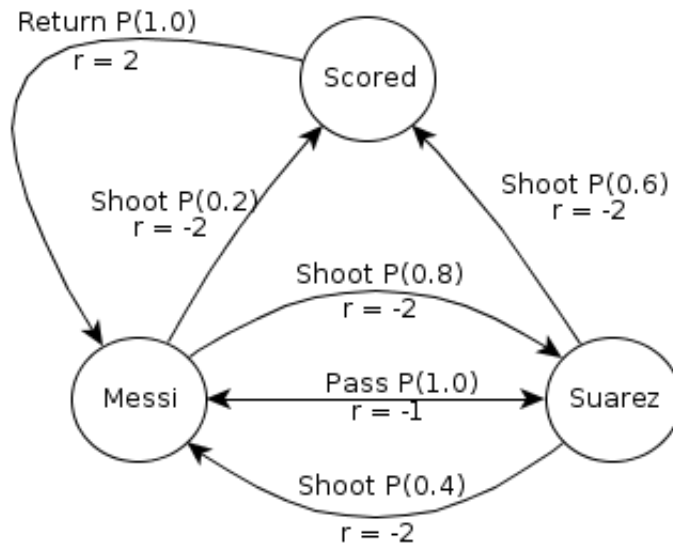


Problem Set X: Reward shaping and policy-based approaches

1. Consider again the two football-playing robots, Messi and Suarez, from an earlier tutorial. Recall that there are three states: *Messi*, *Suarez* (denoting who has the ball), and *Scored* (denoting that a goal has been scored); and the following action descriptions:

- If Messi shoots, he has 0.2 chance of scoring a goal and a 0.8 chance of the ball going to Suarez. Shooting towards the goal incurs a cost of 2 (or a reward of -2).
- If Suarez shoots, he has 0.6 chance of scoring a goal and a 0.4 chance of the ball going to Messi. Shooting towards the goal incurs a cost of 2 (or a reward of -2).
- If either player passes, the ball will reach its intended target with a probability of 1.0. Passing the ball incurs a cost 1 (or a reward of -1).
- If a goal is scored, the only action is to return the ball to Messi, which has a probability of 1.0 and has a reward of 2. Thus the reward for scoring is modelled by giving a reward of 2 when *leaving* the goal state.

The following diagram shows the transition probabilities and rewards:



Consider the following policy update table and policy evaluation table, with discount factor $\gamma = 0.8$:

Iter	$Q(\text{Messi}, P)$	$Q(\text{Messi}, S)$	$Q(\text{Suarez}, P)$	$Q(\text{Suarez}, S)$	$Q(\text{Scored})$
0	0	0	0	0	0
1					
2	-4.194	-4.772	-4.355	-3.993	-1.355

Apply two iterations of policy iteration. Finish both tables and show the working for the policy evaluation and policy update.

What is the policy after two iterations?

Iter	$\pi(Messi)$	$\pi(Suarez)$	$\pi(Scored)$
0	Pass	Pass	Return
1			Return
2			Return

2. Consider a robotic helper at a hospital that delivers items to staff. The robot is given a task to deliver a treatment kit to a medical specialist in a room. The robot has to pickup the kit from the storeroom, but first has to go to get the key for the storeroom. However, it does not know in advanced whether the key will be there.

The robot will receive a reward of +10 for delivering the kit, and a reward of +5 for going to the room to inform the specialist that the key is missing. There are no other rewards.

Consider this as the following map, where S is the starting state, K is the key rack, M is the medical store room, and R is the room where the store is to be delivered.

5		M													
4												R			
3															
2															
1				###											
0		S						K							
	0	1	2	3	4	5	6	7							

Your task: design a potential function for this problem. You can assume that you can know the the position of the agent, the position of S, M, K, and R, and you can see the a variable *Key* with values 0, 1, and 2, where 0 indicates there we do not know if the key is in the room, 1 is the key is in the room, and 2 the key is not in the room, and a Boolean variable *Med* to indicate whether the agent has the medicale kit. Initially, $Key = 0$ and $Med = False$.

3. Using your potential function, perform two different reward shaping updates using Q-learning from state K (4,0) and the agent has found the key.

First, perform for the action *Up* ending in state (4,1).

Then, assume that the *Right* action had been chosen instead of *Up*, ending in state (5,0).

Compare the two updates to see whether your reward shaping function has worked.

Assume that $Q(s, a) = 0$ for all s and a , $\gamma = 0.9$ and $\alpha = 0.2$.