CP8319 Assignment 1

Student Name: Albina Cako

Question 1.

a. The value of r_s that would cause the optimal policy to return the shortest path to the green target square is $r_s = -1$. Below is the calculated value for the optimal solution of each square:

0	1	2	3
-5	2	3	4
2	3	4	5
1	0	-1	-2

b. In this section, all the rewards are added + 2 to them. Below is the updated grid word:

12	11	10	9
-3	10	9	8
10	9	8	7
11	12	13	14

c.

$$V^{\pi}_{old}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s\right]$$

$$V^{\pi}_{new}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t (r_t + c) \mid s_0 = s\right]$$

$$= \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s\right] + c\sum_{t=0}^{\infty} \gamma^t$$

$$V^{\pi}_{new}(s) = V^{\pi}_{old}(s) + \frac{c}{1-\gamma}$$

d. In this case, the optimal policy will keep looping around forever. It will never reach the target because each step will keep gaining a positive reward value. The agent never stops exploring in this case. In the gridworld, all the unshaded squares and the green square would become + ∞, while the red square would be -2.

Below is a visual of how the gridworld would look like:

+ ∞	+ ∞	+ ∞	+ ∞
-2	+ ∞	+ ∞	+ ∞
+ ∞	+ ∞	+ ∞	+ ∞
+ ∞	+ ∞	+ ∞	+ ∞

e. Yes it will depend on gamma. As gamma is closer to 1, just like in part d, the policy will keep looping around forever. This is because the agent focuses on long term goals and keeps exploring. However, as gamma decreases (gets closer to zero) as some point the policy will reach the target (green in the gridworld) in the shortest amount of time. Essentially, as gamma decreases the agent is focusing on short term goals and it will take less risks and try to reach the target in the shorter amount of time.

f. In this case, any vale of rs ≤ -5 would result in the termination in the red square.

Question 2

a. Deterministic:

Policy Iteration Optimal Value Function (V_pi)	[0.59 0.656 0.729 0.656 0.656 0. 0.81 0. 0.729 0.81 0.9 0. 0. 0.9 1. 0.]
Policy Iteration Optimal Policy (p_pi)	[1 2 1 0 1 0 1 0 2 1 1 0 0 2 2 0]

Stochastic:

Policy Iteration Optimal Value Function (V_pi)	[0.062 0.056 0.07 0.051 0.086 0. 0.11 0. 0.141 0.244 0.297 0. 0. 0.377 0.638 0.]
Policy Iteration Optimal Policy (p_pi)	[0 3 0 3 0 0 0 0 3 1 0 0 0 2 1 0]

b. Deterministic:

Value Iteration Optimal Value Function (V_vi)	[0.59 0.656 0.729 0.656 0.656 0. 0.81 0. 0.729 0.81 0.9 0. 0. 0.9 1. 0.]
Value Iteration Optimal Policy (p_vi)	[1 2 1 0 1 0 1 0 2 1 1 0 0 2 2 0]

Stochastic:

Value Iteration Optimal Value Function (V_vi)	[0.062 0.056 0.07 0.051 0.086 0. 0.11 0. 0.141 0.244 0.297 0. 0. 0.377 0.638 0.]
Value Iteration Optimal Policy (p_vi)	[0 3 0 3 0 0 0 0 3 1 0 0 0 2 1 0]

c. A deterministic environment is an environment where the outcome is based on the current state. Therefore, the decisions are more certain and based on our current state. However, in a stochastic environment, there is more uncertainty because we cannot know the outcomes based on the current state. This is why when we switch to a stochastic environment, there are more iterations than in the deterministic one. Our policy is also different, as we do not always know the outcome for the action that we apply. Our action is more random and our policy might not be “perfect”.