

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 ($i = 1, \dots, d$).

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

$(\mathcal{F}_t^i)_{t \in \mathbb{R}_+}$: 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 $x^i \in E^i$.

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

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

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

$$\left\{ \begin{array}{l} \mathbb{T} = \mathbb{R}_+^d : \text{time space. } E = \prod_{i=1}^d E^i : \text{state space.} \\ \mathbb{P} = \prod_{i=1}^d \mathbb{P}^i : \text{probability. } \Omega = \prod_{i=1}^d \Omega^i : \text{path space.} \\ \mathcal{F}_t = \bigotimes_{i=1}^d \mathcal{F}_{t^i}^i : \sigma\text{-field } (t = (t^1, \dots, t^d)) \end{array} \right.$$

つぎに、strategy を定義する。

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

$$(i) \quad \pi(0) = (0, \dots, 0).$$

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

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

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

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

$\mathcal{H} = \{\text{all strategies } \pi\}$.

このとき、期待利得は次のようになる。

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

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

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.**):

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

2. Dynamic allocation index.

Dynamic allocation index を導入しておく。

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

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

3. Deteriorating bandit problem.

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

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

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

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

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

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 $i = 1, \dots, d$ and $t \in \mathbb{R}_+$ it holds that

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

- (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 $X^i = (X_t^i, \mathfrak{F}_t^i, \mathbb{P}^i)_{t \in \mathbb{R}_+}$.

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

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

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

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