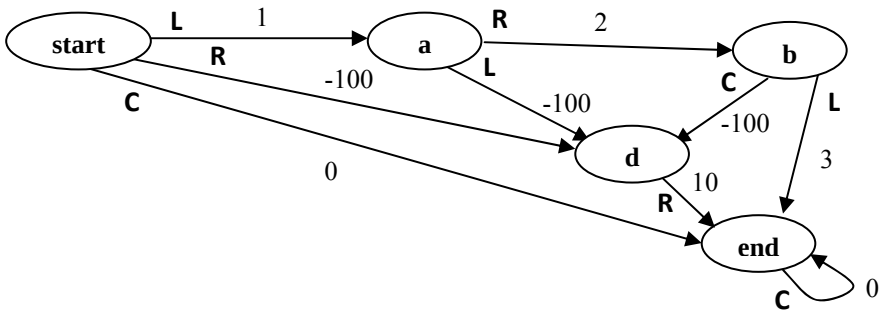
CS 760 - Homework 4

Out: 4/12/10
Due: 4/19/10
50 points

Consider the deterministic reinforcement environment drawn below. The numbers on the arcs are the *immediate* rewards. Let the discount rate equal 0.8 and the probability of taking an *exploration* step be 0.02. The **L/R/C** at the beginning of arcs is the name of the action that arc represents (this information is used in Part 4).



1) Assume you wish to use a Q *table* to represent the Q function. All cells in this table should initially contain 3 (an arbitrary choice). Also assume your RL agent uses 1-step Q-learning. Show the state of your Q table after <u>each</u> 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 → a → b → d → end*
   ii. *start → a → b → end*
   iii. *start → a → d → end*

2) Repeat Part 1, using a fresh Q table (i.e, all cells filled with 3), 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 23-25. Explain your answer.

3) 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.

4) Imagine states are presented using three Boolean-valued features as follows:

start = 001 a = 010 b = 100 c = 110 end = 000

Discuss how you can use a perceptron for each action (**L**, **R**, and **C**) to represent the Q function.

Draw the perceptrons. Assume all weights (and biases) initially have value 3 and that the output units are not thresholded (i.e., their output is simply the weighted sum of their inputs).

Show how you would use SARSA to train these perceptrons for the actions listed in Part 1i using eta = 0.5. Following this first episode, what do these perceptons estimate as the Q values for the initial state's three possible actions?