CS 6300 **HW05: TD and Q Learning** **Ryan Dalby** **Due March 3, 2022**

Please use the LaTeX template to produce your writeups. See the Homework Assignments page on the class website for details. Hand in via gradescope.

Jennifer is currently finishing up in college and is trying to decide what she wants to do with the rest of her life. She assumes her options can be modeled as an MDP, with a discount $\gamma = 1/2$. 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.

Suppose Jennifer doesn't actually know the MDP model. Instead, she watched three of her older siblings go through life. They exhibited the following episodes.

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

1. What are the transition probabilities and rewards that you can know about?

Transition Probabilities	**Rewards**
T(C, x, C)	R(C, x, C)
T(C, x, G)	R(C, x, G)
T(G, x, G)	R(G, x, G)
T(G, x, J)	R(G, x, J)
T(C, y, J)	R(C, y, J)
T(G, y, J)	R(G, y, J)

2. Find the values $V(s)$ using direct estimation.

$$V(C) = \frac{(-200 + -100 + 100) + (-200 + 200) + (-400 + 50 + 25 + 125) + (-400 + 50 + 25 + 25)}{5}$$

$$V(C) = -60$$

$$V(G) = \frac{(100 + 50 + 250) + (100 + 500) + (1000) + (100 + 50 + 50) + (100 + 100) + (200)}{6}$$

$$V(G) = \frac{1300}{3} \approx 433.33$$

3. Use TD Learning instead to find estimates of the values, assuming $\alpha = 1/2^{n-1}$, where n is the sibling number.

   Using updates as follows:

$$V^{\pi}(s) = (1-\alpha)V^{\pi}(s) + \alpha(R(s,a,s') + \gamma V^{\pi}(s'))$$

Episode	Trial	Update
1	1	$V^{\pi}(C) = 0 + 1(-200 + \frac{1}{2}(0)) = -200$
	2	$V^{\pi}(C) = 0 + 1(-200 + \frac{1}{2}(-200)) = -300$
	3	$V^{\pi}(C) = 0 + 1(400 + \frac{1}{2}(0)) = 400$
2	1	$V^{\pi}(C) = \frac{1}{2}(400) + \frac{1}{2}(-400 + \frac{1}{2}(0)) = 0$
	2	$V^{\pi}(G) = \frac{1}{2}(0) + \frac{1}{2}(100 + \frac{1}{2}(0)) = 50$
	3	$V^{\pi}(G) = \frac{1}{2}(50) + \frac{1}{2}(100 + \frac{1}{2}(50)) = \frac{175}{2} = 87.5$
	4	$V^{\pi}(G) = \frac{1}{2}(\frac{175}{2}) + \frac{1}{2}(1000 + \frac{1}{2}(0)) = \frac{2175}{4} = 543.75$
3	1	$V^{\pi}(C) = \frac{3}{4}(0) + \frac{1}{4}(-400 + \frac{1}{2}(\frac{2175}{4})) = -\frac{1025}{32} \approx -32.03$
	2	$V^{\pi}(G) = \frac{3}{4}(\frac{2175}{4}) + \frac{1}{4}(100 + \frac{1}{2}(\frac{2175}{4})) = \frac{16025}{32} \approx 32.03$
	3	$V^{\pi}(G) = \frac{3}{4}(\frac{16025}{32}) + \frac{1}{4}(100 + \frac{1}{2}(\frac{16025}{32})) = \frac{118575}{256} \approx 500.78$
	4	$V^{\pi}(G) = \frac{3}{4}(\frac{118575}{256}) + \frac{1}{4}(200 + \frac{1}{2}(0)) = \frac{118575}{256} \approx 463.18$

4. Use Q learning instead, and extract the estimated optimal policy.

   Using updates as follows:

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha(R(s,a,s') + \gamma \max_{a'} Q(s',a'))$$

Episode	Trial	Update
1	1	$Q(C,x) = 0 + 1(-200 + \frac{1}{2}(\max\{0,0\})) = -200$
	2	$Q(C,x) = 0 + 1(-200 + \frac{1}{2}(\max\{-200,0\})) = -200$
	3	$Q(C,y) = 0 + 1(400 + \frac{1}{2}(0)) = 400$
2	1	$Q(C,x) = \frac{1}{2}(-200) + \frac{1}{2}(-400 + \frac{1}{2}(\max\{0,0\})) = -300$
	2	$Q(G,x) = \frac{1}{2}(0) + \frac{1}{2}(100 + \frac{1}{2}(\max\{0,0\})) = 50$
	3	$Q(G,x) = \frac{1}{2}(50) + \frac{1}{2}(100 + \frac{1}{2}(\max\{50,0\})) = \frac{175}{2} = 87.5$
	4	$Q(G,y) = \frac{1}{2}(0) + \frac{1}{2}(1000 + \frac{1}{2}(0)) = 500$
3	1	$Q(C,x) = \frac{3}{4}(-300) + \frac{1}{4}(-400 + \frac{1}{2}(\max\{\frac{175}{2},500\})) = -\frac{525}{2} = -262.5$
	2	$Q(G,x) = \frac{3}{4}(\frac{175}{2}) + \frac{1}{4}(100 + \frac{1}{2}(\max\{\frac{175}{2},500\})) = \frac{1225}{8} \approx 153.12$
	3	$Q(G,x) = \frac{3}{4}(\frac{1225}{8}) + \frac{1}{4}(100 + \frac{1}{2}(\max\{\frac{1225}{8},500\})) = \frac{6475}{32} \approx 202.34$
	4	$Q(G,x) = \frac{3}{4}(\frac{6475}{32}) + \frac{1}{4}(200 + \frac{1}{2}(0)) = \frac{25825}{128} \approx 201.76$

The optimal policy is:

$$\pi(C) = y$$
$$\pi(G) = y$$