Multi-armed Bandits

1 The UCB algorithm

We now present an algorithm for the multi-armed bandit problem known as the upper confidence bound (UCB) algorithm. These notes closely follow the presentation in Chapter 2 of Bubeck and Cesa-Bianchi, Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems, NOW Publishers, 2012, specialised to the case of Bernoulli bandits for simplicity.

We consider a bandit with K arms, and denote by $X_i(t)$, $i \in \{1, \ldots, K\}$, $t \in \mathbb{N}$, the random reward that would be received if arm i were played in time step t. We assume that the random rewards $X_i(t), t \in \mathbb{N}$ associated with the i^{th} arm are iid Bern(μ_i), and that rewards are mutually independent across arms. Finally, we assume without loss of generality (wlog) that the arms have been ordered so that $\mu_1 > \mu_2 \geq \ldots \geq \mu_K$, but that the ordering, and the parameters $\mu_1, \ldots, \mu_K$, are unknown to the player. The quantities $\Delta_i = \mu_1 - \mu_i$, $i \geq 2$, are termed the arm gaps. They measure how difficult it is to distinguish the best arm from competing arms, and will hence play a role in the best regret we can achieve.

We introduce some more notation. Denote by $I(t) \in \{1, \ldots, K\}$ the arm played in round, or time step, t. We shall use the terms round and time step interchangeably. In general, $I(t)$ will depend on the arms played in previous rounds and the observed rewards. It may involve additional randomness, i.e., the player might adopt a randomised strategy. We assume that the player has access to a source of randomness independent of $X_i(t)$, $i \in \{1, \ldots, K\}$, $t \in \mathbb{N}$, in order to implement such a policy. Denote the number of times arm i has been played in the first t rounds by

$$N_i(t) = \sum_{s=1}^{t} \mathbf{1}(I(s) = i),$$

and the total reward obtained from this arm in the first t rounds by

$$S_i(t) = \sum_{s=1}^{n} X_i(s)\mathbf{1}(I(s) = i),$$

where $\mathbf{1}(A)$ denotes the indicator of the event A. Finally, denote the sample mean reward obtained from the first n plays of arm i by $\hat{\mu}_{i,n}$. Then, we have

$$\hat{\mu}_{i,N_i(t)} = \frac{S_i(t)}{N_i(t)},$$

provided $N_i(t)$ is non-zero; $\hat{\mu}_{i,N_i(t)}$ is undefined if $N_i(t)$ is zero.

We motivate the UCB algorithm before stating it precisely. Suppose t is such that $N_i(t) > 0$ for all i, i.e., each arm i has been played at least once in the first t rounds. The simplest or most naive approach would be to assume that $\hat{\mu}_{i,N_i(t)}$ is an accurate estimate of μ_i, and play the arm with the highest sample mean. How well does this strategy perform? Suppose we first play each arm once, so that sample means are well defined, and subsequently only play the arm with the highest sample mean. Now, with probability $1 - \mu_1$, arm 1 yields zero reward on the first play, whereas with probability at least μ_2, one of the other arms (in fact, the second arm) yields a unit reward. Moreover, the sample mean for this arm will always be strictly positive in future, so the first arm will never be played. Thus, we see that with probability at least $\mu_2(1-\mu_1)$, the first arm is played only once. Hence, the regret up to time T is given by

$$\mathcal{R}(T) \geq \mathbb{E}\Big[\sum_{s=1}^{T} \mu_1 - \mu_{I(s)}\Big] \geq (T-1)\mu_2(1-\mu_1)(\mu_1 - \mu_2).$$

Thus, the regret scales linearly in T for this strategy, whereas both heuristics in the last section yielded sublinear regret. Therefore, treating the sample mean as if it were the true mean (a policy known as *certainty equivalence*) is clearly not optimal.

We need to somehow account for the uncertainty in our estimate. We take the approach of being optimistic in our estimate of the true reward distribution. For a given small value $\delta > 0$, how large could μ_i be to account for the observed rewards, with confidence at least δ? Since the rewards take values in $[0, 1]$, we can use Hoeffding's inequality from earlier. We have

$$\mathbb{P}\big(\mu_i > \hat{\mu}_{i,n} + x\big) \leq \mathbb{P}\Big(\Big|\sum_{i=1}^{n} X_i(t) - n\mu_i)\Big| > nx\Big|\Big) \leq e^{-2nx^2}. \tag{1}$$

It follows that

$$x = \sqrt{\frac{1}{2n}\log\Big(\frac{1}{\delta}\Big)} \;\Rightarrow\; \mathbb{P}\big(\mu_i \le \hat{\mu}_{i,n} + x\big) > 1 - \delta.$$

Thus, the largest, or most optimistic, value consistent with a $(1-\delta)$ confidence interval for μ_i is equal to $\hat{\mu}_{i,n} + \sqrt{\log(1/\delta)/2n}$, after arm i has been played n times; in other words, at time t, this value is $\hat{\mu}_{i,N_i(t)} + \sqrt{\log(1/\delta)/2N_i(t)}$. This is called the upper confidence bound (UCB) associated with arm i. The UCB algorithm plays, at each time instant, the arm with the highest UCB value (breaking ties arbitrarily, if there are any). The last ingredient in the UCB algorithm is to not use a fixed confidence level δ, but to adapt it over time in the correct way. With this rough intuition, we now formally describe the UCB algorithm and analyse its regret. The algorithm will be parametrised by a positive real number, α.

UCB(α) algorithm

1. In the first K rounds, where K is the number of arms, play each arm once, in arbitrary order.

2. At the end of each round $t \ge K$, compute the UCB(α) index of each arm. The index of arm $i \in \{1, \ldots, K\}$ is defined to be

$$\hat{\mu}_{i,N_i(t)} + \sqrt{\frac{\alpha \log t}{2N_i(t)}}.$$

   In round $t+1$, play the arm with the highest index, breaking ties arbitrarily. In other words,

$$I(t+1) \in \operatorname*{arg\,max}_{i=1,\ldots,K} \left\{ \hat{\mu}_{i,N_i(t)} + \sqrt{\frac{\alpha \log T}{2N_i(t)}} \right\}$$

We can now prove the following upper bound on the regret of this algorithm.

Theorem 1 *Consider the multi-armed bandit problem with K arms, where the rewards from the i^{th} arm are iid Bernoulli(μ_i) random variables, and rewards from different arms are mutually indpendent. Assume wlog that*

$\mu_1 > \mu_2 \geq \ldots \geq \mu_K$, and, for $i \geq 2$, let $\Delta_i = \mu_1 - \mu_i$. Let $\mathcal{R}(T)$ denote the regret of the UCB(α) algorithm in the first T rounds. Then, for $\alpha > 1$,

$$\mathcal{R}(T) \leq \sum_{i=2}^{K} \Big(\frac{\alpha+1}{\alpha-1} \Delta_i + \frac{2\alpha \log T}{\Delta_i} \Big).$$

Remarks. The main point to observe is that the regret grows very slowly with T; it only grows logarithmically in T. This shouldn't be too surprising as we already saw a heuristic that achieved logarithmic regret, albeit for a fixed time horizon T, and assuming that the reward parameters μ_1 and μ_2 were known. Secondly, the expression above shows a trade-off inherent to the choice of α. In the long run, when T is large, the second term dominates, and so we would like to choose α as small as possible. But the first term, which is constant for all T, grows with α, and blows up to infinity as α approaches 1 from above. In practice, choosing α a little bit bigger than 1, say $\alpha = 2$, is a good compromise.

Proof. First observe that if $I_{s+1} = i$, then one or more of the following three things must be true:

$$\hat{\mu}_{1,N_1(s)} \leq \mu_1 - \sqrt{\frac{\alpha \log s}{2N_1(s)}}, \tag{2}$$

$$\hat{\mu}_{i,N_i(s)} > \mu_i + \sqrt{\frac{\alpha \log s}{2N_i(s)}}, \tag{3}$$

$$N_i(s) < \frac{2\alpha \log s}{\Delta_i^2}. \tag{4}$$

Indeed, if all three inequalities are false, we have:

$$\begin{aligned} \hat{\mu}_{1,N_1(s)} + \sqrt{\frac{\alpha \log s}{2N_1(s)}} &> \mu_1 \\ &= \mu_i + \Delta_i \\ &\geq \mu_i + \sqrt{\frac{2\alpha \log s}{N_i(s)}} \\ &\geq \hat{\mu}_i + \sqrt{\frac{\alpha \log s}{2N_i(s)}}. \end{aligned} \tag{5}$$

Here, the first inequality holds because (2) is assumed to be false, the equality follows from the definition of Δ_i, the second inequality holds because (4) is false, and the last inequality because (3) is false. However, if (5) is true, then the UCB(α) index of arm 1 is strictly bigger than that of i at the end of round s, and so arm i cannot be played in round $s+1$. By contradiction, at least one of (2)-(4) must be true, as claimed.

We now use this to bound the expected number of times that a suboptimal arm has been played up to a time t. Let

$$u = \Big\lceil \frac{2\alpha \log t}{\Delta_i^2} \Big\rceil,$$

where we have not made it explicit in the notation that u is a function of i and t. Now, for an arbitrary sequence $I(s)$, $s = 1, 2, \ldots, t$, we have

$$N_i(t) = \sum_{s=1}^{t} \mathbf{1}(I(s) = i) \leq u + \sum_{s=1}^{t} \mathbf{1}(N_i(s-1) \geq u \text{ and } I(s) = i).$$

Indeed,equality holds above if the sequence $I(s)$ is such that $N_i(t) \geq u$, whereas the inequality is strict if $N_i(t) < u$. Moreover, the inequality $N_i(s-1) \geq u$ cannot hold for $s \leq u$ (Why?), and so we can rewrite the above as

$$N_i(t) \leq u + \sum_{s=u+1}^{t} \mathbf{1}(N_i(s-1) \geq u \text{ and } I(s) = i). \tag{6}$$

Notice that the above is a statement about sequences that always holds, and does not depend on the distribution of the random variables involved. Moreover, for every $s \leq t$, and for u defined as above, $u \geq (2\alpha \log(s-1))/\Delta_i^2$. Hence, $N_i(s-1) \geq u$ implies that $N_i(s-1) \geq (2\alpha \log(s-1))/\Delta_i^2$. Thus, taking expectations on both sides of (6), we get

$$\begin{aligned}
\mathbb{E}[N_i(t)] &\leq u + \mathbb{E}\Big[\sum_{s=u+1}^{t} \mathbf{1}(I(s) = i \text{ and ineq. (4) is false}) \Big] \\
&\leq u + \sum_{s=u+1}^{t} \mathbb{E}[\mathbf{1}(\text{ineq. (2) or ineq. (3) is true})] \\
&\leq u + \sum_{s=u+1}^{t} \big(\mathbb{P}(\text{ineq. 2 is true}) + \mathbb{P}(\text{ineq. (3) is true}) \big). \qquad (7)
\end{aligned}$$

We will bound the above quantity, by bounding each of the last two probabilities. We bound each of the probabilities using Hoeffding's inequality, and the bounds take the same form:

$$\mathbb{P}\Big(\hat{\mu}_{i,N_i(s)} - \mu_i > \sqrt{\frac{\alpha \log s}{2N_i(s)}}\Big) \leq e^{-\alpha \log s} = s^{-\alpha}.$$

The same bound also applies to the probability of the inequality in (2). Substituting these bounds in (7), we get

$$\begin{aligned}
\mathbb{E}[N_i(t)] &\leq u + \sum_{s=u+1}^{t} 2s^{-\alpha} \\
&\leq u + \int_{u}^{\infty} 2s^{-\alpha} ds \\
&\leq u + \frac{2}{\alpha - 1} u^{-\alpha+1} \\
&\leq u + \frac{2}{\alpha - 1},
\end{aligned}$$

since $u \geq 1$. We've used the fact that $\alpha > 1$ by assumption to obtain the third inequality. Substituting the definition of u in the above, we get

$$\mathbb{E}[N_i(t)] \leq \left\lceil \frac{2\alpha \log t}{\Delta_i^2} \right\rceil + \frac{2}{\alpha - 1} \leq \frac{2\alpha \log t}{\Delta_i^2} + 1 + \frac{2}{\alpha - 1}.$$

Finally, note that a regret of $\Delta_i = \mu_1 - \mu_i$ is incurred every time arm i is played. Using the above bound on the expected number of plays of arm i up to time t, and summing over i, we obtain the claim of the theorem. □

The above theorem gives us an upper bound on the regret achieved by the UCB(α) algorithm. Ignoring the constant term, and noting that we need α to be bigger than 1, we see that the regret grows with T like $2 \log T \sum_{i=2}^{K} (1/\Delta_i)$.

Can some other algorithm do better? Can it achieve a regret growing more slowly with T, say as $\log \log T$, or even bounded by a constant for all T? Our next result says that this is impossible, and that a $\log T$ scaling is the best achievable. Moreover, the constant factor multiplying $\log T$ in the UCB regret bound is close to the best that any algorithm can achieve. There is a variant of the UCB algorithm known as KL-UCB which does achieve the best possible asymptotic growth rate of regret (i.e., the best possible

constant multiplying the $\log T$ term), but it is a bit more complicated and we won't study it.

Definition. A policy or strategy for the multi-armed bandit problem is said to be *strongly consistent* if its regret satisfies $\mathcal{R}(T) = o(T^\alpha)$ for all $\alpha > 0$.

In words, a policy is strongly consistent if its regret grows slower than any fractional power of T. The first heuristic we studied for the multi-armed bandit problem had regret growing as $\mathcal{R}(T) \sim T^{\mu_2/\mu_1}$. Thus, the regret for this heuristic is sublinear, but it is not $o(T^\alpha)$ for any $\alpha \leq \mu_2/\mu_1$. Hence, it is not strongly consistent. On the other hand, the UCB(α) policy is strongly consistent for any $\alpha > 1$ as it was shown that its regret grows only logarithmically in T. Lai and Robbins (Asymptotically efficient adaptive allocation rules, *Advances in Applied Mathematics*, 1985) proved the following about any strongly consistent policy.

Theorem 2 (Lai and Robbins, 1985) *Consider the multi-armed bandit problem with K arms, where the rewards from arm i are iid Bern(μ_i), and rewards from distinct arms are mutually independent. Then, for any strongly consistent policy, the number of times, $N_i(T)$, that a sub-optimal arm i is played up to time T, satisfies*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_i(T)]}{\log T} \geq \frac{1}{K(\mu_i;\mu^*)},$$

where $\mu^* = \max_{i=1}^K \mu_i$ denotes the mean reward from the optimal arm, and $K(q;p)$ is the relative entropy or KL-divergence of a Bern(q) distribution with respect to a Bern(p) distribution.

We won't prove this theorem, but use it to obtain a lower bound on the regret of any policy for the multi-armed bandit problem. Notice that it suffices to restrict ourselves to strongly consistent policies as we know that such policies exist (in fact, we showed that UCB(α) is one such policy), and that any policy which is not strongly consistent has regret growing at least as fast as T^ϵ, for some $\epsilon > 0$. Now, using the theorem of Lai and Robbins, it easily follows that the regret of any policy is bounded below as follows:

$$\liminf_{T\to\infty} \frac{\mathcal{R}(T)}{\log T} = \liminf_{T\to\infty} \frac{\sum_{i:\mu_i<\mu_*}(\mu^* - \mu_i)\mathbb{E}[N_i(T)]}{\log T} \geq \sum_{i:\mu_i<\mu^*} \frac{\mu^* - \mu_i}{K(\mu_i;\mu^*)}. \quad (8)$$

We want to know how this compares with the UCB(α) algorithm, for which we showed that the regret satisfies

$$\limsup_{T\to\infty} \frac{\mathcal{R}(T)}{\log T} \leq \sum_{i:\mu_i<\mu^*} \frac{2}{\mu^* - \mu_i}. \tag{9}$$

In order to compare the bounds in (8) and (9), we invoke Pinsker's inequality, which states for Bern(p) and Bern(q) distributions that $K(q;p) \geq 2(q-p)^2$. The proof is a homework problem. Using this inequality, we see that the upper bound on the regret achieved by UCB is approximately four times as large as the Lai and Robbins lower bound on the best regret achievable by any algorithm. This shows that the UCB algorithm is close to optimal.