

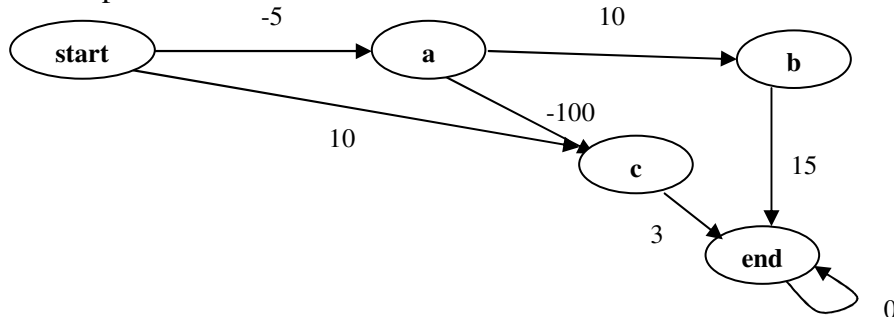
CS 760 - Homework 4

Out: 4/13/09

Due: 4/22/09

50 points

Consider the deterministic reinforcement environment drawn below. The numbers on the arcs are the *immediate* rewards. Let the discount rate equal 0.9 and the probability of taking an *exploration* step be 0.05.



- 1) Assume you wish to use a *Q table* to represent the Q function. All cells in this table should initially contain zero. Also assume your RL agent uses 1-step Q-learning. Show the state of your Q table after each of the following "episodes" (to represent the Q table, you can simply draw a copy of the above graph, but instead of attaching immediate rewards to arcs, attach the Q values). Be sure to show your work.
 - i. $start \rightarrow a \rightarrow b \rightarrow end$
 - ii. $start \rightarrow a \rightarrow c \rightarrow end$
 - iii. $start \rightarrow c \rightarrow end$
 - iv. $start \rightarrow a \rightarrow b \rightarrow end$
- 2) Repeat Part 1, using a fresh Q table (i.e, all cells filled with a zero), but this time use SARSA. For SARSA do you need to use α (a "learning rate" - see Equation 13.10 of Mitchell)? If so, set α as described in Lecture 24, slides 9-11. Explain your answer.
- 3) Repeat Part 1, again using a fresh Q table, but this time use $TD(\lambda)$, with $\lambda = \frac{1}{2}$. Since $TD(\lambda)$ involves a good deal of calculation, you only need to process the first two episodes of Part 1. Do you need to use α here? Explain.
- 4) If you performed RL for a large number of episodes, what policy would Q learning produce? Indicate this policy by copying the above graph and using thick arrows to represent the policy. Briefly explain your answer.
- 5) Repeat Part 4 using SARSA. Show this policy on a *fresh* copy of the graph.