

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 Δ_{2d} be the simplex of dimension $2d$. Define a linear surjection $A : \Delta_{2d} \rightarrow \Theta$ 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 μ_1, \dots, μ_K . The UCB algorithm pulls arm $a_t = \arg \max_k \hat{\mu}_k(t) + \sqrt{\frac{2 \log t}{N_k(t)}}$, 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 $\hat{\mu}_k(t)$ is an estimation of the mean of arm k . **Suppose** that for all $t \in \{1, \dots, T\}$, for all $k \in [K]$, $|\mu_k - \hat{\mu}_k(t)| \leq \sqrt{\frac{2 \log t}{N_k(t)}}$.
 - (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 ν_k supported on $[0, 1]$ with mean μ_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  $a_{T+1} \in \arg \max_k \hat{\mu}_k(T)$ .

```

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 $k \in [K]$,

$$\mathbb{P}[a_{T+1} = k] \leq \exp \left(-4 \left\lfloor \frac{T}{K} \right\rfloor \Delta_k^2 \right).$$

Solution: This is a direct application of Hoeffding's inequality, **as the pulls are here deterministic**.

(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

$$\mathbb{E}[R_T^{\text{simple}}] \leq \tilde{\Delta} + K \tilde{\Delta} \exp \left(-2 \left\lfloor \frac{T}{K} \right\rfloor \tilde{\Delta}^2 \right).$$

Solution: The simple regret can be bounded as:

$$\mathbb{E}[R_T^{\text{simple}}] \leq \tilde{\Delta} + \sum_{k, \Delta_k > \tilde{\Delta}} \Delta_k \exp \left(-2 \left\lfloor \frac{T}{K} \right\rfloor \Delta_k^2 \right).$$

The solution then comes from noticing that the function $\Delta \mapsto \Delta \exp \left(-2 \left\lfloor \frac{T}{K} \right\rfloor \Delta^2 \right)$ is decreasing on the interval $[\sqrt{\frac{2}{\lfloor \frac{T}{K} \rfloor}}, +\infty)$.

- (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.

Solution: Assuming, $K > 1$, take $\tilde{\Delta} = \sqrt{\frac{2 \ln(K)}{\ln(2) \lfloor \frac{T}{K} \rfloor}} \geq \sqrt{\frac{2}{\lfloor \frac{T}{K} \rfloor}}$.

Applying (b) then yields

$$\mathbb{E}[R_T^{\text{simple}}] \leq \tilde{\Delta} + K \tilde{\Delta} \exp\left(-\frac{4}{\ln(2)} \ln(K)\right) \leq 2\tilde{\Delta}.$$

The constant $c = \frac{4}{\sqrt{\ln(2)}}$ then comes by noticing that $\lfloor \frac{T}{K} \rfloor \geq \frac{T}{2K}$ when $T \geq K$ (any valid constant is accepted).

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

$$\begin{aligned} \nu_k^j &= \mathcal{N}(0, 1) \quad \text{for any } k \in [K] \text{ such that } j \neq k \\ \nu_k^k &= \mathcal{N}(\Delta, 1) \quad \text{for any } k \in [K]. \end{aligned}$$

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

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

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

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

Solution: There is at most one arm k_0 such that $\mathbb{P}_i[a_{T+1} = i] > \frac{1}{2}$. Then by noticing that $\sum_{k \neq k_0} [N_k(T)] \leq T$, we can choose $i \neq k_0$ such that $\mathbb{E}_{\nu^0}[N_i(T)] \leq \frac{T}{K-1}$.

- (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 $\text{KL}(\nu_k^0, \nu_k^i)$ (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.

Solution:

$$\sum_k \mathbb{E}_{\nu^0}[N_k(T)] \text{KL}(\nu_k^0, \nu_k^i) \geq \text{KL}(\text{Ber}(\mathbb{E}_{\nu^0}[Z]), \text{Ber}(\mathbb{E}_{\nu^i}[Z])),$$

where Z is a random variable in $[0, 1]$, H_T measurable where H_T is the history of observations + eventual randomisation.

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

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

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

Solution: Taking $Z = \mathbf{1}(a_{T+1} \neq i)$, we then have with the fundamental inequality (and admitted results on the KL):

$$\mathbb{E}_{\nu^0}[N_i(T)] \frac{\Delta^2}{2} \geq 2 (\mathbb{P}_{\nu^0}(a_{T+1} = i) - \mathbb{P}_{\nu^i}(a_{T+1} = i))^2,$$

which leads to

$$\mathbb{P}_{\nu^i}(a_{T+1} = i) \leq \mathbb{P}_{\nu^0}(a_{T+1} = i) + \frac{\Delta}{2} \sqrt{\mathbb{E}_{\nu^0}[N_i(T)]}.$$

Then

$$\begin{aligned} \mathbb{E}_{\nu^i}[R_T^{\text{simple}}] &= (1 - \mathbb{P}_{\nu^i}[a_{T+1} = i])\Delta \\ &\geq (1 - \mathbb{P}_{\nu^0}[a_{T+1} = i])\Delta - \frac{\Delta^2}{2} \sqrt{\mathbb{E}_{\nu^0}[N_i(T)]} \end{aligned}$$

The solution then follows from $\mathbb{P}_{\nu^0}(a_{T+1} = i) \leq \frac{1}{2}$ and $\mathbb{E}_{\nu^0}[N_i(T)] \leq \frac{T}{K-1}$.

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

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

Solution: Take, e.g., $\Delta = \frac{1}{2} \sqrt{\frac{K-1}{T}}$.

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 π : $R_T(\pi) = T \max_k \mu_k - \sum_{t=1}^T \mu_{a_t}$, where the decisions a_t depend on the algorithm π . 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 π with cumulative pseudo-regret $R_T(\pi)$, we can extend it to a simple regret minimisation algorithm $\tilde{\pi}$ such that

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

Solution: $\tilde{\pi}$ follows the decisions of π up to T and then chooses a_{T+1} at random such that

$$a_{T+1} = k \text{ with probability } \frac{N_k(T)}{T}.$$

It then follows from the definitions of the different regrets.

- (b) Admit we have a multi-armed bandits algorithm π (e.g., MOSS) and a constant $c > 0$ such that for any bandit instance $(\nu_k)_{k \in [K]}$ with ν_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 ν_k supported on $[0, 1]$ bounded as

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

Solution: Just apply (a) to π .

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

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

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

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

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 $\ell_t : \Theta \rightarrow \mathbb{R}$ are convex instead of linear?

Solution: Use $g_t = \nabla \ell_t(\theta_t)$.

8. Explain how the above algorithm generalizes (without proving it in details):
 (a) the Online Gradient Descent algorithm;

Solution: Consider $R(\theta) = \frac{1}{2}\|\theta - \theta_1\|^2$ with the Euclidean norm for some $\theta_1 \in \Theta$. Then,

$$D_R(\theta, \theta_t) = \frac{1}{2}\|\theta - \theta_t\|^2$$

and the iteration becomes

$$\theta_{t+1} = \arg \min_{\theta \in \Theta} \left\{ \langle g_t, \theta \rangle + \frac{1}{2\eta} \|\theta - \theta_t\|^2 \right\}$$

which solving the KKT condition becomes $\theta_{t+1} = \text{Proj}_{\Theta}(\theta_t - \eta g_t)$.

(b) the Exponentially Weighted Average algorithm when $\Theta = \Delta_d$.

Solution: Consider $R(\theta) = KL(\theta|\theta_1)$.

9. Show that for any $\theta \in \Theta$: $\langle g_t, \theta_{t+1} - \theta \rangle \leq \frac{1}{\eta} \langle \nabla R(\theta_{t+1}) - \nabla R(\theta_t), \theta - \theta_{t+1} \rangle$.

Solution: From the optimality condition of $\theta_{t+1} = \arg \min_{\theta} \Phi_t(\theta)$ with $\Phi_t(\theta) = \langle g_t, \theta \rangle + \frac{1}{\eta} D_R(\theta, \theta_t)$, we have $\langle \Phi_t(\theta_{t+1}), \theta_{t+1} - \theta \rangle \leq 0$. Computing Φ_t yields the result.

10. Deduce that $\langle g_t, \theta_{t+1} - \theta \rangle \leq \frac{1}{\eta} \left(D_R(\theta, \theta_t) - D_R(\theta, \theta_{t+1}) - \frac{\mu}{2} \|\theta_t - \theta_{t+1}\|^2 \right)$.

Solution: From the definition of the Bregman divergence and the strong convexity of R we have

$$\begin{aligned} D_R(\theta, \theta_t) - D_R(\theta, \theta_{t+1}) &= R(\theta_{t+1}) - R(\theta_t) - \langle \nabla R(\theta_t), \theta_{t+1} - \theta_t \rangle + \langle \nabla R(\theta_{t+1}) - \nabla R(\theta_t), \theta - \theta_{t+1} \rangle \\ &\geq \frac{\mu}{2} \|\theta_t - \theta_{t+1}\|^2 + \langle \nabla R(\theta_{t+1}) - \nabla R(\theta_t), \theta - \theta_{t+1} \rangle. \end{aligned}$$

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 $\|\cdot\|$.

Solution: Cauchy-Schwarz followed by Young's inequality gives the result.

12. Conclude by providing an upper-bound on the regret assuming that $\|g_t\|_* \leq G$ for all t . Optimize over η .

Solution: Summing over t the two upper-bounds and telescoping the sum gives the regret upper bound:

$$\text{Regret}_T(\theta) \leq \frac{D_R(\theta, \theta_1)}{\eta} + \frac{\eta G^2 T}{2\mu} = G \sqrt{\frac{2D_R(\theta, \theta_1)T}{\mu}}$$

for $\eta = \frac{1}{G} \sqrt{\frac{2\mu D_R(\theta, \theta_1)}{T}}$.