Assignment 3: N-armed bandit problem

Pierre Gérard (ULB)
INFO-F-409 - Learning dynamics

December 11, 2016

1 N-Armed Bandit

Like in the slides *Multi-Agent Reinforcement Learning*, all graph have been averaged over 2000 iterations.

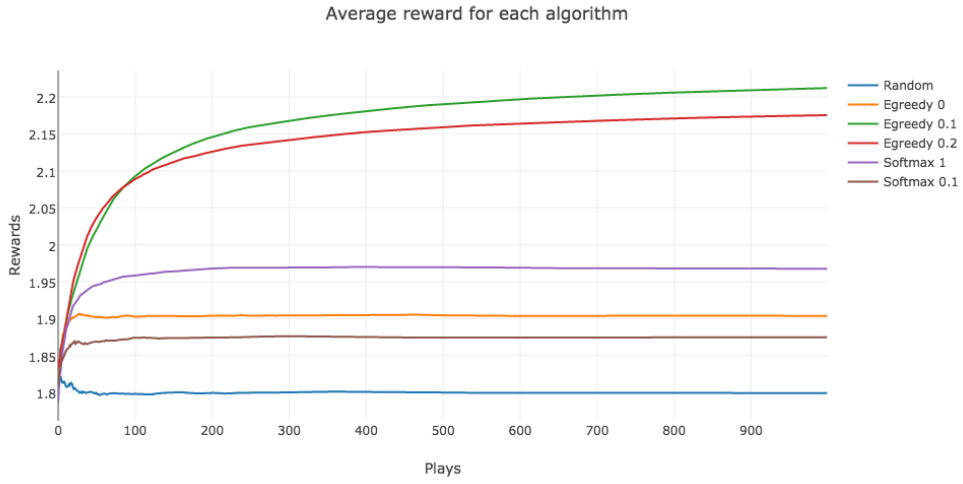
1.1 Exercice 1

1.1.1 Average reward for each algorithm

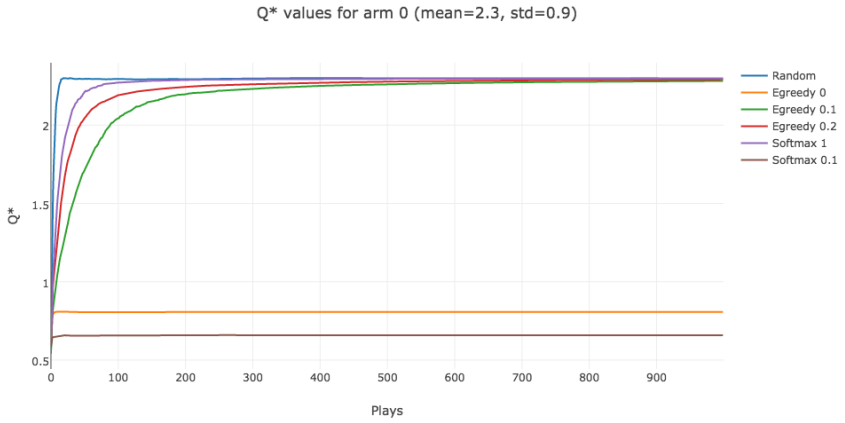
Two interpretations of this graph were possible:

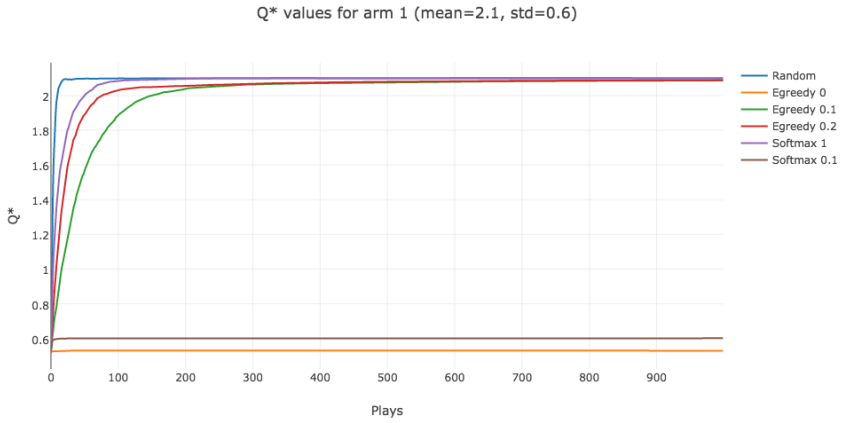
- the overall average reward for this algorithm, including the training. In other word for the play t, an average from 0 to t,
- or the average reward for this algorithm at a precise moment. In other word, what is the reward at t.

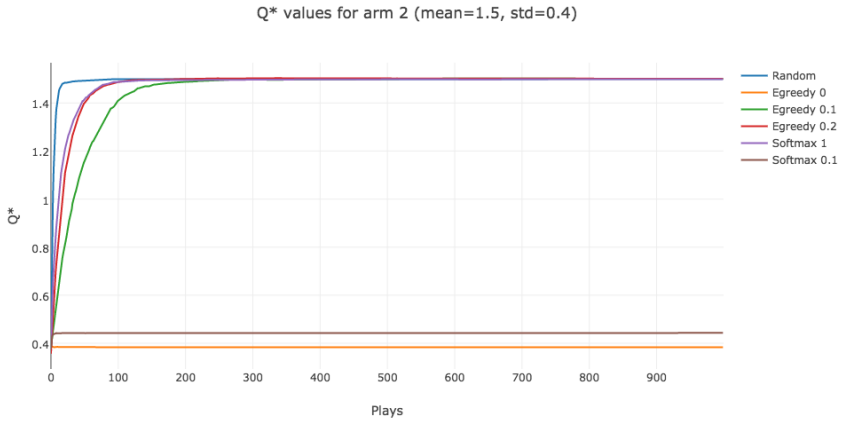
The first one has been chosen because it take into account all "early failure".

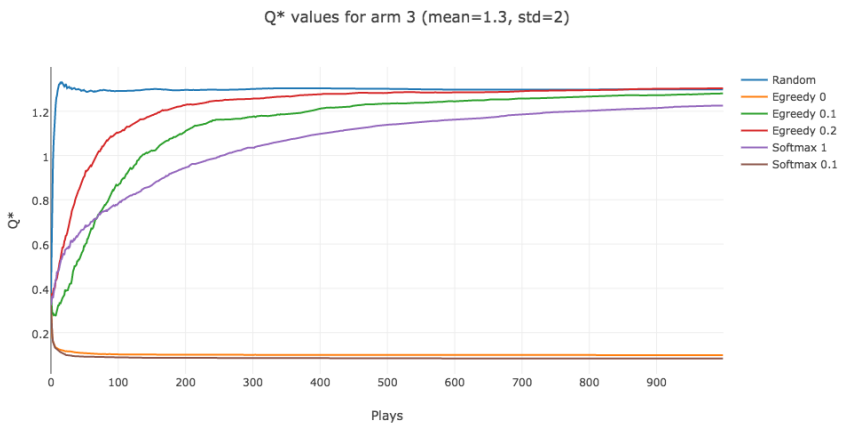


1.1.2 Plot per arm

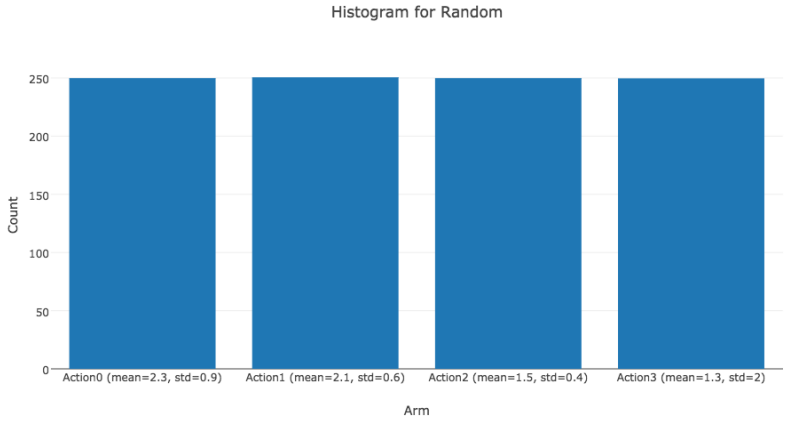


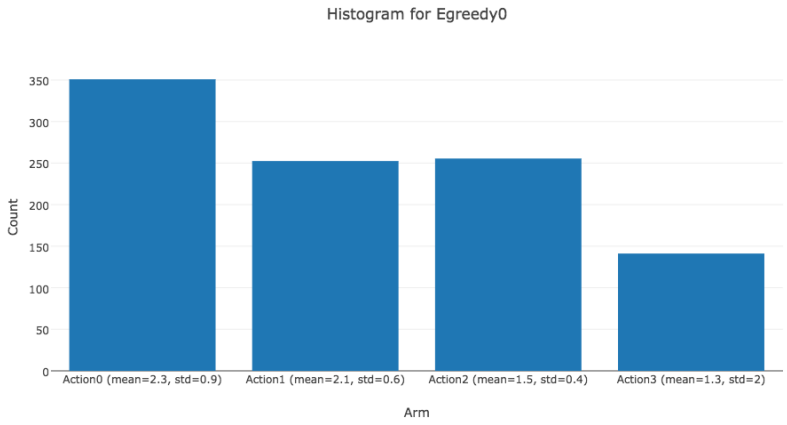


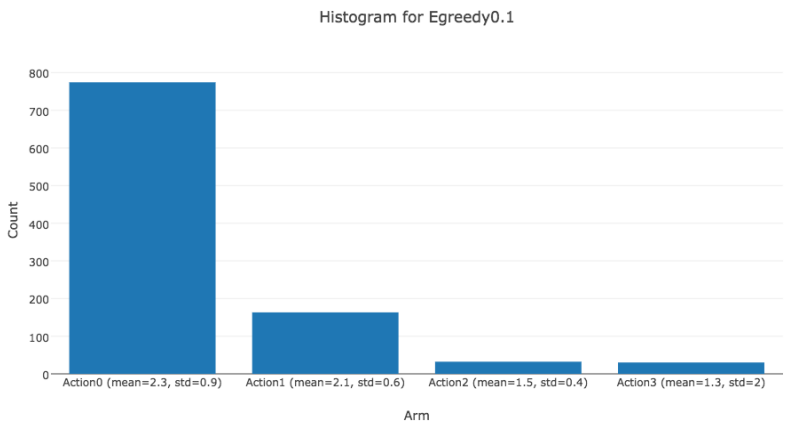


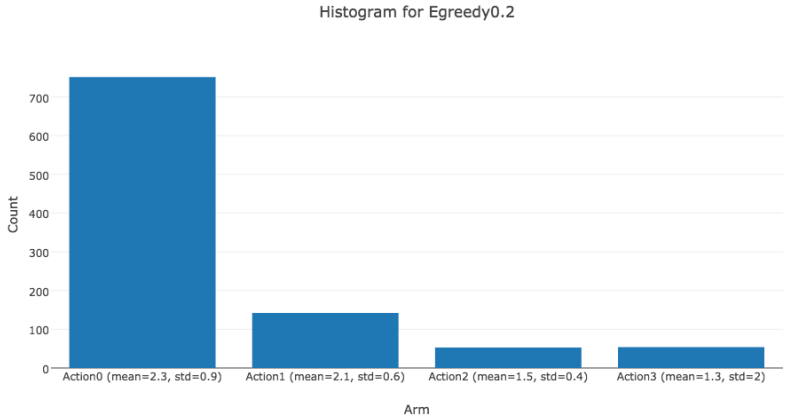


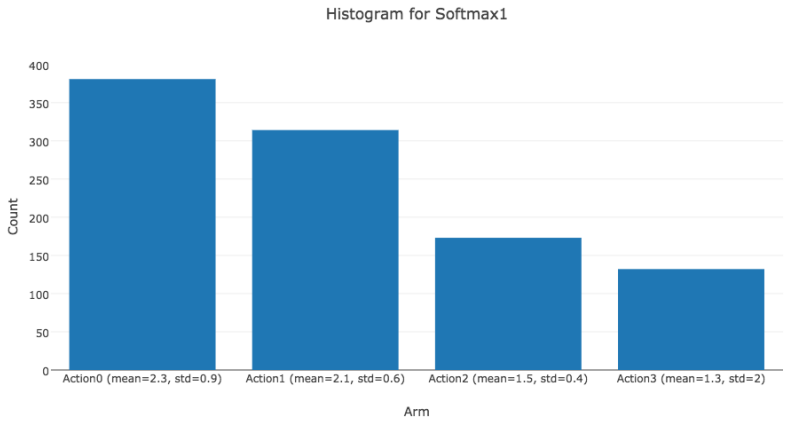
1.1.3 Histogram

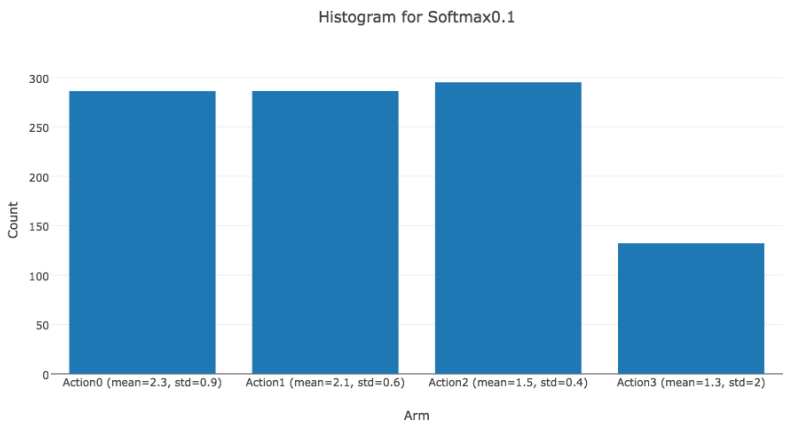





Histogram for Egreedy0.1
Count
800
700
600
500
400
300
200
100
0
Action0 (mean=2.3, std=0.9)
Action1 (mean=2.1, std=0.6)
Action2 (mean=1.5, std=0.4)
Action3 (mean=1.3, std=2)
Arm


Histogram for Egreedy0.2
Count
700
600
500
400
300
200
100
0
Action0 (mean=2.3, std=0.9)
Action1 (mean=2.1, std=0.6)
Action2 (mean=1.5, std=0.4)
Action3 (mean=1.3, std=2)
Arm


Histogram for Softmax1
Count
400
350
300
250
200
150
100
50
0
Action0 (mean=2.3, std=0.9)
Action1 (mean=2.1, std=0.6)
Action2 (mean=1.5, std=0.4)
Action3 (mean=1.3, std=2)
Arm

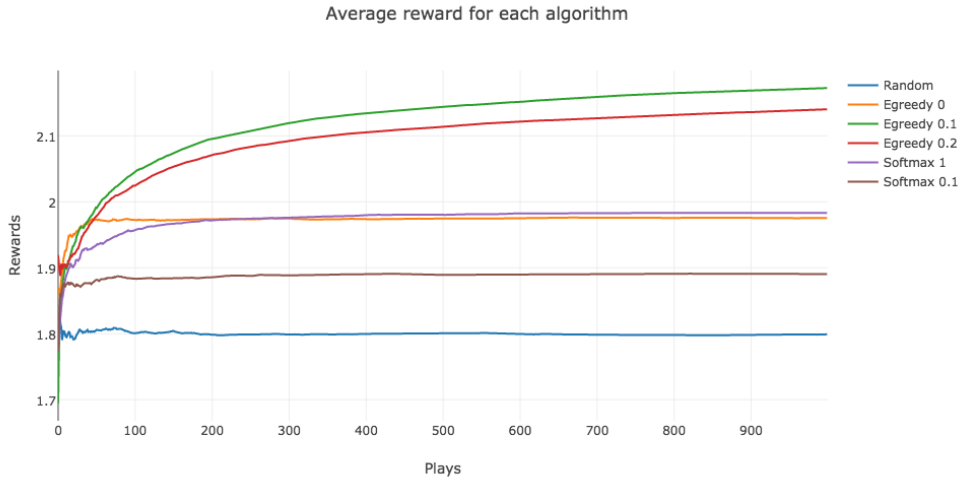


1.1.4 Results

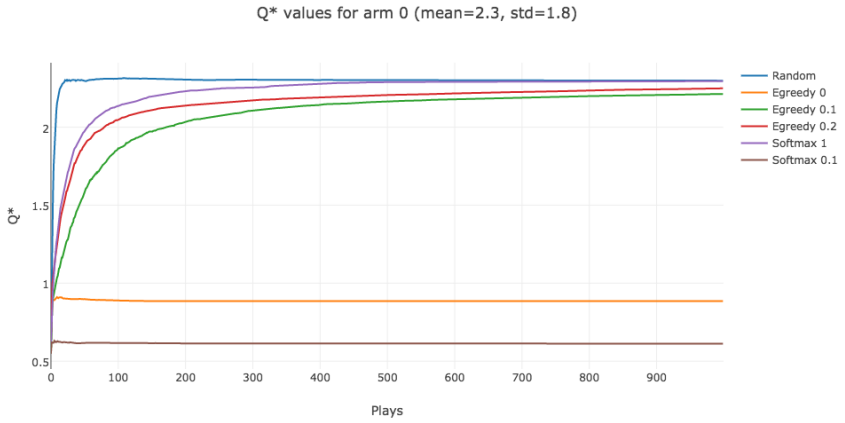
The average reward graph tends to show that the $\epsilon - greedy$ algorithm tends to choose wisely between the arms and yield the best result (except the one with $\epsilon = 0$). It is then followed by softmax with $\tau = 1$ with a poorer performance. $\epsilon - greedy0$ and $softmax\tau 1$ seems not to perform significantly better than the random selector.
The 4 graph representing the evolution of $Q(a)$ for each arm tend to show that all algorithms except $\epsilon - greedy0$ and $softmax\tau 1$ approach the true value $Q*$ in theirs estimations.
The histogram explained the result above. Without surprise, algorithms that tend to yield the best result choose better arm more often.

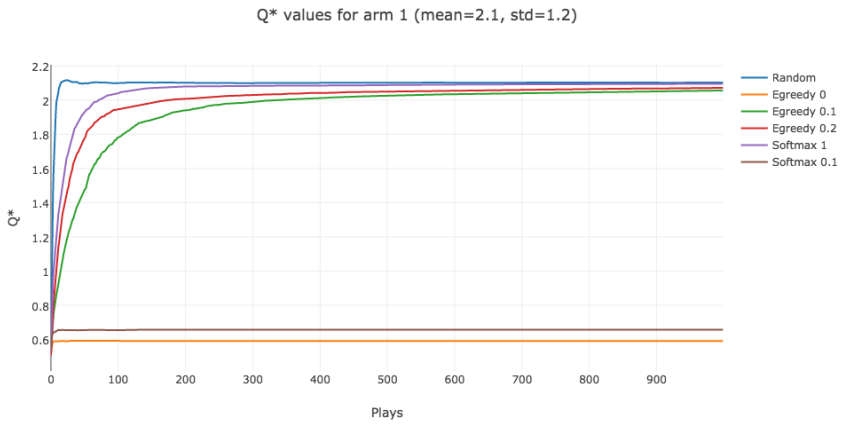
1.2 Exercice 2

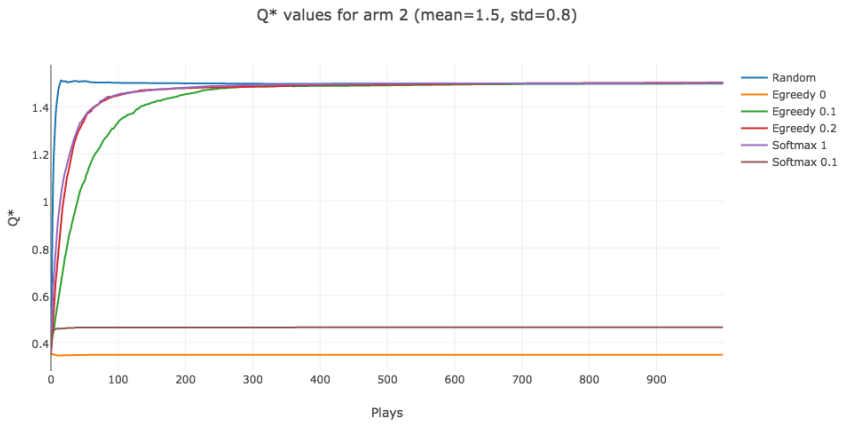
1.2.1 Average reward for each algorithm

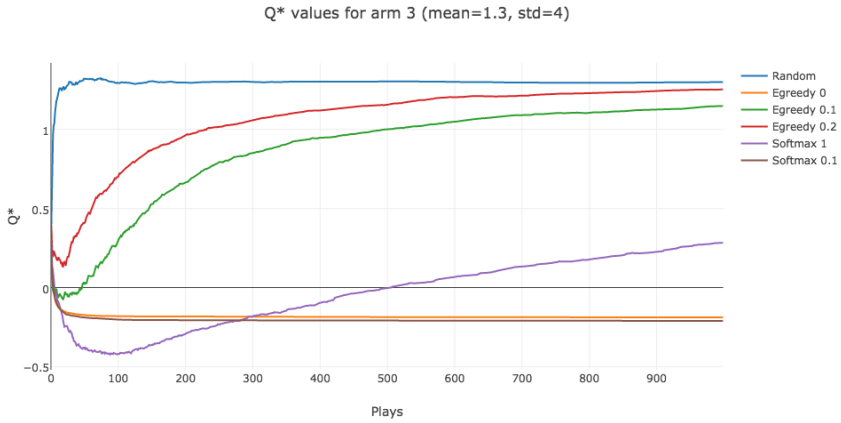


1.2.2 Plot per arm

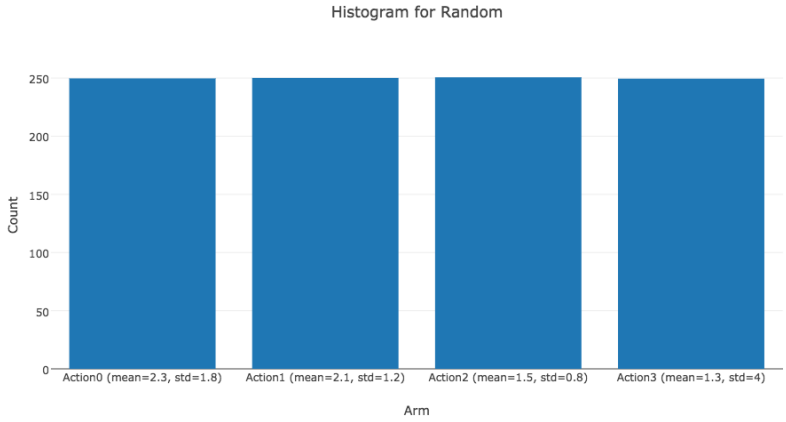


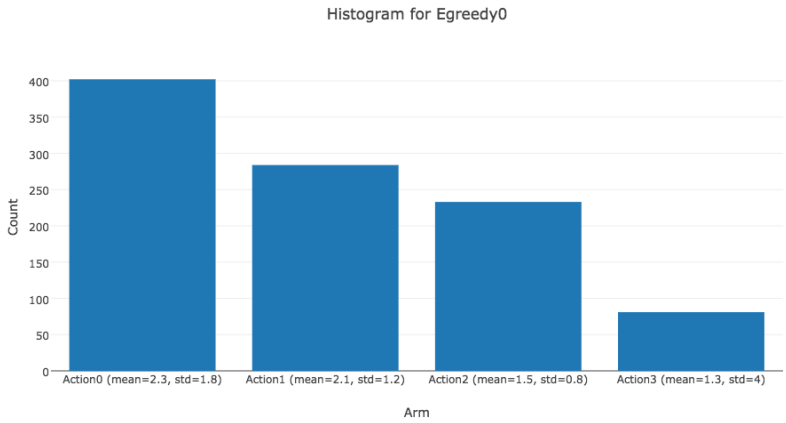


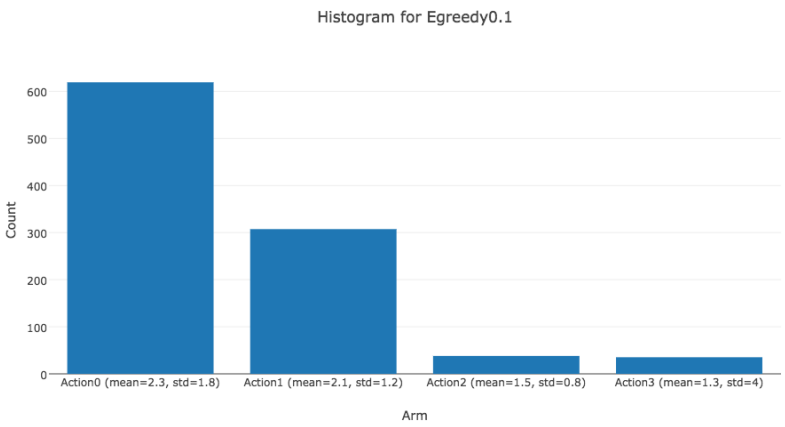


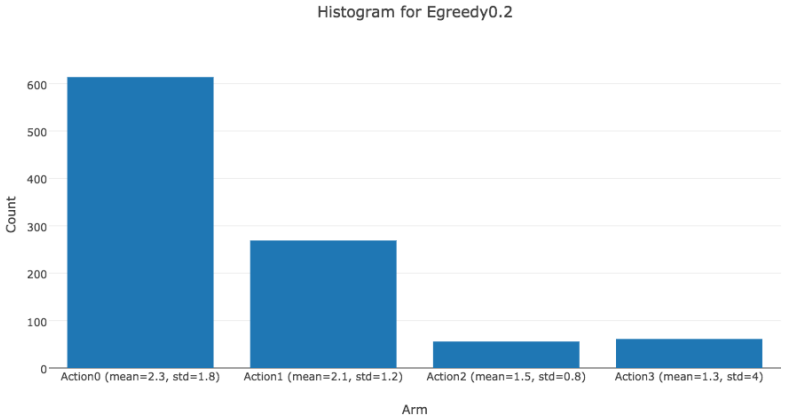


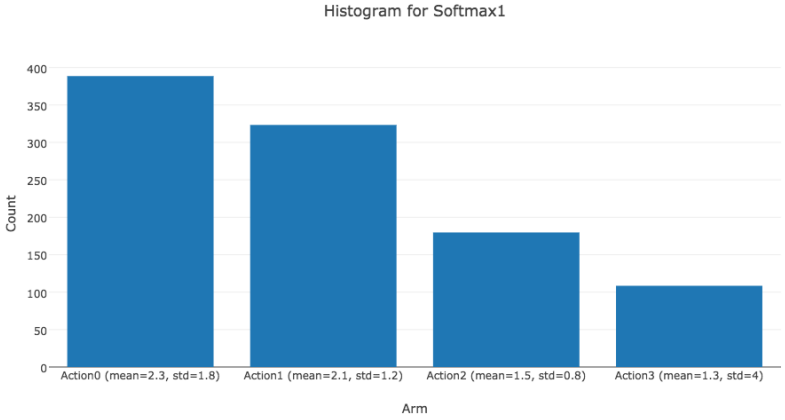
1.2.3 Histogram

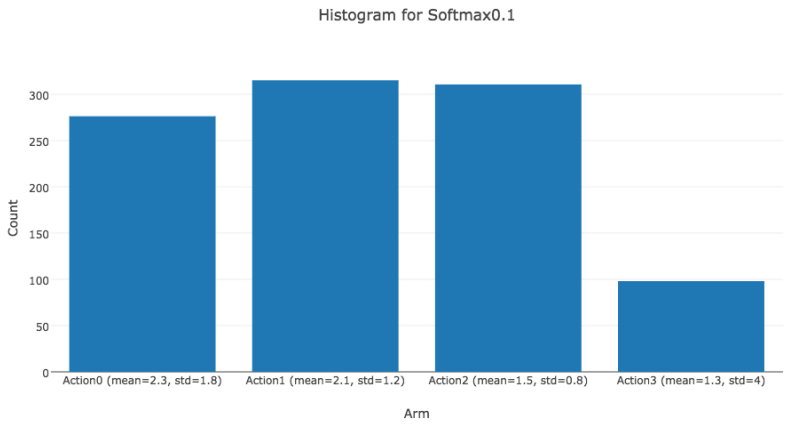





Histogram for Egreedy0.1
Count
600
500
400
300
200
100
0
Action0 (mean=2.3, std=1.8)
Action1 (mean=2.1, std=1.2)
Action2 (mean=1.5, std=0.8)
Action3 (mean=1.3, std=4)
Arm


Histogram for Egreedy0.2
Count
600
500
400
300
200
100
0
Action0 (mean=2.3, std=1.8)
Action1 (mean=2.1, std=1.2)
Action2 (mean=1.5, std=0.8)
Action3 (mean=1.3, std=4)
Arm


Histogram for Softmax1
Count
400
350
300
250
200
150
100
50
0
Action0 (mean=2.3, std=1.8)
Action1 (mean=2.1, std=1.2)
Action2 (mean=1.5, std=0.8)
Action3 (mean=1.3, std=4)
Arm

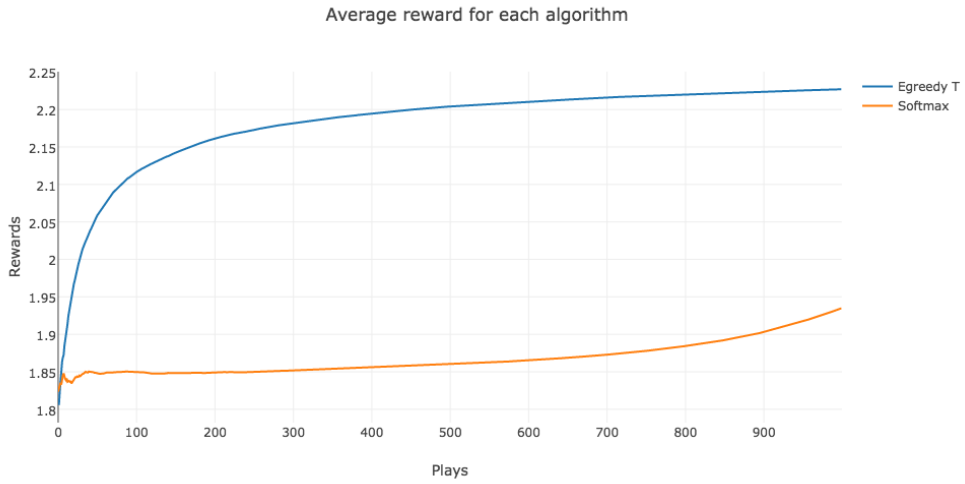


1.2.4 Results

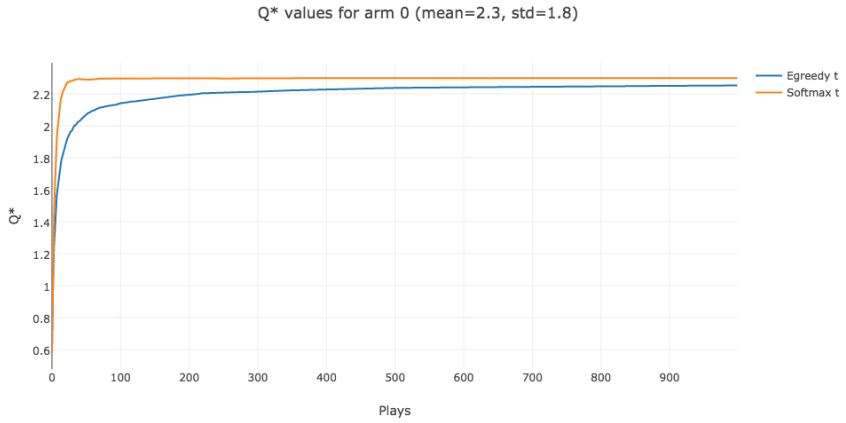
The higher variance has mainly the effect of making it harder for the algorithms to determine $Q*$ and in other word to understand which one of the arm is better. Indeed, the number of plays needed to converge is increased and thus making the average reward on all the play from the beginning slightly smaller. However, the algorithms that yielded result on small variance still yield a similar result on high variance.

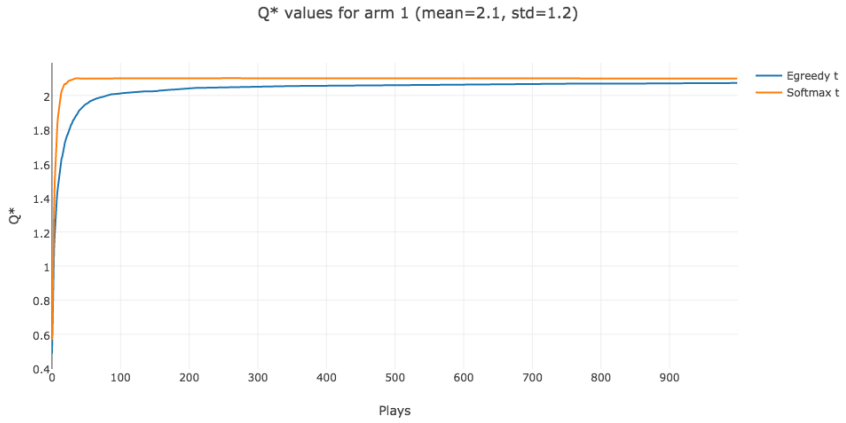
1.3 Exercice 3

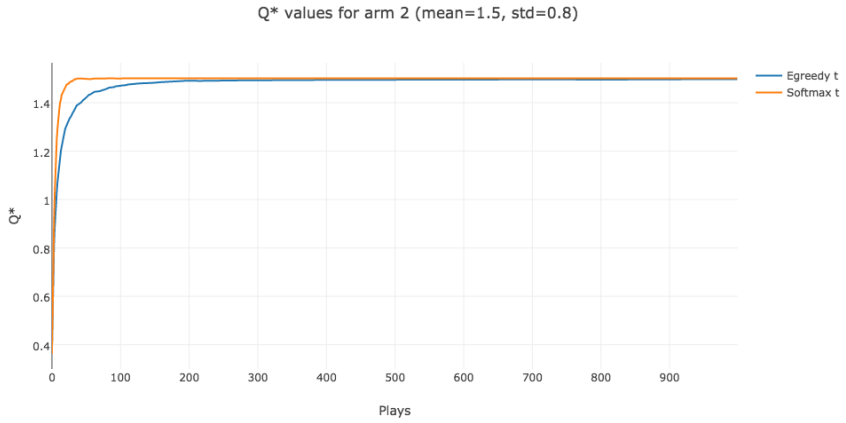
1.3.1 Average reward for each algorithm

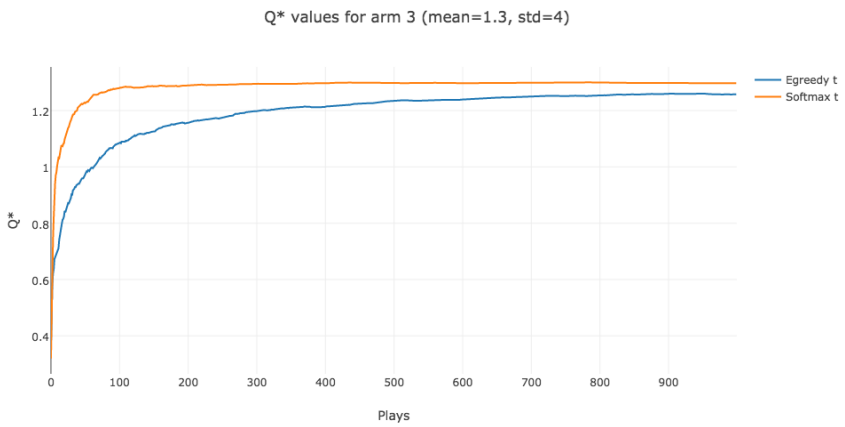


1.3.2 Plot per arm

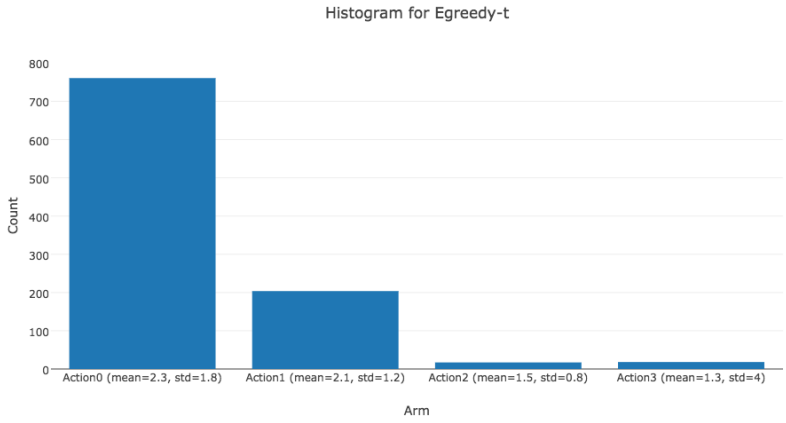


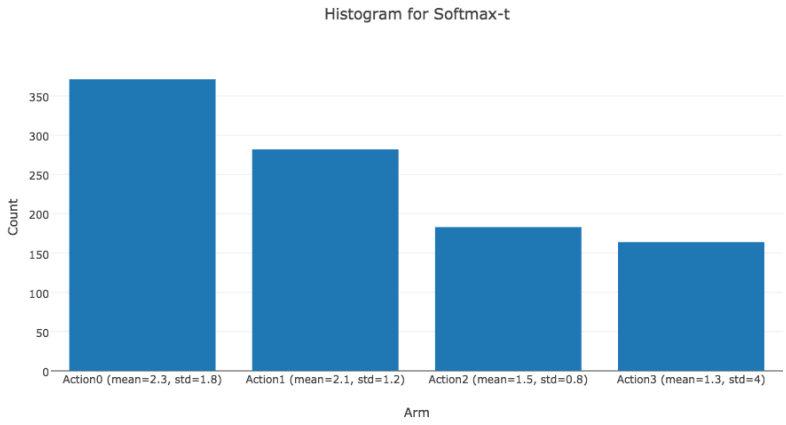






1.3.3 Histogram





1.3.4 Results

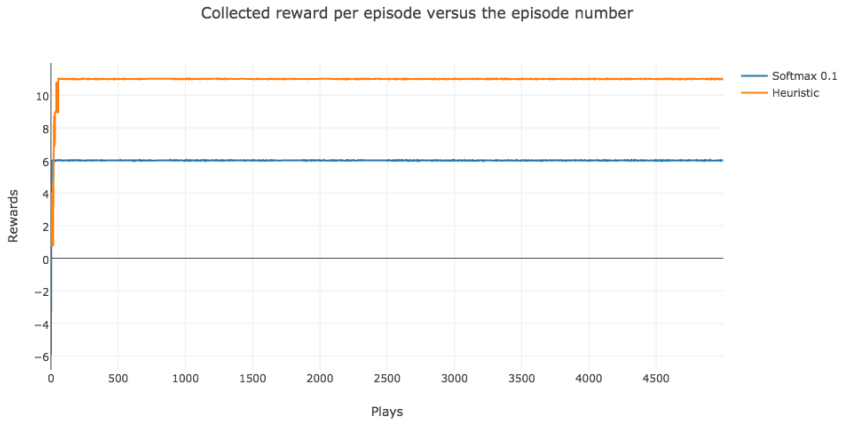
This techniques here is quite interesting because it allows to make a lot of exploration at the beginning and reduce exploring when the number of play increases. If the parameters are chosen

wisely, it would allows to "lock" the algorithms in a optimal state when reached and only exploit this optimal state.
Here we can see that it worked quite well with the $EgreedyT$ but not so well with the $softmaxT$.

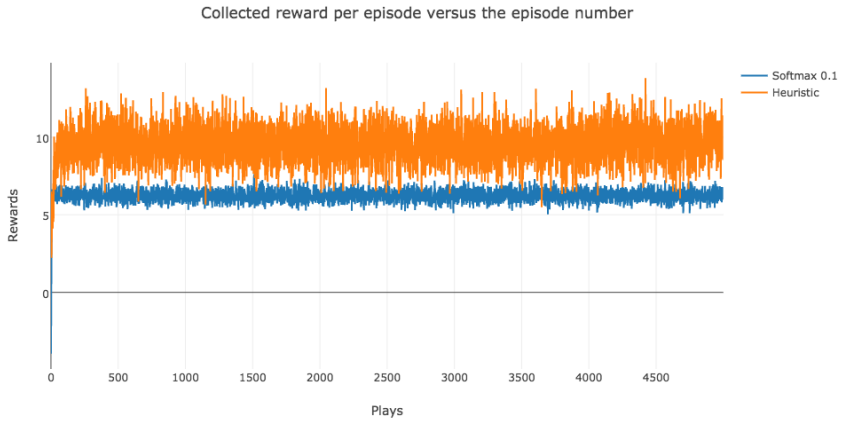
2 Stochastic Reward Game

2.1 Plot

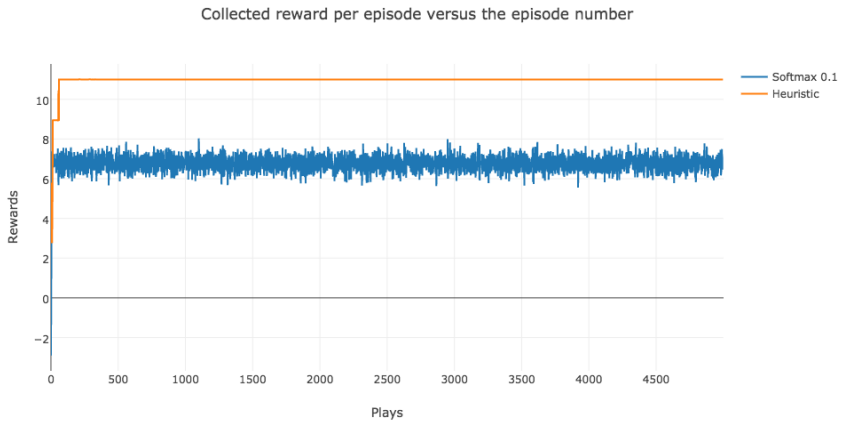
2.1.1 Std=0.2



2.1.2 Std=0.1 and std0=4



2.1.3 Std=0.1 and std1=4



2.1.4 Discussion

The temperature for the first type with simple Boltzmann action selection has been arbitrarily set. According to [1] and what we can see from the early stage on the graph, the two players begin by playing the non nash equilibrium strategy $< 3, 3 >$.Then when the exploration continues, the two player will find the more attractive but still not equilibrium $< 3, 2 >$. Finally, when exploration continus, the two player will find the non-optimal equilibrum strategy $< 2, 2 >$ and remains there forever. There a not willing to go to a strategy where the loss would be huge (-30) and thus will never reach the optimal equilirbrium $< 1, 1 >$
The heuristic is what [1] called *Combined*. It's a combined result of the boltzmann above with a proportion ρ of the MaxQ factor to bias exploration toward actions that have potential to yield higher reward (11 in our case).
By looking at the plot, the heuristic seems to be the best action to choose if one want to reach the optimal equilibrium even if the variance is high for this equilibrium.

2.2 Discussion

2.2.1 JAL vs IL

Independent learners apply Q-learning in the same way as in the bandit problem ignoring the action played by the other player. In the other hand, joint action learners apply Q-learning with the value of their own actions and also with knowledge of other agents actions.
According to [1], if we make the agents independent learner, it will mainly have to effect :

- IL would converge slower to a probability of choosing the best action (optimal equilibrium).
- IL would have a smaller probability to converge to the optimal equilibrium if the penalty are getting more negative.

2.2.2 Always select the action that yield the highest reward

This question could be interpreted in more than one way since the part *"the action that according to them will yield the highest reward"* is not clearly defined. The way is what understand is that one player stop or don't do exploration and fixed his action. In that case the other player will

choose the best response, his best action possible for the action selected by the other player. There is in this case, no garantee of convergence to an equilibrium (e.g. the player 1 choose to play only 3, then player 2 will choose to also play 3 and they will be sticked in $< 3, 3 >$ which is not a equilibrium.

References

[1] Caroline Claus and Craig Boutilier. *The dynamics of reinforcement learning in cooperative multiagent systems*. s 746. 1998, p. 752.