Continuous time multi-armed Markov bandits

千葉大学教養   吉田祐治   （ Yuji Yoshida ）

1. Continuous time multi-armed Markov bandit processes.

直積によって、連続時間 multi-armed bandit problem を定義する。

$\mathbb{R}_+ = [0, \infty)$: time space.

d : number of arms (positive integer).

α : discount factor ($\alpha > 0$).

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

$X^i = (X^i_t, \mathcal{F}^i_t, \mathbb{P}^i)_{t \in \mathbb{R}_+}$: mutually independent Brownian motions with state space E^i.

$(\mathcal{F}^i_t)_{t \in \mathbb{R}_+}$: increasing right continuous family of completed sub-σ-fields of $\mathcal{F}^i$.

$\mathbb{P}^{x^i}$: probability measure on $(\Omega^i, \mathcal{F}^i)$ with initial state $x^i \in E^i$.

$\mathfrak{M}^i$: all $(\mathcal{F}^i_t)_{t \in \mathbb{R}_+}$-adapted stopping times.

f^i : fixed bounded continuous function on E^i.

$X = (X_s)_{s \in \mathbb{T}} = (X^1_{s_1}, \cdots, X^d_{s_d})_{s = (s^1, \cdots, s^d) \in \mathbb{T}}$: d-parameter process with state space E.

$$
\begin{cases}
\mathbb{T} = \mathbb{R}^d_+ : \text{time space.} \quad E = \prod_{i=1}^d E^i : \text{state space.} \\[2mm]
\mathbb{P} = \prod_{i=1}^d \mathbb{P}^i : \text{probability.} \quad \Omega = \prod_{i=1}^d \Omega^i : \text{path space.} \\[2mm]
\mathcal{F}_t = \bigotimes_{i=1}^d \mathcal{F}^i_{t^i} : \sigma\text{-field} \quad (t = (t^1, \cdots, t^d))
\end{cases}
$$

つぎに、strategy を定義する。

Strategy $\pi = \{\pi(t)\}_{t \in \mathbb{R}_+} = \{(\pi^1(t), \cdots, \pi^d(t))\}_{t \in \mathbb{R}_+}$ is $\mathbb{T}$- valued stochastic process on $(\Omega, \mathcal{F})$ s.t. (i) ~ (iv):

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

(ii) $\{\pi^i(t)\}_{t \in \mathbb{R}_+}$ is non-decreasing process for each $i = 1, \cdots, d$.

(iii) $\sum_{i=1}^{d} \pi^i(t) = t$ for all $t \in \mathbb{R}_+$.

(iv) $\{\pi(t) \leq r\} \in \mathcal{F}_r$ for all $t \in \mathbb{R}_+$ and all $r \in \mathbb{T}$.

i.e. $\{\pi^1(t) \leq r^1, \cdots, \pi^d(t) \leq r^d\} \in \mathcal{F}^1_{r^1} \otimes \cdots \otimes \mathcal{F}^d_{r^d}$ for all $t \in \mathbb{R}_+$ and all $(r^1, \cdots, r^d) \in \mathbb{T}$

$\Pi = \{$ all strategies $\pi \}$.

このとき、期待利得は次のようになる。

$$V^\pi(x) = \sum_{i=1}^{d} \mathbb{E}^X\left[\int_0^\infty e^{-\alpha t} f^i(X^i_{\pi^i(t)}) d\pi^i(t)\right] \quad (x \in E): \text{expected value of of total rewards}$$

$$V^*(x) = \sup_{\pi \in \Pi} V^\pi(x) \quad (x \in E): \text{optimal value}.$$

したがって、ここで扱う問題は次のように表わせる。

Continuous time d-armed bandit problem (**C.B.P.**):

To find strategies $\pi^* \in \Pi$ s.t. $V^{\pi^*}(x) = V^*(x)$ $(x \in E)$.

2. Dynamic allocation index.

Dynamic allocation index を導入しておく。

$$v^i(x^i) = \sup_{\tau > 0} \frac{\mathbb{E}^{x^i}\left[\int_0^\tau e^{-\alpha t} f^i(X^i_t) dt\right]}{\mathbb{E}^{x^i}\left[\int_0^\tau e^{-\alpha t} dt\right]} \quad (x^i \in E^i): \text{dynamic allocation index}.$$

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

3. Deteriorating bandit problem.

Deteriolating bandit problem とは、次のものをいう。

$$M^i(t) = \inf_{0 \le r \le t} v^i(X^i_r) \quad (t \in \mathbb{R}_+).$$

$(M^*(s))_{s \in \mathbb{T}}$: deteriorating processes $M^*(s) = \max_{1 \le i \le d} M^i(s^i) \quad (s = (s^1, \cdots, s^d) \in \mathbb{T})$.

Deteriorating bandit problem (**D.B.P.**) :

$$\text{To find strategies } \pi \in \Pi \text{ maximizing } \mathbb{E}^X \left[\int_0^\infty e^{-\alpha t} \sum_{i=1}^d M^i(\pi^i(t)) \, d\pi^i(t) \right] \text{ for } x \in E.$$

4. Main results.

次の結果を、得る。

Theorem 1.

(i) $\exists \, \bar{\pi}$: optimal strategy of (**C.B.P.**).

(ii) $(X_{\bar{\pi}(t)}, \mathcal{F}_{\bar{\pi}(t)})_{t \in \mathbb{R}_+}$ is a standard Markov process.

(iii) $\bar{\pi}$: dynamic allocation index strategy

i.e. for every $i = 1, \cdots, d$ and $t \in \mathbb{R}_+$ it holds that

$\bar{\pi}^i(t)$ increases at t only when $v^*(X_{\bar{\pi}(t)}) = v^i(X^i_{\bar{\pi}^i(t)})$.

(iv) $\bar{\pi}$: optimal strategy of (**D.B.P.**) :

$$V^{\bar{\pi}}(x) = \mathbb{E}^X \left[\int_0^\infty e^{-\alpha t} \sum_{i=1}^d M^i(\bar{\pi}^i(t)) \, d\bar{\pi}^i(t) \right] = \mathbb{E}^X \left[\int_0^\infty e^{-\alpha t} M^*(\bar{\pi}(t)) \, dt \right] \quad (x \in E).$$

5. Bellman's equation.

Bellman 方程式は次のようになる。

$\mathfrak{L}^i$: infinitesimal generator for transition probability of process $X^i = (X_t^i, \mathcal{F}_t^i, \mathbb{P}^i)_{t \in \mathbb{R}_+}$.

$D^i = \{x = (x^1, \cdots, x^d) \in E : v^i(x^i) > v^j(x^j)$ for all $j \neq i \}$.

Theorem 2. $V^* : C^2$-class $\Longleftrightarrow$ (i) and (ii):

(i) $\max_{1 \leq i \leq d} \{ \mathfrak{L}^i V^* - \alpha V^* + f^i \} = 0$ in E.

(ii) $\mathfrak{L}^i V^* - \alpha V^* + f^i = 0$ in D^i.