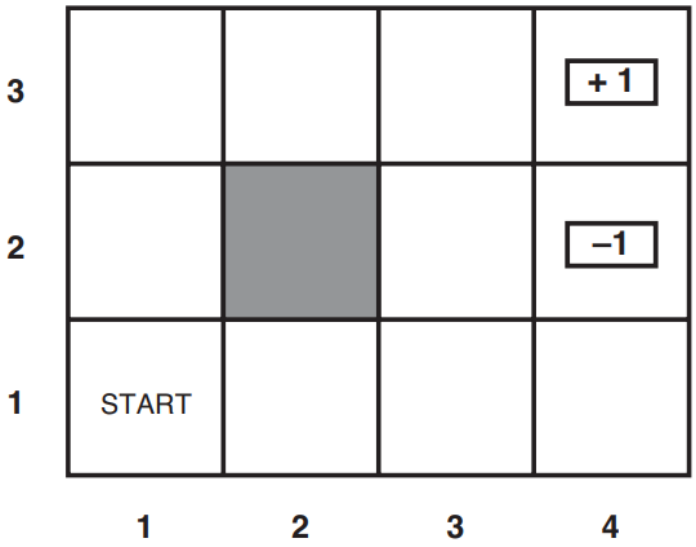
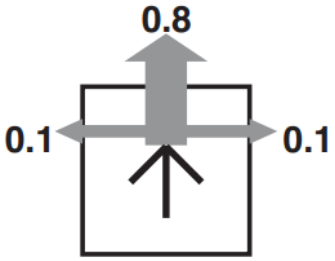
(a) (8 points) For the 4×3 grid world shown in Figure 17.1 of the Russell and Norvig textbook (which is the same grid world as the one discussed in class), calculate which grid cells can be reached from (1,1) by the action sequence [Up, Up, Right, Right, Right] and with what probabilities.



(a)



(b)

Up: 0.1 (1,1) & 0.8 (1,2) & 0.1 (2,1)

Up: 0.02 (1,1) & 0.24 (1,2) & 0.09 (2,1) & 0.64 (1,3) & 0.01 (3,1)

Right: 0.026 (1,1) & 0.258 (1,2) & 0.088 (1,3) & 0.034 (2,1) & 0.073 (3,1) & 0.512 (2,3) & 0.008 (4,1) & 0.001 (3,2)

Right: 0.0284 (1,1) & 0.2178 (1,2) & 0.0346 (1,3) & 0.0276 (2,1) & 0.1728 (2,3) & 0.0346 (3,1) & 0.0073 (3,2) & 0.4097 (3,3) & 0.0656 (4,1) & 0.0016 (4,2)

Right: 0.02462 (1,1) & 0.18054 (1,2) & 0.02524 (1,3) & 0.02824 (2,1) & 0.06224 (2,3) & 0.02627 (3,1) & 0.04443 (3,2) & 0.17994 (3,3) & 0.08688 (4,1) & 0.01368 (4,2) & 0.32792 (4,3)

So, the answer is (1,1), (1,2), (1,3), (2,1), (2,3), (3,1), (3,2), (3,3), (4,1), (4,2), (4,3).

And the corresponding probability is 0.02462 (1,1) & 0.18054 (1,2) & 0.02524 (1,3) & 0.02824 (2,1) & 0.06224 (2,3) & 0.02627 (3,1) & 0.04443 (3,2) & 0.17994 (3,3) & 0.08688 (4,1) & 0.01368 (4,2) & 0.32792 (4,3)

(b) (12 points) The MDP formulation we studied in class included a reward function $R(s, a, s')$ where the reward depends on the triple current state, action, and the outcome state.

i. Show how an MDP with reward function $R(s, a, s')$ can be transformed into a different MDP with reward function $R(s, a)$ such that optimal policies in the new MDP correspond exactly to optimal policies in the original MDP.

$$V^{\pi*}(s) = max\, \Sigma_{s'} P(s'|s,a)[R(s,a,s') + \gamma V^{\pi*}(s')]$$

$$= max[R(s,a) + \gamma \Sigma_{s'} P(s'|s,a) V^{\pi*}(s')]$$

ii. Now, do the same to convert your MDP with $R(s, a)$ into the MDP with $R(s)$ such that the correspondence between optimal policies in the two MDPs is obtained.

$$max[R(s,a) + \gamma \Sigma_{s'} P(s'|s,a) V^{\pi*}(s')] = \underset{a \in A}{argmax}\, R(s) + \gamma \Sigma_{s'} P(s'|s,a) V^{\pi*}(s')$$

iii. Prove that the value of any fixed policy varies linearly with $R(s)$.

$$v^{\pi}(s_t) = R(s_{t+1}) + \gamma v^{\pi}(s_{t+1})$$

$$= R(s_t) + \gamma\, [R(s_{t+2}) + \gamma v^{\pi}(s_{t+2})]$$

$$= R(s_t) + \gamma\, R(s_{t+1}) + \gamma\, [R(s_{t+2}) + \gamma^2 v^{\pi}(s_{t+3})]$$

$$= \sum_{i=0}^{T} \gamma^i\, R(s_{t+i})$$

$$= R(s_t) + \gamma\, R(s_{t+1}) + \gamma^2\, R(s_{t+1}) + \cdots$$

(c) (15 points) Consider an undiscounted MDP having three states (1, 2, 3), with rewards -1, -2, 0, respectively. State 3 is a terminal state. In states 1 and 2 there are two possible actions: a and b. The transition model is as follows:

- In state 1, action a moves the agent to state 2 with probability 0.8 and makes the agent stay put with probability 0.2.
- In state 2, action a moves the agent to state 1 with probability 0.8 and makes the agent stay put with probability 0.2.
- In either state 1 or state 2, action b moves the agent to state 3 with probability 0.1 and makes the agent stay put with probability 0.9.

Given the above MDP, answer the following:

i. What can be determined *qualitatively* about the optimal policy in states 1 and 2?

ii. Apply **policy iteration**, showing each step in full, to determine the optimal policy and the values of states 1 and 2. Assume that the initial policy has action b in both states.

iii. What happens to policy iteration if the initial policy has action a in both states? Does discounting help? Does the optimal policy depend on the discount factor?

i: For state 1:

R(a|s1) = 0.8*(-2) + 0.2*(-1) = -1.6

R(b|s1) = 0.9*(-1) + 0.1*0 = -0.9

For state 2:

R(a|s2) = 0.8*(-1) + 0.2*(-2) = -1.2

R(b|s2) = 0.9*(-2) + 0.1*0 = -1.8

ii: Policy Evaluation:

$$\text{Initialize } V^{\pi}(1), V^{\pi}(2) = 0, for\ any\ \pi\ and\ \gamma = 0.9$$

$$V^{\pi}(s) = \Sigma_{s'} P(s'|s,a)[R(s,\pi(s),s') + \gamma V^{\pi}(s')]$$

$$V^{b}(1) = 0.9 * (-1 + 0.9 * V^{b}(2)) + 0.1 * 0$$

$$V^{b}(2) = 0.9 * (-2 + 0.9 * V^{b}(1)) + 0.1 * 0$$

$$V^{b}(1) = \text{-6.857}$$

$$V^{b}(2) = \text{-7.354}$$

Policy Improvement:

$$\pi(1) = \underset{a \in A}{argmax} \Sigma_{s'} P(s'|s,a)[R(s,\pi(s),s') + \gamma V^{\pi}(s')]$$

$$= \max\Big[\big(\ 0.8 * (-2 + 0.9 * V^{\pi(2)}(2)\ \big) + \ 0.2 * (-1 + 0.9 * V^{a}(1))\Big),$$

$$\big(\ 0.8 * (-2 + 0.9 * V^{\pi(2)}(2)\ \big) + \ 0.2 * (-1 + 0.9 * V^{b}(1))\Big]$$

$$\text{So, } \pi(1) = a$$

$$V^b(2) < V^a(2) = 0$$

$$\pi(2) = \underset{a \in A}{argmax} \Sigma_{s'} P(s'|s,a)[R(s,\pi(s),s') + \gamma V^{\pi}(s')]$$

$$= \max\big[(\ 0.8 * (-1 + 0.9 * V^a(1)\ ) + \ 0.2 * (-2 + 0.9 * V^{\pi(2)}(2))\big),$$

$$(\ 0.8 * (-1 + 0.9 * V^a(1)\ ) + \ 0.2 * (-2 + 0.9 * V^{\pi(2)}(2))]$$

$$\text{So, } \pi(2) = a$$

Policy Evaluation:

$$V^a(1) = \ 0.8 * (-2 + 0.9 * V^a(2)) + \ 0.2 * (-1 + 0.9 * V^a(1))$$

$$V^a(2) = \ 0.8 * (-1 + 0.9 * V^a(1)) + \ 0.2 * (-2 + 0.9 * V^a(1))$$

$$V^a(1) = \text{-13.68}$$

$$V^a(2) = \text{-13.08}$$

Policy Improvement:

$$\pi(1) = \underset{a \in A}{argmax} \Sigma_{s'} P(s'|s,a)[R(s,\pi(s),s') + \gamma V^{\pi}(s')]$$

$$= \max\big[\big(\ 0.8 * (-2 + 0.9 * V^{\pi(2)}(2)\ \big) + \ 0.2 * (-1 + 0.9 * V^a(1))\big),$$

$$\big(\ 0.8 * (-2 + 0.9 * V^{\pi(2)}(2)\ \big) + \ 0.2 * (-1 + 0.9 * V^b(1))]$$

$$\text{So, } \pi(1) = b$$

$$V^b(2) < V^a(2) = 0$$

$$\pi(2) = \underset{a \in A}{argmax} \Sigma_{s'} P(s'|s,a)[R(s,\pi(s),s') + \gamma V^{\pi}(s')]$$

$$= \max\big[(\ 0.8 * (-1 + 0.9 * V^a(1)\ ) + \ 0.2 * (-2 + 0.9 * V^{\pi(2)}(2))\big),$$

$$(\ 0.8 * (-1 + 0.9 * V^a(1)\ ) + \ 0.2 * (-2 + 0.9 * V^{\pi(2)}(2))]$$

$$\text{So, } \pi(2) = b$$

Now Policy Unchanged.

So, the optimal policy in state 1 is b and in state 2 is b.

iii:

- If the initial policy is a, the iteration will convergence faster.
- Discount factor will help convergence. Is there is no discounting, the value will become infinite so that there is no convergence.
- The optimal policy will depend on discount factor, which will influence the horizon agent consider. The smaller discount factor, the shorter agent's sight; vice versa. The optimal policy will change in different horizon.

(d) (15 points) For large state spaces, a parametric approximation of the value function is preferable for RL. Consider the following function approximation for a grid world domain:

$$\hat{V}_\theta(x,y) = \theta_0 + \theta_1 x + \theta_2 y + \theta_3\sqrt{(x-x_g)^2 + (y-y_g)^2}$$

where (x_g, y_g) is the location of the goal state.

A TD update then seeks to adjust the parameters to try to reduce the temporal difference between successive iterations. In other words (using the Widrow-Hoff rule),

$$\theta_i \leftarrow \theta_i + \alpha[R(x,y) + \gamma\hat{V}_\theta(x',y') - \hat{V}_\theta(x,y)]\frac{\partial\hat{V}_\theta(x,y)}{\partial\theta_i}.$$

i. Write out the TD learning update equations for parameters θ_1, θ_2, θ_3, and θ_4.

ii. Consider again the 4×3 grid world shown in Figure 17.1 of the Russell and Norvig textbook. Beginning with initial parameter values of 0, show how the parameters update as the agent executes the action sequence [Up, Up, Right, Right, Right] from (1,1). You may assume that each action succeeds as intended and use $\alpha = 0.75$ and $\gamma = 0.9$.

i:

$$\theta_0 \Leftarrow \theta_0 + \alpha[\mathbb{R}(x,y) + \gamma\hat{v}_\theta(x',y') - \hat{v}_\theta(x,y)]$$

$$\theta_1 \Leftarrow \theta_1 + \alpha[\mathbb{R}(x,y) + \gamma\hat{v}_\theta(x',y') - \hat{v}_\theta(x,y)] * x$$

$$\theta_2 \Leftarrow \theta_2 + \alpha[\mathbb{R}(x,y) + \gamma\hat{v}_\theta(x',y') - \hat{v}_\theta(x,y)] * y$$

$$\theta_3 \Leftarrow \theta_3 + \alpha[\mathbb{R}(x,y) + \gamma\hat{v}_\theta(x',y') - \hat{v}_\theta(x,y)] * \sqrt{(x-4)^2 + (y-3)^2}$$

ii:

$$\hat{v}_\theta(x,y) = \theta_0 + \theta_1 x + \theta_2 y + \theta_3\sqrt{(x-x_g)^2 + (y-y_g)^2}$$

Assumption: the action sequence is move sequence.

- From (1,1) to (1,2):

$$\theta_0 \Leftarrow 0 + 0.75[-0.04 + 0.9 * (0) - 0] = -0.03$$

$$\theta_1 \Leftarrow 0 + 0.75[-0.04 + 0.9 * (0) - 0] * 1 = -0.03$$

$$\theta_2 \Leftarrow 0 + 0.75[-0.04 + 0.9 * (0) - 0] * 1 = -0.03$$

$$\theta_3 \Leftarrow [0 + 0.75[-0.04 + 0.9 * (0) - 0]] * \sqrt{(1-4)^2 + (1-3)^2} = -0.03 * \sqrt{13} = -0.1082$$

- From (1,2) to (1,3):

$$\theta_0 \Leftarrow -0.03 + 0.75\big[-0.04 + 0.9 * (-0.03 - 0.03 * 1 - 0.03 * 3 - 0.1082 * 3\ ) - (-0.03 - 0.03 * 1 - 0.03 * 2 - 0.1082 * \sqrt{10}\ )\big] = -0.0337$$

$$\theta_1 \Leftarrow -0.03 + 0.75\big[-0.04 + 0.9 * (-0.03 - 0.03 * 1 - 0.03 * 3 - 0.1082 * 3\ ) - (-0.03 - 0.03 * 1 - 0.03 * 2 - 0.1082 * \sqrt{10}\ )\big] * 1 = -0.0337$$

$$\theta_2 \Leftarrow -0.03 + 0.75\big[-0.04 + 0.9 * (-0.03 - 0.03 * 1 - 0.03 * 3 - 0.1082 * 3\ ) - (-0.03 - 0.03 * 1 - 0.03 * 2 - 0.1082 * \sqrt{10}\ )\big] * 2 = -0.0375$$

$$\theta_3 \Leftarrow -0.1082 + 0.75\big[-0.04 + 0.9 * (-0.03 \ - 0.03 * 1 - 0.03 * 3 - 0.1082 * 3 \ ) - (-0.03 \ - 0.03 * 1 - 0.03 * 2 - 0.1082 * \sqrt{10} \ )\big] * \sqrt{10} = -0.12$$

- From (1,3) to (2,3):

$$\theta_0 \Leftarrow -0.0337 + 0.75[-0.04 + 0.9 * (-0.0337 \ - 0.0337 * 2 - 0.0375 * 3 - 0.12 * 2 \ ) - (-0.0337 \ - 0.0337 * 1 - 0.0375 * 3 - 0.12 * 3 \ )] = 0.035$$

$$\theta_1 \Leftarrow -0.0337 + 0.75[-0.04 + 0.9 * (-0.0337 \ - 0.0337 * 2 - 0.0375 * 3 - 0.12 * 2 \ ) - (-0.0337 \ - 0.0337 * 1 - 0.0375 * 3 - 0.12 * 3 \ )] * 1 = 0.035$$

$$\theta_2 \Leftarrow -0.0375 + 0.75[-0.04 + 0.9 * (-0.0337 \ - 0.0337 * 2 - 0.0375 * 3 - 0.12 * 2 \ ) - (-0.0337 \ - 0.0337 * 1 - 0.0375 * 3 - 0.12 * 3 \ )] * 3 = 0.1687$$

$$\theta_3 \Leftarrow -0.12 + 0.75[-0.04 + 0.9 * (-0.0337 \ - 0.0337 * 2 - 0.0375 * 3 - 0.12 * 2 \ ) - (-0.0337 \ - 0.0337 * 1 - 0.0375 * 3 - 0.12 * 3 \ )] * 3 = 0.0862$$

- From (2,3) to (3,3):

$$\theta_0 \Leftarrow 0.035 + 0.75[-0.04 + 0.9 * (0.035 + 0.035 * 3 + 0.1687 * 3 + 0.0862 * 1 \ ) - (0.035 + 0.035 * 2 + 0.1687 * 3 + 0.0862 * 2 \ )] = -0.0883$$

$$\theta_1 \Leftarrow 0.035 + 0.75[-0.04 + 0.9 * (0.035 + 0.035 * 3 + 0.1687 * 3 + 0.0862 * 1 \ ) - (0.035 + 0.035 * 2 + 0.1687 * 3 + 0.0862 * 2 \ )] * 2 = -0.2116$$

$$\theta_2 \Leftarrow 0.1687 + 0.75[-0.04 + 0.9 * (0.035 + 0.035 * 3 + 0.1687 * 3 + 0.0862 * 1 \ ) - (0.035 + 0.035 * 2 + 0.1687 * 3 + 0.0862 * 2 \ )] * 3 = -0.2012$$

$$\theta_3 \Leftarrow 0.0862 + 0.75[-0.04 + 0.9 * (0.035 + 0.035 * 3 + 0.1687 * 3 + 0.0862 * 1 \ ) - (0.035 + 0.035 * 2 + 0.1687 * 3 + 0.0862 * 2 \ )] * 2 = -0.1604$$

- From (3,3) to (4,3):

$$\theta_0 \Leftarrow -0.0883 + 0.75[1 + 0.9 * (-0.0883 - 0.2116 \ * 4 \ - 0.2012 \ * 3 + 0 \ ) - (-0.0883 - 0.2116 \ * 3 \ - 0.2012 \ * 3 \ - 0.1604 * 1 \ )] = 0.7387$$

$$\theta_1 \Leftarrow \ -0.2116 \ + 0.75[1 + 0.9 * (-0.0883 - 0.2116 \ * 4 \ - 0.2012 \ * 3 + 0 \ ) - (-0.0883 - 0.2116 \ * 3 \ - 0.2012 \ * 3 \ - 0.1604 * 1 \ )] * 3 = 2.2694$$

$$\theta_2 \Leftarrow -0.2012 \ + 0.75[1 + 0.9 * (-0.0883 - 0.2116 \ * 4 \ - 0.2012 \ * 3 + 0 \ ) - (-0.0883 - 0.2116 \ * 3 \ - 0.2012 \ * 3 \ - 0.1604 * 1 \ )] * 3 = 2.2798$$

$$\theta_3 \Leftarrow -0.1604 + 0.75[1 + 0.9 * (-0.0883 - 0.2116 \ * 4 \ - 0.2012 \ * 3 + 0 \ ) - (-0.0883 - 0.2116 \ * 3 \ - 0.2012 \ * 3 \ - 0.1604 * 1 \ )] * 1 = 0.6666$$