# Exercise sheet n°1

In this session, we consider online learning with experts (see Lecture #1) with linear losses. The losses $\ell_{jt}$ are in $[0,1]$ when not precised otherwise.

**Exercise 1 :**

Consider online learning with experts (see Lecture #1) with linear losses. Show that no strategy satisfies for all sequence $(\ell_{1t},\dots,\ell_{Nt})_t \in ([0,1]^N)^{\mathbb{N}}$:

$$\sum_{t=1}^T \sum_{j=1}^N p_{jt}\ell_{jt} - \sum_{t=1}^T \min_{k\in[N]} \ell_{kt} = o(T).$$

**Exercise 2 :**

Consider online learning with experts (see Lecture #1) with linear losses. Assume in this exercise that $\ell_{jt} \in [m, M]$, with $m, M \in \mathbb{R}$ unknown. How can we tune $\eta$?

We consider in the following EWA with adaptive rates $(\eta_t)_t$:

$$p_{jt} = \frac{e^{-\eta_t \sum_{s=1}^{t-1} \ell_{js}}}{\sum_{k=1}^N e^{-\eta_t \sum_{s=1}^{t-1} \ell_{ks}}}.$$

**1)** Show that if $(\eta_t)$ are non-increasing

$$\frac{1}{N}\sum_{j=1}^N p_{jt} e^{-\eta_t \ell_{jt}} \geq \frac{1}{N^{\frac{\eta_t}{\eta_{t+1}}}} \frac{\left(\sum_{j=1}^N \exp(-\eta_{t+1}\sum_{s=1}^t \ell_{js})\right)^{\frac{\eta_t}{\eta_{t+1}}}}{\sum_{k=1}^N \exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_{ks}\right)}.$$

**Hint:** Use the fact that $x \mapsto x^{\frac{\eta_t}{\eta_{t+1}}}$ is convex.

**2)** Show that if $(\eta_t)$ are non-increasing, then the regret of EWA satisfies:

$$R_T \leq \frac{\ln N}{\eta_T} + \sum_{t=1}^T \delta_t,$$

where $\delta_t = \sum_{j=1}^N p_{jt}\ell_{jt} + \frac{1}{\eta_t}\ln\left(\sum_{j=1}^N p_{jt} e^{-\eta_t \ell_{jt}}\right)$.

**Hint:** Multiply by $\frac{-1}{\eta_t}$ the logarithm of the expression obtained in **1)** to make a telescopic sum appears.

Recall the Bernstein's inequality for a random variable $X \in [m, M]$:

$$\forall \eta > 0,\ \ln \mathbb{E}[e^{\eta X}] \leq \eta \mathbb{E}[X] + \frac{e^{\eta(M-m)} - 1 - \eta(M-m)}{(M-m)^2}\mathrm{Var}\left(X\right).$$

We now consider EWA with $\eta_t = \frac{\ln N}{\sum_{s=1}^{t-1}\delta_s}$, with the convention that $\frac{\ln N}{0} = +\infty$.

**3)** Let $v_t = \sum_{j=1}^N (\ell_{jt} - \sum_{k=1}^N p_{kt}\ell_{kt})^2 p_{jt}$.

(a) Show that $v_t \geq \frac{\eta_t(M-m)}{e^{\eta_t(M-m)} - \eta_t(M-m) - 1}(M-m)\delta_t$.

(b) Deduce that $v_t \geq \frac{2\delta_t}{\eta_t} - \frac{2}{3}(M-m)\delta_t$.

**4)**

(a) Show that $\left(\sum_{t=1}^{T} \delta_t\right)^2 \leq \sum_{t=1}^{T} v_t \ln N + (M-m)(1 + \frac{2}{3}\ln N)\sum_{t=1}^{T} \delta_t$.

(b) Finally, show that $R_T \leq (M-m)\sqrt{T \ln N} + (M-m)(2 + \frac{4}{3}\ln N)$.

**Exercise 3 :**

Consider the $\varepsilon$-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 $\varepsilon$-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 :**

**Concentration for sequences of random length.** Let $X_1, X_2, \ldots$ be a sequence of independent standard Gaussian random variables defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that $T : \Omega \to \{1, 2, 3, \ldots\}$ is another variable and let $\hat{\mu}_T = \sum_{t=1}^{T} \frac{X_t}{T}$ be the empirical mean based on $T$ samples.

**1)** Show that if $T$ is independent from $X_t$ for all $t$, then for any $\delta \in (0,1)$

$$\mathbb{P}\left(\hat{\mu}_T \geq \sqrt{\frac{2\ln(1/\delta)}{T}}\right) \leq \delta.$$

**2)** Now relax the assumption that $T$ is independent from $(X_t)_t$. Let $E_t = \mathbb{1}_{T=t}$ and $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$ be the $\sigma$-algebra generated by the first $t$ samples. Let $\delta \in (0,1)$ and show there exists a random variable $T$ such that for all $t$, $E_t$ is $\mathcal{F}_t$-measurable and

$$\mathbb{P}\left(\hat{\mu}_T \geq \sqrt{\frac{2\ln(1/\delta)}{T}}\right) = 1.$$

**Hint:** You can use the law of the iterated logarithm, which says if $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with zero mean and unit variance, then

$$\limsup_{n\to\infty} \frac{\sum_{t=1}^{n} X_t}{\sqrt{2n\ln\ln n}} = 1 \quad \text{almost surely.}$$

**3)** What is the relation between the above inequality and our concentration lemma for the empirical means in bandits problems? Do **2)** and our lemma contradict? Why?

**Exercise 5 :**

**Phased SE.** Consider the following phased Successive Eliminations algorithm parameterized by $a > 1$.

**Algorithm:** Phased Successive Eliminations
**input:** $T$, $a \geq 1$

$\mathcal{K} \leftarrow [K]$
$\ell \leftarrow 0$
**while** $\mathrm{Card}(\mathcal{K}) > 1$ **do**
  pull all arms in $\mathcal{K}$ $\lceil a^\ell \rceil$ times
  **for** *all* $k \in \mathcal{K}$ *such that* $\hat{\mu}_k + \sqrt{\frac{2\ln T}{N_k(T)}} \leq \max_{i\in\mathcal{K}} \hat{\mu}_i - \sqrt{\frac{2\ln T}{N_i(T)}}$ **do** $\mathcal{K} \leftarrow \mathcal{K} \setminus \{k\}$
  $\ell \leftarrow \ell + 1$
**repeat** pull only arm in $\mathcal{K}$ **until** $t = T$

**1)** Show a regret bound similar to Successive Eliminations algorithm.

**2)** What is the role played by $a$?

**Exercise 6 :**

**Adapting to reward variance.** Let $X_1, \dots, X_N$ be a sequence of i.i.d. random variables with mean $\mu$, variance $\sigma^2$ and bounded support so that $X_t \in [0, M]$ almost surely. Define the estimators

$$\hat{\mu}_N = \frac{1}{N}\sum_{t=1}^{N} X_t$$
$$\hat{\sigma}^2 = \frac{1}{N}\sum_{t=1}^{N} (\hat{\mu} - X_t)^2.$$

We admit in the following the **empirical Bernstein** inequality:

$$\mathbb{P}\left( |\hat{\mu}_N - \mu| \geq \sqrt{\frac{2\hat{\sigma}^2}{N}\ln(3/\delta)} + \frac{3M}{N}\ln(3/\delta) \right) \leq \delta.$$

**1)** Show that the Bernstein inequality in the course implies here

$$\mathbb{P}\left( |\hat{\mu}_N - \mu| \geq \sqrt{\frac{2\sigma^2}{N}\ln(2/\delta)} + \frac{2M}{3N}\ln(2/\delta) \right) \leq \delta$$

Comment on the differences between the two above *Bernstein inequalities*.

**2)** Show that $\hat{\sigma}^2 = \frac{1}{N}\sum_{t=1}^{N}(X_t - \mu)^2 - (\hat{\mu}_N - \mu)^2$.

**3)** Is $\hat{\sigma}^2$ an unbiased estimator of $\sigma^2$? If not, can we easily make it unbiased?

**4)** Show that

$$\mathbb{P}\left(\hat{\sigma}^2 \geq \sigma^2 + \sqrt{\frac{2M^2\sigma^2}{N}\ln(1/\delta)} + \frac{2M^2}{3N}\ln(1/\delta)\right) \leq \delta.$$

**Hint:** Use Bernstein inequality of **1)**.

**5)** **(Hard)** Consider a bandit setting with $K$ arms, bounded rewards $X_k(t) \in [0, M]$ and the variance of the $k$-th arm is $\sigma_k^2$. Design a policy that depends on $M$, but does not need to know $\sigma_i$ a priori, such that there exists a universal constant $C > 0$ with

$$R_T \leq C \sum_{k, \Delta_k > 0} \left(\Delta_k + (M + \frac{\sigma_i^2}{\Delta_i})\ln T\right).$$

**Hint:** Without the stack of rewards model, the empirical Bernstein inequality can be extended (up to some changes) to cases where $N$ is a random variable.

**Exercise 7 :**

This exercise studies the celebrated Thompson sampling algorithm, described below.

In words, Thompson sampling starts with a prior distribution $\boldsymbol{p}_0$ distribution on the (mean) parameters of the bandits instance and at each round $t$, it draws random samples $\theta_k(t)$ from the posterior distribution $\boldsymbol{p}_{t-1}$ on the instance parameters at time $t-1$, which is defined as

$$\boldsymbol{p}_{t-1}(A) = \mathbb{P}\Big((\mu_1, \ldots, \mu_K) \in A \mid \mathcal{F}_{t-1}\Big) \quad \text{for any } A \in B(\mathbb{R}), \tag{1}$$

where $\mathcal{F}_{t-1} = \sigma\Big(U_1, X_{a_1}(1), U_2, X_{a_2}(2), \ldots X_{a_{t-1}}(t-1)\Big)$ and the $U_s$ are random variables uniformly drawn in $[0,1]$, that are independent with all other variables.

**Algorithm:** Thompson sampling
**input:** prior distribution $\boldsymbol{p}_0$
**for** $t = 1, \ldots, T$ **do**
  Sample $\boldsymbol{\theta}(t) \sim \boldsymbol{p}_{t-1}$
  Pull $a_t \in \operatorname{argmax}_{k \in [K]} \theta_k(t)$ `// Ties broken arbitrarily`
  Update $\boldsymbol{p}_t$ as the posterior distribution of the parameters, following Bayes rule.

We note for each time $t \in \mathbb{N}$ and arm $k \in [K]$:

$$S_k(t) = \sum_{s=1}^{t} X_k(s) \mathbb{1}_{a_s = k}.$$

**1)** Consider an instance of Bernoulli bandits, i.e., $\mathcal{D} = \{\text{Bernoulli}(\mu) \mid \mu \in [0,1]\}^K$. Show then that in the case of Bernoulli rewards with a uniform prior, at each time $t \in \mathbb{N}$, $\boldsymbol{p}_{t-1}$ is the joint distribut1ion of $K$ independent Beta distributions, where the $k$-th Beta distribution has parameters $(S_k(t-1)+1, N_k(t-1)-S_k(t-1)+1)$. In other words for any $t \in \mathbb{N}$, the drawn samples $\theta_k(t)$ are independent with each other conditioned on $\mathcal{F}_{t-1}$ and

$$\theta_k(t) \sim \mathrm{Beta}(S_k(t-1)+1, N_k(t-1)-S_k(t-1)+1).$$

**2)** Consider now that the prior is the improper uniform distribution[1] on $\mathbb{R}$ and Gaussian bandits with variance $\sigma^2$, i.e., $\mathcal{D} = \{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}\}^K$.
For any $t \in \mathbb{N}$, what is the distribution of $\boldsymbol{p}_{t-1}$ in this case?

[1] This can be seen as the uniform distribution on $\mathbb{R}$. It is not a proper distribution, since it is not of measure 1, but the Bayes rule can still be applied with it.