Note on the UCB Algorithm

Wen Sun[1]

[1]Department of Computer Science, Cornell University

April 29, 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 $t \in \{0, \ldots, T-1\}$, 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 UCB

1: Play each arm once (note: this only gonna occur constant K total regret;)
2: Set $n_{i,1} = 1$ for all $i \in [K]$
3: **for** $t = 0 \to T - 1$ **do**
4: Compute $\hat{\mu}_t(i)$ for all $i \in [K]$
5: Play $I_t = \arg\max_{i \in [K]} \left(\hat{\mu}_t(i) + \sqrt{\frac{\log(TK/\delta)}{n_{t-1}(i)}} \right)$, and observe r_t
6: **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 $|r| \leq a \in \mathbb{R}^+$. Given N many i.i.d scalars $\{r_i\}_{i=1}^N \overset{iid}{\sim} \nu$ sampled from ν, we have that:*

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

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

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

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 $A_1, A_2, \ldots, A_N$, we have that $\mathbb{P}(A_1 \text{ or } A_2 \ldots \text{ or } A_N) \leq \sum_{i=1}^N \mathbb{P}(A_i)$. 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 $\mathbb{R}^2$. $\mathbb{P}(A_1 \text{ or } A_2 \ldots \text{ or } A_N)$ 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 $\text{area}(A_1 \cup A_2 \cdots \cup A_K) \leq \sum_{i=1}^N \text{area}(A_i)$.

2 Upper Confidence Bound

Below we introduce the optimal algorithm: Upper Confidence Bound (UCB).

At the beginning of iteration t, UCB maintains the average reward of each arm i so far,

$$\hat{\mu}_t(i) = \frac{1}{n_t(i)} \sum_{\tau=0}^{t-1} \mathbf{1}[I_\tau = i] r_\tau, \tag{1}$$

where $n_t(i)$ is the total number of times we pulled arm i so far, i.e., $n_t(i) = \sum_{\tau=0}^{t-1} \mathbf{1}[I_\tau = i]$. In words, for each arm, we keep tracking its current average reward using all the rewards we received from this arm so far.

Alg. 1 summarizes the UCB algorithm. Note that every round when we choose an arm I_t, we do not choose just based on the estimated mean $\hat{\mu}_t(i)$, instead we add a bonus term $\sqrt{\log(KT/\delta)/n_t(i)}$ to the estimated mean $\hat{\mu}_t(i)$. Intuitively, the bonus plus the mean forms the so called Upper Confidence Bound of the true mean μ_i, i.e.,

$\sqrt{\log(TK/\delta)/n_t(i)} + \hat{\mu}_t(i) \geq \mu_i$ (with high probability of course). The UCB algorithm picks the arm that has the largest upper confidence bound. This general strategy is called Optimistic in the Face of Uncertainty. Below we show that UCB achieves the optimal regret $\tilde{O}(\sqrt{KT})$. The reason that it is optimal is that it matches to the lower bound of MAB (one can show that there exists a MAB problem, where *any* algorithm has to suffer regret $\Omega(\sqrt{KT})$).

2.1 Analysis (optional)

Theorem 2 (Regret of UCB). *Fix $\delta \in (0,1)$. Assume that the reward sampled from ν_i is bounded in $[0,1]$ for any $i \in [K]$. Assume T is known. With probability at least $1-\delta$, we have:*

$$\texttt{Regret} = O(2\sqrt{\log(TK/\delta)}\sqrt{KT}) = \tilde{O}(\sqrt{KT}).$$

Proof. Consider a fixed iteration t and a fixed arm $i \in [K]$, applying Hoeffding's inequality with a parameter δ', we know that with probability at least $1-\delta'$, for i and t specifically, we have:

$$|\hat{\mu}_t(i) - \mu_i| \leq \sqrt{\frac{\log(2/\delta')}{n_t(i)}}. \tag{2}$$

Now we are interested in bounding the probability that $\forall t \in [T], \forall i \in [K], |\hat{\mu}_t(i) - \mu_i| \leq \sqrt{\log(2/\delta')/n_t(i)}$. Via union bound again, we have:

$$\mathbb{P}\big(\exists t \in [T], i \in [K], |\hat{\mu}_t(i) - \mu_i| \geq \sqrt{\log(2/\delta')/n_t(i)}\big) \leq \sum_{t=0}^{T-1}\sum_{i=1}^{K} \mathbb{P}(|\hat{\mu}_t(i) - \mu_i| \geq \sqrt{\log(2/\delta')/n_t(i)})$$

$$\leq TK\delta'.$$

Hence, we have:

$$\mathbb{P}\big(\forall t \in [T], i \in [K], |\hat{\mu}_t(i) - \mu_i| \leq \sqrt{\log(2/\delta')/n_t(i)}\big) \geq 1 - TK\delta'.$$

Now let us simply set $\delta' = \delta/(KT)$, we have:

$$Pr\big(\forall t \in [T], i \in [K], |\hat{\mu}_t(i) - \mu_i| \leq \sqrt{\log(2KT/\delta)/n_t(i)}\big) \geq 1 - \delta. \tag{3}$$

The above says that with probability at least $1-\delta$, the confidence interval we have are all valid, for all t and all i, simultaneously!

Let us now assume Eq 3 holds. All the analysis below condition on Eq 3 holds.

Let us consider round t. In round t, we choose I_t because it has the highest upper confidence bound, which means that I_t's upper confidence bound is no smaller than the upper confidence bound of any other arms, including $i^\star$:

$$\hat{\mu}_t(I_t) + \sqrt{\log(TK/\delta)/n_t(I_t)} \geq \hat{\mu}_t(I^\star) + \sqrt{\log(TK/\delta)/n_t(I^\star)}.$$

So the total regret is:

$$\begin{aligned}
\texttt{Regret} &= \sum_{t=0}^{T-1} \mu_{I^\star} - \mu_{I_t} \\
&\leq \sum_{t=0}^{T-1} \hat{\mu}_t(I^\star) + \sqrt{\log(KT/\delta)/n_t(I^\star)} - \mu_{I_t} \\
&\leq \sum_{t=0}^{T-1} \hat{\mu}_t(I_t) + \sqrt{\log(KT/\delta)/n_t(I_t)} - \mu_{I_t} \\
&\leq \sum_{t=0}^{T-1} \hat{\mu}_t(I_t) + \sqrt{\log(KT/\delta)/n_t(I_t)} - (\hat{\mu}_t(I_t) - \sqrt{\log(KT/\delta)/n_t(I_t)}) \\
&\leq 2\sqrt{\log(KT/\delta)} \sum_{t=0}^{T-1} \sqrt{1/n_t(I_t)}.
\end{aligned}$$

The last step is just need to upper bound $\sum_{t=0}^{T-1} \sqrt{1/n_t(I_t)}$. This step is a bit tricky and we can proceed as follows:

$$\begin{aligned}
\sum_{t=0}^{T-1} \sqrt{1/n_t(I_t)} &= \sum_{i=1}^{K} \sum_{t=0}^{T-1} \mathbf{1}[I_t = i] \sqrt{1/n_t(i)} \\
&= \sum_{i=1}^{K} \sum_{t=1}^{n_T(i)} \sqrt{1/t} \\
&\leq \sum_{i=1}^{K} \sqrt{n_T(i)}
\end{aligned}$$

The first equality above basically considers all $I_0, \ldots I_{T-1}$, and cluster them into K groups where the i group contains all steps where arm i is pulled. I.e., the group of time steps that corresponds to arm i is $\{t : \mathbf{1}[I_t = i]\}$. The size of this group should be equal to $n_T(i)$—the total number of times we pulled arm i during the whole learning process. The final inequality uses the trick that $\sum_{i=1}^{N} 1/\sqrt{i} \leq \sqrt{N}$.

Another important fact here is that $\sum_{i=1}^{K} n_T(i) = T$, i.e., the number of times we played for each arm over T rounds must sum up equal to T (because we only play one arm per iteration). Consider function $f(x) = \sqrt{x}$ which is a concave function. By Jensen's inequality, we have:

$$\frac{1}{K} \sum_{i=1}^{K} \sqrt{n_T(i)} \leq \sqrt{\frac{1}{K} \sum_{i=1}^{K} n_T(i)} = \sqrt{T/K}.$$

Hence, we get:

$$\texttt{Regret} \leq 2\sqrt{\log(KT/\delta)} \sum_{t=0}^{T-1} \sqrt{1/n_t(I_t)} \leq 2\sqrt{\log(KT/\delta)} K \sqrt{T/K} = 2\sqrt{\log(TK/\delta)} \sqrt{TK}.$$

□

3 Bibliographic Remarks for MAB

The UCB analysis can be found in Bubeck et al. (2012).

References

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 2012.