Control
and Cybernetics
VOL. 11 (1982) No. 3–4

Modified Policy Iteration Algorithms for Discounted Markov Games

by

KATSUHISA OHNO

and

STANISŁAW ZDRZAŁKA*

Department of Applied Mathematics
and Physics, Faculty of Engineering
Kyoto University
Kyoto, 606 Japan

In the paper, a modification of the Newton-Raphson procedure for two-person zero-sum discounted Markov games is proposed and a sufficient condition for its convergence derived. As an additional result, a new sufficient condition for convergence of the Newton-Raphson procedure is obtained. It is also proved that the *Van der Wal's* [11] generalized policy iteration method converges from any starting point.

1. Introduction and Notation

Is this paper we deal with computational methods for the two-person zero-sum discounted Markov game with finite state space and finite action sets. For this game, a number of algorithms have already appeared in the literature. As regards the concepts employed, one can distinguish three main approaches: successive approximation, policy iteration method and modified policy iteration method (a survey of the existing algorithms with regard to this classification is presented in Section 2). The last mentioned method is of particular interest, since it requires less computational effort than the policy iteration technique, and simultaneously, it preserves a high convergence rate. The only algorithm employing this concept was presented by *Van der Wal* [11]. This algorithm was obtained by modification of the *Hoffman and Karp's* [3] policy iteration method and its convergence was proved under a certain condition imposed on a starting point.

Here, we examine two basic algorithms of the modified policy iteration method. In Section 3, we propose a modification of the Newton–Raphson technique, which

* on leave from the Institute of Engineering Cybernetics, Technical University of Wrocław, 50–370 Wrocław, Poland.

was first used for the discounted Markov game by *Pollatschek and Avi-Itzhak* [6], and derive a condition under which this algorithm converges from any starting point. As an additional results, we obtain a new condition for convergence of the Newton–Raphson technique, which is less restrictive than the one derived by *Pollatschek and Avi-Itzhak*. In Section 4, we prove that *Van der Wal's* [11] algorithm converges from any starting point, thus removing the restriction which was imposed on a starting point in the original proof of convergence.

Now we specify the game and introduce the notation and the basic results. The game under consideration may be regarded as a dynamic system with a state space $S=\{1, ..., N\}$. At discrete times $t=0, 1, ...$, the system is observed and its behaviour is influenced by two players, P_1 and P_2, having opposite aims. For each state $i \in S$ there exist two sets of available actions, $K_i=\{1, ..., k_i\}$ for P_1 and $L_i=\{1, ..., l_i\}$ for P_2. If the system is in state i and players P_1 and P_2 choose actions $k \in K_i$ and $l \in L_i$, respectively, then the first player receives a reward r_i^{kl} from the second one and the system proceeds to a new state j with probability p_{ij}^{kl}, where $\sum_{j \in S} p_{ij}^{kl}=1$ for any i, k, and l. The rewards are discounted at a rate β, $0 \leqslant \beta < 1$, that is the unit reward at time $t=n$ has the value β^n at time $t=0$. The total discounted reward is equal to $\sum_{n=0}^{\infty} \beta^n R_n$, where R_n is a reward obtained at time $t=n$. We consider the criterion of a total expected discounted reward.

The discounted Markov game, as described above, was introduced by *Gilette* [2], who noticed that all results obtained by *Shapley* [1953] for the terminating game remain valid if the condition: $\max_{i,k,l} \sum_{j \in S} p_{ij}^{kl} < 1$, determining the terminating game, is replaced by: $\sum_{j \in S} p_{ij}^{kl}=1$ for any i, k, l, and rewards are discounted at a rate β, $0 \leqslant \beta < 1$.

Of particular interest in the discounted Markov game are stationary strategies. A stationary strategy of P_1 is a N-tuple $f=(f^1, ..., f^N)$, where $f^i=(x_1^i, ..., x_{k_i}^i)$ is a probability distribution on the action set K_i such that, whenever state i is reached, action $k \in K_i$ is chosen with probability x_k^i regardless of the time and history of the game. In a similar way, we define a stationary strategy of P_2, $g=(g^1, ..., g^N)$, where $g^i=(y_1^i, ..., y_{l_i}^i)$ is a probability distribution on L_i.

Shapley [9] has shown that a solution of the discounted Markov game exists in the class of stationary strategies and the value of the game, denoted by $v_\beta=(v_\beta(1), ..., ..., v_\beta(N))$, is the unique solution of the equation

$$v(i)=\max_{f^i} \min_{g^i} \sum_{k \in K_i} \sum_{l \in L_i} x_k^i y_l^i \Big(r_i^{kl}+\beta \sum_{j \in S} p_{ij}^{kl} v(j)\Big), \quad i \in S. \tag{1}$$

The optimal strategies, denoted by f^* and g^*, satisfy

$$v_\beta(i)=\sum_{k \in K_i} \sum_{l \in L_i} x_k^{*i} y_l^{*i} \Big(r_i^{kl}+\beta \sum_{j \in S} p_{ij}^{kl} v_\beta(j)\Big), \quad i \in S.$$

Before presenting the main results of this paper, we propose a simple classification of the algorithms for the discounted Markov game based on the terminology and concepts from the Markov decision process.

2. Algorithms for Discounted Markov Games

Following [11], let us define the operators $L(f, g)$, U_g and U on R^N by:

$$(L(f,g)\,v)(i)=\sum_{k\in K_i}\sum_{l\in L_i} x_k^i y_l^i\Big(r_i^{kl}+\beta\sum_{j\in S} p_{ij}^{kl}\,v(j)\Big),\qquad i\in S;$$

$$U_g\,v=\max_f L(f,g)\,v;\qquad U_f\,v=\min_g L(f,g)\,v;$$

$$Uv=\max_f\min_g L(f,g)\,v,$$

where max and maxmin are taken componentwise. Under this notation, equation (1) becomes

$$v=Uv.$$

In R^N, we define a norm of a vector $v=(v(1),\ldots,v(N))$ as the maximum norm, that is, $\|v\|=\max_i|v(i)|$. The operators defined above have the following properties:[1]

(i) $L(f,g)$, U_g and are monotone.

(ii) $L(f,g)$, U_g and U are strictly contractive with respect to the maximum norm in R^N with the contraction radius β.

(iii) The fixed point of operator U is equal to v_β.

Properties (i) and (ii) can be easily derived. Property (iii) was shown by *Shapley* [9].

The algorithms for solving the discounted Markov game, that is, for solving the equation (1), can be classified to the following three groups. For the sake of clarity of presentation, we show here only the basic algorithms.

a) Successive Approximation

Algorithm: for $n=1, 2, \ldots$, and for given v_0 determine

$$v_n=Uv_{n-1}.$$

This method is a natural consequence of the properties (ii) and (iii) of the operator U. The first algorithm in this group was proposed by *Shapley* [9]. *Charnes and Schroeder* [1] supplied this algorithm with bounds based on the contraction radius of U and the maximum norm of the difference between two successive approximations. *Van der Wal* [10] extended the notion of stopping times, suggested by *Wessels* [13] for Markov decision processes, into the Markov games and obtained a set of operators $\{U_\tau\colon \tau\in T\}$ on R^N, where T is a set of nonzero transition memoryless stopping times. For each $\tau\in T$, the operator U_τ is strictly contractive with a fixed point equal to v_β and yields stationary optimal strategies for P_1 and P_2. In this *way Van der Wal* obtained a set of successive approximation algorithms for the discounted Markov game.

[1] Among real vectors $v, w\in R^N$, $v\leqslant w$ means $v(i)\leqslant w(i)$ for each i and $v<w$ means $v\leqslant w$ and $v\neq w$. An operator T on R^N is said to be monotone if for each $v, w\in R^N$, with $v\leqslant w$, we have $Tv\leqslant Tw$. An operator T on the normed space R^N is said to be strictly contractive if there is α $\alpha\in(0,1)$ such that $\|Tv-Tw\|\leqslant\alpha\|v-w\|$ for all $v, w\in R^N$.

At each iteration of this group of algorithms N matrix games, that is N linear programs, must be solved. The method belongs to the class of iterations in the function space. It does not utilize the information contained in the strategies obtained in each iteration.

b) Policy Iteration Method

For each $v_n \in R^N$, we define:

$$X(v_n)=\{f_{n+1} : U_{f_{n+1}} v_n = Uv_n\},$$

$$Y(v_n)=\{g_{n+1} : U_{g_{n+1}} v_n = Uv_n\},$$

Note that for each $(f_{n+1}, g_{n+1}) \in X(v_n)\times Y(v_n)$ a pair (f^i_{n+1}, g^i_{n+1}) is a pair of optimal strategies in a matrix game with entries $r^{kl}_i+\beta \sum_{j\in S} p^{kl}_{ij} v_n(j)$.

Algorithm A: for $n=1, 2, \ldots$, and for given v_0 determine

Step 1. (policy improvement step) $g_{n+1} \in Y(v_n)$;

Step 2. (value determination step) v_{n+1} satisfying: $v_{n+1}=U_{g_{n+1}} v_{n+1}$.

Algorithm B: for $n=1, 2,\ldots$, and for given v_0 determine

Step 1. $(f_{n+1}, g_{n+1}) \in X(v_n)\times Y(v_n)$;

Step 2. v_{n+1} satysfying: $v_{n+1}=L(f_{n+1}, g_{n+1})\, v_{n+1}$.

Both these variants are extensions of *Howard's* [4] policy iteration method for Markov decision processes. Algorithm A was proposed for the average Markov game by *Hoffman and Karp* [3]. *Rao, Chandrasekaran and Nair* [8] proved its convergence for the discounted Markov game. Algorithm B, known as the Newton–Raphson type algorithm, was presented by *Pollatschek and Avi-Itzhak* [6]. Its convergence was proved under the condition

$$\beta<1-\max_i \sum_{j\in S} (\max_{k,l} p^{kl}_{ij}-\min_{k,l} p^{kl}_{ij}), \tag{2}$$

which is rather too strong to have a practical meaning. It should be mentioned here that *Van der Wal* [11] gave a counterexample showing that this algorithm does not always converage.

In both variants of this method, each policy improvement step requires the solution of N linear programs. In the value determination step, algorithm A requires the solution of a Markov decision process, while algorithm B requires only the solution of a set of linear equations. Algorithms A and B were compared in a number of numerical experiments by *Pollatschek and Avi-Itzhak* and *Rao, Chandrasekaran and Nair*. It occured that the procedure B was far superior to the procedure A as regards the number of iterations and computing time. Both these procedures appeared to be superior to the standard successive approximation algorithm.

c) Modified Policy Iteration Method

Algorithm A': for $n=1, 2, \ldots$, and for given v_0 and m, $m \geqslant 1$, determine

Step 1. (policy improvement step) $w_{n+1} = Uv_n$ and $g_{n+1} \in Y(v_n)$;

Step 2. (value approximation step) $v_{n+1} = U^m_{g_{n+1}} w_{n+1}$.

Algorithm B': for $n=1, 2, \ldots$, and for given v_0 and m, $m \geqslant 1$, determine

Step 1. $w_{n+1} = Uv_n$ and $(f_{n+1}, g_{n+1}) \in X(v_n) \times Y(v_n)$;

Step 2. $v_{n+1} = L(f_{n+1}, g_{n+1})^m w_{n+1}$.

In comparison with the previous method, here, the exact solutions of the equations appearing in the value determination step are replaced by their approximations. In fact, from property (ii) of the operators U_g and $L(f, g)$ it follows that for $m_{=\infty}$ we get algorithms A and B of the policy iteration method. This idea was employed in Markov decision processes by *Van Nunen* [12], *Puterman and Shin* [7] and *Ohno* [5].

Algorithm A, was proposed by *Van der Wal* [11] who also proved its convergence under the condition: $Uv_0 \leqslant v_0$. In Section 4, we prove that this algorithm converges from any starting point. Algorithm B' is examined in Section 3 of this paper.

The computational efforts incurred in Step 1 remain the same as in Step 1 of the policy iteration method. In Step 2, algorithm A' requires the solution of mN maximization problems, while algorithm B' requires only recurrent computing of the values

$$L(f_{n+1}, g_{a+1})^k w_{r+1} = L(f_{n+1}, g_{r+1}) L(f_{n+1}, g_{n+1})^{k-1} w_{n+1}, \qquad k=1, \ldots, m.$$

3. Modified Policy Iteration Method: Algorithm B'

In this section we shall prove a sufficient condition for convergence. We need the following

LEMMA 1. *Let A and B be matrices of the same dimension with entries a_{ij} and b_{ij}, respectively. Denote by $ValC$ the value of the matrix game with matrix C. Then,*

$$\min_{i,j} (a_{ij} - b_{ij}) \leqslant ValA - ValB \leqslant \max_{i,j} (a_{ij} - b_{ij}).$$

This property was suggested by *Pollatschek and Avi-Itzhak* [6]. Its proof is obvious.

THEOREM 1. *Algorithm B' converges from any starting point if*

$$1 - 3\beta + \beta^{m+1} + \beta^{m+2} > 0. \tag{3}$$

Proof: Let us define numbers η_n and ξ_n as

$$\eta_n = \min_i [Uv_n(i) - v_n(i)],$$

$$\zeta_n = \max_i [Uv_n(i) - v_n(i)].$$

From the monotonicity of $L(f, g)$ it follows that

$$\eta_n \beta^k e \leqslant L(f_{n+1}, g_{n+1})^k Uv_n - L(f_{n+1}, g_{n+1})^k v_n \leqslant \xi_n \beta^k e, \qquad k=0, 1, \ldots, \tag{4}$$

where $L(f_{n+1}, g_{n+1})^0 \equiv 1$ and $e=(1, \ldots, 1)$. Using the equality

$$v_{n+1} - Uv_n = \sum_{k=1}^{m} [L(f_{n+1}, g_{n+1})^k Uv_n - L(f_{n+1}, g_{n+1})^k v_n],$$

we obtain from (4)

$$\eta_n \beta \frac{1-\beta^m}{1-\beta} e \leqslant v_{n+1} - Uv_n \leqslant \xi_n \beta \frac{1-\beta^m}{1-\beta} e. \tag{5}$$

In a similar way, using the fact that

$$\begin{aligned} v_{n+1} - v_n = L(f_{n+1}, g_{n+1})^{m+1} v_n - v_n = \\ = \sum_{k=1}^{m+1} [L(f_{n+1}, g_{n+1})^k v_n - L(f_{n+1}, g_{n+1})^{k-1} v_n], \end{aligned}$$

we get from (4)

$$\eta_n \frac{1-\beta^{m+1}}{1-\beta} e \leqslant v_{n+1} - v_n \leqslant \xi_n \frac{1-\beta^{m+1}}{1-\beta} e.$$

By Lemma 1,

$$\begin{aligned} Uv_{n+1}(i) - Uv_n(i) &\leqslant \max_{k,l} \beta \sum_{j \in S} p_{ij}^{kl} (v_{n+1}(j) - v_n(j)) \leqslant \\ &\leqslant \max_{k,l} \beta \sum_{j \in S} p_{ij}^{kl} \xi_n \frac{1-\beta^{m+1}}{1-\beta} e(j) = \xi_n \beta \frac{1-\beta^{m+1}}{1-\beta}, \qquad i \in S, \end{aligned} \tag{6}$$

and

$$Uv_{n+1}(i) - Uv_n(i) \geqslant \eta_n \beta \frac{1-\beta^{m+1}}{1-\beta}, \qquad i \in S. \tag{7}$$

Now, rewriting the difference $Uv_{n+1} - v_{n+1}$ as

$$Uv_{n+1} - v_{n+1} = Uv_{n+1} - Uv_n + Uv_n - v_{n+1}$$

and employing the inequalities (5), (6) and (7), we get

$$\begin{aligned} \left(-\beta \frac{1-\beta^m}{1-\beta} \xi_n + \beta \frac{1-\beta^{m+1}}{1-\beta} \eta_n\right) e \leqslant Uv_{n+1} + \\ - v_{n+1} \leqslant \left(\beta \frac{1-\beta^{m+1}}{1-\beta} \xi_n - \beta \frac{1-\beta^m}{1-\beta} \eta_n\right) e. \end{aligned}$$

These inequalities imply

$$\begin{aligned} \eta_{n+1} &\geqslant c(1-\beta^{m+1}) \eta_n - c(1-\beta^m) \xi_n, \\ \xi_{n+1} &\leqslant -c(1-\beta^m) \eta_n + c(1-\beta^{m+1}) \xi_n, \end{aligned}$$

where $c = \dfrac{\beta}{1-\beta}$. Setting $\eta'_n = -\eta_n$ and $\zeta_n = (\eta'_n, \xi_n)^T$, we obtain

$$\zeta_{n+1} \leqslant A\zeta_n,$$

where

$$A = \begin{bmatrix} c\,(1-\beta^{m+1}) & c\,(1-\beta^m) \\ c\,(1-\beta^m) & c\,(1-\beta^{m+1}) \end{bmatrix}.$$

Therefore,

$$\zeta_{n+1} \leqslant A^{n+1}\,\zeta_0. \tag{8}$$

The sequence $A^{n+1}\,\zeta_0$ converges to zero for any ζ_0 if and only if the eigen-values of the matrix A, denoted by λ_1 and λ_2, satisfy the condition: $|\lambda_1|$, $|\lambda_2| < 1$. We have: $\lambda_1 = \beta^{m+1}$ and $\lambda_2 = c\,(2-\beta^m-\beta^{m+1})$. It is clear that $|\lambda_1| < 1$ always holds, and that $|\lambda_2| < 1$ if and only if the condition (3) is satisfied.

Hence, the inequality (8) implies: $\limsup_{n\to\infty} \xi_n \leqslant 0$, and $\limsup_{n\to\infty} \eta'_n \leqslant 0$. Since $\eta'_n = -\eta_n$ and $\eta_n \leqslant \xi_n$ for every n, we find that

$$0 \leqslant \liminf_{n\to\infty} \eta_n \leqslant \limsup_{n\to\infty} \eta_n,\ \liminf_{n\to\infty} \xi_n \leqslant \limsup_{n\to\infty} \xi_n \leqslant 0.$$

Therefore $\lim_{n\to\infty} \|Uv_n - v_n\| = 0$. Since $\|v_\beta - v_n\| \leqslant \|v_\beta - Uv_n\| + \|Uv_n - v_n\| = \|Uv_\beta - Uv_n\| + \|Uv_n - v_n\| \leqslant \beta\|v_\beta - v_n\| + \|Uv_n - v_n\|$ and $\beta < 1$, this completes the proof. ■

The immediate consequence of Theorem 1 is the following

COROLLARY. *The Newton–Raphson type algorithm* (*policy iteration metod: algorithm B*) *converges from any starting point if*

$$\beta < \frac{1}{3}. \tag{9}$$

This follows from the fact that for $m_{\to\infty}$, the algorithm considered in this section becomes the Newton–Raphosn type algorithm. Hence, for $m_{\to\infty}$ the condition (3) has the form: $1-3\beta > 0$.

In *Van der Wal* [11], an example of the discounted Markov game is shown where the Newton–Raphson procedure does not converge. We shall confine our numerical experiment to this game only.

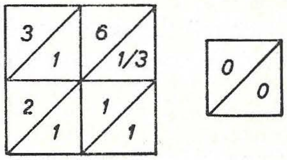

Fig. 1. Two-person zero-sum Markov game with two states

The mentioned game is shown in Fig. 1. The game has two states. Both players have two actions in state 1 and only one in state 2. According to the notation shown in Fig. 1, whenever state 1 is reached and players P_1 and P_2 choose actions 1 and 2, respectively, then P_1 receives the reward of 6 units from P_2, and the system remains in state 1 with probability 1/3 and moves to state 2 with probability 2/3.

In Table 1, we show the paths of convergence by the algortihm B' of modified policy iteration method and by the Newton–Raphson procedure. The convergence of this procedures in examined for two values of the discountfactor β: 3/4 and 1/4. Note that in the first case the conditions (3) and (9) are not satisfied. Here we observe only the behaviour of the seqeunce $v_n(1)$, since $v_n(2)=0$ for any n (we assume that $v_0=0$).

Table 1.

β	$v_n(1)$	$m=1$	$m=2$	$m=3$	$m=4$	$m=\infty$ (NR)
$\frac{3}{4}$	$v_0(1)$	0	0	0	0	0
	$v_1(1)$	5,25	6,9375	8,203125	9,152344	12
	$v_2(1)$	8,203125	7,983398	8,000793	8,004104	4
	$v_3(1)$	8,012655	7,999741		8,000001	12
	$v_4(1)$	8,000821				4
$\frac{1}{4}$	$v_0(1)$	0	0	0	0	0
	$v_1(1)$	3,75	3,9375	3,984375	3,996093	4
	$v_2(1)$	3,984375	3,996093	3,999023	3,999755	
	$v_3(1)$	3,999023	3,999755			

The example demonstrates that the algorithm B' of the modified policy iteration method converges in the case $\beta=3/4$, where the Newton–Raphson procedure oscillates. But it also occurs ($\beta=1/4$), that the Newton–Raphson procedure converges in one iteration, while its modifications need, in the same case, two or three iterations to reach a given neighbourhood of the solution.

4. Modified Policy Iteration Method: Algorithm A′

In this section, we show that the algorithm A' of the modified policy iteration method converges from any starting point.

First, we prove the following

LEMMA 2. *Let w_n and v_n be sequences determined by the algortihm A'. For every v_0 and $\varepsilon>0$ there exists a natural number M such that*

$$w_{n+1}-v_n\leqslant\varepsilon e, \text{ for all } n\geqslant M.$$

Proof: Let ξ_0 be a number such that $w_1-v_0\leqslant\xi_0 e$.
From the monotonicity of $L(f, g)$, we get

$$U_{g_1}w_1-U_{g_1}v_0=L(f', g_1)w_1-L(f_1, g_1)v_0\leqslant L(f', g_1)w_1+$$
$$-L(f', g_1)v_0\leqslant\beta\xi_0 e.$$

By induction,

$$U_{g_1}^{m+1} w_1 - U_{g_1}^{m+1} v_0 - U_{g_1} v_1 - v_1 \leqslant \beta^{m+1} \xi_0 e .$$

This implies

$$w_2 - v_1 = U v_1 - v_1 \leqslant U_{g_1} v_1 - v_1 \leqslant \beta^{m+1} \xi_0 e .$$

By induction,

$$w_{n+1} - v_n \leqslant \beta^{n(m+1)} \xi_0 e \text{ for } n = 0, 1, \ldots,$$

which in view of the fact that $0 \leqslant \beta < 1$, completes the proof. ■

THEOREM 2. *Algorithm A′ converges from any starting point.*

Proof: Choose an arbitrary $\varepsilon > 0$. By Lemma 2, there exists M such that for $n \geqslant M$

$$w_{n+1} - v_n \leqslant \varepsilon e . \tag{10}$$

From the monotonicity of $L(f, g)$, we get

$$U_{g_{n+1}} w_{n+1} - w_{n+1} = L(f', g_{n+1}) w_{n+1} + \\ - L(f_{n+1}, g_{n+1}) v_n \leqslant L(f', g_{n+1}) w_{n+1} - L(f', g_{n+1}) v_n \leqslant \beta \varepsilon e .$$

Adding this inequality to (10), we get

$$U_{g_{n+1}} w_{n+1} - v_n \leqslant (1 + \beta) \varepsilon e .$$

By induction,

$$U_{g_{n+1}}^{m-1} w_{n+1} - v_n \leqslant \sum_{k=0}^{m-1} \beta^k \varepsilon e \leqslant \frac{1}{1-\beta} \varepsilon e .$$

This inequality, and the monotonicity of $L(f, g)$ imply

$$v_{n+1} - w_{n+1} = U_{g_{n+1}}^{m} w_{n+1} - U_{g_{n+1}} v_n = L(f'', g_{n+1}) U_{g_{n+1}}^{m-1} w_{n+1} + \\ - L(f_{n+1}, g_{n+1}) v_{n+1} \leqslant L(f'', g_{n+1}) U_{g_{n+1}}^{m-1} w_{n+1} - L(f'', g_{n+1}) v_{n+1} \leqslant \\ \leqslant \frac{\beta}{1-\beta} \varepsilon e \equiv c \varepsilon e, \text{ for } n \geqslant M. \tag{11}$$

By Lemma 1,

$$w_{n+2} - U w_{n+1} = U v_{n+1} - U w_{n+1} \leqslant \beta c \varepsilon e .$$

Moreover, from (11) it follows that

$$v_{n+2} - w_{n+2} \leqslant c \varepsilon e .$$

An addition of the last two inequalities yields

$$v_{n+2} - U w_{n+1} \leqslant (1 + \beta) c \varepsilon e .$$

By induction, we obtain

$$v_{n+k+1}-U^k w_{n+1}\leqslant \sum_{p=0}^{k} \beta^p c\,\varepsilon\, e, \text{ for } k=0, 1, \ldots,$$

which in view of the properties (ii) and (iii) of U, yields

$$\limsup_{k\to\infty} v_{n+k}\leqslant v_\beta+\frac{c\,\varepsilon}{1-\beta}\, e. \tag{12}$$

Let us rewrite (10) as follows

$$Uv_n-v_n\leqslant \varepsilon\, e, \text{ for } n\geqslant M.$$

This inequality, and Lemma 1 imply

$$U^2 v_n-Uv_n\leqslant \beta\,\varepsilon\, e,$$

and consequently,

$$U^k v_n-U^{k-1} v_n\leqslant \beta^{k-1}\,\varepsilon\, e, \text{ for } k=1, 2, \ldots.$$

Hence,

$$U^k v_n-v_n=\sum_{p=1}^{k} (U^p v_n-U^{p-1} v_n)\leqslant \sum_{p=1}^{k} \beta^{p-1}\,\varepsilon\, e,$$

Letting $k_{\to\infty}$, we obtain

$$v_n\geqslant v_\beta-\frac{\varepsilon}{1-\beta}\, e,$$

which combined with (12), proves the theorem. ■

5. Remarks

In this paper, we have focused our attention on the basic algorithms, avoiding a discussion of such details as upper and lower bounds on a value of a game and stopping rules. The discussion of these questions may be found in the references, as regards the existing algorithms. The algorithm B' of the modified policy iteration method, introduced in this paper, can be supplied with precisely the same upper and lower bounds as those suggested by *Van der Wal* [11] for the algorithm A.

Convergence of the algorithms discussed in Sections 3 and 4 is preserved if we allow the transition probabilities to satisfy the condition $\sum_{j\in S} p_{ij}^{kl}\leqslant 1$, instead of $\sum_{j\in S} p_{ij}^{kl}=1$. These algorithms may also be applied to the terminating Markov game. In this case, the contraction radius of the operators $L(f, g)$, U_g, and U is equal to $\max_{i,k,l} \sum_{j\in S} p_{ij}^{kl}$.

The example shown in this paper demonstrates that the selection of the parameter m, which indicates the number of successive approximations employed in Step 2, is of importance as regards the rate of convergence. In this respect, we are not able to give justified recommendations. We mention only that the convergence of both algorithms, A' and B', is preserved if the parameter m varies with the number of iteration.

In Section 3, we have derived sufficient conditions for convergence of the policy iteration method and its modification in the case where in Step 2 the operator $L(f, g)$ has been employed. It can be verified on numerical examples that the condition (9) is less restrictive than the condition (2), suggested by *Pollatschek and Avi–Itzhak*, Nevertheless, the restriction on the discountfactor β imposed in (9) seems to be still too strong. This remark holds good for the condition (3), as well.

Acknowledgments

We wish to thank the referee for his helpful suggestions and comments.

References

[1] Charnes A., Schroeder R.G. On Some Stochastic Tactical Anti-Submarine Games. *Naval Reseach Logistics Quarterly*, **14** (1967), 291–311.

[2] Gilette D. Stochastic games with zero stop probabilities. Contribution to the theory of games, Vol. III, Ed. by M. Dresher, A.W. Tucker, and P. Wolfe, Princeton, New Jersey, 1957, 179–187.

[3] Hoffman A.K., Karp R.M. On Nonterminating Stochastic Games. *Management Science*, **12** (1966), 359–370.

[4] Howard R.A. Dynamic Programming and Markov Processes. MIT Press, Cambridge, Massachusetts, 1960.

[5] Ohno K. A Unified Approach to Algorithms with Suboptimality Test in Discounted Semi-Markov Decision Processes. *Journal of the Operations Research Society of Japan*, **24** (1981), 296–323.

[6] Pollatschek M.A., Avi-Itzhak B. Algorithms for Stochastic Games with Geometrical Interpretation, *Management Science*, **15** (1969), 399–415.

[7] Puterman M.L., Shin M.C. Modified Policy Iteration Algorithms for Discounted Markov Decision Problems. *Management Science*, **24** (1978), 1127–1137.

[8] Rao S.S., Chandrasekaran R., Nair K.P.K. Algorithms for Discounted Stochastic Games. *Journal of Optimization Theory and Applications*, **11** (1973), 627–637.

[9] Shapley L.S. Stochastic games. *Proc. Nat. Acad. Sci. USA*, **39** (1953), 1095–1100.

[10] Van der Wal J. Discounted Markov Games: Successive Approximation and Stopping Times. *Int. Journal of Game Theory*, 6, Issue **1** (1977), 11–22.

[11] Van der Wal J. Discounted Markov Games: Generalized Policy Iteration Method. *Journal of Optimization Theory and Applications*, **25** (1978), 125–138.

[12] Van Nunen J.A.E.E. A Set of Successive Approximation Methods for Discounted Markov Decision Processes. *Zeitschrift fur Operations Research*, **30** (1976), 203–208.

[13] Wessels J. Stopping Times and Markov Programming. Transactions of the Seventh Prague Conference and 1974 EMS, Prague, Academia, 1977, 575–585.

Received, June 1982.

Algorytmy zmodyfikowanej metody iteracji polityk dla dyskontowych gier Markova

W pracy przedstawiono modyfikację procedury typu Newtona–Raphsona dla dwuosobowych gier Markova o sumie zerowej oraz podano warunki wystarczające zbieżności. Otrzymano również nowy, słabszy warunek wystarczający zbieżności procedury typu Newtona–Raphosna oraz pokazano, że algorytm uogólnionej metody iteracji polityk zaproponowany przez Van der Wala [1978] jest zbieżny dla dowolnego punktu startowego.

Алгоритмы модифицированного метода итераций политик для учетных марковских игр

В работе представлена модификация процедуры типа Ньютона–Равсона для марковских игр двух лиц с нулевой суммой, а также приведены достаточные условия сходимости. Получено также новое ослабленное достаточное условие сходимости процедуры типа Ньютона–Равсона и показано, что алгоритм обобщенного метода итераций политик, предложенный Вандервалем [1978], сходим для произвольной стартовой точки.