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

Kai, Yu
Department of Mathematics, Kyushu University

http://hdl.handle.net/2324/13075
ON OPTIMAL NON-RANDOM STATIONARY POLICIES IN FINITE STATE STOCHASTIC GAMES

By

Yu 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, ··· , 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 ∈ A, b ∈ 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 $\pi$ for Player I is a sequence $(f_1, f_2, ···)$, 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 $\pi$ is denoted by $f^\infty$. A policy and a stationary policy for Player II are defined analogously, and denoted by $\sigma = (g_1, g_2, ···)$ and $\sigma = g^\infty$ respectively.

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

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

where $\beta$ is a fixed discount factor, $0 \leq \beta < 1$, such that a reward at $n$-th day in future is worth $\beta^n$ times now. In the stochastic game, then, Player I wishes to choose $\pi$ so that each component of the vector $I(\pi, \sigma) = (I(\pi, \sigma)_i, i = 1, ··· , N)$ is maximized in some sense, and Player II wishes to choose $\sigma$ so that $I(\pi, \sigma)$ simultaneously minimized in some sense. A policy $\pi^*$ is optimal for Player I if

$$\inf_{\sigma'} \sup_{\pi} I(\pi, \sigma)_i \leq I(\pi^*, \sigma')_i \text{ for all } \sigma' \text{ and } i \in S,$$

and a policy $\sigma^*$ is optimal for Player II if

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

We shall say that the game is strictly determined if

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

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 \(v \equiv (v_1, \ldots, v_N) \in V\),

\[
L_{fg}v \equiv (L_{fg}v)(i), \quad i = 1, \ldots, N,
\]

where \((L_{fg}v)(i) = r(i, f(i), g(i)) + \sum_{j=1}^{N} \alpha_{ij}(f(i), g(i))v_j\), for each \(i \in S\). For each pair \((\pi, \sigma)\) of policies we let

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

where \(L_{f_1g_2}L_{f_2g_2}v \equiv L_{f_1g_2}(L_{f_2g_2}v)\). Denoting the vector \(r(i, f(i), g(i)), i = 1, \ldots, N\) by \(r(f, g)\) and the \(N \times N\) matrix \((\alpha_{ij}(f(i), g(i)))\) by \(P(f, g)\), then, (2.2) can be expressed as follows:

\[
I_n(\pi, \sigma; v) = r(f_1, g_1)
\]

\[
+ \sum_{k=1}^{N-1} \beta^k \prod_{i=1}^{k} P(f_i, g_i)r(f_{k+1}, g_{k+1}) + \beta^N \prod_{i=1}^{N} P(f_i, g_i)v,
\]

where \(\prod_{i=1}^{k} P(f_i, g_i) \equiv P(f_1, g_1) \cdots P(f_k, g_k)\). Similarly \(I(\pi, \sigma)\) is expressed by

\[
I(\pi, \sigma) = r(f_1, g_1) + \sum_{k=1}^{N-1} \beta^k \prod_{i=1}^{k} P(f_i, g_i)r(f_{k+1}, g_{k+1}),
\]

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 \(n \to \infty\).

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

\[
\|I(\pi, \sigma)\| \leq \sum_{k=0}^{\infty} \beta^k R = \frac{R}{1 - \beta}
\]

for any pair \((\pi, \sigma)\).

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

\[
I(\pi, \sigma) - I_n(\pi, \sigma; v)
\]

\[
= \beta^n \prod_{i=1}^{N} P(f_i, g_i) \left\{ r(f_{n+1}, g_{n+1}) + \sum_{k=1}^{n} \beta^k \prod_{i=1}^{k} P(f_{n+i}, g_{n+i}) r(f_{n+k+1}, g_{n+k+1}) - v \right\}.
\]

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 \(\pi \equiv (f_{n+1}, f_{n+2}, \ldots)\) and \(\sigma \equiv (g_{n+1}, g_{n+2}, \ldots)\). Hence, by (a) of Lemma 2.1,
On Optimal Non-Random Stationary Policies in Finite State Stochastic Games

\[ \|a_r(f_{n+1}, g_{n+1}) + \sum_{k=1}^{\infty} \beta^k \prod_{i=1}^{k} P(f_{n+i}, g_{n+i}) r(f_{n+k+1}, g_{n+k+1}) \| \leq \frac{R}{1 - \beta}. \]

Thus

\[ \| I_\alpha(\pi, \sigma; v) - I_\beta(\pi, \sigma; v) \| \leq \beta^n \left( \frac{R}{1 - \beta} + \| v \| \right), \]

which yields that \( I_\alpha(\pi, \sigma; v) \) 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 \( \alpha, \alpha' \geq 0. \)

**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 \( v = (v_1, \ldots, v_N) \in V, \)

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

**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

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

On the other hand, we get

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

and

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

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:

\[ T \equiv ((Tv)(i), i = 1, \ldots, N), \quad v \in V, \]

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 \( u \leq v + \| u - v \|, \) where \( 1 \) is the identity of
V. Since it is clear from the definition that $T$ is monotone,

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

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

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

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

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

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

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

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

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

Then (3.1) is expressed as follows:

$$v^*_i = \min_b \left\{ r(i, f_*(i), b) + \beta \sum_{j=1}^N P_{ij}(f_*(i), b)v_j^* \right\},$$

$$= \max_a \left\{ r(i, a, g_*(i)) + \beta \sum_{j=1}^N P_{ij}(a, g_*(i))v_j^* \right\} \quad \text{for} \quad i \in S.$$

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),

$$v^*_i \leq r(i, f_*(i), g_*(i)) + \beta \sum_{j=1}^N P_{ij}(f_*(i), g_*(i))v_j^*$$

$$= (L_{f_*(i), g_*(i)}v^*)(i),$$

$$v^*_i \geq r(i, f_*(i), g_*(i)) + \beta \sum_{j=1}^N P_{ij}(f_*(i), g_*(i))v_j^*$$

$$= (L_{f_*(i), g_*(i)}v^*)(i), \quad \text{for all} \quad n \geq 1.$$
Since $L_{f,g}^n$ and $L_{f,g}$ are monotone for each $n \geq 1$, as are easily shown by its definition,

$$v^*_n \leq (L_{f,g_1} \cdots L_{f,g_2} \cdots L_{f,g_3} v^*) (i) = I_n (f^*, \sigma : v^*)_i,$$

$$v^n \leq (L_{f,g_1} \cdots L_{f,g_2} \cdots L_{f,g_3} v^n) (i) = I_n (\pi, g^*_n : v^n)_i, \quad \text{for all } n \geq 1,$$

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

$$I (\pi, g^*_n)_i \leq v^*_n \leq I (f^*_n, \sigma)_i.$$

Since $\pi, \sigma$ and $i \in S$ are arbitrary,

$$\sup_{\pi} I (\pi, g^*_n)_i \leq v^*_n \leq \inf_{\sigma} I (f^*_n, \sigma)_i, \quad \text{for all } i \in S.$$

Then we have

$$\inf_{\sigma} \sup_{\pi} I (\pi, \sigma)_i \leq \sup_{\sigma} \inf_{\pi} I (\pi, g^*_n)_i$$

$$\leq \inf_{\sigma} I (f^*_n, \sigma)_i$$

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

Generally it is true that

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

Therefore we have

$$\inf_{\sigma} \sup_{\pi} I (\pi, \sigma)_i = \sup_{\pi} \inf_{\sigma} I (\pi, g^*_n)_i$$

$$= \inf_{\sigma} I (f^*_n, \sigma)_i$$

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

Thus our game is strictly determined, and $f^*_n$ and $g^*_n$ are optimal stationary policies for Player I and Player II respectively.

§ 4. Computational procedures of $\varepsilon$-optimal policies.

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

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

where $T$ is the operator defined in § 2.

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

**Proof.** By the definition of $T$,

$$\|v^n - v^{n+1}\| \leq 2^n \|v^n - T v^n\| \quad \text{for all } n \geq 1,$$

hence $\{v^n\}$ is a Cauchy-sequence. Thus $\{v^n, n = 1, 2, \ldots\}$ 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
it holds that
\[ \|v^n - v^*\| \leq \frac{(1-\beta)^n \varepsilon}{3} + \frac{\beta^n \|v^0 - T v^0\|}{1-\beta} \]
and we can choose \( f_{n(\varepsilon)} \) and \( g_{n(\varepsilon)} \) such that
\[ I(\pi', g_{n(\varepsilon)}(i), \sigma) \leq \sup_{\pi} \inf_{\sigma} I(\pi, \sigma) \]
for all \( \pi' \) and \( i \in S \),
\[ I(f_{n(\varepsilon)}, \sigma) \geq \inf_{\pi} \sup_{\sigma} I(\pi, \sigma) \]
for all \( \sigma' \) and \( i \in S \).

This shows that \( f_{n(\varepsilon)} \) and \( g_{n(\varepsilon)} \) are \( \varepsilon \)-optimal policies for Players I and II respectively.

**Proof.** By Lemma 4.1 and (4.1),
\[ \|v^n - v^*\| \leq \frac{(1-\beta)^n \varepsilon}{3} + \frac{\beta^n \|v^0 - T v^0\|}{1-\beta} \]
Then trivially
\[ \|v^n - v^*\| \leq \frac{(1-\beta)^n \varepsilon}{3} \]
for sufficiently large \( n \) for which (4.2) holds, and similarly
\[ \|v^{n+1} - v^*\| \leq \frac{(1-\beta)^n \varepsilon}{3} \]
By Lemma 2.3 and the definitions of \( \{v^n, n = 1, 2, \ldots\} \) and of \( T \), there exist \( \{a_i \in A, i = 1, \ldots, N\} \) and \( \{b_i \in B, i = 1, \ldots, N\} \) such that
\[ v^n(i) \leq \min_b \left\{ r(i, a_i, b) + \sum_{j=1}^N P_{ij}(a_i, b)v_j \right\} + \frac{(1-\beta)^n \varepsilon}{3} \]
\[ v^{n+1}(i) \geq \max_a \left\{ r(i, a, b_i) + \sum_{j=1}^N P_{ij}(a, b_i)v_j \right\} - \frac{(1-\beta)^n \varepsilon}{3} \]
Now we define the functions \( f_{n(\varepsilon)} : S \to A \) and \( g_{n(\varepsilon)} : S \to B \) by
\[ f_{n(\varepsilon)}(i) \equiv a_i, \quad g_{n(\varepsilon)}(i) \equiv b_i, \quad \text{for} \quad i = 1, \ldots, N. \]
By (4.3), (4.4), (4.5) and (4.6), then we have
\[ r(i, a, g_{n(\varepsilon)}(i)) + \beta \sum_{j=1}^N P_{ij}(a, g_{n(\varepsilon)}(i))v_j^* - (1-\beta)\varepsilon \]
\[ \leq v_i^* \leq r(i, f_{n(\varepsilon)}(i), b) + \beta \sum_{j=1}^N P_{ij}(f_{n(\varepsilon)}(i), b)v_j^* + (1-\beta)\varepsilon, \]
for all \( a \in A, b \in B \) and \( i \in S \). The above inequality (4.7) implies that for any function \( f \) and \( g \)
\[ (L_{f_{n(\varepsilon)}}v^*)(i) - (1-\beta)\varepsilon \leq v_i^* \leq (L_{f_{n(\varepsilon)}}v^*)(i) + (1-\beta)\varepsilon. \]
Then, for any policies \( \pi = (f_1, f_2, \ldots) \), \( \sigma = (g_1, g_2, \ldots) \) and for any integer \( m \geq 1 \),
\[ I_m(\pi, g_n^{\infty} : v^*) = (1 - \beta)\varepsilon (1 + \beta + \cdots + \beta^{m-1}) \]
\[ \leq v^* \leq I_m(f_n^{\infty}, \sigma : v^*) + (1 - \beta)\varepsilon (1 + \beta + \cdots + \beta^{m-1}). \]

By Lemma 2.1 (b)
\[ v^* \leq I_n(\pi, g_n^{\infty} : v^*), \]

for all \( \pi, \sigma \) 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 \( f_n^{\infty} \) and \( g_n^{\infty} \) satisfy the inequalities required in the theorem.

References