1 The Multi-armed Bandit Problem

The Multiarmed Bandit (MAB) problem is essentially the problem of online learning, when we only have access to partial observations with time. More formally, the setting is as follows. The algorithm has access to a set of $K$ actions with time. At each time $t$, the algorithm selects an action $i_t$ in this set. The payoff $g_{i_t,t}$ of action $i_t$ is then revealed to the algorithm. Recall that this is different from our standard online learning setting, in which the algorithm would have had access to the payoff of all the actions at time $t$.

The goal of the algorithm is to select a sequence of actions $i_1, i_2, \ldots$ so as to have low regret to the action with the highest cumulative payoff in hindsight. More formally, the regret of the algorithm to the best action is defined as:

$$ R(T) = \max_i \sum_{t=1}^{T} g_{i,t} - \sum_{t=1}^{T} g_{i_t,t} $$

and the goal of the algorithm is to select a sequence $g_{i_t,t}$ so that this value is as low as possible.

What are some applications of the MAB problem? A popular application is in placing ads on a webpage. Suppose we would like to advertise some products by placing ads on a website; we would like to place only one ad for each user so as to not confuse the user, and there are $K$ possible ads that we could display. One (fairly simplistic) model for this problem is to model it as a MAB problem. Here each action is an ad; when an user arrives, one out of the $K$ ads is displayed; this is equivalent to picking one action out of the available $K$. The payoff of the algorithm is 1 if the user clicks on the displayed ad and 0 otherwise, and the algorithm’s goal is to get as almost as many clicks as that of the ad invoking the highest number of clicks.

In the learning theory literature, we study two kinds of MAB problems.

1. In the Stochastic MAB problem, each action $i$ has associated with it a distribution $D_i$. The support of $D_i$ is the interval $[0,1]$ and the expectation $E_{X \sim D_i}[X]$ is $\mu_i$. At each time $t$, $g_{i,t}$, the payoff of action $i$, is drawn independently from $D_i$; thus the $g_{i,t}$ are iid over $t$. Moreover, at any time $t$, $g_{i,t}$ and $g_{j,t}$ are also independent, for $i \neq j$. Notice that this is a very strong assumption on the payoffs.

2. In the Non-Stochastic MAB problem, we make no probabilistic assumptions on the payoffs.

   We assume that the payoffs of all actions are decided by an adversary at time $t = 0$ – that is, before the algorithm begins. The algorithm’s goal is then to minimize the regret with respect to the action with the highest cumulative payoff in hindsight.

Another feature of algorithms for the MAB problem is that very often (although not always) the algorithms are randomized. Therefore, two possible notions of regret are typically considered. The first one is Expected Regret, where we would like the regret to be low on expectation. The second one is Regret with High Probability, where we would like the regret to be low with probability $1 - \delta$. 


1. Initialize: Pick each action once in order. For action \( j \), let \( \bar{x}_j = g_{j,j} \), and \( n_j = 1 \).

2. For \( t = K + 1, K + 2, \ldots, \)
   (a) Select action \( j_t = j \) that maximizes \( \bar{x}_j + \sqrt{\frac{2\ln t}{n_j}} \).
   (b) Receive payoff \( g_{j_t,t} \).
   (c) Update: \( n_{j_t} = n_{j_t} + 1 \), and \( \bar{x}_{j_t} = (1 - \frac{1}{n_{j_t}})\bar{x}_{j_t} + \frac{1}{n_{j_t}}g_{j_t,t} \).

Figure 1: The Upper Confidence Bounds Algorithm.

for some small \( \delta \). Usually, a bound on the regret whp also implies a small expected regret; however, we still study algorithms which have small expected regret because they are typically simpler and easier to analyze.

1.1 The Stochastic MAB Problem and the UCB Algorithm

First let us look at the Stochastic MAB problem. Recall that there are \( K \) actions, and at each time \( t \), the payoff \( g_{i,t} \) for each action \( i \) is drawn iid from a distribution \( D_i \). Moreover for a specific \( t, g_{i,t} \) and \( g_{j,t} \) are also independent. We also assume that \( D_i \) has support in the interval \([0, 1]\); that is, for all \( i \) and \( t \), \( 0 \leq g_{i,t} \leq 1 \). Moreover, we let \( \mathbb{E}[g_{i,t}] = \mu_i \) for all \( i \), and \( \mu^* = \max_i \mu_i \).

If \( T_i(T) \) denotes the number of times action \( i \) is selected before time \( T \), then, we are interested in the following measure of regret after time \( T \):

\[
R(T) = \mu^*T - \sum_i \mu_i \mathbb{E}[T_i(T)]
\]

One of the simplest algorithms for this problem is the Upper Confidence Bounds (UCB) algorithm. The algorithm is stated in Figure 1.

Notice, that in the UCB Algorithm, \( n_j \) is the number of times action \( j \) has been picked so far, and \( \bar{x}_j \) is the empirical mean of the payoffs obtained from action \( j \) so far. A feature of the algorithm is that it is completely deterministic; therefore, the expectation in the expected regret calculation is over the stochasticity of the payoffs.

Why is the algorithm called Upper Confidence Bound? To understand this, let us take a brief detour to see what confidence bounds are.

**Theorem 1 (Confidence Bounds)** Let \( D \) be a distribution with support in \([0, 1]\), and let \( \mathbb{E}_{X \sim D}[X] = \mu \). Let \( X_1, \ldots, X_m \) be drawn independently from \( D \), and let \( \bar{X} = \frac{1}{m} \sum_i X_i \). Then,

\[
\Pr \left[ \mu \leq \bar{X} + \sqrt{\frac{\ln(1/\delta)}{2m}} \right] \geq 1 - \delta
\]

and

\[
\Pr \left[ \mu \geq \bar{X} - \sqrt{\frac{\ln(1/\delta)}{2m}} \right] \geq 1 - \delta
\]

The quantity \( \bar{X} - \sqrt{\frac{\ln(2/\delta)}{2m}} \) is called the Lower Confidence Bound, while the quantity \( \bar{X} + \sqrt{\frac{\ln(2/\delta)}{2m}} \) is called an Upper Confidence Bound for \( \mu \). The term confidence comes from the fact that we are confident of these inequalities with probability \( 1 - \delta \). (Typically \( \delta \) is very small.)
Now it is easy to see why the algorithm is called Upper Confidence Bounds; at time $t$, the upper bound on the confidence interval of the payoff of action $j$ is exactly $\bar{x}_j + \sqrt{\frac{2 \ln t}{n_j}}$, for $\delta = \frac{1}{t^4}$. The algorithm thus guesses that the payoff of each action $i$ is the upper limit of its confidence interval of the empirical mean payoff for that action; the best action is then selected based on this guess.

We can show the following performance bounds for the UCB algorithm.

**Theorem 2** Suppose that UCB is run on $K$ actions, with arbitrary payoff distributions $D_1, \ldots, D_K$ with support on $[0, 1]$. Moreover, let $\mu^* = \max_i \mu_i$, and let $\Delta_j = \mu^* - \mu_i$. Then, the expected regret of UCB after time $T$ is at most:

$$8 \sum_{i: \mu_i < \mu^*} \left( \frac{\ln T}{\Delta_i^2} \right) + \left( 1 + \frac{\pi^2}{3} \right) \sum_{i=1}^K \Delta_i$$

To prove this theorem, we first prove bounds on $\mathbb{E}[T_i(T)]$ for all $i$. In particular, we show the following:

**Lemma 1** If $\mu_i < \mu^*$, then,

$$\mathbb{E}[T_i(T)] \leq 8 \left( \frac{\ln T}{\Delta_i^2} \right) + \left( 1 + \frac{\pi^2}{3} \right)$$

Before we prove Lemma 1, we state a theorem which is useful in our proof.

**Theorem 3 (Martingale Bounds)** Let $X_1, \ldots, X_n$ be random variables with a common range $[0, 1]$, such that $\mathbb{E}[X_i \mid X_1, \ldots, X_{i-1}] = \mu$. Moreover, let $S_n = X_1 + \ldots + X_n$. Then, for any $a > 0$,

$$\Pr[S_n \geq n\mu + a] \leq e^{-2a^2/n}$$

and

$$\Pr[S_n \leq n\mu - a] \leq e^{-2a^2/n}$$

**Exercise 1** Derive a proof of Theorem 1 using Theorem 3.

To prove Lemma 1, observe that there are essentially two temporal phases with respect to action $i$. The first phase is when $T_i(t) < \frac{8 \ln T}{\Delta_i^2}$, that is, when we have picked action $i$ less than $\frac{8 \ln T}{\Delta_i^2}$ times. The second phase is when $T_i(t) \geq \frac{8 \ln T}{\Delta_i^2}$. To show the bound on $T_i(T)$, we essentially show that in the second phase, we select action $i$ only a very few times in expectation.

**First Phase.** Suppose that for a specific $t \leq T$, $T_i(t - 1) \geq 8 \left( \frac{\ln T}{\Delta_i^2} \right)$. If there is no such $t$, (that is, at time $T - 1$, we are still in the first phase), then $T_i(T) \leq 1 + \frac{8 \ln T}{\Delta_i^2}$, and we are done.

**Second Phase.** Now we analyze what happens when this is not the case; that is, we are in the second phase at time $T - 1$. Let $i^*$ be the action with the highest value of $\mu_i$. For a specific $s$ and any action $i$, we define $\bar{X}_i^s$ to be the empirical mean of the payoffs of $i$ over the first $s$ times $i$ is selected. Moreover, we define $\bar{X}_{i^*}^s = \bar{X}_{i^*}^s$.

Let us fix a time $t$, and suppose that action $i$ is picked $s_i$ times before $t$, and action $i^*$ is picked $s$ times before $t$. Under what conditions do we pick action $i$ over $i^*$ at time $t$? This only happens if:

$$\bar{X}_i^s + \sqrt{\frac{2 \ln t}{s_i}} \leq \bar{X}_{i^*}^s + \sqrt{\frac{2 \ln t}{s}}$$

(1)
Moreover, this holds when at least one of the following equations hold:

\[ \bar{X}_* \leq \mu^* - \sqrt{\frac{2 \ln t}{s}} \]  
\[ \bar{X}_{s_i} \geq \mu_i + \sqrt{\frac{2 \ln t}{s_i}} \]  
\[ \mu^* < \mu_i + 2\sqrt{\frac{2 \ln t}{s_i}} \]  

Otherwise, if none of the Equations 2, 3 and 4 hold, then,

\[ X_* + \sqrt{\frac{2 \ln t}{s}} > \mu^* \geq \mu_i + 2\sqrt{\frac{2 \ln t}{s_i}} > X_{s_i} + \sqrt{\frac{2 \ln t}{s_i}} \]

which implies that the action \( i^* \) is picked instead of \( i \) at time \( t \).

When do Equations 2, 3 and 4 hold?

For Equation 2, observe that we can apply the Martingale Bounds. Set \( X_j \) as \( \frac{1}{s} \) times the payoff when action \( i^* \) is selected for the \( j \)-th time, \( n = s \), and \( S_n = \bar{X}_* \). Recall that as the \( g_{i,t} \) are drawn independently, the conditions in the Martingale bounds are satisfied. Therefore,

\[ \Pr \left( \bar{X}_* \geq \mu^* - \sqrt{\frac{2 \ln t}{s}} \right) \leq e^{-2s \cdot 2 \ln t / s} = t^{-4} \]

that is, at time \( t \), Equation 2 only holds with probability at most \( t^{-4} \).

For Equation 3, we again apply the Martingale Bounds. Here, \( X_j \) is \( \frac{1}{s_i} \) times the payoff when action \( i \) is selected for the \( j \)-th time, \( n = s_i \), and \( S_n = \bar{X}_{s_i} \). Therefore,

\[ \Pr \left( \bar{X}_{s_i} \leq \mu_i + \sqrt{\frac{2 \ln t}{s_i}} \right) \leq e^{-2s_i \cdot 2 \ln t / s_i} \leq t^{-4} \]

that is, at time \( t \), Equation 3 also holds with probability at most \( t^{-4} \).

For Equation 4, we observe that if we are in the second phase for action \( i \), \( s_i \geq 8 \left( \frac{\ln t}{\Delta^2} \right) \), and therefore

\[ \mu^* - \mu_i - 2\sqrt{\frac{2 \ln t}{s_i}} \geq \mu^* - \mu_i > 0 \]

Therefore, Equation 4 never holds, so long as we are in the second phase for action \( i \).

Now we take a look at \( T_i(T) \), when we are in the second phase for action \( i \) at time \( T - 1 \). Observe that:

\[
T_i(T) \leq \left\lceil \frac{8 \ln T}{\Delta^2_i} \right\rceil + \sum_{t=\frac{8 \ln T}{\Delta^2_i}}^{T} 1_{s_i \text{ picked at time } t} \\
\leq \left\lceil \frac{8 \ln T}{\Delta^2_i} \right\rceil + \sum_{t=\frac{8 \ln T}{\Delta^2_i}}^{T} 1_{s \text{, } s_i \text{ s.t Equation 1 holds?}}
\]

In the equation above, the notation \( 1_S \) denotes the indicator function for the event \( S \); namely it is 1 when \( S \) happens and 0 otherwise.

We now compute the expectation of \( T_i(T) \). Notice that:
\[
\mathbb{E}[T_i(T)] \leq \left[ \frac{8 \ln T}{\Delta_i^2} \right] + \sum_{t=\text{lin} T/\Delta_i^2}^{T} \sum_{s=1}^{t} \sum_{s_{-i}=\text{lin} T/\Delta_i^2}^{t} \Pr \left( \bar{X}_s^* + \sqrt{\frac{2 \ln t}{s}} \leq \bar{X}_s^i + \sqrt{\frac{2 \ln t}{s_i}} \right)
\]

Here the bounds on the summations on \( t \) and \( s_i \) occur because we are in the second phase for action \( i \). As the event Equation 1 holds only when either Equation 2 or Equation 3 hold,

\[
\Pr \left( \bar{X}_s^* + \sqrt{\frac{2 \ln t}{s}} \leq \bar{X}_s^i + \sqrt{\frac{2 \ln t}{s_i}} \right) \leq \Pr \left( \bar{X}_s^* \leq \mu^* - \sqrt{\frac{2 \ln t}{s}} \right) + \Pr \left( \bar{X}_s^i \geq \mu_i + \sqrt{\frac{2 \ln t}{s_i}} \right)
\]

\[
\leq 2t^{-4}
\]

\[
\mathbb{E}[T_i(T)] \leq \left[ \frac{8 \ln T}{\Delta_i^2} \right] + \sum_{t=\text{lin} T/\Delta_i^2}^{T} \sum_{s=1}^{t} \sum_{s_{-i}=\text{lin} T/\Delta_i^2}^{t} 2t^{-4}
\]

\[
\leq \left[ \frac{8 \ln T}{\Delta_i^2} \right] + \sum_{t} 2t^{-2}
\]

\[
\leq 1 + \frac{8 \ln T}{\Delta_i^2} + \frac{\pi^2}{3}
\]

Thus, we have that for all \( i \) such that \( \mu_i < \mu^* \),

\[
\mathbb{E}[T_i(T)] \leq 8 \left( \frac{\ln T}{\Delta_i^2} \right) + \left( 1 + \frac{\pi^2}{3} \right)
\]

which proves Lemma 1.

Now we combine the bounds from Lemma 1 to prove Theorem 2. Recall that after time \( T \), the regret with respect to \( \mu^* \) is \( \sum_{i: \mu_i < \mu^*} \mathbb{E}[T_i(T)] \Delta_i \). This regret is at most:

\[
R(T) = \sum_{i: \mu_i < \mu^*} \mathbb{E}[T_i(T)] \Delta_i
\]

\[
\leq \sum_{i: \mu_i < \mu^*} 8 \left( \frac{\ln T}{\Delta_i^2} \right) + \left( 1 + \frac{\pi^2}{3} \right) \sum_i \Delta_i
\]

from which Theorem 2 follows.