DISCOUNTED DYNAMIC PROGRAMMING

By David Blackwell[1]

University of California, Berkeley

1. Introduction. Soon after the appearance of Wald's work in sequential analysis, Richard Bellman recognized the broad applicability of the methods of sequential analysis, named this body of methods dynamic programming, and applied the methods to many problems (see [1] and papers cited there). The first development of a general theory underlying these methods is due to Karlin [6], and a rather complete analysis of the finite case was given by Howard [5]. Dubins and Savage [3] have recently developed a general theory of gambling; the relation of gambling to dynamic programming is not completely clear, but it is certainly close.

Our formulation of a dynamic programming problem is somewhat narrower than Bellman's. For us, a dynamic programming problem is specified by four objects S, A, q, r, where S, A are any non-empty Borel sets, q associates with each pair $(s, a) \in S \times A$ a probability distribution $q(\cdot \mid s, a)$ on S, and r is a bounded Baire function on $S \times A \times S$. We think of S as the set of possible states of some system, and A as the set of acts available to you. Periodically, say once a day, you observe the current state s of the system, then choose an act $a \in A$. Then the system moves to a new state s' (which will be the state you observe tomorrow), selected according to $q(\cdot \mid s, a)$, and you receive a reward $r(s, a, s')$. Your problem is, given the initial state of the system, to maximize your total expected reward over the infinite future.

This total expected reward may well be infinite, for example, if $r \equiv 1$. Or it may well be undefined. For example, if S has two elements 0, 1, A has only a single element, q is deterministic with $0 \to 1$, $1 \to 0$, and the transition $0 \to 1$ yields \$1, while $1 \to 0$ costs \$1, the series of rewards, starting in state 0, is $1 - 1 + 1 - 1 + \cdots$. We shall avoid this problem by introducing a discount factor β, $0 \leqq \beta < 1$, so that unit reward on the nth day is worth only β^{n-1}, and shall try to maximize the total discounted expected reward.

A *plan* π specifies for each $n \geqq 1$ what act to choose on the nth day as a Borel measurable function of the history $h = (s_1, a_1, \cdots, s_n)$ of the system to date or, more generally, π specifies for each h a probability distribution over A. Associated with each π is a bounded function $I(\pi)$ on S, the total expected discounted reward from π, as a function of the initial state of the system. We shall be especially interested in the (non-randomized) *stationary* plans π. A stationary π is defined by a single function f mapping S into A: whenever the system is in state s, you choose act $f(s)$.

Received 24 September 1964.

[1] Prepared with the partial support of the National Science Foundation, Grant GP-2593.

Our main results are

(1) *There need not exist an ϵ-optimal π*, i.e. we give an example in which there is an $\epsilon > 0$ such that for every π there is a π' such that

$$I(\pi') \geqq I(\pi) + \epsilon \text{ for some } s \in S. \text{ (Section 3).}$$

(2) *There always exists a (p, ϵ)-optimal stationary π^**, i.e. for any probability distribution p on S and any $\epsilon > 0$, there is a stationary π^* such that, for every π,

$$p\{I(\pi) > I(\pi^*) + \epsilon\} = 0. \text{ (Theorem 6(b)).}$$

(3) *Not every π need be dominated within ϵ by a stationary π^**, i.e. we give an example of a π and an $\epsilon > 0$ such that, for every stationary π^*,

$$I(\pi^*) < I(\pi) - \epsilon \text{ for some } s \in S. \text{ (Section 5).}$$

(4) *If A is countable, there is an ϵ-optimal stationary π^**, i.e. for every $\epsilon > 0$, there is a stationary π^* such that, for every π,

$$I(\pi) \leqq I(\pi^*) + \epsilon \text{ for all } s \in S. \text{ (Theorem 7(a)).}$$

(5) *If A is finite there is an optimal stationary π^**, i.e. there is a stationary π^* such that, for every π,

$$I(\pi^*) \geqq I(\pi) \text{ for all } s \in S. \text{ (Theorem 7(b)).}$$

(6) If there is an optimal π, there is one which is stationary. (Theorem 6(c)).

2. Probabilistic definitions and notation. By a *Borel set* we mean a Borel subset of some complete separable metric space. A *probability* on a non-empty Borel set X is a probability measure defined over the Borel subsets of X; the set of all probabilities on X is denoted by $P(X)$. For any non-empty Borel sets X, Y, a *conditional probability* on Y given X is a function $q(\cdot \mid \cdot)$ such that for each $x \in X$, $q(\cdot \mid x)$ is a probability on Y and for each Borel set $B \subset Y$, $q(B \mid \cdot)$ is a Baire function on X. The set of all conditional probabilities on Y given X is denoted by $Q(Y \mid X)$. The product space of X and Y will be denoted by XY. The set of bounded Baire functions on X is denoted by $M(X)$. For any $u \in M(XY)$ and any $q \in Q(Y \mid X)$, qu denotes the element of $M(X)$ whose value at $x_0 \in X$ is $qu(x_0) = \int u(x_0, y)\, dq(y \mid x_0)$. For any $p \in P(X)$ and any $u \in M(X)$, pu is the integral of u with respect to p. For any $p \in P(X)$, $q \in Q(Y \mid X)$, pq is the probability on XY such that, for every $u \in M(XY)$, $pq(u) = p(qu)$. Every probability m on XY has a factorization $m = pq$; p is unique and is just the marginal distribution of the first coordinate variable with respect to m; q is not quite unique; it is a version of the conditional distribution of the second coordinate variable given the first. These facts and all others in this section, except the Lemma at the end, are in [7].

We extend the above notation in an obvious way to a finite or countable sequence of non-empty Borel sets $X_1, X_2, \cdots$. If $q_n \in Q(X_{n+1} \mid X_1 \cdots X_n)$ for $n \geqq 1$ and $p \in P(X_1)$, $pq_1 \cdots q_n$ is a probability on $X_1X_2 \cdots X_{n+1}$, $pq_1q_2 \cdots$ is

a probability on the infinite product space $X_1X_2 \cdots$, $q_2q_3 \in Q(X_3X_4 \mid X_1X_2)$, for any $u \in M(X_1X_2 \cdots X_{n+1})$, $n \geqq 1$ and any m, $1 \leqq m \leqq n$, $q_m \cdots q_nu \in M(X_1 \cdots X_m)$, etc.

To avoid further complicating an already involved notation, we introduce an ambiguity as follows: for any function u on Y, we shall use the same symbol u to denote the function v on XY such that $v(x, y) = u(y)$ for all y. Thus, for example, for any $q \in Q(Y \mid X)$, $u \in M(Y)$, $qu \in M(X)$; any $q \in Q(Y \mid X)$ will also denote the element q' of $Q(Y \mid ZX)$ defined by $q'(\cdot \mid z, \cdot) = q(\cdot \mid \cdot)$, etc.

A $p \in P(X)$ is *degenerate* if it is concentrated at some one point $x \in X$; a $q \in Q(Y \mid X)$ is *degenerate* if each $q(\cdot \mid x)$ is degenerate. The degenerate q are exactly those for which there is a Baire function f mapping X into Y for which $q(\{f(x)\} \mid x) = 1$ for all $x \in X$. Any such f will also denote its associated degenerate q, so that, for any $u \in M(XY)$, $fu(x) = u(x, f(x))$ for all $x \in X$.

We shall use the following.

LEMMA [2]. *For any $q \in Q(Y \mid X)$, $u \in M(XY)$, $\epsilon > 0$ there is a degenerate $f \in Q(Y \mid X)$ such that*

$$fu \geqq qu \text{ for all } x \in X \tag{1}$$

and

$$q(\{y\colon u(x_0, y) \geqq u(x_0, f(x_0)) + \epsilon\} \mid x_0) = 0 \text{ for all } x_0 \in X. \tag{2}$$

The Lemma asserts that, in the situation where we observe $x \in X$ then choose $y \in Y$, receiving an income $u(x, y)$, any randomized plan q can be replaced by a non-randomized plan f such that (1) our expected income for each x is at least as large as it was before and (2) with probability 1, for each x, the actual income under q does not exceed the actual income under f by as much as ϵ.

3. Dynamic programming definitions and notation. A *dynamic programming problem* is defined by S, A, q, r, β, where S, A are any non-empty Borel sets, $q \in Q(S \mid SA)$, $r \in M(SAS)$, and $0 \leqq \beta < 1$. A *plan* π is a sequence $(\pi_1, \pi_2, \cdots)$, where $\pi_n \in Q(A \mid H_n)$ and $H_n = SA \cdots S(2n - 1 \text{ factors})$ is the set of possible histories of the system when the nth act must be chosen. A plan π is (non-randomized) *Markov* if each π_n is a degenerate element of $Q(A \mid S)$, i.e. $\pi = (f_1, f_2, \cdots)$, where each f_n is a Baire function from S into A, and is (non-randomized) *stationary* if there is a Baire function f mapping S into A such that $\pi_n = f$ for all n. The stationary plan defined by f is denoted by $f^{(\infty)}$.

Any plan π, together with the law of motion q of the system, defines for each initial state s a conditional distribution on the set $\Omega = ASAS \cdots$ of futures of the system, i.e. it defines an element of $Q(\Omega \mid S)$, namely $e_\pi = \pi_1q\pi_2q \cdots$. Denote the coordinate functions on $S\Omega$ by $\sigma_1, \alpha_1, \sigma_2, \alpha_2, \cdots$ so that our reward on the nth day, as a function of the history of the system, is $r(\sigma_n, \alpha_n, \sigma_{n+1})$, and our total discounted reward is $u = \sum_1^\infty \beta^{n-1}r(\sigma_n, \alpha_n, \sigma_{n+1})$. The expected total discounted reward from π, as a function of the initial state, is then

$$I(\pi) = e_\pi u = \sum_1^\infty \beta^{n-1}\pi_1q \cdots \pi_nqr.$$

(Note the use of ambiguous notation.)

For any $p \varepsilon P(S)$, and any $\epsilon > 0$, π^* will be called (p, ϵ)-*optimal* if $p\{I(\pi) > I(\pi^*) + \epsilon\} = 0$ for every π. π^* will be called ϵ-*optimal* if it is (p, ϵ)-optimal for every p, or, equivalently, if $I(\pi) \leqq I(\pi^*) + \epsilon$ for all π, s, and will be called *optimal* if it is ϵ-optimal for every $\epsilon > 0$ or, equivalently, if $I(\pi) \leqq I(\pi^*)$ for all π, s. (p, ϵ)-optimal stationary plans always exist, but p-optimal, i.e. (p, ϵ)-optimal for every $\epsilon > 0$, ϵ-optimal plans (stationary or not) may not exist, as the following examples show.

EXAMPLE 1. (There are no p-optimal plans). S has a single element, say 0, and A has countably many elements, say 1, 2, 3, $\cdots$. We take $r(0, a, 0) = (a - 1)/a$. There is no π with $I(\pi) = 1/(1 - \beta)$, but $\sup_\pi I(\pi) = 1/(1 - \beta)$.

EXAMPLE 2. (There are no ϵ-optimal plans). We take $S = A =$ unit interval $[0, 1]$. The state of the system remains fixed: $(s, a) \to s$, and the reward r is 1 or 0 according as (s, a) is in a given Borel subset $B \subset SA$ or not. For any $\pi = (\pi_1, \pi_2, \cdots)$, $\{s: \pi_1 q r > 0\}$ is a Borel subset of the projection D of B on S. For B chosen so that D is not a Borel set there is an $s_0 \varepsilon D$ for which $\pi_1 r = 0$, so that $I(\pi)(s_0) \leqq \beta + \beta^2 + \cdots = \beta/(1 - \beta)$. Since there is a π^* with $I(\pi^*)(s_0) = 1/(1 - \beta)$, π is not ϵ-optimal for any $\epsilon < 1$.

4. Existence of (p, ϵ)-optimal π.

THEOREM 1. *For any $p \varepsilon P(S)$ and any $\epsilon > 0$ there is a (p, ϵ)-optimal plan.*

The proof of Theorem 1 is simple but, regrettably, non-constructive. We associate with each π the number $pI(\pi)$, the expected return from π when the initial state has distribution p, denote by v the upper bound over all π of the numbers $pI(\pi)$, choose a sequence $\pi^{(1)}, \pi^{(2)}, \cdots$ of policies with $pI(\pi^{(n)}) \to v$, and set $u = \sup_n I(\pi^{(n)})$.

Let S_n consist of all s for which n is the smallest k with $I(\pi^{(k)}) \geqq u - \epsilon$, and let π^* be the plan which uses $\pi^{(n)}$ for all initial states $s \varepsilon S_n$, i.e.

$$\pi_m^*(\cdot \mid s_1, a_1, \cdots, s_m) = \pi_m^{(n)}(\cdot \mid s_1, a_1, \cdots, s_m) \text{ for } s_1 \varepsilon S_n.$$

Then $I(\pi^*) = I(\pi^{(n)})$ on S_n, and $I(\pi^*) \geqq u - \epsilon$ everywhere. We show that, for every π, $p\{I(\pi) \leqq u\} = 1$, which will show that π^* is (p, ϵ)-optimal. For, take any π and any $\gamma > 0$. The construction above, applied to the sequence $\pi, \pi^{(1)}, \pi^{(2)}, \cdots$ yields a π^{**} with $I(\pi^{**}) \geqq \max(u, I(\pi)) - \gamma$ everywhere. But $pI(\pi^{**}) \leqq v \leqq pu$, while $pI(\pi^{**}) \geqq p \max(u, I(\pi)) - \gamma$, so that $p \max(u, I(\pi)) \leqq pu + \gamma$. Since γ is any positive number, $p \max(u, I(\pi)) \leqq pu$, and $p\{I(\pi) \leqq u\} = 1$.

5. (p, ϵ)-domination by Markov π.

THEOREM 2. *For any $p \varepsilon P(S)$, $\epsilon > 0$, π, there is a Markov π^* which (p, ϵ)-dominates π,* i.e. $p\{I(\pi^*) \geqq I(\pi) - \epsilon\} = 1$.

PROOF. We may suppose that π is already Markov from some point on, say for $n > N$, since any two policies π, π' which agree for the first N days have $\|I(\pi') - I(\pi)\| \leqq \beta^N \|r\|/(1 - \beta)$ where, for any $u \varepsilon M(S)$, $\|u\| = \sup_s |u(s)|$. We now show that, if $\pi = (\pi_1, \cdots, \pi_N, f_{N+1}, \cdots)$ is Markov for $n > N$, for any $\gamma > 0$ there is an f_N mapping S into A with $p\{I(\pi') \geqq I(\pi) - \gamma\} = 1$,

where $\pi' = (\pi_1, \cdots, \pi_{N-1}, f_N, f_{N+1}, \cdots)$. Using this fact N times, with $\gamma = \epsilon/N$, will produce a Markov π^* which (p, ϵ)-dominates π.

To find f_N, we write $I(\pi) = \pi_1 q \cdots \pi_{N-1} q(u + \beta^{N-1}\pi_N q v)$, where $u(s_1, a_1, \cdots, s_N) = \sum_{k=1}^{N-1} \beta^{k-1} r(s_k, a_k, s_{k+1})$ and $v(s_N, a_N, s_{N+1}) = r(s_N, a_N, s_{N+1}) + (\sum_{k=1}^{\infty} \beta^k f_{N+1} q \cdots f_{N+k} q r)(s_N, a_N, s_{N+1})$. It suffices to find f_N for which

$$p\{\pi_1 q \cdots \pi_{N-1} q f_N w = \pi_1 q \cdots \pi_{N-1} q \pi_N w - \gamma\} = 1, \tag{3}$$

where $w = \beta^{N-1} q v \in M(SA)$.

Consider the probability $m = p\pi_1 q \cdots \pi_N$ on $SA \cdots SA$ ($2N$ factors), and denote the coordinate variables by $\sigma_1, \alpha_1, \cdots, \sigma_N, \alpha_N$. For any f_N, $x = \pi_1 q \cdots \pi_{N-1} q f_N w(\sigma_1)$ is a version of $E(w(\sigma_N, f_N(\sigma_N)) \mid \sigma_1)$ and $y = \pi_1 q_1 \cdots \pi_{N-1} q \pi_N w(\sigma_1)$ is a version of $E(w(\sigma_N, \alpha_N) \mid \sigma_1)$. If we choose f_N so that $w(\sigma_N, f(\sigma_N)) \geqq w(\sigma_N, \alpha_N) - \gamma$ with probability 1, we shall have $x \geqq y - \gamma$ with probability 1, which is equivalent to (3). That such an f_N exists follows at once from the Lemma of Section 2 with $X = S$, $Y = A$, q a version of the conditional distribution of α_N given σ_N, $u = w$, and $\epsilon = \gamma$. This completes the proof.

COROLLARY. *For any $p \in P(S)$, $\epsilon > 0$, there is a (p, ϵ)-optimal Markov π^*.*

PROOF. From Theorem 1 there is a $(p, \epsilon/2)$-optimal π and from Theorem 2 there is a Markov π^* which $(p, \epsilon/2)$-dominates π. This π^* is (p, ϵ)-optimal.

In Theorem 2 we cannot replace (p, ϵ)-domination by ϵ-domination. Here is an example.

EXAMPLE 3. (A plan π which cannot be ϵ-dominated by a Markov plan). We take $S = B \cup X$, where B is a Borel subset of the unit square XY whose projection D on X is not a Borel set. A is the unit interval. The law of motion q is degenerate and independent of a: $(x, y) \to x$, $x \to x$. $r(x, a, x) = 1$ if $(x, a) \in B$, $r = 0$ otherwise. Any plan π^* such that $\pi^*(\cdot \mid s_1, a_1, \cdots, s_n)$ is degenerate at y whenever $s_1 = (x, y)$ has $I(\pi^*) = \beta/(1 - \beta)$ on B. For any $\pi = (\pi_1, \pi_2, \cdots)$ for which $\pi_2 \in Q(A \mid S)$, i.e. does not depend on the initial state, the set of $x \in X$ for which $\pi_2 q r > 0$ is a Borel subset of D, so there is an $x_0 \in D$ for which $\pi_2 q r = 0$. For any y_0 with $(x_0, y_0) \in B$, we have

$$I(\pi)(x_0, y_0) \leqq \beta^2/(1 - \beta),$$

so $I(\pi) \leqq I(\pi^*) - \beta$ for some s.

6. Stationary plans and operators. Associated with each Baire function f mapping S into A is a corresponding operator T, mapping $M(S)$ into $M(S)$, defined as follows. For $u \in M(S)$, $Tu = fq(r + \beta u)$, where the u on the right, considered as a function on SAS, depends on the last coordinate only. Tu is our expected income, as a function of the initial state, if we start using $f^{(\infty)}$ but are terminated at the beginning of the second day with a final reward $u(s')$, where s' is the state at termination. $T^n u$ has a similar interpretation, replacing "second" by "n + 1st". The following properties of T, formulated as a theorem, are immediate.

THEOREM 3. (a) *T is monotone*, i.e. *$u \leqq v$ for all s implies $Tu \leqq Tv$ for all s.*

(b) *For any constant c, $T(u + c) = Tu + \beta c$.*

(c) *For any Markov $\pi = (f_1, f_2, \cdots)$, $TI(\pi) = I(f, \pi)$, where (f, π) denotes the Markov plan $(f, f_1, f_2, \cdots)$.*

For any Markov $\pi = (f_1, f_2, \cdots)$ we shall say that f mapping S into A is *π-generated* if there is a partition of S into Borel sets $S_1, S_2, \cdots$ such that $f = f_n$ on S_n; we say that a Markov $\pi' = (g_1, g_2, \cdots)$ is *π-generated* if each g_n is π-generated. We associate with each Markov π the operator U, mapping $M(S)$ into $M(S)$, defined by $Uu = \sup_n T_n u$, where T_n is the operator associated with f_n. The following interpretation of U will be justified later. $U^n u$ is our optimal expected return, over all π-generated Markov π', as a function of the initial state, if we start using π' but are terminated at the beginning of the $n + 1$st day with a final reward $u(s')$, where s' is the state at termination. Here are some basic properties of U.

THEOREM 4. (a) *U is monotone.*

(b) *For any constant c, $U(u + c) = Uu + \beta c$.*

(c) *For any T associated with a π-generated f, $Tu \leqq Uu$.*

(d) *For any $u \in M(S)$ and any $\epsilon > 0$, there is a π-generated f whose associated T satisfies $Tu \geqq Uu - \epsilon$.*

PROOF. (a), (c) are immediate. For (b) we have

$$T_n(u + c) = T_n u + \beta c \leqq Uu + \beta c,$$

so that $U(u + c) \leqq Uu + \beta c$. This inequality, with u replaced by $u + c$, c by $-c$, yields $Uu \leqq U(u + c) - \beta c$, establishing (b). For (d), let S_n consist of all s for which

$$T_i u < Uu - \epsilon \text{ for } i < n,$$

$$T_n u \geqq Uu - \epsilon,$$

and set $f = f_n$ for $s \in S_n$. Then, for any v, $Tv = T_n v$ on S_n, where T is associated with f. In particular, $Tu = T_n u \geqq Uu - \epsilon$ on S_n, so $Tu \geqq Uu - \epsilon$ everywhere.

To justify our informal interpretation of U, note that, for any Markov $\pi' = (g_1, g_2, \cdots)$, the total income from π' with termination on the $n + 1$st day with final payment u is

$$I_n(\pi', u) = T_1' T_2' \cdots T_n' u,$$

where T_i' is the operator associated with g_i. If π' is π-generated, $T_i v \leqq Uv$ for all i, so that $I_n(\pi', u) \leqq U^n u$. To find π' with $I_n(\pi', u) \geqq U^n u - \epsilon$, choose any positive numbers ϵ_i, and choose g_i π-generated so that

$$T_i' U^{n-i} u \geqq U U^{n-i} u - \epsilon_i = U^{n-i+1} u - \epsilon_i.$$

By induction downward on i, starting at $i = n$, we obtain

$$T_i' \cdots T_n' u \geqq U^{n-i+1} u - d_i,$$

where $d_i = \epsilon_i + \beta\epsilon_{i+1} + \cdots + \beta^{n-i}\epsilon_n$. For $i = 1$ we obtain $I_n(\pi', u) \geqq U^n u - d_1$, and the ϵ_i can be chosen so that $d_1 \leqq \epsilon$.

THEOREM 5. *If U is any operator with properties* (a) *and* (b), *U is a contraction with modulus β*, i.e. $\|Uu - Uv\| \leqq \beta\|u - v\|$, *so that, from the Banach fixed-point theorem, U has a unique fixed point u^*, and $\|U^n u - u^*\| \leqq \beta^n\|u - u^*\|$ for all n.*

PROOF. $v \leqq u + \|u - v\|$ yields

$$Uv \leqq U(u + \|u - v\|) = Uu + \beta\|u - v\|,$$

using (a), (b). Interchange u and v to obtain $Uu \leqq Uv + \beta\|u - v\|$, completing the proof.

The principal general results on optimal plans are contained in the following theorem. Related results are given by Dubins and Savage [3], as indicated.

THEOREM 6. (a) *For any Markov $\pi = (f_1, f_2, \cdots)$, denoting by T_n the operator associated with f_n and by $U = \sup T_n$ the operator associated with π, the fixed point u^* of U is the optimal return among π-generated plans: $I(\pi') \leqq u^*$ for every π-generated π', and for every $\epsilon > 0$ there is a π-generated f such that $I(f^{(\infty)}) \geqq u^* - \epsilon$. Any f with $Tu^* \geqq u^* - \epsilon(1 - \beta)$ satisfies this inequality.*

(b) *For any $p \in P(S)$, $\epsilon > 0$, there is a (p, ϵ)-optimal plan which is stationary.*

(c) *For any $\epsilon \geqq 0$, if there is an ϵ-optimal $\pi^* = (\pi_1, \pi_2, \cdots)$, there is an $\epsilon/(1 - \beta)$-optimal plan which is stationary* ([3], Theorem 3.9.6).

(d) *Denote for each $a \in A$ by T_a the operator associated with $f \equiv a$. Any u with $T_a u \leqq u$ for all a is an upper bound on incomes: $I(\pi) \leqq u$ for all π* ([3], Theorems 2.12.1, 3.3.1).

(e) *If for every $\epsilon > 0$ there is an ϵ-optimal plan, then the optimal return u^* is a Baire function and it satisfies the optimality equation $u^* = \sup_a T_a u^*$* ([3], Theorem 3.3.1).

(f) *A π is optimal if and only if its return $I(\pi)$ satisfies the optimality equation.*

PROOF. (a) For any π-generated $\pi' = (g_1, g_2, \cdots)$, we have $I(\pi') = T_1' \cdots T_n' u_n$, where $u_n = I(g_{n+1}, g_{n+2}, \cdots)$ and T_i' is the operator associated with g_i. Since each T_i' is a contraction with modulus β,

$$\|T_1' \cdots T_n' u_n - T_1' \cdots T_n' u^*\| \leqq \beta^n\|u_n - u^*\| \leqq \beta^n([\|r\|/(1 - \beta)] + \|u^*\|).$$

Thus $T_1' \cdots T_n' u^* \to I(\pi')$ as $n \to \infty$. But $T_1' \cdots T_n' u^* \leqq U^n u^* = u^*$, so that $I(\pi') \leqq u^*$. From Theorem 4(d), there is a π-generated f for which $Tu^* \geqq Uu^* - \epsilon' = u^* - \epsilon'$, where $\epsilon' = \epsilon(1 - \beta)$. We verify inductively that

$$T^n u^* \geqq u^* - \epsilon'(1 + \beta + \cdots + \beta^{n-1}) \text{ for all } n \geqq 1.$$

Since $T^n u^* \to I(f^{(\infty)})$, we conclude that

$$I(f^{(\infty)}) \geqq u^* - [\epsilon'/(1 - \beta)] = u^* - \epsilon.$$

(b) From the Corollary to Theorem 2, there is a $(p, \epsilon/2)$-optimal Markov $\pi = (f_1, f_2, \cdots)$. From (a), there is a stationary $f^{(\infty)}$ with $I(f^{(\infty)}) \geqq u^* - (\epsilon/2) \geqq I(\pi) - (\epsilon/2)$, where u^* is the fixed point of the U associated with π. This $f^{(\infty)}$ is (p, ϵ)-optimal.

(c) For any $\pi^* = (\pi_1, \pi_2, \cdots)$, $I(\pi^*) = \pi_1 q(r + \beta w)$, where $w \in M(SAS)$, $w(s, a, s') = I(\pi_{s,a})(s')$ and π_{sa} denotes the plan which π^* specifies, starting with the second day, when the first state and act are s, a, i.e. $\pi_{sa} = (\pi_1', \pi_2', \cdots)$, where

$$\pi_n'(\cdot \mid s_1, a_1, \cdots, s_n) = \pi_{n+1}(\cdot \mid s, a, s_1, a_1, \cdots, s_n).$$

If π^* is ϵ-optimal, $w(s, a, s') \leqq I(\pi^*)(s') + \epsilon$ for all s', so that $I(\pi^*) \leqq \pi_1 q(r + \beta I(\pi^*) + \beta\epsilon) = \pi_1 h$, say. From the Lemma of Section 2, there is an f for which $fh \geqq \pi_1 h$ for all s, so that, for the T corresponding to f, $I(\pi^*) \leqq T(I(\pi^*)) + \beta\epsilon$. By induction on n we obtain $T^n I(\alpha^*) \geqq I(\alpha^*) - \epsilon(\beta + \cdots + \beta^n)$. Letting $n \to \infty$ yields $I(f^{(\infty)}) \geqq I(\pi^*) - \beta\epsilon/(1 - \beta)$. Since π^* is ϵ-optimal, $f^{(\infty)}$ is $\epsilon + [\beta\epsilon/(1 - \beta)] = \epsilon/(1 - \beta)$-optimal.

(d) For any $s_0 \in S$ and any $\epsilon > 0$, there is a stationary $f^{(\infty)}$ such that

$$I(\pi)(s_0) \leqq I(f^{(\infty)})(s_0) + \epsilon \text{ for all } \pi;$$

just choose $f^{(\infty)}(p, \epsilon)$-optimal, where p is concentrated on s_0. $T_a u \leqq u$ for all a implies $Tu \leqq u$ for all T and in particular for the T associated with f. Thus $T^n u$ decreases to $I(f^{(\infty)})$ and $I(f^{(\infty)}) \leqq u$. Then $I(\pi)(s_0) \leqq u(s_0) + \epsilon$. Letting $\epsilon \to 0$ completes the proof.

(e) From (c), the hypothesis implies that there is a $1/n$-optimal stationary plan $f_n^{(\infty)}$ say. With $\pi = (f_1, f_2, \cdots)$, the fixed point u^* of the U associated with π is, from (a), the optimal return among π-generated policies. In particular $u^* \geqq I(f_n^{(\infty)})$, so that $u^* \geqq I(\pi)$ for all π, and u^* is the optimal return. We have $\sup_a T_a u^* \geqq U u^* = u^*$. On the other hand, for any $a \in A$,

$$T_a u^* \leqq T_a(I(f_n^{(\infty)}) + (1/n)) = I(a, f^{(\infty)}) + \beta/n \leqq u^* + \beta/n,$$

where $(a, f^{(\infty)})$ is the Markov policy $(g, f, f, f, \cdots)$ with $g \equiv a$. Letting $n \to \infty$ yields $T_a u^* \leqq u^*$. Thus u^* satisfies the optimality equation.

(f) If $I(\pi^*)$ satisfies the optimality equation, we obtain from (d), with $u = I(\pi^*)$, that π^* is optimal. Conversely, if π^* is optimal, the hypothesis of (e) is satisfied, so that u^*, the optimal return, does satisfy the optimality equation.

REMARKS. (d) is extremely useful in proving optimality; if u is known to be the return from a policy π and u satisfies $T_a u \leqq u$ for all u, (d) implies that π is optimal. The criterion for optimality in (e), (f) was stated in general, without proof, by Bellman [1]. We do not know whether the optimal return always satisfies the optimality equation or whether, even under the hypothesis of (e), the (bounded) solution is unique.

7. Further results. If A is countable, with elements $a_1, a_2, \cdots$, every Markov plan is π^*-generated, where $\pi^* = (g_1, g_2, \cdots)$ and $g_n \equiv a_n$. Conversely, for any pure Markov $\pi = (f_1, f_2, \cdots)$, the study of π-generated plans can be reduced to the countable A case by interpreting act n in state s as the selection of $f_n(s)$. We prefer to keep the original A, and introduce the concept of essential countability as follows. Two acts a and b will be called *equivalent at state* s if

$$r(s, a, \cdot) = r(s, b, \cdot) \text{ and } q(\cdot \mid s, a) = q(\cdot \mid s, b),$$

i.e. if $T_a u(s) = T_b u(s)$ for all $u \in M(S)$. For any Markov $\pi = (f_1, f_2, \cdots)$, A will be called *essentially countable by* π if for every (s, a) there is an n for which $f_n(s)$ is equivalent to a at s. A will be called *essentially finite by* π if there is a partition of S into Borel sets $S_1, S_2, \cdots$ such that for every (s, a) with $s \in S_n$, at least one of the acts $f_1(s), \cdots, f_n(s)$ is equivalent to a at s.

THEOREM 7. (a) *If A is essentially countable by $\pi = (f_1, f_2, \cdots)$, the fixed point u^* of the operator U associated with π is the optimal return. U is identical with the operator* $\sup_a T_a$, *so that u^* is the unique (bounded) solution of the optimality equation. For every $\epsilon > 0$ there is an ϵ-optimal stationary plan.*

(b) *If A is essentially finite by $\pi = (f_1, f_2, \cdots)$, there is an optimal stationary plan.*

PROOF. (a) For any u, s, $T_n u(s) = T_a u(s)$, where $a = f_n(s)$. Thus $Uu \leqq \sup_a T_a u$. But for any $a \in A$, $T_a u(s) = T_n u(s)$ for some n, so that $T_a u(s) \leqq Uu(s)$, and $\sup_a T_a u(s) \leqq Uu(s)$. Thus the operators $\sup_a T_a$, U are identical. Theorem 6 (d) then implies $I(\pi) \leqq u^*$ for all π. From Theorem 6 (a) there is a stationary $f^{(\infty)}$ with $I(f^{(\infty)}) \geqq u^* - \epsilon$. This $f^{(\infty)}$ is ϵ-optimal.

(b) If A is essentially finite, define B_n as the set of all s for which n is the smallest i with $T_i u^*(s) = \sup_n T_n u^*(s)$ (the sequence $\{T_n u^*(s)\}$ contains only finitely many different numbers). Define $f = f_n$ on B_n, so that $Tu^* = Uu^* = u^*$, where T is associated with f. Then u^*, as the fixed point of T, is the return from $f^{(\infty)}$, and $f^{(\infty)}$ is optimal.

We conclude with the extension of the improvement routines given by Howard [5] and Eaton and Zadeh [4] for the case of finite S, A.

THEOREM 8. (a) (*Howard improvement*). *If $I(g, \pi) \geqq I(\pi)$, then $I(g^{(\infty)}) \geqq I(g, \pi) \geqq I(\pi)$.*

(b) (*Eaton-Zadeh improvement*). *For any f, g mapping S into A, define $h = f$ on $I(f^{(\infty)}) \geqq I(g^{(\infty)})$, $h = g$ on $I(g^{(\infty)}) > I(f^{(\infty)})$. Then $I(h^{(\infty)}) \geqq \max(I(f^{(\infty)}), I(g^{(\infty)}))$.*

PROOF. (a) If T is associated with g, we have $TI(\pi) = I(g, \pi) \geqq I(\pi)$, so that

$$T^n I(\pi) \uparrow I(g^{(\infty)}) \text{ and } I(g^{(\infty)}) \geqq I(g, \pi).$$

(b) (Proof by Ashok Maitra). If T_1, T_2, T are associated with f, g, h, we have, for any u,

$$Tu = T_1 u \text{ on } I(f^{(\infty)}) \geqq I(g^{(\infty)})$$

$$Tu = T_2 u \text{ on } I(g^{(\infty)}) > I(f^{(\infty)}).$$

With $u = \max(I(f^{(\infty)}), I(g^{(\infty)}))$, we obtain

$$Tu = T_1 u \geqq T_1 I(f^{(\infty)}) = I(f^{(\infty)}) = u \text{ on } I(f^{(\infty)}) \geqq I(g^{(\infty)}),$$

$$Tu = T_2 u \geqq T_2 I(g^{(\infty)}) = I(g^{(\infty)}) = u \text{ on } I(g^{(\infty)}) > I(f^{(\infty)}).$$

Thus $Tu \geqq u$, so that $I(h^{(\infty)}) \geqq u$.

REFERENCES

[1] BELLMAN, RICHARD (1957). *Dynamic Programming*. Princeton Univ. Press.
[2] BLACKWELL, D. (1964). Memoryless strategies in finite stage dynamic programming. *Ann. Math. Statist.* **35** 863–865.
[3] DUBINS, L. E. and SAVAGE, L. J. (1963). *How to gamble if you must* (dittoed draft).
[4] EATON, J. H. and ZADEH, L. A. (1961). Optimal pursuit strategies in discrete state probabilistic systems. *J. Basic Engineering Ser.* D **84** 23–29.
[5] HOWARD, RONALD A. (1960). *Dynamic Programming and Markov Processes*. Wiley, New York.
[6] KARLIN, S. (1955). The structure of dynamic programming models. *Naval Res. Logist. Quart.* **2** 285–294.
[7] LOÈVE, M. (1960). *Probability Theory*. Van Nostrand, Princeton.