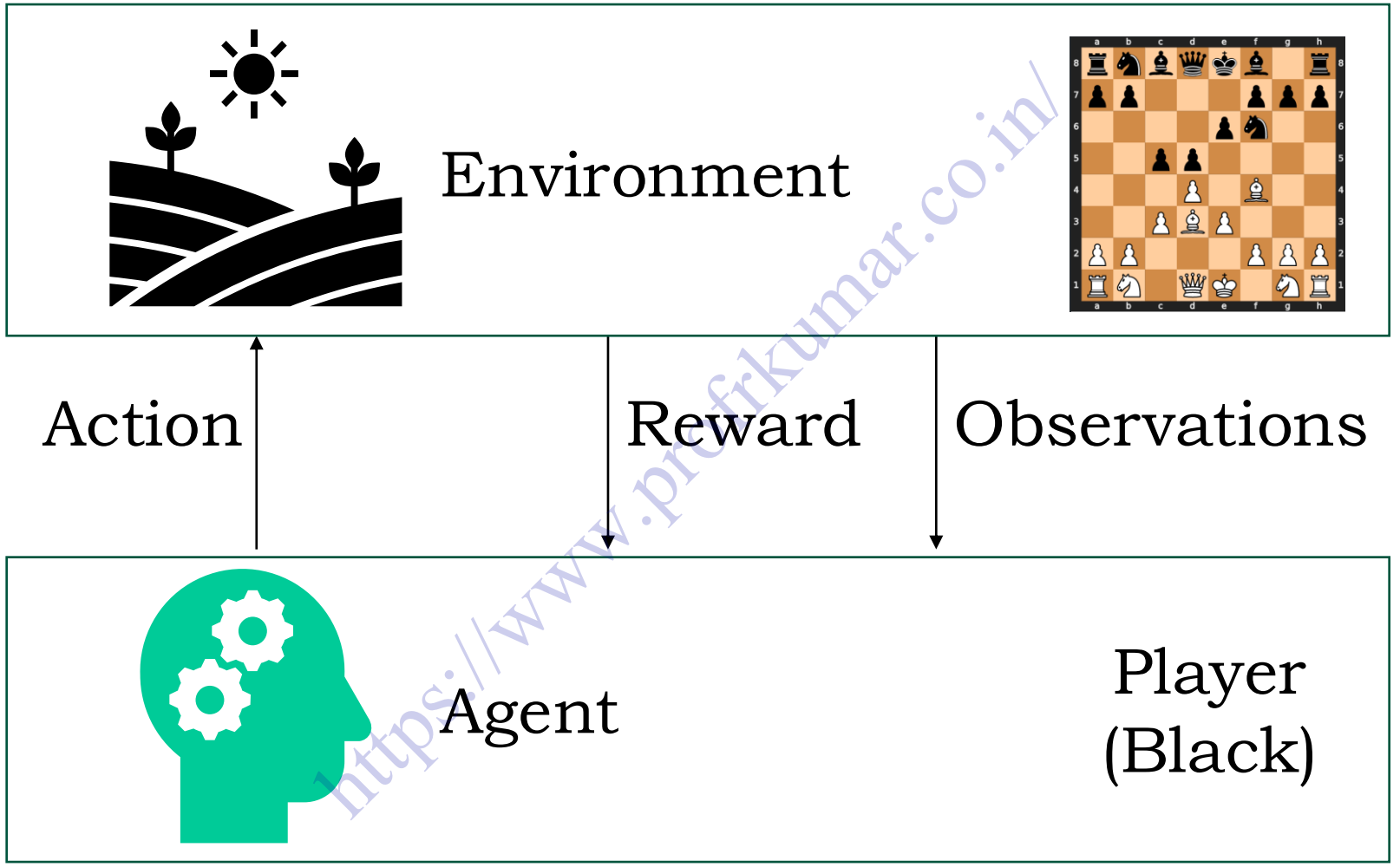
Reinforcement Learning

Introduction



https://www.profrkumar.co.in/

Reinforcement Learning

Markov Decision Process

The framework of Reinforcement Learning (RL).

- Set of states, S
- Set of actions, A
- Transition function, $P(s'|s,a)$
- Reward function, $R(s,a,s')$
- Initial state, s_0
- Discount factor, γ
- Horizon, H

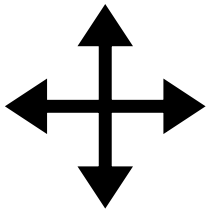
$$p(s_1|s_0,\rightarrow) = 0.95$$
$$p(s_0|s_0,\rightarrow) = 0.05$$

The goal of the agent is to find a policy that will maximize the total reward.

$$S = \{s_0, s_1, s_2, s_3, \dots, s_{14}\}$$

$$A = \{\rightarrow, \leftarrow, \uparrow, \downarrow\}$$





Action
s

Reward:
+10 (green box)
0 (elsewhere)

Episode terminates when monkey reaches banana

https://www.profrkumar.co.in/

Reinforcement Learning

Markov Decision Process

The goal of the agent is to **find a policy that will maximize the total reward.**

$$V(s) = \underbrace{R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots + \gamma^{T-t} R_T}_{\text{Discounted return}}$$

$$V(s) = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots + \gamma^{T-t} R_T$$

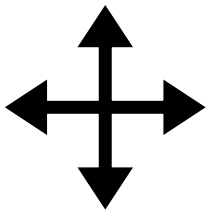
$$V(s) = R_t + \gamma V(s') = r + \gamma V(s')$$

$$S = \{s_0, s_1, s_2, s_3, \ldots, s_{14}\}$$

$$A = \{\rightarrow, \leftarrow, \uparrow, \downarrow\}$$

$V(s) =$

s_0 5.314	s_1 5.905	s_2 6.561	s_3 7.29
s_4 5.905	s_5 6.561	s_6 7.29	s_7 8.1
s_8 6.561	s_9 7.29	s_{10} 8.1	s_{11} 9
s_{12} 7.29	s_{13} 8.1	s_{14} 9	+10



Action
s

Reward:
+10 (green box)
0 (elsewhere)

Episode terminates when monkey reaches banana

https://www.profrkumar.co.in/

Reinforcement Learning

Markov Decision Process

The goal of the agent is to **find a policy that will maximize the total reward.**

$$V(s) = \underbrace{R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots + \gamma^{T-t} R_T}_{\text{Discounted return}}$$

$$V(s) = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots + \gamma^{T-t} R_T$$

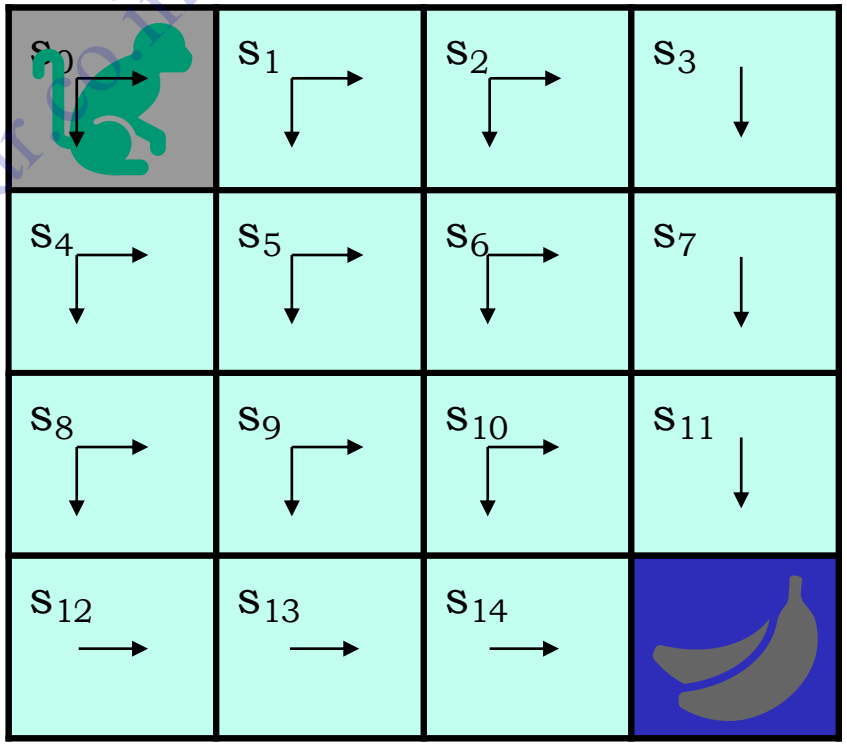
$$V(s) = R_t + \gamma V(s') = r + \gamma V(s')$$

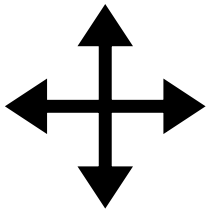
1 **NOTE:** Even this optimal/greedy policy is non – deterministic. Likewise, most policies in real life are **non – deterministic**.

$$S = \{s_0, s_1, s_2, s_3, \ldots, s_{14}\}$$

$$A = \{\rightarrow, \leftarrow, \uparrow, \downarrow\}$$

$\pi^*(s) =$





Action
s

Reward:
+10 (green box)
0 (elsewhere)

Episode terminates when monkey reaches banana

https://www.profrkumar.co.in/

Reinforcement Learning

Markov Decision Process

The goal of the agent is to **find a policy that will maximize the total reward.**

$$\underbrace{V(s) = r + \gamma V(s')}$$

Value function for deterministic transition

$$\underbrace{V^*(s) = \max_a \left(r + \gamma V^*(s')\right)}$$

Optimal value function for deterministic transition

2 This method fails in a stochastic setting, when an action has different probabilities of transitioning to **different state s' by taking an**

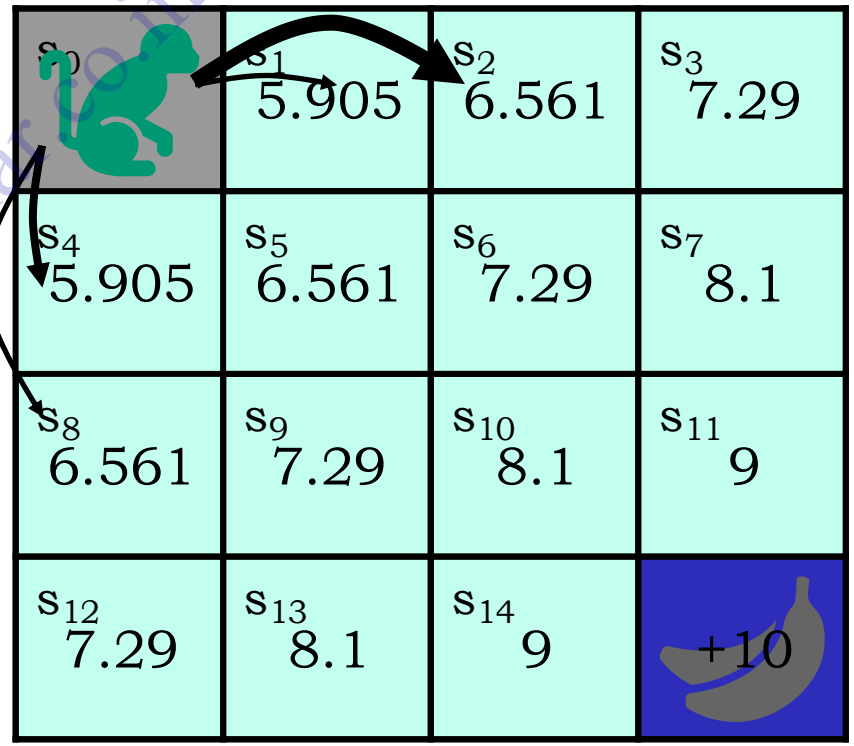
$p(s_1|s_0, \rightarrow) = 0.2$
$p(s_2|s_0, \rightarrow) = 0.8$
$p(s_4|s_0, \downarrow) = 0.58$
$p(s_8|s_0, \downarrow) = 0.42$

Noise in transition or non deterministic transition.

$S = \{s_0, s_1, s_2, s_3, \ldots, s_{14}\}$

$A = \{\rightarrow, \leftarrow, \uparrow, \downarrow\}$

$G_t(s) =$



Action s

Reward:
+10 (green box)
0 (elsewhere)

Episode terminates when monkey reaches banana

http://www.profrkumar.co.in/

Reinforcement Learning

Markov Decision Process

The goal of the agent is to **find a policy that will maximize the total reward.**

$$\underbrace{V^*(s) = \max_a \left(r + \gamma V^*(s')\right)}$$

Optimal value function for deterministic transition

$$V^*(s) = \max_a \mathbb{E}(r + \gamma V^*(s') | s \in S))$$

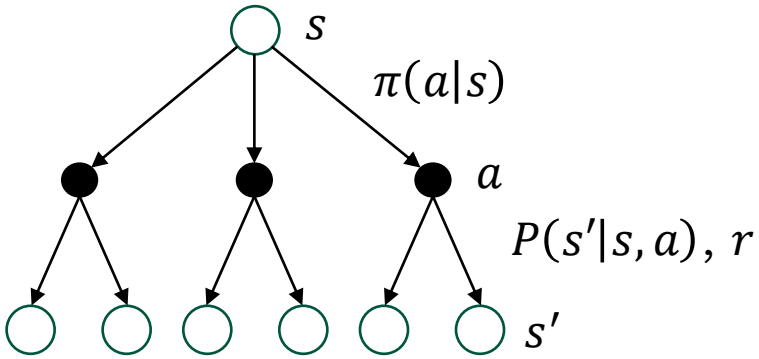
Bellman's Optimality

$$V^*(s) = \max_a \sum P(s'|s,a)\left(r + \gamma V^*(s')\right)$$

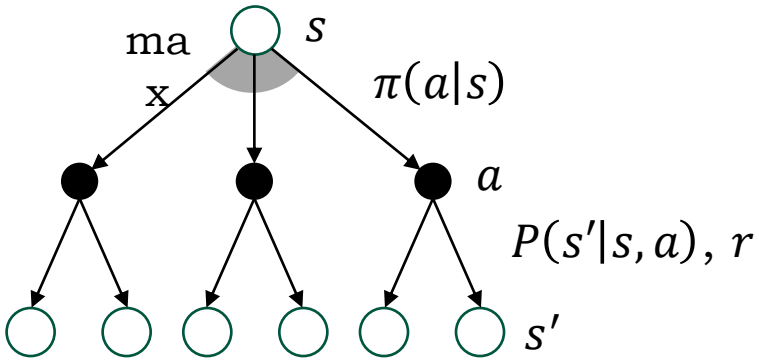
Generalized Bellman's Equation

$$V^{\pi}(s) = \sum_a \pi(a|s) \sum_{s'} P(s'|s,a)\left(r + \gamma V^{\pi}(s')\right)$$

Backup diagram



Backup diagram



https://www.profrkumar.co.in/

Reinforcement Learning

Markov Decision Process: Key Takeaways

The framework of Reinforcement Learning (RL).

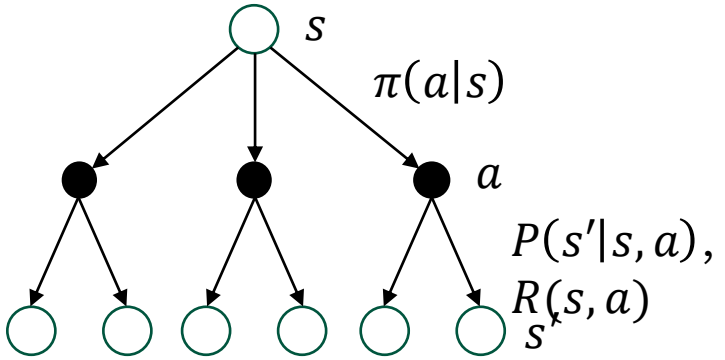
- Set of states, S
- Set of actions, A
- Transition function, $P(s'|s,a)$ } dynamics of environment
- Reward function, $R(s,a,s')$
- Initial state, s_0
- Discount factor, γ
- Horizon, H
- Value function
- Bellman Optimality Equation

Generalized Value Function

$$V^{\pi}(s) = \sum_{a} \pi(a|s) \sum_{s'} P(s'|s,a)\left(r + \gamma V^{\pi}(s')\right)$$

The goal of the agent is to find a policy that will maximize the total reward.

Backup diagram



Bellman's Optimality Equation

$$V^{*}(s) = \max_{a} \sum P(s'|s,a)\left(r + \gamma V^{*}(s')\right)$$

https://www.profrkumar.co.in/