Stat 6341.001: Numerical Linear Algebra and Statistical Computing, Fall 2026 **Topic 3**

UCB

Lecturer: Yun Wei *Scribe: Yun Wei*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

UCB:

Define

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

The upper confidence bound is

$$\overline{f}^{\,t}(\pi) = \widehat{f}^{\,t}(\pi) + \sqrt{\frac{2\log\left(\frac{2\pi^2 t^2 A}{\delta}\right)}{n^t(\pi)}}.$$

Similarly, the lower confidence bound is

$$\underline{f}^{\,t}(\pi) = \widehat{f}^{\,t}(\pi) - \sqrt{\frac{2\log\left(\frac{2\pi^2 t^2 A}{\delta}\right)}{n^t(\pi)}}.$$

With probability at least $1-\delta$,

$$\underline{f}^{\,t}(\pi) \le f^\star(\pi) \le \overline{f}^{\,t}(\pi).$$

The UCB action selection rule is

$$\pi^t = \arg\max_{\pi} \overline{f}^{\,t}(\pi).$$

Initialization:

At $t = 1$,

$$n^t(\pi) = 0, \qquad \forall \pi \in [A],$$

and thus

$$\overline{f}^{\,1}(\pi) = \infty, \qquad \forall \pi \in [A].$$

In the first step, any arm may be selected since it is all infinity. Once a certain arm is selected, then the upper confidence bound $\overline{f}$ for that arm is finite. So at $t = 2$, any remaining arm maybe selected. In the first A steps, each arm will be selected once.

Lemma 0.1 (Confidence width potential lemma)**.** *We have*

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 \lesssim \sqrt{AT}.$$

Proof:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 = \sum_{\pi} \sum_{t=1}^{T} \frac{\mathbf{1}(\pi^t = \pi)}{\sqrt{n^t(\pi)}} \wedge 1.$$

Since

$$n^t(\pi^t) = \sum_{\pi} n^t(\pi) \mathbf{1}(\pi^t = \pi),$$

we can write

$$\frac{1}{\sqrt{n^t(\pi^t)}} = \sum_{\pi} \frac{\mathbf{1}(\pi^t = \pi)}{\sqrt{n^t(\pi)}}.$$

Therefore,

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 = \sum_{\pi} \sum_{t=1}^{T} \frac{\mathbf{1}(\pi^t = \pi)}{\sqrt{n^t(\pi)}} \wedge 1.$$

Re-indexing according to the number of times arm π is selected,

$$= \sum_{\pi} \sum_{t=1}^{n^{T+1}(\pi)} \frac{1}{\sqrt{t-1}} \wedge 1.$$

Using

$$\sum_{t=1}^{n} \frac{1}{\sqrt{t-1}} \wedge 1 \leq 1 + 2\sqrt{n},$$

we obtain

$$\leq \sum_{\pi} \left(1 + 2\sqrt{n^{T+1}(\pi)} \right).$$

Thus,

$$= A + 2 \sum_{\pi} \sqrt{n^{T+1}(\pi)}.$$

By Jensen's inequality, since $F(x) = \sqrt{x}$ is concave,

$$\frac{1}{A} \sum_{\pi} \sqrt{n^{T+1}(\pi)} \leq \sqrt{\frac{1}{A} \sum_{\pi} n^{T+1}(\pi)}.$$

Hence,

$$A + 2 \sum_{\pi} \sqrt{n^{T+1}(\pi)} = A + 2A \left(\frac{1}{A} \sum_{\pi} \sqrt{n^{T+1}(\pi)} \right)$$

$$\leq A + 2A \sqrt{\frac{1}{A} \sum_{\pi} n^{T+1}(\pi)}.$$

Since

$$\sum_{\pi} n^{T+1}(\pi) = T,$$

we get

$$\leq A + 2A \sqrt{\frac{T}{A}} = A + 2\sqrt{AT}.$$

Finally, consider two cases:

$$\text{1. If } A \leq \sqrt{AT} \quad \text{(equivalently, } A \leq T\text{)},$$

then

$$A + 2\sqrt{AT} \leq 3\sqrt{AT}.$$

$$\text{2. If } A > T,$$

then trivially

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 \leq T \leq \sqrt{AT}.$$

Therefore,

$$\boxed{\sum_{t=1}^{T} \frac{1}{\sqrt{n^t(\pi^t)}} \wedge 1 \leq 3\sqrt{AT}}$$

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.