Question 3

(a) For each state we have 4 possible actions, so the number of policies is $4^6 = 4096$.

(b) State transition matrix T:

$$T = \begin{pmatrix} 0.2 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0.8 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$R = (0, 0, 20, 10, 0, -10)^T$$

$$\Rightarrow V^{\pi_0} = (I - \gamma T)^{-1} R = \begin{array}{cc} \begin{pmatrix} 154.19 \\ 175.61 \\ 200 \\ -64.9 \\ -87.8 \\ -100 \end{pmatrix} & \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} \end{array}$$

(C) For state S_1:

$R(S_1, up) = V^0(S_1) = 154.19$

$R(S_1, down) = 0.2V^0(S_1) + 0.8V^0(S_4) = -21.08$

$R(S_1, left) = V^0(S_1) = 154.19$

$R(S_1, right) = 0.2V^0(S_1) + 0.8V^0(S_2) = 171.326$

$\Rightarrow \pi'(S_1) = \underline{right}$

For state S_2:

$R(S_2, up) = V^0(S_2) = 175.61$

$R(S_2, down) = 0.2V^0(S_2) + 0.8V^0(S_5) = -35.12$

$R(S_2, left) = 0.2V^0(S_2) + 0.8V^0(S_1) = 158.474$

$R(S_2, right) = 0.2V^0(S_2) + 0.8V^0(S_3) = 195.122$

$\Rightarrow \pi'(S_1) = \underline{right}$

For state S_3:

$\max\{V^0(S_3), 0.2V^0(S_3) + 0.8V^0(S_2)\}$

$= \max\{V^0(S_3), V^0(S_2)\} = V^0(S_3) = R(S_3, down)/\pi^0(S_3, right)$

$\rightarrow \underline{\pi'(S_3) = down.}$

For state S_4:

$\max\{V^0(S_4), V^0(S_5)\} = V^0(S_4) = R(S_4, up/down/left)$

$\Rightarrow \underline{\pi'(S_4) = down}$

For S_5:

$$\max\{V^0(S_4), V^0(S_5), V^0(S_1), V^0(S_6)\} = V_0(S_2) = R(S_5, up).$$

$$\rightarrow \underline{\pi'(S_5) = up}$$

For S_6:

$$\max_S \{V^0(S_5), V^0(S_6), V^0(S_3)\} = V^0(S_3) = R(S_6, up)$$

$$\rightarrow \underline{\pi'(S_3) = up}$$

(d) We can continue the iterative process in (C), our new transition matrix:

policy evaluation:

$$T' = \begin{pmatrix} 0.2 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0.8 & 0 & 0 & 0.2 \end{pmatrix}$$

$$\Rightarrow V^{\pi'} = (I - \gamma T')^{-1} R = \begin{array}{c} \begin{pmatrix} 154.19 \\ 175.61 \\ 200 \\ 100 \\ 154.19 \\ 163.41 \end{pmatrix} \end{array} \begin{array}{l} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array}$$

policy improvement:

① For S_1: $\max\{V'(S_1), V'(S_2), V'(S_4)\} = V'(S_2) = R(S_1, right)$

$\rightarrow \pi''(S_1) =$ right

② For S_2: $\max\{V'(S_1), V'(S_2), V'(S_3), V'(S_5)\} = V'(S_3) = R(S_2, right)$

$\rightarrow \pi''(S_2) =$ right.

③ For S_3: $\max\{V'(S_3), V'(S_2)\} = V'(S_3) = R(S_3, up/down/right)$

$\rightarrow \pi''(S_3) =$ down

④ For S_4: $\max\{V'(S_4), V'(S_5)\} = V'(S_5) = R(S_4, right)$

$\rightarrow \pi''(S_4) =$ right

⑤ For S_5: $\max\{V'(S_5), V'(S_4), V'(S_6), V'(S_2)\} = V'(S_2) = R(S_5, up)$

$\rightarrow \pi''(S_5) =$ up

⑥ For S_6: $\max\{V'(S_3), V'(S_5), V'(S_6)\} = V'(S_3) = R(S_6, up)$

$\rightarrow \pi''(S_6) =$ up.

$\Rightarrow \pi'' = (right, right, down, right, up, up) \neq \pi'$,

so we keep doing iteration:

policy evaluation: new transition matrix:

$$T'' = \begin{pmatrix} 0.2 & 0.8 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0.8 & 0 \\ 0 & 0.8 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0.8 & 0 & 0 & 0.2 \end{pmatrix}$$

$$\Rightarrow V^{\pi'''} = (I - \gamma T'')^{-1} R = \begin{pmatrix} 154.19 \\ 175.61 \\ 200 \\ 147.58 \\ 154.19 \\ 163.41 \end{pmatrix}$$

policy improvement:

① For S_1: $\max\{V^3(S_1), V^3(S_2), V^3(S_4)\} = V^3(S_2) = R(S_1, \text{right})$

$\rightarrow \pi^3(S_1) = \text{right}$

② For S_2: $\max\{V^3(S_1), V^3(S_2), V^3(S_3), V^3(S_5)\} = V^3(S_3) = R(S_2, \text{right})$

$\rightarrow \pi^3(S_2) = \text{right}$

③ For S_3: $\max\{V^3(S_3), V^3(S_2)\} = V^3(S_3) = R(S_3, \text{up/down/right})$

$\rightarrow \pi^3(S_3) = \text{down}$.

④ For S_4: $\max\{V^3(S_4), V^3(S_5)\} = V^3(S_5) = R(S_4, \text{right})$

$\rightarrow \pi^3(S_4) = \text{right}$

⑤ For S_5: $\max\{V^3(S_5), V^3(S_4), V^3(S_6), V^3(S_2)\} = V^3(S_2) = R(S_5, \text{up})$

$\rightarrow \pi^3(S_5) = \text{up}$

⑥ For S_6: $\max\{V^3(S_3), V^3(S_5), V^3(S_6)\} = V^3(S_3) = R(S_6, \text{up})$

$\rightarrow \pi^3(S_6) = \text{up}$

$\Rightarrow \pi^3 = (\text{right}, \text{right}, \text{down}, \text{right}, \text{up}, \text{up}) = \pi''$

$\Rightarrow$ Now policy iteration algorithm converges

$$\Rightarrow V^* = \begin{pmatrix} 154.19 \\ 175.61 \\ 200 \\ 147.58 \\ 154.19 \\ 163.41 \end{pmatrix}$$

(e) The optimal value function is unique since V^* is defined as the best value that can be achieved by any state:

$$V^*(S) = \max_{\pi} V^{\pi}(S).$$

(f) By (d) we know the optimal policy is:

$$\pi^* = \underset{S_1}{(\text{right}}, \underset{S_2}{\text{right}}, \underset{S_3}{\text{down}}, \underset{S_4}{\text{right}}, \underset{S_5}{\text{up}}, \underset{S_6}{\text{up})}$$

(g) No. Since there could be other policies $\tilde{\pi}$ that also returns the same optimal value function. For instance, $\tilde{\pi} = \underset{S_1}{(\text{right}}, \underset{S_2}{\text{right}}, \underset{S_3}{\text{up}}, \underset{S_4}{\text{right}}, \underset{S_5}{\text{up}}, \underset{S_6}{\text{up})}$

(obtained by changing $\pi^*(S_3)$ into "up", which yields the same reward) is also an optimal policy.

(h). If we scale rewards for each state by a constant factor k, then all deductions would remain unchanged, which means we could yield the same optimal policy, with the corresponding value function scaled by k.

Proof of this proposition:

Let $R' = kR$ (where k is a real constant), then by the new value function under policy π is:

$$V'_\pi = (I - \gamma T^\pi)^{-1} \cdot kR = k \cdot (I - \gamma T^\pi)^{-1} R = k \cdot V_\pi$$

When we do policy improvement for a particular state S, we compute:

$$\max_{S_i} \{ p V_\pi(S_i) + (1-p) V_\pi(S) \} = \max_{S_i} \{ V_\pi(S_i) \}$$

If we scale V_π by k, $\max_{S_i}\{V_\pi(S_i)\}$ wouldn't change, implying the new transition matrix $T^{\pi'}$ is also unchanged. Thus, by induction, the entire policy iteration process would yield the same optimal policy, with the corresponding V^* scaled by a factor k.