Introduction to Stochastic Multi-Armed Bandits

Stefano Tracá, Cynthia Rudin

Thanks to Tianyu Wang for error checking.

1 Problem setup

The name "multi-armed bandit" comes from the name of a gambling machine. You can choose one of the arms (levers) of the machine at each round, and get a reward based on which arm you choose. The rewards for each arm are iid from a distribution, and each arm has its own distribution. If one of the arms is better than the rest it would be good to always pull that arm, but you don't know which one it is! So you need to divide your time between *exploring* arms that you think might be good with *exploiting* arms that you know are good.

There are many applications for MAB, including recommender systems. For instance, the New York Times uses MAB to determine which news article to show you on your telephone. Usually MAB is considered to be an alternative to massive A-B testing. Say you want to optimize the look of your website, but there are many possible website options to consider. To determine which one is the best, you might usually do pairwise hypothesis testing between all pairs. This will take forever, so you might want to run a MAB instead, which in some sense conducts all the tests at once. Clinical trials also can use MAB. We can give many different drugs to the patients, and we can use MAB to find the best drugs, based on the performance of these drugs over the course of the trial, without having to do pairwise tests.

Formally, the stochastic multi-armed bandit problem is a game played in n rounds. At each round t the player chooses an action among a finite set of m possible choices called *arms*. When arm j is played ($j \in \{1, \cdots, m\}$) a random reward $X_j(t)$ is drawn from an unknown distribution. The distribution of $X_j(t)$ does not change with time (the index t is just used to indicate in which turn the reward was drawn). At the end of each turn the player can update her estimate of the mean reward of arm j:

$$\widehat{X}_j(t) = \frac{1}{T_j(t-1)} \sum_{s=1}^{t-1} X_j(s) \mathbb{1}_{\{I_t = j\}}, \tag{1}$$

where $T_j(t-1)$ is the number of times arm j has been played before round t starts, and $\mathbb{1}_{\{I_t=j\}}$ is an indicator function equal to 1 if arm j is played at time t (otherwise its value is 0). After a while, this empirical mean will be close to the arm's mean reward. Updating these estimates after each round will help the player in choosing a good arm in the next round.

At each turn, the player suffers a possible regret from not having played the best arm. If they had chosen the best arm, their reward would have been $X^*(t)$ where notation * means the best arm. The total regret at the end of the game is given by

$$R_n^{\text{(raw)}} = \sum_{t=1}^{n} \sum_{j=1}^{m} [X^*(t) - X_j(t)] \mathbb{1}_{\{I_t=j\}}, \tag{2}$$

where $X^*(t)$ is the reward of the best arm at time t if it would have been played at time t.

We don't usually define the regret this way though when doing theory. We usually assign the regret to be based on the means of the arms distributions. So let's try it again:

$$R_n = \sum_{t=1}^{n} \sum_{j=1}^{m} [\mu_* - \mu_j(t)] \mathbb{1}_{\{I_t=j\}}, \tag{3}$$

The mean regret for having played arm j is given by $\Delta_j = \mu_* - \mu_j$, where μ_* is the mean reward of the best arm and μ_j is the mean reward obtained when playing arm j. So the regret is now:

$$R_n = \sum_{t=1}^{n} \sum_{j=1}^{m} \Delta_j \mathbb{1}_{\{I_t=j\}}, \tag{4}$$

The strategies presented in the following sections aim to minimize the expected cumulative regret $\mathbb{E}[R_n]$, where the expectation is over the random draw of the arms. (The algorithm reacts to these random draws, so the choice of arms I_t also then becomes random.)

$$\mathbb{E}[R_n] = \mathbb{E}\sum_{t=1}^{n}\sum_{j=1}^{m}\Delta_j \mathbb{1}_{\{I_t=j\}}, \tag{5}$$

A complete list of the symbols used can be found in Appendix C.

2 ε-greedy algorithm

Algorithm 2: ε-greedy algorithm

```
Input          : number of rounds $n$, number of arms $m$, a constant $k$ such that $k > \max\{10, \frac{4}{\min_j \Delta_j^2}\}$, sequence
                 $\{\varepsilon_t\}_{t=1}^{n} = \min\left\{1, \frac{km}{t}\right\}$
Initialization: play all arms once and initialize $\widehat{X}_j(m)$ (defined in (1)) for each $j = 1, \cdots, m$
for $t = m+1$ to $n$ do
    With probability $\varepsilon_t$ play an arm uniformly at random (each arm has probability $\frac{1}{m}$ of being selected),
    otherwise (with probability $1 - \varepsilon_t$) play arm $j$ such that
        $\widehat{X}_j(t-1) \geq \widehat{X}_i(t-1)\ \forall i$
    Get reward $X_j(t)$;
    Update $\widehat{X}_j(t)$;
end
```

The first algorithm we consider is called ε-greedy, and it is in Algorithm 2. The idea is very simple: with some small probability, play an arm uniformly at random. Otherwise, pick the arm that we think is the best.

You will notice that there are some interesting terms in the algorithm, defining the choice of ϵ_t. They are chosen that way so that we can get a tight bound on the regret of the algorithm.

Theorem 2.1 shows that the regret of ε-greedy is bounded by terms that are logarithmic in t. You can see this because the ε_t are $\theta\left(1/t\right)$ which means their sum over time is logarithmically bounded (because $\log t$ is the integral of $1/t$), while the $\beta_j(t)$ term is $o\left(1/t\right)$. (To see this, you need the assumptions we made about ε_t.)

Theorem 2.1 (Regret-bound for ε-greedy algorithm - Auer, Cesa-Bianchi, and Fischer, Finite-time analysis of the multiarmed bandit problem, 2002)**.** *The bound on the mean regret $\mathbb{E}[R_n]$ at time n is given by*

$$\mathbb{E}[R_n] \leq \sum_{j=1}^{m}\Delta_j \tag{6}$$

$$+ \sum_{t=m+1}^{n}\sum_{j:\mu_j<\mu_*}\Delta_j\left(\varepsilon_t\frac{1}{m} + (1-\varepsilon_t)\beta_j(t)\right) \tag{7}$$

where

$$\beta_j(t) = k\left(\frac{t}{mke}\right)^{-\frac{k}{10}}\log\left(\frac{t}{mke}\right) + \frac{4}{\Delta_j^2}\left(\frac{t}{mke}\right)^{-\frac{k\Delta_j^2}{4}}. \tag{8}$$

The sum in (6) is the exact mean regret during the initialization of Algorithm 1. For the rounds after the initialization phase, the quantity in the parenthesis of (7) is an upper bound on the probability of playing arm j. In the bound, $\beta_j(t)$ is an upper bound on the probability that our algorithm thinks arm j is the best arm at round t, and $1/m$ is the probability of choosing arm j when the choice is made at random. Proof in Appendix A.

3 The UCB algorithm

The UCB algorithm is also very simple. It creates a confidence interval on the mean reward. At each round, it chooses the arm with the highest upper confidence interval. This is because any arm with a high upper confidence bound could be the best arm.

Algorithm 3: UCB algorithm

```
Input          : number of rounds $n$, number of arms $m$
Initialization: play all arms once and initialize $\widehat{X}_j$ (as defined in (1)) for each $j = 1, \cdots, m$
for $t = m + 1$ to $n$ do
    play arm $j$ with the highest upper confidence bound on the mean estimate:
        $\widehat{X}_j(t-1) + \sqrt{\frac{2\log(t)}{T_j(t-1)}}$;
    Get reward $X_j$;
    Update $\widehat{X}_j$;
end
```

The bound grows logarithmically in n.

Theorem 3.1 (Regret-bound of the UCB algorithm - Auer, Cesa-Bianchi, and Fischer, Finite-time analysis of the multiarmed bandit problem, 2002)**.** *The bound on the mean regret $\mathbb{E}[R_n]$ at time n is given by*

$$\mathbb{E}[R_n] \quad \leq \quad \sum_{j=1}^{m} \Delta_j + \sum_{j:\mu_j<\mu_*} \frac{8}{\Delta_j} \log(n) + \sum_{j=1}^{m} \Delta_j \left(\sum_{t=m+1}^{n} 2t^{-4}(t-1-m)^2 \right). \tag{9}$$

Proof in Appendix B.

A Regret-bound of the ε-greedy algorithm

The mean regret at round n is given by

$$R_n = \sum_{t=1}^{n} \sum_{j=1}^{m} \Delta_j \mathbb{1}_{\{I_t = j\}}, \tag{10}$$

where $\mathbb{1}_{\{I_t=j\}}$ is an indicator function equal to 1 if arm j is played at time t (otherwise its value is 0) and $\Delta_j = \mu^* - \mu_j$ is the difference between the mean of the best arm's reward distribution and the mean of the j's arm reward distribution. By taking the expectation we have that

$$\mathbb{E}[R_n] = \sum_{t=1}^{n} \sum_{j=1}^{m} \Delta_j \mathbb{P}(\{I_t = j\})$$

which can be rewritten as

$$\mathbb{E}[R_n] \quad = \quad \sum_{t=1}^{n} \sum_{j=1}^{m} \Delta_j \left[\varepsilon_t \frac{1}{m} + (1-\varepsilon_t) \mathbb{P}(\widehat{X}_{j,T_j(t-1)} \geq \widehat{X}_{i,T_i(t-1)} \ \ \forall i) \right], \tag{11}$$

where notation $\widehat{X}_{i,T_i(t-1)}$ is the estimated mean for arm i after it has been chosen $T_i(t-1)$ times up to time $t-1$. The first term is the probability that we choose arm j by exploring. We explore with probability ε_t and if we explore, we chose j at random, that is, with probability $1/m$. If we chose j while exploiting, which happens with prob $1-\varepsilon_t$, then its average reward is above that of all the other arms.

For this proof, we assume the rewards are bounded, say between 0 and 1. If they are bounded by something bigger than 1, we would have an extra constant scaling factor in the theorem.

STEP 1: Conditions when we think arm j is the best at time t. If we think arm j is the best at time t, then either we overestimated its mean reward, or we underestimated the reward of the best arm, which is called arm $*$. If neither of those things occurred, arm j's rewards would have been below those of arm $*$ and thus we would not think that arm j is the best when it isn't. In the first inequality below, we consider the probability arm j has average reward above all the other arms, and this is less than the probability that arm j has reward greater than just one of those arms (in particular, arm $*$).

$$\begin{aligned} \mathbb{P}(\widehat{X}_{j,T_j(t-1)} \geq \widehat{X}_{i,T_i(t-1)} \ \ \forall i) \quad &\leq \quad \mathbb{P}(\widehat{X}_{j,T_j(t-1)} \geq \widehat{X}_{*,T_*(t-1)}) &(12)\\ &\leq \quad \mathbb{P}\left(\widehat{X}_{j,T_j(t-1)} \geq \mu_j + \frac{\Delta_j}{2}\right) + \mathbb{P}\left(\widehat{X}_{*,T_*(t-1)} \leq \mu_* - \frac{\Delta_j}{2}\right), &(13) \end{aligned}$$

where the last inequality follows from the fact that either we must have underestimated arm $*$ or overestimated arm j:

$$\left\{\widehat{X}_{j,T_j(t-1)} \geq \widehat{X}_{*,T_*(t-1)}\right\} \subset \left(\left\{\widehat{X}_{*,T_*(t-1)} \leq \mu_* - \frac{\Delta_j}{2}\right\} \cup \left\{\widehat{X}_{j,T_j(t-1)} \geq \mu_j + \frac{\Delta_j}{2}\right\}\right). \tag{14}$$

In fact, suppose that there exist an element $\omega \in \left\{\widehat{X}_{j,T_j(t-1)} \geq \widehat{X}_{*,T_*(t-1)}\right\}$ that does not belong to $\left(\left\{\widehat{X}_{*,T_*(t-1)} \leq \mu_* - \frac{\Delta_j}{2}\right\} \cup \left\{\widehat{X}_{j,T_j(t-1)} \geq \mu_j + \frac{\Delta_j}{2}\right\}\right)$. Then, we would have that

$$\begin{aligned} \omega \quad &\in \quad \left(\left\{\widehat{X}_{*,T_*(t-1)} \leq \mu_* - \frac{\Delta_j}{2}\right\} \cup \left\{\widehat{X}_{j,T_j(t-1)} \geq \mu_j + \frac{\Delta_j}{2}\right\}\right)^C &(15)\\ &= \quad \left\{\widehat{X}_{*,T_*(t-1)} > \mu_* - \frac{\Delta_j}{2}\right\} \cap \left\{\widehat{X}_{j,T_j(t-1)} < \mu_j + \frac{\Delta_j}{2}\right\}, &(16) \end{aligned}$$

but from the intersection of events given in (16) it follows that $\widehat{X}_{*,T_*(t-1)} > \mu_* - \frac{\Delta_j}{2} = \mu_j + \frac{\Delta_j}{2} > \widehat{X}_{j,T_j(t-1)}$ which contradicts $\omega \in \left\{\widehat{X}_{j,T_j(t-1)} \geq \widehat{X}_{*,T_*(t-1)}\right\}$.

Therefore, all elements of $\left\{\widehat{X}_{j,T_j(t-1)} \geq \widehat{X}_{*,T_*(t-1)}\right\}$ belong to $\left(\left\{\widehat{X}_{*,T_*(t-1)} \leq \mu_* - \frac{\Delta_j}{2}\right\} \cup \left\{\widehat{X}_{j,T_j(t-1)} \geq \mu_j + \frac{\Delta_j}{2}\right\}\right)$.

STEP 2: Let us bound the probability of overestimating sub-optimal arm j at time t. Let us consider the first term of (13). The computations for the second term are basically identical.

$$\begin{aligned}\mathbb{P}\left(\widehat{X}_{j,T_j(t-1)} \geq \mu_j + \frac{\Delta_j}{2}\right) &= \sum_{s=1}^{t-1} \mathbb{P}\left(T_j(t-1) = s, \widehat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right)\\ &= \sum_{s=1}^{t-1} \mathbb{P}\left(T_j(t-1) = s | \widehat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right) \mathbb{P}\left(\widehat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right)\\ &\leq \sum_{s=1}^{t-1} \mathbb{P}\left(T_j(t-1) = s | \widehat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right) e^{-\frac{\Delta_j^2}{2}s}, \qquad (17)\end{aligned}$$

where in the last inequality we used the Chernoff-Hoeffding bound. The second term will be small when s is large, so that term will be sufficient to handle whatever the first term brings when s is large. When s is small, the first term could be problematic since it will be large. We're going to separate this sum into large s and small s and handle them separately. Here, small s means less than x_0, where we define it as:

$$x_0 := \frac{1}{2m} \sum_{s=1}^{t} \varepsilon_s.$$

Then

$$(17) \quad \leq \quad \sum_{s=1}^{\lfloor x_0 \rfloor} \mathbb{P}\left(T_j(t-1) = s | \widehat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right) \cdot 1 + \sum_{s=\lfloor x_0 \rfloor + 1}^{t-1} 1 \cdot e^{-\frac{\Delta_j^2}{2}s}. \qquad (18)$$

Here, we split the sum into two pieces and bounded one of the terms by 1.

Let us work on the second term. We will now use the fact that $\sum_{s=\lfloor x_0 \rfloor+1}^{\infty} e^{-bs} \leq \frac{1}{b} e^{-b\lfloor x_0 \rfloor}$, where in our case $b = \frac{\Delta_j^2}{2}$.

$$(17) \quad \leq \quad \sum_{s=1}^{\lfloor x_0 \rfloor} \mathbb{P}\left(T_j(t-1) = s | \widehat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right) + \frac{2}{\Delta_j^2} e^{-\frac{\Delta_j^2}{2}\lfloor x_0 \rfloor}. \qquad (19)$$

Now comes a trick. Let us define $T_j^R(t-1)$ as the number of times arm j is played when we are performing exploration. Note that $T_j^R(t-1) \leq T_j(t-1)$ and that $T_j^R(t-1) = \sum_{s=1}^{t-1} B_s$ where B_s is a Bernoulli r.v. with parameter ε_s/m (this is the probability that we explore times the probability that we choose arm j when exploring, so ϵ_s times $1/m$. In that case, $T_j^R(t-1)$ equals a values less than s but we don't know which one. Luckily we're constructing upper bounds. So we add up all possibilities for it.

$$(17) \quad \leq \quad \sum_{s=1}^{\lfloor x_0 \rfloor} \mathbb{P}\left(T_j^R(t-1) \leq s | \widehat{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right) + \frac{2}{\Delta_j^2} e^{-\frac{\Delta_j^2}{2}\lfloor x_0 \rfloor}. \qquad (20)$$

Now things are good, since the number of times we explore to choose arm j, $T_j^R(t-1)$, does not depend on the estimate of the mean for arm j. The number of terms in the sum is $\lfloor x_0 \rfloor$:

$$(17) \quad \leq \quad \lfloor x_0 \rfloor \mathbb{P}\left(T_j^R(t-1) \leq \lfloor x_0 \rfloor\right) + \frac{2}{\Delta_j^2} e^{-\frac{\Delta_j^2}{2}\lfloor x_0 \rfloor}. \qquad (21)$$

We have that

$$\mathbb{E}[T_j^R(t-1)] = \frac{1}{m} \sum_{s=1}^{t} \varepsilon_s, \quad Var(T_j^R(t-1)) = \sum_{s=1}^{t} \frac{\varepsilon_s}{m}\left(1 - \frac{\varepsilon_s}{m}\right) \leq \frac{1}{m} \sum_{s=1}^{t} \varepsilon_s = \mathbb{E}[T_j^R(t-1)],$$

and, using the Bernstein inequality $\mathbb{P}(S_n \leq \mathbb{E}[S_n] - a) \leq \exp\{-\frac{a^2/2}{\sigma^2 + a/2}\}$ with $S_n = T_j^R(t-1)$ and $a = \frac{1}{2}\mathbb{E}[T_j^R(t-1)]$,

$$\begin{aligned}
\mathbb{P}(T_j^R(t-1) \leq \lfloor x_0 \rfloor) &\leq \mathbb{P}(T_j^R(t-1) \leq x_0) \\
&= \mathbb{P}\left(T_j^R(t-1) \leq \mathbb{E}[T_j^R(t-1)] - \frac{1}{2}\mathbb{E}[T_j^R(t-1)]\right) \\
&\leq \exp\left\{-\frac{\frac{1}{8}(\mathbb{E}[T_j^R(t-1)])^2}{\mathbb{E}[T_j^R(t-1)] + \frac{1}{4}\mathbb{E}[T_j^R(t-1)]}\right\} \\
&= \exp\left\{-\frac{4}{5}\frac{1}{8}\mathbb{E}[T_j^R(t-1)]\right\} \leq \exp\left\{-\frac{1}{5}x_0\right\}. \qquad (22)
\end{aligned}$$

STEP 3: To upper bound (22), let us find a lower bound on $\lfloor x_0 \rfloor$. Let us define $n' = \lfloor km \rfloor + 1$, then

$$\begin{aligned}
x_0 &= \frac{1}{2m}\sum_{s=1}^{t} \varepsilon_s \\
&= \frac{1}{2m}\sum_{s=1}^{t} \min\left\{1, \frac{km}{s}\right\} \\
&= \frac{1}{2m}\sum_{s=1}^{n'} 1 + \frac{km}{2m}\sum_{s=n'+1}^{t}\frac{1}{s} \qquad (23) \\
&= \frac{n'}{2m} + \frac{k}{2}\left(\sum_{s=1}^{t}\frac{1}{s} - \sum_{s=1}^{n'}\frac{1}{s}\right)
\end{aligned}$$

Here we will use some properties of harmonic sequences, namely $\sum_{t=1}^{n}\frac{1}{t} \leq \log n + 1$ and $\sum_{t=1}^{n}\frac{1}{t} > \int_1^{n+1}\frac{1}{t}dt = \ln(n+1)$.

$$\begin{aligned}
&\geq \frac{n'}{2m} + \frac{k}{2}\left(\log(t+1) - (\log(n') + \log(e)\right) \\
&\geq \frac{k}{2}\log\left(\frac{n'}{m}\frac{1}{k}\right) + \frac{k}{2}\log\left(\frac{t}{n'e}\right) \\
&= \frac{k}{2}\log\left(\frac{t}{mke}\right). \qquad (24)
\end{aligned}$$

Remark 1. Note that if t was less than n', then we would have $x_0 = t/2m$, yielding an exponential decay of the bound on the probability of j being the best arm. To see this, $t < n'$ would imply that, using (21) and (22),

$$(21) \leq \frac{t}{2m}\exp\left\{-\frac{1}{5}\frac{t}{2m}\right\} + \frac{2}{\Delta_j^2}\exp\left\{-\frac{\Delta_j^2}{2}\frac{t}{2m}\right\}.$$

Continuing the proof of Theorem 3.1, we obtain a bound on the first term in (13) as follows. Using (24) combined with (22) in (21), we get

$$\frac{k}{2}\left(\frac{t}{mke}\right)^{-\frac{k}{10}}\log\left(\frac{t}{mke}\right) + \frac{2}{\Delta_j^2}\left(\frac{t}{mke}\right)^{-\frac{k\Delta_j^2}{4}}. \qquad (25)$$

STEP 4: Let us bound the probability of underestimating sub-optimal arm j at time t. Since the computations for the second term in (13) are essentially identical, by removing the $1/2$ factor we get this bound on $\mathbb{P}\left(\widehat{X}_{j,T_j(t-1)} \geq \widehat{X}_{i,T_i(t-1)} \quad \forall i\right)$:

$$\beta_j(t) = k\left(\frac{t}{mke}\right)^{-\frac{k}{10}}\log\left(\frac{t}{mke}\right) + \frac{4}{\Delta_j^2}\left(\frac{t}{mke}\right)^{-\frac{k\Delta_j^2}{4}}. \qquad (26)$$

STEP 5: Let us bound the probability of playing suboptimal arm j. We have now an upper bound for $\mathbb{P}(\widehat{X}_{j,T_j(t-1)} \geq \widehat{X}_{i,T_i(t-1)} \quad \forall i)$. We plug this into (11) which yields the following bound on the mean regret at time n:

$$\mathbb{E}[R_n] \leq \sum_{j=1}^{m}\Delta_j + \sum_{t=m+1}^{n}\sum_{j:\mu_j<\mu_*}\Delta_j\left(\varepsilon_t\frac{1}{m} + (1-\varepsilon_t)\beta_j(t)\right)$$

This, combined with the bound $\beta_j(t)$ above, proves the theorem.

B The regret bound of the UCB algorithm

The regret at round n is given by

$$R_n \quad = \quad \sum_{j=1}^{m} \Delta_j + \sum_{t=m+1}^{n} \sum_{j=1}^{m} \Delta_j \mathbb{1}_{\{I_t = j\}}$$

The expected regret $\mathbb{E}[R_n]$ at round n is bounded by

$$\mathbb{E}[R_n] \leq \sum_{j=1}^{m} \Delta_j + \sum_{j=1}^{m} \Delta_j \mathbb{E}[T_j(n)]. \tag{27}$$

where $T_j(n) = \sum_{t=1}^{n} \mathbb{1}_{\{I_t = j\}}$ is the number of times arm j has been chosen up to round n. Recall that

$$\widehat{X}_j = \frac{1}{T_j(t-1)} \sum_{s=1}^{T_j(t-1)} X_j(s). \tag{28}$$

Let's suppose the rewards are bounded, say between 0 and 1.

STEP 1: Let us bound the probability of overestimating or underestimating suboptimal arm j. From the Chernoff-Hoeffding Inequality we have that

$$\mathbb{P}\left(\frac{1}{T_j(t-1)} \sum_{i=1}^{T_j(t-1)} X_{j,i} - \mu_j \leq -\varepsilon \right) \leq \exp\{-2T_j(t-1)\varepsilon^2\},$$

and

$$\mathbb{P}\left(\frac{1}{T_j(t-1)} \sum_{i=1}^{T_j(t-1)} X_{j,i} - \mu_j \geq \varepsilon \right) \leq \exp\{-2T_j(t-1)\varepsilon^2\}. \tag{29}$$

By selecting $\varepsilon = \sqrt{\frac{2\,\log(t)}{T_j(t-1)}}$ we have

$$\mathbb{P}\left(\widehat{X}_j + \sqrt{\frac{2\,\log(t)}{T_j(t-1)}} \leq \mu_j \right) \leq t^{-4}, \tag{30}$$

and

$$\mathbb{P}\left(\widehat{X}_j - \sqrt{\frac{2\,\log(t)}{T_j(t-1)}} \geq \mu_j \right) \leq t^{-4}. \tag{31}$$

STEP 2: Let us bound the number of times we play arm j.

In the following, notice that in (33) the summation starts from $m+1$ because in the first m initialization rounds each arm is played once. Moreover, (34) follows from (33) by assuming that arm j has already been played u times. For each t,

$$\begin{gathered}\left\{ \widehat{X}_{j,T_j(t-1)} + \sqrt{\frac{2\log(t)}{T_j(t-1)}} \geq \widehat{X}_{*,T_*(t-1)} + \sqrt{\frac{2\log(t)}{T_*(t-1)}}, T_j(t-1) \geq u \right\} \subset \\ \left\{ \max_{s_j \in \{u,\ldots,T_j(t-1)\}} \widehat{X}_{j,s_j} + \sqrt{\frac{2\log(t)}{s_j}} \geq \min_{s_* \in \{1,\ldots,T_*(t-1)\}} \widehat{X}_{*,s_*} + \sqrt{\frac{2\log(t)}{s_*}} \right\}\end{gathered} \tag{32}$$

which justifies (36). We also have that (32) is included in

$$\bigcup_{s_*=1}^{T_*(t-1)} \bigcup_{s_j=u}^{T_j(t-1)} \left\{ \widehat{X}_{j,s_j} + \sqrt{\frac{2\log(t)}{s_j}} \geq \widehat{X}_{*,s_*} + \sqrt{\frac{2\log(t)}{s_*}} \right\}.$$

Thus, for any integer u, we may write

$$\begin{aligned}
T_j(n) &= 1 + \sum_{t=m+1}^{n} \mathbb{1}\{I_t = j\} && (33)\\
&= u + \sum_{t=m+1}^{n} \mathbb{1}\{I_t = j, T_j(t-1) \geq u\} && (34)\\
&= u + \sum_{t=m+1}^{n} \mathbb{1}\left\{\widehat{X}_{j,T_j(t-1)} + \sqrt{\frac{2\log(t)}{T_j(t-1)}} \geq \widehat{X}_{*,T_*(t-1)} + \sqrt{\frac{2\log(t)}{T_*(t-1)}}, T_j(t-1) \geq u\right\} && (35)\\
&\leq u + \sum_{t=m+1}^{n} \mathbb{1}\left\{\max_{s_j \in \{u,\dots,T_j(t-1)\}} \widehat{X}_{j,s_j} + \sqrt{\frac{2\log(t)}{s_j}} \geq \min_{s_* \in \{1,\dots,T_*(t-1)\}} \widehat{X}_{*,s_*} + \sqrt{\frac{2\log(t)}{s_*}}\right\} && (36)\\
&\leq u + \sum_{t=m+1}^{n} \sum_{s_*=1}^{T_*(t-1)} \sum_{s_j=u}^{T_j(t-1)} \mathbb{1}\left\{\widehat{X}_{j,s_j} + \sqrt{\frac{2\log(t)}{s_j}} \geq \widehat{X}_{*,s_*} + \sqrt{\frac{2\log(t)}{s_*}}\right\}. && (37)
\end{aligned}$$

STEP 3: Let us rewrite the event of playing arm j as a subset of the union of underestimating or overestimating arm j. When

$$\mathbb{1}\left\{\widehat{X}_j + \sqrt{\frac{2\ \log(t)}{T_j(t-1)}} \geq \widehat{X}_* + \sqrt{\frac{2\ \log(t)}{T_*(t-1)}}\right\} \tag{38}$$

is equal to one, at least one of the following has to be true:

$$\begin{aligned}
\widehat{X}_* &\leq \mu_* - \sqrt{\frac{2\ \log(t)}{T_*(t-1)}}; && (39)\\
\widehat{X}_j &\geq \mu_j + \sqrt{\frac{2\ \log(t)}{T_j(t-1)}}; && (40)\\
\mu_* &< \mu_j + 2\sqrt{\frac{2\ \log(t)}{T_j(t-1)}}. && (41)
\end{aligned}$$

(In fact, suppose none of them hold simultaneously. Then from (39) we would have that $\widehat{X}_* > \mu_* - \sqrt{\frac{2\log(t)}{T_*(t-1)}}$; then, by applying (41) (with opposite verse since we are assuming it does not hold) we get $\widehat{X}_* > \mu_j + 2\sqrt{\frac{2\log(t)}{T_j(t-1)}} - \sqrt{\frac{2\log(t)}{T_*(t-1)}}$ and then from (40) (again, with opposite verse) follows that $\widehat{X}_* > \widehat{X}_j + \sqrt{\frac{2\log(t)}{T_j(t-1)}} - \sqrt{\frac{2\log(t)}{T_*(t-1)}}$ which is in contradiction with (38).) Now, if we set $u = \left\lceil \frac{8}{\Delta_j^2} \log(t) \right\rceil$, for $T_j(t-1) \geq u$,

$$\begin{aligned}
&\mu_* - \mu_j - 2\sqrt{\frac{2\log(t)}{T_j(t-1)}}\\
\geq\ &\mu_* - \mu_j - 2\sqrt{\frac{2\log(t)}{u}}\\
\geq\ &\mu_* - \mu_j - \Delta_j = 0,
\end{aligned}$$

therefore, with this choice of u, (41) can not hold.

STEP 4: Let us bound the expected number of times we play arm j.

Using (37), we have that

$$\begin{aligned}
T_j(n) \leq\ &\left\lceil \frac{8}{\Delta_j^2} \log(n) \right\rceil\\
+\ &\sum_{t=m+1}^{n} \sum_{s_*=1}^{T_*(t-1)} \sum_{s_j=u}^{T_j(t-1)} \mathbb{1}\left\{\widehat{X}_{*,s_*} \leq \mu_* - \sqrt{\frac{2\log(t)}{s_*}}\right\}\\
+\ &\sum_{t=m+1}^{n} \sum_{s_*=1}^{T_*(t-1)} \sum_{s_j=u}^{T_j(t-1)} \mathbb{1}\left\{\widehat{X}_{j,s_j} \geq \mu_j + \sqrt{\frac{2\log(t)}{s_j}}\right\}
\end{aligned}$$

and by taking expectation,

$$\begin{aligned}
\mathbb{E}[T_j(n)] &\leq \left\lceil \frac{8}{\Delta_j^2}\log(n) \right\rceil \\
&+ \sum_{t=m+1}^{n} \sum_{s_*=1}^{T_*(t-1)} \sum_{s_j=u}^{T_j(t-1)} \mathbb{P}\left\{ \widehat{X}_{*,s_*} \leq \mu_* - \sqrt{\frac{2\log(t)}{s_*}} \right\} \\
&+ \sum_{t=m+1}^{n} \sum_{s_*=1}^{T_*(t-1)} \sum_{s_j=u}^{T_j(t-1)} \mathbb{P}\left\{ \widehat{X}_{j,s_j} \geq \mu_j + \sqrt{\frac{2\log(t)}{s_j}} \right\} \\
&\leq \frac{8}{\Delta_j^2}\log(n) + 2\sum_{t=m+1}^{n} t^{-4}(t-1-m)^2.
\end{aligned}$$

where in the last step we create an upper bound for $T_*(t-1)$ and $T_j(t-1)$ by $(t-1-m)$ (cases where we have only played the best arm or arm j).

Therefore, by using (27)

$$\begin{aligned}
\mathbb{E}[R_n] &\leq \sum_{j=1}^{m} \Delta_j \\
&+ \sum_{j:\mu_j<\mu_*} \frac{8}{\Delta_j}\log(n) + \sum_{j=1}^{m} \Delta_j \left(\sum_{t=m+1}^{n} 2\,(t)^{-4}\,(t-1-m)^2 \right).
\end{aligned}$$

Notice that the parenthesis is bound by $\log(n)$.

C Notation summary

- m: number of arms;
- n: number of rounds;
- $X_j(t)$: random reward for playing arm j;
- μ_*: mean reward of the optimal arm ($\mu_* = \max_{1 \leq j \leq m} \mu_j$);
- Δ_j: difference between the mean reward of the optimal arm and the mean reward of arm j ($\Delta_j = \mu_* - \mu_j$);
- $\hat{X}_j$: current estimate of μ_j;
- I_t: arm played at turn t;
- $T_j(t-1)$: number of times arm j has been played before round t starts;
- k: a constant greater than 10 such that $k > \frac{4}{\min_j \Delta_j}$ in Algorithm 1;
- $\beta_j(t)$: upper bound on the probability of considering suboptimal arm j being the best arm at round t when using Algorithm 1;
- n': particular time defined as km in the comparison between Algorithm 1 in Section 2;
- R_n: total regret at round n.