SEQUENTIAL LEARNING
HOME ASSIGNMENT

This homework should be uploaded by **Friday, March 11, 2022** on the website

<p align="center">http://pierre.gaillard.me/teaching/mva.php</p>

The password to upload is `mva2022`. The penalty scale is minus two points (on the final grade over 20 points) for every day of delay. The homework can be done alone or in groups of two students. The code can be done in any langage (`python`, `R`, `matlab`, `julia`...) and should not be returned but the results and the figures must be included into the pdf report.

All questions require a proper mathematical justification or derivation (unless otherwise stated), but most questions can be answered concisely in just a few lines.

Part 1. Rock Paper Scissors

We consider the sequential version of a repeated two-player zero-sum games between a player and an adversary.

Let $L \in [-1, 1]^{M \times N}$ be a loss matrix.

At each round $t = 1, \ldots, T$
- The player chooses a distribution $p_t \in \Delta_M := \{p \in [0,1]^M, \sum_{i=1}^M p_i = 1\}$
- The adversary chooses a distribution $q_t \in \Delta_N$
- The actions of both players are sampled $i_t \sim p_t$ and $j_t \sim q_t$
- The player incurs the loss $L(i_t, j_t)$ and the adversary the loss $-L(i_t, j_t)$.

<p align="center">Setting 1: Setting of a sequential two-player zero sum game</p>

1. Recall M, N and a loss matrix $L \in [-1, 1]^{M \times N}$ that corresponds to the game "Rock paper scissors"[1].

Full information feedback We assume that both players know the matrix L in advance and can compute $L(i, j)$ for any (i, j). In order to implement the exponential weight algorithm, you need a way to sample from the exponential weight distribution. Implement the function `rand_weighted` that takes as input a probability vector $p \in \Delta_M$ and uses a single call to `rand()` to return $X \in [M]$ with $P(X = i) = p_i$. Define a function `EWA_update` that takes as input a learning rate $\eta > 0$, a vector $p_t \in \Delta_M$ and a loss vector $\ell_t \in [-1, 1]^M$ and return the updated vector $p_{t+1} \in \Delta_M$ defined for all $i \in [M]$ by

$$p_{t+1}(i) = \frac{p_t(i) \exp(-\eta \ell_t(i))}{\sum_{j=1}^M p_t(j) \exp(-\eta \ell_t(j))} \, .$$

---

[1]This is a common game where two players choose one of 3 options: (Rock, Paper, Scissors). The winner is decided according to the following: Rock crushes scissors, Paper covers Rock, Scissors cuts paper

2. *Simulation against a fixed adverary.* Consider the game "Rock paper scissors" and assume that the adversary chooses $q_t = (1/3, 1/6, 1/2)$ and samples $j_t \sim q_t$ for all rounds $t \geq 1$.

   (a) What is the loss $\ell_t(i)$ incurred by the player if he chooses action i at time t?

   (b) Simulate an instance of the game for $t = 1, \ldots, T = 100$ for $\eta = 1$. Plot the evolution of the weight vectors $p_1, p_2, \ldots, p_T$. What seems to be the best strategy against this adversary?

   (c) Plot the average loss $\bar{\ell}_t = \frac{1}{t} \sum_{s=1}^{t} \ell(i_s, j_s)$ as a function of t.

   (d) Repeat one simulation for different values of learning rates $\eta \in \{0.01, 0.05, 0.1, 0.5, 1\}$. What are the best η in practice and in theory?

3. *Simulation against an adaptive adversary.*

   (a) Consider a version of OGD with decision set $q_t = \Delta_N$ that chooses action $j_t \sim q_t$ and suffers the expected loss $\ell_t(q_t) = \sum_{j=1}^{N} q_t(j) g_t(j)$ where $g_t(j) = L(i_t, j)$. What is the update of OGD in this case? Define a function `OGD_update` that takes as input a vector q_t, the loss vector g_t, and a learning rate $\eta > 0$ and return $q_{t+1} \in \Delta_N$.

   (b) Repeat the simulation of Question 2.b) when the adversary is choosing p_t with OGD with learning parameter $\eta = 0.05$.

   (c) Plot $\frac{1}{t} \sum_{s=1}^{t} \ell(i_s, j_s)$ as a function of t. What strategy seems to be the best between OGD and EWA?

   Admitted: It is possible to show that if both players play according to a regret minimizing strategy the cumulative loss of the player converges to the value of the game

   $$V = \min_{p \in \Delta_M} \max_{q \in \Delta_q} p^\top L q.$$

   (d) Define $\bar{p}_t = \frac{1}{t} \sum_{s=1}^{t} p_s$. Plot in log log scale $\|\bar{p}_t - (1/3, 1/3, 1/3)\|_2$ as a function of $t = 1, \ldots, 10\,000$. What do you observe?

   Admitted: It is possible to show that $(\bar{p}_t, \bar{q}_t)_{t \geq 1}$ converges almost surely to a Nash equilibrium of the game. This means that if $p \times q$ is a Nash equilibrium, none of the players should change is strategy if the other player does not change hers.

Bandit feedback   Now, we assume that the players do not know the game in advance but only observe the performance $L(i_t, j_t)$ (that we assume here to be in $[0, 1]$) of the actions played at time t. They need to learn the game and adapt to the adversary step by step as more feedback is observed.

4. Explain the main differences between EXP3 and EWA and implement `EXP3`.

5. Repeat Questions 2.b) to 2.d) with `EXP3` instead of `EWA`.

Optional extentions   The following questions are optional.

6. Run a simulation where `EXP3` is competing with an `UCB` algorithm. Who wins between `UCB` and `EXP3`? Repeat question 3.e) with `EXP3.IX`

7. Try different games (not necessarily zero-sum games). In particular, how these algorithms behave for the prisoner's dilemma (see wikipedia)? The prisoner's dilemma is a two-player games that shows why two completely rational individuals might not cooperate, even if it appears that it is in their best interests to do so. The loss matrices are:

$$L^{(player)} = \begin{pmatrix} 1 & 3 \\ 0 & 2 \end{pmatrix} \quad \text{and} \quad L^{(adversary)} = \begin{pmatrix} 1 & 0 \\ 3 & 2 \end{pmatrix}.$$

Part 2. Bernoulli Bandits

We consider a stochastic bandit setting in which the arm rewards have Bernoulli distributions. A random variable X is said to have Bernoulli distribution with parameter p, which we denote by $\mathcal{B}(p)$, if it takes value 0 with probability $1 - p$ and value 1 with probability p. The set $\{1, \ldots, K\}$ is denoted by $[K]$.

Each arm $k \in [K]$ has a reward distribution $\mathcal{B}(p_k)$.

At each round $t = 1, \ldots, T$
  – The player chooses an arm $k_t \in [K]$,
  – The player observes a reward $X_t^{k_t} \sim \mathcal{B}(p_{k_t})$, independent of all other rewards.

Setting 2: Bernoulli bandit

Notations:
  – In this part, the term "regret" refers to the quantity $R_T = \max_{k \in [K]} T p_k - \sum_{t=1}^{T} p_{k_t}$.
  – N_t^k denotes the number of pulls of arm k immediately after time t, i.e. $N_t^k = \sum_{s=1}^{t} \mathbb{I}\{k_s = k\}$.
  – $\hat{\mu}_t^k$ denotes the empirical mean of arm k after time t: $\hat{\mu}_t^k = \frac{1}{N_t^k} \sum_{s=1}^{t} X_s^{k_s} \mathbb{I}\{k_s = k\}$.

A bit of context: why Bernoulli bandits matter. Many applications have binary outcomes, in which the reward then follows a Bernoulli distribution. A prominent example is online advertising, in which a seller shows advertisements to visitors of a website, and a usual goal is to maximize the probability that the visitor clicks on the ad. In its most basic form, this is exactly the bandit interaction described above: the seller (player) chooses an ad (arm) which is displayed to the visitor, and then the seller observes whether there is a click or not (reward). More elaborate models of that interaction take into account prior information that the seller has about the visitor, turning it into a *contextual* bandit, or get rid of the independence assumption, etc.

1. *Follow the leader.* All experiments in this question will be done for $K = 2$, $p = (0.5, 0.6)$.

   (a) Prove that the expected regret of the Follow-the-leader algorithm (FTL) verifies $\mathbb{E} R_T \geq \alpha T$, for some $\alpha > 0$. Recall that FTL pulls at each time the arm with highest empirical mean: $k_{t+1} = \arg\max_{k \in [K]} \hat{\mu}_t^k$.

   (b) Implement FTL.

   (c) For time $T = 100$, plot a histogram of the regret R_T of FTL over 1000 repetitions of the experiment. Explain the figure.

   (d) Plot the mean regret of FTL over 1000 repetitions, as a function of $t \in \{1, \ldots, 1000\}$. Is FTL a good algorithm for stochastic bandits?

2. *UCB.* A random variable is said to be σ^2-sub-Gaussian if for all $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq e^{\frac{1}{2}\sigma^2 \lambda^2}$. The UCB($\sigma^2$) algorithm pulls arm $k_{t+1} = \arg\min_{k \in [K]} \hat{\mu}_t^k + \sqrt{\frac{2\sigma^2 \log(t)}{N_t^k}}$. It is designed to have low regret on σ^2-sub-Gaussian random variables.

   (a) Compute the cumulant generating function, defined for $\lambda \in \mathbb{R}$ by $\phi_X(\lambda) = \log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}]$, for a Bernoulli random variable with parameter p.

   (b) Prove that if a random variable X (not necessarily Bernoulli) verifies $\phi_X''(\lambda) \leq \sigma^2$ for all $\lambda \in \mathbb{R}$, then the random variable is σ^2-sub-Gaussian. Remark: this is not an equivalence (you are not required to prove this).

   (c) Using question 2.b, find σ^2 such that a random variable with distribution $\mathcal{B}(p)$ is σ^2-sub-Gaussian.

(d) Prove that a random variable X supported on $[0,1]$ with mean $p \in [0,1]$ verifies $\phi_X(\lambda) \leq \phi_Y(\lambda)$ for all $\lambda \in \mathbb{R}$, where Y has a $\mathcal{B}(p)$ distribution. Hint: prove that for all $x \in [0,1]$, for all $\lambda \in \mathbb{R}$, $e^{\lambda x} \leq 1 - x + xe^{\lambda}$.

(e) Prove that all random variables supported on $[0,1]$ are sub-Gaussian.

(f) Implement the $\mathtt{UCB}(\sigma^2)$ algorithm.

(g) Plot the mean regret of $\mathtt{UCB}(1/4)$ as a function of time up to $T = 1000$ for $K = 2$, $p = (0.5, 0.6)$, over 1000 repetitions. Compare with the result of question 1.d.

(h) For $K = 2$, $p = (0.6, 0.5)$, $T = 1000$, plot the mean regret of $\mathtt{UCB}(\sigma^2)$ over 1000 repetitions as a function of σ^2, for $\sigma^2 \in \{0, 1/32, 1/16, 1/4, 1\}$. Do it again for $p = (0.85, 0.95)$ and compare the results: does the optimal parameter change? How does it compare to the theoretic parameter?

The results of the question 2.c on Bernoulli distributions can be improved: it is possible to prove that a random variable with distribution $\mathcal{B}(p)$ is σ^2-sub-Gaussian with parameter $\sigma^2(p) = 0$ if $p \in \{0, 1\}$, $\sigma^2(p) = 1/4$ if $p = 1/2$ and $\sigma^2(p) = \frac{1}{2}\frac{p-(1-p)}{\log p - \log(1-p)}$ for $p \in (0,1) \setminus \{1/4\}$.

3. On the same figure, plot the variance of $\mathcal{B}(p)$ and the sub-Gaussian constant $\sigma^2(p)$ described above as a function of $p \in [0,1]$.

4. **(optional)** Prove that a σ^2-sub-Gaussian random variable has variance bounded by σ^2.

Adaptation to the variance. The algorithm $\mathtt{UCB}(\sigma^2)$ uses only the empirical mean of the arms to choose the next arm, except for a parameter σ^2 which has to be chosen such that all arms are σ^2-sub-Gaussian. In particular, all variance information about the distributions is lost. Intuitively an arm with lower variance should require fewer samples in order to know its mean with enough precision.

5. UCB-V. For bounded rewards belonging to $[0, b]$, the algorithm $\mathtt{UCB\text{-}V}(b, \xi, c)$ (V for variance) computes the empirical variance of the arms, $\hat{v}_t^k = \frac{1}{N_t^k} \sum_{s=1}^t \mathbb{I}\{k_s = k\}(X_s^{k_s} - \hat{\mu}_t^k)^2$ and pulls the arm $k_{t+1} = \arg\max_{k \in [K]} \hat{\mu}_t^k + \sqrt{\frac{2\hat{v}_t^k \xi \log t}{N_t^k}} + \frac{3bc\xi}{N_t^k}$. For theoretical regret bounds to hold, ξ should be taken slightly larger than 1 and c larger than a function of ξ, which increases as $\xi \to 1$. **All experiments in this question will be done for $b = 1$, $\xi = 1.1$ and $c = 1$.**

(a) Prove that $N_t^k \hat{v}_t^k = \sum_{s=1}^t \mathbb{I}\{k_s = k\}(X_s^{k_s})^2 - \frac{1}{N_t^k}(\sum_{s=1}^t \mathbb{I}\{k_s = k\}X_s^{k_s})^2$.

(b) Prove that $N_{t+1}^{k_{t+1}} \hat{v}_{t+1}^{k_{t+1}} = N_t^{k_{t+1}} \hat{v}_t^{k_{t+1}} + (X_{t+1}^{k_{t+1}} - \hat{\mu}_t^{k_{t+1}})(X_{t+1}^{k_{t+1}} - \hat{\mu}_{t+1}^{k_{t+1}})$. What is the practical advantage of that formulation?

(c) Implement $\mathtt{UCB\text{-}V}$.

(d) On the same figure, plot the mean regret of $\mathtt{UCB\text{-}V}$ and $\mathtt{UCB}(1/4)$ as a function of time up to $T = 1000$ for $K = 2$, $p = (0.5, 0.6)$, over 1000 repetitions.

(e) Same question for $p = (0.1, 0.2)$ and $p = (0, 0.1)$. Compare to the results of 5.d. When does $\mathtt{UCB\text{-}V}$ improve over $\mathtt{UCB}$?

Algorithms for parametric distributions. $\mathtt{UCB}$ uses only an estimate of the mean, while $\mathtt{UCB\text{-}V}$ uses estimates of the mean and variance. However, Bernoulli distributions have many properties beyond their mean and variance, and these properties are not used by $\mathtt{UCB\text{-}V}$. We can design algorithms that use fully the knowledge that the distribution of the arms are Bernoulli $\mathcal{B}(p)$, with the only unknown being the parameter p. The algorithm $\mathtt{kl\text{-}UCB}$ is designed precisely to take advantage of the knowledge that distributions belong to a so-called one-parameter exponential family, and that algorithm can use fully the Bernoulli assumption. See [2, 1].

6. **(optional)** Implement the `kl-UCB` algorithm for Bernoulli bandits (see [2, 1]), and compare with `UCB` and `UCB-V` on various Bernoulli bandit problems, for examples the settings of questions 5.d and 5.e.

References

[1] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

[2] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376. JMLR Workshop and Conference Proceedings, 2011.

[3] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.