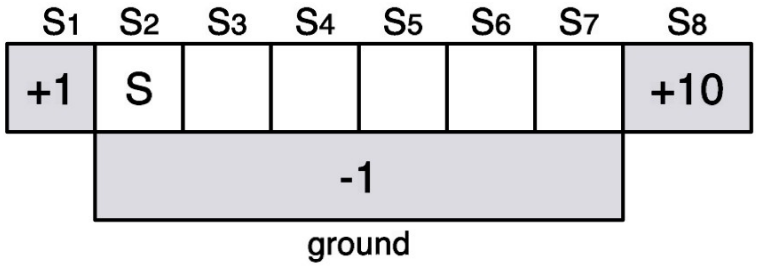
CS188 – Introduction to Artificial Intelligence

Section Handout #5, FORMULATING AND SOLVING MDPS

Klein, Fall 2007

Question 1 (Class)



Consider the above MDP, representing a robot on a balance beam. Each grid square is a state and the available actions are right and left. The agent starts in state s2 , and all states have reward 0 aside from the ends of the grid s1 and s8 and the ground state, which have the rewards shown. Moving left or right results in a move left or right (respectively) with probability p. With probability 1 − p, the robot falls off the beam (transitions to ground, and receives a reward of -1). Falling off, or reaching either endpoint, result in the end of the episode (i.e., they are terminal states). Note that terminal states receive no future reward.

a. For what values of p is the optimal action from s_2 to move right if the discount γ is 1?

b. For what values of γ is the optimal action from s_2 to move right if p = 1?

c. Given initial value estimates of zero, show the results of one, then two rounds of value iteration.

d. We can develop learning updates that involve two actions instead of one. Write down the utility $U^{\pi}(s)$ of a state s under policy π in terms of the next two states s' and s'', given that

$$U^{\pi}(s) = \Sigma_{s'} T(s, \pi(s), s')[R(s, \pi(s), s') + \gamma U^{\pi}(s')]$$

e. Write a two-step-look-ahead value iteration update that involves U (s) and U (s''), where s'' is the state two time steps later. Why would this update not be used in practice?

f. Write a two-step-look-ahead TD-learning update that involves U (s) and U (s'') for the observed state-action-state-action-state sequence s, a, s', a', s''

g. Given initial q-value estimates of zero, show the result of Q-learning with learning rate α = 0.5 after two epsiodes: [s2 , s3 , ground] and [s2 , s3 , s4 , s5 , ground] where the agent always moves right. You need only write down the non-zero entries. For the purposes of Q-learning updates, terminal states should be treated as having a single action die which leads to future rewards of zero. Hint: q-values of terminal states which have been visited should not be zero.

Question 1 (Class)

Golf as an MDP

We formulate golf as an MDP as follows:

State Space : {*Tee, Fairway,Sand, Green*}

Actions : {*Conservative shot, Power shot*}

Initial State : *Tee*

Transition model : (note that action not on this list have probability 0)

s	a	s'	$T(s,a,s')$
Tee	Conservative	Fairway	0.9
Tee	Conservative	Sand	0.1
Tee	Power shot	Green	0.5
Tee	Power shot	Sand	0.5
Fairway	Conservative	Green	0.8
Fairway	Conservative	Sand	0.2
Sand	Conservative	Green	1.0

Rewards:

(note: R(·,·,s) means that the reward is received for transitioning *to* state s, regardless of action taken or previous state)

s	R(·,·,s)
Fairway	-1
Sand	-2
Green	3

a. Consider the policy of always taking the "Conservative Shot". What is the utility of the initial state under this policy?

b. Compute estimates of the utility of each state under the optimal policy using Value Iteration with 3 iterations. Show the utilities of each state at each iteration. Assume we start with all utilities set to 0.