Exercice 1 (Regret cumulé d'UCB)
Nous considérons le problème du bandit stochastique à K bras dans lequel l'agent joue la stratégie UCB suivante : au temps t, l'agent tire le bras

$$d_t \in \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} \quad m_{k,T_k(t-1)} + \sqrt{\frac{\alpha \ln t}{2T_k(t-1)}},$$

avec $\alpha > 2$, $m_{k,s}$ la moyenne empirique des récompenses obtenues du bras k sur ses s premiers tirages, $T_k(t-1)$ le nombre de fois que le bras k a été tiré sur $[1; t-1]$ et n l'horizon du jeu. Nous rappelons que tous les gains des bras sont dans l'intervalle $[0; 1]$.

Soit k^* un bras optimal : $k^* \in \operatorname{argmax}_{k\in\{1,\dots,K\}}\mu_k$, où μ_k est l'espérance des récompenses obtenues du bras k. Soit R_n le regret, i.e. $R_n = n\mu_{k^*} - \mathbb{E}G_n$ avec G_n le gain cumulé de l'agent sur les n pas de temps. L'indice de sous-optimalité du bras k est $\Delta_k = \mu_{k^*} - \mu_k$. Pour tout entier $t \geq 1$ et $k \in \{1, \dots, K\}$, introduisons les événements

$$\mathcal{A}_{k,t} = \left\{ \exists 1 \leq s \leq t, \quad m_{k,s} \geq \mu_k + \sqrt{\frac{\alpha \ln t}{2s}} \right\}$$
$$\mathcal{B}_t = \left\{ \exists 1 \leq s \leq t, \quad m_{k^*,s} + \sqrt{\frac{\alpha \ln t}{2s}} \leq \mu_{k^*} \right\}$$
$$\mathcal{C}_{k,t} = \left\{ T_k(t-1) < \frac{2\alpha \ln n}{\Delta_k^2} \right\}$$

On notera $\mathcal{E}^c$ l'événement complémentaire de $\mathcal{E}$. Dans les questions suivantes, nous considérons que k est un bras sous-optimal ($\Delta_k > 0$).

1) *Montrer que* $d_t \neq k$ *sur l'événement* $\mathcal{A}_{k,t}^c \bigcap \mathcal{B}_t^c \bigcap \mathcal{C}_{k,t}^c$.
2) *En déduire* $T_k(n) \leq \sum_{t=1}^n \mathbb{1}_{\{d_t=k\}\cap\mathcal{C}_{k,t}} + \sum_{t=1}^n \mathbb{1}_{\mathcal{A}_{k,t}\cup\mathcal{B}_t}$.
3) *Montrer* $\mathbb{P}(\mathcal{A}_{k,t}) \leq t^{1-\alpha}$.
4) *De même, déterminer un majorant de* $\mathbb{P}(\mathcal{B}_t)$.
5) *Montrer* $\sum_{t=1}^n t^{1-\alpha} \leq 1 + \frac{1}{\alpha-2}$.
6) *Montrer que pour tout* $u > 0$, *on a* $\sum_{t=1}^n \mathbb{1}_{\{d_t=k\}\cap\{T_k(t)<u\}} < u$.
7) *Déduire des questions précédentes l'inégalité* $R_n \leq \sum_{k:\Delta_k>0} \left(\frac{2\alpha \ln n}{\Delta_k} + 3 + \frac{2}{\alpha-2}\right)$.
8) *En quoi ce résultat diffère-t-il de celui du cours ?*
9) Cette borne diverge lorsque $\alpha \to 2$. *Est-ce le comportement attendu ?*