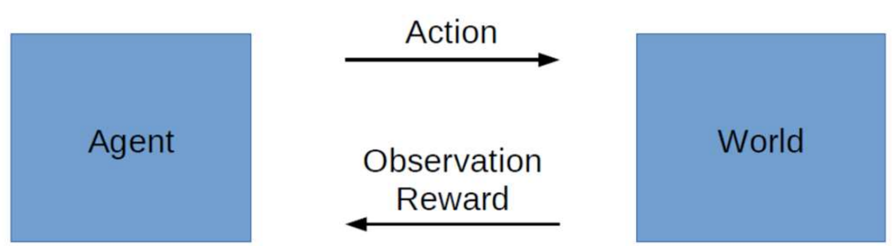
RL Setting

The RL Setting



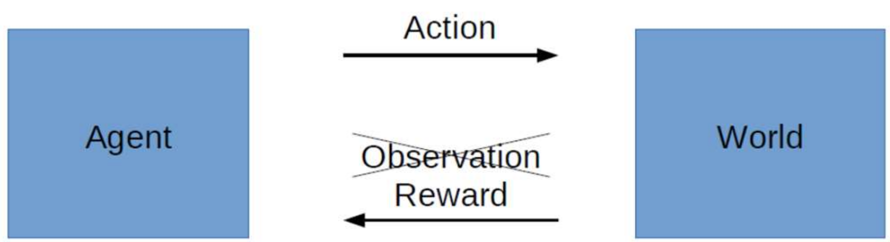
On a single time step, agent does the following:

1. observe some information
2. select an action to execute
3. take note of any reward

Goal of agent: select actions that maximize cumulative reward in the long run

Markov Decision Process-MDP

Let's turn this into an MDP



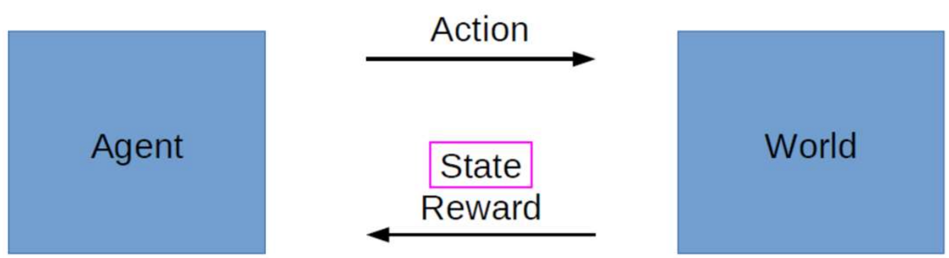
On a single time step, agent does the following:

1. observe some information
2. select an action to execute
3. take note of any reward

Goal of agent: select actions that maximize cumulative reward in the long run

MDP

Let's turn this into an MDP



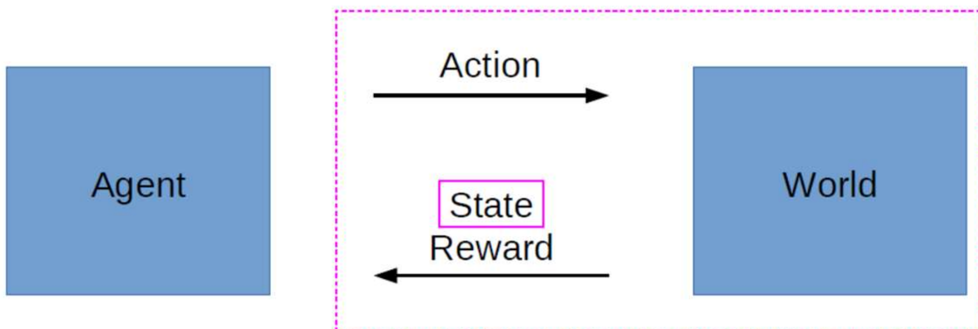
On a single time step, agent does the following:

1. observe state
2. select an action to execute
3. take note of any reward

Goal of agent: select actions that maximize cumulative reward in the long run

MDP

Let's turn this into an MDP



This part is the MDP

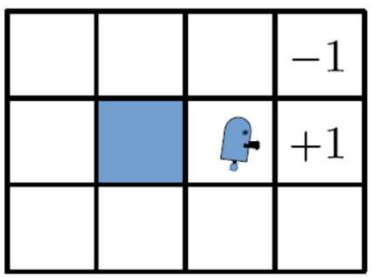
On a single time step, agent does the following:

1. observe state
2. select an action to execute
3. take note of any reward

Goal of agent: select actions that maximize cumulative reward in the long run

Example – Grid World

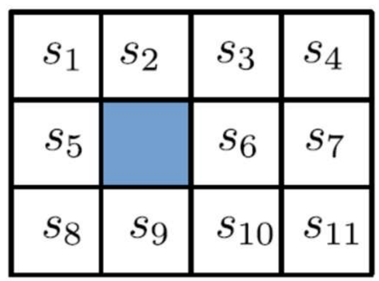
Example: Grid world



Grid world:
– agent lives on grid
– always occupies a single cell
– can move left, right, up, down
– gets zero reward unless in “+1” or “-1” cells

Example

States and actions

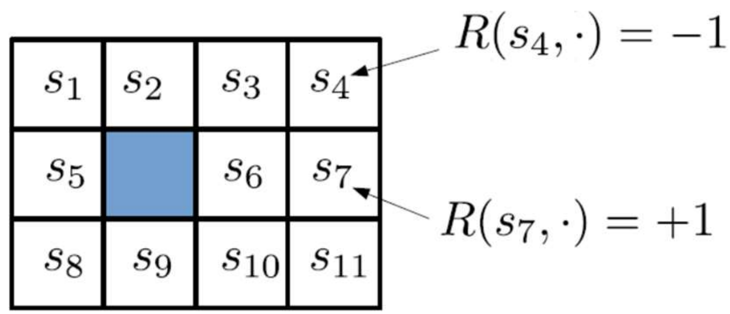


State set: $S = \{s_1, \ldots, s_{11}\}$

Action set: $A = \{left, right, up, down\}$

Example

Reward function



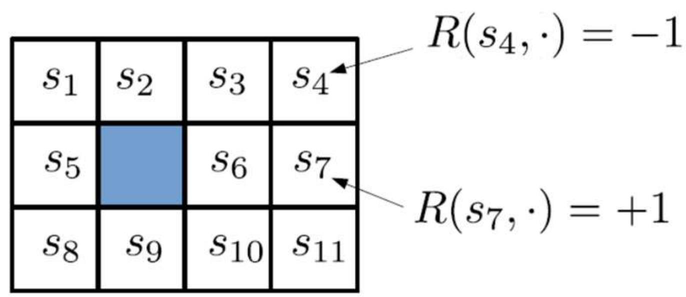
Reward function: $R(s_4, \cdot) = -1$

$R(s_7, \cdot) = +1$

Otherwise: $R(s, \cdot) = 0$

Example

Reward function



Reward function: $R(s_4, \cdot) = -1$

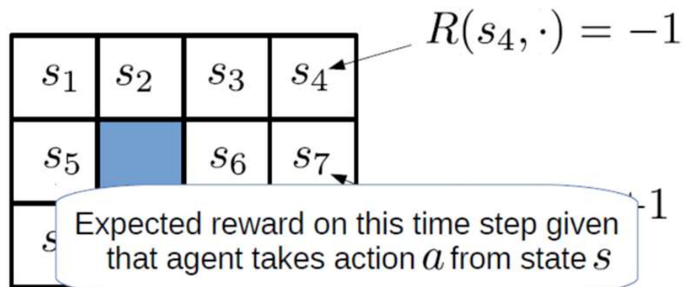
$R(s_7, \cdot) = +1$

Otherwise: $R(s, \cdot) = 0$

In general: $R(s, a) = \mathbb{E}[r_{t+1} | s_t = s, a_t = a]$

Example

Reward function



Expected reward on this time step given that agent takes action a from state s

Reward function: $R(s_4$ -1

$R(s_7,$ $+1$

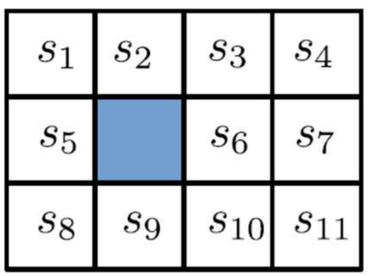
Otherwise: $R(s, \cdot) = 0$

In general: $R(s, a) = \mathbb{E}[r_{t+1} | s_t = s, a_t = a]$

Example

Transition function

Transition model: $P(s_{t+1} = s' | s_t = s, a_t = a)$



For example:

$$P(s_{t+1} = s_4 | s_t = s_3, a_t = left) = 0.1$$
$$P(s_{t+1} = s_2 | s_t = s_3, a_t = left) = 0.7$$
$$P(s_{t+1} = s_6 | s_t = s_3, a_t = left) = 0.1$$
$$P(s_{t+1} = s_3 | s_t = s_3, a_t = left) = 0.1$$

– This entire probability distribution can be written as a table over state, action, next state.

s_t	a_t	s_{t+1}	probability of this transition

MDP

Definition of an MDP

s_1	s_2	s_3	s_4
s_5		s_6	s_7
s_8	s_9	s_{10}	s_{11}

An MDP is a tuple: $\mathcal{M} = \langle S, A, R, P \rangle$

where

State set: $S = \{s_1, \ldots, s_{11}\}$

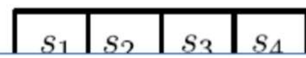
Action set: $A = \{left, right, up, down\}$

Reward function: $R(s, a) = \mathbb{E}[r_{t+1} | s_t = s, a_t = a]$

Transition model: $P(s_{t+1} = s' | s_t = s, a_t = a)$

MDP

Definition of an MDP



Why is it called a *Markov* decision process?

Because we're making the following assumption:

$$P(s_{t+1}|s_t, a_t) = P(s_{t+1}|s_t, a_t, \ldots, s_1, a_1)$$

– this is called the "Markov" assumption

State set: $S = \{s_1, \ldots, s_{11}\}$

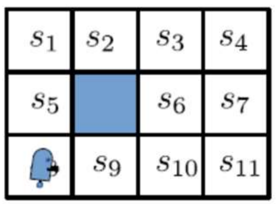
Action set: $A = \{left, right, up, down\}$

Reward function: $R(s, a) = \mathbb{E}[r_{t+1}|s_t = s, a_t = a]$

Transition model: $P(s_{t+1} = s'|s_t = s, a_t = a)$

MDP

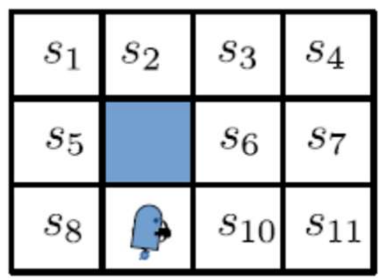
The Markov Assumption

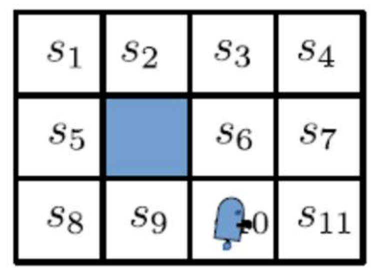


Suppose agent starts in s_8 and follows this path: s_8, s_9, s_{10}

Notice that probability of arriving in s_{11} if agent executes right action does not depend on path taken to get to s_{10} :

$$P(s_{11}|s_{10}, right) = P(s_{11}|s_{10}, right, s_9, right, s_8, right)$$





Different Models

	No Agents	Single Agent	Multiple Agents
State Known	Markov Chain	Markov Decision Process (MDP)	Markov Game (a.k.a. Stochastic Game)
State Observed Indirectly	Hidden Markov Model (HMM)	Partially-Observable Markov Decision Process (POMDP)	Partially-Observable Stochastic Game (POSG)

MDP/POMDP/Dec-POMDP

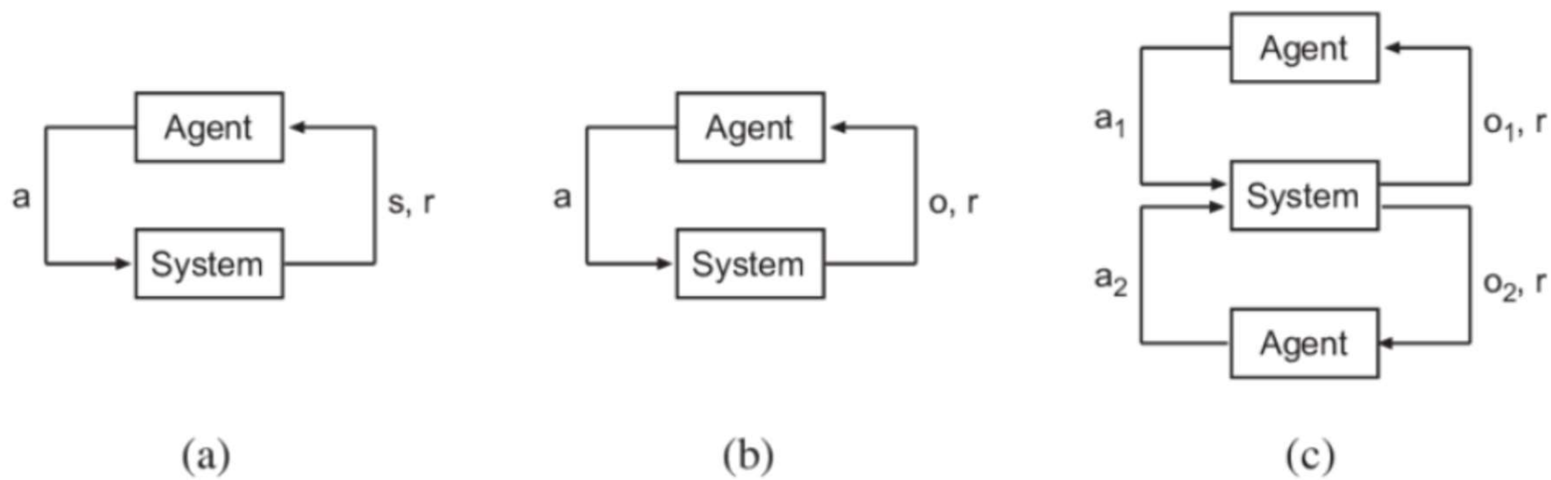
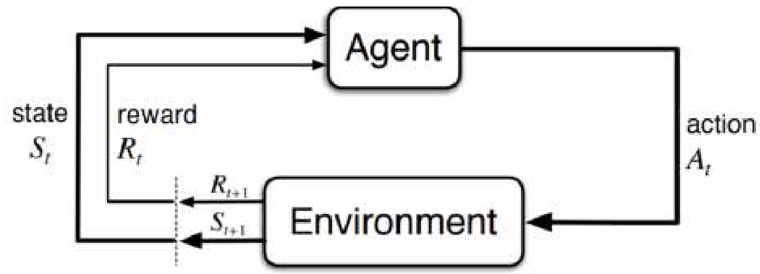


Figure: (a) Markov decision process (MDP) (b) Partially observable Markov decision process (POMDP) (c) Decentralized partially observable Markov decision process with two agents (Dec-POMDP)

Definition

The agent is acting in an environment. How the environment reacts to certain actions is defined by a model which we may or may not know. The agent can stay in one of many states ($s \in S$) of the environment, and choose to take one of many actions ($a \in A$) to switch from one state to another. Which state the agent will arrive in is decided by the transition probabilities between states $P(s'|s, a)$. Once an action is taken, the environment delivers a reward ($r \in R$) as a feedback.

Finite Markov Decision Processes(MDP)



At each step t the agent:

- Receives state S_t / observation O_t and reward R_t
- Executes action A_t

The environment:

- Receives action A_t
- Emits state S_{t+1} / observation O_{t+1} and reward R_{t+1}

Markov property:

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, S_2, \ldots, S_t]$$

"The future is independent of the past given the present"

Daily life trajectory:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots, S_T$$

Markov Property