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, $1, 2, \dots, N$; A is a set of actions available to Player I; B is a set of actions available to Player II; P is the law of motion of the system—it associates with each pair $a \in A, b \in B$ a transition 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 $(f_1, f_2, \dots)$, 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 analogously, and denoted by $\sigma \equiv (g_1, g_2, \dots)$ and $\sigma \equiv g^\infty$ respectively.

Let $X_1, X_2, \dots$ 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

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

where β is a fixed discount factor, $0 \leq \beta < 1$, 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 $I(\pi, \sigma) = (I(\pi, \sigma)_i, i = 1, \dots, N)$ 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 $\inf_{\sigma} \sup_{\pi} I(\pi, \sigma)_i \leq I(\pi^*, \sigma')_i$ for all σ' and $i \in S$, and a policy σ^* is *optimal* for Player II if $\sup_{\pi} \inf_{\sigma} I(\pi, \sigma)_i \geq I(\pi', \sigma^*)_i$ 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 -dimensional real vector space V as follows: for $v \equiv (v_1, \dots, v_N) \in V$,

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

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

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

where $L_{f_1g_1}L_{f_2g_2}v \equiv L_{f_1g_1}(L_{f_2g_2}v)$. Denoting the vector $(r(i, f(i), g(i)), i = 1, \dots, N)$ 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:

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

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

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

LEMMA 2.1. (a) $\sup_{\pi, \sigma} \|I(\pi, \sigma)\| \leq \frac{R}{1-\beta}$, 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 $n \rightarrow \infty$.

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

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

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

$$(2.5) \quad I(\pi, \sigma) - I_n(\pi, \sigma; v) = \beta^n \prod_{l=1}^n P(f_l, g_l) \left\{ r(f_{n+1}, g_{n+1}) + \sum_{k=1}^{\infty} \beta^k \prod_{l=1}^k P(f_{n+l}, g_{n+l}) r(f_{n+k+1}, g_{n+k+1}) - v \right\}.$$

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 ${}^n\pi \equiv (f_{n+1}, f_{n+2}, \dots)$ and ${}^n\sigma \equiv (g_{n+1}, g_{n+2}, \dots)$. Hence, by (a) of Lemma 2.1,

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

Thus

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

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

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 $\alpha, \alpha' \geq 0$.*

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 $v = (v_1, \dots, v_N) \in V$,*

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

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

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

On the other hand, we get

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

and

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

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:

$$Tv \equiv ((Tv)(i), i=1, \dots, N), \quad v \in V,$$

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 $u \leq v + \|u - v\| \mathbf{1}$, where $\mathbf{1}$ is the identity of

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

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

and consequently, $Tu - Tv \leq \beta\|u - v\|\mathbf{1}$. Similarly $Tv - Tu \leq \beta\|u - v\|\mathbf{1}$. Thus we get $\|Tu - Tv\| \leq \beta\|u - v\|$, 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

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

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

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

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

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

Then (3.1) is expressed as follows:

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

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

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

Since $L_{f \cdot g_n}$ and $L_{f_n g \cdot}$ are monotone for each $n \geq 1$, as are easily shown by its definition,

$$\begin{aligned} v_i^* &\leq (L_{f \cdot g_1} L_{f \cdot g_2} \cdots L_{f \cdot g_n} v^*)(i) = I_n(f_*, \sigma : v^*)_i, \\ v_i^* &\geq (L_{f_1 g \cdot} L_{f_2 g \cdot} \cdots L_{f_n g \cdot} v^*)(i) = I_n(\pi, g_* : v^*)_i, \quad \text{for all } n \geq 1, \end{aligned}$$

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

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

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

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

Then we have

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

Generally it is true that

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

Therefore we have

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

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 $\{v^n, n = 1, 2, \dots\}$ by

$$v^n \equiv T v^{n-1}, \quad n = 1, 2, \dots,$$

where T is the operator defined in § 2.

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

PROOF. By the definition of T ,

$$(4.1) \quad \|v^n - v^{n+1}\| \leq \beta^n \|v^0 - T v^0\| \quad \text{for all } n \geq 1,$$

hence $\{v^n\}$ is a Cauchy-sequence. Thus $\{v^n, n = 1, 2, \dots\}$ 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*

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

it holds that

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

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

$$I(\pi', g_{n(\varepsilon)})_i - \varepsilon \leq \sup_{\pi} \inf_{\sigma} I(\pi, \sigma)_i \quad \text{for all } \pi' \text{ and } i \in S,$$

$$I(f_{n(\varepsilon)}, \sigma')_i + \varepsilon \geq \inf_{\sigma} \sup_{\pi} I(\pi, \sigma)_i \quad \text{for all } \sigma' \text{ and } i \in S.$$

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

PROOF. By Lemma 4.1 and (4.1),

$$\begin{aligned} \|v^n - v^*\| &\leq \beta^n \|v^0 - Tv^0\| (1 + \beta + \dots) \\ &= \frac{\beta^n \|v^0 - Tv^0\|}{1 - \beta}. \end{aligned}$$

Then trivially

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} & I_m(\pi, g_{n(\varepsilon)}^\infty : v^*)_i - (1-\beta)\varepsilon(1+\beta+\dots+\beta^{m-1}) \\ & \leq v_i^* \leq I_m(f_{n(\varepsilon)}^\infty, \sigma : v^*)_i + (1-\beta)\varepsilon(1+\beta+\dots+\beta^{m-1}). \end{aligned}$$

By Lemma 2.1 (b)

$$I(\pi, g_{n(\varepsilon)}^\infty)_i - \varepsilon \leq v_i^* \leq I(f_{n(\varepsilon)}^\infty, \sigma)_i + \varepsilon,$$

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 $f_{n(\varepsilon)}^\infty$ and $g_{n(\varepsilon)}^\infty$ 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.