Exercise sheet n°1

In this session, we consider online learning with experts (see Lecture #1) with linear losses. The losses ℓ_{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 $(\ell_{1t},\dots,\ell_{Nt})_t \in ([0,1]^N)^{\mathbb{N}}$:

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

Solution: Consider ℓ_{jt} generated as i.i.d. random variables with distribution Bernoulli(1/2), then for any algorithm $\mathbb{E}[\sum_{t=1}^{T}\sum_{j=1}^{N} p_{jt}\ell_{jt}] = \frac{T}{2}$. Moreover as soon as $N \geq 2$, $\mathbb{E}[\sum_{t=1}^{T}\min_{k\in[N]}\ell_{kt}] \leq \frac{T}{4}$. In consequence, for any algorithm, there exists a sequence such that

$$\sum_{t=1}^{T}\sum_{j=1}^{N} p_{jt}\ell_{jt} - \sum_{t=1}^{T}\min_{k\in[N]}\ell_{kt} \geq \frac{T}{4}.$$

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 η?
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 (η_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 (η_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 $X \in [m, M]$:

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

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 $v_t \geq \dfrac{\eta_t(M-m)}{e^{\eta_t(M-m)} - \eta_t(M-m) - 1}(M-m)\delta_t$.

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

Solution: 1) $\eta_t \geq \eta_{t+1}$ so that $x \mapsto x^{\frac{\eta_t}{\eta_{t+1}}}$ is convex. Jensen inequality then writes for any $(x_j)_j$

$$\sum_{j=1}^N \frac{1}{N} x_j^{\frac{\eta_t}{\eta_{t+1}}} \geq \frac{1}{N^{\frac{\eta_t}{\eta_{t+1}}}}\left(\sum_{j=1}^N x_j\right)^{\frac{\eta_t}{\eta_{t+1}}}.$$

Taking $x_j = p_{jt}^{\frac{\eta_{t+1}}{\eta_t}} e^{-\eta_{t+1}\ell_{jt}}$ yields the result.

2) Taking logarithm of previous expression yields:

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

Summing over t yields, by telescoping,

$$\sum_{t=1}^{T} \frac{1}{\eta_t} \ln\left(\sum_{j=1}^{N} p_{jt} e^{-\eta_t \ell_{jt}}\right) \geq \frac{\ln N}{\eta_0} - \frac{\ln N}{\eta_{T+1}} + \frac{1}{\eta_{T+1}} \ln(\sum_j \exp(-\eta_{T+1} \sum_{s=1}^{T} \ell_{js})) - \frac{1}{\eta_0} \ln(N)$$
$$\geq -\frac{\ln N}{\eta_T} - \min_j \sum_{s=1}^{T} \ell_{js}.$$

This directly yields:

$$\sum_{t=1}^{T} \delta_t \geq -\frac{\ln N}{\eta_T} + \underbrace{\sum_j \sum_t p_j t \ell_{jt} - \min_j \sum_{s=1}^{T} \ell_{js})}_{R_T},$$

which allows to conclude.

3) (a) v_t is the variance of the r.v. given by $X = -\ell_{jt}$ with probability p_{jt}, and $\ln \mathbb{E}[e^{\eta_t X}] - \eta_t \mathbb{E}[X] = \eta_t \delta_t$. This inequality is then a direct application of Bernstein inequality.

(b) For $f : x \mapsto \frac{x}{e^x - x - 1} - \frac{2}{x}$, the previous inequality gives:

$$v_t \geq \frac{2\delta_t}{\eta_t} - f(\eta_t(M-m))(M-m)\delta_t.$$

A functional study then allows to say that $f(x) \geq -\frac{2}{3}$ for $x \geq 0$ (f is increasing on $\mathbb{R}_+$ and $f(0) = -\frac{2}{3}$).

4)(a) We have by telescoping

$$(\sum_{t=1}^{T} \delta_t)^2 = \sum_{t=1}^{T} \left((\sum_{s=1}^{t} \delta_s)^2 - (\sum_{s=1}^{t-1} \delta_s)^2 \right)$$
$$= \sum_{t=1}^{T} \delta_t (\underbrace{\delta_t}_{\leq (M-m) \text{ by direct bound}} + 2 \underbrace{\sum_{s=1}^{t-1} \delta_s}_{= \frac{\ln N}{\eta_t}})$$
$$\leq \sum_{t=1}^{T} (M-m)\delta_t + 2\frac{\delta_t}{\eta_t} \ln N$$
$$\leq (M-m) \sum_{t=1}^{T} \delta_t + \sum_{t=1}^{T} v_t \ln N + \frac{2}{3} \ln(N)(M-m)\delta_t,$$

where the last inequality comes from applying 3)(b).

(b) For $x = \sum_{t=1}^{T} \delta_t$, we just showed a second order inequality of the form: $x^2 \leq a + bx$ with $a = \sum_{t=1}^{T} v_t \ln N$ and $b = (M-m)(1 + \frac{2}{3}\ln N)$. This then yields

$$x \leq \frac{a + \sqrt{a^2 + 4b}}{2} \leq a + \sqrt{b},$$

i.e.,

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

Moreover, as η_t is non-increasing: $\frac{\ln N}{\eta_T} \leq \frac{\ln N}{\eta_{T+1}} = \sum_{t=1}^{T} \delta_t$. So that plugging the above bound in the inequality of question 2) yields

$$R_T \leq 2\sqrt{\sum_{t=1}^{T} v_t \ln N + (M-m)(2 + \frac{4}{3}\ln N)}.$$

We then conclude by noting that $v_t \leq (M-m)^2$.

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

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

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

Solution: With $\Delta_t = \left(\frac{K\ln(t)}{t}\right)^{\frac{1}{3}}$ and $c \in \mathbb{R}_+$,

$$\begin{aligned}
\mathbb{E}[\Delta_{a_t}] &= \sum_k \mathbb{P}(a_t = k)\Delta_k \\
&\leq 2c\Delta_t + \sum_k \mathbb{P}(a_t = k, \Delta_k > 2c\Delta_t) \\
&\leq 2c\Delta_t + \varepsilon_t + \sum_k \mathbb{P}(\hat{\mu}_k(t-1) - \mu_k \geq c\Delta_t) + \mathbb{P}(\mu^* - \hat{\mu}_{k^*} \geq c\Delta_t).
\end{aligned}$$

It just remains to bound $\mathbb{P}(\hat{\mu}_k(t-1) - \mu_k \geq c\Delta_t)$ in $\mathcal{O}\left(\frac{\Delta_t}{K}\right)$. Similarly to the proof of the instance dependent bound, we have for any x_t

$$\mathbb{P}(\hat{\mu}_k(t-1) - \mu_k \geq c\Delta_t) \leq x_t \mathbb{P}(N_k^{\mathrm{R}}(t-1) \leq x_t) + \frac{e^{-2c^2\Delta_t^2 x_t}}{2c^2\Delta_t^2}.$$

Taking $x_t = \frac{3\ln(K^{\frac{1}{3}}/\Delta_t)}{2c^2\Delta_t^2} = \frac{t^{\frac{2}{3}}\ln\left(\frac{t}{\ln(t)}\right)}{2c^2(K\ln(t))^{\frac{2}{3}}}$, it just remains to bound

$$x_t \mathbb{P}(N_k^{\mathrm{R}}(t-1) \leq x_t)$$

Similarly to the instance dependent proof, we can use Bernstein inequality, where here

$$\mathbb{E}[N_k^{\mathrm{R}}(t-1)] = \frac{1}{K}\sum_{s=1}^{t-1} \min\left(1, \frac{(K\ln(s))^{\frac{1}{3}}}{s^{\frac{1}{3}}}\right).$$

In particular, there exist constants c', c'' such that when $\frac{t}{\ln(t)} \geq c''K$, then $\mathbb{E}[N_k^{\mathrm{R}}(t-1)] \geq c'\frac{t^{\frac{2}{3}}\ln(t)^{\frac{1}{3}}}{K^{\frac{2}{3}}}$. Moreover here, c' is bounded away from 0 (e.g. $c' > 0.1$), while we can take c'' as large as possible. So when $\frac{t}{\ln(t)} \geq c''K$, we have for a large enough choice of c that $x_t \leq \frac{1}{2}\mathbb{E}[N_k^{\mathrm{R}}(t-1)]$. Bernstein inequality then yields

$$\begin{aligned} x_t \mathbb{P}(N_k^{\mathrm{R}}(t-1) \leq x_t) &\leq \mathbb{E}[N_k^{\mathrm{R}}(t-1)]\exp(-\frac{\mathbb{E}[N_k^{\mathrm{R}}(t-1)]}{5}) \\ &= \mathcal{O}\left(\frac{t^{\frac{2}{3}}\ln(t)^{\frac{1}{3}}}{K^{\frac{2}{3}}} e^{-\frac{c'}{5}\frac{t^{\frac{2}{3}}\ln(t)^{\frac{1}{3}}}{K^{\frac{2}{3}}}}\right) \end{aligned}$$

In particular, when $\frac{t^{\frac{2}{3}}\ln(t)^{\frac{1}{3}}}{K^{\frac{2}{3}}} \geq \frac{5\ln(t)}{c'}$, this term is in $\mathcal{O}\left(\frac{\Delta_t}{K}\right)$, so that we have the desired instantaneous regret bound when $\frac{t}{\ln(t)} \geq c''K$ for a large enough c'' (recall that c' is bounded away from 0). We can then conclude by noting that if $\frac{T}{\ln(T)} \leq c''K$,

$$R_T \leq T \leq T^{\frac{2}{3}}T^{\frac{1}{3}} \leq T^{\frac{2}{3}}(c''K\ln(T))^{\frac{2}{3}}.$$

If instead $\frac{T}{\ln(T)} > c''K$, let $\tau = \inf\{t \mid \frac{t}{\ln(t)} \geq c''K\}$. It comes

$$\begin{aligned} R_T &\leq \tau - 1 + \sum_{t=\tau}^{T} \mathbb{E}[\Delta_{a_t}] \\ &\leq \tau^{\frac{2}{3}}(K\ln(\tau))^{\frac{1}{3}} + \mathcal{O}\left(\sum_{t=\tau}^{T}\Delta_t\right) \\ &= \mathcal{O}\left(T^{\frac{2}{3}}(K\ln(T))^{\frac{1}{3}}\right). \end{aligned}$$

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

$T : \Omega \to \{1, 2, 3, \dots\}$ 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 $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ be the σ-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 $X_1, X_2, \dots$ 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?

Solution: 1) This is a consequence of Hoeffding inequality, when conditioning on the values of T.

2) Define $T = \{t \in \mathbb{N} \mid \hat{\mu}_T \geq \sqrt{\frac{2\ln(1/\delta)}{t}}\}$. Thanks to the law of iterated algorithm, it is properly defined ($T < +\infty$ almost surely) and satisfies the point **3)**.

3) Although T corresponds to $N_k(T)$ in our concentration lemma, it does not contradict with it. The reason is that $N_k(T)$ is almost surely bounded by T, and the failure probability is bounded by a term scaling with T. Allowing $N_k(T)$ to go to possibly infinite values thus makes the concentration irrelevant.

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 $\text{Card}(\mathcal{K}) > 1$ **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?

Solution: 1) Similar proof technique, but $E[N_k(T)]$ is now bounded by

$$\min \left\{ n = \sum_{\ell=0}^{\ell_k} \lceil a^\ell \rceil \mid n \geq \frac{32 \ln T}{\Delta_k^2} \text{ and } \ell_k \in \mathbb{N} \right\}.$$

By noting $n_\ell = \sum_{i=0}^{\ell} \lceil a^i \rceil$, we have $n_\ell < n_{\ell+1} \leq (a+1) n_\ell + 1$. This then implies with the above bound on $E[N_k(T)]$ that

$$E[N_k(T)] \leq 32(a+1) \frac{\ln T}{\Delta_k^2} + 1.$$

This allows to conclude.

2) a plays a trade-off between

- the incurred regret,
- the frequence at which we need to update the policy (and number of arm switches).

Exercise 6 :

Adapting to reward variance. Let $X_1, \ldots, X_N$ be a sequence of i.i.d. random variables with mean μ, variance σ^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 σ^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 σ_k^2. Design a policy that depends on M, but does not need to know σ_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.

Solution: 1) In the course, the Bernstein inequality reads (after rescaling by $\frac{N}{M}$) as

$$\mathbb{P}\left(|\hat{\mu}_T - \mu| \geq \varepsilon\right) \leq 2\exp(-\frac{N\varepsilon^2}{2\sigma^2 + \frac{2}{3}M\varepsilon}).$$

Taking $\varepsilon = \sqrt{\frac{2\sigma^2}{N}\ln(2/\delta)} + \frac{2M}{3N}\ln(2/\delta)$ allows to conclude. The main difference is the one we admit holds for the estimated variance, while the one we just proved only holds for the true variance (which is generally unknown).

2) By definition

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{N}\sum_{t=1}^{N}(\hat{\mu}_N - X_t)^2 \\ &= \frac{1}{N}\sum_{t=1}^{N} X_t^2 - 2\hat{\mu}_N^2 + \hat{\mu}_N^2 \\ &= \frac{1}{N}\sum_{t=1}^{N}\left(X_t^2 - 2X_t + \mu^2\right) - \left(\hat{\mu}_N^2 + \mu^2 - \frac{2}{N}\sum_{t=1}^{N} X_t\mu\right) \\ &= \frac{1}{N}\sum_{t=1}^{N}(X_t - \mu)^2 - (\hat{\mu}_N - \mu)^2.\end{aligned}$$

3) Obviously, $\mathbb{E}[\hat{\sigma}^2] < \mathbb{E}[(X_t - \mu)^2]$, so it is biased. Actually we have

$$\begin{aligned}\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}[(X_t - \mu)^2] - \mathbb{E}[(\hat{\mu}_N - \mu)^2] \\ &= \mathbb{E}[(X_t - \mu)^2] - \mathrm{Var}\left(\hat{\mu}_N\right) \\ &= \mathbb{E}[(X_t - \mu)^2] - \frac{1}{N}\mathrm{Var}\left(X_t\right) \qquad \text{variance of sum of independent variables is sum of variances} \\ &= \frac{N-1}{N}\mathrm{Var}\left(X_t\right).\end{aligned}$$

$\frac{N}{N-1}\hat{\sigma}^2$ is an unbiased estimator.

4) This is a consequence of the Bernstein inequality **1)** and the fact that $\mathrm{Var}\left((X_t - \mu)^2\right) \leq M^2\sigma^2$.

5) The algorithm to propose a the variant of UCB that pulls at each time step

$$a_t \in \operatorname*{argmax}_k \hat{\mu}_k(t-1) + \underbrace{\sqrt{\frac{2\hat{\sigma}_k^2(t-1)}{N_k(t)}\ln(3t^3)} + \frac{3M}{N_k(t)}\ln(3t^3)}_{:=B_k(t-1)}.$$

We can show similarly to the course that for each time step t,

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

$$\mathbb{P}\left(\hat{\sigma}_k^2(t) \geq \sigma_k^2 + \underbrace{\sqrt{\frac{2M^2\sigma_k^2}{N_k(t)}\ln(1/t^3)} + \frac{2M^2}{3N_k(t)}\ln(1/t^3)}_{:=S_k(t)}\right) \leq \frac{1}{t^2}.$$

So that here for $\delta = t^3$, we have for the good event

$$\mathcal{E}_{k,t} = \Big\{ \hat{\mu}_k(t) - \mu_k \leq B_k(t) \text{ and } \mu^* - \hat{\mu}_{k^*}(t) \leq B_k(t) \\ \text{and } \hat{\sigma}_k^2(t) \leq \sigma_k^2 + S_k(t) \Big\}$$

and $\mathbb{P}(\mathcal{E}_{k,t}) \geq 1 - \frac{2}{t^2}$. In particular if $\mathcal{E}_{k,t}$ holds, we pull the arm k at time $t+1$ if $2B_k(t) \geq \Delta_k$ (similar to UCB proof). Now, we bound $B_k(t)$ using **4)** under the event $\mathcal{E}_{k,t}$.

$$\begin{aligned} B_k(t) &= \sqrt{\frac{2\hat{\sigma}_k^2(t)}{N_k(t)} \ln(3t^3)} + \frac{3M}{N_k(t)} \ln(3t^3) \\ &\leq \sqrt{\frac{2\sigma_k^2 + 2S_k(t)}{N_k(t)} \ln(3t^3)} + \frac{3M}{N_k(t)} \ln(3t^3) \\ &\leq \sqrt{\frac{2\sigma_k^2}{N_k(t)} \ln(3t^3)} + \sqrt{\frac{2S_k(t)}{N_k(t)} \ln(3t^3)} + \frac{3M}{N_k(t)} \ln(3t^3). \end{aligned}$$

Note that

$$\begin{aligned} \sqrt{\frac{2S_k(t)}{N_k(t)} \ln(3t^3)} &\leq \sqrt{\frac{2}{N_k(t)} \sqrt{\frac{2M^2\sigma_k^2}{N_k(t)} \ln(3t^3) \ln(3t^3)} + \sqrt{\frac{4M^2}{3N_k(t)^2} \ln(3t^3)^2} } \\ &= \sqrt{\frac{2M \ln(3t^3)}{N_k(t)} \sqrt{\frac{2\sigma_k^2}{N_k(t)} \ln(3t^3)} + \frac{2M}{\sqrt{3}N_k(t)} \ln(3t^3)} \\ &\leq \sqrt{\frac{\sigma_k^2}{2N_k(t)} + \frac{(1 + \frac{2}{\sqrt{3}})M}{N_k(t)} \ln(3t^3)} \qquad \frac{x+y}{2} \geq \sqrt{xy}. \end{aligned}$$

So that there exists a constant $C > 0$ such that

$$B_k(t) \leq C \sqrt{\frac{\sigma_k^2}{N_k(t)} \ln(3t^3)} + C \frac{M}{N_k(t)} \ln(3t^3).$$

As a consequence, if $\mathcal{E}_{k,t}$ holds, $a_{t+1} = k$ implies that

$$2C \sqrt{\frac{\sigma_k^2}{N_k(t)} \ln(3t^3)} + 2C \frac{M}{N_k(t)} \ln(3t^3) \geq \Delta_k.$$

This is a polynomial inequality, which yields for $a = 2CM$, $b = 2C\sigma_k^2$ and $c = \Delta_k$:

$$\begin{aligned}\sqrt{\frac{\ln(3t^3)}{N_k(t)}} &\geq \frac{b}{2a}\left(\sqrt{1+\frac{4ac}{b^2}}-1\right)\\ &\geq \frac{b}{2a}\min\left(\frac{ac}{b^2},\frac{\sqrt{ac}}{b}\right) && \sqrt{1+z}-z \geq \min(\frac{z}{4},\frac{\sqrt{z}}{2})\\ &= \frac{1}{2}\min\left(\frac{c}{b},\sqrt{\frac{c}{a}}\right).\end{aligned}$$

This directly implies for a large enough constant C'

$$\begin{aligned}N_k(t) &\leq C'\ln(3t^3)\max\left(\frac{M}{\Delta_k},\frac{\sigma_k^2}{\Delta_k^2}\right)\\ &\leq C'\ln(3t^3)\left(\frac{M}{\Delta_k}+\frac{\sigma_k^2}{\Delta_k^2}\right).\end{aligned}$$

From there, we can conclude similarly to the end of the proof in the course for UCB.

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

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

where $\mathcal{F}_{t-1} = \sigma\Big(U_1, X_{a_1}(1), U_2, X_{a_2}(2), \dots X_{a_{t-1}}(t-1)\Big)$ 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 $t = 1,\dots,T$ **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 distribution 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

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

2) Consider now that the prior is the improper uniform distribution[1] on $\mathbb{R}$ and Gaussian bandits with variance σ^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?

Solution: 1) As the prior is continuous, Bayes rule yields that the posterior is also continuous. Its density $p_t(\boldsymbol{\mu})$ is then proportional by Bayes rule to

$$\begin{aligned} p_t(\boldsymbol{\mu}) &\propto \mathbb{P}(\forall k, \sum_{t=1}^{N_k(t)} X_k(t) = S_k(t) \mid (N_k(t), S_k(t), \mu_k)_k) \\ &= \prod_{k=1}^{K} \mathbb{P}(k, \sum_{t=1}^{N_k(t)} X_k(t) = S_k(t) \mid N_k(t), S_k(t), \mu_k) \\ &= \prod_{k=1}^{K} \mathbb{P}(G_{\mu_k, N_k(t)} = S_k(t)), \end{aligned}$$

where $G_{\mu_k, N_k(t)}$ is a binomial r.v. of parameters $(N_k(t), \mu_k)$, so that

$$p_t(\boldsymbol{\mu}) \propto \prod_{k=1}^{K} \binom{N_k(t)}{S_k(t)} \mu_k^{S_k(t)} (1-\mu_k)^{N_k(t) - S_k(t)},$$

which exactly corresponds to the product of independent Beta distributions as described in the question.

2) In that case, we again have that the density of the posterior is proportional to

$$\begin{aligned} p_t(\boldsymbol{\mu}) &\propto \prod_{k=1}^{K} \mathbb{P}(k, \sum_{t=1}^{N_k(t)} X_k(t) = S_k(t) \mid N_k(t), S_k(t), \mu_k) \\ &= \prod_{k=1}^{K} \mathbb{P}(G_{\mu_k, N_k(t)} = S_k(t)), \end{aligned}$$

[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.

where $G_{\mu_k,N_k(t)}$ is a Gaussian r.v. of mean $N_k(t)\mu_k$ and variance $N_k(t)\sigma^2$. This finally gives

$$p_t(\boldsymbol{\mu}) \propto \prod_{k=1}^{K} e^{-\frac{(S_k(t)-N_k(t)\mu_k)^2}{2\sigma^2 N_k(t)}}$$
$$= \prod_{k=1}^{K} e^{-\frac{N_k(t)(S_k(t)/N_k(t)-\mu_k)^2}{2\sigma^2}},$$

so that the posterior corresponds to the joint distribution of K independent Gaussian variables, with mean $S_k(t)/N_k(t)$ and variance $\sigma^2/N_k(t)$.

Exercises seen in class (if time)

Exercise 8 :

1) Give an example where both

(a) $Z_T \xrightarrow{\mathcal{L}} Z$,

(b) f is continuous,

but $\lim_{T\to\infty} \mathbb{E}[f(Z_T)] \neq \mathbb{E}[f(Z)]$.

Definition. *We say that* $(Y_T)_T$ *is* ***uniform asymptotic integrable*** *(uai) if*

$$\lim_{L\to\infty} \lim_{T\to\infty} \mathbb{E}[\|Y_T\| \mathbb{1}_{\|Y_T\|>L}] = 0.$$

2) Show that if f is continuous, $Z_T \xrightarrow{\mathcal{L}} Z$ and $(f(Z_T))_T$ is uai, then

(a) $f(Z_T) \in \mathbb{L}^1$ for T large enough;

(b) $f(Z) \in \mathbb{L}^1$;

(c) $\mathbb{E}[f(Z_T)] \to_{T\to\infty} \mathbb{E}[f(Z)]$.

Hint: for b), use Skorokhod's theorem.

3) Show that if $(Y_T)_T$ is bounded in $\mathbb{L}^p$ for $p > 1$, i.e. $\sup_{T\geq 1} \mathbb{E}[\|Y_T\|^p] = B < +\infty$, then $(Y_T)_T$ is uai.

Solution: 1) $Z_T = (1 - \frac{1}{T})\delta_0 + \frac{1}{T}\delta_T$.

2) a) definition of the limit and using that $\mathbb{E}[\|f(Z_T)\|] \leq L + \mathbb{E}[\|f(Z_T)\| \mathbb{1}_{\|f(Z_T)\|>L}]$.
b) Skorokhod's theorem with Fatou lemma
c) $|\mathbb{E}[f(Z_T)] - \mathbb{E}[f(Z)]| \leq |\mathbb{E}[\varphi_L(f(Z_T))] - \mathbb{E}[f(Z)]| + |\mathbb{E}[\varphi_L(f(Z_T))] - \mathbb{E}[f(Z_T)]|$ for φ_L the clipping operator in $[-L, L]$.

Going to lim sup:

$$\limsup_T |\mathbb{E}[f(Z_T)] - \mathbb{E}[f(Z)]| \leq |\mathbb{E}[\varphi_L(f(Z))] - \mathbb{E}[f(Z)]| + \limsup_T \mathbb{E}[\varphi_L(f(Z_T))] - \mathbb{E}[f(Z_T)]|.$$

The first term is 0 by dominated convergence. The second is to be handled with the uai property, by taking $\limsup_L$.

3) $x^p \gg x$ for x large enough. In particular, $\forall M > 0, \exists L_M, \forall x \geq L_M, x^p \geq Mx$. Then for such L_M,

$$\begin{aligned}\mathbb{E}[\|Y_T\| \mathbb{1}_{\|Y_T\|>L_M}] &\leq \frac{1}{M}\mathbb{E}[\|Y_T\|^p \mathbb{1}_{\|Y_T\|>L_M}] \\ &\leq \frac{B}{M}.\end{aligned}$$

Taking $M \to \infty$, a monotonicity argument concludes.

Exercise 9 :

Sub-Gaussian random variables. Let X be a **centered** random variable in $\mathbb{R}$. Show that affirmations below satisfy the following implications chain: 1. $\Longrightarrow$ 2. $\Longrightarrow$ 3. $\Longrightarrow$ 4. $\Longrightarrow$ 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 $q \in \mathbb{N}^*$, $\mathbb{E}[X^{2q}] \leq 2q!(2\sigma^2)^q$;
4. *Orlicz condition:* $\mathbb{E}[\exp(\frac{X^2}{4\sigma^2})] \leq 4$;
5. *Laplace transform:* for any $\eta \in \mathbb{R}$, $\ln(\mathbb{E}[e^{\eta X}]) \leq \frac{20\sigma^2\eta^2}{2}$.

Solution: 1) $\Longrightarrow$ 2) is Hoeffding inequality for a single random variable.

For 2) $\Longrightarrow$ 3),

$$\begin{aligned}
\mathbb{E}[X^{2q}] &= \int_0^{+\infty} \mathbb{P}(X^{2q} > u)\mathrm{d}u \\
&= \int_0^{+\infty} \mathbb{P}(|X| > u^{\frac{1}{2q}})\mathrm{d}u \\
&\leq 2\int_0^{+\infty} \exp(\frac{-u^{1/q}}{2\sigma^2})\mathrm{d}u \\
&= (2\sigma^2)^q 2q \int_0^{+\infty} \exp(-v)v^{q-1}\mathrm{d}v \qquad\qquad v = \frac{u^{1/q}}{2\sigma^2} \\
&= (2\sigma^2)^q 2q\Gamma(q) \\
&= 2(2\sigma^2)^q q!
\end{aligned}$$

For 3) $\Longrightarrow$ 4), the monotone convergence theorem gives

$$\mathbb{E}[\exp(\frac{X^2}{4\sigma^2})] = \mathbb{E}[\sum_{k=0}^{\infty} \frac{X^{2k}}{(2\sigma^2)^k k!}\frac{1}{2^k}] \leq 2\sum_{k=0}^{\infty}\frac{1}{2^k} = 4.$$

For 4) $\implies$ 5), using the fact that X is centered, we have for any $\eta \in \mathbb{R}$

$$\begin{aligned}
\mathbb{E}[\exp(\eta X)] &= \mathbb{E}[\sum_{k=0}^{\infty} \frac{(\eta X)^k}{k!}] \\
&= 1 + \mathbb{E}[\sum_{k=2}^{\infty} \frac{(\eta X)^k}{k!}] \\
&\leq 1 + \frac{\eta^2}{2}\mathbb{E}[X^2 \exp(|\eta X|)] \\
&\leq 1 + \frac{\eta^2}{2}\exp(2\sigma^2\eta^2)\mathbb{E}[X^2 \exp(\frac{X^2}{8\sigma^2})] && \inf_a(\frac{\eta^2}{2a} + \frac{aX^2}{2}) = \eta|X|, a = \frac{1}{4\sigma^2} \\
&\leq 1 + 2\sigma^2\eta^2 \exp(2\sigma^2\eta^2)\mathbb{E}[\exp(\frac{X^2}{4\sigma^2})] && z \leq \exp(\frac{z}{2}) \\
&\leq (1 + 8\sigma^2\eta^2)\exp(2\sigma^2\eta^2) \leq \exp\left(\frac{20\sigma^2\eta^2}{2}\right) && 1 + z \leq e^z.
\end{aligned}$$

Exercise 10 :

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.

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

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

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

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

Solution: **1)** Observe that $N_k(T) \leq T$, so that

$$\begin{aligned}
\Delta_k\mathbb{E}[N_k(T)] &\leq C_0 + \min\left\{\Delta_k T, \frac{C\ln(T)}{\Delta_k}\right\} \\
&\leq C_0 + \sqrt{C\ln(T)T}.
\end{aligned}$$

2) The finer analysis consists in saying that

$$\begin{aligned}
R_T &= \sum_{k=1}^{K} \Delta_k \mathbb{E}[N_k(T)] \\
&\leq \sum_{k=1}^{K} \min\left\{ \Delta_k \mathbb{E}[N_k(T)], C_0 + \frac{C\ln(T)}{\Delta_k} \right\} \\
&\leq \sum_{k=1}^{K} \sqrt{\mathbb{E}[N_k(T)]}\sqrt{C_0 + C\ln(T)} \\
&\leq \sqrt{C_0 + C\ln(T)} \sqrt{K \sum_{k=1}^{K} \mathbb{E}[N_k(T)]} && \text{Cauchy Schwarz} \\
&\leq \sqrt{KT(C_0 + C\ln(T))}.
\end{aligned}$$

Exercise 11 :

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 $T_1 < T_2 > T_3 < \dots$, 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:

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

Solution: **1)** is by definition of $\tilde{\pi}$.
2) is for the choice $T_\ell = 2^\ell$.
3) directly derives from the choice of n_ℓ.
4) $T_\ell = 2^{2^\ell}$.