ON THE LEARNING ALGORITHM OF 2-PERSON ZERO-SUM MARKOV GAME WITH EXPECTED AVERAGE REWARD CRITERION

Tanaka, Kensuke
Department of Mathematics, Faculty of Science, Niigata, University

https://doi.org/10.5109/13364

出版情報：Bulletin of informatics and cybernetics. 21 (3/4), pp.1-17, 1985-03. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics Vol. 21, No. 3~4, 1985

ON THE LEARNING ALGORITHM OF 2-PERSON ZERO-SUM MARKOV GAME WITH EXPECTED AVERAGE REWARD CRITERION

By

Kensuke TANAKA*

Abstract

We develop a method for learning the optimal strategies of 2-person zero-sum Markov game with expected average reward criterion. To do this, at each stage the players play a modified matrix game with relation to each state, and then receive an information about the result of the game from a teacher. Using the information, the players generate a pair of mixed strategies with relation to each state used at next stage. Then, such a pair of mixed strategies generated by the players converges with probability one and in mean square to a pair of the optimal stationary strategies. Further, when the learning is stopped at some stage by the teacher, the probability of error is estimated.

1. Introduction

This paper is a continuation of our paper [5] with the title "*On the learning algorithm of 2-person zero-sum Markov game*" and is concerned with a learning algorithm. In [5], we showed that a sequence of the mixed strategies generated by some learning algorithm converges to a pair of the optimal stationary strategies of the Markov game with a discount factor under an assumption of incomplete information relating to reward functions and transition probabilities of a system.

However, in some practical problems, it is necessary that we make use of such an algorithm to 2-person zero-sum Markov game with expected average reward criterion. For this reason, it is tried to apply the learning algorithm to such a game under an absence of complete information relating to reward functions and transition probabilities of the system. To construct the learning algorithm under this situation, at each stage two players play the modified matrix game corresponding to all states of the game system and make use of the information about the games in the form of realization of the random variables given by a teacher. Then, we can show that, under some assumption, a pair of the mixed strategies of the players generated by the learning algorithm utilized efficiently the given information converges with probability one and in mean square to a pair of the optimal stationary strategies of the original Markov game. Further, when the teacher makes the players to stop learning at some stage, the pro-

* Department of Mathematics, Faculty of Science, Niigata University, Niigata, Japan

bability, such that the difference of the mixed strategies at this stage from the optimal strategies of the Markov game is greater than $\varepsilon>0$, is estimated. In the proof of such convergence, the idea of regularization for supplying the lack of the strict convexity of the payoff functions in the modified matrix game and an assumption for ensuring the convergence of an approximate game value generated sequentially by the teacher play important roles.

This paper consists of four sections. In section 2, we shall give the knowledge about a 2-person zero-sum Markov game with expected average reward criterion necessary in the paper. In section 3, we shall state the regularization of the modified matrix game and show the properties of optimal mixed strategies by form of the lemmas. In section 4, we shall give the formulation of learning system and show that a pair of the mixed strategies generated by the learning algorithm converges with probability one and in mean square to a pair of the optimal stationary strategies under some conditions. Further, when learning is stopped at some stage, the probability of error is estimated.

2. Preliminaries

In this section, we consider a 2-person zero-sum Markov game with the expected average reward criterion defined by a set of five objects (S, A, B, P, r). Here, S is a finite set labeled $\{1, 2, \cdots, s\}$, called the state space of the game system; $A=\bigcup_{l=1}^{s} A_l$ is called the action space for player I and A_l is a finite set $\{a_l^1, a_l^2, \cdots, a_l^{m_l}\}$, from which player I will choose his action when the state of the game system is $l\in S$; $B=\bigcup_{l=1}^{s} B_l$ is called the action space for player II and B_l is a finite set labeled $\{b_l^1, b_l^2, \cdots, b_l^{n_l}\}$, from which player II will choose his action at state $l\in S$, P is a transition probability which governs the law of the motion of the system, that is, for each triple (l, a_l, b_l) $\in S\times\bigcup_{l=1}^{s} A_l\times\bigcup_{l=1}^{s} B_l$, corresponds to a probability on S; r, a reward function of player I, is a function $r(l, a_l, b_l)$ on $S\times\bigcup_{l=1}^{s} A_l\times\bigcup_{l=1}^{s} B_l$ and $-r$ is a reward function of player II.

In such a game, player I and II observe the state of the system at each stage and classify it to one of the possible states $l\in S$ and then, player I and II choose independently actions $a_l\in A_l$ and $b_l\in B_l$ by the mixed strategies, respectively, without any collaboration with any others. Then, as a consequence of the present state $l\in S$ and the actions $a_l\in A_l$ and $b_l\in B_l$ chosen by the players, player II pays player I reward $r(l, a_l, b_l)$ and the game system moves to a new state $l'\in S$ according to the transition probability $P(l'|l, a_l, b_l)$.

A strategy π for player I is a sequence of $\pi_0, \pi_1, \cdots$, in which each π_n specifies a probability $\pi_n(\cdot|h_n)$ on A_{l_n} under the given history $h_n=(l_0, a_0, b_0, \cdots, a_{n-1}, b_{n-1}, l_n)$ of the system, where l_t, a_t and b_t are the t-th state, the t-th action chosen by player I and t-th action chosen by player II, respectively. Especially, if each element π_n of $\pi=(\pi_0, \pi_1, \pi_2, \cdots)$ depends only on the n-th state of the system, the strategy π is said to be Markov strategy. Moreover, if each element π_n of Markov strategy $\pi=(\pi_0, \pi_1, \cdots)$ is independent of n, the strategy π is said to be stationary. In this case, there is a

map f from S into $P(A)$ such that $\pi_n=f$ for all $n=0, 1, 2, \cdots$ and π is denoted by f, where $P(A)=\bigcup_{l=1}^{s}P(A_l)$ and $P(A_l)$ is the set of all probabilities on A_l. Π denotes the class of all strategies for player I. Strategies, Markov strategies and stationary strategies for player II are defined analogously. Γ denotes the class of all strategies for player II.

Now, we define the expected average gain for player I. When the game system starts from a state $l_0\in S$ and a pair of the strategies (π, σ) is used, then the total expected gain for player I up to the n-th transition is defined to be

$$I_n(\pi, \sigma)=E_{\pi,\sigma}\Big[\sum_{t=0}^{n}r(l_t, a_t, b_t)\,|\,x_0=l_0\Big],$$

where $E_{\pi,\sigma}[\cdot\,|\,x_0=l_0]$ denotes the conditional expectation given $x_0=l_0$ related to a pair of strategies (π, σ). Then, player I wants to maximize

$$\overline{\lim_{n\to\infty}}\frac{I_n(\pi, \sigma)(l_0)}{n+1}$$

and player II wants to minimize

$$\underline{\lim_{n\to\infty}}\frac{I_n(\pi, \sigma)(l_0)}{n+1}.$$

From a point of view of this criterion, an optimal strategy π^* for player I can be defined as follows: for all strategies σ' for player II and all $l_0\in S$,

$$\inf_{\sigma\in\Gamma}\sup_{\pi\in\Pi}\overline{\lim_{n\to\infty}}\frac{I_n(\pi, \sigma)(l_0)}{n+1}\leqq\underline{\lim_{n\to\infty}}\frac{I_n(\pi^*, \sigma')(l_0)}{n+1}.$$

Similarly, for all strategies π' for player I and all $l_0\in S$, an optimal strategy σ^* for player II can be defined as

$$\sup_{\pi\in\Pi}\inf_{\sigma\in\Gamma}\underline{\lim_{n\to\infty}}\frac{I_n(\pi, \sigma)(l_0)}{n+1}\geqq\overline{\lim_{n\to\infty}}\frac{I_n(\pi', \sigma^*)(l_0)}{n+1}.$$

Further, we shall say that the Markov game has a value if for all initial state $l_0\in S$

$$\inf_{\sigma\in\Gamma}\sup_{\pi\in\Pi}\overline{\lim_{n\to\infty}}\frac{I_n(\pi, \sigma)(l_0)}{n+1}=\sup_{\pi\in\Gamma}\inf_{\pi\in\Pi}\underline{\lim_{n\to\infty}}\frac{I_n(\pi, \sigma)(l_0)}{n+1}.$$

This common quantity as a function on S is called the value function of the game.

Let X^s be an s-dimensional vector space. For each $p_l=(p_l(1), p_l(2), \cdots, p_l(m_l))\in P(A_l)$ and $q_l=(q_l(1), q_l(2), \cdots, q_l(n_l))\in P(B_l)$, we define an operator $L(p_l, q_l)\colon X^s\to X^s$ as follows: for each $l\in S$ and $u=(u_1, u_2, \cdots, u_s)\in X^s$,

$$L(p_l, q_l)u_l=r(l, p_l, q_l)+\sum_{l'=1}^{s}u_{l'}P(l'\,|\,l, p_l, q_l),$$

where

$$r(l, p_l, q_l)=\sum_{i=1}^{m_l}\sum_{j=1}^{n_l}r(l, a_l^i, b_l^j)p_l(i)q_l(j)$$

and

$$P(l'\,|\,l, p_l, q_l)=\sum_{i=1}^{m_l}\sum_{j=1}^{n_l}P(l'\,|\,l, a_l^i, b_l^j)p_l(i)q_l(j)$$

Then, since for each $l\in S$ and $u\in X^s$ $L(p_l, q_l)u_l$ is continuous on $P(A_l)\times P(B_l)$ and $P(A_l)\times P(B_l)$ is compact, there exist the probabilities $p_l^*\in P(A_l)$ and $q_l^*\in P(B_l)$ such that, for each $u\in X^s$,

$$\begin{aligned}\min_{q_l} L(p_l^*, q_l)u_l &= \max_{p_l}\min_{q_l} L(p_l, q_l)u_l \\ &= \min_{q_l}\max_{p_l} L(p_l, q_l)u_l \\ &= \max_{p_l} L(p_l, q_l^*)u \\ &= L(p_l^*, q_l^*)u_l .\end{aligned}$$

we can prove the following theorem.

THEOREM 2.1. *If there exist a vector $v^*\in X^s$ and a constant d such that, for all $l\in S$,*

$$d+v_l^* = \max_{p_l}\min_{q_l} L(p_l, q_l)v^*, \tag{2.2}$$

then, the Markov game has a value d, i.e., for all $l_0\in S$,

$$d=\max_{\pi\in\Pi}\min_{\sigma\in\Gamma}\varliminf_{n\to\infty}\frac{I_n(\pi, \sigma)(l_0)}{n+1} = \min_{\sigma}\max_{\pi}\varlimsup_{n\to\infty}\frac{I_n(\pi, \sigma)(l_0)}{n+1}$$

and there exist the optimal stationary strategies p_l^* and q_l^* of the players satisfying (2.1) *with $v^*\in X^s$ instead of $u\in X^s$.*

The proof of the theorem is given in [2].

Under the following assumption, a solution to (2.2) is determined by the method of successive approximations.

ASSUMPTION. There exist an integer $u\geqq 0$, a constant $\alpha(0<\alpha\leqq 1)$ and a state $l^*\in S$ such that, for all $(u+1)$-step Markov strategies $\pi^{u+1}=(f_0, f_1, \cdots, f_u)$ and $\sigma^{u+1}=(g_0, g_1, \cdots, g_u)$ of the players and all initial states $l_0\in S$,

$$\begin{aligned}&P^{u+1}(l^*|l_0, \pi^{u+1}, \sigma^{u+1}) \\ &\quad = \sum_{l_1=1}^{s} P(l_1|l_0, f_0(l_0), g_0(l_0))\cdots\sum_{l_u=1}^{s} P(l_u|l_{u-1}, f_{u-1}(l_{u-1}), \\ &\qquad g_{u-1}(l_{u-1}))P(l^*|l_u, f_u(l_u), g_u(l_u)) \\ &\quad \geqq \alpha > 0 .\end{aligned}$$

THEOREM 2.2. *Under the assumption, make a sequence $\{(d^{(n)}, v^{(n)})\}_{n=1,2,\cdots}$, according to the following algorithm:*

$$\begin{aligned}&V_l^{(n)} = \max_{p_l}\min_{q_l}\{r(l, p_l, q_l)+\sum_{l'=1}^{s} v_{l'}^{(n-1)}P(l'|l, p_l, q_l)\} \\ &d^{(n)} = V_{l^*}^{(n)} \\ &v_l^{(n)} = V_l^{(n)} - d^{(n)}, \qquad n\geqq 1 ,\end{aligned}$$

then $\{(d^{(n)}, v^{(n)})\}$ converges uniformly and exponentially fast to a solution (d, v^*) of (2.2), *where $v^{(0)}\in X^s$ and l^* is given in the assumption.*

This theorem is proved by a similar argument in [3].

We state more detailed results for the learning algorithm used in this paper. Let

$$\nabla^{(n)}(v)=\min_{l\in S}[v_l^{(n+1)}-v_l^{(n)}]$$

$$\Delta^{(n)}(v)=\max_{l\in S}[v_l^{(n+1)}-v_l^{(n)}]$$

$$D^{(n)}(v)=\Delta^{(n)}(v)-\nabla^{(n)}(v)$$

and

$$U^{(n)}(v)=\max_{l\in S}|v_l^{(n+1)}-v_l^{(n)}|.$$

For any $n=Nu+m$ $(1\leqq m\leqq u)$,

$$U^{(n+1)}(V)\leqq U^{(n)}(v)\leqq D^{(n)}(v)\leqq(1-\alpha)^N D^{(m)}(v)\leqq(1-\alpha)^N A,\ n\geqq 1,$$

where

$$A=\max\{D^{(1)}(v),\ D^{(2)}(v),\ \cdots,\ D^{(u)}(v)\}.$$

Then, from the above facts, it follows that

$$\begin{aligned}\max_{l\in S}|v_l^{(n)}-v_l^*|&\leqq\sum_{t=n}^{\infty}U^{(t)}(v)\\&\leqq(1-\alpha)^N\frac{uA}{\alpha}\\&\leqq(1-\alpha)^N B,\end{aligned}\tag{2.3}$$

where $n=Nu+m$, $1\leqq m\leqq u$, and $B=uA/\alpha$.

As mentioned in Theorem 2.1 and Theorem 2.2, it is important to consider a 2-person zero-sum game, which may be called a modified matrix game at each state $l\in S$ with the following payoff function: for all $a_l^i\in A_l$ and $b_l^j\in B_l$,

$$V_l^*(a_l^i,\ b_l^j)=r(l,\ a_l^i,\ b_l^j)+\sum_{l'=1}^{s}v_{l'}^*P(l'\,|\,l,\ a_l^i,\ b_l^j),$$

where $v^*=(v_1^*,\ v_2^*,\ \cdots,\ v_s^*)$ is the vector given in Theorem 2.1. The optimal strategies p_l^* and q_l^* for player I and II in this modified game at each state $l\in S$ correspond to them for the players in the original Markov game with the expected average reward criterion. And the constant d given in Theorem 2.1 is the value of the original Markov game.

3. Regularization of the Modified Matrix Game

From the fact mentioned in section 2, it is important that the players search for the optimal stationary strategies at each state l of the modified matrix game. Here, under an absence of complete information about reward functions and transition probabilities of the game system, the players have to search for the optimal stationary strategies by a teacher's guidance. Then, the teacher gives the players the information related to the optimal strategies by making them play the modified matrix game at each state $l\in S$. To do this, the gradient method is used as the learning algorithm for solving such an optimization problem. But, there exist several difficulties that are connected with the lack of strict convexity of the payoff functions. One method of avoiding these difficulties is to introduce an idea of regularization in this modified matrix game.

Suppose that, in the regularized game at each state $l\in S$, when the strategies a_l^i and b_l^j for the players are chosen by the mixed strategies $p_l=(p_l(1), p_l(2), \cdots, p_l(m_l))$ and $q_l=(q_l(1), q_l(2), \cdots, q_l(n_l))$, the payoff functions of the players are

$$V_l^*(a_l^i, b_l^j)-\frac{\delta_l}{2}(p_l(i)-q_l(j))$$

and

$$-V_l^*(a_l^i, b_l^j)-\frac{\delta_l}{2}(q_l(j)-p_l(i)),$$

respectively, $i=1, 2, \cdots, m_l$, $j=1, 2, \cdots, n_l$, where $\delta_l>0$ is a regularization parameter at state l and

$$V_l^*(a_l^i, b_l^j)=r(l, a_l^i, b_l^j)+\sum_{l'=1}^{s} v_{l'}^* P(l'|l, a_l^i, b_l^j).$$

Then, the expected gain of the regularized game for player I at each state l is given by

$$V_{l,\delta_l}^*(p_l, q_l)=V_l^*(p_l, q_l)-\frac{\delta_l}{2}(\|p_l\|^2-\|q_l\|^2)$$

and, similarly, the expected gain for player II is given by

$$-V_{l,\delta_l}^*(p_l, q_l),$$

where $\|\cdot\|$ is Euclidean norm and

$$V_l^*(p_l, q_l)=L(p_l, q_l)v_l^*.$$

Moreover, we assume that the mixed strategies available to the players at each state l are in ε_l-simplices. i.e., $p_l\in S_{\varepsilon_l}^{m_l}$ and $q_l\in S_{\varepsilon_l}^{n_l}$, where

$$S_\varepsilon^m=\{X=(x_1, x_2, \cdots, x_m),\ x_i\geqq\varepsilon,\ i=1, 2, \cdots, m,$$
$$\sum_{i=1}^{m} x_i=1, \qquad \left(0\leqq\varepsilon\leqq\frac{1}{m}\right)\}.$$

In view of this point, $V_{l,\delta_l}^*(p_l, q_l)$ is strictly convex for any fixed $\delta_l>0$. Thus, the game has a unique saddle point $(p_l^*(\varepsilon_l, \delta_l), q_l^*(\varepsilon_l, \delta_l))$ for any fixed $\varepsilon_l\in[0, \varepsilon_l]$, $\varepsilon_l=\min\{1/m_l, 1/n_l\}$, such that for all $p_l\in S_{\varepsilon_l}^{m_l}$ and $q_l\in S_{\varepsilon_l}^{n_l}$

$$V_{l,\delta_l}^*(p^*(\varepsilon_l, \delta_l), q_l)\geqq V_{l,\delta_l}^*(p^*(\varepsilon_l, \delta_l), q^*(\varepsilon_l, \delta_l))$$
$$\geqq V_{l,\delta_l}^*(p_l, q^*(\varepsilon_l, \delta_l)).$$

$p_l^*(\varepsilon_l, \delta_l)$ and $q_l^*(\varepsilon_l, \delta_l)$ are the optimal strategies for the players in the game at state $l\in S$ restricted by $\varepsilon_l>0$ and $\delta_l>0$.

The following two lemmas play an important role in our learning algorithm.

LEMMA 1. *If, for each state l, the sequence $\{\varepsilon_l(n)\}$ and $\{\delta_l(n)\}$ satisfy*

$$\varepsilon_l(n)\in(0, \hat{\varepsilon}_l), \qquad \hat{\varepsilon}_l=\min\left\{\frac{1}{m_l}, \frac{1}{n_l}\right\}$$
$$\delta_l(n)>0, \qquad \min_{n\to\infty}\varepsilon_l(n)=\min_{n\to\infty}\delta_l(n)=0$$

and

$$\min_{n\to\infty}\frac{\varepsilon_l(n)}{\delta_l(n)}=\mu_l\in[0,\ \infty),$$

then, at state l, *the sequence* $\{p_l^*(\varepsilon_l(n),\ \delta_l(n)),\ q_l^*(\varepsilon_l(n),\ \delta_l(n))\}$ *converges to a saddle point* $(p_l^*,\ q_l^*)$ *of the modified matrix game (depending, generally, on* μ_l*).*

LEMMA 2. *For each state* l, *there exists* $\delta_l'\in(0,\ \infty)$ *and constants* $K_l^{(1)}$, $K_l^{(2)}$ *and* $K_l^{(3)}$ *such that*

$$\|p_l^*(\varepsilon_1,\ \delta_1)-p_l^*(\varepsilon_2,\ \delta_2)\|+\|q_l^*(\varepsilon_1,\ \delta_1)-q_l^*(\varepsilon_2,\ \delta_2)\|$$
$$\leqq K_l^{(1)}|\varepsilon_1-\varepsilon_2|+K_l^{(2)}|\delta_1-\delta_2|+K_l^{(3)}\left|\frac{\varepsilon_1}{\delta_1}-\frac{\varepsilon_2}{\delta_2}\right|$$

for any $\varepsilon_1,\ \varepsilon_2\in[0,\ \hat{\varepsilon}_l]$ *and* $\delta_1,\ \delta_2\in(0,\ \delta_l')$, *where* $\|\cdot\|$ *is Euclidean norm.*

The proofs of these lemmas are given in [1].

4. Formulation and Convergence of the Learning Algorithm

In this section, the teacher does not know the value of the original Markov game and the players do not know the reward functions and the transition probabilities of the system. As the method of learning, at each stage, the teacher makes the approximate value of the original game by the successive method in Theorem 2.2 and the players receive the information about the reward functions in the form of realization of some random variables. And the players construct the mixed strategies used at next stage by a pseudogradient method.

For our aim, we describe the learning algorithms in detail.

At the first stage, the teacher chooses any initial vector $v^{(0)}=(v_1^{(0)},\ v_2^{(0)},\ \cdots,\ v_s^{(0)})\in X^s$. Player I and II play a game at each state $l\in S$ using any mixed strategies $p_l^{(0)}$ and $q_l^{(0)}$, respectively.

Now, let $v^{(n)}=(v_1^{(n)},\ v_2^{(n)},\ \cdots,\ v_s^{(n)})\in X^s$ be an s-dimensional vector constructed by the teacher at the n-th stage. Let $p_l^{(n)}$ and $q_l^{(n)}$ be the mixed strategies of each state l constructed by the players at the n-th stage, respectively. Then, at the $(n+1)$-th stage, the players play the modified matrix game at each state $l\in S$ using the mixed strategies $p_l^{(n)}$ and $q_l^{(n)}$ and suppose that at this game, the players choose the pure strategies $x_l^{(n+1)}$ and $y_l^{(+1)}$, respectively. So the teacher gives the players the information about the payoff function in the form of realization of the random variables $V_l^{(n+1)}(1)$ and $V_l^{(n+1)}(2)$ such that

$$\begin{aligned}&E[V_l^{(n+1)}(1)\,|\,x_l^{(n+1)},\ y_l^{(n+1)}]=V_l^{(n+1)}(x_l^{(n+1)},\ y_l^{(n+1)}),\\&\sigma^2[V_l^{(n+1)}(1)\,|\,x_l^{(n+1)},\ y_l^{(n+1)}]=R_l^{(1)}(x_l^{(n+1)},\ y_l^{(n+1)})<\infty,\end{aligned}\tag{4.1a}$$

and

$$\begin{aligned}&E[V_l^{(n+1)}(2)\,|\,x_l^{(n+1)},\ y_l^{(n+1)}]=-V_l^{(n+1)}(x_l^{(n+1)},\ y_l^{(n+1)})\\&\sigma^2[V_l^{(n+1)}(2)\,|\,x_l^{(n+1)},\ y_l^{(n+1)}]=R_l^{(2)}(x_l^{(n+1)},\ y_l^{(n+1)})<\infty,\end{aligned}\tag{4.1b}$$

where for $v_l^{(n)}$ given in Theorem 2.2,

$$V_l^{(n+1)}(x_l^{(n+1)},\ y_l^{(n+1)})=r(l,\ x_l^{(n+1)},\ y_l^{(n+1)})+\sum_{l'=1}^{s}v_{l'}^{(n)}P(l'\,|\,l,\ x_l^{(n+1)},\ y_l^{(n+1)}).$$

Using this information given by the teacher and a projection operator $\Pi_F\{\cdot\}$ on a closed bounded set F, player I and II construct the mixed strategies used in each state $l\in S$ at the next stage as follows:

$$p_l^{(n+1)}=\Pi_{S^{n_l}_{\varepsilon_l(n+1)}}[p_l^{(n)}+\gamma_l^{(n)}A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})] \tag{4.2a}$$

and

$$q_l^{(n+1)}=\Pi_{S^{n_l}_{\varepsilon_l(n+1)}}[q_l^{(n)}+\gamma_l(n)B^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})] \tag{4.2b}$$

where $\{\varepsilon_l(n)\}$, $\{\delta_l(n)\}$ and $\{\gamma_l(n)\}$ are the sequences of numbers and for $a_l^i\in A_l$ and $b_l^j\in B_l$, $i=1, 2, \cdots, m_l$, $j=1, 2, \cdots, n_l$, $A^{(n+1,l)}(a_l^i, b_l^j)$ is an m_l-dimensional vector whose k-th element is

$$A_k^{(n+1,l)}(a_l^i, b_l^j)=\begin{cases} V_l^{(n+1)}(1)/p_l^{(n)}(i)-\delta_l(n), & k=a_l^i \\ -\dfrac{V_l^{(n+1)}(1)/p_l^{(n)}(i)-\delta_l(n)}{m_l-1}, & k\neq a_l^i \end{cases}$$

and $B^{(n+1,l)}(a_l^i, b_l^j)$ is an n_l-dimensional vector whose k-th element is

$$B_k^{(n+1,l)}(a_l^i, b_l^j)=\begin{cases} V_l^{(n+1)}(2)/q_l^{(n)}(j)-\delta_l(n), & k=b_l^j \\ -\dfrac{V_l^{(n+1)}(2)/q_l^{(n)}(j)-\delta_l(n)}{n_l-1}, & k\neq b_l^j \end{cases}$$

the projection operator $\Pi_F\{\cdot\}$ has the following property represented by the Euclidean norm:

$$\Pi_F\{x\}\in F \text{ and } \|x-y\|\geqq\|\Pi_F\{x\}-y\| \quad \text{for all } x \tag{4.3}$$

and all $y\in F$.

Next, the teacher constructs an s-dimensional vector $v^{(n+1)}$ used at the next stage as follows: at each state $l\in S$,

$$V_l^{(n+1)}=\max_{p_l}\min_{q_l}\{r(l, p_l, q_l)+\sum_{l'=1}^{s}v_{l'}^{(n)}P(l'|l, p_l, q_l)\}$$

$$d^{(n+1)}=V_{l^*}^{(n+1)}$$

$$v_l^{(n+1)}=V_l^{(n+1)}-d^{(n+1)}.$$

Moreover, the learning of the players is, sequentially, continued by the guidance of the teacher.

Then, the following theorems assure the purpose of learning, because a pair of the mixed strategies generated by the above algorithm converges with probability one and in mean square to a saddle point (p_l^*, q_l^*) at each state $l\in S$ of the original Markov game. For simplicity, we use the notations $p_l^{(n)*}=p^*(\varepsilon_l(n), \delta_l(n))$ and $q_l^{(n)*}=q^*(\varepsilon_l(n), \delta_l(n))$.

Now, throughout the paper, we assume the assumption.

THEOREM 4.1. *Suppose that the sequences* $\{\varepsilon_l(n)\}$, $\{\delta_l(n)\}$ *and* $\{\gamma_l(n)\}$ *satisfy the following conditions: for each state* $l\in S$,

(a) $\gamma_l(n)>0$, $\delta_l(n)>0$, $\varepsilon_l(n)\in(0, \hat{\varepsilon}_l)$, $n=0, 1, \cdots$,

$\varepsilon_l(n)\longrightarrow 0$ *and* $\delta_l(n)\longrightarrow 0$ *as* $n\longrightarrow\infty$,

where $\hat{\varepsilon}_l(n)=\min\left\{\frac{1}{m_l}, \frac{1}{n_l}\right\}$,

(b) $\lim_{n\to\infty}\frac{\varepsilon_l(n)}{\delta_l(n)}=\mu_l\in[0, \infty)$,

(c) $\sum_{n=0}^{\infty}\gamma_l(n)\delta_l(n)=\infty$,

(d) $\sum_{n=0}^{\infty}\gamma_l^2(n)/\varepsilon_l(n)<\infty$,

(e) $\sum_{n=1}^{\infty}|\varepsilon_l(n)-\varepsilon_l(n-1)|<\infty$,

(f) $\sum_{n=1}^{\infty}|\delta_l(n)-\delta_l(n-1)|<\infty$,

(g) $\sum_{n=1}^{\infty}|\varepsilon_l(n)/\delta_l(n)-\varepsilon_l(n-1)/\delta_l(n-1)|<\infty$.

Then, at each state $l\in S$ *the sequence* $(p_l^{(n)}, q_l^{(n)})$ *generated by the learning algorithm* (4.2) *converges with propability one as* $n\to\infty$ *to a pair of optimal stationary strategies* (p_l^*, q_l^*) *of the Markov game for a pair of any initiall mixed strategies* $(p_l^{(0)}, q_l^{(0)})\in S_{\varepsilon_l(0)}^{m_l}\times S_{\varepsilon_l(0)}^{n_l}$.

REMARK 1. *Since* $\sum_{n=0}^{\infty}\gamma_l^2(n)/\varepsilon_l(n)<\infty$ *and* $\varepsilon_l(n)\to 0$ *as* $n\to\infty$, $\sum_{n=0}^{\infty}\gamma_l^2(n)<\infty$. *So* $\sum_{n=0}^{\infty}\gamma_l^2(n)\delta_l^2(n)<\infty$.

PROOF. By (4.2a) the (4.3), we get for $n\geqq 0$,

$$\begin{aligned}\|p_l^{(n+1)}-p_l^{(n+1)*}\|^2&\leqq\|p_l^{(n)}+\gamma_l(n)A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})-p_l^{(n+1)*}\|^2\\&=\|p_l^{(n)}-p_l^{(n+1)*}\|^2+2\gamma_l(n)\langle p_l^{(n)}-p_l^{(n+1)*}, A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})\rangle\\&\quad+\gamma_l^2(n)\|A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})\|^2,\end{aligned}\tag{4.4}$$

where $\langle\cdot, \cdot\rangle$ denotes inner product and $\|\cdot\|$ denotes Euclidean norm.

Further, it follows that

$$\begin{aligned}\|p_l^{(n)}-p_l^{(n+1)*}\|^2&=\|p_l^{(n)}-p_l^{(n)*}+p_l^{(n)*}-p_l^{(n+1)*}\|^2\\&\leqq(\|p_l^{(n)}-p_l^{(n)*}\|+\|p_l^{(n)*}-p_l^{(n+1)*}\|)^2\\&\leqq\|p_l^{(n)}-p_l^{(n)*}\|^2+3\sqrt{2}\,\|p_l^{(n)*}-p_l^{(n+1)*}\|,\end{aligned}$$

and

$$\begin{aligned}&2\gamma_l(n)\langle p_l^{(n)}-p_l^{(n+1)*}, A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})\rangle\\&\quad=2\gamma_l(n)\langle p_l^{(n)}-p_l^{(n)*}, A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})\rangle\\&\qquad+2\gamma_l(n)\langle p_l^{(n)*}-p_l^{(n+1)*}, A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})\rangle\\&\quad\leqq 2\gamma_l(n)\langle p_l^{(n)}-p_l^{(n)*}, A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})\rangle\\&\qquad+\sqrt{2}\,\|p_l^{(n)*}-p_l^{(n+1)*}\|+\gamma_l^2(n)\|A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})\|^2.\end{aligned}\tag{4.6}$$

By inserting (4.5) the (4.6) into (4.4), we can obtain

$$\|p_l^{(n+1)}-p_l^{(n+1)*}\|^2\leqq\|p_l^{(n)}-p_l^{(n)*}\|^2+4\sqrt{2}\,\|p_l^{(n+1)*}-p_l^{(n+1)*}\|$$

$$+2\gamma_l(n)\langle p_l^{\{n\}}-p_l^{\{n\}*},\ A^{(n+1,\,l)}(x_l^{\{n+1\}},\ y_l^{\{n+1\}})\rangle$$
$$+2\gamma_l^2(n)\|A^{(n+1,\,l)}(x_l^{\{n+1\}},\ y_l^{\{n+1\}})\|^2. \tag{4.7}$$

Taking conditional expectation of (4.7) for given $p_l^{\{n\}}$ and $q_l^{\{n\}}$, we get

$$\begin{aligned}
&E[\|p_l^{\{n+1\}}-q_l^{\{n+1\}*}\|^2|\,p_l^{\{n\}},\ q_l^{\{n\}}]\\
&\quad\leqq\|p_l^{\{n\}}-p_l^{\{n\}*}\|^2+4\sqrt{2}\,\|p_l^{\{n\}*}-p_l^{\{n+1\}*}\|\\
&\qquad+2\gamma_l(n)E[\langle p_l^{\{n\}}-p_l^{\{n\}*},\ A^{(n+1,\,l)}(x_l^{\{n+1\}},\ y_l^{\{n+1\}})\rangle|\,p_l^{\{n\}},\ q_l^{\{n\}}]\\
&\qquad+2\gamma_l^2(n)E[\|A^{(n+1,\,l)}(x_l^{\{n+1\}},\ y_l^{\{n+1\}})\|^2|\,p_l^{\{n\}},\ q_l^{\{n\}}].
\end{aligned}\tag{4.8}$$

Here, from the definition of $A^{(n+1,\,l)}(x_l^{\{n+l\}},\ y_l^{\{n+1\}})$, it follows that

$$\begin{aligned}
&E[\langle p_l^{\{n\}}-p_l^{\{n\}*},\ A^{(n+1,\,l)}(x_l^{\{n+1\}},\ y_l^{\{n+1\}})\rangle|\,p_l^{\{n\}},\ q_l^{\{n\}}]\\
&=E[E[\langle p_l^{\{n\}}-p_l^{\{n\}*},\ A^{(n+1,\,l)}(x_l^{\{n+1\}},\ y_l^{\{n+1\}})\rangle|\,x_l^{\{n+1\}},\ y_l^{\{n+1\}}]|\,p_l^{\{n\}},\ q_l^{\{n\}}]\\
&=\sum_{i=1}^{m_l}\sum_{j=1}^{n_l}\Big\{(p_l^{\{n\}}(i)-p_l^{\{n\}*}(i))(V_l^{\{n+1\}}(a_l^i,\ b_l^j)/p_l^{\{n\}}(i)-\delta_l(n))\\
&\quad+\sum_{k\neq i}(p_l^{\{n\}}(k)-p_l^{\{n\}*}(k))\Big(-\frac{1}{m_l-1}(V_l^{\{n+1\}}(a_l^i,\ b_l^j)/p_l^{\{n\}}(i)-\delta_l(n))\Big)\Big\}\,p_l^{\{n\}}(i)q_l^{\{n\}}(j)\\
&=\frac{m_l}{m_l-1}\Big\{V_l^{\{n+1\}}(p_l^{\{n\}},\ q_l^{\{n\}})-\frac{\delta_l(n)}{2}(\|p_l^{\{n\}}\|^2-\|q_l^{\{n\}}\|^2)\\
&\quad-V_l^{\{n+1\}}(p_l^{\{n\}*},\ q_l^{\{n\}})+\frac{\delta_l(n)}{2}(\|p_l^{\{n\}*}\|^2-\|q_l^{\{n\}}\|^2)-\frac{\delta_l(n)}{2}\|p_l^{\{n\}}\|^2\\
&\quad-\frac{\delta_l(n)}{2}\|p_l^{\{n\}*}\|^2+\delta_l(n)\langle p_l^{\{n\}*},\ p_l^{\{n\}}\rangle\Big\}\\
&=\frac{m_l}{m_l-1}\Big\{V_{l,\delta_l(n)}^{\{n+1\}}(p_l^{\{n\}},\ q_l^{\{n\}})-V_{l,\delta_l(n)}^{\{n+1\}}(p_l^{\{n\}*},\ q_l^{\{n\}})-\frac{\delta_l(n)}{2}\|p_l^{\{n\}}-p_l^{\{n\}*}\|^2\Big\}\\
&=\frac{m_l}{m_l-1}\Big\{(V_{l,\delta_l(n)}^{\{n+1\}}(p_l^{\{n\}},\ q_l^{\{n\}})-V_{l,\delta_l(n)}^{*}(p_l^{\{n\}},\ q_l^{\{n\}}))\\
&\quad-(V_{l,\delta_l(n)}^{\{n+1\}}(p_l^{\{n\}*},\ q_l^{\{n\}})-V_{l,\delta_l(n)}^{*}(p_l^{\{n\}*},\ q_l^{\{n\}}))\\
&\quad+(V_{l,\delta_l(n)}^{*}(p_l^{\{n\}},\ q_l^{\{n\}})-V_{l,\delta_l(n)}^{*}(p_l^{\{n\}*},\ q_l^{\{n\}}))\\
&\quad-\frac{\delta_l(n)}{2}(\|p_l^{\{n\}}-p_l^{\{n\}*}\|^2)\Big\},
\end{aligned}\tag{4.9}$$

where

$$\begin{aligned}
V_{l,\delta_l(n)}^{\{n+1\}}(p_l^{\{n\}},\ q_l^{\{n\}})=&\sum_{i=1}^{m_l}\sum_{j=1}^{n_l}p_l^{\{n\}}(i)V_l^{\{n+1\}}(a_l^i,\ b_l^j)q_l^{\{n\}}(j)\\
&-\frac{\delta_l(n)}{2}(\|p_l^{\{n\}}\|^2-\|q_l^{\{n\}}\|^2)
\end{aligned}$$

and

$$V_{l,\delta_l(n)}^{*}(p_l^{\{n\}},\ q_l^{\{n\}})=V_l^{*}(p_l^{\{n\}},\ q_l^{\{n\}})-\frac{\delta_l(n)}{2}(\|p_l^{\{n\}}\|^2-\|q_l^{\{n\}}\|^2).$$

Moreover, by (2.3) we can obtain the following inequalities:

$$\begin{aligned}
&\max_{l\in S}|V^{(n+1)}_{l,\delta_l(n)}(p_l^{(n)}, q_l^{(n)})-V^*_{l,\delta_l(n)}(p_l^{(n)}, q_l^{(n)})|\\
&\quad=\max_l|V_l^{(n+1)}(p_l^{(n)}, q_l^{(n)})-V_l^*(p_l^{(n)}, q_l^{(n)})|\\
&\quad\leqq\max_l\sum_{i=1}^{m_l}\sum_{j=1}^{n_l}p_l^{(n)}(i)\Big(\sum_{l'=1}^{s}|v_{l'}^{(n)}-v_{l'}^*|P(l'|l, a_l^i, b_l^j)q_l^{(n)}(j)\Big)\\
&\quad\leqq\max_l|v_l^{(n)}-v_l^*|\\
&\quad\leqq(1-\alpha)^NB,
\end{aligned}\tag{4.10}$$

where $n=Nu+m$, $1\leqq m\leqq u$, for u given in the assumption, and by similar argument,

$$\begin{aligned}
&\max_l|V^{(n+1)}_{l,\delta_l(n)}(p_l^{(n)*}, q_l^{(n)})-V^*_{l,\delta_l(n)}(p_l^{(n)*}, q_l^{(n)})|\\
&\quad\leqq\max_l|v_l^{(n)}-v_l^*|\\
&\quad\leqq(1-\alpha)^NB.
\end{aligned}\tag{4.11}$$

Also, since $p_l^{(n)*}$ and $q_l^{(n)*}$ are the optimal mixed strategies in $S^{m_l}_{\varepsilon_l(n)}\times S^{n_l}_{\varepsilon_l(n)}$ of the regularized matrix game with payoff matrix

$$\{V^*_{l,\delta_l(n)}(a_l^i, b_l^j):\ a_l^i\in A_l \text{ and } b_l^j\in B_l\},$$

it follows that

$$V^*_{l,\delta_l(n)}(p_l^{(n)}, q_l^{(n)})-V^*_{l,\delta_l(n)}(p_l^{(n)*}, q_l^{(n)})\leqq0.\tag{4.12}$$

Hence, inserting (4.9), (4.10) and (4.12) in (4.8), we arrive at the following inequality:

$$\begin{aligned}
&E[\|p_l^{(n+1)}-p_l^{(n+1)*}\|^2|p_l^{(n)}, q_l^{(n)}]\\
&\quad\leqq\Big(1-\frac{m_l}{m_l-1}\gamma_l(n)\delta_l(n)\Big)\|p_l^{(n)}-p_l^{(n)*}\|^2+4\sqrt{2}\|p_l^{(n)*}-p_l^{(n+1)*}\|\\
&\qquad+\frac{4m_l}{m_l-1}\gamma_l(n)(1-\alpha)^NB+2\gamma_l^2(n)\\
&\qquad\cdot E[\|A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})\|^2|p_l^{(n)}, q_l^{(n)}].
\end{aligned}\tag{4.13}$$

Next, we need to get an estimate of $E[\|A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})\|^2p_l^{(n)}, q_l^{(n)}]$, that is, from (2.3) and (4.1)

$$\begin{aligned}
&E[\|A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})\|^2|p_l^{(n)}, q_l^{(n)}]\\
&\quad=E[E[\|A^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)})\|^2|x_l^{(n+1)}, y_l^{(n+1)}]|p_l^{(n)}, q_l^{(n)}]\\
&\quad=E\Big[E\Big[\sum_{k=1}^{s}(A_k^{(n+1,l)}(x_l^{(n+1)}, y_l^{(n+1)}))^2|x_l^{(n+1)}, y_l^{(n+1)}\Big]|p_l^{(n)}, q_l^{(n)}\Big]\\
&\quad=\frac{m_l}{m_l-1}E[E[(V_l^{(n+1)}(1)/p_l^{(n)}(x_l^{(n+1)})-\delta_l(n))^2|x_l^{(n+1)}, y_l^{(n+1)}]|p_l^{(n)}, q_l^{(n)}]\\
&\quad=\frac{2m_l}{m_l-1}\Big\{\sum_{i=1}^{m_l}\sum_{j=1}^{n_l}\frac{R_l^{(1)}(a_l^i, b_l^j)+(V_l^{(n+1)}(a_l^i, b_l^j))^2}{p_l^{(n)}(i)}q_l^{(n)}(j)+\delta_l^2(n)\Big\}
\end{aligned}$$

$$\leqq \frac{2m_l}{m_l-1}\left\{\frac{R_l^{(1)}+(\max|v_l^{(n)}-v_l^*|+M)^2}{\varepsilon_l(n)}+\delta_l^2(n)\right\}$$

$$\leqq \frac{2m_l}{m_l-1}\left\{\frac{R_l^{(1)}+2B+2M^2}{\varepsilon_l(n)}+\delta_l^2(n)\right\} \tag{4.14}$$

where

$$R_l^{(1)}=\sum_{i=1}^{m_l}\sum_{j=1}^{n_l}R_l^{(1)}(a_l^i,\ b_l^j)$$

and M_l is a constant such that, for all $a_l^i\in A_l$ and $b_l^j\in B_l$,

$$|v_l^*(a_l^i,\ b_l^j)|\leqq M_l.$$

Hence, from (4.14), (4.13) can be written as the following:

$$\begin{aligned}
&E[\|p_l^{(n+1)}-p_l^{(n+1)*}\|^2\,|\,p_l^{(n)},\ q_l^{(n)}]\\
&\quad\leqq\left(1-\frac{m_l}{m_l-1}\gamma_l(n)\delta_l(n)\right)\|p_l^{(n)}-p_l^{(n)*}\|^2+4\sqrt{2}\,\|p_l^{(n)*}-p_l^{(n+1)*}\|\\
&\qquad+\frac{4m_lB}{m_l-1}\gamma_l(n)(1-\alpha)^N+\frac{2m_l(R_l^{(1)}+2B+2M^2)}{m_l-1}\times\frac{\gamma_l^2(n)}{\varepsilon_l(n)}\\
&\qquad+\frac{2m_l}{m_l-1}\gamma_l^2(n)\delta_l^2(n).
\end{aligned}\tag{4.15}$$

Similarly, we get

$$\begin{aligned}
&E[\|q_l^{(n+1)}-q_l^{(n+1)*}\|^2\,|\,p_l^{(n)},\ q_l^{(n)}]\\
&\quad\leqq\left(1-\frac{n_l}{n_l-1}\gamma_l(n)\delta_l(n)\right)\|q_l^{(n)}-q_l^{(n)*}\|^2+4\sqrt{2}\,\|q_l^{(n)*}-q_l^{(n+1)*}\|\\
&\qquad+\frac{4n_lB}{n_l-1}\gamma_l(n)(1-\alpha)^N+\frac{2n_l(R_l^{(2)}+2B+2M^2)}{n_l-1}\times\frac{\gamma_l^2(n)}{\varepsilon_l(n)}\\
&\qquad+\frac{4n_l}{n_l-1}\gamma_l^2(n)\delta_l^2(n),
\end{aligned}\tag{4.16}$$

where

$$R_l^{(2)}=\sum_{i=1}^{m_l}\sum_{j=1}^{n_l}R_l^{(2)}(a_l^i,\ b_l^j).$$

Now, putting that

$$c[n+1]=\sum_{l=1}^{s}(\|p_l^{(n+1)}-p_l^{(n+1)*}\|^2+\|q_l^{(n+1)}-q_l^{(n+1)*}\|^2)$$

and

$$d[n+1]=\sum_{l=1}^{s}\left(\frac{m_l-1}{m_l}\|p_l^{(n+1)}-p_l^{(n+1)*}\|^2+\frac{n_l-1}{n_l}\|q_l^{(n+1)}-q_l^{(n+1)*}\|^2\right),$$

there are the constants L_1 and L_2 such that, for each n,

$$L_1d[n+1]\leqq c[n+1]\leqq L_2d[n+1]. \tag{4.17}$$

From Lemma 2, (4.15) and (4.16), there exist a positive integer n_0 and positive constants

$K_l^{(i)}<\infty$, $i=1, 2, 3$, for each $l\in S$ such that, for all $n\geqq n_0$,

$$
\begin{aligned}
&E[d[n+1]\,|\,p_l^{(n)}, q_l^{(n)}, l\in S]\\
&\leqq\sum_{l=1}^{s}\Big\{\Big(1-\frac{m_l}{m_l-1}\gamma_l(n)\delta_l(n)\Big)\Big(\frac{m_l-1}{m_l}\|p_l^{(n)}-p_l^{(n)*}\|^2+\frac{n_l-1}{n_l}\|q_l^{(n)}-q_l^{(n)*}\|^2\Big)\\
&\quad+4\sqrt{2}\Big(\frac{m_l-1}{m_l}\|p_l^{(n)*}-p_l^{(n+1)*}\|+\frac{n_l-1}{n_l}\|q_l^{(n)*}-q_l^{(n+1)*}\|\Big)\\
&\quad+8B\gamma_l(n)(1-\alpha)^N+2(R_l^{(1)}+R_l^{(2)}+4M^2+4B^2)\times\frac{\gamma_l^2(n)}{\varepsilon_l(n)}+4\gamma_l^2(n)\delta_l^2(n)\Big\}\\
&\leqq(1-\gamma(n))d[n]+\sum_{l=1}^{s}\Big\{K_l^{(1)}\,|\,\varepsilon_l(n+1)-\varepsilon_l(n)\,|+K_l^{(2)}\,|\,\delta_l(n+1)-\delta_l(n)\,|\\
&\quad+K_l^{(3)}\Big|\frac{\varepsilon_l(n+1)}{\delta_l(n+1)}-\frac{\varepsilon_l(n)}{\delta_l(n)}\Big|+8B\gamma_l(n)(1-\alpha)^N\\
&\quad+2(R_l^{(1)}+R_l^{(2)}+4M^2+4B^2)\times\frac{\gamma_l(n)}{\varepsilon_l(n)}+4\gamma_l^2(n)\delta_l^2(n)\Big\},
\end{aligned}
\tag{4.18}
$$

where $\gamma(n)=\min_l(\gamma_l(n)\delta_l(n))$ and $n=Nu+r$, $1\leqq r\leqq u$, for u given in the assumption.

Now, (4.18) can be rewritten as follows:

$$
E[d[n+1]\,|\,p_l^{(n)}, q_l^{(n)}, l\in S]\leqq d[n]+\beta[n+1],
\tag{4.19}
$$

where

$$
\begin{aligned}
\beta[n+1]=\sum_{l=1}^{s}\Big\{&K_l^{(1)}\,|\,\varepsilon_l(n+1)-\varepsilon_l(n)\,|+K_l^{(2)}\,|\,\delta_l(n+1)-\delta_l(n)\,|\\
&+K_l^{(3)}\Big|\frac{\varepsilon_l(n+1)}{\delta_l(n+1)}-\frac{\varepsilon_l(n)}{\delta_l(n)}\Big|+8B\gamma_l(n)(1-\alpha)^N\\
&+2(R_l^{(1)}+R_l^{(2)}+4M^2+4B^2)\times\frac{\gamma_l^2(n)}{\varepsilon_l(n)}+4\gamma_l^2(n)\delta_l^2(n)\Big\}.
\end{aligned}
\tag{4.20}
$$

Introducing the notation $D[n]=d[n]+\sum_{k=n+1}^{\infty}\beta[k]$, (4.19) implies that

$$
E[D[n+1]\,|\,p_l^{(n)}, q_l^{(n)}, l\in S]\leqq D[n].
$$

Since $D[n]\geqq 0$, there is a ramdom variable $D\geqq 0$ such that $D[n]\to D$ w.p.1 as $n\to\infty$. Hence, it holds that

$$
d[n]\longrightarrow D\quad\text{w.p.1 as } n\longrightarrow\infty,
$$

because $\sum_{k=n+1}^{\infty}\beta[k]\to 0$ as $n\to\infty$ by the conditions of the sequences $\{\gamma_l(n)\}$, $\{\varepsilon_l(n)\}$ and $\{\delta_l(n)\}$ in the theorem and the remark. Taking the expectation of the both sides of (4.18) and summing the obtained inequalities with respect to n from n_0 to ∞, if follows that

$$
\sum_{n=n_0}^{\infty}\gamma[n]E[d[n]]<\infty.
\tag{4.21}
$$

Here, since $\gamma(n)=\min_l(\gamma_l(n)\delta_l(n))$, by using the conditions (a) and (b) in the theorem, we can prove that

$$\gamma(n)>0, \qquad n=0, 1, \cdots, \tag{4.22}$$
$$\gamma(n) \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty,$$

and

$$\sum_{n=0}^{\infty} \gamma(n)=\infty.$$

Then, from (4.21) and (4.22), there exists a subsequence $\{n_k\}$ of $\{n\}$ such that

$$\lim_{k\to\infty} E[d[n_k]]=0,$$

from which, by Fatou's lemma, we can conclude that $d[n_k]\to 0$ w.p.1 as $k\to\infty$. Therefore, $D=0$ w.p.1, hence also

$$c[n] \longrightarrow 0 \quad \text{w.p.1 as} \quad n \longrightarrow \infty.$$

Thus, the theorem is proved.

Next, when the teacher makes the players to stop learning at $(n+1)$-th stage, the estimate of probability, such that the difference of $p_l^{\{n+1\}}$ and $q_l^{\{n+1\}}$ from the optimal mixed strategies p_l^* and q_l^* of the Markov game is greater than $\varepsilon>0$, is given in the following theorem under the conditions in Theorem 4.1.

THEOREM 4.2. *For any positive number* $\varepsilon>0$ *and* all $n\geqq n_0$

$$\begin{aligned} &P\Big[\sum_{l=1}^{s}(\|p_l^{\{n+1\}}-p_l^*\|+\|q_l^{\{n+1\}}-q_l^*\|)\geqq\varepsilon\Big] \\ &\leqq 2\Big(\frac{S}{\varepsilon}\Big)^2\sum_{l=1}^{s}\Big\{\prod_{k=n_0}^{n}(1-\gamma_l(k)\delta_l(k))E[\|p_l^{\{n_0\}}-p_l^{\{n_0\}*}\|^2 \\ &\quad+\|q_l^{\{n_0\}}-q_l^{\{n_0\}*}\|^2]+\sum_{x=n_0}^{n+1}\prod_{k=x}^{n}(1-\gamma_l(k)\delta_l(k))P_l(x)+Q_l^2(n+1)\Big\}, \end{aligned} \tag{4.23}$$

where, by using N *such that* $x=Nu+r$, $1\leqq r\leqq u$,

$$\begin{aligned} P_l(x)=&4\sqrt{2}\Big(K_l^{(1)}|\varepsilon_l(x)-\varepsilon_l(x-1)|+K_l^{(2)}|\delta_l(x)-\delta_l(x-1)| \\ &+K_l^{(3)}\Big|\frac{\varepsilon_l(x)}{\delta_l(x)}-\frac{\varepsilon_l(x-1)}{\delta_l(x-1)}\Big|\Big)+\Big(\frac{4m_lB}{m_l-1}-\frac{4n_lB}{n_l-1}\Big)\gamma_l(x-1)(1-\alpha)^N \\ &+\Big(\frac{2m_l(R_l^{(1)}+2M^2+2B^2)}{m_l-1}+\frac{2n_l(R_l^{(2)}+2M^2+2B^2)}{n_l-1}\Big) \\ &\times\frac{\gamma_l^2(x-1)}{\varepsilon_l(x-1)}+\Big(\frac{2m_l}{m_l-1}+\frac{2n_l}{n_l-1}\Big)\gamma_l^2(x-1)\delta_l^2(x-1) \end{aligned} \tag{4.24}$$

and

$$Q_l(n+1)=K_l^{(1)}\varepsilon_l(n+1)+K_l^{(2)}\delta_l(n+1)+K_l^{(3)}\Big|\frac{\varepsilon_l(n+1)}{\delta_l(n+1)}-\mu_l\Big|.$$

PROOF. By using Chebyshev's inequality, we get

$$P\Big[\sum_{l=1}^{s}(\|p_l^{\{n+1\}}-p_l^*\|+\|q_l^{\{n+1\}}-q_l^*\|\geqq\varepsilon)\Big]$$

$$\leqq \sum_{l=1}^{s} P[\|p_l^{(n+1)}-p_l^*\|+\|q_l^{(n+1)}-q_l^*\|\geqq \varepsilon/s]$$

$$\leqq \left(\frac{s}{\varepsilon}\right)^2 \sum_{l=1}^{s} E[(\|p_l^{(n+1)}-p_l^*\|+\|q_l^{(n+1)}-q_l^*\|)^2]$$

$$\leqq 2\left(\frac{s}{\varepsilon}\right)^2 \sum_{l=1}^{s} \{E[(\|p_l^{(n+1)}-p_l^{(n+1)*}\|+\|q_l^{(n+1)}-q_l^{(n+1)*}\|)^2]$$

$$+E[(\|p_l^{(n+1)*}-p_l^*\|+\|q_l^{(n+1)*}-q_l^*\|)^2]\} . \tag{4.25}$$

Here, using a similar argument as the proof of Theorem 4.1, it follows that

$$E[(\|p_l^{(n+1)}-p_l^{(n+1)*}\|+\|q_l^{(n+1)}-q_l^{(n+1)*}\|)^2]$$

$$\leqq 2E[\|p_l^{(n+1)}-p_l^{(n+1)*}\|^2+\|q_l^{(n+1)}-q_l^{(n+1)*}\|^2]$$

$$\leqq 2\{(1-\gamma_l(n)\delta_l(n))E[\|p_l^{(n)}-p_l^{(n)*}\|^2+\|q_l^{(n)}-q_l^{(n)*}\|^2]+P_l(n+1)\}$$

$$\leqq 2\left\{\sum_{k=n_0}^{n}(1-\gamma_l(k)\delta_l(k))E[\|p_l^{(n_0)}-p_l^{(n_0)*}\|+\|q_l^{(n_0)}-q_l^{(n_0)*}\|]\right.$$

$$\left.+\sum_{x=n_0}^{n+1}\prod_{k=x}^{n}(1-\gamma_l(k)\delta_l(k))P_l(x)\right\}, \tag{4.26}$$

where

$$\prod_{k=x}^{n}(1-\gamma_l(k)\delta_l(k))=1 \qquad \text{if} \quad x=n+1 .$$

Further, by taking $m\to\infty$ in Lemma 2 with $\varepsilon_1=\varepsilon_l(n+1)$, $\delta_1=\delta_l(n+1)$ and $\varepsilon_2=\varepsilon_l(m)$, $\delta_2=\delta_l(m)$, it follows that

$$E[(\|p_l^{(n+1)*}-p_l^*\|+\|q_l^{(n+1)*}-q_l^*\|)^2]$$

$$\leqq Q_l^2(n+1), \tag{4.27}$$

where $Q_l(n+1)$ is given in the theorem.

Thus, from (4.25), (4.26) and (4.27), the results of the theorem is obtained and the proof is completed.

REMARK 2. *Under the conditions of Theorem* 4.1, *it is easy to show that the right-hand side of* (4.23) *goes to zero as* $n\to\infty$.

To mention the convergence in the mean square, the following lemma is important.

LEMMA 3. *If* $\{u_n\}$ *is a sequence of nonnegative numbers which satisfies the following conditions: for all* $n\geqq n_0$,

$$u(n)\leqq u(n-1)(1-\lambda(n))+\theta(n),$$

$$\lambda(n)\in(0,\ 1], \qquad \sum_{n=n_0}^{\infty}\lambda(n)=\infty \quad \textit{and} \quad \lim_{n\to\infty}\frac{\theta(n)}{\lambda(n)}=0,$$

then, the sequence $\{u(n)\}$ *converges to zero as* $n\to\infty$.

The proof of this lemma is shown in [1].

THEOREM 4.3. *If the conditions* (d)$\sim$(h) *in Theorem* 4.1 *are replaced by the following conditions: for* $\gamma(n)=\min_l(\gamma_l(n)\delta_l(n))$ *and each state* $l\in S$,

(d′) $$\frac{\gamma_l^2(n)\delta_l^2(n)}{\gamma(n)} \longrightarrow 0 \quad as \quad n \longrightarrow \infty,$$

(e′) $$\frac{\gamma_l^2(n)}{\varepsilon_l(n)\gamma(n)} \longrightarrow 0 \quad as \quad n \longrightarrow \infty,$$

(f′) $$\frac{\gamma_l(n)}{\gamma(n)}(1-\alpha)^N \longrightarrow 0 \quad (n=Nu+r,\ 1\leqq r\leqq u) \quad as \quad n \longrightarrow \infty,$$

(g′) $$\frac{1}{\gamma(n)}|\varepsilon_l(n+1)-\varepsilon_l(n)| \longrightarrow 0 \quad as \quad n \longrightarrow \infty,$$

(h′) $$\frac{1}{\gamma(n)}|\delta_l(n+1)-\delta_l(n)| \longrightarrow 0 \quad as \quad n \longrightarrow \infty,$$

(i′) $$\frac{1}{\gamma(n)}\left|\frac{\varepsilon_l(n+1)}{\delta_l(n+1)}-\frac{\varepsilon_l(n)}{\delta_l(n)}\right| \longrightarrow 0 \quad as \quad n \longrightarrow \infty,$$

then, at each state $l\in S$, *the sequence* $\{(p_l^{(n)}, q_l^{(n)})\}$ *constructed by the learning algorithm* (4.2) *converges in mean square to a pair* (p_l^*, q_l^*) *of the optimal stationary strategies of the Markov game.*

PROOF. From a similar argument in the proof of Theorem 4.1, it holds that, for $n=uN+r$, $1\leqq r\leqq u$,

$$\begin{aligned} E[d[n+1]] \leqq (1-\gamma(n))E[d[n]] + \sum_{l=1}^{s}\Big\{ & K_l^{(1)}|\varepsilon_l(n+1)-\varepsilon_l(n)| \\ & + K_l^{(2)}|\delta_l(n+1)-\delta_l(n)| + K_l^{(3)}\left|\frac{\varepsilon_l(n+1)}{\delta_l(n+1)}-\frac{\varepsilon_l(n)}{\delta_l(n)}\right| \\ & + 8B\gamma_l(n)(1-\alpha)^N + 2(R_l^{(1)}+R_l^{(2)}+4M^2+4B^2)\frac{\gamma_l(n)}{\varepsilon_l(n)} \\ & + 4\gamma_l^2(n)\delta_l^2(n)\Big\}, \end{aligned} \tag{4.28}$$

where $\gamma(n)=\min_l(\gamma_l(n)\delta_l(n))$.

Then, making use of Lemma 3 and the conditions of the theorem in (4.28), the theorem is proved.

Now, we consider in detail a case when the sequences in learning algorithm (4.2) are such that, for all states $l\in S$, $\gamma_l(n)\sim 1/n^\alpha$, $\varepsilon_l(n)\sim 1/n^\beta$, $\delta_l(n)\sim 1/n^\sigma$, $(\varepsilon_l(n)/\delta_l(n)-\mu_l)$ $\sim 1/n^\nu$ for $\beta=\sigma$ and $\sim 1/n^{\beta-\alpha}$ for $\beta>\sigma$, where the eqivalence of two sequences means that the ratio of their terms converges to a nonzero constant as $n\to\infty$. From the conditions of Thorem 4.1 and the theorem, it follows that for the convergence of the learning algorithm with probability one, it is sufficient to choose

$$0<\alpha<1, \qquad \beta\geqq\sigma>0, \qquad \nu>0$$
$$1/2<\alpha+\sigma\leqq 1, \qquad 2\alpha-\beta>0$$

and for the mean square convergence

$$0<\alpha<1, \qquad \beta\geqq\sigma>0, \qquad \nu>0$$
$$\alpha+\sigma\leqq 1, \qquad \alpha-\beta-\sigma>0.$$

References

[1] A.Z. NAZIN and A.S. POZNYAK: *Stochastic zero-sum game of two automata,* (in Russian). Avtom. Telemekh., No. 1 (1977), 53-61.

[2] K. TANAKA, S. IWASE and K. WAKUTA: *On Markov games with the expected average reward criterion.* Sci. Rep. Niigata Univ., Ser. A, No. 13 (1976), 31-41.

[3] K. TANAKA and K. WAKUTA: *On Markov games with the expected average reward criterion* (II). Sci. Rep. Niigata Univ., Ser. A, No. 13 (1976), 49-54.

[4] K. TANAKA and H. HOMMA: *On the learning algorithm of 2-person zero-sum game.* Sci. Rep. Niigata Univ., Ser. A, No. 16 (1979), 15-22.

[5] K. TANAKA and H. HOMMA: *On the learning algorithm of 2-person zero-sum Markov game.* Bulletin of Mathematical Statistics, Vol. 19, No. 1-2 (1980), 23-34.

Communicated by N. Furukawa
Received March 20, 1984