Week 8: MDP and Value Iteration

COMP90054 – AI Planning for Autonomy

Key concepts

- Markov Decision Processes (MDPs)
- Solving MDPs:
  - Value Iteration

Classical Planning vs. MDPs

minimise the cost

maximise the reward

Classical Planning (6)	Markov Decision Processes (MDPs)
Set of states S	Set of states S
Initial state s_0	Initial state s_0
Action A(s)	Action A(s)
Transition function s' = f(a, s)	***Transition probabilities*** $P_a(s' \mid s)$ Non-deterministic
Goals $S_G \subseteq S$	***Reward function r(s, a, s') (positive or negative)***
Action costs c(a, s)	
	Discount factor $0 \leq \gamma \leq 1$ (prefer shorter plans over longer plans)

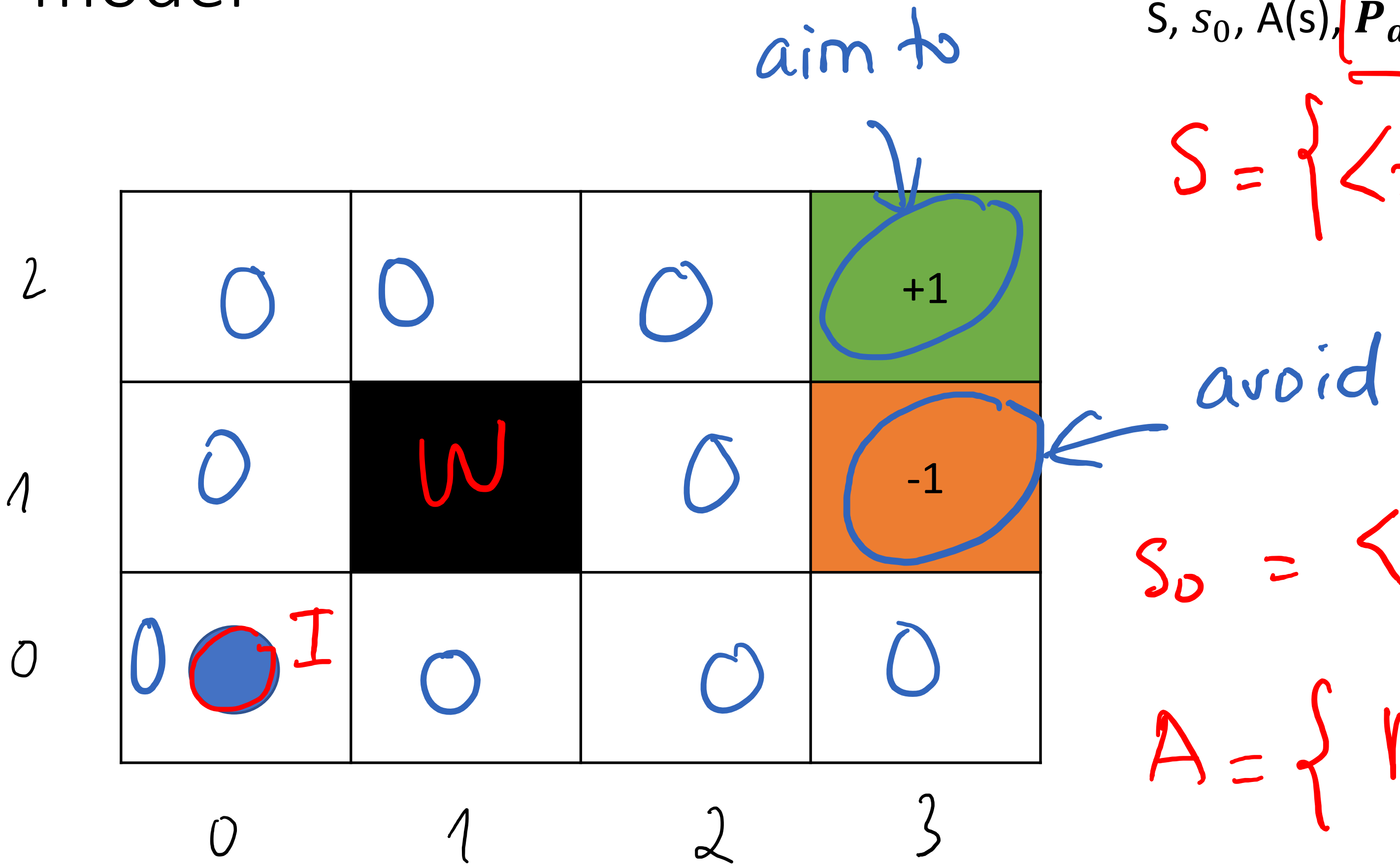
$s_0 \xrightarrow{a} s_1$

(single outcome)

$s_0 \xrightarrow{a\ (P = 0.8)} s_1$

$s_0 \xrightarrow{a\ (P = 0.2)} s_2$

Task 1: Model the Grid MDP example with a formal discounted-reward MDP model



S, s_0, A(s), $\boxed{P_a(s'|s)}$, r(s, a, s'), γ

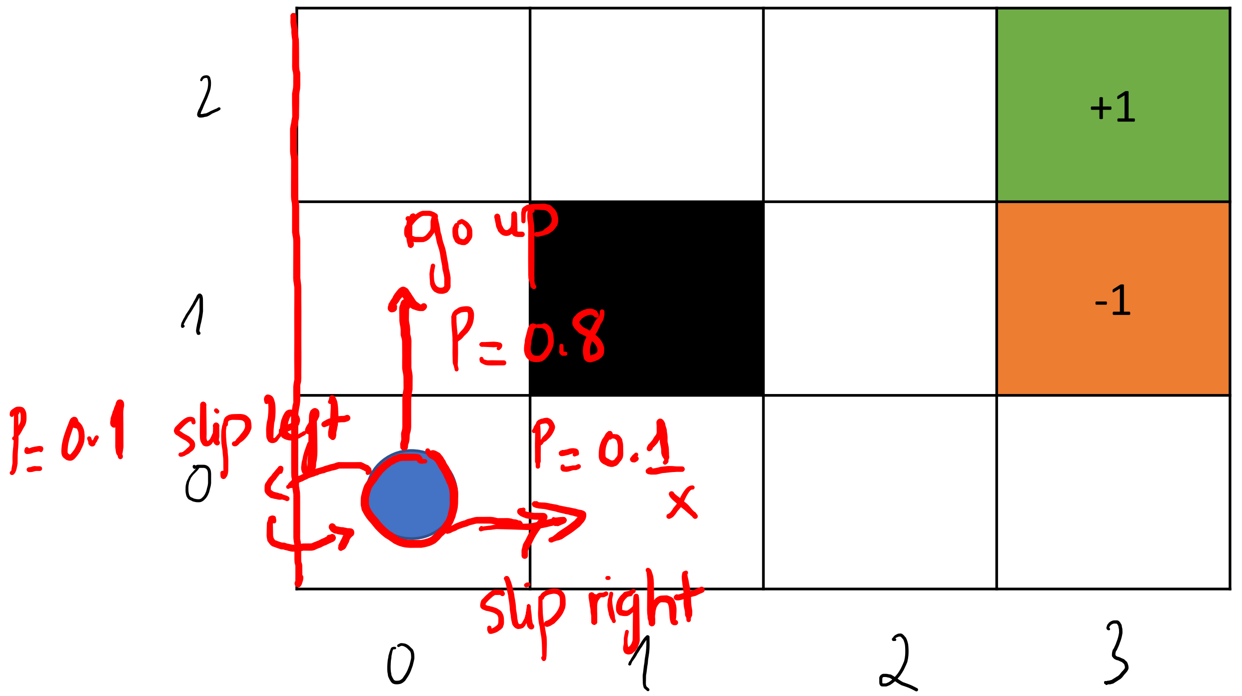
$S = \{ \langle x, y \rangle \mid x \in \{0, 1, 2, 3\}, y \in \{0, 1, 2\}, \langle 1, 1 \rangle \text{ is the wall} \}$

$s_0 = \langle 0, 0 \rangle$

$A = \{ \text{North}, \text{South}, \text{East}, \text{West} \}$

Task 1: Model the Grid MDP example with a formal discounted-reward MDP model

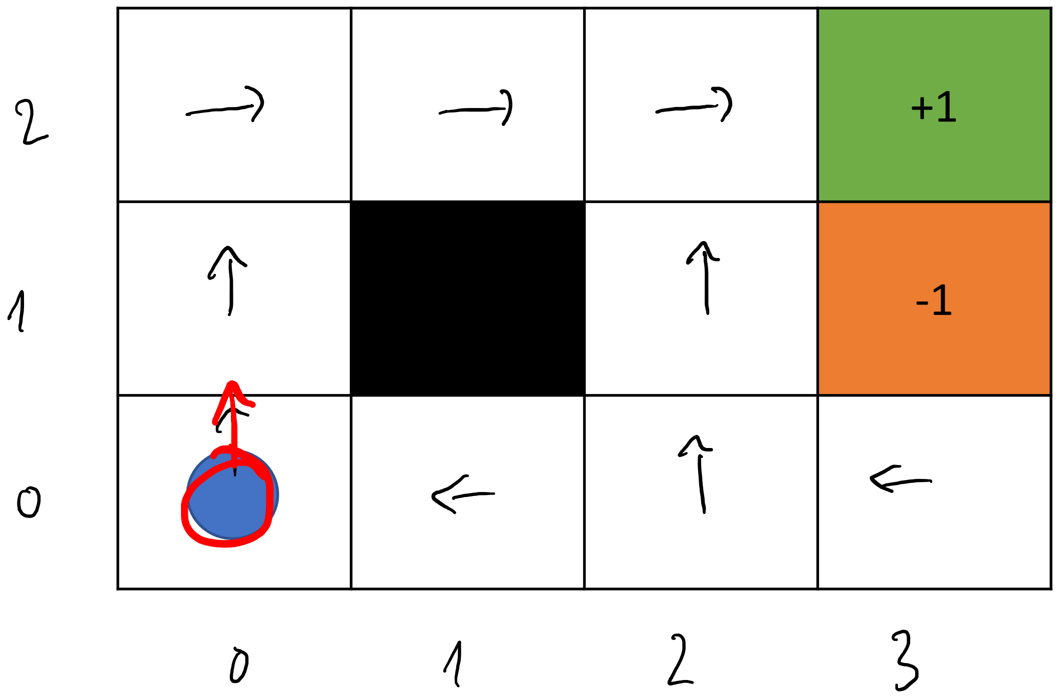
S, s_0, A(s), $\boldsymbol{P_a(s'|s)}$, ***r(s, a, s')***, γ



- Action: North:
  + P= 0.8 : go up.
  + P= 0.1: slip right
  + P= 0.1: slip left
- reward ?

Task 1: Model the Grid MDP example with a formal discounted-reward MDP model

policy.



avoid

Solving MDPs?

- maximise the reward.
- get the expected reward of an action.

Bellman equations

For discounted-reward MDPs the Bellman equation is defined recursively as:

expected reward of action a at state s

Q-value

$$Q(s,a) = \sum_{s' \in S} \underbrace{P_a(s'|s)}_{\text{the probability of action a}} \, [\underbrace{r(s,a,s')}_{\text{immediate reward}} + \underbrace{\gamma \, V(s')}_{\text{discounted future reward}}]$$

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

expected value of being in state s and acting optimally

Solving MDPs? Value Iteration

- Set V_0 to arbitrary value function; e.g., $V_0(s) = 0$ for all s.
- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\ [r(s,a,s') + \gamma\ V_i(s')]$$

Workshop Problem

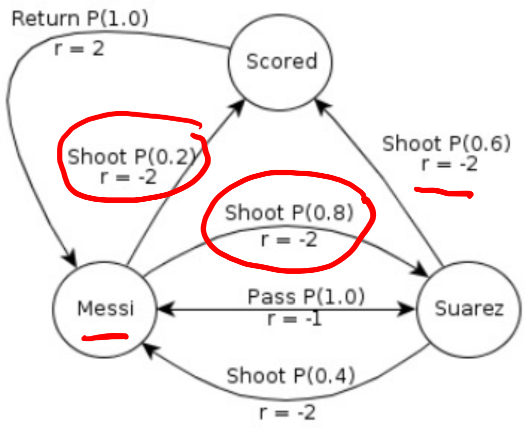
Consider two football-playing robots: Messi and Suarez.

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:



Messi $\xrightarrow[P=0.2]{\text{shoot}}$ Scored

Messi $\xrightarrow[P=0.8]{\text{shoot}}$ Suarez

$P_{shoot}(\underbrace{Suarez}_{\text{to}} \mid \underbrace{Messi}_{\text{from}}) = 0.8$

$S = \{Messi, Suarez, Scored\}$

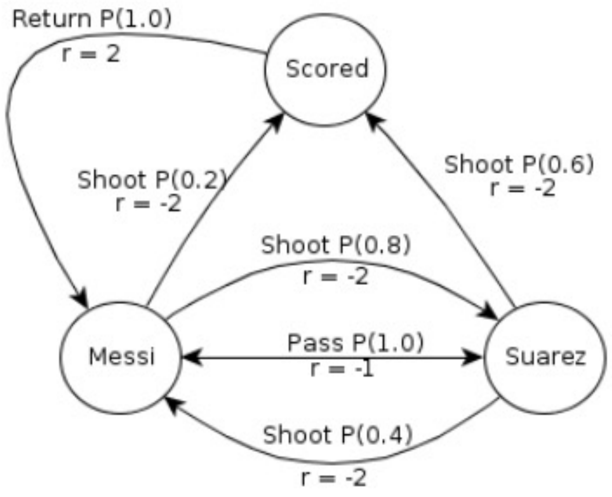
$r(Messi, shoot, Suarez) = -2$

Find the best action for all states?

Workshop Problem

	Iteration 0	Iteration 1	Iteration 2	Iteration 3
V(Messi)	0			
V(Suarez)	0			
V(Scored)	0			

The following diagram shows the transition probabilities and rewards:

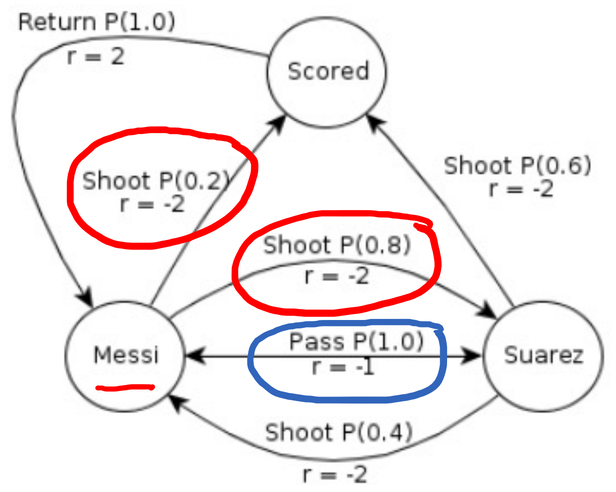


$\boldsymbol{\gamma} = \mathbf{1}$

Iteration 0: Set $V_0(s) = 0$ for all s

Workshop Problem

The following diagram shows the transition probabilities and rewards:



	Iteration 0	Iteration 1	Iteration 2	Iteration 3
V(Messi)	0	-1		
V(Suarez)	0			
V(Scored)	0			

- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\,[r(s,a,s') + \gamma\, V_i(s')]$$

s = Messi
s' = Suarez/Scored
a = shoot/pass
$\gamma = 1$

Iteration 1: $V_1(Messi)$

- shoot

$P_{shoot}(Suarez|Messi)[r(Messi, shoot, Suarez) + \gamma \times V(Suarez)]$

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

$= 0.8[-2 + 1 \times 0] + 0.2[-2 + 1 \times 0] = -2$

- pass

$P_{pass}(Suarez|Messi)[r(Messi, pass, Suarez) + \gamma \times V(Suarez)]$

$= 1 \times [-1 + 1 \times 0] = -1$

max → -1

Workshop Problem

	Iteration 0	Iteration 1	Iteration 2	Iteration 3
V(Messi)	0	-1		
V(Suarez)	0	-1		
V(Scored)	0			

- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

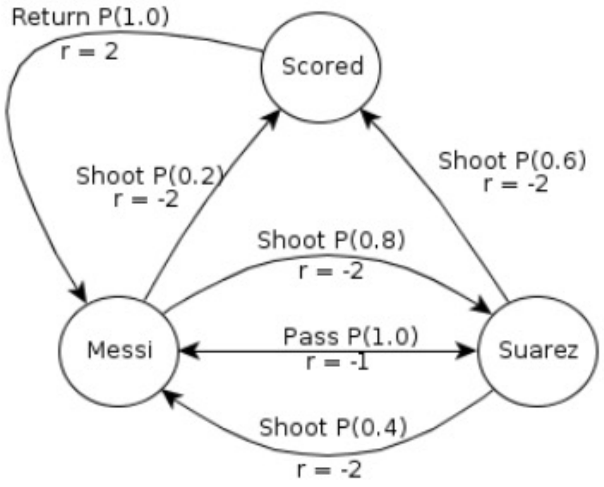
$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\,[r(s,a,s') + \gamma\, V_i(s')]$$

Iteration 1: $V_1(Suarez)$

- shoot

- pass

The following diagram shows the transition probabilities and rewards:



s = Suarez

s' = Messi/Scored

a = shoot/pass

$\boldsymbol{\gamma} = \mathbf{1}$

Workshop Problem

	Iteration 0	Iteration 1	Iteration 2	Iteration 3
V(Messi)	0	-1		
V(Suarez)	0	-1		
V(Scored)	0	2		

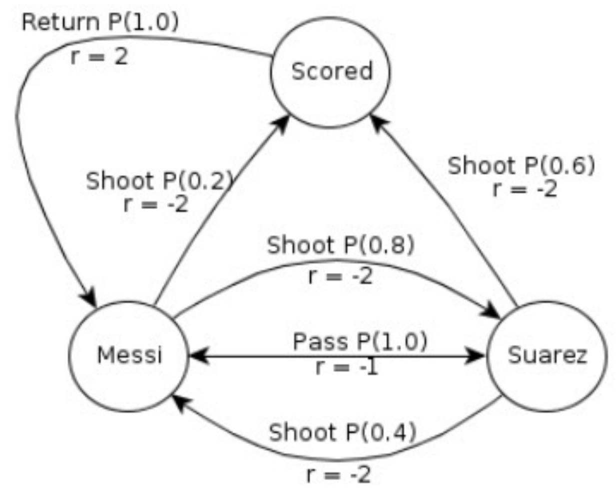
- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\,[r(s,a,s') + \gamma\, V_i(s')]$$

Iteration 1: $V_1(Scored)$

- return

The following diagram shows the transition probabilities and rewards:



s = Scored

s' = Messi

a = return

$\gamma = 1$

Workshop Problem

	Iteration 0	Iteration 1	Iteration 2	Iteration 3
V(Messi)	0	-1	-2	
V(Suarez)	0	-1		
V(Scored)	0	2		

- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

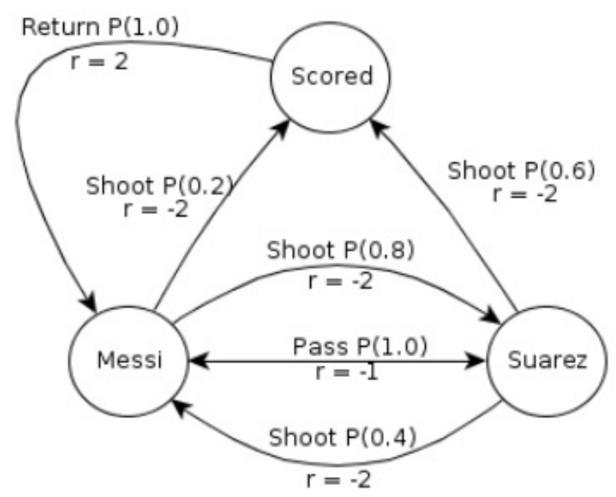
$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\,[r(s,a,s') + \gamma\, V_i(s')]$$

Iteration 2: $V_2(Messi)$

- shoot

- pass

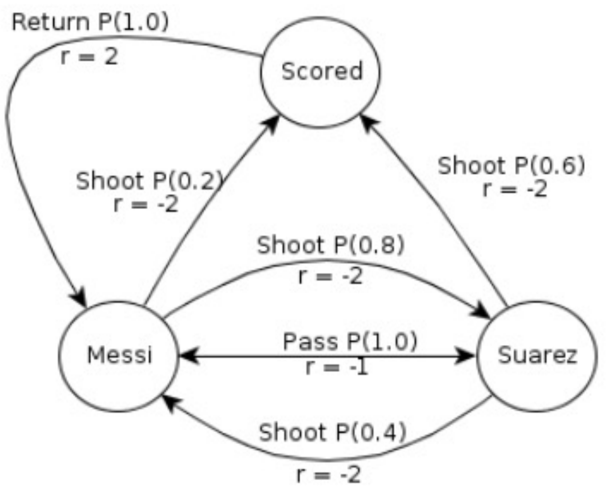
The following diagram shows the transition probabilities and rewards:



Workshop Problem

	Iteration 0	Iteration 1	Iteration 2	Iteration 3
V(Messi)	0	-1	-2	
V(Suarez)	0	-1	-1.2	
V(Scored)	0	2		

The following diagram shows the transition probabilities and rewards:



- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\, [r(s,a,s') + \gamma\, V_i(s')]$$

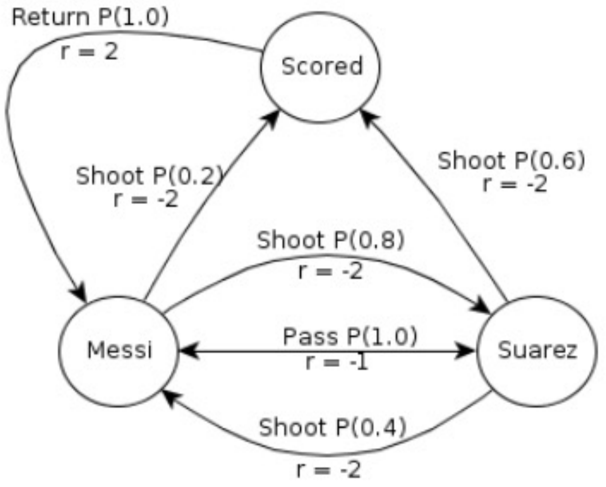
Iteration 2: $V_2(Suarez)$

- shoot

- pass

Workshop Problem

The following diagram shows the transition probabilities and rewards:



	Iteration 0	Iteration 1	Iteration 2	Iteration 3
V(Messi)	0	-1	-2	
V(Suarez)	0	-1	-1.2	
V(Scored)	0	2	1	

- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

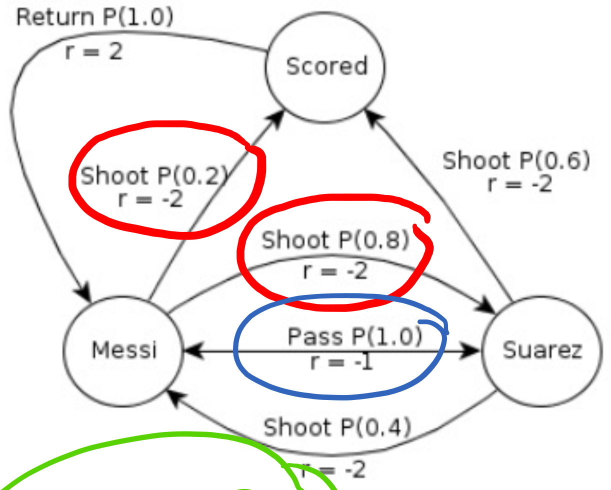
$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\,[r(s,a,s') + \gamma\, V_i(s')]$$

Iteration 2: $V_2(Scored)$

- return

Workshop Problem

The following diagram shows the transition probabilities and rewards:



	Iteration 0	Iteration 1	Iteration 2	Iteration 3
V(Messi)	0	-1	-2	-2.2
V(Suarez)	0	-1	-1.2	
V(Scored)	0	2	1	

- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\,[r(s,a,s') + \gamma\, V_i(s')]$$

-2.2

Iteration 3: $V_3(Messi)$

- shoot $P_{shoot}(Sua|M)[r(M,shoot,S) + \gamma V(Suarez)]$

$+ P_{shoot}(Scored|M)[r(M,shoot,Scored) + \gamma V(Scored)]$

$= 0.8[-2 + 1 \times (-1.2)] + 0.2[-2 + 1 \times (1)]$

$= -2.76$

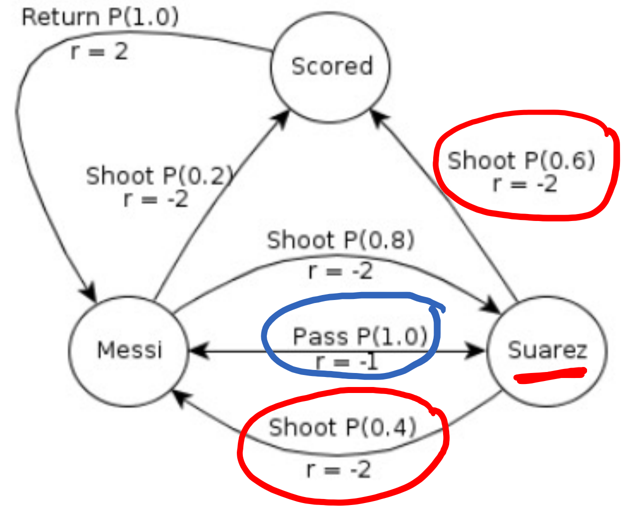
-2.2

- pass

$P_{pass}(Suarez|M)[r(M,pass,S) + \gamma V(Suarez)] = 1 \times [-1 + 1 \times (-1.2)]$

Workshop Problem

The following diagram shows the transition probabilities and rewards:



	Iteration 0	Iteration 1	Iteration 2	Iteration 3
V(Messi)	0	-1	-2	-2.2
V(Suarez)	0	-1	-1.2	-2.2?
V(Scored)	0	2	1	

- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\,[r(s,a,s') + \gamma\, V_i(s')]$$

Iteration 3: $V_3(Suarez)$

- shoot

$$P_{shoot}(Messi \mid Suarez)\,[r(Suarez, shoot, Messi) + \gamma V(Messi)]$$
$$+\ P_{shoot}(Scored \mid Suarez)\,[r(Suarez, shoot, Scored) + \gamma V(Scored)]$$
$$= 0.4[-2 + 1 \times (-2)] + 0.6[-2 + 1 \times (1)] = -2.2$$

max

- pass

$$P_{pass}(Messi \mid Suarez)\,[r(Suarez, pass, Messi) + \gamma V(Messi)]$$
$$= 1 \times [-1 + 1 \times (-2)] = -3$$

Workshop Problem

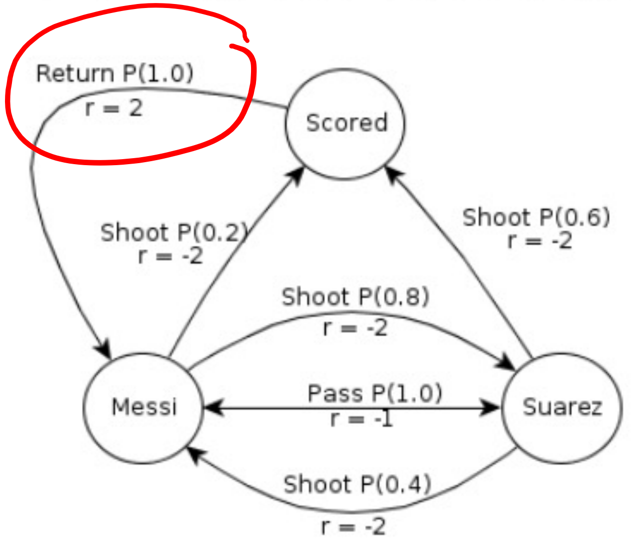
assume: stop

	Iteration 0	Iteration 1	Iteration 2	Iteration 3
V(Messi)	0	-1	-2	-2.2
V(Suarez)	0	-1	-1.2	-2.2
V(Scored)	0	2	1	0

- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\,[r(s,a,s') + \gamma\, V_i(s')]$$

The following diagram shows the transition probabilities and rewards:



Iteration 3: $V_3(Scored)$

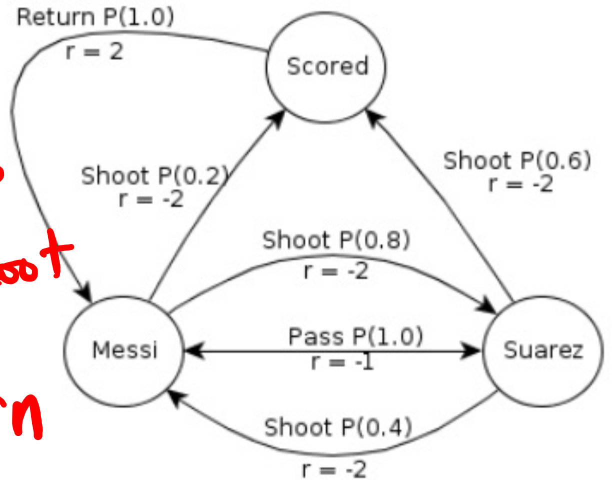
- return

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

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

Workshop Problem

The following diagram shows the transition probabilities and rewards:



	Iteration 0	Iteration 1	Iteration 2	Iteration 3
V(Messi)	0	-1	-2	-2.2
V(Suarez)	0	-1	-1.2	-2.2
V(Scored)	0	2	1	0

from Pass

from Shoot

from Return

If we only have 3 iterations, what actions did we take to maximise the reward?

Messi Pass
Suarez Shoot
Scored Return

Workshop Problem

	Iteration 0	Iteration 1	Iteration 2	Iteration 3
V(Messi)	0	-1	-2	-2.2
V(Suarez)	0	-1	-1.2	-2.2
V(Scored)	0	2	1	0

When to stop the iteration?

The iteration is stopped when Δ reaches some pre-defined threshold θ

(when the largest change in the values between iterations is "small enough")

Algorithm – Value iteration

Input: MDP $M = \langle S, s_0, A, P_a(s' \mid s), r(s, a, s') \rangle$
Output: Value function V

Set V to arbitrary value function; e.g., $V(s) = 0$ for all s

Repeat
  $\boxed{\Delta \leftarrow 0}$
  For each $s \in S$
    $\underbrace{V'(s) \leftarrow \max_{a \in A(s)} \sum_{s' \in S} P_a(s' \mid s)\,[r(s, a, s') + \gamma\, V(s')]}_{\text{Bellman equation}}$
    $\boxed{\Delta \leftarrow \max(\Delta, |V'(s) - V(s)|)}$
  $V \leftarrow V'$
Until $\Delta \leq \theta$

$\theta = 0.01$

$\Delta \leq \theta$

Messi Suarez Scored

① : $\Delta = \max(|-1-0|, |-1-0|, |2-0|) = 2$

② : $\Delta = \max(|-2+1|, |-1.2+1|, |1-2|) = 1$

⋮ $\Delta \leq \theta$