MATHÉMATIQUES
VISION
APPRENTISSAGE

MVA

Sequential Learning

2021-22

Home Assignment

Antoine Desjardins
Alex Fauduet

Table des matières

1 Rock Paper Scissors **2**

1.1 Question 1 2

1.2 Question 2 - Simulation against a fixed adverary 2

1.3 Question 3 - Simulation against an adaptive adversary 8

1.4 Question 4 11

1.5 Question 5 11

2 Bernoulli Bandits **18**

2.1 Question 1 - Follow The Leader 18

2.2 Question 2 - UCB 19

2.3 Question 3 22

2.4 Question 4 23

2.5 Question 5 - UCB-V 23

1 Rock Paper Scissors

1.1 Question 1

The game "Rock, Paper, Scissors" would correspond to $M = N = 3$, with a loss Matrix L.

L	Rock	Paper	Scissors
Rock	0	1	-1
Paper	-1	0	1
Scissors	1	-1	0

1.2 Question 2 - Simulation against a fixed adverary

(a) $l_t(i) = L[i, j_t]$ (simply by disjunction of cases and applying the relevant probability of occurence to each of them).

(b)

The best strategy naturally distinguishes itself as the one played with the highest probability. That strategy is of course to play the arm that wins over the opponent's highest-probability move. In our example that move is scissor (probability 0.5), so we should play rock (blue line in the figure 1). As we can see, in spite of some variations it becomes eventually clear that it is best to play only the luckiest of arms.

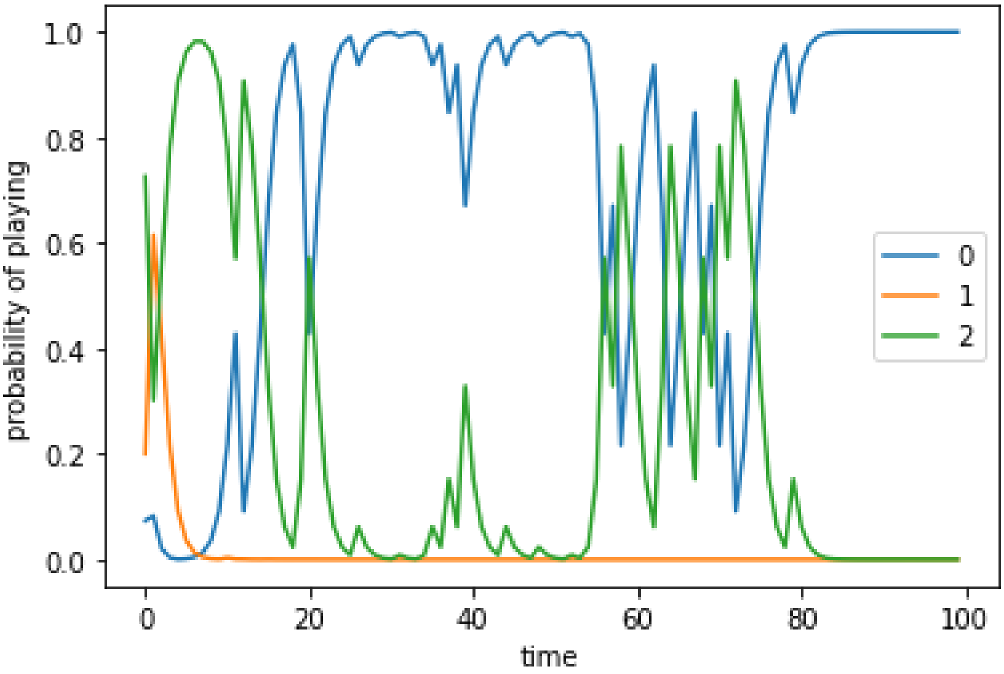

Figure 1 – Evolution of the best strategy in the Rock-Paper-Scissors game

(c)

Coherently, the loss converges towards its asymptotic value of around -0.2 as the best arm gains in popularity.

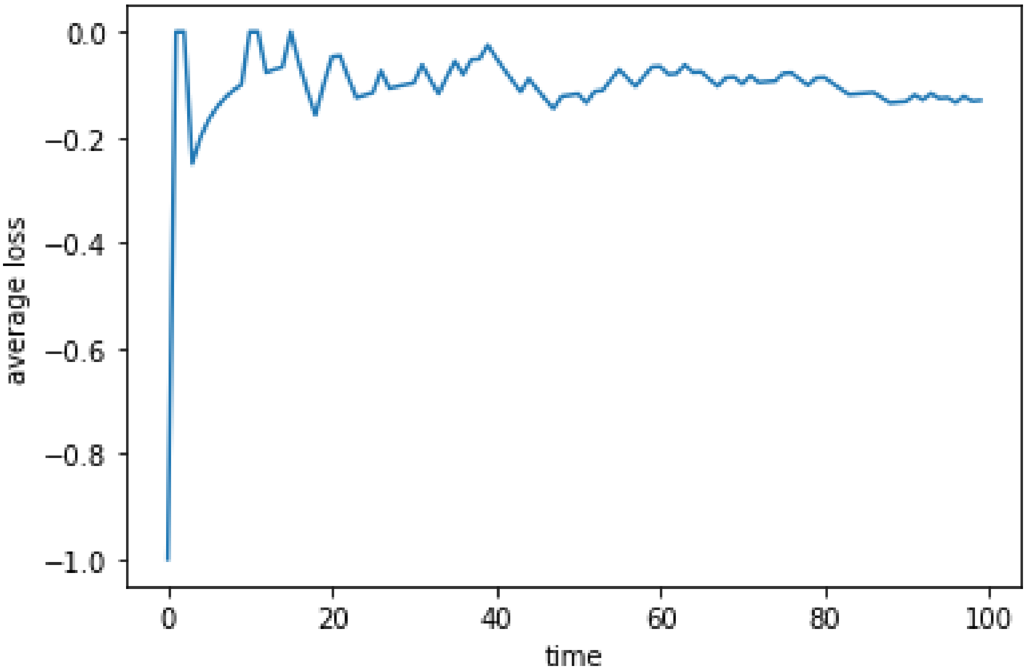

FIGURE 2 – Evolution of the average loss in the Rock-Paper-Scissors game

(d)

The best value of η appears to be, in practice, around 1 as we reach the asymptotic limit of the loss faster. In theory high values are risky as we may have a slow-convergence pendulum movement over topologies such as half-pipes. We can notice that the best strategy emerges faster with $\eta = 1$ too, which of course is coherent with the loss.

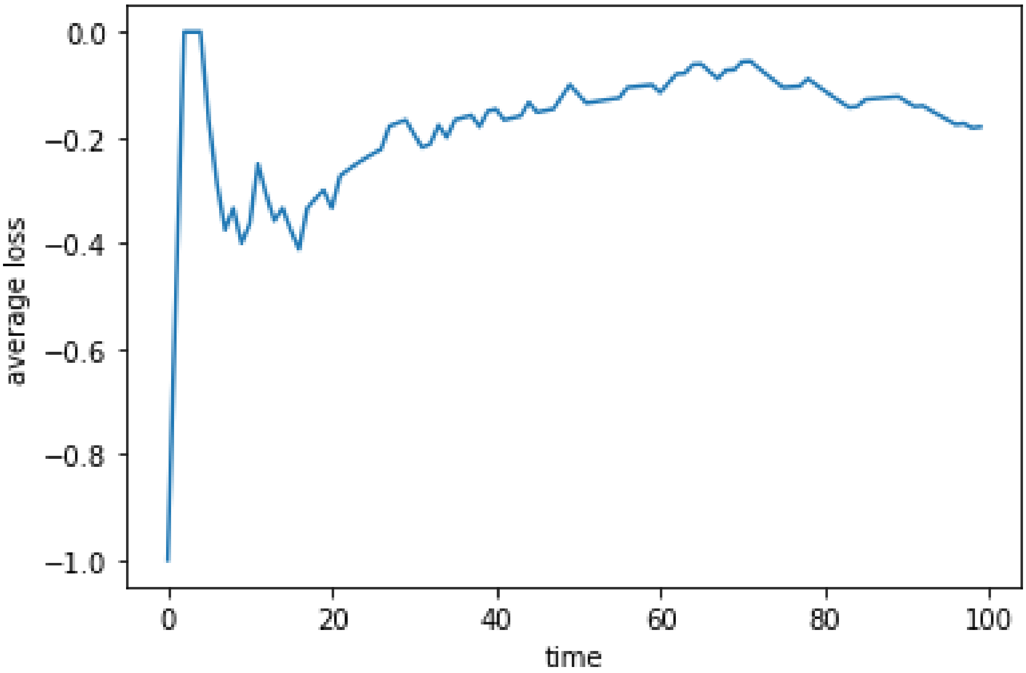

FIGURE 3 – Evolution of the average loss in the Rock-Paper-Scissors game where $\eta = 0.01$

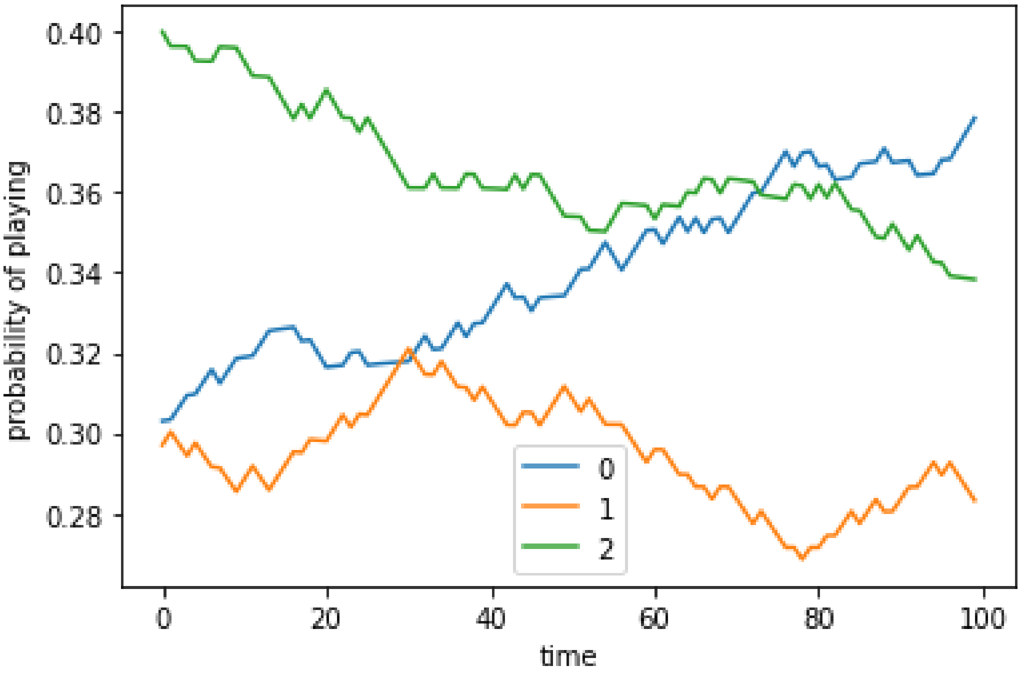

FIGURE 4 – Evolution of the best strategy in the Rock-Paper-Scissors game where $\eta = 0.01$

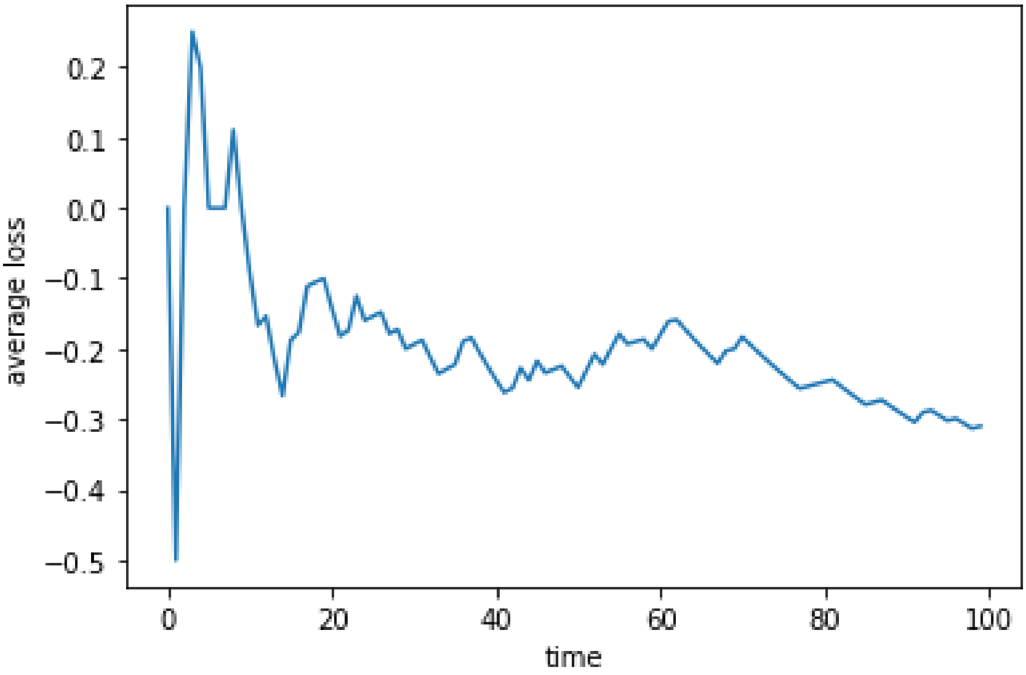

FIGURE 5 – Evolution of the average loss in the Rock-Paper-Scissors game where $\eta = 0.05$

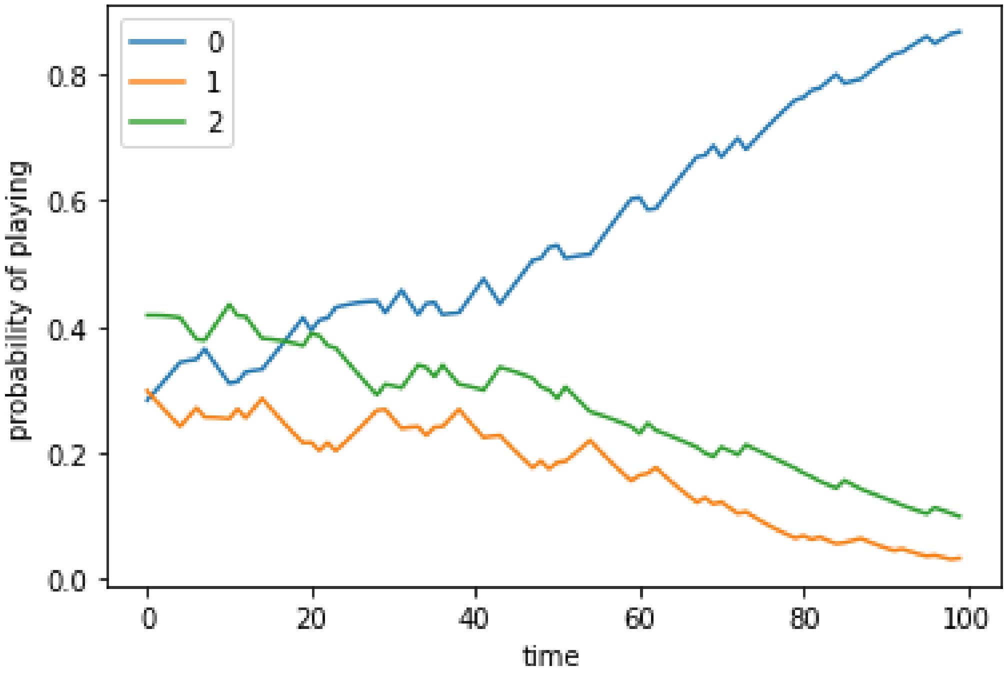

Figure 6 – Evolution of the best strategy in the Rock-Paper-Scissors game where $\eta = 0.05$

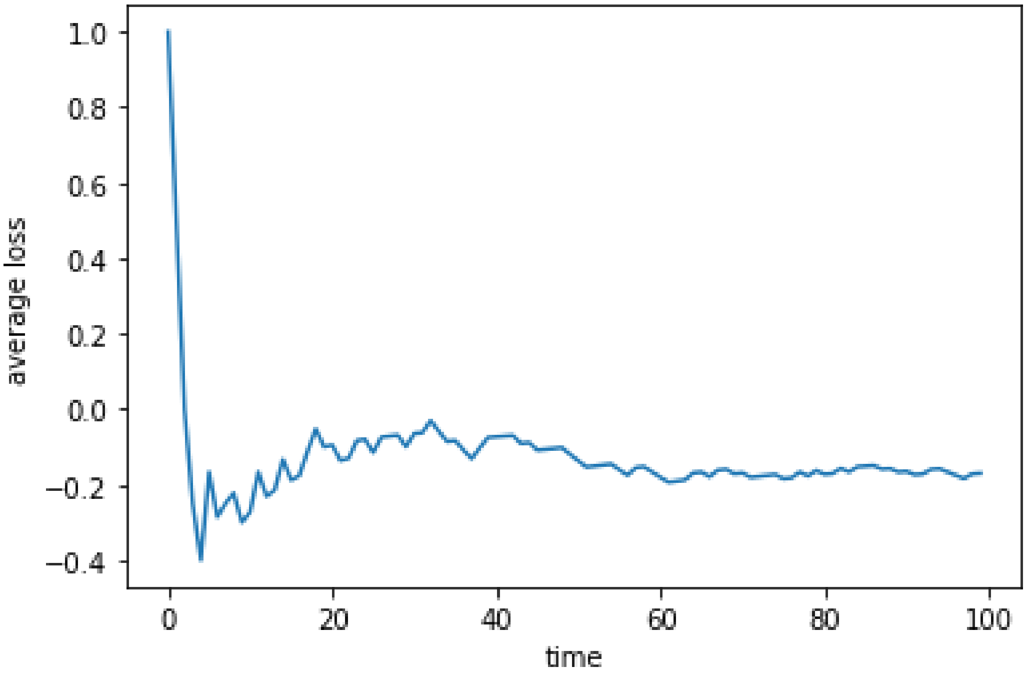

Figure 7 – Evolution of the average loss in the Rock-Paper-Scissors game where $\eta = 0.1$

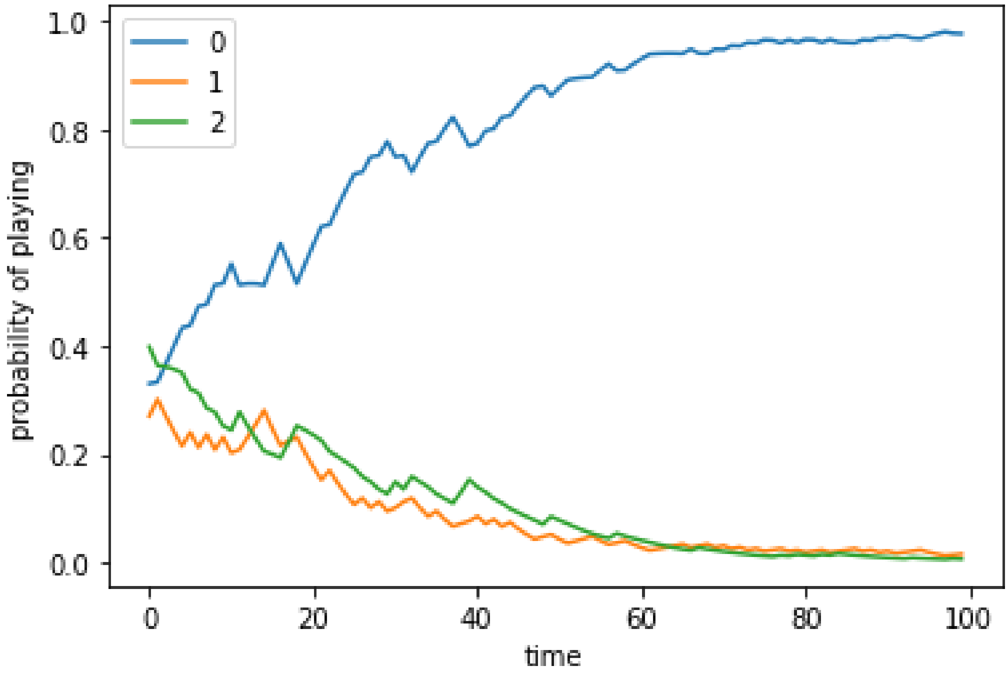

Figure 8 – Evolution of the best strategy in the Rock-Paper-Scissors game where $\eta = 0.1$

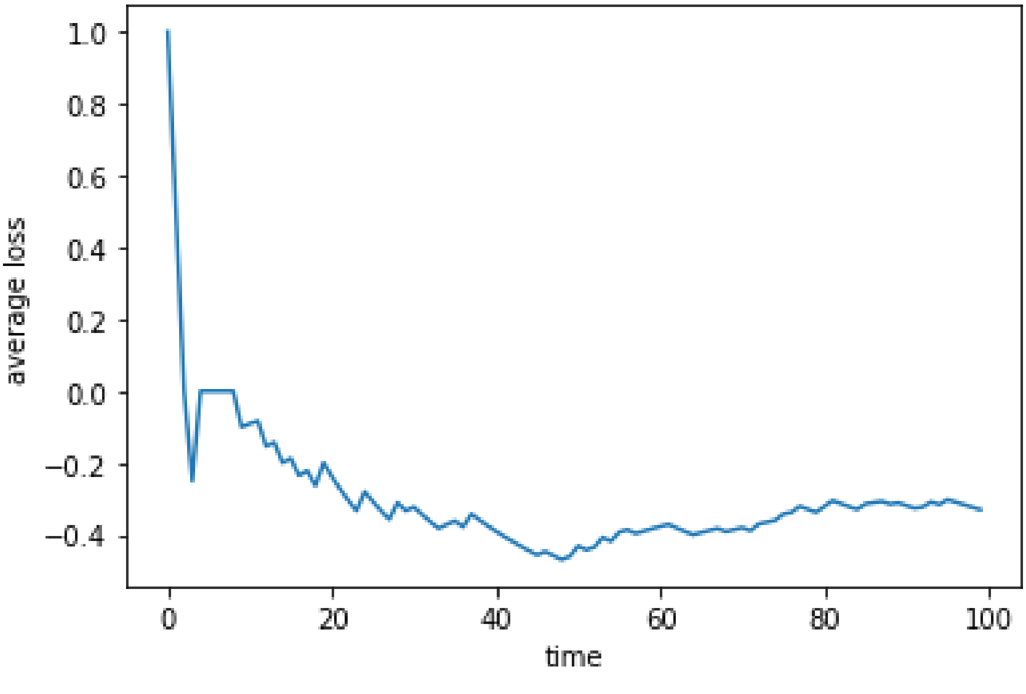

Figure 9 – Evolution of the average loss in the Rock-Paper-Scissors game where $\eta = 0.5$

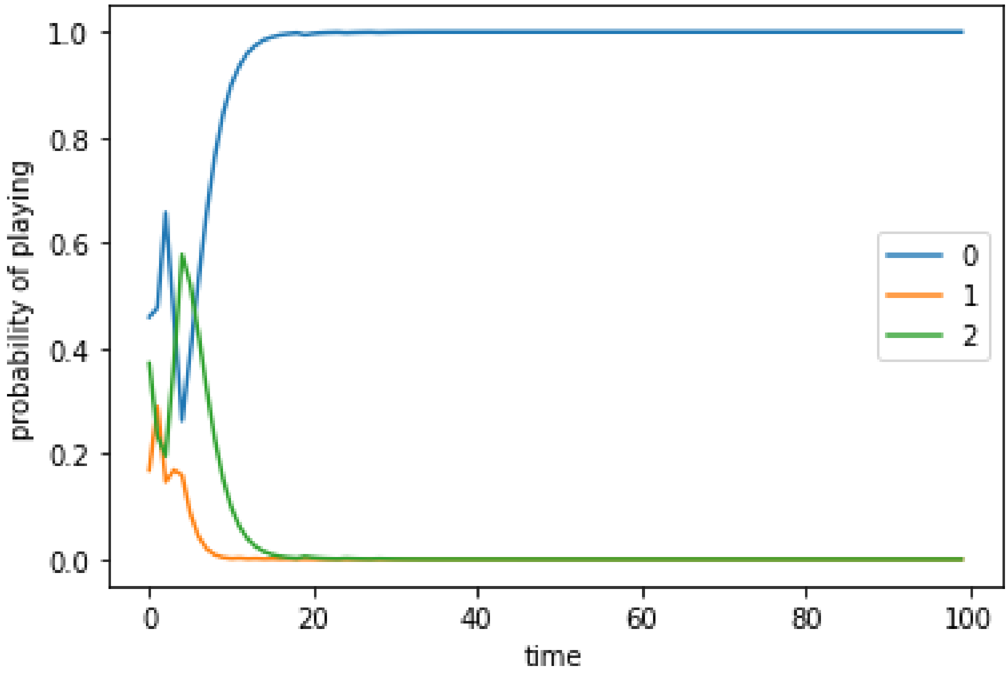

Figure 10 – Evolution of the best strategy in the Rock-Paper-Scissors game where $\eta = 0.5$

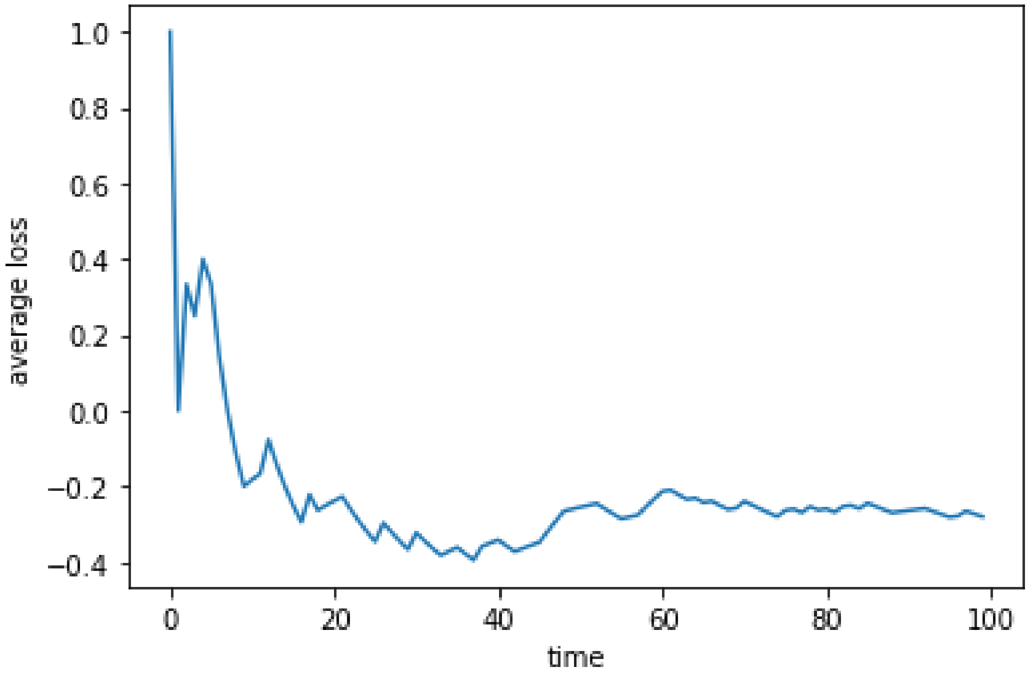

Figure 11 – Evolution of the average loss in the Rock-Paper-Scissors game where $\eta = 1$

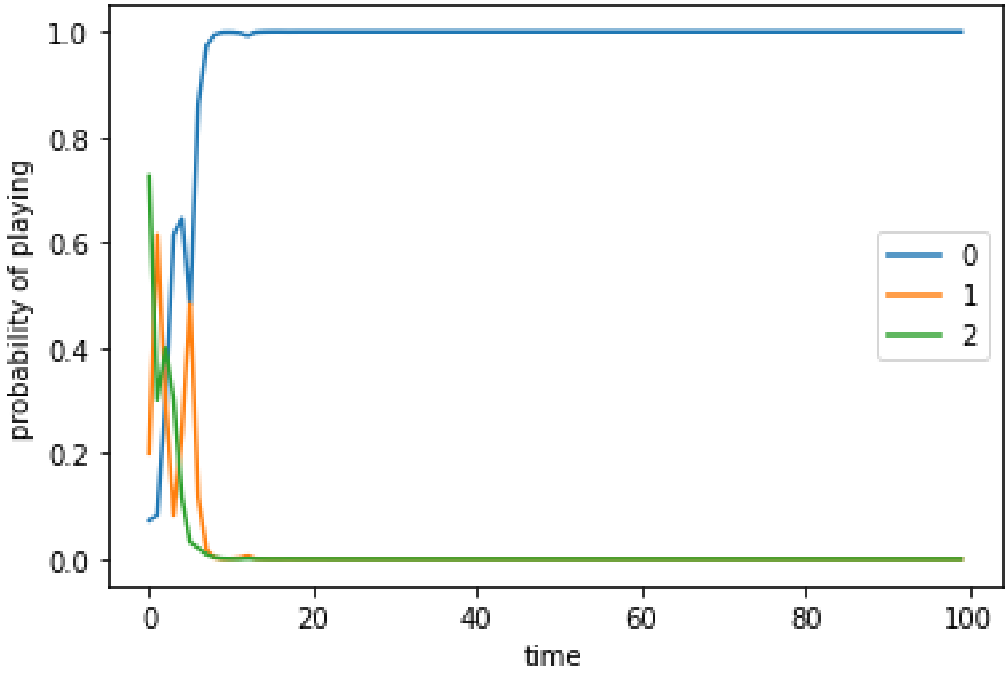

FIGURE 12 – Evolution of the best strategy in the Rock-Paper-Scissors game where $\eta = 1$

1.3 Question 3 - Simulation against an adaptive adversary

(a)

In this case, the update for OGD is

$$q_{t+1} = \Pi_{\Delta_N}(q_t - \eta * \nabla l_t(q_t))$$

$$q_{t+1} = \Pi_{\Delta_N}(q_t - \eta * \frac{d\Sigma_{j=1}^{N} q_t(j)g_t(j)}{dq_t})$$

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

For the projection, we use the method described in *Projection onto the probability simplex : An efficient algorithm with a simple proof, and an application*, Weiran Wang, Miguel Á. Carreira-Perpiñán, that computes the euclidian projection over Δ_N in $O(N \log N)$.

(b)

Figure 13 shows the emergence of the best strategy for the player. The strategy seems unstable and has a noticeable change from the previous one as it is a *mixed strategy*, meaning that we pull the arms with some randomness and a given probability for each.

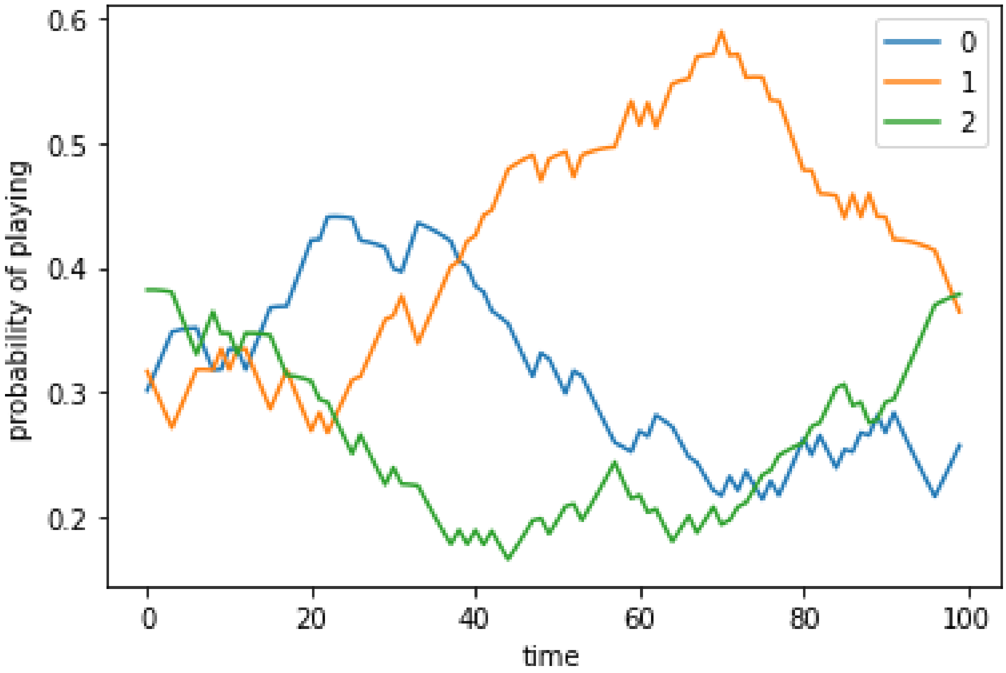

FIGURE 13 – Evolution of the best strategy in the Rock-Paper-Scissors game with OGD where $\eta = 0.05$

Figure 14 has been added to show the strategy of the adversary.

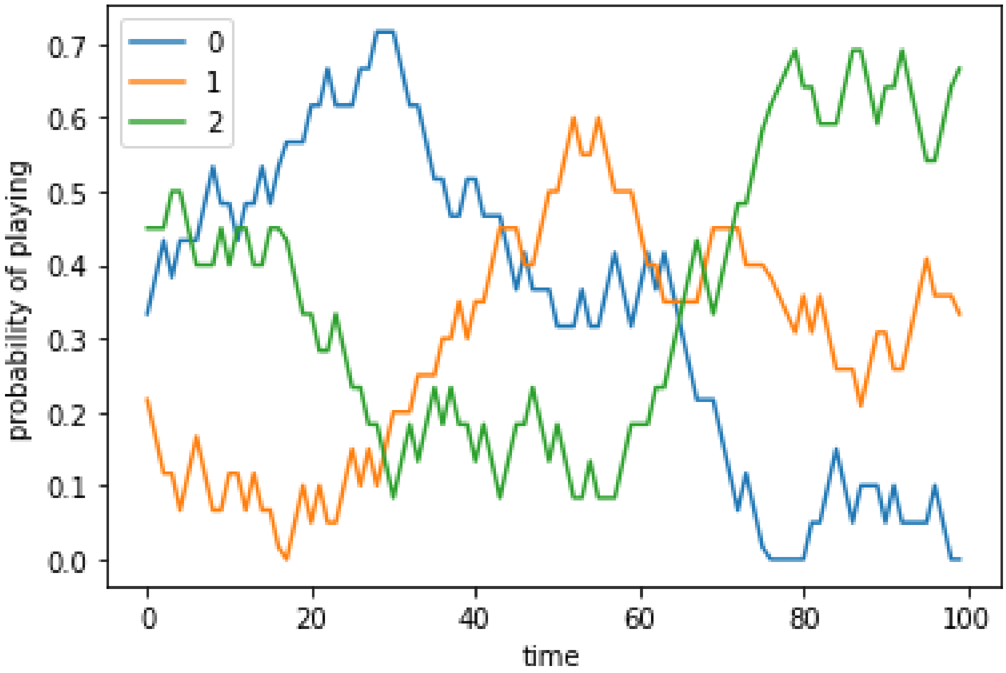

FIGURE 14 – Evolution of the best strategy in the Rock-Paper-Scissors game with OGD where $\eta = 0.05$ for the adversary

(c)

Figure 15 represents the loss suffered by the player during the battle between OGD and EWA. As the loss is negative, it means the player EWA is winning against the more efficient OGD. However the loss is close to zero, and depending on the execution we have witnessed either OGD or EWA winning. One should also notice that the player has a head start as the q_0 of the adversary is very bad. Our conclusion is that both algorithm are very close in terms of performance, and it is difficult to decide on a clear winner.

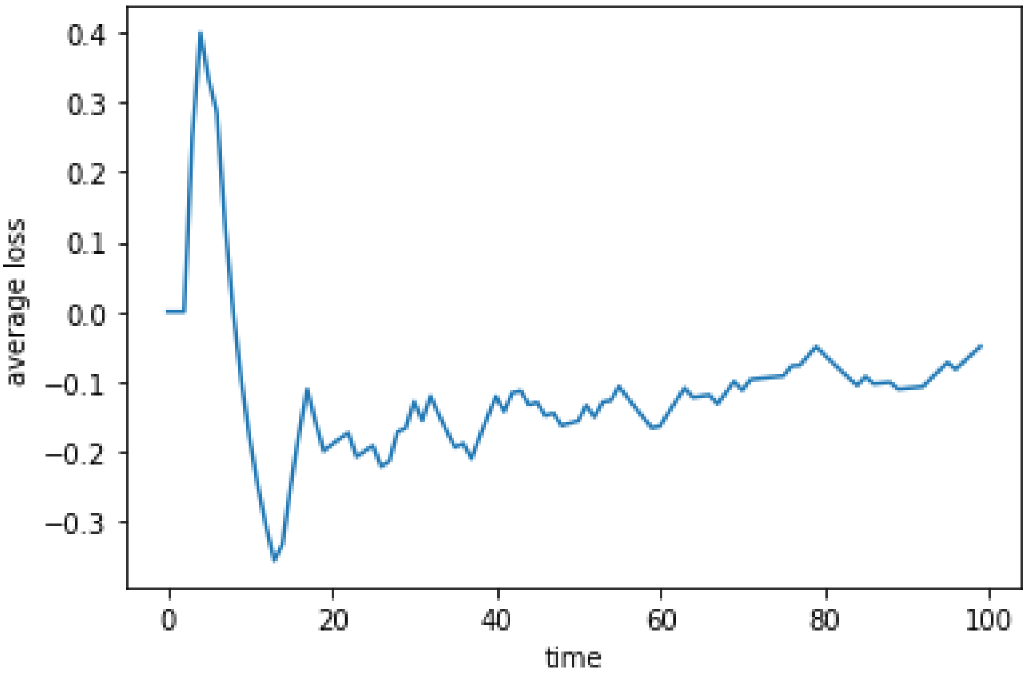

Figure 15 – Evolution of the loss in the Rock-Paper-Scissors game with OGD where $\eta = 0.05$ for the adversary

d

We can observe on figure 17 that the optimal strategy converges towards equivalent probability to play any arms. This allows to be unpredictable and is a well-known *mixed-strategy Nash equilibrium* for the Rock-Paper-Scissors game.

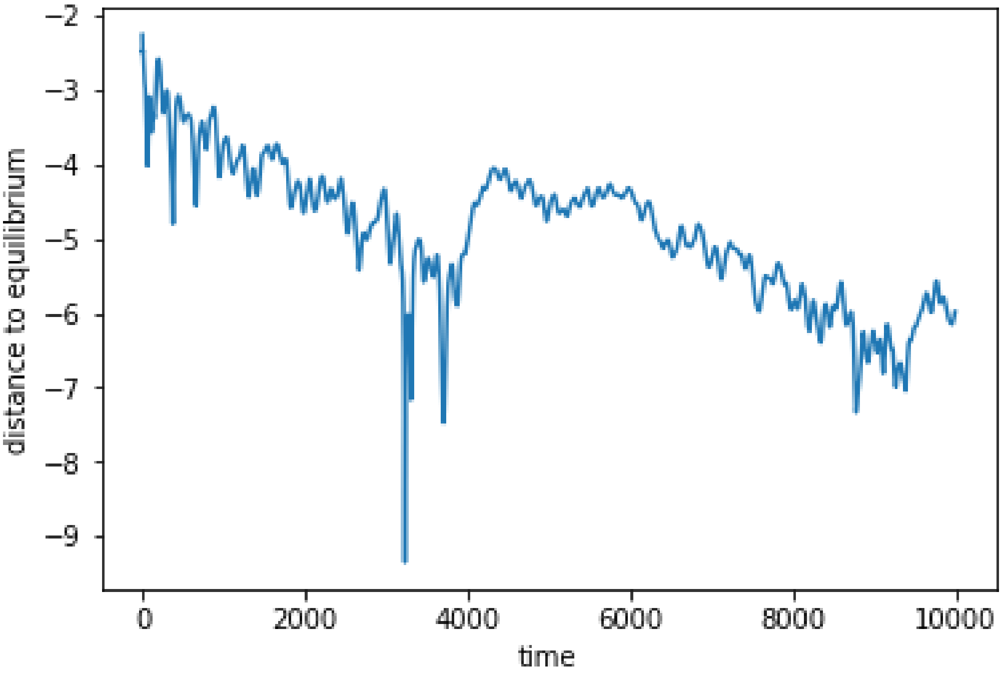

Figure 16 – Evolution of the distance to the Nash equilibrium in log-scale

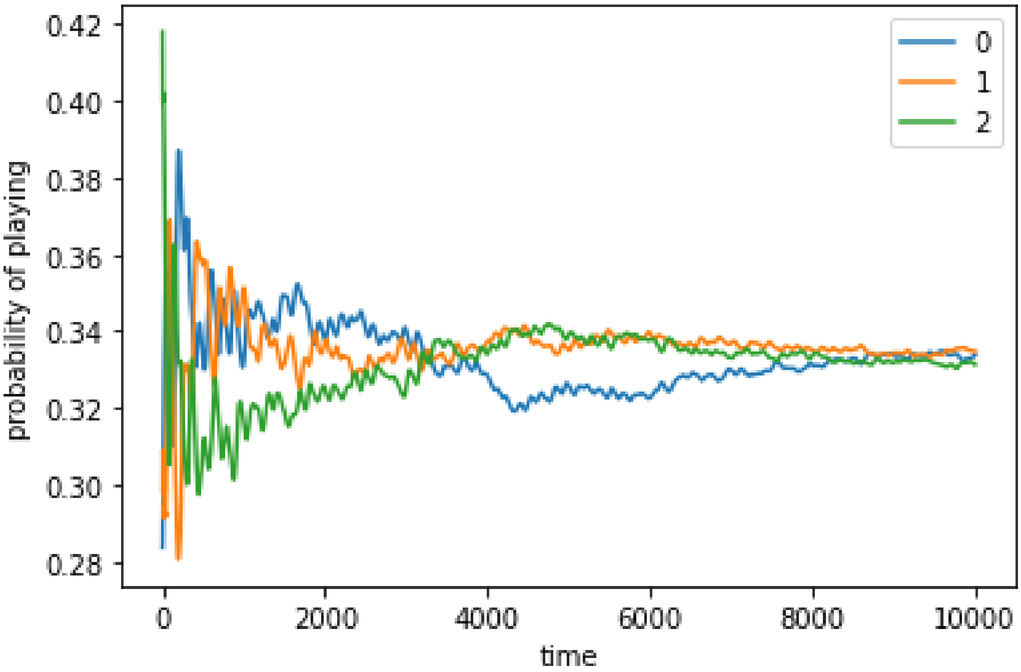

FIGURE 17 – Evolution of the strategy in the Rock-Paper-Scissors game for p_t bar

1.4 Question 4

EXP3 is used to replace EWA when we don't have full information feedback. We just swap the loss in EWA with an unbiased proxy : $l_t(j) \backsim g_t(j) = \frac{l_t(j)}{p_t(j)}\mathbb{I}\{j = k_t\}$.

1.5 Question 5

The results in this section are very similar to EWA, though we have a slightly slower convergence as the player also needs to learn the loss as he plays.

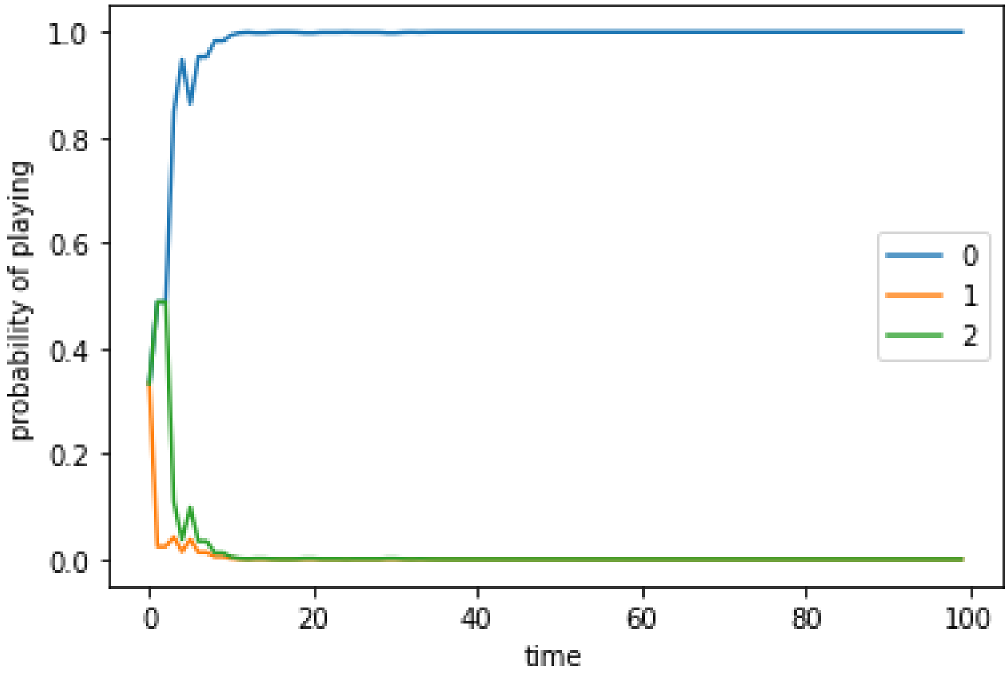

Figure 18 – Evolution of the best strategy in the Rock-Paper-Scissors game for EXP3

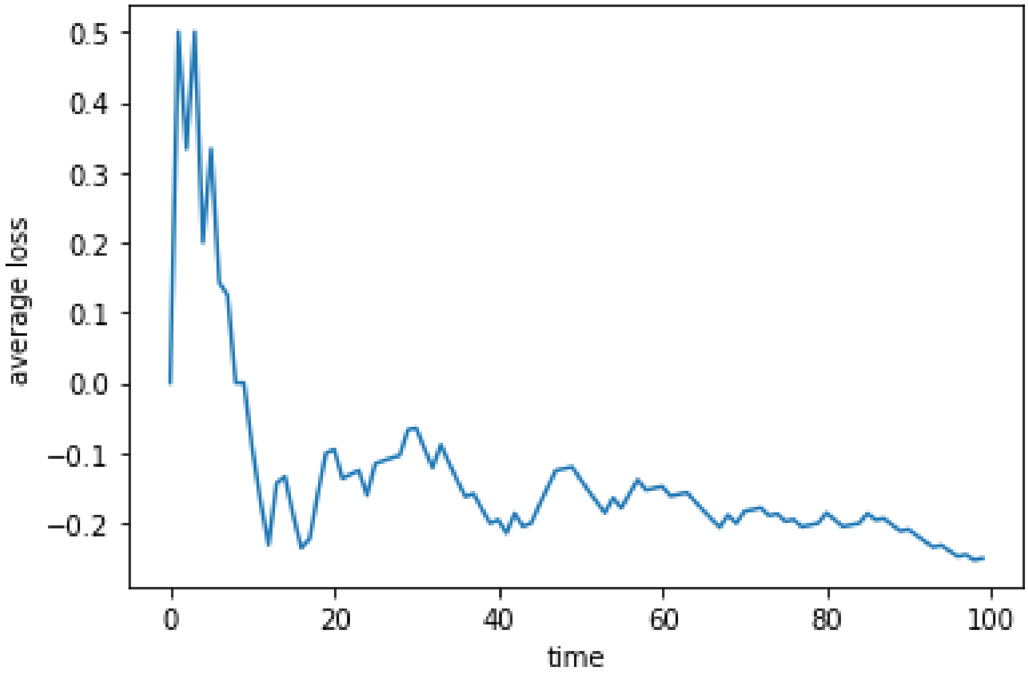

Figure 19 – Evolution of the average loss in the Rock-Paper-Scissors game for EXP3

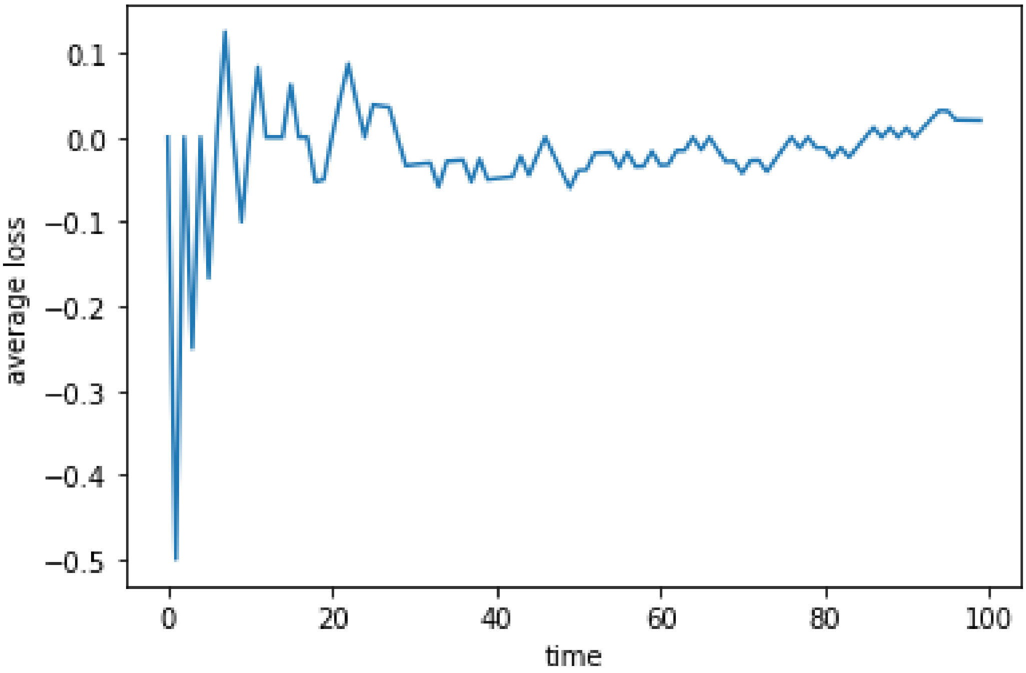

Figure 20 – Evolution of the average loss in the Rock-Paper-Scissors game where $\eta = 0.01$ for EXP3

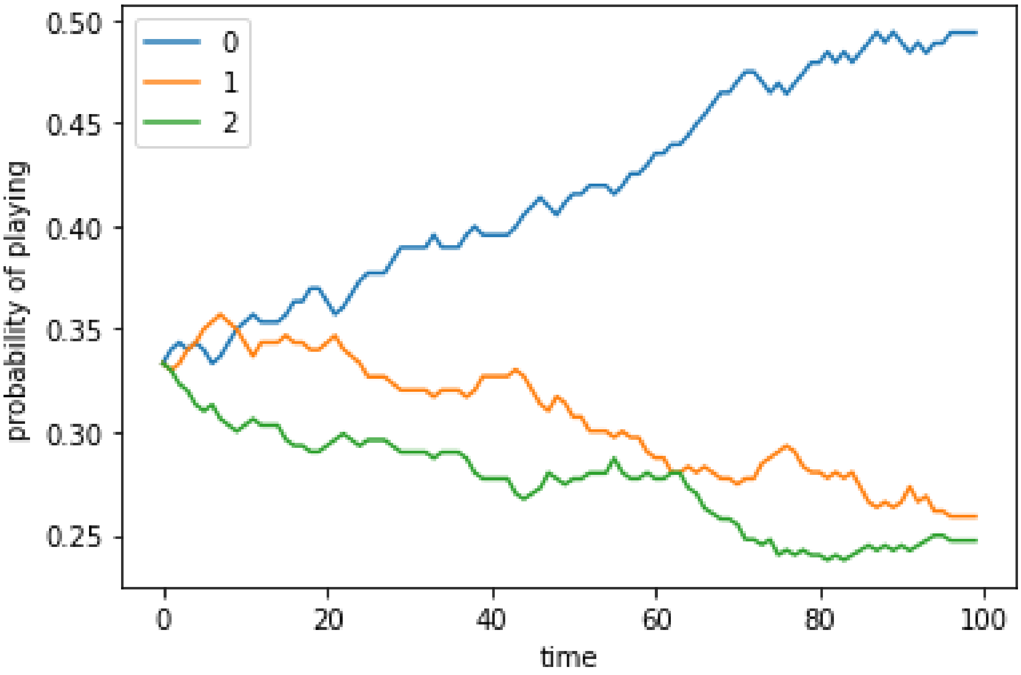

Figure 21 – Evolution of the best strategy in the Rock-Paper-Scissors game where $\eta = 0.01$ for EXP3

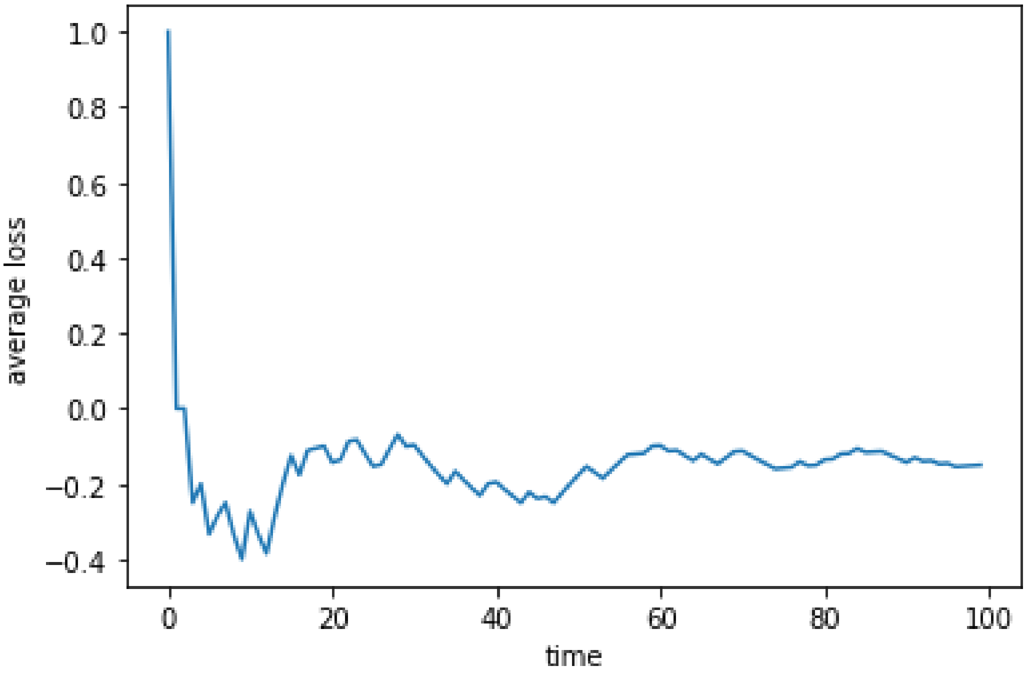

Figure 22 – Evolution of the average loss in the Rock-Paper-Scissors game where $\eta = 0.05$ for EXP3

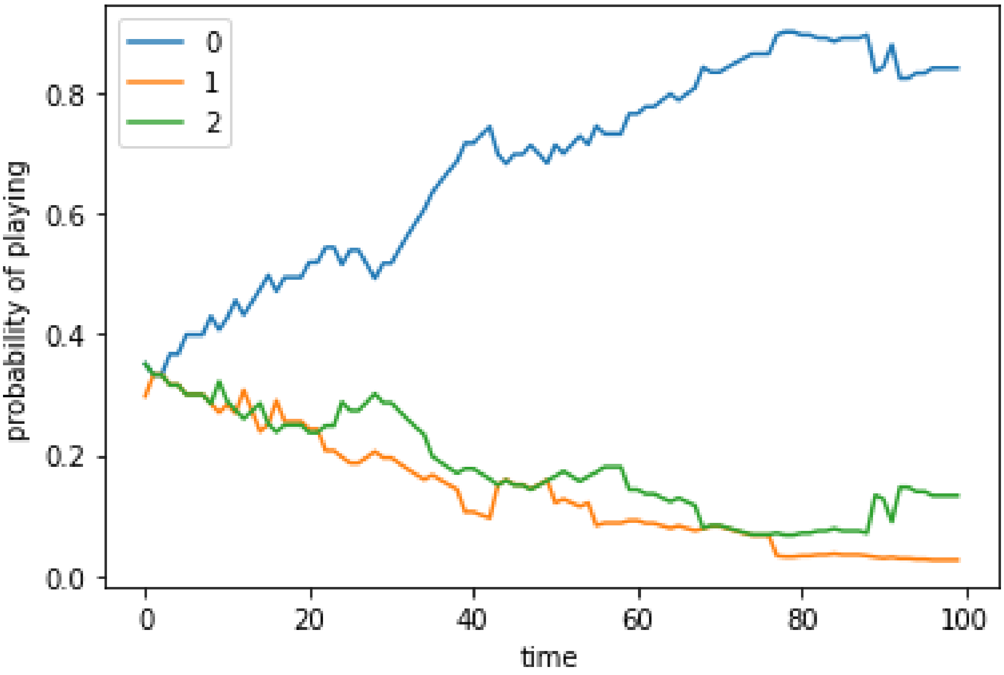

Figure 23 – Evolution of the best strategy in the Rock-Paper-Scissors game where $\eta = 0.05$ for EXP3

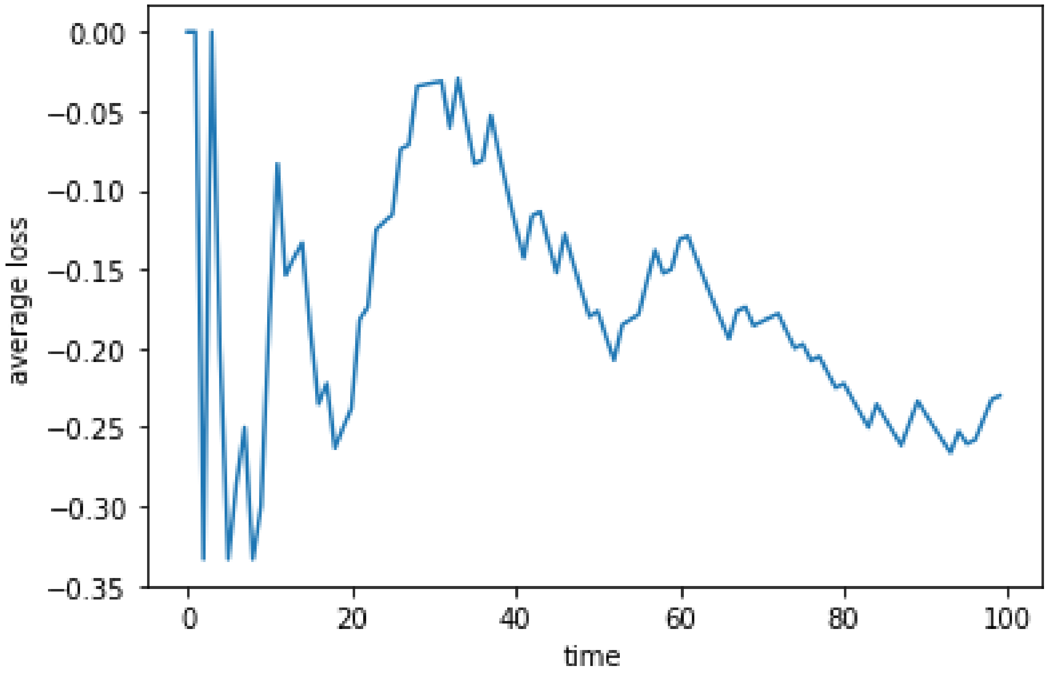

Figure 24 – Evolution of the average loss in the Rock-Paper-Scissors game where $\eta = 0.1$ for EXP3

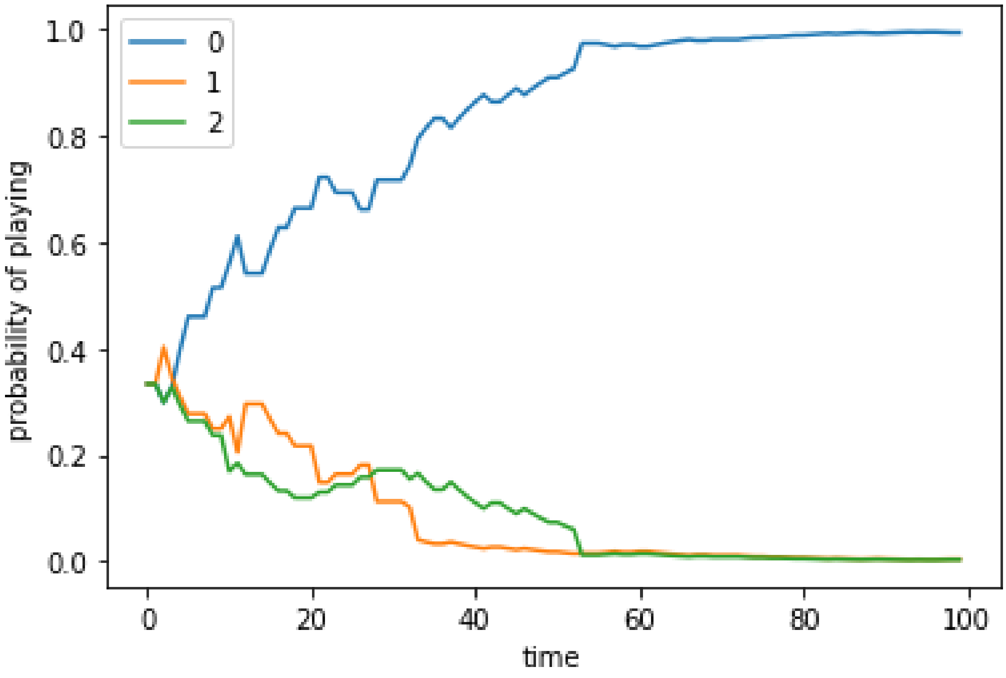

Figure 25 – Evolution of the best strategy in the Rock-Paper-Scissors game where $\eta = 0.1$ for EXP3

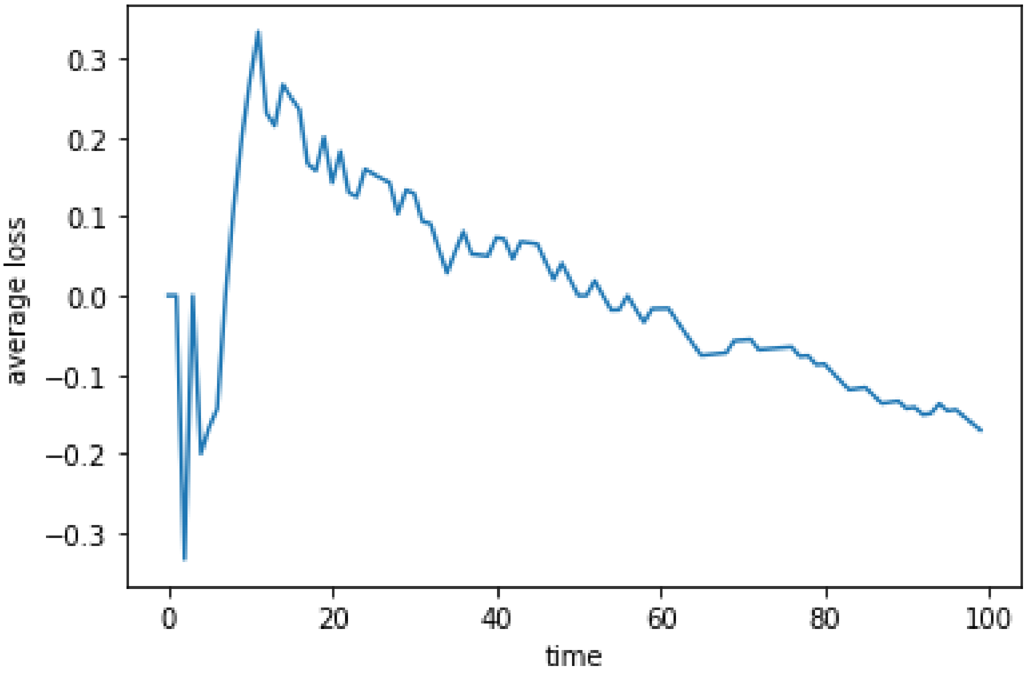

Figure 26 – Evolution of the average loss in the Rock-Paper-Scissors game where $\eta = 0.5$ for EXP3

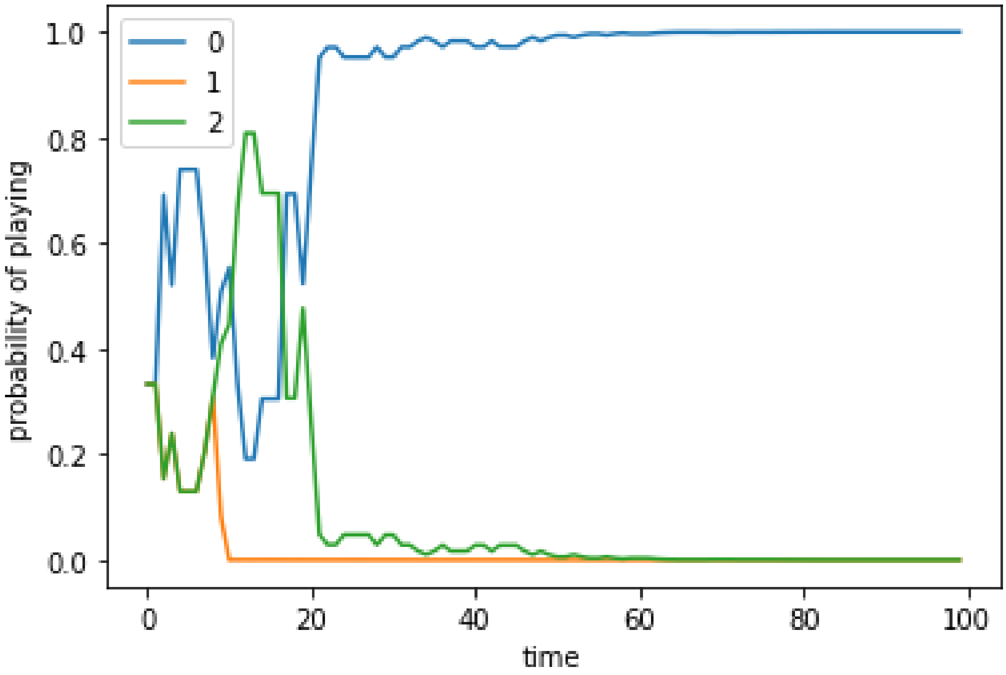

Figure 27 – Evolution of the best strategy in the Rock-Paper-Scissors game where $\eta = 0.5$ for EXP3

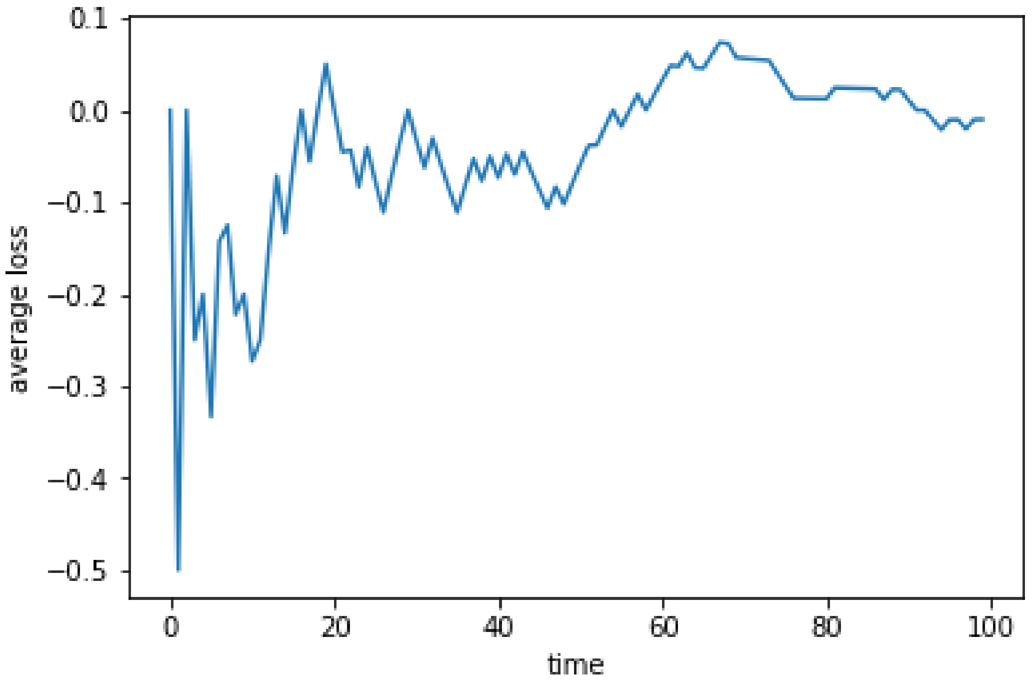

FIGURE 28 – Evolution of the average loss in the Rock-Paper-Scissors game where $\eta = 1$ for EXP3

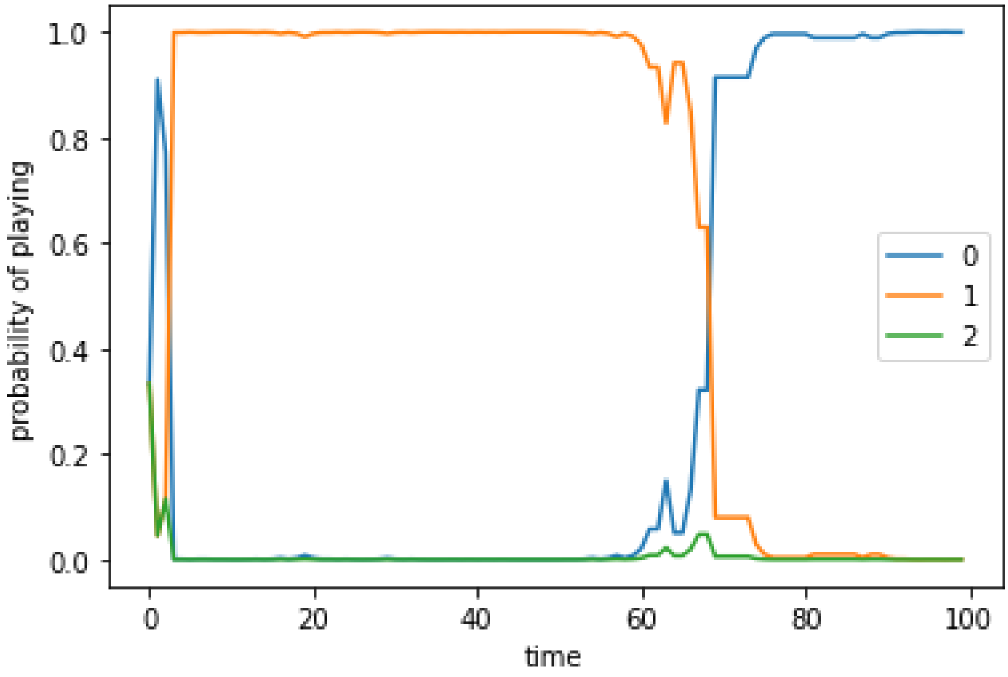

FIGURE 29 – Evolution of the best strategy in the Rock-Paper-Scissors game where $\eta = 1$ for EXP3

2 Bernoulli Bandits

2.1 Question 1 - Follow The Leader

a

Applying "Follow The Leader" here, we can identify 2 phases : as long as the player has gotten unlucky and has not gotten the reward 1 for either arm, the empirical mean of both arms is 0 and the player does not differentiate between them, and pulls the first arm with probability ρ. Once he gets the reward 1 for either one, its empirical mean becomes positive and the player only pulls that arm.

We write τ the random variable representing the duration of the "indifferent phase". Let $t \leq T$.

$$\sum_{t=1}^{T} p_{k_t} = \sum_{t=1}^{\tau} p_{k_t} + \sum_{t=\tau+1}^{T} p_{k_t}$$

$$\mathbb{E}\left[\sum_{t=1}^{T} p_{k_t} | \tau\right] = \sum_{t=1}^{\tau} (\rho p_1 + (1-\rho)p_2) + \left(\rho \sum_{t=\tau+1}^{T} p_1 + (1-\rho) \sum_{t=\tau+1}^{T} p_2\right) = T(\rho p_1 + (1-\rho)p_2)$$

and

$$\mathbb{E}[R_T] = Tp_2 - \mathbb{E}\left[\sum_{t=1}^{T} p_{k_t}\right] = \rho(p_2 - p_1)T$$

We would typically have $\rho = \frac{1}{2}$, giving $\mathbb{E}[R_T] = \frac{p_2 - p_1}{2} T$.

c

On this figure we can see the proportions of the various values of mean regret observed with FTL. This can be done easily as the mean regret is round. It looks very much like a binomial distribution, wich is coherent with the imbued randomness of the decision (and the associated reward) made in the algorithm. We also fin that said histogram is centered on $5 = \frac{0.6-0.5}{2} 100$ which is coherent with the previous theoretical result.

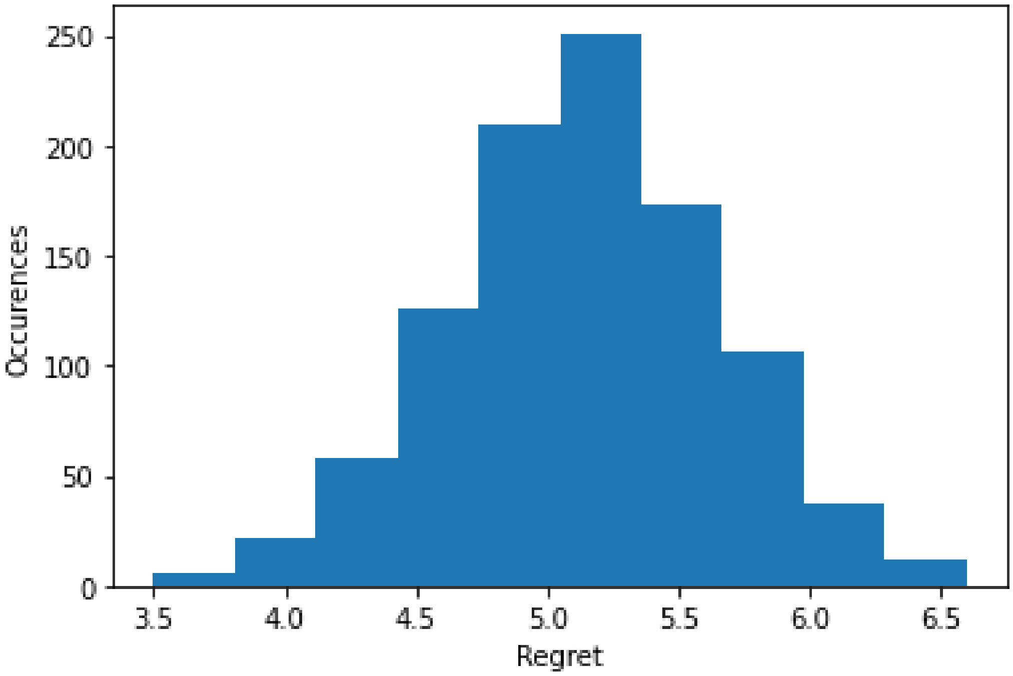

FIGURE 30 – histogram of the regret observed over 1000 repetitions of the experiment on 100 steps of the game

d

Once again, we can see in figure 31 that the mean loss converges to the theoretical value, though it does so relatively slowly : FTL is quite inconsistent, as it can find and pull the best arm instantly or never pull it at all.

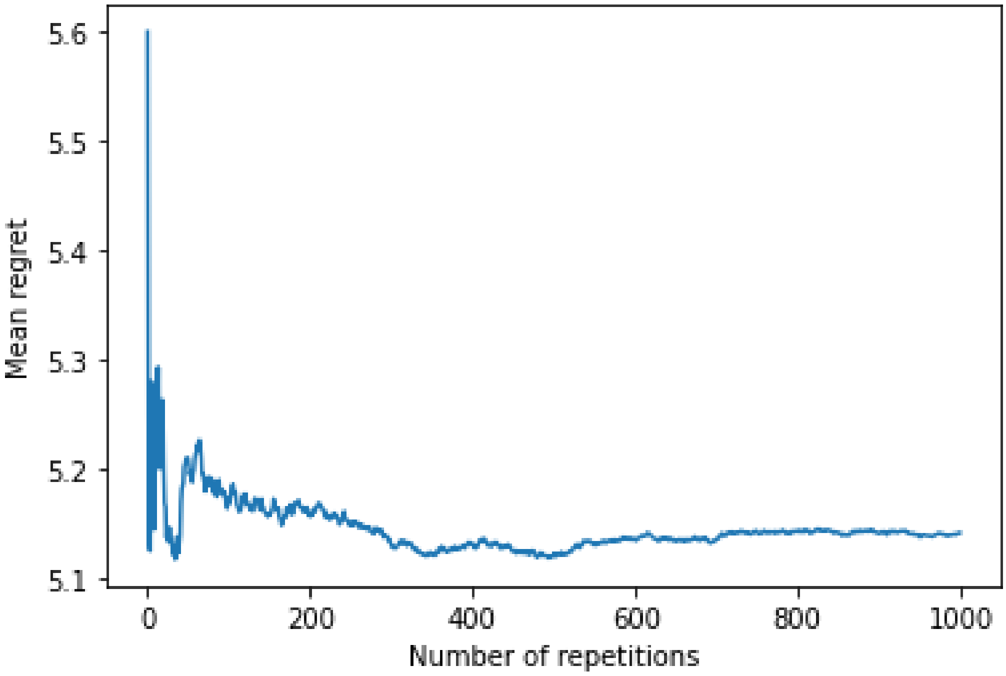

FIGURE 31 – Mean regret of the FTL algorithm averaged over 1000 repetitions as a function of time

FTL is not adapted here, as it will keep pulling the first arm that gave it a positive reward without getting sufficient information (this could be corrected through an exploration phase for example).

2.2 Question 2 - UCB

a

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[x])}\right] = (1-p)e^{-\lambda p} + pe^{\lambda(1-p)} = e^{-\lambda p}\left(1-p+pe^{\lambda}\right)$$

and

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

b

Let's suppose $\Phi''_X(\lambda) \leq \sigma^2$.

By integrating both sides from 0 to λ, and using $\Phi'_X(0) = \mathbb{E}[X - \mathbb{E}[X]] = 0$ (simple enough to verify with the expression developed question 2.4), we get $\Phi'_X(\lambda) \leq \lambda\sigma^2$,

Again from 0 to λ, using $\Phi_X(0) = \log 1 = 0$, $\Phi_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$.

Finally using the exponential on both sides, we find $\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[x])}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}$

c

Thanks to the result in the previous question, we only need to find σ^2 such that $\phi''_X(\lambda) \leq \sigma^2$ with X being a binomial random variable of parameter λ.
Thanks to question a, we have computed that

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

$$\Phi'_X(\lambda) = \frac{e^\lambda p}{1-p+e^\lambda p} - p$$

$$\Phi''_X(\lambda) = \frac{e^\lambda p(1-p+e^\lambda p) - (e^\lambda p)^2}{(1-p+e^\lambda p)^2} = p(1-p)\frac{e^\lambda}{(1-p+e^\lambda p)^2}$$

Let us denote $f(\lambda) = \frac{e^\lambda}{(1-p+e^\lambda p)^2}$.

$$f'(\lambda) = \frac{e^\lambda(1-p+pe^\lambda)^2 - e^\lambda 2(1-p+pe^\lambda)pe^\lambda}{(1-p+pe^\lambda)^4} = \frac{e^\lambda(1-p-pe^\lambda)}{(1-p+pe^\lambda)^3}$$

Thus, f is increasing on $(-\infty; \log\frac{1-p}{p}]$ and decreasing on $[\log\frac{1-p}{p}; +\infty)$

We deduce that Φ''_X is maximal on $\lambda^* = \log\frac{1-p}{p}$ and for all σ^2 such that $\sigma^2 \geq \Phi''_X(\log\frac{1-p}{p}) = \frac{1}{4}$, X is σ^2-sub-gaussian.

d

Let $x \in [0,1]$ and $\lambda \in \mathbb{R}$. Using the convexity inequality,

$$e^{\lambda x} = e^{x\lambda+(1-x)0} \leq xe^\lambda + (1-x)e^0 = 1-x+xe^\lambda$$

$$\frac{\mathbb{E}[e^{\lambda(X-p)}]}{\mathbb{E}[e^{\lambda(Y-p)}]} = \frac{\mathbb{E}[e^{\lambda(X-p)}]}{e^{-\lambda p}(1-p+pe^\lambda)} = \mathbb{E}\left[\frac{e^{\lambda X}}{1-p+pe^\lambda}\right] \leq \mathbb{E}\left[\frac{1-X+Xe^\lambda}{1-p+pe^\lambda}\right] = \mathbb{E}\left[\frac{1-p+pe^\lambda}{1-p+pe^\lambda}\right] = 1$$

where the inequality is obtained using the previous point.
Then applying the log on both sides, we prove $\Phi_X(\lambda) \leq \Phi_Y(\lambda)$.

e

All random variables supported on $[0,1]$ have mean in $[0,1]$ (since all possible values are comprised in that interval and we are performing a linear combination of those). Naming X such a random variable, thanks to question (d) we know that $\phi_X(\lambda) \leq \phi_Y(\lambda)\forall\lambda \in \mathbb{R}$, where Y is a binomial variable of parameter the mean of variable X. Thanks to question (b) and (c) we know that Y is $\frac{1}{4}$-sub-gaussian, and as such verifies $\forall\lambda \in \mathbb{R},\ \ \Phi_Y(\lambda) \leq \frac{\sigma^2\frac{1}{4}}{2}$.

The inequality from the previous question gives $\forall\lambda \in \mathbb{R},\ \ \Phi_X(\lambda) \leq \frac{\sigma^2(\frac{1}{4})^2}{2}$, and thus X is $\frac{1}{4}$-sub-gaussian.

g

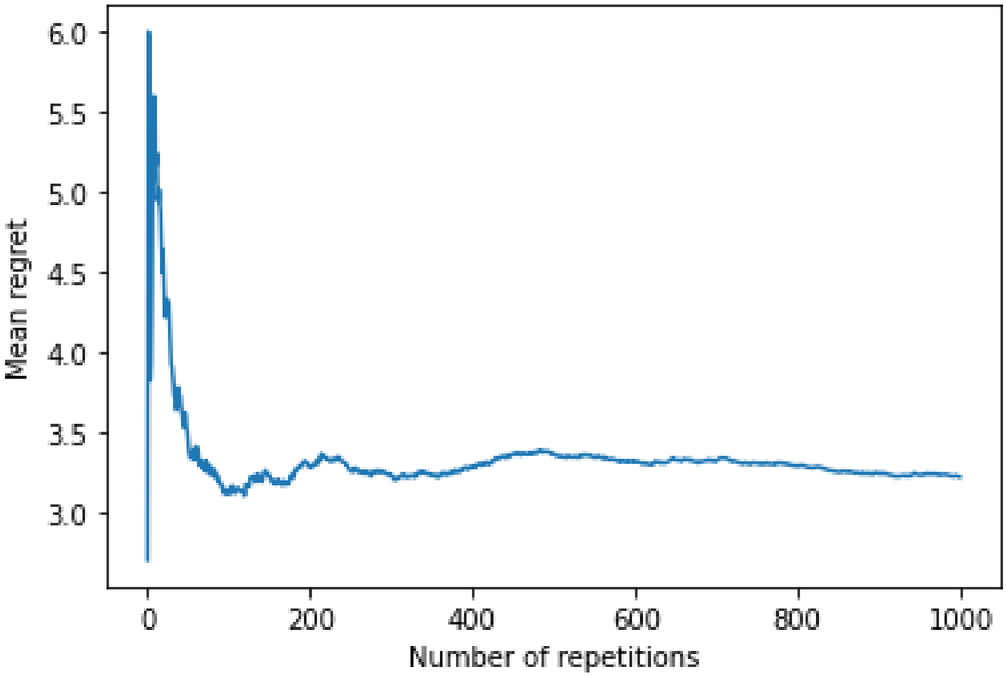

Figure 32 – Mean regret of the UCB algorithm averaged over 1000 repetitions as a function of time

The mean regret is much lower than with FTL, as UCB bases its decisions on a lot more information.

h

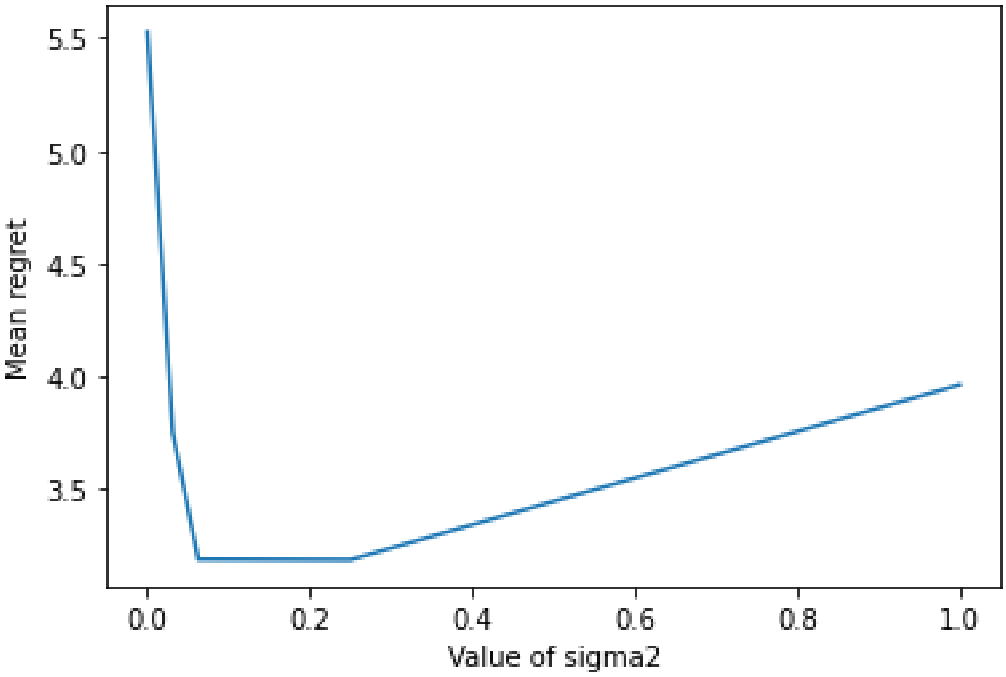

Figure 33 – Mean regret as a function of chosen parameter σ^2 for UCB for $p = (0.5, 0.6)$

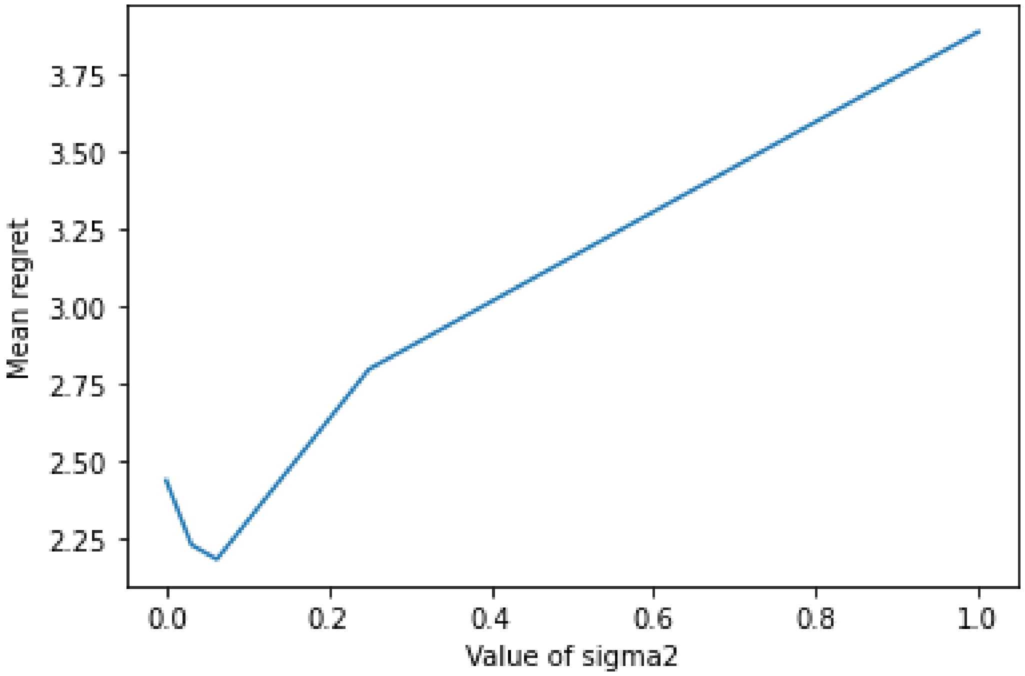

FIGURE 34 – Mean regret as a function of chosen parameter σ^2 for UCB for $p = (0.85, 0.95)$

It seems that the best σ^2 for $p = (0.5, 0.6)$ is close to 0.25, whereas for $p = (0.85, 0.95)$ it is close to 0.

This is confirmed by the following question, as the first configuration as both arms close to $\frac{1}{2}$, and the second one close to 1.

2.3 Question 3

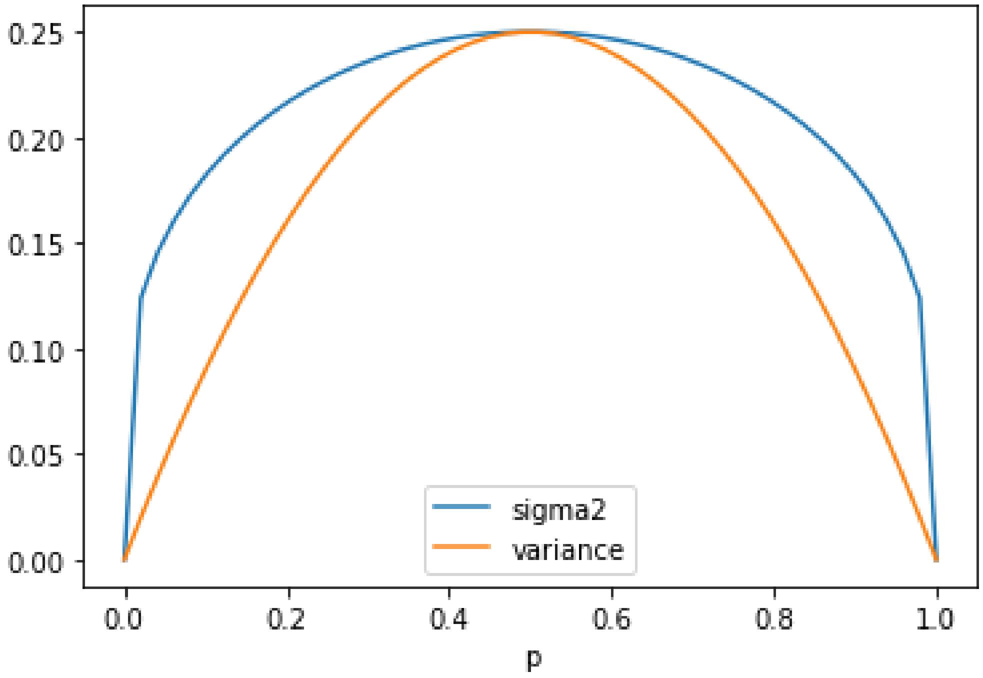

FIGURE 35 – Values for the best σ^2 and variance $p(1-p)$ as functions of p

2.4 Question 4

Let $\lambda \in \mathbb{R}$

$$\Phi'_X(\lambda) = \frac{\mathbb{E}\left[(X - \mathbb{E}[X])e^{\lambda(X - \mathbb{E}[X])}\right]}{\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right]}$$

$$\Phi''_X(\lambda) = \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2 e^{\lambda(X - \mathbb{E}[X])}\right]}{\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right]} - \left(\frac{\mathbb{E}\left[(X - \mathbb{E}[X])e^{\lambda(X - \mathbb{E}[X])}\right]}{\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right]}\right)^2$$

$$\Phi''_X(0) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] - \left(\mathbb{E}\left[(X - \mathbb{E}[X])\right]\right)^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

So if we have $\forall \lambda \in \mathbb{R}, \;\; \Phi''_X(\lambda) \leq \sigma^2$, we can deduce in particular that $\mathbb{E}[(X - \mathbb{E}[X])^2] \leq \sigma^2$

2.5 Question 5 - UCB-V

a

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

b

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

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

This formulation avoids computationally intensive calculations to update the empirical variance, and instead builds on the value of the previous iteration. That also allows the algorithm to "forget" results from more that one iteration before.

d

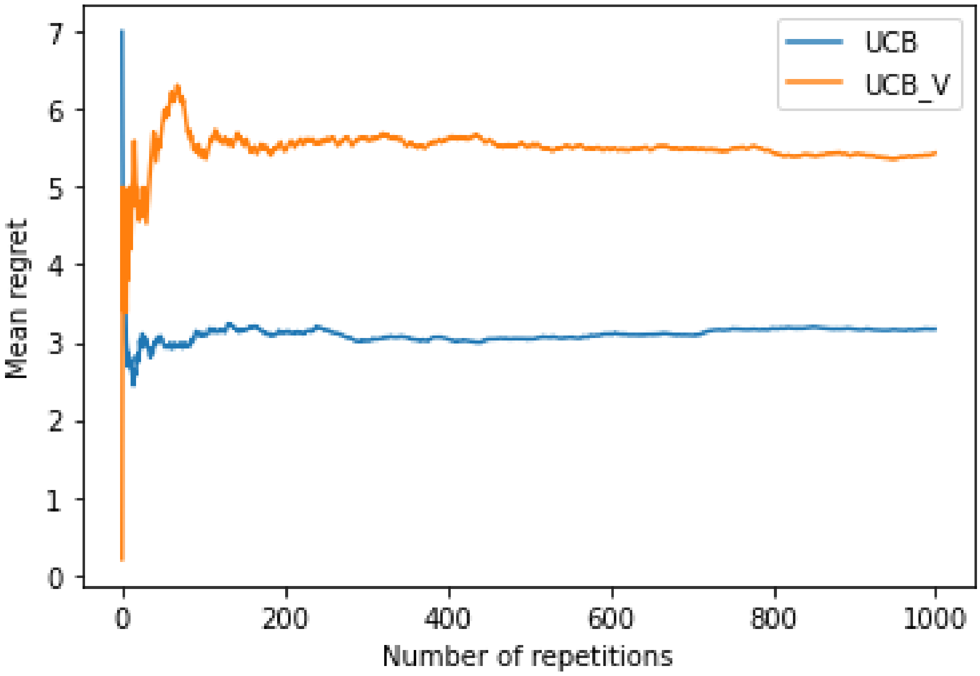

FIGURE 36 – Mean regret of UCB and UCBV for $p = (0.5, 0.6)$

Here, UCB-V performs worse than UCB. This is not surprising, as we've seen that for $p = (0.5, 0.6)$, the best σ^2 is close to $\frac{1}{4}$.

e

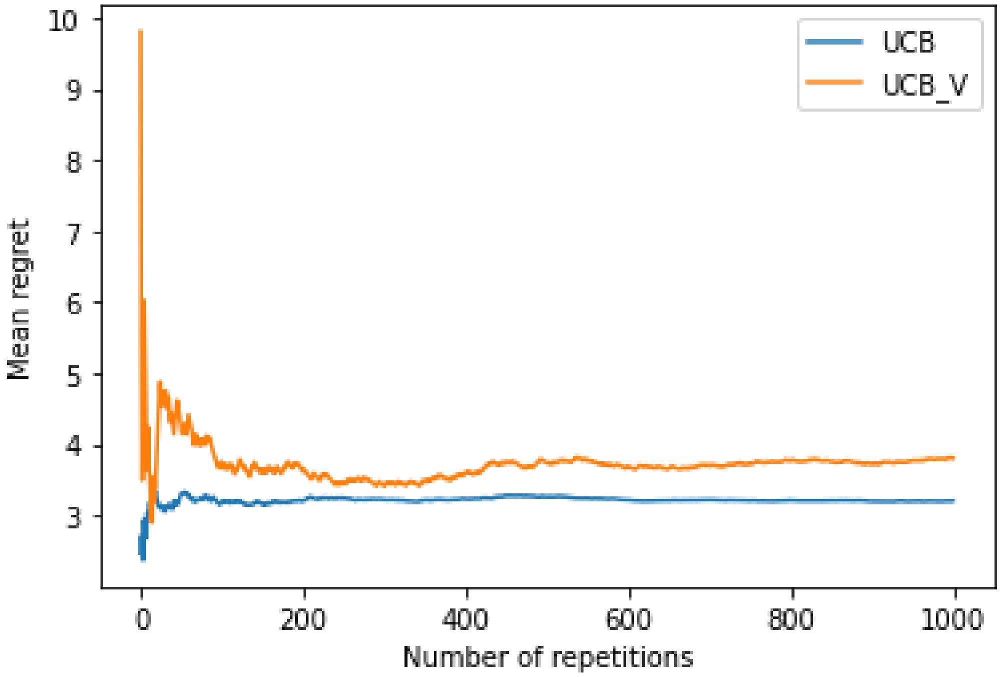

FIGURE 37 – Mean regret of UCB and UCBV for $p = (0.1, 0.2)$

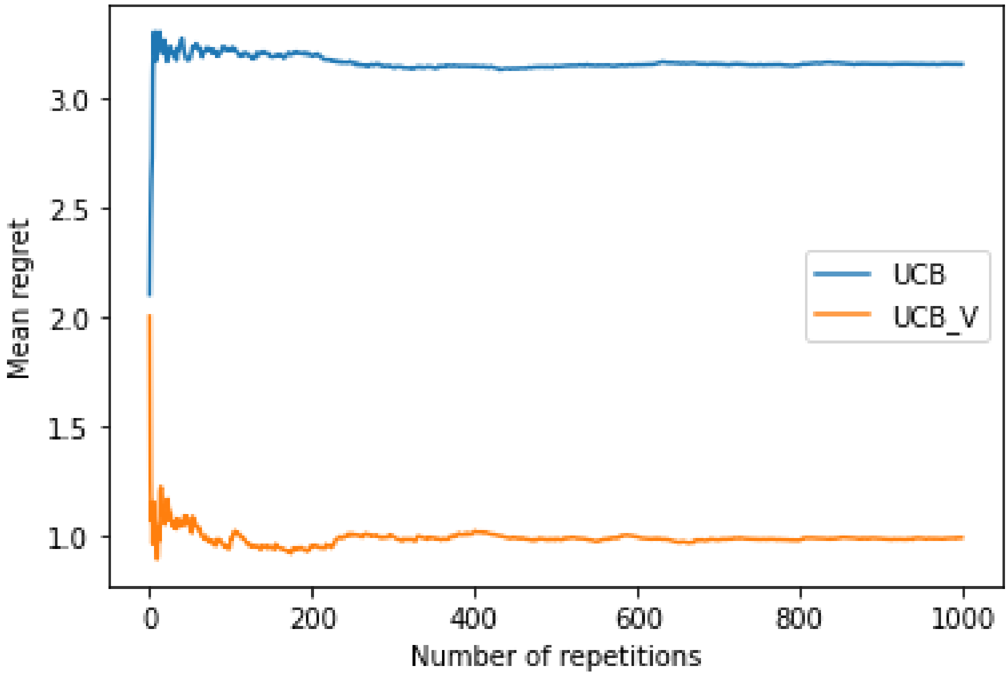

FIGURE 38 – Mean regret of UCB and UCBV for $p = (0, 0.1)$

As $\sigma^2 = \frac{1}{4}$ gets further and further away from the best σ^2, UCB-V gets progressively better than UCB, as it is able to provide a better approximation.

UCB-V also has the advantage of not requiring previous information on the distribution of the arms.