**ECE 8803: Online Decision Making in Machine Learning** **Fall 2022**

# Lecture 17: October 27

*Lecturer: Guanghui Wang*

**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.*

Last class, we introduced the multi-armed bandit (MAB) problem in a formal framework, discussed the metric of pseudo-regret at length, and briefly introduced the *upper-confidence-bound* (UCB) algorithm and motivated at a high level why it trades off exploration and exploitation.

## 17.1. Recap: The MAB problem, pseudo-regret, and UCB

The *$K$-armed multi-armed bandit problem* involves taking one out of $K$ actions at every round $t$ based on past reward feedback. Concretely, we take action $A_t$ at round $t$, and get a reward of $G_{t,A_t}$. Our goal is to maximize the *expected reward* $\mathbb{E}\left[\sum_{t=1}^T G_{t,A_t}\right]$; equivalently to minimize what we call the *pseudo-regret*, with respect to the best action that we could have taken in hindsight, i.e. $\overline{R}_T := T\mu^* - \mathbb{E}\left[\sum_{t=1}^T G_{t,A_t}\right]$.

We assume that the optimal arm is unique, for simplicity[1], and denote it by $a^*$ Then, that the pseudo-regret has an equivalent expression as below:

$$\overline{R}_T = \sum_{a \neq a^*} \Delta_a \cdot \mathbb{E}\left[N_a(T)\right], \tag{17.1}$$

where $N_a(T)$ denotes the number of times[2] that a suboptimal arm $a$ was sampled, and $\Delta_a := \mu^* - \mu_a$ denotes the suboptimality gap. This has an intuitive meaning: the greater the number of times we sample a suboptimal arm, the higher the regret will be; the more suboptimal that arm is, the higher the regret will be.

Unlike in the setting of *full-information feedback* that we were studying earlier, now we only get to see the reward of the action that we picked, $G_{t,A_t}$, at every round. We have seen in the last two lectures that this type of *limited-information feedback* can have catastrophic implications for the greedy algorithm, which simply picks the arm with the best sample mean at any given round. On the other hand, only exploring is also suboptimal as we do not at all *exploit* the useful information in the reward data that we acquire. Thus, effective algorithms in the MAB environment will trade off elements of exploration and exploitation.

[1] This does not affect the results and proof that we show in today's lecture: it just makes the details a little more complicated.

[2] Recall that the reason that there is an expectation is because the number of times a suboptimal arm can be sampled will in general depend on the *random* realization of the rewards as well as the choice of algorithm.

One of the algorithms that does this best is the *upper-confidence-bound* (UCB) algorithm, which is detailed below:

$$A_t = \arg\max_a \left[\widehat{\mu}_{a,t-1} + \sqrt{\frac{\log(1/\delta)}{2N_{t-1}(a)}}\right]. \tag{17.2}$$

We denote the RHS of the above for each $a$ as $\mathsf{UCB}(a,t)$ as shorthand; then, we have $A_t = \arg\max_a \mathsf{UCB}(a,t)$. Above, $\delta$ is a parameter that will dictate the width of the confidence interval, as well as the probability that the true mean lies within the interval: we will discuss this more at length now. In the meantime, we explain why the objective of maximizing $\mathsf{UCB}(a,t)$ does, indeed, incentivize both exploitation and exploration. Notice that a) the term $\widehat{\mu}_{a,t-1}$ encourages picking arms with a larger sample mean obtained thus far, and b) the term $\sqrt{\frac{2\log(1/\delta)}{N_{t-1}(a)}}$ encourages picking arms that have been pulled relatively infrequently thus far (because it is inversely proportional to $N_{t-1}(a)$).

From a cognitive perspective, UCB encapsulates the principle of *optimism in the face of uncertainty*: we take an optimistic perspective on arms that we have seen very few times thus far. We talked about clinical trials/drug discovery as a prototypical example of the MAB problem in class. It may additionally be fun to think about an example in your life in which you have multiple choices that you're attempting to learn about, limited-information feedback, and an exploration-exploitation tradeoff. Do you think you would use the optimism principle?

### 17.1.1 The upper-confidence principle and the meaning of $\delta$

To see this picture at work, we now consider the value of the hyperparameter $\delta$ and its connection to the notion of a confidence interval. We discuss this *informally* here, and briefly touch upon formal caveats at the end. Suppose that we had seen $n$ samples of arm $a$ before round $t$. Then, an application of Hoeffding's lemma tells us that

$$\begin{aligned}\mathbb{P}\left[\widehat{\mu}_{a,t-1} - \mu_a > \sqrt{\frac{\log(1/\delta)}{2n}}\right] &\leq \exp\left(-\frac{2n \cdot \log(1/\delta)}{2n}\right) \\ &= \delta.\end{aligned}$$

This tells us that with probability at least $1-\delta$, our sample mean of arm $a$ would would not be *too* much larger than the true mean given by $\mu_a$. This is called a $(1-\delta)$-*confidence interval*, and essentially is the reason why the UCB algorithm is named so! Furthermore, the extent of closeness decays with the number of times the arm is sampled, $n$, in a $1/\sqrt{n}$ fashion. In other words, the $(1-\delta)$-confidence intervals *shrink in width* as more samples are drawn of a particular arm.

### 17.1.2 Properties of UCB that we will use

We demonstrated the performance of UCB on various random examples. One such snapshot is shown in Figure 17.1. In particular, we noticed the following two patterns across all the examples:

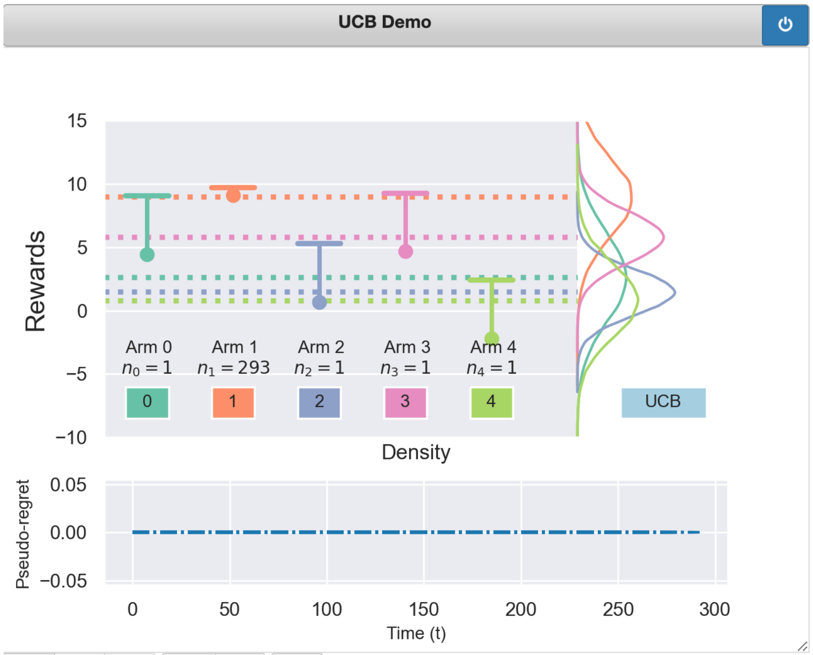

Figure 17.1: Snapshot of demo of UCB algorithm with parameter $\delta = 1/T^2$ (this choice is explained in this lecture) on a randomly chosen 5-armed bandit instance. Here, arm 1 is the best arm and is pulled disproportionately often by UCB. This demo is borrowed from the lectures in UC Berkeley's DS102 course, Fall 2019 iteration, and was originally created by graduate student Karl Krauth.

- For the optimal arm, the *true* mean $\mu^*$ is always contained within its confidence interval; in other words, regardless of how many rounds have been played, we always have $\mu^* < \mathsf{UCB}(a^*, t)$. Thus, even when the optimal arm has been sampled many times, its UCB always lies slightly above its *true mean*. This property is clearly visible in Figure 17.1 even after 293 samples of the optimal arm (1 in this example), and will prove to be useful to analyze UCB.

- After suboptimal arms have been sampled a minimal number of times, their *upper confidence bounds* tend to lie below the true mean of the optimal arm $\mu^*$. This minimal number of times tends to depend on how big the gap was between $\mu^*$ and $\mu_a$ in the first place: the larger the gap, the fewer number of times arm $a$ needs to be sampled before we observe this behavior. You can see this manifest in the extreme example in Figure 17.1: arms 2 and 4 are more suboptimal than arms 3 and 0 with respect to the optimal arm 1 — so even after just 1 pull, their UCB's will be below the optimal arm mean $\mu_1$. This is not the case for arms 3 and 0, which will require more pulls to be conclusively ruled out.

In fact, this is essentially the reason for why regret tends to have an inverse dependence on the suboptimality gap between arms. We will shortly define these properties mathematically, and use them to show that UCB incurs a very low pseudo-regret (that also turns out to be the best that you can do!).

## 17.2. Proof of UCB pseudo-regret

For the rest of this lecture, we will show the following pseudo-regret bound on UCB:

**Theorem 1** *UCB with the choice $\delta = 1/T^2$ achieves pseudo-regret*

$$\overline{R}_T \leq 3 \sum_{a \neq a^*} \Delta_a + \sum_{a \neq a^*} \frac{4 \log T}{\Delta_a}$$

Notice that Theorem 1 shows that the pseudo-regret of UCB scales only *logarithmically* in the number of rounds $T$. This is of course much better than the greedy and "only explore" aproaches, which each incurred linear regret in $T$; but it is also much better than the briefly reviewed explore-then-commit and $\epsilon$-greedy algorithms, which turn out to incur $\mathcal{O}(T^{2/3})$ regret in the worst case.

We will now prove Theorem 1. This proof is very inspired by the treatment in Chapter 7 of Lattimore and Szepesvári (2020), which is worth a concurrent read along with this lecture note in order to further internalize the proof details. We will briefly mention places where the notes differ slightly from Chapter 7 of Lattimore and Szepesvári (2020).

We will show that $\mathbb{E}[N_a(T)]$ is not too large for a suboptimal arm $a$; in particular, we will show that

$$\mathbb{E}[N_a(T)] \leq 3 + \frac{4 \log T}{\Delta_a^2} \tag{17.3}$$

for each value of $a \neq a^*$. It is easily verified that plugging this upper bound into Equation (17.1)

In the first $K$ rounds, we select each arm in a round-robin fashion as we do need to see at least one sample of each. After this initial period, our critical observation is that the suboptimal arm $a$ can *only* be pulled on round $t$ if one of the following "bad events" occurs:

- The UCB index of arm $a$ turns out to be *larger* than the optimal mean value, i.e. $\mathsf{UCB}(a, t) > \mu^*$.

- The UCB index of the optimal arm $a^*$ turns out to be *smaller* than its true mean value given by $\mu^*$, i.e. $\mathsf{UCB}(a^*, t) < \mu^*$

What happens if neither of these events is true? It will simply mean that $\mathsf{UCB}(a^*, t) > \mathsf{UCB}(a, t)$ and arm $a$ *cannot* be picked on that round. To visualize why this is the case, see Figure 17.2. You can see from this figure that if neither of the "bad events" occurs, we have

$$\mathsf{UCB}(a, t) < \mu^* < \mathsf{UCB}(a^*, t), \tag{17.4}$$

and so arm $a$ will not be picked at round $t$.

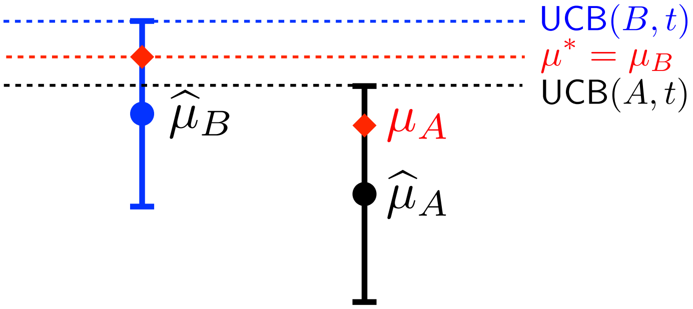

Figure 17.2: Depiction of the "good event" that occurs when both Properties 1 and 2 hold. This is a schematic of our drug discovery example where drug B is better than drug A, thus $a^* = B$. First, we see that Property 1 holds as $\mu^* = \mu_B < \mathsf{UCB}(B,t)$ (i.e. the red line is below the blue line). Next, we see that Property 2 holds as $\mathsf{UCB}(A,t) < \mu^*$, i.e. the black line is below the red line. It is easy to see that therefore, the black line is below the blue line and drug B will be selected under this "good event".

Finally, it will be convenient to also notate the UCB indexes according to the $n^{th}$ time the arm is sampled. In particular, let $\tau_{a,n}$ denote the epoch at which arm $a$ is sampled for the $n^{th}$ time (note that this epoch will be, in general, random). Then, we write

$$\mathsf{UCB}_n(a) := \mathsf{UCB}(a, \tau_n) = \widehat{\mu}_a(n) + \sqrt{\frac{\log(T^2)}{2n}},$$

where we define $\widehat{\mu}_a(n)$ to be the sample mean accrued from $n$ iid samples of reward from arm $a$. Note that we have also substituted the value of $\delta := 1/T^2$.

### 17.2.1 The two "good properties" in math

We now make idea precise. We will show here that the pseudo-regret will be very low when *both* of the following good events hold:

**Property 1:** The UCB index of the optimal arm $a^*$ is *always* greater than its true mean value $\mu^*$; in other words, we define $\mathcal{A}_1$ to be the event that

$$\mathsf{UCB}_n(a^*) > \mu^* \text{ for all } n = 1, \ldots, T.$$

To get some intuition for why we may hope for this property to be true, recall that $\mathsf{UCB}_n(a^*) = \widehat{\mu}_{a^*}(n) + \sqrt{\frac{\log(T^2)}{2n}} = \widehat{\mu}_{a^*}(n) + \sqrt{\frac{\log T}{n}}$. For large values of $n$, we would hope for $\widehat{\mu}_{a^*} \approx \mu^*$ and for the inequality to be true; for small values of $n$, we would hope for the large value of the confidence width $\sqrt{\frac{\log T}{n}}$ to lead to the property being true anyway. It turns out that Property 1 holds with very high probability; in fact, at least $1 - 1/T$. We will use Hoeffding's lemma to show this in the next lecture.

**Property 2:** The UCB index of a suboptimal arm $a$ is smaller than the mean value of the *optimal* arm $\mu^*$ after arm $a$ has been sampled $n_a := \frac{4 \log T}{\Delta_a^2}$ times; in other words, we define $\mathcal{A}_2$ to be the event that

$$\mathsf{UCB}_n(a) < \mu^* \text{ for all } n = n_a, \ldots, T.$$

It turns out that Property 2 also holds with high probability; in fact, at least $1 - 1/T$. It is useful to think about why Property 2 can only be guaranteed once arm $a$ has been sampled sufficiently often. Remember your observations from the demo: initially, a suboptimal arm $a$ was sampled even if its reward was lower, either because its confidence width was very large *or* because its mean was overestimated from few samples. Once more samples are taken, you observed that both of these adverse effects disappeared: a) the confidence width will shrink as more samples are taken, and b) the sample means observed become closer to the true mean. It turns out that the value of $n_a$ is chosen carefully as the tipping point at which these adverse effects sufficiently disappear so that Property 2 holds.

It is now easy to see that when both Properties 1 and 2 hold, we have $N_a(T) \leq n_a$. On the other hand, if one of these properties *does not hold*, the worst-case number of times arm $a$ may be sampled is given by $T$. Thus, we can *upper-bound* the pseudo-regret by

$$\mathbb{E}[N_a(T)] \leq \mathbb{P}\left[\text{Properties 1 and 2 hold}\right] \cdot n_a + \mathbb{P}\left[\text{one of Properties 1 } \textit{or} \text{ 2 does not hold}\right] \cdot T.$$

The first term above is the "good one": $n_a := \frac{4 \log T}{\Delta_a^2}$, which matches the second term in Equation (17.3). It remains to bound the second term by showing that *both* Property 1 and 2 are very likely to hold. We will now show that the probability that either one of Properties 1 or 2 does not hold is at most $\frac{2}{T}$. Together with the fact that each arm must additionally be sampled at least once before any of this theory applies, this leads to the second term being equal to 3. This will complete our proof. It will turn out that we will use Hoeffding's lemma to do this, as we have assumed that the rewards are bounded between 0 and 1: we will do this and complete this proof at the beginning of next lecture.

## 17.3. Bibliographical notes

See (Lattimore and Szepesvári, 2020, Chapter 7) for excellent bibliographical notes. This proof appears there, but was originally conceived in this relatively accessible form by Auer et al. (2002) (original analyses by Lai and Robbins (1985); Agrawal (1995) are *asymptotic* in $T$ and much more complex, using subtle sequential statistics concepts).

## References

Rajeev Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

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

T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.