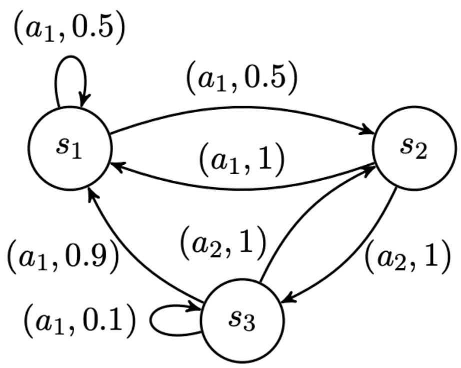
Homework 4
Due April 5th by 11:59pm



(a) Transition graph

R	a_1			a_2		
	s_1	s_2	s_3	s_1	s_2	s_3
s_1	-1	0.5	—	—	—	—
s_2	10	—	—	—	—	-1
s_3	0	—	-1	—	10	—

(b) Reward table for the MDP

Figure 3: Example of a 3-state, 2-action MDP with (action, probability) on arcs

Consider the MDP above. There are three states: s1, s2, and s3. There are two actions: a1 and a2. Edges are labeled with (action, probability) pairs. For example, taking action a1 in state s1 leads back to s1 with probability 0.5 and to s2 with probability 0.5.

The table to the right gives rewards. The left column is the starting state. For each action, the rewards are given for each destination state. For example, taking action a1 in state s2 with you return to s1 yields a reward of 10.

Suppose you start with a Q-table initialized to all zeroes. Show the value of the Q-table after each of the following transitions, taken in order, with the a learning rate and discount factor of 0.5. Note that the starting Q-table for the second update is the table after the first update.

S1, A1, S2
S2, A2, S3

Show your work for partial credit.

What is the optimal policy for this MDP? Explain briefly why.