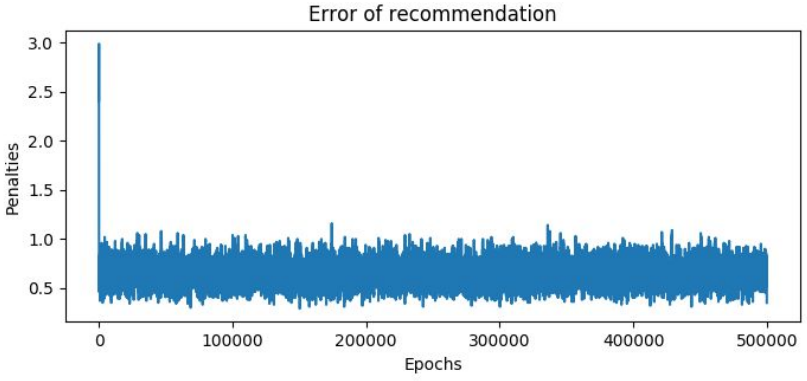
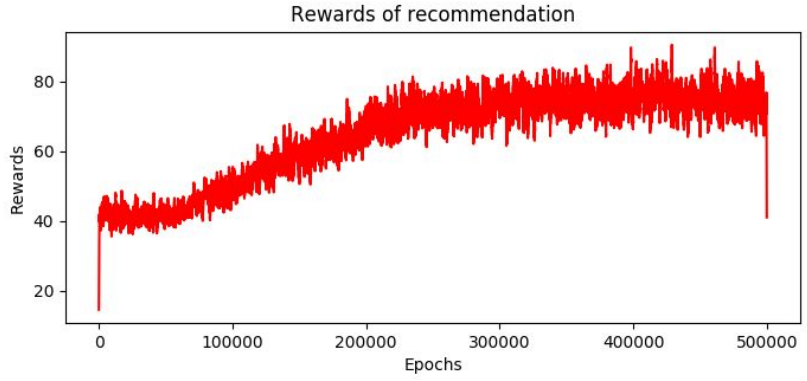
Report 14/04 : Reinforcement learning for Cache-Friendly Recommendations

Jade Bonnet and Clément Bernard

0) Previous mistakes

One mistake was the plot of rewards. Indeed, we only added the rewards when it was -5 (that corresponded to the penalty). We changed that and then used a variable window to average and filter the penalties and rewards. Here is the result (gamma = 0.9) :

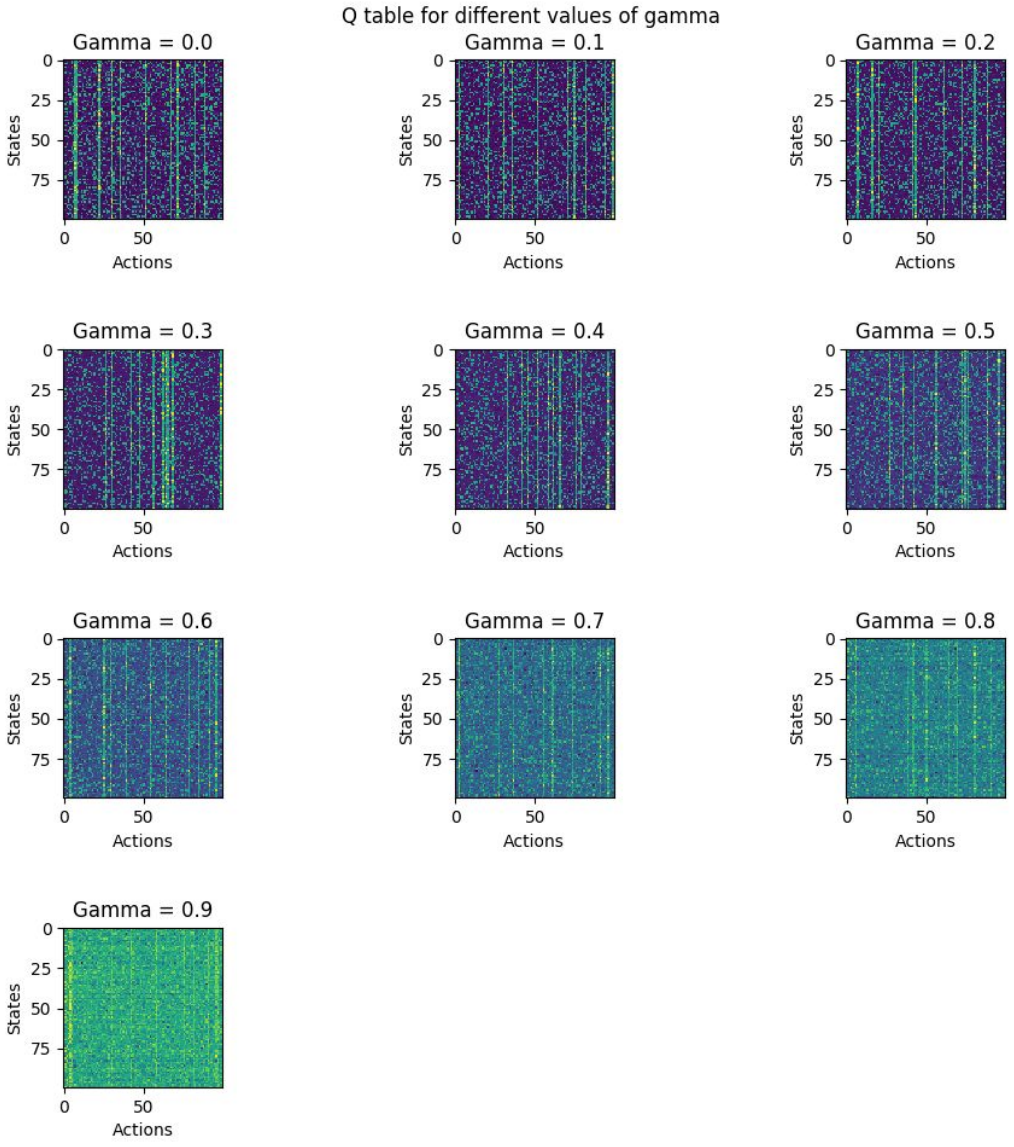




Penalties and rewards for gamma = 0.9

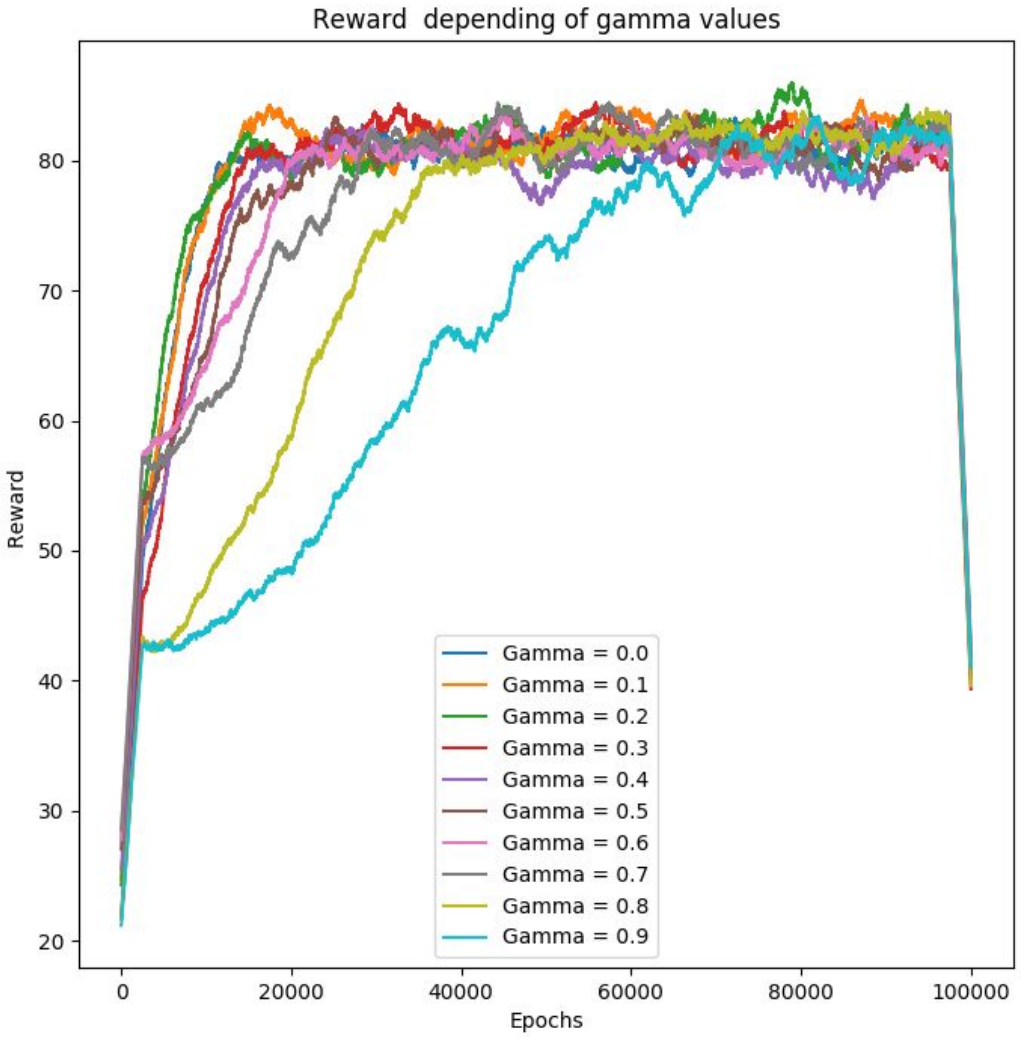
1) **Try to change the value of gamma**

Here is our result when we implement 100 000 epochs for different values of gamma.



Q_tables for different gamma values for 100 000 epochs

What we don't see is that, for instance with gamma = 0.9, it will converge but with a lot more epochs. The final values will also be different (that is to say it won't be like for gamma = 0 where the q_table values are between -5 and 10).

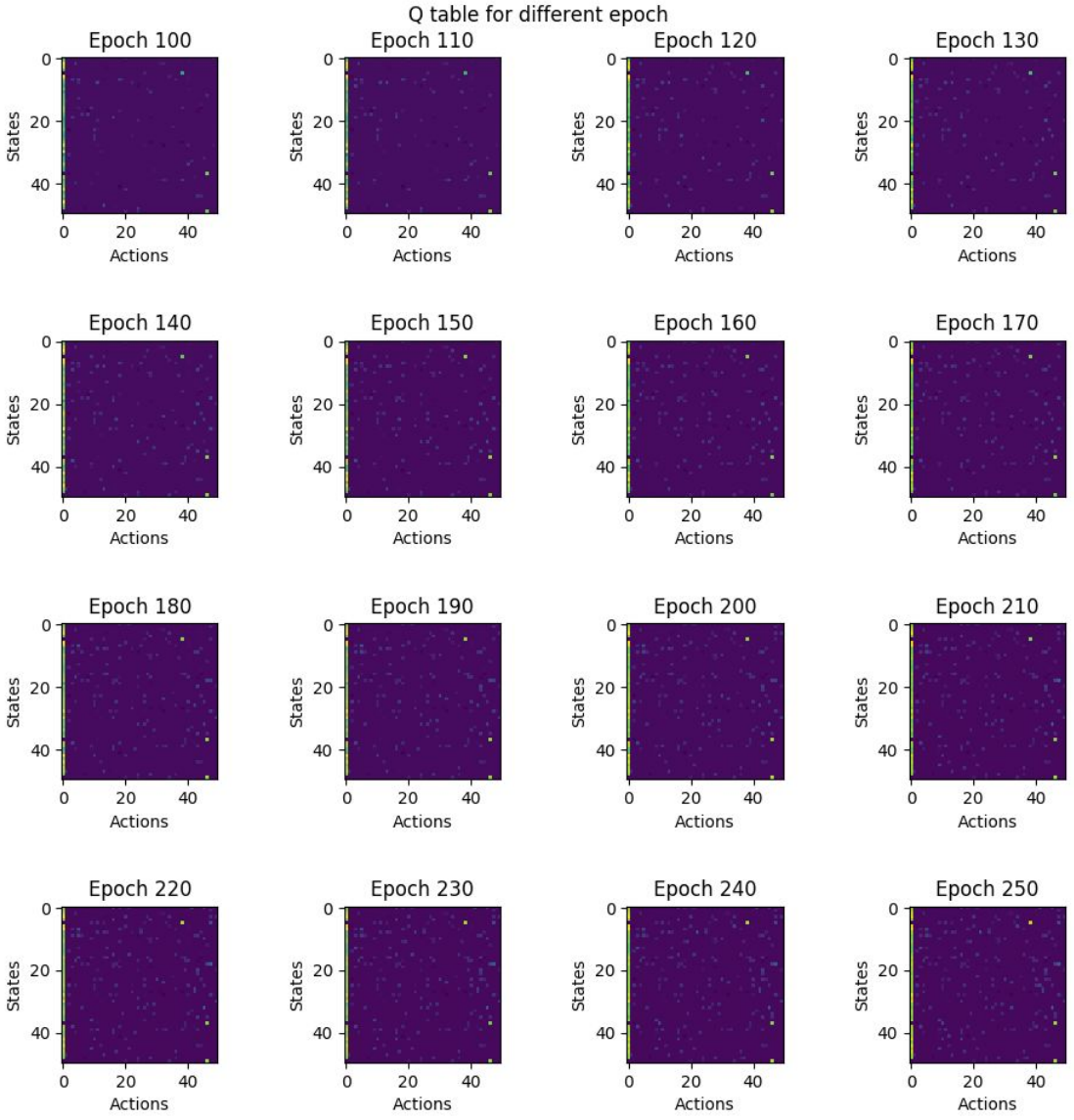


Averaged rewards for 100 000 epochs with different gamma values

We can see with this graph that the rewards converge slowly for high values of gamma.

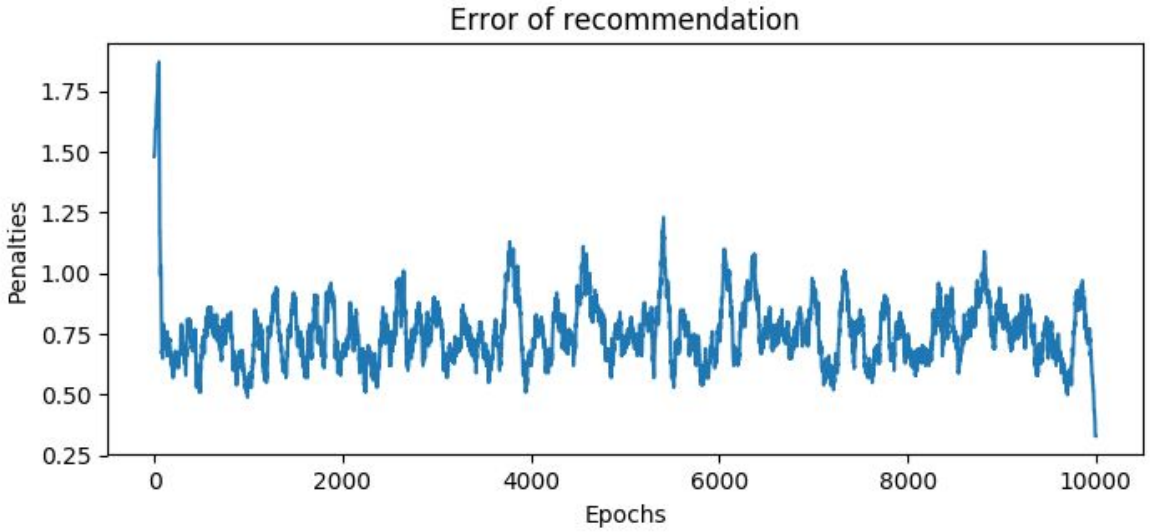
2) Try to see why the penalties and rewards converge around 200 epochs

We tried to see with a plot if our algorithm can fix quickly which action is best (with 200 epochs) and then if it will only increase the values of the good actions and decrease those of bad actions. Nevertheless, we didn't see this phenomenon as soon as 200 epochs. Here is the plot :



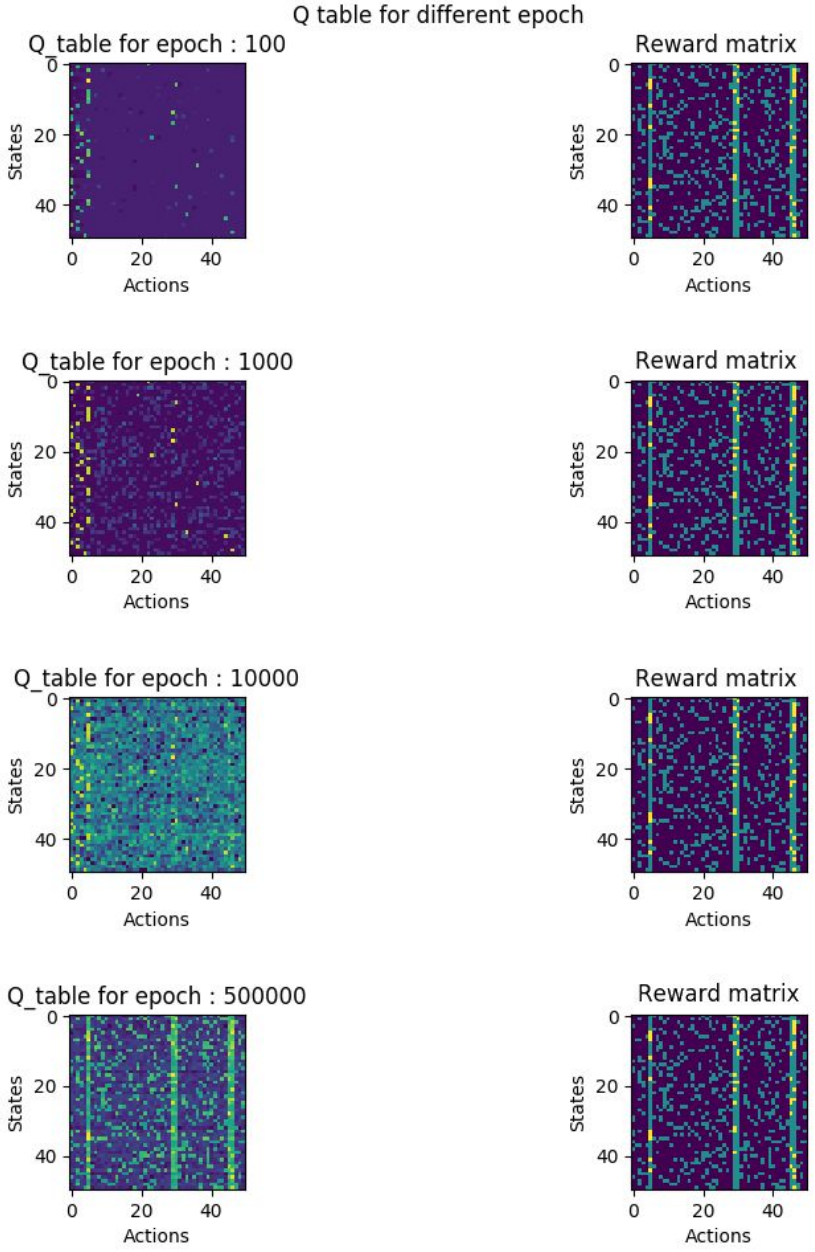
Q tables for different low epochs

We still have this plot of penalty :



Penalty for 10 000 epochs by averaging with window of size 100

Therefore we tried to compare the tables with the “reward table” (q_table obtained for gamma = 0).
Nevertheless for an epoch of 100, we can’t see really the pattern of the reward table.
We aim to continue to search on this point for next week.



Comparaison q_table with reward matrix

3) **Try to see if actions with currently bad rewards can be recommended if it will lead to good futur rewards**

We didn't play with it yet.

4) **Try to change the values of rewards**

This was not our priority so we didn’t play with it.