# Discounted semi-Markov decision processes : linear programming and policy iteration

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

STATISTICS AND OPERATION RESEARCH GROUP

Memorandum COSOR 74-01

# Discounted semi-Markov decision processes: linear programming and policy iteration

by

J. Wessels

and

J.A.E.E. van Nunen

Eindhoven, January 1974

Abstract.

For semi-Markov decision processes with discounted rewards we derive the well known results regarding the structure of optimal strategies (nonrandomized, stationary Markov strategies) and the standard algorithms (linear programming, policy iteration). Our analysis is completely based on a primal linear programming formulation of the problem.

## § 1. Introduction.

In this memorandum discounted semi-Markov problems as discussed e.g. by Jewell [4], and De Ghellinck and Eppen [3], will be treated.
We consider (semi-) Markov decision processes with a finite set of states, $S := \{1,2,\ldots,N\}$. For any $i \in S$ a finite set $K(i)$ of allowed decisions is available. If $k \in K(i)$ has been chosen the probability for finding the system in state $j$ at the next decision point is equal to $p_{ij}^k$ ($p_{ij}^k \geq 0$, $\sum_{j=1}^{N} p_{ij}^k \leq 1$). If this occurs, the interdecision time has a probability distribution function $F_{ij}^k$, $t = 0$ is a decision moment. At the decision moment a(n) (expected) reward $r^k(i)$ is earned. The Markov model appears when $F_{ij}^k$ has the form:

$$F_{ij}^k(t) = \begin{cases} 0 & \text{for } t < 1 \\ 1 & \text{for } t \geq 1 \end{cases}.$$

The goal is to maximize the total discounted expected reward over an infinite time horizon.
For these problems it is possible to give a linear programming formulation (e.g. see d'Epenoux [2], De Ghellinck and Eppen [3], Howard [5]).
Also Howard's policy iteration algorithm (see e.g. [4], [5]) can be used to find optimal solutions for this type of problems. The relationship between linear programming and policy iteration is well known. Mine and Osaki [7] discussed this relation for (semi-) Markov decision processes with and without discounting. Here we derived the known results concerning the structure of optimal strategies (nonrandomized Markov or memoryless strategies) and the relation between the standard algorithms (linear programming and policy iteration) completely form a primal linear programming formulation of the problem.

§ 2. Semi-Markov decision processes with discounting.

Let the initial state or initial distribution $\{\pi_j(0)\}$ $(\pi_j(0) \geq 0, \sum_{j\in S} \pi_j(0) = 1)$ be given. Then an arbitrary decision rule determines the stochastic process. This decision rule may be eventually randomized and non Markov, hence basing decisions on the complete past of the process. For a given decision rule let $\pi_i^k(n,t)$, for the n-th decision point, be the joint probability that state i is observed, that decision k is made, and that this n-th decision point occurs not later than time t.
For $t \geq 0$, $j \in S$, the $\pi_j^k(n,t)$ will satisfy the following recurrence relation:

$$(1) \qquad \sum_{k\in K(j)} \pi_j^k(n,t) = \begin{cases} \pi_j(0) & \text{if } n = 0 \\ \sum_{i\in S} \sum_{\ell\in K(i)} p_{ij}^{\ell} \int_0^t \pi_i^{\ell}(n-1, t-\tau)\, dF_{ij}^{\ell}(\tau) , & \\ & n = 1,2,\dots . \end{cases}$$

We assume that $F_{ij}^k(0) < 1$ for all $i,j \in S$ and for all $k \in K(i)$. This assumption guarantees that the expected number of transitions in a finite interval $(0,T)$ is finite. Now we can state the following lemma.

Lemma. For any decision rule converges the total expected discounted reward (using a discount rate $\beta > 0$)

$$(2) \qquad \sum_{n=0}^{\infty} \{ \sum_{j\in S} \sum_{k\in K(j)} r^k(j) \int_0^{\infty} e^{-\beta t}\, d\,\pi_j^k(n,t)\}$$

absolutely, and the sum is uniformly bounded by $\pm \frac{r^*}{1-\delta}$ with

$$r^* := \max_{j,k} |r^k(j)| ,$$

$$\delta := \max_{i,j,k} \int_0^{\infty} e^{-\beta t}\, d\, F_{ij}^k(t) .$$

Proof.

$$\sum_{n=0}^{\infty} \{ \sum_{j\in S} \sum_{k\in K(j)} |r^k(j)| \int_0^{\infty} e^{-\beta t} \, d\pi_j^k(n,t)\} \leq$$

$$\leq \sum_{n=0} r^*\{ \sum_{j\in S} \sum_{k\in K(j)} \int_0^{\infty} e^{-\beta t} \, d\pi_j^k(n,t)\} \, .$$

Consider:

$$\sum_{j\in S} \sum_{k\in K(j)} \int_0^{\infty} e^{-\beta t} \, d\pi_j^k(n,t) = \sum_{j\in S} \int_0^{\infty} e^{-\beta t} \, d \Big( \sum_{k\in K(j)} \pi_j^k(n,t)\Big) =$$

$$= \sum_{j\in S} \int_0^{\infty} e^{-\beta t} \, d\Big( \sum_{i\in S} \sum_{\ell\in K(i)} p_{ij}^{\ell} \int_0^{t} \pi_i^{\ell}(n-1\, , t-\tau) d\, F_{ij}^{\ell}(\tau)\Big) =$$

$$= \sum_{j\in S} \sum_{i\in S} \sum_{\ell\in K(i)} p_{ij}^{\ell} \int_0^{\infty} e^{-\beta t} \, d\, F_{ij}^{\ell}(t) \int_0^{\infty} e^{-\beta t} \, d\, \pi_i^{\ell}(n-1,t) \leq$$

$$\leq \delta \sum_{i\in S} \sum_{\ell\in K(i)} \int_0^{\infty} e^{-\beta t} \, d\, \pi_i^{\ell}(n-1,t) \leq \delta^n \, .$$

$0 \leq \delta < 1$, as a consequence of the assumption $F_{ij}^k(0) < 1$, $i,j \in S$, $k \in K(i)$. So

$$\sum_{n=0}^{\infty} \{ \sum_{j\in S} \sum_{k\in K(j)} |r^k(j)| \int_0^{\infty} e^{-\beta t} \, d\pi_j^k(n,t)\} \leq \frac{r^*}{1-\delta} \, . \qquad \square$$

The problem is to determine a decision rule for which (2) is maximal. Where, as a consequence of the absolute convergence of (2), it is also possible to write (2) as

$$(3) \qquad \sum_{j\in S} \sum_{k\in K(j)} r^k(j) \sum_{n=0}^{\infty} \int_0^{\infty} e^{-\beta t}\, d\pi_j^k(n,t) \ .$$

§ 3. Linear programs and the structure of optimal strategies.

(3) depends only on the decision rule through

$$x_j^k := \sum_{n=0}^{\infty} \int_0^{\infty} e^{-\beta t}\, d\pi_j^k(n,t)$$

Hence, with

$$(4) \qquad \pi_j^k(n) := \int_0^{\infty} e^{-\beta t}\, d\pi_j^k(n,t) \ ,$$

the problem can also be formulated as:
Determine the decision rule for which

$$(5) \qquad \sum_{j\in S} \sum_{k\in K(j)} r^k(j) \sum_{n=0}^{\infty} \pi_j^k(n)$$

is maximal.
As a consequence of (1) the $\pi_j^k(n)$ will satisfy the recurrence relation:

$$(6) \qquad \sum_{k\in K(j)} \pi_j^k(n) = \begin{cases} \pi_j(0) \ , & n = 0 \\ \sum_{i\in S} \sum_{\ell\in K(i)} p_{ij}^{\ell}\, \eta_{ij}^{\ell}\, \pi_i^{\ell}(n-1) \ , & n = 1,2,\ldots \end{cases}$$

where

$$\eta_{ij}^{\ell} := \int_0^{\infty} e^{-\beta t}\, dF_{ij}^{\ell}(t) \qquad \text{(hence } 0 \le \eta_{ij}^{\ell} < 1) \ .$$

Lemma. Every nonnegative solution $\{\pi_j^k(n)\}$ of (6) may be considered as the transforms (4) of the $\pi_j^k(n,t)$ corresponding with a Markov decision rule.

Proof. A Markov decision rule which satisfies the requirements is constructed in the following way:
Select at the n-th decision point with probability $d_j^k(n)$, decision $k \in K(j)$ if at this point state j has been observed, with

$$d_j^k(n) := \frac{\pi_j^k(n)}{\sum_{k \in K(j)} \pi_j^k(n)} .$$

It is easy to verify that the $\{\pi_j^k(n,t)\}$ related with this decision rule have transforms $\{\pi_j^k(n)\}$. □

As a consequence of this lemma it is permitted to consider only Markov strategies.
Furthermore, the lemma legitimates to formulate the problem as follows:

Maximize (5), subject to (6) and the nonnegativity constraints $\pi_j^k(n) \geq 0$.

This problem is a linear programming problem with an infinite number of constraints and variables.

As a second step we define:

$$x_j^k := \sum_{n=0}^{\infty} \pi_j^k(n) .$$

Then we have the transformed problem I (see also e.g. [3], [7]):

$$
\text{I} \quad \begin{cases} \max \sum_{j \in S} \sum_{k \in K(j)} r^k(j) x_j^k \\ \text{subject to} \\ \sum_{k \in K(j)} x_j^k = \pi_j(0) + \sum_{i \in S} \sum_{\ell \in K(i)} q_{ij}^{\ell} \, x_i^{\ell} , \quad j \in S \\ x_j^k \geq 0 , \quad j \in S, \; k \in K(j) \end{cases}
$$

where

$$
q_{ij}^{\ell} := p_{ij}^{\ell} \, \eta_{ij}^{\ell} .
$$

Now, problem I is a standard linear programming problem.

Lemma. If $\pi_j(0) > 0$, $j \in S$, then there exists a one to one correspondence between basic feasible solutions of I and nonrandomized stationary Markov strategies (see [3], [7]).

To prove this lemma we remark that $\pi_j(0) > 0$ implies $\sum_{k \in K(j)} x_j^k > 0$. Hence for any basic feasible solution there is for each $j \in S$ exactly one $k(j) \in K(j)$ with $x_j^k > 0$ and $x_j^k = 0$ for $k \neq k(j)$. Conversely, given a nonrandomized stationary Markov strategy denoted by $f := (k(1),k(2),\ldots,k(N))$, the system of equations:

$$
x_j^{k(j)} - \sum_{i \in S} q_{ij}^{k(i)} \cdot x_i^{k(i)} = \pi_j(0) , \quad j \in S ,
$$

has a unique solution $\{x_j^{k(j)}\}$ with $x_j^{k(j)} > 0$.
This follows from the system's diagonal dominance ($0 \leq \eta_{ij}^{\ell} < 1$) or from the fact that $Q^n \to 0$ as $n \to \infty$, where Q is an $N \times N$ matrix with elements $q_{ij}^{k(i)}$ □

Furthermore, it is permitted to take $\pi_j(0) = \frac{1}{N}$, in the linear programming problem I, since an optimal solution of the linear programming problem I remains optimal if the $\pi_j(0)$ in the right hand side are changed (see Gass [8]).

Theorem. In order to find an optimal decision rule it is permitted to restrict the attention to nonrandomized stationary Markov strategies. This optimal decision rule can be found by solving the linear programming problem I, with arbitrary $\{\pi_j(0)\}$, $\pi_j(0) > 0$.

This theorem follows from the fact that for any decision rule the transformes of the corresponding $\pi_j^k(n,t)$ yield a feasible solution of I, while conversely an optimal basic solution of I corresponds to a nonrandomized stationary Markov strategy. The fact that a restriction to nonrandomized stationary Markov strategies is permitted is also proved in another way by Denardo [1].

Remark. The total expected discounted reward, if the process starts in state i and the optimal strategy $f^*$ is used ($v_i(f^*)$), can be found by solving the dual problem.

## § 4. Policy iteration.

It is easy to find a basic feasible solution: select for each $j \in S$ one $k(j) \in K(j)$, then $\{x_j^{k(j)}\}_{j\in S}$ form a basic solution and the corresponding nonrandomized stationary Markov strategy is $f = \{k(1),k(2),\ldots,k(N)\}$. Whether this basis yields an optimal solution or not may be checked by constructing the price vector, as usual in linear programming (see Gass [8]). Thus the problem is to find a linear combination of the N equality constraints of I and the reward equation, such that the elements corresponding to the $x_j^{k(j)}$ equal 0, i.e. look for $v_j(f)$, $j \in S$, with

$$(8) \qquad v_j(f) - \sum_{i\in S} q_{ji}^{k(j)}\, v_i(f) - r^{k(j)}(j) = 0 \ , \quad j \in S \ .$$

From (8) we see that $v_j(f)$ is the total expected discounted reward if the system's initial state is j and strategy f ($j \to k(j)$) is followed (dynamic programming equations).
For the other elements of the reward equation we have:

$$v_j(f) - \sum_{i\in S} q_{ji}^{k}\, v_i(f) - r^{k}(j) \ , \quad j \in S,\ k \in K(j) \ .$$

If for all $j \in S$ and $k \in K(j)$

$$v_j(f) - \sum_{i\in S} q_{ji}^{k}\, v_i(f) - r^{k}(j) \geq 0 \ ,$$

then strategy f ($j \to k(j)$) is optimal.
For the object value W we then have:

$$W := \sum_{j \in S} \pi_j(0) \cdot v_j(f) \ .$$

If for some $k \in K(j)$ and $j \in S$

$$v_j(f) - \sum_{i \in S} q_{ji}^k \, v_i(f) - r_i^k < 0 \ , \tag{9}$$

then a better solution is possible by selecting for each $j \in S$ one $k(j) \in K(j)$ for which (9) holds. When for some $j \in S$ such a choice is not available, the old $k(j) \in K(j)$ for which (8) holds is selected again. This yields a new and better basic feasible solution.
That the new basic feasible solution and so the corresponding strategy $g \in K := \{K(1) \times K(2) \times \ldots \times K(N)\}$ is better than the old one $f \in K$ can be shown as follows:
Let

$$v(f) = r(f) + Q(f)v(f)$$

$$v(g) = r(g) + Q(g)v(g)$$

and

$$\gamma(g,f) = r(g) + Q(g)v(f) - r(f) - Q(f)v(f) \ ,$$

where for simplicity a vector notation is used. Now from (8) en (9) it will be clear that $\gamma(g,f) \geq 0$

$$\Delta v := v(g) - v(f) = \gamma(g,f) + Q(g)v(g) - Q(g)v(f) =$$

$$= \gamma(g,f) + Q(g)\Delta v \ .$$

So,

$$\Delta v = (I - Q(g))^{-1} \gamma(g,f) \geq 0 \quad \text{and} \neq 0 \ .$$

This implies that, when this improvement procedure is applied, an old strategy will never appear again as long as real improvement is possible. This leads to the following algorithm:

i) select a nonrandomized strategy;
ii) solve for this strategy the set of equations (8);
iii) if possible select a better strategy based on (9) and goto ii); if such a strategy is not available, stop:

Consequently such an algorithm will converge, in a finite number of iterations, to an optimal strategy $f^*$ with object value:

$$W := \sum_{j \in S} \pi_j(0) \cdot v_j(f^*) .$$

This means that Howard's policy iteration method, that can be derived with dynamic programming (see [6], [4]), follows straightforward from the primal linear programming formulation, using the possibility of changing more basic variables in each iteration step (see also [7]).

References.

[1] E.V. Denardo, "Contraction mappings in the theory underlying dynamic programming", Siam Review 9 (1967), 165-177.

[2] F. d'Epenoux, "Sur un problème de production et de stockage dans l'Aléatoire", Rev. Franç. Rech. Opérationnelle 14 (1960), 3-16.

[3] G.T. de Ghellinck and G.D. Eppen, "Linear programming solutions for separable Markovian decision problems", Management Sci. 13 (1967), 371-394.

[4] R.A. Howard, "Dynamic programming and Markov processes", M.I.T. Press, Cambridge, Massachusetts (1960).

[5] ----------, "Dynamic probabilistic systems", Volume II, John Wiley & Sons, Inc., New York (1971).

[6] W.S. Jewell, "Markov-renewal programming: I, Formulation finite return models", Operations Res. 11 (1963), 938-948.

[7] H. Mine and S. Osaki, "Markovian decision processes", American Elsevier, New York (1970).

[8] S.I. Gass, "Linear programming methods and applications", McGraw-Hill Book Company, New York (1958).