The method of successive approximations for the discounted Markov game

TECHNOLOGICAL UNIVERSITY EINDHOVEN

Department of Mathematics

STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 75-02

The method of successive approximations for the discounted Markov game

by

J. van der Wal

Eindhoven, March 1975

The method of successive approximations for the discounted Markov game

by

J. van der Wal

Abstract. This paper presents a number of successive approximation algorithms for the repeated two-person zero-sum game called Markov game using the criterion of total expected discounted rewards. As Wessels [12] did for Markov decision processes stopping times are introduced in order to simplify the proofs. It is shown that each algorithm provides upper and lower bounds for the value of the game and nearly optimal stationary strategies for both players.

1. Introduction and notations

We are concerned with a dynamic system with a finite state space $S := \{1,\ldots,N\}$. The behaviour of the system is influenced by two players, P_1 and P_2, having opposite aims. For each $x \in S$ two finite nonempty sets of actions exist, one for each player, denoted by K_x for P_1 and L_x for P_2.

At times $t = 0,1,2,\ldots$ both players select an action out of the set available to them. As a joint result of the state x of the system and the two selected actions, k for P_1 and ℓ for P_2, the system moves to a new state y with probability $p(y|x,k,\ell)$, $\sum_{y\in S} p(y|x,k,\ell) = 1$, and P_1 will receive some (possibly negative) expected amount from P_2 denoted by $r(x,k,\ell)$.

As Zachrisson [15] did, we will call these two-person zero-sum games Markov games. Most authors however, following Shapley [10] use the term stochastic games.

A strategy d for $P_1(P_2)$ in this game is any function that specifies for each time $t = 0,1,2,\ldots$ and for each state $x \in S$ the probability $d(a|x,n,h_n)$ that action $a \in K_x(L_x)$ will be taken as a function of x,n and the history h_n. By the history h_n upto time n we mean the sequence $h_n = (x_0,k_0,\ell_0,\ldots,x_{n-1},k_{n-1},\ell_{n-1})$ of prior states and actions (h_0 is the empty sequence).

If all $d(a|x,n,h_n)$ are independent of n and h_n the strategy is called stationary. A policy f(g) for $P_1(P_2)$ will be defined as any function such that f(x)(g(x)) is a probability distribution on $K_x(L_x)$ for all $x \in S$. Thus a stationary strategy prescribes the same policy for each time t and we will denote it by $f^{(\infty)}(g^{(\infty)})$. We will use the letters π and ρ to denote a strategy for P_1 and P_2 respectively. In the following the symbols k, f and π will be used for P_1 and the symbols ℓ, g and ρ for P_2 only. We will consider the discounted Markov game, i.e. we will discount future income at a rate β, with $0 \leq \beta < 1$.

Let $V(\pi,\rho)$ denote the N-columnvector with x-th component equal to the total expected discounted reward for P_1 when the game starts in state x, P_1 plays strategy π and P_2 plays ρ.

Shapley [10] has shown that this game has a value, denoted by the N-column vector v_β and that both players have stationary optimal strategies, denoted by $f^{*(\infty)}$ and $g^{*(\infty)}$, i.e. Shapley has shown that

$$\inf_{\rho} V(f^{*(\infty)},\rho) = V(f^{*(\infty)},g^{*(\infty)}) = v_\beta = \sup_{\pi} V(\pi,g^{*(\infty)}) .$$

Let e denote the N-columnvector with all elements equal to unity: $e = (1,\ldots,1)^T$. A strategy π_ε for P_1 (ρ_ε for P_2) will be called ε-optimal if $V(\pi_\varepsilon,\rho) \geq V_\beta - \varepsilon.e$ for all ρ ($V(\pi,\rho_\varepsilon) \leq V_\beta + \varepsilon.e$ for all π), $\varepsilon \geq 0$. An 0-optimal strategy is called optimal.

We are looking for techniques for the solution (the determination of both upper and lower bounds on v_β and ε-optimal strategies) of the discounted Markov game. One method has been suggested by Hoffman and Karp [4] (their algorithm was originally given for the Markov game with the average reward per unit time criterion but can be applied for the discounted game as well). Another method can be found in Pollatschek and Avi-Itzhak [8]. However, these authors only prove convergence of their Newton-Raphson (Howard) technique under very strong conditions.

In this report we will introduce stopping times as suggested by Wessels [12] for Markov decision processes in order to develop a number of successive approximation algorithms (section 2). This approach has also the advantage of simplifying the proofs. In section 3 we show that a special class of stopping times generates algorithms providing upper and lower bounds on v_β and ε-optimal strategies which are stationary.

One of the algorithms we will obtain is the standard successive approximation algorithm given by McQueen [6] for Markov decision processes. Some of the other algorithms are presented for Markov decision processes by Hastings [2], Reetz [9] and Van Nunen [7].

2. Stopping times

In this section we will use stopping times as Wessels [12] did for the discounted Markov decision process and the results we obtain will be very similar.

Definition 1. A map τ from S^∞ into the set of integers between 0 and ∞ (bounds included) is called a stopping time if and only if

$$\tau^{\leftarrow}(n) = B \times S^\infty, \qquad \text{with } B \subset S^{n+1} .$$

This means: if $\tau(x_0,\ldots,x_n,x_{n+1},\ldots) = n$, then for all $y_k \in S$, $k \geq n+1$, $\tau(x_0,\ldots,x_n,y_{n+1},\ldots) = n$ as well.

Definition 2. A stopping time τ is called nonzero if and only if $\tau(\alpha) > 0$, for each $\alpha \in S^\infty$.

Let τ be a stopping time and π and ρ be arbitrary strategies, let X_τ be a random variable denoting the state of the system at "time τ" if $\tau < \infty$ and let $X_\tau := 1$ if $\tau = \infty$, and let X_0 denote the starting state, the state of the system at $t = 0$. Now a notation will be introduced for the expected discounted reward for P_1 if the Markov game will be terminated at "time τ" with P_1 obtaining a final payoff $v(y)$ if $X_\tau = y$, when $X_0 = x$ and strategies π and ρ are used. By termination and "time τ" we mean termination as soon as a path $(x_0,\ldots,x_n)$ has occurred such that $\tau(x_0,\ldots,x_n,x_{n+1},\ldots) = n$, $n = 0,1,\ldots$

Definition 3. Let τ be a stopping time and let π and ρ be arbitrary strategies, then the operator $L_\tau(\pi,\rho)$ on $\mathbb{R}^N$ is defined by

$$(L_\tau(\pi,\rho)v)(x) = E[\sum_{n=0}^{\tau-1} \beta^n q_n + \beta^\tau v(X_\tau) \mid X_0 = x,\pi,\rho], \qquad x \in S$$

(where E denotes expectation and q_n is a random variable denoting the reward for P_1 at time n.)

Definition 4. Let τ be a stopping time, then the operator U_τ on $\mathbb{R}^N$ is defined by

$$U_\tau v = \sup_\pi \inf_\rho L_\tau(\pi,\rho)v$$

where the sup inf is taken componentwise.

Theorem 1.

i) $L_\tau(\pi,\rho)$ is a monotone mapping.

ii) $L_\tau(\pi,\rho)$ is strictly contracting for nonzero τ with respect to supnorm in $\mathbb{R}^N$ with contraction radius $\max_{x\in S} E(\beta^\tau \mid X_0 = x,\pi,\rho)$.

iii) U_τ is a monotone mapping.

iv) U_τ is strictly contracting for nonzero τ with respect to supnorm in $\mathbb{R}^N$. The contraction radius r_τ of U_τ satisfies

$$r_\tau \le \max_{x\in S} \sup_\pi \sup_\rho E(\beta^\tau \mid X_0 = x,\pi,\rho)$$

and

$$r_\tau \ge \max_{x\in S} \max\{\sup_\pi \inf_\rho E(\beta^\tau \mid X_0 = x,\pi,\rho),\ \inf_\pi \sup_\rho E(\beta^\tau \mid X_0 = x,\pi,\rho)\}.$$

Proof. i) and iii) are obvious, and the proof of ii) is straightforward.

iv) For arbitrary v and w in $\mathbb{R}^N$ we have,

$$U_\tau v(x) \le U_\tau(w + \|v - w\|e)(x) =$$

$$= \sup_\pi \inf_\rho E[\sum_{n=0}^{\tau-1} \beta^n q_n + \beta^\tau(w(X_\tau) + \|v - w\|) \mid X_0 = x,\pi,\rho] \le$$

$$\le \sup_\pi \inf_\rho E[\sum_{n=0}^{\tau-1} \beta^n q_n + \beta^\tau w(X_\tau) \mid X_0 = x,\pi,\rho] +$$

$$+ \sup_\pi \sup_\rho E(\beta^\tau \mid X_0 = x,\pi,\rho)\|v - w\| =$$

$$= U_\tau w(x) + \sup_\pi \sup_\rho E(\beta^\tau \mid X_0 = x,\pi,\rho)\|v - w\| .$$

Similarly we show

$$U_\tau w(x) \le U_\tau v(x) + \sup_\pi \sup_\rho E(\beta^\tau \mid X_0 = x,\pi,\rho)\|v - w\| .$$

Hence U_τ is strictly contracting with respect to supnorm in $\mathbb{R}^N$ for all nonzero τ, and we have obtained an upper bound on r_τ. The lower bound is found by taking $v = v_0.e$ and $w = 0$ and considering the cases $v_0 \to +\infty$ and $v_0 \to -\infty$.□

Remark 1. Counter examples can be constructed showing that r_τ is neither necessarily equal to the lower bound nor necessarily equal to the upper bound given in theorem 1 iv) (see Van der Wal [11]).

Shapley [10] has shown that the value of the game v_β, which is obviously the fixed point of the operator U_τ with $\tau \equiv \infty$, is also equal to the fixed point of the operator U_τ with $\tau \equiv 1$. As a consequence of theorem 1 iv) U_τ has a unique fixed point for all nonzero τ. Fortunately these fixed points are all equal to v_β. This is stated in the following theorem.

Theorem 2. U_τ has the unique fixed point v_β for any nonzero τ.

Proof. U_τ has a unique fixed point for any nonzero τ thus we only have to show $U_\tau v_\beta = v_\beta$. The value v_β and the optimal strategies $f^{*(\infty)}$ and $g^{*(\infty)}$ satisfy

$$V(\pi,g^{*(\infty)}) \le v_\beta \le V(f^{*(\infty)},g^{*(\infty)}) \le V(f^{*(\infty)},\rho) \quad \text{for all } \pi \text{ and } \rho .$$

With $V(\pi,\rho) = L_{\tau\equiv\infty}(\pi,\rho)0$ it follows that

$$\inf_\rho L_{\tau\equiv\infty}(f^{*(\infty)},\rho)0 = v_\beta = \sup_\pi L_{\tau\equiv\infty}(\pi,g^{*(\infty)})0 .$$

Now let P_1 use the fixed stationary strategy $f^{*(\infty)}$. Then we obtain a Markov decision process and we may apply theorem 3.1c) in Wessels [12]. There is stated for any nonzero τ

$$\inf_\rho L_\tau(f^{*(\infty)},\rho)\inf_\rho L_{\tau\equiv\infty}(f^{*(\infty)},\rho)0 = \inf_\rho L_{\tau\equiv\infty}(f^{*(\infty)},\rho)0$$

or

$$\inf_\rho L_\tau(f^{*(\infty)},\rho)v_\beta = v_\beta .$$

Similarly we find

$$\sup_\pi L_\tau(\pi,g^{*(\infty)})v_\beta = v_\beta .$$

As a consequence we have

$$v_\beta = \inf_\rho L_\tau(f^{*(\infty)},\rho)v_\beta \le \sup_\pi \inf_\rho L_\tau(\pi,\rho) = U_\tau v_\beta \le \sup_\pi L_\tau(\pi,g^{*(\infty)})v_\beta = v_\beta .$$

Thus $U_\tau v_\beta = v_\beta$ for all nonzero τ. □

Knowing that for nonzero τ all U_τ have fixed point v_β, we are interested in those operators U_τ for which $U_\tau v$ can be computed relatively easily. In general there will exist no stationary optimal strategies for a "τ-step" Markov game with payoff $v \in \mathbb{R}^N$. However, it turns out that for special stopping times τ we only need to consider stationary strategies.

Definition 5. A nonzero stopping time τ is called transition memoryless if and only if a subset T of S^2 exists such that

$$\tau(\alpha) = n \leftrightarrow (\alpha_k,\alpha_{k+1}) \notin T \quad \text{for } k = 0,\ldots,n-2,\ (\alpha_{n-1},\alpha_n) \in T .$$

Theorem 3. If τ is nonzero and transition memoryless, then for any $v \in \mathbb{R}^N$ stationary strategies $f^{(\infty)}$ and $g^{(\infty)}$ exist such that for all π and ρ

$$L_\tau(\pi,g^{(\infty)})v \le L_\tau(f^{(\infty)},g^{(\infty)})v \le L_\tau(f^{(\infty)},\rho)v .$$

Proof. We will define a new infinite horizon Markov game with $\bar{S}$, the new state space, being the union of two representations of S: $S^* := \{x^* \mid x \in S\}$ and $S_* := \{x_* \mid x \in S\}$ and with $K_{x^*} := K_{x_*} := K_x$ and $L_{x^*} := L_{x_*} := L_x$. Furthermore, define for all $x_*,y_* \in S_*$, $x^*,y^* \in S^*$

$$p(x_* \mid x_*,k,\ell) := 1,\ r(x_*,k,\ell) := (1-\beta)v(x),\ k \in K_x,\ \ell \in L_x ,$$

$$p(y^* \mid x^*,k,\ell) := p(y \mid x,k,\ell) \quad \text{if } (x,y) \notin T ,$$

$$p(y_* \mid x^*,k,\ell) := p(y \mid x,k,\ell) \quad \text{if } (x,y) \in T ,$$

$$r(x^*,k,\ell) := r(x,k,\ell),\ k \in K_x,\ \ell \in L_x$$

and for $x,y \in \bar{S}$: $p(y \mid x,k,\ell) := 0$ if not already defined otherwise. For the Markov game defined above optimal stationary strategies exist (Shapley [10]). The part of such a strategy, which concerns the states $x^* \in S^*$, constitutes a stationary optimal strategy for the "τ-step" game with final payoff v. Hence the theorem has been proved. □

3. Successive approximation

In this section we show that each nonzero transition memoryless stopping time generates a successive approximation algorithm.

Let τ be a nonzero transition memoryless stopping time. Define the sequence of vectors $\{v_{\tau,n}\}_{n=0}^{\infty} \subset \mathbb{R}^N$ by

$$v_{\tau,0} = 0$$

$$v_{\tau,n} = U_\tau v_{\tau,n-1}, \quad n = 1,2,\ldots .$$

Let $f_{\tau,n}^{(\infty)}$ and $g_{\tau,n}^{(\infty)}$ be optimal strategies for the "τ-step" game with final payoff $v_{\tau,n-1}$, $n = 1,2,\ldots$

Moreover, define $\lambda_{\tau,n}$, $\mu_{\tau,n}$, $a_{\tau,n}$ and $b_{\tau,n}$, $n = 1,2,\ldots$ by

$$\lambda_{\tau,n} := \min_{x \in S} \{v_{\tau,n}(x) - v_{\tau,n-1}(x)\},$$

$$\mu_{\tau,n} := \max_{x \in S} \{v_{\tau,n}(x) - v_{\tau,n-1}(x)\},$$

$$a_{\tau,n} := \begin{cases} \max\limits_{x,g} E[\beta^\tau \mid X_0 = x, f_{\tau,n}^{(\infty)}, g^{(\infty)}] & \text{if } \lambda_{\tau,n} < 0, \\ \min\limits_{x,g} E[\beta^\tau \mid X_0 = x, f_{\tau,n}^{(\infty)}, g^{(\infty)}] & \text{if } \lambda_{\tau,n} \geq 0, \end{cases}$$

$$b_{\tau,n} := \begin{cases} \min\limits_{x,f} E[\beta^\tau \mid X_0 = x, f^{(\infty)}, g_{\tau,n}^{(\infty)}] & \text{if } \mu_{\tau,n} < 0, \\ \max\limits_{x,f} E[\beta^\tau \mid X_0 = x, f^{(\infty)}, g_{\tau,n}^{(\infty)}] & \text{if } \mu_{\tau,n} \geq 0. \end{cases}$$

Now we state the following theorem.

Theorem 4. For nonzero transition memoryless stopping times τ the following estimates hold:

i) $$v_{\tau,n} + \frac{a_{\tau,n}\lambda_{\tau,n}}{1 - a_{\tau,n}} \cdot e \leq v_\beta \leq v_{\tau,n} + \frac{b_{\tau,n}\mu_{\tau,n}}{1 - b_{\tau,n}} \cdot e .$$

And for all π and ρ

ii) $$V(f_{\tau,n}^{(\infty)}, \rho) \geq v_{\tau,n} + \frac{a_{\tau,n}\lambda_{\tau,n}}{1 - a_{\tau,n}} \cdot e .$$

iii) $$V(\pi, g_{\tau,n}^{(\infty)}) \leq v_{\tau,n} + \frac{b_{\tau,n}\mu_{\tau,n}}{1 - b_{\tau,n}} \cdot e .$$

Proof. We first show ii). Let g be an arbitrary policy. We have (by definition)

$$L_\tau(f^{(\infty)}_{\tau,n}, g^{(\infty)})v_{\tau,n-1} \geq v_{\tau,n} \geq v_{\tau,n-1} + \lambda_{\tau,n}e ,$$

and

$$L_\tau(f^{(\infty)}_{\tau,n}, g^{(\infty)})(v_{\tau,n-1} + \lambda_{\tau,n}e)(x) =$$

$$= L_\tau(f^{(\infty)}_{\tau,n}, g^{(\infty)})v_{\tau,n-1}(x) + E[\beta^\tau \mid X_0 = x,\ f^{(\infty)}_{\tau,n}, g^{(\infty)}]\lambda_{\tau,n} ,$$

and by definition of $a_{\tau,n}$:

$$E[\beta^\tau \mid X_0 = x,\ f^{(\infty)}_{\tau,n}, g^{(\infty)}]\lambda_{\tau,n} \geq a_{\tau,n}\lambda_{\tau,n} .$$

Hence

$$L^p_\tau(f^{(\infty)}_{\tau,n}, g^{(\infty)})v_{\tau,n-1} \geq L^{p-1}_\tau(f^{(\infty)}_{\tau,n}, g^{(\infty)})(v_{\tau,n-1} + \lambda_{\tau,n}e) \geq$$

$$\geq L^{p-2}_\tau(f^{(\infty)}_{\tau,n}, g^{(\infty)})L_\tau(f^{(\infty)}_{\tau,n}, g^{(\infty)})(v_{\tau,n-1} + \lambda_{\tau,n}e) \geq$$

$$\geq L^{p-2}_\tau(f^{(\infty)}_{\tau,n}, g^{(\infty)})(v_{\tau,n-1} + \lambda_{\tau,n}e + a_{\tau,n}\lambda_{\tau,n}e) \geq$$

$$\geq \ldots \geq v_{\tau,n} + (a_{\tau,n} + \ldots + a^{p-1}_{\tau,n})\lambda_{\tau,n}e .$$

Therefore

$$V(f^{(\infty)}_{\tau,n}, \rho) \geq \min_g V(f^{(\infty)}_{\tau,n}, g^{(\infty)}) = \min_g \lim_{p\to\infty} L^p_\tau(f^{(\infty)}_{\tau,n}, g^{(\infty)})v_{\tau,n-1} \geq$$

$$\geq v_{\tau,n} + \frac{a_{\tau,n}\lambda_{\tau,n}}{1 - a_{\tau,n}} \cdot e .$$

The first inequality follows from the fact that for Markov decision processes stationary optimal strategies exist (see e.g. Blackwell [1] or Wessels and Van Nunen [14]). The equality follows from lemma 1.1 in Wessels [21]. With $v_\beta \geq \min_g V(f^{(\infty)}_{\tau,n}, g^{(\infty)})$ follows

$$v_\beta \geq v_{\tau,n} + \frac{a_{\tau,n}\lambda_{\tau,n}}{1 - a_{\tau,n}} \cdot e .$$

Similarly we show iii) and the second inequality in i). □

Remark 2. These bounds are practically identical to those given by Wessels and Van Nunen [13] for Markov decision processes.
Hinderer [3] has given many estimates for the special case $\tau \equiv 1$ for finite stage Markov decision processes. Some of these estimates may be extended for infinite horizon Markov games.
Since

$$\frac{a_{\tau,n}\lambda_{\tau,n}}{1 - a_{\tau,n}} \quad \text{and} \quad \frac{b_{\tau,n}\mu_{\tau,n}}{1 - b_{\tau,n}}$$

tend to zero if n tends to infinity, we can construct for nonzero transition memoryless stopping times τ an algorithm of the following form.

Algorithm (τ).

STEP 0: Define $v_{\tau,0}(x) := 0$ for $x = 1,\ldots,N$. Select $\varepsilon > 0$.

STEP 1: Compute $v_{\tau,n} := U_\tau v_{\tau,n-1}$ for $n = 1,\ldots,M$, where M is the smallest integer with

$$\frac{b_{\tau,n}\mu_{\tau,n}}{1 - b_{\tau,n}} - \frac{a_{\tau,n}\lambda_{\tau,n}}{1 - a_{\tau,n}} \leq \varepsilon \ .$$

STEP 2: Find stationary strategies $f^{(\infty)}_{\tau,M}$ and $g^{(\infty)}_{\tau,M}$ satisfying for all π and ρ

$$L_\tau(\pi, g^{(\infty)}_{\tau,M})v_{\tau,M-1} \leq L_\tau(f^{(\infty)}_{\tau,M}, g^{(\infty)}_{\tau,M})v_{\tau,M-1} \leq L_\tau(f^{(\infty)}_{\tau,M}, \rho)v_{\tau,M-1} \ .$$

We now have quite a number of algorithms, however only a few of them are of practical interest. Often the amount of work which has to be done in order to compute $v_{\tau,n}$ from $v_{\tau,n-1}$ will be tremendous.
However, there exist special nonzero transition memoryless stopping times, for which, in order to compute $v_{\tau,n}(x)$ from $v_{\tau,n-1}$, it is only necessary to compare the (mixed) actions which may be taken in state x, and one does not have to consider actions in other states.
We will give four of these algorithms which are already known from discounted Markov decision processes.

Algorithms.

i) $\tau \equiv 1$. The standard successive approximation method with $a_{\tau,n} = b_{\tau,n} = \beta$ for all n. The estimates have been given for discounted Markov decision processes by MacQueen [5].

ii) $\tau^{\leftarrow}(m) = \{\alpha \in S^{\infty} \mid \alpha_0 > \alpha_1 > \ldots > \alpha_{m-1}, \alpha_{m-1} \leq \alpha_m\}$. In this case $v_{\tau,n}$ can be computed recursively by

$$v_{\tau,n}(x) = \max_{f(x)} \min_{g(x)} \sum_{k \in K_x} f^k(x) \sum_{\ell \in L_x} g^{\ell}(x)[r(x,k,\ell) + \beta \sum_{y \geq x} p(y|x,k,\ell) v_{\tau,n-1}(y) + \beta \sum_{y<x} p(y|x,k,\ell) v_{\tau,n}(y)],$$

$x = 1,\ldots,N$. Where $f^k(x)$ ($g^{\ell}(x)$) denotes the probability that action $k(\ell)$ will be selected in state x according to policy f(g).

iii) $\tau^{\leftarrow}(m) = \{\alpha \in S^{\infty} \mid \alpha_0 = \alpha_1 = \ldots = \alpha_{m-1}, \alpha_{m-1} \neq \alpha_m\}$. Here $v_{\tau,n}$ is given by

$$v_{\tau,n}(x) = \max_{f(x)} \min_{g(x)} \cdot \left[\frac{\sum_{k \in K_x} f^k(x) \sum_{\ell \in L_x} g^{\ell}(x)[r(x,k,\ell) + \beta \sum_{y \neq x} p(y|x,k,\ell) v_{\tau,n-1}(y)]}{1 - \beta \sum_{k \in K_x} f^k(x) \sum_{\ell \in L_x} g^{\ell}(x) p(x|x,k,\ell)} \right],$$

$x = 1,\ldots,N$.

iv) $\tau^{\leftarrow}(m) = \{\alpha \in S^{\infty} \mid \alpha_0 \geq \alpha_1 \geq \ldots \geq \alpha_{m-1}, \alpha_{m-1} < \alpha_m\}$. This algorithm is a combination of the algorithms ii) and iii).

$v_{\tau,n}$ is given by

$$v_{\tau,n}(x) = \max_{f(x)} \min_{g(x)} \sum_{k \in K_x} f^k(x) \sum_{\ell \in L_x} g^{\ell}(x) \cdot [r(x,k,\ell) + \beta \sum_{y>x} p(y|x,k,\ell) v_{\tau,n-1}(y) + \beta \sum_{y<x} p(y|x,k,\ell) v_{\tau,n}(y)] \cdot [1 - \beta \sum_{k \in K_x} f^k(x) \sum_{\ell \in L_x} g^{\ell}(x) p(x|x,k,\ell)]^{-1}.$$

Algorithms ii), iii) and iv) were introduced for discounted Markov decision processes by Van Nunen [7] inspired by algorithms of Hastings [2] (algorithm ii)) and Reetz [9] (algorithm iii)). Van Nunen also shows that it is quite difficult to compare these four algorithms, giving examples demonstrating that the decision which algorithm should be prefered depends on the specific structure of the problem under consideration.

Remark 3. In the algorithm we suggest to execute STEP 1 until

$$\frac{b_{\tau,n}\mu_{\tau,n}}{1 - b_{\tau,n}} - \frac{a_{\tau,n}\lambda_{\tau,n}}{1 - a_{\tau,n}} \leq \varepsilon .$$

For algorithm i) this criterion is quite useful since $a_{\tau,n} = b_{\tau,n} = \beta$ for all n. If however, $a_{\tau,n}$ and $b_{\tau,n}$ have to be computed, as for algorithms ii), iii) and iv), it might be more sensible to use upper and lower bounds on $a_{\tau,n}$ and $b_{\tau,n}$. For instance in the algorithms ii), iii) and iv) we might replace $a_{\tau,n}$ by β if $\lambda_{\tau,n} < 0$ and by 0 if $\lambda_{\tau,n} \geq 0$ and $b_{\tau,n}$ by 0 if $\mu_{\tau,n} < 0$ and by β if $\mu_{\tau,n} \geq 0$. We might also continue the execution of STEP 1 until

$$(*) \qquad \max\{|\lambda_{\tau,n}|, \mu_{\tau,n}, \mu_{\tau,n} - \lambda_{\tau,n}\} \leq \varepsilon\beta^{-1}(1 - \beta) .$$

It can be shown that (*) implies

$$\frac{b_{\tau,n}\mu_{\tau,n}}{1 - b_{\tau,n}} - \frac{a_{\tau,n}\lambda_{\tau,n}}{1 - a_{\tau,n}} \leq \varepsilon .$$

If in the case of algorithm iii) or iv) we have for all x, k and ℓ $p(x|x,k,\ell) \geq c > 0$, we might replace $a_{\tau,n}$ by $\frac{\beta - \beta c}{1 - \beta c}$ if $\lambda_{\tau,n} < 0$ and by 0 if $\lambda_{\tau,n} \geq 0$, and $b_{\tau,n}$ by 0 if $\mu_{\tau,n} < 0$ and by $(\beta - \beta c)/(1-\beta c)$ if $\mu_{\tau,n} \geq 0$. Note that:

$$\max_{x,f,g} E[\beta^{\tau} \mid X_0 = x, f^{(\infty)}, g^{(\infty)}] \leq (1 - c)\beta + (1 - c)c.\beta^2 + ..$$

$$..+ (1 - c)c^k\beta^{k+1} +... = \frac{\beta - \beta c}{1 - \beta c} .$$

In this case we might also continue STEP 1 until

$$\max\{|\lambda_{\tau,n}|, \mu_{\tau,n}, \mu_{\tau,n} - \lambda_{\tau,n}\} \leq \varepsilon\beta^{-1}(1 - c)^{-1}(1 - \beta) .$$

Then one may show that after termination

$$\frac{b_{\tau,n}\mu_{\tau,n}}{1 - b_{\tau,n}} - \frac{a_{\tau,n}\lambda_{\tau,n}}{1 - a_{\tau,n}} \leq \varepsilon$$

will hold.

4. Some final remarks

We only considered the case of a discount factor $0 \leq \beta < 1$ and $\sum_{y \in S} p(y|x,k,\ell) = 1$ for all x, k and ℓ. Another approach could have been to demand $\sum_{y \in S} p(y|x,k,\ell) < 1$ for all x,k,ℓ and to use the criterion of total rewards. The difficulties we would encounter can be overcome by defining an extra absorbing state $0 \notin S$ with $r(0,k,\ell) \equiv 0$ and defining

$$p(0|x,k,\ell) = 1 - \sum_{y \in S} p(y|x,k,\ell) \ .$$

Furthermore we should redefine the stopping times on $\bar{S} := S \cup \{0\}$. The operator L_τ and U_τ should work on $\mathbb{R}^N$ again (no extra component corresponding to state 0) and the expression $E(\beta^\tau \mid X_0 = x,\pi,\rho)$ should be replaced by $P(X_\tau \in S \mid X_0 = x,\pi,\rho)$ (the probability that the game has not yet been absorbed in state 0 at time τ).

This approach can be used for the discounted game where the time between two subsequent action points is not equal to unity but has a probability distribution: the discounted semi-Markov game. In that case we may define

$$p'(y|x,k,\ell) := p(y|x,k,\ell)\beta(x,y,k,\ell) \ ,$$

where $\beta(x,y,k,\ell)$ denotes the expected discount factor when actions k and ℓ are taken in state x and the system moves to state y.

An interesting situation arises if in each state one of the players has only one action available: the perfect information case. Then the amount of work needed to compute $v_{\tau,n}$ from $v_{\tau,n-1}$ becomes essentially the same as for a Markov decision process of the same size. Another advantage is that we may also use a suboptimality test as introduced by MacQueen [6]. This is shown in [11]. For algorithm i) ($\tau \equiv 1$) the test can be performed with hardly any extra work.

References

[1] Blackwell, D.; Discounted dynamic programming. Ann. Math. Statist. 36 (1965), 226-235.

[2] Hastings, N.A.J.; Some notes on dynamic programming and replacement. Oper. Res. Q. 19 (1968), 453-464.

[3] Hinderer, K.; Estimates for finite stage dynamic programs. Institut für Mathematische Stochastik, Universität Hamburg.

[4] Hoffman, A.J. and Karp, R.M.; On nonterminating stochastic games. Management Science 12 (1966), 359-370.

[5] MacQueen, J.; A modified dynamic programming method for Markovian decision problems. J. Math. Anal. Appl. 14 (1966), 38-43.

[6] MacQueen, J.; A test for suboptimal actions in Markovian decision problems. O.R. 15 (1967), 559-561.

[7] Nunen, J.A.E.E. van; Improved successive approximation methods for discounted Markov decision processes. To appear in Colloquia Mathematica Societatis Janos Bolyai 12 (A. Prekopa ed.) North-Holland publ. co - Amsterdam.

[8] Pollatschek, M.A. and Avi-Itzhak, B.; Algorithms for stochastic games with geometrical interpretation. Management Science 15 (1969), 399-415.

[9] Reetz, D.; Solution of a Markovian Decision Problem by Successive Overrelaxation. Zeitschrift Operat. Res. 17 (1973), 29-32.

[10] Shapley, L.S.; Stochastic games. Proc. Nat. Acad. Sci. USA 39 (1953), 1095-1100.

[11] Wal, J. van der; The solution of Markov games by successive approximation. Master's thesis. Technological University Eindhoven. February 1975 (Department of Mathematics).

[12] Wessels, J.; Stopping times and Markov programming. To appear in Proceedings of the 1974 European meeting of Statisticians and 7th Prague Conference.

[13] Wessels, J. and Nunen, J.A.E.E. van; A principle for generating optimization procedures for discounted Markov decision processes. To appear in Colloquia Mathematica Societatis Janos Bolyai 12 (A. Prekopa ed.) North-Holland publ. co - Amsterdam.

[14] Wessels, J. and Nunen, J.A.E.E. van; Discounted semi-Markov decision processes: linear programming and policy iteration. Statistica Neerlandica 29 (1975), 1-7.

[15] Zachrisson, L.E.; Markov games. Annals of Mathematics Studies No. 52 (Princeton, New Yersey, 1964), 211-253.