COMP 440 Homework 3

Tony Chen(xc12) and Adam Wang(sw33)

September 2016

1 Policy evaluation and pacman

- Assume $\lambda = 1.0$:

	1	2	3	4	5
V^{π_0}	20	20	20	20	20
V^{π_1}	50	40	30	20	20
V^{π_2}	60	50	40	30	20
V^*	60	50	40	30	20

- If such λ exists, it must satisfy:

$$20 \quad > \quad 10 + 10\lambda + 10\lambda^2 + 20\lambda^3$$

and

$$20 \quad > \quad 10 + 10\lambda + 10\lambda^2 + 10\lambda^3 + 20\lambda^4$$

We can see that for any $0 \leq \lambda < 0.5$, for example, $\lambda = 0.25$, π_0 is strictly better than the other two.

- If such λ exists, it must satisfy:

$$10 + 10\lambda + 10\lambda^2 + 20\lambda^3 \quad > \quad 20$$

and

$$10 + 10\lambda + 10\lambda^2 + 20\lambda^3 \quad > \quad 10 + 10\lambda + 10\lambda^2 + 10\lambda^3 + 20\lambda^4$$

The first inequality requires $\lambda > 0.5$ while the second requires $\lambda < 0.5$, so none such λ exists.

- If such λ exists, it must satisfy:

$$10 + 10\lambda + 10\lambda^2 + 10\lambda^3 + 20\lambda^4 \quad > \quad 20$$

and

$$10 + 10\lambda + 10\lambda^2 + 10\lambda^3 + 20\lambda^4 \quad > \quad 10 + 10\lambda + 10\lambda^2 + 20\lambda^3$$

We can see that for any $0.5 < \lambda \leq 1$, for example, $\lambda = 0.75$, π_2 is strictly better than the other two.

2 Policy iteration

- Since staying in state 1 and 2 will have negative reward, the agent should try to get into state 3 through action b. Because b has a low success rate and state 2 has a lower reward than state 1, the agent should try to get into state 1 first through action a if it is currently in state 2, then keep applying action b to try to get into state 3.

- $\pi \leftarrow \{b, b\}$
  Step $1^{(1)}$(evaluation): $\{V(1) = 0.1V(3) + 0.9(-1 + V(1)), V(2) = 0.1V(3) + 0.9(-2 + V(2)), V(3) = 0\} \rightarrow \{V(1) = -9, V(2) = -18, V(3) = 0\}$
  Step $2.1^{(1)}$(update): $\{Q(1, a) = 0.8(-2 - 18) + 0.2(-1 - 9) = -18, Q(1, b) = 0 + 0.9(-1 - 9) = -9\} \rightarrow \pi(1)$ unchanged
  Step $2.2^{(1)}$(update): $\{Q(2, a) = 0.8(-1 - 9) + 0.2(-2 - 18) = -12, Q(2, b) = 0 + 0.9(-2 - 18) = -18\} \rightarrow \pi(2)$ change to a
  $\pi \leftarrow \{b, a\}$
  Step $1^{(2)}$(evaluation): $\{V(1) = 0.1V(3) + 0.9(-1 + V(1)), V(2) = 0.8(-1 + V(1)) + 0.2(-2 + V(2)), V(3) = 0\} \rightarrow \{V(1) = -9, V(2) = -10.5, V(3) = 0\}$ Step $2.1^{(2)}$(update): $\{Q(1, a) = 0.8(-2 - 10.5) + 0.2(-1 - 9) = -12, Q(1, b) = 0 + 0.9(-1 - 9) = -9\} \rightarrow \pi(1)$ unchanged
  Step $2.2^{(2)}$(update): $\{Q(2, a) = 0.8(-1 - 9) + 0.2(-2 - 10.5) = -10.5, Q(2, b) = 0 + 0.9(-2 - 10.5) = -11.25\} \rightarrow \pi(2)$ unchanged
  So the resulting policy is b for state 1 and a for state 2.

- $\pi \leftarrow \{a, a\}$
  Step $1^{(1)}$(evaluation): $\{V(1) = 0.8(-2 + V(2)) + 0.2(-1 + V(1)), V(2) = 0.8(-1 + V(1)) + 0.2(-2 + V(2)), V(3) = 0\}$
  This cannot be solved because the two equations are inconsistent unless we set both $V(1)$ and $V(2)$ as infinity.

- Yes, discount factor $\lambda < 1$ will allow policy iteration to work with initial policy as a.
  $\lambda = 0.9$:
  $\pi \leftarrow \{a, a\}$
  Step $1^{(1)}$(evaluation): $\{V(1) = 0.8(-2 + 0.9V(2)) + 0.2(-1 + 0.9V(1)), V(2) = 0.8(-1 + 0.9V(1)) + 0.2(-2 + 0.9V(2)), V(3) = 0\} \rightarrow \{V(1) = -\frac{1170}{77}, V(2) = -\frac{1140}{77}\}$
  Step $2.1^{(2)}$(update): $\{Q(1, a) = 0.8(-2 - \frac{1140}{77}) + 0.2(-1 - \frac{1170}{77}) = -\frac{6423}{385}, Q(1, b) = 0 + 0.9(-1 - \frac{1170}{77}) = -\frac{11223}{770}\} \rightarrow \pi(1)$ change to b
  Step $2.2^{(2)}$(update): $\{Q(2, a) = 0.8(-1 - \frac{1170}{77}) + 0.2(-2 - \frac{1140}{77}) = -\frac{6282}{385}, Q(2, b) = 0 + 0.9(-2 - \frac{1140}{77}) = -\frac{5823}{385}\} \rightarrow \pi(2)$ change to b

The policy for $\lambda = 0.9$ is $\{b, b\}$.

$\lambda = 0.1$:

$\pi \leftarrow \{a, a\}$

Step $1^{(1)}$(evaluation): $\{V(1) = 0.8(-2 + 0.1V(2)) + 0.2(-1 + 0.1V(1)), V(2) = 0.8(-1 + 0.1V(1)) + 0.2(-2 + 0.1V(2)), V(3) = 0\} \rightarrow \{V(1) = -\frac{310}{159}, V(2) = -\frac{220}{159}\}$

Step $2.1^{(2)}$(update): $\{Q(1, a) = 0.8(-2 - \frac{220}{159}) + 0.2(-1 - \frac{310}{159}) = -\frac{2621}{795}, Q(1, b) = 0 + 0.9(-1 - \frac{310}{159}) = -\frac{1407}{530}\} \rightarrow \pi(1)$ change to b

Step $2.2^{(2)}$(update): $\{Q(2, a) = 0.8(-1 - \frac{310}{159}) + 0.2(-2 - \frac{220}{159}) = -\frac{2414}{795}, Q(2, b) = 0 + 0.9(-2 - \frac{220}{159}) = -\frac{807}{265}\} \rightarrow \pi(2)$ unchanged

The policy for $\lambda = 0.1$ is $\{b, a\}$.