CMPUT 366, Winter 2021
Assignment #4

Due: Monday, April 19, 2021, 11:59pm
Total points: 55

For this assignment use the following consultation model:

1. you can discuss assignment questions and exchange ideas with other *current* CMPUT 366 students;
2. you must list all members of the discussion in your solution;
3. you may **not** share/exchange/discuss written material and/or code;
4. you must write up your solutions individually;
5. you must fully understand and be able to explain your solution in any amount of detail as requested by the instructor and/or the TAs.

Anything that you use in your work and that is not your own creation must be properly cited by listing the original source. Failing to cite others' work is plagiarism and will be dealt with as an academic offence.

First name: Vicky

Last name: Zhao

CCID: ziwei11 @ualberta.ca

Collaborators: _____

1. **(Markov Decision Processes)**

   Two coins are placed in a tray; each coin is either heads up or tails up with equal probability, independent of the other coin. A robot arm is above **one** of the coins. At each time step, the arm can perform one of the following operations:

   - `Flip` the coin that it is above: The coin under the arm will be randomly set to either heads up or tails up with equal probability. This operation costs 1 unit of battery power.
   - `Move` to be above the other coin; this operation costs 2 units of battery power.
   - `Call` the Electricity Fairy, who will reward the robot with 20 units of battery power if both coins are heads up, reward the robot with 10 units of battery power if both tails are up, or fine the robot 5 units of electricity if the coins mismatch. After rewarding or fining the robot, the Electricity Fairy will then flip both coins before leaving.

   (a) **[10 points]** Represent this scenario formally as a Markov Decision Process, treating changes in battery power as the reward signal.

   Note: You do not need to include the total battery charge as part of the state. Changes in battery charge are reward only. (I.e., the battery never runs out; its charge can become negative and the robot can keep operating.)

See next page.

   (b) **[2 points]** Is this a continuing or episodic scenario? Justify your answer.

This is a continuing scenario, because the battery never runs out.

   (c) **[3 points]** If you answered episodic to the previous question, describe how to treat the scenario as continuing. If you answered continuing to the previous question, describe how to treat the scenario as episodic.

We can set the battery to be a battery which may run out of power. In this situation, the scenario is episodic because it can end after some finite number T of time steps in a special terminal state S_T.

2. **(Bellman Action-Value Equation)**

   (a) **[4 points]** Give an expression for $q_\pi(s, a)$ as an expectation of random variables, in the same form as equation (4.3) in the text.

$$q_\pi(s,a) = E_\pi\left[R_{t+1} + \gamma \sum_{A_{t+1}} \pi(A_{t+1} \mid S_{t+1}) q_\pi(S_{t+1}, A_{t+1}) \,\middle|\, S_t = s,\ A_t = a \right]$$

1.(a)

robot arm location ↓ left coin head(up) ↓ right coin head (down) ↓

states: {(left, up, down), (left, up, up), (left, down, down), (left, down, up)
(right, up, down), (right, up, up), (right, down, down), (right, down, up)}

actions: {Flip, Move, Call}

rewards: {-1, -2, 20, 10, -5}

S	a	S'	P(S'|S, a)	r(S, a, S')
(left, up, down)	Flip	(left, down, down)	0.5	-1
(left, up, up)	Flip	(left, down, up)	0.5	-1
(left, down, up)	Flip	(left, up, up)	0.5	-1
(left, down, down)	Flip	(left, up, down)	0.5	-1
(right, up, down)	Flip	(right, up, up)	0.5	-1
(right, up, up)	Flip	(right, up, down)	0.5	-1
(right, down, up)	Flip	(right, down, down)	0.5	-1
(right, down, down)	Flip	(right, down, up)	0.5	-1
(left, up, down)	Flip	(left, up, down)	0.5	-1
(left, up, up)	Flip	(left, up, up)	0.5	-1
(left, down, up)	Flip	(left, down, up)	0.5	-1
(left, down, down)	Flip	(left, down, down)	0.5	-1
(right, up, down)	Flip	(right, up, down)	0.5	-1
(right, up, up)	Flip	(right, up, up)	0.5	-1
(right, down, up)	Flip	(right, down, up)	0.5	-1
(right, down, down)	Flip	(right, down, down)	0.5	-1

(left, up, up)	Move	(right, up, up)	1	-2
(right, up, up)	Move	(left, up, up)	1	-2
(left, up, down)	Move	(right, up, down)	1	-2
(right, up, down)	Move	(left, up, down)	1	-2
(left, down, up)	Move	(right, down, up)	1	-2
(right, down, up)	Move	(left, down, up)	1	-2
(left, down, down)	Move	(right, down, down)	1	-2
(right, down, down)	Move	(left, down, down)	1	-2
(left, up, up)	Call	(left, down, down)	1	+20
(right, up, up)	Call	(right, down, down)	1	+20
(left, down, down)	Call	(left, up, up)	1	+10
(right, down, down)	Call	(right, up, up)	1	+10
(left, up, down)	Call	(left, down, up)	1	-5
(left, down, up)	Call	(left, up, down)	1	-5
(right, up, down)	Call	(right, down, up)	1	-5
(right, down, up)	Call	(right, up, down)	1	-5

(b) **[4 points]** Give an expression for $q_\pi(s,a)$ as a weighted sum, in the same form as equation (4.4) in the text.

$$q_\pi(s,a) = \sum_{s'}\sum_{r} p(s',r|s,a)\left[r + \gamma \sum_{a'} \pi(a'|s')\, q_\pi(s',a')\right]$$

3. **(Policy Improvement)** Consider the following MDP with actions $\mathcal{A} = \{a, b\}$, states $\mathcal{S} = \{W, X, Y, Z\}$, and the following dynamics. All unspecified transitions have probability 0:

$$p(X,4|W,a) = 0.5 \qquad p(Z,0|X,a) = 0.8 \qquad p(Z,1|Y,a) = 1$$
$$p(Y,2|W,a) = 0.5 \qquad p(Z,25|X,a) = 0.2 \qquad p(Z,5|Y,b) = 1$$
$$p(Z,3|W,b) = 1 \qquad p(Z,0|X,b) = 1 \qquad p(Z,0|Z,a) = 1$$
$$p(Z,0|Z,b) = 1$$

(a) **[4 points]** Consider the policy $\pi(b|s) = .5, \pi(a|s) = .5$ for all states $s \in \mathcal{S}$; i.e., a policy that randomizes between the two actions uniformly. Using a discount rate of $\gamma = 0.8$, what is the value $v_\pi(s)$ for each state $s \in \mathcal{S}$? (Hint: What must the value of state Z be under any policy?)

$v_\pi(W) = 0.5\times0.5\times(4+0.8\times v_\pi(X)) + 0.5\times0.5\times(2+0.8\times v_\pi(Y)) + 0.5\times1\times(3+0.8\times v_\pi(Z))$

$v_\pi(X) = 0.5\times0.8\times(0+0.8\times v_\pi(Z)) + 0.5\times0.2\times(25+0.8\times v_\pi(Z)) + 0.5\times1\times(0+0.8\times v_\pi(Z))$

$v_\pi(Y) = 0.5\times1\times(1+0.8\times v_\pi(Z)) + 0.5\times1\times(5+0.8\times v_\pi(Z))$

$v_\pi(Z) = 0.5\times1\times(0+0.8\times v_\pi(Z)) + 0.5\times1\times(0+0.8\times v_\pi(Z))$

$$\Rightarrow \begin{cases} v_\pi(W) = 4.1 \\ v_\pi(X) = 2.5 \\ v_\pi(Y) = 3 \\ v_\pi(Z) = 0 \end{cases}$$

(b) **[8 points]** Construct a policy π' that strictly improves upon π: That is, $v_{\pi'}(s) \geq v_\pi(s)$ for all $s \in \mathcal{S}$, and $v_{\pi'}(s) > v_\pi(s)$ for at least one $s \in \mathcal{S}$. Appeal to the Policy Improvement Theorem to show that your new policy is indeed a strict improvement.

We set $\pi(a|s)=1$ for action a which has higher reward r than others,

so suppose policy π': $\pi'(a|W)=1$, $\pi'(a|X)=1$, $\pi'(b|Y)=1$, $\pi'(a|Z)=1$

According to Policy Improvement Theorem, if $q_\pi(s,\pi'(s)) \geq v_\pi(s)\ \forall s\in\mathcal{S}$, then $v_{\pi'}(s) \geq v_\pi(s)\ \forall s\in\mathcal{S}$

$q_\pi(W,\pi'(W)) = v_\pi(W) = 1\times0.5\times[4+0.8\times v_\pi(X)] + 1\times0.5\times[2+0.8\times v_\pi(Y)]$

$q_\pi(X,\pi'(X)) = v_\pi(X) = 1\times0.8\times[0+0.8\times v_\pi(Z)] + 1\times0.2\times[25+0.8\times v_\pi(Z)]$

$q_\pi(Y,\pi'(Y)) = v_\pi(Y) = 1\times1\times[5+0.8\times v_\pi(Z)]$

$q_\pi(Z,\pi'(Z)) = v_\pi(Z) = 1\times1\times[0+0.8\times v_\pi(Z)]$

$$\Rightarrow \begin{cases} q_\pi(W,\pi'(W)) = 7 > 4.1 \\ q_\pi(X,\pi'(X)) = 5 > 2.5 \\ q_\pi(Y,\pi'(Y)) = 5 > 3 \\ q_\pi(Z,\pi'(Z)) = 0 \geq 0 \end{cases}$$

Thus, $v_{\pi'}(s) \geq v_\pi(s)$ for all $s\in\mathcal{S}$ because $q_\pi(s,\pi'(s)) \geq v_\pi(s)$ for all $s\in\mathcal{S}$ according to Policy Improvement Theorem, and $v_{\pi'}(s) > v_\pi(s)$ for at least one $s\in\mathcal{S}$, π' strictly improves upon π.

4. **(Monte Carlo Prediction)** Consider an MDP with actions $\mathcal{A} = \{a, b, c\}$, states $\mathcal{S} = \{W, X, Y, Z\}$ (with terminal state Z), and unknown dynamics. Suppose that you have used a policy π to generate 4 episodes with the following trajectories:

$\langle S_0 = W, A_0 = a, R_1 = 0, S_1 = X, A_1 = a, R_2 = 10, S_2 = Y, A_2 = b, R_3 = 0, S_3 = Z\rangle,$
$\langle S_0 = W, A_0 = a, R_1 = -10, S_1 = X, A_1 = b, R_2 = 0, S_2 = Z\rangle,$
$\langle S_0 = W, A_0 = b, R_1 = 2, S_1 = Y, A_1 = c, R_2 = 6, S_2 = Z\rangle,$
$\langle S_0 = W, A_0 = b, R_1 = 0, S_1 = Y, A_1 = c, R_2 = 12, S_2 = Z\rangle.$

(a) **[8 points]** Use first-visit Monte Carlo prediction to estimate $v_\pi(s)$ for every $s \in \mathcal{S}$. Assume undiscounted rewards (i.e., $\gamma = 1$).

$S = \{W, X, Y, Z\}$

episode	W	X	Y	Z
1	10	10	0	
2	-10	0		
3	8		6	
4	12		12	

Thus, $V_\pi(W) = \frac{10 + (-10) + 8 + 12}{4} = 5$,

$V_\pi(X) = \frac{10 + 0}{2} = 5$,

$V_\pi(Y) = \frac{0 + 6 + 12}{3} = 6$,

$V_\pi(Z) = 0$.

5. **(Temporal Difference Control)** Consider an MDP with actions $\mathcal{A} = \{a, b, c\}$, states $\mathcal{S} = \{W, X, Y\}$, and unknown dynamics.

Suppose that you have previously computed the following estimated action values:

$Q(W, a) = 0$	$Q(X, a) = 8$	$Q(Y, a) = 0$
$Q(W, b) = 0$	$Q(X, b) = 7$	$Q(Y, b) = 0$
$Q(W, c) = 0$	$Q(X, c) = 16$	$Q(Y, c) = 0$

Now suppose that starting from state $S_t = W$, the current behaviour policy selects action $A_t = a$, leading to reward $R_{t+1} = 4$ and a transition to state $S_{t+1} = X$. The behaviour policy then selects action $A_{t+1} = a$.

(a) **[6 points]** What is the updated estimate for $Q(W, a)$ according to the Q-learning algorithm? Assume a step size of $\alpha = 0.5$ and a discount rate of $\gamma = 1$.

$Q(W,a) \leftarrow Q(W,a) + \alpha[R + \gamma \max_c Q(X,c) - Q(W,a)]$

$0 + 0.5[4 + 1 \times 16 - 0] = 10$

Thus, the updated estimate for $Q(W,a)$ is 10 according to the Q-learning algorithm.

(b) **[6 points]** What is the updated estimate for $Q(W, a)$ according to the Sarsa algorithm? Assume a step size of $\alpha = 0.5$ and a discount rate of $\gamma = 1$.

$Q(W,a) \leftarrow Q(W,a) + \alpha[R + \gamma Q(X,a) - Q(W,a)]$

$0 + 0.5[4 + 1 \times 8 - 0] = 6$

Thus, the updated estimate for $Q(W,a)$ is 6 according to the Sarsa algorithm.

Submission

The assignment you downloaded from eClass is a single ZIP archive which includes this document as a PDF *and* its LaTeX source.

Each assignment is to be submitted electronically via eClass by the due date. **Your submission must be a a single PDF file containing your answers.**

To generate the PDF file with your answers you can do any of the following:

- insert your answers into the provided LaTeX source file between `\begin{answer}` and `\end{answer}`. Then run the source through LaTeX to produce a PDF file;
- print out the provided PDF file and legibly write your answers in the blank spaces under each question. Make sure you write as legibly as possible for we cannot give you any points if we cannot read your hand-writing. Then scan the pages and include the scan in your ZIP submission to be uploaded on eClass;
- use your favourite text processor and type up your answers there. Make sure you number your answers in the same way as the questions are numbered in this assignment.