Homework Sequential Learning

Clémence GRISLAIN - clemence.grislain@eleves.enpc.fr

March 10, 2023

1 Part1 - Bandit convex optimization

1. It is **oblivious** as the loss function at time t does not depend on the previous actions taken by the learner at previous times $\{\theta_i\}_{i<t}$.

2.
$$R_T = \sum_{t=1}^{T} l_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=1}^{T} l_t(\theta)$$

3. (a)
$$\nabla \hat{l}_t(\hat{\theta}_t) = \nabla \mathbb{E}_v(l_t(\hat{\theta}_t + \delta v)$$
$$= \nabla \int_{-1}^{1} l_t(\hat{\theta} + \delta v) p(v) dv$$

where $p(v)$ is the density, which, in case of uniform distribution on [-1,1] is $p(v) = \frac{1}{2}$

$$= \frac{1}{2} \int_{-1}^{1} \nabla l_t(\hat{\theta}_t + \delta v) dv \quad \text{as the integral is finite}$$
$$= \frac{1}{2} \left(\frac{1}{\delta} l_t(\hat{\theta}_t + \delta) - \frac{1}{\delta} l_t(\hat{\theta}_t - \delta) \right)$$

For d=1, we have $u_t \in \mathbb{S}_1 = \{-1, 1\}$ and as it is uniformly distributed $\mathbb{P}(u_t = 1) = \mathbb{P}(u_t = -1) = \frac{1}{2}$

$$= \mathbb{P}(u_t = 1) \frac{1}{\delta} l_t(\hat{\theta}_t + \delta u_t) u_t \Big|_{u_t = +1} + \mathbb{P}(u_t = -1) \frac{1}{\delta} l_t(\hat{\theta}_t + \delta u_t) u_t \Big|_{u_t = -1}$$
$$= \mathbb{E} \left(\frac{1}{\delta} l_t(\hat{\theta}_t + \delta u_t) u_t \right) \quad \square$$

(b) $\forall \theta \in \Theta$,
$$|\hat{l}_t(\theta) - l_t(\theta)| = |\mathbb{E}_v(l_t(\theta + \delta v) - l_t(\theta)|$$

by convexity of the function l_t

$$|\mathbb{E}_v(l_t(\theta + \delta v) - l_t(\theta))| \leq |\nabla l_t(\theta) \mathbb{E}_v(\delta v)|$$
$$\implies |\hat{l}_t(\theta) - l_t(\theta)| \leq |l_t(\theta)| \times |\mathbb{E}_v(\delta v)| \leq G \times \delta \quad \square$$

as $v \in \mathbb{S}_1 \implies |v| \leq 1$

4. (a) If OGD was applied on the losses h_t we would have
$$\theta_{t+1} = Proj_{\Theta_\delta} \left(\hat{\theta}_t - \frac{d\eta}{\delta} l_t(\theta_t) u_t \right)$$
$$\theta_{t+1} = Proj_{\Theta_\delta} \left(\hat{\theta}_t - \nabla h_t(\hat{\theta}_t) \right)$$

(b) The function $h_t(.)$ is convex, thus it verifies $\forall \theta_\delta^* \in \Theta$

$$\sum_{t=1}^{T} h_t(\hat{\theta}_t) - h_t(\theta_\delta^*) \leq \sum_{t=1}^{T} \nabla h_t(\hat{\theta}_t).(\hat{\theta}_t - \theta_\delta^*)$$

$$\leq \sum_{t=1}^{T} (\nabla \hat{l}_t(\hat{\theta}_t) + \xi_t).(\hat{\theta}_t - \theta_\delta^*)$$

$$\leq \sum_{t=1}^{T} (\frac{d}{\delta} l_t(\theta_t) u_t).(\hat{\theta}_t - \theta_\delta^*)$$

let's define $z_{t+1} = \hat{\theta}_t - \frac{d\eta}{\delta} l_t(\theta_t) u_t$

$$\leq \sum_{t=1}^{T} \frac{1}{\eta} (\hat{\theta}_t - z_{t+1}).(\hat{\theta}_t - \theta_\delta^*)$$

using the inequality $||x-y||^2 = ||y||^2 + ||x||^2 - 2 < x, y >$ we have

$$\leq \frac{1}{2\eta} \sum_{t=1}^{T} \underbrace{||\hat{\theta}_t - z_{t+1}||^2}_{=\eta^2 ||\frac{d}{\delta} l_t(\theta_t) u_t||^2 \leq \frac{d^2\eta^2}{\delta^2}} + \underbrace{||\hat{\theta}_t - \theta_\delta^*||^2}_{\leq ||z_t - \theta_\delta^*||^2 \text{ as } Proj_{\Theta_\delta} \text{ convex}} - ||\theta_\delta^* - z_{t+1}||^2$$

$$\leq \frac{d^2 T}{2\delta^2} \eta + \frac{1}{2\eta} \sum_{t=1}^{T} ||z_t - \theta_\delta^*||^2 - ||\theta_\delta^* - z_{t+1}||^2$$

$$\leq \frac{d^2 T}{2\delta^2} \eta + \frac{1}{2\eta} ||\theta_\delta^* - \theta_1||^2$$

$$\leq \frac{d^2 T}{2\delta^2} \eta + \frac{D^2}{2\eta} \quad \square$$

(c) $\forall t, \forall \theta_\delta^* \in \Theta$

$$\mathbb{E}(\hat{l}_t(\hat{\theta}_t)) - \hat{l}_t(\theta_\delta^*) = \mathbb{E}(\hat{l}_t(\hat{\theta}_t) - \hat{l}_t(\theta_\delta^*))$$

$$= \mathbb{E}(h_t(\hat{\theta}_t) - h_t(\theta_\delta^*)) + \mathbb{E}(< \xi_t, \hat{\theta}_t - \theta_\delta^* >)$$

$$= \mathbb{E}(h_t(\hat{\theta}_t) - \hat{h}_t(\theta_\delta^*)) + \mathbb{E}(E_{u_t}(< \xi_t, \hat{\theta}_t - \theta_\delta^* >) | u_t)$$

$$= \mathbb{E}(h_t(\hat{\theta}_t) - h_t(\theta_\delta^*)) + \mathbb{E}(< \underbrace{E_{u_t}(\xi_t)}_{=0}, \hat{\theta}_t - \theta_\delta^* > | u_t)$$

as $\hat{\theta}_t - \theta_\delta^*$ are independent from u_t.

$$= \mathbb{E}(h_t(\hat{\theta}_t) - h_t(\theta_\delta^*))$$

$$\Longrightarrow \sum_{t=1}^{T} \mathbb{E}(\hat{l}_t(\hat{\theta}_t)) - \hat{l}_t(\theta_\delta^*) \leq \frac{d^2 T}{2\delta^2} \eta + \frac{D^2}{2\eta} \quad \square$$

5.

$$\mathbb{E}(R_t) = \mathbb{E}\left(\sum_{t=1}^{T} l_t(\theta_t) - \min_{\theta \in \Theta} \sum_{t=1}^{T} l_t(\theta) \right)$$

Let's $\theta^* = \arg\min_{\theta \in \Theta} \sum_{t=1}^{T} l_t(\theta)$

$$\mathbb{E}(R_t) = \mathbb{E}\left(\sum_{t=1}^{T} l_t(\theta_t) - \sum_{t=1}^{T} l_t(\theta^*) \right)$$

$$= \sum_{t=1}^{T} \mathbb{E}\left(l_t(\theta_t) - l_t(\theta^*)\right)$$

$$= \sum_{t=1}^{T} \mathbb{E}\left(l_t(\theta_t) - \hat{l}_t(\hat{\theta}_t) + \hat{l}_t(\hat{\theta}_t) - \hat{l}_t(\theta^*) + \hat{l}_t(\theta^*) - l_t(\theta^*)\right)$$

$$= \underbrace{\sum_{t=1}^{T} \mathbb{E}\left(\hat{l}_t(\hat{\theta}_t) - \hat{l}_t(\theta^*)\right)}_{\leq \frac{Td^2}{2\delta^2}\eta + \frac{D^2}{2\eta}} + \sum_{t=1}^{T} \mathbb{E}\left(l_t(\theta_t) - \hat{l}_t(\hat{\theta}_t) + \hat{l}_t(\theta^*) - l_t(\theta^*)\right)$$

$$\leq \frac{Td^2}{2\delta^2}\eta + \frac{D^2}{2\eta} + \sum_{t=1}^{T} \mathbb{E}\left(l_t(\theta_t) - l_t(\hat{\theta}_t) + l_t(\hat{\theta}_t) - \hat{l}_t(\hat{\theta}_t) + \hat{l}_t(\theta^*) - l_t(\theta^*)\right)$$

$$\leq \frac{Td^2}{2\delta^2}\eta + \frac{D^2}{2\eta} + \sum_{t=1}^{T} \mathbb{E}\left(|l_t(\theta_t) - l_t(\hat{\theta}_t)| + \underbrace{|l_t(\hat{\theta}_t) - \hat{l}_t(\hat{\theta}_t)|}_{\leq \delta G} + \underbrace{|\hat{l}_t(\theta^*) - l_t(\theta^*)|}_{\leq \delta G}\right)$$

and we have $|l_t(\theta_t) - l_t(\hat{\theta}_t)| \leq |\nabla l_t(\hat{\theta}_t)| \times |\delta u_t| \leq \delta G$

$$\implies \mathbb{E}(R_t) \leq \frac{Td^2}{2\delta^2}\eta + \frac{D^2}{2\eta} + 3\delta TG \quad \square$$

6. By optimizing the parameters η and δ, we obtain a regret in $T^{3/4}$. More preciely, by setting $\delta = T^{-1/4} \times \sqrt{\frac{dD}{3G}}$ and $\eta = T^{-3/4} \times \sqrt{\frac{D}{3Gd}}D$ we obtain $\mathbb{E}(R_t) \leq \sqrt{12 \times DdG} \times T^{3/4}$.

7. (a) With d=2 and $\Theta = \{||\theta|| \leq 1\}$ we obtain the following curve for the cumulative regrets R_t for $t \in [1, T]$ with η and δ computed for $T = 1000$ with the formula above: We draw the

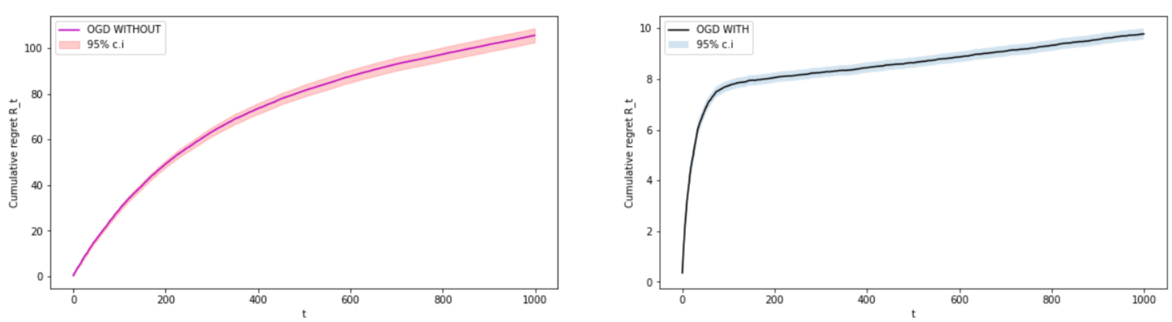

Figure 1: Cumulative regrets R_t of OGD and OGD without gradient

95% confidence interval over the 100 runs.

(b) We observe that both algorithms' cumulative reward R_T depend linearly of the parameter d. As a consequence, the higher the dimension of the search space Θ the higher the cumulative regrets R_T.

We draw the 95% confidence interval over the 100 runs.

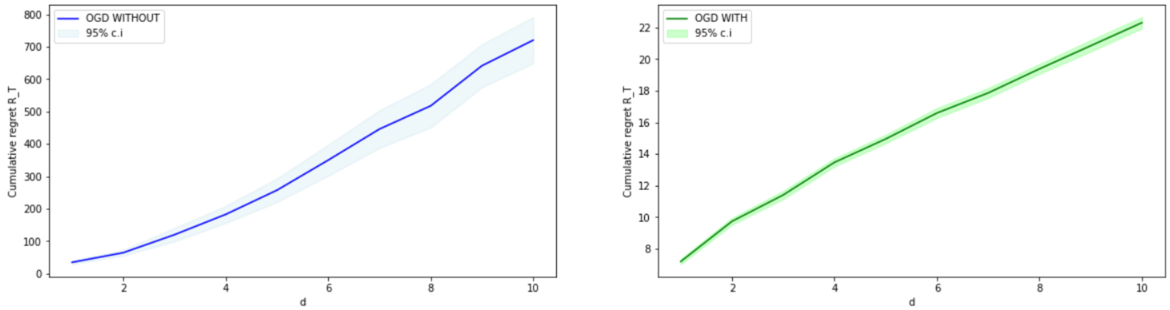

Figure 2: Cumulative regrets R_T for T=1000 of OGD and OGD without gradient with $d \in [1, 10]$

2 Part 2 - Stochastic Best Arm Identification

Without losing any generality, we suppose that the best arm is the first one $k^* = 1$.

1. (a) We recall that if $\{X_i\}_{\mathbb{N}}$ are iid and σ^2-sub-Gaussian, we have that

$$\mathbb{P}\left(\bar{X}_t - \mathbb{E}(X_1) \geq \epsilon\right) \leq e^{-\frac{1}{2}\frac{\epsilon^2}{\sigma^2}t}$$

We suppose that the the reward X_t^k are iid and 1-sub-Gaussian. Thus we have that the probability of error is

$$\mathbb{P}\left(\bigcup_{k=2}^{K}(\hat{\mu}_{T,k} > \hat{\mu}_{T,k^*})\right)$$

$$\leq \sum_{k=2}^{K}\mathbb{P}\left(\hat{\mu}_{T,k} > \hat{\mu}_{T,k^*}\right)$$

$$\leq \sum_{k=2}^{K}\mathbb{P}\left(\left(\hat{\mu}_{T,k^*} \leq \mu_{k^*} - \frac{\Delta_k}{2}\right) \cup \left(\hat{\mu}_{T,k} \geq \mu_k + \frac{\Delta_k}{2}\right)\right)$$

$$\leq \sum_{k=2}^{K}\mathbb{P}\left(\hat{\mu}_{T,k^*} - \mu_{k^*} \leq -\frac{\Delta_k}{2}\right) + \mathbb{P}\left(\hat{\mu}_{T,k} - \mu_k \geq \frac{\Delta_k}{2}\right)$$

using that $\forall k$ the X_t^k are 1-sub-Gaussian

$$\leq 2\sum_{k=2}^{K}e^{-\frac{1}{2}(\frac{\Delta_k}{2})^2\frac{T}{K}} = 2\sum_{k=2}^{K}e^{-\frac{\Delta_k^2}{8}\frac{T}{K}} \quad \square$$

2. (a) The probability that the best arm is discarded at the end of the first phase is

$$\mathbb{P}(\underset{k \in A_1 = \{1..K\}}{argmin} \hat{X}_{k,n_1} = k^*) = \mathbb{P}\left(\bigcap_{k=2}^{K}(\hat{X}_{k,n_1} > \hat{X}_{k^*,n_1})\right)$$

these events are independent as the reward $X_{k_t}^t$ is independent of all other rewards.

$$= \prod_{k=2}^{K}\mathbb{P}\left(\hat{X}_{k,n_1} > \hat{X}_{k^*,n_1}\right)$$

$$\leq \prod_{k=2}^{K} 2 \times e^{-\frac{\Delta_k^2}{8}n_1}$$

By analogy with the previous question

$$\leq 2^{K-1} \times e^{-\frac{n_1}{8}\sum_{k=2}^{K}\Delta_k^2} \quad \square$$

(b) From the Central limit theorem, we have that a 95% confidence interval is

$$\bar{B}_n \pm 1.96 \times \sqrt{\frac{\bar{B}_n(1-\bar{B}_n)}{n}}$$

as $\bar{B}_n$ is the empiric mean and $\bar{B}_n(1-\bar{B}_n)$ the empiric variance of the Bernoulli law.

(c) For K = 20 Bernoulli arms with $\mu_1 = 0.5$ and $\mu_k = 0.4$ for $k \geq 2$, for $T \geq \{100, 500, 2000\}$ we obtain the following probability of error with 95% confidence interval based on the above formula:

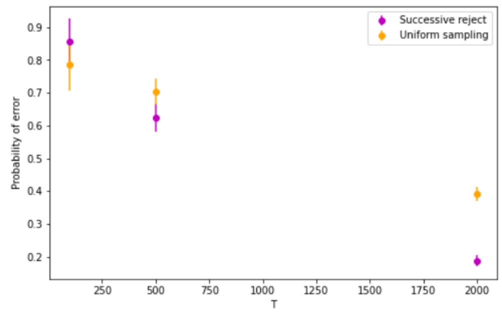

Figure 3: Probability of error of succesive reject and uniform sanpling algorithms for $T \in \{100, 500, 2000\}$

We observe that the uniform sampling is better than the successive reject for small value of T, as we can see on the experiment with $T = 100$. Yet, as soon as we increase T, the successive reject algorithm outperforms the uniform sampling. We see that for $T = 2000$, the successive reject algorithm probability of error gets smaller than 0.2 and has a very small 95% confidence interval. Whereas, even though the uniform sampling probability of error also achieves small 95% confidence interval, the probability of error is ≈ 0.4.

Fixed Confidence

1. (a) We implemented a stochastic bandit with all arm distributions are Gaussian with variance 1, with $K = 10$ such arms, with means (0.5, 0.4, 0.4, 0.3, . . . , 0.3) and use $\delta = 0.01$ for probability of error upper bound. To evaluate the UCB algorithm in this stochastic bandit setting, we use pseudo cumulative regrets

$$\bar{R}_t = t \times \mu_{k^*} - \sum_{s=1}^{t} \mu_{k_t}$$

Ans obtain the following curve of $\bar{R}_t$ ove 95% confidence interval over the 100 runs: We can

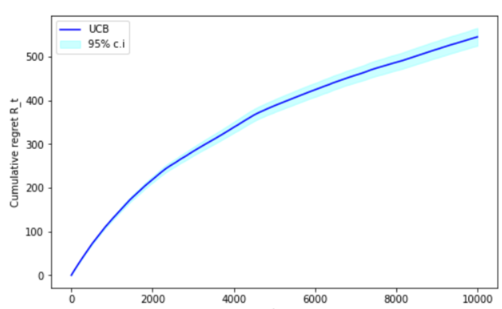

Figure 4: Pseudo cumulative regret R_t of UCB algorithm for $t \in [1, T]$, with T=10000

observe that, as seen in the course, the pseudo cumulative regrets $\bar{R}_t$ is in $\sqrt{t}$.

(b) We implemented the UCB and uniform sampling algorithm with the GLRT stopping criterion and observe over 50 runs the stopping time of each algorithm. On the Figure 2 we see that the UCB algorithm has much higher stopping time than the uniform sampling algorithm whose median stopping time is superior to 10^6 steps whereas Uniform sampling algorithm has a median stopping time of 60000.

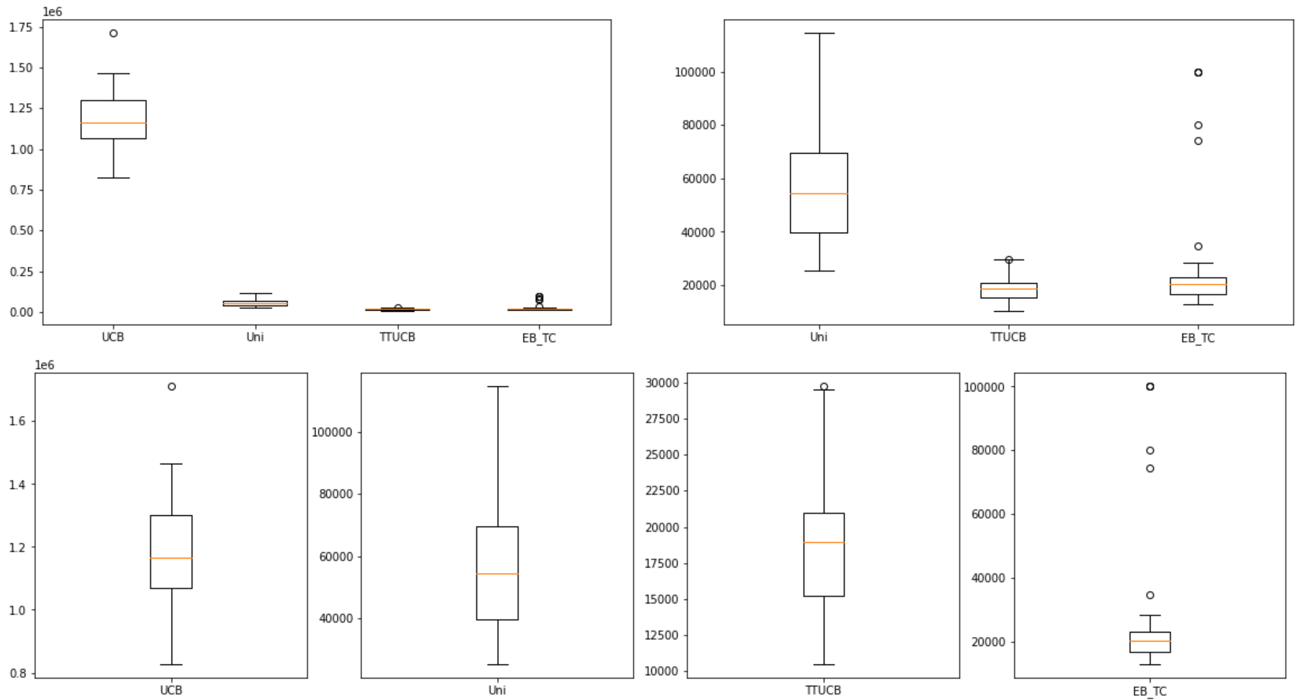

Figure 5: Stopping times box plot over 50 runs of UCB, Uniform sampling, TTUCB and EB_TC algorithms

2. In addition with the UCB and Uniform sampling algorithms, we implemented algorithms based on "Challenger" in order to reach faster the GLRT stopping criterion. As a result, both algorithms based on this method, TTUCB and EB_TC, achieve lower stopping time than the uniform sampling and UCB, with median stopping times $\approx$ 20000 steps. Yet, TTUCB seems better than EB_TC due to the exploration term when choosing B_t. In fact, it happens that EB_TC gets stuck into infinite loops (in the experiments, we stopped the algorithm at $t > 100000$, hence the outliers). If at the beginning the best arm $k^* = 1$ has bad luck and a small empirical mean and both $k = 2$ and $k = 3$ have good empirical means such that the $B_t = 2$ or $B_t = 3$, EB_TC will no more focus on the best arm $k^* = 1$ but only on B_t and the challenger $C_t = 3$ or $C_t = 2$. At this point, the algorithm will only focus on trying to dissociate the arm 2 from the arm 3 and never play any other arm. As the two arms have the same means, it will never end. The exploration term in TTUCB provides such infinite loop to occur.