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

Lecture 16: October 25

Lecturer: Bhuvesh Kumar

Last class, we introduced the multi-armed bandit (MAB) problem in a formal framework, discussed exploration vs exploitation and discussed the metric of pseudo-regret at length.

16.1. Recap: The MAB problem, and pseudo-regret

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:

$$\overline{R}_T = \sum_{a \neq a^*} \Delta_a \cdot \mathbb{E}\left[N_a(T)\right], \tag{16.1}$$

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 lecture 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: Lecture 11, for example, showed an example for which $N_a(T)$ can be very different for different

16.1.1 Suboptimal Algorithms

Now, we quickly recall the suboptimal algorithms that we had introduced last lecture, and mention without proof their pseudo-regret guarantees. Homework 3 has you explore and show these guarantees for yourself in a hands-on manner.

- Greedy (only exploit): $A_t = \arg\max_a \widehat{\mu}_{a,t-1}$. This turns out to incur linear pseudo-regret in T because of the very bad and non-vanishing possibility of pulling the suboptimal arm all the time.
- Only explore: select all arms in a round-robin fashion. This can easily be verified to incur linear pseudo-regret as well.
- Explore-then-commit for T_0 rounds: the homework has you show that you can expect sub-linear regret with this, with some caveats.
- ϵ-greedy with a randomizing schedule given by $\epsilon_{tt\geq 1}$: the homework has you show that you can also expect sub-linear regret with this, but the guarantee may not be optimal.

16.2. The upper-confidence-bounding algorithm

ETC and ϵ-greedy both constitute heuristic approaches to trading off exploration and exploitation in our algorithm. While they make some headway in ensuring sublinear pseudoregret, they turn out to not be optimal. The reason, at a high level, is because they do not adapt the tradeoff of exploration-vs-exploitation to the properties of the data at hand. For example, it is intuitive to see that the case where Drug A is 20% effective and drug B is 70% effective will require more exploration than a hypothetical case in which Drug A is 10% effective and Drug B is 90% effective, simply because the former case has more randomness, and the mean efficacies of each drug are "closer" to one another — so we need to collect more samples of each to offset possible adverse effects due to randomness in the rewards.

We will now explore an algorithm that turns out to trade off exploration and exploitation in precisely this way. This is called the upper-confidence-bound algorithm, and is formally defined below.

Definition 1 *At round t, let $\widehat{\mu}_{a,t-1}$ denote the sample mean of arm a, and $N_{t-1}(a)$ denote the number of times that arm a was sampled until then. (Note that both of these are random variables owing to the random reward feedback.) Then, UCB selects the action A_t as:*

$$A_t = \arg\max_a \left[\widehat{\mu}_{a,t-1} + \sqrt{\frac{\log(1/\delta)}{2N_{t-1}(a)}} \right]. \tag{16.2}$$

We denote the RHS of the above for each a as $\mathsf{UCB}(a,t)$ as shorthand; then, we have $A_t = \arg\max_a \mathsf{UCB}(a,t)$. δ 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.

16.2.1 Mathematical principle: incentivizing exploration and exploitation

It is instructive to look at Equation 16.2 and ask what factors would lead to the objective becoming large for a given arm $a \in 1, \cdots, K$. Essentially, there are two factors:

- *Large sample mean:* the larger the values of, $\widehat{\mu}_{a,t-1}$, larger UCB would be. Thus the term $\widehat{\mu}_{a,t-1}$ encourages picking arms with a larger sample mean obtained thus far, i.e. this term encourages exploitation.
- *Small number of samples thus far:* Notice the inverse dependence on $N_{t-1}(a)$. If this is small, i.e. arm a has been sampled very few times until now, it makes sense to increase the objective to incentivize exploration. Thus, the term $\sqrt{\frac{2\log(1/\delta)}{N_{t-1}(a)}}$ encourages picking arms that have been pulled relatively infrequently thus far.

Clearly, the choice of δ crucially determines the operating point on the exploration exploitation tradeoff. It is instructive to consider two extremes:

- The case where $\delta = 1$, which yields the greedy algorithm.
- The case where $\delta = 0$, which can be verified to do pure exploration (as only the second term will always dominate)

Thus, larger values of would encourage exploitation, while lower values of would encourage exploration. We will now see a statistical interpretation of the value of δ through the notion of confidence intervals.

16.2.2 Cognitive principle: Optimism in the face of uncertainty

There is also an interesting cognitive principle at work with the UCB algorithm, which is the principle of *optimism in the face of uncertainty*: when we don't know much about an action, we take the *upper-confidence-bound*, rather than the *lower-confidence-bound*, in a display of optimism, i.e. we take an optimistic perspective on arms that we have seen very few times thus far. There is some preliminary evidence that humans tend to make their decisions in this way in the face of uncertainty (but this remains a matter of extensive debate among cognitive scientists and psychologists).

16.2.3 The upper-confidence principle and the meaning of δ

We now consider the value of the hyperparameter δ 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 μ_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.

This formalizes the notion of confidence intervals, and will be important for the proof technique that we will introduce in the next lecture. Moreover, it lends interpretational value to the hyperparameter δ of choice. If $\delta = 1$, all bets are off: the confidence widths shrink to 0, and we cannot guarantee anything! This constitutes the overly high-risk "greedy" approach that purely uses the sample means. On the other hand, if $\delta = 0$, we want to be excessively certain about the sample means and have confidence widths that accommodate the true means with probability 1! This will never be possible unless we make the widths arbitrarily large, and turns out to lead to a very conservative approach of over-exploring

16.2.4 Demonstration of UCB's Performance

Next lecture, we will see how to set this parameter δ, and also show a remarkable ability of the UCB algorithm to automatically tailor the trade-off between exploration and exploitation to the instance. We spend the rest of this lecture demonstrating the performance of UCB. One such snapshot is shown in Figure 16.1.

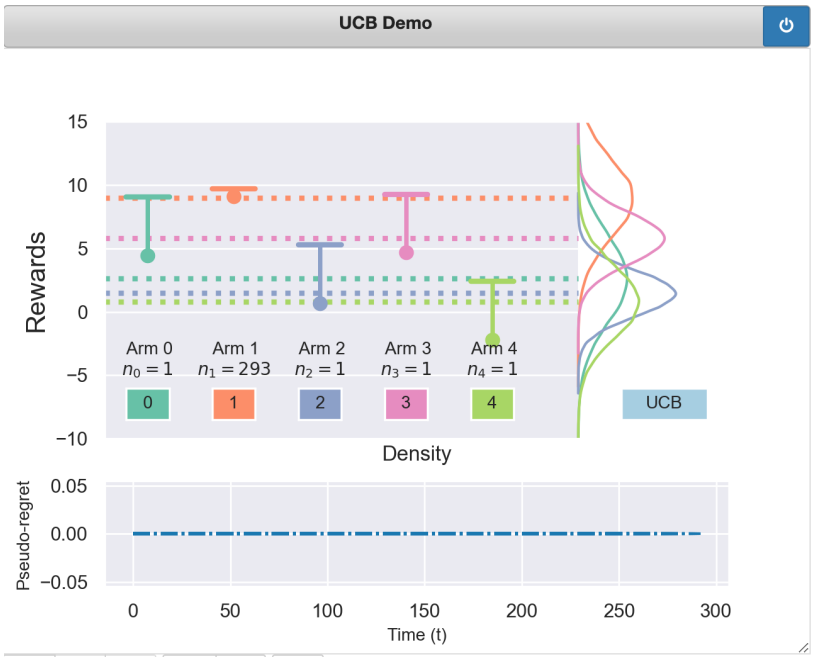

Figure 16.1: Snapshot of demo of UCB algorithm with parameter $\delta = 1/T^2$ (this choice is explained in this lecture) on a randomly chosen 5-armed bandit instance. Here, arm 1 is the best arm and is pulled disproportionately often by UCB. This demo is borrowed from the lectures in UC Berkeley's DS102 course, Fall 2019 iteration, and was originally created by graduate student Karl Krauth.

The demo shows consistently nice behavior with UCB: on all cases, it finds the optimal arm fairly quickly and while it does sample one or two of the suboptimal arms once in a while, it does so very rarely. It in fact turns out that we can show that the pseudo-regret of

UCB is given by

$$\overline{R}_T \leq \mathcal{O}\left(\sum_{a \neq a^*} \frac{\log T}{\Delta_a}\right)$$

where $\Delta_a := \mu^* - \mu_a$ denotes the suboptimality gap.

16.2.5 Properties of UCB that we will use

Now, how do we turn our observations into an analysis that reflects this? The discussion in Section 16.2.3 offers some initial hints: we will, repeatedly, use the fact that the sample means are concentrated around the true means in a way that depends on the number of times the arm has been sampled thus far. In particular, we notice that across all the random examples that we demonstrate in class, the following patterns present themselves:

We demonstrated the performance of UCB on various random examples.

In particular, we noticed the following two patterns across all the examples:

- For the optimal arm, the *true* mean μ^* 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 16.1 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 μ^*. This minimal number of times tends to depend on how big the gap was between μ^* and μ_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 16.1: 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 μ_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.

16.3. Proof of UCB pseudo-regret

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

Theorem 2 *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 2 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 ϵ-greedy algorithms, which turn out to incur $\mathcal{O}(T^{2/3})$ regret in the worst case.

We will now prove Theorem 2. 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

$$\mathbb{E}[N_a(T)] \leq 3 + \frac{4 \log T}{\Delta_a^2} \tag{16.3}$$

for each value of $a \neq a^*$. It is easily verified that plugging this upper bound into Equation (16.1)

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 μ^*, 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 16.2. You can see from this figure that if neither of the "bad events" occurs, we have

$$\mathsf{UCB}(a, t) < \mu^* < \mathsf{UCB}(a^*, t), \tag{16.4}$$

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

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 $\delta := 1/T^2$.

16.3.1 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 μ^*; in other words, we define $\mathcal{A}_1$ to be the event that

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

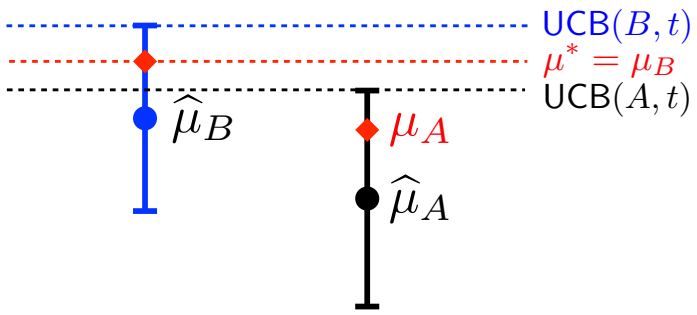

Figure 16.2: Depiction of the "good event" that occurs when both Properties 1 and 2 hold. This is a schematic of our drug discovery example where drug B is better than drug A, thus $a^* = B$. First, we see that Property 1 holds as $\mu^* = \mu_B < \mathsf{UCB}(B,t)$ (i.e. the red line is below the blue 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".

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 μ^* 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 (16.3). 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.

16.4. 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.