Unsupervised Learning
Exercise 4 Report

clement.despiney@epitech.eu - baptiste.terras@epitech.eu

1. Goal

The exercise goal was to use a stochastic method to move an agent in 1D world. Some of the positions of the 1D world were assigned rewards randomly at set intervals. The agent collects the reward for the position it is in. The agent's final goal is to follow a policy that maximizes its expected total reward for each period in between the reset of the rewards.

The default policy was to simply move left as much as possible without caring about the consequences or rewards passed by. It simply went to the leftmost position, collecting rewards on the way once. If the agent were 'lucky' the leftmost position would have a reward and the agent would collect it until the reset. This agent's averaged reward on each interval was around 15-16 usually. The goal for the exercise was to find a policy that would give an average over 20.

2. Our policy

Our policy (located in `/ex4/clement_despiney_baptiste_terras_policy.py`) reach an average of 27 per run. We exceed the cutoff of 20 by almost a third, and the default policy by more than half.

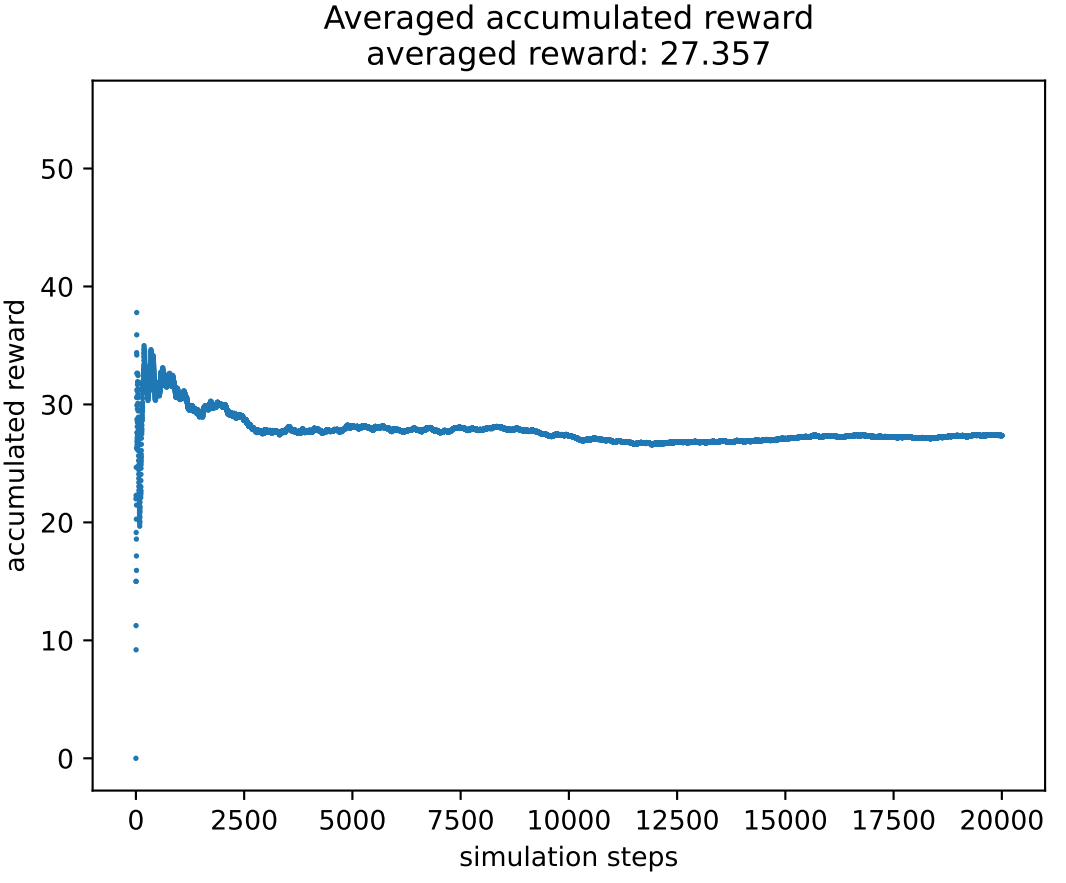
Our policy used a value function updated by the following equation:

$$V(s_0) = \max[r_0 + \gamma V(s_1)]$$

We set γ at 0.5 for this run and it gave satisfactory results.

This equation simply "spreads" the information about the rewards in the entire 1D world. By simply moving on to a position with a higher value function than the previous position, we can very easily maximize the accumulated reward. The big caveat being that there is not enough time to completely survey the world, therefore we might enter a local minimum.

We also chose to differentiate two periods for the agent. First, a 5 move period of random walk to survey the local positions. A second exploitation period until the reset where it exploits the maximum rewarding position using the value function. Then, as the world rewards reset, we go to the first exploratory period again.


Averaged accumulated reward
averaged reward: 27.357
accumulated reward
0
10
20
30
40
50
0
2500
5000
7500
10000
12500
15000
17500
20000
simulation steps