# ON THE LEARNING ALGORITHM OF 2-PERSON ZERO-SUM MARKOV GAME

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

Homma, Hisafumi
Department of Mathematics, Faculty of Science, Niigata University

https://doi.org/10.5109/13137

出版情報：統計数理研究. 19 (1/2), pp.23-34, 1980-03. 統計科学研究会
バージョン：
権利関係：

# ON THE LEARNING ALGORITHM OF 2-PERSON ZERO-SUM MARKOV GAME

By

**Kensuke TANAKA* and Hisafumi HOMMA***

(Received September 5, 1979)

## 1. Introduction

This paper is a continuation of our paper [2] with the title "On the learning algorithm of 2-person zero-sum game" and is concerned with a learning algorithm. In [2], we tried to find the learning algorithm utilized efficiently the given information at each stage to get a pair of the optimal mixed strategies under an assumption of incomplete information relating to payoff matrix.

However, in view of practical problem, it does not seem general enough to us that such an algorithm is applied to 2-person zero-sum game. For this reason, it is necessary to extend the algorithm to more general game. In this paper, it is shown that the learning algorithm is extended to 2-person zero-sum Markov game under an assumption of incomplete information relating to reward functions and transition probabilities of states. To construct the algorithm under this situation, at each stage two players play the games corresponding to all states of the system and use the information with relation those results given by a teacher. Moreover, we 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 mixed strategies. In the proof of convergence, the idea of regularization for supplying the lack of the strict convexity plays an important role. Such an attempt to learn the optimal strategies of a Markov game has been hardly found, as far as the present authers are aware, except this paper.

This paper consists of four sections. In section 2, we shall state the knowledge about 2-person zero-sum Markov game necessary in this paper. In section 3, we shall state the regularization of the dummy game. 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 to a pair of the optimal stationary strategies.

## 2. Preliminaries

In this section, we give a formulation of Markov game. We determine "2-person zero-sum Markov game" by a sixtuple $(S, A, B, p, r, \alpha)$. Here $S$ is a finite set labeled $\{1, 2, \cdots, s\}$, the set of states of a system; $A$ is a finite set labeled

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

$\{a^1, a^2, \cdots a^{m_1}\}$, the set of actions available to player I; $B$ is a finite set labeled $\{b^1, b^2, \cdots, b^{m_2}\}$, the set of actions available to player II; $p$ is a transition probability function which governs the law of motion of the system and is a function $p(\cdot \mid l, a, b)$ on $S$ for each triple $(l, a, b) \in S \times A \times B$; $r$, a reward for player I, is a function $r(l, a, b)$ on $S \times A \times B$ and $-r$ is a reward function for player II; $\alpha$, a discount factor, is a positive number.

In this game, player I and player II observe the state of the system at each stage and classify it onto one of the possible states $l \in S$ and then player I and player II choose actions by the mixed strategies, respectively. As a consequence of the present state $l \in S$ and the actions $a \in A$ and $b \in B$ chosen by the players, player II pays player I reward $r(l, a, b)$ unit of money and the system moves to a new state $l' \in S$, which is governed by the transition probability function $p(l' \mid l, a, b)$.

A strategy $\pi$ for player I is a sequence of $\pi_1, \pi_2, \cdots$, where on the $n$-th stage $\pi_n$ specifies the action to be chosen by player I using a probability distribution $\pi_n(\cdot \mid h_n)$ on $A$ depending to each histry $h_n = (l_0, a_0, b_0, l_1, a_1, b_1, \cdots, a_{n-1}, b_{n-1}, l_n)$ of the system. A strategy $\pi$ is, particularly, said to be stationary if there is a map $f$ from $S$ into $P(A)$ such that $\pi_n = f$ for all $n$, where $P(A)$ is the set of all probabilities on $A$. $\Pi$ denotes the class of all strategies for player I. Strategies and stationary strategies for player II are defined analogously. $\Gamma$ denotes the class of all strategies for player II.

Now, we define the expected discounted gain for player I. When the system starts from a state $l_0 \in S$ and a pair $(\pi, \sigma)$ of the strategies for the players is used, the total expected discounted gain for player I is defined to be

$$\phi(l_0, \pi, \sigma) = E^{\pi, \sigma}\left[\sum_{n=0}^{\infty} \alpha^n r(l_n, a_n, b_n)\right].$$

Then, player I wants to maximize $\phi(l_0, \pi, \sigma)$ and, at the same time, player II wants to minimize $\phi(l_0, \pi, \sigma)$.

A strategy $\pi^*$ is optimal for player I if, for all $\sigma' \in \Gamma$ and $l_0 \in S$,

$$\inf_{\sigma \in \Gamma} \sup_{\pi \in \Pi} \phi(l_0, \pi, \sigma) \leqq \phi(l_0, \pi^*, \sigma').$$

A strategy $\sigma^*$ is also optimal for player II if for all $\pi' \in \Pi$ and $l_0 \in S$

$$\sup_{\pi \in \Pi} \inf_{\sigma \in \Gamma} \phi(l^0, \pi, \sigma) \geqq \phi(l_0, \pi', \sigma^*).$$

We say that the Markev game is strictly determined if, for all initial state $l_0 \in S$,

$$\sup_{\pi \in \Pi} \inf_{\sigma \in \Gamma} \phi(l_0, \pi, \sigma) = \inf_{\sigma \in \Gamma} \sup_{\pi \in \Pi} \phi(l_0, \pi, \sigma).$$

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 $u = (u_1, u_2, \cdots, u_s) \in X^s$, we define $\|u\| = \max_i |u_i|$. $(X^s, d)$ is a complete metric space, where $d(v, u) = \|v - u\|$ for each $v$ and $u \in X^s$.

Now, for each $p=(p_1, p_2, \cdots, p_{m_1})\in P(A)$ and $q=(q_1, q_2, \cdots, q_{m_2})\in P(B)$, we define an operator $L(p, q)$: $X^s\to X^s$ as follows: for each $l\in S$ and $u\in X^s$,

$$L(p, q)u(l)=r(l, p, q)+\alpha\sum_{l'=1}^{s}u_{l'}p(l'|l, p, q),$$

where

$$r(l, p, q)=\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}r(l, a^i, b^j)p_iq_j$$

and

$$p(l'|l, p, q)=\sum_{j=1}^{m_2}\sum_{i=1}^{m_1}p(l'|l, a^i, b^j)p_iq_j.$$

Then, since $P(A)$ and $P(B)$ are compact and $L(p, q)u(l)$ is a continuous function on $P(A)\times P(B)$, there exist maps $p^*$ and $q^*$ from $S$ into $P(A)$ and $P(B)$, respectively, such that for each $l\in S$ and $u\in X^s$,

$$\begin{aligned}\min_q L(p^*, q)u(l)&=\max_p\min_q L(p, q)u(l)\\&=\min_q\max_p L(p, q)u(l)\\&=\max_p L(p, q^*)u(l).\end{aligned}\tag{2.1}$$

Moreover, we can define an operator $T$: $X^s\to X^s$ as follows: for each $l\in S$ and $u\in X^s$,

$$Tu(l)=\max_p\min_q L(p, q)u(l).$$

This operator is a contraction mapping on $X^s$ since $0<\alpha<1$. Since $X^s$ is a complete metric space, $T$ has a unique fixed point in $X^s$ by the Banach's fixed point theorem. Let $v^*$ be the unique fixed point of $T$. Then it holds that, for each $l\in S$.

$$\begin{aligned}v^*&=\max_p\min_q L(p, q)v^*(l)\\&=\max_p\min_q\left\{r(l, p, q)+\alpha\sum_{l'=1}^{s}v_{l'}^*p(l'|l, p, q)\right\},\end{aligned}\tag{2.2}$$

where $v^*=(v_1^*, v_2^*, \cdots, v_s^*)$.

From (2.1) and (2.2), we can obtain, for each $l\in S$,

$$\begin{aligned}v^*&=\min_q L(p^*, q)v^*(l)\\&=\max_p\min_q L(p, q)v^*(l)\\&=\min_q\max_p L(p, q)v^*(l)\\&=\max_p L(p, q^*)v^*(l)\\&=L(p^*, q^*)v^*(l).\end{aligned}\tag{2.3}$$

Hence, we have $v^* \in X^s$ as the value function of the game and $p^*$ and $q^*$ are the optimal stationary strategies for player I and player II, respectively. In order to solve the fixed point $v^* \in X^s$, we choose $v^{(0)} \in X^s$ arbitrarily and construct $v^{(n)}$ sequentially as follows: for each $l \in S$,

$$v_l^{(n+1)} = \max_p \min_q \left\{ r(l, p, q) + \alpha \sum_{l'=1}^{s} v_{l'}^{(n)} p(l' | l, p, q) \right\},$$

where $v^{(n)} = (v_1^{(n)}, v_2^{(n)}, \cdots, v_s^{(n)})$.
Then, we get the following inequality: for all $n = 1, 2, \cdots$,

$$\| v^{(n)} - v^* \| \leqq \alpha^n \| v^{(0)} - v^* \|, \tag{2.4}$$

where $\| \cdot \|$ is maximum norm in $X^s$.

As mentioned above, at each state $l \in S$ we need to consider a 2-person zero-sum game, which may be called a dummy game, with the following payoff function:

$$r(l, a, b) + \alpha \sum_{l'=1}^{s} v_{l'}^* p(l' | l, a, b), \qquad \text{for all } a \in A \text{ and } b \in B.$$

The optimal strategies $p^*$ and $q^*$ for player I and player II in the dummy game correspond to them for the players in the original Markov game, respectively. The value function of the dummy game corresponds also to one of the original Markov game.

### 3. Regularization of the dummy game

From the fact mentioned in section 2, a teacher makes two players search the optimal stationary strategies $p^*$ and $q^*$ of the original Markov game by a learning of the dummy game. Then, the gradient method used as the learning algorithm for solving such an optimization problem encounters several difficulties that are connected mainly with the lack of strict convexity of the payoff function. One of the possible method of avoiding these difficulties is to introduce an idea of regularization in the dummy game.

Suppose that in regularized game at each state $l \in S$, when the strategies $a$ and $b$ for two players are chosen by the mixed strategies $p$ and $q$, respectively, the payoff functions of the players are

$$v_l^*(a, b) - \frac{\delta}{2}(p_a - q_b)$$

and

$$-v_l^*(a, b) - \frac{\delta}{2}(q_b - p_a),$$

respectively ($a = 1, 2, \cdots, m_1$; $b = 1, 2, \cdots, m_2$), where $\delta > 0$ is the regularization parameter and

$$v_l^*(a, b) = r(l, a, b) + \alpha \sum_{l'=1}^{s} v_{l'}^* p(l' | l, a, b).$$

Then, the expected gain for player I at each state $l$ is given by

$$v^*_{l,\delta}(p, q)=v^*_l(p, q)-\frac{\delta}{2}(\|p\|^2-\|q\|^2) \tag{3.1}$$

and the expected gain for player II is $-v^*_{l,\delta}(p, q)$, where $\|\cdot\|$ is Euclidean norm.

Moreover, we assume that the mixed strategies available to players are in $\varepsilon$-simplices, i.e., $p\in S^{m_1}_\varepsilon$ and $q\in S^{m_2}_\varepsilon$, where

$$S^m_\varepsilon=\left\{x=(x_1, x_2, \cdots, x_s);\ x_i\geqq\varepsilon,\ i=1, 2, \cdots, m,\ \sum_{i=1}^{m} x_i=1,\ \left(0\leqq\varepsilon\leqq\frac{1}{m}\right)\right\}.$$

Thus, this game is 2-person zero-sum game restricted by the fixed $\varepsilon$ and $\delta$. Since $v^*_{l,\delta}(p, q)$ is strictly convex for any fixed $\delta>0$, the game has a unique saddle point $(p^*(l, \varepsilon, \delta), q^*(l, \varepsilon, \delta))$, for any fixed $\varepsilon\in[0, \hat{\varepsilon}]$, $\hat{\varepsilon}=\min\left(\frac{1}{m_1}, \frac{1}{m_2}\right)$, such that

$$v^*_{l,\delta}(p^*(l, \varepsilon, \delta), q)\geqq v^*_{l,\delta}(p^*(l, \varepsilon, \delta), q^*(l, \varepsilon, \delta))\geqq v^*_{l,\delta}(p, q^*(l, \varepsilon, \delta)) \tag{3.2}$$

for all $p\in S^{m_1}_\varepsilon$ and $q\in S^{m_2}_\varepsilon$. Then, the strategies $p^*(l, \varepsilon, \delta)$ and $q^*(l, \varepsilon, \delta)$ are the optimal strategies for the players in the game.

The following two lemmas show the connection of the regularized game with the dummy game and the properties of the saddle point $(p^*(l, \varepsilon, \delta), q^*(l, \varepsilon, \delta))$ as a function of $(\varepsilon, \delta)$ that plays an important role in our learning algorithm.

LEMMA 1. *If the sequences* $\{\varepsilon[n]\}$ *and* $\{\delta[n]\}$ *satisfy*

$$\varepsilon[n]\in(0, \hat{\varepsilon}),\qquad \delta[n]>0,\qquad \lim_{n\to\infty}\varepsilon[n]=\lim_{n\to\infty}\delta[n]=0,$$

*and*

$$\lim_{n\to\infty}\frac{\varepsilon[n]}{\delta[n]}=\mu\in[0, \infty),$$

*then, at each state* $l$, *the sequence* $\{(p^*(l, \varepsilon[n], \delta[n]), q^*(l, \varepsilon[n], \delta[n]))\}$ *converges to a saddle point* $(p^*(l), q^*(l))$ *of the dummy game (depending, generally, on* $\mu$*).*

LEMMA 2. *There exist* $\delta'\in(0, \infty)$ *and constants* $K_1$, $K_2$ *and* $K_3$ *such that for all states* $l\in S$,

$$\|p^*(l, \varepsilon_1, \delta_1)-p^*(l, \varepsilon_2, \delta_2)\|+\|q^*(l, \varepsilon_1, \delta_1)-q^*(l, \varepsilon_2, \delta_2)\|$$

$$\leqq K_1|\varepsilon_1-\varepsilon_2|+K_2|\delta_1-\delta_2|+K_3\left|\frac{\varepsilon_1}{\delta_1}-\frac{\varepsilon_2}{\delta_2}\right| \tag{3.3}$$

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

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

## 4. Construction and convergence of the learning algorithm

In this section, a pseudogradient method is used as the algorithm of learning the optimal mixed strategies $p^*(l, \varepsilon, \delta)$ and $q^*(l, \varepsilon, \delta)$ at each state $l\in S$ in the regularized game for the fixed $\varepsilon>0$ and $\delta>0$. A teacher makes the players learn

the optimal strategies of the original Markov game by the following learning algorithm:

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

Now, let $v^{(n)}=(v_1^{(n)}, v_2^{(n)}, \cdots, v_s^{(n)})$ be the value of game generated by a teacher at the $n$-th stage. Further, let $p^{(n)}(l)$ and $q^{(n)}(l)$ be the mixed strategies of each state $l \in S$ obtained by the players at the $n$-th stage, respectively. Then, at the $(n+1)$-th stage, the players play the game at each state $l \in S$ using the mixed strategies $p^{(n)}(l)$ and $q^{(n)}(l)$ and suppose that as a result of this play, the players choose pure strategies $x_{n+1}(l)$ and $y_{n+1}(l)$, respectively. So a teacher makes the players know a value of payoff function:

$$v_l^{(n+1)}(x_{n+1}(l), y_{n+1}(l))=r(l, x_{n+1}(l), y_{n+1}(l)) + \alpha \sum_{l'=1}^{s} v_{l'}^{(n)} p(l' \mid l, x_{n+1}(l), y_{n+1}(l)) .$$

Using this information given by a teacher, player I and player II construct the mixed strategies of each $l$ used at next stage as follows:

$$p^{(n+1)}(l)=\Pi_{S_{\varepsilon[n+1]}^{m_1}}\{p^{(n)}(l)+\gamma[n+1]A^{(l)}(x_{n+1}(l), y_{n+1}(l))\} \tag{4.1a}$$

and

$$q^{(n+1)}(l)=\Pi_{S_{\varepsilon[n+1]}^{m_2}}\{q^{(n)}(l)-\gamma[n+1]B^{(l)}(x_{n+1}(l), y_{n+1}(l))\} . \tag{4.1b}$$

where $\{\varepsilon[n]\}$, $\{\delta[n]\}$ and $\{\gamma[n]\}$ are the sequences of numbers: for $a^i$ and $b^j$, $i=1, 2, \cdots, m_1$, $j=1, 2, \cdots, m_2$, $A^{(l)}(x_{n+1}(l), y_{n+1}(l))$ is an $m_1$-dimensional vector whose elements are

$$A_k^{(l)}(a^i, b^j)=\begin{cases} \dfrac{v_l^{(n+1)}(a^i, b^j)}{p_i^{(n)}(l)}-\delta[n+1], & k=a^i \\ -\dfrac{1}{m_1-1}\left(\dfrac{v_l^{(n+1)}(a^i, b^j)}{p_i^{(n)}(l)}-\delta[n+1]\right), & k \neq a^i \end{cases}$$

and $B^{(l)}(x_{n+1}(l), y_{n+1}(l))$ is an $m_2$-dimensional vector whose elements are

$$B_k^{(l)}(a^i, b^j)=\begin{cases} \dfrac{v_l^{(n+1)}(a^i, b^j)}{q_j^{(n)}(l)}+\delta[n+1], & k=b^j \\ -\dfrac{1}{m_2-1}\left(\dfrac{v_l^{(n+1)}(a^i, b^j)}{q_j^{(n)}(l)}+\delta[n+1]\right), & k \neq b^j ; \end{cases}$$

$\Pi_S\{\cdot\}$ is a projection operator on the closed bounded set $S$ that has the property represented by Euclidean norm

$$\Pi_S\{x\} \in S \quad \text{and} \quad \|x-y\| \geqq \|\Pi_S\{x\}-y\|$$

for all $x$ and $y \in S$.

Next, a teacher constructs the value of game used at the next stage as follows: at each state $l \in S$,

$$v_l^{(n)} = \max_p \min_q \left\{ r(l, p, q) + \alpha \sum_{l'=1}^{s} v_{l'}^{(n)} p(l' \mid l, p, q) \right\}.$$

Moreover, the learning of the players is, sequentially, continued by the guidance of a 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 $p^{(n)*}(l) = p^{(n)*}(l, \varepsilon[n], \delta[n])$ and $q^{(n)*}(l) = q^{(n)*}(l, \varepsilon[n], \delta[n])$.

THEOREM 1. *Suppose that the sequences* $\{\varepsilon[n]\}$, $\{\delta[n]\}$ *and* $\{\gamma[n]\}$ *satisfy the following conditions*:

(a) $\gamma[n] > 0$, $\delta[n] > 0$, $\varepsilon[n] \in (0, \hat{\varepsilon})$, $n = 1, 2, \cdots$, $\delta[n] \to 0$ *as* $n \to \infty$,

(b) $\lim_{n\to\infty} \dfrac{\varepsilon[n]}{\delta[n]} = \mu < \infty$,

(c) $\sum_{n=1}^{\infty} \gamma[n]\delta[n] = \infty$,

(d) $\sum_{n=1}^{\infty} \gamma^2[n]\delta^2[n] < \infty$,

(e) $\sum_{n=1}^{\infty} \dfrac{\gamma^2[n]}{\varepsilon[n-1]} < \infty$,

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

(g) $\sum_{n=1}^{\infty} |\delta[n] - \delta[n-1]| < \infty$,

(h) $\sum_{n=1}^{\infty} \left| \dfrac{\varepsilon[n]}{\delta[n]} - \dfrac{\varepsilon[n-1]}{\delta[n-1]} \right| < \infty$.

*Then, at each state* $l$, *the sequence* $\{(p^{(n)}(l), q^{(n)}(l))\}$ *generated by the learning algorithm* (4.1) *converges with probability one as* $n \to \infty$ *to a pair of the original Markov game for a pair of any initial mixed strategies* $(p^{(0)}(l), q^{(0)}(l)) \in S^{m_1}_{\varepsilon[0]} \times S^{m_2}_{\varepsilon[0]}$.

PROOF. By (4.1a) and (4.2),

$$\begin{aligned} \|p^{(n+1)}(l) - p^{(n+1)*}(l)\|^2 \leqq{} & \|p^{(n)}(l) - p^{(n)*}(l)\|^2 + 4\sqrt{2}\,\|p^{(n)*}(l) - p^{(n+1)*}(l)\| \\ & + 2\gamma[n+1]\langle p^{(n)}(l) - p^{(n)*}(l),\ A^{(l)}(x_{n+1}(l),\ y_{n+1}(l))\rangle \\ & + 2\gamma^2[n+1]\,\|A^{(l)}(x_{n+1}(l),\ y_{n+1}(l))\|^2 \end{aligned} \tag{4.3}$$

where $\langle \cdot, \cdot \rangle$ denotes inner product and $\|\cdot\|$ denotes Euclidean norm. Taking conditional expectation of (4.3) for given $p^{(n)}(l)$ and $q^{(n)}(l)$, we get

$$\begin{aligned} & E[\|p^{(n+1)}(l) - p^{(n+1)*}(l)\|^2 \mid p^{(n)}(l),\ q^{(n)}(l)] \\ & \quad \leqq \|p^{(n)}(l) - p^{(n)*}(l)\|^2 + 4\sqrt{2}\,\|p^{(n)*}(l) - p^{(n+1)*}(l)\| \\ & \qquad + 2\gamma[n+1]E[\langle p^{(n)}(l) - p^{(n)*}(l),\ A^{(l)}(x_{n+1}(l),\ y_{n+1}(l))\rangle \mid p^{(n)}(l),\ q^{(n)}(l)] \\ & \qquad + 2\gamma^2[n+1]E[\|A^{(l)}(x_{n+1}(l),\ y_{n+1}(l))\|^2 \mid p^{(n)}(l),\ q^{(n)}(l)]. \end{aligned} \tag{4.4}$$

Here, for each $n$ and $l\in S$, it holds that by (3.1)

$$
\begin{aligned}
&E[\langle p^{(n)}(l)-p^{(n)*}(l),\ A^{(l)}(x_{n+1}(l),\ y_{n+1}(l))\rangle \mid p^{(n)}(l),\ q^{(n)}(l)] \\
&\quad=\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\Big\{\sum_{k=1}^{m_1}(p_k^{(n)}(l)-p_k^{(n)*}(l))A_k^{(l)}(a^i,\ b^j)\Big\}p_i^{(n)}(l)q_j^{(n)}(l) \\
&\quad=\frac{m_1}{m_1-1}\Big\{v^{(n+1)}_{\delta[n+1],\,l}(p^{(n)}(l),\ q^{(n)}(l))-v^{(n+1)}_{\delta[n+1],\,l}(p^{(n)*}(l),\ q^{(n)}(l)) \\
&\qquad-\frac{1}{2}\delta[n+1]\sum_{i=1}^{m_1}(p_i^{(n)}(l)-p_i^{(n)*}(l))^2\Big\} \\
&\quad=\frac{m_1}{m_1-1}\Big\{v^{*}_{\delta[n+1],\,l}(p^{(n)}(l),\ q^{(n)}(l))-v^{*}_{\delta[n+1],\,l}(p^{(n)*}(l),\ q^{(n)}(l)) \\
&\qquad+v^{(n+1)}_{\delta[n+1],\,l}(p^{(n)}(l),\ q^{(n)}(l))-v^{*}_{\delta[n+1],\,l}(p^{(n)}(l),\ q^{(n)}(l)) \\
&\qquad+v^{*}_{\delta[n+1],\,l}(p^{(n)*}(l),\ q^{(n)}(l))-v^{(n+1)}_{\delta[n+1],\,l}(p^{(n)*}(l),\ q^{(n)}(l)) \\
&\qquad-\frac{1}{2}\delta[n+1]\sum_{i=1}^{m_1}(p_i^{(n)}(l)-p_i^{(n)*}(l))^2\Big\},
\end{aligned}
\tag{4.5}
$$

where

$$
\begin{aligned}
v^{*}_{\delta[n+1],\,l}(p^{(n)}(l),\ q^{(n)}(l))&=r(l,\ p^{(n)}(l),\ q^{(n)}(l)) \\
&\quad+\alpha\sum_{l'=1}^{s}v^{*}_{l'}p(l'\mid l,\ p^{(n)}(l),\ q^{(n)}(l)) \\
&\quad-\frac{1}{2}\delta[n+1](\|p^{(n)}(l)\|^2-\|q^{(n)}(l)\|^2)
\end{aligned}
$$

and

$$
\begin{aligned}
v^{(n+1)}_{\delta[n+1],\,l}(p^{(n)}(l),\ q^{(n)}(l))&=r(l,\ p^{(n)}(l),\ q^{(n)}(l)) \\
&\quad+\alpha\sum_{l'=1}^{s}v^{(n)}_{l'}p(l'\mid l,\ p^{(n)}(l),\ q^{(n)}(l)) \\
&\quad-\frac{1}{2}\delta[n+1](\|p^{(n)}(l)\|^2-\|q^{(n)}(l)\|^2).
\end{aligned}
$$

Also, for each $n$ and $l\in S$, it holds that by (3.1)

$$
\begin{aligned}
&E[\|A^{(l)}(x_{n+1}(l),\ y_{n+1}(l))\|^2 \mid p^{(n)}(l),\ q^{(n)}(l)] \\
&\quad=\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sum_{k=1}^{m_1}(A_k^{(l)}(a^i,\ b^j))^2p_i^{(n)}(l)q_j^{(n)}(l) \\
&\quad=\frac{m_1}{m_1-1}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\Big(\frac{v_l^{(n+1)}(a^i,\ b^j)}{p_i^{(n)}(l)}-\delta[n+1]\Big)^2p_i^{(n)}(l)q_j^{(n)}(l) \\
&\quad\leqq\frac{2m_1}{m_1-1}\Big\{\frac{1}{\varepsilon[n]}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(v_l^{(n+1)}(a^i,\ b^j))^2q_j^{(n)}(l)+\delta^2[n+1]\Big\}.
\end{aligned}
\tag{4.6}
$$

Moreover, we can obtain the following inequalities: for each $n$

$$\max_l |v^{(n+1)}_{\delta[n+1],l}(p^{(n)}(l), q^{(n)}(l))-v^*_{\delta[n+1],l}(p^{(n)}(l), q^{(n)}(l))| \leqq \alpha^{n+1}M \tag{4.7}$$

and, for each $n$, state $l\in S$, $a^i\in A$ and $b^j\in B$,

$$\begin{aligned}|v_l^{(n+1)}(a^i, b^j)| &\leqq |r(l, a^i, b^j)|+\alpha\sum_{l'=1}^{s}|v_{l'}^{(n)}|\,p(l'|l, a^i, b^j)\\ &\leqq N+\alpha N+\alpha^2N+\cdots+\alpha^nN\\ &\leqq \frac{N}{1-\alpha},\end{aligned} \tag{4.8}$$

where

$$M=\max_l |v_l^*-v_l^{(0)}|$$

and

$$N=\max_{l,a,b}(|r(l, a, b)|, v_l^{(0)}).$$

Hence, from (4.4), (4.5), (4.6), (4.7) and (4.8), we have

$$\begin{aligned}&E[\|p^{(n+1)}(l)-p^{(n+1)*}(l)\|^2|\,p^{(n)}(l), q^{(n)}(l)]\\ &\leqq\left(1-\frac{m_1}{m_1-1}\gamma[n+1]\delta[n+1]\right)\|p^{(n)}(l)-p^{(n)*}(l)\|^2\\ &\quad+4\sqrt{2}\,\|p^{(n)*}(l)-p^{(n+1)*}(l)\|+\frac{4m_1}{m_1-1}M\gamma[n+1]\alpha^{n+1}\\ &\quad+\frac{4m_1^2}{m_1-1}\frac{\gamma^2[n+1]}{\varepsilon[n]}\left(\frac{N}{1-\alpha}\right)^2+\frac{4m_1}{m_1-1}\gamma^2[n+1]\delta^2[n+1]\\ &\quad+\frac{2m_1}{m_1-1}\gamma[n+1]\{v^*_{\delta[n+1],l}(p^{(n)}(l), q^{(n)}(l))-v^*_{\delta[n+1],l}(p^{(n)*}(l), q^{(n)}(l))\}.\end{aligned} \tag{4.9}$$

Similarly, we have

$$\begin{aligned}&E[\|q^{(n+1)}(l)-q^{(n+1)*}(l)\|^2|\,p^{(n)}(l), q^{(n)}(l)]\\ &\leqq\left(1-\frac{m_2}{m_2-1}\gamma[n+1]\delta[n+1]\right)\|q^{(n)}(l)-q^{(n)*}(l)\|^2\\ &\quad+4\sqrt{2}\,\|q^{(n)*}(l)-q^{(n+1)*}(l)\|+\frac{4m_2}{m_2-1}M\gamma[n+1]\alpha^{n+1}\\ &\quad+\frac{4m_2^2}{m_2-1}\frac{\gamma^2[n+1]}{\varepsilon[n]}\left(\frac{N}{1-\alpha}\right)^2+\frac{4m_2}{m_2-1}\gamma^2[n+1]\delta^2[n+1]\\ &\quad+\frac{2m_2}{m_2-1}\gamma[n+1]\{v^*_{\delta[n+1],l}(p^{(n)}(l), q^{(n)*}(l))-v^*_{\delta[n+1],l}(p^{(n)}(l), q^{(n)}(l))\}.\end{aligned} \tag{4.10}$$

Now, putting that

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

and

$$d[n+1]=\sum_{l=1}^{s}\left(\frac{m_1}{m_1-1}\|p^{(n+1)}(l)-p^{(n+1)*}(l)\|^2+\frac{m_2}{m_2-1}\|q^{(n+1)}(l)-q^{(n+1)*}(l)\|^2\right),$$

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

$$L_1 d[n+1]\leqq c[n+1]\leqq L_2 d[n+1].$$

From Lemma 2, (4.9), (4.10) and the definition of the optimal strategies $p^{(n)*}(l)$ and $q^{(n)*}(l)$ in each state $l\in S$, there exist a positive integer $n_0$ and positive constants $K_i<\infty$ $(i=1, 2, \cdots, 6)$ such that, for all $n\geqq n_0$,

$$\begin{aligned}E[d[n+1]\,|\,p^{(n)}(l),\ q^{(n)}(l),\ l=1, 2, \cdots, s]&\\ \leqq(1-L_1\gamma[n+1]\delta[n+1])d[n]+K_1|\varepsilon[n+1]-\varepsilon[n]|+K_2|\delta[n+1]-\delta[n]|&\\ +K_3\left|\frac{\varepsilon[n+1]}{\delta[n+1]}-\frac{\varepsilon[n]}{\delta[n]}\right|+K_4\alpha^{n+1}+K_5\frac{\gamma^2[n+1]}{\varepsilon[n]}+K_6\gamma^2[n+1]\delta^2[n+1].&\end{aligned}\tag{4.11}$$

Also, (4.11) can be rewritten as follows:

$$E[d[n+1]\,|\,p^{(n)}(l),\ q^{(n)}(l),\ l=1, 2, \cdots, s]\leqq d[n]+\beta[n+1],\tag{4.12}$$

where

$$\begin{aligned}\beta[n+1]=K_1|\varepsilon[n+1]-\varepsilon[n]|+K_2|\delta[n+1]-\delta[n]|+K_3\left|\frac{\varepsilon[n+1]}{\delta[n+1]}-\frac{\varepsilon[n]}{\delta[n]}\right|&\\ +K_4\alpha^{n+1}+K_5\frac{\gamma^2[n+1]}{\varepsilon[n]}+K_6\gamma^2[n+1]\delta^2[n+1].&\end{aligned}$$

Introducing the notation $D[n]=d[n]+\sum_{k=n+1}^{\infty}\beta[k]$, (4.12) implies that

$$E[D[n+1]\,|\,p^{(n)}(l),\ q^{(n)}(l),\ l=1, 2, \cdots, s]\leqq D[n].$$

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

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

because $\beta[n]\to 0$ as $n\to\infty$ by the conditions (d)$\sim$(h). Taking the expectations of the both sides of (4.11) and summing the obtained inequalities with respect to $n$ from $n_0$ to $\infty$, it follows that

$$\sum_{n=n_0}^{\infty}\gamma[n+1]\delta[n+1]E[d[n]]>\infty.\tag{4.13}$$

Then, from (4.13) and the condition (c), there exist a subsequence $\{n_k\}$ 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

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

Thus, the theorem is proved.

THEOREM 2. *If the conditions* (d)∼(h) *in Theorem* 1 *are replaced by the conditions*

(d′) $\gamma[n+1]\delta[n+1]\to 0$ *as* $n\to\infty$,

(e′) $\dfrac{\gamma[n+1]}{\varepsilon[n]\delta[n+1]}\to 0$ *as* $n\to\infty$,

(f′) $\dfrac{1}{\gamma[n+1]\delta[n+1]}|\varepsilon[n+1]-\varepsilon[n]|\to 0$ *as* $n\to\infty$,

(g′) $\dfrac{1}{\gamma[n+1]\delta[n+1]}|\delta[n+1]-\delta[n]|\to 0$ *as* $n\to\infty$,

(h′) $\dfrac{1}{\gamma[n+1]\delta[n+1]}\left|\dfrac{\varepsilon[n+1]}{\delta[n+1]}-\dfrac{\varepsilon[n]}{\delta[n]}\right|\to 0$ *as* $n\to\infty$,

*then, at each state* $l\in S$, *the sequence* $\{(p^{(n)}(l), q^{(n)}(l))\}$ *converges in mean square to a pair of the optimal stationary strategies* $(p^*(l), q^*(l))$.

The proof of this theorem can be shown by the similar argument to Theorem 2 in [1] by noting that

$$\lim_{n\to\infty}\frac{\alpha^{n+1}}{\gamma[n+1]\delta[n+1]}=0 .$$

We can obtain an upper bound for the rate of convergence of the learning algorithm (4.1) by the following theorem.

THEOREM 3. *Suppose that the conditions of Theorem* 2 *are satisfied and that there exist* $r\in(0, 1)$ *and* $t>r$ *such that*

$$\liminf_{n\to\infty} n^r\gamma[n]\delta[n]>0 ,$$

*and*

$$\limsup_{n\to\infty} n^t\chi[n]\in(0, \infty) ,$$

*where*

$$\chi[n]=\frac{\gamma^2[n]}{\varepsilon[n-1]}+\gamma^2[n]\delta^2[n]+|\varepsilon[n]-\varepsilon[n-1]|$$
$$+|\delta[n]-\delta[n-1]|+\left|\frac{\varepsilon[n]}{\delta[n]}-\frac{\varepsilon[n-1]}{\delta[n-1]}\right|+\alpha^n .$$

*Then, there exist the constants* $C$ *and* $C'\in(0, \infty)$ *such that*

$$E\left[\sum_{i=1}^{s}(\|p^{(n)}(l)-p^{(n)*}(l)\|^2+\|q^{(n)}(l)-q^{(n)*}(l)\|^2)\right]$$
$$\leqq\frac{C}{n^{t-r}}+C'\left\{\varepsilon^2[n]+\delta^2[n]+\left(\frac{\varepsilon[n]}{\delta[n]}-\mu\right)^2\right\} .$$

Since $\lim_{n\to\infty} n^t\alpha^n=0$ for $t>0$, the theorem holds by Lemma 3 in [1].

We consider in detail a case when the sequence in learning algorithm (4.1) are such that $\gamma[n]\sim 1/n^{\beta}$, $\varepsilon[n]\sim 1/n^{\gamma}$, $\delta[n]\sim 1/n^{\sigma}$, $(\varepsilon[n]/\delta[n]-\mu)\sim 1/n^{\nu}$ for $\gamma=\sigma$ and $\sim 1/n^{\gamma-\sigma}$ for $\gamma>\sigma$, where the equivalence of two sequences means that the ratio of their terms converges as $n\to\infty$ to a nonzero constant. From the conditions of Theorem 1 and 2, it follows that for the convergence of the algorithm with probability one it is sufficent to choose $\alpha$, $\beta$, $\gamma$, $\sigma$ and $\nu$ such that

$$0<\alpha<1,\quad \gamma\geqq\sigma>0,\quad \nu>0,\quad \frac{1}{2}<\beta+\sigma\leqq 1,\quad 2\beta-\gamma>1,$$

and for the mean square convergence,

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

## References

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

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

[3] TANAKA, K. and WAKUTA, K., *On continuous time Markov games with countable state space*, J. Oper. Res. Soc. Japan, No. 1 (1978), 17-28.