ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XVI, 3 (1979)

A. NOWAK (Katowice)

DISCOUNTED DYNAMIC PROGRAMMING ON EUCLIDEAN SPACES

The aim of this paper is to give sufficient conditions for the existence of stationary optimal and ε-optimal policies in a discounted dynamic programming problem. We assume that the set of actions A is a subset of an n-dimensional Euclidean space R^n. Our model is the same as that one studied by Furukawa [5] and Hinderer (stationary case in [6]). It is a generalization of the model of Blackwell [3]. A more general model, with the discount factor depending on states and actions, was investigated by Schäl [7], [9]. The proof of existence of an optimal policy is based on a selection theorem, which is established in Section 2 of this paper. Similar results were obtained by Freedman [4], Furukawa [5], and Schäl [7]-[9].

1. Notation and definitions. A *standard Borel space* (abbreviated to an *SB-space*) is a Borel subset of a Polish space, endowed with the induced topology and the Borel σ-field. Throughout this section, X and Y are non-empty SB-spaces. By XY we denote the Cartesian product of X and Y. We always consider XY with the product topology and with the product σ-field.

By $P(X)$ we mean the set of all probability measures on X, and by $Q(Y|X)$ — the set of all transition probabilities from X to Y. For $p \in P(X)$ and $q \in Q(Y|X)$, pq denotes the product probability measure on XY. This notation extends to a finite or infinite sequence of SB-spaces $X_1, X_2, \ldots$ If $p \in P(X_1)$ and $q_n \in Q(X_{n+1}|X_1X_2 \ldots X_n)$ for $n \geqslant 1$, then

$$pq_1 \ldots q_{n-1} \in P(X_1X_2 \ldots X_n), \quad q_1q_2 \ldots \in Q(X_2X_3 \ldots | X_1), \quad \ldots$$

By δ_x we denote a measure from $P(X)$ such that $\delta_x(\{x\}) = 1$.

By N, R, and R^n we mean the set of all positive integers, the real line, and the n-dimensional Euclidean space, respectively. The set of all bounded, real-valued, Borel measurable functions on X is denoted by $M(X)$. The set $M(X)$ is a Banach space with the norm

$$\|u\| := \sup_{x \in X} |u(x)|, \quad u \in M(X).$$

For $u, v \in M(X)$, $u \leqslant v$ means $u(x) \leqslant v(x)$ for all $x \in X$.

A *multifunction* φ from X to Y is a function defined on X, the values of which are non-empty subsets of Y. A multifunction φ is called *closed* (*compact*) *valued* if, for each $x \in X$, $\varphi(x)$ is closed (compact). A function $f\colon X \to Y$ is a *measurable selection* of φ if it is measurable and $f(x) \in \varphi(x)$ for all $x \in X$.

2. Selection theorem. In this section we prove a selection theorem which is useful for many optimization problems.

Throughout this section, X is a non-empty SB-space, φ is a multifunction from X to R^n such that its graph

$$G := \{(x, y) \in XR^n : y \in \varphi(x)\}$$

is a Borel subset of XR^n, and u is a real-valued, bounded from above, Borel measurable function on G. We are interested in the measurability of the function

$$v(x) := \sup_{y \in \varphi(x)} u(x, y), \qquad x \in X, \tag{2.1}$$

and in the existence of a measurable selection f of φ such that

$$u(x, f(x)) \geqslant v(x) - \varepsilon, \qquad x \in X, \tag{2.2}$$

for given $\varepsilon \geqslant 0$.

THEOREM 2.1. *If φ is closed valued and, for each $x \in X$, $u(x, \cdot)$ is upper semi-continuous on $\varphi(x)$, then v is Borel measurable and for any $\varepsilon > 0$ there exists a measurable selection f of φ satisfying* (2.2). *The second assertion is also true for $\varepsilon = 0$ if we assume that φ is compact valued.*

Proof. The proof of the measurability of v is based on the Novikov theorem (see [1], Theorem 1.5):

If B is a Borel subset of XR^n such that all its x-sections

$$B_x := \{y \in R^n : (x, y) \in B\}$$

are closed, then the projection of B on X,

$$\mathrm{proj}_X(B) := \{x \in X : (x, y) \in B \text{ for some } y \in R^n\},$$

is a Borel subset of X.

It is sufficient to show that

$$Z_c := \{x \in X : v(x) \geqslant c\}$$

is a Borel subset of X for all $c \in R$. By (2.1),

$$Z_c = \bigcap_{m \in N} \left\{x \in X : u(x, y) \geqslant c - \frac{1}{m} \text{ for some } y \in \varphi(x)\right\} = \bigcap_{m \in N} \mathrm{proj}_X(B_m),$$

where

$$B_m := \left\{(x, y) \in G : u(x, y) \geqslant c - \frac{1}{m}\right\}, \qquad m \in N.$$

B_m is a Borel subset of XR^n and, by the upper semi-continuity of $u(x, \cdot)$, all x-sections of B_m are closed. In virtue of Novikov's theorem, $\mathrm{proj}_X(B_m)$ is a Borel subset of X for all $m \in N$. Thus Z_c is Borel.

The second part of the proof is based on the following selection theorem (see [1], Theorem 1.6):

If ψ is a closed-valued multifunction from X to R^n with the graph being a Borel subset of XR^n, then there exists a measurable selection of ψ.

For $\varepsilon > 0$ we define a new multifunction from X to R^n as follows:

$$\psi_\varepsilon(x) := \{y \in \varphi(x) : u(x, y) \geqslant v(x) - \varepsilon\}.$$

In order to complete the proof it suffices to find a measurable selection of ψ_ε. Since $u(x, \cdot)$ is upper semi-continuous, ψ_ε is closed valued. If φ is compact valued, then $u(x, \cdot)$ attains its supremum on $\varphi(x)$, and the multifunction ψ_0 is well defined. By the measurability of G, u and v, the graph of ψ_ε,

$$\{(x, y) \in XR^n : y \in \psi_\varepsilon(x)\} = \{(x, y) \in G : u(x, y) \geqslant v(x) - \varepsilon\},$$

is a Borel subset of XR^n. Thus there exists a measurable selection f of ψ_ε.

The problem of finding a measurable selection of φ satisfying (2.2) with $\varepsilon > 0$ is treated by Schäl ([8], Theorem 1) and Freedman [4]. They consider a multifunction φ whose values are subsets (not necessarily closed) of a metric space Y. Schäl assumes that φ is separable, i.e. Y contains a denumerable dense subset Y' such that $Y' \cap \varphi(x)$ is dense in $\varphi(x)$ for $x \in X$, and u is a Carathéodory map. Freedman gives a version of Theorem 2.1 with X and Y compact, graph of φ and $\{(x, y) \in XY : u(x, y) > c\}$ being F_σ in XY for $c \in R$.

The existence of a measurable selection f of φ such that $u(x, f(x)) = v(x)$, $x \in X$, is studied, e.g., by Furukawa [5], and Schäl [8], [9]. The second part of Theorem 2.1 is a generalization of a result of Furukawa ([5], Theorem 4.1). A similar result is obtained by Schäl ([9], Theorem 12.1) under assumptions that φ is measurable, and u is the limit of a decreasing sequence of Carathéodory maps.

3. Dynamic programming model. A *discounted dynamic programming model* is given by a 6-tuple $(S, A, \varphi, q, r, \beta)$ of the following meaning:

(i) S is a non-empty SB-space, the set of states of a system.

(ii) A is a Borel subset of R^n, the set of actions.

(iii) φ is a multifunction from S to A, $\varphi(s)$ is the set of actions feasible to us at the state s. We assume that the graph of φ,

$$G := \{(s, a) \in SA : a \in \varphi(s)\},$$

is a Borel subset of SA, and there exists a measurable selection of φ.

(iv) q is a transition probability from G to S, the law of motion of the system.

(v) r is a measurable real-valued function on G, the reward function. We assume that r is bounded from above.

(vi) $0 < \beta < 1$ — the discount factor.

When the system is at the state s and we take the action $a \in \varphi(s)$, we receive a reward $r(s, a)$, and the system moves to a new state s', according to the probability distribution $q(\cdot \mid s, a)$. The process is then repeated from the state s'. Future rewards are discounted with the constant factor β. We intend to maximize the expectation of the total discounted reward over the infinite future.

We define sets of histories recursively: $H_1 := S$, $H_{n+1} := GH_n$, $n \in N$. A policy π is a sequence $\{\pi_1, \pi_2, \ldots\}$, where $\pi_n \in Q(A \mid H_n)$ and $\pi_n(\varphi(s_n) \mid h) = 1$ for $h = (s_1, a_1, s_2, \ldots, s_n)$ from H_n, $n \in N$. If we use a policy π, then we choose the n-th action according to the probability distribution $\pi_n(\cdot \mid h)$, where h is a history of the system up to time n. Any measurable selection f of φ defines a policy $\{\pi_n\}$:

$$\pi_n(\cdot \mid s_1, a_1, \ldots, s_n) := \delta_{f(s_n)}, \qquad (s_1, a_1, \ldots, s_n) \in H_n, \quad n \in N.$$

Such a policy is called *stationary* and is denoted by $f^{(\infty)}$. If we use a policy $f^{(\infty)}$ and the system is at the state s, then we take the action $f(s)$ independently both of the time and the history.

Any policy $\pi = \{\pi_n\}$ determines the transition probability

$$e_\pi := \pi_1 q \pi_2 q \ldots \in Q(ASAS \ldots \mid S)$$

(first we must extend q to a transition probability from SA to S, and each π_n — to a transition probability from $SASA \ldots S$ ($2n-1$ factors) to A). An expected reward corresponding to a policy π is given by

$$v_\pi(s_1) := \int_{ASAS\ldots} \Big(\sum_{n \in N} \beta^{n-1} r(s_n, a_n)\Big) e_\pi\big(d(a_1, s_2, a_2, \ldots) \mid s_1\big), \qquad s_1 \in S$$

(put $r(s, a) := 0$ for $(s, a) \in SA \setminus G$).

The optimal reward function v is defined by

$$v(s) := \sup_\pi v_\pi(s), \qquad s \in S.$$

The function v is universally measurable and satisfies the optimality equation

$$v(s) = \sup_{a \in \varphi(s)} \Big(r(s, a) + \beta \int_S v(s') q(ds' \mid s, a)\Big), \qquad s \in S \tag{3.1}$$

(see [10], Theorems 7.1 and 8.2).

A policy π^* is *optimal* if $v_{\pi^*} \geqslant v_\pi$ for all policies π. For $\varepsilon > 0$, a policy π^* is called *ε-optimal* if $v_{\pi^*} \geqslant v_\pi - \varepsilon$ for all π. Our problem is to find an optimal (or ε-optimal) stationary policy.

4. Criteria of optimality and ε-optimality. With a dynamic programming problem we associate some operators defined on the set of all measurable, bounded from above functions $u : S \to R$. We put

$$Lu(s, a) := r(s, a) + \beta \int_S u(s')\, q(ds' \mid s, a), \qquad (s, a) \in G,$$

$$L_f u(s) := Lu\big(s, f(s)\big),$$

$$\tilde{L}_f u(s) := \beta \int_S u(s')\, q\big(ds' \mid s, f(s)\big), \qquad s \in S,$$

where f is a measurable selection of φ. Denote by L_f^n the n-th iteration of L_f.

We state as a lemma some properties of L_f and $\tilde{L}_f$. They follow immediately from well-known properties of the integral.

LEMMA 4.1 (cf. [10], Theorem 5.1). *Let u_1 and u_2 be measurable, real-valued and bounded from above functions on S. Then*

(i) $L_f(u_1 + u_2) = L_f u_1 + \tilde{L}_f u_2$;

(ii) *for $c \in R$, $\tilde{L}_f c = \beta c$;*

(iii) *L_f and $\tilde{L}_f$ are monotone: $u_1 \leqslant u_2$ implies $L_f u_1 \leqslant L_f u_2$, and $\tilde{L}_f u_1 \leqslant \tilde{L}_f u_2$;*

(iv) $\lim_m L_f^m 0 = v_{f^{(\infty)}}$;

(v) $L_f v_{f^{(\infty)}} = v_{f^{(\infty)}}$.

We have the following criterion of optimality of a stationary policy:

THEOREM 4.1 (cf. [9], Theorem 5.3). *Assume that the optimal reward function v is measurable. Then a policy $f^{(\infty)}$ is optimal if and only if*

$$Lv\big(s, f(s)\big) = \sup_{a \in \varphi(s)} Lv(s, a), \qquad s \in S. \tag{4.1}$$

Proof. In virtue of the optimality equation, condition (4.1) is equivalent to

$$L_f v = v. \tag{4.2}$$

If $f^{(\infty)}$ is an optimal policy, then $v_{f^{(\infty)}} = v$ and, by Lemma 4.1 (v), condition (4.2) is satisfied.

Now assume that a measurable selection f of φ satisfies (4.2). By Lemma 4.1 we obtain inductively

$$v = L_f^m v = L_f^m 0 + \tilde{L}_f^m v \leqslant L_f^m 0 + \beta^m K, \qquad m \in N,$$

where K is the upper bound of v. If we pass to the limit, then $v \leqslant v_{f^{(\infty)}}$. Since $v \geqslant v_\pi$ for all policies π, $f^{(\infty)}$ is optimal.

Let $\varepsilon > 0$. The following theorem gives a criterion of ε-optimality of a stationary policy:

THEOREM 4.2. *Assume that v is measurable. If a measurable selection f of φ satisfies*

$$Lv(s, f(s)) \geqslant \sup_{a \in \varphi(s)} Lv(s, a) - \varepsilon(1-\beta), \quad s \in S, \tag{4.3}$$

then $f^{(\infty)}$ is ε-optimal.

Proof. By the optimality equation, we can rewrite condition (4.3) in the equivalent form

$$L_f v \geqslant v - \varepsilon(1-\beta). \tag{4.4}$$

In order to prove ε-optimality of $f^{(\infty)}$, we have to show that $v_{f^{(\infty)}} \geqslant v - \varepsilon$. Proceeding inductively, by (4.4) and Lemma 4.1 we obtain

$$L_f^m v \geqslant v - \varepsilon(1-\beta)(1+\beta+\ldots+\beta^m) > v - \varepsilon, \quad m \in N.$$

On the other hand,

$$L_f^m v = L_f^m 0 + \tilde{L}_f^m v \leqslant L_f^m 0 + \beta^m K, \quad m \in N$$

(see the proof of Theorem 4.1). Consequently,

$$L_f^m 0 + \beta^m K \geqslant v - \varepsilon, \quad m \in N.$$

Passing to the limit we obtain $v_{f^{(\infty)}} \geqslant v - \varepsilon$, which completes the proof.

5. Existence of optimal and ε-optimal policies. In this section we give sufficient conditions for the existence of stationary optimal and ε-optimal policies. We assume:

A1. For every $u \in M(S)$, $s \in S$, the function

$$w(s, \cdot) := \int_S u(s') q(ds' | s, \cdot) \tag{5.1}$$

is continuous on $\varphi(s)$.

A2. The reward function r is bounded and, for each $s \in S$, $r(s, \cdot)$ is upper semi-continuous on $\varphi(s)$.

THEOREM 5.1 (cf. [7], Theorem 8.2). *If a discounted dynamic programming problem satisfies assumptions* A1 *and* A2, *the set of actions A is a closed subset of R^n, and φ is closed valued, then for any $\varepsilon > 0$ there exists a stationary ε-optimal policy.*

Proof. First we prove that the optimal reward v is a measurable function. Let T be the operator defined on $M(S)$ by

$$Tu(s) := \sup_{a \in \varphi(s)} Lu(s, a), \qquad s \in S.$$

With the help of T we can rewrite the optimality equation (3.1) in the form $Tv = v$. Under assumptions A1 and A2, Lu is a bounded measurable function and, for each $s \in S$, $Lu(s, \cdot)$ is upper semi-continuous on $\varphi(s)$. In virtue of Theorem 2.1, $Tu \in M(S)$. For any $u_1, u_2 \in M(S)$,

$$|Tu_1(s) - Tu_2(s)| \leqslant \sup_{a \in \varphi(s)} |Lu_1(s, a) - Lu_2(s, a)| \leqslant \beta \|u_1 - u_2\|, \qquad s \in S.$$

Thus T is a contraction. By the Banach fixed-point theorem, there exists $u_0 \in M(S)$ such that $Tu_0 = u_0$. Since for bounded r the function v is a unique bounded solution of the optimality equation, we have $u_0 = v$ (see [10], Theorem 8.2).

Now φ and Lv satisfy the assumptions of Theorem 2.1. Hence there exists a measurable selection f of φ such that (4.3) is satisfied. By Theorem 4.2, $f^{(\infty)}$ is a stationary ε-optimal policy.

THEOREM 5.2 (cf. [5], Theorem 4.2, and [9], Theorem 15.2). *If a dynamic programming problem satisfies assumptions* A1 *and* A2, *and φ is compact valued, then there exists a stationary optimal policy.*

Proof. We have already proved that the optimal reward v is a measurable function (see the proof of Theorem 5.1). Note that φ and Lv satisfy the assumptions of Theorem 2.1. Thus there exists a measurable selection f of φ which satisfies (4.1). In virtue of Theorem 4.1, $f^{(\infty)}$ is a stationary optimal policy.

Schäl ([7], Theorem 8.2) gave a version of Theorem 5.1 with A an SB-space, φ separable, and r a Carathéodory map. Furukawa ([5], Theorem 4.2) has obtained Theorem 5.2 under the stronger assumption that A is a compact subset of R^n, and $r(s, \cdot)$ is continuous. Schäl ([9], Theorem 15.2) has proved a similar result, assuming that A is an SB-space, φ is separable (or φ is measurable and q satisfies a stronger continuity condition than A1), and r is the limit of a decreasing sequence of Carathéodory maps.

Our results can easily be generalized to a non-stationary Markovian decision model with finite or infinite horizon. Under similar assumptions there exists a Markovian optimal (ε-optimal) policy.

6. Appendix. We give sufficient conditions for the transition probability q to satisfy assumption A1. They are based on the following Scheffé theorem:

Let $(X, \mathcal{X}, \mu)$ be a measurable space with a σ-finite measure μ, and

let p and p_m be measurable real-valued non-negative functions on X such that

$$\int_X p(x)\mu(dx) = \int_X p_m(x)\mu(dx) = 1, \quad m \in N.$$

If $\lim_m p_m(x) = p(x)$ μ-a.e., then

$$\lim_m \int_X |p(x) - p_m(x)|\mu(dx) = 0$$

(see [2], p. 223).

Assume that the transition probability q satisfies the following condition:

There exist a σ-finite measure μ on S and a measurable non-negative function $p: GS \to R$ such that

$$q(B|s, a) = \int_B p(s, a, s')\mu(ds')$$

for all Borel subsets $B \subset S$.

If $p(s, \cdot, s')$ is continuous on $\varphi(s)$ for $s' \in S$, then the function $w(s, \cdot)$ defined by (5.1) is also continuous on $\varphi(s)$.

For the proof, let a, $a_m \in \varphi(s)$ for $m \in N$, and

$$\lim_m a_m = a.$$

By the Scheffé theorem,

$$\lim_m |w(s, a) - w(s, a_m)| \leqslant \lim_m \|u\| \int_S |p(s, a, s') - p(s, a_m, s')|\mu(ds') = 0.$$

Hence $w(s, \cdot)$ is continuous on $\varphi(s)$.

References

[1] V. I. Arkin and V. L. Levin (В. И. Аркин и В. Л. Левин), *Выпуклость значений векторных интегралов, теоремы измеримого выбора и вариационные задачи*, Успехи математических наук 27 (1972), p. 21-77.

[2] P. Billingsley, *Convergence of probability measures*, Wiley, New York 1968.

[3] D. Blackwell, *Discounted dynamic programming*, Ann. Math. Statist. 36 (1965), p. 226-235.

[4] D. A. Freedman, *The optimal reward operator in special classes of dynamic programming problems*, Ann. Prob. 2 (1974), p. 942-949.

[5] N. Furukawa, *Markovian decision processes with compact action spaces*, Ann. Math. Statist. 43 (1972), p. 1612-1622.

[6] K. Hinderer, *Foundations of non-stationary dynamic programming with discrete time-parameter*, Lectures Notes in Operations Research and Mathematical Systems 33, Springer, Berlin 1970.

[7] M. Schäl, *On continuous dynamic programming with discrete time-parameter*, Z. Wahrscheinlichkeitstheorie verw. Geb. 21 (1972), p. 279-288.

[8] — *A selection theorem for optimization problem*, Archiv der Mathematik 25 (1974), p. 219-224.

[9] — *Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal*, Z. Wahrscheinlichkeitstheorie verw. Geb. 32 (1975), p. 179-196.

[10] R. E. Strauch, *Negative dynamic programming*, Ann. Math. Statist. 37 (1966), p. 871-890.

INSTITUTE OF MATHEMATICS
SILESIAN UNIVERSITY
40-007 KATOWICE

Received on 22. 11. 1976

A. NOWAK (Katowice)

PROGRAMOWANIE DYNAMICZNE Z DYSKONTEM W PRZESTRZENIACH EUKLIDESOWYCH

STRESZCZENIE

W pracy rozpatrywane jest programowanie dynamiczne z dyskontem, z przestrzenią decyzji będącą podzbiorem n-wymiarowej przestrzeni euklidesowej. Podane są warunki wystarczające dla istnienia stacjonarnych polityk optymalnych i ε-optymalnych.
