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

Lecture 18: Nov 1

Lecturer: Guanghui Wang

Last class, we dug in deeper into the *upper-confidence-bound* (UCB) algorithm, for which the action choice at round t is given by

$$A_t = \arg\max_a \left[\widehat{\mu}_{a,t-1} + \sqrt{\frac{\log(1/\delta)}{2N_{t-1}(a)}} \right]. \tag{18.1}$$

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, 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. We claimed last lecture that the choice $\delta := 1/T^2$ results in UCB having this nice pseudo-regret guarantee:

$$\overline{R}_T \leq 3 \sum_{a \neq a^*} \Delta_a + \sum_{a \neq a^*} \frac{4 \log T}{\Delta_a} = \mathcal{O}(\log T). \tag{18.2}$$

Last lecture, we identified two properties that, if UCB satisfied, would directly lead to this guarantee on pseudo-regret. Today, we will complete this understanding by showing that these two properties hold *with high probability* over the randomness in the realized rewards. We will also see that this performance of UCB is, in fact, the best that we can do. Finally, we will briefly discuss where the multi-armed bandit paradigm is used in practice today.

18.1. Recap: The two key properties of UCB

Recall that the pseudo-regret of *any* algorithm has the following intuitive expression:

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

where $N_a(T)$ denotes the number of times a suboptimal arm a was sampled and Δ_a denotes its suboptimality gap in reward with respect to the optimal arm a^*. Consequently, it suffices to show that UCB will satisfy the following upper bounds on the number of times each suboptimal arm is pulled, i.e.

$$\mathbb{E}[N_a(T)] \leq 3 + \frac{4 \log T}{\Delta_a^2} \text{ for all } a \neq a^*.$$

Last lecture, we essentially showed that if two "good events" held:

- **Property 1:** The *true* mean of the optimal arm lies within the confidence interval specified by UCB, i.e. $\mu^* \leq \mathsf{UCB}(a^*, t)$ on all rounds indexed by $t \geq 1$. Intuitively, this says that our confidence intervals should contain the true mean within them; we are not heavily *underestimating* the true best mean reward.
- **Property 2:** The upper-confidence-bound of a suboptimal arm a lies *below* the true mean of the optimal arm, i.e. $\mathsf{UCB}(a, t) \leq \mu_a$ as long as t is large enough such that $N_a(t) \geq n_a := \frac{4 \log T}{\Delta_a^2}$. Intuitively, this says that the confidence interval of a suboptimal a shrinks sufficiently, once it has been pulled sufficiently often, that it will lie below the true mean of the optimal arm.

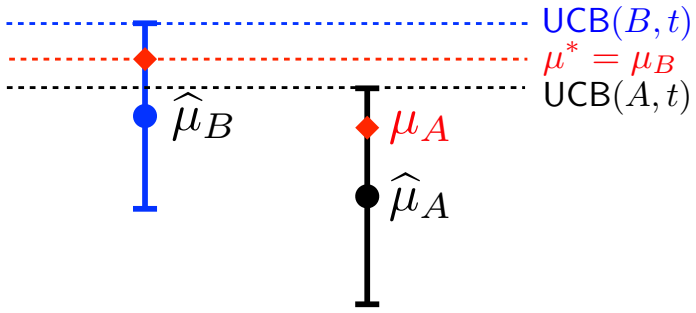

Figure 18.1: Depiction of the "good event" that occurs when both Properties 1 and 2 hold. This is a schematic of our drug discovery example where drug B is better than drug A, thus $a^* = B$. First, we see that Property 1 holds as $\mu^* = \mu_B < \mathsf{UCB}(B, t)$ (i.e. the red line is below the blue 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".

Figure 18.1 depicts these two properties, and shows that in this case the suboptimal arm a will not be pulled, as its UCB is dominated by that of a^*. We saw in the demo of UCB that these properties tended to hold throughout the execution of UCB. We then showed that if Properties 1 and 2 both hold, we can guarantee that $N_a(T) \leq \frac{4 \log T}{\Delta_a^2}$. On the other hand, if one of Properties 1 or 2 does not hold, "all bets are off"; nevertheless, each arm can be sampled at most T times. Consequently, we have

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

where we have included the additive factor of 1 to factor for the fact that each arm is pulled once at the very beginning. The second inequality uses the coarse upper bound

$\mathbb{P}[\text{Properties 1 and 2 hold}] \leq 1$. To complete our proof argument, we will now show that $\mathbb{P}[\text{one of Properties 1 } \textit{or } 2 \text{ does not hold}] \leq \frac{2}{T}$, and so both Properties 1 and 2 are very high-probability events when T is large. Plugging in this high-probability bound above completes our proof.

18.1.1 Showing that Property 1 holds (with high probability)

We will critically use the fact that the rewards (and therefore, their means) are bounded between 0 and 1 for this part of the proof. To understand why we should expect Property 1 to hold, we can then appeal to the Hoeffding bound. We recall that if arm a^* has been sampled n times, we can write the UCB index as $\mathsf{UCB}_n(a^*) := \widehat{\mu}_a(n) + \sqrt{\frac{\log(1/\delta)}{2n}}$. For our choice $\delta := 1/T^2$, this becomes equal to $\mathsf{UCB}_n(a^*) := \widehat{\mu}_{a^*}(n) + \sqrt{\frac{\log T}{n}}$. Essentially, for any number of arm pulls n, we require that

$$\mu^* \leq \mathsf{UCB}_n(a^*) := \widehat{\mu}_{a^*}(n) + \sqrt{\frac{\log T}{n}}$$

$$\implies \widehat{\mu}_{a^*}(n) \geq \mu^* - \sqrt{\frac{\log T}{n}}$$

holds with very high probability. Thus, the sample mean should not be *too much* below the actual mean of the optimal arm a^*. This is a situation that is apt for Hoeffding's lemma, substituting which gives us

$$\mathbb{P}\left[\widehat{\mu}_{a^*}(n) < \mu^* - \sqrt{\frac{\log T}{n}}\right] \leq \exp\left(-\frac{2n \cdot \log T}{n}\right) = \exp\left(-2\log T\right) = \frac{1}{T^2}$$

for each value of $n = 1, \ldots, T$. Taking a union bound over $n = 1, \ldots, T$ gives us that event $\mathcal{A}_1$ *does not* hold with probability at most $1/T$; therefore event $\mathcal{A}_1$ holds with probability at least $1 - 1/T$. Note that Property 1 is an "any-time" property, and so this last step should remind you of the any-time properties you showed for FTL in HW 1, Problem 2.

18.1.2 Showing that Property 2 holds (with high probability)

Essentially, we need to show that $\mathsf{UCB}_n(a) \leq \mu^*$ provided that $n \geq n_a$. First, we will show that the sample mean $\widehat{\mu}_a(n)$ does not heavily *overestimate* the true mean μ_a, i.e. we want

$$\widehat{\mu}_a(n) \leq \mu_a + \sqrt{\frac{\log T}{n}}$$

with high probability. This turns out to be the case; applying Hoeffding's lemma in the same way as above (but to the upper tail of the random variable $\widehat{\mu}_a(n)$ this time) gives us

$$\mathbb{P}\left[\widehat{\mu}_a(n) > \mu_a + \sqrt{\frac{\log T}{n}}\right] \leq \exp\left(-\frac{2n \cdot \log T}{n}\right) = \exp\left(-2\log T\right) = \frac{1}{T^2}.$$

Thus, as before, the event that $\widehat{\mu}_a(n) \leq \mu_a + \sqrt{\frac{\log T}{n}}$ for all $n = 1, \ldots, T$ holds with probability at least $1 - 1/T$. Therefore, for all $n \geq n_a$, we will have $\widehat{\mu}_a(n) \leq \mu_a + \sqrt{\frac{\log T}{n}}$, and so $\mathsf{UCB}_n(a) \leq \mu_a + 2\sqrt{\frac{\log T}{n}} = \mu_a + \sqrt{\frac{4\log T}{n}}$.

To show further that $\mathsf{UCB}_n(a) \leq \mu^*$ under this event, we need the number of pulls n to be sufficiently large so that the confidence width term, $\sqrt{\frac{4\log T}{n}}$, does not overtake the suboptimality gap $\mu^* - \mu_a := \Delta_a$. In other words, we need:

$$\sqrt{\frac{4\log T}{n}} \leq \Delta_a$$
$$\implies n \geq n_a := \frac{4\log T}{\Delta_a^2}.$$

Thus, as soon as we cross the threshold of rounds at which $n \geq n_a$, we will get

$$\mathsf{UCB}_n(a) = \mu_a + \sqrt{\frac{4\log T}{n}} \leq \mu_a + \Delta_a =: \mu^*,$$

which is exactly Property 2. This shows that Property 2 *does not* hold with property at most $1/T$, just like Property 1. Taking a union bound over both properties yields that one of Properties 1 or 2 is violated with very low probability, at most $2/T$. This is exactly what we needed, and it completes our proof of UCB!

18.1.3 Technical notes and subleties

I will point out two technical matters to wrap up our discussion of UCB. First, the proof as described above does not absolutely require the rewards to be bounded between 0 and 1. One simple extension is to consider a more general interval $[a, b]$, in which case all of the Hoeffding bounds (and therefore confidence intervals used in UCB) will scale with the range $b - a$. More pertinently, we only needed an *exponential tail bound* of the form of Hoeffding's lemma to hold to prove Properties 1 and 2 hold with high probability. Such a tail bound can be obtained, e.g. for Gaussian rewards (as we saw in the introductory review lecture); but also for many other reward distributions that need not be bounded. A random variable that satisfies such an exponential tail bound is called a *sub-Gaussian* random variable; Gaussian and bounded rewards are two classic examples, but there are many others. See the bibliographical notes for details on these types of random variables.

Second, it is important to note the re-indexing that we did to define the UCB's $\mathsf{UCB}_n(a)$ for n pulls of arm a. This is to ensure that we can apply Hoeffding's lemma *with n fixed, not being a random variable itself*, and then take a union bound over all $n = 1, \ldots, T$. A more direct approach could have been to try and upper bound, e.g. the probability $\mathbb{P}\left[\widehat{\mu}_{a,t} > \mu_a + \sqrt{\frac{\log T}{N_a(t)}}\right]$. However, we cannot in general do this, as $N_a(t)$ is itself a random variable and its values are itself influenced by the past observed reward realizations (i.e. they are not independent). There are situations in which these more sophisticated probabilities can be characterized, but these use more advanced sequential statistics theory (notably, the theory of martingale sequences) and will not be touched upon in depth in this class.

18.2. Can we do better? A lower bound sketch

It turns out that UCB is, in a certain sense, *optimal* for the multi-armed bandit problem. The intuition is that any algorithm actually *needs* to take those n_a samples of each suboptimal

arm a in order to distinguish whether it is in fact optimal or not. This would then lead to a *lower bound* on the pseudo-regret of *any* algorithm meant for the multi-armed bandit problem, constituting a fundamental limit on how well we can do[1].

This type of a lower bound is subtle and its formal proof uses advanced techniques from information theory. I will provide a recording of this proof that is optional viewing. In this lecture, we expand (informally) on how this intuition leads to a lower bound, and how to define a lower bound in the first place. For simplicity, we provide the intuition and main idea for the 2-armed case, but all of the ideas extend to the more general K-armed case.

18.2.1 The main idea: hypothesis testing

Let us consider our drug discovery example with drugs A and B; the MAB problem wants us to identify and use the better drug for most of the patients, and minimizing pseudo-regret enables us to achieve this. Suppose that Drug A has efficacy μ_A and Drug B has efficacy μ_B; define $\Delta := |\mu_A - \mu_B|$. Then, the result on UCB tells us that we can guarantee pseudo-regret at most $3\Delta + \frac{4 \log T}{\Delta}$, and the leading factor is $\frac{4 \log T}{\Delta}$.

We will now reason about why Drugs A and B have to be tested out some minimal number of times to be distinguished from one another. To make this concrete, let us fix the efficacy of Drug B to be $\mu_B = 0.4$, and consider two candidate scenarios for the efficacy of Drug A:

- **Null hypothesis:** Efficacy of Drug A, $\mu_A = 0.2$. In this scenario, Drug B is the optimal drug and $\Delta = 0.4 - 0.2 = 0.2$.

- **Alternative hypothesis:** Efficacy of Drug B, $\mu'_A = 0.6$. In this scenario, Drug A is the optimal drug and $\Delta = 0.6 - 0.4 = 0.2$.

In aggregate, our two candidate instances (or *hypotheses*) are given by $\boldsymbol{\mu} := (\mu_A = 0.2, \mu_B = 0.4)$ and $\boldsymbol{\mu} := (\mu'_A = 0.6, \mu_B = 0.4)$. Importantly, note that both of them have the same suboptimality gap Δ. What we notice about these candidate instances is that they have the same mean reward behavior for Drug B; but in one case Drug A is worse and in the other case Drug A is better. Moreover, no matter how many times we sample Drug B, we will get *no useful information* to test these two scenarios, as Drug B has the same mean reward in both cases!

Therefore, even if the true scenario is the "null hypothesis", where Drug A is less effective, we need to sample that suboptimal drug A a minimal number of times to be able to rule out the alternative hypothesis that Drug A is the more effective drug. We saw that $\mathcal{O}\left(\frac{\log T}{\Delta^2}\right)$ samples were *sufficient* to ensure that the alternative hypothesis is ruled out with high probability over our collected observations; it turns out that this many samples is also *necessary* (for suitable values of μ_A, μ_B that are bounded away from $\{0, 1\}$). Essentially, this is because the Hoeffding-type bounds we have used are actually tight and can be matched by lower bounds, as I alluded to in early lectures (but we do not prove this here).

1. A different example of such a fundamental limit was contained in HW 2, Problem 3 for the OLO problem. You saw there that for any algorithm, there existed an adversary that forced $\mathcal{O}(\sqrt{T})$ regret.

18.2.2 How to formalize an "information-theoretic" lower bound

It is crucial to note that the above reasoning works primarily because we need our algorithm to work effectively *under both the null and alternative hypothesis*. Of course, if we only needed our algorithm to work under scenarios of the null-hypothesis type, even if we did not know Δ beforehand, we would never need to test Drug A out; we could run with Drug B at the very beginning! Conversely, if we only needed to cover the alternate-hypothesis scenario, we would never need to test Drug B out. This means that there certainly exist algorithms that would do much better than $\mathcal{O}(\log T/\Delta)$ *for null scenarios and alternate scenarios respectively*. However, they would do *much* worse than $\mathcal{O}(\log T/\Delta)$ for the scenario for which they are not intended[2], making them significantly less appealing[3].

Consequently, lower bounds are defined over algorithms that guarantee "reasonable" performance over all possible situations. We denote an algorithm by $\mathcal{A}$, and its incurred regret on an instance $\boldsymbol{\mu}$ by $\overline{R}_T(\mathcal{A};\boldsymbol{\mu})$. Then, an informal definition of algorithms satisfying "reasonable" performance is given by

$$\mathbb{A} := \{\mathcal{A} : \overline{R}_T(\mathcal{A};\boldsymbol{\mu}) = \mathcal{O}(T^p) \text{ for all instances } \boldsymbol{\mu} \text{ and for any value of } p > 0 \ \}.$$

This is a weaker requirement than the logarithmic regret property that we just showed; note, for example, that UCB satisfies this property. (Clearly, algorithms that are not in $\mathbb{A}$ will have worse lower bounds (i.e. incur more pseudo-regret) than the ones that we will provide now.)

A typical lower-bound statement is then described below.

Theorem 1 (Informal, stated for 2 arms) *Consider any algorithm $\mathcal{A} \in \mathbb{A}$. Then, for* any *instance $\boldsymbol{\mu}$ we have*

$$\lim_{T\to\infty} \frac{\overline{R}_T(\mathcal{A};\boldsymbol{\mu})}{\log T} \geq \frac{1}{\Delta},$$

where we defined $\Delta := |\mu_1 - \mu_2|$.

We end this informal discussion of lower-bound ideas with a final note: we have crucially used the fact that samples from one of the options (Drug B) gives us *no side information* about the other option (Drug A). This may be an overly pessimistic model in situations where we may have prior information about either of the options, or about how they impact one another. This motivates the design of a very different family of algorithms that naturally incorporates side information, popularly called *Thompson sampling*. Thompson sampling is significantly more difficult to analyze than UCB (and we will not be analyzing it), but it has a nice intuitive principle and often yields superior empirical performance to UCB. Both algorithms are widely used in practice. We will discuss Thompson sampling in detail next lecture.

18.3. Bibliographical notes

- Notice that the hyperparameter $\delta := 1/T^2$ in UCB requires knowledge of the horizon (number of rounds) T. It is possible to use a version of the "doubling trick" (which

2. Lattimore (2015) describes the ensuring "Pareto frontier" between these hypotheses in an appealing way.
3. The greedy algorithm is one example of such an algorithm: think of why!

you saw in HW 1, Problem 3) to set the hyperparameter adaptively without knowing the number of rounds T in advance. Moreover, it is possible to set more sophisticated choices of $\{\delta_t\}_{t\geq 1}$ to improve the leading constant in front of $\log T/\Delta_a^2$ to 1 rather than 4; for more discussion on this, see (Lattimore and Szepesvári, 2020, Chapter 8).

- Consider a 2-armed bandit problem; the upper bound $\log T/\Delta^2$ becomes vacuous as $\Delta \to 0$. On the other hand, in this situation, the pseudo-regret would be at most ΔT. Considering the *minimum* of the above two upper bounds, and doing a sweep over all values of Δ, leads to the so-called *minimax* upper bound on pseudo-regret given by $\overline{R}_T = \mathcal{O}(\sqrt{KT\log(T)})$. (The reason this is called a *minimax* upper bound is because it is worst-case over all possible reward means for the arms.) It is possible to remove the $\log T$ factor by yet another modification to the UCB algorithm with data-adaptive hyperparameters $\{\delta_t\}_{t\geq 1}$ for the confidence interval. The ensuing strategy is called "minimax optimal strategy in the stochastic case" (MOSS); see (Lattimore and Szepesvári, 2020, Chapter 9) for further details.

- In the case of Bernoulli rewards, a modification to the confidence intervals in UCCB that is *instance-dependent* can be made. Accordingly, the factor $\log T/\Delta^2$ can be improved to $\log T/D_{\mathsf{KL}}(\mu_2, \mu_1)$; where $D_{\mathsf{KL}}(\cdot,\cdot)$ denotes the KL-divergence between two probability distributions (which you will learn about if you take an information theory course). In situations where, for example, one of μ_1 or μ_2 is very close to $\{0, 1\}$, this turns out to yield a significant improvement over UCB, which is based on the Hoeffding bound. See (Lattimore and Szepesvári, 2020, Chapter 10) for more details on this algorithm, which is commonly called KL-UCB.

- Chapters 14 and 16 in Lattimore and Szepesvári (2020) contain the essential proof idea for the lower bound that we have sketched out above. If you have prior background in information theory, you may enjoy reading these chapters to understand how a lower bound proof is formally constructed: the above is only an informal discussion. Chapter 15 also contains a rudimentary introduction to information theory, but is not a substitute for a full course in the area.

- In class, we also discussed the successive elimination algorithm. In HW3, problem 3, you can derive the regret bound by your own.

References

Tor Lattimore. The pareto regret frontier for bandits. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 208–216, 2015.

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