EE 290 Theory of Multi-armed Bandits and Reinforcement Learning Lecture 3 - 1/26/2021

Lecture 3

Lecturer: Jiantao Jiao *Scribe: Xinyang Geng and Yide Shentu*

1 Recap

ETC (explore-then-commit) algorithm: Pull all arms same number of times, then commit to arm with highest empirical mean reward.

$$R(T) = T\mu^* - \sum_{t=1}^{T} A_t \Rightarrow \mathrm{E}[R(T)] \lesssim T^{2/3}(K\log(T))^{1/3}$$

The problem with the **ETC** algorithm: pull every arm a constant number of times. In some cases although some arms have already yield very bad results, they will still be pulled exactly N times.

2 Adaptive exploration v.s. Non-adaptive exploration

Adaptive Exploration: the choice of next arm depends on the reward history.
Non-adaptive Exploration: pull every arm the same number of times which has been specified without seeing the reward.
Intuition: Stop pulling an arm when we are confident that it is not good enough.
Under clean event, we pull some arms and get the corresponding rewards:

Arm 1	$R_{1,1}$	$R_{1,2}$	$R_{1,3}$	...	$\sim P_1$
Arm 2	$R_{1,1}$	$R_{1,2}$	$R_{1,3}$	...	$\sim P_2$
Arm 3	$R_{1,1}$	$R_{1,2}$	$R_{1,3}$	...	$\sim P_3$
Arm 4	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

Reward of an arm $R_{a,i}$ is sampled from the corresponding distribution P_a.

Definition 1. *The true expected reward of bandit arm a is* $\mu_a \triangleq \mu(a) = \mathrm{E}_{\sim P_a}[R_{a,i}]$.

Definition 2. *At time t, the number of times arm a has been pulled is* $n_t(a)$.

With the notation defined above, we can write the empirical average reward of arm a as:

$$\bar{\mu}_t(a) = \frac{1}{n_t(a)} \sum_{j=1}^{n_t(a)} R_{a,j}$$

Theorem 3. *Under the setting of* ***Adaptive Exploration****, assume we pulled arm a* $n_t(a)$ *times:*

$$P(|\bar{\mu}_t(a) - \mu(a)| \leq r_t(a)) \geq 1 - \frac{1}{T^4}$$
$$\textit{where } r_t(a) \triangleq \sqrt{\frac{2\log(T)}{n_t(a)}}$$

Remark
One may want to apply Hoeffding's inequality here to proof theorem 3, but in this case the number of times arm a has been pulled is no longer a pre-decided value. $n_t(a)$ belongs to a very complex distribution since it is also depend on the underlying policy. Therefore one can no longer apply the Hoeffding's inequality here since the Hoeffding's inequality assumes that the sample is drawn independently from a distribution with a fixed number of times.

Proof
Let's define a new symbol:

$$\bar{v}_j(a) \triangleq \frac{1}{j}\sum_{k=1}^{j} R_{a,j}$$

for all fixed j $\leq$ T,

$$\mathbb{P}\left\{|\bar{v}_j(a) - \mu(a)| \leq \sqrt{\frac{2\log(T)}{j}}\right\} \geq 1 - \frac{2}{T^4}$$

by Hoeffding's inequality. (for every fixed j this is true) apply the union bound for all arms and all possible choice of j with the equation above we can obtain:

$$\begin{aligned}\mathbb{P}\left\{\forall a \in A, \forall j \leq T, |\bar{v}_j(a) - \mu(a)| > \sqrt{\frac{2\log(T)}{j}})\right\} &\leq \sum_a \sum_j \mathbb{P}\left\{|\bar{v}_j(a) - \mu(a)| > \sqrt{\frac{2\log(T)}{j}}\right\} \\ &\leq \sum_a \sum_j \frac{2}{T^4} \\ &\leq T^2 \frac{2}{T^4} = \frac{2}{T^2}.\end{aligned}$$

Since we only pull T times, we can at most pull T arms. Therefore

$$\mathbb{P}\left\{\forall a \in A, \forall j \leq T, |\bar{v}_j(a) - \mu(a)| \leq \sqrt{\frac{2\log(T)}{j}})\right\} \geq 1 - \frac{2}{T^4}$$

□

3 Successive elimination

Definition 4. *UCB: Upper Confidence Bound; LCB: Lower Confidence Bound*

In our special case:

$$UCB_t(a) \triangleq \hat{\mu}_t(a) + r_t(a)$$
$$LCB_t(a) \triangleq \hat{\mu}_t(a) - r_t(a)$$

Algorithm 1 Successive elimination

```
Initiate all arms to be "active"
while There are more than one "active" arm do
    Try all active arms
    for Each active arm a do
        if ∃a', UCB_t(a) < LCB_t(a') then
            Deactivate arm a
        end if
    end for
end while
```

Theorem 5. *Under the setting of successive elimination, the pseudo-regret*

$$R(T) \lesssim \sqrt{KT\log(T)}$$

Proof
Let a^* be the optimal arm. Consider a suboptimal arm a, such that $\mu(a) < \mu(a^*)$. Look at last time t when we did not deactivate arm a. From the arm elimination condition we can see that

$$\mu(a^*) - \mu(a) \leq 2[r_t(a^*) + r_t(a)]$$

($UCB_t(a)$ has to be larger than $LCB_t(a^*)$)
Since t is the **last** time we did not deactivate arm a, $n_t(a) = n_T(a)$ (after time t, arm a will be deactivated), $|n_t(a) - n_t(a^*)| \leq 1$.
Plug in $r_t(a^*)$ and $r_t(a)$ we have:

$$\begin{aligned}\Delta(a) = \mu(a^*) - \mu(a) \leq 2[r_t(a^*) + r_t(a)] &= 2\left[\sqrt{\frac{2\log(T)}{n_t(a^*)}} + \sqrt{\frac{2\log(T)}{n_t(a)}}\right] \\ &\leq 2\left[\sqrt{\frac{2\log(T)}{n_T(a)}} + \sqrt{\frac{2\log(T)}{n_T(a)}}\right] \\ &\lesssim \sqrt{\frac{\log(T)}{n_T(a)}}\end{aligned}$$

Intuition: arms got pulled large number of times can not be too bad.

$$R(T) = \sum_{t=1}^{T}(\mu^* - \mu(A_t)) = \sum_{a=1}^{K}\Delta(a)n_T(a)$$

Plug in $\Delta(a) \lesssim \sqrt{\frac{\log(T)}{n_T(a)}}$

$$
\begin{aligned}
R(T) &\lesssim \sum_{a=1}^{K} \sqrt{\log(T)}\sqrt{n_T(a)} \\
&\leq K \sum_{a=1}^{K} \sqrt{\log(T)} \frac{\sqrt{n_T(a)}}{K} \\
&\overset{\text{Jensen}}{\leq} K\sqrt{\log T}\sqrt{\sum_{a=1}^{K} \frac{n_T(a)}{K}} \\
&\overset{\sum_{a=1}^{K} n_T(a)=T}{=} K\sqrt{\log(T)}\sqrt{\frac{T}{K}} = \sqrt{KT\log(T)}
\end{aligned}
$$

□