EE 290 Theory of Multi-armed Bandits and Reinforcement Learning Lecture 2 - 1/21/2021

Lecture 2: Analysis of finite-arm i.i.d.-reward bandit

Lecturer: Jiantao Jiao *Scribe: Marsalis Gibson, Haotian Gu*

In this lecture, we discuss stochastic bandits (review the problem setting and introduce notation) and discuss one simple algorithm that solves the problem.

Notation	Description
A	the set of arms / actions that a learner can choose from during the game
$a \in A$	denotes a unique arm / action
$T \in \mathbb{Z}$	the total number of rounds the learner will play
$t \in [0, T]$	a specific round of the game
A_t	the action chosen by the learner during round t
P_a	the reward distribution for arm a
$\nu = (A, \{P_a\}_{a \in A})$	a combination of the set of actions and the set of each action's reward distribution for a specific problem
X_t	the reward given for the learner's action after round t
$K \in \mathbb{Z}$	total number of arms for the game
$\mu(a)$	the expected reward of arm a. It is also the mean of the reward distribution $P(a)$
μ^*	the expected reward of the optimal action/arm

1 Finite-arm i.i.d. Reward Bandit

We consider a basic model of the bandit problem with i.i.d. rewards. In the bandit problem, there are sequential interactions between a learner and its environment. On each round of the problem (every round t), the learner chooses an action A_t from a finite set of arms A and then based on the learner's action, the environment samples a reward $X_t \sim P_{A_t}$ for the learner, where P_{A_t} is the reward distribution for action A_t. There are K arms in total.

There are two properties for this problem.

- It is memoryless, meaning the reward after round t does not depend on any action or reward before round t once the action at round t is given: $P_{X_t|A_1,X_1,A_2,X_2,...,X_{t-1},A_t} := P_{A_t}$
- It is casual, meaning that the action of the learner during round T is influenced by the actions and rewards of previous rounds. $P_{A_t|A_1,X_1,A_2,X_2,...,A_{t-1},X_{t-1}} := \pi_t(\cdot|A_1, X_1, A_2, X_2, ..., A_{t-1}, X_{t-1})$

2 Designing Algorithms to Minimize Regret

If I design an algorithm to solve the finite-arm iid reward bandit problem, how well should it perform?

To analyze an algorithm's performance, we compare the algorithm's reward to the expected reward you'd get by playing the optimal arm for every round. This comparison is known as the regret. The goal is to design an algorithm that minimizes regret.

Definition 1 (Best Arm Benchmark)**.** *Always playing the optimal arm every round is the best outcome for the learner and results in the best performance for the game. Denote the expected reward of the optimal arm as μ^*. We call $T \cdot \mu^*$ the the best arm benchmark [1].*

Definition 2 (Random Regret)**.** *Random regret is defined as the difference between the best arm benchmark and our cumulative reward during the game:*

$$T\mu^* - \sum_{t=1}^{T} X_t \tag{1}$$

With **random regret**, we take the accumulation of rewards the learner received during the game and compare that value against the best arm benchmark. On the other hand, **pseudo regret** uses the expected reward of each arm played during the game and sums them together to get the algorithm's performance.

Definition 3 (Pseudo Regret)**.** *Pseudo regret is the difference between the best arm benchmark and the sum of the expected rewards of each arm played:*

$$R(T) = T\mu^* - \sum_{t=1}^{T} \mu_{A_t} \geq 0 \tag{2}$$

In analyzing the first bandit algorithm, we're going to compute the algorithm's worst case expected pseudo regret, or just simply **worst case expected regret**.

Definition 4 (Worst Case Expected Regret)**.** *Worst case expected regret is defined as the worst-case expected pseudo regret:*

$$\sup_{\nu \in \mathcal{V}} \mathbb{E}_\nu [T\mu^* - \sum_{t=1}^{T} \mu_{A_t}] \tag{3}$$

3 The Explore-then-Commit Algorithm and its Analysis

In this section, we are going to propose the first bandit algorithm in this course: **Explore-then-Commit**, and analyze its **worst-case expected regret**.

The algorithm is based on a simple idea: explore arms uniformly (at the same rate), regardless of what has been observed previously, and pick an empirically best arm for exploitation. The idea can be translated directly to Algorithm 1.

Algorithm 1 Explore-then-Commit

Input: K arms with unknown reward distributions $\{P_a, a = 1 \cdots, K\}$, T rounds, exploration parameter $N \in \mathbb{N}$
Phase 1: Exploration in the first $N \cdot K$ rounds:
for $1 \leq a \leq K$ **do**
  for $1 \leq t \leq N$ **do**
    Play arm a, and get reward $\hat{r}_{t,a} \overset{\text{i.i.d}}{\sim} P_a$.
  end for
  Calculate $\hat{\mu}(a) = \frac{1}{N} \sum_{t=1}^{N} \hat{r}_{t,a}$.
end for.
Phase 2: Exploitation in the remaining $T - N \cdot K$ rounds:
Play $\hat{a} = \arg\max_a \hat{\mu}(a)$ in all remaining rounds.

Note that in Algorithm 1, we dedicate an initial segment of rounds to exploration, and the remaining rounds to exploitation. Intuitively, for a fixed number of total rounds T, if we choose a large N, which means

longer exploration, then we may become more confident when we exploit, but the regret from exploration will increase; on the other hand, a smaller N results in lower regret from exploration, but we are more likely to pick the sub-optimal arm in exploitation. This is a clear illustration of the exploration-exploitation trade-off we discussed in the first lecture. Later we will see from the analysis that, choosing an appropriate N to balance exploration and exploitation is the key to get a desired regret bound.

Before the formal regret analysis, without loss of generality, we assume that **for every arm, the reward is bounded between 0 and 1 with probability 1.** Meanwhile, we introduce the following well-known result, called **Hoeffding's inequality**.

Theorem 5. *Let* $X_1, X_2, X_3, \cdots \overset{i.i.d}{\sim} P$, $\mathbb{E}[X_i] = \mu$, *and* $X_i \in [0,1]$ *with probability 1. Then*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq 2\exp(-2n\epsilon^2). \tag{4}$$

Remark Theorem 5 presented here is a special case of the general Hoeffding's inequality. For more detailed discussions of Hoeffding's inequality, please see Chapter 2.6, [2].

Under the assumptions of Theorem 5, we notice that for a fixed ϵ, the probability that the empirical mean deviates from the expectation by ϵ will decay exponentially fast, as the number of samples n increases. In other words, it characterizes how fast the empirical mean will eventually 'concentrate' to a small neighborhood of the expectation. Hence, Hoeffding's inequality belongs to a broad class of inequalities, called **concentration inequalities**.

Now we are ready to analyze the worst-case expected regret.

Theorem 6. *With* $N = (T/K)^{2/3}(\ln T)^{1/3}$, *Algorithm 1 achieves worst-case expected regret bound* $\mathbb{E}[R(T)] \leq O(T^{2/3}(K\ln T)^{1/3})$.

Proof We prove the theorem by breaking the regret into different parts under different events:

Step 1: Probability of 'Clean Event'. Define $r = \sqrt{\frac{2\ln T}{N}}$, and define the **clean event**

$$\mathcal{E} := \{\text{for all arms } a \in \{1, \cdots, K\},\ |\hat{\mu}(a) - \mu(a)| \leq r\}.$$

By Theorem 5, for each arm a,

$$\mathbb{P}\left(|\hat{\mu}(a) - \mu(a)| \geq r\right) \leq 2\exp(-2N \cdot r^2) = \frac{2}{T^4}.$$

Hence, by the union bound,

$$\mathbb{P}(\mathcal{E}^c) \leq \mathbb{P}\left(\exists a \in \{1, \cdots, K\},\ |\hat{\mu}(a) - \mu(a)| \geq r\right) \leq \frac{2K}{T^4}.$$

Consequently,

$$\mathbb{P}(\mathcal{E}) \geq 1 - \frac{2K}{T^4}. \tag{5}$$

Remark In Step 1, we define the 'clean event' to be the case when every empirical mean $\hat{\mu}(a)$ stays close to the true expectation $\mu(a)$, and provide a lower bound for $\mathbb{P}(\mathcal{E})$ using Hoeffding's inequality. Under $\mathcal{E}$, we expect that the arm chosen in the exploitation phase is 'good enough', and that the regret from exploitation can be nicely bounded. We will make this intuition formal in Step 2.

Step 2: Regret under 'Clean Event'. Now assume $\mathcal{E}$ holds. Let $\hat{a} = \arg\max_a \hat{\mu}(a)$ be the arm we choose to exploit, and let $a^* = \arg\max_a \mu(a)$ be the true optimal arm. Then, by definition

$$\hat{\mu}(\hat{a}) \geq \hat{\mu}(a^*), \quad \mu(\hat{a}) \leq \mu(a^*).$$

Meanwhile, under $\mathcal{E}$,

$$\mu(\hat{a}) + r \geq \hat{\mu}(\hat{a}), \quad \hat{\mu}(a^*) \geq \mu(a^*) - r.$$

These imply

$$\mu(\hat{a}) + r \geq \hat{\mu}(\hat{a}) \geq \hat{\mu}(a^*) \geq \mu(a^*) - r.$$

Hence,

$$0 \leq \mu(a^*) - \mu(\hat{a}) \leq 2r = 2\sqrt{\frac{2\ln T}{N}}. \tag{6}$$

Therefore, under $\mathcal{E}$, the regret from the exploration phase is upped bounded by NK, which is a trivial bound. And the regret from the exploitation phase is upped bounded by $2r(T - NK)$, since there are in total $T - NK$ rounds of exploitation and the arm $\hat{a}$ we exploit must satisfy (6). Putting them together, we get

$$R(T) \leq NK + 2\sqrt{\frac{2\ln T}{N}}(T - NK) \text{ under } \mathcal{E}. \tag{7}$$

Step 3: Regret under 'Dirty Event'. Now assume the complement of 'clean event' $\mathcal{E}^c$ holds. Note that this case can only happen with a small probability by (5). Thus, in this case, we just employ a trivial bound

$$R(T) \leq T \text{ under } \mathcal{E}^c. \tag{8}$$

Step 4: Put Everything Together. Combining Step 1-3 gives us

$$\begin{aligned}
\mathbb{E}[R(T)] &= \mathbb{E}[R(T)\mathbb{1}(\mathcal{E})] + \mathbb{E}[R(T)\mathbb{1}(\mathcal{E}^c)] \\
&\leq \mathbb{E}[R(T)\mathbb{1}(\mathcal{E})] + T \cdot \frac{2K}{T^4} \\
&\leq NK + 2\sqrt{\frac{2\ln T}{N}}(T - NK) + \frac{2K}{T^3} \\
&\leq NK + 2T\sqrt{\frac{2\ln T}{N}} + \frac{2K}{T^3}.
\end{aligned}$$

Now set $N = (T/K)^{2/3}(\ln T)^{1/3}$ and we will get $\mathbb{E}[R(T)] \leq O(T^{2/3}(K\ln T)^{1/3})$. □

Remark The proof is mainly following the presentation in Section 1.2 of [1]. In some other literature, for example, Chapter 6.1, [3], the regret bound for Algorithm 1 is claimed to be $O(T^{1/2})$. The difference is that in those proofs, it is required to know the so called **suboptimality gap** (performance gap between the best arm and the second best arm) in prior, in order to design N, which is a seldom reasonable assumption.

This completes the worst-case expected regret analysis for Algorithm 1. It is interesting to see that, with such a naive algorithm, the regret bound is sublinear with respect to T, which indicates that as $T \to \infty$, we will almost surely choose the optimal arm.

References

[1] A. Slivkins, "Introduction to multi-armed bandits," *arXiv preprint arXiv:1904.07272*, 2019.

[2] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[3] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.