Assignment 2

March 3, 2022

1. **Complete Exercise 3.1, but all your MDPs can be things relevant to your own interests (i.e., they don't have to be 'as different as possible' as the text says). Justify all your decisions: the reward sources, reward magnitudes, actions available, state representation, and the size of your state space. Justify your choices. You don't have to be exact about the state space (though that is good); just showing some awareness of size and its implications will be enough. Here is the framework we use:**

   (a) Tic-tac-toe

   Tic-tac-toe is a game for two players who take turns marking the spaces in a three-by-three grid with X or O. The player who succeeds in placing three of their marks in a horizontal, vertical, or diagonal row is the winner.

   i. State Space: Every possible configuration of Xs and Os on the three-by-three board.

   ii. Action Space:

   $$\{(0,0),(0,1),(0,2),(1,0),(1,1),(1,2),(2,0),(2,1),(2,2)\}$$

   Each element in the action space is a position on grid.

   iii. Reward Function:
   The reward of Tic-Tac-Toe is a winning combo of three consecutive characters, which awards the agent +1. The agent's loss would be punished by -1,

   iv. What constitutes an "episode"? When do episodes terminate, and why?
   Each play of game is an episode. After there is no improvement on the win-rate, the episodes terminate.

   (b) self-driving cars

   i. State Space:
   The sensor information about surroundings.

   ii. Action Space:
   Accelerator, steering wheel, and brake

   iii. Reward Function:
   Success to get to the destination will get positive rewards. Have an accident will get negative rewards.

   iv. What constitutes an "episode"? When do episodes terminate, and why? Each ride to destination is an episode. After there is no weight update in policy, the episodes terminate.

   (c) Dota

   i. State Space:
   Hero's position, hero's item and hero's ability usage.

   ii. Action Space: Mouse click and keyboard.

   iii. Reward Function: Kill the enemy, push a tower, win the game get positive reward. Killed by enemy or loss the game get negative reward.

   iv. What constitutes an "episode"? When do episodes terminate, and why? Each play of game is an episode. After there is no improvement on the win-rate, the episodes terminate.

2. **Is the reinforcement learning framework adequate to usefully represent all goal-directed learning tasks? Can you think of any clear exceptions? (3.2)**

   No. The reinforcement learning agent need to know actions and rewards in each state. If some tasks does not has clear action and reward for some state. Then the reinforcement learning framework is not adequate.

3. **Consider the problem of driving. You could define the actions in terms of the accelerator, steering wheel, and brake, that is, where your body meets the machine. Or you could define them farther out—say, where the rubber meets the road, considering your actions to be tire torques. Or you could define them farther in—say, where your brain meets your body, the actions being muscle twitches to control your limbs. Or**

you could go to a really high level and say that your actions are your choices of where to drive. What is the right level, the right place to draw the line between agent and environment? On what basis is one location of the line to be preferred over another? Is there any fundamental reason for preferring one location over another, or is it a free choice? (3.3)

We should define the action in terms of the accelerator, steering wheel, and brake, that is, where your body meets the machine.

The action an agent performs should be reliable, Consider if we define the actions being muscle twitches to control your limbs. However, it is hard for a human to control your muscle precisely. When we do the exploration part, we cannot say this muscle sends 5N to some direction and that muscle sends 10N to another direction. Therefore under this choice, the system is not reliable.

At the same time, the action space could not be too large. If the action space is too large, then it will take us more time to explore on the action space.

4. **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 (3.1). 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? (3.5)**

   I think this is an exploration issue, since the agent have not found the goal state and therefore it does not showing any improvement in escaping from the maze.

   The potential solution could be set the reward of each non-goal state be -1. This will let those state that the agent visit a lot get worse state values. Hence the agent will behave to find other way and eventually find the goal.

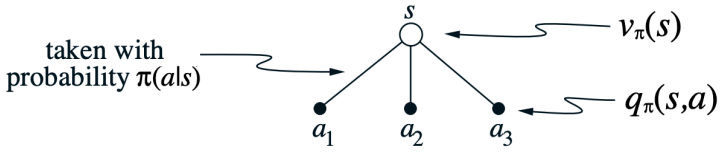
5. **Broken Vision System Imagine that you are a vision system. When you are first turned on for the day, an image floods into your camera. You can see lots of things, but not all things. You can't see objects that are occluded, and of course you can't see objects that are behind you. After seeing that first scene, do you have access to the Markov state of the environment? Suppose your camera was broken that day and you received no images at all, all day. Would you have access to the Markov state then? (3.6)**

   A state signal should construct and maintain on the basis of immediate sensations together with the previous state or some other memory of past sensations.

   For the first scene, we have access to the Markov state of the environment. A Markov state should not be excepted to inform the agent of everything about the environment, it only need to provide necessary information to make decisions. Since the agent can see lots of things, then it could take that as a state signal and make decisions and gain rewards.

   If my camera was broken, then this would not be considered as a Markov state. If the camera is broken, then it will unable to record any of the images or videos in current sensation. Hence we cannot determine the action we should perform and the reward we gain.

6. **The value of a state depends on the the values of the actions possible in that state and on how likely each action is to be taken under the current policy. We can think of this in terms of a small backup diagram rooted at the state and considering each possible action**:



   Give the equation corresponding to this intuition and diagram for the value at the root node, $v_\pi(s)$, in terms of the value at the expected leaf node, $q_\pi(s,a)$, given $S_t = s$. This expectation depends on the policy, π. Then give a second equation in which the expected value is written out explicitly in terms of $\pi(a|s)$ such that no expected value notation appears in the equation. (3.12)

$$A = \{a_1, a_2, a_3\}$$

$$v_\pi(s) = \sum_{a \in A} \mathbb{E}[R_t | s_t = s, a_t = a] \cdot \pi(s,a)$$

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \cdot q_\pi(s,a)$$

7. **Give a definition of v_* in terms of q_* (3.16).**

$$v_*(s) = \max_{a \in A(s)} q_*(s, a)$$

8. **Give a definition of q_* in terms of v_* (3.17).**

$$q_*(s, a) = \sum p(s', r|s, a)[r + v_*(s')]$$

9. **Give a definition of π_* in terms of q_* (3.18).**

$$\pi_*(s) = \arg\max_a q_*(s, a)$$

Since $\pi_*(s)$ is a probability, hence after we find the max action we should set them with equal probability.

10. **Give a definition of π_* in terms of v_* (3.19).**

$$\pi_*(s) = \arg\max_a \sum p(s', r|s, a)[r + v_*(s')]$$

11. **For each of the following, type out the equation, and then explain in words the computation that takes place:**

(a) The Bellman Equation for v_π

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]$$

Where a is an action taken from the action space $A(s)$, s' is the next state taken from state space S, the reward r are taken from R, the policy π is mapping from each state, $s \in S$, $a \in A(s)$, to the probability $\pi(a|s)$ of taking action a in state s. Then the value of each state $v_\pi(s)$ is the expected return from s following policy π.

When we compute $v_\pi(s)$, we find the probability of take an action at state s under policy π, and for all possible next state s', we find the probability of landing in that state, then find the sum of immediate reward of s' and $v_\pi(s')$. The equation $\sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]$ calculate the expected reward after we perform action a and the whole equation calculate the expected return when starting in s and following π

(b) The Bellman Equation for q_π

$$q_\pi(s, a) = \sum p(s', r|s, a)[r + \gamma v_\pi(s')]$$

Where a is an action taken from the action space $A(s)$, s' is the next state taken from state space S, the reward r are taken from R. $v_\pi(s)$ is the expected return from s following policy π.

when we compute $q_\pi(s, a)$, we find the transition probability and times the sum of immediate reward and discounted future reward.Then we get the expected return starting from s, taking the action a, and thereafter following policy π

12. **Complete Exercise 3.8, but for $\gamma = 0.7$ and $R_1 = 1, R_2 = 3, R_3 = -2, R_4 = 10, R_5 = -1, R_6 = -1$, and $R_7 = 2$, with $T = 7$. What are $G_0, \cdots, G_7$?**

$$\begin{aligned}
G_7 &= 0 \\
G_6 &= R_7 + \gamma G_7 = 2 \\
G_5 &= R_6 + \gamma G_6 = 0.4 \\
G_4 &= R_5 + \gamma G_5 = -0.72 \\
G_3 &= R_4 + \gamma G_4 = 9.496 \\
G_2 &= R_3 + \gamma G_3 = 4.6472 \\
G_1 &= R_2 + \gamma G_2 = 6.25304 \\
G_0 &= R_1 + \gamma G_1 = 5.377128
\end{aligned}$$

13. Complete Exercise 3.14, but for this grid of value estimates instead (use the center state valued at +2.7):

2.2	4	3.3	3.2	1.5
3.6	9.6	5.2	7.4	2.6
1.7	3.3	2.7	2.7	1.2
0.1	0.9	0.9	0.6	-0.2
-1.1	-0.5	-0.4	-0.6	-1.2

$$\frac{1}{4}(0+0.9*2.7)+\frac{[}{1}]4(0+0.9*5.2)+\frac{1}{4}(0+0.9*0.9)+\frac{1}{4}(0+0.9*3.3)$$
$$=\frac{1}{4}\cdot 2.43+\frac{1}{4}\cdot 4.68+\frac{1}{4}\cdot 0.81+\frac{1}{4}\cdot 2.97$$
$$=2.7225$$