# Sequential Learning

Final Examination

The duration of the exam is 2 hours. A single two-sided sheet of handwritten notes (with any content) is allowed. Answers can be written in French or English.
This exam is made of 3 parts. The first part contains varied questions on the course. Parts 2 and 3 are exercices on adversarial and stochastic online learning respectively.

## Part 1. Appetizers

1. Let $g_1, \dots, g_T \in [0,1]^d$ and $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1\}$ and let $\ell_t(\theta) = \langle \theta, g_t \rangle$. Let $\theta_1, \dots, \theta_T \in \Theta$ be the predictions of an algorithm that aims at minimizing the regret $R_T := \sum_{t=1}^T \ell_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=1}^T \ell_t(\theta)$.
   (a) Show that there exists a sequence $(g_t)$ such that $\theta_t \in \arg\min_{\theta \in \Theta} \sum_{s=1}^{t-1} \langle \theta, g_s \rangle$ incurs linear regret $R_T \geq (1 - 1/d)T$.
   (b) Let $\Delta_{2d}$ be the simplex of dimension $2d$. Define a linear surjection $A : \Delta_{2d} \to \Theta$ such that for any $\theta \in \Theta$ and $g_t \in [0,1]^d$ there exists $p \in \Delta_{2d}$ and $h_t \in [0,1]^{2d}$ (to be specified) with $\theta = Ap$ and $\ell_t(\theta) = \langle p, h_t \rangle$.
   (c) Give the pseudo code of the exponentially weighted average algorithm to minimize the regret with respect to $p \in \Delta_{2d}$.
   (d) Give the order in $d$ and $T$ of the associated regret upper-bound on $R_T$ and tell what would be the difference for online gradient descent.

2. What is the difference between a distribution-dependent and a distribution-free regret bound? What are the two corresponding bounds achieved by the Upper-Confidence-Bound algorithm?

3. (UCB) Consider a stochastic bandit with $K$ arms, distributions with support in $[0,1]$ and means $\mu_1, \dots, \mu_k$. The UCB algorithm pulls arm $a_t = \arg\max \widehat{\mu}_k(t) + \sqrt{\frac{2 \log t}{N_k(t)}}$, where $N_k(t) = \sum_{s=1}^{t-1} \mathbb{I}\{a_s = k\}$ is the number of pulls of arm $k$ before $t$ and $\widehat{\mu}_k(t)$ is an estimation of the mean of arm $k$. **Suppose** that for all $t \in \{1, \dots, T\}$, for all $k \in [K]$, $|\mu_k - \widehat{\mu}_k(t)| \leq \sqrt{\frac{2 \log t}{N_k(t)}}$.
   (a) Show that
   $$N_k(a_t) \leq \frac{8 \log t}{\Delta_k^2}\,.$$
   (b) Prove an upper bound on the regret $R_T = T \max_{k \in [K]} \mu_k - \sum_{t=1}^T \mu_{a_t}$.

## Part 2. Simple Regret minimization

We consider the stochastic bandit setting: an algorithm sequentially interacts with $K$ arms ($K > 1$), where each arm $k \in \{1, \dots, K\}$ is described by a distribution $\nu_k$ supported on $[0, 1]$ with mean $\mu_k$.

In opposition to the classical objective of regret minimization, we here consider the pure exploration problem of simple regret minimization: at the end of the game, the algorithm returns a final decision $a_{T+1} \in \{1, \dots, K\}$ **that can be randomized** and aims at minimizing:

$$R_T^{\text{simple}} = \max_k \mu_k - \mu_{a_{T+1}}.$$

In this part, we study the Uniform Exploration Algorithm (UE) described by Algorithm (1) below.

While $t \leq T$ do
  For $k = 1$ to $K$ do
    – Pull arm $k$
  End for
End While
Return $a_{T+1} \in \arg\max_k \widehat{\mu}_k(T)$.

Algorithm 1: Uniform Exploration

4. In this question, we will bound the expected regret $\mathbb{E}[R_T^{\text{simple}}]$ of UE.
   (a) Note in the following $\Delta_k = \max_j \mu_j - \mu_k$. Prove that for any arm $k \in [K]$,
   $$\mathbb{P}[a_{T+1} = k] \leq \exp\left(-4 \left\lfloor \frac{T}{K} \right\rfloor \Delta_k^2\right).$$
   (b) For $T \geq K$, show that for any $\tilde{\Delta} \geq \sqrt{\frac{2}{\lfloor \frac{T}{K} \rfloor}}$, the expected simple regret of UE can be bounded as
   $$\mathbb{E}[R_T^{\text{simple}}] \leq \tilde{\Delta} + K\tilde{\Delta} \exp\left(-2 \left\lfloor \frac{T}{K} \right\rfloor \tilde{\Delta}^2\right).$$
   (c) Taking a well chosen value of $\tilde{\Delta}$ with the above bound, show that we can bound the expected simple regret of UE as follows for $T \geq K$:
   $$\mathbb{E}[R_T^{\text{simple}}] \leq c\sqrt{\frac{K \ln(K)}{T}},$$
   where $c$ is a universal constant to specify.

5. In this question, we will prove lower bounds on the regret of any algorithm. For this question, we consider Gaussian distributions. Let $\Delta > 0$, we consider in the following $K + 1$ bandit instances $(\nu^j)_{0 \leq j \leq K}$, where
   $$\nu_k^j = \mathcal{N}(0, 1) \quad \text{for any } k \in [K] \text{ such that } j \neq k$$
   $$\nu_k^k = \mathcal{N}(\Delta, 1) \quad \text{for any } k \in [K].$$

We write $\mathbb{E}_{\nu^i}[\cdot]$ (respectively $\mathbb{P}_{\nu^i}[\cdot]$) for the expectation (respectively probability) when the algorithm plays on problem $\nu^i$.

(a) Justify that for any algorithm there exists $i \in [K]$ such that both hold:

$$\mathbb{E}_{\nu^0}[N_i(T)] \leq \frac{T}{K-1} \qquad \text{and} \qquad \mathbb{P}_i[a_{T+1} = i] \leq \frac{1}{2}, \tag{1}$$

and that $\mathbb{E}_{\nu^i}[R_T^{\text{simple}}] = (1 - \mathbb{P}_{\nu^i}[a_{T+1} = i])\Delta$.

(b) Recall the fundamental inequality, which relates the expectations $\mathbb{E}_{\nu^0}[Z]$ and $\mathbb{E}_{\nu^i}[Z]$ with $\mathbb{E}_{\nu^0}[N_k(T)]$ and $\mathrm{KL}(\nu_k^0, \nu_k^i)$ (Kullback-Leibler divergence between the Gaussian distributions of arm $k$ under the two bandit problems) for all $k \in [K]$, where $Z$ is some random variable satisfying conditions to specify.

(c) Admit that for Gaussian distributions $\nu = \mathcal{N}(\mu, 1)$ and $\nu' = \mathcal{N}(\mu', 1)$, the Kullback-Leibler divergence is given by $\mathrm{KL}(\nu, \nu') = \frac{(\mu-\mu')^2}{2}$. Also admit that for two Bernoulli distributions of parameter $p, q$, $\mathrm{KL}(\mathrm{Ber}(p), \mathrm{Ber}(q)) \geq 2(p-q)^2$.

Then show, using the fundamental inequality, that for any algorithm, with $i$ satisfying Equation (1):

$$\mathbb{E}_{\nu^i}[R_T^{\text{simple}}] \geq \frac{\Delta}{2} - \sqrt{\frac{T}{K-1}}\frac{\Delta^2}{2}.$$

(d) Prove that there exists a universal constant $C$ such that for any algorithm, there exists a Gaussian bandit problem $\nu$ with mean rewards in $[0, 1]$ such that

$$\mathbb{E}_{\nu}[R_T^{\text{simple}}] \geq C\sqrt{\frac{K-1}{T}}. \tag{2}$$

A lower bound similar to Equation (2) can be shown when the distributions have bounded support in $[0, 1]$. We will now show that this lower bound can be reached by some algorithm.

6. Denote $R_T(\pi)$ the cumulative pseudo-regret of a bandit algorithm $\pi$: $R_T(\pi) = T\max_k \mu_k - \sum_{t=1}^T \mu_{a_t}$, where the decisions $a_t$ depend on the algorithm $\pi$. Similarly, we now denote the simple regret of an algorithm $\tilde{\pi}$ as $R_T^{\text{simple}}(\tilde{\pi})$ in the following to avoid any confusion.

(a) Show that for any multi-armed bandits algorithm $\pi$ with cumulative pseudo-regret $R_T(\pi)$, we can extend it to a simple regret minimisation algorithm $\tilde{\pi}$ such that

$$\mathbb{E}[R_T^{\text{simple}}(\tilde{\pi})] = \frac{\mathbb{E}[R_T(\pi)]}{T}.$$

(b) Admit we have a multi-armed bandits algorithm $\pi$ (e.g., MOSS) and a constant $c > 0$ such that for any bandit instance $(\nu_k)_{k\in[K]}$ with $\nu_k$ supported on $[0, 1]$: $\mathbb{E}[R_T(\pi)] \leq c\sqrt{KT}$.

Then show that some algorithm $\tilde{\pi}$ has a simple regret for any bandit instance $(\nu_k)_{k\in[K]}$ with $\nu_k$ supported on $[0, 1]$ bounded as

$$\mathbb{E}[R_T^{\text{simple}}(\tilde{\pi})] \leq c\sqrt{\frac{K}{T}}.$$

## Part 3. Online Mirror Descent

Let $\Theta \subseteq \mathbb{R}^d$ be a compact convex decision space. We consider the following setting. At each $t \geq 1$, the learner chooses $\theta_t \in \Theta$, the environment chooses a vector $g_t \in \mathbb{R}^d$ and reveals it to the learner. The goal of the learner is to minimize his linear regret

$$\text{Regret}_T(\theta) = \sum_{t=1}^{T} \langle g_t, \theta_t - \theta \rangle \qquad \text{for all } \theta \in \Theta\,.$$

Let $R : \Theta \to \mathbb{R}$ be a sub-differentiable mirror-map, which is $\mu$-strongly convex with respect to some norm $\|\cdot\|$. We consider the Online Mirror Descent algorithm defined by:

$$\theta_{t+1} = \operatorname*{arg\,min}_{\theta \in \Theta} \left\{ \langle g_t, \theta \rangle + \frac{1}{\eta} D_R(\theta, \theta_t) \right\}$$

where $\eta > 0$ and $D_R(x, y) = R(x) - R(y) - \langle \nabla R(y), x - y \rangle$ is the Bregman divergence associated with $R$.

7. How would you adapt the above algorithm if the losses $\ell_t : \Theta \to \mathbb{R}$ are convex instead of linear?
8. Explain how the above algorithm generalizes (without proving it in details):
   (a) the Online Gradient Descent algorithm;
   (b) the Exponentially Weighted Average algorithm when $\Theta = \Delta_d$.
9. Show that for any $\theta \in \Theta$: $\langle g_t, \theta_{t+1} - \theta \rangle \leq \frac{1}{\eta} \langle \nabla R(\theta_{t+1}) - \nabla R(\theta_t), \theta - \theta_{t+1} \rangle$ .
10. Deduce that $\langle g_t, \theta_{t+1} - \theta \rangle \leq \frac{1}{\eta} \Big( D_R(\theta, \theta_t) - D_R(\theta, \theta_{t+1}) - \frac{\mu}{2} \|\theta_t - \theta_{t+1}\|^2 \Big)$ .
11. Show that $\langle g_t, \theta_{t+1} - \theta_t \rangle \leq \frac{\mu}{2\eta} \|\theta_t - \theta_{t+1}\|^2 + \frac{\eta}{2\mu} \|g_t\|_*^2$ , where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.
12. Conclude by providind an upper-bound on the regret assuming that $\|g_t\|_* \leq G$ for all $t$. Optimize over $\eta$.