ELEN E6885 Introduction to Reinforcement Learning HW2

Tong Wu, tw2906

Problem 1

Part 1

True

Part 2

False

Part 3

True

Part 4

True

Part 5

False

Problem 2

Part 1

The policy iteration use policy evaluation and policy improvement to process the optimisation. Two methods will each do an iteration until the policy is converged. Since each policy evaluation is started with the previous policy, so the speed of convergence will increased, which means use less iteration can get convergence. The value iteration include the Bellman optimality equation into the update rule and one policy evaluation.

Part 2

Problem 2.2

For policy π,

$$\pi = 1 \quad \text{when} \quad a = a^*(s)$$

$$a^*(s) = \arg\max \left(R_s^a + \gamma \Sigma_{s'} P_{ss'}^a V(s') \right)$$

So for the first step of evaluation:

$$V'(s) = \sum_a \pi \left(R_s^a + \gamma \sum_{s'} P_{ss'}^a V(s') \right)$$

$$= R_s^{a(s)} + \gamma \sum_{s'} P_{ss'}^a V(s')$$

$$= \max \left(R_s^a + \gamma \sum_{s'} P_{ss'}^a V(s') \right)$$

Which will generate the same value function as one iteration of the value iteration.

Problem 3

Part 1

Problem 3.1

$$\gamma = 0.5$$

$$\begin{bmatrix} v(A) \\ v(B) \\ v(C) \\ v(D) \end{bmatrix} = \begin{bmatrix} R(A) \\ R(B) \\ R(C) \\ R(D) \end{bmatrix} + \gamma \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.2 & 0 & 0.5 & 0.3 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} V(A) \\ V(B) \\ V(C) \\ V(D) \end{bmatrix}$$

$R(A) = 0.5 \times (-2) \times 2 = -2$

$R(B) = 0.2 \times (-2) + 0.5 \times (-4) + 0.3 \times (-2) = -3$

$R(C) = -2$

$R(D) = 0$

$$\begin{bmatrix} v(A) \\ v(B) \\ v(C) \\ v(D) \end{bmatrix} = \begin{bmatrix} -2 \\ -3 \\ -2 \\ 0 \end{bmatrix} + 0.5 \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.2 & 0 & 0.5 & 0.3 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} V(A) \\ V(B) \\ V(C) \\ V(D) \end{bmatrix}$$

$$= \begin{bmatrix} -3.97 \\ -3.90 \\ -2 \\ 0 \end{bmatrix}$$

Part 2

Problem 3.2

For initial states $= 0$, $\gamma = 1$, $K = 1$:

$V(A) = -2 \quad V(B) = -3 \quad V(C) = -2$

For $K = 2$

$V(A) = 0.5 \times (-4) + 0.5 \times (-5) = -4.5$

$V(B) = 0.3 \times (-2) + 0.2 \times (-4) + 0.5 \times (-6) = -4.4$

$V(C) = 1 \times (-2) = -2$

Part 3

Problem 3.3

K=3, For A:

$q(A,a_1) = -2+1\times(-4.4) = -6.4$

$q(A,a_2) = -2+(-4.5) = -6.5$

$\max(q(A,a)) = -6.4$

For B:

$q(B,a_1) = -4+(-2) = -6$

$q(B,a_2) = -2+(-4.5) = -6.5$

$q(B,a_3) = -2+0 = -2$

$\max(q(B,a)) = -2$

For C:

$\max(q(C,a)) = -2$

Problem 4

Part 1

Problem 4.1.

For $T^*(v) = \max R^a + \gamma P^a v$, which is the Bellman optimality backup.

For any number of n,

$T^*(V_n) = V_{(n+1)}$, $V_{(n+1)} = V_n$

For the case n+1 and n+2,

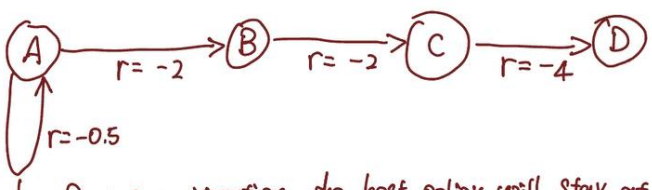
$V_{(n+1)} = T^*(V_n) = T_{(n+1)} = T^*(V_{n+1}) = V_{(n+2)}$

According to this, the each step will not change the best policy.

Part 2

Problem 4.2

For example:

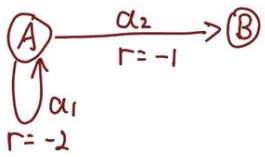


For the first two iteration, the best policy will stay at state A, while after that, the best policy changed.

Problem 5

Problem 5

For example, for the MDP model



For both a_1 and a_2

$$V(A) = a_1(-2 + V(A)) + a_2(-1)$$

Which shows that it cannot guarantee to convergence Vπ.

Problem 6

Part 1

Problem 6.1

(2, 1) -> right

(2, 2) -> right

(1, 1) -> up

(1, 2) -> left.

Part 2

Problem 6.2

For round 1:

$V(1,2) = 0$

$V(2,1) = 0$

For round 2:

$V_2(1,2) = 0.8 \times 0.9 \times 0.8 \times 5 + 0.1 \times (-5) = 2.4$

$V_2(2,1) = 0.8 \times 0.9 \times 0.8 \times 5 + 0 = 2.9$

Part 3

Problem 6.3.

For state (1,1), $r = 1$

$$V_{(1,1)} = \frac{(0+0+1\times(-5)) + (0+0+0+1\times 5) + (0+0+0+1\times 5)}{3}$$

$$= \frac{-5+5+5}{3}$$

$$= \frac{5}{3}$$

For state (2,2)

$$V_{(2,2)} = \frac{5+5}{2} = 5$$

Part 4

Problem 6.4

For $\alpha = 0.1$, $\gamma = 0.9$, $V_0 = 0$

For trails 1):

$V_{(1,1)} = 0 + 0.1 \times (0.9 \times 0) = 0$

$V_{(1,2)} = 0 + 0.1 \times (-5 + 0.9 \times 0) = -0.5$

$V_{(1,3)} = 0 + 0.1 \times (0.9 \times 0) = 0$

For trails 2):

$V_{(1,1)} = 0 + 0.1 \times (0 + 0.9 \times (-0.5)) = 0.045 \times 0.1 = -0.045$

$V_{(1,2)} = -0.5 + 0.1 \times (0 + 0.9 \times 0 + 0.5) = -0.45$

$V_{(2,2)} = 0 + 0.1 \times (5 + 0.9 \times 0) = 0.5$

$V_{(2,3)} = 0 + 0.1 \times (0 + 0.9 \times 0) = 0$

TD-0 learning has update for these states that non-zero.