Stat 6341.001: Numerical Linear Algebra and Statistical Computing, Fall 2026 **Topic 2**

Multi-Armed Bandit: ε-Greedy and UCB

Lecturer: Yun Wei *Scribe: Sudakshina Singha Roy*

1 Introduction

We briefly recall the multi-armed bandit setup and the ε-Greedy algorithm from the previous lecture.

1.1 Multi-Armed Bandit Problem

for $t = 1, \ldots, T$ do
  select decision $\pi^t \in \Pi := \{1, \ldots, A\} = [A]$
  observe reward $r^t \in \mathbb{R}$

Here, t denotes the current time step, T is the total number of time steps, and A is the total number of available actions.

Let $f^*(\pi)$ denote the mean reward of action π, and let $\pi^* = \arg\max_{\pi \in \Pi} f^*(\pi)$ be an optimal action.

Regret is defined as:

$$\mathrm{Reg} = \sum_{t=1}^{T} f^*(\pi^*) - \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left[f^*(\pi^t) \right].$$

The first term is the total mean reward we could obtain by always choosing the optimal action π^*, whereas the second term is the expected total mean reward obtained by the algorithm. Thus, the goal is to minimize the regret.

1.2 The ε-Greedy Algorithm

Since the true mean reward function f^* is unknown, we estimate it from the rewards observed so far.
For an action π, define $n^t(\pi) = \sum_{s<t} \mathbf{1}_{\{\pi^s = \pi\}}$, where $n^t(\pi)$ is the number of times action π has been selected before time t.

The empirical estimate of $f^*(\pi)$ is

$$\widehat{f}^t(\pi) = \frac{1}{n^t(\pi)} \sum_{s<t} r^s \mathbf{1}_{\{\pi^s = \pi\}}.$$

The indicator function is defined as: $\mathbf{1}_{\{\pi^s = \pi\}} = \begin{cases} 1, & \pi^s = \pi, \\ 0, & \pi^s \neq \pi. \end{cases}$

The ε-greedy algorithm chooses with probability $1 - \varepsilon$,

$$\widehat{\pi}^t = \arg\max_{\pi \in \Pi} \widehat{f}^t(\pi)$$

and with probability ε,

$$\pi^t \sim \mathrm{unif}\,[A],$$

Thus, with probability $1-\varepsilon$ the algorithm uses the currently best estimated action, while with probability ε it explores.

2 Continuation of the ε-Greedy Algorithm

Proposition 2.1. *Assume that $f^*(\pi) \in [0,1]$ and the reward r^t is 1-sub-Gaussian (always think of it as Gaussian with variance 1). Then, for any finite T, by choosing ε appropriately, the ε-greedy algorithm ensures that with probability at least $1-\delta$,*

$$\mathrm{Reg} \lesssim A^{1/3}T^{2/3}\log^{1/3}\left(\frac{AT}{\delta}\right)$$

Consequently,

$$E\left[\mathrm{Reg}\right] \lesssim A^{1/3}T^{2/3}\log^{1/3}\left(AT\right)$$

To prove the proposition, let us consider the following lemma.

Lemma 2.2. *With probability at least $1-\delta$,*

$$\max_{\pi}\left|\widehat{f}^t(\pi) - f^*(\pi)\right| \lesssim \sqrt{\frac{A\log(AT/\delta)}{\varepsilon t}}$$

Comment: The left-hand side measures the largest estimation error over all actions. As t becomes larger, more observations are collected, so the estimation error becomes smaller. For a proof of the lemma, see Section 2.2 of [1].

Proof of Proposition 2.1

We begin with

$$\mathrm{Reg} = \sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t}\left[f^*(\pi^*) - f^*(\pi^t)\right].$$

Using the definition of the ε-greedy algorithm, this can be decomposed as

$$\mathrm{Reg} = \sum_{t=1}^{T}(1-\varepsilon)\left[f^*(\pi^*) - f^*(\widehat{\pi}^t)\right] + \varepsilon\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim\mathrm{unif}([A])}\left[f^*(\pi^*) - f^*(\pi^t)\right].$$

Since $f^*(\pi)\in[0,1]$,$f^*(\pi^*) - f^*(\pi^t) \le 1$. Therefore,

$$\mathrm{Reg} \le \sum_{t=1}^{T}\left[f^*(\pi^*) - f^*(\widehat{\pi}^t)\right] + \varepsilon T.$$

Now decompose

$$f^*(\pi^*) - f^*(\widehat{\pi}^t) = \left[f^*(\pi^*) - \widehat{f}^t(\pi^*)\right] + \left[\widehat{f}^t(\pi^*) - \widehat{f}^t(\widehat{\pi}^t)\right] + \left[\widehat{f}^t(\widehat{\pi}^t) - f^*(\widehat{\pi}^t)\right].$$

Because $\widehat{\pi}^t = \arg\max_\pi \widehat{f}^t(\pi)$, we have $\widehat{f}^t(\pi^*) - \widehat{f}^t(\widehat{\pi}^t) \le 0$. Hence,

$$f^*(\pi^*) - f^*(\widehat{\pi}^t) \le 2\max_{\pi}\left|f^*(\pi) - \widehat{f}^t(\pi)\right|.$$

Therefore,

$$\text{Reg} \leq 2\sum_{t=1}^{T} \max_{\pi} \left| f^*(\pi) - \widehat{f}^t(\pi) \right| + \varepsilon T.$$

Using the lemma 2.2, with probability at least $1-\delta$

$$\text{Reg} \lesssim 2\sum_{t=1}^{T} \sqrt{\frac{A\log(AT/\delta)}{\varepsilon t}} + \varepsilon T.$$

Using the fact $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T} - 1$, we obtain

$$\text{Reg} \lesssim 2\sqrt{\frac{A\log(AT/\delta)}{\varepsilon}}\sqrt{T} + \varepsilon T.$$

Comment: The first term is associated with the estimation error, while the second term comes from random exploration.

To balance these two terms, choose $\varepsilon \propto \left(\frac{A\log(AT/\delta)}{T}\right)^{1/3}$.

Plugging this choice into the previous bound gives

$$\text{Reg} \lesssim A^{1/3}T^{2/3}\log^{1/3}\left(\frac{AT}{\delta}\right).$$

This leads to the following question: can we improve the algorithm? The ε-greedy algorithm explores with probability ε, even for those well-estimated bad arms/actions. Can we choose actions in a more adaptive way instead of exploring uniformly?

3 Upper Confidence Bound (UCB) Algorithm

Suppose that at each time t we can construct a lower and upper confidence bound

$$\underline{f}^t, \overline{f}^t : \Pi \to \mathbb{R}$$

such that, with probability at least $1-\delta$,

$$f^*(\pi) \in \left[\underline{f}^t(\pi), \overline{f}^t(\pi)\right]$$

for every $t \in [T]$ and every $\pi \in [A]$.

The UCB algorithm chooses $\pi^t = \arg\max_{\pi\in\Pi} \overline{f}^t(\pi)$. Thus, at time t, the chosen action is the one having the largest upper confidence bound.

Lemma 3.1 (Instantaneous regret of UCB)**.** *For t, suppose $f^*(\pi) \in \left[\underline{f}^t(\pi), \overline{f}^t(\pi)\right]$*
Then $\pi^t = \arg\max_{\pi} \overline{f}^t(\pi)$ has the following property:

$$f^*(\pi^*) - f^*(\pi^t) \leq \overline{f}^t(\pi^t) - f^*(\pi^t) \leq \overline{f}^t(\pi^t) - \underline{f}^t(\pi^t).$$

Proof. Because π^t maximizes the upper confidence bound, $\overline{f}^t(\pi^*) \le \overline{f}^t(\pi^t)$. Also, because the confidence interval contains the true mean reward, $f^*(\pi^*) \le \overline{f}^t(\pi^*)$. Combining these gives $f^*(\pi^*) \le \overline{f}^t(\pi^t)$. Therefore,

$$f^*(\pi^*) - f^*(\pi^t) \le \overline{f}^t(\pi^t) - f^*(\pi^t).$$

Finally, $f^*(\pi^t) \ge \underline{f}^t(\pi^t)$, so $\overline{f}^t(\pi^t) - f^*(\pi^t) \le \overline{f}^t(\pi^t) - \underline{f}^t(\pi^t)$.

□

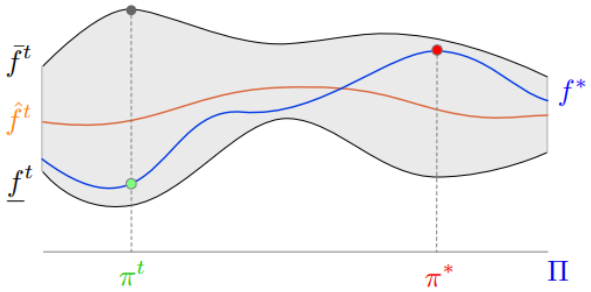

Figure 1: Illustration of the UCB algorithm.

Lemma 3.2 (Construction of confidence bounds)**.** *With probability 1-δ,*

$$\max \left| \widehat{f}^t(\pi) - f^*(\pi) \right| \le \sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi)}}.$$

We skip the proof. Then by the lemma,

$$\underline{f}^t(\pi) := \widehat{f}^t(\pi) - \sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi)}} \le f^*(\pi) \le \widehat{f}^t(\pi) + \sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi)}} =: \overline{f}^t(\pi)$$

Proposition 3.3 (UCB regret)**.** *Assume $f^*(\pi) \in [0,1]$ and r^t is 1-sub-Gaussian. Then, with probability at least $1-\delta$,*

$$\mathrm{Reg} \lesssim \sqrt{AT\log(AT/\delta)}.$$

Comment: This gives a $T^{1/2}$ dependence, compared with the $T^{2/3}$ dependence of the ε-greedy bound. A smaller power of T is better because the regret grows more slowly. Thus, the UCB algorithm has a better regret bound than the ε-greedy algorithm.

Proof of Proposition 3.3. Using lemma 3.1 and our construction of the confidence bounds, the regret can be bounded by the sum of the confidence widths: with probability 1-δ,

$$f^*(\pi^*) - f^*(\pi^t) \le \overline{f}^t(\pi^t) - \underline{f}^t(\pi^t) \le 2\sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi^t)}}$$

On the other hand,

$$f^*(\pi^*) - f^*(\pi^t) \le 1.$$

Thus, with probability 1-δ,

$$f^*(\pi^*) - f^*(\pi^t) \le 2\sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi^t)}} \wedge 1$$

where $a \wedge b = \min\{a, b\}$.

Thus with probability 1-δ,

$$\begin{aligned}\mathrm{Reg} &= \sum_{t=1}^{T} \left[f^*(\pi^*) - f^*(\pi^t)\right] \\ &\lesssim \sum_{t=1}^{T} \left(2\sqrt{\frac{2\log(2T^2A/\delta)}{n^t(\pi^t)}} \wedge 1\right).\end{aligned}$$

here, the "$\wedge 1$" term appears because we can write off the regret for early rounds where $n^t(\pi^t) = 0$ as 1.

To be continued. □

References

[1] D. J. Foster and A. Rakhlin, *Foundations of reinforcement learning and interactive decision making*, arXiv preprint arXiv:2312.16730, 2023. Sections 2.2 and 2.3.