Exercise session n°2 : stochastic bandits

Exercise 1 :

Sub-Gaussian random variables. Let X be a **centered** random variable in $\mathbb{R}$. Show that affirmations below satisfy the following implications chain: 1. $\implies$ 2. $\implies$ 3. $\implies$ 4. $\implies$ 5.

1. *Laplace transform:* for any $\eta \in \mathbb{R}$, $\ln(\mathbb{E}[e^{\eta X}]) \leq \frac{\sigma^2\eta^2}{2}$;
2. *Concentration:* for any $\varepsilon > 0$, $\max\{\mathbb{P}(X \geq \varepsilon), \mathbb{P}(X \leq -\varepsilon)\} \leq \exp(\frac{-\varepsilon^2}{2\sigma^2})$;
3. *Moment condition:* for any $q \in \mathbb{N}^*$, $\mathbb{E}[X^{2q}] \leq q!(4\sigma^2)^q$;
4. *Orlicz condition:* $\mathbb{E}[\exp(\frac{X}{8\sigma^2})] \leq 2$;
5. *Laplace transform:* for any $\eta \in \mathbb{R}$, $\ln(\mathbb{E}[e^{\eta X}]) \leq \frac{24\sigma^2\eta^2}{2}$.

Exercise 2 :

Doubling trick. This exercise analyses a meta-algorithm based on the doubling trick that converts a policy depending on the horizon to a policy with similar guarantees that does not. Let $\mathcal{B}$ be an arbitrary set of bandits. Suppose you are given a policy (algorithm) $\pi = \pi(T)$ designed for $\mathcal{B}$ that accepts the horizon T as a parameter and has a regret guarantee of

$$\max_{1 \leq t \leq T} R_t(\pi(n), \nu) \leq f_T(\nu), \quad \forall \nu \in \mathcal{B}.$$

For a fixed sequence of integers $T_1 < T_2 > T_3 < \ldots$, we define the algorithm $\tilde{\pi}$ that first runs $\pi(T_1)$ on $[\![1, T_1]\!]$; then runs **independently** $\pi(T_2)$ on $[\![T_1, T_1 + T_2]\!]$; etc. So $\tilde{\pi}$ runs $\pi(T_i)$ on $[\![\sum_{j=1}^{i-1} T_j, \sum_{j=1}^{i} T_j]\!]$ and does not require a prior knowledge of T.

1) For a fixed $T \in \mathbb{N}$, let $\ell_{\max} = \min\{\ell \in \mathbb{N}^* \mid \sum_{i=1}^{\ell} T_i \geq T\}$. Prove that for any $\nu \in \mathcal{B}$, the regret of $\tilde{\pi}$ on ν is at most

$$R_T(\tilde{\pi}, \nu) \leq \sum_{\ell=1}^{\ell_{\max}} f_{T_\ell}(\nu).$$

2) (Distribution free bound) Suppose that $f_T(\nu) \leq \sqrt{T}$. Show that for a good choice of n_ℓ, for any $\nu \in \mathcal{B}$ and $T \in \mathbb{N}$:

$$R_T(\tilde{\pi}, \nu) \leq \frac{1}{\sqrt{2} - 1}\sqrt{T}.$$

3) (Instance dependent bound) Suppose that $f_T(\nu) \leq g(\nu)\ln(T)$ for some function g. Show that with the same choice of sequence n_ℓ as in b), we can bound the regret for any $\nu \in \mathcal{B}$ and $T \in \mathbb{N}$ as:

$$R_T(\tilde{\pi}, \nu) \leq g(\nu)\frac{\ln(T)^2}{2\ln(2)}.$$

4) Can you suggest a sequence of n_ℓ such that for some universal constant $C > 0$, the regret of $\tilde{\pi}$ can be bounded for any $\nu \in \mathcal{B}$ and $T \in \mathbb{N}$ as:

$$R_T(\tilde{\pi}, \nu) \leq Cg(\nu)\ln(T).$$

Exercise 3 :

Consider the ε-greedy algorithm with $\varepsilon_t = \min\left(1, \frac{(K\ln(t))^{\frac{1}{3}}}{t^{\frac{1}{3}}}\right)$ for any $t \in \mathbb{N}$. Show that for a large enough universal constant $C > 0$, the regret of ε-greedy satisfies

$$R_T \leq CT^{\frac{2}{3}}(K\ln(T))^{\frac{1}{3}}.$$

Hint: Bound the instantaneous regret $\mathbb{E}[\Delta_{a_t}]$.

Exercise 4 :

Distribution free bound. Let $\mathcal{B}$ be an arbitrary set of bandits. Suppose you are given a policy (algorithm) $\pi = \pi(T)$ designed for $\mathcal{B}$ that has the following guarantees

$$\mathbb{E}[N_k(T)] \leq C_0 + C\frac{\ln(T)}{\Delta_k^2}, \quad \forall \nu \in \mathcal{B}, \forall T \in \mathbb{N},$$

for some constants C_0, C.

(a) First, show that it directly implies the following distribution free bound:

$$R_T \leq KC_0 + K\sqrt{CT\ln(T)}.$$

(b) Show, with a refined analysis, that we even have the following bound

$$R_T \leq \sqrt{KT(C_0 + C\ln(T))}.$$