CISC856 - Reinforcement Learning

Assignment 1

April 18, 2022

Exercise 1 (10 pts)

Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specifically, suppose that, for any time step, the true values of actions 1 and 2 are respectively 0.1 and 0.2 with probability 0.5 (case A), and 0.9 and 0.8 with probability 0.5 (case B). If you are not able to tell which case you face at any step, what is the best expectation of success you can achieve and how should you behave to achieve it? Now suppose that on each step you are told whether you are facing case A or case B (although you still don't know the true action values). This is an associative search task. What is the best expectation of success you can achieve in this task, and how should you behave to achieve it?

Exercise 2 (10 pts)

Assume that

$$G_t = R_{t+1} + \gamma G_{t+1}$$

If you assume that the reward $R_t = 1$. Demonstrate that:

$$G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$$

What is the significance of this result for Reinforcement Learning?

Exercise 3 (10 pts)

In class we discussed the concept of "Exponential Weighted Average". Demonstrate that the right hand side of

$$Q_{n+1} = (1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$$

is an exponential weighted average. Explain why.

Exercise 4 (10 pts)

In class we discussed the problem of how to deal with a *markovian* environment. Consider the following grid:

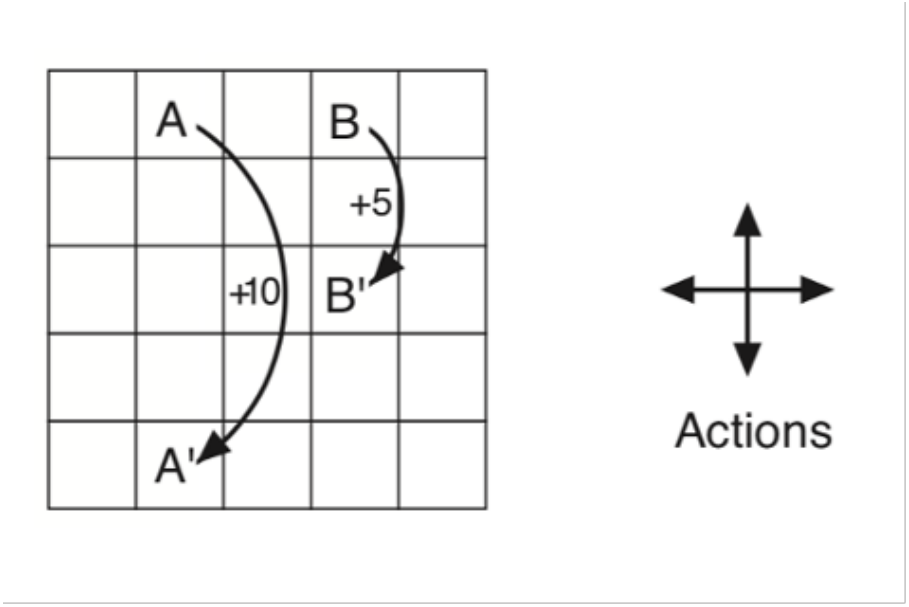

Figure 1: Grid world

In it, for all the states (cells in the grid), each one of the actions (north, south, east, and west) is chosen with probability $\frac{1}{4}$. The agent then moves with probability 1 to the chosen direction. While moving, if the agent hits a wall, it cannot move and it receives a reward of -1; if moving to a cell in the grid, the reward is 0.; if it reaches cell A $(1, 2)$, it moves to cell A' $(5, 2)$ and receives a reward +10; and, if it reaches cell B $(1, 4)$ it moves to cell B' $(3, 4)$ and receives a reward +5.

Write a program (in Python 3) to find the state-value for each one of the states for discount rates $\gamma \in \{0.75, 0.85, 0.9\}$.