Machine Learning Theory (CSC 482A/581B) - Lecture 21

Guest Lecturer: Bingshan Hu

1 Multi-Armed Bandit Learning

We will start to talk about the multi-armed bandit problem, which can be viewed as a form of online learning in which Learner only receives *partial information* at the end of each round. In an n-armed bandit problem, there are n arms. In each round, Learner selects one arm to pull and observes the reward associated with *only the pulled arm*. Note that the rewards associated with the other arms remain hidden from Learner. In this sense, Learner receives only partial information. The goal of Learner is to accumulate as much reward as possible over a finite sequence of pulls.

As Learner only receives partial information at the end of each round, a successful learning algorithm needs to make a good balance between *exploration* (pulling an under-sampled arm that might yield better reward) and *exploitation* (pulling the arm with the highest observed reward so far). Let $[n]$ be the arm set and n be the number of arms. Let x_i^t be the reward obtained of pulling arm i in round t. For simplicity, we assume $x_i^t \in [0,1]$. The general multi-armed bandit learning protocol can be described as follows:

Algorithm 1 Multi-Armed Bandit Learning Protocol

1: **for** $t = 1, 2, \cdots, T$ **do**
2: Algorithm $\mathcal{A}$ pulls an arm $i_t \in [n]$
3: Simultaneously, Nature generates a reward vector $(x_1^t, x_2^t, \cdots, x_n^t) \in [0,1]^n$
4: Algorithm $\mathcal{A}$ observes reward $x_{i_t}^t$
5: **end for**

The cumulative reward of an algorithm $\mathcal{A}$ over T rounds is

$$G_T[\mathcal{A}] = \sum_{t=1}^{T} x_{i_t}^t \quad .$$

The cumulative reward of arm i over T rounds is

$$G_T[i] = \sum_{t=1}^{T} x_i^t \quad .$$

Then the *regret* of an algorithm $\mathcal{A}$ over T rounds is

$$\begin{aligned} R_T[\mathcal{A}] &= \max_{i\in[n]} G_T[i] - G_T[\mathcal{A}] \\ &= \max_{i\in[n]} \sum_{t=1}^{T} x_i^t - \sum_{t=1}^{T} x_{i_t}^t \quad . \end{aligned}$$

The goal of a multi-armed bandit algorithm is typically to minimize the above regret either in expectation or with high probability or in a worst case sense, which depends on how the reward vectors are assumed to be generated. In this course we will only cover the following bandit settings.

Stochastic setting. In this setting, each arm $i \in [n]$ is associated with an (unknown) probability distribution p_i on $[0, 1]$, and the rewards for arm i are assumed to be drawn i.i.d. from p_i.

Adversarial setting. We make no probabilistic assumptions on the rewards x_i^t. The rewards can be generated by an adversary.

This lecture will focus on the stochastic multi-armed bandit problem while the adversarial setting will be covered in the coming lecture.

2 Stochastic Multi-Armed Bandit Problems

For simplicity, we assume there is no dependency among arms.

The rewards of arm i are i.i.d. according to a fixed probability distribution p_i on $[0, 1]$. However, Learner does not know the mean reward of arm i. Let x_i^t be the reward of arm i at round t.

Let μ_i be the (unknown) mean reward for arm i, i.e., $\mu_i = \mathbb{E}_{x_i^t \sim p_i}[x_i^t]$.

Let $i_t \in [n]$ be the arm pulled by Learner at round t.

In the stochastic setting, it makes sense to bound the regret of an algorithm in expectation over the draw of the rewards. The *expected regret* of algorithm $\mathcal{A}$ over T rounds is given by

$$\mathbb{E}[R_T[\mathcal{A}]] = \mathbb{E}\left[\max_{i\in[n]} \sum_{t=1}^{T} x_i^t - \sum_{t=1}^{T} x_{i_t}^t\right] \quad .$$

In practice, however, the expected regret is hard to work with as the max term is inside the expectation. Therefore, we often minimize the *pseudo-regret* instead. The pseudo-regret is given by

$$\bar{R}_T[\mathcal{A}] = \max_{i\in[n]} \mathbb{E}\left[\sum_{t=1}^{T} x_i^t - \sum_{t=1}^{T} x_{i_t}^t\right] \quad , \tag{1}$$

which pulls the max term outside the expectation.

Clearly, we always have $\bar{R}_T[\mathcal{A}] \leq \mathbb{E}[R_T[\mathcal{A}]]$, which means the pseudo-regret is upper bounded by the expected regret. However, an upper bound on the pseudo-regret does not imply an upper bound on the expected regret. (Nishant will talk more about this later: in adversarial bandit setting).

In order to minimize the pseudo-regret, one essentially wants to find the optimal arm, i.e., the arm with the highest mean reward. Various strategies have been proposed to do when we have samples from the arms.

The Upper Confidence Bound (UCB) algorithm was devised by Auer et al. (2002). It provides a good theoretical guarantee and can be implemented easily. The basic idea behind UCB algorithm is to construct a confidence interval for each arm at each round. Instead of pulling the arm with the highest sample mean, Learner pulls the arm with the highest upper confidence bound. We will talk about all these details soon.

3 UCB and Regret Bound

To introduce UCB, let's define the following first.

Let $O_i(t)$ be the number of times that arm i has been pulled until the end of round t, i.e., $O_i(t) := \sum_{s=1}^{t} \mathbf{1}(i_s = i)$.

Let $\hat{\mu}_{i,O_i(t)}$ denote the sample (empirical) mean of arm i until the end of round t, i.e., the averaged reward among $O_i(t)$ pulls: $\hat{\mu}_{i,O_i(t)} := \frac{1}{O_i(t)} \sum_{s=1}^{t} x_i^s \cdot \mathbf{1}(i_s = i)$. Note that if an arm is not pulled at round t, the indicator function will return 0.

Let $D_i(t) := \sqrt{\frac{2\ln(t)}{O_i(t-1)}}$ be the confidence radius of arm i at round t. Then the *confidence interval* of arm i at round t is defined as $[\hat{\mu}_{i,O_i(t-1)} - D_i(t), \hat{\mu}_{i,O_i(t-1)} + D_i(t)]$. Note that when constructing the confidence interval at round t, we can only use the information obtained until the end of round $t-1$. That is why we use $O_i(t-1)$ and $\hat{\mu}_{i,O_i(t-1)}$ instead of $O_i(t)$ and $\hat{\mu}_{i,O_i(t)}$. The *upper confidence bound* $\bar{\mu}_i(t)$ is defined as the upper bound of the confidence interval, i.e., $\bar{\mu}_i(t) := \hat{\mu}_{i,O_i(t-1)} + D_i(t)$.

Algorithm 2 presents UCB in detail.

Algorithm 2 UCB

1: **for** $t = 1, 2, \cdots, T$ **do**
2: $\quad \forall i \in [n]$, compute the upper confidence bound $\bar{\mu}_i(t) = \hat{\mu}_{i,O_i(t-1)} + \sqrt{\frac{2\ln(t)}{O_i(t-1)}}$
3: $\quad$ Pull arm $i_t \in \arg\max_{i \in [n]} \bar{\mu}_i(t)$
4: $\quad$ Observe reward $x_{i_t}^t$
5: **end for**

For simplicity, let i^* be the unique optimal arm and $\mu^* := \mu_{i^*}$. Let $\Delta_i := \mu_{i^*} - \mu_i$ be the gap of mean reward between the best arm i^* and any sub-optimal arm i.

Theorem 1 (Auer et al. (2002))**.** *The pseudo-regret $\bar{R}_T$ of UCB is at most*

$$\bar{R}_T[UCB] \leq \sum_{i:\Delta_i > 0} \frac{8\ln(T)}{\Delta_i} + \left(1 + \frac{\pi^2}{3}\right) \Delta_i \quad .$$

Before going into the proof of Theorem 1. Now we come back to the pseudo-regret and rewrite it as

$$\begin{aligned}
\bar{R}_T[UCB] &= \max_{i \in [n]} \mathbb{E}\left[\sum_{t=1}^{T} x_i^t - \sum_{t=1}^{T} x_{i_t}^t\right] \\
&= \max_{i \in [n]} \mathbb{E}\left[\sum_{t=1}^{T} x_i^t\right] - \mathbb{E}\left[\sum_{t=1}^{T} x_{i_t}^t\right] \\
&= \max_{i \in [n]} T\mu_i - \sum_{t=1}^{T} \mathbb{E}\left[x_{i_t}^t\right] \\
&= T\mu^* - \sum_{t=1}^{T} \mathbb{E}\left[\mu_{i_t}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} O_i(T)\right] \cdot \mu^* - \mathbb{E}\left[\sum_{i=1}^{n} O_i(T) \cdot \mu_i\right] \\
&= \sum_{i:\Delta_i > 0} \mathbb{E}[O_i(T)] \cdot \Delta_i \quad .
\end{aligned} \tag{2}$$

Now we only need to upper bound $\mathbb{E}[O_i(T)]$ for all $i \in [n]$ such that $\Delta_i > 0$.

Proof of Theorem 1. Fix any sub-optimal arm i such that $\Delta_i > 0$, and let L_i be a positive integer that will be chosen later. Then we have

$$\begin{aligned}
\mathbb{E}[O_i(T)] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(i_t = i)\right] \quad \text{(Regret decomposition)} \\
&= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(i_t = i, O_i(t-1) \leq L_i)\right]}_{(\omega) \leq L_i} + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(i_t = i, O_i(t-1) > L_i)\right] \\
&\leq L_i + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(i_t = i, O_i(t-1) > L_i)\right] \quad .
\end{aligned} \tag{3}$$

Term (ω) can be trivially bounded by L_i via bounding the indicator function directly. Each time when arm i is pulled, the number of pulls of arm i will increment by one. After the number of pulls of arm i hits L_i, the indicator function in term (ω) cannot return 1 anymore.

Remark 1. Usually, if $O_i(t-1) \leq L_i$, we say that sub-optimal arm i is in the *under-sampled* regime while if $O_i(t-1) > L_i$, we say sub-optimal arm i is in the *sufficiently sampled* regime. Later, we will show that when a sub-optimal arm in the sufficiently sampled regime, we can use Hoeffding's inequality.

If a sub-optimal arm i is pulled, it means its upper confidence bound $\bar{\mu}_i(t)$ must be no smaller than that of the best arm. Otherwise, sub-optimal arm i cannot be pulled. Therefore, we have

$$\begin{aligned}
\mathbb{E}[O_i(T)] &\leq L_i + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(\bar{\mu}_i(t) \geq \bar{\mu}_{i^*}(t), O_i(t-1) > L_i)\right] \\
&= L_i + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left(\hat{\mu}_{i,O_i(t-1)} + \sqrt{\tfrac{2\ln(t)}{O_i(t-1)}} \geq \hat{\mu}_{i^*,O_{i^*}(t-1)} + \sqrt{\tfrac{2\ln(t)}{O_{i^*}(t-1)}}, O_i(t-1) > L_i\right)\right] \quad ,
\end{aligned} \tag{4}$$

where the equality simply uses the definition of $\bar{\mu}_i(t)$ and $\bar{\mu}_{i^*}(t)$.

Note that both $O_i(t-1)$ and $O_{i^*}(t-1)$ are random variables, but both of them have a deterministic lower bound and a deterministic upper bound. Then, we have

$$\begin{aligned}
\mathbb{E}[O_i(T)] &\leq L_i + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left(\max_{L_i \leq s_i \leq t-1}\left(\hat{\mu}_{i,s_i} + \sqrt{\tfrac{2\ln(t)}{s_i}}\right) \geq \min_{0 < s^* \leq t-1}\left(\hat{\mu}_{i^*,s^*} + \sqrt{\tfrac{2\ln(t)}{s^*}}\right)\right)\right] \\
&\leq L_i + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s_i=L_i}^{t-1}\sum_{s^*=1}^{t-1} \mathbf{1}\underbrace{\left(\hat{\mu}_{i,s_i} + \sqrt{\frac{2\ln(t)}{s_i}} \geq \hat{\mu}_{i^*,s^*} + \sqrt{\frac{2\ln(t)}{s^*}}\right)}_{(\alpha)}\right] \quad ,
\end{aligned} \tag{5}$$

where the last inequality uses the union bound to pull the max and the min outside the indicator function.

Now we decompose the indicator function again. If condition (α) holds, it means at least one of the following three in (6), (7), and (8) must hold: (Why? It can be proved by using contradiction, simply reverse all the inequalities in (6), (7), and (8). Then you will find it contradicted with condition (α))

$$\hat{\mu}_{i,s_i} \geq \mu_i + \sqrt{\frac{2\ln(t)}{s_i}} \quad \text{(Sub-optimal } i \text{ is over estimated.)} \tag{6}$$

$$\hat{\mu}_{i^*,s^*} \leq \mu_{i^*} - \sqrt{\frac{2\ln(t)}{s^*}} \quad \text{(Best arm } i^* \text{ is under estimated.)} \tag{7}$$

The above two terms are nice for us as we can use Hoeffding's inequality.

$$\mu_{i^*} - \mu_i < 2\sqrt{\frac{2\ln(t)}{s_i}} \quad \text{(Only true when } s_i < \frac{8\ln(t)}{\Delta_i^2}\text{. But this term might cause trouble.)} \tag{8}$$

How to deal with (8)? It is very simple. Just tune L_i to make it never happen. Note that $s_i \geq L_i$ as s_i starts from L_i (Recall the deterministic lower bound of $O_i(t-1)$). Set $L_i = \left\lceil \frac{8\ln(T)}{\Delta_i^2} \right\rceil \geq \frac{8\ln(T)}{\Delta_i^2}$. Then (8) can never be true as it is impossible to have $\frac{8\ln(T)}{\Delta_i^2} \leq L_i \leq s_i < \frac{8\ln(t)}{\Delta_i^2} \leq \frac{8\ln(T)}{\Delta_i^2}$.

After setting $L_i = \left\lceil \frac{8\ln(T)}{\Delta_i^2} \right\rceil$, from (5) we have

$$\begin{aligned}
\mathbb{E}[O_i(T)] &\leq L_i + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s_i=L_i}^{t-1}\sum_{s^*=1}^{t-1} \mathbf{1}\underbrace{\left(\hat{\mu}_{i,s_i} + \sqrt{\frac{2\ln(t)}{s_i}} \geq \hat{\mu}_{i^*,s^*} + \sqrt{\frac{2\ln(t)}{s^*}}\right)}_{(\alpha)}\right] \\
&\leq L_i + \sum_{t=1}^{T}\sum_{s_i=L_i}^{t-1}\sum_{s^*=1}^{t-1} \mathbb{P}\underbrace{\left(\hat{\mu}_{i,s_i} \geq \mu_i + \sqrt{\frac{2\ln(t)}{s_i}}\right)}_{(6)} + \mathbb{P}\underbrace{\left(\hat{\mu}_{i,s^*} \leq \mu_{i^*} - \sqrt{\frac{2\ln(t)}{s^*}}\right)}_{(7)}
\end{aligned} \tag{9}$$

From Hoeffding's inequality, for each s_i and t we have

$$\mathbb{P}\underbrace{\left(\hat{\mu}_{i,s_i} \geq \mu_i + \sqrt{\frac{2\ln(t)}{s_i}}\right)}_{(6)} \leq e^{-2\cdot s_i \cdot \frac{2\ln(t)}{s_i}} = \frac{1}{t^4} \tag{10}$$

Similarly, we have

$$\mathbb{P}\underbrace{\left(\hat{\mu}_{i^*,s^*} \leq \mu_{i^*} - \sqrt{\frac{2\ln(t)}{s^*}}\right)}_{(7)} \leq e^{-2\cdot s^* \cdot \frac{2\ln(t)}{s^*}} = \frac{1}{t^4} \tag{11}$$

By plugging (10) and (11) to (9) we have

$$\begin{aligned}
\mathbb{E}[O_i(T)] &\leq L_i + \sum_{t=1}^{T}\sum_{s_i=L_i}^{t-1}\sum_{s^*=1}^{t-1}\left(\tfrac{1}{t^4} + \tfrac{1}{t^4}\right) \\
&\leq L_i + \sum_{t=1}^{T}\tfrac{2}{t^2} \\
&\leq \tfrac{8\ln(T)}{\Delta_i^2} + 1 + \tfrac{\pi^2}{3}
\end{aligned} \tag{12}$$

Note $L_i = \left\lceil \frac{8\ln(T)}{\Delta_i^2} \right\rceil \leq \frac{8\ln(T)}{\Delta_i^2} + 1$ and $\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}$.

By applying (12) to the definition of pseudo-regret (2) we have

$$\begin{aligned} \bar{R}_T[UCB] &= \sum_{i:\Delta_i>0} \mathbb{E}[O_i(T)] \cdot \Delta_i \\ &\leq \sum_{i:\Delta_i>0} (\frac{8\ln(T)}{\Delta_i^2} + 1 + \frac{\pi^2}{3}) \cdot \Delta_i \\ &= \sum_{i:\Delta_i>0} \frac{8\ln(T)}{\Delta_i} + (1 + \frac{\pi^2}{3}) \cdot \Delta_i \quad , \end{aligned} \tag{13}$$

which concludes the proof. □

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.