Reinforcement Learning: Midterm 2 April 15, 2020

Name: _____

Start time: _____ Finish time: _____ Total minutes_____

Answer the questions in the answer spaces provided on the question sheets. If you run out of room for an answer, note in the answer space that it is continued on another page, and continue on a blank sheet.

No use of a computing device is allowed other than as a timer, clock, music player, simple (non-programmable) scientific calculator (for example, https://www.calculator.net/scientific-calculator.html), or for word processing (editing this PDF file or writing your answers in some word processor). This is a closed-book exam. You may have one 8.5x11 single-sided handwritten page of notes.

If you think something about a question is open to interpretation, write any assumptions you've made as part of answering the question.

Be concise in your answers; you need not try to fill in all or even most of the space provided for an answer. Show your work, though, since partial credit may be awarded.

You have four contiguous hours to complete this exam starting from when you look at any page other than the first. The four hours is much more than I think you need, but should allow for any needed time for dealing with computer issues in creating your final PDF.

Submit the completed exam to Gradescope no later than 1 PM, April 15, 2020. You may submit:

- An edited version of the PDF with answers added (PDF editor on iPad for example, with handwritten answers, or Preview on Mac with text annotation).
- A scanned paper version of the PDF (that you've handwritten answers on). Make sure the scanned version is legible before you submit it.
- A PDF output of some word processor. For example, you can write your answers in LaTeX, or Microsoft Word (with Equations), or Google Docs (with Auto-Latex addon). You need not repeat the question: just provide your answer.

Good luck!

5 points 1. Circle all of the following methods that use bootstrapping to estimate values:

A. Q-learning
B. Sarsa
C. Expected Sarsa
D. Tree Backup
E. Monte Carlo Tree Search

6 points 2. Circle all of the following statements that are true about Monte Carlo Tree Search (MCTS):

A. MCTS works only for episodic tasks
B. MCTS does planning in the background
C. MCTS does planning at decision time
D. MCTS can be used with a random rollout policy
E. MCTS can be used with a non-random rollout policy
F. The intent of MCTS is to select an action better than that of the underlying rollout policy

5 points 3. When using Temporal Difference learning, why is it better to learn action values (Q-values) rather than state values (V-values)?

5 points 4. Briefly describe what *bootstrapping* is in the context of Reinforcement Learning.

24 points

5. You are given an environment with 1 state, x, and 2 actions, b and c. T is the terminal state. Your TD algorithm generates the following episode using the policy π when interacting with its environment:

Timestep	Reward	State	Action
0		x	b
1	16	x	c
2	12	x	b
3	16	T	

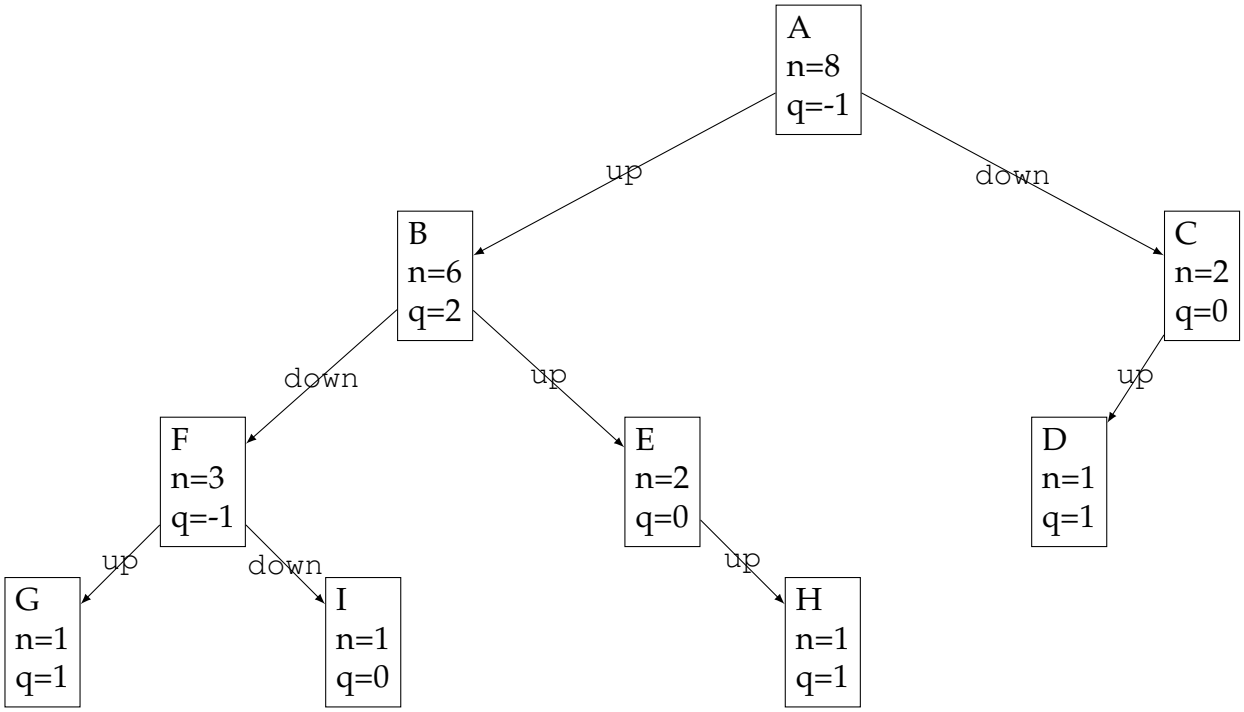
- The policy π is given by: $\pi(b|x) = 0.9, \pi(c|x) = 0.1$
- The current values of q are: $q(x, b) = 1$ and $q(x, c) = 2$.
- the discount factor, γ, is $\frac{1}{2}$.
- the step size, α, is 0.1

Show the values of $q(x, b)$ and $q(x, c)$ after their *first* update using 1-step Sarsa, 2-step Sarsa, 2-step Expected Sarsa, and 2-step Tree Backup. Note: you should update $q(x, b)$ and $q(x, c)$ only once per learning algorithm. **Show your work** and carry out your calculations to *two* decimal places.

Learning Algorithm	$q(x, b)$ after its first update	$q(x, c)$ after its first update
1-step Sarsa	_____	_____
2-step Sarsa	_____	_____
2-step Expected Sarsa	_____	_____
2-step Tree Backup	_____	_____

10 points

6. You are using Monte Carlo Tree Search to decide on the next action for a two-person competitive game with 2 actions at each state (`up` and `down`). It is player 1's turn to play in state A. The state of the tree so far is as follows (each node consists of state identifier, n value, and q value):



Remember that the formula for the UCT value for a node, v, is:

$$UCT(v) = \frac{q(v)}{n(v)} + c\sqrt{\frac{\ln n(v.parent)}{n(v)}}$$

Assume the constant c in the UCT formula is 0.5.

(a) What is the node that is next selected (show your work)?

(b) What are the details of the node that gets expanded from the selected node?

(c) Assuming that the simulation (rollout) from the expanded node gives a value of 1 (that is, player 1 wins), backup that value to all of the affected nodes.

(d) Assume that after this final rollout, we've run out of time to run the MCTS simulation and must now choose an action for player 1 from state A. What action will be chosen and why?

5 points 7. What's the main difference between the Dyna-Q and Dyna-Q+ algorithms?

4 points 8. Direct reinforcement learning updates a *policy* based on *interactions with the environment*. Planning (indirect reinforcement learning), however, updates a _____ based on _____.

4 (bonus) 9. According to `https://ourworldindata.org/coronavirus`, the number of Covid-19 deaths in Italy approximately doubled from March 16, 2020 (370) to March 22, 2020 (795). **Using *the Rule of 72***, approximate the daily percentage increase.