Week 9: Temporal difference learning

COMP90054 – AI Planning for Autonomy

Key concepts

- Q-learning and SARSA
- On-policy vs off-policy learning

Model-based vs Model-free

- Model-based: Know the transition probability $P_a(s'|s)$ and reward function r(s, a, s’)
    - E.g: Value Iteration
- Model-free: Don’t know the transition probability and reward function
    - E.g: SARSA, Q-learning

Q-learning vs. SARSA

Q-learning (Off-policy)	SARSA (On-policy)
$Q(s,a) = Q(s,a) + \alpha[r + \gamma \max_{a' \in A(s')} Q(s',a') - Q(s,a)]$	$Q(s,a) = Q(s,a) + \alpha[r + \gamma Q(s',\pi(s')) - Q(s,a)]$
Update rule: Not update based on the policy. Update Q-function based on the assumption that the next action would be the action with the maximum Q. Optimistic: the greedy action will be chosen while in fact, the policy may choose an action other than the best	Update rule: Updated based on the policy. We know the action that it will execute next (whether it is best or not) when performing the update
• Learning from prior experience • The main advantage of off-policy approaches is that they can use samples from sources other than their own policy.	• Learning on the job • The main advantage of on-policy approaches is that they can learn optimal behaviour while operating in their environment.

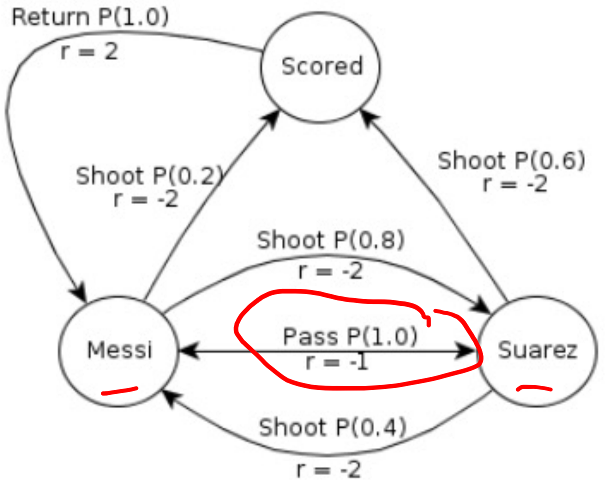
Problem 2: Q-learning

Q - Table

$Q(\text{Messi}, \text{Pass}) = -0.4$

action state	Pass	Shoot	Return
Messi	-0.4	-0.8	-
Suarez	-0.7	-0.2	-
Scored	-	-	1.2

The following diagram shows the transition probabilities and rewards:



In the next step of the episode, from the state 'Suarez', Suarez passes the ball to Messi. Show the Q-learning update for this action using a discount factor $\gamma = 0.9$ and learning rate $\alpha = 0.4$

Note: Assume that this is a model-free problem, so the transition probabilities are not accessible to your algorithm.

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma \max_{a' \in A(s')} Q(s',a') - Q(s,a)]$$

$$\underset{s}{\text{Suarez}} \xrightarrow[a]{\text{pass}} \underset{s'}{\text{Messi}}$$

$$Q(\text{Suarez}, \text{pass}) = Q(\text{Suarez}, \text{pass}) + \alpha\left[r + \gamma \max_{a' \in A(\text{Messi})} Q(\text{Messi}, a') - Q(\text{Suarez}, \text{pass})\right]$$

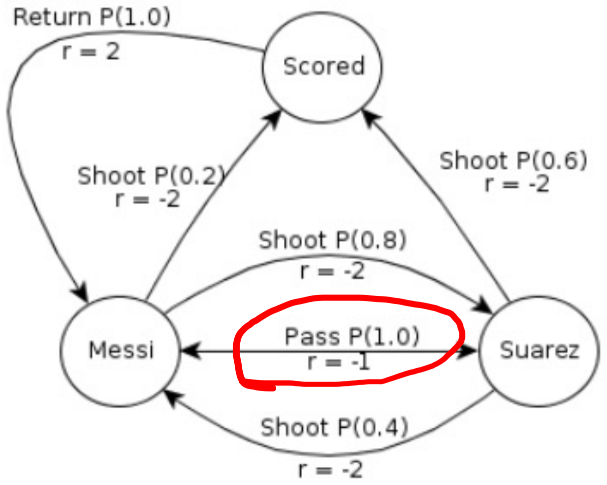
$$= -0.7 + 0.4[-1 + 0.9(-0.4) - (-0.7)]$$

$$= -0.964$$

Finite number of actions and states

Problem 3: SARSA

The following diagram shows the transition probabilities and rewards:



	Pass	Shoot	Return
Messi	~~-0.4~~	-0.8	-
Suarez	-0.7	-0.2	-
Scored	-	-	1.2

Consider again being in the state 'Suarez', Suarez passes the ball to Messi and then Messi decides to shoot. Show the SARSA update for the Pass action using a discount factor $\gamma = 0.9$ and learning rate $\alpha = 0.4$ and assuming a' (the next action to be execute) is ***Shoot***. Compare to the Q-learning update. What is different?

$$Q(s,a) = Q(s,a) + \alpha[r + \underline{\gamma\, Q(s',\pi(s'))} - Q(s,a)]$$

$$\underset{s}{\text{Suarez}} \xrightarrow{\text{pass}} \underset{s'}{\text{Messi}} \xrightarrow{\text{shoot}}$$

$$Q(\text{Suarez}, \text{pass}) = \underline{Q(\text{Suarez}, \text{pass})} + \alpha\left[r + \gamma\, \underline{Q(\text{Messi}, \pi(\text{Messi}))} - Q(\text{Suarez}, \text{pass})\right]$$

$$= -0.7 + 0.4\left[-1 + 0.9\, Q(\text{Messi}, \text{shoot}) - (-0.7)\right]$$

$$= -0.7 + 0.4\left[-1 + 0.9 \times (-0.8) - (-0.7)\right]$$

$$= -1.108$$

Problem 3: N-step TD (SARSA)

N-step TD will not be examinable

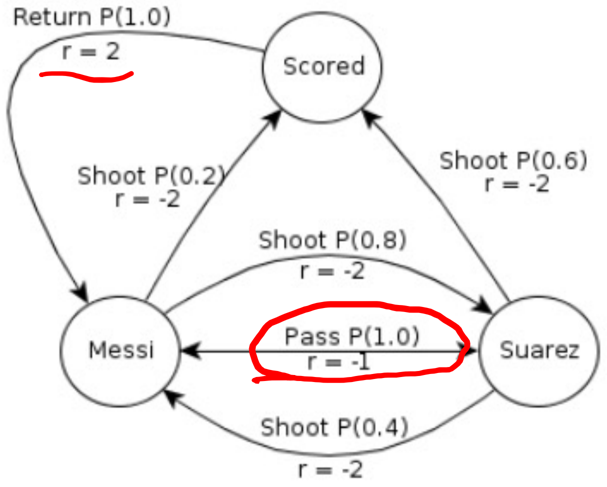
Problem 3: N-step TD (SARSA)

Q- Table

	Pass	Shoot	Return
Messi	-0.4	-0.8	-
Suarez	-0.7	-0.2	-
Scored	-	-	1.2

The following diagram shows the transition probabilities and rewards:

P, r



Given the following trace from a historical game feed from last season: "Suarez passes the ball to Messi, 1 Messi dribbles around all of his opponents, 2 shoots and scores yet another goal! Barcelona F.C 10 - 0 Real Madrid! The ball is 3 returned to Messi for kickoff. After he passes the ball to Suarez, 4 the referee blew the final whistle. End of the game, the ball is taken by Messi to remember the match forever."

Show the 3-step SARSA update for the above feed. Discount factor $\gamma = 0.9$ and learning rate $\alpha = 0.4$

$$Q(s,a) = Q(s,a) + \alpha[\, G_t^n - Q(s,a)]$$

$$G_t^3 = r_t + \gamma\, r_{t+1} + \gamma^2\, r_{t+2} + \gamma^3 V(\, r_{t+3})$$

3- step

$$\text{Suarez} \xrightarrow{\text{pass}} \text{Messi} \xrightarrow{\text{shoot}} \text{Scored} \xrightarrow{\text{return}} \text{Messi} \xrightarrow{\text{pass}} \text{Suarez}$$

$$G^3 = r(\text{Suarez}, \text{pass}, \text{Messi}) + \gamma\, r(\text{Messi}, \text{shoot}, \text{Scored}) + \gamma^2 r(\text{Scored}, \text{return}, \text{Messi}) + \gamma^3 V(\text{Messi}, \text{pass}, \text{Suarez})$$

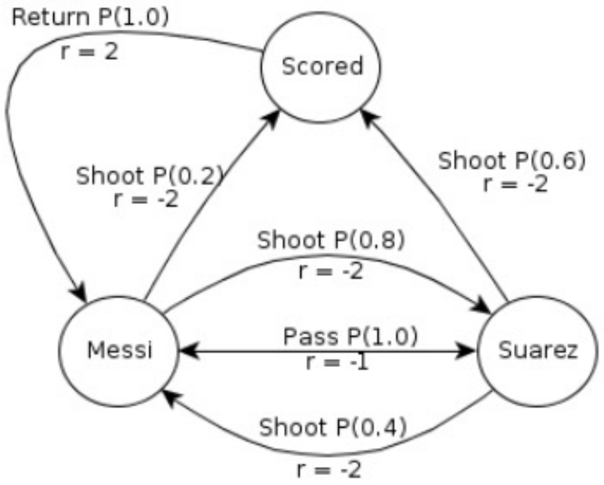
$$= -1 + 0.9 \times (-2) + (0.9)^2 \times 2 + (0.9)^3 \times (-0.4)$$

$$= -1.4716$$

Problem 3: N-step TD (SARSA)

	Pass	Shoot	Return
Messi	-0.4	-0.8	-
Suarez	-0.7	-0.2	-
Scored	-	-	1.2

The following diagram shows the transition probabilities and rewards:



Given the following trace from a historical 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! The ball is returned to Messi for kickoff. After he passes the ball to Suarez, the referee blew the final whistle. End of the game, the ball is taken by Messi to remember the match forever."

Show the 3-step SARSA update for the above feed. Discount factor $\gamma = 0.9$ and learning rate $\alpha = 0.4$

$\text{Suarez} \xrightarrow{\text{pass}} \text{Messi}$

s

$$\underline{Q(s,a)} = Q(s,a) + \alpha[G_t^n - Q(s,a)] \qquad G_t^3 = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 V(r_{t+3})$$

$$G^3 = -1.4716$$

$$Q(\text{Suarez}, \text{pass}) = \underline{Q(\text{Suarez}, \text{pass})} + \alpha[G^3 - Q(\text{Suarez}, \text{pass})]$$

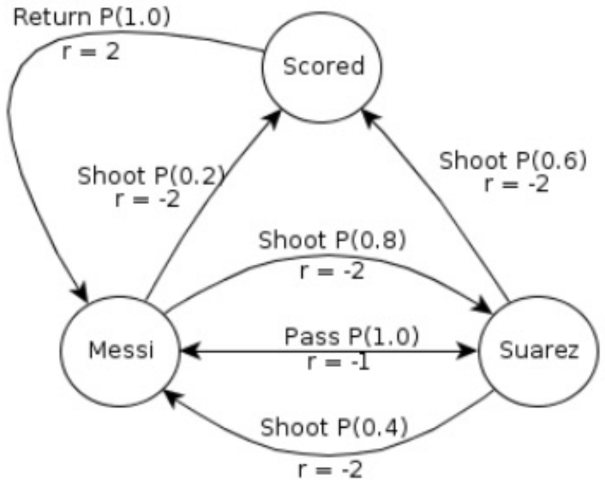
$$= -0.7 + 0.4 \times [-1.4716 - (-0.7)]$$

$$= -1.00864$$

Problem 3: N-step TD (SARSA)

	Pass	Shoot	Return
Messi	-0.4	-0.8	-
Suarez	-0.7	-0.2	-
Scored	-	-	1.2

The following diagram shows the transition probabilities and rewards:



3-step vs 1-step:

- 3-step can converge much faster than 1-step.

https://gibberblot.github.io/rl-notes/single-agent/n-step.html