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	30	30	30	30
V^{π_1}	50	50	40	30	30
V^{π_2}	60	60	50	40	30
V^*	60	60	50	40	30

- This table will be used for the following three sub-problems:

	1	2	3	4	5
V^{π_0}	20λ	$10 + 20\lambda$	$10 + 20\lambda$	$10 + 20\lambda$	$10 + 20\lambda$
V^{π_1}	$10\lambda + 10\lambda^2 + 10\lambda^3 + 20\lambda^4$	$10 + 10\lambda + 10\lambda^2 + 20\lambda^3$	$10 + 10\lambda + 20\lambda^2$	$10 + 20\lambda$	$10 + 20\lambda$
V^{π_2}	$10\lambda + 10\lambda^2 + 10\lambda^3 + 10\lambda^4 + 20\lambda^5$	$10 + 10\lambda + 10\lambda^2 + 10\lambda^3 + 20\lambda^4$	$10 + 10\lambda + 10\lambda^2 + 20\lambda^3$	$10 + 10\lambda + 20\lambda^2$	$10 + 20\lambda$

If such λ exists, it must enable every cell in row one to be greater than or equal to the cells in row two and three of the same column, at least one column of row one is better than row two, and at least one column of row one is better than row three.

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:

	1	2	3	4	5
V^{π_0}	5	15	15	15	15
V^{π_1}	3.36	13.44	13.75	15	15
V^{π_2}	3.34	13.36	13.44	13.75	15

- If such λ exists, it must satisfy:

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

and

$$10\lambda + 10\lambda^2 + 10\lambda^3 + 20\lambda^4 \quad > \quad 10\lambda + 10\lambda^2 + 10\lambda^3 + 10\lambda^4 + 20\lambda^5$$

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 enable every cell in row three to be greater than or equal to the cells in row one and two of the same column, at least one column of row three is better than row one, and at least one column of row three is better than row two.
  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:

	1	2	3	4	5
V^{π_0}	15	25	25	25	25
V^{π_1}	23.67	31.56	28.75	25	25
V^{π_2}	25.25	33.67	31.56	28.75	25

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 $\{b, a\}$.

3 MDPs and peeking blackjack

```
========== START GRADING
—— START PART writeupValid
—— END PART writeupValid [took 0:00:00.024574, 0/0 points]
—— START PART 3.1.1-0
—— END PART 3.1.1-0 [took 0:00:00.000947, 2/2 points]
—— START PART 3.1.1-1
—— END PART 3.1.1-1 [took 0:00:00.000711, 3/3 points]
—— START PART 3.1.2
—— END PART 3.1.2 [took 0:00:00.003898, 15/15 points]
—— START PART 3.1.3
—— END PART 3.1.3 [took 0:00:00.000384, 5/5 points]
—— START PART 3.1.4
—— END PART 3.1.4 [took 0:00:00.005044, 10/10 points]
—— START PART 3.1.5
—— END PART 3.1.5 [took 0:00:00.012610, 10/10 points]
—— START PART 3.2.1
—— END PART 3.2.1 [took 0:00:00.000276, 15/15 points]
—— START PART 3.2.2
—— END PART 3.2.2 [took 0:00:00.004059, 5/5 points]
```

========== END GRADING [65/65 points]

Total max points: 65

Counter example in submission.py

4 Robot couriers and Markov decision problems

- discount factor = 1
  $V_0 = \{1 : 0, 2 : 0, 3 : 0\}$
  $V_1 = \{1 : 8.0, 2 : 16.0, 3 : 7.0\}$
  $V_2 = \{1 : 17.75, 2 : 29.94, 3 : 17.88\}$
  $V_3 = \{1 : 29.66, 2 : 43.42, 3 : 30.91\}$
  $V_4 = \{1 : 42.97, 2 : 56.78, 3 : 44.14\}$
  Optimal policy: $\{1 : r, 2 : r, 3 : r\}$

- discount factor = 0.75
  $V_0 = \{1 : 0, 2 : 0, 3 : 0\}$
  $V_1 = \{1 : 8.0, 2 : 16.0, 3 : 7.0\}$
  $V_2 = \{1 : 15.31, 2 : 26.20, 3 : 14.41\}$
  $V_3 = \{1 : 21.36, 2 : 33.59, 3 : 21.53\}$
  Optimal policy: $\{1 : c, 2 : r, 3 : r\}$

- discount factor = 0.5
  $V_0 = \{1 : 0, 2 : 0, 3 : 0\}$
  $V_1 = \{1 : 8.0, 2 : 16.0, 3 : 7.0\}$
  $V_2 = \{1 : 12.88, 2 : 22.47, 3 : 11.75\}$
  $V_3 = \{1 : 15.50, 2 : 25.60, 3 : 14.36\}$
  Optimal policy: $\{1 : c, 2 : r, 3 : c\}$

- discount factor = 0.1
  $V_0 = \{1 : 0, 2 : 0, 3 : 0\}$
  $V_1 = \{1 : 8.0, 2 : 16.0, 3 : 7.0\}$
  $V_2 = \{1 : 8.98, 2 : 16.75, 3 : 7.95\}$
  Optimal policy: $\{1 : c, 2 : c, 3 : c\}$