Yiping Guo s3177105

1.1 (a) Since this question is a deterministic MDP, use the Bellman equation to calculate the optimal value function for each state:

$$V(s) = R(s) + \gamma \frac{max}{a} T(s, a, s_1) V(s_1)$$

For goal state:

V(G) = 1

(b) Because every state has a reward of zero, let's iterate backward

V(G) = 1 + 0 where R(s) = 1 and $\gamma \frac{max}{a} T(s_{goal}, a_{goal}, s_{goal-1}) V(s_{goal-1}) = 0$

$$V(s_{goal-1}) = R(s_{goal-1}) + \gamma \frac{max}{a} T(s_{goal-1}, a_{goal-1}, s_{goal-2}) V(s_{goal-2}) = 0 + 0$$

Hence, the discount factor γ does not affect the optimal value function formula because it is always included in the zero part.

Since the discount factor γ does not appear in the optimal policy formula, optimal policy does NOT depend on the value of the discount factor γ

(c) $V(s) = R(s) + \gamma \frac{max}{a} T(s, a, s_1) V(s_1) = c + \gamma \frac{max}{a} T(s, a, s_1) c$

V(G) = 1 + c

It does not change the optimal policy because c is a constant and V(s) will only depend on the discount factor γ and transformation function T

(d) $V(s) = R(s) + \gamma \frac{max}{a} T(s, a, s_1) V(s_1) = a(c + R(s_1)) + \gamma \frac{max}{a} T(s, a, s_1) V(s_1)$

$= ac + aR(s_1) + \gamma \frac{max}{a} T(s, a, s_1) V(s_1)$

$= ac + aR(s_1) + \gamma \frac{max}{a} T(s, a, s_1)(R(s_1) + \gamma \frac{max}{a} T(s_1, a_1, s_2) V(s_2))$

$= ac + (a + \gamma \frac{max}{a} T(s, a, s_1)) R(s_1) + \gamma^2 \frac{max}{a} T(s, a, s_1) \frac{max}{a} T(s_1, a_1, s_2) V(s_2)$

V(G) = $ac \sum_{k=0}^{t} a^k$

Yes, this reward equation changes the optimal policy as can be seen from the above formulas. For instance, when a = 1, c = 1 the equalition is related to both s_1 and s_2. So it is not MDP anymore. (which only relate to s_1

$$V(s) = 1 + (1 + \gamma \frac{max}{a} T(s, a, s_1)) R(s_1) + \gamma^2 \frac{max}{a} T(s, a, s_1) \frac{max}{a} T(s_1, a_1, s_2) V(s_2)$$

1.2 (a) total discounted return

$G_t = \sum_{t=0}^{\infty} \gamma^t r_t = 0 + \sum_{t=1}^{\infty} 1 = 0$ (when t = 0)

(b) $G_t = \sum_{t=0}^{\infty} \gamma^t r_t = \frac{\gamma^2}{1-\gamma} + \sum_{t=1}^{\infty} 0 = \frac{\gamma^2}{1-\gamma}$ (when t = 0)

From the lecture notes of week 7, we know the optimal action from state s_0 is the optimal policy from state s_0

Use value iteration:

$$V(s) = R(s) + \gamma \frac{max}{a} T(s, a, s_1) V(s_1)$$

For a_1 , $R(s) = 1$ except $R(s_0)$, hence $V(s) = 1$

For a_2 , $R(s) = 0$ except $R(s_0)$, hence $V(s) = 0$

So the optimal action from s_0 is a_1