AI Planning for Autonomy

Sample Solutions for Problem Set XII: MDPs and Reinforcement Learning

1. The difference between Sarsa and Q-learning is that Q-learning is "off-policy" learning, while Sarsa is "on-policy" learning. Essentially, this means that Sarsa chooses its action using the same policy used to choose the previous action, and then uses this difference to update its Q-function; while Q-learning simply chooses the next value based on the maximum Q-value.

2. 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.

3. 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

4. For the TD updates, these can be calculated as:

$$\begin{aligned} V(Messi) &= V(Messi) + r(shoot) + \gamma \cdot V(Scored) - V(Messi) \\ &= -2.0 + (-2) + 0.9 \cdot 1.0 - (-2.0) \\ &= -4 + 0.9 + 2.0 \\ &= -1.1 \end{aligned}$$

$$\begin{aligned} V(Suarez) &= V(Suarez) + r(pass) + \gamma \cdot V(Messi) - V(Suarez) \\ &= -1.2 + (-1) + 0.9 \cdot -2.0 - (-1.2) \\ &= -2.2 - 1.8 + 1.2 \\ &= -2.8 \end{aligned}$$

$$\begin{aligned} V(Scored) &= V(Scored) + r(return) + \gamma \cdot V(Messi) - V(Scored) \\ &= 1.0 + 2 + 0.9 \cdot -2.0 - 1.0 \\ &= 3 - 1.8 - 1.0 \\ &= 0.2 \end{aligned}$$