Problem Set VII: Value & Policy Iteration

Aim The purpose of this workshop is to help you get a better understanding of MDPs, value iteration, and policy iteration.

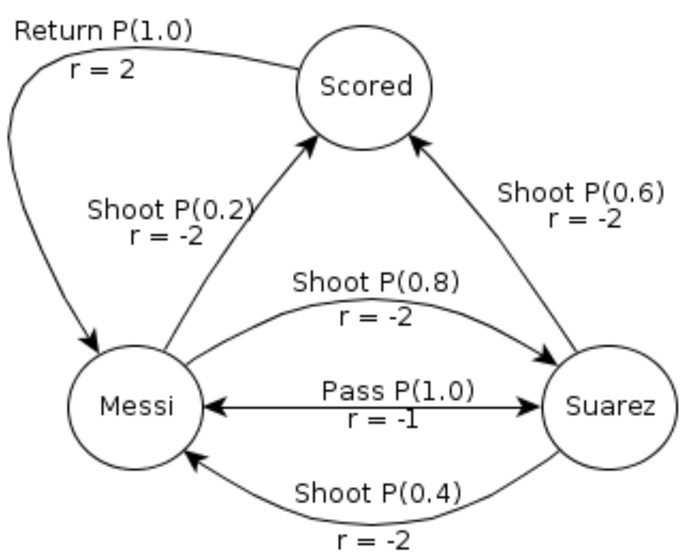
Consider two football-playing robots: Messi and Suarez.

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:



Tasks

1. Assume that we have calculated the following *non-optimal* value function V for this problem using value iteration with $\gamma = 1.0$, after iteration 2 we arrive at the following:

Iteration		0	1	2	3
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.

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

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

Iter	$Q^{\pi}(Messi, P)$	$Q^{\pi}(Messi, S)$	$Q^{\pi}(Suarez, P)$	$Q^{\pi}(Suarez, S)$	$Q^{\pi}(Scored)$
0	0	0	0	0	0
1					
2	-4.194	-5.465	-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

Additional Tasks for Personal Study

To improve your understanding of value iteration, try completing the first question of Project 3 at `http://ai.berkeley.edu/reinforcement.html#Q1`. You can download all the necessary files to complete this task.

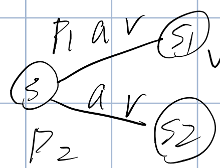
Hints In order to help you complete the task during the workshop, here are some useful hints:

1. The functions that you need to change:
   (a) __init__
   (b) computeQValueFromValue
   (c) computeActionFromValue
2. The files you need to take a look:
   (a) util.py (Counter())
   (b) mdp.py (isTerminal(), getStates() ,getPossibleActions(), etc.)
   (c) other files: (gridworld.py, learningAgent.py) not directly related
3. How to test your code:
   (a) python autograder.py -q q1 (testing by autograde)
   (b) python gridworld.py -a value -i 5 (result after 5 iteration)
   (c) python gridworld.py -a value -i 100 -k 10 (how value iteration works)

MDP

- S
- $A(S)$
- $P_a(s'|s)$
- $r(s, a, s')$

$\pi(s) = a$



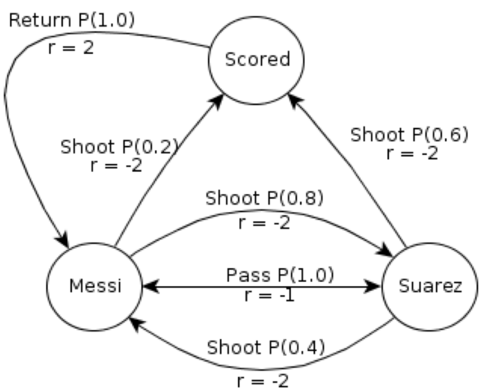
$P_1(r+V)$

$P_2(r+V)$

$$\pi(s) = \underset{a \in A}{\operatorname{argmax}}\ Q(s, a)$$

$$Q(s, a) = \sum_{s'} P_a(s'|s)\,[r + \gamma V(s')]$$

$$V(s) = \max_{a \in A} Q(s, a)$$



it	0	1	2	3
V(Messi)	0	-1	-2	-2.2
V(Suarez)	0	-1	-1.2	-2.2
V(scored)	0	2	1	0

① Calculate the Q

$V(\text{Messi})$

pass $\rightarrow Q(\text{Messi}, \text{pass}) = P_{pass}(\text{Suaraz} \mid \text{Mess})\,[r(M, p, S) + \gamma V(S_a)]$

$= 1[-1 + 1(-1.2)] = -2.2$

shoot $\rightarrow Q(\text{Messi}, \text{shoot}) =$

$P(\text{Squarez} \mid M)\,[r(M, S, \text{Suarez}) + \gamma V(\text{Suarez})]$

$+ P_{shoot}(\text{Scored} \mid M)\,[r(M, S, \text{scored}) + \gamma V(\text{Scored})]$

$= 0.8[-2 + 1(-1.2)] + 0.2[\cancel{-2} + 1(1)] = -2.76$

Messi should pass, because pass has high Q

pass $\quad Q(S, \text{pass}) = P_{pass}(\text{Messi} \mid \text{Suaraz})\,[r(S_u, P, M)$

$$V(\text{Suarez}) \xrightarrow{\text{shoot}}$$

$$+\gamma V(\text{Messi})$$

$$= 1[-1+1(-2)] = -3.$$

$$Q(\text{su}, \text{shoot}) = P_{\text{shoot}}(\text{Messi} \mid \text{suarez})$$

$$[r(\text{suarez}, \text{shoot}, M) + \gamma V(\text{Messi})]$$

$$+ P_{\text{shoot}}(\text{Scored} \mid \text{Suarez})[r(\text{Suarez}, \text{shoot}$$

$$+ \gamma V(\text{scored})]] = 0.4[-2+1(-2)] +$$

$$0.6[-2+1(1)]$$

$$= -1.6 + -0.6$$

$$= -2.2$$

$$V(\text{scored}) = P_{\text{return}}(M \mid \text{scored})[r + \gamma V(\text{Messi})]$$

$$= 1[2+1(-2)] = 0$$

Stop iteration when the values don't change or the iteration meets the maximum times.

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

$$V^{\pi}(Messi) = Q(Messi, pass) = P(Suarez \mid Messi)[r + \gamma V(Su)]$$
$$= 1[-1 + \gamma V(Sua)]$$

$$V^{\pi}(Suarez) = Q(Suarez, pass) = P_{pass}(Messi \mid Suarez)$$
$$[r + \gamma V(Messi)]$$
$$1[-1 + \gamma V(Messi)]$$

$$\begin{cases} M = -1 + \gamma S \\ S = -1 + \gamma M \\ SC = 2 + \gamma M \end{cases}$$

$$V(scored) = Q(scored, return) =$$
$$1[2 + \gamma M$$

$$S = \frac{1}{\gamma - 1} = -5$$

$$M = \frac{1}{\gamma - 1} = -5$$

$$SC = -2.$$

Iteration, we will never use these values for next

Iter	$Q^\pi(Messi, P)$	$Q^\pi(Messi, S)$	$Q^\pi(Suarez, P)$	$Q^\pi(Suarez, S)$	$Q^\pi(Scored)$
0	0	0	0	0	0
~~1~~	-5	-5.52	-5	-4.56	-2
2	-4.194	-5.465	-4.355	-3.993	-1.355

$$Q(M, shoot) = P_{shoot}(Suarez \mid Messi)[r + \gamma V(Suarez)]$$

$$+ P_{shoot}(Scored \mid Messi)[r + \gamma V(Scored)]$$

$$= -5.52$$