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

By

Kensuke TANAKA* and Hisafumi HOMMA*

(Received October 31, 1978)

1. Introduction

In view of practical problems, the methods of learning a pair of optimal strategies, which is called a solution of game, have been investigated by many authors (see, for example, [1], [2], [3] and [4]). Especially, a type of the learning algorithm in [4] is pseudogradient one which uses the idea of regularization for supplying the lack of the strict convexity of the payoff function.

But their algorithm does not seem general enough to us in point of utilizing the given information. For this reason, we propose a learning algorithm which is an extension of their algorithm. Then, we show that a pair of the mixed strategies generated by our algorithm converges with probability one (w. p. 1) and in mean square to a pair of the optimal mixed strategies under some conditions. Moreover, we give an upper bound for the rate of convergence of our algorithm.

2. Formulation of 2-person zero-sum game

In this section, we consider a zero-sum game with two players (player I and player II) consisting of an infinite number of repeated single games under an assumption of incomplete information relating to payoff matrix. Player I and player II have, respectively, m and l available pure strategies. If player I has chosen a pure strategy with the number i and player II has chosen one with the number j, then their payoffs consist of ξ and η, respectively, which are random variables such that

$$\mathrm{E}[\xi \mid i, j] = -\mathrm{E}[\eta \mid i, j] = r_{ij}, \quad |r_{ij}| < \infty, \tag{1}$$

$$\mathrm{E}[\xi^2 \mid i, j] = R_{ij}^{(1)} < \infty, \quad \mathrm{E}[\eta^2 \mid i, j] = R_{ij}^{(2)} < \infty. \tag{2}$$

In general, the payoff matrix $A=(r_{ij})$ is assumed to be known in advance. But, in this game, the payoff matrix is not known and in each single game the players receive information relating to the payoff matrix A in the form of realization of ξ and η.

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

Now, suppose that the players choose their pure strategies by probabilities $p=(p_1, p_2, \cdots, p_m)^T$ and $q=(q_1, q_2, \cdots, q_l)^T$, respectively, where the notation T denotes transpose of vector. Then the expected payoff for player I is

$$V(p, q)=p^T Aq=\sum_{i=1}^{m}\sum_{j=1}^{l} p_i r_{ij} q_j \tag{3}$$

and one for player II is $-V(p, q)$. From well-known min-max theorem, there is a pair (p^*, q^*) of optimal mixed strategies such that

$$V(p^*, q)\geqq V(p^*, q^*)\geqq V(p, q^*) \tag{4}$$

for all mixed strategies p and q. Then the value $V(p^*, q^*)$ is called a value of game. The problem of each player is to find an optimal mixed strategy while he plays a single game repeatedly.

3. Regularization of game

One of the most natural methods for solving the optimization problem is the gradient method. However, a direct application of it in the game encounters several difficulties that are connected mainly with the lack of strict convexity of the payoff function. One of the possible methods of avoiding these difficulties is to introduce an idea of regularization in the game.

Suppose that in the regularized game, when the strategies of player I and player II are i and j, respectively, the payoffs of player I and player II are $\xi-\frac{\delta}{2}(p_i-q_j)$ and $\eta-\frac{\delta}{2}(q_j-p_i)$, respectively ($i=1, \cdots, m$; $j=1, \cdots, l$), where $\delta>0$ is the regularization parameter. Then the expected payoff for player I is given by

$$V_\delta(p, q)=V(p, q)-\frac{\delta}{2}(\|p\|^2-\|q\|^2) \tag{5}$$

and one for player II is $-V_\delta(p, q)$, where $\|\cdot\|$ is Euclidean norm.

Suppose, moreover, that the mixed strategies available to players are in ε-simplices, i. e., $p\in S_\varepsilon^m$, $q\in S_\varepsilon^l$, where

$$S_\varepsilon^k=\left\{x=(x_1, \cdots, x_k);\ x_i\geqq\varepsilon(i=1, \cdots, k), \sum_{i=1}^{k} x_i=1,\ \left(0\leqq\varepsilon\leqq\frac{1}{k}\right)\right\}.$$

Thus, for any fixed ε and δ, this game is 2-person zero-sum game.

The optimal strategies for the players in this game are $p^*(\varepsilon, \delta)$ and $q^*(\varepsilon, \delta)$ such that

$$V_\delta(p^*(\varepsilon, \delta), q)\geqq V_\delta(p^*(\varepsilon, \delta), q^*(\varepsilon, \delta))\geqq V_\delta(p, q^*(\varepsilon, \delta)) \tag{6}$$

for all $p\in S_\varepsilon^m$ and $q\in S_\varepsilon^l$. It is not difficult to show that there is a saddle point of the function $V_\delta(p, q)$ for any fixed $\delta\geqq 0$, $\varepsilon\in[0, \hat{\varepsilon}]$ ($\hat{\varepsilon}=\min(1/m, 1/l)$). Further, since $V_\delta(p, q)$ is

strictly convex for $\delta>0$, it has a unique saddle point $(p^*(\varepsilon, \delta), q^*(\varepsilon, \delta))$.

The following two lemmas show the connection of the regularized game with the original game and the properties of the saddle point $(p^*(\varepsilon, \delta), q^*(\varepsilon, \delta))$ as a function of (ε, δ) that are important in order to prove the theorems.

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

$$\varepsilon[n]\in(0, \hat{\varepsilon}), \quad \delta[n]>0,$$

$$\lim_{n\to\infty}\varepsilon[n]=\lim_{n\to\infty}\delta[n]=0$$

and

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

then the squence $\{p^*(\varepsilon[n], \delta[n]), q^*(\varepsilon[n], \delta[n]))\}$ *converges to a saddle point* (p^*, q^*) *of the original game (depending, generally, on* μ*).*

LEMMA 2. *For any payoff matrix* A, *there exist* $\delta'\in(0, \infty)$ *and constants* K_1, K_2, K_3 *such that*

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

$$\leqq K_1|\varepsilon_1-\varepsilon_2|+K_2|\delta_1-\delta_2|+K_3\left|\frac{\varepsilon_1}{\delta_1}-\frac{\varepsilon_2}{\delta_2}\right| \tag{7}$$

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

The proofs of these lemmas are given in [4].

4. Learning algorithm

We use a pseudogradient procedure as the algorithm of learning the optimal mixed strategies $p^*(\varepsilon, \delta)$, $q^*(\varepsilon, \delta)$ in the regularized game for fixed $\varepsilon>0$ and $\delta>0$, and we define the learning algorithm as follows:

At the first stage, player I and player II play independently a single game using any mixed strategies $p[0]$ and $q[0]$, respectively.

Now, let $p[n]$ and $q[n]$ be the mixed strategies obtained by the players at the n-th stage, respectively. Then, the players play a single game using the mixed strategies $p[n]$ and $q[n]$, respectively, at the $(n+1)$-th stage, and suppose that at this play the players choose pure strategies x_{n+1} and y_{n+1} and gain the payoffs $\xi[n+1]$ and $\eta[n+1]$, respectively. At next stage player I and player II use the mixed strategies constructed by, respectively,

$$p[n+1]=\pi_{S^m_{\varepsilon[n+1]}}\{p[n]+\gamma[n+1]A(x_{n+1}, y_{n+1})\} \tag{8a}$$

and

$$q[n+1]=\pi_{S^l_{\varepsilon[n+1]}}\{q[n]+\gamma[n+1]B(x_{n+1}, y_{n+1})\}, \tag{8b}$$

where $\{\varepsilon[n]\}$, $\{\delta[n]\}$ and $\{\gamma[n]\}$ are sequences of numbers; for $i=1,\cdots, m$ and $j=1,\cdots, l$, $A(i, j)$ is m-dimensional vector whose elements are

$$A_k(i, j)=\begin{cases}\dfrac{\xi[n+1]}{p_i[n]}-\delta[n+1], & k=i\\ -\dfrac{1}{m-1}\left(\dfrac{\xi[n+1]}{p_i[n]}-\delta[n+1]\right), & k\neq i,\ k=1,\cdots, m,\end{cases}$$

and $B(i, j)$ is l-dimensional vector whose elements are

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

$\pi_S\{\cdot\}$ is a projection operator on the closed bounded set S that has the property

$$\pi_S\{x\}\in S \text{ and } \|x-y\|\geqq\|\pi_S\{x\}-y\| \tag{9}$$

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

The following theorems give sufficient conditions for a pair of the mixed strategies generated by above algorithms to converge with probability one and in mean square to a saddle poin (p^*, q^*) of the original game. For simplicity, we use the notations $p^*[n]=p^*(\varepsilon[n], \delta[n])$ and $q^*[n]=q^*(\varepsilon[n], \delta[n])$.

THEOREM 1. *The sequence* $\{(p[n], q[n])\}$ *generated by the learning algorithm* (8) *converes with probability one as* $n\to\infty$ *to a saddle point* (p^*, q^*) *of the original game for any initial condition* $(p[0], q[0])\in S^m_{\varepsilon[0]}\times S^l_{\varepsilon[0]}$, *if the sequences* $\{\varepsilon[n]\}$, $\{\delta[n]\}$ *and* $\{\gamma[n]\}$ *satisfy*

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

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

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

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

PROOF. By (8a) and (9),

$$\|p[n+1]-p^*[n+1]\|^2 \leqq \|p[n]-p^*[n]\|^2+4\sqrt{2}\,\|p^*[n+1]-p^*[n]\| +2\gamma[n+1]\langle p[n]-p^*[n],\, A(x_{n+1},\, y_{n+1})\rangle +2\gamma^2[n+1]\,\|A(x_{n+1},\, y_{n+1})\|^2, \tag{10}$$

where $\langle \cdot\,,\,\cdot \rangle$ denotes inner product. Taking conditional expectation of (10) for given $p[n]$ and $q[n]$, we get

$$\begin{aligned}&\mathrm{E}\{\|p[n+1]-p^*[n+1]\|^2 \mid p[n],\, q[n]\}\\ &\leqq \|p[n]-p^*[n]\|^2+4\sqrt{2}\,\|p^*[n+1]-p^*[n]\|\\ &\quad +2\gamma[n+1]\mathrm{E}\{\langle p[n]-p^*[n],\, A(x_{n+1},\, y_{n+1})\rangle \mid p[n],\, q[n]\}\\ &\quad +2\gamma^2[n+1]\mathrm{E}\{\|A(x_{n+1},\, y_{n+1})\|^2 \mid p[n],\, q[n]\}.\end{aligned} \tag{11}$$

Well, for each n, it holds that by (1), (3) and (5)

$$\begin{aligned}&\mathrm{E}\{\langle p[n]-p^*[n],\, A(x_{n+1},\, y_{n+1})\rangle \mid p[n],\, q[n]\}\\ &=\sum_{i=1}^{m}\sum_{j=1}^{l}\left\{\sum_{k=1}^{m}(p_k[n]-p_k^*[n])A_k(i,\, j)\right\}p_i[n]q_j[n]\\ &=\frac{m}{m-1}\sum_{i=1}^{m}\sum_{j=1}^{l}(p_i[n]-p_i^*[n])\left(\frac{r_{ij}}{p_i[n]}-\delta[n+1]\right)p_i[n]q_j[n]\\ &=\frac{m}{m-1}\Big\{V(p[n],\, q[n])-V(p^*[n],\, q[n])\\ &\qquad -\delta[n+1]\left(\sum_{i=1}^{m}p_i^2[n]-\sum_{i=1}^{m}p_i[n]p_i^*[n]\right)\Big\}\\ &=\frac{m}{m-1}\Big\{V_{\delta[n+1]}(p[n],\, q[n])-V_{\delta[n+1]}(p^*[n],\, q[n])\\ &\qquad -\frac{1}{2}\delta[n+1]\|p[n]-p^*[n]\|^2\Big\}\,,\end{aligned} \tag{12}$$

and that by (2)

$$\begin{aligned}&\mathrm{E}\{\|A(x_{n+1},\, y_{n+1})\|^2 \mid p[n],\, q[n]\}\\ &=\sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{m}A_k^2(i,\, j)p_i[n]q_j[n]\\ &\leqq\frac{2m}{m-1}\sum_{i=1}^{m}\sum_{j=1}^{l}\left(\frac{R_{ij}^{(1)}}{p_i^2[n]}+\delta^2[n+1]\right)p_i[n]q_j[n]\\ &\leqq\frac{2m}{m-1}\left\{\frac{1}{\varepsilon[n]}\sum_{i=1}^{m}\sum_{j=1}^{l}R_{ij}^{(1)}\,q_j[n]+\delta^2[n+1]\right\}.\end{aligned} \tag{13}$$

Hence, from (11), (12) and (13) we have

$$\begin{aligned}&\mathrm{E}\{\|p[n+1]-p^*[n+1]\|^2 \mid p[n],\, q[n]\}\\ &\leqq\left(1-\frac{m}{m-1}\gamma[n+1]\,\delta[n+1]\right)\|p[n]-p^*[n]\|^2+4\sqrt{2}\,\|p^*[n+1]-p^*[n]\|\end{aligned}$$

$$+\frac{4m}{m-1}\frac{\gamma^2[n+1]}{\varepsilon[n]}\sum_{i=1}^{m}\sum_{j=1}^{l}R_{ij}^{(1)}q_j[n]+\frac{4m}{m-1}\gamma^2[n+1]\,\delta^2[n+1]$$

$$+\frac{2m}{m-1}\gamma[n+1]\left\{V_{\delta[n+1]}(p[n],\,q[n])-V_{\delta[n+1]}(p^*[n],\,q[n])\right\}. \quad (14)$$

Similarly, we have

$$\mathrm{E}\{\|q[n+1]-q^*[n+1]\|^2\,|p[n],\,q[n]\}$$

$$\leqq\left(1-\frac{l}{l-1}\gamma[n+1]\delta[n+1]\right)\|q[n]-q^*[n]\|^2+4\sqrt{2}\,\|q^*[n+1]-q^*[n]\|$$

$$+\frac{4l}{l-1}\frac{\gamma^2[n+1]}{\varepsilon[n]}\sum_{i=1}^{m}\sum_{j=1}^{l}R_{ij}^{(2)}p_i[n]+\frac{4l}{l-1}\gamma^2[n+1]\,\delta^2[n+1]$$

$$+\frac{2l}{l-1}\gamma[n+1]\left\{V_{\delta[n+1]}(p[n],\,q^*[n])-V_{\delta[n+1]}(p[n],\,q[n])\right\}. \quad (15)$$

Now, we put

$$c[n]=\|p[n]-p^*[n]\|^2+\|q[n]-q^*[n]\|^2$$

and

$$d[n]=\frac{m-1}{m}\|p[n]-p^*[n]\|^2+\frac{l-1}{l}\|q[n]-q^*[n]\|^2.$$

Then, there are constants L_1, L_2 such that for each n

$$L_1d[n]\leqq c[n]\leqq L_2d[n]$$

From Lemma 2, (14) and (15), there exist a number n_0 and positive constants $K_i<\infty$ $(i=1,\cdots,5)$ such that, for $n\geqq n_0$

$$\mathrm{E}\{d[n+1]\,|p[n],\,q[n]\}\leqq d[n]-\gamma[n+1]\delta[n+1]c[n]$$

$$+K_1|\varepsilon[n+1]-\varepsilon[n]|+K_2|\delta[n+1]-\delta[n]|+K_3\left|\frac{\varepsilon[n+1]}{\delta[n+1]}-\frac{\varepsilon[n]}{\delta[n]}\right|$$

$$+K_4\frac{\gamma^2[n+1]}{\varepsilon[n]}+K_5\gamma^2[n+1]$$

$$+2\gamma[n+1]\{V_{\delta[n+1]}(p[n],\,q^*[n])-V_{\delta[n+1]}(p^*[n],\,q[n])\}.$$

And, using the definition of a saddle point, we obtain

$$\mathrm{E}\{d[n+1]\,|p[n],\,q[n]\}\leqq(1-L_1\gamma[n+1]\delta[n+1])d[n]+\beta[n+1], \quad (16)$$

where

$$\beta[n+1]=K_1|\varepsilon[n+1]-\varepsilon[n]|+K_2|\delta[n+1]-\delta[n]|+K_3\left|\frac{\varepsilon[n+1]}{\delta[n+1]}-\frac{\varepsilon[n]}{\delta[n]}\right|$$

$$+K_4\frac{\gamma^2[n+1]}{\varepsilon[n]}+K_5\gamma^2[n+1]\delta^2[n+1].$$

From (16), we have

$$\mathrm{E}\{d[n+1]\,|\,p[n],\ q[n]\}\leqq d[n]+\beta[n+1]. \tag{17}$$

Now, we introduce the sequence $D[n]=d[n]+\sum_{k=n+1}^{\infty}\beta[k]$, for which (17) implies that

$$\mathrm{E}\{D[n+1]\,|\,p[n],\ q[n]\}\leqq D[n].$$

Since $D[n]\geqq 0$, it follows that there is a constant $D\geqq 0$ such that $D[n]\longrightarrow D$ w.p.1 as $n\rightarrow\infty$, hence also by conditions $(d)\sim(h)$,

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

Taking the expectations of the both sides of (16) and summing the obtained inequalities with respect to n from n_0 to ∞, it follows that

$$\sum_{n=n_0}^{\infty}\gamma[n+1]\delta[n+1]\mathrm{E}\{d[n]\}<\infty.$$

Then, by the condition (c) there exists a subsequence $\{n_k\}$ such that

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

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

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

Thus, Theorem 1 is proved by Lemma 1.

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

$$(d')\quad \gamma[n+1]\delta[n+1]\longrightarrow 0 \quad as \quad n\longrightarrow\infty,$$

$$(e')\quad \frac{\gamma[n+1]}{\varepsilon[n]\delta[n+1]}\longrightarrow 0 \quad as \quad n\longrightarrow\infty,$$

$$(f')\quad \frac{1}{\gamma[n+1]\delta[n+1]}\left|\varepsilon[n+1]-\varepsilon[n]\right|\longrightarrow 0 \quad as \quad n\longrightarrow\infty,$$

$$(g')\quad \frac{1}{\gamma[n+1]\delta[n+1]}\left|\delta[n+1]-\delta[n]\right|\longrightarrow 0 \quad as \quad n\longrightarrow\infty,$$

$$(h')\quad \frac{1}{\gamma[n+1]\delta[n+1]}\left|\frac{\varepsilon[n+1]}{\delta[n+1]}-\frac{\varepsilon[n]}{\delta[n]}\right|\longrightarrow 0 \quad as \quad n\longrightarrow\infty,$$

then the sequence $\{(p[n],\ q[n])\}$ *converges in mean square to saddle point* $(p^*,\ q^*)$ *of the original game.*

The following theorem gives an upper bound for the rate of convergence of the algorithm (8) in mean square.

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

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

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

where

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

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

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

The proofs of Theorem 2 and Theorem 3 can be shown by the similar argument to Theorem 2 and Lemma 3 of [4], respectively.

We consider in detail a case when the sequences in the algorithm (8) are such that $\gamma[n]\sim 1/n^\alpha$, $\varepsilon[n]\sim 1/n^\beta$, $\delta[n]\sim 1/n^\sigma$, $(\varepsilon[n]/\delta[n]-\mu)\sim 1/n^\nu$ for $\beta=\sigma$ and $\sim 1/n^{\beta-\sigma}$ for $\beta>\sigma$, 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 Theorem 2, it follows that for the convergence of the algorithm with probability one it is sufficient to choose α, β, σ and ν such that

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

and for mean square convergence,

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

References

[1] TSYPKIN, Ya. Z., Adaptation and learning in automatic systems, Academic Press, New York and London, 1971.

[2] CRAWFORD, V. P., *Learning the optimal strategy in a zero-sum game*, Econometrika, Vol. 42, No. 5 (1974), 885–891.

[3] SADOVSKY, A. L., *Monotone iterative algorithm of a solution of matrix games*, Dokl. Akad. Nauk SSSR, Vol. 238, No. 3 (1978), 538–540, (in Russian).

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