Sequential Learning - Home Assignment

Etienne GAUTHIER

1 Bernoulli Bandits

In what follows, $K = 2$ and $p = (0.5, 0.6)$ unless stated otherwise.

1. (a) Let $\mathcal{E}_{i,j} := \{X_1(1) = i, X_2(2) = j\}$ for $i, j \in \{0, 1\}$. We have:

$$\begin{aligned}\mathbb{E}[R_T] &= \mathbb{E}[\mathbb{E}[R_T \mid X_1(1), X_2(2)]] \\ &= \mathbb{E}[R_T \mid \mathcal{E}_{1,0}]\mathbb{P}(\mathcal{E}_{1,0}) + \sum_{\substack{i,j\in\{0,1\} \\ (i,j)\neq(1,0)}} \mathbb{E}[R_T \mid \mathcal{E}_{i,j}]\mathbb{P}(\mathcal{E}_{i,j}).\end{aligned}$$

The second term is positive. To lower bound the first term, we note that if $\mathcal{E}_{1,0}$ holds then the player always selects arm 1 because $\hat{\mu}_1(t) > 0$ and $\hat{\mu}_2(t) = 0$ for $t = 3, ..., T$. Therefore, by independance of $X_1(1)$ and $X_2(2)$:

$$\begin{aligned}\mathbb{E}[R_T] &\geq (T \times 0.6 - T \times 0.5) \times 0.5 \times (1 - 0.6) \\ &\geq \alpha T\end{aligned}$$

for some $\alpha > 0$.

(b) (c)

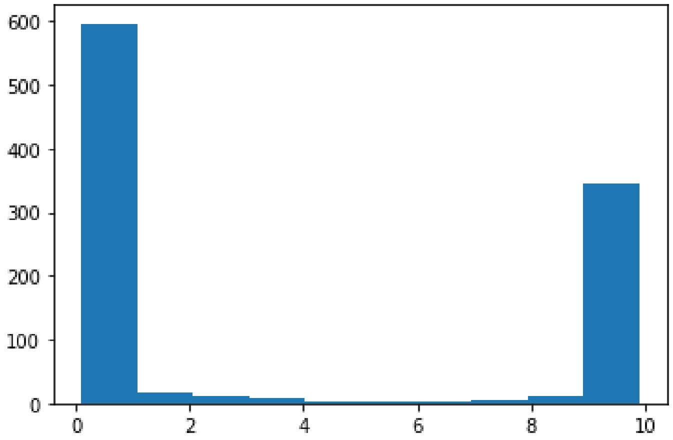


Figure 1: Histogram of regrets R_T for $T = 100$ and 1000 repetitions.

We note that FTL has a very high momentum when the two random variables have means close to each other: once an arm is selected after a few rounds, it is very unlikely that the player will select the other arm later. This is because if the player selects one of the two arms after a few rounds, it means that during the first rounds the outcomes of the other arm were low. Therefore, the empirical mean of this other arm is low. So the player is unlikely to select it again in the future.
Most of the time, the player always selects the arm with with the higher mean after a few rounds ; this results in a very low regret.
Still, it is common to have a regret which is almost equal to $10 = 100 \times (0.6 - 0.5)$. This happens when the outcomes of arm 1 are mostly equal to 0 during the first rounds. So the regret is almost equal to 10 because in this case the player will mostly select arm 2, except for a few rounds at the beginning of the experiment.
We also note that, in a few experiments, FTL yields a regret which is neither close to 0 nor to 10. This happens rarely, namely when after a few rounds the player starts selecting the same arm, but then changes. This can happen in extreme situations where for example the player always selects arm 1 after a few rounds but the outcomes obtained with arm 1 turn out to be almost all equal to 0 (which is very rare) and so the empirical mean of arm 1 gets lower and lower, until it becomes lower than the empirical mean of arm 2, and so the player finally starts selecting arm 2.

(d)

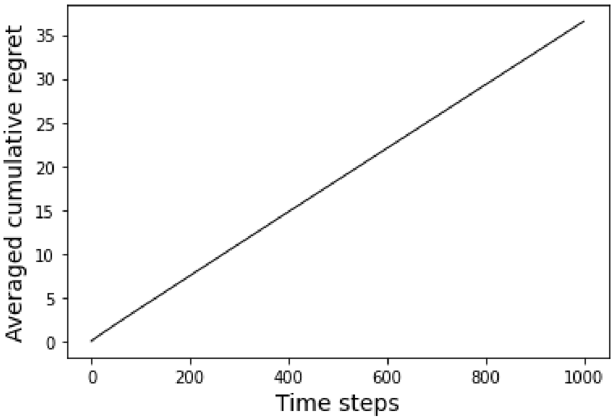


Figure 2: Regret as a function of time with FTL, averaged over 1000 episodes.

As we can see the mean regret is not sublinear. This is coherent with question 1.(a). Therefore FTL is not a good algorithm for stochastic bandits in our setting.

2. (a) Let X be a Bernoulli random variable with parameter p. The cumulant generating function of X is given by:

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

for all $\lambda \in \mathbb{R}$.

(b) Let X be a random variable such that $\phi''_X(\lambda) \leq \sigma^2$ for all $\lambda \in \mathbb{R}$. We have $\phi_X(0) = 0$ and $\phi'_X(\lambda) = \mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}(X - \mathbb{E}[X])]$ so $\phi'_X(0) = 0$, hence the fundamental theorem of calculus yields

$$\begin{aligned}\phi_X(\lambda) &= \int_0^\lambda \int_0^u \phi''_X(t)dtdu \\ &\leq \int_0^\lambda u\sigma^2 du \\ &= \frac{1}{2}\sigma^2\lambda^2.\end{aligned}$$

So X is σ^2-sub-Gaussian.

(c) Let X be a Bernoulli random variable with parameter $p \in [0,1]$. First, we assume $p \neq 0$ and $p \neq 1$. We have

$$\phi''_X(\lambda) = \frac{\frac{1-p}{pe^\lambda}}{(1+\frac{1-p}{pe^\lambda})^2}.$$

Let us prove that $\phi''_X(\lambda) \leq (\frac{1}{2})^2$ for all $\lambda \in \mathbb{R}$. The function

$$\begin{aligned} f\colon \mathbb{R} &\longrightarrow]0,+\infty[\\ \lambda &\longmapsto \frac{1-p}{pe^\lambda}\end{aligned}$$

is bijective since $p \neq 0$ and $p \neq 1$. So it is sufficient to upper bound the values of the functionw

$$\begin{aligned} g\colon]0,+\infty[&\longrightarrow \mathbb{R} \\ x &\longmapsto \frac{x}{(1+x)^2}.\end{aligned}$$

By derivating g we easily see that the maximum of g occurs when $x = 1$ and it is equal to $\frac{1}{4}$. So by question 2.(b) we deduce that all Bernoulli random variables with parameter $p \neq 0$ and $p \neq 1$ are $(\frac{1}{2})^2$-sub-Gaussian. This is also trivially true when $p = 0$ or $p = 1$, so finally we conclude that all Bernoulli random variables are $(\frac{1}{2})^2$-sub-Gaussian.

(d) Let X be a random variable supported on $[0,1]$ with mean $p \in [0,1]$, and let Y be a Bernoulli random variable with parameter p. First note that $e^{\lambda x} \leq 1 - x + xe^\lambda$ for all $x \in [0,1]$ and $\lambda \in \mathbb{R}$. This can be easily seen by derivating $f_x(\lambda) = 1 - x + xe^\lambda - e^{\lambda x}$ for a fixed $x \in [0,1]$. Indeed, f_x has a minimum in 0 which is equal to 0. Therefore we have

$$\begin{aligned}\int_0^1 e^{\lambda x}\mathbb{P}(X = x)dx &\leq \int_0^1 (1 - x + xe^\lambda)\mathbb{P}(X = x)dx \\ &= 1 - p + pe^\lambda\end{aligned}$$

so $e^{-\lambda p}\int_0^1 e^{\lambda x}\mathbb{P}(X=x)dx \leq e^{-\lambda p}(1-p+pe^{\lambda})$ for all $\lambda \in \mathbb{R}$. By taking the log and using question 2.(a) we deduce that $\phi_X(\lambda) \leq \phi_Y(\lambda)$ for all $\lambda \in \mathbb{R}$.

(e) By question 2.(d) and 2.(c) we deduce that all random variables supported on $[0,1]$ are $\frac{1}{4}$-sub-Gaussian.

(f) (g)

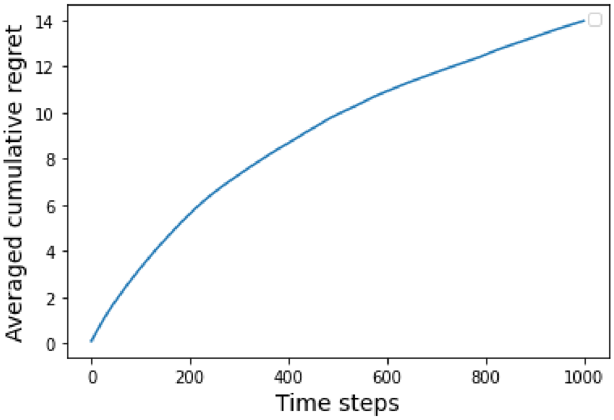
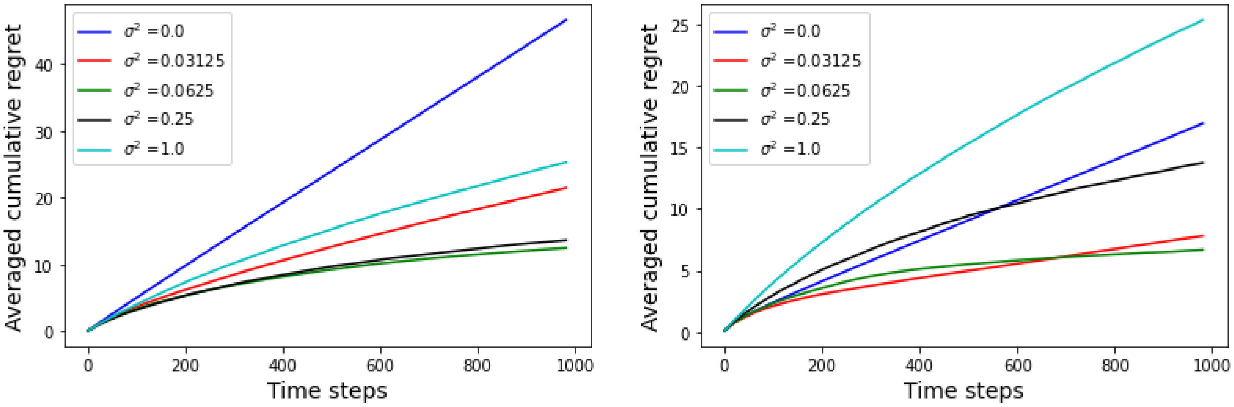


Figure 3: Regret as a function of time with UCB(1/4), averaged over 1000 episodes.

With UCB(1/4), the mean regret is sublinear contrary to FTL, cf question 1.(d).

(h)



(a) $p=(0.5,0.6)$

(b) $p=(0.85,0.95)$

Figure 4: Regret as a function of time with UCB(σ^2), averaged over 1000 episodes.

In the long term, the optimal parameter remains the same, and it is equal to $\sigma^2=\frac{1}{16}$.

3. For

$$\sigma^2(p) = \begin{cases} 0 & \text{if } p \in \{0,1\} \\ \frac{1}{4} & \text{if } p = \frac{1}{2} \\ \frac{1}{2}\frac{p-(1-p)}{\log p - \log(1-p)} & \text{otherwise} \end{cases}$$

we have the following figure:

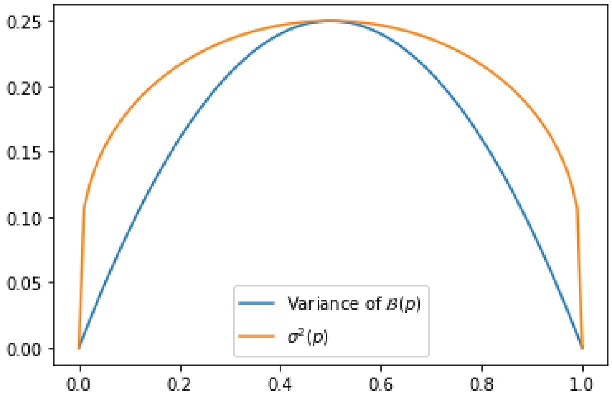


Figure 5: Variance of $\mathcal{B}(p)$ and $\sigma^2(p)$.

4. Let X be a σ^2-sub-Gaussian variable. We define $f(\lambda) = \mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}]$ and $g(\lambda) = e^{\frac{1}{2}\sigma^2\lambda^2}$ for all $\lambda \in \mathbb{R}$. By definition we have $f \leq g$. Now, f and g are both (at least) twice differentiable so by Taylor's theorem:

$$f(\lambda) = f(0) + \frac{f'(0)}{1!}(\lambda - 0) + \frac{f''(0)}{2!}(\lambda - 0)^2 + o_{\lambda \to 0}(\lambda^2)$$

and

$$g(\lambda) = g(0) + \frac{g'(0)}{1!}(\lambda - 0) + \frac{g''(0)}{2!}(\lambda - 0)^2 + o_{\lambda \to 0}(\lambda^2).$$

It is immediate to check that $f(0) = g(0) = 1$, $f'(0) = g'(0) = 0$, $f''(0) = \text{Var}(X)$ and $g''(0) = \sigma^2$. Since $f - g \leq 0$ we have

$$\frac{1}{\lambda^2}(f(\lambda) - g(\lambda)) \leq 0$$

for all $\lambda \neq 0$ and so by taking $\lambda \underset{\neq}{\longrightarrow} 0$ we conclude that $\text{Var}(X) \leq \sigma^2$.

5. (a) We have:

$$\begin{aligned}
N_k(t)\hat{v}_k(t) &= \sum_{s=1}^{t} \mathbb{1}_{a_s=k}(X_{a_s}(s) - \hat{\mu}_k(t))^2 \\
&= \sum_{s=1}^{t} \mathbb{1}_{a_s=k} \left(X_{a_s}(s) - \frac{1}{N_k(t)} \sum_{i=1}^{t} \mathbb{1}_{a_i=k} X_{a_i}(i) \right)^2 \\
&= \sum_{s=1}^{t} \mathbb{1}_{a_s=k} X_{a_s}(s)^2 - \frac{2}{N_k(t)} \sum_{s=1}^{t} 1_{a_s=k} X_{a_s}(s) \sum_{i=1}^{t} \mathbb{1}_{a_i=k} X_{a_i}(i) \\
&+ \frac{1}{N_k(t)^2} \sum_{s=1}^{t} \mathbb{1}_{a_s=k} \left(\sum_{i=1}^{t} \mathbb{1}_{a_i=k} X_{a_i}(i) \right)^2 \\
&= \sum_{s=1}^{t} \mathbb{1}_{a_s=k} X_{a_s}(s)^2 - \frac{1}{N_k(t)} \left(\sum_{s=1}^{t} \mathbb{1}_{a_s=k} X_{a_s}(s) \right)^2
\end{aligned}$$

because $\sum_{s=1}^{t} \mathbb{1}_{a_s=k} = N_k(t)$ by definition.

(b) We will write $k := a_{t+1}$. So we have $N_k(t+1) = N_k(t) + 1$. We will also write $S_k(t) := \sum_{s=1}^{t} \mathbb{1}_{a_s=k} X_{a_s}(s)$. By question 5.(a) we have:

$$\begin{aligned}
\Delta_k(t+1) &:= N_k(t+1)\hat{v}_k(t+1) - N_k(t)\hat{v}_k(t) \\
&= X_k(t+1)^2 - \frac{1}{N_k(t)+1}(S_k(t) + X_k(t+1))^2 + \frac{1}{N_k(t)} S_k(t)^2 \\
&= X_k(t+1)^2 - \frac{1}{N_k(t)+1}(S_k(t)^2 + 2X_k(t+1)S_k(t) + X_k(t+1)^2) + \frac{1}{N_k(t)} S_k(t)^2 \\
&= X_k(t+1)^2 \left(1 - \frac{1}{N_k(t)+1}\right) - \frac{2}{N_k(t)+1} X_k(t+1)S_k(t) + \frac{1}{N_k(t)(N_k(t)+1)} S_k(t)^2 \\
&= X_k(t+1)^2 \left(1 - \frac{1}{N_k(t)+1}\right) \\
&+ \left(\frac{1}{N_k(t)(N_k(t)+1)} - \frac{1}{N_k(t)} - \frac{1}{N_k(t)+1} \right) X_k(t+1)S_k(t) + \frac{1}{N_k(t)(N_k(t)+1)} S_k(t)^2 \\
&= X_k(t+1)^2 - \frac{X_k(t+1)}{N_k(t)} S_k(t) - \frac{X_k(t+1)}{N_k(t)+1} \left(S_k(t) + X_k(t+1)\right) \\
&+ \frac{1}{N_k(t)(N_k(t)+1)} S_k(t) \left(S_k(t) + X_k(t+1)\right) \\
&= \left(X_k(t+1) - \frac{1}{N_k(t)} S_k(t) \right) \left(X_k(t+1) - \frac{1}{N_k(t)+1} (S_k(t) + X_k(t+1)) \right) \\
&= \left(X_k(t+1) - \hat{\mu}_k(t)\right) \left(X_k(t+1) - \hat{\mu}_k(t+1)\right).
\end{aligned}$$

Therefore we proved that:

$$N_{a_{t+1}}(t{+}1)\hat{v}_{a_{t+1}}(t{+}1) = N_{a_{t+1}}(t)\hat{v}_{a_{t+1}}(t) + \left(X_{a_{t+1}}(t+1) - \hat{\mu}_{a_{t+1}}(t)\right) \left(X_{a_{t+1}}(t+1) - \hat{\mu}_{a_{t+1}}(t+1)\right).$$

This formulation is useful in practice because at each round t we can update $\hat{v}_{a_t}(t)$ in $\mathcal{O}(1)$ time instead of $\mathcal{O}(t)$.

(c) (d)

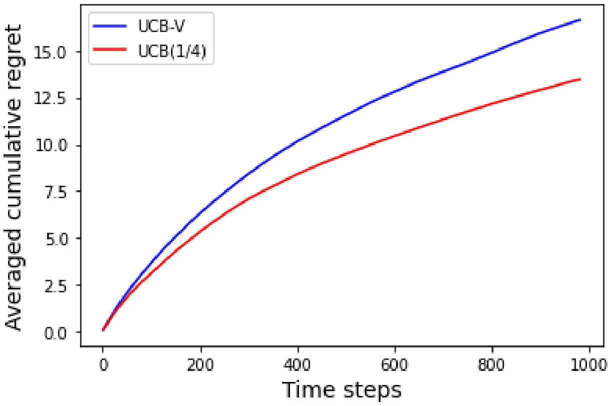
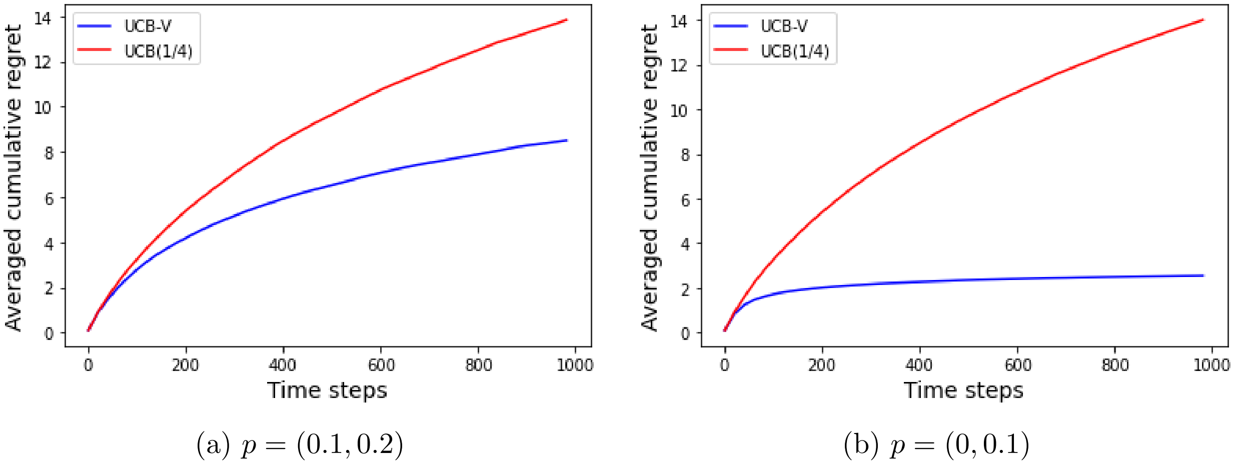


Figure 6: Regret as a function of time with UCB(1/4) and UCB-V, averaged over 1000 episodes.

(e)



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

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

Figure 7: Regret as a function of time with UCB(1/4) and UCB-V, averaged over 1000 episodes.

6. (a) We proceed by induction on t. Assume $f_k^t(x) \propto x^{S_k(t)}(1-x)^{N_k(t)-S_k(t)}$ for all $x \in [0,1]$, and for all $k \in \{0,1\}$. Let $k := a_{t+1}$. We update

$$\begin{aligned} f_k^{t+1}(x) &\propto f_k^t(x) \times \mathbb{P}(X_k(t+1) \mid \mu_k = x) \\ &\propto x^{S_k(t)}(1-x)^{N_k(t)-S_k(t)} \times x^{X_k(t+1)}(1-x)^{1-X_k^{t+1}} \\ &= x^{S_k(t+1)}(1-x)^{N_k(t+1)-S_k(t+1)} \end{aligned}$$

because $N_k(t+1) = N_k(t) + 1$ since $a_{t+1} = k$. The other arm is not updated, that is to say we keep $f_{1-k}^{t+1}(x) = f_{1-k}^{t}(x) \propto x^{S_{1-k}(t)}(1-x)^{N_{1-k}(t)-S_{1-k}(t)}$. But since $a_{t+1} = k$, we have $S_{1-k}(t+1) = S_{1-k}(t)$ and $N_{1-k}(t+1) = N_{1-k}(t)$. So we can rewrite $f_{1-k}^{t+1}(x) \propto x^{S_{1-k}(t+1)}(1-x)^{N_{1-k}(t+1)-S_{1-k}(t+1)}$. This shows that the distribution of arm k at time t is a Beta distribution with parameters $(S_k(t)+1, N_k(t)-S_k(t)+1)$, for any $k \in \{0,1\}$.

(b) (c)

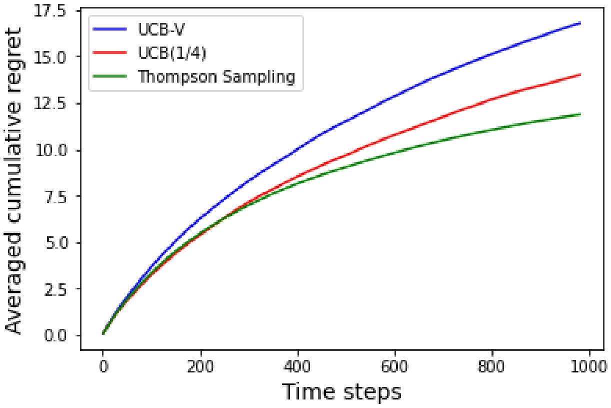
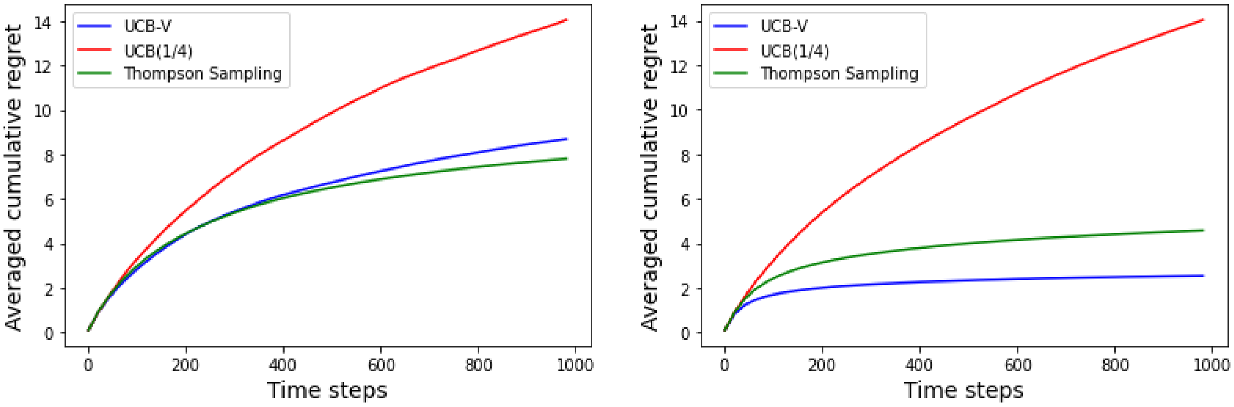


Figure 8: Regret as a function of time with UCB(1/4), UCB-V and Thompson Sampling, averaged over 1000 episodes.



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

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

Figure 9: Regret as a function of time with UCB(1/4), UCB-V and Thompson Sampling, averaged over 1000 episodes.

Thompson Sampling is better than optimistic algorithms UCB and UCB-V for $p = (0.5, 0.6)$ and $p = (0.1, 0.2)$. But UCB-V performs better for $p = (0, 0.1)$.

2 Rock Paper Scissors

Full information feedback

1. Here $M = N = 3$ and the loss matrix is given by $L = \begin{array}{cc} & \begin{array}{ccc} \text{R} & \text{P} & \text{S} \end{array} \\ & \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} \begin{array}{c} \text{R} \\ \text{P} \\ \text{S} \end{array} \end{array}$.

2. (a) With EWA updates: at each time step t, $q_t = \delta_{j_t}$ where $j_t \in \text{argmax}(p_t L)$. Therefore, the loss $\ell_t(i)$ incurred by the player if he chooses action i at times t is equal to L_{i,j_t} with $j_t \in \text{argmax}(p_t L)$.

(b) Concerning the implementation, I initialized $p_1 = (1/3, 1/3, 1/3)$. If $\text{argmax}(p_t L)$ contains multiple elements, then the adversary selects a random action in $\text{argmax}(p_t L)$.

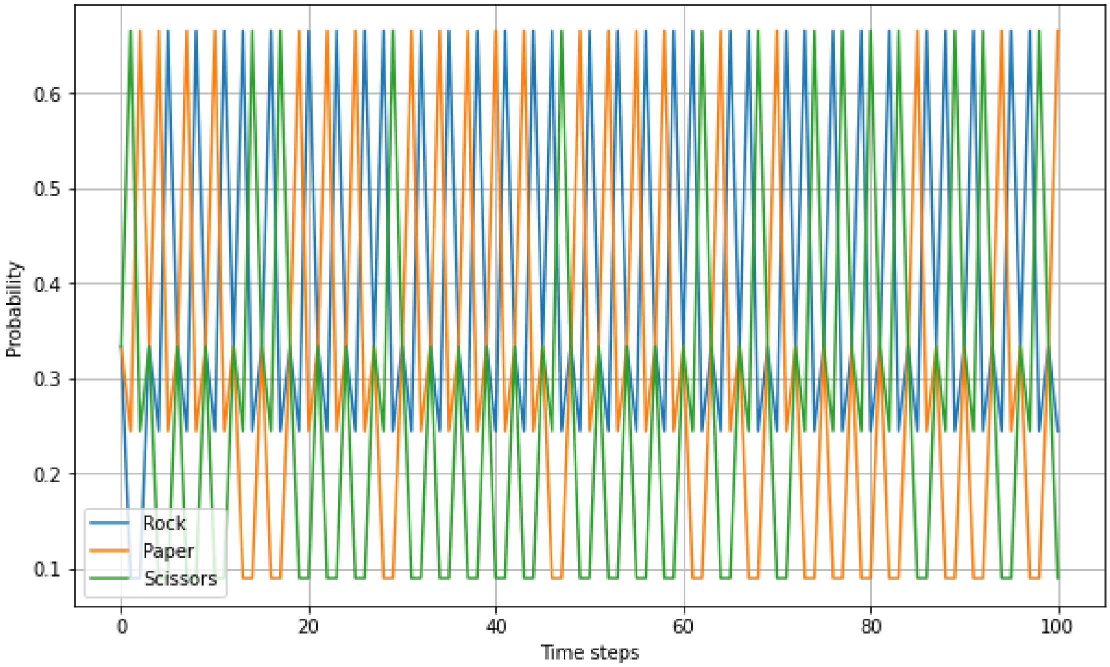
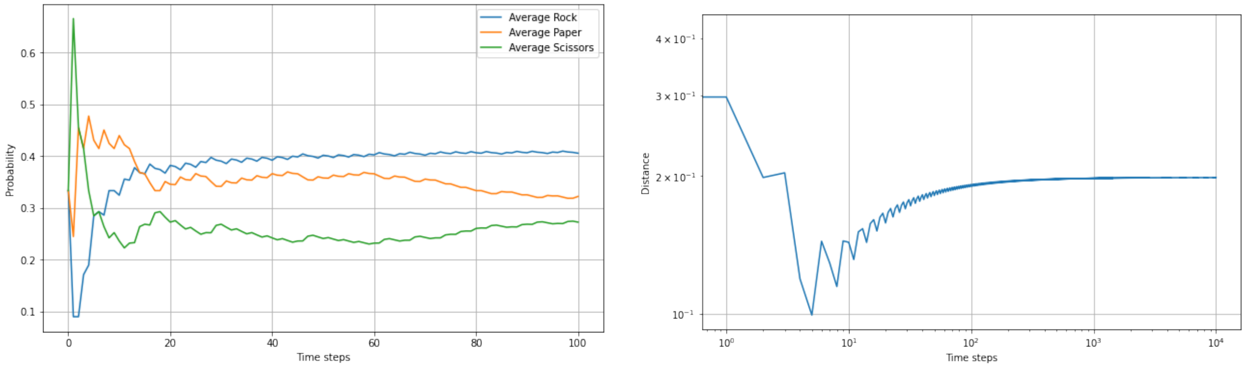


Figure 10: Instance of the game with $T = 100$ and $\eta = 1$.

(c)



(a) Average weights.

(b) $\|\bar{p}_t - (1/3, 1/3, 1/3)\|_2$ in log log scale.

Figure 11: Evolution of the average probabilities for an instance of the game with $T = 10000$ and $\eta = 1$.

We observe that $\|\bar{p}_t - (1/3, 1/3, 1/3)\|_2$ converges to some value > 0.

(d)

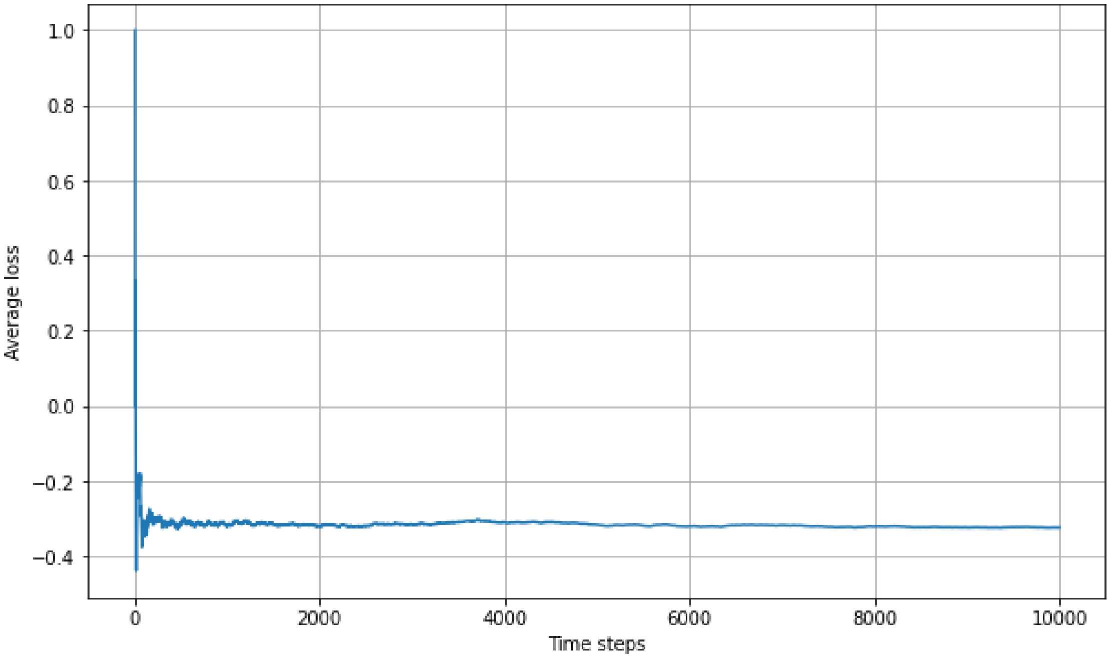
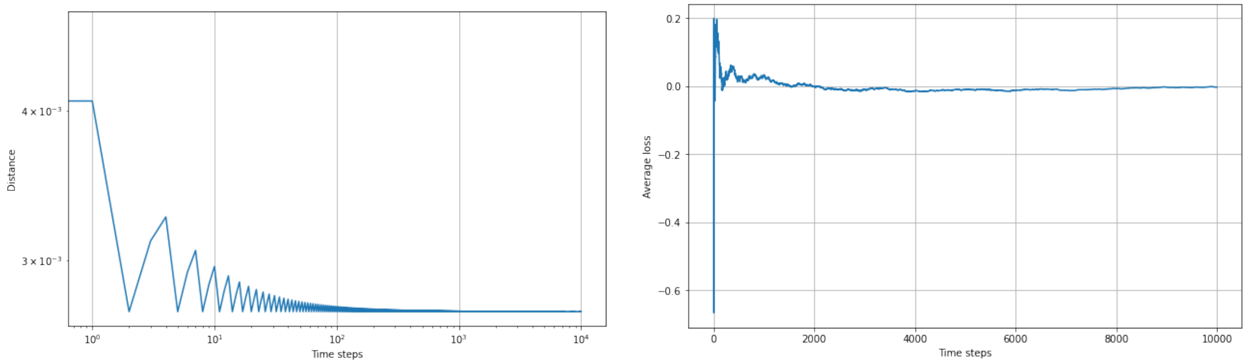


Figure 12: Evolution of the average loss $\bar{\ell}_t = \frac{1}{t}\sum_{s=1}^{t} \ell(i_s, j_s)$ for an instance of the game with $T = 10000$ and $\eta = 1$.

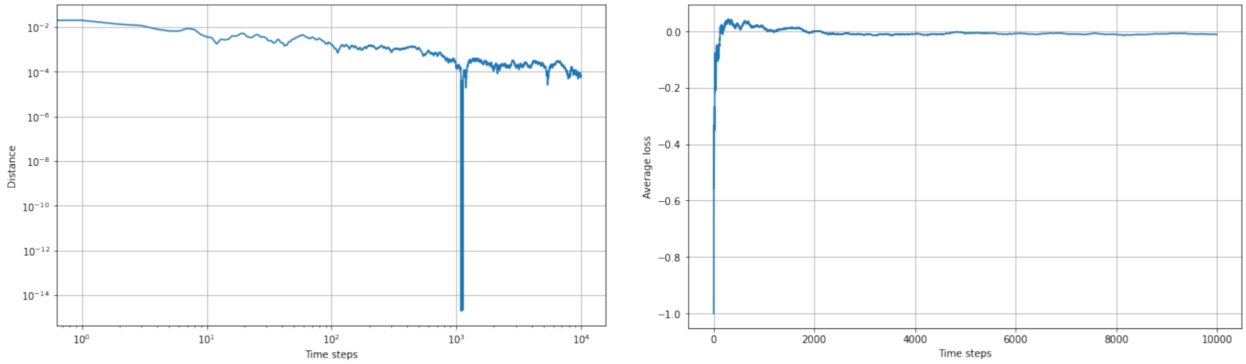
(e)



(a) $\|\bar{p}_t - (1/3, 1/3, 1/3)\|_2$ in log log scale.

(b) Average loss $\bar{\ell}_t = \frac{1}{t}\sum_{s=1}^{t} \ell(i_s, j_s)$.

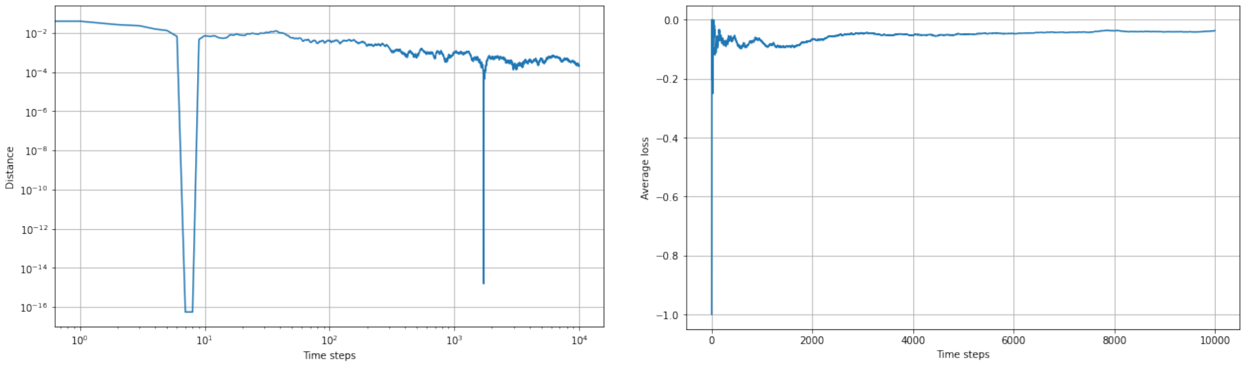
Figure 13: Evolution of the average weight vector and the average loss for an instance of the game with $T = 10000$ and $\eta = 0.01$.



(a) $\|\bar{p}_t - (1/3, 1/3, 1/3)\|_2$ in log log scale.

(b) Average loss $\bar{\ell}_t = \frac{1}{t}\sum_{s=1}^{t} \ell(i_s, j_s)$.

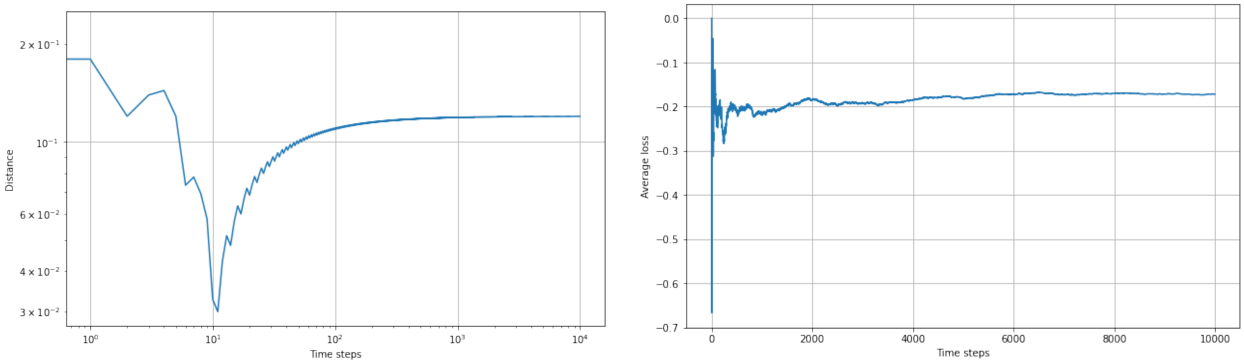
Figure 14: Evolution of the average weight vector and the average loss for an instance of the game with $T = 10000$ and $\eta = 0.05$.



(a) $\|\bar{p}_t - (1/3, 1/3, 1/3)\|_2$ in log log scale.

(b) Average loss $\bar{\ell}_t = \frac{1}{t}\sum_{s=1}^{t} \ell(i_s, j_s)$.

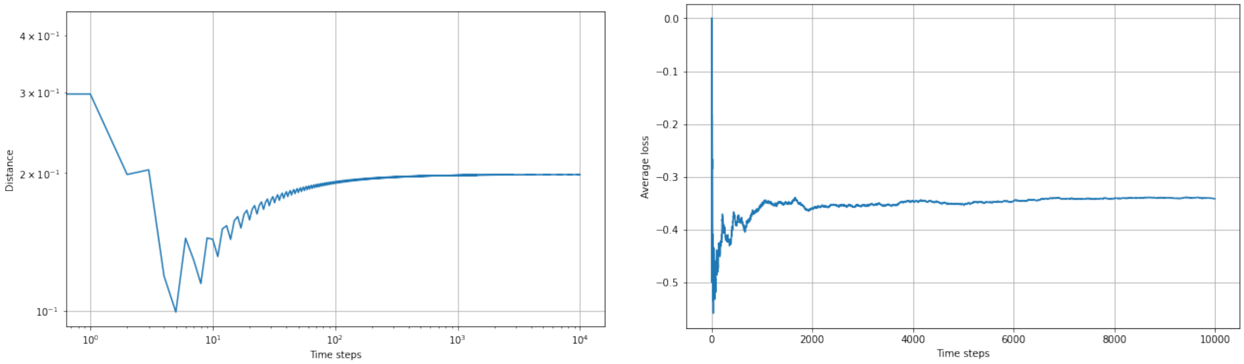
Figure 15: Evolution of the average weight vector and the average loss for an instance of the game with $T = 10000$ and $\eta = 0.1$.



(a) $\|\bar{p}_t - (1/3, 1/3, 1/3)\|_2$ in log log scale.

(b) Average loss $\bar{\ell}_t = \frac{1}{t}\sum_{s=1}^{t} \ell(i_s, j_s)$.

Figure 16: Evolution of the average weight vector and the average loss for an instance of the game with $T = 10000$ and $\eta = 0.5$.



(a) $\|\bar{p}_t - (1/3, 1/3, 1/3)\|_2$ in log log scale.

(b) Average loss $\bar{\ell}_t = \frac{1}{t}\sum_{s=1}^{t} \ell(i_s, j_s)$.

Figure 17: Evolution of the average weight vector and the average loss for an instance of the game with $T = 10000$ and $\eta = 1$.

In theory, the best learning rate here is $\eta = \sqrt{\frac{\log 3}{10000}} \approx 0.01$. But in practice for this game there is no major difference in the convergence speed of the average weight vector, when it converges. We see that the convergence might not always be smooth for some instances of the game, due to randomness. Also, a higher learning rate yields a lower average loss here.

Bandit feedback

3. In the context of bandit feedback, the player can use EXP3 algorithm. Ideally, we would like to reuse EWA but this is not possible since the player does not observe $\ell_t(i)$ for all $i \neq i_t$, where i_t is the action chosen by the player at time t. The idea of EXP3 is to replace $\ell_t(i)$ with an unbiased estimate that is observed by the player. We choose

$$\hat{\ell}_t(i) = \frac{\ell_t(i)}{p_t(i)} \mathbb{1}_{i=i_t}.$$

4. With the loss matrix $L = \begin{array}{cccc} \text{R} & \text{P} & \text{S} & \\ \left[\begin{array}{c} 0 \\ -1 \\ 1 \end{array}\right. & \begin{array}{c} 1 \\ 0 \\ -1 \end{array} & \left.\begin{array}{c} -1 \\ 1 \\ 0 \end{array}\right] & \begin{array}{c} \text{R} \\ \text{P} \\ \text{S} \end{array} \end{array}$, we obtain the following results.

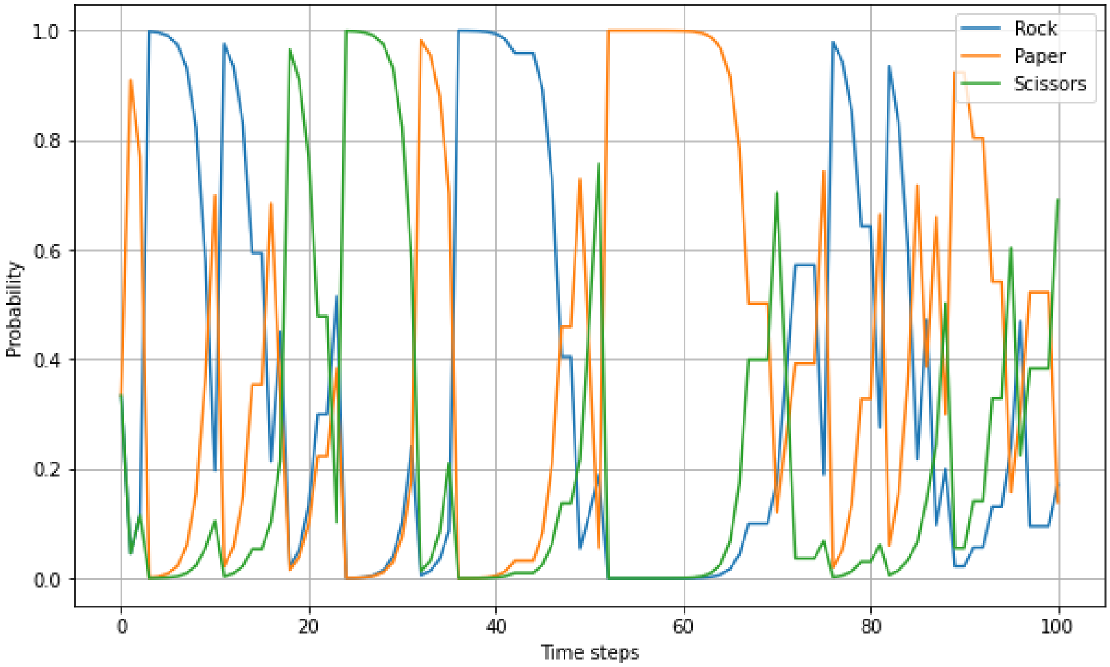
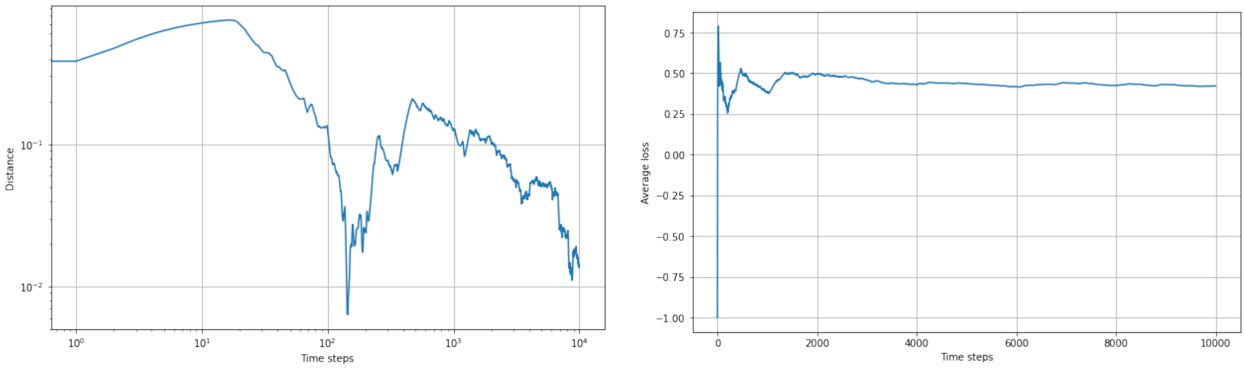


Figure 18: Instance of the game with $T = 100$ and $\eta = 1$.



(a) $\|\bar{p}_t - (1/3, 1/3, 1/3)\|_2$ in log log scale.

(b) Average loss $\bar{\ell}_t = \frac{1}{t}\sum_{s=1}^{t} \ell(i_s, j_s)$.

Figure 19: Evolution of the average weight vector and the average loss for an instance of the game with $T = 10000$ and $\eta = 1$.

With a rescaled loss matrix $L_s = \begin{array}{cc} \begin{array}{ccc} \text{R} & \text{P} & \text{S} \end{array} & \\ \begin{bmatrix} 0.5 & 1 & 0 \\ 0 & 0.5 & 1 \\ 1 & 0 & 0.5 \end{bmatrix} & \begin{array}{c} \text{R} \\ \text{P} \\ \text{S} \end{array} \end{array}$, we obtain:

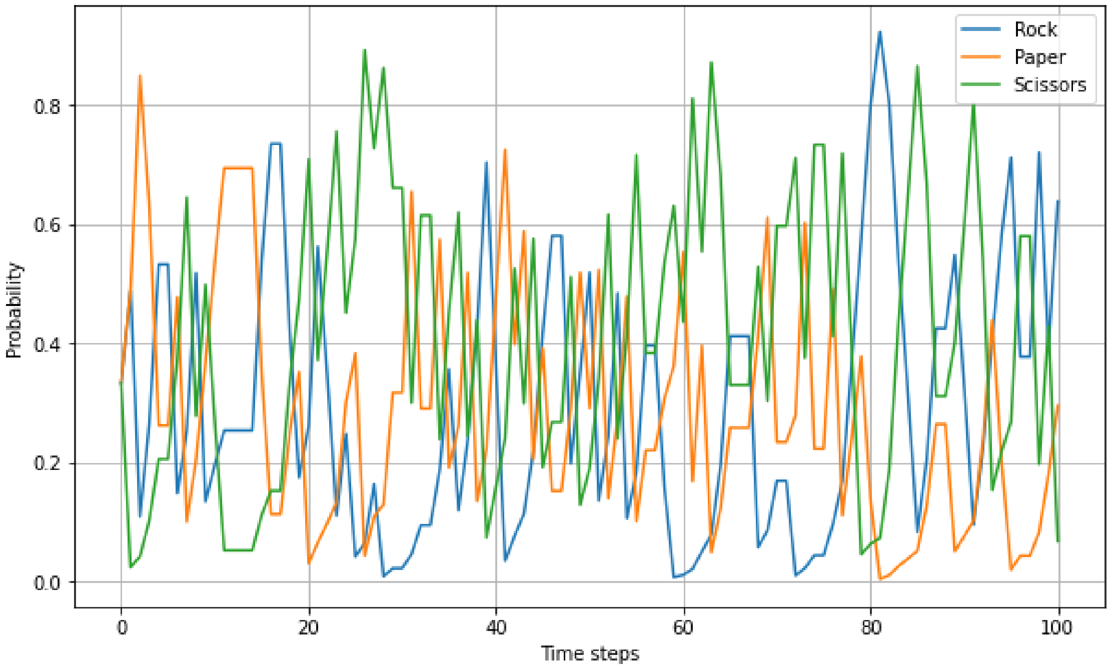
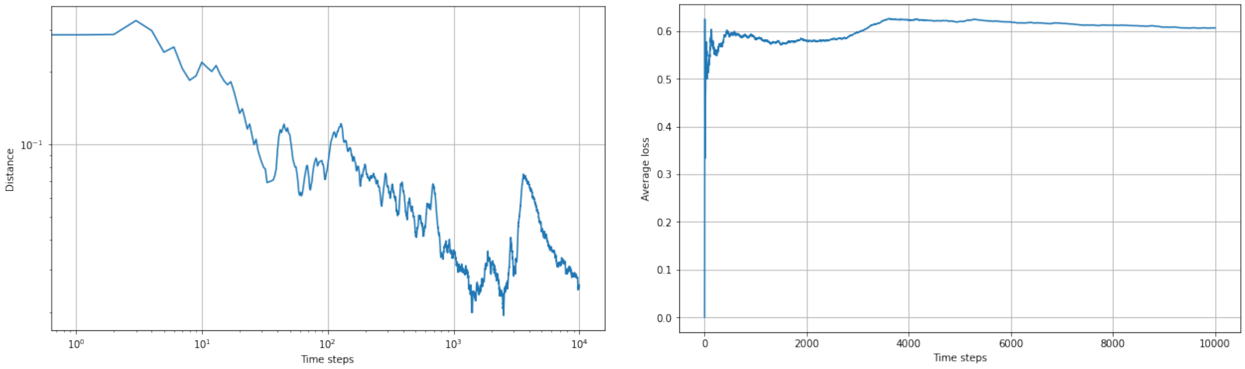


Figure 20: Instance of the game with $T = 100$ and $\eta = 1$.



(a) $\|\bar{p}_t - (1/3, 1/3, 1/3)\|_2$ in log log scale.

(b) Average loss $\bar{\ell}_t = \frac{1}{t}\sum_{s=1}^{t} \ell(i_s, j_s)$.

Figure 21: Evolution of the average weight vector and the average loss for an instance of the game with $T = 10000$ and $\eta = 1$.