1: MDP

The following MDP represents a simple transportation problem.

- States: s1, s2, s3, s4, s5, s6, s7; s1 is the start state
- Actions: In states s1, s2 and s3, two actions are possible: WALK and TELEPORT
  States s4, s5, s6 and s7 are terminal/absorbing states.
- Reward: R(s1) = R(s2) = R(s3) = -0.1
  R(s4) = -4
  R(s5) = 2
  R(s6) = 5
  R(s7) = -2

The transition probabilities are given by the following tables:

ACTION: WALK		Ending State						
		s1	s2	s3	s4	s5	s6	s7
Starting State	s1	0.5	0.5	0	0	0	0	0
	s2	0	0	0	0.5	0.5	0	0
	s3	0	0	0.5	0	0	0.5	0

ACTION: TELEPORT		Ending State						
		s1	s2	s3	s4	s5	s6	s7
Starting State	s1	0	0.5	0.5	0	0	0	0
	s2	0	0	0.25	0.75	0	0	0
	s3	0	0	0	0	0	0.5	0.5

Assuming $V_0(s_1) = 0, V_0(s_2) = 0, V_0(s_3) = 0$ and for the terminal states:
$V_0(s_4) = R(s_4) = -4, V_0(s_5) = R(s_5) = 2, V_0(s_6) = R(s_6) = 5, V_0(s_7) = R(s_7) = -2$

(a) After one step of value iteration, what is the utility of $V_1(s_1)$, $V_1(s2)$ and $V_1(s_3)$? Please show all of your work.

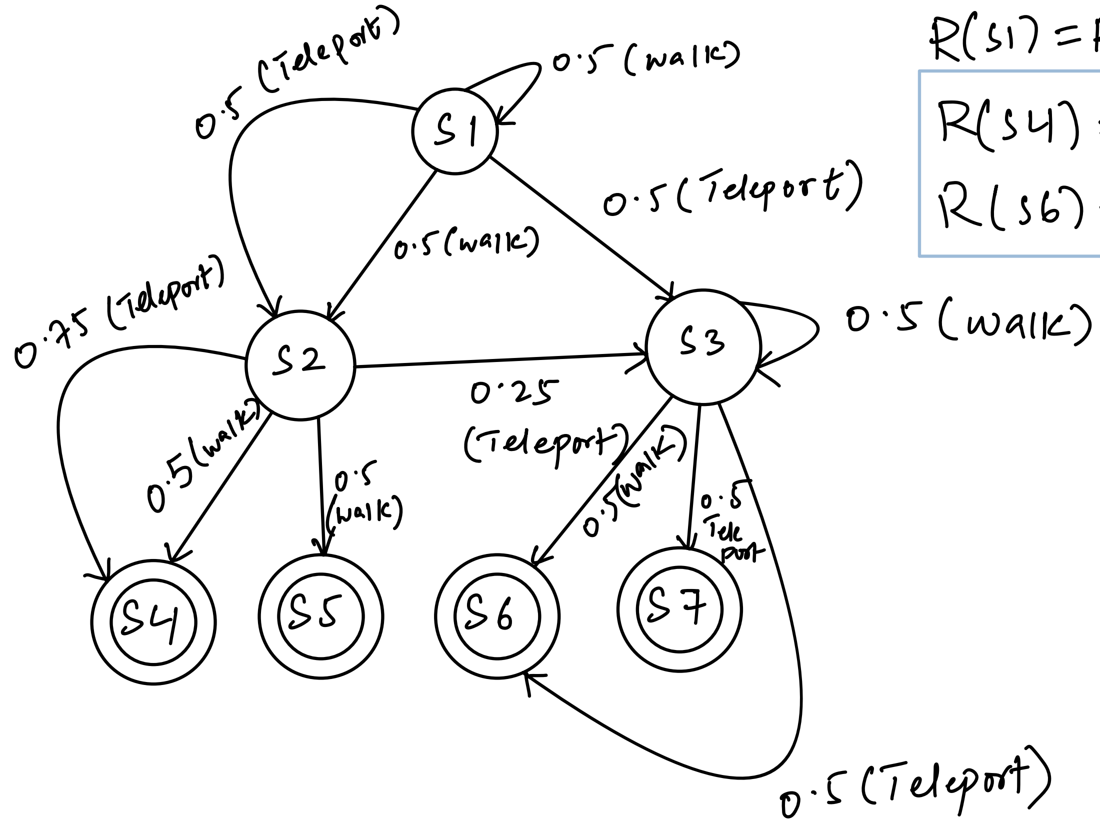
(b) After two steps of value iteration. What is the utility of $V_2(s_1)$, $V_2(s_2)$ and $V_2(s_3)$? Please show all of your work.

(c)(Programming) Please write a program to perform value iteration on this problem. Recall that the value iteration algorithm iterates and updates the utilities using:

$$U_{t+1}(S_i) = \max_a \left(R(S_i) + \gamma \sum_{S_j} P(S|S_i, a) \times U_t(S_j) \right)$$

- What is the utility of the states s_1, s_2, and s_3 after two iterations with $\gamma = 1$? Verify that the program output matches the result in question b.
- Assuming $\gamma = 0.95$, what is the utility of the states s_1, s_2, and s_3 when the algorithm converges? What is the optimal policy for this MDP?

Utility: sum of (discounted) reward
V: Maximize expected utility



$R(s1) = R(s2) = R(s3) = -0.1$

$$\boxed{\begin{array}{ll} R(s4) = -4 & R(s5) = 2 \\ R(s6) = 5 & R(s7) = -2 \end{array}}$$

Terminal.

$V_0(s1) = V_0(s2) = V_0(s3) = 0$ # initial expected utility (max sum of rewards)

$V_0(s4) = R(s4) = -4$, $V_0(s5) = R(s5) = 2$, $V_0(s6) = R(s6) = 5$, $V_0(s7) = R(s7) = -2$

a) # iteration 1

$$V_1(s_1) = \max_a \left(R(s_1) + 1 \sum_{s_j} P(s \mid s_1, a) * V_0(s_j) \right)$$

$$= \max_a \left(\begin{bmatrix} -0.1 + (0.5 \cdot 0) + (0.5 \cdot 0) \\ -0.1 + (0.5 \cdot 0) + (0.5 \cdot 0) \end{bmatrix} \right)$$

walk

Teleport

$$V_1(s1) = -0.1$$

$$V_1(s2) = \max_a \left(R(s2) + 1 \sum_{s_j} P(s \mid s_2, a) * V_0(s_j) \right)$$

$$= \max_a \left(\begin{bmatrix} -0.1 + (0.5 \cdot -4) + (0.5 \cdot 2) \\ -0.1 + (0.75 \cdot -4) + (0.25 \cdot 0) \end{bmatrix} \right)$$

$$V_1(s2) = -1.1$$

$$V_1(s3) = \max_a \left(R(s3) + 1 \sum_{s_j} P(s \mid s_3, a) * V_0(s_j) \right)$$

$$= \max_a \left(\begin{bmatrix} -0.1 + (0.5 \cdot 5) + (0.5 \cdot 0) \\ -0.1 + (0.5 \cdot 5) + (0.5 \cdot -2) \end{bmatrix} \right)$$

$$V_1(s3) = 2.4$$

b) # iteration # 2

$$V_2(s_1) = \max_a \left(R(s_1) + 1 \sum_{s_j} P(s \mid s_1, a) * V_1(s_j) \right)$$

$$= \max_a \left(\begin{bmatrix} -0.1 + (0.5 \cdot -0.1) + (0.5 \cdot -1.1) \\ -0.1 + (0.5 \cdot -1.1) + (0.5 \cdot 2.4) \end{bmatrix} \right) \begin{matrix} \text{Walk} \\ \text{Teleport} \end{matrix}$$

$$V_2(s_1) = 0.55$$

$$V_2(s_2) = \max_a \left(R(s_2) + 1 \sum_{s_j} P(s \mid s_2, a) * V_1(s_j) \right)$$

$$= \max_a \left(\begin{bmatrix} -0.1 + (0.5 \cdot -4) + (0.5 \cdot 2) \\ -0.1 + (0.75 \cdot -4) + (0.25 \cdot 2.4) \end{bmatrix} \right) \begin{matrix} \text{Walk} \\ \text{Teleport} \end{matrix}$$

$$V_2(s_1) = -1.1$$

$$V_2(s_3) = \max_a \left(R(s_3) + 1 \sum_{s_j} P(s \mid s_3, a) * V_1(s_j) \right)$$

$$= \max_a \left(\begin{bmatrix} -0.1 + (0.5 \cdot 5) + (0.5 \cdot 2.4) \\ -0.1 + (0.5 \cdot 5) + (0.5 \cdot -2) \end{bmatrix} \right) \begin{matrix} \text{Walk} \\ \text{Teleport} \end{matrix}$$

$$V_2(s_3) = 3.6$$

(c)

(i) same as part (b)

(ii) S1 - utility : 1.45 Policy : Teleport

S2 - utility : -1.05 Policy : Walk

S3 - utility : 4.33 Policy : Walk