Lecture 14

Instructor: Haipeng Luo

1 Stochastic Multi-armed Bandit

In the last two lectures we have seen algorithms for the *adversarial* multi-armed bandit problem when there is no assumption on how the loss vectors are generated. On the other hand, there is a also huge literature on the *stochastic* version of the multi-armed bandit problem, where each arm represents an unknown distribution and each pull of the arm generates an independent sample of the corresponding distribution.

While the problem setting is clearly just a special case of its adversarial version, the goal for stochastic bandit is usually to derive regret bounds that are distribution-dependent and in some situations stronger than the worst-case $\mathcal{O}(\sqrt{TK})$ bound. Moreover, although in the full information setting the stochastic assumption makes the problem much easier (indeed, FTL would solve the problem already), in the bandit setting, due to the lack of feedback the problem is still quite challenging even with the stochastic assumption.

Formally, we assume that for each action a, there is a unknown distribution $\mathcal{D}_a$ with mean $\mu(a)$ such that $\ell_1(a), \ldots, \ell_T(a)$ are independent samples of $\mathcal{D}_a$. Let $a^\star = \operatorname{argmin}_a \mu(a)$ be the optimal action in terms of the expected loss. For this problem we usually care about the a slightly different version of regret, called *pseudo-regret*, defined as

$$\bar{\mathcal{R}}_T = \mathbb{E}\left[\sum_{t=1}^{T} \left(\ell_t(a_t) - \ell_t(a^\star)\right)\right]$$

which is the expected regret against the fixed action $a^\star$ (instead of the empirically best action $\operatorname{argmin}_a \sum_t \ell_t(a)$), where the expectation is over the randomness of both the environment and the algorithm. Clearly the pseudo-regret can also be simplified as

$$\bar{\mathcal{R}}_T = \mathbb{E}\left[\sum_{t=1}^{T} \left(\mu(a_t) - \mu(a^\star)\right)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K} \Delta_a \mathbf{1}\{a_t = a\}\right] = \sum_{a:\Delta_a>0} \Delta_a \mathbb{E}[n_T(a)]$$

where $\Delta_a = \mu(a) - \mu(a^\star)$ is called the suboptimality gap of action a and $n_t(a) = \sum_{\tau=1}^{t} \mathbf{1}\{a_\tau = a\}$ is the number of times action a has been pulled up to round t. Therefore, to analyze an algorithm in this setting, it boils down to bounding the term $\mathbb{E}[n_T(a)]$.

In the stochastic setting, the tradeoff between exploration and exploitation is perhaps even more intuitive. Let $\widehat{\mu}_t(a) = \frac{1}{n_t(a)} \sum_{\tau=1}^{t} \mathbf{1}\{a_\tau = a\}\ell_\tau(a)$ be the empirical mean of action a up to round t. Since the environment is stochastic, $\widehat{\mu}_t(a)$ could be a good approximation of $\mu(a)$ if $n_t(a)$ is large enough. Therefore, on one hand we want to exploit by picking the empirically best action $\operatorname{argmin}_a \widehat{\mu}_t(a)$, but on the other hand we also need to explore so that all actions are picked frequently enough and $\widehat{\mu}_t(a)$ is truly a good approximation of $\mu(a)$.

2 First Attempt: Explore-then-exploit

The simplest strategy to balance the tradeoff is to first perform pure exploration for a while, and then do pure exploitation and commit to a single action for the rest of the time. Formally, let T_0 be the number of exploration rounds to be specified later. The explore-then-exploit strategy is as follows:

1. For the first T_0 rounds, pick each action for T_0/K times (in an arbitrary order);
2. For the remaining $T - T_0$ rounds, always pick $\widehat{a} = \operatorname{argmin}_a \widehat{\mu}_{T_0}(a)$.

One can then show the following regret bound.

Theorem 1. *The pseudo-regret of explore-then-exploit is bounded as*

$$\bar{\mathcal{R}}_T \leq \sum_{a:\Delta_a>0} \left(\frac{T_0}{K} + 2T \exp\left(-\frac{T_0 \Delta_a^2}{8K} \right) \right) \Delta_a.$$

Proof. It suffices to prove that $\mathbb{E}[n_T(a)] \leq \frac{T_0}{K} + 2T \exp\left(-\frac{T_0\Delta_a^2}{8K}\right)$ for all a with $\Delta_a > 0$. Indeed, by the algorithm it is clear that

$$\mathbb{E}[n_T(a)] = \frac{T_0}{K} + (T - T_0)\mathbb{E}[\mathbf{1}\{\widehat{a} = a\}] = \frac{T_0}{K} + (T - T_0)\Pr(\widehat{a} = a),$$

and also $\Pr(\widehat{a} = a) \leq \Pr(\widehat{\mu}_{T_0}(a) \leq \widehat{\mu}_{T_0}(a^\star))$. Next note that if $\widehat{\mu}_{T_0}(a) \leq \widehat{\mu}_{T_0}(a^\star)$ happens, then one of the following two rare events must happen

$$\widehat{\mu}_{T_0}(a) \leq \mu(a) - \Delta_a/2$$
$$\widehat{\mu}_{T_0}(a^\star) \geq \mu(a^\star) + \Delta_a/2$$

since otherwise $\widehat{\mu}_{T_0}(a) > \mu(a) - \Delta_a/2 = \mu(a^\star) + \Delta_a/2 > \widehat{\mu}_{T_0}(a^\star)$. Now recall $\widehat{\mu}_{T_0}(a)$ ($\widehat{\mu}_{T_0}(a^\star)$) is the average of T_0/K i.i.d. samples of a distribution with mean $\mu(a)$ ($\mu(a^\star)$), and thus by Hoeffding's inequality (included at the end of the section) we know that the probability of each of the above two events is bounded by $\exp\left(-\frac{T_0\Delta_a^2}{8K}\right)$. A union bound then implies that $\Pr(\widehat{a} = a) \leq 2\exp\left(-\frac{T_0\Delta_a^2}{8K}\right)$, completing the proof. □

How should we choose the parameter T_0? For simplicity, let's consider the case when there are only two actions so that the optimal T_0 is such that $\frac{T_0}{2} + 2T \exp\left(-\frac{T_0\Delta^2}{16}\right)$ is minimized (Δ is the only non-zero gap). Direct calculations show that the optimal T_0 is $\frac{16}{\Delta^2} \ln\left(\frac{T\Delta^2}{4}\right)$ and the bound becomes

$$\frac{8}{\Delta}\left(1 + \ln\left(\frac{T\Delta^2}{4}\right)\right), \tag{1}$$

which only has a logarithmic dependence in T and is an instance-dependent bound that is better than the worst-case $\mathcal{O}(\sqrt{TK})$ bound as long as Δ is not too small. Note that this does not contradict with the lower bound $\Omega(\sqrt{TK})$ that we discussed previously. Indeed, recall that in the proof of the lower bound, the construction of the environment is also stochastic, but the gap is as small as $1/\sqrt{T}$.

The instance-dependent bounds we have seen before (mainly for the expert problem) are never worse than the worst-case bound, but bound (1) can actually be arbitrarily large if Δ is too small. However, while smaller Δ indeed increases the difficulty of distinguishing the best action from the suboptimal one, at the same time it also means that the picking the suboptimal action is not too terrible – it only incurs a regret Δ per round. This means bound (1) is loose for small Δ, but one can simply tighten it as

$$\min\left\{T\Delta,\ \frac{8}{\Delta}\left(1 + \ln\left(\frac{T\Delta^2}{4}\right)\right)\right\},$$

which is at most $\mathcal{O}(\sqrt{T \ln T})$ (by maximizing over Δ), meaning that the bound is never much worse than the one by using Exp3 or other adversarial bandit algorithms.

Bound (1) is in fact close to optimal for a fixed Δ, so this simple tradeoff between exploration and exploitation does work pretty well, at least in theory. However, the big caveat is that T_0 has to be tuned according to the suboptimality gaps, which are clearly unknown in practice. Moreover if T_0 is independent of the gap, one can show that the pseudo-regret can in fact be as large as $\Omega(T^{\frac{2}{3}})$. In the next section we will discuss an algorithm that addresses this issue completely.

Lemma 1 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_T \in [-B, B]$ for some $B > 0$ be independent random variables such that $\mathbb{E}[X_t] = 0$ for all $t \in [T]$, then we have for all $\delta \in (0, 1)$,*

$$\Pr\left(\sum_{t=1}^{T} X_t \geq B\sqrt{2T \ln \tfrac{1}{\delta}}\right) \leq \delta.$$

3 The UCB Algorithm

The classic algorithm for stochastic multi-armed bandit is the UCB (Upper Confidence Bound) algorithm [Auer et al., 2002], although since we use "losses" instead of "rewards" (which was used traditionally in [Auer et al., 2002]), the algorithm that we will discuss here is actually LCB (Lower Confidence Bound). For convention, we will still call it the UCB algorithm.

UCB applies a very important principle called "optimism in face of uncertainty", which is useful in many other stochastic problems with bandit feedback. The main idea of the principle is the following: among all plausible environments that are consistent with the data observed, assume the most favorable one is the true environment and act accordingly.

Let's apply this principle to stochastic multi-armed bandit. At time t, we have gather empirical averages $\widehat{\mu}_{t-1}(a)$ for each action a. What are the plausible environments, that is, the plausible values of the means $\mu(a)$, given this information? In light of Hoeffding's inequality, with high probability the mean $\mu(a)$ should be in the confidence interval (ignoring constants and logarithmic terms)

$$\left[\widehat{\mu}_{t-1}(a) - 1/\sqrt{n_{t-1}(a)}, \widehat{\mu}_{t-1}(a) + 1/\sqrt{n_{t-1}(a)}\right].$$

Having these plausible environments, we will then ask which is the most favorable one. Since our goal here is to suffer as less loss as possible, the best scenario is thus when $\mu(a)$ is exactly $\widehat{\mu}_{t-1}(a) - 1/\sqrt{n_{t-1}(a)}$ (called the lower confidence bound) for each a. Finally, we will simply be optimistic and assume that this is indeed the true environment and act according to it, which in this case will mean picking the action with the smallest lower confidence bound.

Formally, with constants and logarithmic terms carefully chosen, we define the lower confidence bound for action a at time t as

$$\mathrm{LCB}_t(a) = \widehat{\mu}_{t-1}(a) - 2\sqrt{\frac{\ln T}{n_{t-1}(a)}}.$$

Then at time t the UCB algorithm simply picks

$$a_t = \operatorname*{argmin}_{a \in [K]} \mathrm{LCB}_t(a).$$

First of all, note that $n_{t-1}(a)$ is initially 0, leading to negative infinity for $\mathrm{LCB}_t(a)$, which means the algorithm will be forced to pick each action once for the first K rounds. Afterwards, the two terms in $\mathrm{LCB}_t(a)$ are essentially playing the role of exploitation and exploration respectively since it suggests picking action with low empirical mean but penalized by how many times it has been selected. Whenever a suboptimal action is picked, its lower confidence bound will most likely go up and as a result it is less likely to be picked again in the future, which means optimism drives exploration. (Indeed, think about what happens to a pessimistic strategy that picks the action with the lowest *upper* confidence bound instead).

Notice that in contrast to randomized algorithms such as Exp3, both UCB and the explore-then-exploit strategy are deterministic algorithms – there is no randomness from the algorithms themselves. Importantly, UCB does not need to know the gaps Δ_a and is a very simple and practical algorithm (even the $\ln T$ term in $\mathrm{LCB}_t(a)$ can in fact be replaced by $\ln t$ to make the algorithm truly parameter-free). We finally prove the following bound for UCB that is in the same spirit of Eq. (1).

Theorem 2. *The pseudo-regret of UCB is bounded as*

$$\bar{\mathcal{R}}_T \leq \sum_{a:\Delta_a > 0} \left(\frac{16 \ln T}{\Delta_a} + 2\Delta_a\right)$$

Proof. Again it suffices to bound $\mathbb{E}[n_T(a)]$ by $\frac{16 \ln T}{\Delta_a^2} + 2$. Intuitively, for the first small number of rounds n (to be specified later), the concentration bounds are loose and there is nothing much to say. Therefore, we simply ignore these rounds and bound $\mathbb{E}[n_T(a)]$ by

$$n + \sum_{t=n+1}^{T} \Pr\left(a_t = a \text{ and } n_{t-1}(a) > n\right).$$

Note that n is similar to the number of pure exploration rounds T_0/K in the proof of explore-then-exploit, but the important thing is that n is merely for the analysis and is not a parameter of the algorithm. Similarly, the event $a_t = a$ happens only if one of the following two rare events happens

$$\begin{aligned} \text{LCB}_t(a^\star) &\geq \mu(a^\star) \\ \text{LCB}_t(a) &\leq \mu(a^\star) \end{aligned}$$

since otherwise $\text{LCB}_t(a) > \mu(a^\star) > \text{LCB}_t(a^\star)$ and a will not be picked according to the algorithm. Therefore we have by a union bound, $\Pr\left(a_t = a \text{ and } n_{t-1}(a) > n\right)$ is bounded by

$$\Pr\left(\text{LCB}_t(a^\star) \geq \mu(a^\star)\right) + \Pr\left(\text{LCB}_t(a) \leq \mu(a^\star) \text{ and } n_{t-1}(a) > n\right).$$

The first term, which is equivalent to

$$\Pr\left(\widehat{\mu}_{t-1}(a^\star) - \mu(a^\star) \geq 2\sqrt{\frac{\ln T}{n_{t-1}(a^\star)}}\right),$$

could be seemingly bounded by applying Hoeffding's inequality directly. However, one trap here is that $n_{t-1}(a)$ is actually also a random variable depending on the samples we observe. To deal with this subtle issue, we can imagine that there is a (infinite) sequence $X_1(a), X_2(a), \ldots$ of independent samples of $\mathcal{D}_a$ for each action a, and at time t the observed loss $\ell_t(a_t)$ is the $n_t(a_t)$-th sample of this sequence, that is, $\ell_t(a_t) = X_{n_t(a_t)}(a_t)$. With $\widetilde{\mu}_m(a) = \frac{1}{m}\sum_{k=1}^m X_m(a)$ being the average of the first m samples of this sequence, we then have $\widehat{\mu}_{t-1}(a) = \widetilde{\mu}_{n_{t-1}(a)}(a)$ and

$$\begin{aligned} &\Pr\left(\widehat{\mu}_{t-1}(a^\star) - \mu(a^\star) \geq 2\sqrt{\frac{\ln T}{n_{t-1}(a^\star)}}\right) \\ \leq\ &\Pr\left(\exists\, k \in [t-1] \ \text{ s.t. } \widetilde{\mu}_k(a^\star) - \mu(a^\star) \geq 2\sqrt{\frac{\ln T}{k}}\right) \\ \leq\ &\sum_{k=1}^{t-1} \Pr\left(\widetilde{\mu}_k(a^\star) - \mu(a^\star) \geq 2\sqrt{\frac{\ln T}{k}}\right), \end{aligned}$$

where each term is the last summation can now be bounded by $1/T^2$ using Hoeffding's inequality since k is fixed, and the summation is bounded by $1/T$.

For the second term $\Pr\left(\text{LCB}_t(a) \leq \mu(a^\star) \text{ and } n_{t-1}(a) > n\right)$, note that it is equivalent to

$$\begin{aligned} &\Pr\left(\widehat{\mu}_{t-1}(a) - 2\sqrt{\frac{\ln T}{n_{t-1}(a)}} \leq \mu(a^\star) \text{ and } n_{t-1}(a) > n\right) \\ =\ &\Pr\left(\Delta_a - 2\sqrt{\frac{\ln T}{n_{t-1}(a)}} \leq \mu(a) - \widehat{\mu}_{t-1}(a) \text{ and } n_{t-1}(a) > n\right) \end{aligned}$$

and thus by picking $n = \lfloor \frac{16 \ln T}{\Delta_a^2} \rfloor$, it is bounded by

$$\Pr\left(2\sqrt{\frac{\ln T}{n_{t-1}(a)}} \leq \mu(a) - \widehat{\mu}_{t-1}(a)\right),$$

which by the exact same argument as before is further bounded by $1/T$. This finishes the proof. □

It can be shown that the above bound for UCB is very close to optimal. Moreover, even though the bound can be arbitrarily large when the gaps are small, one can still show that the worst-case pseudo-regret for UCB is of order $\mathcal{O}(\sqrt{TK \ln T})$ (see Homework 3).

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.