Your Name:
CMPUT 366 / 609 (circle one)

Midterm Exam CMPUT 366/609
Instructors A. White & R. Sutton, Fall 2017

INTELLIGENT SYSTEMS

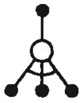
Write your name at the top of this page. Please also indicate if you are a student of **CMPUT 366 or CUMPUT 609** at the top of the page. This is an in-class closed-book exam. No books, computers, or calculators are allowed. You are allowed to bring one page of notes, but they must be handwritten by you personally.

There are 8 questions, most with multiple parts, for a total of 67 points (6+6+3+3+5+6+22+16) points. You can mark, write, or sketch your answers directly on the exam. Sufficient blank space is left for answering each question, but feel free to use the backs of pages if needed. You must turn in your exam within the 80 minute class period. Partial credit will be given for incomplete or partially correct answers *if you show your work.*

Read each question carefully and answer all parts—it will save you points.

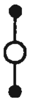
1. (6 pts total) Draw lines connecting the corresponding algorithm names, update diagrams, and update rules:

Sarsa

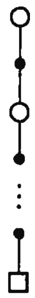


$Q(S_t, A_t) \leftarrow Q(S_t, A_t)+\alpha[R_{t+1}+\gamma \max_a Q(S_{t+1}, a)-Q(S_t, A_t)]$

Monte Carlo
for v_π



$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$



Q -learning

$Q(S_t, A_t) \leftarrow Q(S_t, A_t)+\alpha[R_{t+1}+\gamma Q(S_{t+1}, A_{t+1})-Q(S_t, A_t)]$

This question has three parts, each of which can be answered concisely, but be prepared to explain and justify your concise answer.

2. (6 pts total) Suppose you have a policy π and its action-value function, q_π, then you greedify q_π to produce the deterministic policy π':

$$\pi'(s) = \arg\max_a q_\pi(s,a) \qquad \text{for all } s \in \mathcal{S}.$$

(a) What do you know about the relationship between π and π'?

(b) Now suppose you notice that π' is the same as π. What then do you know about the two policies?

(c) Now suppose you notice that π' is different from π. Do you know anything more about the two policies other than what you reported in part (a)?

3. (3 pts) **Multiple choice:** In TD methods, a larger discount parameter γ, $0 < \gamma < 1$, means

a) a closer approximation to the dynamic-programming solution

b) more concern for immediate rewards relative to later rewards

c) less concern for immediate rewards relative to later rewards

d) both a) and b)

e) both a) and c)

4. (3 points) The goal of reinforcement learning can be seen as producing a ____________, which maps from ____________ to ____________.

5. (5 pts) Suppose the discount rate γ is 0.5 and the following sequence of rewards is observed: $R_1 = 2$, $R_2 = 4$, $R_3 = -8$, $R_4 = 12$, followed by the terminal state. What are the following returns?

G_4?

G_3?

G_2?

G_1?

G_0?

6. (6 pts) Let $X_1, X_2, X_3, \ldots$ be a time sequence of random numbers, each a sample from the same distribution, and let E_t be an estimate of the mean of the distribution. In particular, let it be the sample average of the first $t-1$ numbers. In the space below, specify an incremental way of computing E_t whose memory and per-time-step computation does not increase with t. First specify any initialization of variables, then specify the update of E_{t+1} from E_t and X_t. For full marks, give the update in the form of our standard learning rule ($NewEst \leftarrow OldEst + \cdots$).

Initialization:

Update:

Will the method you define above work well if the distribution from which the samples are drawn changes slowly over time? (Yes or No and justify your answer)

7. All about v_π (22 pts total)

Consider the value function v_π for a stochastic policy π and a continuing finite Markov decision process with discounting.

(a) [3 pts] Give an equation *defining* $v_\pi(s)$ in terms of the subsequent rewards $R_{t+1}, R_{t+2}, \ldots$ that would follow if the MDP were in state s at time t. (If you choose to write it in terms of the return, G_t, define your return notation in terms of the underlying rewards.)

(b) [3 pts] Sketch the update diagram (also sometimes known as a backup diagram) for the dynamic programming algorithm (expected update) for v_π. Find a place to attach the labels s, a, r, and s'.

(c) [4 pts] What is the Bellman equation for v_π? Write it in an explicit form in terms of $p(s', r|s, a)$ so that no expected value notation appears.

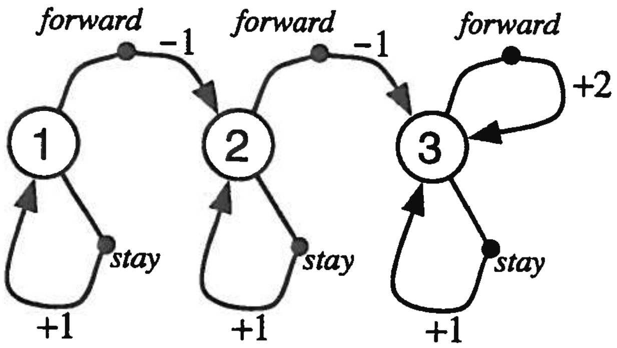
(d) [4 pts] Consider the simplest dynamic-programming algorithm for computing v_π. An array $V(s)$ is initialized to zero. Then there are repeated sweeps through the state space, with one update to an array element done for each state in each sweep. What is that DP update?

(e) [4 pts] Consider the simplest temporal-difference learning method for estimating v_π from experience. An array $V(s)$ is initialized to zero. Then an infinite sequence of experience, $S_0, A_0, R_1, S_1, A_1, R_2, S_2, \ldots$, is processed, with one update to $V(S_t)$ done for each transition. What is the equation for that TD update?

(f) [4 pts] Now consider the simplest Monte Carlo learning method for estimating v_π from *episodic* experience. An array $V(s)$ is initialized to zero. Then an infinite sequence of episodes is experienced, where an individual episode of experience is denoted $S_0, A_0, R_1, S_1, A_1, R_2, S_2, \ldots, R_T, S_T$, where T is the final time step of the episode, S_T is the terminal state, and the value of all terminal states is taken to be zero. When an episode is processed, one update to $V(S_t)$ is made for each time step $t < T$ in the episode. What is the equation for that Monte-Carlo update? (If you choose to write it in terms of the return, define your return notation in terms of the underlying rewards.)

8. Markov Decision Process (16 points total)

Consider the **continuing** finite MDP in the figure below. There are three states, (1, 2, and 3), and two actions, *forward* and *stay*. The *forward* action takes the agent to a higher numbered state (except in state 3), and the *stay* action keeps the agent in the same state. The effect of all actions is deterministic. The expected rewards on each transition are as indicated in the figure.



(a) (6 points) Suppose the discount factor is $\gamma = \frac{1}{2}$. What then is the optimal value function and the optimal deterministic policy?

$v_*(1) =$ $\pi_*(1) =$

$v_*(2) =$ $\pi_*(2) =$

$v_*(3) =$ $\pi_*(3) =$

(b) (6 points) Suppose the discount factor is $\gamma = \frac{3}{4}$. What then is the optimal value function and the optimal deterministic policy?

$v_*(1) =$ $\pi_*(1) =$

$v_*(2) =$ $\pi_*(2) =$

$v_*(3) =$ $\pi_*(3) =$

(c) (4 points) Suppose you are doing learning by the tabular TD(0) algorithm. You start with all value estimates $V(s) = 0$, and you observe the following partial trajectory (sequence of states, actions and rewards, where the state numbers are bolded):

1, *forward*, -1, **2**, *stay*, +1, **2**

Assuming the step size is $\alpha = 0.5$, and $\gamma = 0.5$, show the updates that are performed.