Bandits with Oblivious Experts

Source: Bandit Algorithms, L & S, Chap 18.2.

Setting: Time horizon = n

k-armed bandit, adversarial environment

M experts

an expert recommends arms, i.e.,

For expert m: $(e_{m1}\ e_{m2} \cdots e_{mk})$ is a probability distribution over the k arms, with $0 \le e_{mi} \le 1$, $\sum_{i=1}^{n} e_{mi} = 1$.

At time t:

$$E^{(t)} = \begin{pmatrix} e_{11}^{(t)} & \cdots & \cdots & e_{1k}^{(t)} \\ e_{21}^{(t)} & e_{22}^{(t)} & \cdots & e_{2k}^{(t)} \\ \vdots & & & \\ e_{m1}^{(t)} & e_{m2}^{(t)} & \cdots & e_{mk}^{(t)} \\ \vdots & & & \\ e_{M1}^{(t)} & \cdots & \cdots & e_{Mk}^{(t)} \end{pmatrix}$$

(k arms →, M experts ↓)

Oblivious Expert: Experts recommendations can vary with time, BUT experts cannot learn

from past observations of this bandit player.

e.g.: Experts are pre-trained and static:
$E^{(t)} = E$.

Player's Selection of Action:

Player chooses a distribution over Experts, i.e.,

* $Q_t = (q_1^{(t)} \ldots\ldots q_M^{(t)})$, a prob. dist. over $\{1, 2, \ldots, M\}$ experts.

* Expert m recommends arm j from m^{th} row of $E^{(t)}$ ($E^{(t)}$ known to player).

* Effectively, player plays dist P_t over arms $\{1, 2, \ldots, k\}$, where

$$P_t = Q_t \cdot E^{(t)} = (p_1^{(t)} \ldots\ldots p_k^{(t)}).$$

Player chooses policy $\pi = (\pi_1, \pi_2 \ldots\ldots \pi_n)$

where π_t : History$(t-1) \longmapsto \{1, \ldots, M\}$,

i.e., at time t, chooses a dist. over experts (which in turn, induces a dist. over arms).

Adversarial Setting:

① At begining of time t: $(A_1, X_1, \ldots, A_{t-1}, X_{t-1})$ (action at time 1 → reward at time 1 $\in [0,1]$.)
known to adversary, player.

Player has policy π (known to player and adversary)

② Adversary chooses arm rewards $(x_{t1}, \ldots, x_{tk})$, with $x_{tj} \in [0,1]$ (or equivalently, chooses loss $y_{tj} = 1 - x_{tj}$).

③ Players uses (secret) randomness to sample an expert according to dist. Q_t, and further uses (secret) randomnesr to sample an arm from the chosen expert's dist. Equivalently, player samples arm j from P_t,

i.e $A_t \sim P_t = Q_t E^{(t)}$

④ Player observes reward $X_t = x_{tA_t}$

Regret: $R_n(\pi, x) = \max_{1 \le m \le M} \sum_{t=1}^{n} E_m^{(t)} x_t - E\left[\sum_{t=1}^{n} X_t\right]$,

i.e., with respect to the best fixed expert in hindsight.

Exp 4: Exponential Weighting for Exploration and Exploitation with Experts.

Input: n, k, M, η (see text for a slightly more general version).

Initialize: Choose experts equally likely, i.e.,

$$Q_1 = \left(\frac{1}{M} \quad \frac{1}{M} \quad \cdots \quad \frac{1}{M}\right).$$

for each $t \ge 1$, $t = 1, 2, \ldots, n$:

⑨ Experts reveal $E^{(t)}$ (to player and adversary)

(b) Player chooses arm $A_t \sim Q_t E^{(t)}$

(c) Player sees reward $X_t = x_{tA_t}$

(d) Importance Sampling: $\hat{Y}_{tj} = \left(\frac{\mathbb{1}_{\{A_t = j\}}}{P_{tj}} \right) Y_t$,

where $Y_t = (1 - X_t)$ and $\hat{X}_{tj} = (1 - \hat{Y}_{tj})$

(e) $\tilde{X}_t = E^{(t)} \hat{X}_t$ ($\tilde{X}_t$: M-dim. vector; $\hat{X}_t$: k-dim. vector) (expert-level estimate of rewards).

(f) Update Q_t: $Q_{t+1,m} = \left(\frac{e^{\eta \tilde{X}_{tm}} \cdot Q_{tm}}{\sum_l e^{\eta \tilde{X}_{tl}} \cdot Q_{tl}} \right)$,

$m = 1, 2, \cdots, M$

Theorem (18.1 in textbook): $\eta = \sqrt{\frac{2 \ln(M)}{nk}}$. Then,

$$R_n(\pi, x) \leq \sqrt{2nk \ln(M)}$$

Proof: We use the following corollary (that is from the proof of Exp 3, see Thm 11.2 in textbook). This is a sharpened version that we saw at the end of step (B) in the proof of Exp 3.

Lemma (18.2 in text): Fix any expert $m^* \in \{1,2,\ldots,M\}$. Then,

$$\left(\sum_{t=1}^{n} \tilde{x}_{tm^*} - \sum_{t=1}^{n} \sum_{m=1}^{M} Q_{tm} \tilde{x}_{tm} \right)$$

$$\leq \frac{\ln(M)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \sum_{m=1}^{M} Q_{tm} \left(1 - \tilde{x}_{tm}\right)^2$$

the improvement by a factor of 2 follows from a tighter analog of Step (B) in Exp 3, available in Thm 11.2 in textbook.

At the begining of time t, history is:

$$I_t = \left\{ E^{(1)}, A_1, X_1, E^{(2)}, A_2, X_2, \ldots, E^{(t-1)}, A_{t-1}, X_{t-1}, E^{(t)} \right\}$$

Let m^* be the best arm in hindsight; then,

$$m^* = \underset{1 \le m \le M}{\operatorname{argmax}} \sum_{t=1}^{n} E_m^{(t)} x_t$$

$x_t = \begin{pmatrix} x_{t1} \\ \vdots \\ x_{tK} \end{pmatrix}$ = arm rewards at time t

$E_m^{(t)}$ → m^{th} row of $E^{(t)}$

From Lemma above,

$$\sum_{t=1}^{n} \tilde{x}_{tm^*} - \sum_{t=1}^{n} \sum_{m=1}^{M} Q_{tm} \tilde{x}_{tm} \le \frac{\ln(M)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \sum_{m=1}^{M} Q_{tm} (1 - \tilde{x}_{tm})^2$$

→ (*)

Recall that $\hat{x}_{ti}$ is unbiased, i.e.,

$$\mathbb{E}\left[\hat{x}_{ti} \mid I_t\right] = x_{ti}$$

$$\Rightarrow \mathbb{E}\left[\tilde{x}_t \mid I_t\right] = \mathbb{E}\left[E^{(t)} \hat{x}_t \mid I_t\right]$$

$\tilde{x}_t$: M-dim. column vector. $E^{(t)}$: (M×K) matrix. $\hat{x}_t$: K-dim column vector

$$= E^{(t)} \underbrace{\mathbb{E}\left[\hat{x}_t \mid I_t\right]}_{\text{unbiased}} = E^{(t)} \cdot x_t$$

Taking $\mathbb{E}[\cdot]$ on both sides of (*):

$$R_n(\pi, x) \le \frac{\ln(M)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{n} \sum_{m=1}^{M} \mathbb{E}\left[Q_{tm} \left(1 - \tilde{x}_{tm}\right)^2 \right]$$

Recall: $\hat{y}_{ti} = (1 - \hat{x}_{ti})$, $y_{ti} = (1 - x_{ti})$, $\tilde{y}_{tm} = (1 - \tilde{x}_{tm})$

($\hat{y}_{ti}$: time, arm; y_{ti}: loss, arm; $\tilde{y}_{tm}$: reward, time, expert)

Also, $\tilde{y}_t = E^{(t)} \hat{y}_t$, $\hat{y}_{ti} = \left(\frac{\chi_{\{A_t = i\}} \, y_{ti}}{P_{ti}} \right)$

$$\therefore \mathbb{E}\left[\tilde{y}_{tm}^2 \mid I_t \right] = \mathbb{E}\left[\left(\frac{E^{(t)}_{m A_t} \, y_{t A_t}}{P_{t A_t}} \right)^2 \mid I_t \right]$$

$$= \sum_{j=1}^{k} \left(\cdot \right) \cdot P_{tj} = \sum_{j=1}^{k} \frac{\left(E^{(t)}_{mj} \, y_{tj} \right)^2}{P_{tj}}$$

(P_{tj}: $P(A_t = j)$)

$$\le \sum_{j=1}^{k} \frac{E^{(t)}_{mj}}{P_{tj}} \qquad \left(\because 0 \le E^{(t)}_{mj}, \, y_{tj} \le 1 \right)$$

$$\Rightarrow \mathbb{E}\left[\sum_{m=1}^{M} Q_{tm}\left(1-\tilde{X}_{tm}\right)^2\right] = \mathbb{E}\left[\sum_{m=1}^{M} Q_{tm}\tilde{Y}_{tm}^2\right]$$

$$\leq \mathbb{E}\left[\sum_{m=1}^{M} Q_{tm}\,\mathbb{E}\left[\tilde{Y}_{tm}^2 \,\middle|\, I_t\right]\right]$$

$$\leq \mathbb{E}\left[\sum_{m=1}^{M} Q_{tm} \cdot \sum_{j=1}^{K} \frac{E_{mj}^{(t)}}{P_{tj}}\right]$$

$$= \mathbb{E}\left[\sum_{j=1}^{K} \frac{\overbrace{\sum_{m=1}^{M} Q_{tm}\cdot E_{mj}^{(t)}}^{P_{tj}}}{P_{tj}}\right]$$

$$= K$$

$$\Rightarrow R_n(\pi, x) \leq \frac{\ln(M)}{\eta} + \frac{\eta n K}{2}$$

Choose $\eta = \sqrt{\frac{\ln(M)}{nK}} \Rightarrow R_n(\pi, x) \leq \sqrt{2nK\ln(M)}$

∎