ON OPTIMAL NON-RANDOM STATIONARY POLICIES IN FINITE STATE STOCHASTIC GAMES

Kai, Yu
Department of Mathematics, Kyushu University

https://doi.org/10.5109/13075

出版情報：統計数理研究. 15 (3/4), pp.93-99, 1973-03. Research Association of Statistical Sciences
バージョン：
権利関係：

ON OPTIMAL NON-RANDOM STATIONARY POLICIES IN FINITE STATE STOCHASTIC GAMES

By

Yû KAI*

(Received October 2, 1972)

§ 1. Introduction.

A stochastic game is determined by five objects: S, A, B, P, r. Here S is a state space of N points, $1, 2, \cdots, N$; A is a set of actions available to Player I; B is a set of actions available to Player II; P is the low of motion of the system—it associates with each pair $a \in A$, $b \in B$ a trasition probability $P_{ij}(a, b)$ for $i, j \in S$; and r, the immediate reward, is a function on $S \times A \times B$. Throughout this paper we are concerned with the non-random Markov policies only, then a policy π for Player I is a sequence $(f_1, f_2, \cdots)$, where each f_n is a mapping from S into A, and the policy chooses an action $f_n(i)$ in a state i at n-th day; a policy is said to be *stationary* if $f_n = f$ for some mapping f from S into A for all n, and in this case π is denoted by f^∞. A policy and a stationary policy for Player II are defined analogiously, and denoted by $\sigma \equiv (g_1, g_2, \cdots)$ and $\sigma \equiv g^\infty$ respectively.

Let $X_1, X_2, \cdots$ be a Markov process on the state space S. The expected total reward with initial state i from a pair (π, σ) of policies for Player I and II is given by

$$I(\pi, \sigma)_i \equiv E\Big[\sum_{n=1}^{\infty} \beta^{n-1} r(X_n, f_n(X_n), g_n(X_n)) \,|\, \pi \text{ and } \sigma \text{ are used and } X_1 = i\Big],$$

where β is a fixed discount factor, $0 \leqq \beta < 1$, such that a reward at n-th day in future is worth β^n times now. In the stochastic game, then, Player I wishes to choose π so that each component of the vector $I(\pi, \sigma) = (I(\pi, \sigma)_i,\ i = 1, \cdots, N)$ is maximized in some sense, and Player II wishes to choose σ so that $I(\pi, \sigma)$ simultaneously minimized in some sense. A policy π^* is *optimal* for Player I if $\inf_\sigma \sup_\pi I(\pi, \sigma)_i \leqq I(\pi^*, \sigma')_i$ for all σ' and $i \in S$, and a policy σ^* is *optimal* for Player II if $\sup_\pi \inf_\sigma I(\pi, \sigma)_i \geqq I(\pi', \sigma^*)_i$ for all π' and $i \in S$. We shall say that the game is *strictly determined* if $\sup_\pi \inf_\sigma I(\pi, \sigma)_i = \inf_\sigma \sup_\pi I(\pi, \sigma)_i$ for all $i \in S$.

Throughout this paper we impose the following assumptions: (A1) A and B are compact convex sets; (A2) $P_{ij}(a, b)$ is a continuous and concave-convex function on $A \times B$ for each pair $i, j \in S$; (A3) $r(i, a, b)$ is bounded on $S \times A \times B$, i.e. $\sup_{i,a,b} |r(i, a, b)| \equiv R < \infty$, and for each fixed $i \in S$, is a continuous concave-convex function on $A \times B$.

---

* Department of Mathematics, Kyushu University, Fukuoka.

Under the assumptions stated above this paper shows that the game is strictly determined and that both players have optimal stationary policies. Furthermore a computational procedures for finding ε-optimal policies is given. Kushner and Chamberlain [1] treated these problems in the case where the policies feasible to both players were restricted to the stationary ones.

§ 2. Some lemmas.

In this section we shall prove several lemmas concerning the expected total reward by virtue of new-defined operators. For each pair (f, g), where f is a mapping from S into A and g is from S into B, we define an operator L_{fg} on N-dimentional real vector space V as follows: for $v\equiv(v_1, \cdots, v_N)\in V$,

$$L_{fg}v\equiv((L_{fg}v)(i),\ i=1, \cdots, N)\,, \tag{2.1}$$

where $(L_{fg}v)(i)\equiv r(i, f(i), g(i))+\beta\sum_{j=1}^{N}P_{ij}(f(i), g(i))v_j$, for each $i\in S$. For each pair (π, σ) of policies we let

$$I_n(\pi, \sigma\,;\ v)\equiv L_{f_1g_1}L_{f_2g_2}\cdots L_{f_ng_n}v\,, \qquad v\in V\,, \tag{2.2}$$

where $L_{f_1g_1}L_{f_2g_2}v\equiv L_{f_1g_1}(L_{f_2g_2}v)$. Denoting the vector $(r(i, f(i), g(i)),\ i=1, \cdots, N)$ by $r(f, g)$ and the $N\times N$ matrix $(P_{ij}(f(i), g(i)))$ by $P(f, g)$, then, (2.2) can be expressed as follows:

$$\begin{aligned} I_n(\pi, \sigma\,;\ v)&=r(f_1, g_1)\\ &\quad+\sum_{k=1}^{n-1}\beta^k\prod_{l=1}^{k}P(f_l, g_l)r(f_{k+1}, g_{k+1})+\beta^n\prod_{l=1}^{n}P(f_l, g_l)v\,, \end{aligned} \tag{2.3}$$

where $\prod_{l=1}^{k}P(f_l, g_l)\equiv P(f_1, g_1)\cdots P(f_k, g_k)$. Similarly $I(\pi, \sigma)$ is expressed by

$$I(\pi, \sigma)=r(f_1, g_1)+\sum_{k=1}^{\infty}\beta^k\prod_{l=1}^{k}P(f_l, g_l)r(f_{k+1}, g_{k+1})\,. \tag{2.4}$$

LEMMA 2.1. (a) $\sup_{\pi,\sigma}\|I(\pi, \sigma)\|\leqq\frac{R}{1-\beta}$, *where* $\|v\|=\max_i|v_i|$ *for* $v\in V$.

(b) *For any* $v\in V$, $I_n(\pi, \sigma\,;\ v)$ *converges to* $I(\pi, \sigma)$ *as* $n\to\infty$.

PROOF. (a) Since by (A3) $\sup_{i,a,b}|r(i, a, b)|=R<\infty$, from (2.4)

$$\|I(\pi, \sigma)\|\leqq\sum_{k=0}^{\infty}\beta^kR=\frac{R}{1-\beta}\qquad\text{for any pair } (\pi, \sigma)\,.$$

(b) From (2.3) and (2.4), for any $\pi=(f_1, f_2, \cdots)$ and $\sigma=(g_1, g_2, \cdots)$,

$$\begin{aligned} &I(\pi, \sigma)-I_n(\pi, \sigma\,;\ v)\\ &=\beta^n\prod_{l=1}^{n}P(f_l, g_l)\Big\{r(f_{n+1}, g_{n+1})+\sum_{k=1}^{\infty}\beta^k\prod_{l=1}^{k}P(f_{n+l}, g_{n+l})r(f_{n+k+1}, g_{n+k+1})-v\Big\}\,. \end{aligned} \tag{2.5}$$

Here, it is noted that the term in the brace in the righthand side of (2.5) expresses the expected total reward from the pair of policies ${}^n\pi\equiv(f_{n+1}, f_{n+2}, \cdots)$ and ${}^n\sigma\equiv(g_{n+1}, g_{n+2}, \cdots)$. Hence, by (a) of Lemma 2.1,

$$\left\| r(f_{n+1}, g_{n+1}) + \sum_{k=1}^{\infty} \beta^k \prod_{l=1}^{k} P(f_{n+l}, g_{n+l}) r(f_{n+k+1}, g_{n+k+1}) \right\| \leqq \frac{R}{1-\beta} .$$

Thus

$$\|I(\pi, \sigma) - I_n(\pi, \sigma ; v)\| \leqq \beta^n \left(\frac{R}{1-\beta} + \|v\| \right),$$

which yields that $I_n(\pi, \sigma ; v)$ converges to $I(\pi, \sigma)$ as $n \to \infty$.

LEMMA 2.2. *If both of $f(a, b)$ and $g(a, b)$ are concave-convex functions on $A \times B$, then $\alpha f(a, b) + \alpha' g(a, b)$ is a concave-convex function on $A \times B$ for $\alpha, \alpha' \geqq 0$.*

PROOF. This Lemma is clear from the definition of the concave-convex function on $A \times B$.

Next we give a minimax lemma useful for our stochastic game.

LEMMA 2.3. *For any vector $v = (v_1, \cdots, v_N) \in V$,*

$$\text{(2.6)} \qquad \max_a \min_b \left\{ r(i, a, b) + \beta \sum_{j=1}^{N} P_{ij}(a, b) v_j \right\}$$
$$= \min_b \max_a \left\{ r(i, a, b) + \beta \sum_{j=1}^{N} P_{ij}(a, b) v_j \right\} \qquad \textit{for all} \quad i \in S .$$

PROOF. By the assumptions (A1), (A2), (A3) and Lemma 2.2, $r(i, a, b) + \beta \sum_{j=1}^{N} P_{ij}(a, b)(v_j + \|v\|)$ is a continuous concave-convex function on $A \times B$ for each $i \in S$. Then, by the general minimax theorem (cf. [4]), it holds that

$$\text{(2.7)} \qquad \max_a \min_b \left\{ r(i, a, b) + \beta \sum_{j=1}^{N} P_{ij}(a, b)(v_j + \|v\|) \right\}$$
$$= \min_b \max_a \left\{ r(i, a, b) + \beta \sum_{j=1}^{N} P_{ij}(a, b)(v_j + \|v\|) \right\} \qquad \text{for all} \quad i \in S .$$

On the other hand, we get

$$\text{(2.8)} \qquad \max_a \min_b \left\{ r(i, a, b) + \beta \sum_{j=1}^{N} P_{ij}(a, b)(v_j + \|v\|) \right\}$$
$$= \max_a \min_b \left\{ r(i, a, b) + \beta \sum_{j=1}^{N} P_{ij}(a, b) v_j \right\} + \beta \|v\| ,$$

and

$$\text{(2.9)} \qquad \min_b \max_a \left\{ r(i, a, b) + \beta \sum_{j=1}^{N} P_{ij}(a, b)(v_j + \|v\|) \right\}$$
$$= \min_b \max_a \left\{ r(i, a, b) + \beta \sum_{j=1}^{N} P_{ij}(a, b) v_j \right\} + \beta \|v\| , \qquad \text{for all} \quad i \in S .$$

Thus, from (2.7), (2.8) and (2.9), we get (2.6), which completes the proof.

Now we define an operator T on V as follows:

$$Tv \equiv ((Tv)(i),\ i = 1, \cdots, N), \qquad v \in V ,$$

where $(Tv)(i) \equiv \max_a \min_b \left\{ r(i, a, b) + \beta \sum_{j=1}^{N} P_{ij}(a, b) v_j \right\}$ for every $i \in S$.

LEMMA 2.4. *The operator T is a contraction mapping on V, and has an unique fixed point $v^* \in V$, i.e. $Tv^* = v^*$.*

PROOF. For vectors $u, v \in V$, plainly $u \leqq v + \|u - v\| \mathbf{1}$, where $\mathbf{1}$ is the identity of

V. Since it is clear from the definition that T is monotone,

$$Tu \leqq T(v+\|u-v\|\mathbf{1}) = Tv+\beta\|u-v\|\mathbf{1},$$

and consequently, $Tu-Tv \leqq \beta\|u-v\|\mathbf{1}$. Similarly $Tv-Tu \leqq \beta\|u-v\|\mathbf{1}$. Thus we get $\|Tu-Tv\| \leqq \beta\|u-v\|$, which shows that T is a contraction mapping on V because of the discount factor β.

Since V is the Banach space with the supremum norm, T has an unique fixed point $v^* \in V$ by virtue of the Banach fixed-point theorem. Thus the Lemma is proved.

§ 3. Optimal stationary policies.

In this section we give our main theorem, the existence of optimal stationary policies. The proof of it is very constructive.

THEOREM 3.1. *The game is strictly determined, and players I and II have optimal stationary policies.*

PROOF. By the assumptions (A1), (A2), (A3) and lemmas 2.2, 2.3, it holds that

$$\begin{aligned} v_i^* &= \max_a \min_b \left\{r(i, a, b)+\beta\sum_{j=1}^N P_{ij}(a, b)v_j^*\right\} \\ &= \min_b \max_a \left\{r(i, a, b)+\beta\sum_{j=1}^N P_{ij}(a, b)v_j^*\right\} \qquad \text{for all} \quad i\in S, \end{aligned}$$

and furthermore there exist two sequences $\{a_i\in A, i=1, \cdots, N\}$ and $\{b_i\in B, i=1, \cdots, N\}$ such that

$$\begin{aligned} v_i^* &= \min_b \left\{r(i, a_i, b)+\beta\sum_{j=1}^N P_{ij}(a, b)v_j^*\right\} \\ &= \max_a \left\{r(i, a, b_i)+\beta\sum_{j=1}^N P_{ij}(a, b)v_j^*\right\} \qquad \text{for all} \quad i\in S. \end{aligned} \tag{3.1}$$

We now define the functions f_* from S into A and g_* from S into B by

$$f_*(i) \equiv a_i, \qquad g_*(i) \equiv b_i \qquad \text{for each} \quad i \in S.$$

Then (3.1) is expressed as follows:

$$\begin{aligned} v_i^* &= \min_b \left\{r(i, f_*(i), b)+\beta\sum_{j=1}^N P_{ij}(f_*(i), b)v_j^*\right\} \\ &= \max_a \left\{r(i, a, g_*(i))+\beta\sum_{j=1}^N P_{ij}(a, g_*(i))v_j^*\right\} \qquad \text{for} \quad i\in S. \end{aligned} \tag{3.2}$$

Let fix $i\in S$ arbitrary. For any policies $\pi=(f_1, f_2, \cdots)$ and $\sigma=(g_1, g_2, \cdots)$, by (2.1) and (3.2),

$$\begin{aligned} v_i^* &\leqq r(i, f_*(i), g_n(i))+\beta\sum_{j=1}^N P_{ij}(f_*(a), g_n(i))v_j^* \\ &= (L_{f_* g_n}v^*)(i), \\ v_i^* &\geqq r(i, f_n(i), g_*(i))+\beta\sum_{i=1}^N P_{ij}(f_n(i), g_*(i))v_j^* \\ &= (L_{f_n g_*}v^*)(i), \qquad \text{for all} \quad n \geqq 1. \end{aligned}$$

Since $L_{f_* g_n}$ and $L_{f_n g_*}$ are monotone for each $n \geqq 1$, as are easily shown by its definition,

$$v_i^* \leqq (L_{f_* g_1} L_{f_* g_2} \cdots L_{f_* g_n} v^*)(i) = I_n(f_*, \sigma : v^*)_i ,$$

$$v_i^* \geqq (L_{f_1 g_*} L_{f_2 g_*} \cdots L_{f_n g_*} v^*)(i) = I_n(\pi, g_* : v^*)_i , \qquad \text{for all} \quad n \geqq 1 ,$$

where $I_n(\pi, \sigma : v) \equiv (I_n(\pi, \sigma : v)_i,\ i=1, \cdots, N)$. By Lemma 2.1 (b), $I_n(f_*^\infty, \sigma : v^*)$ converges to $I(f_*^\infty, \sigma)$ and $I_n(\pi, g_*^\infty : v^*)$ to $I(\pi, g_*^\infty)$ as $n \to \infty$. Hence

$$I(\pi, g_*^\infty)_i \leqq v_i^* \leqq I(f_*^\infty, \sigma)_i .$$

Since π, σ and $i \in S$ are arbitrary,

$$\sup_\pi I(\pi, g_*^\infty)_i \leqq v_i^* \leqq \inf_\sigma I(f_*^\infty, \sigma)_i , \qquad \text{for all} \quad i \in S .$$

Then we have

$$\begin{aligned} \inf_\sigma \sup_\pi I(\pi, \sigma)_i &\leqq \sup_\pi I(\pi, g_*^\infty)_i \\ &\leqq \inf_\sigma I(f_*^\infty, \sigma)_i \\ &\leqq \sup_\pi \inf_\sigma I(\pi, \sigma)_i \qquad \text{for all} \quad i \in S . \end{aligned}$$

Generally it is true that

$$\inf_\sigma \sup_\pi I(\pi, \sigma)_i \geqq \sup_\pi \inf_\sigma I(\pi, \sigma)_i \qquad \text{for all} \quad i \in S .$$

Therefore we have

$$\begin{aligned} \inf_\sigma \sup_\pi I(\pi, \sigma)_i &= \sup_\pi I(\pi, g_*^\infty)_i \\ &= \inf_\sigma I(f_*^\infty, \sigma)_i \\ &= \sup_\pi \inf_\sigma I(\pi, \sigma)_i \qquad \text{for all} \quad i \in S . \end{aligned}$$

Thus our game is strictly determined, and f_*^∞ and g_*^∞ are optimal stationary policies for Player I and Player II respectively.

§ 4. Computational procedures of ε-optimal policies.

Let v^0 be any vector of V and we shall define the sequence $\{v^n, n=1, 2, \cdots\}$ by

$$v^n \equiv T v^{n-1} , \qquad n = 1, 2, \cdots ,$$

where T is the operator defined in § 2.

LEMMA 4.1. *The sequence $\{v^n, n=1, 2, \cdots\}$ converges to v^* which is the fixed point of the operator T.*

PROOF. By the definition of T,

$$\|v^n - v^{n+1}\| \leqq \beta^n \|v^0 - T v^0\| \qquad \text{for all} \quad n \geqq 1 , \tag{4.1}$$

hence $\{v^n\}$ is a Caucy-sequence. Thus $\{v^n, n=1, 2, \cdots\}$ converges to v^*, for V is a Banach space and T has an unique fixed point v^*.

THEOREM 4.1. *Fix any $\varepsilon > 0$. Then, for sufficiently large n such that*

(4.2)
$$\beta^n \leqq \frac{(1-\beta)^2\varepsilon}{3\|v^0-Tv^0\|},$$

it holds that

$$\|v^n-v^*\| \leqq \frac{(1-\beta)\varepsilon}{3},$$

and we can choose $f_{n(\varepsilon)}$ *and* $g_{n(\varepsilon)}$ *such that*

$$I(\pi', g^{\infty}_{n(\varepsilon)})_i-\varepsilon \leqq \sup_{\pi}\inf_{\sigma} I(\pi, \sigma)_i \qquad \textit{for all} \quad \pi' \textit{ and } i\in S,$$

$$I(f^{\infty}_{n(\varepsilon)}, \sigma')_i+\varepsilon \geqq \inf_{\sigma}\sup_{\pi} I(\pi, \sigma)_i \qquad \textit{for all} \quad \sigma' \textit{ and } i\in S.$$

[*This shows that* $f^{\infty}_{n(\varepsilon)}$ *and* $g^{\infty}_{n(\varepsilon)}$ *are* ε*-optimal policies for Players I and II respectively*].

PROOF. By Lemma 4.1 and (4.1),

$$\|v^n-v^*\| \leqq \beta^n\|v^0-Tv^0\|(1+\beta+\cdots)$$
$$=\frac{\beta^n\|v^0-Tv^0\|}{1-\beta}.$$

Then trivially

(4.3)
$$\|v^n-v^*\| \leqq \frac{(1-\beta)\varepsilon}{3}$$

for sufficiently large n for which (4.2) holds, and similarly

(4.4)
$$\|v^{n+1}-v^*\| \leqq \frac{(1-\beta)\varepsilon}{3}.$$

By Lemma 2.3 and the definitions of $\{v^n, n=1, 2, \cdots\}$ and of T, there exist $\{a_i\in A, i=1, \cdots, N\}$ and $\{b_i\in B, i=1, \cdots, N\}$ such that

(4.5)
$$v_i^{n+1} \leqq \min_b \Big\{r(i, a_i, b)+\beta\sum_{j=1}^{N}P_{ij}(a_i, b)v_j^n\Big\}+\frac{(1-\beta)\varepsilon}{3},$$

(4.6)
$$v_i^{n+1} \geqq \max_a \Big\{r(i, a, b_i)+\beta\sum_{j=1}^{N}P_{ij}(a, b_i)v_j^n\Big\}-\frac{(1-\beta)\varepsilon}{3}.$$

Now we define the functions $f_{n(\varepsilon)}: S\to A$ and $g_{n(\varepsilon)}: S\to B$ by

$$f_{n(\varepsilon)}(i)\equiv a_i, \qquad g_{n(\varepsilon)}(i)\equiv b_i, \qquad \text{for} \quad i=1, \cdots, N.$$

By (4.3), (4.4), (4.5) and (4.6), then we have

(4.7)
$$r(i, a, g_{n(\varepsilon)}(i))+\beta\sum_{j=1}^{N}P_{ij}(a, g_{n(\varepsilon)}(i))v_j^*-(1-\beta)\varepsilon$$
$$\leqq v_i^* \leqq r(i, f_{n(\varepsilon)}(i), b)+\beta\sum_{j=1}^{N}P_{ij}(f_{n(\varepsilon)}(i), b)v_j^*+(1-\beta)\varepsilon,$$

for all $a\in A$, $b\in B$ and $i\in S$. The above inequality (4.7) implies that for any function f and g

$$(L_{fg_{n(\varepsilon)}}v^*)(i)-(1-\beta)\varepsilon \leqq v_i^* \leqq (L_{f_{n(\varepsilon)}g}v^*)(i)+(1-\beta)\varepsilon.$$

Then, for any policies $\pi=(f_1, f_2, \cdots)$, $\sigma=(g_1, g_2, \cdots)$ and for any integer $m\geqq 1$,

$$I_m(\pi, g^{\infty}_{n(\varepsilon)} : v^*)_i-(1-\beta)\varepsilon(1+\beta+ \cdots +\beta^{m-1})$$
$$\leqq v_i^* \leqq I_m(f^{\infty}_{n(\varepsilon)}, \sigma : v^*)_i+(1-\beta)\varepsilon(1+\beta+ \cdots +\beta^{m-1}) .$$

By Lemma 2.1 (b)

$$I(\pi, g^{\infty}_{n(\varepsilon)})_i-\varepsilon \leqq v_i^* \leqq I(f^{\infty}_{n(\varepsilon)}, \sigma)_i+\varepsilon ,$$

for all π, σ and $i \in S$. By appealing to the proof of Theorem 3.1 we have $v_i^* = \sup_{\pi} \inf_{\sigma} I(\pi, \sigma)_i = \inf_{\sigma} \sup_{\pi} I(\pi, \sigma)_i$ for all $i \in S$. Thus we find that $f_{n(\varepsilon)}$ and $g_{n(\varepsilon)}$ satisfy the inequalities required in the theorem.

References

[1] H. J. Kushner and S. G. Chamberlain, *Finite state stochastic games: Existence theorems and Computational procedures,* IEEE. Trans. Automatic Control, Vol. AC-14, No. 3, June. 1969.

[2] R. E. Strauch, *Negative dynamic programming,* Ann. Math. Statist. Vol. 37, 1966.

[3] T. Parthasarathy and T. S. E. Raghavan, *Some topics in Two-Person Games,* American Elsevier, New York, 1971.

[4] S. Karlin, *Mathematical Methods and Theory in Games: Programming and Economics,* Vol. II, Addison-Wesley, London, 1959.