Multi-armed Bandits and Upper Confidence Bound Algorithm

Introduction to Artificial Intelligence with Mathematics
Lecture Notes

Ganguk Hwang

Department of Mathematical Sciences
KAIST

Multi-armed Bandits

- We consider a bandit with K arms.
- Denote by $X_i(t), i \in \{1, 2, \cdots, K\}, t \in \mathbb{N}$, the random reward that we would get if arm i were played at time t.
- For simplicity, for each i we assume that $X_i(t)$'s are independent and indentically distributed and $X_i(t) \in [0, 1]$.
- $\mu_i, i \in \{1, 2, \cdots, K\}$ are the expected rewards from arms and they are unknown to us. Let $\mu^* = \max_i \mu_i$.
- $\Delta_i := \mu^* - \mu_i, i \geq 1$ are called the arm gaps.

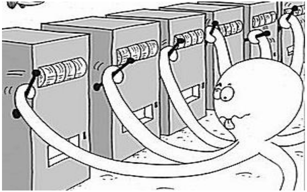

Figure: Multi-armed Bandit (source: MS research)

Denote by $I(t) \in \{1, 2, \cdots, K\}$ the arm played at time t and let

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

which is the number of times arm i has been played until time t,

$$R_i(t) := \sum_{s=1}^{t} X_i(s) 1\{I(s) = i\}$$

which is the total reward from arm i until time t, and

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

which is the estimated average reward from arm i until time t.

With K arms we want to maximize the cumulative reward $\sum_{i=1}^{K} R_i(t)$.

- The simplest one is to assume that the estimated average reward is a good accurate of μ_i.
- We take the arm with the largest estimated average reward.
- However, we need to consider the uncertainty in the estimation.
- To deal with the uncertainty, the Upper Confidence Bound (UCB) algorithm is proposed.
- The UCB algorithm is based on Hoeffding's inequality.

Theorem 1 (Hoeffding's Inequality)

Suppose that $X_1, \cdots, X_m$ are independent random random variables with $a_i \le X_i \le b_i$. Then

$$P\left\{\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{m}\sum_{i=1}^{m} E[X_i] > \epsilon\right\} \le \exp\left(\frac{-2\epsilon^2 m^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right).$$

To prove Hoeffding's inequality we need the following lemma.

Lemma 2 (Hoeffding's Lemma)

Given a random variable X with $a \le X \le b$ and $E[X] = 0$, for any $s > 0$ we have

$$E[e^{sX}] \le e^{\frac{s^2(b-a)^2}{8}}.$$

Proof of Hoeffding's Lemma: Given any x such that $a \le x \le b$, define $\lambda \in [0,1]$ by $\lambda = \frac{b-x}{b-a}$. Then, we have $x = b - \lambda(b-a) = \lambda a + (1-\lambda)b$ and

$$e^{sx} = e^{s\lambda a + s(1-\lambda)b} \le \lambda e^{sa} + (1-\lambda)e^{sb} = \frac{b-x}{b-a}e^{sa} + \frac{x-a}{b-a}e^{sb}.$$

Using the above and the fact that $E[X] = 0$

$$\begin{aligned} E[e^{sX}] &\le E\left[\frac{b-X}{b-a}e^{sa} + \frac{X-a}{b-a}e^{sb}\right] = \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &= (1-p)e^{sa} + pe^{sb} = (1-p+pe^{s(b-a)})e^{sa} \\ &= (1-p+pe^{s(b-a)})e^{-ps(b-a)} \end{aligned}$$

where $p = \frac{-a}{b-a} \in [0,1]$ ($a \le 0$ because $E[X] = 0$).

Let $u = s(b-a)$ and define

$$\phi(u) = -ps(b-a) + \log(1-p+pe^{s(b-a)}) = -pu + \log(1-p+pe^{u}).$$

From our previous inequality, we have that

$$E[e^{sX}] \leq e^{\phi(u)}.$$

To obtain the upper bound for $\phi(u)$, by Taylor's theorem there is some $z \in [0, u]$ such that

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{1}{2}u^2\phi''(z) \leq \phi(0) + u\phi'(0) + \frac{1}{2}u^2 \sup_v \phi''(v).$$

It can be easily checked that $\phi(0) = 0, \phi'(0) = 0$, and $\phi''(u) \leq \frac{1}{4}$ (c.f. $\phi''(u)$ is cancave for $v = e^u > 0$). It then follows that

$$\phi(u) \leq \frac{1}{2}u^2\frac{1}{4} = \frac{1}{8}s^2(b-a)^2$$

and hence

$$E[e^{sX}] \leq e^{\phi(u)} \leq e^{\frac{1}{8}s^2(b-a)^2}.$$

Proof of Hoeffding's Inequality: Let $Z_i = X_i - E[X_i]$. Then $E[Z_i] = 0$ and $a_i - E[X_i] \le Z_i \le b_i - E[X_i]$. For $s, t > 0$,

$$
\begin{aligned}
P\left\{\sum_{i=1}^{m} Z_i > t\right\} &= P\left\{\exp\left(s\sum_{i=1}^{m} Z_i\right) > \exp(st)\right\} \le \frac{1}{\exp(st)} E\left[\prod_{i=1}^{m} e^{sZ_i}\right] \\
&= \frac{1}{\exp(st)} \prod_{i=1}^{m} E\left[e^{sZ_i}\right] \le \exp(-st) \prod_{i=1}^{m} \exp\left(\frac{1}{8}s^2(b_i - a_i)^2\right) \\
&= \exp\left(\frac{1}{8}s^2\sum_{i=1}^{m}(b_i - a_i)^2 - st\right).
\end{aligned}
$$

By letting $s = \frac{4t}{\sum_{i=1}^{m}(b_i - a_i)^2}$ which minimizes the RHS, we obtain

$$
P\left\{\sum_{i=1}^{m} Z_i > t\right\} \le \exp\left(\frac{-2t^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right).
$$

The theorem immediately follows by letting $t = \epsilon m$.

Note that, using a similar argument we can also show that

$$P\left\{\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{m}\sum_{i=1}^{m} E[X_i] < -\epsilon\right\} \leq \exp\left(\frac{-2\epsilon^2 m^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right).$$

We now explain the UCB algorithm. We first observe from the above Hoeffding's inequality that

$$P\{\mu_i > \hat{\mu}_{i,n} + x\} \leq e^{-2nx^2}.$$

If we let $x = \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$ for sufficiently small $\delta > 0$, we obtain

$$P\{\mu_i \leq \hat{\mu}_{i,n} + x\} > 1 - \delta.$$

That is, with high probability $(1-\delta)$ the upper bound for μ_i is

$$\hat{\mu}_{i,n} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}.$$

In the UCB algorithm, we do not use a fixed conficence level δ, but to adapt it over time in the correct way as follows.

UCB(α) Algorithm

- In the first K time epochs, play each arm once in arbitrary order.
- At the end of each time epoch $t \geq K$, compute the UCB(α) index of each arm which is given by ($\alpha > 1$)

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

- At time epoch $t+1$, play the arm with the highest index, breaking ties arbitrarily. That is,

$$I(t+1) \in \arg\max_i \hat{\mu}_{i,N_i(t)} + \sqrt{\frac{\alpha \log t}{2N_i(t)}}.$$

Let

$$\mathcal{R}(t) = t\mu^* - \sum_{i=1}^{K} E[R_i(t)]$$

which is called the regret until time t. Then the UCB(α) Algorithm has the following upper bound:

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

- The regret grows very slowly with time t; it only grows logarithmically in t.
- When t is large, the second term dominates, so we choose α as small as possible. However, the first term blows up to infinity as α goes down to 1. This is a trade-off in the choice of α.
- In practice, we use $\alpha = 2$ (a little bigger than 1).