The Annals of Statistics
1974 Vol. 2, No. 1, 219-222

ON STATIONARY POLICIES—THE GENERAL CASE[1]

BY MICHAEL ORKIN

University of California, Berkeley and
Case Western Reserve University

A recent result of Blackwell states that in a positive dynamic programming problem with countable state space, if there is an optimal policy, then there is a stationary optimal policy. We extend this result by allowing the state space to be Borel and by proving that if there is an optimal policy, then for any probability measure μ on the state space there is a stationary policy which is optimal on a set of μ measure 1.

1. Introduction. In [3] Blackwell considered positive dynamic programming problems with countable state space and showed that if there exists an optimal policy, then there exists a stationary optimal policy. In this paper we extend Blackwell's result to the general case by proving:

THEOREM 1. *In a positive dynamic programming problem with Borel state space, S, if there is an optimal policy, then for every probability measure μ on S, there exists a stationary policy which is optimal on a set of μ measure* 1.

We note that our results also hold if the state space and constraint set are allowed to be analytic. Also our formulation of the problem, while consistent with [3], is slightly less general than [2] or [1]. However, our results hold in this context as well.

2. The dynamic programming model. A positive dynamic programming problem is specified by three objects S, A, r, where S, considered the set of states of some system, is a Borel subset of a complete separable metric space, A is a Borel subset of $S \times \Pi(S)$, where $\Pi(S)$ is the set of all probability measures on S, endowed with the weak * topology and the corresponding Borel σ-field (for details about $\Pi(S)$ see [5]), and where r is a nonnegative bounded Borel measurable function (called a reward function) defined on $A \times S$. Also, we assume for each $x \in S$ the corresponding x-section $A_x = \{P \in \Pi(S) \mid (x, P) \in A\}$ is nonempty.

When you are at $x \in S$ you select in a measurable way (as a function of the past states and selections) any $P \in A_x$, and move to a new state $y \in S$ selected at random according to P. You then receive the reward $r(x, P, y)$ and proceed as before. If you select P and you are in state x, your expected reward for this move is $\int_S r(x, P, y)P(dy)$.

A policy, σ, is a sequence $\sigma_1, \sigma_2, \cdots$, where σ_n tells you how to act on the nth

Received September 1972.
[1] This research was supported by NSF Grant GP-33908X.
AMS 1970 *subject classifications*. Primary 49C15, 90A05; Secondary 49A99, 60G45.
Key words and phrases. Positive dynamic programming, optimal policy stationary policy, measurable policy, martingale.

move as a Borel measurable function of the previous history $h = (x_1, P_1, \cdots, x_{n-1}, P_{n-1}, x_n)$ of the system. Starting in a fixed initial state, your object is to find a policy which maximizes your total expected reward over the infinite future. For a discussion allowing a more general class of reward functions and policies and a more general state space, see [4].

A policy σ in which σ_n is a function only of the current state is called stationary; equivalently, any Borel function from S to $\Pi(S)$ whose graph is in A defines a stationary policy: when x is the current state, choose $P = f(x)$. If f defines a stationary policy, we denote this policy by f^∞. Let $W(x, \sigma)$ be your expected income starting at x and using σ, and let $U(x) = \sup W(x, \sigma)$, the sup being taken over all policies. We assume that for all x, $U(x) < \infty$. A policy σ is called optimal at x if $W(x, \sigma) = U(x)$ and optimal if it is optimal at every x.

3. Main results. Assume that for the problem (S, A, r) there is an optimal policy $\tilde{\sigma}$.

LEMMA 1. *For each $x \in S$, denote by $\tilde{A}_x$ the set of P in A_x for which the following equation is true*:

$$U(x) = \int_S r(x, P, y) + U(y)P(dy) . \tag{1}$$

Then

(a) *The set $\tilde{A} = \{(x, P) \mid P \in \tilde{A}_x\}$ is Borel.*
(b) *For each x, $\tilde{A}_x$ is nonempty.*
(c) *In the reduced problem $(S, \tilde{A}, r)$, $U(x) =$ the original U for each $x \in S$.*
(d) *In the reduced problem there is an optimal policy.*

PROOF. Since there is an optimal policy $\tilde{\sigma}$, $U(x) = W(\tilde{\sigma}, x)$, and therefore $U(x)$ is Borel measurable (in general, the function U need not be Borel, although it is universally measurable, cf. [4], [6]). Also, the function $h(x, P) = \int r(x, P, y) + U(y)P(dy)$ is Borel measurable (see Lemma on page 266 in [6]). Thus, $\tilde{A} = \{(x, P) \mid U(x) = h(x, P)\}$ is the intersection of the graphs of two Borel measurable functions and hence is a Borel set, proving (a). For (b), claim $\tilde{\sigma}_1(x)$ satisfies (1) for each x. To see this, suppose $P = \tilde{\sigma}_1(x)$ and note that $U(x) = W(x, \tilde{\sigma}) = \int r(x, P, y) + E(\sum_{i=2}^{\infty} r_i \mid y)P(dy)$, where $r_i =$ reward from the ith move. But $E(\sum_{i=2}^{\infty} r_i \mid y) = U(y)$ almost surely; otherwise, we could improve $\tilde{\sigma}$ by playing optimally in y after the first move (on a set of positive measure), which would contradict the fact that $\tilde{\sigma}$ is optimal.

For (c) (d), claim the following policy yields U: use $\tilde{\sigma}$ whenever possible, play arbitrarily if you reach a position where $\tilde{\sigma}$ is not available (something is always available, e.g., $\tilde{\sigma}_1$). Using this policy, with probability 1 $\tilde{\sigma}$ will always be available. If not, there is a first n, say n_0 such that with positive probability $\tilde{\sigma}_{n_0}$ will not be available on the n_0th move. But this cannot happen; an argument similar to the one for (b) shows you could then improve $\tilde{\sigma}$ after the n_0th move, contradicting the fact that $\tilde{\sigma}$ is optimal. □

Reduce the problem so that (1) is always satisfied. Lemma 1 says we can do

this and that the new U = the original U at every $x \in S$, and there is an optimal policy (which we shall still call $\bar{\sigma}$).

LEMMA 2. *Let $x_0 \in S$ and assume $U(x_0) > 0$. Let $\varepsilon > 0$ with $U(x_0) - \varepsilon > \varepsilon$. Then there exists a Borel set $B \subset S$ with $x_0 \in B$ and $\sup_{x \in B} U(x) < \infty$, and a stationary policy f^∞ such that the policy $\bar{\sigma}$: use f^∞ while in B and $\bar{\sigma}$ when not in B, satisfies:*

$$W_B(x_0, \bar{\sigma}) > U(x_0) - \varepsilon , \tag{2}$$

where if $x \in B$, $W_B(x, \sigma)$ is your expected income up to the first exit from B starting in x and using σ, and

$$W(x_0, \bar{\sigma}) = U(x_0) . \tag{3}$$

PROOF. Theorem 1 of [2] shows there exists a stationary policy f^∞ for which $W(x_0, f^\infty) > U(x_0) - \varepsilon/2$. Since r is nonnegative and $U(x) < \infty$ for each $x \in S$, we can find a Borel set $\hat{B}$ containing x_0 such that $W_{\hat{B}}(x_0, f^\infty) > U(x_0) - \varepsilon/2$ and $\sup_{x \in \hat{B}} U(x) < \infty$. Let $B = \{x \in \hat{B} \mid W_{\hat{B}}(x, f^\infty) > \varepsilon/2\}$. Then $\sup_{x \in B} U(x) < \infty$ and $x_0 \in B$ (remember $U(x_0) - \varepsilon > \varepsilon$). Also, $W_B(x_0, f^\infty) > U(x_0) - \varepsilon$, since $U(x_0) - \varepsilon/2 < W_{\hat{B}}(x_0, f^\infty) \leqq W_B(x_0, f^\infty) + \varepsilon/2$ Prob ($\hat{B} - B$ is hit before $\hat{B}$ is left for the first time) $\leqq W_B(x_0, f^\infty) + \varepsilon/2$.

For (3), let I_n be your income from the first n moves and let $(X_1 \equiv x_0, X_2, \cdots)$ be the history of states of the system, using $\bar{\sigma}$. Since (1) always holds, the sequence $V_n = I_{n-1} + U(X_n)$ is a martingale, so that (proceeding as in [3]) for all n, $E(V_n) = U(x_0)$, i.e.,

$$E(I_{n-1}) + EU(X_n) = U(x_0) . \tag{4}$$

Letting $n \to \infty$ gives

$$W(x_0, \bar{\sigma}) + a = U(x_0)$$

where $a = \lim_{n \to \infty} EU(X_n)$. We must show $a = 0$. For this note that

$$E(I^* - I_{n-1} \mid X_1, \cdots, X_n) \geqq E(F \mid X_1, \cdots, X_n) \quad \text{a.s.} , \tag{5}$$

where $I^* = \lim_{n \to \infty} I_n$ and F is our future income after $X_1, \cdots, X_n$ up to and including the first re-entry into B ($F = 0$ if $X_n \in B$). Since we are playing optimally up to the first re-entry into B

$$U(X_n) \leqq E(F \mid X_1, \cdots, X_n) + MP_n \quad \text{a.s.} , \tag{6}$$

where $M = \sup_{x \in B} U(x)$ (remember $M < \infty$) and $P_n = \text{Prob}\,(X_i \in B \text{ for some } i \geqq n \mid X_1, \cdots, X_n)$. Combining (5) and (6) and taking expectations yields

$$EU(X_n) \leqq E(I^* - I_{n-1}) + ME(P_n) . \tag{7}$$

Also,

$$E(I^* - I_{n-1} \mid X_1, \cdots, X_n) \geqq (\varepsilon/2)P_n \quad \text{a.s.} , \tag{8}$$

since $x \in B \Rightarrow W_{\hat{B}}(x, f^\infty) > \varepsilon/2$ and since we are playing optimally when not in B. Taking expectations and letting $n \to \infty$ in (8) yields $E(P_n) \to 0$ and letting $n \to \infty$ in (7) then yields $EU(X_n) \to 0$, i.e., $a = 0$.

LEMMA 3. *For each state x there exists a stationary policy which is optimal at x.*

PROOF. If $U(x) = 0$ then any policy is optimal at x. Suppose then x_0 is such that $U(x_0) > 0$. Make the Lemma 1 reduction so (1) is always satisfied. Pick a sequence $\varepsilon_n \downarrow 0$ with $U(x_0) - \varepsilon_1 > \varepsilon_1$. Let B_1 be a Borel set satisfying Lemma 2 for $\varepsilon = \varepsilon_1$, with f_1^∞ the corresponding stationary policy. For each $x \in B_1$, reduce A_x to the single action $f_1(x)$. For this reduced problem, (3) shows that the new $U(x_0)$ is the same as the original $U(x_0)$. Now, in the reduced problem find a set $B_2 \supset B_1$ and a function f_2 (which must coincide with f_1 on B_1) satisfying Lemma 2 for $\varepsilon = \varepsilon_2$. Reduce the problem again by allowing $f_2(x)$ to be the only action available at $x \in B_2$ and use (3) again to see that $U(x_0)$ remains the same, etc. We get a sequence of Borel sets $B_1, B_2, \cdots$, and a Borel measurable function f defined on $B = \bigcup_{n=1}^\infty B_n$. Extend f arbitrarily so that it is defined and measurable on S. Now (2) gives $W_{B_n}(x_0, f^\infty) > U(x_0) - \varepsilon_n$ for all n. Thus, $W_B(x_0, f^\infty) = U(x_0)$, so $W(x_0, f^\infty) = U(x_0)$, completing the proof.

THEOREM 1. *Let μ be a probability measure on S. Then there exists a stationary policy which is optimal on a set of μ measure* 1.

PROOF. Introduce a new state x^* with only one action available at x^*; namely μ. Define $r(x^*, \mu, y) \equiv 0$. Since $\tilde{\sigma}$ was optimal for (S, A, r), its corresponding extension must be optimal for $(S \cup \{x^*\}, A \cup \{(x^*, \mu)\}, r)$. Lemma 3 says there is an optimal stationary strategy σ^* at x^*. But σ^* must be optimal on a set of μ measure 1; otherwise, we could modify σ^* after the first move (which must be μ) and do better, contradicting the fact that σ^* is optimal at x^*.

Acknowledgment. We wish to thank Professor David Blackwell for several helpful discussions about this problem.

REFERENCES

[1] BLACKWELL, D. (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.
[2] BLACKWELL, D. (1965). Positive dynamic programming. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 415–418.
[3] BLACKWELL, D. (1971). On stationary policies. *J. Roy. Statist. Soc.* **133** 33–37.
[4] BLACKWELL, D., FREEDMAN, D. and ORKIN, M. The optimal reward operator in dynamic programming. To appear.
[5] DUBINS, L. and FREEDMAN, D. (1965). Measurable sets of measures. *Pacific J. Math.* **14** 1211–1222.
[6] SUDDERTH, W. (1971). On measurable gambling problems. *Ann. Math. Statist.* **42** 260–269.

DEPARTMENT OF MATHEMATICS
CASE WESTERN RESERVE UNIVERSITY
CLEVELAND, OHIO 44106