J. Appl. Prob. **13**, 507–518 (1976)
Printed in Israel

ON OPTIMAL POLICIES AND MARTINGALES IN DYNAMIC PROGRAMMING

ULRICH RIEDER, *University of Hamburg*

Abstract

A martingale approach to a dynamic program with general state and action spaces is taken. Several necessary and sufficient conditions are given for a policy to be optimal. The results comprehend and modify different criteria of optimality given for dynamic programming problems. Finally, two applications are stated.

DYNAMIC PROGRAMMING; GAMBLING; EXCESSIVE FUNCTION; MARTINGALE; OPTIMAL POLICY; THRIFTY POLICY; EQUALIZING POLICY

1. Introduction and summary

Dubins and Savage (1965) and Sudderth (1972) have given an elegant characterization of optimal strategies for general gambling models. In the present paper we take a martingale approach to dynamic programming problems and derive similar characterizations of optimal policies in dynamic programs with general state and action spaces. The impetus for our martingale approach came from Blackwell (1970) and Hordijk (1974).

The analysis is based only on methods of martingale theory. However, most of the results can also be proved by dynamic programming methods (cf. Rieder (1975)).

In Section 2 we formulate the dynamic programming model. Assumptions are stated that ensure the existence of the conditional expected rewards; in that connection we remark that the reward function is not assumed to be bounded from above or from below. Section 3 deals with excessive functions relative to the dynamic program. Some examples and properties of excessive functions are listed and an optional sampling theorem for decision processes is proved. In Section 4 we give several necessary and sufficient conditions for a general (not necessarily Markov) policy to be (s, ε)-optimal (resp. s-optimal), e.g. it is shown that a policy is s-optimal if and only if it is s-thrifty and s-equalizing. Thrifty policies and equalizing policies are characterized in Section 5 and Section 6, respectively. The results are closely related to results in Dubins and Savage

Received in revised form 14 January 1976.

(1965) and Sudderth (1972). They comprehend and modify different criteria of optimality given by Blackwell ((1965), Theorem 6f), Hinderer ((1970), Theorems 17.1, 17.6 and 17.7), Chow, Robbins and Siegmund ((1971), Theorem 4.10), Hordijk ((1974), Theorems 4.6, 6.3 and 6.5), Ross ((1974), Propositions 1 and 2) and Schäl ((1975), Theorems 5.2 and 5.3). In particular, the results are true both in the positive and negative case. The final Section 7 contains two applications of our new results. One is to generalize a theorem of Blackwell (1970), the other is to give rather weak sufficient conditions guaranteeing the existence of an optimal policy. We remark that our methods can also be used to derive similar characterizations for non-stationary non-Markovian dynamic programs.

2. The dynamic programming model

In the present paper we consider a stationary dynamic program $((S, \mathcal{S}), (A, \mathcal{A}), D, q, r)$ of the following meaning:

(i) The *state space* $(S, \mathcal{S})$ and the *action space* $(A, \mathcal{A})$ are assumed to be standard Borel spaces.

(ii) The set D is a measurable subset of $S \times A$ which contains the graph of a measurable map from S into A. For any $s \in S$ the non-empty and measurable s-section $D(s)$ of D is called the *set of all admissible actions* if the system is in state s.

(iii) q is a transition probability from D to S, the so-called *transition law* during a single stage.

(iv) The *reward function* r is a measurable map from D into the set $\mathbf{R}$ of real numbers.

Remark. In this paper we do not consider any discount factor. If the reward function for the time period $(n, n+1]$ has the form

$$\beta(s_1, a_1, s_2) \cdots \beta(s_{n-1}, a_{n-1}, s_n) r(s_n, a_n)$$

where β is a bounded measurable function from $D \times S$ into the set of the non-negative real numbers and may be interpreted as a (generalized) discount factor (cf. Schäl (1975)), the results of this paper can be easily extended to this more general case.

We write $H_1 = S$, $H_{n+1} = D \times H_n$, $n \in \mathbf{N}$. $\mathbf{N}$ is the set of the positive integers. As usual, a *randomized policy* $\pi = (\pi_n)$ is defined as a sequence of transition probabilities π_n from H_n to A such that $\pi_n(h, D(s_n)) = 1$ for all $h = (s_1, a_1, s_2, \cdots, s_n) \in H_n$, $n \in \mathbf{N}$. Let Δ denote the set of all randomized policies. We write F for the set of all *decision functions*, i.e. of all measurable maps $f: S \to A$ whose graphs belong to D. Then a (deterministic) *Markov policy* is a sequence (f_n) with $f_n \in F$, $n \in \mathbf{N}$. A (deterministic) *stationary policy* is a Markov policy (f_n) where $f_n = f$ is independent of n. For such a policy we write f^∞.

Given some initial state s, any policy π defines a probability measure $P_{\pi s}$ on the product space $H = S \times A \times S \times A \cdots$ endowed with the product σ-algebra $\mathcal{H} = \mathcal{S} \otimes \mathcal{A} \otimes \mathcal{S} \otimes \mathcal{A} \cdots$ and thus a random process $(\zeta_1, \alpha_1, \zeta_2, \alpha_2, \cdots)$ (cf. Hinderer (1970), p. 80) where ζ_n and α_n denote the projection from H onto the nth state space and the nth action space, respectively. Then $\chi_n = (\zeta_1, \alpha_1, \zeta_2, \cdots, \zeta_n)$ describes the history at time n.

In order that the total expected rewards are well-defined we have to impose some convergence assumptions. Let $i \in \{-, +\}$. Define

$$W_i(s) = \sup_{\pi \in \Delta} E_{\pi s}\left[\sum_1^\infty r_n^i\right] \qquad s \in S,$$

where $r_n = r \circ (\zeta_n, \alpha_n)$ and where the expectation is taken with respect to $P_{\pi s}$, $c^+ = \max(0, \pm c)$ for any extended real number c. We shall make use of the following assumption.

Assumption A^i. $W_i(s) < \infty$, $s \in S$.

Throughout the paper we impose a general assumption.

General Assumption. Either A^+ or A^- holds.

Our general assumption implies that the total reward $\Sigma_1^\infty r_n$ is $P_{\pi s}$-quasi-integrable for any policy π and $s \in S$ and therefore the following functions

$$V_\pi(s) = E_{\pi s}\left[\sum_1^\infty r_n\right] \qquad \pi \in \Delta, \quad s \in S$$

are well-defined.

Formulations simplify considerably by the use of the following isotone operators. Denote by M_i, $i \in \{-, +\}$ the set of all measurable functions u from S into the set $\bar{\mathbb{R}}$ of the extended real numbers such that $u^i < \infty$ and $E_{\pi s} u^i \circ \zeta_n < \infty$, $n \in \mathbb{N}$, $s \in S$, $\pi \in \Delta$. If Assumption A^i holds, the operators L_f, $f \in F$ and U are well-defined on M_i by

$$L_f u(s) = r(s, f(s)) + \int q(s, f(s), ds') u(s')$$

$$Uu(s) = \sup_{f \in F} L_f u(s), \quad s \in S.$$

3. Excessive functions and martingales

A function $v \in M_i$ is called *excessive* for the dynamic program if

$$v \geqq V_\pi, \quad \pi \in \Delta, \tag{3.1}$$

$$v \geqq Uv. \tag{3.2}$$

We use $\mathscr{E}_i$ to denote the set of all excessive functions. The definition above is a slight modification of the concept introduced by Hordijk (1974); in particular Hordijk assumes the reward structure to be a charge structure and considers only Markov policies. Strauch (1966) has used another definition of an excessive function.

A measurable map $\tau: H \to \mathbb{N} + \{\infty\}$ is called a *stopping time* relative to the sequence of σ-algebras $\mathscr{F}_n = \sigma(\chi_n)$ if $[\tau = n] \in \mathscr{F}_n$ for all $n \in \mathbb{N}$. We write T for the set of all stopping times. For later use we set $\mathscr{F}_\infty = \sigma(\zeta_1, \alpha_1, \zeta_2, \cdots)$.

Throughout this paper we shall assume that the set $\mathscr{E}_i$ is not empty. This assumption holds in many practical problems. Let us consider three *examples.*

(i) The so-called *value-function* $V(s) = \sup_{\pi \in \Delta} V_\pi(s)$, $s \in S$, need not belong to $\mathscr{E}_i$. But, if V is measurable then $V \in \mathscr{E}_i$ (cf. Hinderer (1971)) and V is the smallest function in $\mathscr{E}_i$, i.e. $V \leqq v$ for all $v \in \mathscr{E}_i$. The function V is measurable, e.g., if there exists a policy π such that $V_\pi = V$. We recall that V is always universally measurable (cf. Hinderer (1971)).

(ii) The *limit function* $V_\infty = \liminf_n V_n$, where $V_n(s) = \sup_{\pi \in \Delta} E_{\pi s}[\sum_{k=1}^{n} r_k]$, $n \in \mathbb{N}$, $s \in S$, satisfies the inequality $V_\infty \geqq V_\pi$ for all π. Thus, if V_∞ is measurable and if

$$\liminf_n L_f V_n \geqq L_f V_\infty \quad \text{for all} \quad f \in F \tag{3.3}$$

holds, then $V_\infty \in \mathscr{E}_i$. (For a proof of (3.2) use the relation $V_{n+1} \geqq L_f V_n$, $f \in F$.) Under Assumption A^- the Relation (3.3) is satisfied by Fatou's Lemma.

(iii) Let g be a measurable map from S into $\mathbb{R}$. Then $V^*: S \to \bar{\mathbb{R}}$

$$V^*(s) = \sup_{\pi \in \Delta} \sup_{\tau \in T} E_{\pi s}\left[\sum_{1}^{\tau - 1} r_n + (g \circ \zeta_\tau 1_{\mathbb{N}} \circ \tau)\right]^+, \quad s \in S$$

belongs to $\mathscr{E}_i$, provided that the expectations on the right exist and that $V \in M_i$.

Now we list some *properties* of excessive functions.

Remark 3.1. If v is the smallest excessive function in $\mathscr{E}_i$ and $Uv \in M_i$, then $v = Uv$.

Remark 3.2. If $v \in \mathscr{E}_i$ and $v = L_f v$ for some $f \in F$, then $v = Uv$.

Remark 3.3. If $v \in \mathscr{E}_i$ then $\liminf_n E_{\pi s} v \circ \zeta_n \geqq 0$ on $\{s: V_\pi(s) \in \mathbb{R}\}$, $\pi \in \Delta$.

Remark 3.4. If A^- holds and if $v \in \mathscr{E}_-$ then $\lim_n E_{\pi s} v^- \circ \zeta_n = 0$, $s \in S$, $\pi \in \Delta$.

Remark 3.5. (cf. Hordijk (1974), Theorem 2.17.) Let $v \in M_i$ such that $v(s) < \infty$, $s \in S$. Then $v \in \mathscr{E}_i$ iff $v \geqq Uv$ and $\lim_n E_{\pi s} v \circ \zeta_n \geqq 0$ on $\{s: V_\pi(s) > -\infty\}$ for all π.

[†] Let 1_B denote the indicator function of the set B.

For the remainder of this paper let $v \in \mathscr{E}_i$ be fixed and suppose that Assumption A' holds. Then the random variables X_n

$$X_n = \sum_{k=1}^{n-1} r_k + v \circ \zeta_n, \quad n \in \mathbb{N}, \; X_\infty = \sum_{k=1}^{\infty} r_k$$

are $P_{\pi s}$-quasi-integrable for any $s \in S$, $n \in \mathbb{N} + \{\infty\}$ and any policy π and it is not difficult to establish that

$$V_\pi(s) \leqq E_{\pi s} X_{n+1} \leqq E_{\pi s} X_n \leqq v(s) \qquad n \in \mathbb{N}. \tag{3.4}$$

Now let us consider a fixed π and s such that $V_\pi(s) > -\infty$. Then $(X_n, n \in \mathbb{N} + \{\infty\})$ is a $P_{\pi s}$-*supermartingale* relative to $(\mathscr{F}_n, n \in \mathbb{N} + \{\infty\})$. The limit

$$X'_\infty = \lim_n X_n$$

exists $P_{\pi s}$-a.s., $(X_n, n \in \mathbb{N}, X'_\infty)$ is a $P_{\pi s}$-supermartingale and $X'_\infty \geqq X_\infty$ $P_{\pi s}$-a.s. (cf. Neveu (1965), Proposition IV.5.4). Therefore we get

$$V_\pi(s) \leqq E_{\pi s} X'_\infty \leqq E_{\pi s} X_n \leqq v(s) \qquad n \in \mathbb{N}. \tag{3.5}$$

Perhaps the reader should be reminded that X'_∞ need not coincide with X_∞ (cf. also Theorem 6.3).

Example 3.6. Let $S = \mathbb{N}$; $A = \{0, 1\}$; $D = S \times A$; $q(1, \cdot, 1) = 1$, $q(s, 0, s+1) = 1$ and $q(s, 1, 1) = 1$ for all $s > 1$; $r(s, 0) = 0$ and $r(s, 1) = 1 - 1/s$ for all $s \in S$. Then $v := 1 \in \mathscr{E}_\infty$. If $f(s) = 0$, $s \in S$ then for all $s \geqq 2$, $X'_\infty = 1$ and $X_\infty = 0$ $P_f \infty_s$-a.s.

For $\tau \in T$ define $X_\tau(h) = X_{\tau(h)}(h)$, $h \in H$. X_τ is a random variable and $E_{\pi s} X_\tau^- \leqq E_{\pi s} X_\infty^- < \infty$ if $V_\pi(s) > -\infty$. The following theorem may be regarded as an *optional sampling theorem* for decision processes.

Theorem 3.7 (cf. Hordijk (1974), Theorem 2.20). Let $\pi \in \Delta$ and $s \in S$ such that $V_\pi(s) > -\infty$. If τ_1, $\tau_2 \in T$ are stopping times with $\tau_1 \leqq \tau_2$, then

$$E_{\pi s} X_{\tau_1} \geqq E_{\pi s} X_{\tau_2}.$$

In particular,

$$v(s) \geqq E_{\pi s} X_\tau \geqq V_\pi(s), \quad \tau \in T.$$

The *proof* follows from Neveu ((1965), Proposition IV.5.5), since the $P_{\pi s}$-supermartingale $(X_n, n \in \mathbb{N} + \{\infty\})$ is closed at the right.

4. Optimal policies

Throughout this and the next two sections let $\pi \in \Delta$ and $s \in S$ be fixed such that $-\infty < V_\pi(s) \leqq v(s) < \infty$. In the present paper we are concerned with the following concept of optimality.

Let $\varepsilon \geqq 0$. The policy π is called (s, ε)-*optimal* (relative to v) iff $V_\pi(s) \geqq v(s) - \varepsilon$. The policy π is called (s, ε)-*thrifty* iff $\lim_n E_{\pi s} X_n \geqq v(s) - \varepsilon$. The policy π is called (s, ε)-*equalizing* iff $V_\pi(s) \geqq \lim_n E_{\pi s} X_n - \varepsilon$. π is called s-*optimal* (s-*thrifty*, s-*equalizing*) iff π is $(s, 0)$-optimal, $((s, 0)$-thrifty, $(s, 0)$-equalizing).

The notions 'thrifty' and 'equalizing' are adapted from Dubins and Savage (1965) (cf. also Sudderth (1972)). In this section we derive necessary and sufficient conditions for the policy π to be s-optimal. s-thrifty and s-equalizing policies are characterized in later sections.

As a consequence of (3.4) we obtain the following theorem.

Theorem 4.1 (cf. Sudderth (1972), Theorem 8). Let $\varepsilon \geqq 0$. π is (s, ε)-optimal iff π is (s, ε_1)-thrifty and (s, ε_2)-equalizing for some ε_1, ε_2 such that $\varepsilon_1 + \varepsilon_2 \leqq \varepsilon$.

Theorem 4.2. The following statements are equivalent:

(i) π is s-optimal;

(ii) π is s-thrifty and s-equalizing;

(iii) $(X_n, n \in \mathbb{N} + \{\infty\})$ is a $P_{\pi s}$-martingale;

(iv) $v(s) = E_{\pi s} X_\tau$, $\tau \in T$.

Proof. (i) $\Rightarrow$ (ii): obvious from Theorem 4.1. (ii) $\Rightarrow$ (iii): this implication is easily deduced from (3.4) and the fact that $(X_n, n \in \mathbb{N} + \{\infty\})$ is a $P_{\pi s}$-supermartingale. (iii) $\Rightarrow$ (iv): Since the $P_{\pi s}$-martingale $(X_n, n \in \mathbb{N} + \{\infty\})$ is closed at the right, the statement follows from Neveu ((1965), Proposition IV.5.5). (iv) $\Rightarrow$ (i): immediate with $\tau = \infty$.

The following result may be regarded as a slight generalization of Theorem 4.2.

Theorem 4.3. π is s-optimal iff π is s-thrifty and $\lim_n E_{\pi s} v^+ \circ \zeta_n = 0$.

Proof. Since $-\infty < V_\pi(s) \leqq v(s) < \infty$ we get

$$\lim_n E_{\pi s} X_n = V_\pi(s) + \lim_n E_{\pi s} v \circ \zeta_n. \tag{4.1}$$

In view of (4.1) the 'if' direction is obvious. The 'only if' direction is proved as follows. By Theorem 4.2 the policy π is s-thrifty. Now suppose that Assumption A^+ holds. From Theorem 4.2 (iii) one obtains

$$E_{\pi s} v^+ \circ \zeta_n \leqq E_{\pi s}\left[\sum_{k=n}^{\infty} r_k^+\right].$$

The passage to limit $n \to \infty$ yields the desired assertion. If Assumption A^- is satisfied, the statement is derived from (4.1) and Remark 3.4.

Other criteria of optimality are based on the *optimality equation* (or Bellman

equation) and on the function V_π instead of v. These characterizations are contained in the next theorem. Note that the policy π is an arbitrary, not necessarily Markov, policy.

Theorem 4.4. Let v be the smallest excessive function in $\mathscr{E}_t$. Let $\pi \in \Delta$ such that $V_\pi(s) < \infty$, $s \in S$. The following statements are equivalent:

(i) π is s-optimal for every $s \in S$;

(ii) $V_\pi \in \mathscr{E}_t$;

(iii) $V_\pi \geqq UV_\pi$ and $\lim_n E_{\pi' s} V_\pi \circ \zeta_n \geqq 0$ on $\{s: V_\pi(s) > -\infty\}$ for all $\pi' \in \Delta$.

Proof. The equivalence '(i) $\Leftrightarrow$ (ii)' is obvious since v is the smallest function in $\mathscr{E}_t$. '(ii) $\Leftrightarrow$ (iii)' follows from Remark 3.5.

In the *positive case*, i.e. if $r \geqq 0$, the condition '$\lim_n E_{\pi' s} V_\pi \circ \zeta_n \geqq 0$' is always satisfied.

5. Thrifty policies

We write

$$\varepsilon_1(s) = v(s) - E_{\pi s}[r_1 + v \circ \zeta_2]$$

and if $n \geqq 1$

$$\varepsilon_n(h) = v(s_n) - E_{\pi s}[r_n + v \circ \zeta_{n+1} \mid \chi_n = h], \quad h \in H_n.$$

Since $v \in \mathscr{E}_t$ the function ε_n is $(P_{\pi s})_{\chi_n}$-a.s. non-negative, where $(P_{\pi s})_{\chi_n}$ denotes the distribution of χ_n under $P_{\pi s}$. Theorem 5.1 is a close cousin to Theorems 9 and 10 in Sudderth (1972).

Theorem 5.1. (a) Let $\varepsilon \geqq 0$. π is (s, ε)-thrifty iff $E_{\pi s}[\Sigma_1^\infty \varepsilon_n \circ \chi_n] \leqq \varepsilon$. (b) The following four statements are equivalent.

(i) π is s-thrifty;

(ii) $v(s) = E_{\pi s} X_n$, $n \in \mathbb{N}$;

(iii) $\varepsilon_n = 0$ $(P_{\pi s})_{\chi_n}$-a.s., $n \in \mathbb{N}$;

(iv) $(X_n, n \in \mathbb{N})$ is a $P_{\pi s}$-martingale.

Proof. By induction on n we conclude $E_{\pi s} X_n = v(s) - E_{\pi s}[\Sigma_{k=1}^n \varepsilon_k \circ \chi_k]$. Then the Assertion (a) is obtained from the definition of (s, ε)-thrifty. The proof of (b) follows easily by (3.4) and Part (a).

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

Corollary 5.2. Let $\pi = (f_n)$ be a Markov policy.

(a) π is thrifty iff $v = L_{f_n} v$ $(P_{\pi s})_{\zeta_n}$-a.s., $n \in \mathbb{N}$.

(b) Suppose $v = Uv$. Then π is s-thrifty iff for any $n \in \mathbf{N}$ and for $(P_{\pi s})_{\zeta_n}$-almost all $s' \in S$ the point $f_n(s')$ is a maximum point of the map

$$a \to [r(s', a) + \int q(s', a, dt)v(t)], \quad a \in D(s').$$

The criterion given in Corollary 5.2(b) is frequently used in dynamic programming problems. The additional assumption '$v = Uv$' is required only for the "if' direction.

Corollary 5.3. Let $f \in F$. Then f^∞ is s-thrifty for every $s \in S$ if and only if $v = L_f v$.

Now we intend to generalize Theorem 5.1 by using stopping times. The following subset of T will be of interest. Let $T_{\pi s}$ denote the set of all stopping times $\tau \in T$ such that

$$\limsup_n \int_{\{\tau > n\}} X_n dP_{\pi s} \leqq \int_{\{\tau = \infty\}} X_\infty dP_{\pi s}. \tag{5.1}$$

Remark 5.4. Since $(X_n, n \in \mathbf{N} + \{\infty\})$ is a $P_{\pi s}$-supermartingale, (5.1) is equivalent to

$$\lim_n \int_{\{\tau > n\}} X_n dP_{\pi s} = \int_{\{\tau = \infty\}} X_\infty dP_{\pi s}.$$

Let us discuss some *properties* of the set $T_{\pi s}$.

Any bounded stopping time belongs to $T_{\pi s}$. If $\tau \in T$ and $P_{\pi s}(\tau < \infty) = 1$, then $\tau \in T_{\pi s}$ if and only if $\lim_n \int_{\{\tau > n\}} X_n dP_{\pi s} = 0$. In the *negative case*, i.e. if $r \leqq 0$, the stopping time $\tau = \infty$ belongs to $T_{\pi s}$, if $v \leqq 0$. In the *absorbing case*, i.e. if there exists a measurable subset G of S such that $q(s, a, S - G) = 0$, $r(s, a) = 0$ for $s \in G$, $a \in D(s)$, the stopping time $\tau = \inf\{n \in \mathbf{N}: \zeta_n \in S - G\}$ belongs to $T_{\pi s}$, if $v(s) = 0$ on G.

Let $f \in F$ and $q_f(s, s') = q(s, f(s), s')$, $s, s' \in S$. If there exists an absorbing set G_f (not necessarily independent of f) relative to q_f, if $r(s, a) \leqq 0$ for $s \in G_f$, $a \in D(s)$ and if $v(s) \leqq 0$ for $s \in G_f$, then $\tau = \inf\{n \in \mathbf{N}: \zeta_n \in S - G_f\}$ belongs to $T_f \infty_s$. For example, if S is countable G_f can be taken as the set of recurrent states of the Markov chain with transition law q_f.

We are now ready to state the main result of this section.

Theorem 5.5. π is s-thrifty iff $v(s) = E_{\pi s} X_\tau$, $\tau \in T_{\pi s}$.

Proof. The 'if' direction follows from Theorem 5.1(b) with $\tau = n \in T_{\pi s}$. Let π be s-thrifty. By Theorem 5.1 $(X_n, n \in \mathbf{N})$ is a $P_{\pi s}$-martingale. Hence we infer

$$v(s) = \int_{\{\tau \leqq n\}} X_\tau dP_{\pi s} + \int_{\{\tau > n\}} X_n dP_{\pi s} \quad n \in \mathbf{N}, \ \tau \in T.$$

Letting $n \to \infty$ we get $v(s) = E_{\pi s} X_\tau$ if $\tau \in T_{\pi s}$.

An immediate consequence of Theorem 5.5 is

Theorem 5.6. Let π be s-thrifty. π is s-optimal iff there exists a sequence (τ_n) of stopping times $\tau_n \in T_{\pi s}$ such that $\lim_n E_{\pi s} X_{\tau_n} = V_\pi(s)$.

We close this section with a useful characterization of the set $T_{\pi s}$. As usual, we write $\tau \wedge n = \min(\tau, n)$.

Theorem 5.7. Let π be s-thrifty and $\tau \in T$. Then $\tau \in T_{\pi s}$ iff the sequence $(X^+_{\tau \wedge n})$ is uniformly $P_{\pi s}$-integrable and $X'_\infty = X_\infty$ $P_{\pi s}$-a.s. on $\{\tau = \infty\}$.

Proof. Since $(X_n, n \in \mathbb{N}, X'_\infty)$ is a $P_{\pi s}$-supermartingale and $X'_\infty \geqq X_\infty$ $P_{\pi s}$-a.s. we get $X'_\infty = X_\infty$ on $\{\tau = \infty\}$. $\tau \wedge n$ belongs to $T_{\pi s}$ and thus by Theorem 5.5 we have

$$v(s) = E_{\pi s} X_{\tau \wedge n} = E_{\pi s} X_\tau \quad n \in \mathbb{N}.$$

Hence $(X_{\tau \wedge n}, n \in \mathbb{N}, X_\tau)$ is a $P_{\pi s}$-martingale. By Neveu (1965), Proposition IV.5.1, $(X_{\tau \wedge n})$ is uniformly $P_{\pi s}$-integrable, and thus by definition $(X^+_{\tau \wedge n})$ is uniformly $P_{\pi s}$-integrable. To show that the conditions are sufficient we apply Fatou's Lemma to the sequence $(X_{\tau \wedge n})$ and obtain

$$\limsup_n E_{\pi s} X_{\tau \wedge n} \leqq E_{\pi s}\left(\limsup_n X_{\tau \wedge n}\right) = E_{\pi s} X_\tau$$

which completes the proof.

From the proof of Theorem 5.7 it follows that $\tau \in T_{\pi s}$ iff the sequence $(X_{\tau \wedge n})$ is uniformly $P_{\pi s}$-integrable and $X'_\infty = X_\infty$ $P_{\pi s}$-a.s. on $\{\tau = \infty\}$.

6. Equalizing policies

In negative dynamic programming problems every policy is s-equalizing. More generally, the condition (EN) of Hinderer (1970), the condition (C^+) of Hinderer (1971) and the condition (C) of Schäl (1975) guarantee that all policies are equalizing. (This can be proved by Theorem 6.1.) In positive problems, this is not necessarily true (not even for thrifty policies), e.g. in Example 3.6 the stationary policy f^∞ is thrifty, but not s-equalizing for $s > 1$. However, if S is finite and if Assumption A^+ holds then every s-thrifty stationary policy is s-equalizing.

Theorem 6.1. (a) Let $\varepsilon \geqq 0$. π is (s, ε)-equalizing iff $\lim_n E_{\pi s} v \circ \zeta_n \leqq \varepsilon$. (b) π is s-equalizing iff $\tau = \infty$ belongs to $T_{\pi s}$.

Proof. By Remark 3.3 (a) follows from (4.1). (b) is obvious by Remark 5.4.

Theorem 6.2 generalizes Theorem 4.10 in Chow, Robbins and Siegmund (1971) to more general dynamic programming problems.

Theorem 6.2. Let π be s-thrifty. Then π is s-optimal iff the sequence (X_n^+) is uniformly $P_{\pi s}$-integrable and $X'_\infty = X_\infty$ $P_{\pi s}$-a.s.

The *proof* is a consequence of Theorems 4.2, 5.7 and 6.1(b).

If the policy π is s-equalizing then $V_\pi(s) = E_{\pi s}X_\infty = E_{\pi s}X'_\infty$ by (3.5). The converse fails in general. But the converse is true if the sequence $(v^+ \circ \zeta_n)$ is uniformly $P_{\pi s}$-integrable. This additional condition is satisfied in Sudderth (1972). Note that $\lim_n v \circ \zeta_n \geqq 0$ $P_{\pi s}$-a.s. exists.

Theorem 6.3. Let the sequence $(v^+ \circ \zeta_n)$ be uniformly $P_{\pi s}$-integrable. Then the following statements are equivalent:

(i) π is s-equalizing;
(ii) $V_\pi(s) = E_{\pi s}X'_\infty$;
(iii) $P_{\pi s}(X_\infty = X'_\infty) = 1$;
(iv) $\lim_n v \circ \zeta_n = 0$ $P_{\pi s}$-a.s.

Proof. From (3.5) and Fatou's Lemma we conclude

$$\limsup_n E_{\pi s} v \circ \zeta_n = E_{\pi s}\left(\limsup_n v \circ \zeta_n\right). \tag{6.1}$$

Then the statements are immediate.

We remark that Assumption A^+ need not imply that the sequence $(v^+ \circ \zeta_n)$ is uniformly $P_{\pi s}$-integrable. The following theorem yields a slight modification of Theorems 11, 12, and 13 in Sudderth (1972).

Theorem 6.4. Let the sequence $(v^+ \circ \zeta_n)$ be uniformly $P_{\pi s}$-integrable and let $\varepsilon > 0$.

(a) π is (s,ε)-equalizing iff $E_{\pi s}X'_\infty \leqq V_\pi(s) + \varepsilon$.
(b) If $P_{\pi s}(\lim_n v \circ \zeta_n \leqq \varepsilon) = 1$ then π is (s, ε)-equalizing.
(c) If π is (s, ε^2)-equalizing then $P_{\pi s}(\lim_n v \circ \zeta_n \leqq \varepsilon) \geqq 1 - \varepsilon$.

Proof. From (6.1) we get $\lim_n E_{\pi s}X_n = E_{\pi s}X'_\infty$. Then (a) follows by definition. In view of (a) the proof of (b) (resp. (c)) is similar to the proof of Theorem 12 (resp. Theorem 13) of Sudderth (1972) and is therefore omitted.

7. Applications

Throughout this section, S is assumed to be countable. Then the value-function V belongs to $\mathscr{E}_i$. Here we suppose $v = V$. A policy π is called *optimal* if $V_\pi(s) = V(s)$ for every $s \in S$.

The following result has been proved by Blackwell (1970), Ornstein (1969) and Hordijk (1974) in the positive case.

Theorem 7.1. Assume A^+. If there exists an optimal Markov policy then there exists an optimal stationary policy.

Proof. Let $\pi = (f_n)$ be the optimal Markov policy. Without loss of generality we can assume that $V_\pi(s) = V(s) > -\infty$, $s \in S$. Let us consider the positive stationary dynamic program $(S, (A, \mathcal{A}), D', q, r')$ with

$$D' = \{(s, a) \in D : V(s) = r(s, a) + \sum_{t \in S} q(s, a, t) V(t)\}$$

$$r'(s, a) = r^+(s, a), (s, a) \in D'.$$

In virtue of Theorem 4.2 and Corollary 5.2 the graph of f_1 belongs to D' and thus the dynamic program above is well-defined. Without loss of generality we can assume in view of Corollary 5.2 that π belongs to Δ'. Therefore we conclude

$$V(s) \leqq V'(s) \leqq W_+(s) < \infty \quad s \in S. \tag{7.1}$$

Let $\varepsilon > 0$. According to a theorem of Ornstein (1969) there exists a stationary policy f^∞ such that

$$f(s) \in D'(s), \quad s \in S \tag{7.2}$$

$$V'_{f^\infty}(s) \geqq (1 - \varepsilon) V'(s), \quad s \in S. \tag{7.3}$$

Corollary 5.3 and (7.2) tell us that f^∞ is s-thrifty for $s \in S$. By (7.1) and (7.3) we get $E_{f^\infty, s} V \circ \zeta_n \leqq (1 - \varepsilon)^{-1} E_{f^\infty, s} V'_{f^\infty} \circ \zeta_n \to 0 (n \to \infty)$ and by Theorem 6.1, f^∞ is s-equalizing for $s \in S$. Consequently f^∞ is optimal.

Now we consider a stationary policy f^∞. Theorem 5.6 enables us to state sufficient conditions for the s-optimality of f^∞.

Theorem 7.2. Assume A^+. Let f^∞ be s-thrifty and $V_{f^\infty}(s) > -\infty$. If there exists a sequence (τ_n) of stopping times $\tau_n \in T_{f^\infty, s}$ such that $\tau_n \geqq n$ and

$$\limsup_n E_{f^\infty, s} [V \circ \zeta_{\tau_n} 1_{\mathbb{N}} \circ \tau_n] \leqq 0 \tag{7.4}$$

then f^∞ is s-optimal, i.e. $V_{f^\infty}(s) = V(s)$.

Proof. Assumption A^+ implies $V(s) < \infty$. From (7.4) and Fatou's Lemma one obtains $\lim_n E_{f^\infty, s} X_{\tau_n} \leqq V_{f^\infty}(s)$. The assertion follows from Theorems 5.5 and 5.6.

The Condition (7.4) is satisfied e.g. if there exists an absorbing subset G_f of S such that $r(s, a) \leqq 0$ for $s \in G_f$, $a \in D(s)$

$$\sup_{j \notin G_f} V(j) < \infty \quad \text{and} \quad \lim_n P_{f^\infty, s}(\tau_n < \infty) = 0,$$

where $\tau_n = \inf\{k \geqq n : \zeta_k \in S - G_f\}$ (cf. Section 5).

References

BLACKWELL, D. (1965) Discounted dynamic programming. *Ann. Math. Statist.* **36,** 226–235.

BLACKWELL, D. (1970) On stationary policies. *J. R. Statist. Soc.* A **133,** 33–37.

CHOW, Y. S., ROBBINS, H. AND SIEGMUND, D. (1971) *Great Expectations: The Theory of Optimal Stopping.* Houghton-Mifflin Company, Boston.

DUBINS, L. E. AND SAVAGE, L. J. (1965) *How to Gamble if You Must: Inequalities for Stochastic Processes.* McGraw-Hill, New York.

HINDERER, K. (1970) Foundations of non-stationary dynamic programming with discrete time parameter. *Lecture Notes in Operations Research and Mathematical Systems.* **33,** Springer, Berlin.

HINDERER, K. (1971) Instationäre dynamische Optimierung bei schwachen Voraussetzungen über die Gewinnfunktionen. *Abh. Math. Sem. Univ. Hamburg* **36,** 208–223.

HORDIJK, A. (1974) *Dynamic Programming and Markov Potential Theory.* Mathematical Centre Tracts No. 51, Amsterdam.

NEVEU, J. (1965) *Mathematical Foundations of the Calculus of Probability.* Holden-Day, San Francisco.

ORNSTEIN, D. (1969) On the existence of stationary optimal strategies. *Proc. Amer. Math. Soc.* **20,** 563–569.

RIEDER, U. (1975) On stopped decision processes with discrete time parameter. *Stoch. Proc. Appl.* **3,** 365–383.

ROSS, S. M. (1974) Dynamic programming and gambling models. *Adv. Appl. Prob.* **6,** 593–600.

SCHÄL, M. (1975) Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. *Z. Wahrscheinlichkeitsth.* **32,** 179–196.

STRAUCH, R. E. (1966) Negative dynamic programming. *Ann. Math. Statist.* **37,** 871–890.

SUDDERTH, W. (1972) On the Dubins and Savage characterization of optimal strategies. *Ann. Math. Statist.* **43,** 498–507.