Reinforcement Learnig: Homework 1

Raphaël Avalos

November 7, 2018

1 Dynamic Programming

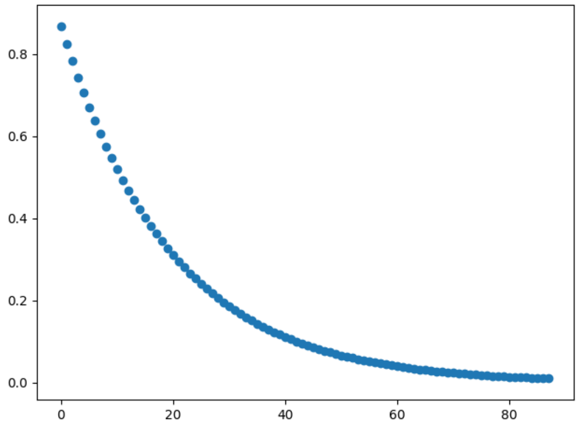
1.1 Question 1

The optimal policy π^* is easy to find because their is only 3 $(state, action)$ that have a reward. And their is only three steps.

$$\pi^* = [1, 1, 2]$$

1.2 Question 2

Figure 1: $\parallel v^k - v^* \parallel_\infty$



The value iteration find the same policy π^* and:

$$v^* = [15.204, 16.361, 17.819]$$

1.3 Question 3

The exact policy iteration returned the same policy.
To compare both algorithm we used the *timeit* module of python.

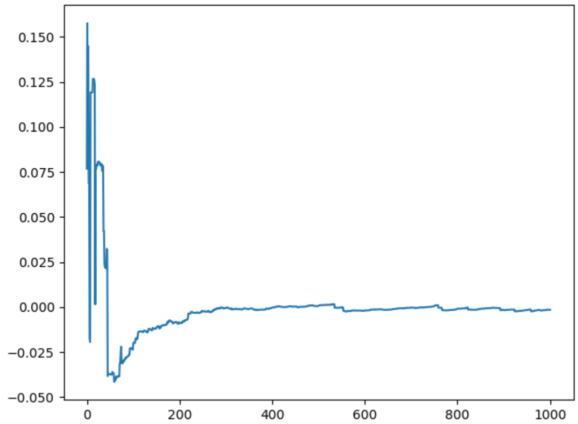
	Mean of 100 runs
VI	0.00208620
PI	0.00179925

- Value Iteration
  - Pros: each iteration is very computationally efficient.
  - Cons: convergence is only asymptotic.
- Policy Iteration
  - Pros: converge in a finite number of iterations (often small in practice).
  - Cons: each iteration requires a full policy evaluation and it might be expensive.

2 Reinforcement Learning

2.1 Question 4

Figure 2: $J_n - J^\pi$

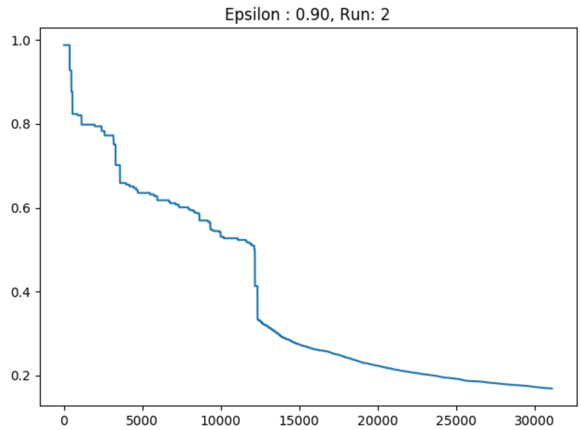


2.2 Question 5

The parameters choosed for the *Q learning algorithm* are the following.

- $\gamma = 0.95$
- $\alpha_n(x,a) = \frac{1}{n}$ because it is easier to make it independent of (x,a) and we know that it satisfies the usual stochastic approximation requirements.
- ϵ represent the tradeoff between exploration and exploitation. We decided to go with $\epsilon = 0.90$

Figure 3: $\parallel v^k - v^* \parallel_\infty$



2.3 Question 6

The optimal policy of a MDP is not affected by the the change of the initial distribution if all the states are still visited an infinit number of time.