NEGATIVE DYNAMIC PROGRAMMING[1]

By Ralph E. Strauch

University of California, Berkeley

1. Introduction. A dynamic programming problem is determined by four objects, S, A, q, and r. S and A are non-empty Borel sets, q is a regular conditional probability on S given $S \times A$, and r is a Baire function on $S \times A \times S$. We interpret S as the set of states of some system, and A as the set of actions available at each state. (The set of actions available is assumed to be independent of the state.) When the system is in state s and we take action a, we move to a new state s' selected according to $q(\cdot \mid s, a)$, and we receive a return $r(s, a, s')$. The process is then repeated from the new state s', and we wish to maximize the total expected return over the infinite future.

A *policy* π is a sequence $\pi_1, \pi_2, \cdots$, where π_n is a regular conditional probability on A given $h = (s_1, a_1, \cdots, a_{n-1}, s_n)$, the history of the system up to the nth stage. Given that we have experienced history h up to the nth stage, we choose the nth action according to $\pi_n(\cdot \mid h)$. Certain types of policies are of special interest. A *random semi-Markov policy* is one in which π_n depends only on s_1 and s_n, and a *random Markov policy* is one in which π_n depends only on s_n. A *non-random* policy is one in which each π_n is degenerate, i.e. is a measurable function from histories to actions. A *semi-Markov policy* is a sequence $f_1, f_2, \cdots$, where each f_n is a measurable function from $S \times S$ to A, and $f_n(s_1, s_n)$ is the action we take at the nth stage if we start in state s_1 and the nth state is s_n. A *Markov policy* is a sequence $f_1, f_2, \cdots$ where each f_n is a measurable function from S to A and $f_n(s)$ is the action we choose at the nth stage if the nth state is s. (Notice that the term *Markov policy* and *semi-Markov policy* refer to non-random policies, and are modified by the adjective *random* if the elements of the policies are probability distributions.) A *stationary policy* is a Markov policy in which $f_n = f$ for some measurable f from S to A and all n.

If $\pi = \{f_1, f_2, \cdots\}$ is a Markov policy, the function of g is *π-generated* if there exists a measurable partition $S_1, S_2, \cdots$ of S such that $g = f_n$ on S_n. A Markov policy $\pi' = \{g_1, g_2, \cdots\}$ is *π-generated* if each g_n is *π-generated*.

Associated with each π is a Baire function on S, $I(\pi)(s)$, the total expected return starting from s and using π. This total return may well be infinite, or may be undefined. There are, however, three cases in which the problem is well defined, which may be described as follows:

(a) *The discounted case.* If the return function r is bounded, and we discount our future return with a discount factor β, $0 \leqq \beta < 1$, so that a return of one

Received 17 May 1965.

[1] This paper is the author's doctoral dissertation at the University of California, Berkeley, and was written with the partial support of a National Science Foundation Cooperative Graduate Fellowship, and the Office of Naval Research, contract NONR 222-43.

[2] Now at The RAND Corporation.

unit n stages in the future is worth β^n now, then the actual total return, hence the total expected return, will be bounded. This case has been studied by Blackwell in [3].

(b) *The positive bounded case.* In this case we assume that the return function is non-negative and bounded, and that the structure of the problem is such that the expected return from any policy is bounded by R, where R is a positive number independent of the policy chosen. This case has been studied by Blackwell in [2].

(c) *The negative case.* In this case we assume that the return function is non-positive. The total return is thus well defined, but may be $-\infty$. The problem is clearly of interest only if there is a policy with finite return, but we find it convenient not to require this at the outset. A more natural setting for this case might be to assume a non-negative cost function, and ask how to minimize cost, but our formulation has the advantage of being consistent with the other two cases.

The purpose of this paper is to make a study of the negative case similar to those made by Blackwell of the other two cases in [2] and [3], and to answer in all three cases some questions, primarily concerned with measurability, raised by Blackwell in [2] and [3].

Our main results are the following: Every policy can be replaced by a semi-Markov policy which dominates it in the discounted or negative cases, or which ϵ-dominates it in the positive case (Section 4). In all cases, the optimal return, $\sup_\pi I(\pi)$ is absolutely measurable, and satisfies the optimality equation. For any probability p on S and $\epsilon > 0$, there exists a policy π^* such that $p\{I(\pi^*) \geqq \sup_\pi I(\pi) - \epsilon\} = 1$. In the negative case, if there is an optimal policy, there is one which is stationary (Sections 7 and 8), and if A is finite, there is an optimal policy (Section 9). Not every Markov policy is dominated by a stationary policy, but every sequence of Markov policies is ϵ-dominated by a Markov policy (Section 6). In the negative and discounted cases, if we are given two policies and at each stage use the one which would be better if we were to continue to use it from that point onward, then the resulting policy is as good as either of the two given ones (Section 9).

Our methods differ from those used by Blackwell in the following manner: His general method of proof may loosely be described as neglecting the tails, while ours may loosely be described as improving on the tails. That is, he asks how to behave if we are going to play for a large finite number of stages, then stop, receiving no terminal return. We ask how to behave if we are going to play for some finite number of stages, then receive some previously fixed terminal return. The former method works in the discounted and positive bounded cases, but not in the negative case, while the latter works in the discounted and negative cases, but not in the positive bounded case.

Throughout the paper, we shall indicate the cases for which each result holds by use of the letters D, P, and N. If the result is previously known in a particular case, the identifying letter will be enclosed in parentheses, and a proof will be

given only if our method of proof differs substantially from Blackwell's, or if it follows from a minor modification of the proof in the negative case. Throughout the paper, we shall denote the completion of a proof by □.

2. Probabilistic definitions and notation. In this section we develop the probabilistic notation to be used throughout the paper. We follow [3] as closely as possible. All facts in this section with the exception of Lemma 2.1 are contained in [9].

A Borel set X is a Borel subset of a complete separable metric space. Unless otherwise indicated, measurable means measurable with respect to the σ-field of Borel subsets of X. This measurability structure will not be explicitly indicated. A probability on a non-empty Borel set X is a probability measure defined on the Borel subsets of X, and the set of all probabilities on X is denoted by $P(X)$. If X and Y are non-empty Borel sets, a (regular) *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 subset B of Y, $q(B \mid \cdot)$ is a Baire function on X. We denote the cartesian product of X and Y by XY. Every probability $m \in P(XY)$ has a factorization $m = pq$, where $p \in P(X)$ is the marginal distribution of the first coordinate variable under m, and $q \in Q(Y \mid X)$ is a version of the conditional distribution of the second coordinate variable given the first.

If X is a non-empty Borel set, then $M(X)$ will have one of two possible meanings, depending on the case under consideration. In the discounted or positive bounded cases, $M(X)$ will denote the set of all bounded Baire functions on X, and in the negative case, $M(X)$ will denote the set of non-positive, extended real-valued Baire functions on X. Unless otherwise noted, when no case is specifically mentioned statements made about elements of M will be valid for either definition. If $u, v \in M(X)$, $u \geqq v$ means $u(x) \geqq v(x)$ for all $x \in X$, and in the discounted and positive bounded cases, $\|u\| = \sup u(x), x \in X$. For any $p \in P(X)$, $u \in M(X)$, pu is the integral of u with respect to p. For any $u \in M(XY)$ and any $q \in Q(Y \mid X)$, qu denotes the element of $M(X)$ whose value at $x \in X$ is given by

$$qu(x) = \int_Y u(x, y)\, dq(y \mid x).$$

For any $p \in P(X)$, $q \in Q(Y \mid X)$ pq is the probability on XY such that for all $u \in M(XY)$, $pq(u) = p(qu)$.

The above notation extends in an obvious way to a finite or infinite 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 $p \in P(X_1)$, then $pq_1 \cdots q_n \in P(X_1 \cdots X_{n+1})$, $pq_1q_2 \cdots \in P(X_1X_2 \cdots)$, $q_2q_3 \in Q(X_3X_4 \mid X_1X_2)$ and for any $u \in M(X_1 \cdots X_{n+1})$, $m \leqq n$, $q_m \cdots q_nu \in M(X_1 \cdots X_m)$, etc. To avoid further complicating the notation we shall use the following convention: 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, if $q \in Q(Y \mid X)$, $u \in M(Y)$, then $qu \in M(X)$, and $q \in Q(Y \mid X)$ will also denote the element $q' \in Q(Y \mid ZX)$ such that $q'(\cdot \mid z, \cdot) = q(\cdot \mid \cdot)$, etc.

A $p \varepsilon P(X)$ is *degenerate* if for some $x \varepsilon X$, $p\{x\} = 1$, and will sometimes be denoted by $\delta(x)$. A $q \varepsilon Q(Y \mid X)$ is *degenerate* if $q(\cdot \mid x)$ is degenerate for each x, and this happens if and only if there is a measurable function f from X to Y such that $q(\cdot \mid x) = \delta(f(x))$ for each x. We will also denote by f the associated degenerate q, so that for $u \varepsilon M(XY)$, $fu(x) = u(x, f(x))$ for all $x \varepsilon X$.

We shall need the following lemma.

LEMMA 2.1. *For any $q \varepsilon Q(Y \mid X)$ and any $u \varepsilon M(XY)$, there is a degenerate $f \varepsilon Q(Y \mid X)$ such that $fu \geqq qu$.*

That is, if we observe x, and then choose $y \varepsilon Y$ according to $q(\cdot \mid x)$ and receive a return of $u(x, y)$, any random plan q can be replaced by a non-random plan f with the property that our expected income under f is at least as great as it was under q for each x. The lemma is contained in Lemma 2 of [3], and the proof is immediate from Theorem 2 of [4].

3. The dynamic programming problem. A *dynamic programming problem* is defined by four elements, S, A, q, r, and in the discounted case, a discount factor β. The *set of states* S and *set of actions* A are non-empty Borel sets. The *law of motion* q is an element of $Q(S \mid SA)$. The *return function* $r \varepsilon M(SAS)$, with the following additional restrictions. In the positive bounded case $r \geqq 0$, and in the negative case we assume $r > -\infty$ and $qr > -\infty$, i.e. both the actual return and the expected return from any action at any state are finite. In the discounted case, $0 \leqq \beta < 1$, and in the positive bounded and negative cases, $\beta = 1$. (We introduce the discount factor $\beta = 1$ in these cases only to allow a common notation throughout the paper.) A *policy* π is a sequence $(\pi_1, \pi_2, \cdots)$ where $\pi_n \varepsilon Q(A \mid H_n)$ and $H_n = SA \cdots S$ ($2n - 1$ factors) is the set of possible histories of the process when the nth action must be chosen. A policy π is (non-random) *Markov* if each π_n is a degenerate element of $Q(A \mid S)$, i.e., if $\pi = \{f_1, f_2, \cdots\}$ where each f_n is a measurable function from S to A, and is *random Markov* if each $\pi_n \varepsilon Q(A \mid S)$. It is (non-random) *semi-Markov* if each π_n is a degenerate element of $Q(A \mid SS)$, and is *random semi-Markov* if each $\pi_n \varepsilon Q(A \mid SS)$. The interpretation is as follows: a Markov policy is one such that the action taken at the nth stage depends only on the nth state and the integer n, while a semi-Markov policy is one such that the action taken at the nth stage depends only on the initial state, the nth state, and the integer n. A *stationary* policy is a Markov policy such that $f_n = f$ for some f and all n. The stationary policy defined by f will be denoted by $f^{(\infty)}$.

For any policies π and σ, let $\pi^n\sigma = \{\pi_1, \cdots, \pi_n, \sigma_{n+1}, \cdots\}$ denote the policy which follows π for n stages then switches to σ. If π is Markov or random Markov, we let ${}^n\pi = \{\pi_{n+1}, \pi_{n+2}, \cdots\}$ denote the policy which π defines from the $n + 1$st stage onward. In particular, ${}^0\pi = \pi$. We shall denote the nth policy in a sequence of policies by π^n.

Any policy π, together with the law of motion q, defines a conditional probability on the set $X = ASAS \cdots$ of futures of the system given the initial state s, i.e. it defines

$$e_\pi = \pi_1 q \pi_2 q \cdots \varepsilon Q(X \mid S).$$

Any return function r defines a total discounted return function on SX given by

$$\rho(s, x) = \sum_{n=1}^{\infty} r(s_n, a_n, s_{n+1})\beta^{n-1}$$

and an expected return function on S given by

$$I(\pi) = e_\pi \rho = \sum_{n=1}^{\infty} \beta^{n-1}\pi_1 q \cdots \pi_n q r.$$

For any $v \in M(S)$, we shall denote by $I_n(\pi, v)$ the expected return if we terminate after the nth stage and receive a terminal reward $v(s_{n+1})$ at the terminal state. Thus

$$I_n(\pi, v)(s_1) = e_\pi[\sum_{j=1}^{n} \beta^{j-1} r(s_j, a_j, s_{j+1}) + \beta^n v(s_{n+1})].$$

We shall denote $I_n(\pi, 0)$ by $I_n(\pi)$. It is clear that if $u \leqq v$ then $I_n(\pi, u) \leqq I_n(\pi, v)$ for all π and all n. In the discounted case it is clear that for any u, $v \in M(S)$, $\|I_n(\pi, u) - I_n(\pi, v)\| \leqq \beta^n \|u - v\|$. In particular, if π and π' are policies such that $\pi_n = \pi_n'$ for $n \leqq N$ then $\|I(\pi) - I(\pi')\| \leqq 2\beta^N \|r\|/(1 - \beta)$. From the dominated and monotone convergence theorems it is clear that

LEMMA 3.1. D. $I_n(\pi) \to I(\pi)$;
P. $I_n(\pi) \uparrow I(\pi)$;
N. $I_n(\pi) \downarrow I(\pi)$.

For any $p \in P(S)$ and $\epsilon > 0$, we say that π^* is (p, ϵ)-*optimal* if $p\{I(\pi^*) \geqq \sup_\pi I(\pi) - \epsilon\} = 1$. The above set is in general not Borel, however we show in Section 7 that it is in the completion of the Borel sets with respect to p, hence the statement has meaning. (Note that this definition of (p, ϵ)-optimality is stronger than that given by Blackwell in [3].) We say π^* is ϵ-*optimal* if $I(\pi^*) \geqq \sup_\pi I(\pi) - \epsilon$. We say that π^* is *p-optimal* or *optimal* if the corresponding statements above hold for $\epsilon = 0$, and that π^* (p, ϵ)-*dominates*, ϵ-*dominates*, *p-dominates*, or *dominates* π if the corresponding statements hold with $\sup I(\pi)$ replaced by $I(\pi)$.

4. Semi-Markov policies are enough. It seems clear that we should be able to restrict our attention to Markov policies. We are interested in maximizing our total return, so at any stage we want to maximize our total return from that stage onward, and should be able to do this knowing only our present state, without regard to our past history. In other words, we would expect that, given any policy π, and any $\epsilon > 0$ there would exist a Markov policy π' such that $I(\pi) \leqq I(\pi') + \epsilon$. Blackwell has shown in [3] that this is not the case. We reproduce his example here with slight modification.

EXAMPLE 4.1. Let $X = Y = (0, 1)$, and let $S = B \cup X \cup \{0\}$, where B is a Borel subset of the unit square XY whose projection D on X is not Borel. Let $A = (0, 1)$. The law of motion q is degenerate and independent of a: $q(\cdot \mid (x, y), a) = \delta(x)$, $q(\cdot \mid x, a) = \delta(0)$ and $q(\cdot \mid 0, a) = \delta(0)$; $r(s, a, s') = 1$ if $s \in X$ and $(s, a) \in B$, and $r = 0$ otherwise. (The example as stated is for the positive or discounted case, but applies with obvious modification to the negative case.) Any policy π^* such that $\pi_2^*(\cdot \mid s_1, a_1, s_2)$ is degenerate at y whenever $s_1 = (x, y)$ has $I(\pi^*) = \beta$ on B, but for any π 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 $\pi_2 qr > 0$ is a Borel subset of D. Pick $x_0 \in D$ such that $\pi_2 qr = 0$. For any y_0 with $(x_0, y_0) \in B$, we have $I(\pi)(x_0, y_0) = 0$, hence there does not exist a Markov π such that $I(\pi^*) \leqq I(\pi) + \epsilon$ for any $\epsilon < 1$. The same example shows that there need not exist an ϵ-optimal policy, since for any $x \in D$ there exists a π such that $I(\pi)(x) = 1$, but there does not exist a π' such that $I(\pi')(x) = 1$ for all $x \in D$.

We do not, however, need to remember the entire past, but only the initial state. According to the next theorem, given any policy π, there is a random semi-Markov policy π^* such that $I(\pi^*) = I(\pi)$ for any return function r, so that we can replace the given policy π with π^* without knowing the return function, and be sure of the same expected return. If the initial state s is not arbitrary, but is chosen according to some $p \in P(S)$, we can replace π with a random Markov π^{**} with no loss in expected return.

THEOREM 4.1. *D, P, N. Let π be any policy, $p \in P(S)$. Then there exists a random semi-Markov policy π^* and a random Markov policy π^{**} such that $I(\pi) = I(\pi^*)$ and $pI(\pi) = pI(\pi^{**})$ for any return function r.*

PROOF. Let $\pi_n{}^*$ be the conditional distribution of a_n given s_n and s_1 under e_π, and let $\pi_n{}^{**}$ be the conditional distribution of a_n given s_n under pe_π. We need the following lemma.

LEMMA 4.1. *D, P, N.*

(a) *For any n, and any $r \in M(SSAS)$,*

$$e_\pi r(s_1, s_n, a_n, s_{n+1}) = e_{\pi^*} r(s_1, s_n, a_n, s_{n+1})$$

(b) *For any n, and any $r \in M(SAS)$,*

$$pe_\pi r(s_n, a_n, s_{n+1}) = pe_{\pi^{**}} r(s_n, a_n, s_{n+1}).$$

PROOF OF THE LEMMA. We shall prove (a), the proof of (b) is similar. The lemma is true for $n = 1$, since $\pi_1{}^* = \pi_1$, hence

$$e_\pi r(s_1, a_1, s_2) = \pi_1 qr = \pi_1{}^* qr = e_{\pi^*} r(s_1, a_1, s_2).$$

Now assume the lemma true for $n < N$. All expectations are under the conditional probability e_π.

$$\begin{aligned} e_\pi r(s_1, s_N, a_N, s_{N+1}) &= E[r(s_1, s_N, a_N, s_{N+1}) \mid s_1] \\ &= E\{E(r(s_1, s_N, a_N, s_{N+1}) \mid s_1, s_N) \mid s_1\} \\ &= E\{u(s_1, s_N) \mid s_1\} \\ &= e_\pi u(s_1, s_N) \end{aligned}$$

where $u(s_1, s_N) = E[r(s_1, s_N, a_N, s_{N+1}) \mid s_1, s_N] = \pi_N{}^* qr(s_1, s_N, a_N, s_{N+1})$ by the properties of conditional distributions. But $u(s_1, s_N) = v(s_1, s_{N-1}, a_{N-1}, s_N) \in M(SSAS)$, and so by the induction hypothesis, $e_\pi u(s_1, s_N) = e_{\pi^*} u(s_1, s_N)$.

Thus

$$\begin{aligned} e_\pi r(s_1, s_N, a_N, s_{N+1}) &= e_\pi u(s_1, s_N) \\ &= e_{\pi^*}\pi_n^* qr(s_1, s_N, a_N, s_{N+1}) \\ &= e_{\pi^*} r(s_1, s_N, a_N, s_{N+1}). \end{aligned}$$ □

We now return to the proof of the theorem. From Lemma 4.1 it follows that, for any return function r, $I_n(\pi) = I_n(\pi^*)$, and $pI_n(\pi) = pI_n(\pi^{**})$ for all n. From Lemma 3.1 it follows that $I(\pi) = I(\pi^*)$ and $pI(\pi) = pI(\pi^{**})$. □

Theorem 4.1 is the only theorem which applies independently of the return function r. Throughout the remainder of the paper we will assume the return function r is fixed.

Suppose we have two policies, π and σ, with the property that for any n, or at least for n sufficiently large, it is better to use π for n stages and then switch to σ than it is to use σ from the beginning. Then, in the discounted and negative case, it is better to use π forever than to use σ forever.

THEOREM 4.2. *D, N. If π and σ are policies for which there exists an n_0 such that for $n \geqq n_0$, $I(\pi^n\sigma) \geqq I(\sigma)$ then $I(\pi) \geqq I(\sigma)$.*

PROOF. *N.* $I_n(\pi) \geqq I(\pi^n\sigma) \geqq I(\sigma)$ for $n \geqq n_0$, and $I_n(\pi) \downarrow I(\pi)$ hence $I(\pi) \geqq I(\sigma)$. *D.* $\|I(\pi) - I(\pi^n\sigma)\| \leqq 2\beta^n\|r\|/(1-\beta)$, hence $I(\pi) \geqq I(\sigma)$. □

We show by example that Theorem 4.2 may fail in the positive case.

EXAMPLE 4.2. Let $S = \{1, 2, \cdots\}$ and $A = \{0, 1\}$. Let $r(s, 0) = 1 - 1/s$, $r(s, 1) = 0$, $q(\cdot \mid s, 0) = \delta(1)$ for all s, $q(\cdot \mid 1, 1) = \delta(1)$ and $q(\cdot \mid s, 1) = \delta(s+1)$ if $s > 1$. State 1 is a terminal state. From any other state s we can either move to 1 and receive $1 - 1/s$ or move to $s+1$ and receive nothing. Let $\pi = \{f_1, f_2, \cdots\}$ and $\sigma = \{g_1, g_2, \cdots\}$ be Markov policies such that $f_n(s) = 1$ and $g_n(s) = 0$ for all n and all s. Then $I(\sigma)(s) = 1 - 1/s$, $I(\pi^n\sigma)(1) = 0$, $I(\pi^n\sigma)(s) = 1 - 1/(s+n)$ for $s > 1$, hence $I(\pi^n\sigma) \geqq I(\sigma)$ for all n, but $I(\pi) = 0$.

We can use Theorem 4.2 to obtain

THEOREM 4.3. *D, N. For any policy π and any $p \in P(S)$ there exists a semi-Markov policy τ and a Markov policy σ such that $I(\tau) \geqq I(\pi)$ and $pI(\sigma) \geqq pI(\pi)$.*

PROOF. We shall give the construction for τ. The construction for σ is similar. Because of Theorem 4.1, we may assume π is random semi-Markov. For each n, using Lemma 2.1, we can find f_n mapping SS into A such that

$$f_n \textstyle\sum_{k=n}^{\infty} \beta^{k-1} q\pi_{n+1}q \cdots \pi_k qr \geqq \pi_n \sum_{k=n}^{\infty} \beta^{k-1} q\pi_{n+1} \cdots \pi_k qr.$$

When $k = n$, $q\pi_{n+1}q \cdots \pi_k qr = r$ by convention.

Let $\tau = \{f_1, f_2, \cdots\}$. Then

$$\begin{aligned} I(\pi) &= \pi_1 \textstyle\sum_{j=1}^{\infty} \beta^{j-1} q\pi_2 \cdots \pi_j qr \\ &\leqq f_1 \textstyle\sum_{j=1}^{\infty} \beta^{j-1} q\pi_2 \cdots \pi_j qr = I(\tau^1\pi) \end{aligned}$$

and for $n \geqq 1$,

$$\begin{aligned} I(\tau^n\pi) &= (\textstyle\sum_{j=1}^{n} \beta^{j-1} f_1 q \cdots f_j q r) + (f_1 q \cdots f_n q \pi_{n+1} \textstyle\sum_{j=n+1}^{\infty} \beta^{j-1} q \pi_{n+2} q \cdots \pi_j q r) \\ &\leqq (\textstyle\sum_{j=1}^{n} \beta^{j-1} f_1 q \cdots f_j q r) + (f_1 q \cdots f_n q f_{n+1} \textstyle\sum_{j=n+1}^{\infty} \beta^{j-1} q \pi_{n+2} q \cdots \pi_j q r) \\ &= I(\tau^{n+1}\pi). \end{aligned}$$

Hence $I(\pi) \leqq I(\tau^1\pi) \leqq I(\tau^2\pi) \leqq \cdots$, so by Theorem 4.2, $I(\tau) \geqq I(\pi)$. The construction of σ is similar except that we start with the assumption that π is Markov and find f_n mapping S into A satisfying the same inequality. □

Even without Theorem 4.2, we are able to obtain a slightly weaker result in the positive case.

THEOREM 4.4. *P. For any policy π, any $\epsilon > 0$ and any $p \in P(S)$, there exists a semi-Markov policy τ and a Markov policy σ such that $I(\tau) \geqq I(\pi) - \epsilon$ and $pI(\sigma) \geqq pI(\pi) - \epsilon$.*

PROOF. Again because of Theorem 4.1, we may assume that π is random semi-Markov. Using Lemma 2.1, we can find, for $m > 0$, $n \leqq m$ an f_{mn} mapping SS into A such that $f_{mn} \sum_{j=n}^{m} q\pi_{n+1} \cdots \pi_j q r \geqq \pi_n \sum_{j=n}^{m} q\pi_{n+1} \cdots \pi_j q r$. Let $\tau^m = \{f_{m1}, f_{m2}, \cdots, f_{mm}, f; f, \cdots\}$ where f is an arbitrary measurable function from SS to A. Then

$$\begin{aligned} I_m(\pi) &= \textstyle\sum_{j=1}^{m} \pi_1 q \cdots \pi_j q r \\ &\leqq f_{m1} \textstyle\sum_{j=1}^{m} q\pi_2 \cdots \pi_j q r \\ &\leqq \cdots \\ &\leqq (\textstyle\sum_{j=1}^{n} f_{m1} q \cdots f_{mj} q r) + (f_{m1} q \cdots f_{mn} \textstyle\sum_{j=n+1}^{m} q\pi_{n+1} \cdots \pi_j q r) \\ &\leqq \cdots \\ &\leqq \textstyle\sum_{j=1}^{m} f_{m1} q \cdots f_{mj} q r \\ &= I_m(\tau^m) \leqq I(\tau^m). \end{aligned}$$

So that $I_m(\pi) \leqq I(\tau^m)$ for all m, hence $I(\pi) \leqq \liminf_m I(\tau^m)$. Let

$$S_m = \{I(\tau^j) < I(\pi) - \epsilon \quad \text{for} \quad j < m, I(\tau^m) \geqq I(\pi) - \epsilon\}$$

and let $\tau = \tau^m$ on S_m. Then τ is semi-Markov, and for $s \in S_m$, $I(\tau)(s) = I(\tau^m)(s) \geqq I(\pi) - \epsilon$, hence $I(\tau) \geqq I(\pi) - \epsilon$. In the same manner, we can assume π is random Markov, and construct a sequence σ^m of Markov policies such that $I_m(\pi) \leqq I(\sigma^m)$. We then let $\sigma = \sigma^{m_0}$, where m_0 is any m such that $pI(\sigma^m 0) \geqq pI(\pi) - \epsilon$.

We do not know if the ϵ can be eliminated in Theorem 4.4.

5. The operators T and U. With any measurable f from S to A we associate the operator T from $M(S)$ to $M(S)$ defined by

$$Tu(s) = \int r(s, f(s), t) + \beta u(t)\, dq(t \mid s, f(s)).$$

We may interpret $Tu(s)$ as the expected return if we are in state s, take action $f(s)$, and receive a terminal return of $u(t)$ at the resulting state t. The following properties of T, which we state as a theorem, are immediate from the definition and from well known properties of the integral.

THEOREM 5.1. *D, P, N unless otherwise indicated.*

(a) *T is monotone, $u \leqq v$ implies $Tu \leqq Tv$.*

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

(c) *$T(\sup_j u_j) \geqq \sup_j Tu_j$.*

(d) *If $u_j \downarrow u$ then $Tu_j \downarrow Tu$ and if $u_j \uparrow u$ then $Tu_j \uparrow Tu$ on $\{\sup_j Tu_j > -\infty\}$.*

(e) *D. T is a contraction, $\|Tu - Tv\| \leqq \beta\|u - v\|$;*

P. T is positive, $u \geqq 0$ implies $Tu \geqq 0$;

N. T is negative, $u \leqq 0$ implies $Tu \leqq 0$.

(f) *For any π, $TI(\pi) = I(f, \pi)$, where $(f, \pi) = (f, \pi_1, \pi_2, \cdots)$ is the policy which uses f followed by π. In particular, $TI(f^{(\infty)}) = I(f^{(\infty)}) = \lim_{n\to\infty} T^n 0$.*

(g) *If $\pi = \{f_1, f_2, \cdots\}$ is a Markov policy, and T_n is the operator associated with f_n, then $I_n(\pi, v) = T_1 \cdots T_n v$.*

For any Markov $\pi = (f_1, f_2, \cdots)$, we say that the function f from S into A is *π-generated* if there exists a partition of S into Borel sets $S_1, S_2, \cdots$ such that $f = f_n$ on S_n. We say that the Markov policy $\pi' = \{g_1, g_2, \cdots\}$ is *π-generated* if each g_n is π-generated. We associate with π the operator U from $M(S)$ into $M(S)$ defined by $Uu = \sup_n T_n u$, where T_n is the operator associated with f_n.

We may interpret $Uu(s)$ as the optimal expected return over π-generated f if we are in state s, take action $f(s)$ and receive a terminal return $u(t)$ at the resulting state t. We list as a theorem some properties of U.

THEOREM 5.2. *D, P, N. unless otherwise indicated.*

(a) *U is monotone, $u \leqq v$ implies $Uu \leqq Uv$.*

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

(c) *$U(\sup_j u_j) \geqq \sup_j Uu_j$.*

(d) *If $u_j \downarrow u$ then $\inf_j Uu_j \geqq Uu$ and if $u_j \uparrow u$ then $\sup_j Uu_j \leqq Uu$ with equality on $\{\sup_j T_n u_j > -\infty$ for all $n\}$.*

(e) *D. U is a contraction, $\|Uu - Uv\| \leqq \beta\|u - v\|$;*

P. U is positive, $u \geqq 0$ implies $Uu \geqq 0$;

N. U is negative, $u \leqq 0$ implies $Uu \leqq 0$.

(f) *For any π-generated f with associated T, and any $u \in M(S)$, $Tu \leqq Uu$.*

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

PROOF. (a), (b), and (e) follow from the corresponding properties of the T_n's; (c) follows from (d), by considering the functions $u_j' = \max u_k$, $1 \leqq k \leqq j$; (f) is clear from the definition, since $Tu(s) = T_n u(s) \leqq Uu(s)$ for $s \in S_n$. For (d), if $u_j \downarrow u$, then using (d) of Theorem 5.1,

$$\inf_j Uu_j = \inf_j \sup_n T_n u_j \geqq \sup_n \inf_j T_n u_j = Uu,$$

while if $u_j \uparrow u$, we have

$$\sup_j Uu_j = \sup_j \sup_n T_n u_j = \sup_n \sup_j T_n u_j .$$

But for each n, $\sup_j T_n u_j \leqq T_n u$, with equality on $\{\sup_j T_n u_j > -\infty\}$, hence $\sup_j U u_j \leqq \sup_n T_n u = Uu$ with equality on $\{\sup_j T_n u_j > -\infty$ for all $n\}$. For (g), we let $S_n = \{s \mid T_i u < Uu - \epsilon$ for $i < n$, $T_n u \geqq Uu - \epsilon\}$, and set $f = f_n$ on S_n. Then for $s \varepsilon S_n$, $Tu(s) = T_n u(s) \geqq Uu(s) - \epsilon$. □

6. Markov policies. If $\hat{\pi}$ is any Markov policy, i.e. any countable collection of measurable functions from S to A, let $G(\hat{\pi}) = \{\pi \mid \pi$ is $\hat{\pi}$-generated$\}$, and let U be the operator associated with $\hat{\pi}$ by Theorem 5.2. Associated with $G(\hat{\pi})$ are three functions which are of interest:

(1) $\lim_{n\to\infty} U^n 0$,

(2) $\sup_{\pi \varepsilon G(\hat{\pi})} I(\pi)$,

(3) $\sup_{f^{(\infty)} \varepsilon G(\hat{\pi})} I(f^{(\infty)})$.

$U^n 0$ is the best we can do using $\hat{\pi}$-generated policies if we terminate at the nth stage with no terminal return; hence (1) is the limit of the optimal return from finite stage play among $\hat{\pi}$-generated policies. (2) represents the optimal return from infinite stage play among $\hat{\pi}$-generated policies, while (3) is the optimal return among $\hat{\pi}$-generated stationary policies.

LEMMA 6.1. (D) (P) N.

$$\lim_{n\to\infty} U^n 0 \geqq \sup_{\pi \varepsilon G(\hat{\pi})} I(\pi) \geqq \sup_{f^{(\infty)} \varepsilon G(\hat{\pi})} I(f^{(\infty)}).$$

PROOF. Lim $U^n 0$ exists by Theorem 5.2 (e), and for any n, and any $\pi \varepsilon G(\hat{\pi})$, $U^n 0 \geqq I_n(\pi)$, hence $\lim U^n 0 \geqq I(\pi)$ and the first inequality follows. The second inequality is obvious. □

Blackwell has shown in [2] and [3] that in the discounted and positive bounded cases equality holds throughout. In fact, in the discounted case even more is true. For any $\epsilon > 0$, there exists a $\hat{\pi}$-generated f such that $I(f^{(\infty)}) + \epsilon \geqq \sup\{I(\pi) \mid \pi \varepsilon G(\hat{\pi})\}$. It is not known if this is true in the positive bounded case. In the negative case however, all three may be different, as the following example shows:

EXAMPLE 6.1. $S = \{0, 1, 2, \cdots\}$, $A = \{3, 4, \cdots\}$, $r(3, a, 2) = r(2, a, 0) = -1$, $r = 0$ otherwise. The transition function is given by $q(\cdot \mid 0, a) = \delta(0)$, $q(\cdot \mid 1, a) = \delta(a)$, $q(\cdot \mid 2, a) = a^{-1}\delta(0) + (1 - a^{-1})\delta(2)$, and $q(\cdot \mid s, a) = \delta(s - 1)$ for $s > 2$. From state 1 we can move to any state $s > 2$, and from s we move to $s - 1$ until we reach state 2. If we are in state 2 and take action a we move to state 0 with probability a^{-1}, and remain in state 2 with the remaining probability. State 0 is a terminal state, and once we reach it we remain forever. The only movements with non-zero return are transitions from state 3 into state 2 and from state 2 into state 0. Let $\hat{\pi}_n = f_n \equiv n + 2$, so that all Markov π are $\hat{\pi}$-generated.

If we are going to terminate at the end of n stages with no terminal return, then in state 1 we can pick an a so large that we shall not reach state 2 before termination, and in state 2 we can pick on a so large that we shall remain there until termination with probability arbitrarily close to one. Thus

$$\begin{aligned} \lim_{n\to\infty} U^n 0(s) &= 0 \qquad \text{if} \quad s = 0, 1, 2, \\ &= -1 \qquad \text{otherwise.} \end{aligned}$$

In infinite play, starting in state 1 we shall eventually move into state 2 regardless of what we do, but from state 2 we can pick a sequence of actions a_n such that $\prod_{n=1}^{\infty}(1 - a_n^{-1})$, the probability we remain in state 2 forever, is arbitrarily close to one. Thus

$$\begin{aligned} \sup_{\pi \varepsilon G(\hat{\pi})} I(\pi)(s) &= 0 \qquad \text{if} \quad s = 0 \text{ or } 2 \\ &= -1 \qquad \text{otherwise.} \end{aligned}$$

Using any stationary policy however, we shall eventually move from state 2 to state 0 with probability one. Thus

$$\begin{aligned} \sup_{f^{(\infty)} \varepsilon G(\hat{\pi})} I(f^{(\infty)})(s) &= 0 \qquad \text{if} \quad s = 0 \\ &= -1 \qquad \text{if} \quad s = 2 \\ &= -2 \qquad \text{otherwise.} \end{aligned}$$

In the negative case, the functions (2) and (3) need not even be measurable, as the following example shows. (This fact was discovered by David Blackwell.)

EXAMPLE 6.2. Let $S = \{(x, r) \mid 0 \leqq x, r \leqq 1, r \text{ rational}\} \cup \{t\}$ and let A be the positive integers. Let $r_1, r_2, \cdots$ be an enumeration of the rationals between 0 and 1, and let $W_1, W_2, \cdots$ be a sequence of Borel subsets of the unit interval. The transition law is given by

$$\begin{aligned} q(\cdot \mid (x, r), a) &= \delta(x, r_a) \qquad \text{if } r_a < r \text{ and } x \varepsilon W_a \\ &= \delta(t) \qquad \text{otherwise,} \\ q(\cdot \mid t, a) &= \delta(t) \end{aligned}$$

and the return function is given by $r(s, a, t) = -1$ if $s \neq t$ and $r = 0$ otherwise. Let $\hat{\pi}_n \equiv n$ for all n, so that all Markov policies are $\hat{\pi}$-generated, and let

$$D = \{(x, 1) \mid \sup_{\pi \varepsilon G(\hat{\pi})} I(\pi)(x, 1) = 0\}.$$

Then D is a set sifted through the sieve $\{W_a\}$, and the sets $\{W_a\}$ may be chosen so that D is not Borel ([11], chapter 7). For any $(x, 1) \varepsilon D$, we may choose f such that $I(f^{\infty})(x, 1) = 0$, hence neither of the functions (2) or (3) is Borel measurable.

Much of this section is devoted to obtaining bounds for (2) in the negative case. The results are known in the discounted and positive bounded cases, and much of the section is closely related to the work of Dubins and Savage on gambling [6].

Using the terminology of [6], we say that U *conserves* $v \varepsilon M(s)$ if $Uv \geqq v$. From Theorem 5.2 (c) and (d) it is clear that if U conserves v_j, $j = 1, 2, \cdots$, then U conserves $\sup_j v_j$, and if U conserves v, then U conserves $\lim_{n\to\infty} U^n v$. In fact, if v is bounded below, then $\lim U^n v$ is a fixed point of U.

THEOREM 6.1. *(D), N. If U conserves v, and $\epsilon > 0$ then there exists a $\pi \varepsilon G(\hat{\pi})$ such that $I(\pi) \geqq v - \epsilon$.*

PROOF. Pick ϵ_n, $n = 1, 2, \cdots$, such that $\sum_{n=1}^{\infty} \epsilon_n \leqq \epsilon$. Then, from Theorem

5.2 (g), we can find a $\hat{\pi}$-generated f_n such that the associated operator T_n satisfies $T_n v \geqq v - \epsilon_n$. Let $\pi = \{f_1, f_2, \cdots\}$. Then

$$I_n(\pi, v) = T_1 \cdots T_n v \geqq v - \textstyle\sum_{i=1}^{n} \epsilon_i \geqq v - \epsilon,$$

so that $\liminf I_n(\pi, v) \geqq v - \epsilon$. In the discounted case, $I_n(\pi, v) \to I(\pi) \geqq v - \epsilon$, while in the negative case,

$$I(\pi) = \lim_{n\to\infty} I_n(\pi) \geqq \liminf_{n\to\infty} I_n(\pi, v) \geqq v - \epsilon. \qquad \square$$

The theorem is clearly false in the positive bounded case, since if U conserves v and $\beta = 1$ then U conserves $v + c$ for any constant c. Hence for any v, and c sufficiently large, the theorem will fail.

Let $u^* = \lim_{n\to\infty} U^n 0$. In the negative case, we see from Theorem 5.2 (d) that $Uu^* \leqq u^*$. If however, u^* is a fixed point of U, we can combine Lemma 6.1 and Theorem 6.1 to obtain

COROLLARY 6.1. *(D), (P), N. If $u^* = Uu^*$, then $u^* = \sup\{I(\pi) \mid \pi \in G(\hat{\pi})\}$.*

We thus wish to find function v such that the resulting policy π will be an improvement on $\hat{\pi}$. For any $\pi \in G(\hat{\pi})$, we define

$$v_\pi' = \sup_{n\geqq 0} I({}^n\pi);$$

$$v_\pi = \lim_{n\to\infty} U^n v_\pi'.$$

(Recall that if $\pi = (f_1, f_2, \cdots)$, then ${}^n\pi = (f_{n+1}, f_{n+2}, \cdots)$.)

LEMMA 6.2. *(D), N. U conserves v_π' and v_π.*

PROOF.

$$\begin{aligned} Uv_\pi' &= U(\sup_{n\geqq 0} I({}^n\pi)) \\ &\geqq \sup_{n\geqq 0} UI({}^n\pi) \\ &\geqq \sup_{n\geqq 1} UI({}^n\pi) \\ &\geqq \sup_{n\geqq 1} T_n I({}^n\pi) \\ &= \sup_{n\geqq 1} I({}^{n-1}\pi) \\ &= v_\pi'. \end{aligned}$$

Hence U conserves v_π', and by Theorem 5.2 (a) and (d), U conserves $U^n v_\pi'$ for all n, and U conserves v_π. $\square$

In the discounted case, U is a contraction operator, and $\sup\{I(\pi) \mid \pi \in G(\hat{\pi})\}$ is its unique fixed point ([3]), so that $v_\pi = \sup\{I(\pi) \mid \pi \in G(\hat{\pi})\}$ for all $\pi \in G(\hat{\pi})$. In the negative case this is not true. In either case however, we can prove the following:

THEOREM 6.2. *(D), N. Let $\pi^j, j = 1, 2, \cdots$, be any sequence of Markov policies. Then for any $\epsilon > 0$ there exists a Markov π^* such that $I(\pi^*) > \sup_j I(\pi_j) - \epsilon$.*

PROOF. Find $\hat{\pi}$ such that each $\pi^j \in G(\hat{\pi})$. We can do this, for example, by letting $\hat{\pi}$ be a reordering of the f_n^j's. Let $v_j = v_{\pi^j}$, and let $v = \sup_j v_j$. Then $v \geqq \sup_j I(\pi^j)$ and U conserves v, so we can find π^* with $I(\pi^*) \geqq v - \epsilon \geqq \sup_j I(\pi^j) - \epsilon$. $\square$

We remark that in the negative case if U conserves v and $\sup\{v(s) \mid s \in S\} = c < 0$, then U also conserves $v - c \in M(S)$, so that in the preceding constructions we can always choose v so that the resulting policy π satisfies $\sup\{I(\pi)(s) \mid s \in S\} = 0$.

Example 6.1 shows that in the negative case, every Markov policy is not dominated by a stationary policy. We say a Markov policy $\pi^* = (f_1, f_2, \cdots)$ is *N-stationary* if $f_n = f$ for some f and $n \leqq N$.

THEOREM 6.3. (D), N. *If $\pi \in G(\hat{\pi})$, then for any N and any $\epsilon > 0$ there exists a $\pi^* \in G(\hat{\pi})$ such that π^* is N-stationary and $I(\pi^*) \geqq I(\pi) - \epsilon$.*

PROOF. Find $\pi' = \{f_1, f_2, \cdots\} \in G(\hat{\pi})$ such that $I(\pi') \geqq v_\pi - \epsilon/2$, and find a $\hat{\pi}$-generated f such that the associated T satisfies $Tv_\pi \geqq v_\pi - \epsilon/2N$. Let $\pi^* = \{g_1, g_2, \cdots\}$ where $g_n = f$ for $n \leqq N$ and $g_n = f_{n-N}$ for $n > N$. Then

$$I(\pi^*) = T^N I({}^N\pi^*) = T^N I(\pi') \geqq T^N v_\pi - \epsilon/2 \geqq v_\pi - \epsilon. \qquad \square$$

THEOREM 6.4. (D), N. *If there exists $\pi^* \in G(\hat{\pi})$ such that $I(\pi^*) = \sup\{I(\pi) \mid \pi \in G(\hat{\pi})\}$, then $I(f^{(\infty)}) = I(\pi^*)$, where f is the first element of π^*.*

PROOF. Let T be the operator associated with f, then

$$I(\pi^*) = TI({}^1\pi^*) \leqq TI(\pi^*) = I(f, \pi^*) \leqq I(\pi^*)$$

so that $TI(\pi^*) = I(\pi^*)$, and $T^n I(\pi^*) = I(\pi^*)$. In the discounted case $T^n I(\pi^*) \to I(f^{(\infty)})$, while in the negative case, $T^n I(\pi^*) \leqq T^n 0 \downarrow I(f^{(\infty)})$. $\square$

We say that u is *excessive* for U if $Uu \leqq u$. Blackwell has shown that in the discounted case if $u \in M(S)$ and u is excessive for U then $u \geqq I(\pi)$ for all $\pi \in G(\hat{\pi})$ ([3]), while in the positive bounded case, if $u \geqq 0$ and u is excessive for U, then $u \geqq I(\pi)$ for all $\pi \in G(\hat{\pi})$ ([2]). We have been unable to obtain similar simple conditions in the negative case. The following theorem is a straightforward generalization of the theorem given in [1].

THEOREM 6.5. (D), (P), N. *Suppose u is excessive for U. Then $u \geqq I(\pi)$ for all $\pi \in G(\hat{\pi})$ if and only if $\sup_{n \geqq 0} I_n(\pi, u) \geqq I(\pi)$ for all $\pi \in G(\hat{\pi})$.*

PROOF. $I_0(\pi, u) = u$, hence the necessity of the condition is trivial. Conversely,

$$I_n(\pi, u) = I_{n-1}(\pi, T_n u) \leqq I_{n-1}(\pi, Uu) \leqq I_{n-1}(\pi, u)$$

where T_n is the operator associated with the nth element of π. Therefore $I_n(\pi, u) \leqq u$ for all n and the result follows. $\square$

7. Absolute measurability of the optimal return. Throughout this and the next section we let $v^* = \sup_\pi I(\pi)$. Blackwell has shown in [3] that if for every $\epsilon > 0$ there exists an ϵ-optional policy, then v^* is measurable and satisfies the optimality equation $v^* = \sup_{a \in A} T_a v^*$ where T_a is the operator associated with $f \equiv a$. In general however, v^* is not measurable and there does not exist an ϵ-optimal policy for every $\epsilon > 0$. In Example 4.1, there does not exist an ϵ-optimal policy for any $\epsilon < 1$, because if $\beta = 1$ say, $v^*(s) = 1$ if $s \in B \cup D$, $v^*(s) = 0$ otherwise. Since D is not measurable, neither is v^*.

In this section we will show that v^* is absolutely measurable, i.e. measurable with respect to the completion of every $p \in P(S)$, and in the next section we will

show that there always exists a (p, ϵ)-optimal policy and that v^* satisfies the optimality equation.

Let $X = ASAS \cdots$ be the set of futures of the system, and let Σ^* be the smallest σ-field in $P(X)$ such that νB is a Σ^*-measurable function of $\nu \in P(X)$ for every Borel subset B of X. Then $P(X)$ is a Borel set, and Σ^* the σ-field of its Borel subsets ([5]).

LEMMA 7.1. *If u is any Baire function on SX such that u is bounded or u is of constant sign then νu is measurable on $SP(X)$.*

PROOF. If u is bounded, the lemma is immediate from 2.2 of [5]. If u is of constant sign then u is the monotone limit of bounded Baire functions. □

Let $\Gamma = \{(s, \nu) \mid \nu = e_\pi(s)$ for some policy $\pi\}$, i.e. Γ is the set of pairs $(s, \nu) \in SP(X)$ such that ν is the probability on X induced by some policy π at s.

LEMMA 7.2. *Γ is a Borel subset of $SP(X)$.*

PROOF. Every $\nu \in P(X)$ has a factorization as $\nu = \nu_1\nu_2 \cdots$ where

$$\nu_1 \in P(A),$$

$$\nu_2 \in Q(S \mid A),$$

$$\nu_{2n+1} \in Q(A \mid AS \cdots AS) \ (2n \text{ factors}),$$

$$\nu_{2n} \in Q(S \mid AS \cdots ASA) \ (2n - 1 \text{ factors}).$$

We shall index the coordinates of $(s, x) \in SX$ by $s = s_1, x = (a_1, s_2, a_2, \cdots)$. Then $\nu = e_\pi(s_1)$ for some π if and only if $\nu_2(\cdot \mid a_1) = q(\cdot \mid s_1, a_1)$ for almost all a_1 with respect to ν and $\nu_{2n}(\cdot \mid a_1, s_2, \cdots a_n) = q(\cdot \mid s_n, a_n)$ for almost all $(a_1, s_2, \cdots, a_n)$ with respect to ν and all $n \geqq 2$.

Let w_{nm}, $m \geqq 1$ be a countable subset of $M(X_n)$ which separates points of $P(X_n)$, where $X_n = AS \cdots AS$ ($2n$ factors) for $n \geqq 1$. Let

$$\Gamma_{1m} = \{(s_1, \nu) \mid \int w_{1m}(a_1, s_2)\, d\nu(a_1, s_2) = \iint w_{1m}(a_1, s_2)\, dq(s_2 \mid s_1, a_1)\, d\nu(a_1)\}$$

and for $n \geqq 2$ let $\Gamma_{nm} = \{(s_1, \nu) \mid \nu w_{nm} = \nu q w_{nm}\}$. Then so $\Gamma = \bigcap_{n=1}^{\infty} \bigcap_{m=1}^{\infty} \Gamma_{nm}$ is Borel. □

THEOREM 7.1. *D, P, N. v^* is absolutely measurable.*

PROOF. By Lemma 7.1, $v(s, \nu) = \nu(\sum_{n=1}^{\infty} r(s_n, a_n, s_{n+1})\beta^{n-1})$ is a measurable function of (s, ν). But $v^*(s) = \sup_{(s,\nu)\in\Gamma} v(s, \nu)$. Let $B_\lambda = \Gamma \cap \{(s, \nu) \mid v(s, \nu) > \lambda\}$. Then B_λ is Borel, and $C_\lambda = \{s \mid v^*(s) > \lambda\}$, which is the projection of B_λ, is analytic, hence absolutely measurable. □

8. Optimality.

THEOREM 8.1. *For any $p \in P(S)$, $\epsilon > 0$,*

D. there exists a (p, ϵ)-optimal stationary policy f^∞,

P. there exists a (p, ϵ)-optimal semi-Markov policy $\hat{\pi}$,

N. there exists a (p, ϵ)-optimal Markov policy π^*.

The statement of Theorem 8.1 in the discounted case is the same as that of Theorem 6(b) of [3]. Our definition of (p, ϵ)-optimality is, however, stronger than that given in [3]. Recall that π is (p, ϵ)-optimal if $p\{I(\pi) \geqq v^* - \epsilon\} = 1$.

PROOF. We shall first show the existence of a (p, ϵ)-optimal π. Since v^* is absolutely measurable, we can find a Borel $N \subset S$ such that $pN = 0$ and a measurable function v_0 such that for $s \notin N$, $v_0(s) = v^*(s)$. Let

$$\Gamma_\epsilon = \Gamma \cap [\{(s, \nu) \mid s \notin N, v(s, \nu) > v_0(s) - \epsilon\} \cup \{(s, \nu) \mid s \in N\}].$$

For each s, Γ_ϵ has a non-empty s-section, so by Theorem 6.3 of [10], we can find a Borel $N' \subset S$ with $pN' = 0$, and a measurable function γ from S into $P(X)$ such that for $s \notin N'$, $\gamma(s) \in \Gamma_\epsilon$. For any Borel subset B of X, the map $(s, \nu) \to \nu(B)$ is measurable in (s, ν). Hence the map $s \to \gamma(s)(B)$ is measurable in s. Clearly for each $s \in S$, $\gamma(s)$ is a probability on X, so that if $\mu(B \mid s) = \gamma(s)(B)$, then $\mu \in Q(X \mid S)$. We can then factor μ as $\mu = \mu_1\mu_2 \cdots$, where

$$\mu_1 \in Q(A \mid S),$$

$$\mu_{2n} \in Q(S \mid SA \cdots SA) \ (2n \text{ factors}),$$

$$\mu_{2n+1} \in Q(A \mid SA \cdots S) \ (2n + 1 \text{ factors}).$$

For $s_1 \notin N'$, $\mu_{2n}(\cdot \mid s_1) = q(\cdot \mid s_n, a_n)$ since $\gamma(s_1) \in \Gamma$. We can now define π by

$$\pi_n = \mu_{2n-1} \quad \text{if} \quad s \notin N',$$

$$\pi_n = \pi_n' \quad \text{if} \quad s \in N',$$

where $\pi' = (\pi_1', \pi_2', \cdots)$ is an arbitrary policy. For $s \notin N \cup N'$, $I(\pi)(s) = v(s, \gamma(s)) \geqq v^*(s) - \epsilon$, and $p(N \cup N') = 0$, hence π is (p, ϵ)-optimal.

Blackwell's definition ([3]) of (p, ϵ)-optimality is that σ is (p, ϵ)-optimal if for all π, $p\{I(\sigma) \geqq I(\pi) - \epsilon\} = 1$, so that the theorem in the discounted case now follows from Theorem 6(b) of [3]. To complete the proof in the positive case, we find π such that π is $(p, \epsilon/2)$-optimal, then from Theorem 4.4, we can find a semi-Markov policy $\hat{\pi}$ such that $I(\hat{\pi}) \geqq I(\pi) - \epsilon/2$.

To complete the proof in the negative case, let us assume that $pv^* > -\infty$. We may do this without loss of generality since if $pv^* = -\infty$, we can replace p with $p' \in P(S)$ such that $p'\{s \mid v^*(s) > -\infty\} = 1$, p' is equivalent to (has the same null sets as) p restricted to $\{s \mid v^*(s) > -\infty\}$, and $p'v^* > -\infty$. Then any policy is (p, ϵ)-optimal if and only if it is (p', ϵ)-optimal. From Theorem 4.3 and the result just proved, for each $m \geqq 1$ we can find a Markov policy π^m such that $pI(\pi^m) \geqq pv^* - 1/m$. Let $v_1 = \sup_m I(\pi^m)$. Then $v_1 \leqq v^*$ and $pv_1 = pv^*$, hence $p\{v_1 = v^*\} = 1$. From Theorem 6.2 it follows that there exists a Markov policy π^* such that $I(\pi^*) \geqq v_1 - \epsilon$. □

We next show that v^* satisfies the optimality equation. Let T_a be the operator defined in Section 5 for $f \equiv a$.

THEOREM 8.2. *D, P, N.*

$$v^*(s) = \sup_a T_a v^*(s) \qquad \textit{for all} \qquad s \in S.$$

In fact, in the discounted case v^* is the unique bounded solution and in the positive bounded case v^* is the smallest non-negative solution.

PROOF. For any $s \in S$, $\epsilon > 0$, we can find a Markov policy π such that $I(\pi)(s) \geqq v^*(s) - \epsilon$, by letting $p = \delta(s)$ and applying Theorem 8.1. Then, if a is the action which π_1 chooses at s,

$$v^*(s) - \epsilon \leqq I(\pi)(s) = T_a I(^1\pi)(s) \leqq T_a v^*(s).$$

We can do this for each s and each $\epsilon > 0$; hence $v^* \leqq \sup_a T_a v^*$. For any s, any a, and $\epsilon > 0$, we can find a π (not necessarily Markov in the positive case) such that $I(\pi) \geqq v^* - \epsilon$ almost surely with respect to $q(\cdot \mid s, a)$. Then

$$v^*(s) \geqq I(a, \pi)(s) = T_a I(\pi)(s) \geqq T_a v^*(s) - \epsilon$$

where (a, π) is the policy which uses a, then uses π from the resulting state. Thus $v^* \geqq \sup_a T_a v^*$, since s, a, and ϵ were arbitrary, and v^* satisfies the optimality equation, $v^* = \sup_a T_a v^*$.

In the discounted case, let w be any other bounded solution. We do not assume that w is absolutely measurable, but only that $T_a w$ is defined for each $a \in A$. If $T_a w$ is defined, so is $T_a(w + c) = T_a w + \beta c$. In particular, $v^* \leqq w + \|v^* - w\|$, so that

$$T_a v^* \leqq T_a w + \beta \|v^* - w\| \qquad \text{for all} \qquad a,$$

$$v^* = \sup_a T_a v^* \leqq \sup_a T_a w + \beta \|v^* - w\| = w + \beta \|v^* - w\|.$$

Reversing the roles of v^* and w we obtain $\|v^* - w\| \leqq \beta \|v^* - w\|$, which implies $v^* = w$.

In the positive case, let $w \geqq 0$ be any other solution, and suppose $w(s) < v^*(s)$ for some s. It follows from Theorem 8.1 that we can find a Markov policy $\pi = (f_1, f_2, \cdots)$ with $w(s) < I(\pi)(s)$. If T_n is the operator associated with f_n, then

$$w(s) < I(\pi)(s) = \lim_{n\to\infty} T_1 \cdots T_n 0(s) \leqq \lim_{n\to\infty} T_1 \cdots T_n w(s) \leqq w(s)$$

which is a contradiction. □

We close this section with the following analogs of Theorems 6.4 and 6.5:

THEOREM 8.3. *(D), N. If there exists an optimal policy π^* then there exists an optimal stationary policy $f^{(\infty)}$.*

PROOF. According to Theorem 4.3, we may assume π^* is semi-Markov. Let f be the first element of π^*, and T be the associated operator. Then

$$\begin{aligned} v^*(s) &= I(\pi^*)(s) = \int r(s, f(s), t) + I(\pi_s{}^*)(t)\, dq(t \mid s, a) \\ &\leqq \int r(s, f(s), t) + v^*(t)\, dq(t \mid s, a) \\ &= Tv^*(s) = I(f, \pi^*) \leqq v^*(s) \end{aligned}$$

where $\pi_s{}^*$ is the policy which π^* determines starting at the second stage when the initial state is s. Thus $Tv^* = v^*$. In the negative case $v^* = T^n v^* \leqq T^n 0 \downarrow I(f^{(\infty)}) \leqq v^*$ while in the discounted case $v^* = T^n v^* \to I(f^{(\infty)})$ so that $I(f^{(\infty)}) = v^*$. □

THEOREM 8.4. *(D) (P) N. Suppose* $T_a u \leqq u$ *for all* a. *Then* $I(\pi) < u$ *for all* π *if and only if* $\sup_{n \geqq 0} I_n(\hat{\pi}, u) \geqq I(\hat{\pi})$ *for all Markov policies* $\hat{\pi}$.

PROOF. From Theorem 8.1 it follows that $\sup_\pi I(\pi) = \sup \{I(\hat{\pi}) \mid \hat{\pi} \text{ Markov}\}$. The result then follows from Theorem 6.5. □

As in the case of Theorem 6.5, this is much weaker than Blackwell's results ([2] and [3]) in the discounted and positive cases, but we have been unable to obtain anything stronger in the negative case.

9. Additional Results and Comments. We say that actions a and b are *equivalent at state* s if $r(s, a, \cdot) = r(s, b, \cdot)$ and $q(\cdot \mid s, a) = q(\cdot \mid s, b)$. We say that A is *essentially finite* by $\pi^* = \{f_1, f_2, \cdots\}$ 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 actions $f_1(s), \cdots, f_n(s)$ is equivalent to a at s.

THEOREM 9.1. *(D), N. If* A *is essentially finite by* $\pi^* = (f_1, f_2, \cdots)$ *and* U *is the operator associated with* π^*, *then*

(a) $U^n 0 \to v^* = \sup_\pi I(\pi)$;

(b) *there exists a stationary optimal policy* $f^{(\infty)}$.

PROOF. *(N)* Let $v_n = U^n 0$, and let $v^* = \lim v_n$, which exists since $v_n \geqq v_{n+1}$. For any π^*-generated π, $I_n(\pi) \leqq U^n 0 = v_n$; hence $I(\pi) \leqq v^*$. By the assumption of essential finiteness, all Markov policies are π^*-generated, and we have shown that

$$\sup_\pi I(\pi) = \sup_{\pi \text{ Markov}} I(\pi),$$

hence $I(\pi) \leqq v^*$ for all π.

For any $s \in S$, $v_n(s) = T_{m(n)} v_{n-1}(s)$, for some $m(n)$ such that $T_{m(n)} v_{n-1}(s) = U v_{n-1}(s)$. Such an $m(n)$ exists by the assumption of essential finiteness. Moreover, we can always choose $m(n) \leqq k$ for $s \in S_k$, and then the sequence $m(1)$, $m(2)$, $\cdots$ contains only finitely many distinct elements, so contains one, say m, infinitely often. Since v_n decreases in n, it follows from Theorem 5.1 (a) and (d) that

$$\begin{aligned} v^*(s) &= \lim_{n \to \infty} v_n(s) \\ &= \lim_{n \to \infty} T_{m(n)} v_{n-1}(s) \\ &= \lim_{n \to \infty} T_m v_{n-1}(s) \\ &= T_m v^*(s) \leqq U v^*(s). \end{aligned}$$

But s was arbitrary, hence $U v^* \geqq v^*$. We can find a measurable partition $B_1, B_2, \cdots$ of s such that $U v^* = T_m v^*$ on B_m. Let $f = f_m$ on B_m, and let T be the operator associated with f. Then $T v^* = U v^* \geqq v^*$, and $I_n(f^{(\infty)}) = T^n 0 \geqq T^n v^* \geqq v^*$. It follows that $I(f^{(\infty)}) = v^*$ and $f^{(\infty)}$ is optimal. □

The Howard improvement routine, proved for finite S, A in [8], and in the general discounted case in [3], is valid in the negative case.

THEOREM 9.2. *(D), N. For any* π, *if* $I(f, \pi) \geqq I(\pi)$ *then* $I(f^{(\infty)}) \geqq I(\pi)$.

PROOF. (N) Let T be the operator associated with f, then $I_n(f^{(\infty)}) \geqq T^n I(\pi) \geqq I(\pi)$. □

Example 4.2 shows that Theorem 9.2 may fail in the positive case.

Our final result is an improvement routine valid in the discounted and negative cases. Given any policies σ and τ, the policy π, obtained from them by choosing at each stage the one which would be better if we were going to use it for the infinite future, is better than either σ or τ.

THEOREM 9.3. *D, N. Let σ and τ be policies, and define π by*

$$\begin{aligned} \pi_n &= \sigma_n \quad \text{if} \quad (s_1, a_1, \cdots, a_{n-1}, s_n) \in B_n \\ &= \tau_n \quad \text{if} \quad (s_1, a_1, \cdots, a_{n-1}, s_n) \in B_n^c, \end{aligned}$$

where $B_n = \{(s_1, a_1, \cdots, a_{n-1}, s_n) \mid u_n > v_n\}$, *and*

$$u_n(s_1, a_1, \cdots, a_{n-1}, s_n) = \sum_{j=n}^{\infty} \beta^{j-1} \sigma_n q \cdots \sigma_j q r;$$

$$v_n(s_1, a_1, \cdots, a_{n-1}, s_n) = \sum_{j=n}^{\infty} \beta^{j-1} \tau_n q \cdots \tau_j q r.$$

Then $I(\pi) \geqq \max(I(\sigma), I(\tau))$.

PROOF. We extend the notation $I_n(\pi, w)$ defined in Section 3 in the following way: For $w_{n+1} \in M(SA \cdots AS)(2n + 1$ factors), we define

$$I_n(\pi, w_{n+1}) = e_\pi[\sum_{j=1}^{n} \beta^{j-1} r(s_j, a_j, s_{j+1}) + \beta^n w_{n+1}(s_1, a_1, \cdots s_{n+1})].$$

With this definition, $I_n(\pi, w_{n+1})$ is our expected return if we use π, terminate after the nth stage, and receive a terminal return w_{n+1} which depends upon the history we have experienced rather than on the terminal state. If we let $w_n = \max(u_n, v_n)$, we see that $I_0(\pi, w_1) = w_1 = \max(I(\sigma), I(\tau))$, and in general, $I_{n-1}(\pi, w_n)$ is our expected return if we follow π for $n - 1$ stages, then switch to the better of σ or τ at the nth stage. On B_n, $\pi_n = \sigma_n$ and $w_n = u_n$, and on B_n^c, $\pi_n = \tau_n$ and $w_n = v_n$, hence

$$\begin{aligned} w_n &= \pi_n(r + \beta u_{n+1}) \leqq \pi_n(r + \beta w_{n+1}) \quad \text{on} \quad B_n \\ &= \pi_n(r + \beta v_{n+1}) \leqq \pi_n(r + \beta w_{n+1}) \quad \text{on} \quad B_n^c \end{aligned}$$

so that

$$I_{n-1}(\pi, w_n) \leqq I_{n-1}(\pi, [\pi_n(r + \beta w_{n+1})]) = I_n(\pi, w_{n+1}),$$

and $I_{n-1}(\pi, w_n) \geqq \max(I(\sigma), I(r))$ for all n. In the discounted case, $I_{n-1}(\pi, w_n) \to I(\pi)$, and in the negative case, $I_{n-1}(\pi, w_n) \leqq I_{n-1}(\pi, 0) \downarrow I(\pi)$. □

Stated in terms of Markov or stationary policies, the theorem becomes

COROLLARY 9.1. *(D), N. Let $\sigma = (f_1, f_2, \cdots)$ and $\tau = (g_1, g_2, \cdots)$ be Markov policies, and define $\pi = (h_1, h_2, \cdots)$ by*

$$\begin{aligned} h_n &= f_n \quad \textit{if } I({}^{n-1}\sigma) > I({}^{n-1}\tau) \\ &= g_n \quad \textit{otherwise.} \end{aligned}$$

Then $I(\pi) \geqq \max(I(\sigma), I(\tau))$.

COROLLARY 9.2. (D), N. (Eaton-Zadeh). *Let* $f^{(\infty)}$ *and* $g^{(\infty)}$ *be stationary policies, and define* h *by*

$$h = f \quad \text{if } I(f^{(\infty)}) \geqq I(g^{(\infty)})$$

$$= g \quad \text{otherwise.}$$

Then $I(h^{(\infty)}) \geqq \max(I(f^{(\infty)}), I(g^{(\infty)}))$.

Corollary 9.2 was proved in the negative case for finite S and A in [7] and in the discounted case in [3]. The following example shows that Corollary 9.2, hence Theorem 9.3, may fail in the positive case.

EXAMPLE 9.1. Let S, A, q, and r be as in Example 4.2. Define f by $f(s) = 0$ for s even, and $f(s) = 1$ for s odd. Let $g = 1 - f$. Then

$$I(f^{(\infty)})(1) = I(g^{(\infty)})(1) = 0;$$

$$I(f^{(\infty)})(s) = 1 - 1/s \quad \text{and} \quad I(g^{(\infty)}) = 1 - 1/(s+1) \quad \text{for} \quad s \text{ even};$$

$$I(f^{(\infty)})(s) = 1 - 1/(s+1) \quad \text{and} \quad I(g^{(\infty)}) = 1 - 1/s \quad \text{for} \quad s > 1 \quad \text{odd}.$$

Hence $h \equiv 1$ and $I(h^{(\infty)}) \equiv 0$. This example shows that no routine can be devised to produce a stationary policy better than two given policies in the positive case, since any stationary policy would have to have $h(s) = 0$ for some s in order to have a non-zero return, and would be less than $\max(I(f^{(\infty)}), I(g^{(\infty)}))$ at that s.

We do not know if there are any theorems similar to Theorems 6.2 or 9.2 which hold in the positive case, and as a result do not know if *semi-Markov* can be replaced by *Markov* or *stationary* in Theorem 8.1. We do not know if Theorems 6.3 or 8.3 are valid in the positive case. Theorems 4.2 and 6.4 are false as stated in the positive case, and we do not know if any similar theorems are true.

In conclusion we wish to express our gratitude to Professor David Blackwell whose guidance and advice throughout the preparation of the thesis were invaluable.

REFERENCES

[1] BLACKWELL, DAVID (1961). On the functional equation of dynamic programming. *J. Math Anal. Appl.* **2** 273–276.

[2] BLACKWELL, DAVID (1964). Positive bounded dynamic programming. Univ. of California, Berkeley. (Mimeographed.)

[3] BLACKWELL, DAVID (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.

[4] BLACKWELL, D. and RYLL-NARDZEWSKI, C. (1963). Non-existence of everywhere proper conditional distributions. *Ann. Math. Statist.* **34** 223–225.

[5] DUBINS, LESTER and FREEDMAN, DAVID. (1965). Measurable sets of measures. *Pacific J. Math.* **14** 1211–1222.

[6] DUBINS, LESTER E. and SAVAGE, LEONARD J. (1965). *How to Gamble If You Must.* McGraw-Hill, New York.

[7] EATON, J. H. and ZADEH, L. A. (1962). Optimal pursuit strategies is discrete state probabilistic systems. *J. Basic Eng. Ser. D* **84** 23–29.

[8] HOWARD, RONALD A. (1960). *Dynamic Programming and Markov Processes*. Wiley, New York.
[9] LOÈVE, MICHEL (1963). *Probability Theory* (3rd ed.) Van Nostrand, Princeton.
[10] MACKEY, GEORGE W. (1957). Borel structure in groups and their duals. *Trans Amer. Math. Soc.* **85** 134–165.
[11] SIERPINSKI, W. (1952). *General Topology*. Univ. of Toronto Press.