# Sequential Learning (Homework)
Nicolas Nguyen & Kayané Robach

Mars 2022

## Part 1 - Rock Paper Scissors

### 1) Loss matrix

At each step, the player chooses one of 3 actions (rock or paper or scissors). The same holds for the adversary. Therefore $\boxed{M = N = 3}$.

The loss matrix is given by :

$$\boxed{L = \begin{pmatrix} 0 & +1 & -1 \\ -1 & 0 & +1 \\ +1 & -1 & 0 \end{pmatrix}}$$

### 2) Simulation against a fixed adversary

#### a) Vector loss

The loss $l_t(i)$ incurred by the player at time $i$ is $L_{i\cdot}$, the $i^{th}$ row of matrix $L$. Then if the adversary chooses action $j_t$, the loss incurred by the player is $L(i, j_t)$

#### b) Simulation for a fixed adversary

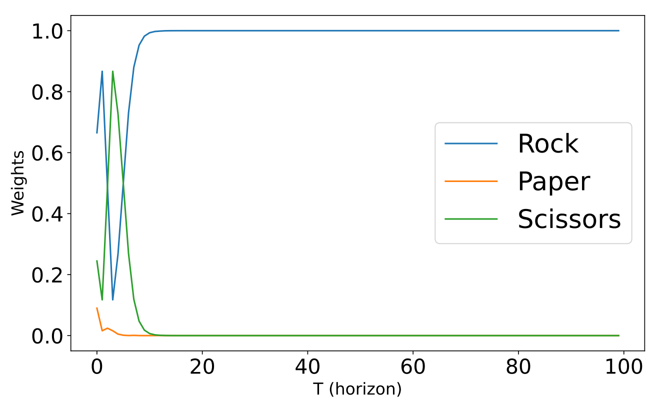



Figure 1: Evolution of the weight vectors for $T = 100$ using $EWA$ for parameter $\eta = 1$ against a fixed adversary $q_t = (\frac{1}{3}, \frac{1}{6}, \frac{1}{2})$.

Figure 1 shows that $EWA$ adapts to the fixed adversary by chosing Rock with probability one as far as $t$ grows. This is because the adversary chooses Scissors with a probability $\frac{1}{2}$ for all $t$.

#### c) Average loss over time

Figure 2 shows that using EWA leads to positive average gains as $t$ grows ($i.e.$ negative average losses).

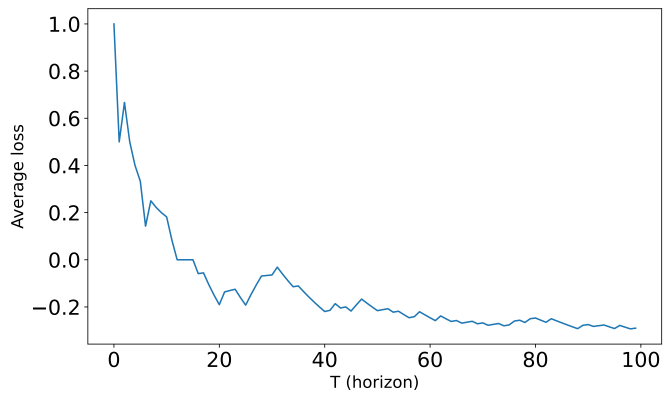



Figure 2: Average loss over time using $EWA$ with parameter $\eta = 1$ until $T = 100$, against a fixed adversary $q_t = (\frac{1}{3}, \frac{1}{6}, \frac{1}{2})$.

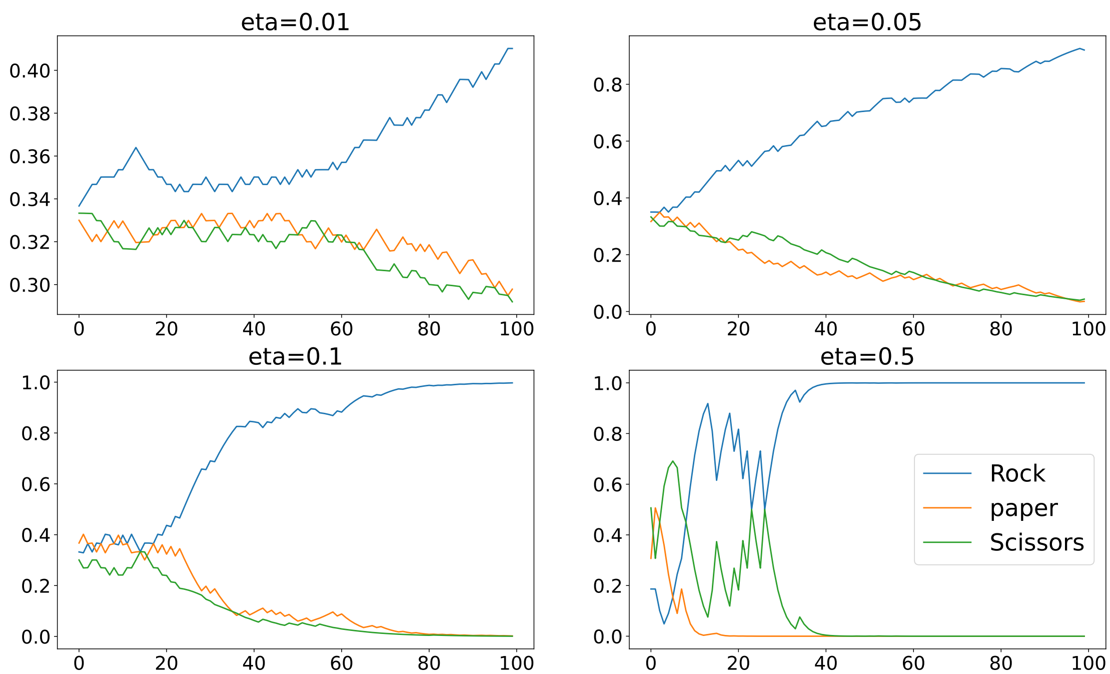



Figure 3: Results showing one instance of playing $EWA$ for different learning rates against a fixed adversary $q_t = (\frac{1}{3}, \frac{1}{6}, \frac{1}{2})$

### d) EWA for different learning rates

Figure 3 shows results playing one instance of $EWA$ for different learning parameters. We can see that the less $\eta$ is, the longer the algorithm takes time to clearly choose one action. Since the updates are proportional to $e^{-\eta \times loss}$, small $\eta$ tend to not penalise much "bad" actions. Small $\eta$ favors exploitation, meanwhile large $\eta$ lead to a faster convergence to the best action in this setting.

Theoretically, the best parameter should be given by $\boxed{\eta_{opt} = \sqrt{\frac{\log(K)}{T}}}$ so in this case $\eta \sim 0.10$ for $T = 100$. In practice higher values of $\eta$ can achieve a smaller cumulative regret and average loss as we can see on Figure 4.

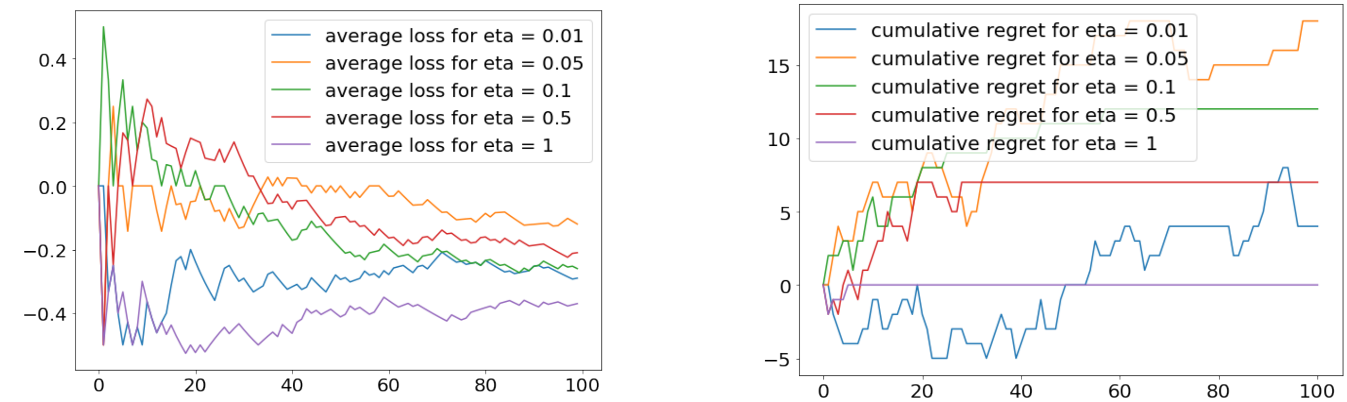
## 3) Simulation against an adaptative adversary (OGD)

### a) OGD update

For a loss defined by $l_t(q_t) = \sum_{j=1}^{N} q_t(j) g_t(j)$, we got $\nabla l_t(q_t) = g_t$ and the update of OGD becomes :

$$\boxed{q_{t+1} = \Pi_{\Delta_K}(q_t - \eta g_t)}$$

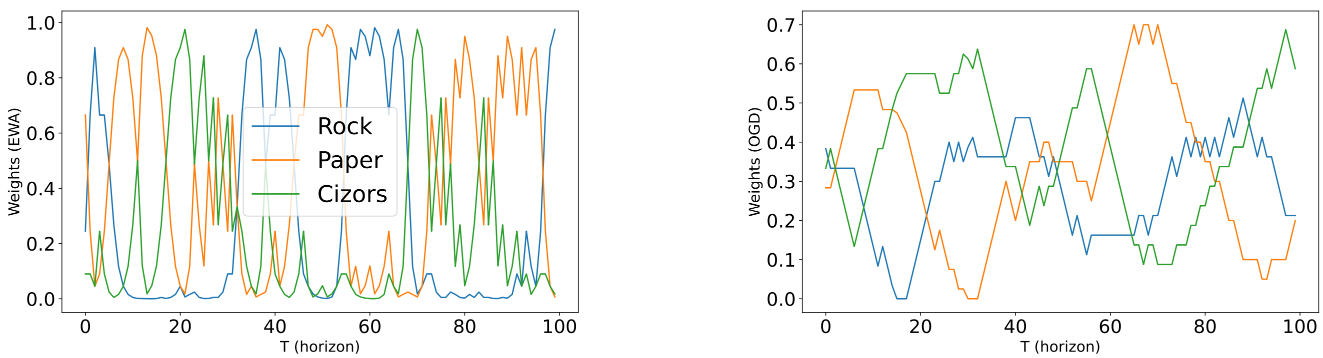
Where $\Pi_{\Delta_K}$ defines the projection onto the K-simplex (here $K = 3$).




(a) Average loss for different $\eta$.

(b) Cumulative regrets for different $\eta$.

Figure 4: Results showing that, in practice, $\eta = 1$ minimizes the average loss and cumulative regrets.




(a) Weights associated to *EWA*.

(b) Weights associated to *OGD*.

Figure 5: Evolution of the weight vectors for EWA (left) and OGD (right) for $T = 100$ using EWA ($\eta_{EWA} = 1$) against an adversary playing *OGD* ($\eta_{OGD} = 0.05$).

### b) Simulation of one instance of EWA vs OGD

Figure 5 shows that playing against an adaptative adversary is harder than playing against a fixed one : the weights do not converge anymore to a fixed strategy as in Figure 1. The weights seem to follow a cyclic pattern, as if each strategy want to adapt to the adversary strategy in a cyclic way.

### c) Average loss when playing against OGD

Figure 6 shows a positive loss asymptotically, which shows that **OGD seems to be a better strategy than EWA** in this setting.

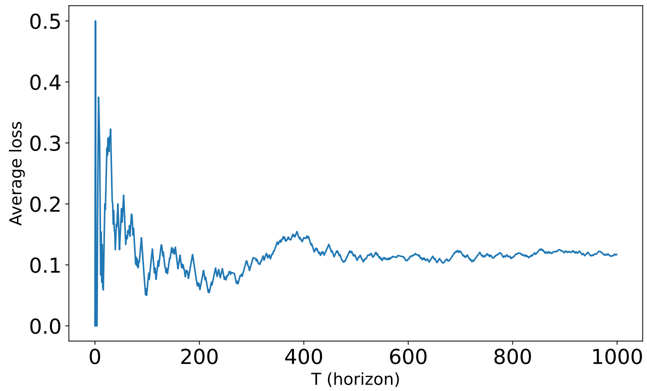



Figure 6: Average loss when playing against $OGD$ for $T = 1000$.

### d) Average weights

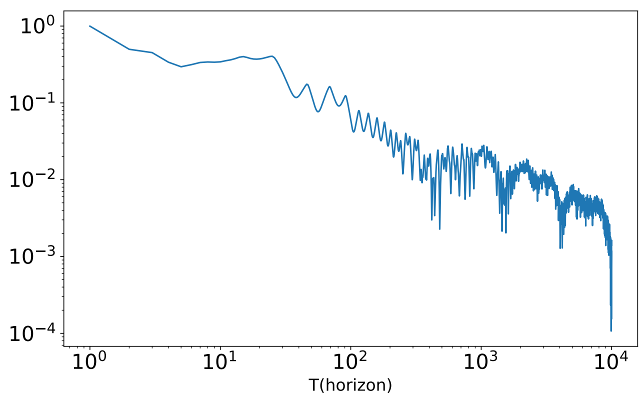



Figure 7: plot of $||\bar{p}_t - (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})||_2$ for $T = 10000$ in log-log scale.

Figure 7 shows that the average weights $\bar{p}_t$ tends (w.r.t. $||.||_2$) to the weights $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ as $t \to +\infty$. Since the game is symmetric each strategy would converge toward $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ which is a Nash equilibrium (indeed, there is no incentive to deviate from this strategy for any of the players ).

## Bandit feedback

### 4) EXP3 and EWA

EXP3 corresponds to the adaptation of EWA algorithm in a setup where players do not know the game in advance, that is to say **they only observe the loss of the actions played at time t** : $i_t$ and $j_t$. Therefore coordinate k of EWA weight at time t+1 :

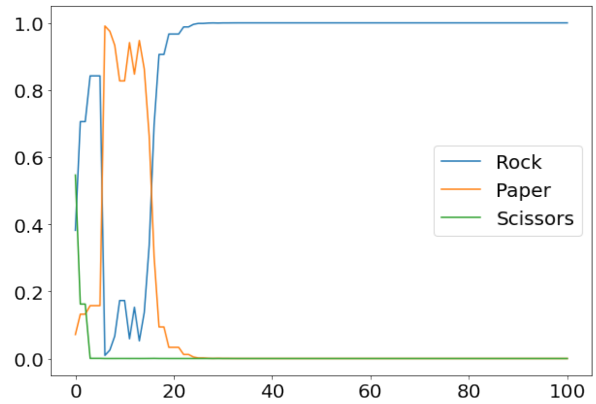
$$\frac{e^{-\eta \sum_{s=1}^{t} l_s(k)}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t} l_s(j)}}$$

cannot be computed since player does not observe $l_s(k)$ for k different than the action played.
Thus we estimate the loss with an unbiased estimator :

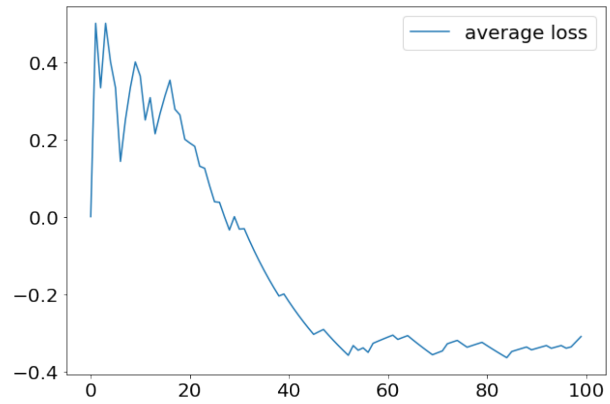
$$\hat{l}_t(k) = \frac{l_t(k)}{p_t(k)} \mathbb{1}_{\{k=i_t\}}$$

hence we define the weights in the EXP3 algorithm as :

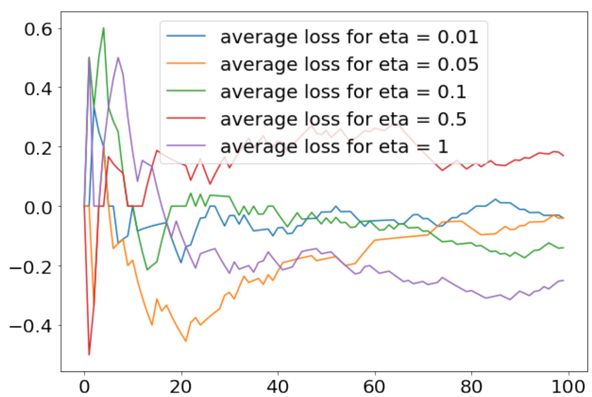
$$\frac{e^{-\eta \sum_{s=1}^{t} \hat{l}_s(k)}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t} \hat{l}_s(k)}}$$



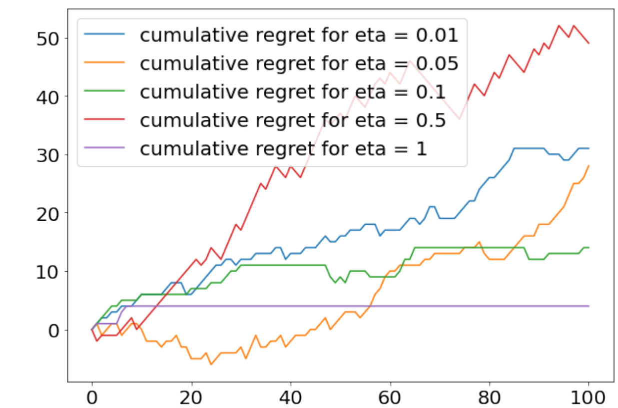

(a) Weights associated to *EXP3*.




(b) Average loss when playing *EXP3*.




(c) Average loss for different $\eta$ when playing *EXP3*.




(d) Cumulative regret for different $\eta$ when playing *EXP3*.

Figure 8: Bandit feedback when playing EXP3 against a fixed adversary.

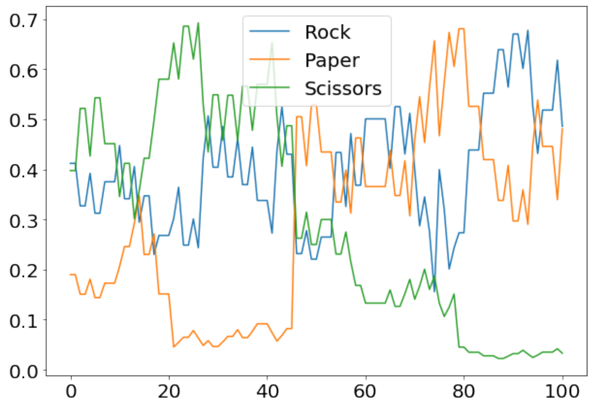
### 5) Simulation against a fixed adversary with EXP3 update

We can see on figure 8 that in our setting (players do not have full information feedback) the EXP3 algorithm against a fixed adversary has a similar behavior to the EWA algorithm in the previous context (Figure 1). Again, $\eta = 1$ seems to be the optimal learning rate value in practice.

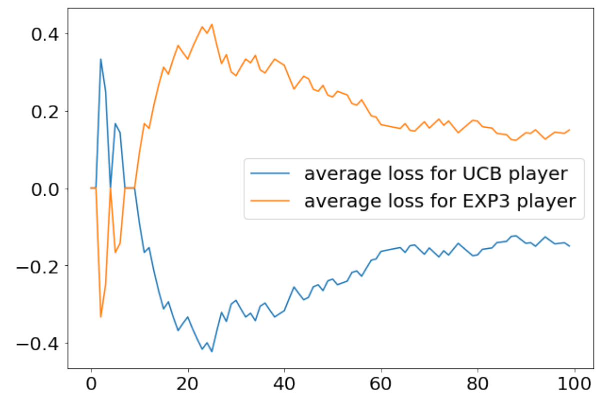
## Optional extensions

### 6) EXP3 and UCB

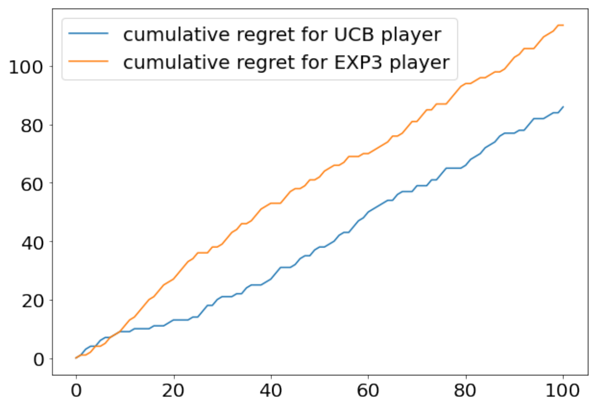
On Figure 9 we can see the results of the simulation of EXP3 competing with UCB.




(a) Weights associated to *EXP3*.




(b) Average loss for *UCB* competing with *EXP3*.

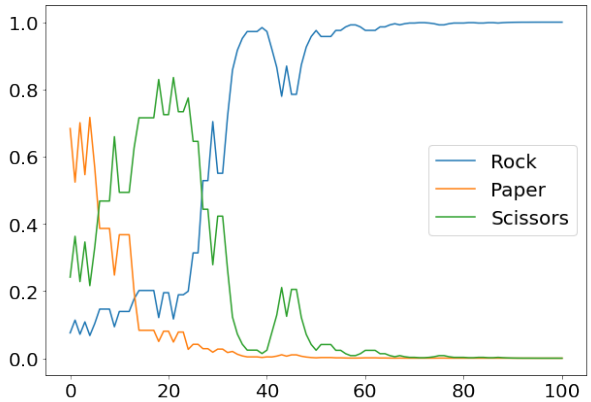



(c) Cumulative regret for *UCB* competing with *EXP3*.

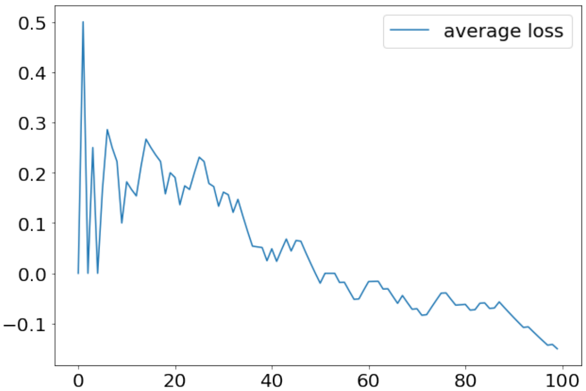
Figure 9: Results of the competition *EXP3* vs. *UCB* shows that both algorithms are good competitors in this setting.

For a random initialisation of the weight vector for *EXP3* algorithm against UCB we saw that UCB wins most of the time. Nonetheless weights do not converge for $T = 100$.

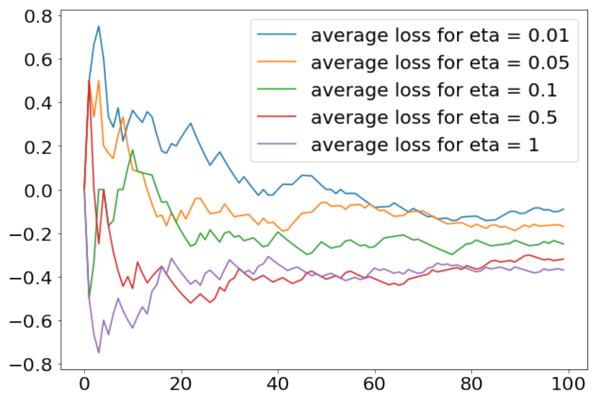
On figure 10 we can see the results of the simulation of *EXP3.IX* against a fixed adversary for $\gamma = 0.5$. Results are similar to the ones we got from *EXP3* (figure 8) except that the variance of the average loss and the cumulative regret is smaller with *EXP3.IX* compared to *EXP3*.



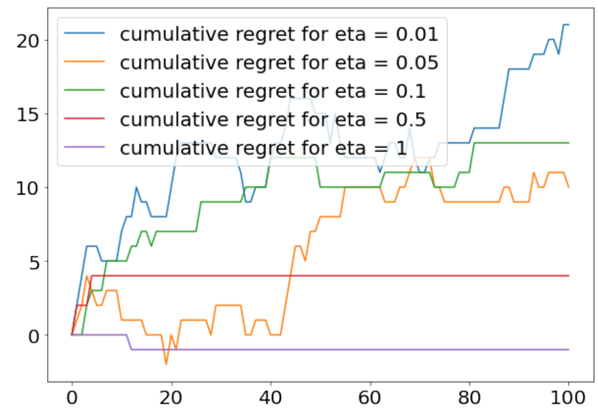

(a) Weights associated to *EXP3.IX*.




(b) Average loss when playing *EXP3.IX*.




(c) Average loss for different $\eta$ when playing *EXP3.IX*.

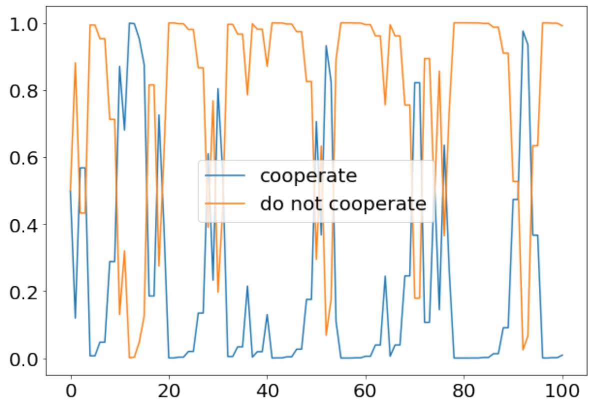



(d) Cumulative regret for different $\eta$ when playing *EXP3.IX*.

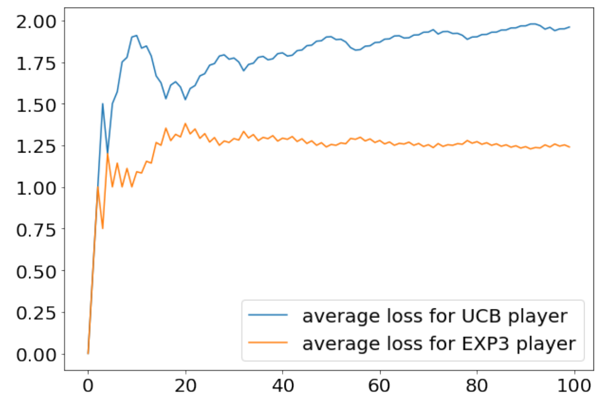
Figure 10: Results for bandit feedback with EXP3.IX against a fixed adversary for $\gamma = 0.5$

### 7) Prisoner's dilemma

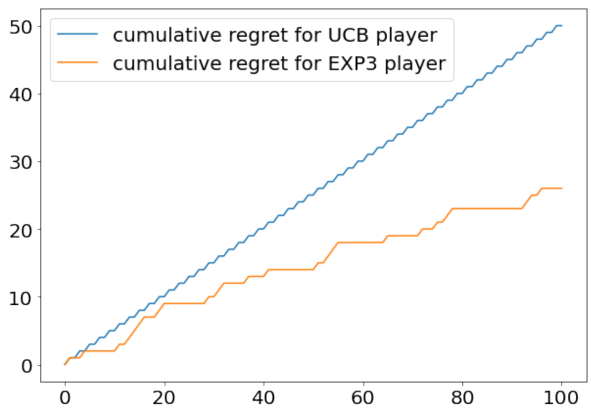
Game : *UCB* competing with *EXP3* ($\eta = 1$) for the prisoner dilemma.




(a) Weights associated to *EXP3*.




(b) Average loss for *UCB* competing with *EXP3*.

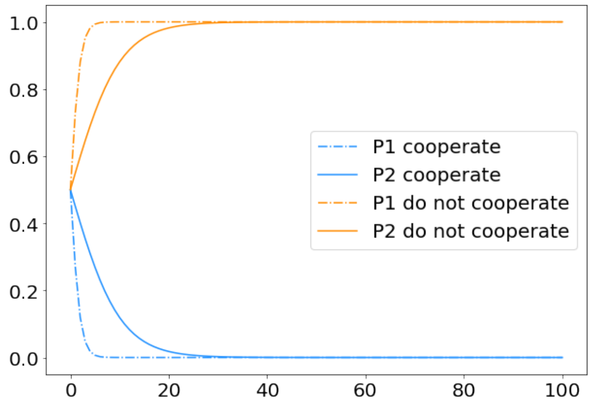



(c) Cumulative regret for *UCB* competing with *EXP3*.

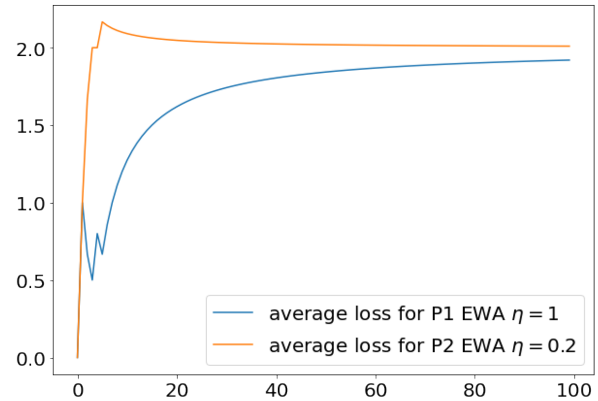
Figure 11: Prisoner's dilemma results when players use *UCB* and *EXP3* algorithms.

Figure 11 show the results for the prisoner dilemma between *UCB* and *EXP3* starting with uniform weight $(\frac{1}{2}, \frac{1}{2})$. *EXP3* weights do not converge and the algorithm wins against *UCB* in terms of cumulative regret.

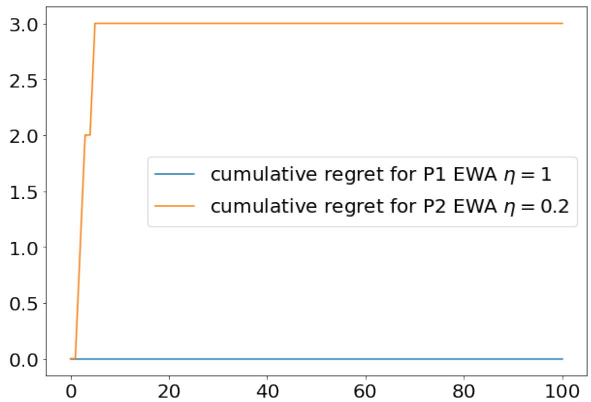
Game : *EWA* (P1, $\eta = 1$) competing with *EWA* (P2, $\eta = 0.2$) for the prisoner dilemma.




(a) Weights for both *EWA* players.




(b) Average loss when playing *EWA* with different $\eta$.




(c) Cumulative regret when playing *EWA* with different $\eta$.

Figure 12: Prisoner's dilemma results when both players use *EWA* algorithms.

Figure 12 show the results for the prisoner dilemma between two *EWA* players ; P1 with $\eta = 1$ and P2 with $\eta = 0.2$, starting with uniform weight $(\frac{1}{2}, \frac{1}{2})$. Weights quickly converge for both players toward the pure nash equilibrium that consist in no cooperation. The player 1 with $\eta = 1$ wins in terms of cumulative regret. This is an example where the nash equilibrium does not lead to an optimal outcome ; while there is no incentive to deviate from there strategy, both players would have been better off cooperating.

# Part 2 - Bernoulli bandits

## 1) Follow the leader

### a) Lower bound on the pseudo-regret

Suppose the previous setting with 2 arms ($K = 2$) with respecting means $\mu_1 = \frac{1}{2}$ and $\mu_2 = \frac{3}{5}$.
At time $t = 0$ (when we initialize the alogorithm by pulling all arms), with probability $\frac{1}{2} \times \frac{2}{5} = \frac{1}{5}$, the rewards are :

$$\begin{cases} 1 & \text{for arm 1} \\ 0 & \text{for arm 2} \end{cases}$$

According to FTL, we pull arm 1. It yields to $\hat{\mu}_1^1 > 0$ and $\hat{\mu}_2^1 = 0$. By induction, FTL keeps pulling arm 1 in any case.

We can thus give a lower bound on the expected regret (we call it pseudo regret according to the course).

$$\begin{aligned} \bar{R_T} &= \mathbb{E}\left[T.p_k - \sum_{t=1}^{T} p_{k_t}\right] \\ &= \sum_{k=1}^{K} \mathbb{E}\left[N_k(T)\Delta_k\right] \\ &= \mathbb{E}\left[N_1(T)(\frac{3}{5} - \frac{1}{2})\right] \qquad \text{since arm 2 is optimal} \\ &\geq \frac{1}{5}(T-1)\frac{1}{10} \end{aligned}$$

Where we call $N_k(T) = \sum_{i=1}^{T} \mathbb{1}_{\{k_t = k\}}$ *i.e.* the number of times arm $k$ is pulled before $T$, and $\Delta_k = p - p_k$.

Therefore in the case of Bernoulli stochastic bandits, $\boxed{\exists \alpha > 0, \bar{R_T} \geq \alpha T}$ (linear regret). Since we want a sub-linear regret, *FTL* is not a good algorithm in the case of stochastic bandits.

### c) Histogram of the regrets

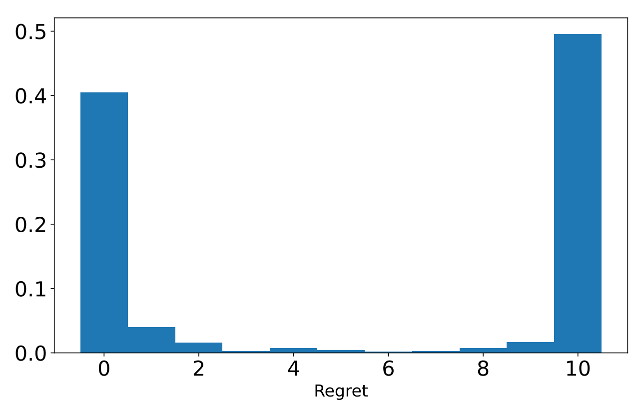



Figure 13: Histogram of the regret for $p = [0.5, 0.6]$

As we can see in Figure 13, the regret take mostly 2 values : 0 and 10.

- At $t = 1$ (when we pull both arms), if arm (1) gives 0 and arm (2) gives 1, FTL continues pulling arm (2), which is the best arm in expectation so the regret is 0.
- On the symmetric case as we saw before, if arm (1) gives arm 1 and (2) gives 0, FTL continues pulling arm (1) which is not the optimal arm, and yields to a regret of $0.6 \times T - 0.5 \times T = 10$ when $T = 100$.
- The other cases ($R_T \in ]0, 10[$) correspond to the case when both arm rewards 0 or 1 at first step.

This interpretation can be verified in the extreme case, for example when $p = [0.1, 0.9]$ (see Figure 14) since we first pull rewards $(1, 0)$ with probability $\frac{1}{100}$.

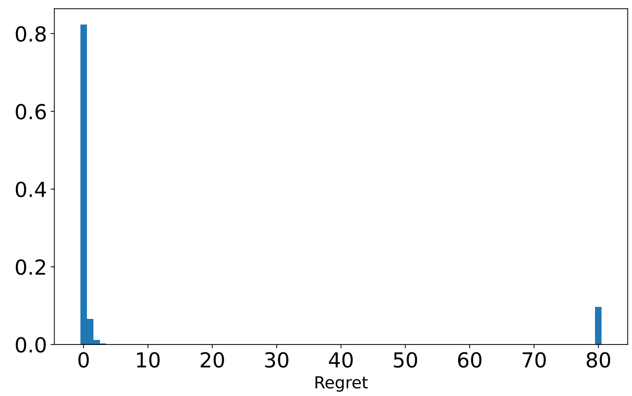



Figure 14: Histogram of the regret for $p = [0.1, 0.9]$

**d) Mean regret over different horizons**

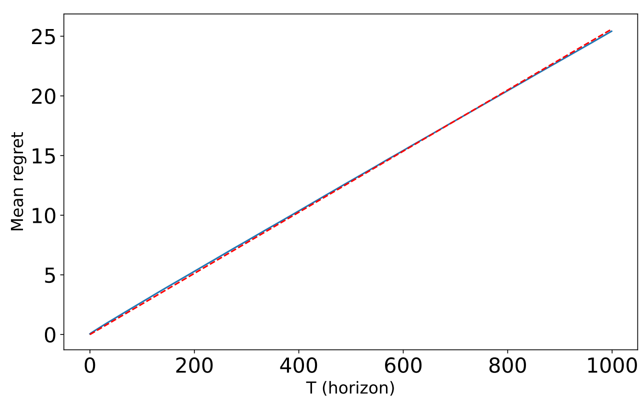



Figure 15: Pseudo-regret (averaged over 1000 iterations) depending on the horizon $T$ for $p = [0.5, 0.6]$. The red line represents the linear regression to show the linearity of the regret.

Averaging the regret confirms the bad behavior of FTL in the bandit setting. The pseudo-regret is indeed linear (as expected in question 1), see Figure 15.
Since we want a sublinear regret (in order to perform as good as the optimal arm asymptotically in expectation), **FTL is not a good algorithm in bandit setting.**

## 2) UCB

### a) Cumulant generative function for Bernoulli r.v.

Let $X \sim \mathcal{B}(p), p \in [0,1]$.

$$\begin{aligned}\forall \lambda \in \mathbb{R}, \quad \phi_X(\lambda) &= \log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}(X)}\right] \\ &= \log(\ e^{\lambda(1-p)}p + e^{\lambda(0-p)}(1-p)\ ) \\ &= \log(\ e^{-}(p(e^{\lambda}-1)+1)\ ) \\ &= -\lambda p + \log(\ 1-p+pe^{\lambda}\ )\end{aligned}$$

$$\boxed{\forall \lambda \in \mathbb{R}, \quad \phi_X(\lambda) = -\lambda p + \log(\ 1-p+pe^{\lambda}\ )}$$

### b) Bounded second derivative implies sub-gaussian

Let $X$ a r.v. for which $\phi_X \in \mathcal{C}^2(\mathbb{R})$. Let $\lambda \in \mathbb{R}$.
A second-order Taylor expansion in 0 yields to :

$$\begin{aligned}\phi_X(\lambda) &= \phi_X(0) + \lambda\phi'_X(0) + \int_0^{\lambda} \phi''_X(t)(\lambda - t)\, dt \\ &\leq \log \mathbb{E}[1] + \lambda\phi'_X(0) + \sigma^2 \int_0^{\lambda} \lambda - t\, dt \\ &\leq \lambda\phi'_X(0) + \sigma^2\frac{\lambda^2}{2}\end{aligned}$$

But

$$\phi'_X(\lambda) = \frac{\frac{d}{d\lambda}\mathbb{E}\left(e^{\lambda(X-\mathbb{E}(X)}\right)}{\mathbb{E}\left(e^{\lambda(X-\mathbb{E}(X)}\right)}$$

In a neighbourhood of 0, we can write the Fourier transform of $X - \mathbb{E}(X)$ (characteristic function) as a series (assumnig that in a neighborhood of 0, X admits moment at any order) :

$$\mathbb{E}\left(e^{\lambda(X-\mathbb{E}(X)}\right) = \mathbb{E}\left(\sum_{n=0}^{+\infty} \frac{\lambda^n (X-\mathbb{E}(X))^n}{n!}\right) = \sum_{n=0}^{+\infty} \frac{\lambda^n \mathbb{E}\left((X-\mathbb{E}(X))^n\right)}{n!}$$

In a neighbourhood of 0, this series is $\mathcal{C}^{\infty}$ and we can write :

$$\begin{aligned}\psi_X(\lambda) := \frac{d}{d\lambda}\mathbb{E}(e^{\lambda(X-\mathbb{E}(X))}) &= \sum_{n=0}^{+\infty} \frac{n\lambda^{n-1}\mathbb{E}((X-\mathbb{E}(X))^n)}{n!} \\ &= \sum_{n=1}^{+\infty} \frac{\lambda^{n-1}\mathbb{E}((X-\mathbb{E}(X))^n)}{(n-1)!}\end{aligned}$$

And then

$$\psi_X(0) = \mathbb{E}(X - \mathbb{E}(X)) = 0$$

This conclude the fact that

$$\boxed{\phi_X(t) \leq \sigma^2\frac{\lambda^2}{2}} \tag{1}$$

### c) Range of the variance

Let's compute the second derivative of the cumulant generating function for a Bernoulli r.v. $X$ with parameter $p \in ]0,1[$ ; Recall we have $\forall \lambda \in \mathbb{R}$,

$$\begin{aligned}
&\phi(\lambda) = -\lambda p + \log(e^\lambda p + (1-p)) \quad \text{which is twice differentiable} \\
\implies &\phi'(\lambda) = -p + p\frac{e^\lambda}{e^\lambda p + (1-p)} \\
\implies &\phi''(\lambda) = p(1-p)\frac{e^\lambda}{(e^\lambda p + (1-p))^2)} := \psi(\lambda)
\end{aligned}$$

A basic function study of $\psi$ shows that it atteins its maximum when $\lambda_{max} = \log(\frac{1-p}{p})$ if $p \neq 0$ and $p \neq 1$, so in this case

$$\forall \lambda \in \mathbb{R}, \psi(\lambda) = \phi''(\lambda) \leq \frac{1}{4}$$

In the case $p = 0$ or $p = 1$, $\forall \lambda \in \mathbb{R}$ $\phi''(\lambda) = 0$ and $\sigma^2 = 0$. Therefore, according to 2b), $\phi''(\lambda) \leq \frac{1}{4} \implies X$ is $\frac{1}{4}$ sub-gaussian (so if we take $\frac{1}{4} \leq \sigma^2$, $X$ is still $\sigma^2$-subgaussian).

$$\boxed{\begin{cases} \sigma^2 = 0 & \text{for } p \in \{0,1\} \\ \sigma^2 = \frac{1}{4} & \text{for } p \in (0,1) \end{cases}}$$

#### d) Bound by Bernoulli cumulant function

For a r.v. $X$ on $[0,1]$ with $\mathbb{E}(X) = p$,

$$\begin{aligned}
\mathbb{E}[e^{\lambda(X-p)}] &= e^{-\lambda p}\mathbb{E}[e^{\lambda X}] \\
&\overset{(\star)}{\leq} e^{-\lambda p}(1 - \mathbb{E}[X] + \mathbb{E}[Xe^\lambda]) \\
&\leq e^{-\lambda p}(1 - p + pe^\lambda)
\end{aligned}$$

So $\log \mathbb{E}[e^{\lambda(X-p)}] \leq -\lambda p + \log(1 - p + pe^\lambda) \leq \phi_Y(\lambda)$ according to 2a), *i.e.* $\boxed{\phi_X(\lambda) \leq \phi_Y(\lambda)}$

($\star$) We used the fact that $\forall x \in [0,1], \forall \lambda \in \mathbb{R}, e^{\lambda x} \leq 1 - x + xe^\lambda$. In fact, consider the function study of $g_\lambda(x) = 1 - x + xe^\lambda - e^{\lambda x}$ which is differentiable, and remark that $\forall \lambda \in \mathbb{R}, g_\lambda(0) = 0$ and $g_\lambda(1) = 0$. Then distinguish the cases $\lambda > 0, \lambda = 0$ and $\lambda < 0$ ; they yield to the same conclusion, $\forall x \in [0,1], \forall \lambda \in \mathbb{R}$ $g_\lambda(x) \geq 0$.

#### e) All bounded r.v. are subgaussian

Let $X$ such a random variable. Then remark that $\mathbb{E}(X) = p \in [0,1]$ so $\phi_X(\lambda) \leq \phi_Y(\lambda)$ by question 2d), with $Y \sim \mathbb{B}(p)$. USing 2c) :

- For $p \in \{0,1\}$ $Y$ is 0-sub-gaussian. Then, for $\sigma^2 \geq 0$, $\phi_Y(\lambda) \leq \frac{1}{2}\sigma^2\lambda^2$ so $\phi_X(\lambda) \leq \frac{1}{2}\sigma^2\lambda^2 \implies X$ is $\sigma^2$ sub-gaussian.
- For $p \in (0,1)$ $Y$ is $\frac{1}{4}$-sub-gaussian. Then, for $\sigma^2 \geq \frac{1}{4}$, $\phi_Y(\lambda) \leq \frac{1}{2}\sigma^2\lambda^2$ so $\phi_X(\lambda) \leq \frac{1}{2}\sigma^2\lambda^2 \implies X$ is $\sigma^2$ sub-gaussian.

#### g) Mean regret of UCB

Figure 16 clearly shows that in this setting, $UCB(\frac{1}{4})$ outperforms FTL. The pseudo-regrets shown in this figure confirms the theory, because we can proove a worst-case upper-bound on the pseudo-regret of UCB in $\mathcal{O}(\sqrt{KTlog(T)})$.

#### h) Influence on the variance

Figure 17 shows that the performances are dependant of the variance $\sigma^2$. UCB($\frac{1}{4}$) has the lowest mean regret for the weight vector $[0.6, 0.5]$, which is coherent with what we knew. However for the case $p = [0.85, 0.95]$, **the optimal parameter changes** : the mean regret seems to be minimal for $\sigma^2 = \frac{1}{16}$. As $\sigma^2$ grows, the confidence interval gets wider hence the bad performances of UCB algorithm : it incites to follow sub-optimal arms.

### 3)

Figure 18 shows that the function $\sigma^2(.)$ upper-bounds the variance of the Bernoulli r.v.

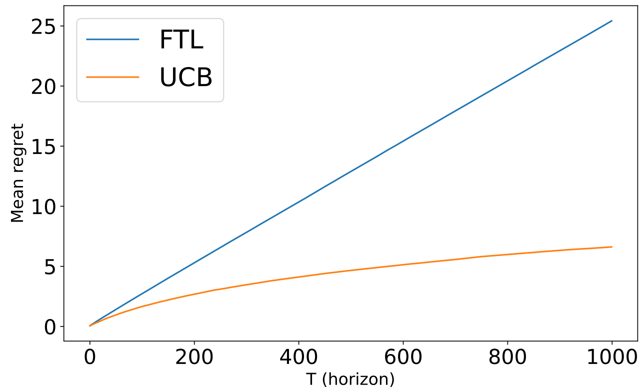



Figure 16: Mean regret for different horizons for FTL (blue) and UCB with $\sigma^2 = \frac{1}{4}$ (orange). The pseudo regrets have been averaged over 1000 iterations.

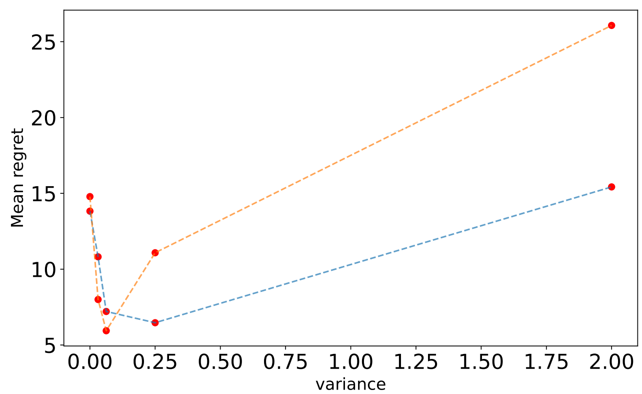



Figure 17: Mean regret for $\sigma^2 \in \{0, \frac{1}{32}, \frac{1}{16}, \frac{1}{4}, 1\}$ for $T = 1000$ averaged over 1000 iterations. $p = [0.6, 0.5]$ (blue), $p = [0.85, 0.95]$ (orange).

## 4) Sub-gaussianity implies bounded variance

Let $X$ a $\sigma^2$ sub-gaussian r.v. so that $\mathbb{E}(e^{\lambda X}) \leq e^{\frac{\lambda^2 \sigma^2}{2}}$. A taylor expansion yields to :

$$\begin{aligned} \mathbb{E}(e^{\lambda X}) &= \sum_{n=0}^{+\infty} \frac{\lambda^n}{n!} \mathbb{E}(X^n) \\ &\leq e^{\frac{\lambda^2 \sigma^2}{2}} = \sum_{n=0}^{+\infty} \frac{1}{n!} \left( \frac{\lambda^2 \sigma^2}{2} \right)^n \end{aligned}$$

Therefore

$$\begin{aligned} & \mathbb{E}(X)t + \mathbb{E}(X^2)\frac{t^2}{2} \leq \frac{\sigma^2 t^2}{2} + o(t^2) \\ \iff & \mathbb{E}(X) + \mathbb{E}(X^2)\frac{t}{2} \leq \frac{\sigma^2 t}{2} + \frac{o(t^2)}{t} \end{aligned}$$

Letting $t \to 0^+$ yields to $\mathbb{E}(X) \leq 0$. The symmetric case when $t \to 0^-$ yields to $\mathbb{E}(X) \geq 0$, hence $\mathbb{E}(X) = 0$.

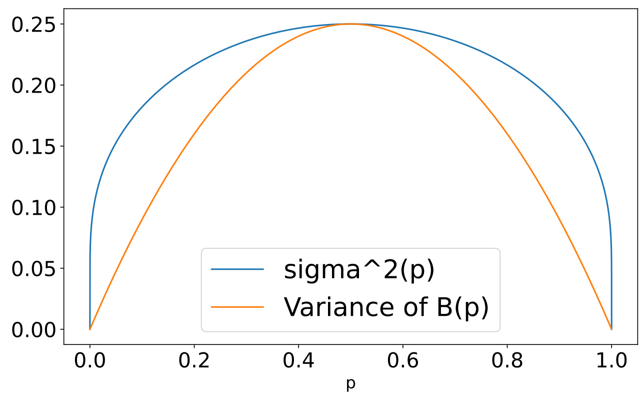



Figure 18: $\sigma^2(p)$ (blue) and $\mathbb{V}(\mathcal{B}(p))$ (orange) for different values of $p$.

Then by dividing the first expression by $t^2$ and using that $\mathbb{E}(X) = 0$ :

$$\begin{aligned} &\mathbb{E}(X^2)\frac{t^2}{2} \leq \frac{\sigma^2 t^2}{2} + o(t^2) \\ \iff &\mathbb{E}(X^2) \leq \sigma^2 + \frac{2o(t^2)}{t^2} \end{aligned}$$

letting $t \to 0$ concludes that $\mathbb{V}(X) \leq \sigma^2$.

## 5) Adaptation to the variance

### a)

Let's start from the definition of $\hat{v_t}^k$ :

$$\begin{aligned} \hat{v_t}^k N_t^k &= \sum_{s=1}^t \mathbb{1}_{\{k_s=k\}} \left(X_s^{k_s} - \hat{\mu_t}^k\right)^2 \\ &= \sum_{s=1}^t \mathbb{1}_{\{k_s=k\}} \left(X_s^{k_s}\right)^2 + \left(\hat{\mu_t}^k\right)^2 - 2X_s^{k_s}\hat{\mu_t}^k \\ &= \sum_{s=1}^t \mathbb{1}_{\{k_s=k\}} \left(X_s^{k_s}\right)^2 + \left(\hat{\mu_t}^k\right)^2 N_t^k - 2\hat{\mu_t}^k \sum_{s=1}^t X_s^{k_s}\mathbb{1}_{\{k_s=k\}} \end{aligned}$$

But

$$\begin{aligned} \left(\hat{\mu_t}^k\right)^2 N_t^k - 2\hat{\mu_t}^k \sum_{s=1}^t X_s^{k_s}\mathbb{1}_{\{k_s=k\}} &= \left(\frac{1}{N_t^k}\sum_{s=1}^t \mathbb{1}_{\{k_s=k\}} X_s^{k_s}\right)^2 N_t^k - 2\left(\frac{1}{N_t^k}\sum_{s=1}^t \mathbb{1}_{\{k_s=k\}} X_s^{k_s}\right)\left(\sum_{s=1}^t X_s^{k_s}\mathbb{1}_{\{k_s=k\}}\right) \\ &= \frac{1}{N_t^k}\left(\sum_{s=1}^t \mathbb{1}_{\{k_s=k\}} X_s^{k_s}\right)^2 - 2\frac{1}{N_t^k}\left(\sum_{s=1}^t \mathbb{1}_{\{k_s=k\}} X_s^{k_s}\right)\left(\sum_{s=1}^t \mathbb{1}_{\{k_s=k\}} X_s^{k_s}\right) \\ &= -\frac{1}{N_t^k}\left(\sum_{s=1}^t \mathbb{1}_{\{k_s=k\}} X_s^{k_s}\right)^2 \end{aligned}$$

Which permits to conclude :

$$\boxed{\hat{v_t}^k N_t^k = \sum_{s=1}^t \mathbb{1}_{\{k_s=k\}} \left(X_s^{k_s}\right)^2 - \frac{1}{N_t^k}\left(\sum_{s=1}^t \mathbb{1}_{\{k_s=k\}} X_s^{k_s}\right)^2}$$

**b)**

$$
\begin{aligned}
N_{t+1}^{k_{t+1}}\hat{v}_{t+1}^{k_{t+1}} &= \sum_{s=1}^{t+1}\mathbb{1}_{\{k_s=k_{t+1}\}}\left(X_s^{k_s}\right)^2 - \frac{1}{N_{t+1}^{k_{t+1}}}\left(\sum_{s=1}^{t+1}X_s^{k_s})^2\right) \\
&= \sum_{s=1}^{t}\mathbb{1}_{\{k_s=k_{t+1}\}}\left(X_s^{k_s}\right)^2 + \left(X_{t+1}^{k_{t+1}}\right)^2 - \hat{\mu}_{t+1}^{k_{t+1}}\left(\sum_{s=1}^{t}\mathbb{1}_{\{k_s=k_{t+1}\}}\right) - \hat{\mu}_{t+1}^{k_{t+1}}X_{t+1}^{k_{t+1}} \\
&= N_t^{k_{t+1}}\hat{v}_t^{k_{t+1}} + \frac{1}{N_t^{k_{t+1}}}\left(\sum_{i=1}^{t}\mathbb{1}_{\{k_s=k_{t+1}\}}X_s^{k_s}\right)^2 + \left(X_{t+1}^{k_{t+1}}\right)^2 - \hat{\mu}_{t+1}^{k_{t+1}}\hat{\mu}_t^{k_{t+1}}N_t^{k_{t+1}} - \hat{\mu}_{t+1}^{k_{t+1}}X_{t+1}^{k_{t+1}} \\
&= N_t^{k_{t+1}}\hat{v}_t^{k_{t+1}} + \hat{\mu}_t^{k_{t+1}}\hat{\mu}_t^{k_{t+1}}N_t^{k_{t+1}} + (X_{t+1}^{k_{t+1}})^2 - \hat{\mu}_{t+1}^{k_{t+1}}\hat{\mu}_t^{k_{t+1}}N_t^{k_{t+1}} - \hat{\mu}_{t+1}^{k_{t+1}}X_{t+1}^{k_{t+1}} \\
&= N_t^{k_{t+1}}\hat{v}_t^{k_{t+1}} + \hat{\mu}_t^{k_{t+1}}N_t^{k_{t+1}}\left(\hat{\mu}_t^{k_{t+1}} - \hat{\mu}_{t+1}^{k_{t+1}}\right) + \left(X_{t+1}^{k_{t+1}}\right)^2 - \hat{\mu}_{t+1}^{k_{t+1}}X_{t+1}^{k_{t+1}}
\end{aligned}
$$

Now

$$
\begin{aligned}
\hat{\mu}_t^{k_{t+1}}N_t^{k_{t+1}}\left(\hat{\mu}_t^{k_{t+1}} - \hat{\mu}_{t+1}^{k_{t+1}}\right) &= \hat{\mu}_t^{k_{t+1}}N_t^{k_{t+1}}\left(\frac{1}{N_t^{k_{t+1}}}\sum_{s=1}^{t}\mathbb{1}_{\{k_s=k_{t+1}\}}X_s^{k_s} - \frac{1}{N_{t+1}^{k_{t+1}}}\sum_{s=1}^{t}\mathbb{1}_{\{k_s=k_{t+1}\}}X_s^{k_s}\right) \\
&= \hat{\mu}_t^{k_{t+1}}\left(\sum_{s=1}^{t}\mathbb{1}_{\{k_s=k_{t+1}\}}X_s^{k_s} - \frac{N_t^{k_{t+1}}}{N_{t+1}^{k_{t+1}}}\sum_{s=1}^{t+1}\mathbb{1}_{\{k_s=k_{t+1}\}}X_s^{k_s}\right) \\
&= \hat{\mu}_t^{k_{t+1}}\left(\sum_{s=1}^{t}\mathbb{1}_{\{k_s=k_{t+1}\}}X_s^{k_s} - \frac{N_t^{k_{t+1}}}{N_t^{k_{t+1}}+1}\sum_{s=1}^{t+1}\mathbb{1}_{\{k_s=k_{t+1}\}}X_s^{k_s}\right) \\
&= \hat{\mu}_t^{k_{t+1}}\left(\sum_{s=1}^{t}\mathbb{1}_{\{k_s=k_{t+1}\}}X_s^{k_s} - \sum_{s=1}^{t+1}\mathbb{1}_{\{k_s=k_{t+1}\}}X_s^{k_s}X_s^{k_s} + \frac{1}{N_{t+1}^{k_{t+1}}}\sum_{s=1}^{t+1}\mathbb{1}_{\{k_s=k_{t+1}\}}X_s^{k_s}\right) \\
&= \hat{\mu}_t^{k_{t+1}}\left(-X_{t+1}^{k_{t+1}} + \hat{\mu}_{t+1}^{k_{t+1}}\right)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
N_{t+1}^{k_{t+1}}\hat{v}_{t+1}^{k_{t+1}} &= N_t^{k_{t+1}}\hat{v}_t^{k_{t+1}} + \hat{\mu}_t^{k_{t+1}}\left(-X_{t+1}^{k_{t+1}} + \hat{\mu}_{t+1}^{k_{t+1}}\right) + \left(X_{t+1}^{k_{t+1}}\right)^2 - \hat{\mu}_{t+1}^{k_{t+1}} \\
&= N_t^{k_{t+1}}\hat{v}_t^{k_{t+1}} + \hat{\mu}_t^{k_{t+1}}\left(-X_{t+1}^{k_{t+1}} + \hat{\mu}_{t+1}^{k_{t+1}}\right) + X_{t+1}^{k_{t+1}}\left(X_{t+1}^{k_{t+1}} - \hat{\mu}_{t+1}^{k_{t+1}}\right)
\end{aligned}
$$

Therefore

$$
\boxed{N_{t+1}^{k_{t+1}}\hat{v}_{t+1}^{k_{t+1}} = N_t^{k_{t+1}}\hat{v}_t^{k_{t+1}} + \left(X_{t+1}^{k_{t+1}} - \hat{\mu}_t^{k_{t+1}}\right)\left(X_{t+1}^{k_{t+1}} - \hat{\mu}_{t+1}^{k_{t+1}}\right)}
$$

This formulation is practical since it permits to update $\hat{v}_{t+1}^{k_{t+1}}$ in an online fashion with just the rewards $X_{t+1}$.

### d) UCB and UCB-V

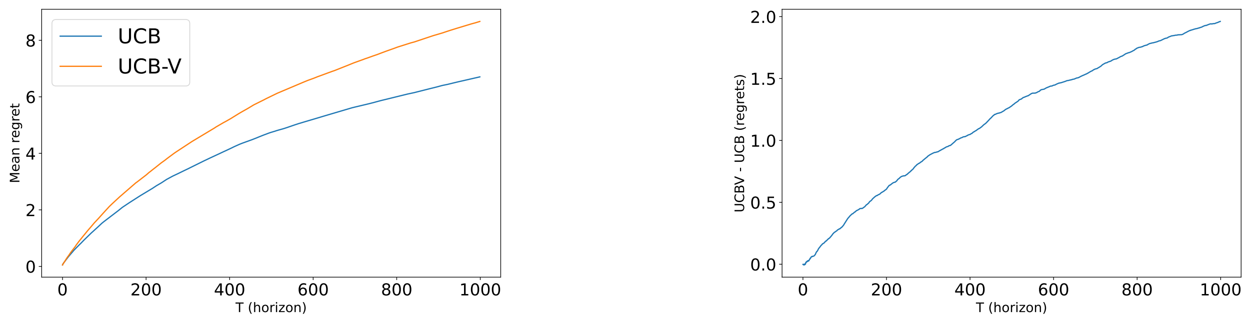
Figure 19 shows that in this setting ($p = [0.5, 0.6]$), UCB($\frac{1}{4}$) performs better than UCB-V.

### e) UCB and UCB-V (for different parameters)

Since we know that the variance of $X \sim \mathcal{B}(p)$ is $p(1-p)$ which is maximum for $p = \frac{1}{2}$, the confidence interval is bigger for the case $p = [0.5, 0.6]$ than the case $p = [0.1, 0.2]$. UCB-V, by estimating the variances, improves over UCB in low variance cases such as for $p = [0.1, 0.2]$ as we can see comparing figures (a) from 19 and 20. This is quite intuitive since the expected regret upper bound of the UCB-V algorithm depends positively on the variance.
On figure 20, we can see that the case $p = [0, 0.1]$ has 0 regret for both UCB and UCB-V algorithms since the optimal arm always yields to a reward of 1 with a variance of 0, and we always choose arm (1).

## 6) KL-UCB

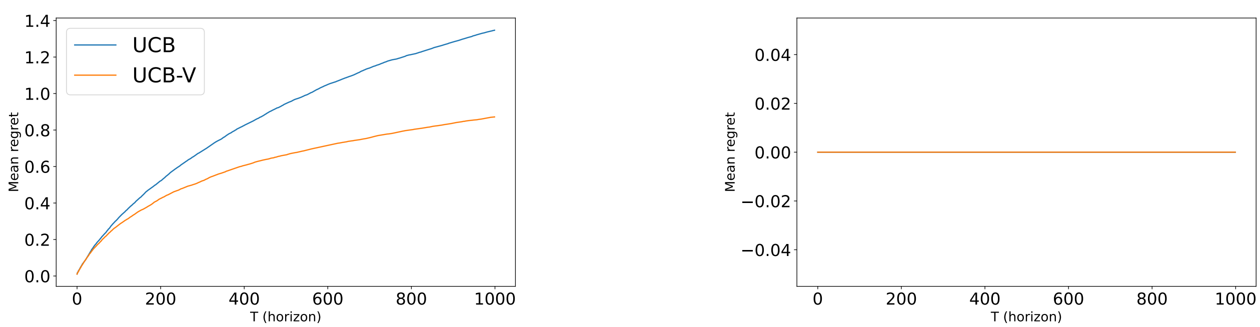
...




(a) Mean regrets for $UCB(\frac{1}{4})$ and UCB-V algorithms.

(b) Expected regret difference of UCB-V and UCB.

Figure 19: Comparison of $UCB(\frac{1}{4})$ and UCB-V for $K = 2$ and $p = [0.5, 0.6]$ over 1000 iterations.




(a) $p = [0.1, 0.2]$

(b) $p = [0, 0.1]$

Figure 20: Mean regrets of $UCB(\frac{1}{4})$ (blue) and UCB-V (orange) averaged over 1000 iterations for $K = 2$.