CS5446 AI Planning and Decision Making

Semester 1, AY2022-23
Assignment 3

Agrawal, Shubhankar
A0248330L

Sagar, Sanchit
A0232478Y

October 30, 2022

1 Homework Assignment

1.1 Homework Problem 1: Reinforcement Learning

1.1.1 Parameter Updates for TD Learning

The utility of the state for TD learning is given by:

$$\hat{U}(x,y) = \theta_0 + \theta_1 x + \theta_2 y + \theta_3 \sqrt{x^2 - y^2 + 3xy}$$

The error function is given by

$$\varepsilon_j(s) = \frac{(\hat{U}_\theta(s) - u_j(s))^2}{2}$$

The parameter updates are given by

$$\theta_i \leftarrow \theta_i + \alpha * (R(s,a,s') + \gamma \hat{U}_\theta(s') - \hat{U}_\theta(s)) \frac{\partial(\hat{U}_\theta(s))}{\partial \theta_i}$$

As per the differentials, we have:

$$\frac{\partial(\hat{U}_\theta(s))}{\partial \theta_0} = 1$$
$$\frac{\partial(\hat{U}_\theta(s))}{\partial \theta_1} = x$$
$$\frac{\partial(\hat{U}_\theta(s))}{\partial \theta_2} = y$$
$$\frac{\partial(\hat{U}_\theta(s))}{\partial \theta_3} = \sqrt{x^2 - y^2 + 3xy}$$

Thus the final parameter updates for each theta are given by:

$$\theta_0 \leftarrow \theta_0 + \alpha * (R(s,a,s') + \gamma \hat{U}_\theta(s') - \hat{U}_\theta(s))$$
$$\theta_1 \leftarrow \theta_1 + \alpha * (R(s,a,s') + \gamma \hat{U}_\theta(s') - \hat{U}_\theta(s)) * x$$
$$\theta_2 \leftarrow \theta_2 + \alpha * (R(s,a,s') + \gamma \hat{U}_\theta(s') - \hat{U}_\theta(s)) * y$$
$$\theta_3 \leftarrow \theta_3 + \alpha * (R(s,a,s') + \gamma \hat{U}_\theta(s') - \hat{U}_\theta(s)) * \sqrt{x^2 - y^2 + 3xy}$$

1.1.2 MOVE Updates

The equation for utility is as given:

$$\hat{U}(x,y) = \theta_0 + \theta_1 x + \theta_2 y + \theta_3 \sqrt{x^2 - y^2 + 3xy}$$

We have all $\theta_i = 1$ in the beginning. The reward is given by R(s, a, s′) = $\sqrt{3}$

Before the action, the utilities are given as:

$$\hat{U}(1,1) = 1 + 1(1) + 1(1) + 1\sqrt{1 - 1 + 3}$$
$$\hat{U}(1,1) = 3 + 1.73 = 4.73$$
$$\hat{U}(2,1) = 1 + 1(2) + 1(1) + 1\sqrt{4 - 1 + 6}$$
$$\hat{U}(2,1) = 4 + 3 = 7$$

We have a learning rate(α) of 0.1 and a discount factor(γ) of 0.9

As a result the parameter updates are:

$$\theta_0 = 1 + 0.1(\sqrt{3} + 0.9 * 7 - 4.73)$$
$$\theta_0 = 1.33$$
$$\theta_1 = 1 + 0.1(\sqrt{3} + 0.9 * 7 - 4.73) * 1$$
$$\theta_1 = 1.33$$
$$\theta_2 = 1 + 0.1(\sqrt{3} + 0.9 * 7 - 4.73) * 1$$
$$\theta_2 = 1.33$$
$$\theta_3 = 1 + 0.1(\sqrt{3} + 0.9 * 7 - 4.73) * \sqrt{1 - 1 + 3}$$
$$\theta_3 = 1.57$$

The new utilities are thus:

$$\hat{U}(1,1) = 1.33 + 1.33(1) + 1.33(1) + 1.57\sqrt{1 - 1 + 3}$$
$$\hat{U}(1,1) = 6.71$$
$$\hat{U}(2,1) = 1.33 + 1.33(2) + 1.33(1) + 1.57\sqrt{4 - 1 + 6}$$
$$\hat{U}(2,1) = 10.03$$

1.1.3 Extended Grid World

We have an extended grid world problem where we do not know the dimensions of the grid. Moreover, there are many obstacles as well as multiple terminal states with rewards of +1 or -1. This extended grid world is represented in Figure 1 below.

The features that can be used to model this generalized learning problem are as follows:

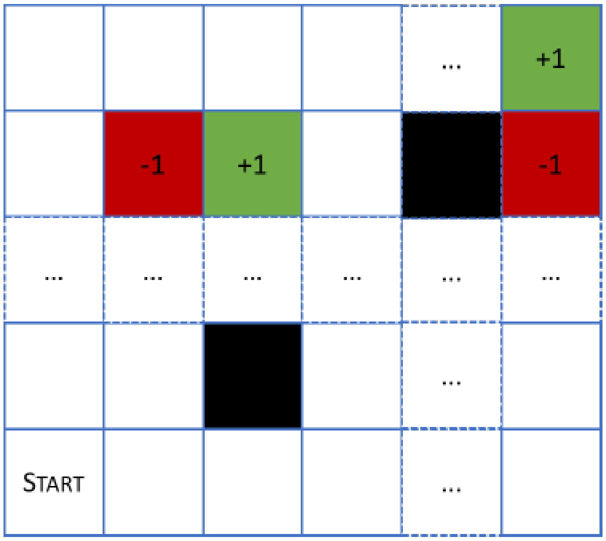

Figure 1: Extended Grid World

1. Number of adjacent +1 terminal states
    - The higher this feature, the better state the agent is in (given a certain adjacency metric).
2. Number of adjacent -1 terminal states
    - The lower this feature, the better state the agent is in (given a certain adjacency metric).
3. Manhattan Distance to nearest +1 terminal state
    - The lower this feature, the better state the agent is in.
4. Manhattan Distance to nearest -1 terminal state
    - The higher this feature, the better state the agent is in.
5. Number of adjacent obstacles
    - The lower this feature, the better state the agent is in.
6. Number of shortest paths to closest +1 terminal state
    - The higher this feature, the better state the agent is in.
7. Number of shortest paths to closest -1 terminal state
    - The lower this feature, the better state the agent is in.

1.2 Homework Problem 2: Partially Observable Markov Decision Process

1.2.1 Belief Updates

With the given a problem with 2 states s_1 and s_2, and a single action a. The transition probability functions are defined as follows:

$$P(s_1|s_1, a) = 0.3$$
$$P(s_2|s_1, a) = 0.7$$
$$P(s_1|s_2, a) = 0.6$$
$$P(s_2|s_2, a) = 0.4$$

The current beliefs are given by:

$$b(s_1) = 0.6$$
$$b(s_2) = 0.4$$

Additionally, for an observation o_1, we have

$$P(o_1|s_1) = 80\%$$
$$P(o_1|s_2) = 20\%$$

We receive evidence $e = o_1$

The belief update in a POMDP is given by:

$$b'(s') = \alpha P(e'|s') * \sum_s P(s'|s, a)b(s)$$

Here is the alpha is the normalizing factor. Thus, we calculate the unnormalized beliefs for the states and then normalize it to get the final values.

$$Unnormalized\, b'(s1) = P(o_1|s_1) * [P(s1|s1, a) * b(s1) + P(s1|s2, a) * b(s2)]$$
$$Unnormalized\, b'(s1) = 0.8 * [0.3 * 0.6 + 0.6 * 0.4] = 0.336$$
$$Unnormalized\, b'(s2) = P(o_1|s_2) * [P(s2|s1, a) * b(s1) + P(s2|s2, a) * b(s2)]$$
$$Unnormalized\, b'(s2) = 0.2 * [0.7 * 0.6 + 0.4 * 0.4] = 0.116$$

$$b'(s1) = 0.336/(0.336 + 0.116) = 0.743$$
$$b'(s2) = 0.116/(0.336 + 0.116) = 0.257$$

1.2.2 Camera Human Tracking

We are given a problem where we have a camera onboard a mobile robot. This camera can move and tilt to view different parts of the environment. If a human appears in the environment, the camera is supposed to track the person's movement.

Our assumptions are as follows:

- There is always one person in the environment
- The robot is stationary

Moreover, we have the following values:

- Positive values indicate pan right or tilt up
- Negative values indicate pan left or tilt down
- Each pan works by 45°

- Each tilt works by 15°
- There is only a 20% probability a pan right leads to a new view
- There is only a 40% chance the correct tilt angle is obtained

In order to model the problem as a POMDP, we take the following assumptions:

- The pan and tilt resulting in changes in degrees, we assume that the total scope for movement is one complete circle i.e. 360 °(It's a 360 °camera).
- The 20% chance of obtaining the right pan angle on pan right does not imply the same on pan left. Thus we assume pan left always succeeds.
- For the remaining 80% of the time when it tries to pan right, it is unsuccessful and remains in the current state (view).
- Similarly, for the remaining 60% of the time when it tries to tilt in any direction, it is unsuccessful and remains in the same state (view).

We model this problem as a POMDP as follows:

State S(Pan, Tilt, x, y) The state is given by the value of two variables for the camera, the pan value and the tilt value and two variables for the human, which represent the relative position of the human with respect to the correct view. This represents the view the camera has, and the position of the human. We assume that the total possible movement is over 360 degrees, thus the total possible Pan, Tilt combinations are 360*360. For the x and y values, we assume they take possible values of $-1, 0, 1$ each where 0 represents the human in the current view (Pan, Tilt) combination, 1 represents that the human is to the right/top (depending on whether it is x or y) of the current view and -1 represents that the human is to the (left/bottom) of the current view.

Actions *Pan, Tilt* The two actions are given as follows:

Pan(n) Where n can take a value of +1 or -1 to pan +45°or -45°respectively.

Tilt(n) Where n can take a value of +1 or -1 to tilt +15°or -15°respectively.

Transition Function The transition function is given as follows:

T((p, t, x, y), Pan(+1), (p+45, t, x, y)) = 0.2 20% chance that pan right leads to correct view.

T((p, t, x, y), Pan(+1), (p, t, x, y)) = 0.8 80% chance that pan right stays in the same view.

T((p, t, x, y), Pan(-1), (p-45, t, x, y)) = 1 Pan left always gives correct view.

T((p, t, x, y), Tilt(+1), (p, t+15, x, y)) = 0.4 40% chance correct tilt angle is obtained.

T((p, t, x, y), Tilt(+1), (p, t, x, y)) = 0.6 60% chance tilt fails and camera remains in the same view.

T((p, t, x, y), Tilt(-1), (p, t-15, x, y)) = 0.4 40% chance correct tilt angle is obtained.

T((p, t, x, y), Tilt(-1), (p, t, x, y)) = 0.6 60% chance tilt fails and camera remains in the same view.

At the extremes, further pan and tilt would not lead to more movements.

To model the observations:

Observation Space The observation space is the images captured by the camera along with the human (if present) in the image.

Observation Function The observation function is the probability of capturing the human's image given the current state (Pan, Tilt) of the camera as well as the position of the human.

Putting this in functions, we have:

O(Human | p, t, 0, 0) = 1 If human is in correct view, we always observe him/her.

O(No Human | p, t, 0, 0) = 0 If human is in correct view, we always observe him/her.

O(Human | p, t, $x_{/0}$**,** $y_{/0}$**) = 0** If human is not in correct view, we never observe him/her.

O(No Human | p, t, $x_{/0}$**,** $y_{/0}$**) = 1** If human is not in correct view, we never observe him/her.