

# SEQUENTIAL LEARNING

## HOME ASSIGNMENT

This homework should be uploaded by **Friday, March 21, 2025** on the website

<https://sequential-learning.github.io/>

The password to upload is **mva2025**. The penalty scale is minus two points (on the final grade over 20 points) for every day of delay. If the upload does not succeed (for some reason), send it by email to pierre.gaillard@inria.fr and remy.degenne@inria.fr but only after you have tried the upload. It can be done alone or in groups of two students. The report can be written in English or in French.

The homework should consist of a **single jupyter notebook for Part 1**, together with a **single pdf report for Part 2** (the code should not be returned for this part).

All questions require a proper mathematical justification or derivation (unless otherwise stated), but most questions can be answered concisely in just a few lines. No question should require lengthy or tedious derivations or calculations.

## Part 1. Portfolio selection

Complete the notebook

[https://sequential-learning.github.io/docs/dm/notebook\\_portfolio.ipynb](https://sequential-learning.github.io/docs/dm/notebook_portfolio.ipynb)

Data is available at

[https://sequential-learning.github.io/docs/dm/data\\_portfolio.zip](https://sequential-learning.github.io/docs/dm/data_portfolio.zip)

## Part 2. Stochastic Best Arm Identification

In the best arm identification setting, an algorithm interacts with the environment in the standard bandit way: at each time, it selects an arm, then observes the corresponding reward. The goal of the algorithm is to find the arm with highest mean, as quickly as possible.

Each arm  $k \in [K]$  has a reward distribution  $\nu_k$  with mean  $\mu_k$ . There is a unique best arm  $k^* = \arg \max_k \mu_k$ .  
 At each round  $t = 1, \dots, \tau$

- The player chooses an arm  $k_t \in [K]$ ,
- The player observes a reward  $X_t^{k_t} \sim \nu_{k_t}$ , independent of all other rewards.

At the stopping time  $\tau$ , the algorithm recommends an arm  $\hat{k} \in [K]$

Setting 1: Best arm identification

A good algorithm should make few mistakes and stop quickly. With the notations of Figure 1, the probability of mistake of the algorithm on problem  $\nu$  is  $\mathbb{P}_\nu(\hat{k} \neq k^*)$ . The possibly random time at which it stops and returns an answer is  $\tau$ .

Notations:  $N_t^k = \sum_{s=1}^t \mathbb{I}\{k_s = k\}$  is the number of times arm  $k$  was sampled up to time  $t$ .  $\hat{\mu}_{t,k} = \frac{1}{N_t^k} \sum_{s=1}^t \mathbb{I}\{k_s = k\} X_s^{k_s}$  is the empirical mean of arm  $k$ .

**Fixed Budget** In *fixed budget* best arm identification, we are given a time  $T$  and set  $\tau = T$ . That is, the algorithm can sample a total number of  $T$  arms ( $T$  known in advance), then must stop and return an answer. We are interested in algorithms with low probability of mistake.

1. *Uniform sampling.* The uniform sampling algorithm pulls all arms  $T/K$  times (we suppose that  $T$  is a multiple of  $K$ ).
  - (a) Prove that the probability of error of uniform sampling is at most  $2 \sum_{k=2}^K \exp(-\frac{T}{K} \frac{\Delta_k^2}{8})$ , where  $\Delta_k = \mu_{k^*} - \mu_k$ .

---

**Algorithm 1:** Successive rejects

---

Input: budget  $T$ , number of arms  $K$ .

Let  $A_1 = \{1, \dots, K\}$ ,  $\overline{\log}(K) = \frac{1}{2} + \sum_{k=2}^K \frac{1}{k}$ ,  $n_0 = 0$  and for  $j \in \{1, \dots, K-1\}$ ,

$$n_j = \left\lceil \frac{1}{\overline{\log}(K)} \frac{T-K}{K+1-j} \right\rceil.$$

**for each phase**  $j = 1, 2, \dots, K-1$  **do**

    For each  $i \in A_j$ , select arm  $i$  during  $n_j - n_{j-1}$  rounds.

    Let  $A_{j+1} = A_j \setminus \arg \min_{k \in A_j} \hat{X}_{k,n_j}$ , where  $\hat{X}_{k,n_j}$  is the empirical mean of arm  $k$  after  $n_j$  observations.  
 (we only remove one element from  $A_j$ ; if there is a tie, select randomly the arm to dismiss among the worst arms).

**end**

---

2. The *Successive Rejects* algorithm is described in Algorithm 1. In each phase, it samples all arms uniformly, and at the end of a phase it discards the worse arm.
  - (a) Give a bound on the probability that the best arm is discarded at the end of the first phase.
  - (b) Let  $B = \mathbb{I}\{\hat{k} \neq k^*\}$  be the Bernoulli random variable with value 1 if the algorithm makes a mistake and 0 otherwise. Its expectation on the bandit problem  $\nu$  is  $\mathbb{P}_\nu(\hat{k} \neq k^*)$ . Suppose that we run  $n$  parallel experiments, and that in each experiment  $i \in [n]$  we run successive rejects on the same bandit  $\nu$  and compute the corresponding  $B_i = \mathbb{I}\{\hat{k} \neq k^*\}$ . Give a confidence interval for  $\mathbb{P}_\nu(\hat{k} \neq k^*)$ .
  - (c) Implement successive rejects and uniform sampling. Plot the probability of error of both algorithms for  $K = 20$  Bernoulli arms with  $\mu_1 = 0.6$  and  $\mu_k = 0.5$  for  $k \geq 2$ , for  $T \in \{100, 500, 2000\}$ . Plot confidence intervals and make sure to use enough experiments to get intervals with smaller width than the error probability.

**Fixed Confidence** In *fixed confidence* best arm identification, we consider only  $\delta$ -correct algorithms, which satisfy  $\mathbb{P}_\nu(\hat{k} \neq k^*) \leq \delta$  for all tuples of distributions in our model. We are then looking for such algorithms with minimal expected stopping time  $\mathbb{E}_\nu[\tau]$ .

All arm distributions in this section will be Gaussian with variance 1. All experiments will use 10 such arms, with means  $(0.5, 0.6, 0.5, 0.4, \dots, 0.4)$ . All experiments will use  $\delta = 0.01$ .

*Stopping rule.* Let  $\hat{*}_t = \arg \max_{k \in [K]} \hat{\mu}_{t,k}$ . All algorithms will use the same stopping rule: stop when

$$\inf_{k \in [K] \setminus \{\hat{*}_t\}} \frac{1}{2} \frac{(\hat{\mu}_{t,\hat{*}_t} - \hat{\mu}_{t,k})^2}{\frac{1}{N_t^{\hat{*}_t}} + \frac{1}{N_t^k}} > \log \frac{1}{\delta} + 3 \log(1 + \log t).$$

$\tau$  is the first time such that this condition is satisfied. This is a heuristic approximation of the GLRT stopping rule seen in the course (the quantity on the left is equal to  $\inf_{\lambda: *_{\lambda} \neq \hat{*}_t} \sum_{k=1}^K N_t^k \frac{1}{2} (\hat{\mu}_{t,k} - \lambda_k)^2$ ).

1. In order to get a sub-linear regret, a regret minimization algorithms has to sample mostly the best arm. It could thus be transformed into a best arm identification algorithm.

(a) Implement the UCB algorithm for regret minimization, which samples the arm

$k_t = \arg \max \hat{\mu}_{t-1,k} + \sqrt{\frac{2 \log t}{N_{t-1}^k}}$ . Plot the mean over 100 experiments of the regret of UCB on the Gaussian bandit problem described above, for  $t$  from 1 to 10000.

(b) Implement a best arm identification algorithm that samples like UCB, stops according to the rule presented above, and returns the most played arm. Implement another algorithm that samples all arms uniformly, stops according to the rule presented above, and returns the arm with highest empirical mean. On a box plot, compare the stopping times of both algorithms.

2. Regret minimization algorithms don't explore enough to be good identification methods. We can modify them to explore more: this is the idea of *Top-Two* algorithms. A Top-Two algorithm computes at each time  $t$  a *leader*  $B_t \in [K]$  and a *challenger*  $C_t \in [K] \setminus \{B_t\}$ , and then with probability  $\beta \in (0, 1)$  it samples the leader ( $k_t = B_t$ ), and it samples the challenger with probability  $1 - \beta$ . It stops according to the rule presented above, and recommends the arm with highest empirical mean. We will use  $\beta = 1/2$ .

(a) Implement the TTUCB algorithm: its leader is the UCB arm  $B_t = \arg \max \hat{\mu}_{t-1,k} + \sqrt{\frac{2 \log t}{N_{t-1}^k}}$  and its challenger is the arm for which it is hardest to say that it is worse,  $C_t = \arg \min_{k \neq B_t} \frac{1}{2} \frac{(\hat{\mu}_{t,B_t} - \hat{\mu}_{t,k})^2}{\frac{1}{N_t^{B_t}} + \frac{1}{N_t^k}}$ .

(b) We added an exploration mechanism, the challenger, and it is possible that using UCB for the leader is not necessary anymore: we could simply select the empirical best arm, without the exploration bonus of UCB.

Implement the so-called EB-TC algorithm which uses  $B_t = \arg \max \hat{\mu}_{t-1,k}$  and  $C_t = \arg \min_{k \neq B_t} \frac{1}{2} \frac{(\hat{\mu}_{t,B_t} - \hat{\mu}_{t,k})^2}{\frac{1}{N_t^{B_t}} + \frac{1}{N_t^k}}$ .

(c) On a box plot, compare the stopping times of all 4 algorithms (UCB, uniform, TTUCB and EB-TC) and comment the results.

3. (Bonus question) Thompson Sampling is another regret minimization algorithm. Implement a Top Two algorithm which uses Thompson Sampling. Give a pseudo-code of your algorithm. Compare that algorithm to the others on various bandit problems.