9 The Upper Confidence Bound Algorithm: Minimax Optimality (†)

We proved that the variants of UCB analyzed in the last two chapters have a worst case regret of $R_n = O(\sqrt{Kn\log(n)})$. The factor of $\sqrt{\log(n)}$ can be removed by modifying the confidence level of the algorithm. The directly named Minimax Optimal Strategy in the Stochastic case algorithm (MOSS) was the first to make this modification and is presented below. MOSS again depends on prior knowledge of the horizon, a requirement that may be relaxed as we explain in the notes.

The term **minimax** is used because, except for constant factors, the worst case bound proven in this chapter cannot be improved on by any algorithm. The lower bounds are deferred to Part IV.

Input n and K
Choose each arm once
Subsequently choose

$$A_t = \operatorname{argmax}_i \hat{\mu}_i(t-1) + \sqrt{\frac{4}{T_i(t-1)} \log^+\left(\frac{n}{KT_i(t-1)}\right)},$$

where $\log^+(x) = \log\max\{1, x\}$.

Algorithm 6: MOSS

THEOREM 9.1 *For any* 1*-subgaussian bandit, the regret of Algorithm 6 is bounded by*

$$R_n \leq 34\sqrt{Kn} + \sum_{i=1}^{K} \Delta_i \,.$$

Before the proof we state and prove a strengthened version of Corollary 5.1.

THEOREM 9.2 *Let* $X_1, X_2, \ldots, X_n$ *be a sequence of independent* 1*-subgaussian random variables and* $S_t = \sum_{s=1}^{t} X_s$*. Then,*

$$\mathbb{P}\left(\textit{exists } t \leq n : S_t \geq \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2n}\right) . \tag{9.1}$$

The bound in Eq. (9.1) is the same as the bound on $\mathbb{P}(S_n \geq \varepsilon)$ that appears in a simple reformulation of Corollary 5.1, so this new result is strictly stronger.

Proof From the definition of subgaussian random variables and Lemma 5.2,

$$\mathbb{E}[\exp(\lambda S_n)] \leq \exp\left(\frac{n\sigma^2\lambda^2}{2}\right).$$

Then choosing $\lambda = \varepsilon/(n\sigma^2)$ leads to

$$\begin{aligned}\mathbb{P}(\text{exists } t \leq n : S_t \geq \varepsilon) &= \mathbb{P}\left(\max_{t\leq n} \exp(\lambda S_t) \geq \exp(\lambda\varepsilon)\right) \\ &\leq \frac{\mathbb{E}[\exp(\lambda S_n)]}{\exp(\lambda\varepsilon)} \leq \exp\left(\frac{n\sigma^2\lambda^2}{2} - \lambda\varepsilon\right) = \exp\left(-\frac{\varepsilon^2}{2n\sigma^2}\right).\end{aligned}$$

The novel step is the first inequality, which follows from the maximal inequality (Theorem 3.5) and the fact that $\exp(\lambda S_t)$ is a supermartingale with respect to the filtration generated by $X_1, X_2, \ldots, X_n$ (Exercise 9.1). □

Before the proof of Theorem 9.1 we need one more lemma to bound the probability that the index of the optimal arm ever drops too far below the actual mean of the optimal arm. The proof of this lemma relies on a tool called the **peeling device**, which is an important technique in probability theory and has many applications beyond bandits. For example, it can be used to prove the law of the iterated logarithm.

LEMMA 9.1 *Let $\delta \in (0,1)$ and $X_1, X_2, \ldots$ be independent and 1-subgaussian and $\hat{\mu}_t = \frac{1}{t}\sum_{s=1}^{t} X_s$. Then for any $\Delta > 0$,*

$$\mathbb{P}\left(\textit{exists } s \geq 1 : \hat{\mu}_s + \sqrt{\frac{4}{s}\log^+\left(\frac{1}{s\delta}\right)} + \Delta \leq 0\right) \leq \frac{16\delta}{\Delta^2}.$$

Proof Let $S_t = t\hat{\mu}_t$. Then

$$\begin{aligned}&\mathbb{P}\left(\text{exists } s \geq 1 : \hat{\mu}_s + \sqrt{\frac{4}{s}\log^+\left(\frac{1}{s\delta}\right)} + \Delta \leq 0\right) \\ &= \mathbb{P}\left(\text{exists } s \geq 1 : S_s + \sqrt{4s\log^+\left(\frac{1}{s\delta}\right)} + s\Delta \leq 0\right) \\ &\leq \sum_{k=0}^{\infty} \mathbb{P}\left(\text{exists } s \in [2^k, 2^{k+1}] : S_s + \sqrt{4s\log^+\left(\frac{1}{s\delta}\right)} + s\Delta \leq 0\right) \\ &\leq \sum_{k=0}^{\infty} \mathbb{P}\left(\text{exists } s \leq 2^{k+1} : S_s + \sqrt{4\cdot 2^k \log^+\left(\frac{1}{2^{k+1}\delta}\right)} + 2^k\Delta \leq 0\right) \\ &\leq \sum_{k=0}^{\infty} \exp\left(-\frac{\left(\sqrt{2^{k+2}\log^+\left(\frac{1}{2^{k+1}\delta}\right)} + 2^k\Delta\right)^2}{2^{k+2}}\right).\end{aligned}$$

In the first inequality we used the union bound, but rather than applying it on every time step as we did in the proof of Theorem 8.1, we apply it on a geometric grid. The second step is straightforward, but important because it sets up to apply Theorem 9.2. The rest is purely algebraic.

$$\sum_{k=0}^{\infty} \exp\left(-\frac{\left(\sqrt{2^{k+2}\log^+\left(\frac{1}{2^{k+1}\delta}\right)} + 2^k\Delta\right)^2}{2^{k+2}}\right) \leq \delta \sum_{k=0}^{\infty} 2^{k+1}\exp\left(-\Delta^2 2^{k-2}\right)$$
$$\leq \frac{8\delta}{\Delta^2} + \int_0^{\infty} 2^{s+1}\exp\left(-\Delta^2 2^{s-2}\right) ds \leq \frac{16\delta}{\Delta^2}\,,$$

where the first inequality follows since $(a+b)^2 \geq a^2 + b^2$ for $a, b \geq 0$ and the second last step follows by noting that the integrand is unimodal and has a maximum value of $8\delta/\Delta^2$. For such functions f one may bound $\sum_{k=a}^{b} f(k) \leq \max_{s\in[a,b]} f(s) + \int_a^b f(s)ds$. □

Proof of Theorem 9.1 As usual, we assume without loss of generality that the first arm is optimal, so $\mu_1 = \mu^*$. Define a random variable Δ that measures how far below the index of the optimal arm drops below its true mean.

$$\Delta = \left(\mu_1 - \min_{s\leq n}\left(\hat{\mu}_{1s} + \sqrt{\frac{4}{s}\log^+\left(\frac{n}{Ks}\right)}\right)\right)^+ .$$

Using the basic regret decomposition (Lemma 4.2) and splitting the actions based on whether or not their suboptimality gap is smaller or larger than 2Δ leads to

$$R_\nu(n) = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[T_i(n)] \leq 2n\Delta + \sum_{i:\Delta_i>2\Delta} \Delta_i \mathbb{E}[T_i(n)]$$
$$\leq 2n\Delta + 8\sqrt{Kn} + \sum_{i:\Delta_i>\max\left\{2\Delta, 8\sqrt{K/n}\right\}} \Delta_i \mathbb{E}[T_i(n)]\,.$$

The first term is easily bounded using Proposition 2.3 and Lemma 9.1.

$$\mathbb{E}[2n\Delta] = 2n\mathbb{E}[\Delta] = 2n\int_0^{\infty} \mathbb{P}\left(\Delta \geq x\right) dx \leq 2n\int_0^{\infty} \min\left\{1, \frac{16K}{nx^2}\right\} dx = 16\sqrt{Kn}\,.$$

For suboptimal arm i define

$$\kappa_i = \sum_{s=1}^{n} \mathbb{I}\left\{\hat{\mu}_{is} + \sqrt{\frac{4}{s}\log^+\left(\frac{n}{Ks}\right)} \geq \mu_i + \Delta_i/2\right\}\,.$$

The reason for choosing κ_i in this way is that for arms i with $\Delta_i > 2\Delta$ it holds that the index of the optimal arm is always larger than $\mu_i + \Delta_i/2$ so κ_i is an upper bound on the number of times arm i is played, $T_i(n)$. If $\Delta_i \geq 8(K/n)^{1/2}$,

then the expectation of $\Delta_i \kappa_i$ is bounded using Lemma 8.1 by

$$\begin{aligned}
\Delta_i \mathbb{E}[\kappa_i] &\leq \frac{1}{\Delta_i} + \Delta_i \mathbb{E}\left[\sum_{s=1}^{n} \mathbb{I}\left\{\hat{\mu}_{is} + \sqrt{\frac{4}{s}\log^+\left(\frac{n\Delta_i^2}{K}\right)} \geq \mu_i + \Delta_i/2\right\}\right] \\
&\leq \Delta_i + \frac{8}{\Delta_i}\left(2\log^+\left(\frac{n\Delta_i^2}{K}\right) + \sqrt{2\pi \log^+\left(\frac{n\Delta_i^2}{K}\right)} + 2\right) \\
&\leq \Delta_i + \sqrt{\frac{n}{K}}\left(2\log 8 + \sqrt{2\pi \log 8} + 2\right) \leq \Delta_i + 10\sqrt{\frac{n}{K}}\,,
\end{aligned}$$

where the first inequality follows by replacing the s in the logarithm with $1/\Delta_i^2$ and adding the $\Delta_i \times 1/\Delta_i^2$ correction term to compensate for the first Δ_i^{-2} rounds where this doesn't actually hold. Then we use Lemma 8.1 and the monotonicity of $x \to 1/x \log^+(ax^2)$ for $p \in [0,1]$ and $ax^2 \geq e^2$. The last inequality follows by naively bounding $2\log 8 + \sqrt{2\pi \log 8} + 2 \leq 10$. Then

$$\begin{aligned}
\sum_{i:\Delta_i > \max\left\{2\Delta, 8\sqrt{K/n}\right\}} \Delta_i \mathbb{E}[T_i(n)] &\leq \sum_{i:\Delta_i > \max\left\{2\Delta, 8\sqrt{K/n}\right\}} \Delta_i \mathbb{E}[\kappa_i] \\
&\leq \sum_{i:\Delta_i > \max\left\{2\Delta, 8\sqrt{K/n}\right\}} \left(\Delta_i + 10\sqrt{\frac{n}{K}}\right) \leq 10\sqrt{nK} + \sum_{i=1}^{K} \Delta_i\,.
\end{aligned}$$

Combining all the results we have $R_n \leq 34\sqrt{Kn} + \sum_{i=1}^{K} \Delta_i$. □

9.1 Notes

1 One may also prove an asymptotic upper bound on the regret of MOSS that is rather close to optimal. Specifically, one can show that

$$\limsup_{n\to\infty} \frac{R_n}{\log(n)} \leq \sum_{i:\Delta_i > 0} \frac{4}{\Delta_i}\,.$$

By modifying the algorithm slightly it is even possible to replace the 4 with a 2 and so recover the optimal asymptotic regret. The trick is to increase g slightly and replace the 4 in the exploration bonus by 2. The major task is then to re-prove Lemma 9.1, which is done by replacing the intervals $[2^k, 2^{k+1}]$ with smaller intervals $[\xi^k, \xi^{k+1}]$ where ξ is tuned subsequently to be fractionally larger than 1. This procedure is explained in detail by Garivier [2013]. When the reward distributions are actually Gaussian there is a more elegant technique that avoids peeling altogether (Exercise 9.4).

2 Although it is not obvious from the bounds proven in this chapter, all versions of MOSS can be arbitrarily worse than UCB in some regimes. This unpleasantness is hidden by both the minimax and asymptotic optimality criteria, which highlights the importance of fully finite-time upper and lower bounds. The counter-example witnessing the failure is quite simple. Let the rewards for all

arms be Gaussian with unit variance and $n = K^3$, $\mu_1 = 0$, $\mu_2 = -\sqrt{K/n}$ and $\mu_i = -1$ for all $i > 2$. From Theorem 8.1 we have that

$$R_n^{\text{UCB}} = O(K \log K)\,,$$

while it turns out that MOSS has a regret of

$$R_n^{\text{MOSS}} = \Omega(\sqrt{Kn}) = \Omega(K^2)\,.$$

A rigorous proof of this claim is quite delicate, but we encourage readers to try to understand why it holds intuitively.

3 The easy way to deal with this problem is to replace the index used by MOSS with a less aggressive confidence level.

$$\hat{\mu}_i(t-1) + \sqrt{\frac{4}{T_i(t-1)} \log^+\left(\frac{n}{T_i(t-1)}\right)}\,. \tag{9.2}$$

The resulting algorithm is never worse than UCB and you will show in Exercise 9.3 that it has has a distribution free regret of $O(\sqrt{nK \log(K)})$. An algorithm that does almost the same thing in disguise is called Improved UCB, which operates in phases and eliminates arms for which the upper confidence bound drops below a lower confidence bound for some arm [Auer and Ortner, 2010]. In practice this algorithm does not perform very well and it is not asymptotically optimal, but the analysis highlights the role of the confidence level in the regret.

4 Overcoming the weakness of MOSS without sacrificing minimax optimality is possible by using an adaptive confidence level that tunes the amount of optimism to match the instance. One of the authors has proposed two ways to do this using the following indices.

$$\hat{\mu}_i(t-1) + \sqrt{\frac{2(1+\varepsilon)}{T_i(t-1)} \log\left(\frac{n}{t}\right)}\,. \tag{9.3}$$

$$\hat{\mu}_i(t-1) + \sqrt{\frac{2}{T_i(t-1)} \log\left(\frac{n}{\sum_{j=1}^K \min\{T_i(t-1), \sqrt{T_i(t-1)T_j(t-1)}\}}\right)}\,.$$

The first of these algorithms is called the Optimally Confident UCB [Lattimore, 2015b] while the second is AdaUCB Lattimore [2018]. Both algorithms are minimax optimal up to constant factors and never worse than UCB. The latter is also asymptotically optimal. If the horizon is unknown, then AdaUCB can be modified by replacing n with t. It remains a challenge to provide a straightforward analysis for these algorithms.

5 There is a hidden cost of pushing too hard to reduce the expected regret, which is that the variance of the regret can grow significantly. We analyze this trade-off formally in a future chapter, but sketch the intuition here. Consider the two-armed case with suboptimality gap Δ and Gaussian noise. Then the

regret of a carefully tuned algorithm is approximately

$$R_n = O\left(n\Delta\delta + \frac{1}{\Delta}\log\left(\frac{1}{\delta}\right)\right),$$

where δ is a parameter of the policy that determines the likelihood that the optimal arm is misidentified. The choice of δ that minimizes the expected regret depends on Δ and is approximately $1/(n\Delta^2)$. With this choice the regret is

$$R_n = O\left(\frac{1}{\Delta}\left(1 + \log\left(n\Delta^2\right)\right)\right).$$

Of course Δ is not known in advance, but it can be estimated online so that the above bound is actually realizable by an adaptive policy that does not know Δ in advance (Exercise 9.3). The problem is that with the above choice the second moment of the regret will be at least $\delta(n\Delta)^2 = n$, which is uncomfortably large. On the other hand, choosing $\delta = (n\Delta)^{-2}$ leads to a marginally larger regret of

$$R_n = O\left(\frac{1}{\Delta}\left(\frac{1}{n} + \log\left(n^2\Delta^2\right)\right)\right).$$

The second moment for this choice, however, is $O(\log^2(n))$. A discussion of these issues, including empirical results, may also be found in the article by Audibert et al. [2007].

9.2 Bibliographic remarks

The MOSS algorithm is due to Audibert and Bubeck [2009], while an anytime modification is by Degenne and Perchet [2016]. The proof that a modified version of MOSS is asymptotically optimal may be found in the article by Ménard and Garivier [2017]. Optimally Confidence UCB and its friends are by one of the authors Lattimore [2015b, 2016b, 2018]. The idea to modify the confidence level has been seen in several places, with the earliest by Lai [1987] and more recently by Honda and Takemura [2010]. Kaufmann [2018] also used a confidence level like in Eq. (9.2) to derive an algorithm based on Bayesian upper confidence bounds.

9.3 Exercises

9.1 Let $X_1, X_2, \ldots, X_n$ be adapted to filtration $\mathbb{F} = (\mathcal{F}_t)_t$ with $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = 0$ almost surely. Prove that $M_t = \exp(\lambda \sum_{s=1}^{t} X_s)$ is a $\mathbb{F}$-supermartingale for any $\lambda \in \mathbb{R}$.

9.2 Let $\Delta_{\min} = \min_{i:\Delta_i > 0} \Delta_i$. Show there exists a universal constant $C > 0$

such that the regret of MOSS is bounded by

$$R_n \leq \frac{CK}{\Delta_{\min}} \log^+ \left(\frac{n\Delta_{\min}^2}{K} \right) + \sum_{i=1}^{K} \Delta_i \,.$$

9.3 Suppose we modify the index used by MOSS to be

$$\hat{\mu}_i(t-1) + \sqrt{\frac{4}{T_i(t-1)} \log^+ \left(\frac{n}{T_i(t-1)} \right)} \,.$$

(a) Show that for all 1-subgaussian bandits this new policy suffers regret at most

$$R_n \leq C \left(\sum_{i:\Delta_i > 0} \Delta_i + \frac{1}{\Delta_i} \log^+(n\Delta_i^2) \right) ,$$

where $C > 0$ is a universal constant.

(b) Under the same conditions as the previous part show there exists a universal constant $C > 0$ such that

$$R_n \leq C\sqrt{Kn \log(K)} + \sum_{i=1}^{K} \Delta_i \,.$$

(c) Repeat parts (a) and (b) using the index

$$\hat{\mu}_i(t-1) + \sqrt{\frac{4}{T_i(t-1)} \log^+ \left(\frac{t}{T_i(t-1)} \right)} \,.$$

9.4 Let $g(t) = at + b$ with $b > 0$ and

$$u(x,t) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2t} \right) - \frac{1}{\sqrt{2\pi t}} \exp\left(-2ab - \frac{(x-2b)^2}{2t} \right)$$

(a) Show that $u(x,t) > 0$ for $x \in (-\infty, g(t))$ and $u(x,t) = 0$ for $x = g(t)$.

(b) Show that $u(x,t)$ satisfies the heat equation:

$$\frac{\partial}{\partial t} u(x,t) = \frac{1}{2} \frac{\partial^2}{\partial x^2} u(x,t) \,.$$

(c) Let B_t be a standard Brownian motion, which for any fixed t has density with respect to the Lebesgue measure.

$$p_t(x) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2t} \right) .$$

Define $\tau_g = \min\{t : B_t = g(t)\}$ be the first time the Brownian motion hits the boundary. Put on your physicists hat (or work hard) to argue that

$$\mathbb{P}\left(\tau_g \geq t\right) = \int_{-\infty}^{g(t)} u(x,t) dx \,.$$

(d) Let $q_g(t)$ be the density of time τ with respect to the Lebesgue measure so that $\mathbb{P}(\tau_g \leq t) = \int_0^t q(t)dt$. Show that

$$q_g(t) = \frac{g(0)}{\sqrt{2\pi t^3}} \exp\left(-\frac{g(t)^2}{2t}\right)$$

(e) In the last part we established the exact density of the hitting time of a Brownian motion approaching a linear boundary. We now generalize this to nonlinear boundaries, but at the cost that now we only have a bound. Suppose that $f : [0, \infty) \to [0, \infty)$ is concave and differentiable and let $\lambda_t : \mathbb{R} \to \mathbb{R}$ be the tangent to f at t given by $\lambda_t(x) = f(t) + f'(t)(x - t)$. Let $\tau_f = \min\{t : B_t = f(t)\}$ and $q_f(t)$ be the density of τ_f. Show that

$$q_f(t) \leq q_{\lambda_t}(t)\,.$$

(f) Suppose that $X_1, X_2, \ldots$ is a sequence of independent standard Gaussian random variables. Show that

$$\mathbb{P}\left(\text{exists } t \leq \infty : \sum_{s=1}^{t} X_s \geq f(t)\right) \leq \int_0^n \frac{\lambda_t(0)}{\sqrt{2\pi t^3}} \exp\left(-\frac{f(t)^2}{2t}\right) dt\,.$$

(g) Let $h : (0, \infty) \to (1, \infty)$ be a concave monotone increasing function such that $\sqrt{\log(h(a))}/h(a) \leq c/a$ for constant $c > 0$ and $f(t) = \sqrt{2t \log h(1/t\delta)} + t\Delta$. Show that

$$\mathbb{P}\left(\text{exists } t \leq \infty : \sum_{s=1}^{t} X_s \geq f(t)\right) \leq \frac{2c\delta}{\sqrt{\pi}\Delta^2}\,.$$

(h) Show that $h(a) = 1 + (1 + a)\sqrt{\log(1 + a)}$ satisfies the requirements of the previous part with $c = 11/10$.

(i) Use your results to modify MOSS for the case when the rewards are Gaussian. Compare the algorithms empirically.

(j) Prove for your modified algorithm that

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

The above exercise has several challenging components and assumes prior knowledge of Brownian motion and its interpretation in terms of the heat equation. We recommend the book by Lerche [1986] as a nice reference on hitting times for Brownian motion against concave barriers. The equation you derived in part (d) is called the Bachelier-Levy formula and the technique for doing so is the method of images. The use of this theory in bandits was introduced by one of the authors [Lattimore, 2018], which readers might find useful when working through these questions.

9.5 In the last exercise you modified MOSS to show asymptotic optimality when the noise is Gaussian. This is also possible for subgaussian noise. Follow the advice in the notes of this chapter to adapt MOSS so that for all 1-subgaussian bandits it holds that

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

while maintaining the property that $R_n \leq C\sqrt{Kn}$ for universal constant $C > 0$.