CS 6300 **HW5: Policy Iteration and TD Learning** **Due February 28, 2017**

Please use the LaTeX template to produce your writeups. See the Homework Assignments page on the class website for details. Hand in at: `https://webhandin.eng.utah.edu/index.php`.

1 Policy Iteration (30pts)

The card game of high-low is played with an infinite deck whose only cards are 2, 3, and 4 in equal proportion. You start with one of the cards showing, and say either *high* or *low*. Then a new card is flipped, and you compare the value of the new card to that of the old card.

- If you are right, you get the value of the new card.
- If the new card has the same value, you don't get any points.
- If you are wrong, the game is done.

If you are not done, the new card then becomes the reference card for drawing the next new card. You accumulate points as above until you are wrong and the game ends.

1. Formulate high-low as an MDP, by listing the states, actions, transition rewards, and transition probabilities.

s	a	s'	$T(s, a, s')$
2	*high*	*win*	2/3
2	*high*	*equal*	1/3
2	*low*	*lose*	2/3
2	*low*	*equal*	1/3
3	*high*	*win*	1/3
3	*high*	*equal*	1/3
3	*high*	*lose*	1/3
3	*low*	*win*	1/3
3	*low*	*equal*	1/3
3	*low*	*lose*	1/3
4	*high*	*lose*	2/3
4	*high*	*equal*	1/3
4	*low*	*win*	2/3
4	*low*	*equal*	1/3

Transition Model

s'	$R(s')$
win	$\{2, 3, 4\}$
equal	0
lose	terminate

Rewards

2. You will be doing one iteration of policy iteration. Assume the initial policy $\pi_0(s) = high$.

(a) Perform policy evaluation to solve for the utility values $V^{\pi_0}(s)$ for the appropriate states s. Please solve these equations analytically.

$$V^{\pi_0}(s) = \sum_{s'} T(s, \pi_0, s') \left[R(s, \pi_0(s), s') + \gamma V^{\pi_0}(s') \right]$$

We can also calculate the expected value of drawing a new card by taking the average of the values since they're all equally likely. We assume $\gamma = 1$

$$V_{k+1}^{\pi_0}(2) = \frac{1}{3}\left(0 + V_k^{\pi_0}(2)\right) + \frac{1}{3}\left(3 + V_k^{\pi_0}(3)\right) + \frac{1}{3}\left(4 + V_k^{\pi_0}(4)\right)$$

But we don't know the values for $V_k^{\pi_0}(3)$ or $V_k^{\pi_0}(4)$, so we can solve those first and then plug them back into the above equation. We get

$$\begin{aligned} V_k^{\pi_0}(3) &= \frac{1}{3}(\text{terminate}) + \frac{1}{3}\left(0 + V_k^{\pi_0}(3)\right) + \frac{1}{3}\left(4 + V_k^{\pi_0}(4)\right) \\ V_k^{\pi_0}(4) &= \frac{2}{3}(\text{terminate}) + \frac{1}{3}\left(0 + V_k^{\pi_0}(4)\right) \end{aligned}$$

Where we can ignore the termination states as we want to *maximize* our output. We can solve for $V_k^{\pi_0}(4)$ and then $V_k^{\pi_0}(3)$ to give

$$\begin{aligned} V_k^{\pi_0}(4) &= 0 \\ V_k^{\pi_0}(3) &= \frac{1}{3}\left(0 + V_k^{\pi_0}(3)\right) + \frac{1}{3}\left(4 + V_k^{\pi_0}(4)\right) \\ &= \frac{1}{3}V_k^{\pi_0}(3) + \frac{4}{3} \\ &= 2 \end{aligned}$$

Finally, by solving for $V_k^{\pi_0}(2)$

$$\begin{aligned} V_{k+1}^{\pi_0}(2) &= \frac{1}{3}\left(0 + V_k^{\pi_0}(2)\right) + \frac{1}{3}\left(3 + V_k^{\pi_0}(3)\right) + \frac{1}{3}\left(4 + V_k^{\pi_0}(4)\right) \\ &= \frac{9}{2} \end{aligned}$$

(b) Perform policy improvement to find the next policy $\pi_1(s)$.

The equation for calculating the next policy is

$$\pi_{k+1}(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma V_k^{\pi_k}(s') \right]$$

We can use this to find the next policy update for each state, giving

$$\pi_1(2) = \underset{a}{\text{argmax}} \left\{ \begin{array}{ccccc} \frac{1}{3}\left(0+\frac{9}{2}\right) & = & 3/2 & (low) \\ \frac{1}{3}\left(0+\frac{9}{2}\right)+\frac{1}{3}\left(3+2\right)+\frac{1}{3}\left(4+0\right) & = & 27/6 & (high) \end{array} \right.$$

$$\pi_1(2) = high$$

$$\pi_1(3) = \underset{a}{\text{argmax}} \left\{ \begin{array}{ccccc} \frac{1}{3}\left(0+2\right)+\frac{1}{3}\left(2+\frac{9}{2}\right) & = & 17/6 & (low) \\ \frac{1}{3}\left(0+2\right)+\frac{1}{3}\left(4+0\right) & = & 2 & (high) \end{array} \right.$$

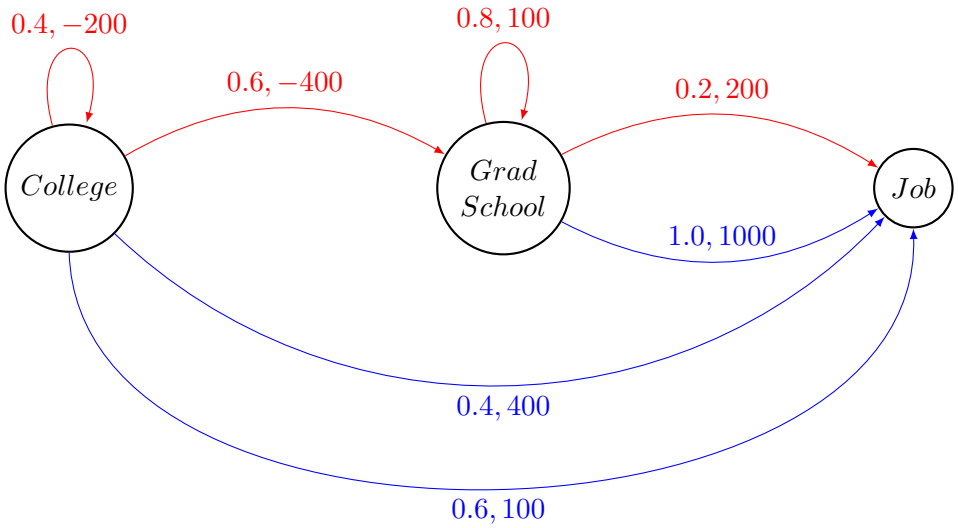
$$\pi_1(3) = low$$

$$\pi_1(4) = \underset{a}{\text{argmax}} \left\{ \begin{array}{ccccc} \frac{1}{3}\left(2+\frac{9}{2}\right)+\frac{1}{3}\left(3+2\right)+\frac{1}{3}(0+0) & = & 23/6 & (low) \\ \frac{1}{3}(0+0) & = & 0 & (high) \end{array} \right.$$

$$\pi_1(4) = high$$

2 Life as a Student

Jennifer is currently finishing up in college and is trying to decide what she wants to do with the rest of her life. She sketches her options as a known MDP:



At each point in her life she can choose to either continue in school (action x) or try to get a job (action y). Her three states are **C**ollege, **G**rad School and **J**ob. **J** is a terminal state. Transition probabilities t and immediate rewards r are shown on the arcs as t, r. The discount $\gamma = 0.5$.

1. Using value iteration, compute values for each state. Add or remove extra lines for iteration as needed.

i	$V_i^*(C)$	$V_i^*(G)$
0	0.0	0.0
1	220.0	1000.0
2	220.0	1000.0

The non-linear equation for *Value Iteration* is defined as

$$V_{k+1}^*(s) = \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k^*(s')\right]$$

We can calculate the value of V for two iterations of $College$.

$$V_1^*(C) = \max_a \begin{cases} 0.4(-200+0) + 0.6(-400+0) & = -320 \quad (x) \\ 0.4(400+0) + 0.6(100+0) & = \quad 220 \quad (y) \end{cases}$$

$$V_2^*(C) = \max_a \begin{cases} 0.4(-200+0.5(220)) + 0.6(-400+0.5(220)) & = -210 \quad (x) \\ 0.4(400+0.5(0)) + 0.6(100+0.5(0)) & = \quad 220 \quad (y) \end{cases}$$

2. Suppose Jennifer didn't actually know the values of the MDP. Instead, she watched three of her older siblings go through life. They exhibited the following trajectories:

Sibling 1	**Sibling 2**	**Sibling 3**
$C, x, -200$	$C, x, -400$	$C, x, -400$
$C, x, -200$	$G, x, \quad 100$	$G, x, \quad 100$
$C, y, \quad 400$	$G, x, \quad 100$	$G, x, \quad 100$
J	$G, y, \quad 1000$	$G, x, \quad 200$
	J	J

(a) According to direct estimation, what are the transition probabilities $T(C, x, C)$, $T(C, x, G)$ and $T(C, y, J)$?

$T(C, x, C)$	$T(C, x, G)$	$T(C, y, J)$
$4/9$	$5/9$	$1/1$

(b) According to direct estimation, what are the rewards $R(C, x, C)$, $R(C, x, G)$ and $R(C, y, J)$?

$R(C, x, C)$	$R(C, x, G)$	$R(C, y, J)$
-300	120	400

(c) Use TD Learning instead to find estimates of the value function $V^{\pi}(s)$, assuming $\alpha = 0.5$.

The equation for Temporal Difference Learning is

$$V^{\pi}(s) \leftarrow (1 - \alpha)V^{\pi}(s) + \alpha \left[R(s, \pi, s') + \gamma V^{\pi}(s') - V^{\pi}(s)\right]$$