7 The Upper Confidence Bound Algorithm

We now describe the celebrated upper confidence bound (UCB) algorithm, which offers several advantages over the 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 upper confidence bound 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. Recall that if $X_1, X_2, \ldots, X_n$ are independent and 1-subgaussian with mean μ and $\hat{\mu} = \sum_{t=1}^{n} X_t/n$, then by Eq. (5.6), for any $\delta \in (0,1)$,

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

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

$$\mathrm{UCB}_i(t-1,\delta) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}\,. \tag{7.2}$$

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.

1: **Input** K and δ
2: Choose each action once
3: For rounds $t > K$ choose action

$$A_t = \mathrm{argmax}_i\, \mathrm{UCB}_i(t-1,\delta)$$

Algorithm 2: UCB(δ) algorithm

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 algorithm first chooses each arm once, which is necessary because the term inside the square root is undefined when $T_i(t-1) = 0$. The value inside the argmax is called the **index** of arm i. Generally speaking, an **index algorithm** chooses the arm in each round that maximizes some value (the index), which usually only depends on 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** (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 behavior. 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 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 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

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

Then choosing the arm with the largest upper confidence bound leads to the situation where arms that have not been played very often are chosen only once their true mean could reasonably be larger than those of arms that have been played often. The term inside the logarithm is called the **confidence level**. 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 itself turns out to be a delicate problem and we will spend quite a lot of time analyzing various choices in future chapters.

The basic difficulty is that there is a trade-off between choosing δ very small and ensuring optimism with very high probability, and the cost of being excessively optimistic about suboptimal arms. Note that optimism is only really required for the optimal arm because this ensures that once the suboptimal arms have been proven to have means less than optimal, then the optimal arm is all that remains.

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 2 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 4.4 by assuming that the reward in round t is

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

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,

$$R_n = \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[T_i(n)\right] . \qquad \text{(Lemma 4.2)}$$

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 behavior of the UCB algorithm. Let G_i be the 'good' event defined by

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

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 payoff of the optimal arm. We will show two things:

1 If G_i occurs, then $T_i(n) \le u_i$.
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 we have:

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

which means that $A_t = \mathrm{argmax}_j \mathrm{UCB}_j(t-1) \neq i$ and so a contradiction is obtained. 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,

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

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

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

Then using a union bound and the concentration bound for sums of independent subgaussian random variables in Corollary 5.1 we obtain:

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

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 5.1,

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

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 2, on any $\nu \in \mathcal{E}_{\mathrm{SG}}^K(1)$ 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 4.2), 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.

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 $\mathcal{E}_{\mathrm{SG}}^{K}(1)$. In Chapter 9 we will also see a more complicated variant of Algorithm 2 that shaves the logarithmic term from the upper bound given above.

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 setup 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 payoffs 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 can outperform the parameter-free UCB, though only by a relatively small margin. If, however, the commitment time must be chosen without the knowledge of Δ, for Δ getting large, or for Δ being bounded, n getting large, UCB arbitrarily outperforms ETC. As it happens, a variant of UCB introduced in the next chapter actually outperforms even the optimally tuned ETC.

7.2 Notes

1 The choice of $\delta = 1/n^2$ led to an easy analysis, but comes with two disadvantages. First of all, it turns out that a slightly smaller value of δ improves the regret (and empirical performance). Secondly, the dependence on n means the horizon must be known in advance, which is often not reasonable. Both of these issues are resolved in the next chapter where δ is chosen to be smaller and to depend on the current round t rather than n. None-the-less – as promised – Algorithm 2 with $\delta = 1/n^2$ does achieve a regret bound similar to the ETC strategy, but without requiring knowledge of the gaps.

2 The assumption that the rewards generated by each arm are independent can be relaxed significantly. All of the results would go through by assuming there

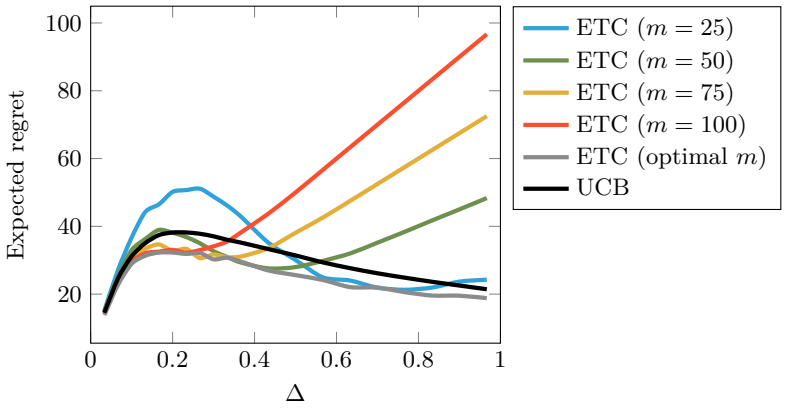

Figure 7.1 Experiment showing universality of UCB relative to fixed instances of explore-then-commit

exists a mean reward vector $\mu \in \mathbb{R}^K$ such that

$$\mathbb{E}[X_t \mid X_1, A_1, \ldots, A_{t-1}, X_{t-1}, A_t] = \mu_{A_t} \text{ a.s.}. \tag{7.11}$$

$$\mathbb{E}[\exp(\lambda(X_t - \mu_{A_t})) \mid X_1, A_1, \ldots, A_{t-1}, X_{t-1}, A_t] \leq \exp(\lambda^2/2) \text{ a.s.}. \tag{7.12}$$

Eq. (7.11) is just saying that the conditional mean of the reward in round t only depends on the chosen action. Eq. (7.12) ensures that the tails of X_t are conditionally subgaussian. That everything still goes through is proven using martingale techniques, which we develop in detail in Chapter 20.

3 So is the optimism principle universal? Does it always lead to policies with strong guarantees in more complicated settings? Unfortunately the answer turns out to be no. The optimism principle usually leads to reasonable algorithms when *(i)* any action gives feedback about the quality of that action and *(ii)* no action gives feedback about the value of other actions. When *(i)* is violated even sublinear regret may not be guaranteed. When *(ii)* is violated an optimistic algorithm may avoid actions that lead to large information gain and low reward, even when this tradeoff is optimal. An example where this occurs is provided in Chapter 25 on linear bandits. Optimism can work in more complex models as well, but sometimes fails to appropriately balance exploration and exploitation.

4 When thinking about future outcomes, humans, as well as other higher animals, often have higher expectations than what is warranted by past experience and/or conditions of the environment. This phenomenon, a form of **cognitive bias**, is known as the **optimism bias** in the psychology and behavioral economics literature and is in fact "one of the most consistent, prevalent, and robust biases documented in psychology and behavioral economics" [Sharot, 2011a]. While much has been written about this bias in these fields and one of the current explanations of why the optimism bias is so prevalent is that it helps exploration, to our best knowledge, the connection to the deeper mathematical

justification of optimism, pursued here and in other parts of this book, has so far escaped the attention of researchers in all the relevant fields.

7.3 Bibliographical remarks

The use of confidence bounds and the idea of optimism first appeared in the work by Lai and Robbins [1985] (for the curious, it is the same Robbins). They analyzed the asymptotics for various parametric bandit problems (see the next chapter for more details on this). The first version of UCB is by Lai [1987]. Other early work is by Katehakis and Robbins [1995], who gave a very straightforward analysis for the Gaussian case and Agrawal [1995], who noticed that all that was needed is an appropriate sequence of upper confidence bounds on the unknown means. In this way, their analysis is significantly more general than what we have done here. These researchers also focussed on the asymptotics, which at the time was the standard approach in the statistics literature. The UCB algorithm was independently discovered by Kaelbling [1993], although with no regret analysis or clear advice on how to tune the confidence parameter. The version of UCB discussed here is most similar to that analyzed by Auer et al. [2002a] under the name UCB1, but that algorithm used t rather than n in the confidence level (see the next chapter). Like us, they prove a finite-time regret bound. However, rather than considering 1-subgaussian environments, Auer et al. [2002a] considers bandits where the payoffs are confined to the $[0, 1]$ interval, which are ensured to be $1/2$-subgaussian. See Exercise 7.2 for hints on what must change in this situation. The basic structure of the proof of our Theorem 7.1 is essentially the same as that of Theorem 1 of Auer et al. [2002a]. The worst-case bound in Theorem 7.2 appeared in the book by Bubeck and Cesa-Bianchi [2012], which also popularized the subgaussian setup. We did not have time to discuss the situation where the subgaussian constant is unknown. There have been several works exploring this direction. If the variance is unknown, but the noise is bounded, then one can replace the subgaussian concentration bounds with an empirical Bernstein inequality [Audibert et al., 2007]. For details see Exercise 7.7. If the noise has heavy tails, then a more serious modification is required as discussed in Exercise 7.8 and the note that follows.

We found the article by Sharot [2011a] on optimism bias from the psychology literature quite illuminating. Readers who are willing to dive deeper in this literature may also find the book by the same author useful [Sharot, 2011b]. Optimism bias is also known as "unrealistic optimism", a term that is most puzzling to us – what bias is ever realistic? The background of this is explained by Jefferson et al. [2017].

7.4 Exercises

7.1 In this exercise we investigate one of the more annoying challenges when analyzing sequential algorithms. Let $X_1, X_2, \ldots$ be a sequence of independent standard Gaussian random variables defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that $T : \Omega \to \{1, 2, 3, \ldots\}$ is another random variable and let $\hat{\mu} = \sum_{t=1}^{T} X_t / T$ be the empirical mean based on T samples.

(a) Show that if T is independent from X_t for all t, then

$$\mathbb{P}\left(\hat{\mu} - \mu \geq \sqrt{\frac{2\log(1/\delta)}{T}}\right) \leq \delta\,.$$

(b) We now relax the assumption that T is independent. Let $E_t = \mathbb{I}\{T = t\}$ be the event that $T = t$ and $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$ be the σ-algebra generated by the first t samples. Show there exists a T such that for all $t \in \{1, 2, 3, \ldots\}$ it holds that E_t is $\mathcal{F}_t$-measurable and

$$\mathbb{P}\left(\hat{\mu} - \mu \geq \sqrt{\frac{2\log(1/\delta)}{T}}\right) = 1 \qquad \text{for all } \delta \in (0, 1)\,.$$

(c) Show that

$$\mathbb{P}\left(\hat{\mu} - \mu \geq \sqrt{\frac{2\log(T(T+1)/\delta)}{T}}\right) \leq \delta\,. \tag{7.13}$$

For part (b) above you may find it useful to apply the law of the iterated logarithm, which says if $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with zero mean and unit variance, then

$$\limsup_{n\to\infty} \frac{\sum_{t=1}^{n} X_t}{\sqrt{2n \log\log n}} = 1 \qquad \text{almost surely}\,.$$

This result is especially remarkable because it relies on no assumptions other than zero mean and unit variance. A thoughtful reader might wonder if Eq. (7.13) might still be true if $\log(T(T+1))/\delta$ were replaced by $\log(\log(T)/\delta)$. It almost can, but the proof of this fact is more sophisticated. For more details see the paper by Garivier [2013] or Exercise 23.6.

7.2 In this chapter we assumed the payoff distributions were 1-subgaussian (that is, $\nu \in \mathcal{E}_{\mathrm{SG}}^{K}(1)$). The purpose of this exercise is to relax this assumption.

(a) First suppose that $\sigma^2 > 0$ is a known constant and that $\nu \in \mathcal{E}_{\mathrm{SG}}^{K}(\sigma^2)$. Modify the UCB algorithm and state and prove an analogue of Theorems 7.1 and 7.2 for this case.

(b) Now suppose that $\nu = (\nu_i)_i$ is chosen so that ν_i is σ_i-subgaussian where $(\sigma_i^2)_i$ are known. Modify the UCB algorithm and state and prove an analogue of Theorems 7.1 and 7.2 for this case.

(c) If you did things correctly, the regret bound in the previous part should not depend on the values of $\{\sigma_i^2 : \Delta_i = 0\}$. Explain why not.

7.3 Recall from Chapter 4 that the pseudo-regret is defined to be the random variable

$$\bar{R}_n = \sum_{t=1}^{n} \Delta_{A_t} \,.$$

The UCB policy in Algorithm 2 depends on confidence parameter $\delta \in (0, 1]$ that determines the level of optimism. State and prove a bound on the pseudo-regret of this algorithm that holds with probability $1 - f(n, K)\delta$ where $f(n, K)$ is a function that depends on n and K only. More precisely show that for bandit $\nu \in \mathcal{E}_{\mathrm{SG}}^K(1)$ that

$$\mathbb{P}\left(\bar{R}_n \geq g(n, \nu, \delta)\right) \leq f(n, K)\delta \,,$$

where g and f should be as small as possible (there are trade-offs – try and come up with a natural choice).

7.4 This exercise is about the empirical behavior of UCB.

(a) Implement Algorithm 2.
(b) Reproduce Fig. 7.1.
(c) Explain the shape of the ETC curves. In particular, when $m = 50$ we see a bump, a dip, and then a linear asymptote as Δ grows. Why does the curve look like this?
(d) Design an experiment to determine the practical effect of the choice of δ. There are many interesting regimes where this is interesting. For example:
  (1) Suppose you have a Gaussian bandit with two arms and means $\mu_1 = 0$ and $\mu_2 = -\Delta$. Let $n = 1000$ and try to determine the optimal value of δ for UCB as a function of Δ.
  (2) What happens if you have more arms. For example, $\mu_1 = 0$ and $\mu_i = -\Delta$ for $i > K$. How does the optimal choice of δ change as K increases?
  (3) Justify your results pseudo-theoretically (that is, provide a theoretically motivated justification for the results, but no proof).

7.5 Fix a 1-subgaussian K-armed bandit environment and a horizon n. Consider the version of UCB that works in phases of exponentially increasing length of $1, 2, 4, \ldots$. In each phase, the algorithm uses the action that would have been chosen by UCB at the beginning of the phase (see Algorithm 3 below).

(a) State and prove a bound on the regret for this version of UCB.
(b) Compare your result with Theorem 7.1.
(c) How would the result change if the kth phase had a length of $\lceil \alpha^k \rceil$ with $\alpha > 1$?

Input K and δ
Choose each arm once
for $\ell = 1, 2, \ldots$ **do**
    Compute $A_\ell = \operatorname{argmax}_i \mathrm{UCB}_i(t-1, \delta)$
    Choose arm A_ℓ exactly 2^ℓ times
end for

Algorithm 3: A phased version of UCB

7.6 Let $\alpha > 1$ and consider the version of UCB that first plays each arm once. Thereafter it operates in the same way as UCB, but rather than playing the chosen arm just once, it plays it until the number of plays of that arm is a factor of α larger (see Algorithm 4 below).

(a) State and prove a bound on the regret for version of UCB with $\alpha = 2$ (doubling counts).
(b) Compare with the result of the previous exercise and with Theorem 7.1. What can you conclude?
(c) Repeat the analysis for $\alpha > 1$. What is the role of α?
(d) Implement these algorithms and compare them empirically to UCB(δ).

Input K and δ
Choose each arm once
for $\ell = 1, 2, \ldots$ **do**
    Let $t_\ell = t$
    Compute $A_\ell = \operatorname{argmax}_i \mathrm{UCB}_i(t_\ell - 1, \delta)$
    Choose arm A_ℓ until round t such that $T_i(t) \geq \alpha T_i(t_\ell - 1)$
end for

Algorithm 4: A phased version of UCB

The algorithms of the last two exercises may seem ridiculous. Why would you wait before updating empirical estimates and choosing a new action? There are at least two reasons:

- It can happen that the algorithm does not observe its rewards immediately, but rather they appear asynchronously after some delay. Alternatively many bandits algorithms may be operating simultaneously and the results must be communicated at some cost.
- If the feedback model has a more complicated structure than what we examined so far, then even computing the upper confidence bound just once can be quite expensive. In these circumstances it's comforting to know that the loss of performance by updating the statistics only rarely is not too severe.

7.7 Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with mean μ and variance σ^2 and bounded support so that $X_t \in [0, b]$ almost surely. Let $\hat{\mu} = \sum_{t=1}^n X_t/n$ and $\hat{\sigma}^2 = \sum_{t=1}^n (\hat{\mu} - X_t)^2/n$. The **empirical Bernstein** inequality says that for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(|\hat{\mu} - \mu| \geq \sqrt{\frac{2\hat{\sigma}^2}{n} \log\left(\frac{3}{\delta}\right)} + \frac{3b}{n} \log\left(\frac{3}{\delta}\right)\right) \leq \delta\,.$$

(a) Show that $\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^n (X_t - \mu)^2 - (\hat{\mu} - \mu)^2$.
(b) Show that $\mathbb{V}[(X_t - \mu)^2] \leq b^2\sigma^2$.
(c) Use Bernstein's inequality (Exercise 5.16) to show that

$$\mathbb{P}\left(\hat{\sigma}^2 \geq \sigma^2 + \sqrt{\frac{2b^2\sigma^2}{n} \log\left(\frac{1}{\delta}\right)} + \frac{2b^2}{3n} \log\left(\frac{1}{\delta}\right)\right) \leq \delta\,.$$

(d) Suppose that $\nu = (\nu_i)_{i=1}^K$ is a bandit where $\mathrm{Supp}(\nu_i) \subset [0, b]$ and the variance of the ith arm is σ_i^2. Design a policy that depends on b, but not σ_i^2 such that

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

where $C > 0$ is a universal constant.

If you did things correctly, then the policy you derived in Exercise 7.7 should resemble UCB-V by Audibert et al. [2007]. The proof of the empirical Bernstein also appears there or in the papers by Mnih et al. [2008] and Maurer and Pontil [2009].

7.8 Let $n \in \mathbb{N}^+$ and $(A_i)_{i=1}^k$ be a partition of $[n]$ so that $\cup_{i=1}^k A_i = [n]$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$. Suppose that $\delta \in (0, 1)$ and $X_1, X_2, \ldots, X_n$ is a sequence of independent random variables with mean μ and variance σ^2. The **median-of-means estimator** $\hat{\mu}_M$ of μ the median of $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k$ where $\hat{\mu}_i = \sum_{t \in A_i} X_t/|A_i|$ is the mean of the data in the ith block.

(a) Show that if $k = \left\lfloor \min\left\{\frac{n}{2}, 8\log\left(\frac{e^{1/8}}{\delta}\right)\right\}\right\rfloor$ and A_i are chosen as equally sized as possible, then

$$\mathbb{P}\left(\hat{\mu}_M + \sqrt{\frac{192\sigma^2}{n} \log\left(\frac{e^{1/8}}{\delta}\right)}\right) \leq \delta\,.$$

(b) Use the median-of-means estimator to design an upper confidence bound algorithm such that for all $\nu \in \mathcal{E}_{\mathbb{V}}^K(\sigma^2)$

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

where $C > 0$ is a universal constant.

This exercise shows that unless one cares greatly about constant factors, then the subgaussian assumption can be relaxed to requiring only finite variance. The result is only possible by replacing the standard empirical estimator with something more robust. The median-of-means estimator is only one way to do this. In fact, the empirical estimator can be made robust by truncating the observed rewards and applying the empirical Bernstein concentration inequality. The disadvantage of this approach is that choosing the location of truncation requires prior knowledge about the approximate location of the mean. Another approach is **Catoni's estimator**, which also exhibits excellent asymptotic properties [Catoni, 2012]. Yet another idea is to minimize the Huber loss [Sun et al., 2017]. This latter paper is focussing on linear models, but the results still apply in one dimension. The application of these ideas to bandits was first made by Bubeck et al. [2013a], where the reader will find more interesting results. Most notably, that things can still be made to work even if the variance does not exist. In this case, however, there is a price to be paid in terms of the regret. The median-of-means estimator is due to Alon et al. [1996]. In case the variance is also unknown, then it may be estimated by assuming a known bound on the **kurtosis**, which covers many classes of bandits (Gaussian with arbitrary variance, exponential and many more), but not some simple cases (Bernoulli). The policy that results from this procedure has the benefit of being invariant under the transformations of shifting or scaling the losses [Lattimore, 2017].