On the use of information in Markov decision processes

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computing Science

STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum-COSOR 81-20

On the use of information in Markov decision processes

by

Jan van der Wal and Jaap Wessels

Eindhoven, the Netherlands

December 1981

ON THE USE OF INFORMATION IN MARKOV DECISION PROCESESSES

by

Jan van der Wal and Jaap Wessels,

Eindhoven

Abstract. This paper gives a systematic treatment of results about the existence of various types of nearly-optimal strategies (Markov, stationary) in countable state total reward Markov decision processes. For example the following questions are considered: do there exist optimal stationary strategies, uniformly nearly-optimal stationary strategies or uniformly nearly-optimal Markov strategies.

1. INTRODUCTION.

This paper deals with the existence of certain types of (nearly-)optimal strategies for the total reward Markov decision process (MDP) with countable state space.

Ever since SHAPLEY [1953] obtained the first result in this direction: the existence of an optimal stationary strategy for the contracting MDP with finite state and action spaces, there has been an almost continuous process of extending certain existence results to more general models (a.o. BLACKWELL [1965], STRAUCH [1966], ORNSTEIN [1969]). Here we will try to give a systematic treatment of the various results for the total reward MDP with countable state space. To be more precise: we consider the question of what kind of information about the history of the process is needed to take good decisions. For example, if stationary strategies

AMS subject classification scheme (1979): 90C47.

Key Words and phrases: Markov decision processes, stationary strategies, Markov strategies.

are sufficient, then one only needs as information the present state, if however we have to consider Markov strategies then we need the present state and the time as information.

The generalization to uncountable state spaces involves techniques of a different nature (measure theory, selection theorems, analytic sets). We will not consider this topic here.

Now let us introduce in a kind of semi-formal way some of the basic notations and definitions for the MDP. For a more extensive and formal introduction see e.g. VAN DER WAL [1981a]. So, consider a dynamic system with countable state space S and arbitrary action space A endowed with some σ-field $\mathcal{A}$ containing all one-point sets. If in state $i \in S$ action $a \in A$ is taken, two things happen: a (possibly negative) reward $r(i,a)$ is earned and a transition is made to state j, $j \in S$, with probability $p(i,a,j)$, where $\Sigma_j\, p(i,a,j) = 1$. The functions $r(i,.)$ and $p(i,.,j)$ are assumed to be $\mathcal{A}$-measurable.

We will distinguish four sets of strategies, namely, the set Π of all randomized and history dependent strategies satisfying the usual measurability conditions with respect to the history of the process, the set RM of all randomized Markov strategies, the set M of all nonrandomized Markov strategies or shortly Markov strategies and F the set of all nonrandomized stationary strategies, shortly stationary strategies. So $F \subset M \subset RM \subset \Pi$. The elements of F will also be called policies and are treated as functions on S. A Markov strategy is often denoted by the sequence $(f_0, f_1, \ldots)$ of functions from S into A specifying the action to be taken at each time.

For each strategy $\pi \in \Pi$ and each initial state $i \in S$, we define in the usual way a probability measure $\mathbb{P}_{i,\pi}$ on $(S \times A)^{\infty}$ and a stochastic process $\{(X_n,A_n),\ n = 0,1,\ldots\}$, where X_n denotes the state of the system at time n and A_n the action chosen at time n. Expectations with respect to $\mathbb{P}_{i,\pi}$ will be denoted by $\mathbb{E}_{i,\pi}$.

Now, the total expected reward, when the process starts in state i and strategy π is used, can be defined by

$$(1.1) \qquad v(i,\pi) := \mathbb{E}_{i,\pi} \sum_{n=0}^{\infty} r(X_n,A_n),$$

whenever the expectation at the right hand side is well-defined. To quarantee this, the following assumption will be made.

GENERAL CONVERGENCE CONDITION. For all $i \in S$ and all $\pi \in \Pi$

$$(1.2) \qquad u(i,\pi) := \mathbb{E}_{i,\pi} \sum_{n=0}^{\infty} r^{+}(X_n,A_n) < \infty \ ^{1)} .$$

A somewhat weaker condition would be

$$(1.3) \qquad \mathbb{E}_{i,\pi} \left[\sum_{n=0}^{\infty} r(X_n,A_n) \right]^{+} < \infty \text{ for all } i \in S ,\ \pi \in \Pi.$$

Throughout this paper, however, we will assume the General Convergence Condition to hold. This condition allows for the interchange of expectation and summation in (1.1) and implies

$$(1.4) \qquad \lim_{n\to\infty} v_n(i,\pi) = v(i,\pi) ,$$

where

$$(1.5) \qquad v_n(i,\pi) := \mathbb{E}_{i,\pi} \sum_{k=0}^{n-1} r(X_k,A_k) .$$

---

1 For any real-valued function f the functions f^{+} and f^{-} are defined by $f^{+} = \max\{0,f\}$ and $f^{-} = \min\{0,f\}$.

Note that under condition (1.3) the interchange need not be allowed.

The *value* of the total reward MDP is defined by

$$(1.6) \qquad v^*(i) := \sup_{\pi \in \Pi} v(i,\pi) \ .$$

Further, we will also consider the related MDP where the negative rewards are neglected, so $r(i,a)$ is replaced by $r^+(i,a)$. For this MDP the value function is denoted by u^*:

$$(1.7) \qquad u^*(i) = \sup_{\pi \in \Pi} u(i,\pi) \ .$$

The rest of the paper is organized as follows. First, in section 2, some basic results and concepts are presented which will be frequently used in the following sections. Section 3 gives some results on the existence of stationary optimal strategies. Section 4 considers the existence of (uniformly) nearly-optimal stationary strategies. In section 5 the question whether the existence of an optimal strategy implies the existence of a stationary optimal one is considered. Section 6 deals with uniformly nearly-optimal Markov strategies.

This introductory section is concluded with some notations and a remark. If the argument i corresponding to the state is deleted the function on S is meant. For example, $v(\pi)$ and v^* are the functions with i-th coordinate $v(i,\pi)$ and $v^*(i)$ respectively. Frequently, these functions will be treated as columnvectors.

The following notations for policies will be very useful. Let f be any policy, then the immediate reward function $r(f)$ and the transition

probability function P(f), which will be treated as a columnvector and a matrix respectively, are defined by

(1.8) $\quad r(f)(i) := r(i,f(i)) \ , \ i \in S \ .$

(1.9) $\quad P(f)(i,j) := p(i,f(i),j) \ , \ i,j \in S.$

Further, we define on suitable subsets of the set of functions on S the operators L(f) and U by

(1.10) $\quad L(f)v := r(f) + P(f)v \ .$

(1.11) $\quad Uv := \sup_{f \in F} L(f)v \ .$

Finally, note that the most intensively studied total reward MDP, the discounted model (see e.g. BLACKWELL [1965]), can be made to fit in our model by the introduction of an extra absorbing state, $*$ say, and redefinition of the transition probabilities by

$$\tilde{p}(i,a,j) := \beta p(i,a,j) \ , \ i,j \in S$$

$$\tilde{p}(i,a,*) := 1 - \beta \ , \ i \in S \ .$$

With one-state rewards

$$\tilde{r}(i,a) := r(i,a) \ , \ i \in S$$

$$\tilde{r}(*,a) := 0 \ .$$

2. *SOME BASIC RESULTS AND CONCEPTS*

The general question we are interested in is: what kind of strategies do we have to consider, or, in different terms, what information about

past and present is relevant for a good control of the system.

A first partial answer for the general case of a countable state space and arbitrary action space has been given by DERMAN and STRAUCH [1966].

LEMMA 2.1. Let $\pi \in \Pi$ be some arbitrary strategy and $i \in S$ be some arbitrary initial state, then there exists a strategy $\hat{\pi} \in RM$ such that $v(i,\hat{\pi}) = v(i,\pi)$.

So this lemma states that for each initial state you only need to consider randomized Markov strategies. And thus the relevant information needed for choosing actions consists of the initial state, the present state and the time.

The proof of the lemma is a construction of $\hat{\pi}$ and, actually, very simple. The strategy $\hat{\pi}$ is defined in such a way that π and $\hat{\pi}$ have the same marginal distributions at each time instant with respect to the state-action combination.

A second result in this general setting has been obtained by VAN HEE [1978]. He proved the following result.

LEMMA 2.2.

$$\sup_{\pi \in M} v(i,\pi) = v^*(i) \text{ for all } i \in S.$$

So, again for a fixed initial state, Markov strategies are (almost) as good as any and using randomized actions is not necessary. The term "almost" refers to the fact that in principle it might still occur that an optimal strategy within RM exists whereas no optimal Markov strategy exists. As we will see in section 5, however, the term "almost" is

superfluous, since if there is an optimal strategy then there is also a stationary one, so there is certainly a Markov strategy.

Note that Derman and Strauch's as well as Van Hee's result is only pointwise, i.e. for fixed starting state. Lemma 2.1 certainly need not hold uniformly in the initial state. That Lemma 2.2 holds even uniformly, i.e. that for each $\varepsilon > 0$ a strategy $\pi \in M$ exists satisfying

$$v(\pi) \geq v^* - \varepsilon f$$

for some nonnegative function f on S, will be seen in section 6.

A third basic result is the following.

LEMMA 2.3.

$$v^* = Uv^*.$$

So v^* satisfies the optimality equation $v = Uv$. The solution to this equation need not be unique, but the fact that v^* is a solution allows in some cases for simple proofs of the existence of uniformly (nearly-) optimal strategies of certain type.

Nex, we will formulate two concepts which will play a crucial role in our analysis, particularly in analyzing whether some stationary strategy is optimal. Together these concepts, as will be seen in section 3, exploit Lemma 2.3.

DEFINITION 2.4. A policy f is called *conserving* if $L(f)v^* = v^*$.

DEFINITION 2.5. A policy f is called *equalizing* if

$$\limsup_{n\to\infty} \mathbb{E}_f\, v^*(X_n) \leq 0 .$$

One easily argues that $L(f)v^*$ and $\mathbb{E}_f v^*(X_n)$ are properly defined as for all $\varepsilon > 0$ there exists a strategy $\pi \in \Pi$ such that $v^* \geq v(\pi) \geq v^* - \varepsilon e$ [1] . So, for example, $L(f)v^*$ and $L(f)v(\pi)$ are almost equal, and $L(f)v(\pi)$, being the total expected reward for the strategy: play f first and then start with π, is well-defined by the General Convergence Condition.

The concepts conserving and equalizing have been first used by DUBINS and SAVAGE [1965] for gambling problems. For MDP's they have been introduced by HORDIJK [1974]. For the relevance of these concepts in more general decision processes, see GROENEWEGEN [1981] or GROENEWEGEN and WESSELS [1980].

We will complete this section with the definition of various types of (nearly-)optimality.

A strategy π is called *optimal* if $v(\pi) = v$.

A strategy π is called ε-*optimal for initial state* i if $v(i,\pi) \geq v^*(\pi) - \varepsilon$.

A strategy π is called εv-*optimal* (where v is a nonnegative function on S) if $v(\pi) \geq v^* - \varepsilon v$.

The latter definition describes some sort of uniform nearly-optimality.

In the sequel various functions v will appear.

3. *OPTIMAL STATIONARY STRATEGIES*

For stationary strategies the only relevant information needed to choose the actions is the present state of the system. So it is important to have rather general conditions which allow for the consideration of

---

[1] e denotes the unitfunction on S: $e(i) = 1$ for all $i \in S$.

stationary strategies only. This will be the subject of the sections 3 and 4. The present section deals with optimal strategies, whereas section 4 considers the existence of nearly-optimal stationary strategies.

As remarked in the preceding section, the concepts "conserving" and "equalizing" are very useful for proving the existence of optimal stationary strategies. The following theorem characterizes optimal stationary strategies.

THEOREM 3.1. (HORDIJK [1974]).

A stationary strategy f is optimal if and only if f is conserving and equalizing.

PROOF. The if part of the proof follows immediately from $v_n(f) \to v(f)$ (cf. formula (1.4)) and

$$v_n(f) = L^n(f)0 = L^n(f)v^* - \mathbb{E}_f v^*(X_n) = v^* - \mathbb{E}_f v^*(X_n). \qquad \square$$

By specifying conditions quaranteeing conservingness and equalizingness, one obtains the following corollary.

COROLLARY 3.2. For finite A, there exists in each of the following 4 cases an optimal stationary strategy

(i) S finite and discounted rewards (SHAPLEY [1953]).

(ii) r bounded and discounted rewards.

(iii) $v^* \leq 0$ (for example as a result of $r \leq 0$) (STRAUCH [1966]).

(iv) There exists a system of Liapunov functions of order 2, i.e. a pair ℓ_1, ℓ_2 of nonnegative functions on S satisfying for all f

$$\ell_1 \geq |r(f)| + P(f)\ell_1$$

$$\ell_2 \geq \ell_1 + P(f)\ell_2$$

(see HORDIJK [1974] and VAN HEE, HORDIJK and VAN DER WAL [1977]).

The condition "A if finite" can be replaced by any other condition quaranteeing sup $L(f)v = \max L(f)v$ for functions v on S (or for v^* only). For example, "A is compact and r(i,a) and p(i,a,j) are continuous in a".

Note that various other contracting models, such as the models in HARRISON [1972], VAN NUNEN [1976] and VAN NUNEN and WESSELS [1977] can be transformed into a "standard" discounted MDP (see e.g. VAN DER WAL [1981a, chapter 5]). So case (ii) extends to these models as well.

The following result is not a direct application of Theorem 3.1, but requires a simple intermediate step.

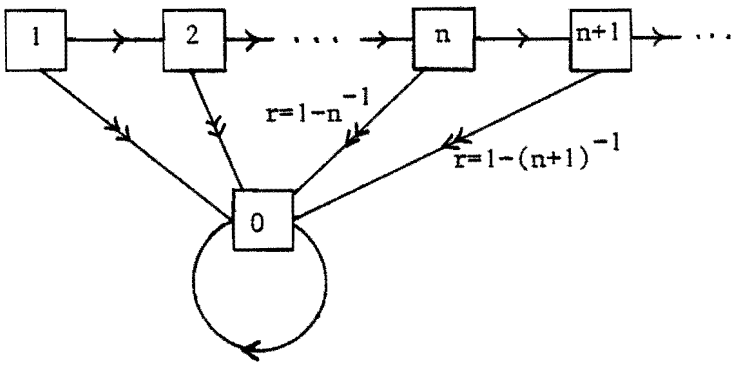
THEOREM 3.3. (see e.g. KALLENBERG [1980] and VAN DER WAL [1981a]). If S and A are finite, then an optimal stationary strategy exists.

PROOF. Let $\{\beta_n\}$ be a sequence of discountfactors tending to 1. By Corollary 3.2(i) there is for any β_n an optimal policy. Since the policy set F is finite, there is a policy f and a subsequence of $\{\beta_n\}$, also tending to 1, for which f is optimal. As for any π the total expected β-discounted reward converges to the total expected (undiscounted) reward if $\beta \uparrow 1$, this policy is also optimal for the undiscounted MDP. □

We will conclude this section with two examples in which optimal stationary strategies do no exist.

The first example shows that in general finiteness of the action space does not quarantee the existence of optimal strategies.

EXAMPLE 3.4. (STRAUCH [1966, Example 4.2]).



$S = \{0,1,2,\ldots\}$, $A = \{1,2\}$. All rewards and transition probabilities are zero except the following:

$$p(n,1,n+1) = 1 \ , \ n = 1,2,\ldots$$

$$p(n,2,0) = 1 \ , \ n = 1,2,\ldots$$

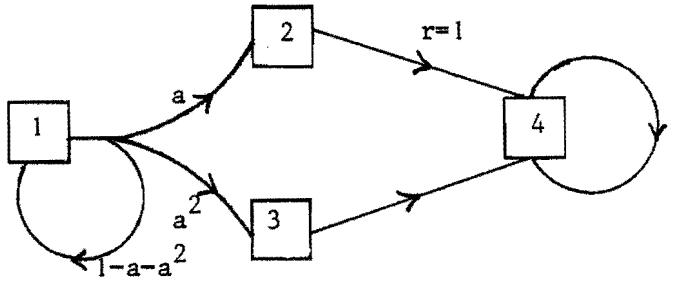
$$p(0,a,0) = 1 \ , \ a = 1,2$$

$$r(n,2) = 1 - n^{-1}, \ n = 1,2,\ldots$$

Clearly $v^*(n) = 1$ for $n = 1,2,\ldots$, but for all π we have $v(n,\pi) < 1$ for all $n = 1,2,\ldots$.

The following example, due to Bather, shows that the condition "S and A are finite" in Theorem 3.3 can not be weakened to "S is finite and A is compact with continuity of r and p".

EXAMPLE 3.5. (BATHER [1973]).



$S = \{1,2,3,4\}$, $A = [0,\frac{1}{2}]$. All $r(i,a)$ and $p(i,a,j)$ are zero except for

$$r(2,a) = 1 \ , \ p(1,a,2) = a \ , \ p(1,a,3) = a^2$$

$$p(1,a,1) = 1 - a - a^2 \ , \ p(2,a,4) = p(3,a,4) = p(4,a,4) = 1, a \in A.$$

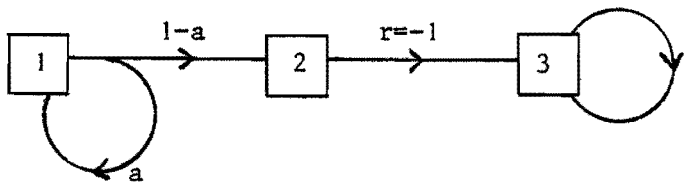
Clearly $v^*(1) = 1$. But for any $\pi \in \Pi$ we have $v(1,\pi) < 1$.

4. *UNIFORMLY NEARLY-OPTIMAL STATIONARY STRATEGIES*

In Theorem 3.1 we have seen that the conditions of conservingness and equalizingness together imply the optimality of a stationary strategy.

In this section we consider the conservingness condition and a fairly strong kind of equalizingness condition separated. As we will see, each of these two conditions implies the existence of an, in a sense, uniformly nearly-optimal stationary strategy. These results will be given in Theorems 4.2 and 4.3. Before giving these results, first an example is presented which shows that finiteness of the state space is in general not sufficient for the existence of nearly-optimal stationary strategies.

EXAMPLE 4.1.



$S = \{1,2,3\}$, $A = [0,1)$. All $r(i,a)$ and $p(i,a,j)$ are zero, except

$$p(1,a,1) = a \ , \ p(1,a,2) = 1-a \ , \ p(2,a,3) = p(3,a,3) = 1 \ ,$$

$$r(2,a) = -1 \ , \ a \in A.$$

Here $v^*(1) = 0$, but $v(1,f) = -1$ for all $f \in F$.

As the following theorem shows, conditions on the action space are far more useful.

THEOREM 4.2. (VAN DER WAL [1981a]).

If in each state i for which $v^*(i) \leq 0$ there exists a conserving action [1)], then

[1] An action a in state i is called conserving, if $r(i,a) + \Sigma_j p(i,a,j)v(j) = v^*(i)$.

there exists an εu^*-optimal stationary strategy, i.e. there exists an f satisfying

(4.1) $\quad v(f) \geq v^* - \varepsilon u^*$.

PROOF. Only a brief outline will be given. For details see VAN DER WAL [1981b].

Define S^- and S^+ by $S^- := \{i \in S | v^*(i) \leq 0\}$, $S^+ := \{i \in S | v^*(i) > 0\}$. By assumption, there exist conserving actions on S^-. As one may show, fixing a conserving policy on S^- does not affect the value. Next, the MDP is embedded on S^+ (the policy on S^- is held fixed). The embedded MDP now has $v^* > 0$. Then this positively valued MDP is transformed into an MDP with nonnegative immediate rewards. For this MDP a uniformly nearly-optimal stationary strategy exists in the sense of (4.1) (see ORNSTEIN [1969] and (4.3) below). Finally, it can be shown that this policy, combined with the fixed conserving actions on S^-, satisfies (4.1) for the original MDP. □

This theorem generalizes the following two results.

(4.2) If $r \leq 0$ (so $u^* = 0$) and A is finite or compact, then an optimal stationary strategy exists (cf. also Corollary 3.2(iii)).

(4.3) If $r \geq 0$, then an εv^*-optimal strategy exists (see ORNSTEIN [1969]).

The strong equalizingness condition which will be considered in Theorem 4.3 is in fact a condition on the tail of the income streams. In order to be able to formulate this condition, some definitions are needed.

Denote by Φ the set of nondecreasing sequences $\varphi = (\varphi_0, \varphi_1, \ldots)$ with

$\varphi_0 \geq 1$ and $\varphi_n \to \infty$.

Define

$$z_\varphi(\pi) := \mathbb{E}_\pi \sum_{n=0}^{\infty} \varphi_n |r(X_n, A_n)|$$

$$z_\varphi^* := \sup_\pi z_\varphi(\pi).$$

Now, an MDP is called *uniformly strongly convergent*, if a sequence $\varphi \in \Phi$ exists for which $z_\varphi^* < \infty$. (See for the introduction of this kind of conditions VAN HEE, HORDIJK and VAN DER WAL [1977] or VAN DER WAL [1981a, chapter 4]).

This condition implies - and is actually equivalent to the condition - that the sum of the absolute rewards from time n onwards tends to zero in a uniform way if $n \to \infty$.

THEOREM 4.3. (VAN HEE and VAN DER WAL [1977, Theorem 7] ór VAN DER WAL [1981a, Theorem 4.11]).
Let the MDP be uniformly strongly convergent for $\varphi \in \Phi$, i.e. $z_\varphi^* < \infty$, then an εz_φ^*-optimal stationary strategy exists, i.e. there exists an f satisfying

$$v(f) \geq v^* - \varepsilon z_\varphi^* .$$

The following results can be seen as special cases of this theorem.

(i) r is bounded and rewards are discounted, then an εe-optimal stationary strategy exists (if the discountfactor β is incorporated in the transition probabilities, then take $\varphi_n = s^n$ with $1 < s < \beta^{-1}$; this implies $z_\varphi^* \leq (1 - \beta s)^{-1} \sup_{i,a} |r(i,a)|e$).

(ii) There exists a nonnegative function μ and constants $C > 0$ and $0 < \varphi < 1$ satisfying for all f

$$|r(f)| \leq C\mu \text{ and } P(f)\mu \leq \varphi\mu ,$$

then an $\varepsilon\mu$-optimal stationary strategy exists (cf. WESSELS [1977], VAN NUNEN [1976] and VAN NUNEN and WESSELS [1977]).

As remarked before, model (ii) can be transformed into a standard discounted model.

(iii) There exists a system of Liapunov functions (ℓ_1,ℓ_2) of order 2 (cf. Corollary 3.2(iv)), then an $\varepsilon\ell_2$-optimal stationary strategy exists (cf. VAN HEE, HORDIJK and VAN DER WAL [1977, theorem 7.2]).

Note that Theorems 4.2 and 4.3 imply that the relevant information consists of the starting state and the present state. If, moreover, u^* in Theorem 4.2 or z_φ^* in Theorem 4.3 is bounded, then an εe-optimal stationary strategy exists, and in that case the only relevant information is the present state.

Also note that Theorems 4.2 and 4.3 show that for a finite set of initial states always a uniformly ε-optimal stationary strategy exists.

5. *OPTIMAL STRATEGIES*

An interesting question is the following. Suppose an optimal strategy exists. Does there exist an optimal stationary strategy?

This question has been considered by STRAUCH [1966] for the negative dynamic programming case and by ORNSTEIN [1969] for the positive case. They showed that in these two cases the answer is affirmative.

Negative case; $r \leq 0$. Here the argument is simple. If an optimal strategy exists, then certainly there are conserving actions in each state, so a conserving policy exists. Since $v^* \leq 0$, this policy is equalizing, whence by Theorem 3.1 also optimal.

Positive case; $r \geq 0$. The optimal strategy uses essentially only conserving actions. So, eliminating all nonconserving actions in each state does not affect the value. By Ornstein's theorem (see (4.3)), a stationary strategy exists satisfying $v(f) \geq \alpha v^*$ for some $\alpha > 0$. But, since f is also conserving (the nonconserving actions being eliminated), even $v(f) = v^*$ (see ORNSTEIN [1969]).

Only recently these partial results have been extended to the case with both positive and negative immediate rewards.

THEOREM 5.1. (see VAN DER WAL [1981b]).
If an optimal strategy exists, then also an optimal stationary strategy exists.

PROOF. The proof, which is heavily based on Ornstein's result for the positive case, can be found in VAN DER WAL [1981a].

So restricting the strategy set from RM to M or from M to F does not affect the existence of an optimal strategy.

6. *UNIFORMLY NEARLY-OPTIMAL MARKOV STRATEGIES*

Until now, we have been formulating conditions for the existence of stationary optimal or nearly-optimal strategies. If only stationary strategies need to be considered, then the information needed to choose the action consists only of the present state and in some cases the initial state. If one cannot restrict the attention to stationary strategies, then, given the initial state, Markov strategies are always sufficient (cf. Lemma 2.2). In this section, we are interested in the question whether the initial state

is important or, in different terms, whether a uniformly nearly-optimal Markov strategy exists.

Therefore, let us first fix some $\varepsilon > 0$ and define a sequence of policies $\{f_n, n = 0,1,\ldots\}$ satisfying

$$L(f_n)v^* \geq v^* - \varepsilon 2^{-n} e \ . \tag{6.1}$$

Clearly such a sequence always exists.

Now, let π be the Markov strategy $(f_1,f_2,\ldots)$.

Then

$$\begin{aligned} v(\pi) &= \lim_{n\to\infty} v_n(\pi) = \lim_{n\to\infty} L(f_1)L(f_2) \ldots L(f_n)0 \\ &= \lim_{n\to\infty} \{L(f_1)L(f_2) \ldots L(f_n)v^* - P(f_1)P(f_2) \ldots P(f_n)v^*\} \qquad (6.2) \\ &\geq v^* - \varepsilon e - \limsup_{n\to\infty} \mathbb{E}_\pi v^*(X_n) \ . \end{aligned}$$

The Markov strategy π might be called εe-conserving. And we see from (6.2) that if π is equalizing (cf. Definition 2.5), then π will be εe-optimal.

So we have the following theorem.

THEOREM 6.1. (cf. STRAUCH [1966, Theorem 8.1]).

Let $\pi = (f_1,f_2,\ldots)$ be a Markov strategy satisfying (6.1) and let $\limsup_{n\to\infty} \mathbb{E}_\pi v^*(X_n) \leq 0$, then $v(\pi) \geq v^* - \varepsilon e$.

Thus, in the negative dynamic programming case εe-optimal Markov strategies exist.

In general, however, εe-optimal Markov strategies need not exist, see e.g. Example 2.26 in VAN DER WAL [1981a]. In this example not even an εe-optimal randomized Markov strategy exists. Also εu^*-optimal Markov

strategies do not exist in general as is immediate from negative dynamic programming. So, in the negative case εe-optimal, but not necessarily εu^*-optimal, Markov strategies exist, whereas in the positive case εu^*-optimal, but not necessarily εe-optimal, Markov strategies exist.

For the case of both positive and negative immediate rewards these partial results can be combined into the following theorem.

THEOREM 6.2.

For each $\varepsilon > 0$ a Markov strategy π exists satisfying

$$v(\pi) \geq v^* - \varepsilon(e + u^*).$$

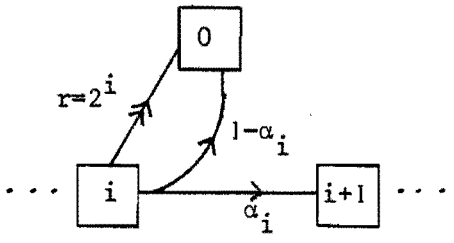
PROOF. The proof uses the same ideas as the proof of Theorem 4.2 and can be found in VAN DER WAL [1981c]. □

From this we see that the only relevant information needed to choose the decisions in the MDP are the time, the present state and, in case u^* is not bounded, the initial state.

7. *SOME REMARKS*

In Theorems 4.2 and 6.2 the function u^* indicates the type of near-optimality of the strategy. Is it possible to replace u^* by an essentially smaller function? Consider the following example.

EXAMPLE 7.1. (cf. VAN DER WAL [1981a, Example 2.25]).



$S = \{0,1,2,\ldots\}$, $A = \{1,2\}$. All rewards and transition probabilities are 0 except for the following. For $i \geq 1$

$r(i,2) = 2^i$, $p(i,2,0) = 1$, $p(i,1,0) = 1-\alpha_i$, $p(i,1,i+1) = \alpha_i$ and $p(0,a,0) = 1$, $a \in A$.

Now let α_i be equal to $(1 + \gamma_i)/2(1 + \gamma_{i+1})$ with $\gamma_i \downarrow 0$.
Then for $i \geq 1$

$$v^*(i) = \sup\{2^i , \alpha_i 2^{i+1} , \alpha_i\alpha_{i+1}2^{i+2} ,\ldots\}$$

$$= \sup\left\{2^i , \frac{1 + \gamma_i}{1 + \gamma_{i+1}} 2^i , \frac{1 + \gamma_i}{1 + \gamma_{i+2}} 2^i ,\ldots\right\}$$

$$= 2^i(1 + \gamma_i).$$

Any stationary strategy that is $\frac{1}{2}u^*$-optimal takes action 2 in infinitely many states. And in those states i where action 2 is taken roughly a fraction γ_i of $v^*(i)$ is lost. The slower γ_i tends to 0 the closer the loss function comes to u^* in the sense that, if $i \to \infty$, the loss goes to ∞ almost as fast as $u^*(i)$. So errors expressed in u^* seem to be about as good as possible.

This covers the case of Theorem 4.2. Extending Example 7.1 in the same way as Example 2.25 is extended to 2.26 in VAN DER WAL [1981a], the argument can be extended to the case of Theorem 6.2.

Our second remark concerns a special type of history dependent strategies, called *tracking* strategies. These strategies have been introduced in HILL [1979]. For tracking strategies the selection of the action may depend only on the present state and the number of times this state has been visited previously.

When using Markov strategies, you can take a better action at each time you come back to a state. Intuitively this is the way time is used in a Markov strategy. Thinking of Markov strategies in this way if seems that also tracking strategies should be good.

So we conjecture that Theorem 6.2 also holds for tracking strategies.

References

BATHER, J. [1973], Optimal decision procedures for finite Markov chains. Part II, Adv. Appl. Prob. 5, 521-540.

BLACKWELL, D. [1965], Discounted dynamic programming. Ann. Math. Statist. 36, 226-235.

DERMAN, C. and R. STRAUCH [1966], A note on memoryless rules for controlling sequential control processes. Ann. Math. Statist. 37, 276-278.

DUBINS, L. and L. SAVAGE [1965], How to gamble if you must: inequalities for stoachastic processes. Mc. Graw-Hill, New York.

GROENEWEGEN, L. [1981], Characterization of optimal strategies in dynamic games. Math. Centre Tract 90, Mathematisch Centrum, Amsterdam.

GROENEWEGEN, L., and J. WESSELS [1980], Conditions for optimality in multi-stage stochastic programming problems. In recent results in stochastic programming, eds. P. KALL and A. PREKOPA. Springer-Verlag, Berlin, 41-57.

HARRISON, J. [1972], Discrete dynamic programming with unbounded rewards. Ann. Math. Statist. 43, 636-644.

HEE, K. VAN [1978], Markov strategies in dynamic programming. Math. Oper. Res. 3, 37-41.

HEE, K. VAN, A. HORDIJK and J. VAN DER WAL [1977], Succesive approximations for convergent dynamic programming. In Markov decision theory, eds. H. TIJMS and J. WESSELS, Math. Centre Tract 93, Mathematisch Centrum, Amsterdam, 183-211.

HEE, K. VAN and J. VAN DER WAL [1977], Strongly convergent dynamic programming: some results. In Dynamische Optimierung, ed. M. SCHÄL, Bonner Math. Schriften nr. 98, Bonn, 165-172.

HILL, T. [1979], On the existence of good Markov strategies. Transactions Amer. Math. Soc. 247, 157-176.

HORDIJK, A. [1974], Dynamic programming and Markov potential theory. Math. Centre Tract 51, Mathematisch Centrum, Amsterdam.

KALLENBERG, L. [1980], Linear programming and finite Markovian control problems. Doctoral dissertation, Univ. of Leiden.

NUNEN, J. VAN [1976], Contracting Markov decision processes. Math. Centre Tract 71, Mathematisch Centrum, Amsterdam.

NUNEN, J. VAN and J. WESSELS [1977], Markov decision processes with unbounded rewards. In Markov decision theory, eds. H. TIJMS and J. WESSELS, Math. Centre Tract 93, Mathematisch Centrum, Amsterdam, 1-24.

ORNSTEIN, D. [1969], On the existence of stationary optimal strategies. Proc. Amer. Math. Soc. 20, 563-569.

SHAPLEY, L. [1953], Stochastic games. Proc. Nat. Acad. Sci. 39, 1095-1100.

STRAUCH, R. [1966], Negative dynamic programming. Ann. Math. Statist. 37, 871-889.

WAL, J. VAN DER [1981a], Stochastic dynamic programming, Math. Centre Tract 139, Mathematisch Centrum, Amsterdam.

WAL, J. VAN DER [1981b], On stationary strategies. Eindhoven Univ. of Technology, Dept. of Math. and Comp. Sci. Memorandum-COSOR 81-14.

WAL, J. VAN DER [1981c], On uniformly nearly-optimal Markov strategies. Eindhoven Univ. of Technology, Dept. of Math. and Comp. Sci., Memorandum-COSOR 81-16.

WESSELS, J. [1977], Markov programming by successive approximations with respect to weighted supremum norms. J. Math. Anal. Appl. 58, 326-335.