8 The Upper Confidence Bound Algorithm: Asymptotic Optimality

Asymptotics is concerned with the behavior of an algorithm as the number of rounds gets large beyond any limits. Asymptotics is studied as a sanity check: We expect all algorithms to perform well in the face of a large amount of data. This of course leaves the question of how well we can expect algorithms to perform, a question that is looked at in detail in Part IV. Keen readers should feel free to peek ahead at the first few chapters of this part to get a sense of how this is done.

The topic of this chapter is a refinement of Algorithm 2 that has a dual goal: On the one hand, the refinement resolves the issue of knowing the horizon in advance, while, on the other hand, it also makes the algorithm achieve the best possible asymptotic performance in a sense discussed in Chapter 16. Algorithm 5 gives the pseudocode of the refined algorithm.

Input K
Choose each arm once
Subsequently choose

$$A_t = \operatorname{argmax}_i \left(\hat{\mu}_i(t-1) + \sqrt{\frac{2 \log f(t)}{T_i(t-1)}} \right)$$

where $f(t) = 1 + t \log^2(t)$

Algorithm 5: Asymptotically optimal UCB

The regret bound for Algorithm 5 is more complicated than what we presented for Algorithm 2 (see Theorem 7.1). The important thing is that the dominant terms have the same order, but the constant multiplying the dominant term is smaller. The significance of this is that the long-term behavior of the algorithm is controlled by this constant, so if the long term behavior is interesting, it is worth putting the effort into reducing this constant.

THEOREM 8.1 *For any 1-subgaussian bandit, the regret of Algorithm 5 satisfies*

$$R_n \leq \sum_{i:\Delta_i > 0} \inf_{\varepsilon \in (0, \Delta_i)} \Delta_i \left(1 + \frac{5}{\varepsilon^2} + \frac{2}{(\Delta_i - \varepsilon)^2} \left(\log f(n) + \sqrt{\pi \log f(n)} + 3 \right) \right) . \tag{8.1}$$

Furthermore,

$$\limsup_{n\to\infty} \frac{R_n}{\log(n)} \leq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i} \,. \tag{8.2}$$

Before the proof, let us present a simpler version of the above bound, avoiding all these epsilons and infimums that make for a confusing theorem statement. By choosing $\varepsilon = \Delta_i/2$ we see that the regret of Algorithm 5 is bounded by

$$R_n \leq \sum_{i:\Delta_i>0} \left(\Delta_i + \frac{1}{\Delta_i} \left(8 \log f(n) + 8\sqrt{\pi \log f(n)} + 44 \right) \right) . \tag{8.3}$$

Even more concretely, there exists some universal constant $C > 0$ such that

$$R_n \leq C \sum_{i:\Delta_i>0} \left(\Delta_i + \frac{\log(n)}{\Delta_i} \right) ,$$

which by the same argument as in the proof of Theorem 7.2 leads a worst-case bound of $R_n \leq C \sum_{i=1}^K \Delta_i + 2\sqrt{CnK \log(n)}$.

Taking the limit of the ratio of the bound in (8.3) and $\log(n)$ does not result in the same constant as in the theorem, which is the main justification for introducing the epsilons in the first place. We shall see in Chapter 15 that the asymptotic bound on the regret given in (8.2), which is derived from (8.1) by choosing $\varepsilon = \log^{-1/4}(n)$, is unimprovable in a strong sense.

We start with a useful lemma that helps us bound the number of times the index of a suboptimal arm will be larger than some thresh-hold above its mean.

LEMMA 8.1 *Let $X_1, X_2, \ldots$ be a sequence of independent 1-subgaussian random variables, $\hat{\mu}_t = \frac{1}{t}\sum_{s=1}^t X_s$, $\varepsilon > 0$ and*

$$\kappa = \sum_{t=1}^n \mathbb{I}\left\{ \hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \varepsilon \right\} , \qquad \kappa' = u + \sum_{t=\lceil u \rceil}^n \mathbb{I}\left\{ \hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \varepsilon \right\} ,$$

where $u = 2a\varepsilon^{-2}$. Then it holds $\mathbb{E}[\kappa] \leq \mathbb{E}[\kappa'] \leq 1 + \frac{2}{\varepsilon^2}(a + \sqrt{\pi a} + 1)$.

The intuition for this result is as follows. Since the X_i are 1-subgaussian and independent we have $\mathbb{E}[\hat{\mu}_t] = 0$, so we cannot expect $\hat{\mu}_t + \sqrt{2a/t}$ to be smaller than ε until t is at least $2a/\varepsilon^2$. The lemma confirms that this is indeed of the right order as an estimate for $\mathbb{E}[\kappa]$.

Proof By Corollary 5.1 we have

$$\mathbb{E}[\kappa] \le \mathbb{E}[\kappa'] = u + \sum_{t=\lceil u \rceil}^{n} \mathbb{P}\left(\hat{\mu}_t + \sqrt{\frac{2a}{t}} \ge \varepsilon\right) \le u + \sum_{t=\lceil u \rceil}^{n} \exp\left(-\frac{t\left(\varepsilon - \sqrt{\frac{2a}{t}}\right)^2}{2}\right)$$
$$\le 1 + u + \int_u^{\infty} \exp\left(-\frac{t\left(\varepsilon - \sqrt{\frac{2a}{t}}\right)^2}{2}\right) dt = 1 + \frac{2}{\varepsilon^2}(a + \sqrt{\pi a} + 1)$$

as required. □

Proof of Theorem 8.1 As usual, we start with the basic regret decomposition.

$$R_n = \sum_{i:\Delta_i > 0} \Delta_i \mathbb{E}[T_i(n)]\,.$$

The rest of the proof revolves around bounding $\mathbb{E}[T_i(n)]$. Let i be the index of some sub-optimal arm (so that $\Delta_i > 0$). The main idea is to decompose $T_i(n)$ into two terms. The first measures the number of times the index of the optimal arm is less than $\mu_1 - \varepsilon$. The second term measures the number of times that $A_t = i$ and its index is larger than $\mu_1 - \varepsilon$.

$$\begin{aligned} T_i(n) &= \sum_{t=1}^{n} \mathbb{I}\{A_t = i\} \\ &\le \sum_{t=1}^{n} \mathbb{I}\left\{\hat{\mu}_1(t-1) + \sqrt{\frac{2\log f(t)}{T_1(t-1)}} \le \mu_1 - \varepsilon\right\} \\ &\quad + \sum_{t=1}^{n} \mathbb{I}\left\{\hat{\mu}_i(t-1) + \sqrt{\frac{2\log f(t)}{T_i(t-1)}} \ge \mu_1 - \varepsilon \text{ and } A_t = i\right\}. \end{aligned} \tag{8.4}$$

The proof of the first part of the theorem is completed by bounding the expectation

of each of these two sums. Starting with the first, we again use Corollary 5.1:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{I}\left\{\hat{\mu}_1(t-1)+\sqrt{\frac{2\log f(t)}{T_1(t-1)}}\le\mu_1-\varepsilon\right\}\right]\\
&\quad=\sum_{t=1}^{n}\mathbb{P}\left(\hat{\mu}_1(t-1)+\sqrt{\frac{2\log f(t)}{T_1(t-1)}}\le\mu_1-\varepsilon\right)\\
&\quad\le\sum_{t=1}^{n}\sum_{s=1}^{n}\mathbb{P}\left(\hat{\mu}_{1,s}+\sqrt{\frac{2\log f(t)}{s}}\le\mu_1-\varepsilon\right)\\
&\quad\le\sum_{t=1}^{n}\sum_{s=1}^{n}\exp\left(-\frac{s\left(\sqrt{\frac{2\log f(t)}{s}}+\varepsilon\right)^2}{2}\right)\\
&\quad\le\sum_{t=1}^{n}\frac{1}{f(t)}\sum_{s=1}^{n}\exp\left(-\frac{s\varepsilon^2}{2}\right)\le\frac{5}{\varepsilon^2}\,.
\end{aligned}$$

The first inequality follows from the union bound over all possible values of $T_1(t-1)$. The last inequality is an algebraic exercise (cf. Exercise 8.1). The function $f(t)$ was chosen precisely so this bound would hold. If $f(t)=t$ instead, then the sum would diverge. Since $f(n)$ appears in the numerator below we would like f to be large enough that its reciprocal is summable and otherwise as small as possible. For the second term in (8.4) we use Lemma 8.1 to get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{I}\left\{\hat{\mu}_i(t-1)+\sqrt{\frac{2\log f(t)}{T_i(t-1)}}\ge\mu_1-\varepsilon\text{ and }A_t=i\right\}\right]\\
&\quad\le\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{I}\left\{\hat{\mu}_i(t-1)+\sqrt{\frac{2\log f(n)}{T_i(t-1)}}\ge\mu_1-\varepsilon\text{ and }A_t=i\right\}\right]\\
&\quad\le\mathbb{E}\left[\sum_{s=1}^{n}\mathbb{I}\left\{\hat{\mu}_{i,s}+\sqrt{\frac{2\log f(n)}{s}}\ge\mu_1-\varepsilon\right\}\right]\\
&\quad=\mathbb{E}\left[\sum_{s=1}^{n}\mathbb{I}\left\{\hat{\mu}_{i,s}-\mu_i+\sqrt{\frac{2\log f(n)}{s}}\ge\Delta_i-\varepsilon\right\}\right]\\
&\quad\le 1+\frac{2}{(\Delta_i-\varepsilon)^2}\left(\log f(n)+\sqrt{\pi\log f(n)}+1\right)\,.
\end{aligned}$$

The first part of the theorem follows by substituting the results of the previous two displays into (8.4). The second part follows by choosing $\varepsilon=\log^{-1/4}(n)$ and taking the limit as n tends to infinity. □

8.1 Notes

1 The improvement to the constants comes from making the confidence interval slightly smaller, which is made possible by a more careful analysis. The main trick is the observation that we do not need to show that $\hat{\mu}_{1,s} \geq \mu_1$ for all s with high probability, but instead that $\hat{\mu}_{1,s} \geq \mu_1 - \varepsilon$ for small ε. This idea buys quite a lot and we will see it repeatedly in subsequent chapters.

2 The choice of $f(t) = 1 + t\log^2(t)$ looks quite odd. As we pointed out in the proof, things would not have gone through had we chosen $f(t) = t$. With a slightly messier calculation we could have chosen $f(t) = t\log^\alpha(t)$ for any $\alpha > 0$. If the rewards are actually Gaussian, then a more careful concentration analysis allows one to choose $f(t) = t$ or even some slightly slower growing function [Katehakis and Robbins, 1995, Lattimore, 2016a, Garivier et al., 2016b].

8.2 Bibliographic remarks

Lai and Robbins [1985] designed policies for which Eq. (8.2) held and proved lower bounds showing that no 'reasonable' policy can improve on this bound for any problem, where 'reasonable' means that they suffer subpolynomial regret on all problems. We will discuss these issues in great detail in Part IV where we address lower bounds. The policy proposed by Lai and Robbins [1985] was based on upper confidence bounds, but was not a variant of UCB. The asymptotics for variants of the policy presented here were given first by Katehakis and Robbins [1995] and Agrawal [1995]. Neither of these articles gave finite-time bounds like what was presented here. When the reward distributions lie in an exponential family, then asymptotic and finite-time bounds with the same flavor to what is presented here are given by Cappé et al. [2013]. There are now a huge variety of asymptotically optimal policies in a wide range of settings. Burnetas and Katehakis [1996] study the general case and give conditions for a version of UCB to be asymptotically optimal. Honda and Takemura [2010, 2011] analyze an algorithm called DMED to derive asymptotic optimality for noise models where the support is bounded or semi-bounded. Kaufmann et al. [2012b] prove asymptotic optimality for Thompson sampling (see Chapter 35) when the rewards are Bernoulli, which is generalized to single parameter exponential families by Korda et al. [2013]. Kaufmann [2018] proves asymptotic optimality for the Bayes UCB class of algorithms for single parameter exponential families. Ménard and Garivier [2017] prove asymptotic optimality and minimax optimality for exponential families (more discussion in Chapter 9).

8.3 Exercises

8.1 [Do the algebra needed at the end of the proof of Theorem 8.1] Show that

$$\sum_{t=1}^{n} \frac{1}{f(t)} \sum_{s=1}^{n} \exp\left(-\frac{s\varepsilon^2}{2}\right) \leq \frac{5}{\varepsilon^2}\,,$$

where $f(t) = 1 + t\log^2(t)$.

First bound $F = \sum_{s=1}^{n} \exp(-s\varepsilon^2/2)$ using a geometric series. Then show that $\exp(-a)/(1-\exp(-a)) \leq 1/a$ holds for any $a > 0$ and conclude that $F \leq \frac{2}{\varepsilon^2}$. Finish by bounding $\sum_{t=1}^{n} 1/f(t)$ using the fact that $1/f(t) \leq 1/(t\log(t)^2)$ and bounding a sum by an integral.

8.2 [One-armed bandits and UCB] Consider the one-armed bandit problem from Exercise 4.9. Notice that this one-armed bandit problem can be formulated as a regular bandit with two actions in which the first action corresponds to playing the machine and the second to not playing it. The noise in this case is 1-subgaussian, which means that the theoretical guarantees of UCB are applicable. For $p = 1$, evaluate

$$\limsup_{n\to\infty} \frac{R_p^{\text{UCB}}(n)}{\log(n)}\,.$$

8.3 [Continuation of Exercise 8.2] The difference between the one and two-armed bandit is that for one-armed bandits the mean of the second arm is known. This additional information is not exploited by UCB. However, we can incorporate this additional information into the definition of UCB as follows: Let $f(t) = 1 + t\log^2(t)$ and define a policy by

$$A_t = \begin{cases} 1, & \text{if } \hat{\mu}_1(t-1) + \sqrt{\frac{2\log f(t)}{T_1(t-1)}} \geq 0\,; \\ 2, & \text{otherwise}\,. \end{cases} \tag{8.5}$$

Prove that the modified UCB algorithm satisfies:

$$\limsup_{n\to\infty} \frac{R_p^{\text{MODIFIED-UCB}}(n)}{\log(n)} \leq \begin{cases} 0, & \text{if } p \geq 1/2; \\ \frac{2}{1-2p}, & \text{if } p < 1/2\,. \end{cases}$$

(**Hint:** Follow the analysis that we gave for UCB, but carefully adapt the proof by using the fact that the index of the second arm is always 0. This will leave you with a finite-time regret guarantee for the modified UCB from which the identity above can be derived.)

8.4 [Continuation of Exercise 8.3] The purpose of this question is to compare UCB and the modified version in (8.5).

(a) Implement a simulator for the one-armed bandit problem and two algorithms. UCB and the modified version analysed in Exercise 8.3.
(b) Use your simulator to estimate the expected regret of each algorithm for a horizon of $n = 1000$ and $p \in \{0, 1/20, 2/20, \ldots, 19/20, 1\}$.
(c) Plot your results with p on the x-axis and the estimated expected regret on the y-axis. Don't forget to label the axis and include error bars and a legend.
(d) Explain the results. Why do the curves look the way they do?
(e) In your plot, for what values of p does the worst-case expected regret for each algorithm occur? What is the worst-case expected regret for each algorithm?

8.5 Let $\sigma^2 \in [0, \infty)^K$ be known and suppose that the reward is $X_t \sim \mathcal{N}(\mu_{A_t}, \sigma^2_{A_t})$. Design an algorithm (that depends on σ^2) for which the asymptotic regret is

$$\limsup_{n\to\infty} \frac{R_n}{\log(n)} = \sum_{i:\Delta_i>0} \frac{2\sigma_i^2}{\Delta_i}\,.$$