Last name:_____ First name:_____ SID#:_____

Collaborators:_____

CMPUT 366/609 Assignment 2: Markov Decision Processes 1

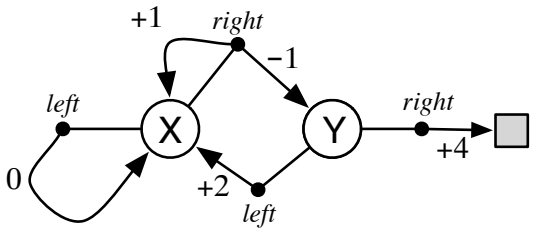
Due: Thursday Sept 28, 11:59pm by gradescope

Policy: Can be discussed in groups (acknowledge collaborators) but must be written up individually

There are a total of 100 points on this assignment, plus 15 extra credit points available.

Be sure to explicitly answer each subquestion posed in each exercise.

Question 1: Trajectories, returns, and values (**15 points total**). This question has six subparts.



Consider the MDP above, in which there are two states, X and Y, two actions, *right* and *left*, and the deterministic rewards on each transition are as indicated by the numbers. Note that if action *right* is taken in state X, then the transition may be either to X with a reward of $+1$ or to Y with a reward of -1. These two possibilities occur with probabilities $3/4$ (for the transition to X) and $1/4$ (for the transition to state Y).
Consider two deterministic policies, π_1 and π_2:

$\pi_1(\mathsf{X}) = \textit{left}$	$\pi_2(\mathsf{X}) = \textit{right}$
$\pi_1(\mathsf{Y}) = \textit{right}$	$\pi_2(\mathsf{Y}) = \textit{right}$

(a) (2 pts.) Show a typical trajectory (sequence of states, actions and rewards) from X for policy π_1:

(b) (2 pts.) Show a typical trajectory (sequence of states, actions and rewards) from X for policy π_2:

(c) (2 pts.) Assuming the discount-rate parameter is $\gamma = 0.5$, what is the return from the initial state for the second trajectory?

$G_0 =$

(d) (2 pts.) Assuming $\gamma = 0.5$, what is the value of state Y under policy π_1?

$v_{\pi_1}(\mathsf{Y}) =$

(e) (2 pts.) Assuming $\gamma = 0.5$, what is the action-value of X,*left* under policy π_1?

$q_{\pi_1}(\mathsf{X}, left) =$

(f) (5 pts) Assuming $\gamma = 0.5$, what is the value of state X under policy π_2?

$v_{\pi_2}(\mathsf{X}) =$

Question 2 [85 points total]. This question has **ten** subparts. The first 9 subparts are questions from SB textbook, second ed. The last subpart (j) is not from SB.

(a) Exercise 3.1 [6 points] (Example RL problems).

(b) Exercise 3.7 [6 points, 3 for each subquestion] (problem with maze running).

(c) Exercise 3.8 [6 points] (computing returns).

(d) Exercise 3.9 [9 points] (computing an infinite return).

(e) Exercise 3.11´ [12 points] (verify Bellman equation in gridworld example)**.** (**This differs from the textbook.**) The Bellman equation (3.13) must hold for each state for the value function v_π shown in Figure 3.3 (see SB text, 2nd ed.). As an example, show numerically that this equation holds for the state just below the center state, valued at -0.4, with respect to its four neighboring states, valued at +0.7, –0.6, –1.2, and -0.4. (These numbers are accurate only to one decimal place.)

(f) Exercise 3.12 [12 points] (Bellman equation for action values, q_π).

(g) Exercise 3.13 [9 points] (Adding a constant reward in a continuing task).

(h) Exercise 3.14 [9 points, 3 for each subquestion, 3 for the example] (Adding a constant reward in an episodic task)

(i) Exercise 3.15 [8 points, 4 points for each equation] (half-backup v_π).

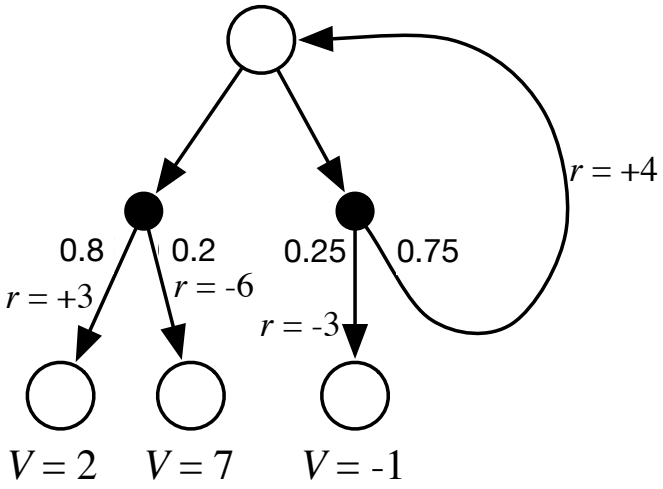
(j) [8 points, 4 for symbolic form, 4 points for numeric answer] Figure 3.6 gives the optimal value of the best state of the gridworld as 24.4, to one decimal place. Use your knowledge of the optimal policy and (3.7) to express this value symbolically, and then to compute it to three decimal places. Hint: Equation (3.9) is also relevant.

Bonus Questions [total 15 points available]. There are **two** bonus questions.

Question 3: Trajectories, returns, and values (**10 Bonus points**)

Consider the following fragment of an MDP graph. The fractional numbers indicate the world?s transition probabilities and the whole numbers indicate the expected rewards. The three numbers at the bottom indicate what you can take to be the value of the corresponding states. The discount is 0.8. What is the value of the top node for the equiprobable random policy (all actions equally likely) and for the optimal policy? Show your work.

$v_\pi =$ $v_* =$



Question 4 [5 bonus points]. Complete Exercise 3.6 (episodic pole balancing)**.** See SB textbook, second ed.