数理解析研究所講究録
第 741 巻 1991 年 15-23

Mutli-armed bandits と ある非協力ゲームの関連について

千葉大学教養部　吉田祐治 (Yuji YOSHIDA

1. d-armed Markov bandit processes.

d : number of arms (positive integer).

$\mathbf{N} = \{ 0,1,2,\cdots \}$: time space.

$(\Omega^i, \mathcal{F}^i, \mathbf{P}^i)$: probability spaces ($i = 1,\cdots,d$).

$X^i = (X_t^i, \mathcal{F}_t^i, \mathbf{P}^i)_{t \in \mathbf{N}}$: mutually independent Markov chains

with Borel state spaces E^i ($i = 1,\cdots,d$).

$\{ \mathcal{F}_t^i \}_{t \in \mathbf{N}}$ is an increasing family of completed sub-σ-fields of $\mathcal{F}^i$

$X = (X_s)_{s \in \mathbf{T}} = (X^1_{s^1}, \cdots, X^d_{s^d})_{s = (s^1,\cdots,s^d) \in \mathbf{T}}$: d-parameter process with state space E

$\mathbf{T} = \mathbf{N}^d$: time space. $E = \prod_{i=1}^d E^i$: state space. $\Omega = \prod_{i=1}^d \Omega^i$: path space.

$\mathbf{P} = \prod_{i=1}^d \mathbf{P}^i$: probability measure. $\mathcal{F}_s = \mathcal{F}^1_{s^1} \otimes \cdots \otimes \mathcal{F}^d_{s^d}$ for $s = (s^1,\cdots,s^d) \in \mathbf{T}$.

T ($\in \mathbf{N}$) : terminal time.

$\mathbf{N}(e,T) = \{ \text{even } t : 0 \le t < T \}$, $\mathbf{N}(o,T) = \{ \text{odd } t : 0 \le t < T \}$, for T ($\in \mathbf{N}$)

β : discount rate ($0 < \beta < 1$). $\mathbf{0} = (0, \cdots, 0) \in \mathbf{T}$. $e_i = (0, \cdots, 0, \overset{i}{1}, 0, \cdots, 0) \in \mathbf{T}$.

f^i, g^i : fixed bounded measurable function on E^i.

h, k : fixed bounded measurable function on E.

2. Strategies.

Strategies when player A moves <u>first</u> and second does player B (first-type strategies

$\pi = \{ \pi(t) \}_{t \in \mathbf{N}} = \{ (\pi^1(t), \cdots, \pi^d(t)) \}_{t \in \mathbf{N}}$ and

$\sigma = \{ \sigma(t) \}_{t \in \mathbf{N}} = \{ (\sigma^1(t), \cdots, \sigma^d(t)) \}_{t \in \mathbf{N}}$

re **T**- valued stochastic sequences on (Ω, $\mathcal{F}$) satisfying the following (i), (ii) and (iii)

i) $\pi(0) = \sigma(0) = \mathbf{0}$.

ii) For all even $t \in \mathbf{N}$ it holds that

$\pi(t+1) = \sigma(t) + e_i$ for some $i = 1,\cdots,d$ and

$\sigma(t+1) = \sigma(t)$.

For all odd $t \in \mathbf{N}$ it holds that

$\sigma(t+1) = \pi(t) + e_i$ for some $i = 1,\cdots,d$ and

$\pi(t+1) = \pi(t)$.

iii) For all $t \in \mathbf{N}$ and all $s \in \mathbf{T}$ it holds that $\{ \pi(t) = s \} \in \mathcal{F}_s$ and $\{ \sigma(t) = s \} \in \mathcal{F}_s$.

S(F) = { all first-type strategies (π, σ) starting from $\mathbf{0}$ },

MS(F) = { all Markov strategies (π, σ) ($\in$ S(F)) },

MS(F ; 1) = { all first-type one-step Markov strategies π }.

Strategies when player B moves first and <u>second</u> does player A (second-type).

$\pi = \{ \pi(t) \}_{t \in \mathbf{N}}$ and $\sigma = \{ \sigma(t) \}_{t \in \mathbf{N}}$

e **T**- valued stochastic sequences on (Ω, $\mathcal{F}$) satisfying the following (i), (ii) and (iii)

(i) $\pi(0) = \sigma(0) = \mathbf{0}$.

(ii) For all even $t \in \mathbf{N}$ it holds that

$\sigma(t+1) = \pi(t) + e_i$ for some $i = 1,\cdots,d$ and

$\pi(t+1) = \pi(t)$.

For all odd $t \in \mathbf{N}$ it holds that

$\pi(t+1) = \sigma(t) + e_i$ for some $i = 1,\cdots,d$ and

$\sigma(t+1) = \sigma(t)$.

(iii) For all $t \in \mathbf{N}$ and all $s \in \mathbf{T}$ it holds that $\{ \pi(t) = s \} \in \mathcal{F}_s$ and $\{ \sigma(t) = s \} \in \mathcal{F}_s$.

S(S) = { all second-type strategies (π, σ) starting from $\mathbf{0}$ },

MS(S) = { all Markov strategies (π, σ) ($\in$ S(S)) } and

MS(S ; 1) = { all second-type one-step Markov strategies σ }.

$D(F;\sigma) = \{ \pi : (\pi, \sigma) \in S(F) \}$, $D(F;\pi) = \{ \sigma : (\pi, \sigma) \in S(F) \}$.

$D(S;\pi) = \{ \sigma : (\pi, \sigma) \in S(S) \}$ and $D(S;\sigma) = \{ \pi : (\pi, \sigma) \in S(S) \}$.

3. Expected rewards in bandit games.

For (π, σ) ($\in$ S(F)), player A's expected values from an initial state x ($\in$ E) to a terminal time T are defined by

$$R_{A,F,T}^{\pi,\sigma}(x) = \mathbf{E}^x\Big[\sum_{i=1}^{d} f^i(X^i_{\pi^i(1)}) (\pi^i(1) - \sigma^i(0))$$

$$+ \beta^2 \sum_{i=1}^{d} f^i(X^i_{\pi^i(3)}) (\pi^i(3) - \sigma^i(2)) + \ldots\ldots\ldots\ldots$$

$$+ \beta^{T-2} \sum_{i=1}^{d} f^i(X^i_{\pi^i(T-1)}) (\pi^i(T-1) - \sigma^i(T-2)) + \beta^T h(X_{\sigma(T)}) \Big] \quad \text{when T is even,}$$

and

$$R_{A,F,T}^{\pi,\sigma}(x) = \mathbf{E}^x\Big[\sum_{i=1}^{d} f^i(X^i_{\pi^i(1)}) (\pi^i(1) - \sigma^i(0))$$

$$+ \beta^2 \sum_{i=1}^{d} f^i(X^i_{\pi^i(3)}) (\pi^i(3) - \sigma^i(2)) + \ldots\ldots\ldots\ldots$$

$$+ \beta^{T-1} \sum_{i=1}^{d} f^i(X^i_{\pi^i(T)}) (\pi^i(T) - \sigma^i(T-1)) + \beta^T h(X_{\pi(T)}) \Big] \quad \text{when T is odd.}$$

For short

(1) $$R_{A,F,T}^{\pi,\sigma}(x) = \mathbf{E}^x\Big[\sum_{t \in \mathbf{N}(e,T)} \beta^t \sum_{i=1}^{d} f^i(X^i_{\pi^i(t+1)}) (\pi^i(t+1) - \sigma^i(t))$$
$$+ \beta^T h(X_{\max\{\pi(T),\sigma(T)\}}) \Big],$$

Then player B's expected values are

(2) $$R_{B,F,T}^{\pi,\sigma}(x) = \mathbf{E}^x\Big[\sum_{t \in \mathbf{N}(o,T)} \beta^t \sum_{i=1}^{d} g^i(X^i_{\sigma^i(t+1)}) (\sigma^i(t+1) - \pi^i(t))$$
$$+ \beta^T k(X_{\max\{\pi(T),\sigma(T)\}}) \Big],$$

Hence we put

(3) $R_{A,F,T}^{*,\sigma}(x) = \sup_{\pi \in D(F ; \sigma)} R_{A,F,T}^{\pi,\sigma}(x)$ for $x \in E$, and

(4) $R_{B,F,T}^{\pi,*}(x) = \sup_{\sigma \in D(F ; \pi)} R_{B,F,T}^{\pi,\sigma}(x)$ for $x \in E$.

Then we shall call the following game when player A moves <u>first</u> first-type bandit games abbreviated as FBG) : For each T, to find strategies $(\pi^*, \sigma^*) \in S(F)$ such that

$$R_{A,F,T}^{\pi^*,\sigma^*} = R_{A,F,T}^{*,\sigma^*} \text{ and } R_{B,F,T}^{\pi^*,\sigma^*} = R_{B,F,T}^{\pi^*,*}.$$

For a second-type strategy $(\pi, \sigma) (\in S(S))$, player A's expected values are defined b

(5) $$R_{A,S,T}^{\pi,\sigma}(x) = \mathbf{E}^x\Big[\sum_{t \in \mathbf{N}(o.T)} \beta^t \sum_{i=1}^{d} f^i(X^i_{\pi^i(t+1)}) (\pi^i(t+1) - \sigma^i(t))$$

$$+ \beta^T h(X_{\max\{\pi(T),\sigma(T)\}}) \Big] .$$

Then player B's expected values are

(6) $$R_{B,S,T}^{\pi,\sigma}(x) = \mathbf{E}^x \Big[\sum_{t \in \mathbf{N}(e,T)} \beta^t \sum_{i=1}^{d} g^i(X_{\sigma^i(t+1)}^{\ i}) (\sigma^i(t+1) - \pi^i(t))$$
$$+ \beta^T k(X_{\max\{\pi(T),\sigma(T)\}}) \Big] .$$

Hence we put

(7) $R_{A,S,T}^{*,\sigma}(x) = \sup_{\pi \in D(S ; \sigma)} R_{A,S,T}^{\pi,\sigma}(x)$ for $x \in E$, and

(8) $R_{B,S,T}^{\pi,*}(x) = \sup_{\sigma \in D(S ; \pi)} R_{B,S,T}^{\pi,\sigma}(x)$ for $x \in E$.

Then we shall call the following game when player A moves second second-type ba games (abbreviated as SBG) : For each T, to find strategies $(\pi^*, \sigma^*) \in S(S)$ such th

$$R_{A,S,T}^{\pi^*,\sigma^*} = R_{A,S,T}^{*,\sigma^*} \text{ and } R_{B,S,T}^{\pi^*,\sigma^*} = R_{B,S,T}^{\pi^*,*}.$$

When the terminal time T = 0, for convenience we define

(9) $R_{A,F,0}^{\pi,\sigma} = h$ and $R_{B,F,0}^{\pi,\sigma} = k$ for all $(\pi, \sigma) \in S(F)$, and

(10) $R_{A,S,0}^{\pi,\sigma} = h$ and $R_{B,S,0}^{\pi,\sigma} = k$ for all $(\pi, \sigma) \in S(S)$.

Lemma 1.

For strategies $(\pi, \sigma) \in S(F)$ and $(\pi', \sigma') \in S(S)$, there exist Markov strategies π_M ; $0, \sigma)$, $\sigma_M \in D(F ; 0, \pi)$, $\pi'_M \in D(S ; 0, \sigma')$ and $\sigma'_M \in D(S ; 0, \pi')$ such that

$$R_{A,F,T}^{\pi_M,\sigma} = R_{A,F,T}^{*,\sigma}, \qquad R_{B,F,T}^{\pi,\sigma_M} = R_{B,F,T}^{\pi,*},$$

$$R_{A,S,T}^{\pi'_M,\sigma'} = R_{A,S,T}^{*,\sigma'} \text{ and } R_{B,S,T}^{\pi',\sigma'_M} = R_{B,S,T}^{\pi',*}.$$

4. A value iteration and optimal strategies.

ITERATION 1.

Subroutine (A) :

(A. 0) Put $U_{A,F,0} = U_{A,S,0} = h$.

(A. F. 1) Put $U_{A,F,1}(x) = \max_{1 \le i \le d} \mathbf{E}^{x}\Big[f^{i}(X_1^{\,i}) + \beta\, U_{A,S,0}(x^1, \cdots, X_1^{\,i}, \cdots, x^d) \Big]$

· $x = (x^1, \cdots, x^d) \in E$. We define a Markov strategy $\pi_1^* \in MS(F\,;\,1)$.

(A. S. 1) Put $U_{A,S,1}(x) = \mathbf{E}^{x}\Big[\beta\, U_{A,F,0}(X_{\sigma_1^*(1)}) \Big]$ for $x \in E$

th $\sigma_1^* \in MS(S\,;\,1)$ given by (B. S. 1).

(A. F. 2) Put $U_{A,F,2}(x) = \max_{1 \le i \le d} \mathbf{E}^{x}\Big[f^{i}(X_1^{\,i}) + \beta\, U_{A,S,1}(x^1, \cdots, X_1^{\,i}, \cdots, x^d) \Big]$

: $x = (x^1, \cdots, x^d) \in E$. We define a Markov strategy $\pi_2^* \in MS(F\,;\,1)$.

(A. S. 2) Put $U_{A,S,2}(x) = \mathbf{E}^{x}\Big[\beta\, U_{A,F,1}(X_{\sigma_2^*(1)}) \Big]$ for $x \in E$

th $\sigma_2^* \in MS(S\,;\,1)$ given by (B. S. 2).

. . . . .

(A. F. r+1) Put $U_{A,F,r+1}(x) = \max_{1 \le i \le d} \mathbf{E}^{x}\Big[f^{i}(X_1^{\,i}) + \beta\, U_{A,S,r}(x^1, \cdots, X_1^{\,i}, \cdots, x^d) \Big]$

r $x = (x^1, \cdots, x^d) \in E$. We define a Markov strategy $\pi_{r+1}^* \in MS(F\,;\,1)$.

(A. S. r+1) Put $U_{A,S,r+1}(x) = \mathbf{E}^{x}\Big[\beta\, U_{A,F,r}(X_{\sigma_{r+1}^*(1)}) \Big]$ for $x \in E$

th $\sigma_{r+1}^* \in MS(S\,;\,1)$ given by (B. S. r+1).

. . . . .

(A. F. T) Put $U_{A,F,T}(x) = \max_{1 \le i \le d} \mathbf{E}^{x}\Big[f^{i}(X_1^{\,i}) + \beta\, U_{A,S,T-1}(x^1, \cdots, X_1^{\,i}, \cdots, x^d) \Big]$

r $x = (x^1, \cdots, x^d) \in E$. We define a Markov strategy $\pi_T^* \in MS(F\,;\,1)$.

(**A. S. T**) Put $U_{A,S,T}(x) = \mathbf{E}^x\left[\beta U_{A,F,T-1}(X_{\sigma_T*(1)}) \right]$ for $x \in E$ with $\sigma_T* \in MS(S; 1)$ given by (B. S. T).

Subroutine (B) :

(**B. 0**) Put $U_{B,F,0} = U_{B,S,0} = k$.

(**B. F. 1**) Put $U_{B,F,1}(x) = \mathbf{E}^x\left[\beta U_{B,S,0}(X_{\pi_1*(1)}) \right]$ for $x \in E$

with $\pi_1* \in MS(F; 1)$ given by (A. F. 1).

(**B. S. 1**) Put $U_{B,S,1}(x) = \max_{1 \le i \le d} \mathbf{E}^x\left[g^i(X_1^{\,i}) + \beta U_{B,F,0}(x^1, \cdots, X_1^{\,i}, \cdots, x^d) \right]$

for $x = (x^1, \cdots, x^d) \in E$. We define a Markov strategy $\sigma_1* \in MS(S; 1)$.

(**B. F. 2**) Put $U_{B,F,2}(x) = \mathbf{E}^x\left[\beta U_{B,S,1}(X_{\pi_2*(1)}) \right]$ for $x \in E$

with $\pi_2* \in MS(F; 1)$ given by (A. F. 2).

(**B. S. 2**) Put $U_{B,S,2}(x) = \max_{1 \le i \le d} \mathbf{E}^x\left[g^i(X_1^{\,i}) + \beta U_{B,F,1}(x^1, \cdots, X_1^{\,i}, \cdots, x^d) \right]$

for $x = (x^1, \cdots, x^d) \in E$. We define a Markov strategy $\sigma_2* \in MS(S; 1)$.

. . . . .

(**B. F. r+1**) Put $U_{B,F,r+1}(x) = \mathbf{E}^x\left[\beta U_{B,S,r}(X_{\pi_{r+1}*(1)}) \right]$ for $x \in E$

with $\pi_{r+1}* \in MS(F; 1)$ given by (A. F. r+1).

(**B. S. r+1**) Put $U_{B,S,r+1}(x) = \max_{1 \le i \le d} \mathbf{E}^x\left[g^i(X_1^{\,i}) + \beta U_{B,F,r}(x^1, \cdots, X_1^{\,i}, \cdots, x^d) \right]$

for $x = (x^1, \cdots, x^d) \in E$. We define a Markov strategy $\sigma_{r+1}* \in MS(S; 1)$.

. . . . .

(**B. F. T**) Put $U_{B,F,T}(x) = \mathbf{E}^x\left[\beta U_{B,S,T-1}(X_{\pi_T*(1)}) \right]$ for $x \in E$

with $\pi_T* \in MS(F; 1)$ given by (A. F. T).

(B. S. T) Put $U_{B,S,T}(x) = \max_{1 \le i \le d} E^x \left[g^i(X_1^i) + \beta\, U_{B,F,T-1}(x^1, \cdots, X_1^i, \cdots, x^d) \right]$

: $x = (x^1, \cdots, x^d) \in E$. We define a Markov strategy $\sigma_T^* \in MS(S ; 1)$.

We define strategies $(\pi^*, \sigma^*) \in MS(F ; T)$ and $(\pi'^*, \sigma'^*) \in MS(S ; T)$ by

1) $(\pi^*, \sigma^*) = [\pi_T^*, \sigma_{T-1}^*, \pi_{T-2}^*, \sigma_{T-3}^*, \cdots, \pi_4^*, \sigma_3^*, \pi_2^*, \sigma_1^*]$ for even T,

2) $(\pi^*, \sigma^*) = [\pi_T^*, \sigma_{T-1}^*, \pi_{T-2}^*, \sigma_{T-3}^*, \cdots, \sigma_4^*, \pi_3^*, \sigma_2^*, \pi_1^*]$ for odd T,

3) $(\pi'^*, \sigma'^*) = [\sigma_T^*, \pi_{T-1}^*, \sigma_{T-2}^*, \pi_{T-3}^*, \cdots, \sigma_4^*, \pi_3^*, \sigma_2^*, \pi_1^*]$ for even T, and

4) $(\pi'^*, \sigma'^*) = [\sigma_T^*, \pi_{T-1}^*, \sigma_{T-2}^*, \pi_{T-3}^*, \cdots, \pi_4^*, \sigma_3^*, \pi_2^*, \sigma_1^*]$ for odd T.

THEOREM 1.

(π^*, σ^*) is an optimal strategy for FBG, and (π'^*, σ'^*) is an optimal strategy for SBG. oreover $U_{A,F,T}$ and $U_{B,F,T}$ are each player's optimal values for FBG and $U_{A,S,T}$ and $U_{B,S,}$ e each player's optimal values for SBG :

(i) $U_{A,F,T} = R_{A,F,T}^{\pi^*,\sigma^*} \ge R_{A,F,T}^{\pi,\sigma^*}$ for every $\pi \in D(F ; \sigma^*)$.

(ii) $U_{B,F,T} = R_{B,F,T}^{\pi^*,\sigma^*} \ge R_{B,F,T}^{\pi^*,\sigma}$ for every $\sigma \in D(F ; \pi^*)$.

(iii) $U_{A,S,T} = R_{A,S,T}^{\pi'^*,\sigma'^*} \ge R_{A,S,T}^{\pi,\sigma'^*}$ for every $\pi \in D(S ; \sigma'^*)$.

(iv) $U_{B,S,T} = R_{B,S,T}^{\pi'^*,\sigma'^*} \ge R_{B,S,T}^{\pi'^*,\sigma}$ for every $\sigma \in D(S ; \pi'^*)$.

For Markov strategies $\pi \in MS(F;1)$ and $\sigma \in MS(S;1)$ we shall introduce the follo semi-linear operators S^{π} and S^{σ} on the space of all bounded measurable functions on E

(15) $$S^{\pi}\phi(x) = \mathbf{E}^{x}\Big[\sum_{i=1}^{d} f^{i}(X^{i}_{\pi^{i}(1)})\,\pi^{i}(1) + \beta\,\phi(X_{\pi(1)})\Big] \text{ for } x \in E, \text{ and}$$

(16) $$S^{\sigma}\phi(x) = \mathbf{E}^{x}\Big[\sum_{i=1}^{d} g^{i}(X^{i}_{\sigma^{i}(1)})\,\sigma^{i}(1) + \beta\,\phi(X_{\sigma(1)})\Big] \text{ for } x \in E$$

for bounded measurable functions ϕ on E. Then

Corollary 1.

For $r = 0, \cdots, T-1$ (i) ~ (iv) hold :

(i) $$U_{A,F,r+1}(x) = \max_{1 \le i \le d} \mathbf{E}^{x}\Big[f^{i}(X^{i}_{1}) + \beta\,U_{A,S,r}(x^{1}, \cdots, X^{i}_{1}, \cdots, x^{d})\Big]$$

for $x = (x^{1}, \cdots, x^{d}) \in E$.

(ii) $$U_{A,S,r+1}(x) = \mathbf{E}^{x}\Big[\beta\,U_{A,F,r}(X_{\sigma_{r+1}^{*}(1)})\Big] \quad \text{for } x \in E,$$

where $\sigma_{r+1}^{*} \in MS(S;1)$ given by (iv).

(iii) $$U_{B,F,r+1}(x) = \mathbf{E}^{x}\Big[\beta\,U_{B,S,r}(X_{\pi_{r+1}^{*}(1)})\Big] \quad \text{for } x \in E,$$

where $\pi_{r+1}^{*} \in MS(F;1)$ given by (i).

(iv) $$U_{B,S,r+1}(x) = \max_{1 \le i \le d} \mathbf{E}^{x}\Big[g^{i}(X^{i}_{1}) + \beta\,U_{B,F,r}(x^{1}, \cdots, X^{i}_{1}, \cdots, x^{d})\Big]$$

for $x = (x^{1}, \cdots, x^{d}) \in E$.

(v) $U_{A,F,r+1} = S^{\pi_{r+1}^{*}} U_{A,S,r}$, $\quad U_{A,S,r+1} = \beta\,P^{\sigma_{r+1}^{*}}\,U_{A,F,r}$,

$U_{B,F,r+1} = \beta\,P^{\pi_{r+1}^{*}}\,U_{B,S,r}$, $\quad U_{B,S,r+1} = S^{\sigma_{r+1}^{*}}\,U_{B,F,r}$.