AI Planning for Autonomy

Sample Solutions for Problem Set VII: Value & Policy Iteration

1. We need to calculate the expected return for each action: pass or shoot.

   If Messi passes:

   $$\begin{aligned} V(Messi) &= P_{pass}(Suarez)[r(Messi, pass, Suarez) + \gamma \cdot V(Suarez)] \\ &= 1 \cdot [-1 + 1 \cdot -1.2] \\ &= 1 \cdot -2.2 \\ &= -2.2 \end{aligned}$$

   If Messi shoots:

   $$\begin{aligned} V(Messi) &= P_{shoot}(Suarez|Messi)[r(Messi, shoot, Suarez) + \gamma \cdot V(Suarez)] + \\ &\quad P_{shoot}(Scored|Messi)[r(Messi, shoot, Scored) + \gamma \cdot V(Scored)] \\ &= 0.8[-2 + 1 \cdot -1.2] + 0.2[-2 + 1 \cdot 1.0] \\ &= -2.56 + (-0.2) \\ &= -2.76 \end{aligned}$$

   Therefore, to maximise our reward, Messi should pass.

2. To calculate $V(Messi)$, we choose the action that maximises our Q-value (expected future discounted reward):

   $$\begin{aligned} V(Messi) &= \max(Q(Messi, pass), Q(Messi, shoot)) \\ &= \max(-2.2, -2.76) \text{ (from previous question)} \\ &= -2.2 \end{aligned}$$

   For *Scored*, there is only one action, which leads directly to the *Messi* state:

   $$\begin{aligned} V(Scored) &= P_{return}(Messi|Scored)[r(Scored, return, Messi) + \gamma \cdot V(Messi)] \\ &= 1[2 + 1 \cdot -2.0] \\ &= 0 \end{aligned}$$

   For Suarez, the situation is similar to Messi:

   $$\begin{aligned} V(Suarez) &= \max(Q(Suarez, pass), Q(Suarez, shoot)) \\ &= \max(P_{pass}(Messi|Suarez)[r(Suarez, pass, Messi) + \gamma \cdot V(Messi), \\ &\quad \Big(P_{shoot}(Messi|Suarez)[r(Suarez, shoot, Messi) + \gamma \cdot V(Messi) + \\ &\quad P_{shoot}(Scored|Suarez)[r(Suarez, shoot, Scored) + \gamma \cdot V(Scored)]\Big) \\ &= \max(1.0[-1 + 1 \cdot -2.0], (0.4[-2 + 1 \cdot 2.0] + 0.6[-2 + 1 \cdot 1.0])) \\ &= \max(-3, (0.4[-2 + 1 \cdot -2.0] + 0.6[-2 + 1 \cdot 1.0])) \\ &= \max(-3, (-1.6 + -0.6)) \\ &= -2.2 \end{aligned}$$

   Thus, the new table is:

Iteration		1	2	3	4
V(Messi)	=	0.0	-1.0	-2.0	-2.2
V(Suarez)	=	0.0	-1.0	-1.2	-2.2
V(Scored)	=	0.0	2.0	1.0	0.0

3. Policy Iteration has two main steps, policy evaluation and policy update. In order to evaluate the given policy:

$$\begin{aligned}
V^\pi(Messi) &= Q^\pi(Messi, Pass) \\
&= P_{pass}(Suarez)[r(Messi, pass, Suarez) + \gamma \cdot V^\pi(Suarez, Pass)] \\
&= \gamma \cdot V^\pi(Suarez, Pass) - 1 \\
V^\pi(Suarez) &= Q^\pi(Suarez, Pass) \\
&= P_{pass}(Messi)[r(Suarez, pass, Messi) + \gamma \cdot V^\pi(Messi, Pass)] \\
&= \gamma \cdot V^\pi(Messi, Pass) - 1 \\
V^\pi(Scored) &= Q^\pi(Scored, return) \\
&= P_{return}(Messi)[r(Scored, return, Messi) + \gamma \cdot V^\pi(Messi, Pass)] \\
&= \gamma \cdot V^\pi(Messi, Pass) + 2
\end{aligned}$$

Then solve a very basic linear algebra about $V^\pi(Messi)$ and $V^\pi(Suarez)$:

$$\begin{aligned}
V^\pi(Messi) &= 1/(\gamma - 1) \\
V^\pi(Suarez) &= 1/(\gamma - 1) \\
V^\pi(Scored) &= 3 + 1/(\gamma - 1)
\end{aligned}$$

Then apply $\gamma = 0.8$, the policy evaluation table would be:

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	-4.772	-4.355	-3.993	-1.355

Then implement two iteration of policy update based on value from the policy evaluation table:

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