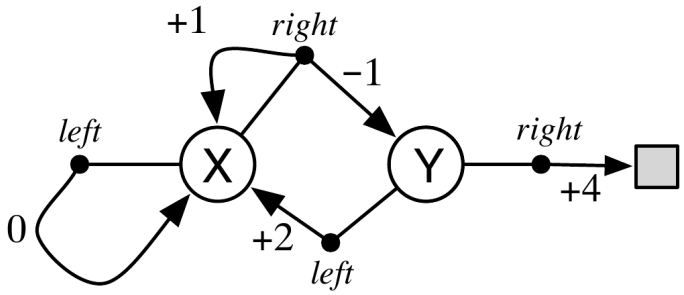
Reinforcement Learning
Prof. S. Harmeling

Monday, 27. April 2020
DUE 23:55h Monday, 04. May 2020

Exercise set #2

Solution should be submitted in teams of two if possible. Due to the current COVID-19 pandemic please submit your solution online using the sciebo file-drop folder. The link will be available in ILIAS. Please submit a single zip file with the following naming scheme: `username1-username2.zip` (e.g. `jadoe101-jodoe108.zip`). Allowed file extensions (of files within the zip file) are: `.pdf`, `.txt`, `.py` and `.ipynb`. Make sure the total file size does not exceed 10 MB.

1. **Trajectories, returns and values (from CMPUT609 [1])**
   Consider the MDP below, in which there are two states, X and Y, two actions, right and left, and the deterministic rewards on each transition are as indicated by the numbers. Note that if action right is taken in state X, then the transition may be either to X with a reward of $+1$ or to Y with a reward of -1. These two possibilities occur with probabilities 0.75 (for the transition to X) and 0.25 (for the transition to state Y).



   Consider two deterministic policies, π_1 and π_2:

   $\pi_1(X) = \textit{left}$ $\quad$ $\pi_2(X) = \textit{right}$
   $\pi_1(Y) = \textit{right}$ $\quad$ $\pi_2(Y) = \textit{right}$

   (a) Show a typical trajectory (sequence of states, actions and rewards) from X for policy π_1:

   5 points

   (b) Show a typical trajectory (sequence of states, actions and rewards) from X for policy π_2:

   10 points

   (c) Assuming the discount-rate parameter is $\gamma = 0.5$, what is the return from the initial state for the second trajectory?

   $G_0 =$

   10 points

   (d) Assuming $\gamma = 0.5$, what is the value of state Y under policy π_1?

   $v_{\pi_1}(Y) =$

   15 points

2. **Questions from Sutton and Barto [2]**

   (a) **Exercise 3.1:** Devise three example tasks of your own that fit into the MDP framework, identifying for each its states, actions, and rewards. Make the three examples as different from each other as possible. The framework is abstract and flexible and can be applied in many different ways. Stretch its limits in some way in at least one of your examples.
   20 points

   (b) **Exercise 3.7:** Imagine that you are designing a robot to run a maze. You decide to give it a reward of +1 for escaping from the maze and a reward of zero at all other times. The task seems to break down naturally into episodes—the successive runs through the maze—so you decide to treat it as an episodic task, where the goal is to maximize expected total reward. After running the learning agent for a while, you find that it is showing no improvement in escaping from the maze. What is going wrong? Have you effectively communicated to the agent what you want it to achieve?
   5 points

   (c) **Exercise 3.8:** Suppose $\gamma = 0.5$ and the following sequence of rewards is received: $R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, R_5 = 2$ with $T = 5$. What are $G_0, G_1, ..., G_5$? Hint: work backwards.
   10 points

   (d) **Exercise 3.9:** Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2$ followed by an infinite sequence of 7s. What are G_1 and G_0?
   10 points

   (e) **Exercise 3.12:** Give an equation for v_π in terms of q_π and π (Bellmann equation for action values).
   5 points

   (f) **Exercise 3.13:** Give an equation for q_π in terms of v_π and the four-argument p.
   10 points

References

[1] CMPUT 609: reinforcement learning for artificial intelligence. `http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/RLAIcourse/2009.html`.

[2] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.