IE613: Online Machine Learning **Jan-Apr 2019**

Assignment 1: Solutions

Question 5. *Show that for any deterministic policy π there exists an environment ν such that $R_T(\pi, \nu) \geq T(1 - 1/K)$ for T rounds and K arms.*

Proof. The cumulative regret for policy π and environment ν is defined as follows:

$$R_T(\pi, \nu) = \max_{i \in [K]} \sum_{t=1}^{T} x_{ti} - \sum_{t=1}^{T} x_{tI_t}$$

where $x_{ti} \in [0, 1]$ is the reward of arm i, I_t is the arm chosen by the policy π and x_{tI_t} is the reward observed in the round t.

As the policy π is deterministic, its action I_t are known a-priori for the round t. Hence there exists an environment ν such that the reward for different arms is as follows:

$$x_{ti} = \begin{cases} 0 & \text{if } i = I_t \\ 1 & \text{otherwise.} \end{cases}$$

For such an environment ν, any deterministic policy π collects zero reward and its cumulative regret is

$$\begin{aligned}
R_T(\pi, \nu) &= \max_{i \in [K]} \sum_{t=1}^{T} x_{ti} - \sum_{t=1}^{T} x_{tI_t} \\
&= \max_{i \in [K]} \sum_{t=1}^{T} x_{ti} && \text{as } \textstyle\sum_{t=1}^{T} x_{tI_t} = 0 \text{ for policy } \pi, \text{ we get} \\
&\geq \mathbb{E}\left[\sum_{t=1}^{T} x_{ti}\right] && \text{because } \max(x) \geq \mathbb{E}[x] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[x_{ti}\right] && \text{by the Linearity of Expectation} \\
&= \sum_{t=1}^{T} \frac{1}{K} \sum_{i=1}^{K} x_{ti} = \sum_{t=1}^{T} \frac{K-1}{K} && \text{from the definition of } x_{ti} \\
\implies R_T(\pi, \nu) &\geq T\left(1 - \frac{1}{K}\right) && \square
\end{aligned}$$

Question 6. *Suppose we had defined the regret by*

$$R_T^{track}(\pi, \nu) = \mathbb{E}\left[\sum_{t=1}^{T} \max_{i \in [K]} x_{ti} - \sum_{t=1}^{T} x_{tI_t}\right]$$

where I_t is the arm chosen by the policy π and x_{tI_t} is the reward observed in the round t. At first sight this definition seems like the right thing because it measures what you actually care about. Unfortunately, however, it gives the adversary too much power. Show that for any policy π (randomized or not) there exists a $\nu \in [0, 1]^{K \times T}$ such that

$$R_T^{track}(\pi, \nu) \geq T\left(1 - \frac{1}{K}\right)$$

Proof. Let $\mathcal{H}_t = \{I_n, X_n\}_{n=1}^t$ be the history of actions and rewards observed till round t. Let policy π uses $\mathcal{H}_t$ and compute a score S_i for each arm $i \in [K]$ and arm is better if its score is high.

Now we define an environment ν such that the reward for different arms is as follows:

$$x_{ti} = \begin{cases} 1 & \text{if } i = \arg\min_{i \in [K]} S_i \\ 0 & \text{otherwise.} \end{cases}$$

where reward 1 is assigned to only the arm that has lowest score and 0 for other arms otherwise. In case of multiple arms having the same score, assign reward 1 to the arbitrarily selected arm.

Now, regret is

$$\begin{aligned} R_T^{track}(\pi, \nu) &= \mathbb{E}\left[\sum_{t=1}^T \max_{i \in [K]} x_{ti} - \sum_{t=1}^T x_{tI_t}\right] \\ &= T - \sum_{t=1}^T \mathbb{E}[x_{tI_t}] && \text{because } \sum_{t=1}^T \max_{i \in [K]} x_{ti} = T \text{ as one arm has reward 1} \\ &= T - \sum_{t=1}^T \mathbb{P}\{I_t = \arg\min_{i \in [K]} S_i\} && \text{by expectation of indicator random variable} \end{aligned}$$

$\mathbb{P}\{I_t = \arg\min_{i \in [K]} S_i\} \leq \frac{1}{K}$ as policy π choose arm I_t that has maximum score

$$\begin{aligned} &\geq T - \sum_{t=1}^T \frac{1}{K} \\ &= T - \frac{T}{K} = T\left(1 - \frac{1}{K}\right) \\ \implies R_T^{track}(\pi, \nu) &\geq T\left(1 - \frac{1}{K}\right) \end{aligned}$$

□

Question 7. *Let $p \in \mathcal{P}_k$ be a probability vector and suppose $\hat{X} : [k] \times \mathbb{R} \to \mathbb{R}$ is a function such that for all $x \in \mathbb{R}^k$, if $A \sim p$,*

$$\mathbb{E}[\hat{X}(A, x_A)] = \sum_{i=1}^k p_i \hat{X}(i, x_i) = x_1.$$

Show there exists an $a \in \mathbb{R}^k$ such that such that $\sum_{i=1}^k a_i p_i = 0$ and $\hat{X}(i, x) = a_i + \frac{\mathbb{I}\{i=1\}x_1}{p_1}$.

Proof. Let $a_i = \hat{X}(i, x_i) - \frac{\mathbb{I}\{i=1\}x_1}{p_1}$ then we have

$$\begin{aligned} \sum_{i=1}^k a_i p_i &= \sum_{i=1}^k p_i \left(\hat{X}(i, x_i) - \frac{\mathbb{I}\{i=1\}x_1}{p_1}\right) && \text{using definition of } a_i \\ &= \sum_{i=1}^k p_i \hat{X}(i, x_i) - \sum_{i=1}^k p_i \frac{\mathbb{I}\{i=1\}x_1}{p_1} \\ &= x_1 - x_1 = 0 && \text{as } \sum_{i=1}^k p_i \hat{X}(i, x_i) = x_1 \text{ and } \mathbb{I}\{i=1\} = 0,\ \forall i \neq 1 \\ \implies \sum_{i=1}^k a_i p_i &= 0 \end{aligned}$$

□