AI Planning for Autonomy

Problem Set XII: MDPs and Reinforcement Learning

Most of the questions on this workshop related to the following description.

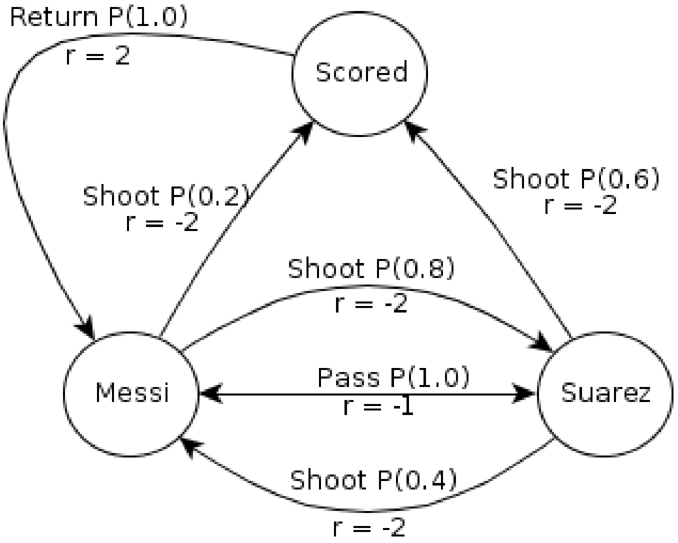
Consider two football-playing robots: Messi and Suarez.

They play a simple two-player cooperate game of football, and you need to write a controller for them. Each player can pass the ball or can shoot at goal.

The football game can be modelled as a discounted-reward MDP with 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:



1. What is the difference between Sarsa and Q-learning?
2. Assume that we have calculated the following *non-optimal* value function V for this problem using value iteration with $\gamma = 1.0$, after 3 iterations we arrive at the following:

Iteration		1	2	3	4
V(Messi)	=	0.0	-1.0	-2.0	
V(Suarez)	=	0.0	-1.0	-1.2	
V(Scored)	=	0.0	2.0	1.0	

If Messi has the ball (the system is in the Messi state), what action should we choose to maximise our reward in the next state: pass or shoot? Assume we are using the values for V after three iterations.

3. Complete the values of these states for iteration 4 using value iteration. Show your working.

4. Given the following trace from a historical soccer game feed from last season:

   “ Suarez passes the ball to Messi, Messi dribbles around all of his opponents, shoots and scores yet another goal! Barcelona F.C 10 - 0 Real Madrid! End of the game, Messi takes the ball to remember the match forever.”

   Perform `TD(0) updates`, using a `discount factor` $\gamma = 0.9$, starting from the **3rd iteration** V values given on the table above.