CSE471 - Homework 5

Kumal Patel

March 17, 2021

Exercise 1.1 (a) $V^*(s) = R(s) + \gamma max_a[\sum_{s'} T(s,a,s')V^*(s')]$

$$V^*(s) = max_a[R(s,a) + \gamma \sum_{s'} T(s,a,s')V^*(s')]$$

(b)

(c)

Exercise 1.2 .

Table 1: Policy for different reward values

↑	←	+10
↑	←	↓
↑	←	←

(a) r = 100

→	→	+10
→	→	↑
→	→	↑

(b) r = -3

→	→	+10
↑	↑	↑
↑	↑	↑

(c) r = 0

←	←	+10
↑	←	↑
↑	←	←

(d) r = 3

(a) Since the reward value is 100 and is larger than the value of agent being in the final state. The policy should represent a set of actions to ensure that the agent stays in the reward state and avoid the final state. And because this environment is stochastic the actions must be perpendicular to direction with 0.1 probability. To ensure the agent doesn't reach the final state the action before should be in the opposite direction.

(b) Since the reward value is -3 the agent would try to avoid this location and just go towards the final state. The policy should represent a set of actions that ensure the agent avoids the reward state and heads towards the final state.

(c) Since the reward value is 0 and the white squares is -1 the agent would want to go to the reward state before going to the final state. The policy should represent a set of actions to ensure this happens.

(d) Since the reward value is 3 the agent would want to stay in this location and avoid the final state. The same reasonings were applied to question a.

Exercise 1.3 $R(Up) = 50 - (\gamma + \gamma^2 + \gamma^3.. + \gamma^{100}) = 50 - \sum_{n=1}^{100} \gamma^n$

$R(Down) = -50 - (\gamma + \gamma^2 + \gamma^3... + \gamma^{100}) = -50 - \sum_{n=1}^{100} \gamma^n$

$50 - \sum_{n=1}^{100} \gamma^n = -50 - \sum_{n=1}^{100} \gamma^n$

$100 - \sum_{n=1}^{100} \gamma^n = - \sum_{n=1}^{100} \gamma^n$

$100 = 2 \sum_{n=1}^{100} \gamma^n$

$50 = \sum_{n=1}^{100} \gamma^n$

$\gamma = 0.984398$

When $\gamma < 0.984398$ the agent will choose the Up action and when $\gamma > 0.984398$ the agent will choose the Down action.

Exercise 1.4 $V^{\pi_0}(cool) = 1 * [1 + 0.5 * V^{\pi_0}(cool)] = 2$

$V^{\pi_0}(warm) = 1 * [-10 + 0.5 * V^{\pi_0}(overheated)] = 0.5$

$\pi_1(cool) = argmax\{slow : 1 * [1 + 0.5 * 2], fast : 0.5[2 + 0.5 * 2] + 0.5[2 + 0.5 * 0.5]\}$

$\pi_1(cool) = argmax\{slow : 2, fast : 2.625\}$

$\pi_1(cool) = fast$

$\pi_1(warm) = argmax\{slow : 0.5 * [1 + 0.5 * 2] + 0.5[1 + 0.5 * 0.5], fast : 1 * [-10 + 0.5 * 0]\}$

$\pi_1(warm) = argmax\{slow : 1.625, fast : -10\}$

$\pi_1(warm) = slow$

$\pi_2(cool) = argmax\{slow : 1 * [1 + 0.5 * 2.625], fast : 0.5[2 + 0.5 * 2.625] + 0.5[2 + 0.5 * 1.625]\}$

$\pi_2(cool) = argmax\{slow : 2.3125, fast : 3.0625\}$

$\pi_2(cool) = fast$

$\pi_2(warm) = argmax\{slow : 0.5 * [1 + 0.5 * 1.625] + 0.5[1 + 0.5 * 2.625], fast : 1 * [-10 + 0.5 * 0]\}$

$\pi_2(warm) = argmax\{slow : 2.0625, fast : -10\}$

$\pi_2(warm) = slow$

Table 2: Policy iteration showing the optimal policy

	cool	**warm**	**overheated**
V^{π_0}	2	0.5	0
V^{π_1}	2.625	1.625	0
V^{π_2}	3.0625	2.0625	0

	cool	**warm**
π_0	slow	fast
π_1	fast	slow
π_2	fast	slow

Exercise 1.5

(a) T(cool, slow, cool) = 1.0
T(cool, fast, cool) = 0.5
T(cool, fast, warm) = 0.5
T(warm, fast, overheated) = 1.0
T(warm, slow, warm) = 0.75
T(warm, slow, cool) = 0.25

R(cool, slow, cool) = 1
R(cool, fast, cool) = 2
R(cool, fast, warm) = 2
R(warm, fast, overheated) = -10
R(warm, slow, warm) = 1
R(warm, slow, cool) = 1

(b) V(cool) = $\frac{-34}{7} = -4.857$
V(warm) = $\frac{-44}{6} = -7.333$

(c)

(d)