Stochastic bandit:
n arms with iid stochastic rewards $(R_{a,i})_{i\in\mathbb{N}}$ wtih mean r_a, $1 \leq a \leq n$.

UCB algorithm:

- At time 1 each arm is sampled once.
- At each time $t > 1$:
  1. Compute an upper confidence bound for each a: $UCB_a(t) = \hat{r}_a(N_a(t)) + \sqrt{\frac{\alpha \log t}{2N_a(t)}}$
  2. Choose $A_{t+1} \in arg\max_a UCB_a(t)$.

Where we denote by $N_a(t)$ the number of times that UCB chooses arm a up to time t and $\hat{r}_a(N_a(t)) = \frac{1}{N_a(t)} \sum_{s=1}^{N_a(t)} R_{a,s}$.

Theorem 1. *If all arms have bounded rewards in* $[0,1]$, $\forall \alpha > 2, \exists C_\alpha > 0$ *s.t.* $\mathbb{E}(N_a(T)) \leq \frac{2\alpha \log T}{(r^* - r_a)^2} + C_\alpha$ *for all suboptimal arm* a.

Proof. Main ingredient of the proof is Hoeffding inequality.

Let $X_1, X_2, \cdots$ be independent variables whose support is bounded: $v_i \leq X_i \leq u_i$ for all i and means $\mathbb{E}(X_i) = m_i$. Then

$$\mathbb{P}(X_1 + \cdots + X_s - (m_1 + \cdots + m_s) \leq -\epsilon) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^s (u_i - v_i)^2}\right)$$

$$\mathbb{P}(X_1 + \cdots + X_s - (m_1 + \cdots + m_s) \geq \epsilon) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^s (u_i - v_i)^2}\right)$$

How to use this here for one arm: $X_i = R_{a,i}$ and $v_i = 0, u_i = 1$.

Therefore by multiplying by s, for any s, Hoeffding says

$$\mathbb{P}(\hat{r}_a(s) \leq r_a - \epsilon) \leq \exp(-2\epsilon^2 s).$$

Let $S = N_a(t)$. Then,

$$\mathbb{P}(UCB_a(t) \leq r_a) = \mathbb{P}(\hat{r}_a(S) \leq r_a - \sqrt{\frac{\alpha \log t}{2S}})$$

Since $S = N_a(t)$ is random and depends on the values of $R_{a,s}$, Hoeffing does not hold for $S = N_a(t)$.

Instead we use the union bound:

$$\begin{aligned}
\mathbb{P}(UCB_a(t) \leq r_a) &= \mathbb{P}\left(\hat{r}_a(S) \leq r_a - \sqrt{\frac{\alpha \log t}{2S}}\right) \\
&\leq \mathbb{P}\left(\exists s \leq t, \hat{r}_a(s) \leq r_a - \sqrt{\frac{\alpha \log t}{2s}}\right) \\
&\leq \sum_{s=1}^{t} \mathbb{P}\left(\hat{r}_a(s) \leq r_a - \sqrt{\frac{\alpha \log t}{2s}}\right) \quad \text{(union bound).}
\end{aligned}$$

Now we can use Hoeffding for all s:

$$\begin{aligned}\mathbb{P}(UCB_a(t) \le r_a) &\le \sum_{s=1}^{t} \exp\left(-2s\frac{\alpha \log t}{2s}\right) \\ &= \sum_{s=1}^{t} \frac{1}{t^\alpha} \\ &= \frac{1}{t^{\alpha-1}}\end{aligned}$$

Finally,

$$\mathbb{P}(UCB_a(t) \le r_a) \le \frac{1}{t^{\alpha-1}}. \tag{1}$$

Similarly, we can define $LCB_a(t) = \hat{r}_a(N_a(t)) - \sqrt{\frac{\alpha \log t}{2N_a(t)}}$. Using the same proof (and Hoeffding for $\ge \epsilon$ instead of $\le -\epsilon$),

$$\mathbb{P}(LCB_a(t) \ge r_a) \le \frac{1}{t^{\alpha-1}}. \tag{2}$$

Now we look at special events. To simplify notation, we assume wlog that arm 1 is optimal and arm 2 is not optimal. Let τ by any stopping time of the algorithm (any time that only depends on the past steps, $1, ..., T$).

$$\begin{aligned}N_2(T) - N_2(\tau) &= \sum_{t=\tau+1}^{T} \mathbf{1}_{\{A_t=2\}} \\ &= \sum_{t=\tau+1}^{T} \mathbf{1}_{\{A_t=2 \wedge UCB_1(t) \le r_1\}} + \sum_{t=\tau+1}^{T} \mathbf{1}_{\{A_t=2 \wedge UCB_1(t) > r_1\}} \\ &= \sum_{t=\tau+1}^{T} \mathbf{1}_{\{A_t=2 \wedge UCB_1(t) \le r_1\}} + \sum_{t=\tau+1}^{T} \mathbf{1}_{\{A_t=2 \wedge UCB_2(t) > r_1\}} \quad \text{(2 was chosen)} \\ &\le \sum_{t=\tau+1}^{T} \mathbf{1}_{\{UCB_1(t) \le r_1\}} + \sum_{t=\tau+1}^{T} \mathbf{1}_{\{UCB_2(t) > r_1\}} \\ &= \sum_{t=\tau+1}^{T} \mathbf{1}_{\{UCB_1(t) \le r_1\}} + \sum_{t=\tau+1}^{T} \mathbf{1}_{\{UCB_2(t) > r_1 \wedge LCB_2(t) < r_2\}} + \sum_{t=\tau+1}^{T} \mathbf{1}_{\{UCB_2(t) > r_1 \wedge LCB_2(t) > r_2\}} \\ &\le \sum_{t=\tau+1}^{T} \mathbf{1}_{\{UCB_1(t) \le r_1\}} + \sum_{t=\tau+1}^{T} \mathbf{1}_{\{UCB_2(t) > r_1 \wedge LCB_2(t) < r_2\}} + \sum_{t=\tau+1}^{T} \mathbf{1}_{\{LCB_2(t) \ge r_2\}}\end{aligned}$$

Let us first study the middle term:

$$\begin{aligned}\mathbf{1}_{\{UCB_2(t) > r_1 \wedge LCB_2(t) < r_2\}} &\le \mathbf{1}_{\left\{\sqrt{\frac{\alpha \log t}{2N_2(t)}} > \frac{r1-r2}{2}\right\}} \\ &= \mathbf{1}_{\left\{N_2(t) < \frac{2\alpha \log t}{(r_1-r_2)^2}\right\}}.\end{aligned}$$

Let us denote $K(t) = \frac{2\alpha \log t}{(r_1 - r_2)^2}$.

We get

$$N_2(T) - N_2(\tau) \leq C + \sum_{t=\tau+1}^{T} \mathbf{1}_{\{N_2(t) < K(t)\}}.$$

Using Equations (1) and (2),

$$\begin{aligned}
\mathbb{E}(C) &= \sum_{t=\tau+1}^{T} \Pr(UCB_1(t) \leq r_1) + \sum_{t=\tau+1}^{T} \Pr(LCB_2(t) \geq r_2) \\
&\leq 2 \sum_{t=1}^{\infty} \frac{1}{t^{\alpha-1}} =: C_\alpha
\end{aligned}$$

notice that C_α is finite if $\alpha > 2$.

Now it is time to choose τ.

Let us consider $\tau = \max\{t \leq T | N_2(t) < K(t)\}$ (exists because $N_2(2) \leq 2 < K(2)$).

if $\tau < T$,

$$\mathbb{E}(N_2(T) \leq \mathbb{E}(N_2(\tau)) + C_\alpha + 0.$$

Moreover, $\mathbb{E}(N_2(\tau)) < K(\tau) < K(T)$. This implies that

$$\mathbb{E} N_2(T) < C_\alpha + \frac{2\alpha \log T}{(r_1 - r_2)^2}.$$

If $\tau = T$, then directly $N_2(\tau) = N_2(T) < K(T)$ so $\mathbb{E}(N_2(T)) < K(T)$.

QED.

□

A direct consequence of this theorem is

$$\begin{aligned}
Reg(UCB, T) &= \mathbb{E}\left(\sum_{t=1}^{T} r_1 - R_{A(t), N_{A(t)}(t)} \right) \\
&= \sum_{a \neq 1} (r_1 - r_a) \mathbb{E} N_a(T) \\
&< n r_1 C_\alpha + \left(\sum_{a \neq 1} \frac{2\alpha}{r_1 - r_a} \right) \log T.
\end{aligned}$$