Sequential Learning

Home Assignment

Part 1. Rock Paper Scissors

1. For the game "Rock paper scissors", $M = N = 3$ and:

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

Full information feedback We assume that both players know the matrix L in advance and can compute $L(i, j)$ for any (i, j).

2. (a) see code `homework.ipynb` available at: `https://github.com/clementgr/sequential-learning`

   (b) see code `homework.ipynb` available at: `https://github.com/clementgr/sequential-learning`

3. *Simulation against a fixed adverary.*

   (a) The loss $\ell_t(i_t)$ incurred by the player if he chooses action i_t at time t is given by $\ell_t(i_t) = L[i_t, j_t]$.

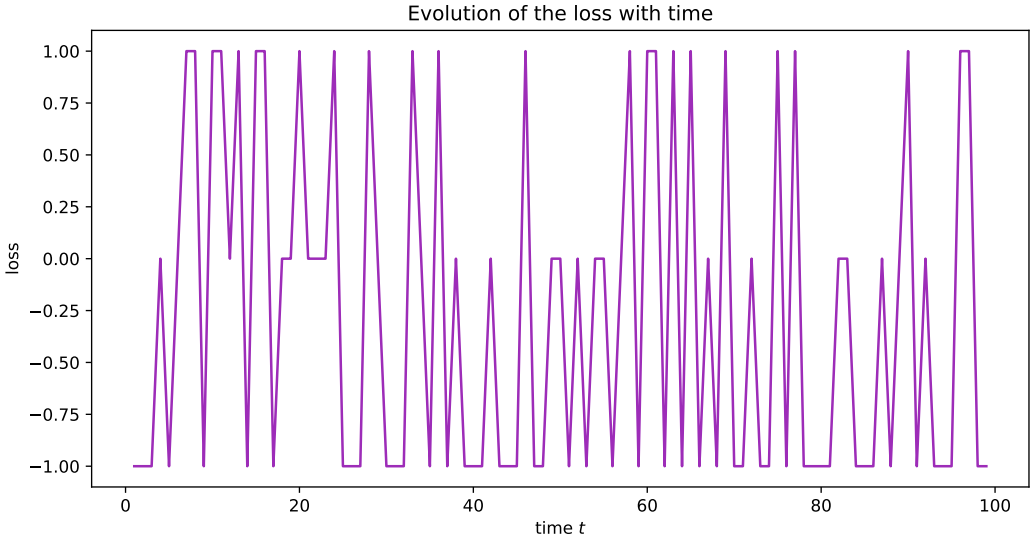

Figure 1: Loss incurred by player p over time for an instance of the game (`fixed EWA`)

(b)

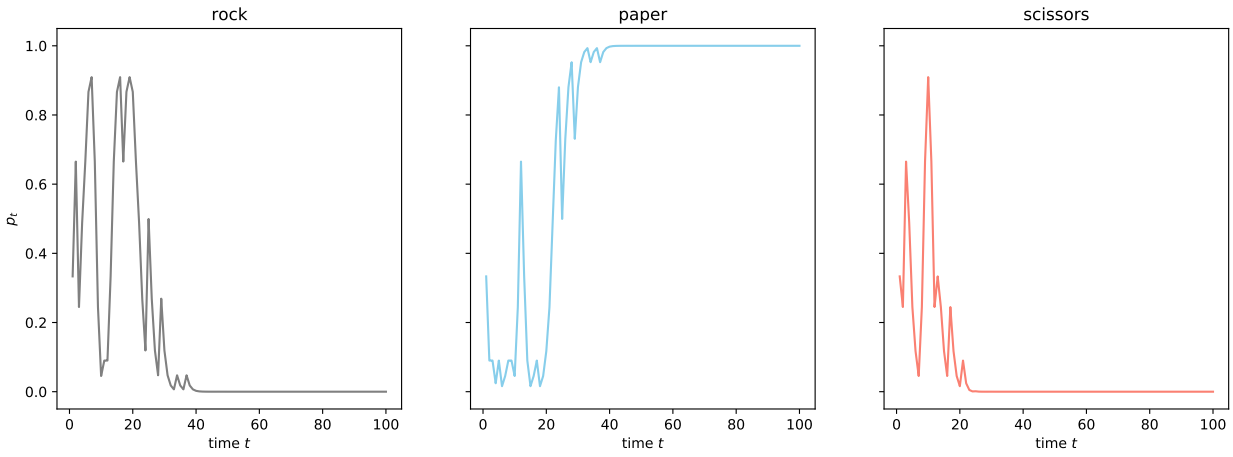

Figure 2: Evolution of the weight vectors $\left(p^{\text{rock}}, p^{\text{paper}}, p^{\text{scissors}}\right)$ over time (`fixed EWA`)

Against this adversary, a good strategy is to play "paper" as – on average – the adversary plays "rock" every two games. Figure 2 shows that experiments support this is a good strategy.

(c)

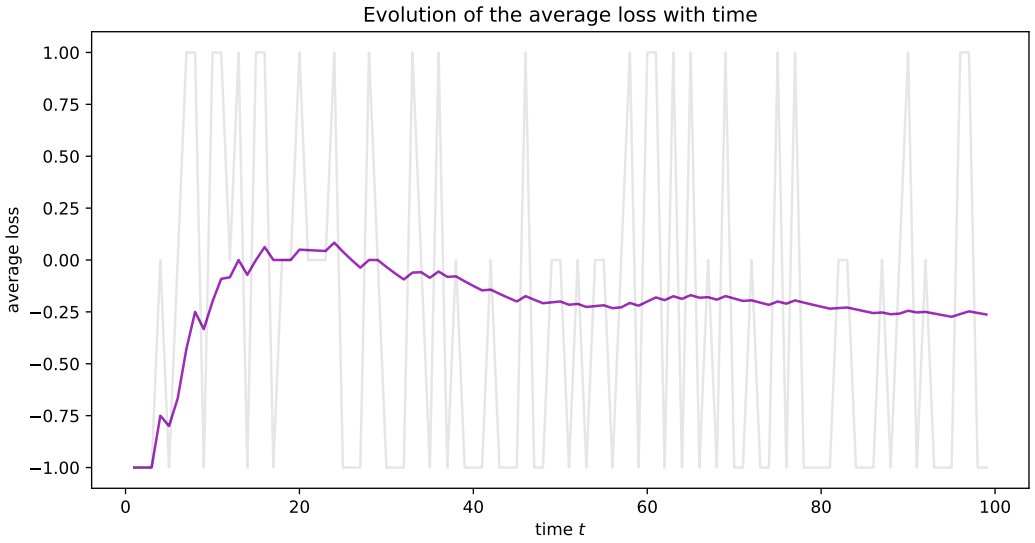

Figure 3: Average loss suffered by player p over time (`fixed EWA`)

(d)

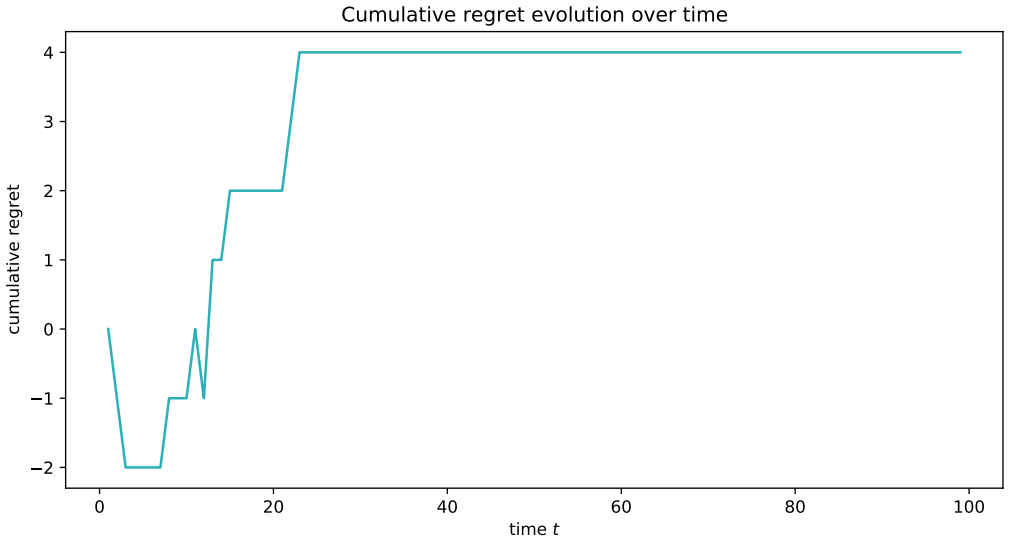

Figure 4: Cumulative regret of player p over time (`fixed EWA`)

As soon as player p start playing "paper" frequently (experimentally, this corresponds to $t' \approx 25$), the regret becomes null such that the cumulative regret stays constant.

(e)

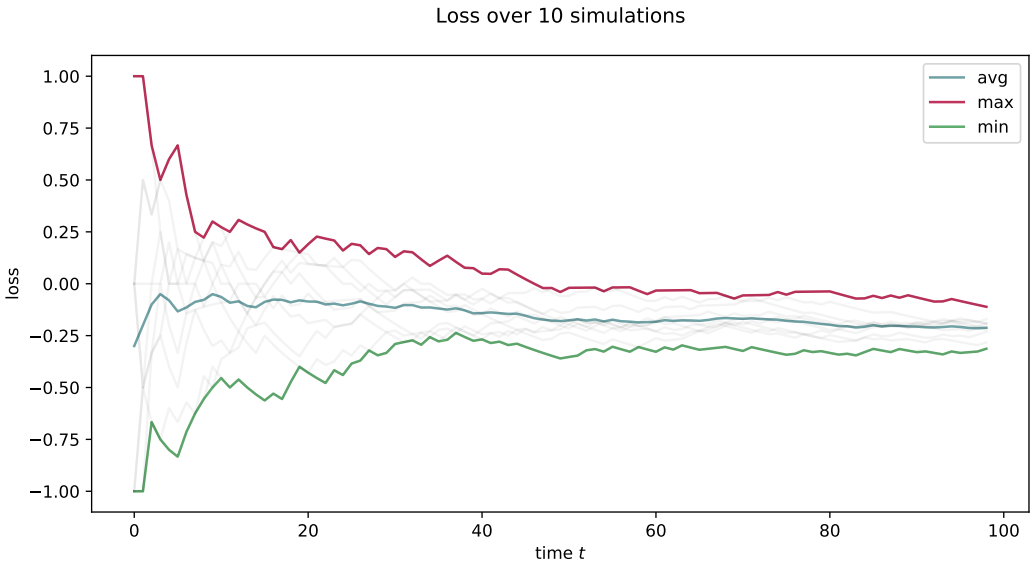

Figure 5: Average loss obtained in average, in maximum and in minimum over 10 simulations (`fixed EWA`)

(f) Let's denote $\eta_1 = 0.01, \eta_2 = 0.05, \eta_3 = 0.1, \eta_4 = 0.5, \eta_5 = 1$.

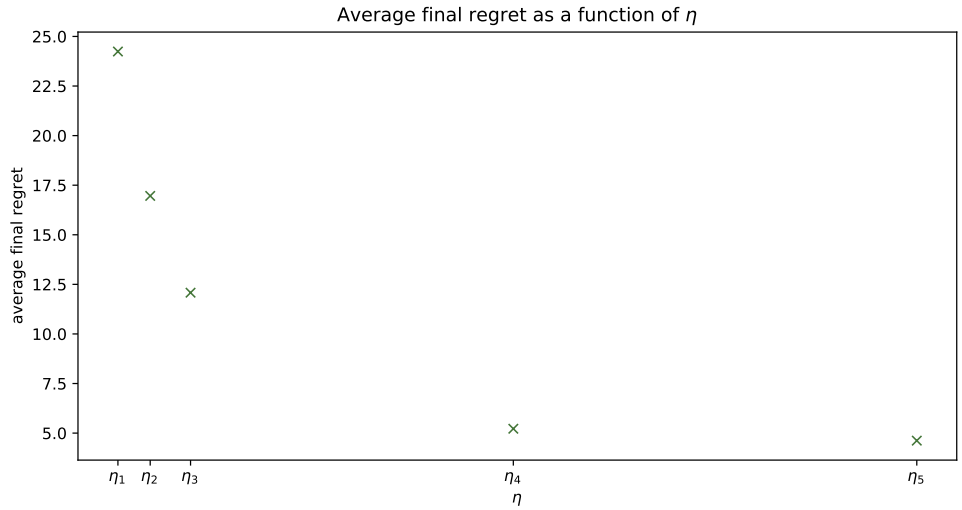

Figure 6: Final regret as a function of η (`fixed EWA`)

The best η in theory is given by:

$$\eta^\star_{\text{th}} = \sqrt{\frac{\log K}{T}}$$

In our case, $K = 3$ and $T = 1000$ give $\eta^\star_{\text{th}} = 0.105$.
In practice, however, the best η is $\eta^\star_{\text{exp}} = 1.0$.

4. *Simulation against an adaptive adversary.*

(a)

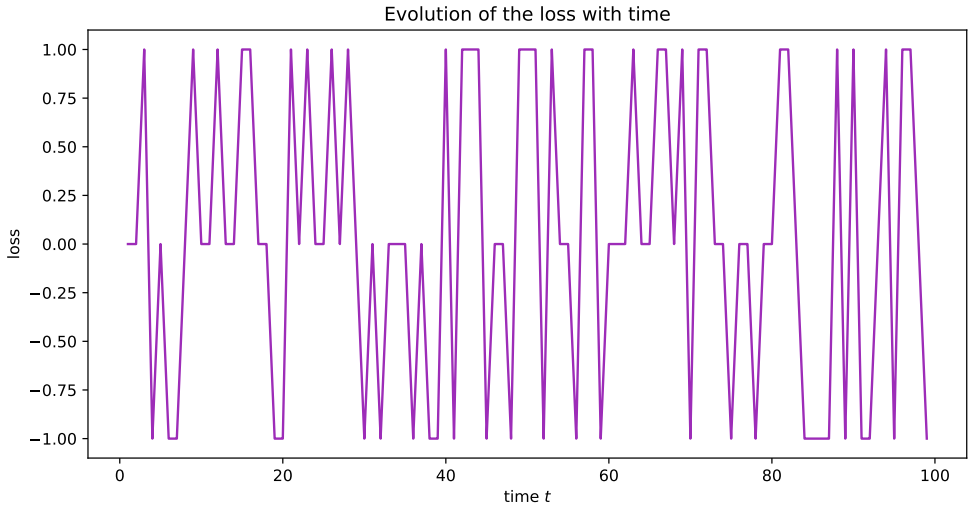

Figure 7: Loss incurred by player p over time for an instance of the game (`adaptative EWA`)

(b)

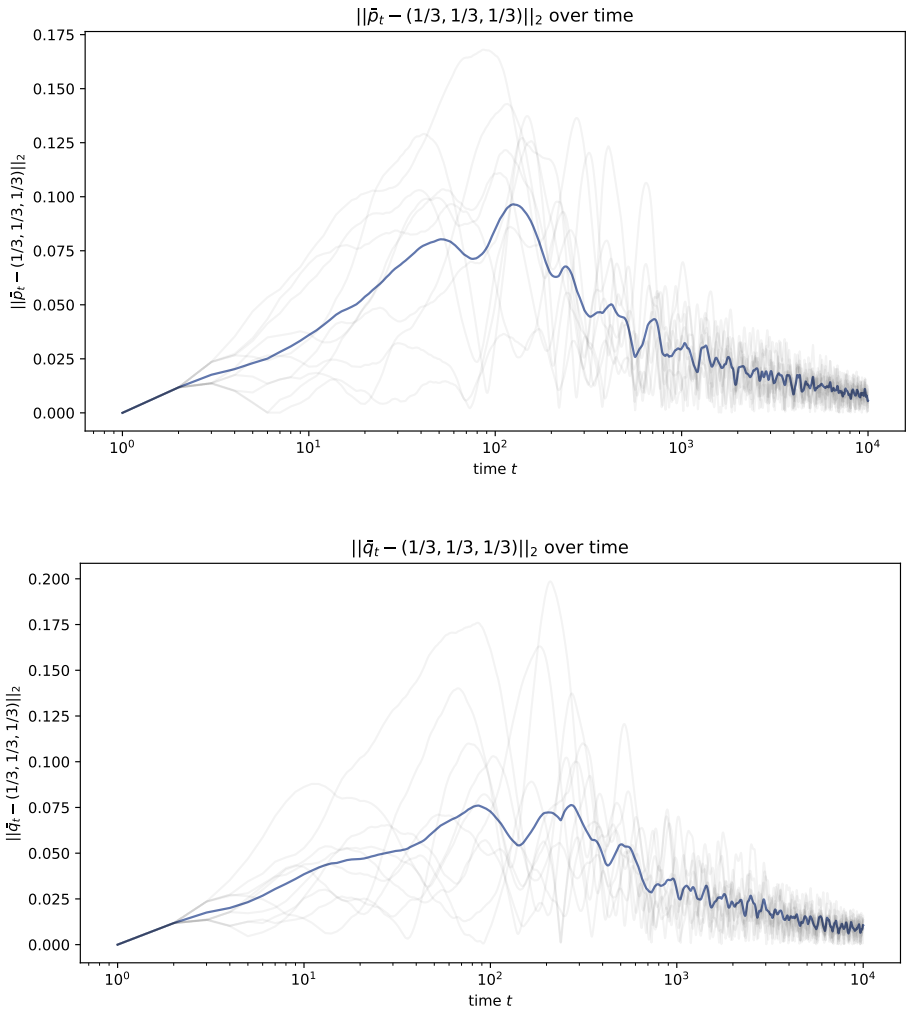

Figure 8: Evolution of $||\overline{p}_t - (1/3, 1/3, 1/3)||_2$ and $||\overline{q}_t - (1/3, 1/3, 1/3)||_2$ over time (`adaptative EWA`)

Experimentally, we see that $(\overline{p}_t, \overline{q}_t)$ converges to $(1/3, 1/3, 1/3)$.
This corroborates the fact that it is possible to show that $(\overline{p}_t, \overline{q}_t)$ converges almost surely to a Nash equilibrium of the game. This Nash equilibrium is $(1/3, 1/3, 1/3)$.

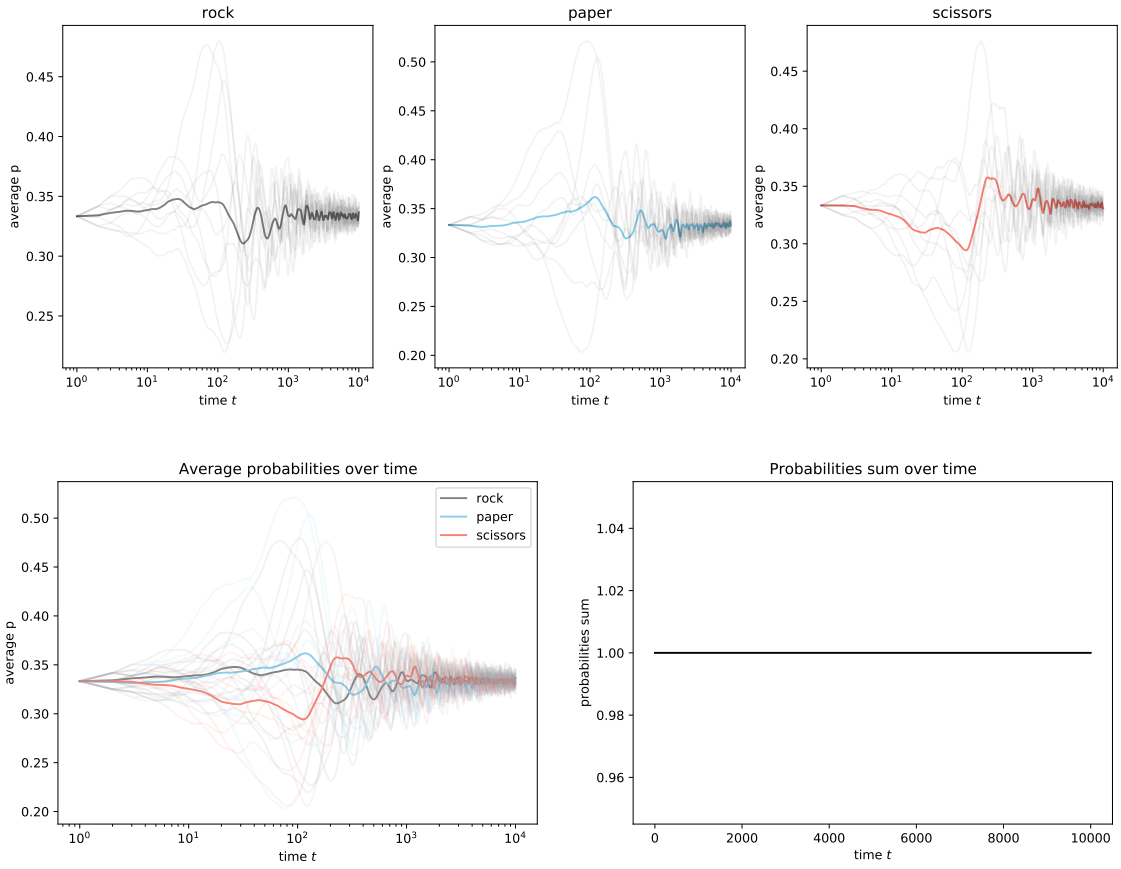

Figure 9: Evolution of $\left(\overline{p}_{\text{rock}}, \overline{p}_{\text{paper}}, \overline{p}_{\text{scissors}}\right)$ over time (`adaptative EWA`)

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 as one goes along.

5. (a) As we assume here to be in $[0, 1]$, we change the matrix L:

$$L = \begin{pmatrix} 0.5 & 1 & 0 \\ 0 & 0.5 & 1 \\ 1 & 0 & 0.5 \end{pmatrix}$$

(b) see code `homework.ipynb` available at: `https://github.com/clementgr/sequential-learning`

6. Fixed `EXP3`

(a) The loss $\ell_t(i_t)$ incurred by the player if he chooses action i_t at time t is given by $\ell_t(i_t) = L[i_t, j_t]$.

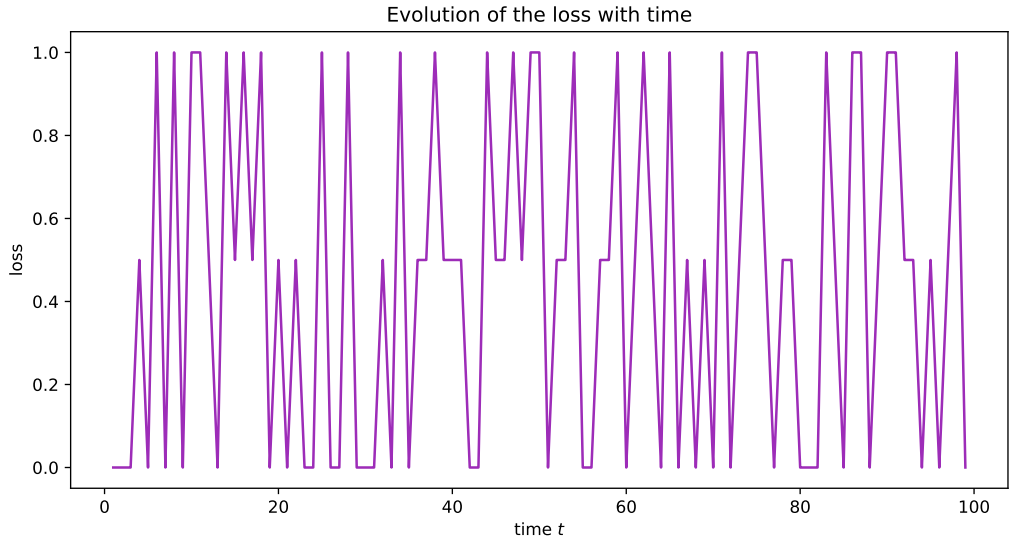

Figure 10: Loss incurred by player p over time for an instance of the game (`fixed EXP3`)

(b)

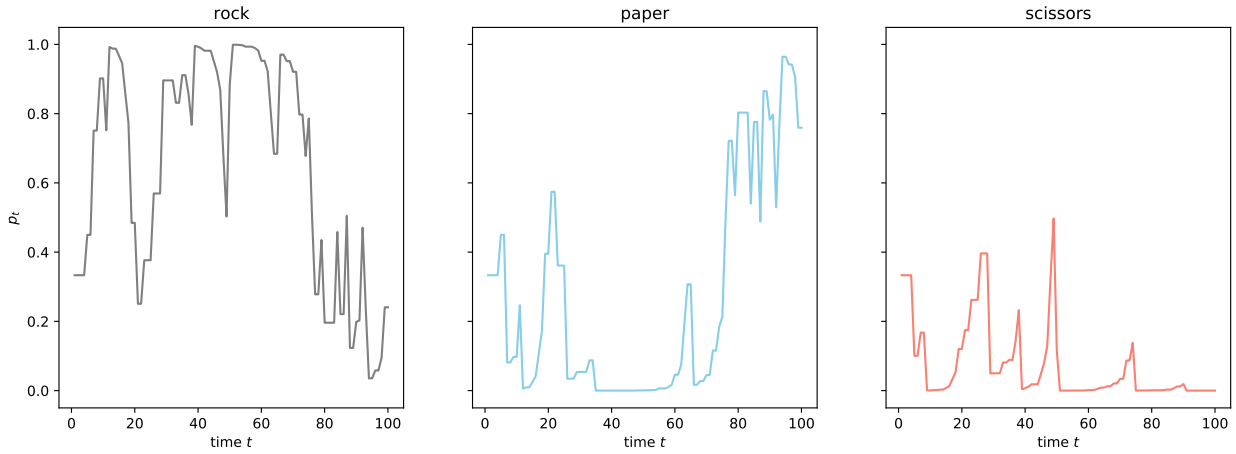

Figure 11: Evolution of the weight vectors $\left(p^{\text{rock}}, p^{\text{paper}}, p^{\text{scissors}}\right)$ over time (`fixed EXP3`)

Here again, against this fixed adversary, a good strategy is to not play "scissors" as – on average – the adversary plays "rock" every two games. Figure 2 shows that experiments support this is a good strategy.

(c)

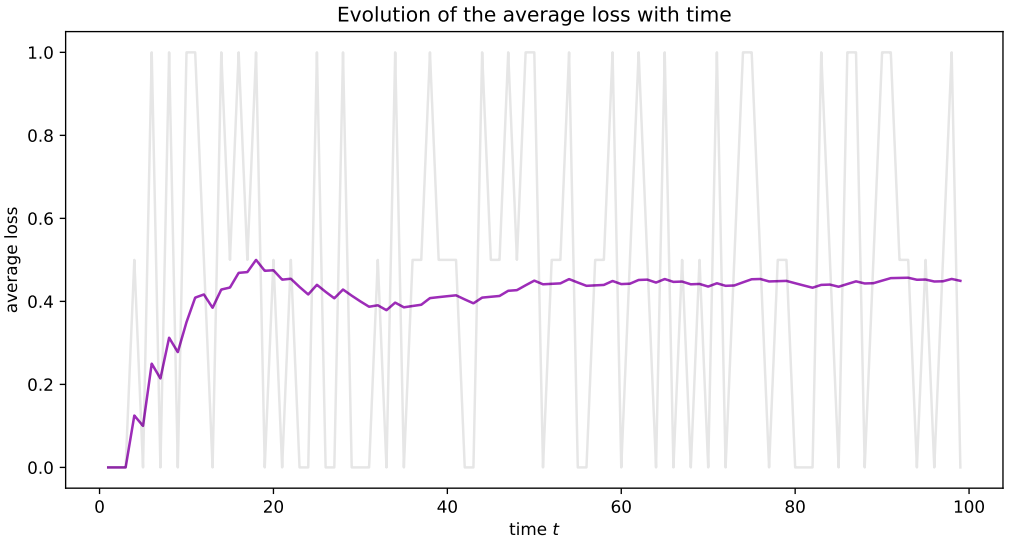

Figure 12: Average loss suffered by player p over time (`fixed EXP3`)

(d)

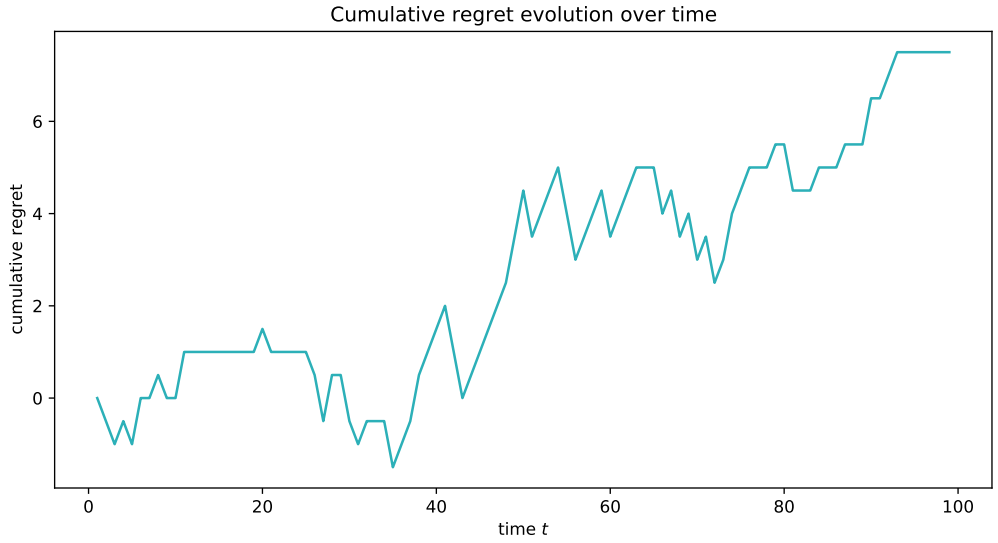

Figure 13: Cumulative regret of player p over time (`fixed EXP3`)

(e)

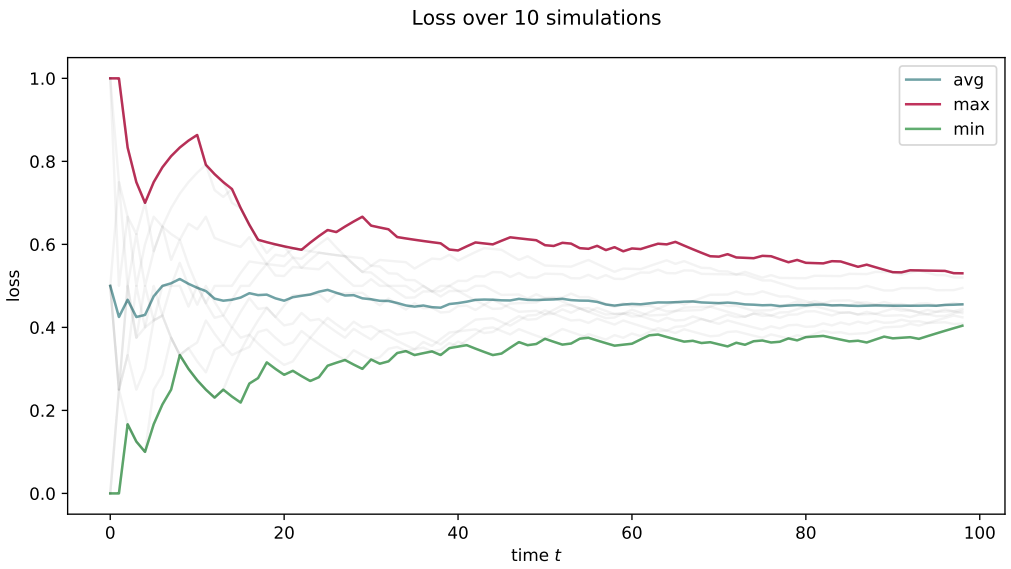

Figure 14: Average loss obtained in average, in maximum and in minimum over 10 simulations (`fixed EXP3`)

(f) Let's denote $\eta_1 = 0.01, \eta_2 = 0.05, \eta_3 = 0.1, \eta_4 = 0.5, \eta_5 = 1$.

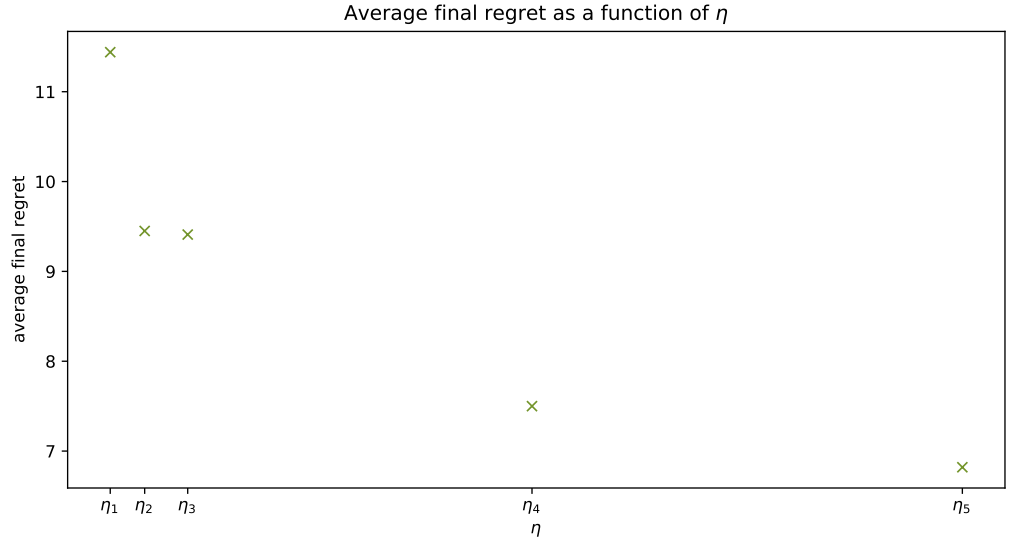

Figure 15: Final regret as a function of η (`fixed EXP3`)

As before, the best theoretical η is $\eta^\star_{\text{th}} = 0.105$ (for $K = 3$ and $T = 1000$).
In practice, however, the best η is again $\eta^\star_{\text{exp}} = 1.0$.

7. Adaptative `EXP3`

(a)

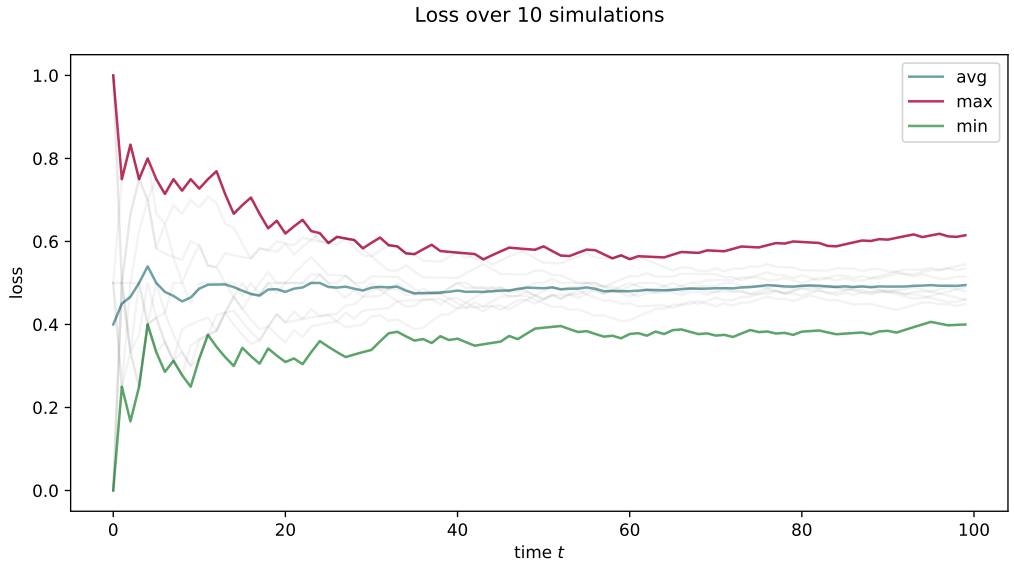

Figure 16: Average loss obtained in average, in maximum and in minimum over 10 simulations (`adaptative EXP3`)

(b)

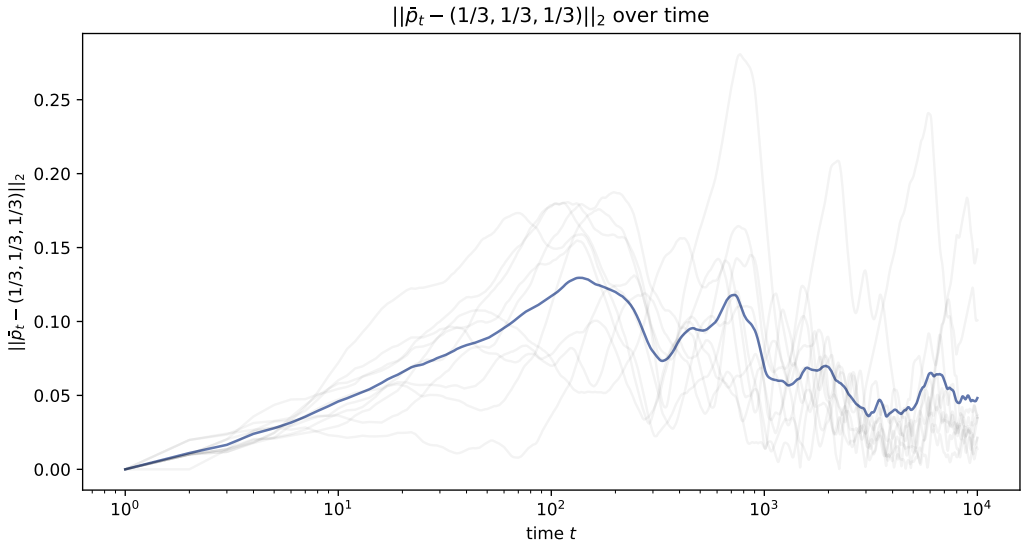

Figure 17: Evolution of $||\overline{p}_t - (1/3, 1/3, 1/3)||_2$ over time (`adaptative EXP3`)

This time, the experiments tend to show that $(\overline{p}_t, \overline{q}_t)$ doesn't seem converges to $(1/3, 1/3, 1/3)$ (or at least not before $T = 1e4$).

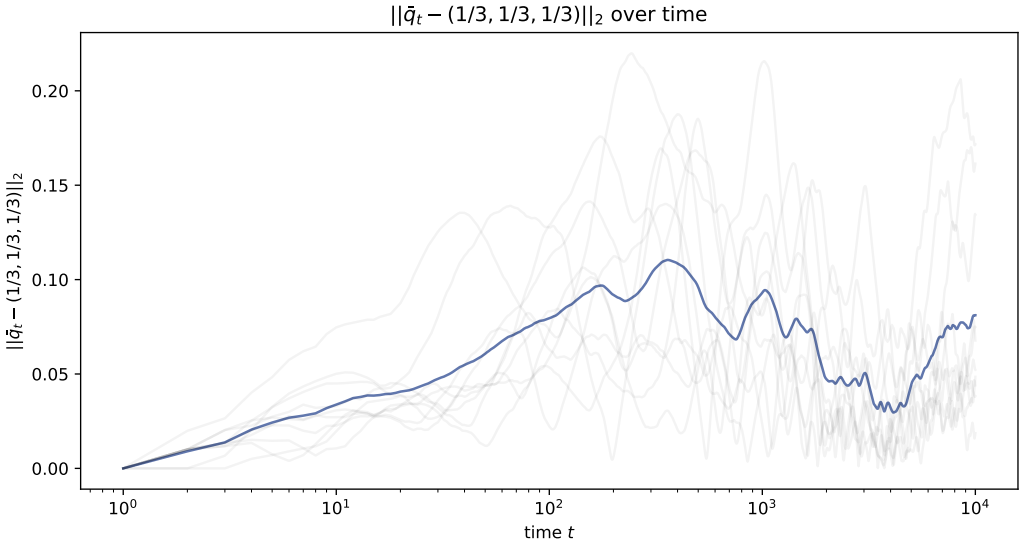

Figure 18: Evolution of $||\overline{q}_t - (1/3, 1/3, 1/3)||_2$ over time (`adaptative EXP3`)

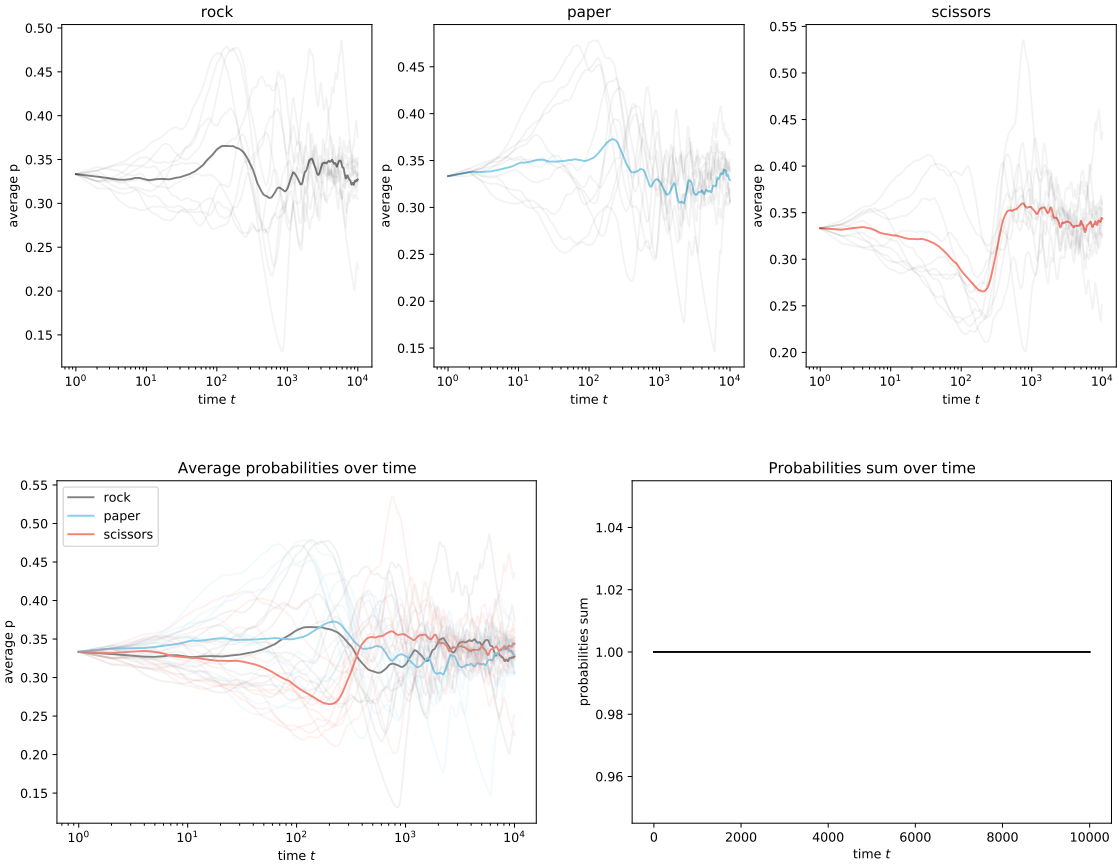

Figure 19: Evolution of $\left(\overline{p}_{\text{rock}}, \overline{p}_{\text{paper}}, \overline{p}_{\text{scissors}}\right)$ over time (`adaptative EXP3`)

Optional extentions The following questions are optional.

8.

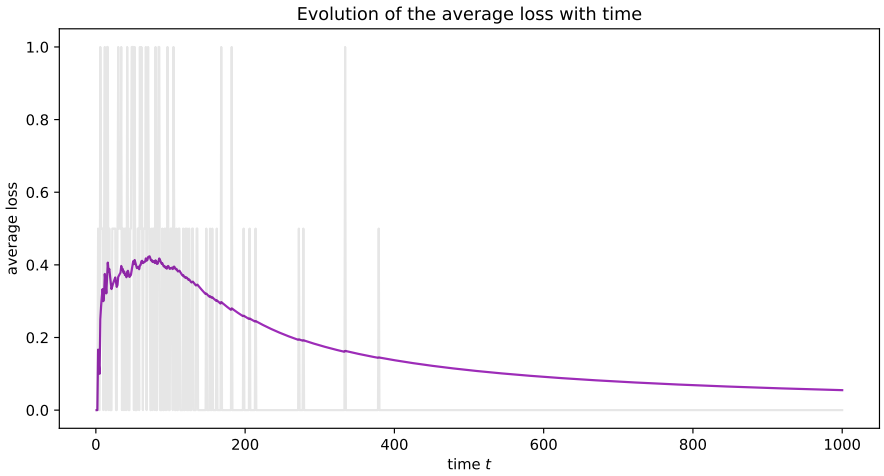

Figure 20: adaptative EXP3 vs. UCB

The loss of the `EXP3` player slowly converges to zero: `UCB` was winning at the beginning, but it didn't manage to remain on top. In the end, neither of the algorithm wins for $T >> 1$.

Part 2. Bernoulli Bandits

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

   (a) Let $K \geq 2$ the number of arms available. Assuming each arm $k \in \{1, \dots, K\}$ has a Bernoulli reward distribution $\mathcal{B}(p_k)$, we denote by $k^\star$ the best arm:

   $$k^\star = \text{argmax}_{k \in \{1,\dots,K\}}(p_k)$$

   Without loss of generalization, we can consider we pull each arm once before following the leader. Then, with probability $1 - p_{k^\star}$, the first pull of arm $k^\star$ gives a 0 reward.

   Now let's introduce the events $\mathcal{E}_k = \{\text{the first pull of arm } k \text{ gives a reward of } 1\}$ and consider:

   $$\mathcal{E} = \bigcup_{k \neq k^\star} \mathcal{E}_k$$

   By independency, we can write:

   $$\mathbb{P}(\mathcal{E}) = \sum_{k \neq k^\star} \mathbb{P}(\mathcal{E}_k) = \sum_{k \neq k^\star} p_k$$

Let's denote by $\Gamma_{k^\star}$ the event $\mathcal{E} \cap \overline{\mathcal{E}}_{k^\star}$. By independency:

$$\gamma_{k^\star} := \mathbb{P}\left(\Gamma_{k^\star}\right) = \mathbb{P}\left(\mathcal{E}\right) \cdot \mathbb{P}\left(\overline{\mathcal{E}}_{k^\star}\right) = \left(1 - p_{k^\star}\right) \sum_{k \neq k^\star} p_k$$

Hence, with probability $\gamma_{k^\star} > 0$ (if we exclude the degenerate case when $p_k = 1$ for any $k \in \{1, \dots, K\}$), for any of the following step t: $\hat{\mu}_t^{k^\star} = 0$ and there exists $k \neq k^\star$ such that $\hat{\mu}_t^k > 0$.

Under these circumstances, FTL never pulls arm $k^\star$ at time t. If we denote by $\Delta^\star = \min_{k \neq k^\star}\left(p_{k^\star} - p_k\right)$, then under $\mathrm{E}_{k\star} = \mathcal{E} \cap \overline{\mathcal{E}}_{k^\star}$, the regret increases by at least $\Delta^\star$ at each step. Therefore:

$$\begin{aligned}\mathbb{E}\left[R_T\right] = \mathbb{E}\Big[R_T \mathbb{1}_{\mathrm{E}_{k\star}}\Big] + \mathbb{E}\left[R_T \mathbb{1}_{\overline{\mathrm{E}}_{k\star}}\right] &\geq \mathbb{E}\Big[(T-K)\Delta^\star \mathbb{1}_{\mathrm{E}_{k\star}}\Big] \\ &\geq (T-K)\Delta^\star \mathbb{P}\left(\mathrm{E}_{k\star}\right)\end{aligned}$$

This shows the expected regret is lienar and verifies $\mathbb{E}\left[R_T\right] \geq \alpha T$ with $\alpha = \Delta^\star \mathbb{P}\left(\mathrm{E}_{k\star}\right)$.

(b) see code `homework.ipynb` available at: `https://github.com/clementgr/sequential-learning`

(c)

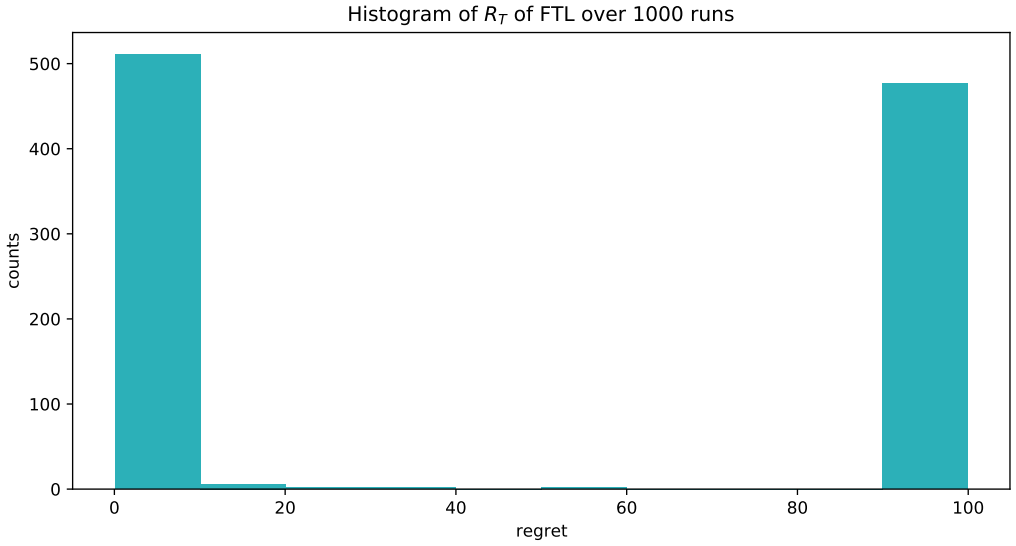

Figure 21: Histogram of the regret R_T of FTL over 1000 repetitions

The histogram displays two main bars. This is expected as:

– if the first pull of each arm gave a reward of 1 for the best arm and a reward of 0 for the worse: then FTL will keep pulling the best arm (question 1.a.) and suffer a null regret (leftmost bar on Figure 21)

– if the first pull of each arm gave a reward of 0 for the best arm and a reward of 1 for the worse: then FTL will keep pulling the worse arm (question 1.a.) and suffer a maximum regret ($R_T = 0.1 * 100 = 10$, rightmost bar on Figure 21)

– small bars between $R_T = 0$ and $R_T = 10$ represent the case where the player suffers a null regret only after a certain amount of time steps t': this happens when both arms gave a reward of 1 after the first pull, then during the following t' pulls, the worse arm's mean estimate was higher than the one of the best arm (because rewards are stochastics). At time t', the worse arm's mean estimate finally becomes lower than the one of the best arm (by law of large numbers).

(d)

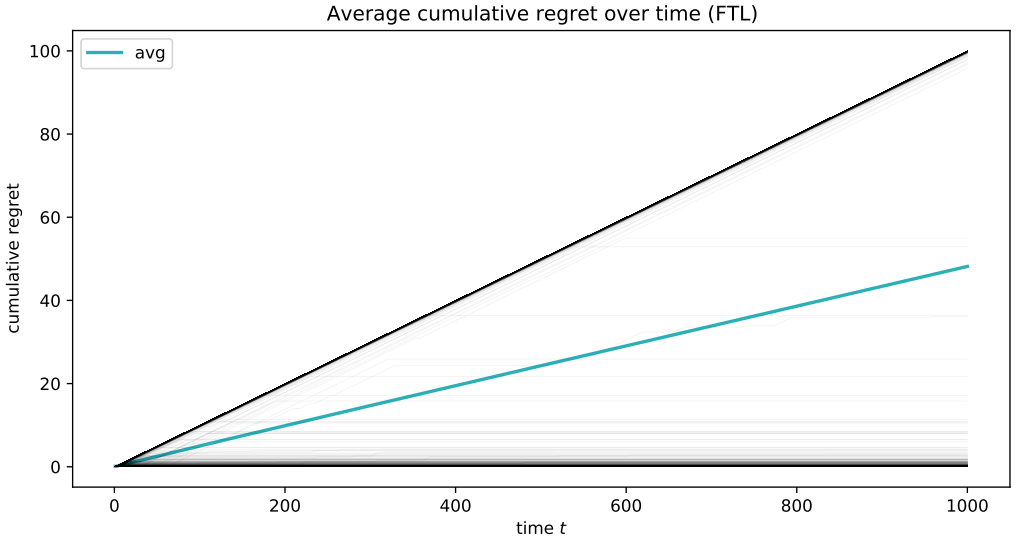

Figure 22: Average cumulative regret of FTL over 1000 repetitions

Figure 22 confirms what we've described in question 1.c. Indeed, we can distinguish two main trends:

– experiments with **null** regret: this happens when the first pull of each arm gave a reward of 1 for the best arm and a reward of 0 for the worse. In that case, FTL will keep pulling the best arm.

– experiments with **maximum** regret: this happens when the first pull of each arm gave a reward of 0 for the best arm and a reward of 1 for the worse. In that case, FTL will keep pulling the worse arm.

– rare experiments highlighted in the previous question.

In the end, the average cumulative reward of FTL over time is **linear** in T: the slope of the average cumulative reward is approximately the mean slop of the 2 edges cases highlighted above:

$$\mathbb{E}[R_T] \approx \left(\frac{p_{\max} - p_{\min}}{2} \right) T$$

Given $p = (0.5, 0.6)$, this gives:

$$\mathbb{E}[R_T] \approx 0.05 \cdot T$$

The numbers reported on Figure 22 confirm this approximation as $R_{1000} \approx 50$.

FTL is **not** a good algorithm for stochastic bandits because of the linear dependance of the cumulative regret in T.

2. *UCB*

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

$$\begin{aligned}
\Phi_X(\lambda) &= \log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \\
&= \log \mathbb{E}\left[e^{\lambda(X-p)}\right] \\
&= \log\left[pe^{\lambda(1-p)} + (1-p)e^{-\lambda p}\right] \\
&= \log\left[e^{-\lambda p}\left(pe^{\lambda} + (1-p)\right)\right] \\
&= -\lambda p + \log\left(pe^{\lambda} + (1-p)\right)
\end{aligned}$$

(b) If we manage to prove that $\Phi_X^{(2)}(\lambda) \leq \sigma^2$ for all $\lambda \in \mathbb{R}$ implies that $\Phi_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$ for all $\lambda \in \mathbb{R}$, we will get the desired result using that exp is an increasing function on $\mathbb{R}$.

Let $\lambda \in \mathbb{R}$ and let's denote φ the function $\varphi : \lambda \mapsto e^{\lambda(X-\mathbb{E}[X])}$. For any $k \in \mathbb{N}^*$, we have:

$$\frac{\partial^k \varphi(\lambda)}{\partial \lambda^k} = (X - \mathbb{E}[X])^k \, \varphi(\lambda)$$

As φ is in L^1, we easily see that for any $k \in \mathbb{N}^*, \varphi^{(k)}(\lambda)$ is in L^1 as well. We can safely swap integral and derivative. Hence:

$$\begin{aligned}
\Phi_X^{(1)}(\lambda) &= \frac{\partial \log \mathbb{E}\left[\varphi(\lambda)\right]}{\partial \lambda} \\
&= \frac{1}{\mathbb{E}\left[\varphi(\lambda)\right]} \cdot \frac{\partial \mathbb{E}\left[\varphi(\lambda)\right]}{\partial \lambda} \\
&= \frac{1}{\mathbb{E}\left[\varphi(\lambda)\right]} \cdot \mathbb{E}\left[\frac{\partial \varphi(\lambda)}{\partial \lambda}\right] \\
&= \frac{\mathbb{E}\left[(X - \mathbb{E}[X])\,\varphi(\lambda)\right]}{\mathbb{E}\left[\varphi(\lambda)\right]}.
\end{aligned} \tag{$\star$}$$

Let $u \in \mathbb{R}^+$. Using the non-decreasing property of the integral, $\Phi_X^{(2)}(\lambda) \leq \sigma^2$ for all $\lambda \in \mathbb{R}$ gives:

$$\begin{aligned}
&\int_0^u \Phi_X^{(2)}(t)dt \leq \int_0^u \sigma^2 dt \\
\Rightarrow \quad &\Phi_X^{(1)}(u) - \Phi_X^{(1)}(0) \leq \sigma^2 u
\end{aligned}$$

Using $(\star)$, we get that $\Phi_X^{(1)}(0) = \mathbb{E}\left[(X - \mathbb{E}[X])\right] = 0$. Hence, for any $u \in \mathbb{R}^+$:

$$\Phi_X^{(1)}(u) \leq \sigma^2 u \tag{$\star\star$}$$

Let $\lambda \in \mathbb{R}^+$. Once again, using the the non-decreasing property of the integral, $(\star\star)$ gives:

$$\begin{aligned} &\int_0^\lambda \Phi_X^{(1)}(t)dt \leq \int_0^\lambda \sigma^2 t dt \\ \Rightarrow \quad &\Phi_X(\lambda) - \Phi_X(0) \leq \frac{\sigma^2\lambda^2}{2} \end{aligned}$$

By definition of Φ_X, we have that $\Phi_X(0) = \log 1 = 0$. Hence, for any $\lambda \in \mathbb{R}^+$:

$$\Phi_X(\lambda) \leq \frac{\sigma^2\lambda^2}{2}$$

The exact same reasoning for any $v \in \mathbb{R}^-$, taking the first integral between v and 0, and the second one between λ and 0, gives, for any $\lambda \in \mathbb{R}^-$:

$$\Phi_X(\lambda) \leq \frac{\sigma^2\lambda^2}{2}$$

In the end, this proves that for any $\lambda \in \mathbb{R}$: $\Phi_X(\lambda) \leq \frac{\sigma^2\lambda^2}{2}$.

As stated previously, we use the fact that exp is non-decreasing on $\mathbb{R}$ and we finally get that:

$$\begin{aligned} e^{\Phi_X(\lambda)} &\leq e^{\frac{1}{2}\sigma^2\lambda^2} \\ \Leftrightarrow \quad \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] &\leq e^{\frac{1}{2}\sigma^2\lambda^2} \end{aligned}$$

which concludes the proof that X is a σ^2-sub-Gaussian random variable.

(c) Let $X \sim \mathcal{B}(p)$ with $p \in [0, 1]$. Using question 2.a, we have, for any $\lambda \in \mathbb{R}$:

$$\Phi_X(\lambda) = -\lambda p + \log\left(pe^\lambda + (1-p)\right)$$

Deriving this expression twice gives:

$$\begin{aligned} \Phi_X^{(2)}(\lambda) &= \frac{pe^\lambda\left[pe^\lambda + (1-p)\right] - \left(pe^\lambda\right)^2}{\left[pe^\lambda + (1-p)\right]^2} \\ &= \frac{pe^\lambda(1-p)}{\left[pe^\lambda + (1-p)\right]^2} \end{aligned}$$

Using that, for any $(a, b) \in \mathbb{R}^2$:

$$ab = \frac{1}{4}(a+b)^2 - \frac{1}{4}(a-b)^2$$

we get that $\frac{ab}{(a+b)^2} \leq \frac{1}{4}$. Hence, with $a = pe^{\lambda}$ and $b = (1-p)$:

$$\Phi_X^{(2)}(\lambda) \leq \frac{1}{4}$$

Now, using question 2.b, we can conclude that X is $\frac{1}{4}$-sub-Gaussian.

(d) Let's start by proving that for any $\lambda \in \mathbb{R}$, for any $x \in [0,1]$, $e^{\lambda x} \leq 1 - x + xe^{\lambda}$.
Let's denote h_λ the function defined as:

$$\begin{aligned} h_\lambda : \quad & [0,1] \to \mathbb{R} \\ & x \mapsto e^{\lambda x} - 1 + x - xe^{\lambda} \end{aligned}$$

We have:

$$h_\lambda^{(2)}(x) = \lambda^2 e^{\lambda x}$$

which is positive for any $\lambda \in \mathbb{R}$ and any $x \in [0,1]$. Hence, h_λ is **convex**. Using the convexity of h_λ, we can write, for any $x \in [0,1]$:

$$h_\lambda(x) \leq x\left(h_\lambda(1) - h_\lambda(0)\right) + h_\lambda(0)$$

Using $h_\lambda(0) = 0$ and $h_\lambda(1) = 0$, we get, for any $x \in [0,1]$: $h_\lambda(x) \leq 0$, which prove that for any $\lambda \in \mathbb{R}$, for any $x \in [0,1]$:

$$e^{\lambda x} \leq 1 - x + xe^{\lambda} \tag{E1}$$

Now, let X be a random variable supported on $[0,1]$ with mean $p \in [0,1]$ and $Y \sim \mathcal{B}(p)$.
For any $\lambda \in \mathbb{R}$, we get:

$$\begin{aligned} \Phi_X(\lambda) - \Phi_Y(\lambda) &= \log \mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] + \lambda p - \log\left(pe^{\lambda} + (1-p)\right) \\ &= \log \mathbb{E}\left[e^{\lambda(X-p)}\right] - \log\left[e^{-\lambda p}\left(pe^{\lambda} + (1-p)\right)\right] \\ &= \log \mathbb{E}\left[e^{-\lambda p}e^{\lambda X}\right] - \log\left[e^{-\lambda p}\left(pe^{\lambda} + (1-p)\right)\right] \\ &= \log\left(\frac{\mathbb{E}\left[e^{\lambda X}\right]}{(pe^{\lambda} + (1-p))}\right) \\ &= \log \mathbb{E}\left[\frac{e^{\lambda X}}{pe^{\lambda} + (1-p)}\right] \end{aligned}$$

As X is supported on $[0,1]$, we can use (E1) and write, for any $\lambda \in \mathbb{R}$: $e^{\lambda X} \leq 1 - X + Xe^{\lambda}$.
Hence:

$$\Phi_X(\lambda) - \Phi_Y(\lambda) \leq \log \mathbb{E}\left[\frac{1 - X + Xe^{\lambda}}{pe^{\lambda} + (1-p)}\right]$$

Eventually, using the linearity of the expectation and the fact that $\mathbb{E}[X] = p$, we get, for any $\lambda \in \mathbb{R}$:

$$\begin{aligned}\Phi_X(\lambda) - \Phi_Y(\lambda) &\leq \log\left(\frac{pe^{\lambda} + (1-p)}{pe^{\lambda} + (1-p)}\right) \\ &\leq 0\end{aligned}$$

which ends the proof.

(e) Let X be a random variable supported on $[0, 1]$. Then $\mathbb{E}[X] \in [0, 1]$.
Let's denote $\mathbb{E}[X]$ by p, and let's consider $Y \sim \mathcal{B}(p)$. Using question 2.b. we get that, for any $\lambda \in \mathbb{R}$:

$$\Phi_X(\lambda) \leq \Phi_Y(\lambda)$$

Hence, using the fact that exp is non-decreasing on $\mathbb{R}$, we can write for any $\lambda \in \mathbb{R}$:

$$\begin{aligned} & e^{\Phi_X(\lambda)} \leq e^{\Phi_Y(\lambda)} \\ \Leftrightarrow \quad & \mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leq \mathbb{E}\left[e^{\lambda(Y - \mathbb{E}[Y])}\right]\end{aligned}$$

Finally, using question 2.c., we know that $Y \sim \mathcal{B}(p)$ is $\frac{1}{4}$-sub-Gaussian:

$$\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leq e^{\frac{1}{2}\left(\frac{1}{4}\right)^2\lambda^2}$$

This proves that X is $\frac{1}{4}$-sub-Gaussian. Thus, all random variables supported on $[0, 1]$ are sub-Gaussian.

(f) see code `homework.ipynb` available at: `https://github.com/clementgr/sequential-learning`

(g)

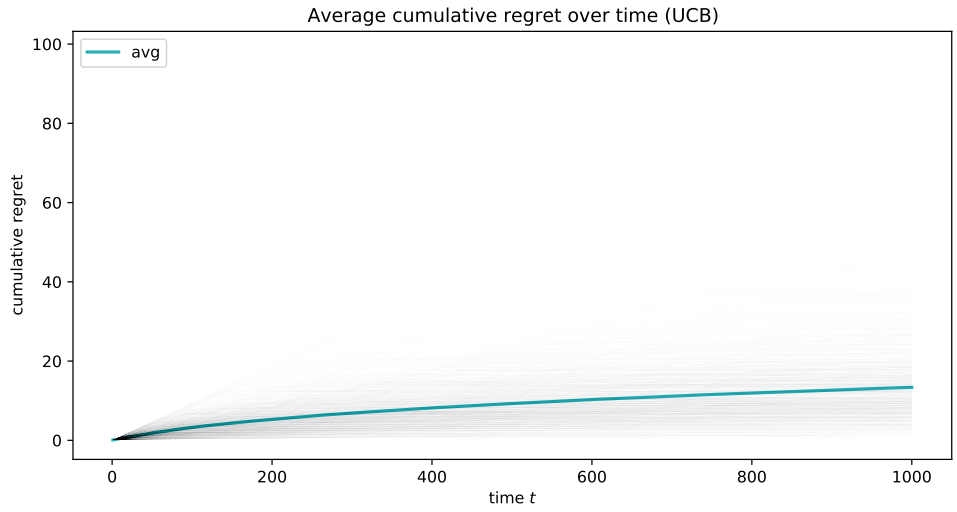

Figure 23: UCB($\frac{1}{4}$)

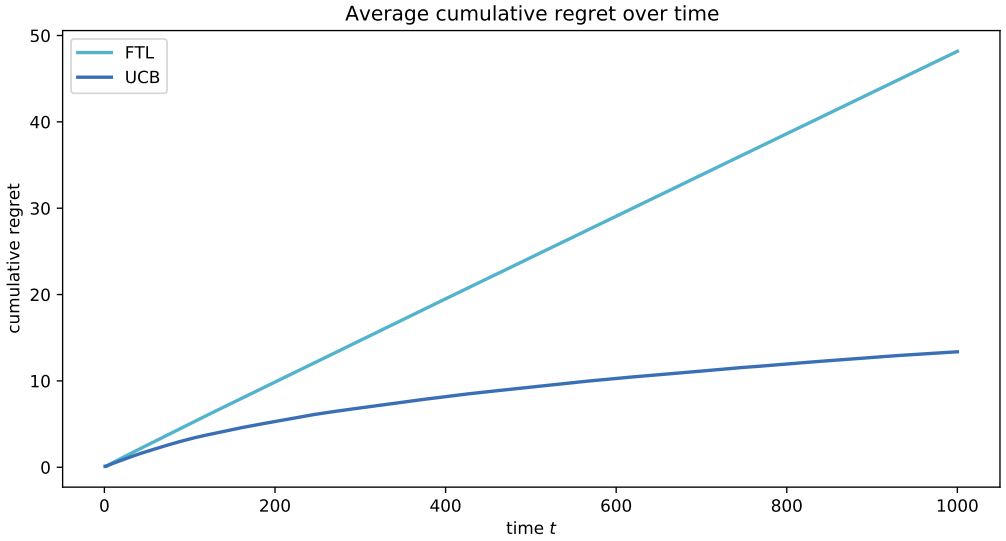

Figure 24: FTL vs. UCB($\frac{1}{4}$)

UCB performs much better than FTL, as can be seen on Figure 24. It gets rid of the linear dependence of the cumulative regret in T.

(h) Let's denote $\sigma_1 = 0, \sigma_2 = \frac{1}{\sqrt{32}}, \sigma_3 = \frac{1}{4}, \sigma_4 = \frac{1}{2}, \sigma_5 = 1$.

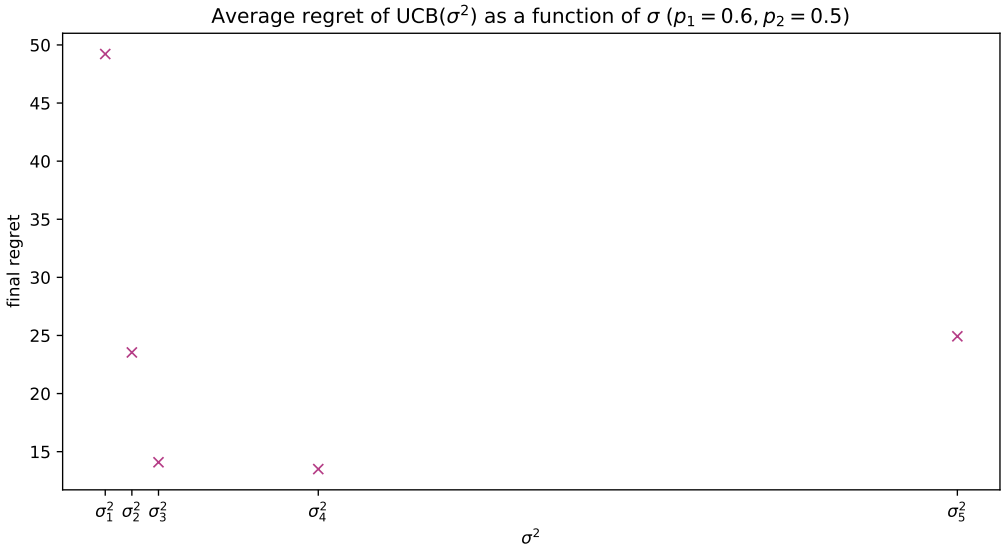

Figure 25: Average regret of UCB(σ^2) as a function of the parameter σ^2 for $p = (0.6, 0.5)$

For $p = (0.6, 0.5)$, the optimal parameter σ^2 is $(\sigma_4)^2 = \frac{1}{4}$.
Given the appearance of Figure 25, there seems to be a trade-off to be found between small and high values of σ^2.

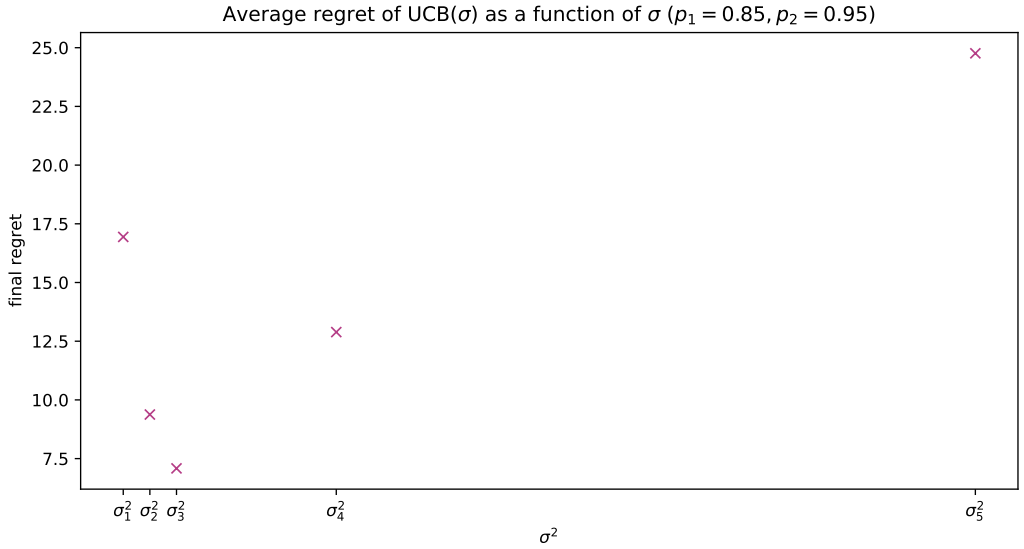

Figure 26: Average regret of UCB(σ^2) as a function of the parameter σ^2 for $p = (0.85, 0.95)$

For $p = (0.85, 0.95)$, the optimal parameter σ^2 **did change**: it's $(\sigma_3)^2 = \frac{1}{16}$.
Here again, Figure 26 reveals there exists a trade-off between small and high values of σ^2.

In theory, the optimal parameter σ^2 equals the maximum of the arms' variances. We thus expect the optimal σ^2 to be smaller when $p = (0.85, 0.95)$ than when $p = (0.6, 0.5)$. Figures 25 and 26 confirm this.

3. $\mathcal{B}(p)$ has variance $p(1-p)$.

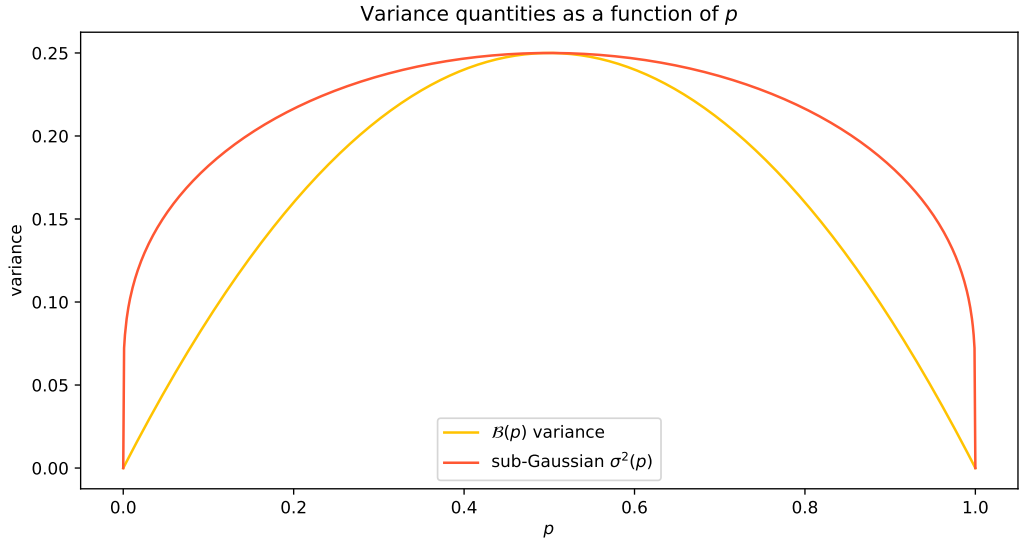

Figure 27: $\mathcal{B}(p)$ variance and $\sigma^2(p)$ sub-Gaussian constant as a function of p

4. **(optional)** (not treated)

Adaptation to the variance. The algorithm $\texttt{UCB}(\sigma^2)$ 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. All experiments in this question were done for $b = 1$, $\xi = 1.2$ and $c = 1$.

   (a) By definition of the empirical variance of the arms, we can write:

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

   By definition of the empirical mean, we can write:

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

   and:

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

   Hence, we have:

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

   which ends the proof.

   (b) Previous computations showed that:

$$N_t^k \hat{v}_t^k = \sum_{s=1}^{t-1} \mathbb{1}_{k_s=k} \left(X_s^{k_s} \right)^2 - N_t^k \left(\hat{\mu}_t^k \right)^2$$

Hence:

$$\begin{aligned} N_t^{k_t}\hat{v}_t^{k_t} - N_{t-1}^{k_t}\hat{v}_{t-1}^{k_t} &= \sum_{s=1}^{t-1}\mathbb{1}_{k_s=k_t}\left(X_s^{k_s}\right)^2 - N_t^{k_t}\left(\hat{\mu}_t^{k_t}\right)^2 - \sum_{s=1}^{t-2}\mathbb{1}_{k_s=k_t}\left(X_s^{k_s}\right)^2 + N_{t-1}^{k_t}\left(\hat{\mu}_{t-1}^{k_t}\right)^2 \\ &= \left(X_t^{k_t}\right)^2 + N_{t-1}^{k_t}\left(\hat{\mu}_{t-1}^{k_t}\right)^2 - N_t^{k_t}\left(\hat{\mu}_t^{k_t}\right)^2 \end{aligned} \tag{E2}$$

By definition, we have that $N_t^{k_t} = N_{t-1}^{k_t} + 1$.
Hence:

$$\hat{\mu}_t^{k_t} = \frac{1}{N_{t-1}^{k_t}+1}\left[N_{t-1}^{k_t}\hat{\mu}_{t-1}^{k_t} + X_t^{k_t}\right]$$

And:

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

Therefore, we have:

$$\begin{aligned} N_{t-1}^{k_t}\left(\hat{\mu}_{t-1}^{k_t}\right)^2 &= \frac{1}{N_{t-1}^{k_t}}\left[\left(N_{t-1}^{k_t}+1\right)\hat{\mu}_t^{k_t} - X_t^{k_t}\right]^2 \\ &= \hat{\mu}_{t-1}^{k_t}\left[\left(N_{t-1}^{k_t}+1\right)\hat{\mu}_t^{k_t} - X_t^{k_t}\right] \\ &= \left(N_{t-1}^{k_t}+1\right)\hat{\mu}_{t-1}^{k_t}\hat{\mu}_t^{k_t} - X_t^{k_t}\hat{\mu}_{t-1}^{k_t} \end{aligned}$$

and we also have:

$$\begin{aligned} N_t^{k_t}\left(\hat{\mu}_t^{k_t}\right)^2 &= \frac{1}{N_{t-1}^{k_t}+1}\left[N_{t-1}^{k_t}\hat{\mu}_{t-1}^{k_t} + X_t^{k_t}\right]^2 \\ &= \hat{\mu}_t^{k_t}\left[N_{t-1}^{k_t}\hat{\mu}_{t-1}^{k_t} + X_t^{k_t}\right] \\ &= N_{t-1}^{k_t}\hat{\mu}_t^{k_t}\hat{\mu}_{t-1}^{k_t} + X_t^{k_t}\hat{\mu}_t^{k_t} \end{aligned}$$

Hence: $N_{t-1}^{k_t}\left(\hat{\mu}_{t-1}^{k_t}\right)^2 - N_t^{k_t}\left(\hat{\mu}_t^{k_t}\right)^2 = \hat{\mu}_{t-1}^{k_t}\hat{\mu}_t^{k_t} + X_t^{k_t}\left(\hat{\mu}_t^{k_t} - \hat{\mu}_{t-1}^{k_t}\right)$

Eventually, putting everything back into (E2):

$$\begin{aligned} N_t^{k_t}\hat{v}_t^{k_t} - N_{t-1}^{k_t}\hat{v}_{t-1}^{k_t} &= \left(X_t^{k_t}\right)^2 + \hat{\mu}_{t-1}^{k_t}\hat{\mu}_t^{k_t} + X_t^{k_t}\left(\hat{\mu}_t^{k_t} - \hat{\mu}_{t-1}^{k_t}\right) \\ &= \left(X_t^{k_t} - \hat{\mu}_{t-1}^{k_t}\right)\left(X_t^{k_t} - \hat{\mu}_t^{k_t}\right) \end{aligned} \tag{E3}$$

which ends the proof.

This formulation is particularly advantageous for reducing computational complexity of UCB-V. Indeed, for any $k \in \{1, \dots K\}$, we can compute $\hat{v}_t^k$ directly from $\hat{v}_{t-1}^k$ using (E3). It only requires to store not only the current estimate $\hat{\mu}_t^k$ like is usually done in UCB, but also the previous estimate $\hat{\mu}_{t-1}^{k_t}$. This saves precious computation time compared to computing $\hat{v}_t^k$ from scratch.

(c) see code `homework.ipynb` available at: `https://github.com/clementgr/sequential-learning`

(d) We've seen in question 2.h. that the best parameter σ^2 for UCB when $p = (0.6, 0.5)$ is $\sigma^2 = \frac{1}{4}$. Hence, we will use this value for σ^2 when comparing UCB to UCB-V.

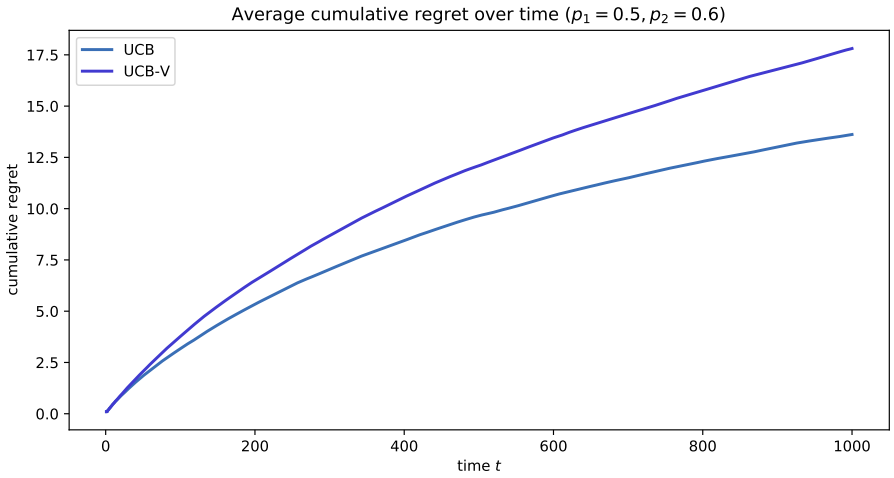

Figure 28: UCB-V vs. UCB($\frac{1}{4}$)

Interestingly, UCB($\frac{1}{4}$) performs slightly better than UCB-V for $p = (0.6, 0.5)$

(e)

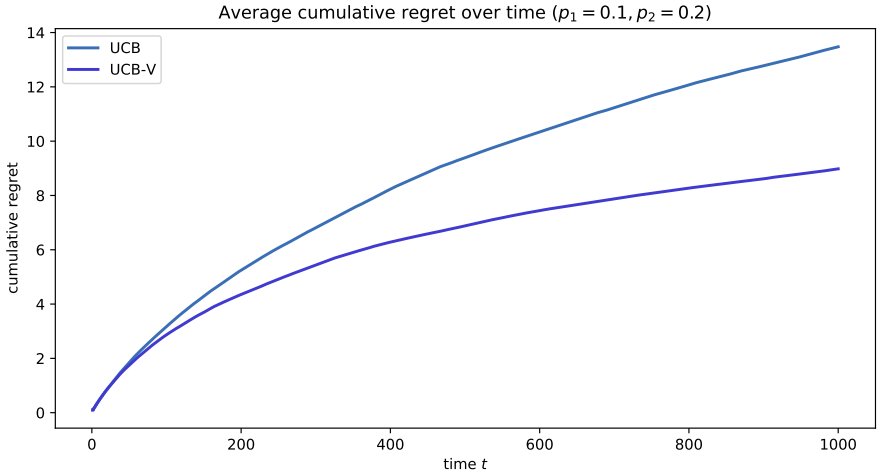

Figure 29: UCB-V vs. UCB($\frac{1}{4}$)

When $p = (0.1, 0.2)$, UCB-V performs slightly better than UCB.

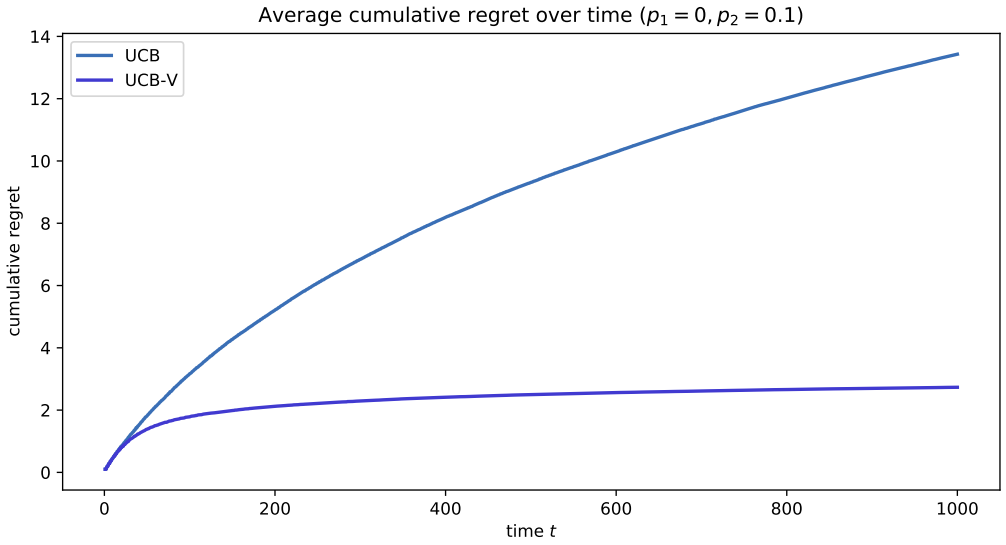

Figure 30: UCB-V vs. UCB($\frac{1}{4}$)

When $p = (0, 0.1)$, UCB-V performs significantly better than UCB.

6. **(optional)** see code `homework.ipynb` available at: `https://github.com/clementgr/sequential-learning`.

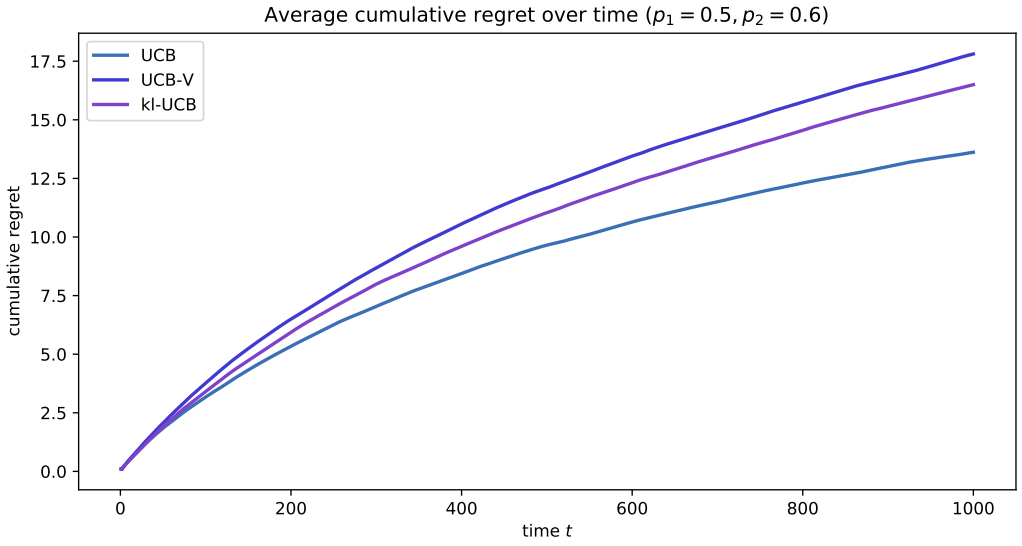

Figure 31: kl-UCB vs. UCB($\frac{1}{4}$) vs. UCB-V on various Bernoulli bandit problems

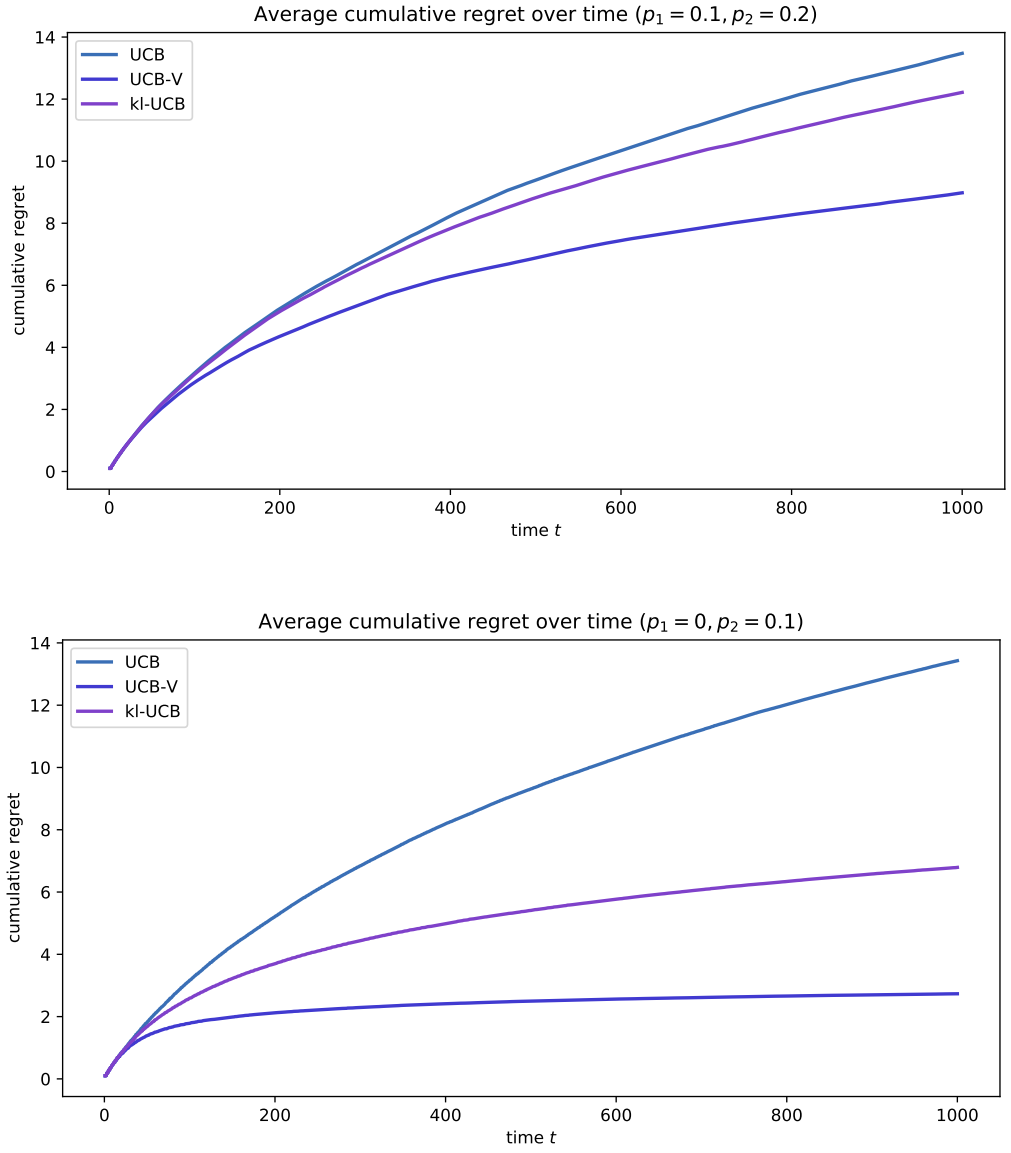

Figure 32: kl-UCB vs. UCB($\frac{1}{4}$) vs. UCB-V on various Bernoulli bandit problems

These plots show that:

- for $p = (0.6, 0.5)$ kl-UCB performs slightly worse than UCB($\frac{1}{4}$) and slightly better than UCB-V
- for $p = (0.1, 0.2)$ kl-UCB performs slightly better than UCB($\frac{1}{4}$) and slightly worse than UCB-V.
- for $p = (0, 0.1)$ kl-UCB performs better than UCB($\frac{1}{4}$) and worse than UCB-V.