数理解析研究所講究録
680巻 1989年 217-221

The Optimal Stopping Problem for Multi-armed Bandit Processes

Yuji Yoshida

吉田 祐治

(Kyushu University)

§ 0. Abstract. This paper deals with the optimal stopping problem for d-armed bandit processes. Under the assumption of independence of arms this paper shows that not only optimal strategies but also optimal stopping times are expressed by the dynamic allocation indices for each arm. Therefore we see that in order to solve this problem it is sufficient to calculate the dynamic allocation indices for each arm. On the other hand this paper reduces this problem to solve d independent one-parameter optimal stopping problems for each arm and shows that the optimal stopping time for the original problem is expressed explicitly as the sum of d smallest optimal stopping times for one-parameter optimal stopping problems.

Moreover this paper gives a necessary and sufficient condition for the finiteness of the smallest optimal stopping time of the original problem. This condition results in the finiteness of the smallest optimal stopping times of d independent one-parameter stopping problems for each arm.

In Markov case this paper shows that the optimal stopping region is equal to Cartesian product of the optimal stopping regions for each arm and also investigates numerical calculation of optimal strategies and optimal stopping times.

§ 1. Preparation.

§ 1.1. Optimal stopping problem.

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

$X = (X_t , \mathcal{F}_t , \mathbb{P})_{t \in \mathbb{N}}$: Markov chain with state space E.

$\mathfrak{M}$: family of all stopping times ($\{ \tau = t \} \in \mathcal{F}_t$ for all $t \in \mathbb{N}$, $\mathbb{P}(\tau < \infty) = 1$).

β : discounted rate ($0 < \beta < 1$).

f : bounded measurable function on E.

c : constant function on E.

$Z = \{ f(X_t) \}_{t \in \mathbb{N}}$: reward process .

$\mathbb{E}^x \left[\sum_{t=0}^{\tau-1} \beta^t f(X_t) + \beta^\tau c \right]$: expected value of reward process starting from initial state x.

Optimal stopping problem ($\mathbb{O}$) :

$$\begin{cases} \text{Find optimal stopping times } \sigma_0 \in \mathfrak{M} : \\ \mathbb{E}^x \left[\sum_{t=0}^{\sigma_0-1} \beta^t f(X_t) + \beta^{\sigma_0} c \right] = \sup_{\tau \in \mathfrak{M}} \mathbb{E}^x \left[\sum_{t=0}^{\tau-1} \beta^t f(X_t) + \beta^\tau c \right] (= U(x)) \text{ for } x \in E. \end{cases}$$

$\sigma_0 = \inf \{ t \in \mathbb{N} : U(X_t) = c \}$: smallest optimal stopping time.

§ 1.2. Dynamic allocation index.

$$\nu(x) = \sup_{\substack{\tau \in \mathfrak{M} \\ \tau \geqq 1}} \frac{\mathbb{E}^x \left[\sum_{r=0}^{\tau-1} \beta^r f(X_r) \right]}{\mathbb{E}^x \left[\sum_{r=0}^{\tau-1} \beta^r \right]} \quad (x \in E) \text{ : dynamic allocation index}$$

$\dot{\tau} = \inf \{ t \geq 1 : \nu(X_t) \leq \nu(X_0) \}$: optimal stopping time.

§ 2. Notations and results.

§ 2.1. Optimal stopping problem for multi-armed bandit processes.

d : number of arms (positive integer).

β : discounted rate ($0 < \beta < 1$).

$X^i = (\mathbb{X}^i_t , \mathcal{F}^i_t , \mathbb{P}^i)_{t \in \mathbb{N}}$: mutually independent Markov chains with state space E^i ($i = 1, \cdots, d$).

f^i : bounded measurable function on E^i ($i = 1, \cdots, d$).

c : constant function on E.

$Z^i = \{ f^i(\mathbb{X}^i_t) \}_{t \in \mathbb{N}}$: machine with arm i (= reward process i) ($i = 1, \cdots, d$) .

$X = (X_s)_{s \in \mathbb{T}} = (\mathbb{X}^1_{s1}, \cdots, \mathbb{X}^d_{sd})_{s = (s^1, \cdots s^d) \in \mathbb{T}}$: d-parameter process with state space E.

$$\begin{cases} \mathbb{T} = \mathbb{N}^d : \text{ time space} \quad E = \prod_{i=1}^{d} E^i : \text{state space} \\ \mathbb{P} = \prod_{i=1}^{d} \mathbb{P}^i : \text{probability} \quad \mathcal{F}_t = \bigotimes_{i=1}^{d} \mathcal{F}^i_{t_i} : \sigma\text{-field } (t = (t_1, \cdots, t_d)) \end{cases}$$

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

$\pi (t) : \Omega \rightarrow \mathbb{T}$ satisfying (i), (ii) and (iii) :

(i) $\pi (0) = (0, \cdots, 0)$.

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

$$\pi (t + 1) = \pi (t) + (0, \cdots, 0, \overset{\overset{i}{\downarrow}}{1}, 0, \cdots, 0) \qquad \text{for some } i = 1, \cdots, d.$$

(iii) $\{ \pi (t) = r \} \in \mathcal{F}_r$ for all $t \in \mathbb{N}$ and all $r \in \mathbb{T}$.

$\pi^i(t)$: number of pulls of arm i up to time t (= time of machine i at time t).

$\Pi = \{$ all strategies π $\}$.

$\mathfrak{M}^{\pi} = \{ \tau : \Omega \to \mathbb{N}$ s.t. $\{ \tau = t \} \cap \{ \pi(t) = r \} \in \mathcal{F}_r$ for $t \in \mathbb{N}$ and $r \in \mathbb{T} \}$

: family of all stopping times along $\pi \in \Pi$.

$$V^{\pi,\tau}(x) = \mathbb{E}^x \left[\sum_{t=0}^{\tau-1} \sum_{i=1}^{d} \beta^t f^i(X^i_{\pi^i(t)}) (\pi^i(t+1) - \pi^i(t)) + \beta^{\tau} c \right] \quad (\pi \in \Pi \text{ and } \tau \in \mathfrak{M}^{\pi})$$

: expected value of reward process.

$V^{*,*}(x) = \sup_{\pi \in \Pi} \sup_{\tau \in \mathfrak{M}^{\pi}} V^{\pi,\tau}(x)$ $(x = (x^1, \cdots, x^d) \in E)$: optimal value.

Optimal stopping problem for d-armed bandit processes ($\mathbb{OB}$):

$$\begin{cases} \text{Find strategies } \pi^* \in \Pi \text{ and stopping times } \tau^* \in \mathfrak{M}^{\pi^*}: \\ V^{\pi^*,\tau^*}(x) = \sup_{\pi \in \Pi} \sup_{\tau \in \mathfrak{M}^{\pi}} V^{\pi,\tau}(x) \ (= V^{*,*}(x)) \text{ for } x \in E \end{cases}$$

Theorem 1 (Bellman equation)

$$V^{*,*}(x) = \max \{ c, \max_{1 \le i \le d} \{ f^i(x^i) + \mathbb{E}^x \left[\beta V^{*,*}(X^i_1) \right] \} \}.$$

§2.2. Optimal strategies and Optimal stopping times.

$\nu^i(x^i)$ ($i = 1, \cdots, d$) : dynamic allocation index with arm i.

$\nu(x) = \max_{1 \le i \le d} \nu^i(x^i)$ $(x = (x^1, \cdots, x^d) \in E)$: maximum index.

Theorem 2

(i) π^* ($\in \Pi$) : index strategy i.e.

for each $t \in \mathbb{N}$ π^* satisfies

$v(X_{\pi^*(t)}) = v^i(X^i_{\pi^{*i}(t)})$ on $\{ \pi(t+1) = \pi(t) + (0, \cdots, 0, \overset{\overset{i}{\vee}}{1}, 0, \cdots, 0) \}$

for some $i = 1, \cdots, d$.

(ii) $\tau^{\pi^*} = \inf \{ t \in \mathbb{N} : v(X_{\pi^*(t)}) \leq (1 - \beta) \cdot c \}$.

(iii) $\mathbb{P}(\tau^{\pi^*} < \infty) = 1$

$\Rightarrow$ π^* is optimal strategy and τ^{π^*} is optimal stopping time for ($\mathcal{OB}$).

$\mathfrak{M}^i$: family of all $\mathcal{F}^i$- adapted stopping times ($\{ \tau \leq t \} \in \mathcal{F}^i_t$ for all $t \in \mathbb{N}$, $\mathbb{P}(\tau < \infty) = 1$).

Optimal stopping problem with arm i ($\mathcal{O}^i$) :

$$\left\{ \begin{array}{l} \text{Find optimal stopping times } \sigma^i_0 \in \mathfrak{M}^i : \\ \mathbb{E}^{x^i}\left[\sum_{t=0}^{\sigma^i_0 - 1} \beta^t f(X^i_t) + \beta^{\sigma_0} c \right] = \sup_{\tau \in \mathfrak{M}^i} \mathbb{E}^{x^i}\left[\sum_{t=0}^{\tau - 1} \beta^t f(X^i_t) + \beta^{\tau} c \right] (= U^i(x^i)) \text{ for } x^i \in E^i. \end{array} \right.$$

$\sigma^i_0 = \inf \{ t \in \mathbb{N} : U^i(X^i_t) = c \}$ ($i = 1, \cdots, d$) : its optimal stopping time.

Theorem 3

$\pi^* \in \Pi$: index strategy .

$\Rightarrow$

(i) $\sigma^i_0 = \inf \{ t \in \mathbb{N} : v^i(X^i_t) \leq (1 - \beta) \cdot c \} = \pi^{*i}(\tau^{\pi^*})$ ($i = 1, \cdots, d$). $\cdots$ machine i

(ii) $\tau^{\pi^*} = \sum_{i=1}^{d} \sigma^i_0 = \inf \{ t \in \mathbb{N} : V^{*,*}(X_{\pi^*(t)}) = c \}$. $\cdots\cdots\cdots\cdots\cdots$ bandit process

(iii) $\mathbb{P}\left[\tau^{\pi^*} < \infty \right] = \prod_{i=1}^{d} \mathbb{P}\left[\sigma^i_0 < \infty \right]$.