Problem: Adaptation to the range for K–armed bandits

So far we only considered K–armed bandit problems $\nu_1, \ldots, \nu_K$ with distributions over a known interval, typically set to $[0, 1]$ with no loss of generality. Can the player learn the range? I.e., minimize the regret when the distributions $\nu_1, \ldots, \nu_K$ are supported on a bounded range $[m, M]$ but the player ignores m and M? The answer is "Yes" and a strategy to do so can be based on the fully adaptive version of the exponentially weighted average strategy studied in class. We will refer to this strategy as FA-EWA in the sequel.

We use our typical notation: at each round, the player picks an arm I_t, a payoff Y_t is drawn at random according to ν_{I_t} given this choice I_t; expectations are denoted by $\mu_1, \ldots, \mu_K$, with maximal value $\mu^\star$; etc.

First case: an element $C \in [m, M]$ is known

We consider an auxiliary strategy outputting probability distributions $p_t = (p_{1,t}, \ldots, p_{K,t})$ over the arms, at round $t \geqslant 1$. We also consider a non-increasing sequence $\gamma_t \in (0, 1/2]$. We draw the arm I_t at random according to the probability distribution q_t defined by

$$q_{j,t} = (1 - \gamma_t) p_{j,t} + \frac{\gamma_t}{K} .$$

The auxiliary strategy is actually given by FA-EWA run on the losses

$$\ell_{j,t} = \frac{-(Y_t - C)\mathbb{1}_{\{I_t = k\}}}{q_{j,t}} - C .$$

This strategy indeed has no knowledge of m and M (but requires an element $C \in [m, M]$).

Some useful (in)equalities. First prove the following statements.

1. For all $j \in \{1, \ldots, K\}$ and all $t \geqslant 1$,

$$|\ell_{j,t} + C| \leqslant \frac{M - m}{\gamma_t / K} .$$

2. Define a filtration $\mathcal{F}$ such that for all $j \in \{1, \ldots, K\}$ and all $t \geqslant 1$,

$$\mathbb{E}[\ell_{j,t} \,|\, \mathcal{F}_{t-1}] = \mu_j .$$

3. For all $j \in \{1, \ldots, K\}$ and all $t \geqslant 1$, we have $\gamma_t \leqslant 1/2$ thus $p_{j,t} \leqslant 2 q_{j,t}$ and

$$\mathbb{E}\big[p_{j,t}(\ell_{j,t} + C)^2\big] \leqslant 2(M - m)^2 .$$

Recall that FA-EWA guarantees that for all ranges $[a, b]$, for all sequences of losses $L_{j,t} \in [a, b]$, for all $T \geqslant 1$,

$$R_T \leqslant 2\sqrt{\sum_{t=1}^{T} v_t \ln N} + 5(b - a) \ln N ,$$

where R_T is some regret and where the v_t are some variance factors.

4. Recall how R_T and v_t are defined; also pin point the slight simplification performed for the sake of readability in the second-order term $4(b - a) \ln N$ compared to what we proved in class.

Substituting the regret bound of FA-EWA

5. Substitute the regret bound of FA-EWA and some of the useful (in)equalities proved above to get

$$\sum_{t=1}^{T} \sum_{j \in \{1,\ldots,K\}} p_{j,t} \ell_{j,t} - \min_{k \in \{1,\ldots,K\}} \sum_{t=1}^{T} \ell_{k,t} \leqslant 2\sqrt{\sum_{t=1}^{T} \sum_{j \in \{1,\ldots,K\}} p_{j,t}(\ell_{j,t} - C)^2 \ln N} + \frac{8(M - m) \ln N}{\gamma_T / K} .$$

6. Note that

$$\sum_{j\in\{1,\ldots,K\}} q_{j,t}\ell_{j,t} = -Y_t$$

and deduce from the previous question a bound on

$$-\sum_{t=1}^{T} Y_t - \min_{k\in\{1,\ldots,K\}} \ell_{k,t}\,.$$

7. Take expectations in the inequality obtained to prove

$$T\mu^\star - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right] \leqslant 3(M-m)\sqrt{KT\ln K} + 10(M-m)\frac{K\ln K}{\gamma_T} + (M-m)\sum_{t=1}^{T}\gamma_t\,.$$

8. Provide a final regret bound of order $\sqrt{T}$.

Second case: getting rid of the knowledge of C

9. How can the strategy above be adapted so that no knowledge of an element $C \in [m, M]$ is required, without degradating too much the regret bound?