CS 6300 **HW05: TD and Q Learning** **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?

   T(C,x,C) = 1/2 R(C,x,C) = -200
   T(C,x,G) = 1/2 R(C,x,G) = -400
   T(C,y,J) = 1 R(C,y,J) = 400
   T(G,x,G) = 4/5 R(G,x,G) = 100
   T(G,x,J) = 1/5 R(G,x,J) = 200
   T(G,y,J) = 1 R(G,y,J) = 1000

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

   Of course $V(J) = 0$. The sum of discounted rewards from each starting point for $V(C)$ is:

$$\begin{aligned} V(C) &= ((-200 - 200/2 + 400/4) + (-200 + 400/2) + (400) + \\ &\quad (-400 + 100/2 + 100/4 + 1000/8) + (-400 + 100/2 + 100/4 + 200/8))/5 \\ &= (-200 + 0 + 400 - 200 - 300)/5 \\ &= -60 \end{aligned}$$

$$\begin{aligned} V(G) &= ((100 + 100/2 + 1000/4) + (100 + 1000/2) + (1000) + \\ &\quad (100 + 100/2 + 200/4) + (100 + 200/2) + (200))/6 \\ &= (400 + 600 + 1000 + 200 + 200 + 200)/6 \\ &= 2600/6 \end{aligned}$$

The following is computed assuming $\gamma = 1$, which will also be considered correct because we didn't clearly cover this case in class.

$$
\begin{aligned}
V(C) &= (0 + 200 + 400 + 800 + 0)/5 = 1400/5 = 280 \\
V(G) &= (1200 + 1100 + 1000 + 400 + 300 + 200)/6 = 4200/6 = 700
\end{aligned}
$$

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

   All values are initialized to 0. The update equation to use, particularized for this problem, is:

$$\begin{aligned} V(s) &\leftarrow (1-\alpha)V(s) + \alpha(R(s,a,s') + \gamma V(s')) \\ &\leftarrow (1 - \frac{1}{2^{n-1}})V(s) + \frac{1}{2^{n-1}}(r_i + \frac{1}{2}V(s')) \end{aligned}$$

   The values below are rounded to two decimal places.

n	i	s, a, r_i	s'	$\frac{1}{2^{n-1}}(r_i + V(s')/2)$	$(1 - \frac{1}{2^{n-1}})V(s)$	$V(C)$	$V(G)$
1	1	$C, x, -200$	C	$-200 = 1 * (-200 + 0/2)$	0	-200	0
	2	$C, x, -200$	C	$-300 = 1 * (-200 - 200/2)$	0	-300	
	3	$C, y, 400$	J	$400 = 1 * (400 + 0/2)$	0	400	
2	1	$C, x, -400$	G	$-200 = (-400 + 0/2)/2$	$200 = 400/2$	0	
	2	$G, x, 100$	G	$50 = (100 + 0/2)/2$	$0 = 0/2$		50
	3	$G, x, 100$	G	$62.5 = (100 + 50/2)/2$	$25 = 50/2$		87.5
	4	$G, y, 1000$	J	$500 = (1000 + 0/2)/2$	$43.75 = 87.5/2$		543.75
3	1	$C, x, -400$	G	$-32.03 = (-400 + 543.75/2)/4$	$0 = (3/4) * 0$	-32.03	
	2	$G, x, 100$	G	$92.97 = (100 + 543.75/2)/4$	$407.81 = (3/4) * 543.75$		500.78
	3	$G, x, 100$	G	$87.60 = (100 + 500.78/2)/4$	$375.59 = (3/4) * 500.78$		463.19
	4	$G, x, 200$	J	$50 = (200 + 0/2)/4$	$347.40 = (3/4) * 463.19$		397.40

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

   All values are initialized to 0. The update equation to use, particularized for this problem, is:

$$\begin{aligned} Q(s,a) &\leftarrow (1-\alpha)Q(s,a) + \alpha(R(s,a,s') + \gamma \max_{a'} Q(s',a')) \\ &\leftarrow (1-\frac{1}{2^{n-1}})Q(s,a) + \frac{1}{2^{n-1}}(r_i + \frac{1}{2}\max_{a'} Q(s',a')) \end{aligned}$$

n	i	s,a,r_i	s'	$\frac{1}{2^{n-1}}(r_i + \max_{a'} Q(s',a')/2)$	$(1-\frac{1}{2^{n-1}})Q(s,a)$	$Q(s,a)$
1	1	$C,x,-200$	C	$-200 = 1*(-200+0/2)$	0	$Q(C,x) = -200$
	2	$C,x,-200$	C	$-200 = 1*(-200+0/2)$	0	$Q(C,x) = -200$
	3	$C,y,400$	J	$400 = 1*(400+0/2)$	0	$Q(C,y) = 400$
2	1	$C,x,-400$	G	$-200 = (-400+0/2)/2$	$-100 = -200/2$	$Q(C,x) = -300$
	2	$G,x,100$	G	$50 = (100+0/2)/2$	$0 = 0/2$	$Q(G,x) = 50$
	3	$G,x,100$	G	$62.5 = (100+50/2)/2$	$25 = 50/2$	$Q(G,x) = 87.5$
	4	$G,y,1000$	J	$500 = (1000+0/2)/2$	$0 = 0/2$	$Q(G,y) = 500$
3	1	$C,x,-400$	G	$-37.5 = (-400+500/2)/4$	$-225 = (3/4)*-300$	$Q(C,x) = -262.5$
	2	$G,x,100$	G	$87.5 = (100+500/2)/4$	$65.625 = (3/4)*87.5$	$Q(G,x) = 153.125$
	3	$G,x,100$	G	$87.5 = (100+500/2)/4$	$114.84375 = (3/4)*153.125$	$Q(G,x) = 202.34375$
	4	$G,x,200$	J	$50 = (200+0/2)/4$	$151.7578125 = (3/4)*202.34375$	$Q(G,x) = 201.7578125$