Exercise 3: The (α, ψ)–UCB algorithm (can be solved after Course #4)

Let $\psi : \mathbb{R} \to \mathbb{R}$ be a convex function such that $\psi(x) = \psi(-x)$ for all $x \in \mathbb{R}$. Consider a bandit model $\mathcal{D}$ such that for all $\nu \in \mathcal{D}$, if X denotes a random variable with distribution ν, then

$$\forall \lambda \geqslant 0, \qquad \max\left\{ \ln \mathbb{E}_\nu\left[e^{\lambda(X-\mathbb{E}[X])}\right], \ln \mathbb{E}_\nu\left[e^{\lambda(\mathbb{E}[X]-X)}\right] \right\} \leqslant \psi(\lambda)\,. \tag{$\star$}$$

For all $x \geqslant 0$, we define the convex conjugate of ψ,

$$\psi^\star(x) = \sup\big\{\lambda x - \psi(\lambda) : \lambda \geqslant 0\big\}\,,$$

and assume that $\psi^\star$ is invertible, with inverse denoted by $(\psi^\star)^{-1}$.

1. Provide such a function ψ for the model $\mathcal{D} = \mathcal{P}\big([0,1]\big)$ of all probability distributions over $[0,1]$. Compute $\psi^\star$ and its inverse.

We generalize the UCB algorithm for stochastic bandits in the following way. We consider the same setting and use the same notation as the ones used in class and in Exercise 3 of the present statement: a stochastic bandit problem is formed by $K \geqslant 2$ probability distributions $\nu_1, \ldots, \nu_K$ in $\mathcal{D}$ with respective expectations μ_k, their maximal expectation is denoted by $\mu^\star$, the gap of arm k is $\Delta_k = \mu^\star - \mu_k$, etc.

(α, ψ)–UCB algorithm

Parameters: $\alpha > 0$ and $\psi : \mathbb{R} \to \mathbb{R}$ with $\psi(x) = \psi(-x)$ for all $x \geqslant 0$

Initialization: Play each arm once, i.e., $I_t = t$ for $t \in \{1, \ldots, K\}$, get a reward $Y_t \sim \nu_t$

For $t \geqslant K+1$,

1. Compute, for all $k \in \{1, \ldots, K\}$,

$$N_k(t-1) = \sum_{s=1}^{t-1} \mathbb{1}_{\{I_s = k\}} \qquad \text{and} \qquad \widehat{\mu}_{k,t-1} = \frac{1}{N_k(t-1)} \sum_{s=1}^{t-1} Y_s \mathbb{1}_{\{I_s = k\}}$$

2. Pick an arm (ties broken arbitrarily)

$$I_t \in \operatorname*{argmax}_{k \in \{1,\ldots,K\}} \left\{ \widehat{\mu}_{k,t-1} + (\psi^\star)^{-1}\left(\frac{\alpha \ln t}{N_k(t-1)} \right) \right\}$$

3. Get a reward $Y_t \sim \nu_{I_t}$ (conditionally to I_t)

We want to upper bound the pseudo-regret of the (α, ψ)–UCB algorithm as follows: for $\alpha > 2$,

$$\overline{R}_T = T\mu^\star - \mathbb{E}\left[\sum_{t=1}^T Y_t\right] \leqslant \sum_{k:\Delta_k > 0} \Delta_k \left(\frac{\alpha}{\psi^\star(\Delta_k/2)} \ln T + \frac{2\alpha}{\alpha - 2} \right). \tag{B}$$

To that end, we first show that for each arm k and $t \geqslant K+1$, an upper confidence bound on μ_k is given by

$$\widehat{\mu}_{k,t-1} + (\psi^\star)^{-1}\left(\frac{\alpha \ln t}{N_k(t-1)} \right).$$

2. Prove that for all $t \geqslant 1$ and all $\lambda \geqslant 0$,

$$\mathbb{E}\left[\exp\Big(-\lambda(Y_t - \mu_k)\, \mathbb{1}_{\{I_t = k\}} \Big) \,\Big|\, \mathcal{F}_{t-1} \right] \leqslant \exp\big(\psi(\lambda)\, \mathbb{1}_{\{I_t = k\}} \big)$$

for a filtration $\mathcal{F} = (\mathcal{F}_t)_{t \geqslant 0}$ to specify explicitly.

Construct an $\mathcal{F}$–adapted supermartingale $(M_t)_{t \geqslant 0}$ based on this inequality.

3. Prove that for all $t \geqslant K+1$, all $\ell \geqslant 1$, and all $\varepsilon > 0$,

$$\mathbb{P}\Big\{\widehat{\mu}_{k,t-1} + \varepsilon \leqslant \mu_k \;\text{ and }\; N_k(t-1) = \ell\Big\} \leqslant \exp\big(-\ell\, \psi^\star(\varepsilon)\big)\,.$$

4. Provide a bound, for $t \geqslant K+1$, on

$$\mathbb{P}\left\{\widehat{\mu}_{k,t-1} + (\psi^\star)^{-1}\left(\frac{\alpha \ln t}{N_k(t-1)}\right) \leqslant \mu_k\right\}.$$

5. Briefly indicate how to bound, for $t \geqslant K+1$,

$$\mathbb{P}\left\{\widehat{\mu}_{k,t-1} - (\psi^\star)^{-1}\left(\frac{\alpha \ln t}{N_k(t-1)}\right) > \mu_k\right\}.$$

To establish the regret bound, we first fix a suboptimal arm j and an optimal arm $a^\star$.

6. Explain why $I_t = j$ for $t \geqslant K+1$ entails one of the following events:

$$\widehat{\mu}_{a^\star,t-1} + (\psi^\star)^{-1}\left(\frac{\alpha \ln t}{N_{a^\star}(t-1)}\right) \leqslant \mu^\star\,,$$

$$\text{or} \qquad \widehat{\mu}_{j,t-1} - (\psi^\star)^{-1}\left(\frac{\alpha \ln t}{N_j(t-1)}\right) > \mu_j\,,$$

$$\text{or} \qquad N_j(t-1) < \frac{\alpha \ln t}{\psi^\star(\Delta_j/2)}\,.$$

7. Establish the regret bound (B).

We conclude this exercice with a discussion of the bound for the model $\mathcal{D} = \mathcal{P}\big([0,1]\big)$.

8. Provide also a distribution-free bound for (α, ψ)–UCB on this model, i.e., a bound over all distributions satisfying $(\star)$. You need first to think of a suitable value for α.