Class Tutorial 14

The Multi-Armed Bandit Problem

An algorithm has K possible actions (a.k.a. , arms) to choose from, and there are T round where both are known in advance. We consider an algorithm that acts according to the following steps.

Multi armed bandit algorithm
Given: K arms, T round In each round $t \in [T]$: 1. Pick an arm a_t 2. Observe a reward $r_t \sim P_{a_t}$ where $r_t \in [0,1]$

Observe that P_{a_t} is an (unknown) distribution from which the reward is sampled, given the arm a_t is chosen at round t.

We use the following notations as well.

- The average reward of arm a is $\mu(a) = E_{r \sim P_a}[r]$.
- The best mean reward is denoted by $\mu^* = \max_a \mu(a)$, and for the optimal arm a^* $\mu(a^*) = \mu^*$.
- The gap of arm a is $\Delta(a) = \mu^* - \mu(a)$.

Objective. The regret (which is an RV (!)) is defined as:

$$R(T) = \sum_{s=1}^{T} E_{r \sim P_{a^*}}[r] - E_{r \sim P_{a_s}}[r] = \sum_{s=1}^{T} \mu^* - \mu(a_s).$$

Our goal is the minimize the expected regret $E[R(T)]$.

Recap: Hoeffding's inequality

1. Let $\{X_i\}_{i=1}^N \in [0,1]$ be i.i.d. RVs, $\mu = E[X_i]$, and define the empirical average as
   - $\bar{\mu}_N = \frac{1}{N}\sum_{i=1}^N X_i$, and let N be a fixed number. Bound $P\left(\bar{\mu}_N - \mu \geq \sqrt{\frac{2\log N}{N}}\right)$ using Hoeffding's inequality
2. Let N be random variable which might depend on $\{X_i\}_{i=1}^N$, and $1 \leq N \leq N_{MAX}$ where N_{MAX} is a known number. Bound Bound $P\left(\bar{\mu}_N - \mu \geq \sqrt{\frac{2\log N_{MAX}}{N}}\right)$ using Hoeffding's inequality.

Solution

1. By Hoeffding's inequality we have that

$$P(\bar{\mu}_N - \mu \geq \epsilon) \leq \exp(-2N\epsilon^2)$$

Setting $\epsilon = \sqrt{\frac{2\log N}{N}}$ we get

$$2\exp(-2N\epsilon^2) = 2\exp(-4\log N) = 2\exp(\log N^{-4}) = \frac{1}{N^4}.$$

Thus, $P\left(\bar{\mu}_N - \mu \geq \sqrt{\frac{2\log N}{N}}\right) \leq \frac{1}{N^4}$

2. As N is a random variable we cannot apply Hoeffding's inequality – for which we need N to be a fixed number. However, since N is bounded, it holds that $N \in \{1,..,N_{MAX}\}$. Thus, the following set of events is equal

$$\{\bar{\mu}_N - \mu \geq \sqrt{\frac{2\log N_{MAX}}{N}}\} = \cup_{i=1}^{N_{MAX}} \{\bar{\mu}_i - \mu \geq \sqrt{\frac{2\log N_{MAX}}{N}}\},$$

(Prove this by showing one set contains the other and vice-versa). Thus, using the union bound,

$$P\left(\left\{\bar{\mu}_N - \mu \geq \sqrt{\frac{2\log N_{MAX}}{N}}\right\}\right)$$

$$= P\left(\cup_{i=1}^{N_{MAX}} \{\bar{\mu}_i - \mu \geq \sqrt{\frac{2\log N_{MAX}}{i}}\right\})$$

$$\leq \sum_{i=1}^{N_{MAX}} P(\bar{\mu}_i - \mu \geq \sqrt{\frac{2\log N_{MAX}}{i}})$$

where i is fixed and not a random variable. As the number of terms in the empirical average is fixed, we can apply Hoeffding's inequality and get

$$P\left(\bar{\mu}_i - \mu \geq \sqrt{\frac{2\log N_{MAX}}{i}}\right) \leq \exp\left(-2i\left(\frac{2\log N_{MAX}}{i}\right)\right) = \frac{1}{N_{MAX}^4}.$$

Thus, $P\left(\left\{\bar{\mu}_N - \mu \geq \sqrt{\frac{2\log N_{MAX}}{N}}\right\}\right) \leq \frac{1}{N_{MAX}^3}.$

Optimism in the face of uncertainty: Upper Confidence Interval (UCB) Algorithm.

Multi armed bandit algorithm
Given: K arms, T round In each round $t \in [T]$: 1. Pick an arm $a_t \in \arg\max_a UCB_t(a)$, where $UCB_t(a) = \bar{\mu}_t(a) + \sqrt{\frac{2\log T}{n_t(a)}}$. 2. Observe a reward $r_t \sim P_{a_t}$ where $r_t \in [0,1]$

For brevity, we denote $\bar{\mu}_t(a) \equiv \bar{\mu}_{n_t(a)}(a)$.

Intution: the bonus quantifies how uncertain we are about the current estimate of arm a and if the bonus is big arm a might be a good arm to play with. After sufficient amount of time we will act by arms with high-reward as the bonus term decreases and $UCB_t(a) \sim \mu(a)$.

1. Define the clean event and the bad event.
2. Bound the probability of the bad event.
3. Bound $\Delta_t \equiv \mu^* - \mu(a_t)$ by $\delta_t(a_t) = \sqrt{\frac{2\log T}{n_t(a_t)}}$ on the clean event.
4. Bound the expected regret.

et a_t be the arm chosen at round t. Then, with high-probability (or when the success event holds) $UCB_t(a_t) \geq \mu^*$ and $UCB_t(a^*) \geq \mu^*$.

Solution

1. In words, the clean event, $\mathcal{A}$, is the event the real averages of all arms $\mu(a)$ are inside the interval $[\bar{\mu}_t(a) - \sqrt{\frac{2\log T}{n_t(a)}}, \bar{\mu}_t(a) + \sqrt{\frac{2\log T}{n_t(a)}}]$ for all $t \in [T]$, where
$$LCB_t(a) = \bar{\mu}_t(a) - \sqrt{\frac{2\log T}{n_t(a)}}$$
$$UCB_t(a) = \bar{\mu}_t(a) + \sqrt{\frac{2\log T}{n_t(a)}}$$
The bad event is the complement of the good event, $\mathcal{A}^C$. As always, $P(\mathcal{A}) + P(\mathcal{A}^C) = 1$. Formally,
$$\mathcal{A} = \{\forall t, a: \mu(a) \in \ [LCB_t(a), UCB_t(a)]\}.$$
2. We need to bound the probability the empirical mean of i.i.d. RVs deviates from its mean, when the number of variables is a random variable as well -- $n_t(a)$ is a random variable. Thus, we apply the result from the second section of previous question (see that we use the symmetric version of Hoeffding's inequality and not the one-sided as there) with another union bound on all arms (while assuming $K \leq T$) and get that for any $t \in [T]$,
$$P(\mathcal{A}^C) \leq \frac{2}{T^2}.$$

3. Start by observing the following holds when the clean event holds.

$$\mu(a) \leq \bar{\mu}_t(a) + \sqrt{\frac{2\log T}{n_t(a)}}$$

$$\mu(a) \geq \bar{\mu}_t(a) - \sqrt{\frac{2\log T}{n_t(a)}}$$

By the optimism of the algorithm,

$$\bar{\mu}_t(a_t) + \sqrt{\frac{2\log T}{n_t(a_t)}} = UCB(a_t) \geq UCB(a^*) \geq \mu^*,$$

and, on the other hand,

$$\bar{\mu}_t(a_t) \leq \mu(a_t) + \sqrt{\frac{2\log T}{n_t(a_t)}}.$$

Combining the two, we get

$$\Delta_t \equiv \mu^* - \mu(a_t) \leq 2\sqrt{\frac{2\log T}{n_t(a_t)}}$$

4. By the tower property, for any $t \in [T]$,

$$\begin{aligned} E[R(t)] &= P(\mathcal{A})E[\,R(t) \mid \mathcal{A}\,] + P(\mathcal{A}^C)E[\,R(t) \mid \mathcal{A}^C\,] \\ &\leq (1-\frac{2}{T^2})E[\,R(t) \mid \mathcal{A}\,] + \frac{2}{T^2}E[\,R(t) \mid \mathcal{A}^C\,] \\ &\leq E[\,R(t) \mid \mathcal{A}\,] + \frac{2}{T^2}E[\,R(t) \mid \mathcal{A}^C\,] \end{aligned}$$

Let $R(t;a)$ be the regret experienced due to acting with arm a at time t. We have that

$$R(t;a) = n_t(a_t)\big(\mu^* - \mu(a_t)\big),$$

And, by definition, $R(t) = \sum_a R(t;a)$.

Given the **clean event** holds, by section 3,

$$\begin{aligned} R(t) &\leq \sum_{a\in[K]} 2n_t(a_t)\sqrt{\frac{2\log T}{n_t(a_t)}} \\ &= O(\sqrt{\log T})\sum_{a\in[K]}\sqrt{n_t(a_t)} \\ &= O\big(\sqrt{\log T}\big)K\frac{1}{K}\sum_{a\in[K]}\sqrt{n_t(a_t)} \\ &= O\big(\sqrt{\log T}\big)K\sqrt{\sum_{a\in[K]}\frac{n_t(a_t)}{K}} \\ &= O\big(\sqrt{\log T}\big)\sqrt{K\sum_{a\in[K]} n_t(a_t)} = O\left(\sqrt{\sqrt{\log T}\,KT}\right) \end{aligned}$$

The fourth relation is by Jensen's inequality in the fourth relation, and the fifth relation is by $\sum_{a\in[K]} n_t(a_t) = t \leq T$.

Given the **bad event** we use the naïve bound $R(t) \leq T$.
Combining all the above we get

$$E[R(t)] \leq O\left(\sqrt{\log T\, KT}\right) + \frac{2}{T} = O\left(\sqrt{\log T\, KT}\right).$$