Northeastern University
College of Engineering

EECE 5698 - ST: Reinforcement Learning Instructor: Prof. Mahdi Imani

HW2 Solution

Problem 1.

Consider the following system with two states $s_k \in \{s^1 = 0, s^2 = 1\}$.
There are two possible actions: a^1 and a^2. The transition probabilities can be expressed as:

$$p(s'|s,a^1)\begin{cases}1 & s=0, s'=0\\ 0 & s=0, s'=1\\ 0 & s=1, s'=0\\ 1 & s=1, s'=1\end{cases} \qquad p(s'|s,a^2)\begin{cases}0 & s=0, s'=0\\ 1 & s=0, s'=1\\ 1 & s=1, s'=0\\ 0 & s=1, s'=1\end{cases}$$

Reward function is as follows: $\begin{cases} moving\ to\ state\ s^2: +1\\ moving\ to\ state\ s^1: 0\\ action\ a^1\ and\ a^2: 0\end{cases}$

Start with a random policy $\pi^0(s^1) = a^1, \pi^0(s^2) = a^1, \gamma = 0.9, \theta = 0.85.$ Use Policy Iteration to compute $\pi^1(s^1),\ \pi^1(s^2)$. Use $V_0(s^1) = V_0(s^2) = 0$, for initialization of Policy Evaluation.

Solution: $\pi^0(s^1) = \pi^0(s^2) = a^1$

$$V_{k+1}(s) = \sum_{s' \in \{s^1, s^2\}} P(s' \mid s, \pi(s))\left[R(s, \pi(s), s') + \gamma V_k(s')\right]$$

$$V_{k+1}(s^1) = \underbrace{P(s^1 \mid s^1, \overbrace{\pi^0(s^1)}^{a^1})}_{1} \left[R(s^1, \overbrace{\pi^0(s^1)}^{a^1}, s^1) + \gamma V_k(s^1)\right]$$

$$+ \underbrace{P(s^2 \mid s^1, \underbrace{\pi^0(s^1)}_{a^1})}_{0} \left[R(s^1, \underbrace{\pi^0(s^1)}_{a^1}, s^2) + \gamma V_k(s^2)\right]$$

$$\underbrace{V_{k+1}(s^1) = R(s^1, a^1, s^1) + \gamma V_k(s^1)}$$

For $s^1 \rightarrow$ $$V_{k+1}(s^2) = \overbrace{P(s^1 \mid s^2, \underbrace{\pi^0(s^2)}_{a^1})}^{0} \left[R(s^2, \overbrace{\pi^0(s^2)}^{a^1}, s^1) + \gamma V_k(s^1)\right]$$

$$+ \underbrace{P(s^2 \mid s^2, \underbrace{\pi^0(s^2)}_{a^1})}_{1} \left[R(s^2, \underbrace{\pi^0(s^2)}_{a^1}, s^2) + \gamma V_k(s^2)\right]$$

$$\underbrace{V_{k+1}(s^2) = R(s^2, a^1, s^2) + \gamma V_k(s^2)}$$

$$V_0(s^1) = V_0(s^2) = 0$$

$$\begin{cases} V_1(s^1) = R(s^1, a^1, s^1) + \gamma V_0(s^1) = 0 + \gamma 0 = 0 \\ V_1(s^2) = R(s^2, a^1, s^2) + \gamma V_0(s^2) = 1 + \gamma 0 = 1 \end{cases}$$

$$\max_{s \in S} |V_1(s) - V_0(s)| = \max\{|V_1(s^1) - V_0(s^1)|, |V_1(s^2) - V_0(s^2)|\}$$
$$= \max\{0, 1\} = 1 > \theta$$

$$V_2(s^1) = R(s^1, a^1, s^1) + \gamma V_1(s^1) = 0 + \gamma 0 = 0$$

$$V_2(s^2) = R(s^2, a^1, s^2) + \gamma V_1(s^2) = 1 + \gamma \times 1 = 1.9$$

$$\max\{|V_2(s^1)-V_1(s^1)|, |V_2(s^2)-V_1(s^2)|\}$$

$$=\max\{0, 0.9\}=0.9 > \theta \leftarrow 0.85 \quad \times$$

$$V_3(s^1)=R(s^1,a^1,s^1)+\gamma V_2(s^1) = 0+\gamma\, 0 = 0$$

$$V_3(s^2)=R(s^2,a^1,s^2)+\gamma V_2(s^2) = 1+\gamma\, 1.9 = 2.7$$

$$\max\{|V_3(s^1)-V_2(s^1)|, |V_3(s^2)-V_2(s^2)|\}$$

$$=\max\{0, 0.81\}=0.81 < \theta \leftarrow 0.85 \quad \checkmark$$

Policy Evaluation step stops.

$$V^{\pi^0}(s^1)=0, \quad V^{\pi^0}(s^2)=2.7$$

Policy Improvement:

Using latest $V(s)$ obtained from policy Evaluation, we have

$$\pi'(s)=\operatorname{argmax} \sum_{s'\in S} P(s'|s,a)\,[R(s,a,s')+\gamma V(s')]$$

$$\begin{aligned}\pi'(s^1)=\operatorname{argmax}\Big\{ & \underbrace{P(s^1|s^1,a^1)}_{1}[\underbrace{R(s^1,a^1,s^1)}_{0}+\gamma \underbrace{V(s^1)}_{0}]+\underbrace{P(s^2|s^1,a^1)}_{0}[\underbrace{R(s^1,a^1,s^2)}_{1}+\gamma \underbrace{V(s^2)}_{2.7}] \\ , & \underbrace{P(s^1|s^1,a^2)}_{0}[\underbrace{R(s^1,a^2,s^1)}_{0}+\gamma \underbrace{V(s^1)}_{0}]+\underbrace{P(s^2|s^1,a^2)}_{1}[\underbrace{R(s^1,a^2,s^2)}_{1}+\gamma \underbrace{V(s^2)}_{2.7}]\Big\}\end{aligned}$$

$$\pi'(s^1)=\operatorname{argmax}\{0, 3.43\}=a^2$$

$$\begin{aligned}\pi'(s^2)=\operatorname{argmax}\Big\{ & \underbrace{P(s^1|s^2,a^1)}_{0}[\underbrace{R(s^2,a^1,s^1)}_{0}+\gamma \underbrace{V(s^1)}_{0}]+\underbrace{P(s^2|s^2,a^1)}_{1}[\underbrace{R(s^2,a^1,s^2)}_{1}+\gamma \underbrace{V(s^2)}_{2.7}] \\ , & \underbrace{P(s^1|s^2,a^2)}_{1}[\underbrace{R(s^2,a^2,s^1)}_{0}+\gamma \underbrace{V(s^1)}_{0}]+\underbrace{P(s^2|s^2,a^2)}_{0}[\underbrace{R(s^2,a^2,s^2)}_{1}+\gamma \underbrace{V(s^2)}_{2.7}]\Big\}\end{aligned}$$

$$\pi'(s^2)=\operatorname{argmax}\{3.43, 0\}=a^1$$

Problem 2.

Consider the problem defined in Problem 1.

a) Given $\begin{bmatrix} V_0(s^1) \\ V_0(s^2) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\gamma = 0.9$, perform Value Iteration method to compute V_1, V_2, V_3 .

b) Compute $\pi(s = 0)$ and $\pi(s = 1)$ associated with V_3.

Solution:

a) $V_0(s^1) = V_0(s^2) = 0$

$$V_{k+1}(s^1) = \max_{a\in\{a^1,a^2\}} \left[\sum_{s'\in S} R(s^1,a,s') + \gamma\, P(s'|s,a)\, V_k(s')\right]$$

$$V_{k+1}(s^1) = \max_{a\in A} \left[P(s^1|s^1,a)\Big(R(s^1,a,s^1)+\gamma\, V_k(s^1)\Big) + P(s^2|s^1,a)\Big(R(s^1,a,s^2)+\gamma\, V_k(s^2)\Big)\right]$$

$$V_{k+1}(s^1) = \max \Big\{\left[P(s^1|s^1,a^1)\Big(R(s^1,a^1,s^1)+\gamma\, V_k(s^1)\Big) + P(s^2|s^1,a^1)\Big(R(s^1,a^1,s^2)+\gamma\, V_k(s^2)\Big)\right],$$
$$\left[P(s^1|s^1,a^2)\Big(R(s^1,a^2,s^1)+\gamma\, V_k(s^1)\Big) + P(s^2|s^1,a^2)\Big(R(s^1,a^2,s^2)+\gamma\, V_k(s^2)\Big)\right]\Big\}$$

$$\underline{V_{k+1}(s^1) = \max\left\{R(s^1,a^1,s^1)+\gamma\, V_k(s^1),\ R(s^1,a^2,s^2)+\gamma\, V_k(s^2)\right\}}$$

$$\underline{V_{k+1}(s^2) = \max\left\{R(s^2,a^1,s^2)+\gamma\, V_k(s^2),\ R(s^2,a^2,s^1)+\gamma\, V_k(s^1)\right\}}$$

$$V_1(s^1) = \max\left\{\underbrace{R(s^1,a^1,s^1)}_{0}+\gamma\underbrace{V_0(s^1)}_{0},\ \underbrace{R(s^1,a^2,s^2)}_{+1}+\gamma\underbrace{V_0(s^2)}_{0}\right\} = 1$$

$$V_1(s^2) = \max\left\{\underbrace{R(s^2,a^1,s^2)}_{1}+\gamma\underbrace{V_0(s^2)}_{0},\ \underbrace{R(s^2,a^2,s^1)}_{0}+\gamma\underbrace{V_0(s^1)}_{0}\right\} = 1$$

$$V_2(s^1) = \max\left\{\underbrace{R(s^1,a^1,s^1)}_{0}+\gamma\underbrace{V_1(s^1)}_{1},\ \underbrace{R(s^1,a^2,s^2)}_{1}+\gamma\underbrace{V_1(s^2)}_{1}\right\} = 1.9$$

$$V_2(s^2) = \max\left\{\underbrace{R(s^2,a^1,s^2)}_{1}+\gamma\underbrace{V_1(s^2)}_{1},\ \underbrace{R(s^2,a^2,s^1)}_{0}+\gamma\underbrace{V_1(s^1)}_{1}\right\} = 1.9$$

$$V_3(s^1) = \max\left\{ \underbrace{R(s^1,a^1,s^1)}_{0} + \gamma \underbrace{V_2(s^1)}_{1.9},\ \underbrace{R(s^1,a^2,s^2)}_{1} + \gamma \underbrace{V_2(s^2)}_{1.9} \right\} = 2.71$$

$$V_3(s^2) = \max\left\{ \underbrace{R(s^2,a^1,s^2)}_{1} + \gamma \underbrace{V_2(s^2)}_{1.9},\ \underbrace{R(s^2,a^2,s^1)}_{0} + \gamma \underbrace{V_2(s^1)}_{1.9} \right\} = 2.71$$

b)

$$\pi(s) = \underset{a \in \{a^1,a^2\}}{\operatorname{argmax}} \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V(s')\right]$$

$$\pi(s^1) = \underset{a \in \{a^1,a^2\}}{\operatorname{argmax}} \left[\sum_{s' \in S} R(s^1,a,s') + \gamma\, P(s'|s,a)\, V_k(s')\right]$$

$$\pi(s^1) = \operatorname{argmax}\Big\{ P(s^1|s^1,a^1)\left(R(s^1,a^1,s^1) + \gamma V_k(s^1)\right) + P(s^2|s^1,a^1)\left[R(s^1,a^1,s^2) + \gamma V_k(s^2)\right],$$

$$P(s^1|s^1,a^2)\left(R(s^1,a^2,s^1) + \gamma V_k(s^1)\right) + P(s^2|s^1,a^2)\left[R(s^1,a^2,s^2) + \gamma V_k(s^2)\right]\Big\}$$

$$\pi(s^1) = \operatorname{argmax}\left\{ \underbrace{R(s^1,a^1,s^1)}_{0} + \gamma \underbrace{V(s^1)}_{2.71},\ \underbrace{R(s^1,a^2,s^2)}_{1} + \gamma \underbrace{V(s^2)}_{2.71} \right\} = a^2$$

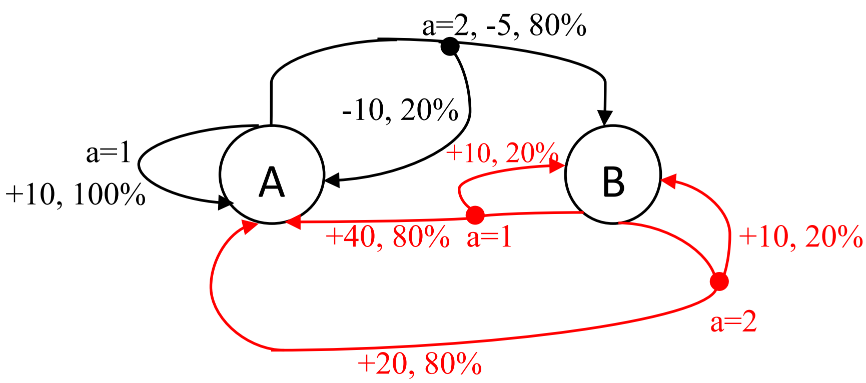
Similarly:

$$\pi(s^2) = \operatorname{argmax}\left\{ \underbrace{R(s^2,a^1,s^2)}_{1} + \gamma \underbrace{V(s^2)}_{2.71},\ \underbrace{R(s^2,a^2,s^1)}_{1} + \gamma \underbrace{V(s^1)}_{1.71} \right\} = a^1$$

Problem 3.

Consider the following MDP having two states: A, B. In each state, there are two possible actions: 1 and 2. The transition model and reward are shown in the diagram below. Apply Policy Iteration to determine the optimal policy and state values of A and B. Assume the initial policy is action 2 for both staters, $\gamma = 0.9$.
For evaluation of policy, you need to solve two set of linear equations for the following form, instead of iterative steps of policy evaluation:

$$V^{\pi}(s) = \sum\nolimits_{s',r} P\big(s' \big| s, \pi(s)\big)[R(s, \pi(s), s') + \gamma V^{\pi}(s')]$$



*Here is an example of transition and reward from the diagram:
In state A, action 2 moves the agent to state B with probability 0.8 with the corresponding reward -5, and make the agent stay at state A with probability 0.2 and corresponding reward -10.

Solution:

$$V^{\pi}(s) = \sum P(s', r | s, a) [r + \gamma V^{\pi}(s')]$$

Policy Evaluation for π^0

$$\begin{cases} V^{\pi^0}(A) = 0.8 [-5 + 0.9 V^{\pi^0}(B)] + 0.2 [-10 + 0.9 V^{\pi^0}(A)] \\ V^{\pi^0}(B) = 0.8 [+20 + 0.9 V^{\pi^0}(A)] + 0.2 [+10 + 0.9 V^{\pi^0}(B)] \end{cases}$$

$$\begin{matrix} 0.82 V^{\pi^0}(A) - 0.72 V^{\pi^0}(B) = -6 \\ -0.72 V^{\pi^0}(A) + 0.82 V^{\pi^0}(B) = 18 \end{matrix} \longrightarrow \begin{cases} V^{\pi^0}(A) = 52.2 \\ V^{\pi^0}(B) = 67.8 \end{cases}$$

Policy Improvement for π^0

$$\pi'(A) = \text{argmax} \left\{ \overbrace{1 [10 + 0.9 V^{\pi^0}(A)]}^{a=1} \; {}_{56.9}, \; \underbrace{0.8 [-5 + 0.9 V^{\pi^0}(B)] + 0.2 [-10 + 0.9 V^{\pi^0}(A)]}_{a=2} \; {}_{52.2} \right\} = 1$$

$$\pi'(B) = \text{argmax} \left\{ \overbrace{0.8 [40 + 0.9 V^{\pi^0}(A)] + 0.2 [+10 + 0.9 V^{\pi^0}(B)]}^{a=1} \; {}_{83.7}, \; \underbrace{0.8 [20 + 0.9 V^{\pi^0}(A)] + 0.2 [10 + 0.9 V^{\pi^0}(B)]}_{a=2} \; {}_{67} \right\} = 1$$

$\pi' \neq \pi^0 \longrightarrow$ Continue

$\pi^1(A) = \pi^1(B) = 1$

Policy Evaluation for π^1

$$\begin{cases} V^{\pi^1}(A) = 1\left[10 + 0.9\, V^{\pi^1}(A)\right] \\ V^{\pi^1}(B) = 0.8\left[40 + 0.9\, V^{\pi^1}(A)\right] + 0.2\left[10 + 0.9\, V^{\pi^1}(B)\right] \end{cases}$$

$$\begin{matrix} 0.1\, V^{\pi^1}(A) = 10 \\ -0.72\, V^{\pi^1}(A) + 0.82\, V^{\pi^1}(B) = 34 \end{matrix} \longrightarrow \begin{cases} V^{\pi^1}(A) = 100 \\ V^{\pi^1}(B) = 129.2 \end{cases}$$

Policy Improvement for π

$$\pi^2(A) = \text{argmax}\left\{ \overbrace{1\left[10 + 0.9\, V^{\pi^1}(A)\right]}^{a=1}, \underbrace{0.8\left[-5 + 0.9\, V^{\pi^1}(B)\right] + 0.2\left[-10 + 0.9\, V^{\pi^1}(A)\right]}_{a=2} \right\} = 2$$

100

105

$$\pi^2(B) = \text{argmax}\left\{ \overbrace{0.8\left[40 + 0.9\, V^{\pi^1}(A)\right] + 0.2\left[-10 + 0.9\, V^{\pi^1}(B)\right]}^{a=1}, \underbrace{0.8\left[20 + 0.9\, V^{\pi^1}(A)\right] + 0.2\left[10 + 0.9\, V^{\pi^1}(B)\right]}_{a=2} \right\} = 1$$

129.2

113

$\pi^2 \neq \pi^1 \rightarrow$ continue

$$V^* = V_5$$

Row-wise

$$\pi^* = \underset{a \in A}{\operatorname{argmax}} \; R_s^a + \gamma \, M(a) \, V^*$$

$$\pi^* = \operatorname{argmax} \left\{ \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2.2738 \\ 3.0431 \end{bmatrix}, \begin{bmatrix} -1 \\ 0.8 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 2.2738 \\ 3.0431 \end{bmatrix} \right\}$$

$$= \operatorname{argmax} \left\{ \begin{bmatrix} 2.7003 \\ 3.2388 \end{bmatrix}, \begin{bmatrix} 1.0464 \\ 3.4696 \end{bmatrix} \right\} = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix}$$

Problem 4.

Consider the following maze with 14 states and a goal. The agent can take one of the following four actions at any given state $A = \{UP, Down, Right, Left\}$

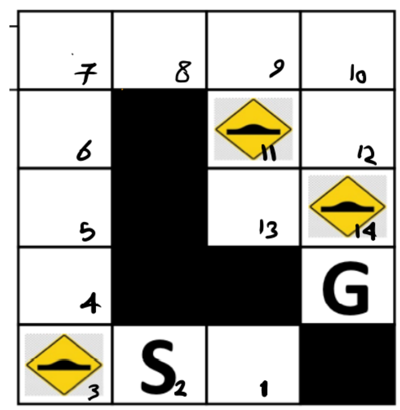
The state transitions are deterministic; for example $P(s'=10 \mid s=12, a=U) = 1$

The reward is as follows:
$$\begin{cases} -1 & \text{taking any action} \\ +200 & \text{moving to goal} \\ -10 & \text{moving to bump} \end{cases}$$

a) Using $\gamma=1$ and $\theta=0.5$, perform Vector-form Value Iteration method with $V_0(s)=0$ to compute V^*.

b) Compute the optimal Policy.

* Show all intermediate state values in maze, without details of calculation.



Solution:

V_0

$\delta = \max |V_1 - V_0| = 199$

V_1

$V(9) = \max\{-1, -11, -1, -1\} = -1$

$V(10) = \max\{-1, -1, -1, -1\} = -1$

$V(11) = \max\{-1, -1, -1, -11\} = -1$

$V(12) = \max\{-1, -11, -1, -11\} = -1$

$V(14) = \max\{\overset{U}{-1}\ \overset{D}{199}\ \overset{R}{-11}\ \overset{L}{-1}\} = 199$

$P(12 \mid 14, U)[R(14, U, 12) + \gamma V_0(12)]$

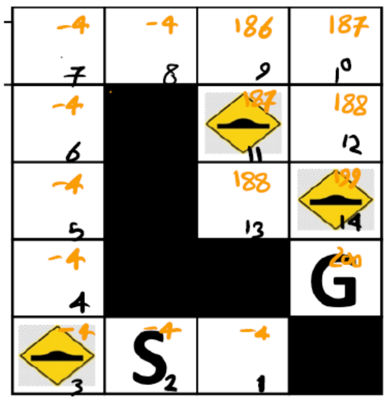
$\delta = 189$

V_2

$\delta = 189$

V_3

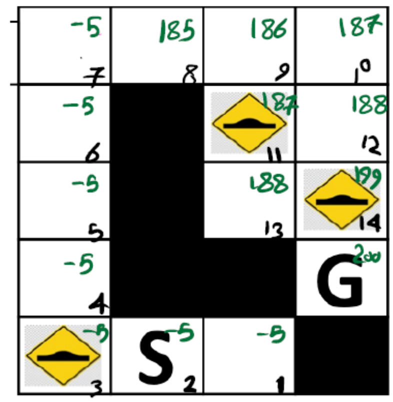
V4



δ=189

δ=189

V5



δ=189

V6

δ=189

V7

δ=189

V8

δ=189

V9

V_{10}

$\delta = 189$

V_{11}

$\delta = 179$

$\delta = 179$

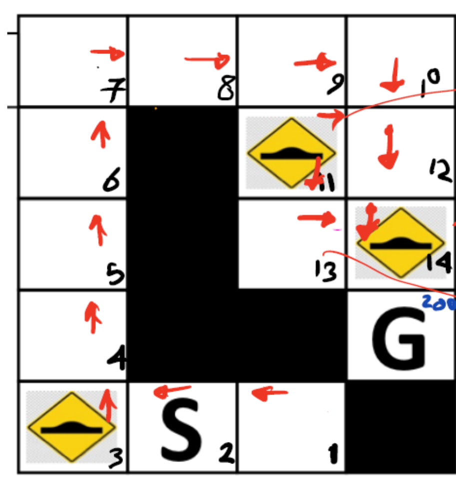
V_{12}

the same

$\delta = 0$

V_{13}

$$V_{13} = V^* \rightarrow \pi^*(s) = \arg\max \sum_{s',r} P(s'|s,a)\left[R(s,a,s') + \gamma V^*(s')\right]$$



U D R L
argmax{185, 187, 187, 176}

U D R L
argmax{186, 188, 187, 176}

U D R L
argmax{187, 199, 188, 187}

U D R L
argmax{176 187 188 187}

Problem 5.

For an MDP defined by state space S, action space A, Reward $R(s, a, s')$ and transition probability $p(s'|s, a)$, write the following:

a) For a given policy π, write

- $V^\pi(s)$ based on V^π
- $V^\pi(s)$ based on Q^π
- $Q^\pi(s,a)$ based on V^π
- $Q^\pi(s,a)$ based on Q^π

b) For the optimal policy π^*, write

- $V^*(s)$ based on V^*
- $V^*(s)$ based on Q^*
- $Q^*(s,a)$ based on V^*
- $Q^*(s,a)$ based on Q^*

- An example of response:

$$v^\pi(s) = \sum_{s'} p\big(s' \big| s, \pi(s)\big)[R(s, \pi(s), s') + \gamma V^\pi(s')]$$

Solution:

a) $$V^{\pi}(s) = \sum_{s'} P(s' | s, \pi(s)) \left[R(s, \pi(s), s') + \gamma V^{\pi}(s') \right]$$

$$V^{\pi}(s) = \sum_{s'} P(s' | s, \pi(s)) \left[R(s, \pi(s), s') + \gamma Q^{\pi}(s', \pi(s')) \right]$$

$$Q^{\pi}(s, a) = \sum_{s'} P(s' | s, a) \left[R(s, a, s') + \gamma V^{\pi}(s') \right]$$

$$Q^{\pi}(s, a) = \sum_{s'} P(s' | s, a) \left[R(s, a, s') + \gamma Q^{\pi}(s', \pi(s')) \right]$$

b) $$V^{*}(s) = \max_{a \in A} \sum_{s'} P(s' | s, a) \left[R(s, a, s') + \gamma V^{*}(s') \right]$$

$$V^{*}(s) = \max_{a \in A} \sum_{s'} P(s' | s, a) \left[R(s, a, s') + \gamma \max_{a' \in A} Q^{*}(s', a') \right]$$

$$Q^{*}(s, a) = \sum_{s'} P(s' | s, a) \left[R(s, a, s') + \gamma V^{*}(s) \right]$$

$$Q^{*}(s, a) = \sum_{s'} P(s' | s, a) \left[R(s, a, s') + \gamma \max_{a' \in A} Q^{*}(s', a') \right]$$