. *If* $\rho \neq 0$ *and* $|\sigma(\rho H)| < 1$, *then* $\rho \varepsilon \sigma(Q)^c$ *and*

$$R_\rho(Q) = \rho^{-1}P^* + \sum_{n=0}^{\infty}(-\rho)^n H^{n+1}. \tag{26}$$

The next lemma shows that $P^* + \rho H$ inherits many of the properties of $R_\rho(Q)$

for all small enough $\rho > 0$. The result enables us to avoid examining the chain structure of P in the sequel.

LEMMA 7. *For all small enough $\rho > 0$, the matrix $P^* + \rho H$ is non-negative, has positive diagonal elements, and is nonsingular.*

PROOF. From Theorem 2 and the fact $R_\rho(Q) = \sum_{i=0}^{\infty} \beta^{i+1} P^i$ for $\rho > 0$, we have

$$I = \lim_{\rho \to 0+} (1 + \rho)^{-1} I \leqq \lim \inf_{\rho \to 0+} R_\rho(Q) = \lim \inf_{\rho \to 0+} [\rho^{-1} P^* + H]$$

so $\rho^{-1} P^* + H$, and hence $P^* + \rho H$, is non-negative and has positive diagonal elements for all small enough $\rho > 0$.

The nonsingularity of $P^* + \rho H$ follows from the fact $\theta(\rho) \equiv \det [P^* + \rho H]$ is a polynomial of degree S and so has finitely many real roots or vanishes identically. The latter possibility cannot occur since $\theta(1) = \det R_1(P - P^*) \neq 0$, which completes the proof.

4. Discount optimality: discrete parameter. In this section we resume consideration of the model of Section 2. We assume throughout that $\|P(g)\| \leqq 1$ for all $g \in F$ without further mention. Also suppose there is a (possibly) negative (real) rate of interest ρ, $-1 < \rho < \infty$. We suppress the dependence of the discount factor $\beta \equiv (1 + \rho)^{-1}$ on ρ in the sequel for simplicity.

The S-vector of expected total discounted returns starting from each state and using the policy $\pi = (f_i)$ is $V_\rho(\pi) = \sum_{N=1}^{\infty} \beta^N P^{N-1}(\pi) r(f_N)$ provided the series $\sum_{N=1}^{\infty} \beta^N P^{N-1}(\pi)$ converges. Since $\rho > -1$, $V_\rho(\pi)$ can be expressed in the form (1) by replacing $P(f)$ and $r(f)$ everywhere in (1) by $\beta P(f)$ and $\beta r(f)$. Thus all results of Section 2 immediately apply to the discount problem. In particular by Corollary 4, the series $\sum_{N=1}^{\infty} \beta^N P^{N-1}(\cdot)$ converges for every policy if and only if it converges for every stationary policy. This is evidently always true if $\rho > 0$. It is also true when $\rho \leqq 0$ is large enough in the *transient case*, i.e., where every stationary policy is transient. The reader is warned that we have found it convenient to depart from the customary definition of $V_\rho(\pi)$ used in the literature, the usual formula being here multiplied by β.

Discount optimality. For each $n = -1, 0, 1, \cdots$, we say π^* is $n^{\pm}$ *discount optimal* if

$$\lim \inf_{\rho \to 0\pm} |\rho|^{-n} [V_\rho(\pi^*) - V_\rho(\pi)] \geqq 0, \quad \text{for all } \pi. \tag{27}$$

The limit inferior is, of course, componentwise. Similarly, we say π^* is $\infty^{\pm}$ *discount optimal* if for some $\rho^* > 0$,

$$V_\rho(\pi^*) - V_\rho(\pi) \geqq 0 \quad \text{for all } \pi \text{ and } \quad 0 < \pm\rho < \rho^*. \tag{28}$$

Notice that n^- discount optimality ($-1 \leqq n \leqq \infty$) is defined, and so will be discussed in the sequel, only in the transient case, even though this presumption will not always be explicitly stated.

It is clear from (27) and (28) that if π^* is $n^{\pm}$ discount optimal ($-1 \leqq n \leqq \infty$), then π^* is $m^{\pm}$ discount optimal for $-1 \leqq m < n$. Thus the sensitivity of $n^{\pm}$ discount optimality increases with n.

We remark that in the transient case, $-1^{\pm}$ discount optimality is uninteresting since then $\lim_{\rho\to 0} |\rho| V_\rho(\pi) = 0$ for all π so every policy is $-1^{\pm}$ discount optimal. It follows that -1^{-} discount optimality is of no interest because it is defined only where every policy is optimal. On the other hand -1^{+} discount optimality is sometimes a useful criterion when there are nontransient policies. Since $\rho V_\rho(\pi)$ $(\rho > 0)$ is the interest one receives in each period with an initial investment of $V_\rho(\pi)$, one can think of $\rho V_\rho(\pi)$ as the average expected reward per period equivalent to the initial lump sum $V_\rho(\pi)$. From this viewpoint, the -1^{+} discount optimality criterion seeks to maximize this equivalent average expected reward as $\rho \downarrow 0$. The main objection to this criterion is that it is not very sensitive. For example, suppose $S = 2$, there are two actions in state 1, and there is only one action in state 2. The immediate reward in state 1 from action 1 is 0 and from action 2 is 10^6. The immediate reward in state 2 is 1. Either action in state 1 moves the process to state 2. Once the process reaches state 2, it stays there. In this example both stationary policies are -1^{+} discount optimal but it is clear that the policy which takes action 2 in state 1 is preferable.

In the transient case a policy is $0^{\pm}$ discount optimal if and only if it maximizes the expected infinite horizon return $V_0(\pi)$. This is because then $V_\rho(\pi)$ is continuous in ρ at $\rho = 0$. The notions of 0^{+} and ∞^{+} discount optimality were introduced by Blackwell [5].

The criteria (27), (28) are concerned with situations where the interest rate approaches zero, or equivalently, the discount factor approaches one. As was suggested in [34] it is also of interest to consider situations in which the discount factor approaches α, $0 < \alpha < \|P(f)\|^{-1}$ for $f \varepsilon F$. This possibility is easily reduced to the situation already discussed by simply replacing each $P(f)$ and $r(f)$ respectively by $\alpha P(f)$ and $\alpha r(f)$. When this device is employed we are in effect expressing a discount factor β, $0 < \beta < \|P(f)\|^{-1}$ for $f \varepsilon F$, differing from α in the form $\beta = (1 + \rho)^{-1}\alpha$. In this event $\rho = \alpha\rho^*$ where ρ^* is the difference between the interest rates determined by β and α. Thus ρ is simply a convenient rescaling of the difference between the interest rates. The criteria (27), (28) are, of course, invariant under such scale changes.

Denote by $D_n^{\pm}$ the set of $f \varepsilon F$ for which f^∞ is $n^{\pm}$ discount optimal, $n = -1, 0, 1, \cdots$. Below we shall characterize these sets and develop algorithms for finding an element of each. To this end, for each $f \varepsilon F$ let $Q(f)$, $P^*(f)$, and $H(f)$ denote the matrices defined in Section 3 associated with the substochastic matrix $P(f)$. Theorem 3 and Lemma 8 below are established in [30] where $P(f)$ is stochastic and $\rho \geqq 0$. We merely state the generalizations here since the proofs are similar to those in [30].

Characterization of discount optimal policies. Since $V_\rho(f^\infty) = R_\rho(Q(f))r(f)$ when $|\sigma(\beta P(f))| < 1$, we may use Theorem 2 to give a Laurent expansion of $V_\rho(f^\infty)$ in ρ about the origin.

THEOREM 3. *If* $f \varepsilon F$, $|\sigma(\rho H(f))| < 1$, *and* $|\sigma(\beta P(f))| < 1$, *then*

$$V_\rho(f^\infty) = \sum_{n=-1}^{\infty} (\pm\rho)^n y_n^{\pm}(f), \tag{29}$$

where $y_{-1}^{\pm}(f) \equiv \pm P^*(f)r(f)$ *and* $y_n^{\pm}(f) \equiv (\mp 1)^n H(f)^{n+1} r(f)$, $n = 0, 1, \cdots$.

Notice that if $-1 < \rho \leqq 0$ in Theorem 3, then $|\sigma(\beta P(f))| < 1$ implies $P(f)$ is transient and $P^*(f) = 0$. Also observe that $y_n^-(f) = (-1)^n y_n^+(f)$ for all n.

If C is a real matrix, we say C is *lexicographically non-negative* written $C \succeq 0$, if the first nonvanishing element of each row of C is positive. Similarly, C is called *lexicographically positive*, written $C \succ 0$, if $C \succeq 0$ and $C \neq 0$. We write $C \succ (\succeq) B$ or $B \prec (\preceq) C$ if $C - B \succ (\succeq) 0$.

For $f \varepsilon F$, let $Y_n^{\pm}(f) = (y_{-1}^{\pm}(f), \cdots, y_n^{\pm}(f))$ for $n \geqq -1$, $Y_n^{\pm}(f) = 0$ for $n < -1$, and $Y^{\pm}(f) = (y_{-1}^{\pm}(f), y_0^{\pm}(f), \cdots)$.

Blackwell [5] has shown that D_∞^+ is nonempty where $P(f)$ is a stochastic matrix for each $f \varepsilon F$. An alternate constructive proof of this result is given in [30]. Essentially the same proofs show that $D_\infty^+(D_\infty^-)$ is nonempty where $P(f)$ is substochastic (transient) for each $f \varepsilon F$. It is also clear that $D_{-1}^{\pm} \supset D_0^{\pm} \supset \cdots \supset D_\infty^{\pm}$ and so are all nonempty. Since $D_n^{\pm}$ is nonempty, it is immediate from (29) that $D_n^{\pm} = \{f : f \varepsilon F,\ Y_n^{\pm}(f) \succeq Y_n^{\pm}(g) \text{ for all } g \varepsilon F\}$ for $n = -1, 0, \cdots$, and $D_\infty^{\pm} = \{f : f \varepsilon F,\ Y^{\pm}(f) \succeq Y^{\pm}(g) \text{ for all } g \varepsilon F\}$.

It is shown in [30] that if $P(f)$ is stochastic for each $f \varepsilon F$ and if $f, g \varepsilon F$, then $Y^{\pm}(f) = Y^{\pm}(g)$ if and only if $Y_S^{\pm}(f) = Y_S^{\pm}(g)$. This result and its proof remains true if $P(f)$ is substochastic for each $f \varepsilon F$. It follows from this fact that $D_S^{\pm} = D_{S+1}^{\pm} = \cdots = D_\infty^{\pm}$.

Example with $F = D_{-1}^{\pm} = D_0^{\pm} = \cdots = D_{S-1}^{\pm} \neq D_S^{\pm}$ *and, for odd* S, $D_S^+ \cap D_S^- = \varnothing$. Let state $S+1$ denote the stopped position. In state s, $1 < s < S+1$, there is only one action and the process moves to state $s+1$ with probability one and receives a reward α_s. In state 1 there are two actions. Action 1 takes the process to state 2 with probability one and earns α_1. Action 2 takes the process to state 2 with probability $\frac{1}{2}$, leaves the process in state 1 with probability $\frac{1}{2}$, and earns 1. Let f^∞ and g^∞ be the stationary policies that take actions 1 and 2 respectively in state 1. Let $V_\rho'(\cdot)$ denote the first component of $V_\rho(\cdot)$. Then $V_\rho'(f^\infty) = \sum_{i=1}^{S} \beta^i \alpha_i = 2\beta(1-\beta)^{S-1}$ where we now let $\alpha_1, \alpha_2, \cdots, \alpha_S$ be the coefficients in the expansion of the last polynomial. Also $V_\rho'(g^\infty) = (2-\beta)^{-1} V_\rho'(f^\infty)$ so

$$|\rho|^{-n}[V_\rho'(f^\infty) - V_\rho'(g^\infty)] = 2\beta^{n+1}(2-\beta)^{-1}(1-\beta)^S|1-\beta|^{-n}.$$

Thus

$$\begin{aligned} \lim_{\rho \to 0\pm} |\rho|^{-n}[V_\rho'(f^\infty) - V_\rho'(g^\infty)] &= 0, && n < S; \\ &= 2(\pm 1)^S, && n = S; \\ &= \infty\,(\pm 1)^S, && n > S; \end{aligned}$$

whence $\{f, g\} = F = D_{-1}^{\pm} = \cdots = D_{S-1}^{\pm} \neq D_S^{\pm}$ and, for odd S, $D_S^+ \cap D_S^- = \varnothing$. We remark that by considering a K state version of this example and adding $S - K \geqq 0$ dummy states, we can achieve $F = D_{-1}^{\pm} = \cdots = D_{K-1}^{\pm} \neq D_K^{\pm} - D_{K+1}^{\pm} = \cdots$.

Policy improvement method for finding $S^{\pm}$ *discount optimal policies.* Suppose

$|\sigma(\beta P(f))| < 1$ for all f and let

$$(30)\quad v_\rho(g,\pi) \equiv r(g) + Q(g)V_\rho(\pi) - \rho V_\rho(\pi) = (1+\rho)[V_\rho(g,\pi) - V_\rho(\pi)].$$

Notice from Lemma 1 that

$$(31)\quad V_\rho(g^\infty) - V_\rho(\pi) = R_\rho(Q(g))v_\rho(g,\pi).$$

The policy improvement algorithm developed in [30] for finding S^+ discount optimal policies depends on the policy improvement method for maximizing $V_\rho(\cdot)$. In order to determine whether a stationary policy f^∞ maximizes $V_\rho(\cdot)$, Theorem 1 tells us to examine whether $v_\rho(g,f^\infty) \leqq 0$ for all $g \varepsilon F$. Since we are here interested in small values of $|\rho|$, it is natural to use (29) to express $v_\rho(g,f^\infty)$ as a Laurent series in ρ.

To this end for f, $g \varepsilon F$, let $r_0(g) = r(g)$, $r_n(g) = 0$ for $n \neq 0$, $y_{-2}^{\pm}(f) = 0$, and for $n \geqq -1$,

$$(32)\quad \psi_n^{\pm}(g,f) = r_n(g) + Q(g)y_n^{\pm}(f) \mp y_{n-1}^{\pm}(f)$$

and $\Psi_n^{\pm}(g,f) = (\psi_{-1}^{\pm}(g,f), \cdots, \psi_n^{\pm}(g,f))$. Also let $\Psi_n^{\pm}(g,f) = 0$ for $n < -1$ and $G_n^{\pm}(f) = \{g : g \varepsilon F, \Psi_n^{\pm}(g,f) > 0\}$ for all n. Observe that $\psi_n^{-}(g,f) = (-1)^n\psi_n^{+}(g,f)$ for all n. Let $\Psi^{\pm}(g,f) = (\psi_{-1}^{\pm}(g,f), \psi_0^{\pm}(g,f), \cdots)$.

The next lemma is obtained by combining Theorem 3 with the above definitions.

LEMMA 8. *If f, $g \varepsilon F$, $|\sigma(\rho H(f))| < 1$, and $|\sigma(\beta P(f))| < 1$, then*

$$(33)\quad v_\rho(g,f^\infty) = \sum_{n=-1}^{\infty} (\pm\rho)^n\psi_n^{\pm}(g,f).$$

REMARK. It is immediate from (33) on setting $g = f$ that $\Psi^{\pm}(f,f) = 0$ for $f \varepsilon F$.

Suppose f, $g \varepsilon F$, and $|\sigma(\beta P(\cdot))| < 1$ for f and g. We say g is an *improvement* of f for ρ if $v_\rho(g,f^\infty) > 0$. If there is no improvement of f for ρ, then $v_\rho(g,f^\infty) \leqq 0$ for all $g \varepsilon F$ so by Theorem 1, f^∞ maximizes $V_\rho(\cdot)$.

It was shown in [30] that $\Psi^{\pm}(g,f) = 0$ if and only if $\Psi_S^{\pm}(g,f) = 0$. By combining this fact with Lemma 8, one sees by a straightforward generalization of results in [30] that the following are equivalent: (i) $f \varepsilon D_S^{\pm}$; (ii) $G_S^{\pm}(f)$ is empty; and (iii) there is no improvement of f for all small enough $\pm\rho > 0$. This characterizes the set $D_S^{\pm}$. One also sees that $G_S^{\pm}(f)$ is the set improvements of f for all small enough $\pm\rho > 0$. Moreover, $g \varepsilon G_S^{\pm}(f)$ implies $Y_S^{\pm}(g) > Y_S^{\pm}(f)$.

The above discussion suggests the following *policy improvement method* for finding an element of $D_S^{\pm}$. Let $f_0 \varepsilon F$ be arbitrary and choose $f_i \varepsilon G_S^{\pm}(f_{i-1})$, $i = 1, \cdots, N$, inductively until an integer N occurs for which $G_S^{\pm}(f_N)$ is empty. Then $f_N \varepsilon D_S^{\pm}$. The procedure for finding an element of D_S^{+} is the one given in [30].

This algorithm is also a method of finding an element of $D_n^{\pm}$ for $n < S$ since $D_n^{\pm} \supset D_S^{\pm}$. However, it is natural to expect that the algorithm can be terminated more rapidly if one merely seeks an element of $D_n^{\pm}$ $(n < S)$. This is the case.

The precise result depends upon obtaining a characterization of $D_n^{\pm}$ analogous to the one given above for $D_S^{\pm}$.

Characterization of $n^{\pm}$ discount optimal policies. For this purpose, we need a preliminary lemma which expresses the difference $Y^{\pm}(g) - Y^{\pm}(f)$ in terms of the test criterion $\Psi^{\pm}(g, f)$.

LEMMA 9. *If f, $g \in F$, then*

$$y_n^{\pm}(g) - y_n^{\pm}(f) = \pm P^*(g)\psi_{n+1}^{\pm}(g, f) + \sum_{k=0}^{n+1} (\mp 1)^k H(g)^{k+1}\psi_{n-k}^{\pm}(g, f), \qquad n = -2, -1, 0, \cdots.$$

PROOF. Since $y_n^-(f) = (-1)^n y_n^+(f)$ and $\psi_n^-(g, f) = (-1)^n \psi_n^+(g, f)$, it suffices to give the proof for $y_n^+(g) - y_n^+(f)$. From Theorem 3, (31), Lemma 8, Theorem 2, and $y_{-2}^{\pm}(g) - y_{-2}^{\pm}(f) = 0$, we have for all small enough $\rho > 0$ that

$$\begin{aligned}
&\sum_{n=-2}^{\infty} \rho^n [y_n^+(g) - y_n^+(f)] \\
&\quad = V_\rho(g^\infty) - V_\rho(f^\infty) = R_\rho(Q(g)\ v_\rho(g, f^\infty) \\
&\quad = [\rho^{-1}P^*(g) + \sum_{n=0}^{\infty} \rho^n (-1)^n H(g)^{n+1}][\sum_{n=-1}^{\infty} \rho^n \psi_n^+(g, f)] \\
&\quad = \sum_{n=-2}^{\infty} \rho^n [P^*(g)\psi_{n+1}^+(g, f) + \sum_{k=0}^{n+1} (-1)^k H(g)^{k+1}\psi_{n-k}^+(g, f)].
\end{aligned}$$

Equating terms of the first and last series completes the proof.

In the remainder of this section, we freely *drop* the superscript $\pm$ for notational simplicity whenever no ambiguity results. Also let $D_{-2} \equiv F$.

REMARK. Observe from Lemma 9 that $\Psi_{n+1}(g, f) = 0$ implies $Y_n(g) = Y_n(f)$. Conversely from (32), $Y_n(g) = Y_n(f)$ implies $\Psi_n(g, f) = \Psi_n(g, g) = 0$. If g^∞ is transient, then by Lemma 9 and the above remarks $\Psi_n(g, f) = 0$ if and only if $Y_n(g) = Y_n(f)$. These facts will be used frequently in the sequel.

THEOREM 4. *Suppose $f \in F$ and $n = -2, -1, \cdots, S - 1$.*

1°. *If $G_{n+1}^{\pm}(f)$ is empty, then $f \in D_n^{\pm}$.*

2°. *If $f \in D_n^{\pm}$, then $G_n^{\pm}(f)$ is empty.*

PROOF. The proof of 1° is by induction on n. The result is trivially true for $n = -2$. Suppose it holds for the integer $n - 1 \geqq -2$, and consider n. Since $G_{n+1}(f)$ is empty, $G_n(f)$ is empty. Thus, by the induction hypothesis $f \in D_{n-1}$. It suffices therefore to show that if $g \in D_{n-1}$, then $\Delta y_n \leqq 0$ where $\Delta y_k \equiv y_k(g) - y_k(f)$.

We have $Y_{n-1}(f) = Y_{n-1}(g)$ so $\Psi_{n-1}(g, f) = \Psi_{n-1}(g, g) = 0$. Thus by Lemma 9 (recall $P^*(g)$ is presumed to vanish when $\pm$ is $-$ below)

$$\Delta y_{n-1} + \rho \Delta y_n = [P^*(g) + \rho H(g)][\psi_n + \rho\psi_{n+1}] - \rho^2 H(g)\psi_{n+1} \tag{34}$$

where $\psi_k \equiv \psi_k(g, f)$ for all k. Since $\Psi_{n+1}(g, f) \leq 0$, we have $\psi_n + \rho\psi_{n+1} \leqq 0$ for all small enough $\rho > 0$ so from (34) and Lemma 7, $\rho \Delta y_n = \Delta y_{n-1} + \rho \Delta y_n \leqq -\rho^2 H(g)\psi_{n+1}$ for all small enough $\rho > 0$. This implies $\Delta y_n \leqq 0$ and completes the proof of 1°.

The fact that 2^0 holds follows by contraposition from Theorem 5 below, which completes the proof.

REMARK 1. Weaker forms of assertion 1^0 of Theorem 4 and of Theorem 5 below were established (with $\pm$ being $+$) for the cases $n = -1$ and $n = 0$ respectively by Blackwell [5] and Veinott [34] by different methods.

REMARK 2. For $n = S - 1$, the assertion 1^0 of Theorem 4 can be strengthened to $f \varepsilon D_S^{\pm}$.

REMARK 3. In the *recurrent case*, i.e., when the diagonal elements of $P^*(f)$ are positive for all $f \varepsilon F$, the converse of 1^0 of Theorem 4 (with $\pm$ being $+$) also holds. To see this notice that if $G_{n+1}^+(f) \neq \varnothing$, there is a smallest index $k \leqq n+1$ for which $G_k^+(f) \neq \varnothing$. Choose $g \varepsilon G_k^+(f)$. Then from Remarks 2 and 3 following Theorem 5, $Y_{k-1}^+(g) > Y_{k-1}^+(f)$, so $f \notin D_n^+$. One immediate consequence of this result is that in the recurrent case $D_{S-1}^+ = D_S^+$.

REMARK 4. In the transient case, the converse of 2^0 of Theorem 4 also holds (c.f., Remark 3). The proof follows the proof of 1^0 of Theorem 4 on noting that $\Delta y_n = H(g)\psi_n \leqq 0$ since $H(g) = R_0(Q(g))$.

Improvement of an $n^{\pm}$ discount non-optimal policy. The next theorem gives a method of choosing a policy that increases $Y_n^{\pm}(f)$ lexicographically.

THEOREM 5. *If $f, g \varepsilon F$ and $\Psi_{n-1}^{\pm}(g, f) = 0$, then $Y_{n-2}^{\pm}(g) = Y_{n-2}^{\pm}(f)$; if also $g \varepsilon G_n^{\pm}(f) \cap G_{n+1}^{\pm}(f)$, then $Y_n^{\pm}(g) > Y_n^{\pm}(f)$, $n = -1, 0, \cdots, S$.*

PROOF. Since $\Psi_{n-1}(g, f) = 0$, by Lemma 9 we have $Y_{n-2}(f) = Y_{n-2}(g)$. Also because $g \varepsilon G_n(f) \cap G_{n+1}(f)$, $\Psi_{n+1}(g, f) > 0$ and $\psi_n > 0$ where $\psi_k \equiv \psi_k(g, f)$ for each k. Thus $\psi_n + \rho\psi_{n+1} > 0$ for all small enough $\rho > 0$. Combining this fact with (34) and Lemma 7, we have that

$$\Delta y_{n-1} + \rho \Delta y_n \geqq 0 \quad \text{for all small enough} \quad \rho > 0, \tag{35}$$

where $\Delta y_k \equiv y_k(g) - y_k(f)$. Moreover, the inequality in (35) is strict, for if not, since $(\psi_n, \psi_{n+1}) > 0$, we have from Lemma 9 that

$$0 = \Delta y_{n-1} = P^*(g)\psi_n \leqq P^*(g)\psi_{n+1}.$$

Thus by (34) for all small enough $\rho > 0$

$$0 = \Delta y_{n-1} + \rho \Delta y_n = [P^*(g) + \rho H(g)]\psi_n + \rho P^*(g)\psi_{n+1} \geqq [P^*(g) + \rho H(g)]\psi_n$$

which is impossible because of Lemma 7 and $\psi_n > 0$. Hence $(\Delta y_{n-1}, \Delta y_n) > 0$, which completes the proof.

REMARK 1. If $\Psi_{n-1}(g, f) = 0$ and $g \varepsilon G_n(f)$, then it is easy to modify g (while retaining these properties) so $g \varepsilon G_{n+1}(f)$. All that is required is to set $g(s) = f(s)$ whenever the sth row of $\Psi_n(g, f)$ vanishes. If this is done then $g \varepsilon G_S(f)$ also.

REMARK 2. If $G_{n-1}(f)$ is empty and $g \varepsilon G_n(f)$, then $\Psi_{n-1}(g, f) = 0$.

REMARK 3. In the recurrent case, $\Psi_{n-1}^+(g, f) = 0$ and $g \varepsilon G_n^+(f)$ imply $Y_{n-1}^+(g) > Y_{n-1}^+(f)$. To see this, observe from Lemma 9 that $\Delta y_{n-1}^+ = P^*(g)\psi_n^+(g, f) > 0$.

REMARK 4. In the transient case, $\Psi_{n-1}(g, f) = 0$ and $g \varepsilon G_n(f)$ imply $Y_{n-1}(g) = Y_{n-1}(f)$ and $Y_n(g) > Y_n(f)$. The first assertion is immediate from Lemma 9.

The second assertion also follows from that result by observing that $\Delta y_n = H(g)\psi_n > 0$ since $H(g) = R_0(Q(g))$.

A computing strategy. The above results suggest an efficient way of carrying out the policy improvement method to find an element of D_N, $-1 \leqq N < S$. We do not consider the case $N = S$ since the procedure we suggest for finding an element of D_{S-1} automatically shows that element to be in D_S also.

An iteration begins with an $f \varepsilon F$ and an n, $-1 \leqq n \leqq N + 1$, for which $G_{n-1}(f)$ is empty. (Initially f is arbitrary, and $n = -1$ which $\pm$ is $+$ and $n = 0$ when $\pm$ is $-$.) If $n = N + 1$ and $G_n(f)$ is empty, then $f \varepsilon D_N$ and we are done. If $n < N + 1$ and $G_n(f)$ is empty, replace n by $n + 1$ and start a new iteration. Finally, if $G_n(f)$ is not empty, choose an element g of that set satisfying $g(s) = f(s)$ whenever the sth row of $\Psi_n(g, f)$ vanishes; replace f by g and n by m where $m = n$ if either n was decreased in the last iteration, or $n = -1$ and $\pm$ is $+$, or $n = 0$ and $\pm$ is $-$, and $m = n - 1$ otherwise; and start a new iteration.

Replacing n by m in the last step of the above iteration is justified as follows. By Theorems 4 and 5, $f \varepsilon D_{m-1}$ so $g \varepsilon D_{m-1}$ and hence $G_{m-1}(g)$ is empty as required to start the next iteration.

The above algorithm first maximizes $y_{-1}(\cdot)$; then it maximizes $y_0(\cdot)$ subject to $Y_{-1}(\cdot)$ fixed at optimum; then it maximizes $y_1(\cdot)$ subject to $Y_0(\cdot)$ fixed at optimum; and so forth. This feature of the method permits the computations to be performed quite efficiently. To illustrate, suppose we have at hand an $f \varepsilon F$, an n for which $G_{n-1}(f)$ is empty, and the matrix $Y_{n-1}(f)$. In order to ascertain whether or not $G_n(f)$ is empty, we must compute $y_n(f)$. An efficient way of doing this will now be given.

Since $\Psi(f, f) = 0$, $Y(f) = (y_{-1}, y_0, \cdots)$ satisfies $(y_{-2} \equiv 0)$

$$\pm y_{n-1} - Q(f)y_n = r_n(f), \qquad n = -1, 0, \cdots \tag{36}$$

The next result implies that $Y(f)$ is the unique solution to (36).

LEMMA 10. *Suppose* $n \geqq -1$, $f \varepsilon F$, *and* $y_{n-1} = y_{n-1}(f)$. *Then* $(y_n, y_{n+1}) = (y_n(f), y_{n+1}(f))$ *satisfies the* nth *and* $(n + 1)$th *equations in* (36). *Conversely, if* (y_n, y_{n+1}) *satisfies those two equations, then* $y_n = y_n(f)$.

PROOF. It suffices to prove the converse. Premultiply the $(n + 1)$th equation by $\pm P^*(f)$ and add the result to the nth equation. Using (15) and the fact $[-Q(f) + P^*(f)]$ is nonsingular completes the proof.

The method for computing $y_n(f)$ given in the above proof is not efficient. We sketch a procedure requiring about one half as much computation. First, determine the recurrent classes and transient states of $P(f)$. Then apply the procedure to be given below for determining the components of $y_n(f)$ for states in each recurrent class. Finally, solve (36) for the components of $y_n(f)$ for the transient states. With this method the total computational effort required to compute $y_n(f)$ is at worst little more than that required to solve a single system of S linear equations in S unknowns.

The above discussion permits us to consider only the case where $P(f) \equiv P$ is

stochastic and irreducible. Let $Q = Q(f)$, $u = r_n(f) \mp y_{n-1}(f)$, $v = r_{n+1}(f)$, $w = y_n(f)$, and $x = y_{n+1}$. Then we must solve

$$-Qw = u \tag{37}$$

$$\pm w - Qx = v. \tag{38}$$

Let Q' denote the first $S - 1$ columns of Q. For any S-vector z, denote by z^S its first $S - 1$ components and by z_S its Sth component. Since $Q1 = 0$, if (w, x) satisfies (37) and (38), then so does $(w, x + \lambda 1)$ for all λ. Thus we can and do assume $x_S = 0$. Similarly, if w satisfies (37), then so does $w + \lambda 1$ for all λ. Thus, (37) and (38), together with $x_S = 0$, are equivalent to

$$-Q'w' = u \tag{37}'$$

$$\pm 1 w_S - Q'x^S = v \mp \begin{pmatrix} w' \\ 0 \end{pmatrix} \tag{38}'$$

where the S-vector w' is defined by $w' \equiv w^S - 1^S w_S$. Application of Gauss' elimination method to (37)′ in effect premultiplies (37)′ by a lower triangular $S \times S$ matrix E (in product form). Since P is irreducible, the last row of EQ' vanishes while its first $S - 1$ rows form an upper triangular nonsingular matrix. This permits w' to be determined. Now if we apply the same steps of the elimination method used in solving (37)′ to the first and last columns in (38)′—in effect, premultiplying (38)′ by E—the last equation in (38)′ can be solved uniquely for w_S because the last row of EQ' vanishes. The desired result is then

$$y_n(f) = \begin{pmatrix} w' \\ 0 \end{pmatrix} + 1 w_S .$$

Fixed points. Recall from Corollary 2 that if $|\sigma(\beta P(f))| < 1$ for all $f \varepsilon F$, then $V_\rho \equiv \max_\pi V_\rho(\pi)$ is the unique fixed point of $\mathfrak{R}_\rho$ where

$$\mathfrak{R}_\rho V \equiv \max_{g \varepsilon F} [\beta r(g) + \beta P(g) V] \tag{39}$$

for each S-vector V. Moreover, $V = V_\rho$ is the largest (smallest) S-vector satisfying $V \leqq \mathfrak{R}_\rho V$ ($V \geqq \mathfrak{R}_\rho V$). We shall now give analogous results for the more sensitive discount criteria which we have studied in this section.

Let $Y = (y_{-1}, y_0, \cdots)$, $TY = (0, y_{-1}, y_0, \cdots)$, and $r^*(g) = (0, r(g), 0, 0, \cdots)$, and define the operator $\mathfrak{R}^\pm$ by

$$\mathfrak{R}^\pm Y \equiv \max_{g \varepsilon F} [r^*(g) \mp TY + P(g) Y]$$

where the maximum is here a lexicographic maximum. Let $Y^\pm = Y^\pm(f)$ where $f \varepsilon D_S{}^\pm$. An application of Lemma 10 yields

COROLLARY 7. *$Y^\pm$ is the unique fixed point of $\mathfrak{R}^\pm$.*

Let $\mathcal{Y}$ be the set of sequences $Y = (y_n)$ for which $\rho^n y_n \to 0$ as $n \to \infty$ for some $\rho > 0$ depending on Y. Evidenctly $\mathcal{Y}$ is a linear space. Then it is easy to see from the corresponding result about V_ρ that $Y = Y^\pm$ is the lexicographically largest (smallest) element of $\mathcal{Y}$ satisfying $Y \preceq \mathfrak{R}^\pm Y$ ($Y \succeq \mathfrak{R}^\pm Y$).

The above results characterize $D_S^{\pm}$ in terms of the fixed point of $\mathfrak{R}^{\pm}$. We now give the analogous characterization of $D_{n-1}^{\pm}$ for $n \leqq S$.

Let $Y_n = (y_{-1}, y_0, \cdots, y_n)$, $TY_n = (0, y_{-1}, \cdots, y_{n-1})$, $r_n{}^*(g) = (0, r(g), 0, \cdots, 0)$ $(n + 2$ columns), and define the operator $\mathfrak{R}_n{}^{\pm}$ by

$$\mathfrak{R}_n{}^{\pm}Y_n = \max_{g \varepsilon F} [r_n{}^*(g) \mp TY_n + P(g)Y_n]$$

where here again the maximum is a lexicographic one. Let $Y_{n-1}^{\pm} = Y_{n-1}^{\pm}(f)$ where $f \varepsilon D_{n-1}^{\pm}$. From Lemma 10 and a slight extension of Theorem 4 we get

COROLLARY 8. *$Y_n = (Y_{n-1}, y_n)$ is a fixed point of $\mathfrak{R}_n{}^{\pm}$ for some y_n if and only if $Y_{n-1} = Y_{n-1}^{\pm}$, $n = 0, 1, \cdots, S - 1$.*

5. The continuous parameter case.

Preliminaries. We now turn to the continuous time parameter version of our Markovian decision process. Our treatment closely parallels and exploits, in so far as possible, the results in the discrete time parameter case. Following Zachrisson [36] and Miller [27]–[29], suppose a system is observed at each time $t \geqq 0$. At each observation the system is found to be in one of S states labeled $1, 2, \cdots, S$ or to have "stopped." Each time the system is observed in state s, an action a is chosen from a finite set A_s of possible actions and a reward rate $r(s, a)$ is received per unit time. The transition rate from state s at time t to state $u (\neq s)$ when action a is taken at time t is denoted by $q(u \mid s, a) \geqq 0$. Similarly, $q(s \mid s, a) \geqq 0$ is the transition rate out of state s and $q(s \mid s, a) - \Sigma_{u \neq s} q(u \mid s, a) \geqq 0$ is the transition rate into the stopped position. Once the process is observed to have stopped, it remains stopped and earns no rewards.

Let $F = \mathbf{X}_{s=1}^{S} A_s$. A *policy* is a function $\pi = (f_t)$ which maps each $t \varepsilon [0, \infty)$ into an element f_t of F and which is Lesbesgue measurable, i.e., the inverse image of each element of F under π is Lesbesgue measurable. Using the policy $\pi = (f_t)$ means that if the system is observed in state s at time t, the action chosen at that time is $f_t(s)$, the sth coordinate of f_t. Let $f^{\infty} = (f)$ be the policy, called *stationary*, with value f everywhere. If π is a policy, denote by π^t its restriction to the interval $[0, t)$. Similarly, f^t is the restriction of f^{∞} to the interval $[0, t)$. If $\pi = (f_t)$ and $\pi^* = (g_t)$ are policies, let (π^t, π^*) denote the policy (h_u) defined by $h_u = f_u$ for $0 \leqq u < t$ and $h_u = g_{u-t}$ for $t \leqq u$, i.e., π is used for t units of time and π^* is used thereafter. Similarly, let ${}^t\pi = (\pi^t, \pi^t, \cdots)$ be the *periodic policy* which uses π^t repeatedly every t units of time.

For any $f \varepsilon F$, let $r(f)$ be the $S \times 1$ column vector whose sth component is $r(s, f(s))$, and let $Q(f)$ be the $S \times S$ infinitesimal generator matrix whose suth element is $q(u \mid s, f(s))$ for $s \neq u$ and whose ssth element is $-q(s \mid s, f(s))$. Evidently $Q(f)1 \leqq 0$ where 1 is a column of $+1$'s. Also let $P(f) \equiv I + Q(f)$.

Given a policy $\pi = (f_t)$, the infinitesimal generators $Q(f_t)$ uniquely determine the transition function of a step-type continuous parameter Markov chain as follows [27]. Since $Q(f_t)$ is bounded and Lesbesgue measurable in $t \geqq 0$, a standard result [6] of differential equations shows that there is a unique $S \times S$ matrix function $P(t, u, \pi) = P(t, u)$ that is absolutely continuous in $u (\geqq t)$ and satisfies

the system of differential equations

$$\frac{\partial}{\partial u} P(t, u) = P(t, u)Q(f_u), \qquad 0 \leqq t \leqq u, \tag{40}$$

a.e. subject to the boundary condition

$$P(t, t) = I, \qquad 0 \leqq t. \tag{41}$$

Moreover,

$$P(t, w) = P(t, u)P(u, w), \qquad 0 \leqq t \leqq u \leqq w. \tag{42}$$

Since the off diagonal elements of $Q(f_u)$ are non-negative, $P(\cdot\,, \cdot)$ is non-negative by a result of Kolmogorov [17, pages 207 ff]. See also [1] and [36, pages 237–9] for simplifications and extensions. The same proof shows the diagonal elements of $P(\cdot\,, \cdot)$ are positive.

Integrating (40) and using (41) gives

$$P(t, u) = I + \int_t^u P(t, w)Q(f_w)\, dw, \qquad 0 \leqq t \leqq u. \tag{43}$$

Thus because $Q(f_w)1 \leqq 0$ and $P(\cdot\,, \cdot) \geqq 0$, we see from (43) that $\|P(t, u)\| \leqq 1$. Also since the integrand in (43) is uniformly bounded,

$$\lim_{h \downarrow 0} \sup_{t \geqq 0} [P(t, t + h) - I] = 0. \tag{44}$$

We call the unique matrix function $P(t, u, \pi)$ defined above the (*substochastic*) *transition function* determined by π. As Miller [27, page 13] has pointed out, we can then invoke a result in Dynkin [13, page 160] to assure that there exists a step-type continuous parameter Markov chain with that same transition function. It will be convenient in the sequel to let $P(t, \pi) = P(0, t, \pi)$.

We say that π is *transient* if $\int_0^\infty P(t, \pi)\, dt$ converges. Since we can restrict attention to step-type Markov chains, one sees (c.f., Miller [27]) by combining Lemma 1.10 from Dynkin [13, page 21] with Fubini's theorem that if $\pi = (f_t)$ is transient the S-vector $V(\pi)$ of expected total returns starting from each state, given by

$$V(\pi) \equiv \int_0^\infty P(t, \pi)r(f_t)\, dt, \tag{45}$$

converges absolutely.

Until further notice we shall *drop* the hypothesis that $\|P(\cdot)\| \leqq 1$, though we *retain* the assumption that $P(\cdot) \geqq 0$. The definition of a transient policy given above is equally valid in this case. Of course, as in the discrete time parameter case, we must then give up our interpretation of the elements of $P(\cdot\,, \pi)$ as probabilities.

We digress briefly to review a few known facts about square matrices that will be needed in the sequel. If B is a square real matrix, then the off diagonal elements of B are non-negative if and only if e^{Bt} is non-negative and has positive diagonal elements for all $t \geqq 0$[1]. If either of these equivalent conditions holds, then $B1 \leqq 0$ if and only if $\|e^{Bt}\| \leqq 1$ for all $t \geqq 0$ [17, page 205].

If Ω is a set of complex numbers, denote by e^{Ω} the range of the exponential function on Ω and by $|\Omega|$ the supremum of the moduli of the elements of Ω. If B is a square complex matrix, then $\|e^B\| \leqq e^{\|B\|}$ and $\sigma(e^B) = e^{\sigma(B)}$ so $|\sigma(e^B)| = |e^{\sigma(B)}|$. For such a matrix it is known that the following are equivalent: (i) $|e^{\sigma(B)}| < 1$, (ii) $\|e^{Bt}\| < 1$ for some $t > 0$, (iii) $e^{Bt} \to 0$ as $t \to \infty$, and (iv) $\int_0^\infty e^{Bt}\,dt$ converges absolutely. These conditions imply that B is nonsingular and $-B^{-1}$ equals the integral in (iv). If also B is real and has non-negative off diagonal elements, the above four conditions are equivalent to: (v) $-B$ is nonsingular and has a non-negative inverse. If further $B1 \leqq 0$, the five conditions given above are equivalent to: (vi) $\|e^{Bt}\| < 1$ for all $t > 0$ and (vii) B is nonsingular. Several of these results follow easily from

$$e^{BT} - I = \int_0^T \frac{d}{dt} e^{Bt}\,dt = \left[\int_0^T e^{Bt}\,dt\right] B, \qquad T \geqq 0. \tag{46}$$

Thus [if $Q(f)1 \leqq 0$], f^∞ is transient if and only if $Q(f)$ satisfies any one of the conditions [(i)–(vii)] (i)–(v) given above.

Equivalent discrete and continuous parameter processes. As Howard [20, page 113] suggests, by choosing an appropriate time unit, we may assume with no loss of generality that $P(f) \equiv I + Q(f) \geqq 0$ for all $f \in F$. (This can be achieved, for example, by multiplying $Q(f)$ and $r(f)$ for all $f \in F$ by a sufficiently small positive number.) We shall impose this assumption in the remainder of this paper without further mention. The assumption permits an equivalence (c.f., [20, page 120]) to be established between the continuous and discrete time parameter cases with common given data $Q(f)$ and $r(f)$.

From (i)–(v) given above and their discrete parameter analogs in Section 2, it is clear that f^∞ is transient in both the continuous and discrete parameter cases or in neither. Moreover, if f^∞ is transient, then $V(f^\infty) = -Q(f)^{-1}r(f)$ is the expected return from f^∞ in *both* the continuous and discrete parameter decision processes. We refer to this observation as the *equivalence principle* in the sequel.

Maximal expected reward. Under the hypothesis that $\|P(f)\| < 1$ for $f \in F$, Howard [20, page 118] showed that there is a stationary policy maximizing $V(\cdot)$ over all stationary policies using his policy improvement method. In this case, Rykov [31] and later, independently Miller [28], showed that $V(\cdot)$ assumes its maximum over all policies among the stationary policies. We show here that these results continue to hold under the weaker hypothesis that every stationary policy is transient.

The next lemma is basic to what follows (c.f., Lemma 1). It is the infinite horizon analog of a result of Miller [29].

LEMMA 11. *If* $\pi = (g_t)$ *and* $\pi^* = (f_t)$ *are transient, then*

$$V(\pi) - V(\pi^*) = \int_0^\infty P(t, \pi) v(g_t, \pi^*)\,dt \tag{47}$$

where $v(g, \pi^*) \equiv r(g) + Q(g)V(\pi^*)$. *If also* $\pi = g^\infty$, *then*

$$V(g^\infty) - V(\pi^*) = [I - P(g)]^{-1} v(g, \pi^*).$$

PROOF. We have for $t \geqq 0$,

$$V(\pi^t, \pi^*) = \int_0^t P(u, \pi) r(g_u)\, du + P(t, \pi) V(\pi^*).$$

A computation shows

$$\frac{d}{dt} V(\pi^t, \pi^*) = P(t, \pi) v(g_t, \pi^*) \text{ a.e. } t \geqq 0.$$

Integrating over the non-negative half line completes the proof.

Now Lemma 2 and Corollaries 1 and 2 continue to hold in the continuous time parameter case because of the equivalence principle. Define V^* to be the maximal expected stationary reward exactly as in the discrete time parameter case.

Let $\mathfrak{R}(t)$ be the operator which assigns to each $V \varepsilon E^S$ the S-vector $\mathfrak{R}(t)V = V(t)$ where $V(\cdot)$ is the unique absolutely continuous solution of Bellman's [2], [4, page 321] differential equation

$$\frac{d}{dt} V(t) = \max_{g \varepsilon F} [r(g) + Q(g)V(t)], \qquad t \geqq 0, \quad V(0) = V. \tag{48}$$

It follows from (48) that

$$\mathfrak{R}(t)\mathfrak{R}(u) = \mathfrak{R}(t + u), \qquad 0 \leqq t, u \text{ and } \mathfrak{R}(0) = I \tag{49}$$

so $\mathfrak{R}(t)$ is a nonlinear one parameter Abelian semi-group [19]. (Actually (49) holds for t, u unrestricted so $\mathfrak{R}(t)$ is a one parameter Abelian group.) By a trivial extension of a result of Miller [29], we may choose $g = f_t$ maximizing the right hand side of (48) so that f_t is piecewise constant. Moreover, $\mathfrak{R}(t)V$ is the maximal expected reward over t units of time where V is the terminal reward. Thus $\mathfrak{R}(t)$ is monotone (nondecreasing).

On defining positive similarity and invariance as in the discrete parameter case and using some obvious definitions, we see that $\tilde{Q}(f) = BQ(f)B^{-1}$, $\tilde{P}(t, \pi) = BP(t, \pi)B^{-1}$, $\tilde{\mathfrak{R}}(t) = B\mathfrak{R}(t)B^{-1}$, and when π is transient, $\tilde{V}(\pi) = BV(\pi)$. Moreover, Lemma 3 holds in the continuous parameter case by the equivalence principle.

LEMMA 12. $\max_\pi \|P(t, \pi)\| \leqq e^{(\alpha-1)t}$ *for every* $t \geqq 0$ *where* $\alpha \equiv \max_g \|P(g)\|$.

PROOF. Since from Miller's theorem [29], there is a piecewise constant policy maximizing $P(t, \cdot)1$, it sufficies to establish the result for each stationary policy f^∞, say. We have

$$\|P(t, f^\infty)\| = \|e^{Q(f)t}\| = e^{-t}\|e^{P(f)t}\| \leqq e^{[\|P(f)\|-1]t},$$

which completes the proof.

Recall the definition of $\mathfrak{R}$ from Section 2.

COROLLARY 9. *Suppose every stationary policy is transient. Then* $\mathfrak{R}(t)V \to V^*$ *as* $t \to \infty$. *Moreover, for each* α *for which* $\max_g |\sigma(P(g))| < \alpha < 1$ *and each* V, *there is a constant* K *such that* $\|\mathfrak{R}(t)V - V^*\| \leqq Ke^{(\alpha-1)t}$ *for* $t \geqq 0$. *Also if* $\mathfrak{R}V \geqq V$ ($\mathfrak{R}V \leqq V$), *then* $\mathfrak{R}(t) \uparrow V^*$ ($\mathfrak{R}(t)V \downarrow V^*$) *as* $t \to \infty$.

PROOF. Because of Lemma 3 and invariance, it suffices to prove the result

under the hypothesis $\max_g \|P(g)\| < \alpha < 1$. To begin with let π' and π respectively maximize the t-period rewards with terminal rewards U and V. Then

$$P(t, \pi)(U - V) \leqq \mathfrak{R}(t)U - \mathfrak{R}(t)V \leqq P(t, \pi')(U - V),$$

so by Lemma 12,

$$\|\mathfrak{R}(t)U - \mathfrak{R}(t)V\| \leqq e^{(\alpha-1)t}\|U - V\|, \qquad t \geqq 0. \tag{50}$$

Letting $U = \mathfrak{R}(u)V$ and using (49) yields

$$\|\mathfrak{R}(t + u)V - \mathfrak{R}(t)V\| \leqq e^{(\alpha-1)t}\|\mathfrak{R}(u)V - V\|.$$

Since $\mathfrak{R}(u)V$ is bounded in $u \geqq 0$ by Lemma 12, it follows from the above inequality that $\mathfrak{R}(t)V \to V'$, say, as $t \to \infty$. Also because of (50), $\mathfrak{R}(t)$ is continuous, whence by (49)

$$V' = \lim_{u\to\infty} \mathfrak{R}(t)\mathfrak{R}(u)V = \mathfrak{R}(t)V'.$$

Thus $d/dt\,\mathfrak{R}(t)V' = 0$ for $t \geqq 0$ so from (48), $V' = \mathfrak{R}V'$. Hence, by Corollary 2, $V' = V^*$, which proves the first assertion of the Corollary. The second assertion of the Corollary follows from (50) on setting $U = V^*$ and using $\mathfrak{R}(t)V^* = V^*$.

Since there is a piecewise constant policy maximizing the t-period expected rewards [29], it suffices to prove the last assertion for the case where F contains a single element g, say. Then $d/dt\,\mathfrak{R}(t)V = e^{Q(g)t}[\mathfrak{R}V - V]$ from which the desired result follows, completing the proof.

COROLLARY 10. *The following four statements are equivalent.*

1^0. *Every (some) stationary policy is transient.*

2^0. *Every (some) periodic policy is transient.*

3^0. *Every (some) policy is transient.*

4^0. *For some* $t > 0$, $\|P(t, \pi)\| < 1$ *for every (some)* π.

If also $\|P(g)\| \leqq 1$ *for all* $g \in F$, *the above are equivalent to*

5^0. *For every* $t > 0$, $\|P(t, \pi)\| < 1$ *for every (some)* π.

PROOF. The proof that 1^0–4^0 are equivalent is an obvious analog of that for the discrete time parameter case in Corollary 4^0.

Suppose now $\|P(g)\| \leqq 1$ for all $g \in F$. Then $5^0 \Rightarrow 4^0$. Moreover, $1^0 \Rightarrow 5^0$ reading with "some". Now on reading with "every", suppose 1^0 holds. Then 5^0 holds for every stationary policy. Combining this fact with Miller's theorem [29] which implies that there is a piecewise constant policy maximizing $P(t, \cdot)1$ (and hence $\|P(t, \cdot)\|$) over all policies, completes the proof.

As in the discrete case, it is evident that the equivalence of 1^0 and 2^0 of Corollary 10 can be restated as follows.

COROLLARY 11.

$$\min_g |e^{\sigma(Q(g))}| \leqq |\sigma(P(t, \pi))|^{1/t} \leqq \max_g |e^{\sigma(Q(g))}|$$

for every $t > 0$ *and* π. *Moreover, the upper (lower) bound is attained for every* $t > 0$ *by any stationary policy* $\pi = g^\infty$ *for which* g *achieves the maximum (minimum) on the right (left).*

It follows from Lemma 11 and Corollary 10 that Theorem 1 continues to hold

in the continuous parameter case. Also Corollary 6 follows from Theorem 1, Corollary 1, and Lemma 2 just as in the discrete parameter case.

Discount optimality. In the remainder of this paper we assume that $\|P(g)\| \leqq 1$ for all $g \varepsilon F$ without further mention. Also suppose there is a (possibly) negative (real) rate of interest ρ, $-\infty < \rho < \infty$.

The S-vector of expected total discounted returns starting from each state and using the policy $\pi = (f_t)$ is

$$V_\rho(\pi) = \int_0^\infty e^{-\rho t} P(t, \pi) r(f_t)\, dt,$$

provided the integral $\int_0^\infty e^{-\rho t} P(t, \pi)\, dt$ converges. For fixed ρ, $V_\rho(\pi)$ can be expressed in the form (45) by replacing $Q(g)$ everywhere in (45) by $Q(g) - \rho I$. Thus all results given above immediately apply to the discount problem. In particular the integral $\int_0^\infty e^{-\rho t} P(t, \cdot)\, dt$ converges for every policy if and only if it converges for every stationary policy. This is always true if $\rho > 0$. It is also true for large enough $\rho \leqq 0$ in the transient case. Moreover, if $\rho > -1$ (which can always be achieved by choosing an appropriate time unit), then $V_\rho(f^\infty) = R_\rho(Q(f))r(f)$ holds for both the continuous and discrete parameter problems when $|\sigma(\beta P(f))| < 1$, or equivalently $|e^{-\rho} e^{\sigma(Q(f))}| < 1$. If these conditions hold for all $f \varepsilon F$, then it is immediate that f^∞ either maximizes $V_\rho(\cdot)$ in both cases jointly or maximizes neither.

We can define discount optimality in the continuous time parameter case by (27) and (28) just as for the discrete time parameter case. Evidently, n^- discount optimality is defined only in the transient case and we assume this whenever discussing this criterion in the sequel even though we do not repeatedly state the fact.

Now there is a stationary $\infty^\pm$ discount optimal policy in the continuous time parameter case (Rykov [31] proved this directly for the case ∞^+ using Blackwell's [5] method), viz., any stationary $\infty^\pm$ discount optimal policy in the corresponding discrete time parameter model. And any such policy is $n^\pm$ discount optimal for $n = -1, 0, 1, \cdots$. It follows without loss of optimality that we can restrict attention to stationary policies. Hence, all results in Sections 3 and 4 on existence and computation of stationary discount optimal policies carry over immediately to the continuous time parameter problem. In particular, Miller's results [28] on 0^+ discount optimality follow at once.

The reduced resolvent. There is, however, an alternate interpretation of the reduced resolvent H_ρ in the continuous time parameter case which we explore briefly. Of course, this result is not necessary for the theory given above, although it does provide an alternate path of development.

Let Q be the infinitesimal generator matrix of a finite state continuous time parameter Markov chain. We retain the innocuous assumption that $P \equiv Q + I \geqq 0$. Also assume, of course, that $Q1 \leqq 0$ (or equivalently, $\|P\| \leqq 1$). Let $P(t) = e^{Qt}$. It is well known [17], [11, page 236] that $P(t)$ converges as $t \to \infty$ to a matrix P^* satisfying $P^*P(t) = P(t)P^* = P^*P^* = P^*$ for $t \geqq 0$; hence, $P^*Q = QP^* = 0$.

The next lemma generalizes results in [24]. See also [12], [14], [19], and [22] for

thorough treatments of resolvents and semi-groups in a more abstract setting. We permit ρ to be complex in the following.

LEMMA 13. *If* $\rho \neq -1$ *and if* $|\sigma(e^{-\rho}[P(1) - P^*])| < 1$, *then* $\rho \in \sigma_0(Q)^c$ *and*

$$H_\rho^{n+1} = \int_0^\infty \frac{t^n}{n!} e^{-\rho t}[P(t) - P^*]\, dt, \qquad n = 0, 1, \cdots. \tag{51}$$

PROOF. Let $\beta = (1 + \rho)^{-1}$. For $T \geqq 0$,

$$\begin{aligned}
&[\rho I - Q + P^*]\left\{\beta P^* + \int_0^T e^{-\rho t}[P(t) - P^*]\, dt\right\} \\
&= P^* + \int_0^T [\rho I - Q]e^{-[\rho I - Q]t}\, dt - P^* \int_0^T \rho e^{-\rho t}\, dt \qquad (52) \\
&= -\int_0^T \frac{d}{dt} e^{-[\rho I - Q]t}\, dt + P^* e^{-\rho T} \\
&= I - e^{-\rho T}[P(T) - P^*].
\end{aligned}$$

Now for rational T, $e^{-\rho T}[P(T) - P^*] = \{e^{-\rho}[P(1) - P^*]\}^T$ which converges to the null matrix as $T \to \infty$ through the rationals because of the hypothesis of the lemma. Since this is so, the right hand side of (52) converges to I as $T \to \infty$ through the reals. Thus the first bracketed matrix on the left of (52) is non-singular so $\rho \in \sigma(Q - P^*)^c \subset \sigma_0(Q)^c$. Premultiplying (52) by $R_\rho(Q - P^*)$ and letting $T \to \infty$ shows

$$R_\rho(Q - P^*) = \beta P^* + \textstyle\int_0^\infty e^{-\rho t}[P(t) - P^*]\, dt.$$

Hence the integral on the right must be H_ρ so (51) holds for $n = 0$. Differentiating (51) (with $n = 0$) with respect to ρ and using (23) completes the proof.

REMARK. We remark that the hypotheses of Lemma 13 are satisfied when $|e^{-\rho}| \leqq 1$ since $\{e^{-\rho}[P(1) - P^*]\}^n = e^{-\rho n}[P(n) - P^*] \to 0$ as $n \to \infty$ whence $|\sigma(e^{-\rho}[P(1) - P^*])| < 1$. This gives a simpler proof of the fact that $[0, \infty) \subset \sigma(Q - P^*)^c$ than the corresponding proof for the discrete time parameter case given in Lemma 4. Notice also that when $\rho = 0$, the analog (25) of (51) in the discrete time parameter case is more complex than (51).

In this paper we have not considered policies in which the action taken at each time is possibly randomized and dependent on the past history of the process. It is known [7], [8], [9], [10], [31] that for a fixed interest rate ρ, there is nothing to be gained from using such policies. Thus, the same is true for $n^\pm$ discount optimality criteria.

Acknowledgment. I am indebted to Alan J. Hoffman for many stimulating comments on an earlier version of Section 2 of this paper. In particular, he obtained Lemma 3 and added the left hand inequality in Corollary 5 using different proofs than we do. He also showed that the inequalities in Corollary 5 can be deduced alternatively from a result of Bellman [4, page 329]. Lemma 3 enabled us to improve our earlier proofs of Corollary 3 and the "every" part of Corollary

4. The left hand inequality in Corollary 5 stimulated our formulation of the "some" part of Corollary 4.

I am also grateful to W. Charles Mylander, III for suggesting that the hypothesis that the norm of ρH be less than one in Theorem 2 (used in [30]) be replaced by the weaker hypothesis that its spectral radius be less than one.

REFERENCES

[1] BELLMAN, R., GLICKSBERG, I. and GROSS, O. (1954). On some variational problems occurring in the theory of dynamic programming. *Rend. Circ. Mat. Palermo.* **3** 375–376.

[2] BELLMAN, R. (1955). Functional equations in the theory of dynamic programming. II. Nonlinear differential equations. *Proc. Nat. Acad. Sci. U.S.A.* **41** 482–485.

[3] BELLMAN, R. (1957). A Markovian decision process. *J. Math. Mech.* **6** 679–684.

[4] BELLMAN, R. (1957). *Dynamic Programming.* Princeton Univ. Press.

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

[6] CODDINGTON, E. A. and LEVINSON, N. (1955). *Theory of Ordinary Differential Equations.* McGraw Hill, New York.

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

[8] DERMAN, C. (1962). On sequential decisions and Markov chains. *Man. Sci.* **9** 16–24.

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

[10] DERMAN, C. and STRAUCH, R. (1966). A note on memoryless rules for controlling sequential control processes. *Ann. Math. Statist.* **37** 276–278.

[11] DOOB, J. (1953). *Stochastic Processes.* Wiley, New York.

[12] DUNFORD, N. and SCHWARTZ, J. T. (1957). *Linear Operators.* I. Wiley, New York.

[13] DYNKIN, E. B. (1961). *Theory of Markov Processes.* (Translated by D. E. Brown and T. Kovary.) Prentice Hall, New Jersey.

[14] DYNKIN, E. B. (1965). *Markov Processes.* **1** (Translated by J. Fabius, V. Greenberg, A. Maitra, and G. Majone.) Academic Press, New York.

[15] EATON, J. H. and ZADEH, L. A. (1962). Optimal pursuit strategies in discrete state probabilistic systems. *Trans. ASME Ser. D, J. Basic Engr.* **84** 23–29.

[16] FELLER, W. (1957). *An Introduction to Probability Theory and Its Applications.* **1** (2nd ed.). Wiley, New York.

[17] FRÉCHET, M. (1938). *Méthode des Fonctions Arbitraires. Théorie des Événements en Chaine dans le cas d'un Nombre Fini d'États Possibles.* Tome 1, Fasc. 3, Livre 2 of E. Borel, *Traité du Calcul des Probabilités et de ses Applications.* Gauthier-Villars, Paris.

[18] HARDY, G. (1949). *Divergent Series.* Clarendon Press, Oxford.

[19] HILLE, E. and PHILIPS, R. S. (1957). *Functional Analysis and Semi-Groups.* (rev. ed.) American Mathematical Society, Providence, R.I.

[20] HOWARD, R. A. (1960). *Dynamic Programming and Markov Processes.* Wiley, New York.

[21] KARLIN, S. (1966). *A First Course in Stochastic Processes.* Academic Press, New York.

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

[23] KEMENY, J. G. and SNELL, J. L. (1960). *Finite Markov Chains.* Van Nostrand, Princeton.

[24] KEMENY, J. G. and SNELL, J. L. (1961). Finite continuous time Markov chains. *Theor. Probability Appl.* **6** 101–105.

[25] KNOPP, K. (1951). *Theory and Application of Infinite Series.* (4th ed.). (Translated by R. C. H. Young.) Hafner, New York.

[26] KNOPP, K. (1945). *Theory of Functions.* Part 1 (Translated by F. Bagemihl.). Dover, New York.

[27] MILLER, B. L. (1967). Finite state continuous-time Markov decision processes with ap-

plications to a class of optimization problems in queueing theory. Technical Report No. 15. Dept. of Operations Research, Stanford Univ.
[28] MILLER, B. L. (1968). Finite state continuous time Markov decision processes with an infinite planning horizon. *J. Math. Anal. Appl.* **22** 552–569.
[29] MILLER, B. L. (1968). Finite state continuous time Markov decision processes with a finite planning horizon. *SIAM J. Control.* **6** 266–280.
[30] MILLER, B. L. and VEINOTT, A. F., JR. (1969). Discrete dynamic programming with a small interest rate. *Ann. Math. Statist.* **40** 366–370.
[31] RYKOV, V. V. (1966). Markov decision processes with finite state and decision spaces. *Theor. Probability Appl.* **11** 302–311.
[32] SHAPLEY, L. S. (1953). Stochastic games. *Proc. Nat. Acad. Sci. U.S.A.* **39** 1095–1100.
[33] STRAUCH, R. E. (1966). Negative Dynamic Programming. *Ann. Math. Statist.* **37** 871–890.
[34] VEINOTT, A. F., JR. (1966). On finding optimal policies in discrete dynamic programming with no discounting. *Ann. Math. Statist.* **37** 1284–1294.
[35] VEINOTT, A. F., JR. (1969). Discrete Dynamic Programming with Sensitive Averaging Optimality Criteria. To appear.
[36] ZACHRISSON, L. (1964). Markov games. In *Advances in Game Theory*. M. Dresher, L. Shapley, and A. Tucker (eds.). Princeton Univ. Press, 211–253.