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

	Classical Planning	Markov Decision Processes (MDPs)	
1	Set of states S	Set of states S	1
2	Initial state s_0	Initial state s_0	2
3	Action A(s)	Action A(s)	3
4	***Transition function s' = f(a, s)***	***Transition probabilities*** $\boldsymbol{P_a(s'\|s)}$	Non-deterministic
5	***Goals*** $\boldsymbol{S_G \subseteq S}$	***Reward function r(s, a, s') (positive or negative)***	
6	***Action costs c(a, s)***		
		Discount factor $0 \leq \gamma \leq 1$ (prefer shorter plans over longer plans)	

$s_1 \xrightarrow{a} s_2$

(single outcome)

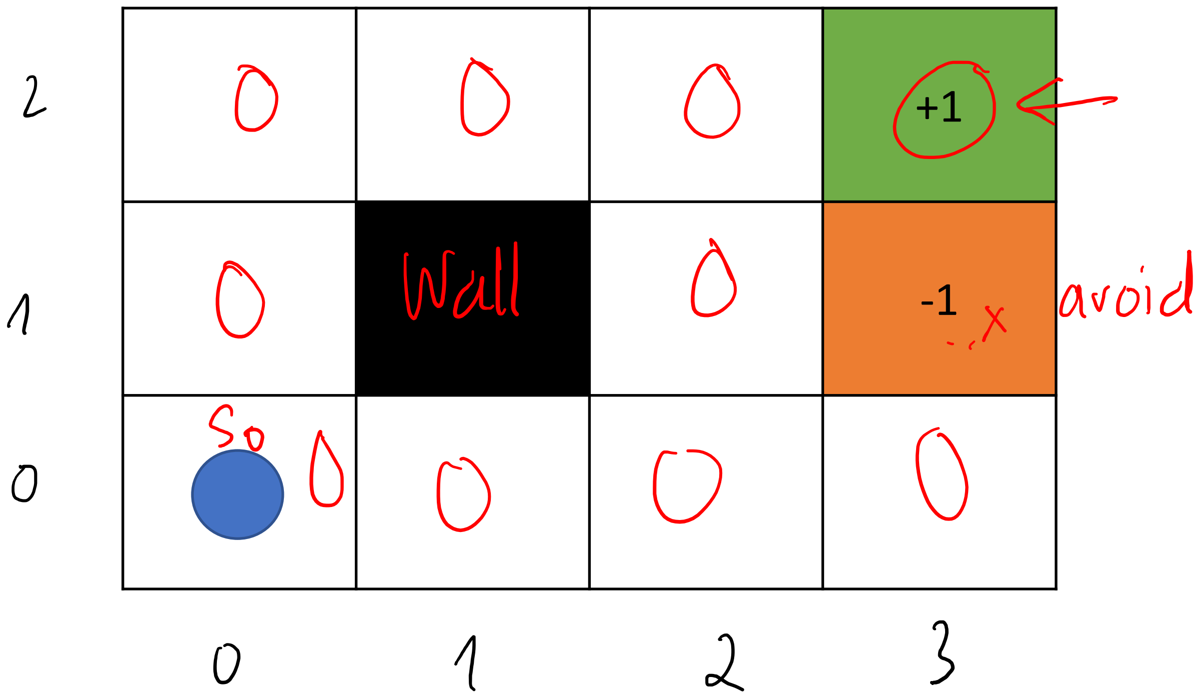
Solution: minimise the cost

$s_1 \xrightarrow{a}$ $\xrightarrow{P=0.2} s_2$, $\xrightarrow{P=0.8} s_3$

maximise the reward

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

(3x 4)



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

maximise the reward

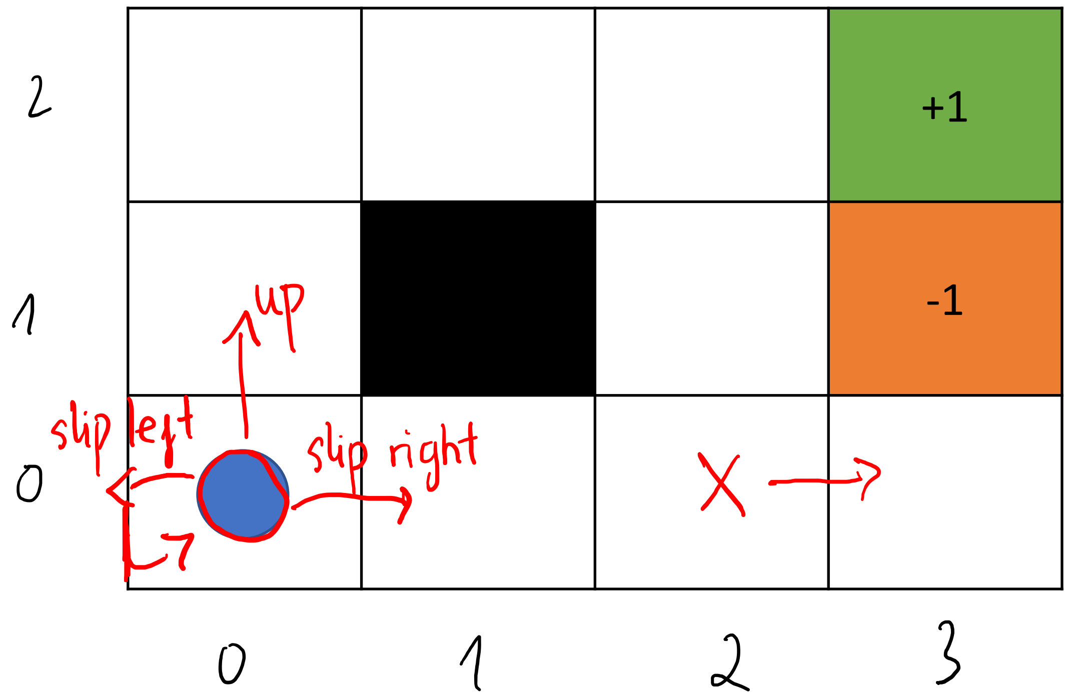
$S = \{ \langle x, y \rangle \mid x = \{0, 1, 2, 3\}, y = \{0, 1, 2\} \}$

$s_0 = \langle 0, 0 \rangle$

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

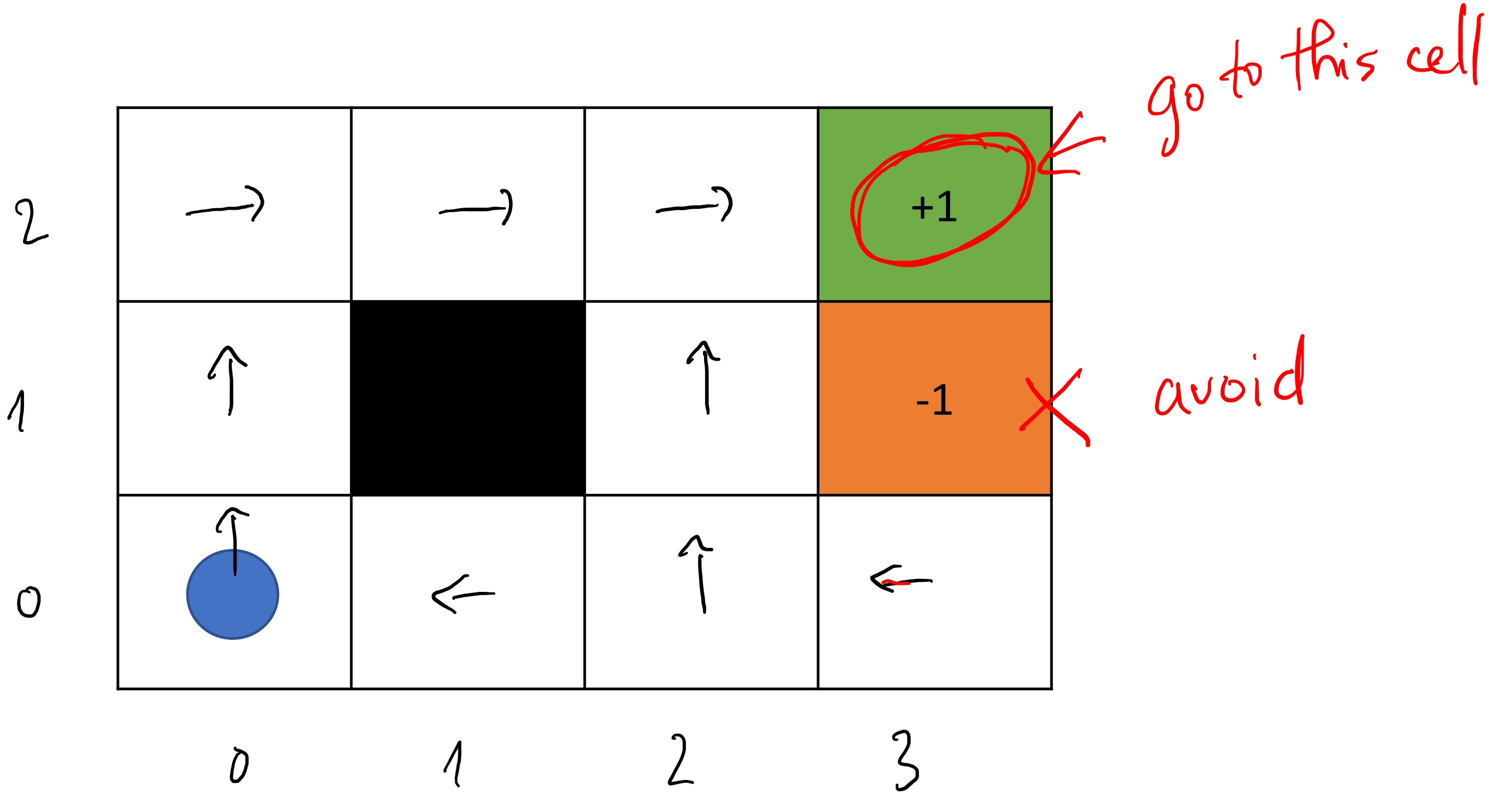
Task 1: Model the Grid MDP example with a formal discounted-reward MDP model

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



- Go North:
  + move up: P= 0.8
  + slip left: P= 0.1
  + slip right: P= 0.1
- Go East:
  + ~~go~~ move right: (P=0.8)
  + slip up: (P=0.1)
  + slip down: (P=0.1)

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



Solving MDPs?

Value Iteration.

aim to maximise the reward

→ find the expected reward of an action

Bellman equations

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

$0 \le \gamma \le 1$

expected reward of action a at state s ←

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

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

↓

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

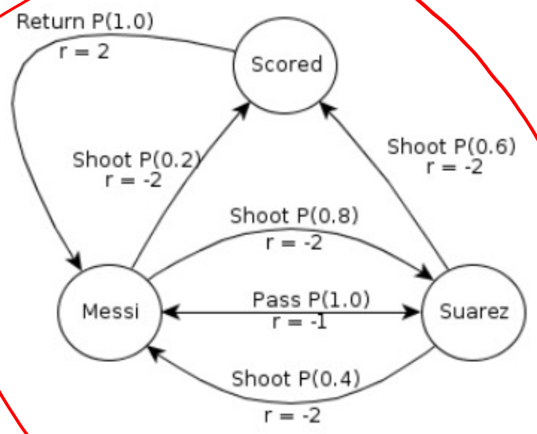
Consider two football-playing robots: Messi and Suarez.

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:



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

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

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

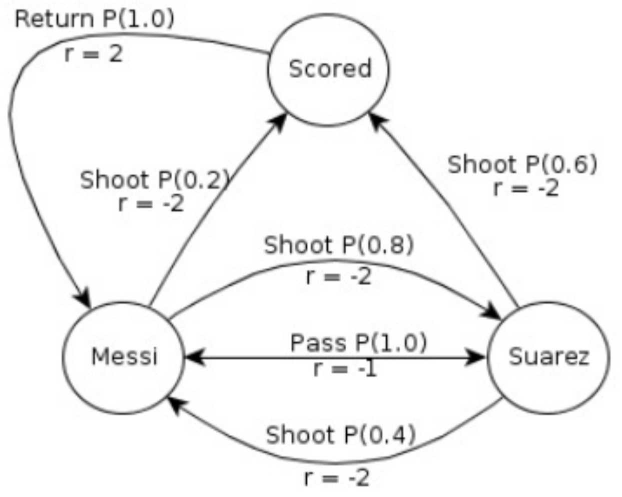
$Messi \xrightarrow{shoot\ P=0.2} Scored$

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

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:

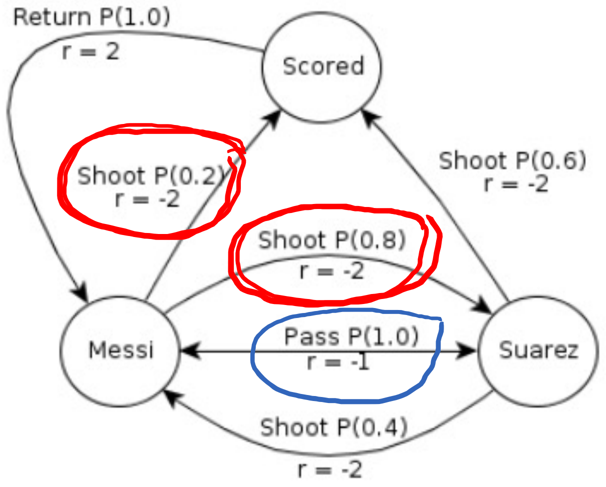


$\boldsymbol{\gamma = 1}$

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

Iteration 1: $V_1(Messi)$

- shoot

$P_{shoot}(Scored|Messi)\,[r(Messi, shoot, Scored) + \gamma \cdot V_0(Scored)]$

$+\ P_{shoot}(Suarez|Messi)\,[r(Messi, shoot, Suarez) + \gamma \cdot V_0(Suarez)]$

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

- pass

$P_{pass}(Suarez|Messi)\,[r(Messi, pass, Suarez) + \gamma \cdot V_0(Suarez)]$

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

max $\rightarrow -1$

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:

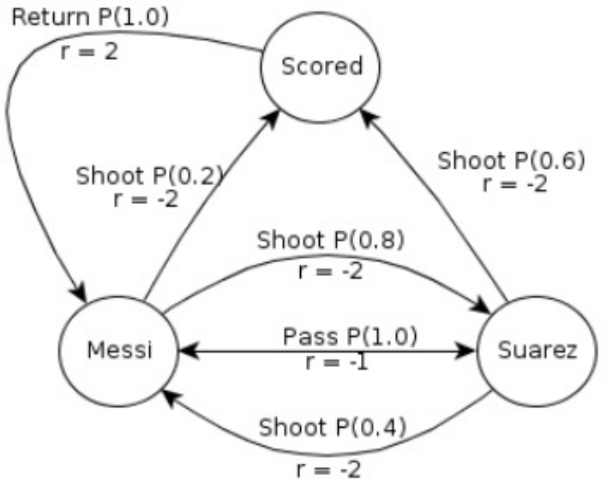
$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\ [r(s,a,s') + \gamma\ V_i(s')]$$

Iteration 1: $V_1(Suarez)$

- shoot

- pass

The following diagram shows the transition probabilities and rewards:



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

Workshop Problem

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

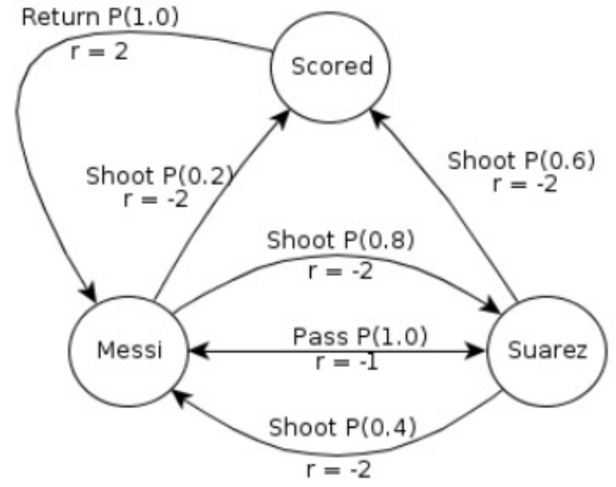
- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

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

Iteration 1: $V_1(Scored)$

- return

The following diagram shows the transition probabilities and rewards:



s = Scored

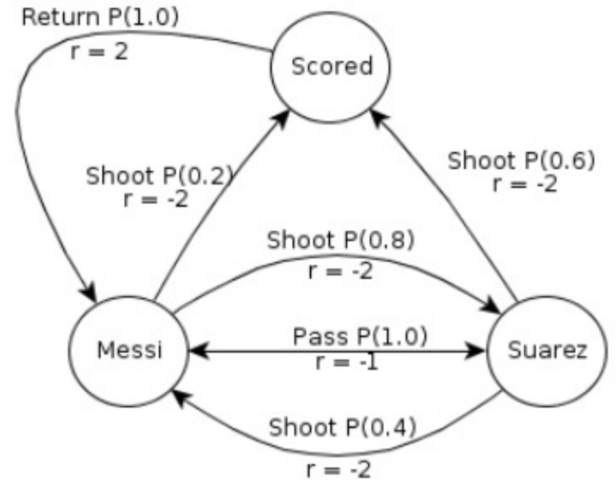
s' = Messi

a = return

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

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		
V(Scored)	0	2		

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

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

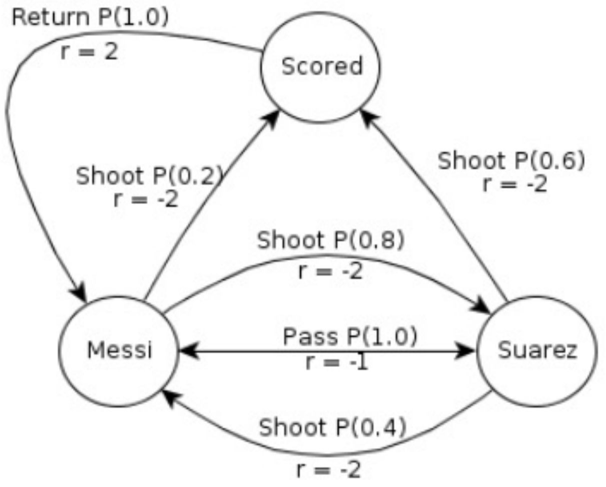
Iteration 2: $V_2(Messi)$

- 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		

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

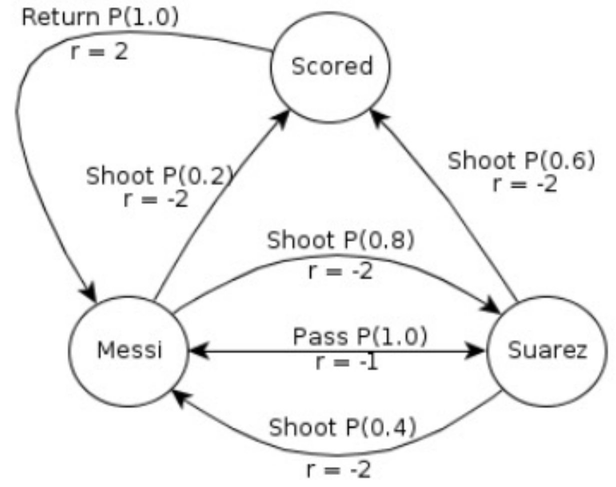
$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\ [r(s,a,s') + \gamma\ V_i(s')]$$

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:

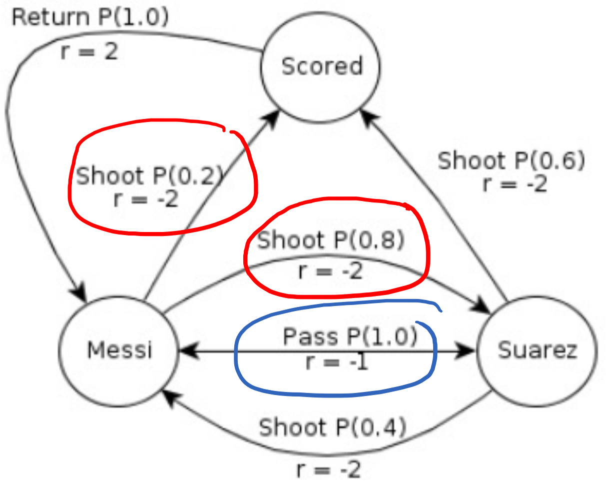
$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s)\,[r(s,a,s') + \gamma\, V_i(s')]$$

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:

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

Iteration 3: $V_3(Messi)$

- shoot

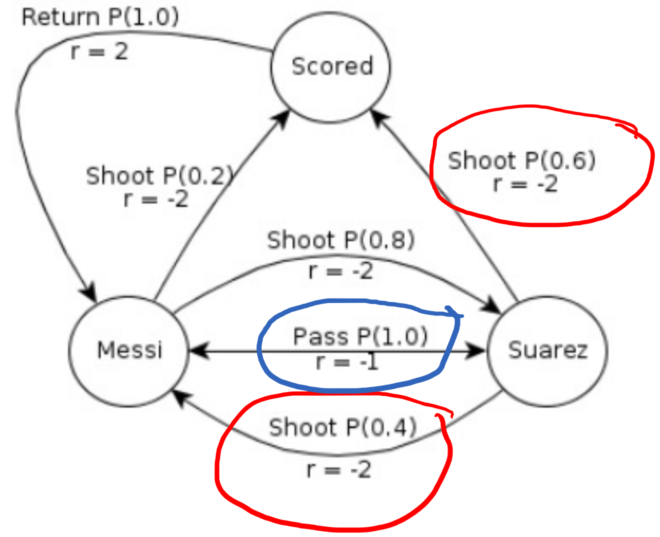
$$P_{shoot}(Scored \mid Messi)\,[r(Messi, shoot, Scored) + \gamma\, \underline{V_2(Scored)}]$$
$$+\, P_{shoot}(Suarez \mid Messi)\,[r(Messi, shoot, Suarez) + \gamma\, \underline{V_2(Suarez)}]$$
$$= 0.2\,[-2 + 1 \times 1] + 0.8\,[-2 + 1 \times (-1.2)] = -2.76$$

- pass

$$P_{pass}(Suarez \mid Messi)\,[r(Messi, pass, Suarez) + \gamma\, V_2(Suarez)]$$
$$= 1 \times [-1 + 1 \times (-1.2)] = -2.2 \rightarrow \text{max}$$

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

$P_{shoot}(Scored \mid Suarez)\,[r(Suarez, shoot, Scored) + \gamma V_2(Scored)]$

$+ P_{shoot}(Messi \mid Suarez)\,[r(Suarez, shoot, Messi) + \gamma V_2(Messi)]$

$= 0.6 \times [-2 + 1 \times 1] + 0.4 \times [-2 + 1 \times (-2)] = -2.2$

- pass

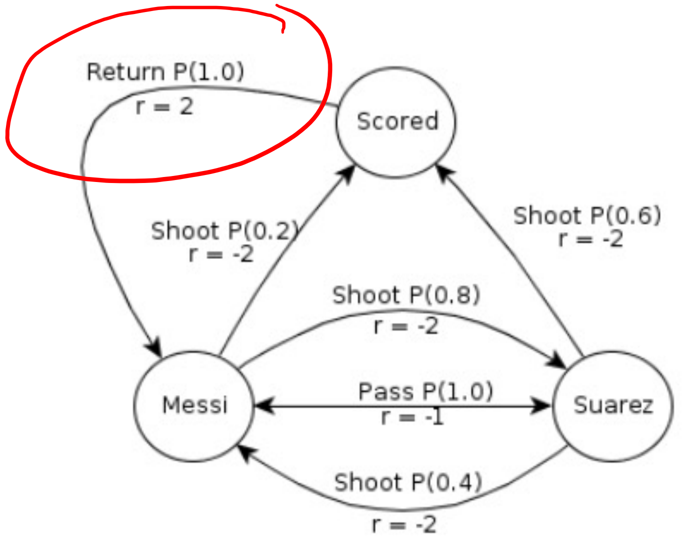
$P_{pass}(Messi \mid Suarez)\,[r(Suarez, pass, Messi) + \gamma V_2(Messi)]$

$= 1 \times [-1 + 1 \times (-2)] = -3$

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

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 3: $V_3(Scored)$

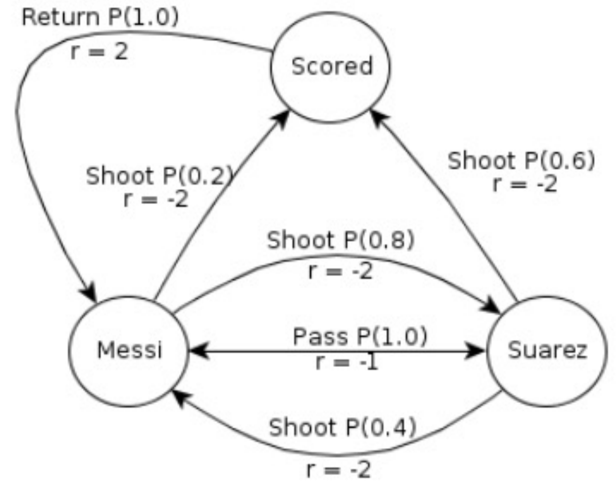
- return

$V_3(Scored) = P_{return}(Messi \mid Scored)\,[r(Scored, return, Messi) + \gamma V_2(Messi)]$

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

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

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

$\Delta \leftarrow 0$

For each $s \in S$

$$\underbrace{V'(s) \leftarrow \max_{a \in A(s)} \sum_{s' \in S} P_a(s' \mid s)\,[r(s,a,s') + \gamma\, V(s')]}_{\text{Bellman equation}}$$

$$\boxed{\Delta \leftarrow \max(\Delta, |V'(s) - V(s)|)}$$

$V \leftarrow V'$

$\boxed{\text{Until } \Delta \leq \theta}$

threshold θ

$\Delta \leq \theta$

$\boxed{\theta = 0.01}$

At iteration 1, $\Delta = \max(1, 1, 2) = 2$

2, $\Delta = \max(1, 0.2, 1) = 1$

3, $\Delta = \max(0.2, 1, 1) = 1$

until $\Delta \leq \theta$