Reinforcement Learning TP1

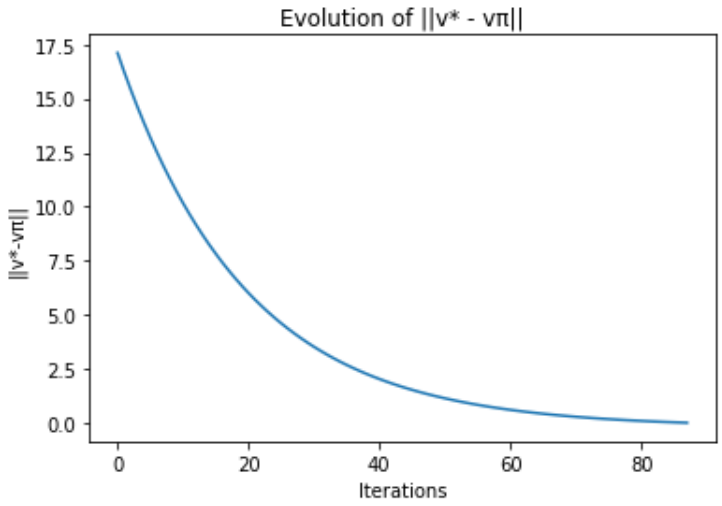
Matthis Maillard

November 2018

1 Exercice 1

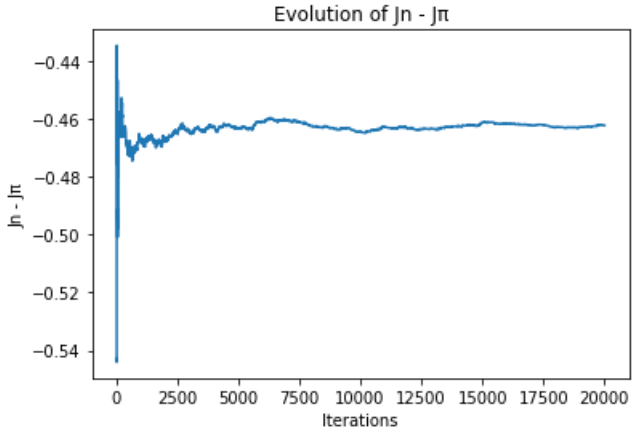
1. The optimal policy is [a1, a1, a2] because the cumulative reward will be maximized if the process is at state s2 and does action a2. Although at state s2, the reward to do action a0 is higher than for a2, as the rewards at state s1 are all 0, the average reward of two iteration is better if the action are a2, a2 than a0, a1.

2. The value iteration algorithm converges in 87 iterations.



3. The policy iteration converges in 3 iterations, so it is far less than for value iteration. The computation time for value iteration is from 3 to 4 times bigger than for policy iteration. The value iteration algorithm is easy to compute but it only converges to an approximate solution of v* while the policy iteration converges to the exact solution and in a small number of iteration. In this problem, the size of states and actions is small but in bigger ones, the computation of inverting the matrix in policy iteration is more problematic and it would increase the computation time.

2 Exercice 2



1. We can see in the figure above that J^n does not converge to $J*$ but the difference seems to converge to a fixed value -0.46. Hence, we can guess that the policy of going to the right when we can and up when we can't is not the optimal policy.

2. I choose to do 10 000 episodes to determine the optimal policy with an epsilon equal to 0.1, so that for 1 iteration out of 10, the next action is drawn uniformly in the possible actions. The learning rate is $\alpha(x_t, a_t) = \frac{1}{N(x_t,a_t}$.We can see in the results below that the policy obtained is the correct one. $||v^* - v^\pi||_\infty$ would tend to zero if we had done more episodes. The cumulative reward quickly looks like a line, meaning that the reward is almost always the same in every episode. As the rewards can only be 1, 0 or -1, we can conclude that the algorithm almost never reaches the box with negative reward and always the positive one.

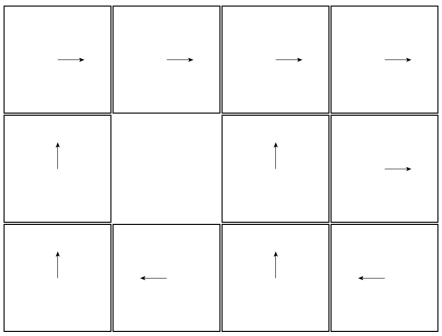
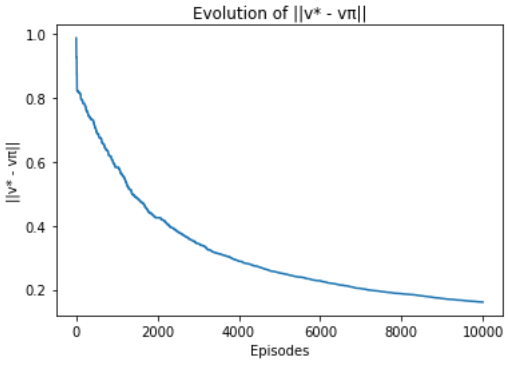
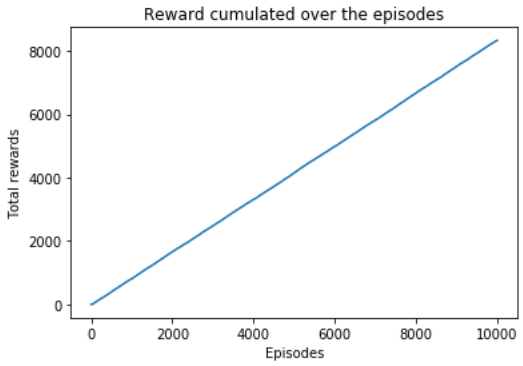


Figure 1: Optimal Policy returned after 10 000 episodes





3. If we change the initial distribution μ_0 in the step() function, the resulting optimal policy is still the same. Only the time to reach it will change.