7 The Upper Confidence Bound Algorithm

The upper confidence bound (UCB) algorithm offers several advantages over the explore-then-commit (ETC) algorithm introduced in the last chapter.

(a) It does not depend on advance knowledge of the suboptimality gaps.
(b) It behaves well when there are more than two arms.
(c) The version introduced here depends on the horizon n, but in the next chapter, we will see how to eliminate that as well.

The algorithm has many different forms, depending on the distributional assumptions on the noise. Like in the previous chapter, we assume the noise is 1-subgaussian. A serious discussion of other options is delayed until Chapter 10.

7.1 The Optimism Principle

The UCB algorithm is based on the principle of **optimism in the face of uncertainty**, which states that one should act as if the environment is as nice as **plausibly possible**. As we shall see in later chapters, the principle is applicable beyond the finite-armed stochastic bandit problem.

Imagine visiting a new country and making a choice between sampling the local cuisine or visiting a well-known multinational chain. Taking an optimistic view of the unknown local cuisine leads to exploration because without data, it could be amazing. After trying the new option a few times, you can update your statistics and make a more informed decision. On the other hand, taking a pessimistic view of the new option discourages exploration, and you may suffer significant regret if the local options are delicious. Just how optimistic you should be is a difficult decision, which we explore for the rest of the chapter in the context of finite-armed bandits.

For bandits, the optimism principle means using the data observed so far to assign to each arm a value, called the **upper confidence bound** that with high probability is an overestimate of the unknown mean. The intuitive reason why this leads to sublinear regret is simple. Assuming the upper confidence bound assigned to the optimal arm is indeed an overestimate, then another arm can only be played if its upper confidence bound is larger than that of the optimal arm, which in turn is larger than the mean of the optimal arm. And yet this cannot

happen too often because the additional data provided by playing a suboptimal arm means that the upper confidence bound for this arm will eventually fall below that of the optimal arm.

In order to make this argument more precise, we need to define the upper confidence bound. Let $(X_t)_{t=1}^n$ be a sequence of independent 1-subgaussian random variables with mean μ and $\hat{\mu} = \frac{1}{n}\sum_{t=1}^n X_t$. By Eq. (5.6),

$$\mathbb{P}\left(\mu \geq \hat{\mu} + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta \qquad \text{for all } \delta \in (0,1)\,. \tag{7.1}$$

When considering its options in round t, the learner has observed $T_i(t-1)$ samples from arm i and received rewards from that arm with an empirical mean of $\hat{\mu}_i(t-1)$. Then a reasonable candidate for 'as large as plausibly possible' for the unknown mean of the ith arm is

$$\text{UCB}_i(t-1,\delta) = \begin{cases} \infty & \text{if } T_i(t-1) = 0 \\ \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} & \text{otherwise}\,. \end{cases} \tag{7.2}$$

Great care is required when comparing (7.1) and (7.2) because in the former the number of samples is the constant n, but in the latter it is a random variable $T_i(t-1)$. By and large, however, this is merely an annoying technicality, and the intuition remains that δ is approximately an upper bound on the probability of the event that the above quantity is an underestimate of the true mean. More details are given in Exercise 7.1.

At last we have everything we need to state a version of the UCB algorithm, which takes as input the number of arms and the error probability δ.

Input k and δ
for $t \in 1, \ldots, n$ **do**
    Choose action $A_t = \text{argmax}_i \text{UCB}_i(t-1,\delta)$
    Observe reward X_t and update upper confidence bounds
end for

Algorithm 3: UCB(δ).

Although there are many versions of the UCB algorithm, we often do not distinguish them by name and hope the context is clear. For the rest of this chapter, we'll usually call UCB(δ) just UCB.

The value inside the argmax is called the **index** of arm i. Generally speaking, an **index algorithm** chooses the arm in each round that maximises some value (the index), which usually only depends on the current time step and the samples from that arm. In the case of UCB, the index is the sum of the empirical mean

of rewards experienced so far and the **exploration bonus**, which is also known as the **confidence width**.

Besides the slightly vague 'optimism guarantees optimality or learning' intuition we gave before, it is worth exploring other intuitions for the choice of index. At a very basic level, an algorithm should explore arms more often if they are (a) promising because $\hat{\mu}_i(t-1)$ is large or (b) not well explored because $T_i(t-1)$ is small. As one can plainly see, the definition in Eq. (7.2) exhibits this behaviour. This explanation is not completely satisfying, however, because it does not explain why the form of the functions is just so.

A more refined explanation comes from thinking of what we expect of any reasonable algorithm. Suppose at the start of round t the first arm has been played much more frequently than the rest. If we did a good job designing our algorithm, we would hope this is the optimal arm, and because it has been played so often, we expect that $\hat{\mu}_1(t-1) \approx \mu_1$. To confirm the hypothesis that arm 1 is optimal, the algorithm had better be highly confident that other arms are indeed worse. This leads quite naturally to the idea of using upper confidence bounds. The learner can be reasonably certain that arm i is worse than arm 1 if

$$\hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} \leq \mu_1 \approx \hat{\mu}_1(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_1(t-1)}}\,, \tag{7.3}$$

where δ is called the **confidence level** and quantifies the degree of certainty. This means that choosing the arm with the largest upper confidence bound leads to a situation where arms are only chosen if their true mean could reasonably be larger than those of arms that have been played often. That this rule is indeed a good one depends on two factors. The first is whether the width of the confidence interval at a given confidence level can be significantly decreased, and the second is whether the confidence level is chosen in a reasonable fashion. For now, we will take a leap of faith and assume that the width of confidence intervals for subgaussian bandits cannot be significantly improved from what we use here (we shall see that this holds in later chapters), and concentrate on choosing the confidence level now.

Choosing the confidence level is a delicate problem, and we will analyse a number of choices in future chapters. The basic difficulty is that δ should be small enough to ensure optimism with high probability, but not so small that suboptimal arms are explored excessively.

Nevertheless, as a first cut, the choice of this parameter can be guided by the following considerations. If the confidence interval fails and the index of an optimal arm drops below its true mean, then it could happen that the algorithm stops playing the optimal arm and suffers linear regret. This suggests we might choose $\delta \approx 1/n$ so that the contribution to the regret of this failure case is relatively small. Unfortunately things are not quite this simple. As we have

already alluded to, one of the main difficulties is that the number of samples $T_i(t-1)$ in the index (7.2) is a random variable, and so our concentration results cannot be immediately applied. For this reason we will see that (at least naively) δ should be chosen a bit smaller than $1/n$.

THEOREM 7.1. *Consider UCB as shown in Algorithm 3 on a stochastic k-armed 1-subgaussian bandit problem. For any horizon n, if $\delta = 1/n^2$, then*

$$R_n \leq 3\sum_{i=1}^{k} \Delta_i + \sum_{i:\Delta_i>0} \frac{16\log(n)}{\Delta_i} .$$

Before the proof we need a little more notation. Let $(X_{ti})_{t\in[n],i\in[k]}$ be a collection of independent random variables with the law of X_{ti} equal to P_i. Then define $\hat{\mu}_{is} = \frac{1}{s}\sum_{u=1}^{s} X_{ui}$ to be the empirical mean based on the first s samples. We make use of the third model in Section 4.6 by assuming that the reward in round t is

$$X_t = X_{T_{A_t}(t)A_t} .$$

Then we define $\hat{\mu}_i(t) = \hat{\mu}_{iT_i(t)}$ to be the empirical mean of the ith arm after round t. The proof of Theorem 7.1 relies on the basic regret decomposition identity,

$$R_n = \sum_{i=1}^{k} \Delta_i \mathbb{E}\left[T_i(n)\right] . \qquad \text{(Lemma 4.5)}$$

The theorem will follow by showing that $\mathbb{E}\left[T_i(n)\right]$ is not too large for suboptimal arms i. The key observation is that after the initial period where the algorithm chooses each action once, action i can only be chosen if its index is higher than that of an optimal arm. This can only happen if at least one of the following is true:

(a) The index of action i is larger than the true mean of a specific optimal arm.
(b) The index of a specific optimal arm is smaller than its true mean.

Since with reasonably high probability the index of any arm is an upper bound on its mean, we don't expect the index of the optimal arm to be below its mean. Furthermore, if the suboptimal arm i is played sufficiently often, then its exploration bonus becomes small and simultaneously the empirical estimate of its mean converges to the true value, putting an upper bound on the expected total number of times when its index stays above the mean of the optimal arm. The proof that follows is typical for the analysis of algorithms like UCB, and hence we provide quite a bit of detail so that readers can later construct their own proofs.

Proof of Theorem 7.1 Without loss of generality, we assume the first arm is optimal so that $\mu_1 = \mu^*$. As noted above,

$$R_n = \sum_{i=1}^{k} \Delta_i \mathbb{E}\left[T_i(n)\right] . \tag{7.4}$$

The theorem will be proven by bounding $\mathbb{E}[T_i(n)]$ for each suboptimal arm i. We make use of a relatively standard idea, which is to decouple the randomness from the behaviour of the UCB algorithm. Let G_i be the 'good' event defined by

$$G_i = \left\{\mu_1 < \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right\} \cap \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2}{u_i}\log\left(\frac{1}{\delta}\right)} < \mu_1\right\},$$

where $u_i \in [n]$ is a constant to be chosen later. So G_i is the event when μ_1 is never underestimated by the upper confidence bound of the first arm, while at the same time the upper confidence bound for the mean of arm i after u_i observations are taken from this arm is below the pay-off of the optimal arm. We will show two things:

1 If G_i occurs, then arm i will be played at most u_i times: $T_i(n) \le u_i$.

2 The complement event G_i^c occurs with low probability (governed in some way yet to be discovered by u_i).

Because $T_i(n) \le n$ no matter what, this will mean that

$$\mathbb{E}\left[T_i(n)\right] = \mathbb{E}\left[\mathbb{I}\left\{G_i\right\} T_i(n)\right] + \mathbb{E}\left[\mathbb{I}\left\{G_i^c\right\} T_i(n)\right] \le u_i + \mathbb{P}\left(G_i^c\right) n\,. \tag{7.5}$$

The next step is to complete our promise by showing that $T_i(n) \le u_i$ on G_i and that $\mathbb{P}\left(G_i^c\right)$ is small. Let us first assume that G_i holds and show that $T_i(n) \le u_i$, which we do by contradiction. Suppose that $T_i(n) > u_i$. Then arm i was played more than u_i times over the n rounds, and so there must exist a round $t \in [n]$ where $T_i(t-1) = u_i$ and $A_t = i$. Using the definition of G_i,

$$\begin{aligned}
\mathrm{UCB}_i(t-1,\delta) &= \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}} && \text{(definition of } \mathrm{UCB}_i(t-1,\delta)\text{)}\\
&= \hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} && \text{(since } T_i(t-1) = u_i\text{)}\\
&< \mu_1 && \text{(definition of } G_i\text{)}\\
&< \mathrm{UCB}_1(t-1,\delta)\,. && \text{(definition of } G_i\text{)}
\end{aligned}$$

Hence $A_t = \mathrm{argmax}_j\, \mathrm{UCB}_j(t-1,\delta) \neq i$, which is a contradiction. Therefore if G_i occurs, then $T_i(n) \le u_i$. Let us now turn to upper bounding $\mathbb{P}\left(G_i^c\right)$. By its definition,

$$G_i^c = \left\{\mu_1 \ge \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right\} \cup \left\{\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \ge \mu_1\right\}. \tag{7.6}$$

The first of these sets is decomposed using the definition of $\mathrm{UCB}_1(t,\delta)$,

$$\left\{\mu_1 \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right\} \subseteq \left\{\mu_1 \geq \min_{s\in[n]} \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\}$$
$$= \bigcup_{s\in[n]} \left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\}.$$

Then using a union bound and the concentration bound for sums of independent subgaussian random variables in Corollary 5.5, we obtain:

$$\mathbb{P}\left(\mu_1 \geq \min_{t\in[n]} \mathrm{UCB}_1(t,\delta)\right) \leq \mathbb{P}\left(\bigcup_{s\in[n]} \left\{\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right\}\right)$$
$$\leq \sum_{s=1}^{n} \mathbb{P}\left(\mu_1 \geq \hat{\mu}_{1s} + \sqrt{\frac{2\log(1/\delta)}{s}}\right) \leq n\delta\,. \tag{7.7}$$

The next step is to bound the probability of the second set in (7.6). Assume that u_i is chosen large enough that

$$\Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq c\Delta_i \tag{7.8}$$

for some $c \in (0,1)$ to be chosen later. Then, since $\mu_1 = \mu_i + \Delta_i$, and using Corollary 5.5,

$$\mathbb{P}\left(\hat{\mu}_{iu_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1\right) = \mathbb{P}\left(\hat{\mu}_{iu_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}}\right)$$
$$\leq \mathbb{P}\left(\hat{\mu}_{iu_i} - \mu_i \geq c\Delta_i\right) \leq \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right).$$

Taking this together with (7.7) and (7.6), we have

$$\mathbb{P}\left(G_i^c\right) \leq n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right).$$

When substituted into Eq. (7.5), we obtain

$$\mathbb{E}\left[T_i(n)\right] \leq u_i + n\left(n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)\right). \tag{7.9}$$

It remains to choose $u_i \in [n]$ satisfying (7.8). A natural choice is the smallest integer for which (7.8) holds, which is

$$u_i = \left\lceil \frac{2\log(1/\delta)}{(1-c)^2\Delta_i^2} \right\rceil.$$

This choice of u_i can be larger than n, but in this case Eq. (7.9) holds trivially

since $T_i(n) \le n$. Then, using the assumption that $\delta = 1/n^2$ and this choice of u_i leads via (7.9) to

$$\mathbb{E}[T_i(n)] \le u_i + 1 + n^{1-2c^2/(1-c)^2} = \left\lceil \frac{2\log(n^2)}{(1-c)^2\Delta_i^2} \right\rceil + 1 + n^{1-2c^2/(1-c)^2} . \quad (7.10)$$

All that remains is to choose $c \in (0,1)$. The second term will contribute a polynomial dependence on n unless $2c^2/(1-c)^2 \ge 1$. However, if c is chosen too close to 1, then the first term blows up. Somewhat arbitrarily we choose $c = 1/2$, which leads to

$$\mathbb{E}\left[T_i(n)\right] \le 3 + \frac{16\log(n)}{\Delta_i^2} .$$

The result follows by substituting the above display in Eq. (7.4). □

As we saw for the ETC strategy, the regret bound in Theorem 7.1 depends on the reciprocal of the gaps, which may be meaningless when even a single suboptimal action has a very small suboptimality gap. As before, one can also prove a sublinear regret bound that does not depend on the reciprocal of the gaps.

THEOREM 7.2. *If $\delta = 1/n^2$, then the regret of UCB, as defined in Algorithm 3, on any $\nu \in \mathcal{E}^k_{\mathrm{SG}}(1)$ environment, is bounded by*

$$R_n \le 8\sqrt{nk\log(n)} + 3\sum_{i=1}^{k} \Delta_i .$$

Proof Let $\Delta > 0$ be some value to be tuned subsequently, and recall from the proof of Theorem 7.1 that for each suboptimal arm i, we can bound

$$\mathbb{E}[T_i(n)] \le 3 + \frac{16\log(n)}{\Delta_i^2} .$$

Therefore, using the basic regret decomposition again (Lemma 4.5), we have

$$\begin{aligned} R_n &= \sum_{i=1}^{k} \Delta_i \mathbb{E}\left[T_i(n)\right] = \sum_{i:\Delta_i<\Delta} \Delta_i \mathbb{E}\left[T_i(n)\right] + \sum_{i:\Delta_i\ge\Delta} \Delta_i \mathbb{E}\left[T_i(n)\right] \\ &\le n\Delta + \sum_{i:\Delta_i\ge\Delta} \left(3\Delta_i + \frac{16\log(n)}{\Delta_i}\right) \le n\Delta + \frac{16k\log(n)}{\Delta} + 3\sum_i \Delta_i \\ &\le 8\sqrt{nk\log(n)} + 3\sum_{i=1}^{k} \Delta_i , \end{aligned}$$

where the first inequality follows because $\sum_{i:\Delta_i<\Delta} T_i(n) \le n$ and the last line by choosing $\Delta = \sqrt{16k\log(n)/n}$. □

The additive $\sum_i \Delta_i$ term is unavoidable because no reasonable algorithm can avoid playing each arm once (try to work out what would happen if it did not). In any case, this term does not grow with the horizon n and is typically negligible.

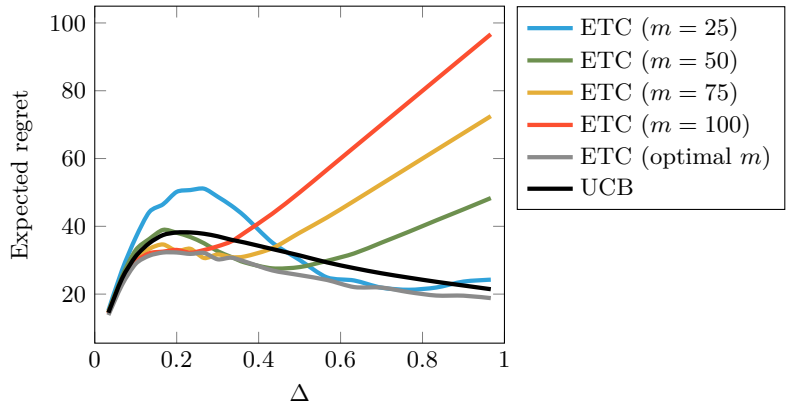

Figure 7.1 Experiment showing universality of UCB relative to fixed instances of ETC

As it happens, Theorem 7.2 is close to optimal. We will see in Chapter 15 that no algorithm can enjoy regret smaller than $O(\sqrt{nk})$ over all problems in $\mathcal{E}^k_{\mathrm{SG}}(1)$. In Chapter 9 we will also see a more complicated variant of Algorithm 3 that shaves the logarithmic term from the upper bound given above.

Experiment 7.1 We promised that UCB would overcome the limitations of ETC by achieving the same guarantees but without prior knowledge of the suboptimality gaps. The theory supports this claim, but just because two algorithms have similar theoretical guarantees does not mean they perform the same empirically. The theoretical analysis might be loose for one algorithm and maybe not the other, or by a different margin. For this reason it is always wise to prove lower bounds (which we do later) and compare the empirical performance, which we do (very briefly) now.

The set-up is the same as in Fig. 6.1, which has $n = 1000$ and $k = 2$ and unit variance Gaussian rewards with means 0 and $-\Delta$ respectively. The plot in Fig. 7.1 shows the expected regret of UCB relative to ETC for a variety of choices of commitment time m. The expected regret of ETC with the optimal choice of m (which depends on the knowledge of Δ and that the pay-offs are Gaussian, cf. Fig. 6.1) is also shown.

The results demonstrate a common phenomenon. If ETC is tuned with the optimal choice of commitment time for each choice of Δ, then it outperforms the parameter-free UCB, though only by a relatively small margin. If, however, the commitment time must be chosen without the knowledge of Δ, then ETC will usually not outperform UCB. As it happens, a variant of UCB introduced in the next chapter actually outperforms even the optimally tuned ETC.