Adaptive Zero-Sum Stochastic Game for Two Finite Markov Chains

A. S. Poznyak* and K. Najim**
*CINVESTAV-IPN, Control Automatico, A.P. 14-740, C.P.07300 Mexico D.F., Mexico,
fax: (52-5)747-70-89 or 747-70-02, e-mail: apoznyak@ctrl.cinvestav.mx
**Process Control Laboratory, E.N.S.I.G.C., Chemin de la loge, 31078 Toulouse cedex, France,
fax: (+33) 5 62 25 23 18, e-mai: Kaddour.Najim@ensigct.fr

Abstract

A two finite Markov chains repeated zero-sum stochastic game with unknown transition matrices and payoffs is considered. The control objective is to obtain the equilibrium point based only on current measurements. The behavior of each players is modelled by a finite controlled Markov chain. A novel adaptive policy is developed based on Lagrange multipliers involved into "learning through reinforcement" procedure. A regularized Lagrange function and a new normalization procedure are introduced. The saddle-point of this function is shown to be unique. The convergence properties are proved and the order of almost sure convergence is estimated as $\left(n^{-\frac{1}{3}}\right)$.

Keywords: Repeated stochastic game; adaptive control; controlled Markov chains„ learning, reinforcement.

1 Introduction

A lot of engineering, business and economic problems can be formulated as stochastic games (see, for example [1]). Finite *stochastic games* can be regarded as competitive *Markov Decision Processes* [2] where there are two or more controllers (players). There is an extensive literature on the existence of optimal strategies [5], [14], [11] and [6]. Several studies have been dedicated to the development of algorithms for the computation of equilibrium points and optimal strategies in stochastic games [10], [12], [4] and [13]. The term "algorithm" was interpreted broadly enough [10]. An *adaptive* or *learning algorithm* (recursive) can be defined as a procedure which forms a new estimate, incorporating new information (realizations), from the old estimate using a fixed amount of computations and memory.

This paper presents a novel algorithm for two finite Markov chains repeated zero-sum stochastic game with incomplete information on the transitions and payoffs. The characteristics (statistics, etc.) of the game are *a priori* unknown. Only the realizations (states and realized payoffs) of the game are available at each stage. Each player is modelled by a finite controlled irreducible Markov chain [9]. The control objective corresponds to the optimization of the random limiting average payoff which captures the long-run average performance. The algorithm presented in this paper is adaptive in the sense that it provides learning control policies for both players (Markov chains). It is based on the Lagrange multipliers approach and a new normalization procedure in order to preserve the probability measure.

2 System Description

Introduce the following notations:

$X_k = \left(x^k(1), ..., x^k(K_k)\right)$ is the finite state spaces associated respectively to the first player $(k = 1)$ and to the second $(k = 1)$ player;

$U_k = \left(u^k(1), ..., u^k(N_k)\right)$ is respectively the set of actions;

$\{\eta_n\}$ is a sequence of Borel functions (below, payoffs)

$$\eta_n = \eta_n\left(\omega, x_n^1, x_n^2, u_n^1, u_n^2\right)$$
$$\omega \in \Omega, x_n^1 \in X_1, x_n^2 \in X_2, u_n^1 \in U_1, u_n^2 \in U_2$$

given at a probability space (Ω, F, P) with a fixed flow $\mathsf{F}_n \subseteq \mathsf{F}_{n+1}$ of σ-algebras. Here x_n^k, u_n^k $(k = 1, 2)$ are the current states and control actions.

The behavior of each player is modelled by a finite homogeneous controlled Markov chain. These Markov chains are assumed to be *irreducible* (completely ergodic, that is, each state can be visited infinitely often) [9].

The game is played in stages $(n = 1, 2, ...)$. The play starts at stage 1 in the initial states x_1^1 and x_1^2. Each of the players is allowed to randomize over pure action choices $(u_1^1 \in U_1$ and $u_1^2 \in U_2)$. These choices

induce immediate random payoffs η_1 and so on. The first player tries to maximize the payoff he can guarantee himself by playing appropriate. The second player is interested in the minimization of the payoff which he at most has to pay.

Definition 1 $\{\eta_n\}$ $(n=1,2...)$ *is said to be the* ***payoff function*** *sequence for the first player (loss function for the second one) if* $|\eta_n| \overset{a.s.}{\leq} \sigma^+ < \infty$ *and*

$$E\left\{\eta_n \mid x_n^1 = x^1(i), x_n^2 = x^2(j); \right.$$
$$\left. u_n^1 = u^1(l), u_n^2 = u^2(r)\right\} = v_{ij}^{lr}$$
$$i = \overline{1, K_1}, j = \overline{1, K_2}, l = \overline{1, N_1}, r = \overline{1, N_2}$$

The play moves to new states (x_2^1 and x_2^2) according to the transition probabilities

$$\Pi^k = \left[\pi_{ij}^{k,l}\right] \ (k=1,2)$$

The element

$$\pi_{ij}^{k,l}(i=1,...,K_k; j=1,...,K_k \text{ and } l=1,...,N_k)$$

represents at time n $(n=1,2,...)$ the probability of transition from state $x^k(i)$ to state $x^k(j)$ under the action $u^k(l) \in U^k$, that is,

$$\pi_{ij}^{k,l} = P\left\{x_{n+1}^k = x^k(j) \mid x_n^k = x^k(i), u_n^k = u^k(l)\right\}$$

A sequence of random stochastic matrices $\left\{d_n^k\right\}$ is said to be **an admissible randomized (mixed) control strategy $\mathcal{D}^k$ for the k^{th} player** if

1) it is causal (independent on the future), that is,

$$d_n^k = \left[d_n^{k,il}\right]_{i=1,...,K_k; l=1,...,N_k}$$

is F_{n-1}-measurable, where

$$\mathsf{F}_{n-1} := \sigma\left(x_1^k, u_1^k, ; ....; x_{n-1}^k, u_{n-1}^k; k=1,2\right)$$

is the σ-algebra generated by the random variables $\left(x_1^k, u_1^k, ; ....; x_{n-1}^k, u_{n-1}^k\right)$ up to time $n-1$;

2) the random variables $\left(u_1^k, ..., u_{n-1}^k\right)$ represent the realizations of the control actions, taking values on the finite sets $U^k = \left\{u^k(1), ..., u^k(N_k)\right\}$, and satisfying the following property:

$$d_n^{k,il} = \Pr\left\{u_n^k = u^k(l) \mid x_n^k = x^k(i) \wedge \mathsf{F}_{n-1}\right\}$$

Denote by Σ^k the class of all **randomized strategies for the k^{th} player**, that is, $\Sigma^k = \left\{\left\{d_n^k\right\}\right\}$.The conditional transition probability matrix $\Pi^k(d_n^k)$ can be defined as

$$\Pi^k(d_n^k) = \left[\pi_n^{k,ij}(d_n)\right]_{i=1,...,K_k; j=1,...,K_k}$$
$$\pi_n^{k,ij}(d_n^k) := P\left\{x_{n+1}^k = x^k(j) \mid x_n^k = x^k(i) \wedge \mathsf{F}_{n-1}\right\}$$
$$\overset{a.s.}{=} \sum_{l=1}^{N_k} \pi_{ij}^{k,l} d_n^{k,il}$$

represents the law of motion among states. The *average payoff* v_n^1 of the 1^{th} player at the n^{th} stage is given by

$$v_n^1 = V\left(d_n^1, d_n^2\right)$$
$$:= \sum_{i=1}^{K_1}\sum_{j=1}^{K_2}\sum_{l=1}^{N_1}\sum_{r=1}^{N_2} v_{ij}^{lr} p_n^1(i) p_n^2(j) d_n^{1,il} d_n^{2,jr}$$

where $p_n^k(i_k)$ is the probability for the k^{th} player to be in the state $x^k(i_k)$ at time n verifying

$$p_{n+1}^k(j) = \sum_{i=1}^{K_k} \pi_n^{k,ij}(d_n^k) p_n^k(i)$$
$$= \sum_{i=1}^{K_k}\left(\sum_{l=1}^{N_k} \pi_{ij}^{k,l} d_n^{k,il}\right) p_n^k(i)$$

Introduce the following functions

$$\Phi_n := \frac{1}{n}\sum_{t=1}^{n} \xi_t$$

and

$$V_n := \frac{1}{n}\sum_{t=1}^{n} v_t = \frac{1}{n}\sum_{t=1}^{n} V\left(d_t^1, d_t^2\right)$$

Lemma 2 *Under the accepted assumptions*

$$\Phi_n = V_n + o_\omega\left(n^{1-\varepsilon}\right), \ \varepsilon > 0$$

with probability one.

Proof: It follows immediately from Lemma 3 for $v_t := \eta_n$, and $a_n := n$, and Lemma 4 for $\eta_n := n^{1-\varepsilon}$ (Appendix A, in [9]). ■

The *strategy* of the k^{th} player consists of any admissible matrix sequences $\mathsf{D}^k = \left\{d_n^k\right\}$ $(k=1,2)$ with $\mathsf{F}_{n-1}-$measurable elements.

Definition 3 *The pair of strategies $\bar{\mathsf{D}}^1$ and $\bar{\mathsf{D}}^2$ is said to be the non-cooperative equilibrium strategy (in the Nash sense) if for any admissible D^1 and D^2*

$$\widetilde{V}\left(\mathsf{D}^1, \bar{\mathsf{D}}^2\right) \overset{a.s.}{\leq} \widetilde{V}\left(\bar{\mathsf{D}}^1, \bar{\mathsf{D}}^2\right)$$
$$:= \min_{x_1^1, x_2^2} \liminf_{n\to\infty} V_n \overset{a.s.}{\leq} \widetilde{V}\left(\bar{\mathsf{D}}^1, \mathsf{D}^2\right)$$

Remark 4 *It is evident (Flesch et al., 1997) that in the case of irreducible controlled finite Markov chains (Poznyak et al., 1999), the set of all equilibrium strategies $\bar{\mathsf{D}}^1$ and $\bar{\mathsf{D}}^2$ contains the set of stationary strategies $\left\{\bar{d}^k\right\}$ $(k=1,2)$ where $\left(\bar{d}^1, \bar{d}^2\right)$ represents a saddle-point (may be not unique) of the function $V\left(d^1, d^2\right)$ which satisfies*

$$V\left(d^1, \bar{d}^2\right) \leq V\left(\bar{d}^1, \bar{d}^2\right) \leq V\left(\bar{d}^1, d^2\right)$$

for any $d^k \in \mathbf{D}_\varepsilon^k$ *where the simplex* $\mathbf{D}_\varepsilon^k$ *is defined as follows*

$$\mathbf{D}_\varepsilon^k := \left\{ d^k = \left\| d^{k,il} \right\| \left(i = \overline{1, K_k}; l = \overline{i, N_k} \right) : d^{k,il} \geq \varepsilon \geq 0, \sum_{l=1}^{N_k} d^{k,il} = 1 \right\}$$

This fact follows from: i) the continuity property of the function $V\left(d^1, d^2\right)$*, ii) the compactness of the sets* $\mathbf{D}_\varepsilon^k$*, and iii) Von Neumann theorem.*

3 Problem Setting

Introduce the following variable change [9]

$$c^{k,il} := p^k(i)\, d^{k,il} \ (k = 1, 2)$$

We are concerned with a constrained optimization problem: $c^{k,il}$ must verify (for each $k = 1, 2$ and $j = \overline{1, K_k}$) the constraint:

$$\sum_{l=1}^{N_k} c^{k,jl} = \sum_{i=1}^{K_k} \sum_{l=1}^{N_k} \pi_{ij}^{k,l} c^{k,il}$$

Consider the following regularized Lagrange function ($\delta > 0$)

$$\begin{gathered} L_\delta := L_\delta(\mathbf{c}^1, \mathbf{c}^2, \boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2) \\ = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \sum_{l=1}^{N_1} \sum_{r=1}^{N_2} v_{ij}^{lr} c^{1,il} c^{2,il} \\ - \sum_{k=1}^{2} (-1)^k \sum_{j=1}^{K_k} \lambda_j^k \left[\sum_{l=1}^{N_k} c^{k,jl} - \sum_{i=1}^{K_k} \sum_{l=1}^{N_k} \pi_{ij}^{k,l} c^{k,il} \right] \\ + (\delta/2) \sum_{k=1}^{2} (-1)^k \left(\sum_{i=1}^{K_k} \sum_{l=1}^{N_k} \left(c^{k,il}\right)^2 - \sum_{j=1}^{K_k} \left(\lambda_j^k\right)^2 \right) \\ \mathbf{c}^k := \left(c^{k,11}, ..., c^{k,1N_k}; ...; c^{k,K_k 1}, ..., c^{k,K_k N_k}\right)^T \\ \boldsymbol{\lambda}^k := \left(\lambda_1^k, ..., \lambda_{K_k}^k\right)^T \in R^{K_k} \end{gathered}$$

given in the set $\left(S_0^{K_1 N_1} \times R^{K_1}\right) \times \left(S_0^{K_2 N_2} \times R^{K_2}\right)$ with

$$S_{\varepsilon=0}^{K_k N_k} := \left\{ \begin{array}{c} \mathbf{c}^k \in R^{N_k K_k} \mid c^{k,il} \geq \varepsilon \geq 0 \\ \sum\limits_{i=1}^{K_k} \sum\limits_{l=1}^{N_k} c^{k,il} = 1 \ (i = \overline{1, K_k}; l = \overline{1, N_k}) \end{array} \right\}$$

The saddle-point of this regularized (augmented) Lagrange function will be denoted by

$$\begin{gathered} \left(\mathbf{c}_\delta^{1,*}, \mathbf{c}_\delta^{2,*}, \boldsymbol{\lambda}_\delta^{1,*}, \boldsymbol{\lambda}_\delta^{2,*}\right) := \\ \arg \max_{\mathbf{c}^1 \in S_0^{K_1 N_1}} \min_{\mathbf{c}^2 \in S_0^{K_2 N_2}} \min_{\boldsymbol{\lambda}^1 \in R^{K_1}} \max_{\boldsymbol{\lambda}^2 \in R^{K_2}} L_\delta \end{gathered} \tag{1}$$

The function $L_{\delta > 0}$ is strictly concave with respect to $\mathbf{c}_\delta^{1,*}$ and $\boldsymbol{\lambda}_\delta^{2,*}$, and is strictly convex with respect to $\mathbf{c}_\delta^{2,*}$ and $\boldsymbol{\lambda}_\delta^{1,*}$. It follows that its saddle-point is unique and possess the Lipshitz property with respect to the parameter δ:

$$\begin{gathered} \sum_{k=1}^{2} \left(\left\| \mathbf{c}_{\delta_1}^{k,*} - \mathbf{c}_{\delta_2}^{k,*} \right\| + \left\| \boldsymbol{\lambda}_{\delta_1}^{k,*} - \boldsymbol{\lambda}_{\delta_2}^{k,*} \right\| \right) \\ \leq Const\, |\delta_1 - \delta_2|, \ k = 1, 2 \end{gathered}$$

It has been shown (see Theorem 2 of section 4.3 in [8]) that if $\delta \to 0$, the saddle-point $\left(\mathbf{c}_\delta^{1,*}, \mathbf{c}_\delta^{2,*}, \boldsymbol{\lambda}_\delta^{1,*}, \boldsymbol{\lambda}_\delta^{2,*}\right)$ converges to the solution $\left(\mathbf{c}_0^{1,*}, \mathbf{c}_0^{2,*}, \boldsymbol{\lambda}_0^{1,*}, \boldsymbol{\lambda}_0^{2,*}\right)$ of the optimization problem (1) with $\delta = 0$, which has the minimal norm (in the case of an irreducible chain this point is also unique):

$$\begin{gathered} \mathbf{c}_\delta^{k,*} \underset{\delta \to 0}{\to} \mathbf{c}_0^{k,**} := \arg \min_{\mathbf{c}_0^{1,*} \in S_0^{K_1 N_1}, \boldsymbol{\lambda}^{k,*} \in R^{K_k}} \\ \left(\left\| \mathbf{c}_0^{k,*} \right\|^2 + \left\| \boldsymbol{\lambda}_0^{k,*} \right\|^2 \right) \end{gathered} \tag{2}$$

(the minimization is done over all saddle-points of the nonregularized Lagrange functions).

Now the main problem considered in this paper can be formulated as follows: *for the given zero-sum stochastic game with incomplete information, find an adaptive non-cooperative equilibrium strategies* $\bar{\mathsf{D}}^1$ *and* $\bar{\mathsf{D}}^2$ *using only the on-line (current) information.* To achieve this objective, the recursive procedures (adaptive policies)

$$\mathbf{c}_{n+1}^k = \mathbf{c}_{n+1}^k \left(x_n^k, u_n^k, \eta_n, x_{n+1}^k, \mathbf{c}_n^k; k = 1, 2 \right)$$

should be constructed generating the sequences $\left\{\mathbf{c}_n^k\right\}$ which converges in some probability sense to the solutions $c_0^{k,**}$ defined by (2). In this process the player's actions are selected according to the randomized (mixed) rule

$$d_n^{k,il} = c_n^{k,il} / \sum_{l=1}^{N_k} c_n^{k,il}$$

(the denominator must be strictly positive). The need of randomization is intuitively obvious when playing against an intelligent opponent.

4 Adaptive Strategy

Consider the following iterative procedure:

Step 1. Use the available data

$$\begin{array}{rcl} x_n^k & = & x^k(\alpha_k), u_n^k = u^k(\beta_k), \eta_n \\ x_{n+1}^k & = & x^k(\gamma_k), \mathbf{c}_n^k \ \left(c_n^{k,il} > 0\right), \boldsymbol{\lambda}_n \end{array}$$

to built the following functions

$$\xi_n^k := \eta_n - (-1)^k \left(\lambda_{\alpha_k}^k - \lambda_{\gamma_k}^k \right) + (-1)^k \delta_n c_n^{k,\alpha_k \gamma_k}$$

and use the following normalization procedure

$$\begin{array}{c}\zeta_n^1 := 1-\left(a_n^1\xi_n^1+b_n^1\right)/c_n^{1,\alpha_k\gamma_k}\\ \zeta_n^2 := \left(a_n^2\xi_n^2+b_n^2\right)/c_n^{2,\alpha_k\gamma_k}\end{array} \tag{3}$$

$$a_n^k := \left(2\frac{(\sigma+2\lambda_n^+)}{\varepsilon_n}+\frac{N_kK_k}{N_kK_k-1}\delta_n\right)^{-1}$$
$$b_n^k := a_n^k\left(\sigma+2\lambda_n^+\right)$$

Step 2. Calculate the elements $c_{n+1}^{k,il}$ and $\lambda_{n+1}^{k,j}$ as

$$c_{n+1}^{k,il} = \begin{cases} c_n^{k,\alpha_k\gamma_k}+\gamma_n^{k,c}(1-c_n^{k,\alpha_k\gamma_k}-\zeta_n^k) \\ (i=\alpha_k, j=\gamma_k)\wedge l=\beta_k \\ c_n^{k,il}-\gamma_n^{k,c}\left(c_n^{k,il}-\dfrac{\zeta_n^k}{N_kK_k-1}\right) \\ (i\neq\alpha_k, j\neq\gamma_k)\vee l\neq\beta_k \end{cases}$$

$\gamma_n^{k,c}\in[0,1]$, and the Lagrange multipliers are adjusted as follows:

$$\lambda_{n+1}^{k,j} = \left[\lambda_n^{k,j}+\gamma_n^{\lambda}\psi_n^{k,j}\right]_{-\lambda_{n+1}^+}^{\lambda_{n+1}^+}$$

$$\psi_n^{k,j} = (-1)^k\left[\sum_{l=1}^N c_n^{k,jl}-\chi\left(x^k(j)=x_{n+1}^k\right)+\delta_n\lambda_n^{k,j}\right]$$

$$[y]_{-\lambda_{n+1}^+}^{\lambda_{n+1}^+} := \begin{cases} y & \text{if } y\in\left[-\lambda_{n+1}^+,\lambda_{n+1}^+\right] \\ \lambda_{n+1}^+ & \text{if } y>\lambda_{n+1}^+ \\ -\lambda_{n+1}^+ & \text{if } y<-\lambda_{n+1}^+ \end{cases} \tag{4}$$

Step 3. Construct the stochastic matrices

$$d_{n+1}^{k,il} = c_{n+1}^{k,il}\left(\sum_{r=1}^{N_k}c_{n+1}^{k,ir}\right)^{-1}\quad\left(i=\overline{1,K_k}, l=\overline{1,N_k}\right)$$

and according to

$$\Pr\left\{u_{n+1}^k=u^k(l)\mid x_{n+1}^k=x^k(\gamma_k)\wedge\mathcal{F}_n\right\}=d_{n+1}^{k,\gamma l}$$

generate randomly new discrete random variables u_{n+1}^k as in learning stochastic automata implementation [7], and get the new observation (realization) η_{n+1} which corresponds to the transition to states x_{n+1}^k.

Step 4 return to Step 1.

The positive sequences $\{\varepsilon_n\}$, $\{\delta_n\}$, $\left\{\lambda_n^+\right\}$, $\left\{\gamma_n^{k,c}\right\}$ and $\left\{\gamma_n^{\lambda}\right\}$ will be defined in what follows.

The normalization procedure (3) is an affine transformation. It is a kind of mapping or projection scheme for obtaining a new variable, namely ζ_n^k which belongs to the unit segment in order to preserve the probability measure. In view of Lemma 2 in Chapter 2 [9] it follows that if $\gamma_n^{k,c}\in[0,1]$, $\mathbf{c}_n^k\in S_{\varepsilon_n}^{K_kN_k}$, i.e., $c_n^{\alpha\beta}\geq\varepsilon_n$ and $\delta_n\downarrow 0$, then $\zeta_n^k\in\left[\zeta_n^{k,-},\zeta_n^{k,+}\right]\subset[0,1]$ where

$$\begin{cases} \zeta_n^{1,-}=1-a_n^1\delta_n,\ \zeta_n^{1,+}=a_n^1\delta_n/(N_1K_1-1) \\ \zeta_n^{2,-}=a_n^2\delta_n,\ \zeta_n^{2,+}=1-a_n^2\delta_n/(N_2K_2-1) \\ \mathbf{c}_{n+1}^k\in S_{\varepsilon_{n+1}}^{K_kN_k}\end{cases}$$

5 Convergence Analysis

Introduce the following Lyapunov function

$$W_n := \sum_{k=1}^2\left(\left\|\mathbf{c}_n^k-\mathbf{c}_{\delta_n}^{k,*}\right\|^2+\left\|\boldsymbol{\lambda}_n^k-\boldsymbol{\lambda}_{\delta_n}^{k,*}\right\|^2\right)$$

starting from

$$n\geq\inf_{t\geq 1}\left\{t:\max_k\left\|\boldsymbol{\lambda}_{\delta_n}^{k,*}\right\|\leq\lambda_n^+\right\}$$

Theorem 5 *If the suggested adaptive control is used with the design parameters ε_n, δ_n, λ_n^+, $\gamma_n^{k,c}$ and γ_n^{λ} satisfying the following conditions*

$$0<\delta_n\downarrow 0,\ 0<\lambda_n^+\uparrow\infty,\ \gamma_n^{k,c}\in(0,1)$$

$$\gamma_n^{\lambda}=\frac{N_kK_k}{N_kK_k-1}\gamma_n^{k,c}a_n^k,\ \sum_{n=1}^{\infty}\gamma_n^{k,c}\varepsilon_n\delta_n\left(\lambda_n^+\right)^{-1}=\infty$$

then

1) if

$$\sum_{n=1}^{\infty}\mu_n<\infty$$

where

$$\mu_n := (\delta_n-\delta_{n+1})^2\lambda_n^+\left(\min_k\gamma_n^{k,c}\varepsilon_n\delta_n\right)^{-1}+\max_k\left(\gamma_n^{k,c}\right)^2$$
$$+|\delta_n-\delta_{n+1}|\max_k\gamma_n^{k,c}\left(1+\left(1+\delta_n\lambda_n^+\right)^2\varepsilon_n/\lambda_n^+\right)$$

then, for any fixed distribution of the initial states of the Markov chains associated to the players, the control policy converges with probability one to the equilibrium point of the game, that is, with probability one

$$W_n\to 0$$

2) if

$$\frac{\mu_n\lambda_n^+}{\varepsilon_n\delta_n\min_k\gamma_n^{k,c}}\to 0$$

then, we obtain the convergence in the mean squares sense:

$$E\{W_n\}\to 0$$

Proof: The regularized Lagrange function (**??**), is strictly concave with respect to $\mathbf{c}^1$ and $\boldsymbol{\lambda}^2$ and strictly convex with respect to $\mathbf{c}^2$ and $\boldsymbol{\lambda}^1$, hence (see Lemma 1 Chapter 4 in [8])

$$\begin{array}{c}\left(\boldsymbol{\lambda}^1-\boldsymbol{\lambda}_{\delta}^{1,*}\right)^T\nabla_{\boldsymbol{\lambda}^1}L_{\delta}\left(\mathbf{c}^1,\mathbf{c}^2,\boldsymbol{\lambda}^1,\boldsymbol{\lambda}^2\right) \\ -\left(\mathbf{c}^1-\mathbf{c}_{\delta}^{1,*}\right)^T\nabla_{\mathbf{c}^1}L_{\delta}\left(\mathbf{c}^1,\mathbf{c}^2,\boldsymbol{\lambda}^1,\boldsymbol{\lambda}^2\right) \\ +\left(\mathbf{c}^2-\mathbf{c}_{\delta}^{2,*}\right)^T\nabla_{\mathbf{c}^2}L_{\delta}\left(\mathbf{c}^1,\mathbf{c}^2,\boldsymbol{\lambda}^1,\boldsymbol{\lambda}^2\right) \\ -\left(\boldsymbol{\lambda}^2-\boldsymbol{\lambda}_{\delta}^{2,*}\right)^T\nabla_{\boldsymbol{\lambda}^2}L_{\delta}\left(\mathbf{c}^1,\mathbf{c}^2,\boldsymbol{\lambda}^1,\boldsymbol{\lambda}^2\right) \\ \geq(\delta/2)\sum_{k=1}^2\left(\left\|\mathbf{c}^k-\mathbf{c}_{\delta}^{k,*}\right\|^2+\left\|\boldsymbol{\lambda}^k-\boldsymbol{\lambda}_{\delta}^{k,*}\right\|^2\right)\end{array} \tag{5}$$

Making use of the identity

$$E\left\{Z \mid x_n = x(\alpha) \wedge \mathsf{F}_{n-1}\right\} = E\left\{Z\frac{\chi(x_n=x(\alpha))}{p(\alpha)} \mid \mathsf{F}_{n-1}\right\},\ p(\alpha) > 0$$

it follows

$$\begin{array}{c} E\left\{\zeta_n^1 \mathbf{e}\left(x_n^1 \wedge u_n^1\right) \mid \mathsf{F}_{n-1}\right\} \overset{a.s.}{=} \\ 1 - a_n^1 \frac{\partial}{\partial \mathbf{c}^1} L_\delta\left(\mathbf{c}^1, \mathbf{c}^2, \boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2\right) - b_n^1 \mathbf{e}^{N_1 K_1} \\ E\left\{\zeta_n^2 \mathbf{e}\left(x_n^2 \wedge u_n^2\right) \mid \mathsf{F}_{n-1}\right\} \overset{a.s.}{=} \\ a_n^2 \frac{\partial}{\partial \mathbf{c}^2} L_\delta\left(\mathbf{c}^1, \mathbf{c}^2, \boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2\right) + b_n^2 \mathbf{e}^{N_2 K_2} \end{array}$$

$$E\left\{\boldsymbol{\psi}_n^k \mid \mathsf{F}_{n-1}\right\} \overset{a.s.}{=} (-1)^k \frac{\partial}{\partial \lambda^k} L_\delta\left(\mathbf{c}^1, \mathbf{c}^2, \boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2\right)$$

$$\boldsymbol{\psi}_n^k := \left(\psi_n^1, ..., \psi_n^{K_k}\right)^T$$

where the $N_k K_k-$dimensional vector $\mathbf{e}(x_n \wedge u_n)$ is defined as follows

$$\mathbf{e}\left(x_n^k \wedge u_n^k\right) := \left[\begin{array}{l} \chi\left(x_n^k = x^k(1), u_n^k = u^k(1)\right) \\ ... \\ \chi\left(x_n^k = x^k(K_k), u_n^k = u^k(1)\right) \\ ... \\ \chi\left(x_n^k = x^k(1), u_n^k = u^k(l)\right) \\ ... \\ \chi\left(x_n^k = x^k(K_k), u_n^k = u^k(l)\right) \\ ... \\ \chi\left(x_n^k = x(K_k), u_n^k = u(N_k)\right) \end{array}\right]$$

To make the analysis easy, let us rewrite the algorithm in a vector form

$$\begin{array}{c} \mathbf{c}_{n+1}^k = \mathbf{c}_n^k + \gamma_n^{k,c}\left(\mathbf{e}\left(x_n^k \wedge u_n^k\right) - \mathbf{c}_n^k \right. \\ \left. + \zeta_n^k \frac{\mathbf{e}^{N_k K_k} - N_k K_k \mathbf{e}\left(x_n^k \wedge u_n^k\right)}{N_k K_k - 1}\right) \end{array}$$

$$\boldsymbol{\lambda}_{n+1}^k = \left[\boldsymbol{\lambda}_n^k + \gamma_n^\lambda \boldsymbol{\psi}_n^k\right]_{-\lambda_{n+1}^+}^{\lambda_{n+1}^+}$$

$$\boldsymbol{\psi}_n^k = (-1)^k \left[\left(\begin{array}{c} \sum_{l=1}^{N_k} c_n^{k,1l} \\ . \\ . \\ \sum_{l=1}^{N_k} c_n^{k,K_k l} \end{array}\right) + \delta_n \boldsymbol{\lambda}_n^k - \mathbf{e}\left(x_{n+1}^k\right)\right]$$

where $\mathbf{e}\left(x_{n+1}^k\right)$ is defined as follows: for $x_{n+1}^k = x^k(i_k)$

$$\mathbf{e}\left(x_{n+1}^k\right) := \left[\underbrace{0, ..., 0, 1}_{i_k}, 0, ..., 0\right]^T \in R^{K_k}$$

Therefore, using the property of the projection operator (4)

$$\begin{array}{c} \left\| \left[\boldsymbol{\lambda}_n^k + \gamma_n^{k,\lambda} \boldsymbol{\psi}_n^k\right]_{-\lambda_{n+1}^+}^{\lambda_{n+1}^+} - \boldsymbol{\lambda}_{\delta_{n+1}}^{k,*} \right\| \\ \leq \left\| \boldsymbol{\lambda}_n^k + \gamma_n^\lambda \boldsymbol{\psi}_n^k - \boldsymbol{\lambda}_{\delta_{n+1}}^{k,*} \right\| \end{array}$$

and taking into account that

$$\left\|\boldsymbol{\psi}_n^k\right\| \leq Const\ \gamma_n^\lambda\left(1 + \delta_n \lambda_n^+\right)$$

it follows

$$\begin{array}{c} W_{n+1} \leq W_n + Const \cdot \mu_{1,n}\sqrt{W_n} + Const \cdot \mu_{2,n} + \varphi_n \\ \mu_{1,n} := |\delta_n - \delta_{n+1}| \\ \mu_{2,n} := \sum_{k=1}^{2}\left(\gamma_n^{k,c}\right)^2 + (\delta_n - \delta_{n+1})^2 + \\ + \left(1 + \delta_n \lambda_n^+\right)^2 \left(\gamma_n^\lambda\right)^2 \\ + |\delta_n - \delta_{n+1}| \left(\sum_{k=1}^{2} \gamma_n^{k,c} + \gamma_n^\lambda\left(1 + \delta_n \lambda_n^+\right)\right) \end{array}$$

$$\begin{array}{c} \varphi_n := 2\sum_{k=1}^{2} \gamma_n^{k,c}\left(\mathbf{c}_n^k - \mathbf{c}_{\delta_n}^{k,*}\right)^T \left(\mathbf{e}\left(x_n^k \wedge u_n^k\right) - \mathbf{c}_n^k \right. \\ \left. + \zeta_n^k \frac{\mathbf{e}^{N_k K_k} - N_k K_k \mathbf{e}\left(x_n^k \wedge u_n^k\right)}{N_k K_k - 1}\right) \\ + 2\gamma_n^\lambda \sum_{k=1}^{2}\left(\boldsymbol{\lambda}_n^k - \boldsymbol{\lambda}_{\delta_n}^{k,*}\right)^T \boldsymbol{\psi}_n^k \end{array}$$

Notice that

$$\left(\mathbf{c}_n^k - \mathbf{c}_{\delta_n}^{k,*}\right)^T \frac{\mathbf{e}^{N_k K_k}}{N_k K_k - 1} = 0$$

The assumptions of this theorem, and the strict convex property (5) lead to

$$E\left\{\varphi_n \mid \mathsf{F}_{n-1}\right\} \overset{a.s.}{=} \leq -\gamma_n^\lambda \delta_n W_n$$

The last inequality implies

$$\begin{array}{c} E\left\{W_{n+1} \mid \mathsf{F}_{n-1}\right\} \overset{a.s.}{\leq} \left(1 - \gamma_n^\lambda \delta_n\right) W_n \\ + Const\left(\mu_{1,n}\sqrt{W_n} + \mu_{2,n}\right) \end{array}$$

In view of

$$2\mu_{1,n}\sqrt{W_n} \leq \mu_{1,n}^2 \rho_n^{-1} + W_n \rho_n$$

(which is valid for any $\rho_n > 0$) for $\rho_n := \gamma_n^\lambda \delta_n$ one establishes

$$2\mu_{1,n}\sqrt{W_n} \leq \left(\mu_{1,n}\right)^2 \left(\gamma_n^\lambda \delta_n\right)^{-1} + W_n \gamma_n^\lambda \delta_n$$

In view of the following estimation

$$\left(\mu_{1,n}\right)^2 \left(\gamma_n^\lambda \delta_n\right)^{-1} + \mu_{2,n} \leq Const \cdot \mu_n$$

from the previous inequality we finally, obtain

$$E\left\{W_{n+1} \mid \mathsf{F}_{n-1}\right\} \overset{a.s.}{\leq} \left(1 - \gamma_n^\lambda \delta_n\right) W_n + Const \cdot \mu_n \quad (6)$$

Observe that $a_n^k = O\left(\varepsilon_n / \lambda_n^+\right)$. It follows from (6) that $\{W_n, \mathsf{F}_n\}$ is a nonnegative quasimartingale. Appealing to the assumptions of this theorem, and Robbins-Siegmund theorem [9])the convergence with probability one is obtained . ■

Corollary 6 *If the design parameters of the adaptive algorithm belong to the following class of parameters*

$$\varepsilon_n := \frac{\varepsilon_0}{1+n^{\varepsilon}\ln n}$$
$$\left(\varepsilon_0 \in \left[0, \left(\max_k N_k K_k\right)^{-1}\right), \varepsilon \geq 0\right)$$

$$\delta_n := \frac{\delta_0}{n^{\delta}} \ \ (\delta_0, \delta > 0)$$
$$\lambda_n^+ := \lambda_0^+ \left(1 + n^{\lambda}\ln n\right) \ \ \left(\lambda_0^+ > 0, \ \lambda \geq 0\right)$$
$$\gamma_n^{k,c} := \frac{\gamma_0^k}{n^{\gamma}} \ \ \left(\gamma_0^k \in (0,1), \ \gamma \geq 0\right)$$

with

$$\gamma + \varepsilon + \lambda + \delta \leq 1, \ \delta \leq \varepsilon$$

then

1. *the convergence with probability one is achieved if*
$$1 - \varepsilon - \lambda + \delta > \gamma > 1/2$$

2. *the mean squares convergence is ensured if*
$$\min\left\{2 - 2\left(\gamma + \varepsilon + \lambda\right), \gamma - \varepsilon - \lambda - \delta, 1 + 2\left(\delta - \lambda\right)\right\} > 0$$

Theorem 7 *Assume that the conditions of the previous theorems are fulfilled, then for the class of parameters defined in the previous corollary, there exists* $\nu > 0$ *such that*

$$W_n \overset{a.s.}{=} o\left(\frac{1}{n^{\nu}}\right)$$
$$0 < \nu < \min\left\{1 - \gamma - \varepsilon - \lambda + \delta; \ 2\gamma - 1; 2\delta\right\}$$
$$:= \nu^*(\gamma, \varepsilon, \delta, \lambda)$$

The proof of this theorem is essentially that of Theorem 2 of chapter 2 in [8].

The optimal convergence rate is given by the next corollary.

Corollary 8 *The following optimal parameters* $\varepsilon^*, \delta^*, \lambda^*$ *and* γ^*

$$\varepsilon = \varepsilon^* = \delta = \delta^* = \frac{1}{6}, \ \lambda = \lambda^* = 0, \gamma = \gamma^* = \frac{2}{3}$$

lead to the maximum convergence rate

$$\nu^*(\gamma^*, \varepsilon^*, \delta^*, \lambda^*) = 2\gamma^* - 1 = \frac{1}{3}$$

References

[1] Gardner R. *Games for Business and Economics.* John Willey & Sons, Inc. 1995.

[2] Filar, J. and Vrieze, K. (1979). *Competitive Markov Decision Processes.* Springer-Verlag, Berlin.

[3] Flesch, J., Thuijsman, F. and Vrieze, O. J. (1997). Markov strategies are better than stationary strategies. *Report M 97-09*, Maastricht University, Mathematics, The Netherlands.

[4] Krausz, A. and Rieder, U. (1997). Markov Games with Incomplete Information. *Mathematical Methods of Operations Research*, 46, 263-279.

[5] Mertens, J. F. and Neyman, A. (1981). Stochastic games. *International Journal of Game Theory*, 10, 53-66.

[6] Mertens, J. F., Sorin, S. and Zamir, S. (1994). On stochastic games. Core Discussion Paper 9420, 9421, 9422, Université Catholique de Louvain.

[7] Poznyak, A. S. and Najim, K. (1997). *Learning automata and stochastic optimization.* Springer-Verlag, Berlin.

[8] Poznyak, A. S. and Najim, K. (1999). Adaptive control of constrained finite Markov chains, *Automatica, 35,* May.

[9] Poznyak, A. S., Najim, K. and Gomez, E. (1999). *Self-learning control of finite Markov chains.* Marcel Dekker, New York.

[10] Raghavan, T. E. S. and Filar, J. A. (1991). Algorithms for stochastic games - A survey. *ZOR - Methods and Models for Operations Research*, 35, 437-472.

[11] Raghavan, T. E. S., Ferguson, T. S., Parthasarathy, T. and Vrieze, O. J. (1991). *Stochastic games and related topics.* Kluwer Academic Publishers, London.

[12] Thuijsman, F. and Vrieze O. J. (1993). Stationary ε-optimal strategies in stochastic games. *OR Spectrum*, 15, 9-15.

[13] Thuijsman, F. (1997). A survey on optimality and equilibria in stochastic games. *Ten years LNMB,* Klein et al. (eds.), CWI-Tract n^o. 122, (Center for Mathematics and Computer Science, Amsterdam).

[14] Vrieze O. J. and Thuijsman, F. (1987). Stochastic games and optimal stationary strategies. *Methods of Operations Research*, eds, Domschke, W., Krabs, W., Lehn, J. and Sppelluci, 513-529.