

Sample Solutions for Problem Set VII: Value Iteration

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

If Messi passes:

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

If Messi shoots:

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

Therefore, to maximise our reward, Messi should pass.

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

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

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

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

For Suarez, the situation is similar to Messi:

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

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