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

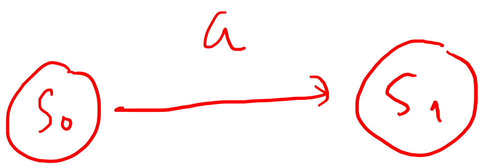
aim to maximise the reward.

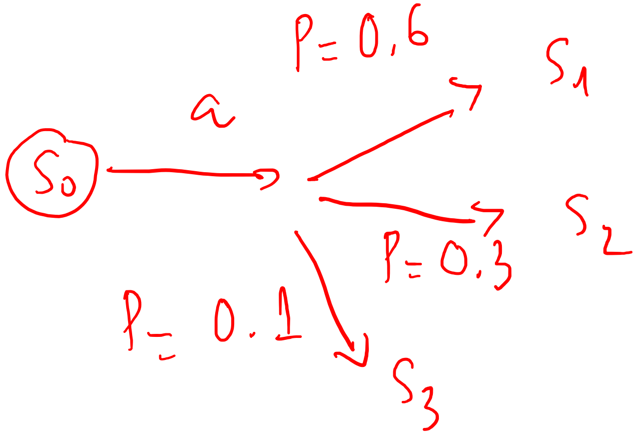
minimise the cost

deterministic

Classical Planning	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*** $\boldsymbol{P_a(s'\|s)}$
Goals $\boldsymbol{S_G \subseteq S}$	***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)

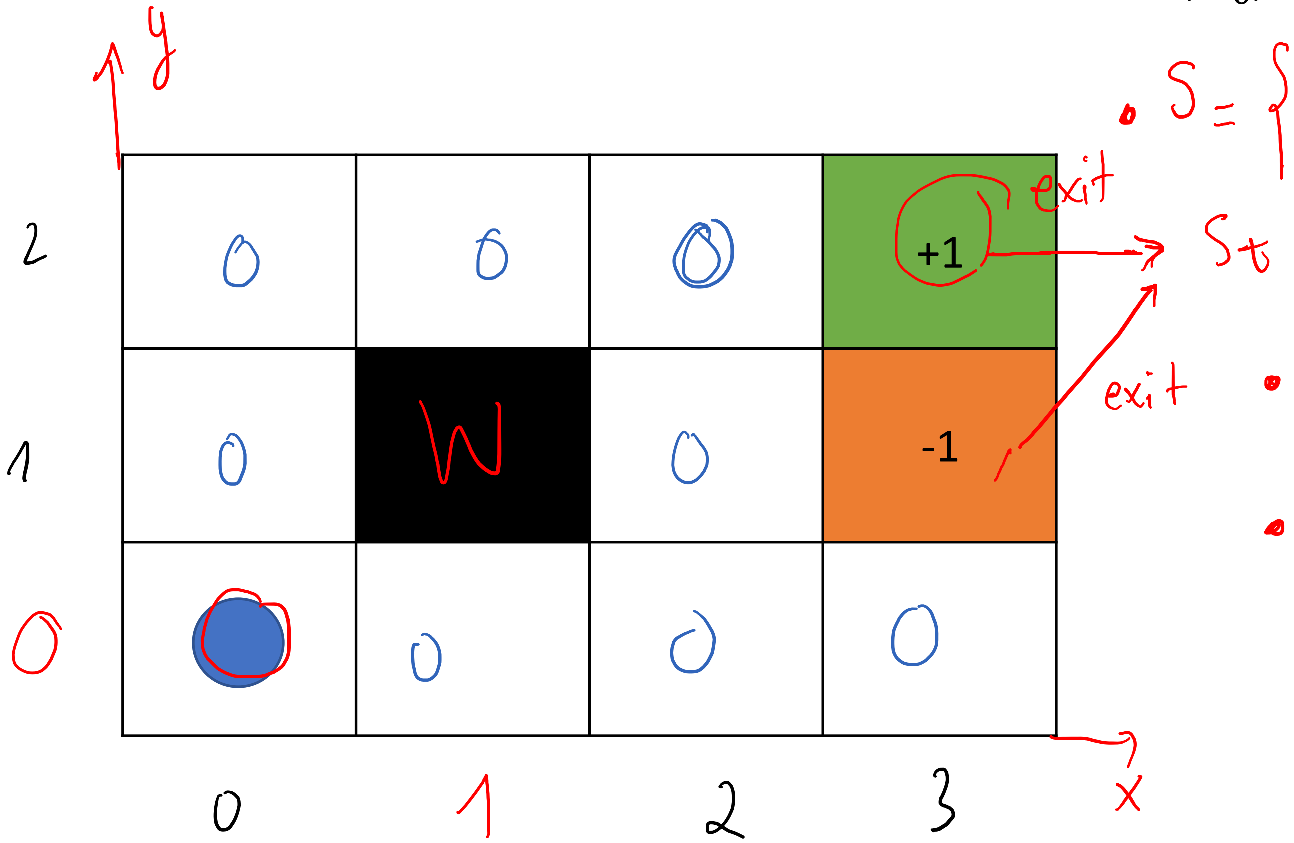
non-deterministic





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

S, s_0, A(s), $P_a(s'|s)$, **r(s, a, s')**, γ



- $S = \{\langle x,y\rangle \mid x \in \{0,..,3\}, y \in \{0,..,2\}\}$
  $\cup \{s_t\} \setminus \{\langle 1,1\rangle\}$
- $s_0 = \langle 0,0\rangle$
- $r(s,a,s') = 0$
  for $s, s' \in S$, $a \in A$.

except:

$r(\langle 3,2\rangle, exit, s_t) = 1$

$r(\langle 3,1\rangle, exit, s_t) = -1$

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

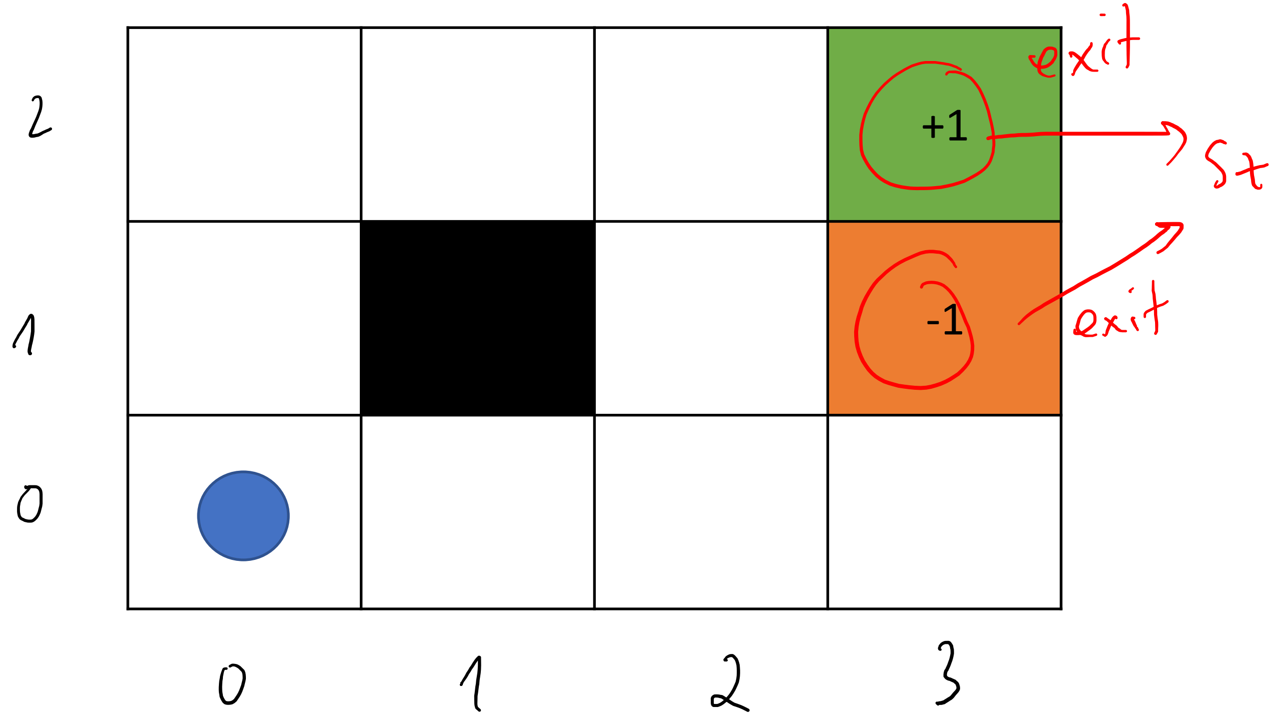
$A(s) = \{\text{North}, \text{South}, \text{East}, \text{West}\}$

for $s \in S$

except:

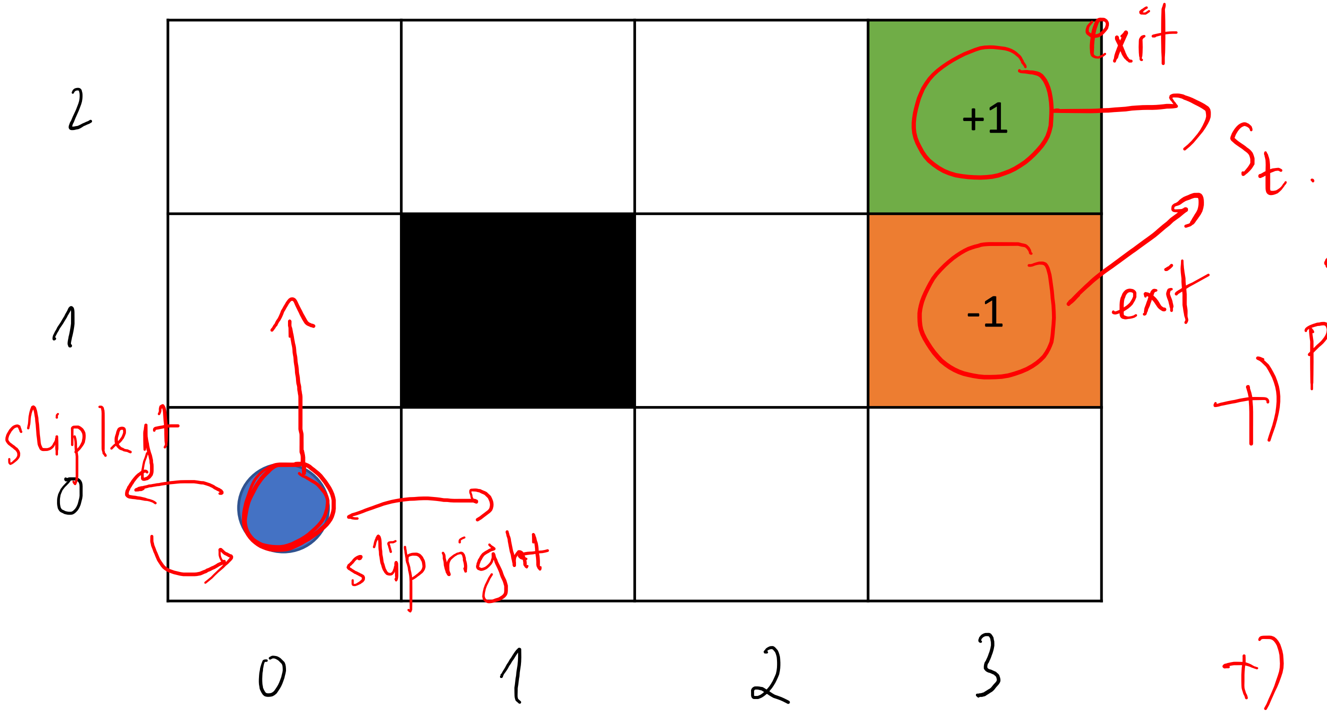
$A(\langle 3, 2 \rangle) = \{\text{exit}\}$

$A(\langle 3, 1 \rangle) = \{\text{exit}\}$



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



$P_{exit}(s_t \mid \langle 3,2\rangle) = 1$

$P_{exit}(s_t \mid \langle 3,1\rangle) = 1.$

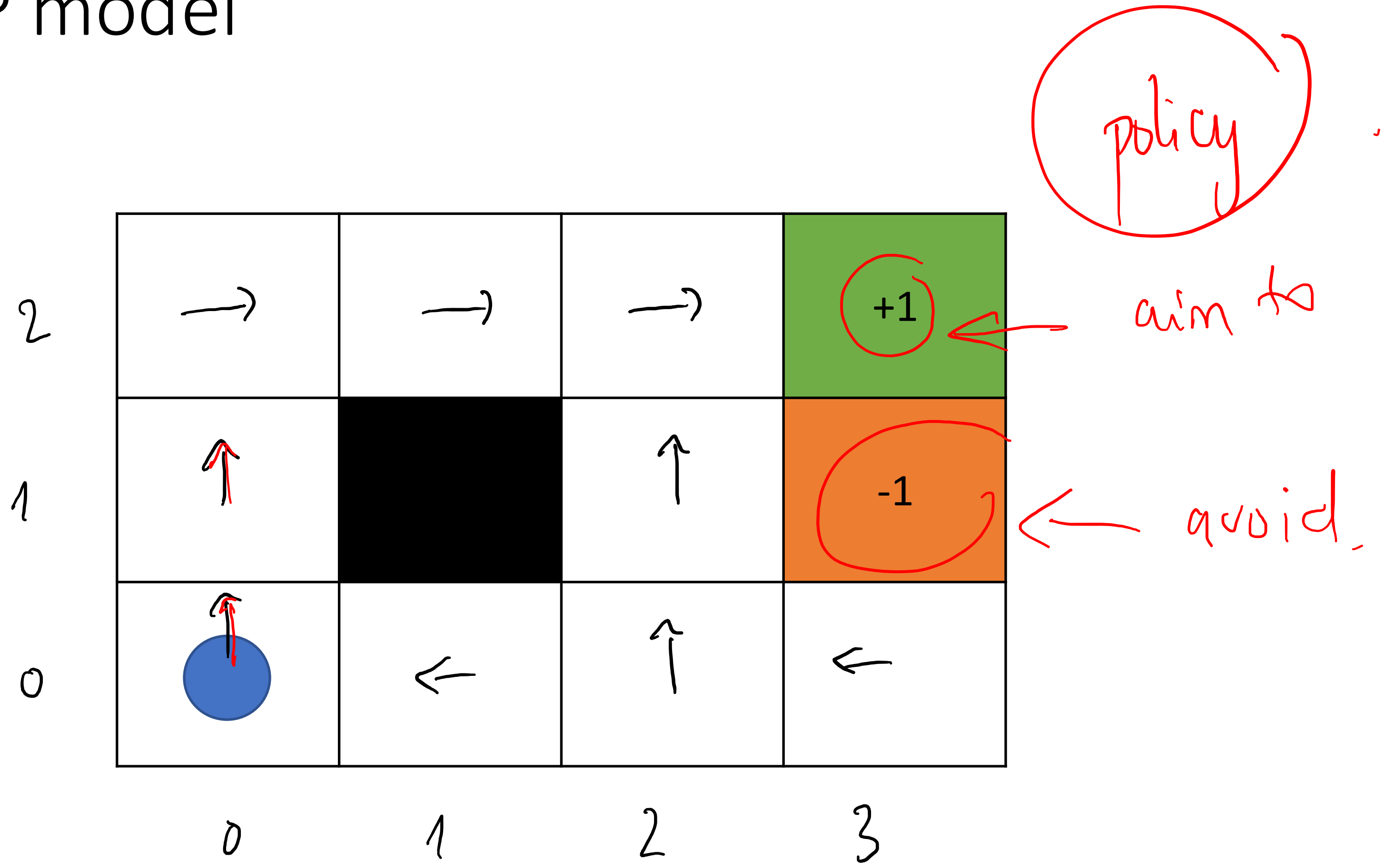
• Go North from $\langle 0,0\rangle$

+) $\underbrace{P_{North}(\langle 0,1\rangle \mid \langle 0,0\rangle)}_{\text{go up.}} = 0.8$

+) $\underbrace{P_{North}(\langle 1,0\rangle \mid \langle 0,0\rangle)}_{\text{slip right}} = 0.1$

+) $\underbrace{P_{North}(\langle 0,0\rangle \mid \langle 0,0\rangle)}_{\text{slip left}} = 0.1$

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



Solving MDPs?

Bellman equations

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

expected reward of action a at state s

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

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

expected value of being in state s and acting optimally

maximise the Q-value.

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

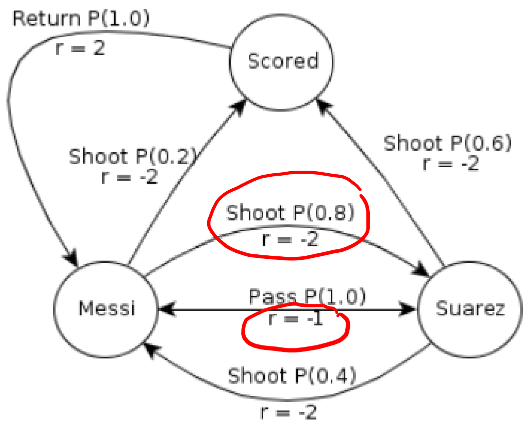
Consider two football-playing robots: Messi and Suarez.

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:



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

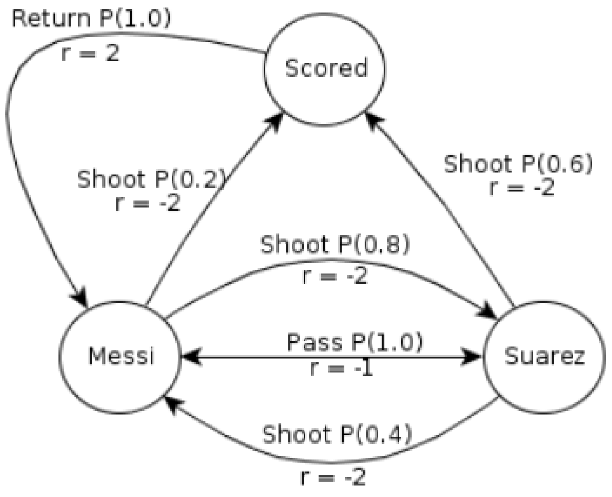
$$r(Messi, pass, Suarez) = -1.$$

Workshop Problem

S = { Messi, Suarez, Scored }

	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:



$\gamma = 1$

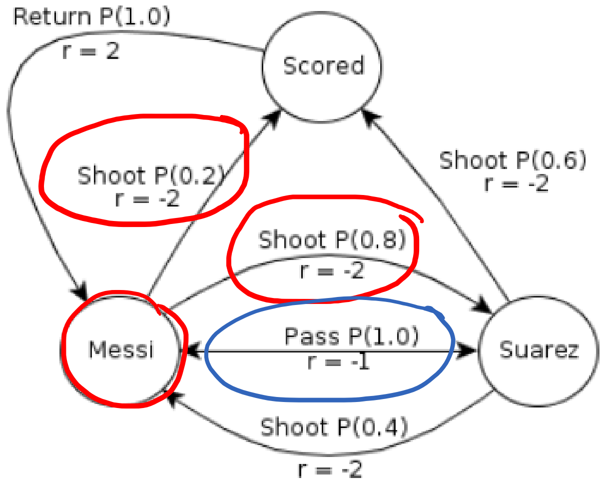
Iteration 0: Set $V_0(s) = 0$ for all s

Init

Workshop Problem

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

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

s = Messi

s' = Suarez/Scored

a = shoot/pass

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

Iteration 1: $V_1(Messi)$

- shoot

$$Q(\text{Messi, shoot}) = P_{\text{shoot}}(\text{Suarez}|\text{Messi})\,[r(\text{Messi, shoot, Suarez}) + \gamma V(\text{Suarez})] + P_{\text{shoot}}(\text{Scored}|\text{Messi})\,[r(\text{Messi, shoot, Scored}) + \gamma V(\text{Scored})]$$

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

- pass

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

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

$\rightarrow$ max

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:

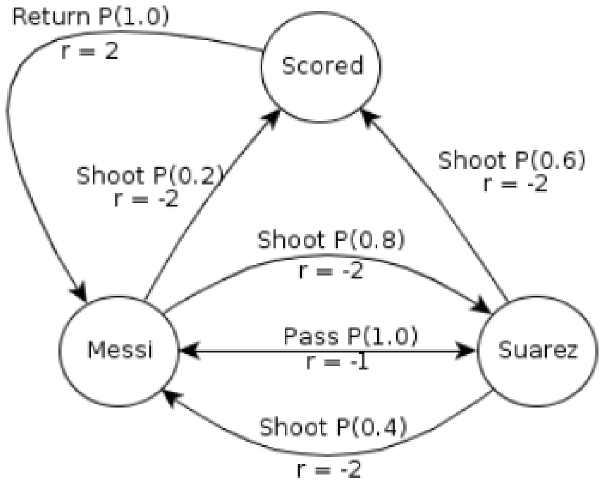
$$V_{i+1}(s) := \max_{a \in A(s)} \sum_{s' \in S} P_a(s'|s) \, [r(s,a,s') + \gamma \, V_i(s')]$$

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		

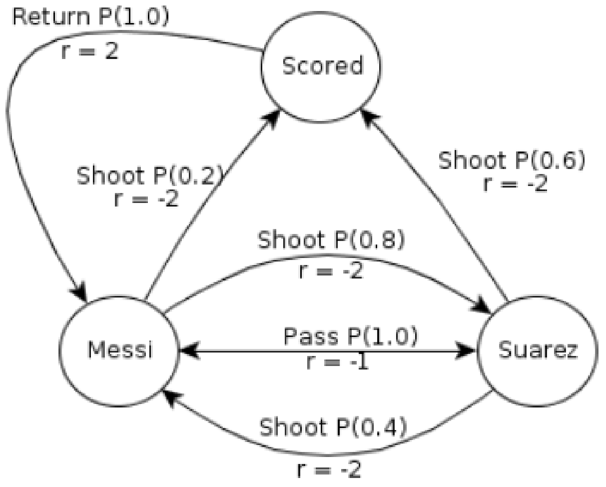
- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

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

Iteration 1: $V_1(Scored)$

- return

The following diagram shows the transition probabilities and rewards:



s = Scored

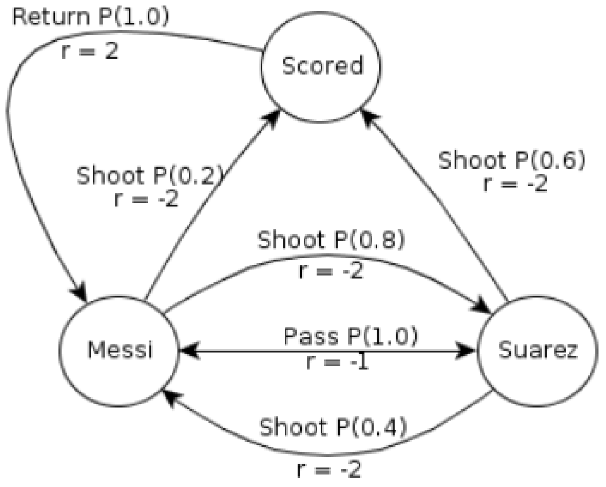
s' = Messi

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:

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

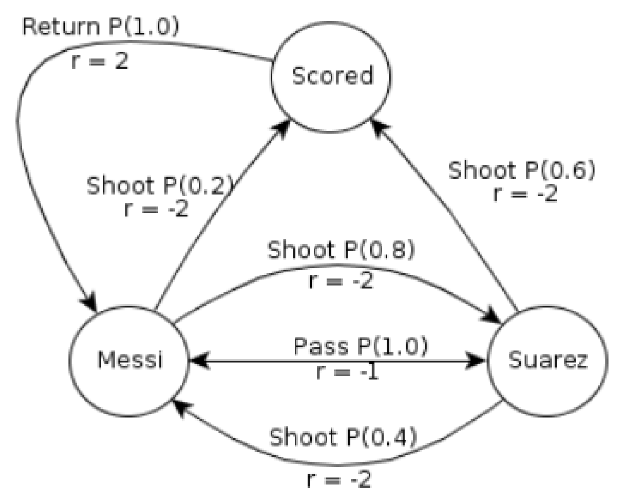
Iteration 2: $V_2(Messi)$

- 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

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

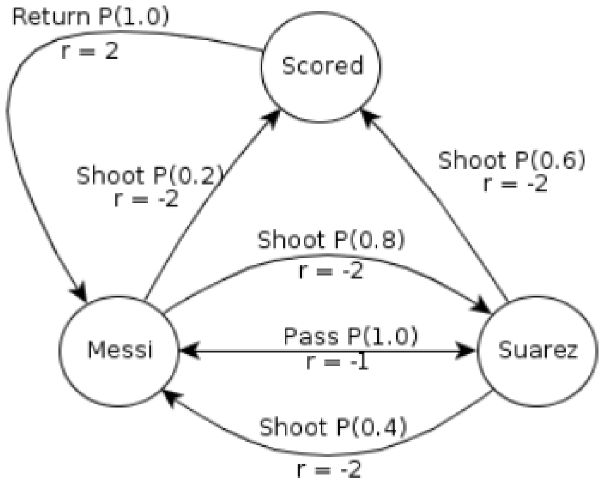
- Set V_{i+1} to result of Bellman's **right hand side** using V_i in place of V:

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

Iteration 2: $V_2(Scored)$

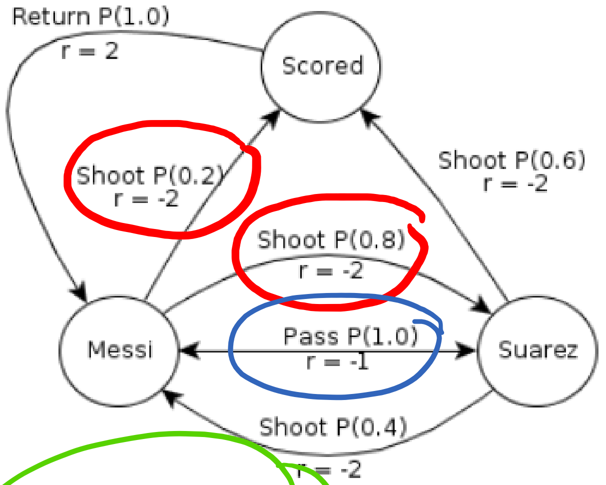
- return

The following diagram shows the transition probabilities and rewards:



Workshop Problem

The following diagram shows the transition probabilities and rewards:



-2.2

	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

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

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

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

$= -2.76$

- pass

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

-2.2

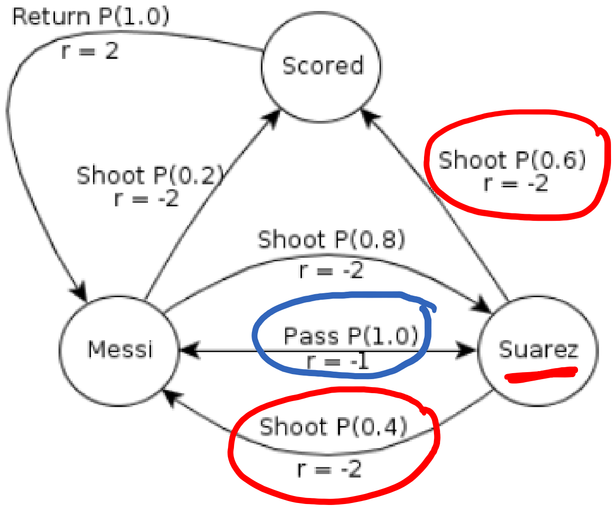
Workshop Problem

	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	

- 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: $V_3(Suarez)$

max

- shoot

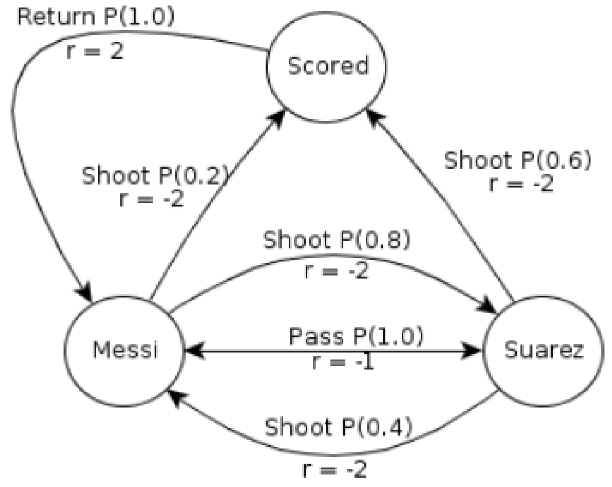
$$P_{shoot}(Messi \mid Suarez)\,[r(Suarez, shoot, Messi) + \gamma V(Messi)]$$
$$+ P_{shoot}(Scored \mid Suarez)\,[r(Suarez, shoot, Scored) + \gamma V(Scored)]$$
$$= 0.4\,[-2 + 1 \times (-2)] + 0.6\,[-2 + 1 \times (1)] = -2.2$$

- pass

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

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

- 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: $V_3(Scored)$

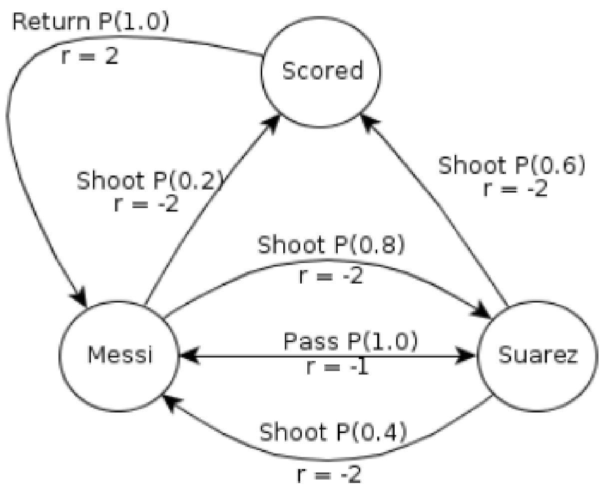
- return

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

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

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

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

Messi Pass
Suarez Shoot
Scored Return

actions have the max Q-value.

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 $s \in S$

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

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

$V \leftarrow V'$

Until $\Delta \leq \theta$

$\theta = 0.001$.

- Iter 1: $\Delta = \max(|-1-0|, |-1-0|, |2-0|) = 2$.
- Iter 2: $\Delta = \max(|-2+1|, |-1.2+1|, |1-2|) = 1$.

⋮

stop when $\Delta \leq \theta$.