Markov Decision Process

~ Jiya Bhagat

What is MDP ?

Markov Decision Process(MDP) : "Reinforcement learning deals with knowledge based on current and its future prediction of next state with optimal way", The direction of this movement towards solution is proposed by agent under Markov Decision Process.

It is a discrete-time stochastic control process.

MDP is set of (s, a, T, R, γ)

Here

- Agent (A): Agent has always a policy [ ] according to which it always take Actions.
- Action (a): The work or activity done by Agent.
- State (s): With every Action the Agent moves its previous state to next state.
- Rewards (r) : At some States the Agent get some rewards.
- Environment: It is environment in which Agent do Actions. Its mapping of previous state to next state and rewards generally known as transition dynamics.

  Transition probability express the current and next state with action.

MARKOV DECISION PROCESS (MDP)

Example: The working of robot is assumed that it seek the can at plane surface and collect it.

MARKOV DECISION PROCESS (MDP)

Example: The working of robot is assumed that it seek the can at plane surface and collect it.

Action

- Seek of can (a,)
- Collection of can (a_2)
- plug in for charge (a_3)

MARKOV DECISION PROCESS (MDP)

Example: The working of robot is assumed that it seek the can at plane surface and collect it.

Action

- Seek of can (a,)
- Collection of can (a_2)
- plug in for charge (a_3)

State

- Battery (L)
- Battery (H)

MARKOV DECISION PROCESS (MDP)

Example: The working of robot is assumed that it seek the can at plane surface and collect it.

Action

- Seek of can (a,)
- Collection of can (a_2)
- plug in for charge (a_3)

State

- Battery (L)
- Battery (H)

State	Reward
By collection of can	+1
At low battery	-3

MARKOV DECISION PROCESS (MDP)

Example: The working of robot is assumed that it seek the can at plane surface and collect it.

Action

- Seek of can (a_1)
- Collection of can (a_2)
- plug in for charge (a_3)

State

- Battery (L)
- Battery (H)

State	Reward
By collection of can	+1
At low battery	-3

Note: State Transition Matrix in term of probability

Policy of Machine - π

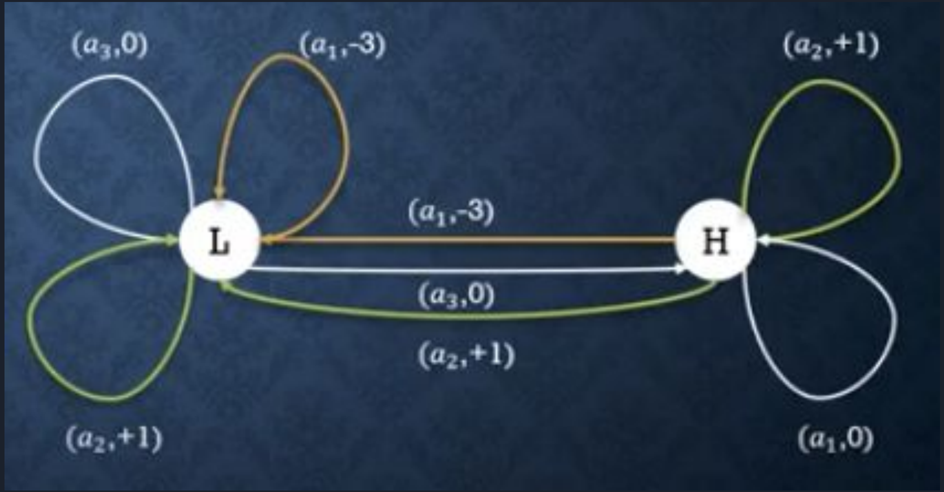
MARKOV DECISION PROCESS (MDP)

Action

- Seek of can (a_1)
- Collection of can (a_2)
- plug in for charge (a_3)

State

- Battery Low (L)
- Battery High (H)



State	Reward
By collection of can	+1
At low battery	-3

STATE TRANSITION DIAGRAM

Action

- Seek of can (a_1)
- Collection of can (a_2)
- plug in for charge (a_3)

STATE TRANSITION DIAGRAM

Action

- Seek of can (a_1)
- Collection of can (a_2)
- plug in for charge (a_3)

Rewards

STATE TRANSITION DIAGRAM

Action

- Seek of can (a_1)
- Collection of can (a_2)
- plug in for charge (a_3)

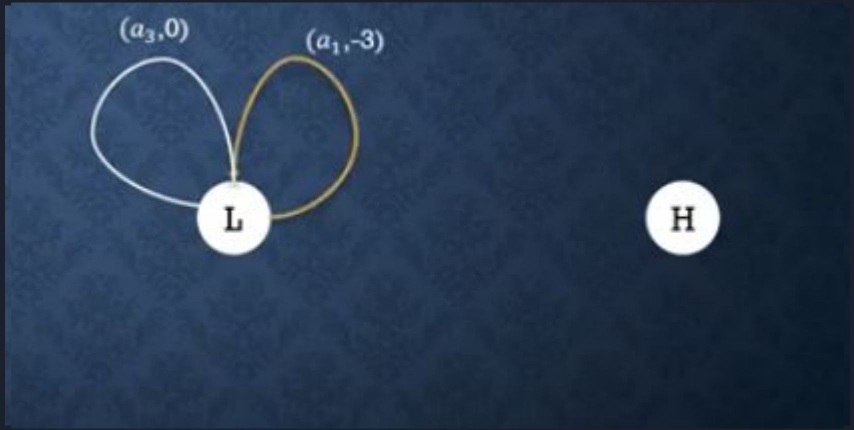
Rewards

0

STATE TRANSITION DIAGRAM

Action

- Seek of can (a_1)
- Collection of can (a_2)
- plug in for charge (a_3)



Rewards

0
0 - 3 = -3

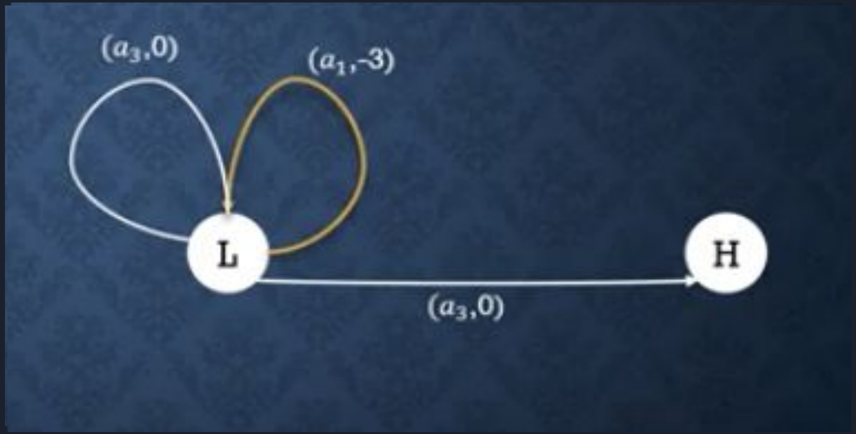
STATE TRANSITION DIAGRAM

Action

- Seek of can (a_1)
- Collection of can (a_2)
- plug in for charge (a_3)

Rewards

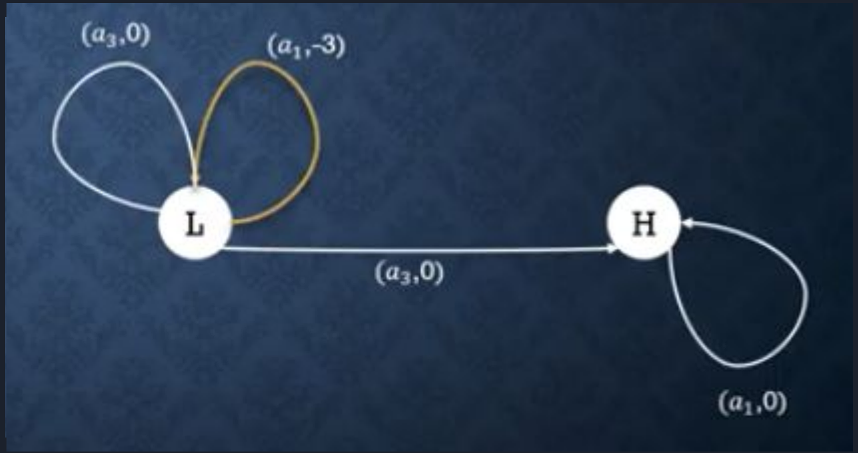
0
0 - 3 = -3
-3 + 0 = -3



STATE TRANSITION DIAGRAM

Action

- Seek of can (a_1)
- Collection of can (a_2)
- plug in for charge (a_3)



Rewards

0
0 - 3 = -3
-3 + 0 = -3
-3 + 0 = -3

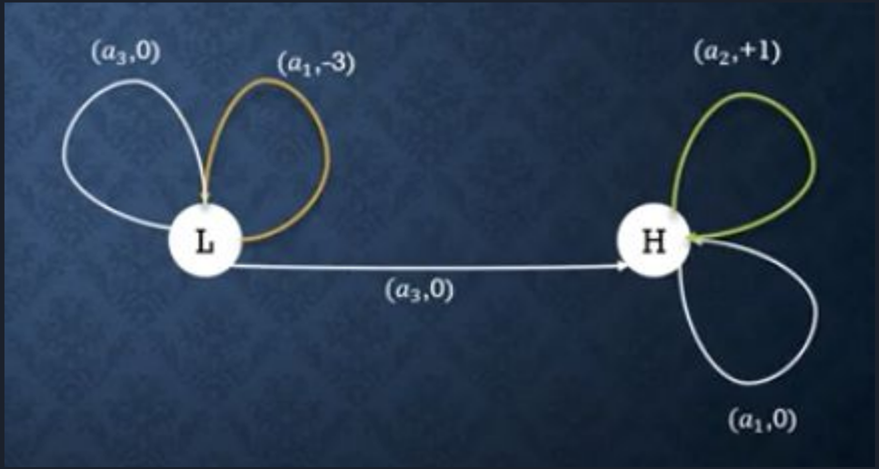
STATE TRANSITION DIAGRAM

Action

- Seek of can (a_1)
- Collection of can (a_2)
- plug in for charge (a_3)

<ins>*Rewards*</ins>

0
0 - 3 = -3
-3 + 0 = -3
-3 + 1 = -2



MARKOV DECISION PROCESS (MDP)

$$P(s' \mid s, a) = \sum_{r \in R} P(s', r \mid s, a)$$

Total discounted return:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \ldots + \gamma^{N-1} R_{t+N}$$

Where

R is the reward it may be positive or negative

γ is discount factor its value is between 0-1