On uniformly nearly-optimal Markov strategies

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computing Science

STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 81-16

On uniformly nearly ε-optimal Markov strategies

by

Jan van der Wal

Eindhoven, The Netherlands

October 1981

ON UNIFORMLY NEARLY-OPTIMAL MARKOV STRATEGIES

by

Jan van der Wal

Abstract

In this paper the following result is proved. In any total reward countable state Markov decision process a Markov strategy π exists which is uniformly nearly-optimal in the following sense: $v(\pi) \geq v^* - \varepsilon(e+u^*)$. Here v^* denotes the value function of the process, u^* denotes the value of the process if all negative rewards are neglected, and e is the unit function.

1. Introduction

Consider a Markov decision process (MDP) with countable state space S and arbitrary action space A, with a σ-field containing all one-point sets. If in state $i \in S$ action $a \in A$ is taken two things happen: a (possibly negative) immediate reward $r(i,a)$ is earned and the system moves to a new state j, $j \in S$, with probability $p(i,a,j)$, where $\Sigma_j\, p(i,a,j) = 1$. The functions $r(i,a)$ and $p(i,a,j)$ are assumed to be measurable in a. Three sets of strategies will be distinguished, namely the set Π of all (possibly randomized and history dependent) strategies satisfying the usual measurability conditions, the set M of all nonrandomized Markov strategies, and the set F of all nonrandomized stationary strategies. So $F \subset M \subset \Pi$. The set of all functions from S into A, also called policies, will be denoted by F as well. For each strategy $\pi \in \Pi$ and each initial state one may 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 $\mathbb{P}_{i,\pi}$ are denoted by $\mathbb{E}_{i,\pi}$.

Now the total expected reward $v(i,\pi)$ when the process starts in $i \in S$ and strategy $\pi \in \Pi$ is used can be defined

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

whenever the expectation at the right hand side is well-defined. To guarantee this the following assumption is made.

General convergence condition

For all $i \in S$ and $\pi \in \Pi$

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

where

$$r^+(i,a) := \max\{0, r(i,a)\} \quad , \quad i \in S, \ a \in A .$$

The value of the total reward MDP is defined by

$$(1.3) \qquad v^*(i) := \sup_{\pi \in \Pi} v(i,\pi) .$$

Further the value of the MDP where only the positive rewards are counted is denoted by u^*, so

$$(1.4) \qquad u^*(i) = \sup_{\pi \in \Pi} u(i,\pi) .$$

Van Hee [1978] proved that under the general convergence condition one can restrict, pointwise, the attention to Markov strategies, i.e.

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

In Van der Wal [1981] it is proved that if in each state i for which $v^*(i) \leq 0$ a conserving action exists (i.e. an action a satisfying

$$r(i,a) + \Sigma_j \, p(i,a,j) \, v^*(j) = v^*(i) \quad) ,$$

then for each $\varepsilon > 0$ a stationary strategy exists satisfying for all $i \in S$

$$(1.5) \qquad v(i,f) \geq v^*(i) - \varepsilon u^*(i) \ .$$

In this paper the following result will be proved.

Theorem 1.

For each $\varepsilon > 0$ a Markov strategy π exists satisfying for all $i \in S$

$$(1.6) \qquad v(i,\pi) \geq v^*(i) - \varepsilon - \varepsilon u^*(i) \ .$$

This theorem extends Van Hee's result in showing that there exist not only pointwise nearly-optimal Markov strategies but even uniformly nearly-optimal Markov strategies. Further note that for this theorem to hold neither conditions on the action space nor conditions on the reward structure are needed.

In Van Hee, Hordijk and Van der Wal [1977] an example is given which shows that a Markov strategy $\bar{\pi}$ satisfying

$$v(\bar{\pi}) \geq v^* - \varepsilon(e + |v^*|)$$

need not exist.

Further it is clear from negative dynamic programming (then $u^* = 0$) that also a Markov strategy π satisfying

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

need not exist. This suggests that the statement in Theorem 1 is fairly strong.

The proof of the theorem is similar to the proof of (1.5) in Van der Wal [1981]. It will be given in Section 2.

First we introduce a few more notations. If in an expression the argument corresponding to the state is deleted, then the function on S is meant. So for example v^* is the function with i-th coordinate $v^*(i)$. Often these functions are treated as columnvectors.

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

$$(1.7) \qquad r(f)(i) = r(i, f(i)) \quad , \quad i \in S \quad ,$$

$$(1.8) \qquad P(f)(i,j) = p(i, f(i), j) \quad , \quad i,j \in S \ .$$

Further we define the operators L(f) on suitable subsets of real-valued functions on S by

$$(1.9) \qquad L(f)v = r(f) + P(f)v \quad .$$

Finally define S^- and S^+ by

$$S^- := \{i \in S \mid v^*(i) \leq 0\}$$

$$S^+ := \{i \in S \mid v^*(i) > 0\} \ .$$

2. The proof of Theorem 1

Roughly the proof goes as follows.

First choose some $\varepsilon > 0$ and define policies f_n, $n = 0,1,\ldots$ satisfying

$$(2.1) \qquad L(f_n)v^* \geq v^* - \varepsilon 2^{-(n+1)}e \ .$$

These policies constitute a Markov strategy $\hat{\pi}$, $\hat{\pi} = (f_0.f_1,\ldots)$.
Let now $\hat{\Pi}$ be the set of all strategies which on S^- act according to $\hat{\pi}$.
I.e., if at some time n the system occupies some state $i \in S^-$, then the action $f_n(i)$ has to be taken. Then, as will be shown,

$$(2.2) \qquad \sup_{\pi \in \hat{\Pi}} v(\pi) \geq v^* - \varepsilon e \ .$$

So fixing nearly conserving actions on S^- in the manner described above has not much influence on what can be gained.

Next we construct a new MDP with state space $S \times T$, $T = \{0,1,\ldots\}$ and the state dependent action sets. To be more precise, the action set in the states (i,t), $i \in S^-$, is taken to be the singleton $\{f_t(i)\}$ whereas in the states (i,t), $i \in S^+$ the action set is not restricted. Further, only transitions from states (i,t) to states $(j,t+1)$ are possible, So one might say that time is included in the state definition.

After some manipulation this newly defined MDP satisfies the conditions in Van der Wal [1981]. I.e., in each state where the value is nonpositive there is a conserving action (as the action space there is a singleton). Hence, in this model there exists a uniformly nearly-optimal stationary strategy in the sense of (1.5). This strategy corresponds to a Markov

strategy in $\hat{\Pi}$ for the original MDP and this strategy will satisfy (1.6) (for a slightly larger ε).

To start with let $\hat{\pi} = (f_0, f_1, \ldots)$ be the Markov strategy defined by (2.1), and define for $t = 1,2,\ldots$

$$\hat{\pi}^t = (f_t, f_{t+1}, \ldots) \ .$$

Further define τ to be the time of the first switch from S^- to S^+ or vice versa:

$$\tau(i_0,i_1,\ldots) = \begin{cases} \text{if } i_0 \in S^- \text{ then } \inf\{n \mid i_n \in S^+\} \\ \\ \text{if } i_0 \in S^+ \text{ then } \inf\{n \mid i_n \in S^-\} \end{cases}$$

for all $i_1, i_2, \ldots \in S$, and where $\inf \emptyset := \infty$.

Then we need the following result:

Lemma 2.1

For all $i \in S^-$

$$(2.3) \qquad \hat{v}(i \ , \hat{\pi}^t) := \mathbb{E}_{i,\hat{\pi}^t}\left[\sum_{n=0}^{\tau-1} r(X_n,A_n) + v^*(X_\tau)\right] \geq v^*(i) - \varepsilon 2^{-t}$$

Proof: Define the following MDP characterized by $\hat{S}$, $\hat{A}$, $\hat{p}$ and $\hat{r}$, with $\hat{S} = S^-$, $\hat{A} = A$ and

$$(2.4) \qquad \begin{cases} \hat{p}(i,a,j) = p(i,a,j) \quad , \quad i,j \in S^- \ , \\ \\ \hat{r}(i,a) \ = r(i,a) + \sum_{j \in S^+} p(i,a,j)\, v^*(j) \quad , \quad i \in S^- \ . \end{cases}$$

Clearly the expression at the left hand side of (2.3) is equal to the total expected reward for strategy $\hat{\pi}^t$ in the transformed MDP. Denoting all objects in the transformed MDP by a hat we have (with $\hat{v}(\hat{\pi}^t)$ defined on S^- only)

$$\hat{v}(\hat{\pi}^t) = \lim_{n\to\infty} \hat{L}(f_t)\ \hat{L}(f_{t+1})\ldots\hat{L}(f_{t+n})\ 0$$

$$\geq \liminf_{n\to\infty} \hat{L}(f_t)\ \hat{L}(f_{t+1})\ldots\hat{L}(f_{t+n})\ v^* \quad (\text{since } v^* \leq 0 \text{ on } S^-)$$

$$\geq \liminf_{n\to\infty} \hat{L}(f_t)\ldots\hat{L}(f_{t+n-1})\ v^* - \varepsilon 2^{-t-n-1} e$$

$$\geq \ldots \geq v^* - \varepsilon 2^{-t} e \ .$$

□

Next, let $\hat{\Pi}$ be the set of strategies using $\hat{\pi}$ on S^-, then

Lemma 2.2

$$\sup_{\pi\in\hat{\Pi}} v(\pi) \geq v^* - 2\varepsilon e \ .$$

Proof: The line of proof is very similar to the one in the proof of Lemma 3.3 in Van der Wal [1981].

Let $\pi^{(n)}$, $n = 1,2,\ldots$ be a strategy satisfying

$$v(i,\pi^{(n)}) \geq v^*(i) - \delta 2^{-n} \text{ for all } i \in S^+ \ .$$

Then also

$$\mathbb{E}_{i,\pi(n)}\left[\sum_{k=0}^{\tau-1} r(X_k,A_k) + v^*(X_\tau)\right] \geq v^*(i) - \delta 2^{-n} \quad .$$

Now let π^* be the strategy which on S^+ uses $\pi^{(k)}$ during the k-th stay in S^+, assuming a restart upon re-entry, and on S^- uses $\hat{\pi}$ not resetting the clock at a (re)entry time.

We will show that this strategy π^* satisfies

$$v(\pi^*) \geq v^* - 2\varepsilon e - \delta e \ .$$

Therefore define τ_n to be the time of the n-th switch from S^+ to S^- or vice versa, i.e. let $\xi = (i_0,i_1,\dots)$ be any path in S^∞ then $\tau_1(\xi) = \tau(\xi)$ and $\tau_n(\xi) = \tau(i_{\tau_{n-1}(\xi)}, i_{\tau_{n-1}(\xi)+1},\dots)$, $n \geq 2$. Then, as $\tau_n \geq n$,

$$v(\pi^*) = \lim_{n\to\infty} \mathbb{E}_{i,\pi^*} \sum_{k=0}^{\tau_n-1} r(X_k,A_k) \ .$$

Now assume $i \in S^+$, then for $n = 1,2,\dots$

$$\mathbb{E}_{i,\pi^*} \sum_{k=0}^{\tau_{2n+1}-1} r(X_k,A_k) \geq \mathbb{E}_{i,\pi^*}\left[\sum_{k=0}^{\tau_{2n+1}-1} r(X_k,A_k) + v^*(X_{\tau_{2n+1}})\right] =$$

$$= \mathbb{E}_{i,\pi^*} \sum_{k=0}^{\tau_{2n}-1} r(X_k,A_k) + \sum_{j\in S^+} \mathbb{P}_{i,\pi^*}(\tau_{2n} < \infty, X_{\tau_{2n}} = j) \cdot$$

$$\cdot\ \mathbb{E}_{j,\pi(n+1)}\left[\sum_{k=0}^{\tau-1} r(X_k,A_k) + v^*(X_\tau)\right] \geq$$

$$\geq \mathbb{E}_{i,\pi^*} \sum_{k=0}^{\tau_{2n}-1} r(X_k,A_k) + \sum_{j\in S^+} \mathbb{P}_{i,\pi^*}(\tau_{2n} < \infty,\ X_{\tau_{2n}} = j)(v^*(j) - \delta 2^{-n-1} \geq$$

$$\geq \mathbb{E}_{i,\pi^*} \sum_{k=0}^{\tau_{2n}-1} r(X_k,A_k) + v^*(X_{\tau_{2n}}) - \delta 2^{-n-1} =$$

$$= \mathbb{E}_{i,\pi^*} \sum_{k=0}^{\tau_{2n-1}-1} r(X_k,A_k) + \sum_{j,t} \mathbb{P}_{i,\pi^*}(\tau_{2n-1} = t\ ,\ X_{\tau_{2n-1}} = j) \cdot$$

$$\cdot\ \mathbb{E}_{j,\hat{\pi}^t}\left[\sum_{k=0}^{\tau-1} r(X_k,A_k) + v^*(X_\tau)\right] - \delta 2^{-n-1} \geq$$

$$\geq \mathbb{E}_{i,\pi^*} \sum_{k=0}^{\tau_{2n-1}-1} r(X_k,A_k) + \sum_{j,t} \mathbb{P}_{i,\pi^*}(\tau_{2n-1} = t, X_{\tau_{2n-1}} = j) \cdot$$

$$\cdot\ (v^*(j) - \varepsilon 2^{-t}) - \delta 2^{-n-1} \geq$$

$$\geq \mathbb{E}_{i,\pi^*}\left[\sum_{k=0}^{\tau_{2n-1}-1} r(X_k,A_k) + v^*(X_{\tau_{2n-1}})\right] - \delta 2^{-n-1} - \varepsilon 2^{-2n+1} \geq$$

$$\geq \ldots \geq \mathbb{E}_{i,\pi^*}\left[\sum_{k=0}^{\tau-1} r(X_k,A_k) + v^*(X_\tau)\right] - \delta[2^{-n-1} + 2^{-n} + \ldots + 2^{-2}] +$$

$$- \varepsilon[2^{-2n+1} + 2^{-2n+3} + \ldots + 2^{-1}] > v^*(i) - \delta - \tfrac{2}{3}\varepsilon\ .$$

Hence $v(i,\pi^*) \geq v^*(i) - \delta - \frac{2}{3}\varepsilon$ for all $i \in S^+$.

Further one may show that for all $i \in S^-$

$$v(i,\pi^*) \geq \mathbb{E}_{i,\hat{\pi}} \left[\sum_{k=0}^{\tau-1} r(X_k,A_k) + v^*(X_\tau) - \delta - \frac{2}{3}\varepsilon \right] \tag{2.5}$$

$$\geq v^*(i) - \delta - \frac{5}{3}\varepsilon > v^*(i) - \delta - 2\varepsilon \ .$$

Since $\delta > 0$ can be chosen arbitrarily the proof is complete.

□

The estimates are not as sharp as possible. E.g., in (2.5) the term $\frac{2}{3}\varepsilon$ can be replaced by $\frac{1}{3}\varepsilon$ since all actions on S^- are taken $\tau \geq 1$ units of time later. By being more careful, and not using Lemma 2.1, one may even prove that

$$\sup_{\pi \in \Pi} v(\pi) \geq v^* - \varepsilon e \ .$$

But for our purpose Lemma 2.2 is quite sufficient.

So the value of the MDP has not been affected very much by the restriction to strategies from $\hat{\Pi}$.

Now let us extend the process with a time parameter. So consider the MDP with state space $S \times T$, $T = \{0,1,\ldots\}$, action space A and further

$$\begin{cases} r((i,t),a) = r(i,a) \ , \\ p((i,t),a,(j,t+1)) = p(i,a,j) \quad , \quad i,j \in S, \ a \in A, \ t \in T \ , \\ p((i,t),a,(j,s)) = 0 \text{ if } s \neq t+1 \ . \end{cases}$$

The part of this new MDP with initial states $\{(i,0)\}$ now corresponds to the original MDP. In order that a strategy for the initial states $\{(i,0)\}$ yields a strategy in $\hat{\Pi}$ we restrict the action space in the states (i,t) with $i \in S^-$ to the singletons $\{f_t(i)\}$. Then, as one easily verifies,

(cf. Lemma 2.2)

$$v^*((i,t)) \geq v^*(i) - 2^{1-t}\varepsilon \ .$$

This newly defined MDP does not yet satisfy the condition that in each state with nonpositive value a conserving action exists, which is needed to be able to apply result (1.5). To have this condition satisfied the immediate rewards are slightly increased in the states (i,t) with $i \in S^+$. Define

$$\begin{cases} \tilde{r}((i,t),a) := r((i,t),a) + 2^{1-t}\varepsilon \ , & i \in S^+ \ , \\ \tilde{r}((i,t),a) := r((i,t),a) \ , & i \in S^- \ . \end{cases}$$

Then clearly (we use tildes for objects in the MDP with rewards $\tilde{r}$) for all $i \in S^+$

$$\tilde{v}^*((i,t)) \geq v^*((i,t)) + 2^{1-t}\varepsilon \geq v^*(i) > 0 \ .$$

So, if $\tilde{v}((i,t)) \leq 0$ then $i \in S^-$ and thus the action set in state (i,t) is a singleton whence this action is also conserving. Thus the result in Van der Wal [1981] applies, stating the existence of a stationary strategy, g say, satisfying

$$(2.6) \qquad \tilde{v}(g) \geq \tilde{v}^* - \delta\tilde{u}^* \ ,$$

where $\delta > 0$ is some arbitrarily chosen constant. This stationary strategy g corresponds to a Markov strategy $\pi_g \in \hat{\Pi}$ for the original MDP, namely the strategy $\pi_g = (g_0, g_1, \ldots)$ with $g_t(i) = g((i,t))$.

As we will show π_g satisfies

$$v(\pi_g) \geq v^* - \varepsilon_0(e + u^*)$$

where ε_0 depends on ε and δ.

Therefore observe that

$$(2.7) \qquad \tilde{v}((i,0),g) \leq v(i,\pi_g) + \sum_{t=0}^{\infty} 2^{1-t}\varepsilon = v(i,\pi_g) + 4\varepsilon$$

and that similarly

$$(2.8) \qquad \tilde{u}^* \leq u^* + 4\varepsilon e \ .$$

Also, by Lemma 2.2,

$$(2.9) \qquad \tilde{v}^*((i,0)) \geq v^*((i,0)) \geq v^*(i) - 2\varepsilon \ .$$

So by subsequently using (2.7), (2.6), (2.9) and (2.8) we obtain for all $i \in S$

$$v(i,\pi_g) \geq \tilde{v}((i,0),g) - 4\varepsilon \geq \tilde{v}^*((i,0)) - \delta\tilde{u}^*((i,0)) - 4\varepsilon$$

$$\geq v^*(i) - 2\varepsilon - \delta u^*(i) - 4\delta\varepsilon - 4\varepsilon \geq v^*(i) - \varepsilon_0(1 + u^*(i)) \ ,$$

with $\varepsilon_0 = \max\{6\varepsilon + 4\delta\varepsilon, \delta\}$.

Clearly ε_0 can be made arbitrarily small by a suitable choice of ε and δ so the proof of Theorem 1 is complete.

References

Hee, K.M. van (1978), Markov strategies in dynamic programming, Math.Oper. Res.3, 37 - 41.

Hee, K.M. van, A. Hordijk and J. van der Wal (1977). Successive approximations for convergent dynamic programming, in Markov decision theory, eds. H.C. Tijms and J. Wessels, Mathematical Centre Tract 93, Mathematisch Centrum, Amsterdam, 183-211.

Wal, J. van der (1981), On uniformly nearly-optimal stationary strategies, Eindhoven University of Technology, Dept. of Maths., Memorandum COSOR 81-11.