Note on Multi-Armed Bandits

Wen Sun[1]

[1]Department of Computer Science, Cornell University

April 26, 2021

1 Preliminaries

1.1 Problem Definition

Multi-armed bandit (MAB) problem is formally defined as follows:

We consider a MAB with $K \in \mathbb{N}^+$ many arms (i.e., actions), where each arm $i \in \{1, 2, \ldots, K\}$ has its own reward distribution ν_i. Denote μ_i as the mean of ν_i, and define

$$\mu^\star = \max_{i \in [K]} \mu_i, \quad i^\star = \arg\max_{i \in [K]} \mu_i.$$

Note that the reward distributions and the means of reward distributions are all unknown. Instead, at any time step, after pulling an arm i, we only receive a reward r sampled from ν_i. One can imagine that if we pull an arm i enough times, then the average reward can serve as a good estimation of the expectation, i.e., μ_i.

At each time step $t \in \{1, \ldots, T\}$, the learner chooses some arm $I_t \in \{1, \ldots, K\}$. We are interested in the learner's regret, defined below:

$$\text{Regret} = T\mu^\star - \sum_{t=1}^{T} \mu_{I_t},$$

where μ_{I_t} is the expected reward of the chosen arm indexed by I_t. The goal is to design the learner such that it achieves sub-linear regret, e.g., $\sqrt{T}$.

1.2 Explore-Exploit Dilemma

Note that only information the learner knows beforehand is just the number of total arms, i.e., K here. Every round, the learner needs to make a decision in terms of just pulling the best arm so far (i.e., exploitation), or trying some other arms that have not been tried not enough times yet (i.e., exploration).

1.3 Statistical tools: Concentration Inequalities

The only concentration inequality we are going to use in this note is the Hoeffding's inequality. Hoeffding's inequality gives us a sense of how the empirical mean can deviate from the true mean in terms of the number of samples.

Algorithm 1 Explore and Exploit (N)

for $k = 1$ to K **do**
  for $i = 1$ to N **do**
    Pull arm k.
    Receive reward $r_i \sim \nu_k$.
  end for
  Compute arm k's average reward $\hat{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} r_i$
end for
for $t > KN$ **do**
  Pull the best empirical arm $\hat{i} = \arg\max_{i\in\{1,\dots,K\}} \hat{\mu}_i$.
end for

Theorem 1 (Hoeffding's Inequality)**.** *Consider a one-dimension distribution ν with expectation μ, where any sample from μ is bounded, i.e., $r \sim \mu$ must have $|r| \le a \in \mathbb{R}^+$. Given N many i.i.d scalars $\{r_i\}_{i=1}^N \overset{iid}{\sim} \nu$ sampled from ν, we have that:*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} r_i/N - \mu\right| \ge \epsilon\right) \le 2\exp\left(-\frac{2N\epsilon^2}{a^2}\right).$$

In other words, with probability at least $1-\delta$, we have:

$$\left|\sum_{i=1}^{N} r_i/N - \mu\right| \le \sqrt{\frac{a^2\log(2/\delta)}{2N}} = O\left(\sqrt{\log(1/\delta)/N}\right).$$

Namely, from Hoeffding's inequality, we know that with high probability, our empirical mean estimation $\sum_{i=1}^{N} r_i/N$ is approaching to the true mean μ in the rate of $1/\sqrt{N}$.

Remark Proving the Hoeffding's inequality is out of the scope of this class. Here all we need to know is that Hoeffding's inequality is an off-shelf statistical tool that builds confidence interval for our mean estimate.

Union bound Another statistical tool that we will leverage is the union bound, i.e., given N events $A_1, A_2, \dots, A_N$, we have that $\mathbb{P}(A_1 \text{ or } A_2 \dots \text{ or } A_N) \le \sum_{i=1}^{N}\mathbb{P}(A_i)$. The inequality can be extended to any number of events. Again we are not going to prove that. The intuition behind this is that think about A_i as a set in $\mathbb{R}^2$. $\mathbb{P}(A_1 \text{ or } A_2 \dots \text{ or } A_N)$ represents the area covered by the set $A_1 \cup A_2 \cdots \cup A_K$. Since there might be overlapping between these sets, we have that $\text{area}(A_1 \cup A_2 \cdots \cup A_K) \le \sum_{i=1}^{N} \text{area}(A_i)$.

2 Algorithm 1: Explore and Exploit

In this section, we study perhaps the simplest MAB algorithm called Explore and Exploit. In high level, as the name of the algorithm suggested, we will first perform enough rounds of exploration, till we have a good estimation of the expectation μ_i for all $i \in \{1, \dots, K\}$. Then we will exploit: afterwards, we simply are going to keep pulling the arm that has the highest estimated expectation.

Alg. 1 summarizes the algorithm. Alg. 1 takes an integer N as input, for the first KN many rounds, it simply tries every arm for N many rounds, and uses the sampled rewards to estimate the expectation μ_i for all $i \in [K]$. After NK rounds, the algorithm picks the best empirical arm so far, i.e., $\hat{i} = \arg\max_{i\in[K]} \hat{\mu}_i$. Afterwards, the algorithm just performs exploitation, i.e., playing $\hat{i}$ forever. Intuitively, if N is large, i.e., if we play each arm enough

many times, we will have $\hat{\mu}_i \approx \mu_i$ for all i. Then there may be a high chance that $\hat{i} = i^\star$. Below, we are going to tune the parameter N, so that the Explore and Exploit algorithm guarantees a sub-linear regret.

Theorem 2 (Regret of Exploit-Explore)**.** *Fix any $\delta \in (0, 1/K)$. Assume reward sampled from ν_i is bounded in $[0, 1]$ for any $i \in [K]$. After T many rounds where we assume T is bigger than KN, we have that with probability at least $1 - K\delta$,*

$$\texttt{Regret} = \tilde{O}(K^{1/3}T^{2/3}).$$

.

Note here $\tilde{O}$ ignores all constants that does not depend on K or T, and all the log term $\log(1/\delta)$.

Proof. For each arm k, from Hoeffding's inequality, we know that with probability at least $1 - \delta$:

$$|\hat{\mu}_k - \mu_k| \leq \sqrt{\frac{\log(2/\delta)}{2N}}.$$

In other words, with probability at most δ, the confidence interval above will fail to capture μ_k.

Now we ask ourselves, what's the probability that $|\hat{\mu}_i - \mu_i| \leq \sqrt{\frac{\log(2/\delta)}{2N}}$ holds for all $k \in [K]$. We can ask the negation question, i.e., what is the probability where there exists a $k \in [K]$ such that $|\hat{\mu}_i - \mu_i| \geq \sqrt{\frac{\log(2/\delta)}{2N}}$ (i.e., the confidence interval for k is not valid)? We can apply union bound here:

$$\mathbb{P}\left(\exists i \in [K], |\hat{\mu}_i - \mu_i| \geq \sqrt{\frac{\log(2/\delta)}{2N}}\right) \leq \sum_{i=1}^{K} \mathbb{P}\left(|\hat{\mu}_i - \mu_i| \geq \sqrt{\frac{\log(2/\delta)}{2N}}\right) \leq K\delta,$$

since Hoeffding's inequality tells us that the probability of the confidence interval of arm i failing is at most δ.

This immediately means that:

$$\mathbb{P}\left(\forall i \in [K], |\hat{\mu}_i - \mu_i| \leq \sqrt{\frac{\log(2/\delta)}{2N}}\right) = 1 - \mathbb{P}\left(\exists i \in [K], |\hat{\mu}_i - \mu_i| \geq \sqrt{\frac{\log(2/\delta)}{2N}}\right) \geq 1 - K\delta.$$

In other words, with probability at least $1 - K\delta$, for ALL $i \in [K]$, we must have:

$$|\hat{\mu}_i - \mu_i| \leq \sqrt{\frac{\log(2/\delta)}{2N}}.$$

Below we bound regret by conditioning on the event that all confidence intervals are valid.

Note that since we are performing pure exploration in the first KN many rounds, the largest possible total regret we could have in the first KN rounds is KN (recall the reward is always bounded in $[0, 1]$), i.e.,

$$\texttt{Regret}_{explore} \leq KN.$$

Now let us consider total regret between round [KN+1, T], i.e., the regret from the exploitation phase.

Recall $\hat{i} = \arg\max_{i \in [K]} \hat{\mu}_i$, and $i^\star = \arg\max_{i \in [K]} \mu_i$, i.e., $\hat{i}$ is the arm with the highest estimated expected reward—the empirical estimator, while $i^\star$ is the best arm. Let us compute the gap between $\mu_{\hat{i}}$ and $\mu_{i^\star}$, i.e., the gap between the expected reward of our estimator $\hat{i}$, and the highest expected reward. First, since $\hat{i}$ has the highest estimated mean, i.e., $\hat{i} = \arg\max_{i \in [K]} \hat{\mu}_i$, we must have

$$\hat{\mu}_{\hat{i}} \geq \hat{\mu}_{i^\star}. \tag{1}$$

On the other hand, we know that with probability at least $1 - K\delta$, we have:

$$\mu_{\hat{i}} \leq \mu_{\hat{i}} + \sqrt{\frac{\log(2/\delta)}{2N}}, \tag{2}$$

$$\hat{\mu}_{i^\star} \geq \mu_{i^\star} - \sqrt{\frac{\log(2/\delta)}{2N}}. \tag{3}$$

Now combine Inequalities 1&2&3, we have:

$$\mu_k + \sqrt{\frac{\log(2/\delta)}{2N}} \geq \mu_{i^\star} - \sqrt{\frac{\log(2/\delta)}{2N}},$$

which leads us to:

$$\mu_{i^\star} - \mu_{\hat{i}} \leq 2\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Now we can bound the total regret from the exploitation phase:

$$\texttt{Regret}_{exploit} = \sum_{i=NK}^{T} (\mu_{i^\star} - \mu_{\hat{i}}) \leq 2(T - NK)\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Hence, the total regret is:

$$\begin{aligned}\texttt{Regret} &= \texttt{Regret}_{explore} + \texttt{Regret}_{exploit} = KN + 2T\sqrt{\frac{\log(2/\delta)}{2N}} - 2NK\sqrt{\frac{\log(2/\delta)}{2N}} \\ &\leq KN + 2T\sqrt{\frac{\log(2/\delta)}{2N}}.\end{aligned}$$

Now let us minimize $KN + 2T\sqrt{\frac{\log(2/\delta)}{2N}}$ by choosing the right N. Again, we are going to compute the gradient with respect to N, set the gradient to zero, and solve for N. If we do that, we will roughly get (here we are very sloppy on constants and log-terms):

$$N = (T/K)^{2/3}.$$

which gives us the regret:

$$\texttt{Regret} = \tilde{O}(K^{1/3}T^{2/3}).$$

Finally, recall that we computed the regret by *assuming the confidence intervals are all valid*. The probability of the confidence intervals being not valid (i.e., there exists at least one arm whose confidence interval does not contain its true expectation μ) is at most $K\delta$ via the union bound argument. Thus, the regret holds with probability $1 - K\delta$ as least.

□

References