

Your Name:


Midterm Exam CMPUT 366/609
R. Sutton, instructor, Fall 2013
INTELLIGENT SYSTEMS


Write your name at the top of this 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 69 points (6 + 3 + 3 + 4 + 3 + 22 + 22 + 6). 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, backup diagrams, and update rules:

Q-learning  $V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$

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

TD(0)  $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$

2. (3 pts) **Multiple choice:** In learning methods, a larger step size, α , usually means
- a) more rapidly approaching the final performance level
 - b) greater error before reaching the final performance level
 - c) less residual error at the final performance level
 - d) reduced risk of divergence
 - e) both a) and d)
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. (4 pts total) In *batch updating*, a fixed training set of data (experience interacting with an MDP) is presented over and over again to a learning algorithm, with updates made only after processing the whole set. For the tabular case and a sufficiently small fixed step size α , each algorithm will converge to a characteristic solution independent of α .
- True or False:** Comparing TD and Monte Carlo solutions according to their Mean Square Error (MSE):
- (a) **T F** (2 pts) TD solutions will usually have lower MSE on new data
 - (b) **T F** (2 pts) TD solutions will never have lower MSE on the training data
5. (3 pts) In this course we have considered both learning methods and planning methods for solving MDPs. What is the difference between learning methods and planning methods?

6. 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}, \dots 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.)

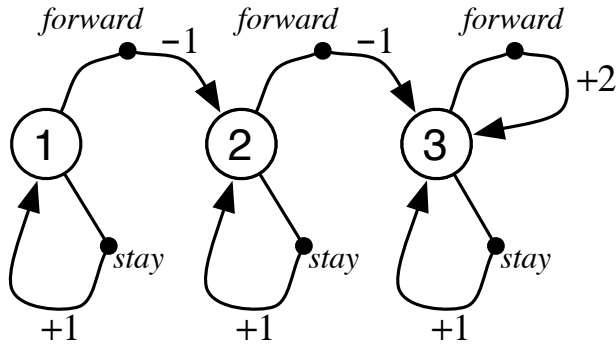
- (b) [3 pts] Sketch the backup diagram for the dynamic programming algorithm (full backup) for v_π . Find a place to attach the labels s , a , r , and s' .

- (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. 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, $S_0, A_0, R_1, S_1, A_1, R_2, S_2, \dots$, 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 $S_0, A_0, R_1, S_1, A_1, R_2, S_2, \dots, R_T, S_T$, 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 $t < T$. 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.)

7. Markov Decision Process (22 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, 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?

$$\begin{array}{ll} v_*(1) = & \pi_*(1) = \\ v_*(2) = & \pi_*(2) = \\ v_*(3) = & \pi_*(3) = \end{array}$$

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

$$\begin{array}{ll} v_*(1) = & \pi_*(1) = \\ v_*(2) = & \pi_*(2) = \\ v_*(3) = & \pi_*(3) = \end{array}$$

- (c) (6 points) For what range of values for γ is the optimal policy to select *forward* in all states?

- (d) (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.