The action elimination algorithm for Markov decision processes

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

PROBABILITY THEORY, STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 76-20

The action elimination algorithm for Markov decision processes

by

N.A.J. Hastings and J.A.E.E. van Nunen

Eindhoven, November 1976

The Netherlands

The action elimination algorithm for Markov decision processes

by

N.A.J. Hastings* and J.A.E.E. van Nunen**

Abstract

An efficient algorithm for solving Markov decision problems is proposed. The value iteration method of dynamic programming is used in conjunction with a test for nonoptimal actions. The algorithm applies to problems with undiscounted or discounted returns with infinite or finite planning horizon. In the finite horizon case the discount factor may exceed unity. The nonoptimality test, which is an extension of Hastings test for the undiscounted reward case, is used to identify actions which cannot be optimal at the current stage. As convergence proceeds the proportion of such actions increases producing major computational savings. For problems with discount factor less than one the test is shown to be tighter than that of MacQueen.

1. Introduction

We consider a finite Markov decision chain with or without discounting. The state space is S, where the states are labeled $i = 1,2,\ldots,N$. If the system is in state $i \in S$ at time n an action k has to be selected from a nonempty finite set K_i. As a consequence of this action $k \in K_i$ we earn a(n) (expected) reward r(i k) and the system moves to state $j \in S$ at time $n + 1$ with probability $p(i,j,k)$. We assume $\sum_j p(i,j,k) = 1$.

The Cartesian product of all sets K_i is the policy space Δ. For any policy $\delta \in \Delta$ we denote by $P(\delta)$ the transition probability matrix and by $r(\delta)$ the column vector of rewards. Rewards earned in the n-th period are discounted by a factor $\beta > 0$ (eventually $\beta \geq 1$). Our goal is to find a strategy that maximizes the total expected reward over a time horizon $T \in \mathbb{N} \cup \{\infty\}$, and to determine the corresponding optimal reward vector v_T. Here, a strategy π_T for a T-horizon problem is a sequence of policies

* Monash University, Melbourne, Australia (from 1 July 1977)

** Graduate School of Management, Delft, The Netherlands.

$\pi_T := (\delta_1, \delta_2, \ldots, \delta_T)$. Note that we restrict the considerations, as is allowed, to nonrandomized strategies. For $T = \infty$ it is even permitted to consider only stationary strategies i.e. $\pi := \pi_\infty := (\delta, \delta, \delta, \ldots)$. The optimal value vector v_T^* can be computed by the value iteration algorithm of dynamic programming. For finite horizon problems we refer to Hinderer [4] and Hübner [6]. For $T = \infty$ we refer to e.g. Hastings [1] or Van Nunen [9]. In the latter situation dynamic programming yields in the limit policies which can be used to constitute stationary strategies that are optimal.

As indicated in e.g. [1], [4], [6], [10] convergence is monitored by using upper and lowerbounds on the optimal return vector v_T^*. These bounds are used to construct sub-optimality tests, see for example references [8], [3], [2], [10]. The test proposed here increases the efficiency of the dynamic programming method considerably a nonoptimal action for a given stage (iteration) is one which does not form part of an optimal policy for that stage. Until now, in the discounted case, tests have been devised whereby only those actions which can be identified as being nonoptimal for all subsequent stages are eliminated. For the average reward situation Hastings [3] proposed to eliminate actions for one or more stages after which they may reenter the state space. Here we extend this idea to Markov decision processes which may be undiscounted or discounted, may have a finite or infinite time horizon and in the finite horizon case may have a discount factor that is allowed to be greater than one.

2. The test

Let $f(n,i)$ be the maximum total expected return generated when the system starts in state $i \in S$ and continues for n-stages. Then

$$(1) \qquad f(n,i) := \max_{k \in K_i} \left[r(i,k) + \beta \sum_{j \in S} p(i,j,k) f(n-1,j) \right]$$

where $f(0)$ is given and $\beta > 0$. The value iteration algorithm computes $f(n,i)$ for $i \in S$ and $n = 1,2,\ldots,T$.

Define

$$(2) \qquad f(n,i,k) := r(i,k) + \beta \sum_{j \in S} p(i,j,k) f(n-1,j) \ ;$$

$$(3) \qquad y(n,i,k) := f(n,i) - f(n,i,k) \geq 0 \ ;$$

$$(4) \qquad \theta_u(n) := \max_{i \in S} [f(n,i) - f(n-1,i)] \ ;$$

$$(5) \qquad \theta_\ell(n) := \min_{i \in S} [f(n,i) - f(n-1,i)] \ ;$$

$$(6) \qquad \varphi(n) := \beta(\theta_u(n) - \theta_\ell(n)) \ ;$$

$$(7) \qquad H(m,n,i,k) := y(n,i,k) - \sum_{\ell=n}^{m-1} \varphi(\ell) \qquad m > n \ .$$

Note that

$$(8) \qquad H(m+1,n,i,k) \leq H(m,n,i,k) \ .$$

In the test we will use, any action $k \in K_i$ is nonoptimal for state $i \in S$ at value iteration stage m if

$$H(m,n,i,k) > 0 \ .$$

3. Basic properties

Lemma 1

a) $\varphi(m) \leq \beta\varphi(m-1)$

b) $f(n+1,i,k) - f(n,i,k) \leq \beta\theta_u(n)$

c) $f(n+1,i) - f(n,i) \geq \beta\theta_\ell(n)$

d) $y(m,i,k) \geq H(n,i,k) \qquad$ for $m > n$

e) $H(m,n,i,k) \geq y(n,i,k) - \frac{1-\beta^{m-n}}{1-\beta} \varphi(n) \ , \qquad m > n$

Proof. Part a is a direct consequence of Hübner [6] theorem The second part of the lemma follows from

$$f(n+1,i,k) - f(n,i,k) = r(i,k) + \beta \sum_{j\in S} p(i,j,k) f(n,j) - r(i,k) - \beta \sum_{j\in S} p(i,j,k) f(n-1,j) .$$

$$= \beta \sum_{j\in S} p(i,j,k)[f(n,j) - f(n-1,j)]$$

$$\leq \beta\theta_u(n)$$

consider

$$f(n+1,i) - f(n,i) \geq r(i,k_0) + \beta \sum_{j\in S} p(i,j,k_0) f(n,j) - r(i,k_0) - \beta \sum_{j\in S} p(i,j,k_0) f(n-1,j)$$

$$= \beta \sum_{j\in S} p(i,j,k_0)[f(n,j) - f(n-1,j)] \geq \beta\theta_\ell(n) ,$$

with k_0 that action in K_i for which the maximum in $f(n,i)$ is attained. This proves part e).
Since

$$y(m,i,k) = f(m,i) - f(m,i,k) \geq f(m-1,i) + \beta\theta_\ell(m-1) - f(m-1,i,k) - \beta\theta_u(m-1) = y(m-1,i,k) - \varphi(m-1)$$

the result d) follows by iterating stagewise.
The final statement of the lemma is a direct consequence of part a of this lemma and the definition of $H(m,n,i,k)$. □

Theorem 1

a) Action k at state i is nonoptimal at stage $m > n$ if $H(m,n,i,k) > 0$.

b) Action k at state i is nonoptimal at stage m > n if

$$y(n,i,k) - \frac{1-\beta^{m-n}}{1-\beta}\varphi(n) > 0$$

c) Action k at state i is nonoptimal for all subsequent stages if

$$y(n,i,k) - \frac{1-\beta^{T-n}}{1-\beta}\varphi(n) > 0 , \qquad T > n .$$

Proof. The proof follows from the foregoing lemma. Part b) and c) can also be found in Hübner [6]. □

Remark. For $\beta = 1$ the term $\frac{1-\beta^{m-n}}{1-\beta}$ has to be replaced by $(m - n)$. For $T = \infty$ the theorem makes sense only if $\beta < 1$. However, the condition can be weakened see Hübner [6] or Porteus [11].

Since in our test actions are eliminated which are nonoptimal for perhaps only one stage, it will be clear that the first stage at which our test eliminates an action for the first time will in general be much earlier than the first stage at which e.g. the MacQueen test [8] or the Hastings and Mello test [2] eliminates that action.
This follows directly from the foregoing theorem.

Corollary 1. For $0 < \beta < 1$ our test is tighter than MacQueen's test and the Hastings and Mello test for eliminating optimal actions.

Proof. MacQueen based his test on part c) of theorem 1, with $T = \infty$.
So in his test an action k is nonoptimal in state i if

$$y(n,i,k) - \frac{1}{1-\beta}\varphi(n) > 0 .$$

In our test an action is eliminated for the first time if $y(n,i,k) > \varphi(n)$.
Clearly

$$\varphi(n) < \frac{\varphi(n)}{1-\beta} \qquad \text{for } 0 < \beta < 1 .$$

Since the MacQueen test is tighter than the Hastings and Mello test the corollary is proved.

Remark. Note that for $\beta \to 1$ the relative power of our test will be greater since $(1 - \beta)^{-1} \to \infty$ as $\beta \to 1$.

4. Computational method

To illustrate the computational method we give a flow chart of the test. Before drawing such a flow chart we have to give some more preliminaries. We assume the terminal values $f(0) = 0$ and apply the test from stage two onwards. We set the test quantity $T(n,i,k)$ at zero at stage 1.
An action fails the test if its test quantity (called flag) is positive or if its flag is "nonoptimal". If the action fails at stage n its trial value $f(n,i,k)$ is then not evaluated at that stage. For an action which passed the test at stage n, the flag $T(n,i,k)$ could be reset to

$$T(n,i,k) := \begin{cases} \text{"nonoptimal" if } y(n,i,k) - \dfrac{1 - \beta^{T-n}}{1 - \beta}\varphi(n) > 0 , \\ y(n,i,k) \quad \text{else} . \end{cases}$$

For an action which fails the test at stage $n - 1$, the flag $T(n,i,k)$ is given by

$$T(n,i,k) := \begin{cases} \text{"nonoptimal" if } T(n-1,i,k) = \text{"nonoptimal"} , \\ T(n - 1,i,k) - \varphi(n - 1) . \end{cases}$$

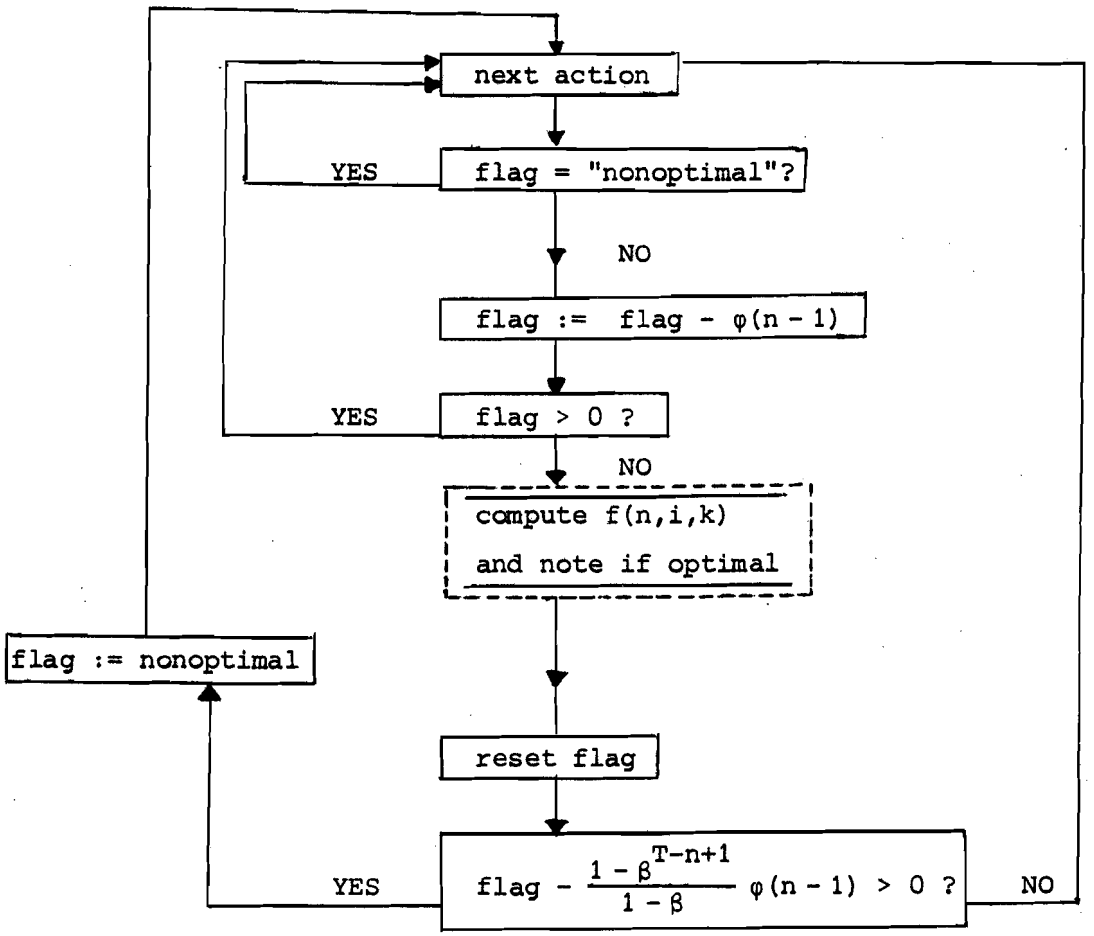
However as in [3], to avoid the making of a second pass it is preferable to use by resetting the "flag" after an action passed the test

$$f(n - 1,i,k) + \beta\theta_{\ell}(n - 1) - f(n,i,k)$$

instead of

$$y(n,i,k) := f(n,i) - f(n,i,k) .$$

The effect of the test is to reduce the number of times that the time consuming step of evaluating $f(n,i,k)$ is carried out. (This step is marked by a dotted line).
The flow chart of the action elimination algorithm has the following structure.



5. Numerical example*

The extreme efficiency of the test will be shown by applying it to Howard's automobile replacement problem [5 p.p. 54-59] with discountfactor $\beta = 0.97$. We use the dynamic programming algorithm of MacQueen [7]. We compare the number of actions eliminated by the Hastings and Mello test [2] with the number of eliminated actions by the test proposed in this paper. In the first test only actions which are non-optimal for the whole future are eliminated. We start the dynamic programming algorithm with a starting vector with all components equal to zero i.e. $f(0,i) = 0$ for all $i \in S$.

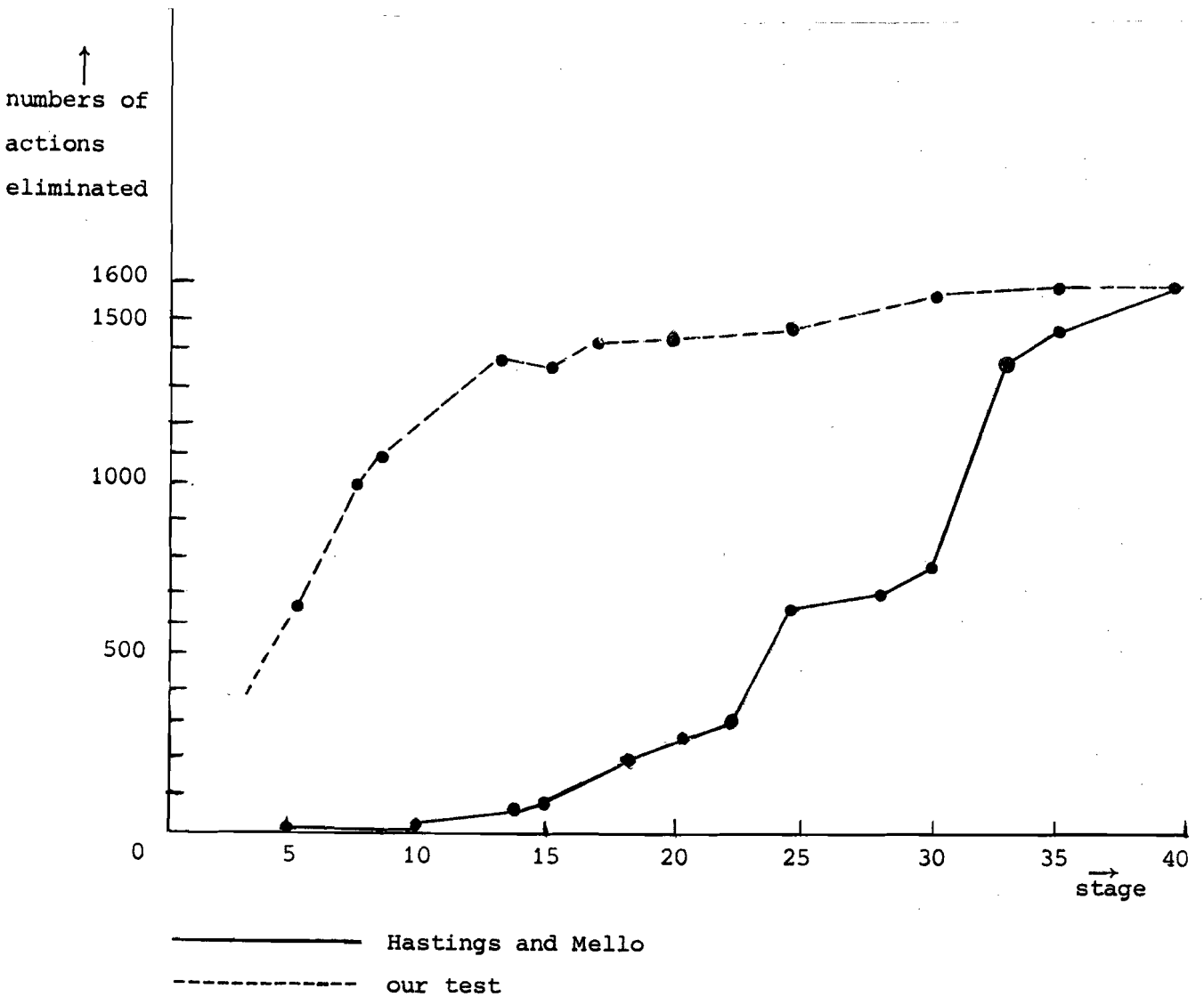
In figure 1 we see that the difference between the number of actions that are eliminated is significant. From iteration 8 until iteration 22 this difference is even over 1000 actions.

---

* The authors are grateful to mr. K. van der Hoeven for computational support.

figure 1

Application to the automobile Replacement problem



7. Some extensions and remarks

In this note we have assumed the equal row sum property. However, the same ideas can be used for a nonoptimality test, in the case that this assumption is released. We then have to exploit more sophisticated bounds for the values f(n + m,i). These bounds are described for example in Porteus [11] or Van Nunen [10].

In [9] and [10] a whole set of successive approximation algorithms for

Markov decision problems containing the Jacobi-, the Gauss Seidel- and overrelaxation algorithms is developed. The nonoptimality test can be incorporated in those algorithms as well.
It is known see e.g. Hübner [6], Porteus [11] that the contraction factor is sometimes even smaller than the discount factor β. In that case the nonoptimality test can be refined by using the more sophisticated contraction factor.
For infinite horizon problems (in the equal row sum case) with respect to the total reward criterion convergence of $f(n,i)$ is only guaranteed if $\beta < 1$. However, for finite horizon problems β is allowed to be greater then or equal to one. If the equal row sum property is not satisfied, convergence of the total expected reward may occur for $\beta \geq 1$, see Porteus [11] or Van Nunen [10].

References

[1] Hastings, N.A.J., "Dynamic Programming with Management Applications", Butterworths, London and Crane-Russak, New York, 1973.

[2] Hastings, N.A.J. and Mello, J.M.C., "Test for suboptimal actions in discounted Markov programming", Management Sci. 19, 1973, pp. 1019-1022.

[3] Hastings, N.A.J., "A test for nonoptimal actions in undiscounted finite Markov decision chains", Management Sci. 23, 1976, pp.

[4] Hinderer, K., "Estimates for finite state dynamic programs", (preprint 1974) to appear in J. Math. Anal. Appl.

[5] Howard, R.A., "Dynamic programming and Markov processes", Wiley, New York-London, 1960.

[6] Hübner , G., "Improved procedures for eliminating suboptimal actions in Markov programming by the use of contraction properties", to appear in: Transactions of the seventh Prague Conf. 1974.

[7] MacQueen, J., "A modified dynamic programming method for Markovian decision problems", J. Math. Anal. Appl. 14, 1966, pp. 38-43.

[8] MacQueen, J., "A test for suboptimal actions in Markovian decision problems", Oper. Res. 15, 1967, pp. 559-561.

[9] Van Nunen, J.A.E.E. and Wessels, J., "A principle for generating optimization procedures for discounted Markov decision processes". In: "Progress in Operations Research" ed. by A. Prékopa. North-Holland Publ. Company 1976. pp. 683-695.

[10] Van Nunen, J.A.E.E., "Contracting Markov Decision Processes", Math. Centre Tract no. 71, Math. Centre, Amsterdam, 1976.

[11] Porteus, E.L., "Bounds and transformations for discounted finite Markov decision chains". Oper. Res. 23, 1975, pp. 761-784.