REINFORCEMENT LEARNING

Lecture 1 : Introduction to Reinforcement Learning

Ibrahim Sammour

December | 2023

MACHINE LEARNING

- A branch of artificial intelligence.
- Enable computers to learn from data or interactions.
- Identify patterns in data to make accurate predictions or decisions.

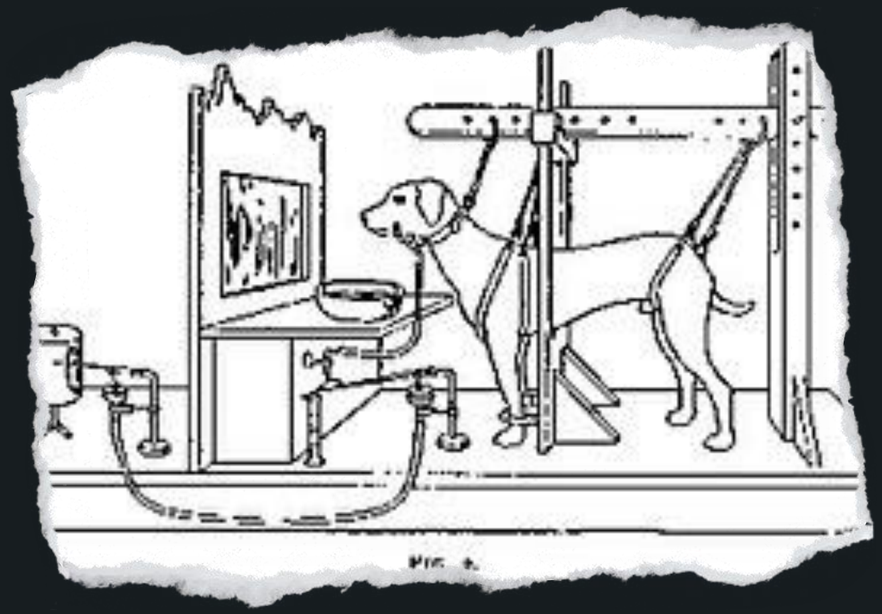
- Different types of ML:
  - Supervised learning (using labeled data).
  - Unsupervised learning (analyzing patterns in unlabeled data).
  - Semi-supervised learning (combining labeled and unlabeled data).
  - **Reinforcement learning** (learning through interaction with an environment).

HISTORICAL BACKGROUND

PAVLOV RESEARCH 1897

- Study the digestive processes of animals over long periods.
- The dogs began to salivate in the presence of the technician who normally fed them

IVAN PAVLOV



Unconditioned Stimulus (US)
Food

Unconditional Response (UR)
Salivation

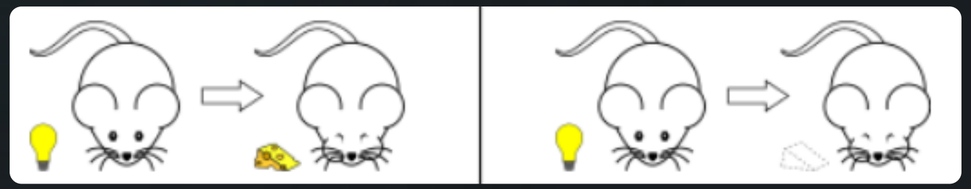
Conditioned Stimulus (CS)
Bell

Conditioned Response (CR)
Salivation

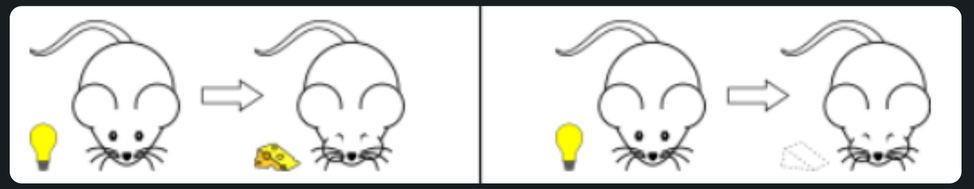
HISTORICAL BACKGROUND

PAVLOV RESEARCH 1897

Forward conditioning (Delay)

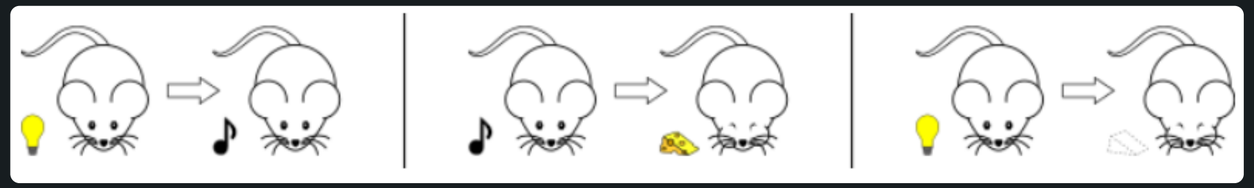


Forward conditioning (Trace)

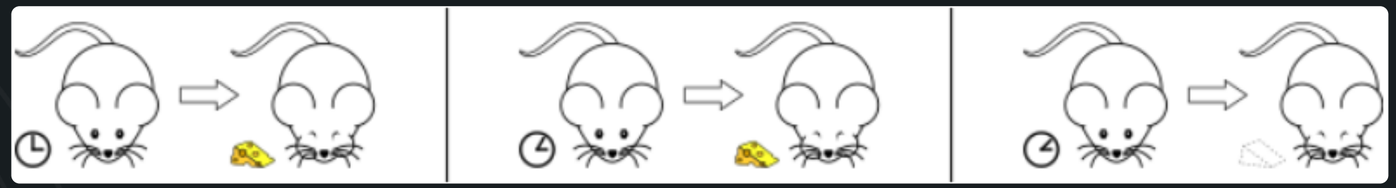


Simultaneous conditioning

Higher-order conditioning



Temporal conditioning



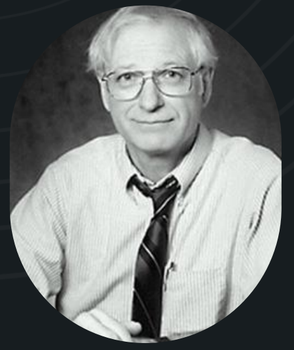
HISTORICAL BACKGROUND

RESCORLA-WAGNER MODEL (1972)

- An animal only learns when events violate its expectations.
- The animal learns gradually.
- Developed a mathematical model:
  - Impact of each stimulus.
  - Impact of combined stimulus (high-order conditioning).
  - Strengths of stimuli reach a maximum with trial and error

$$\Delta V_X^{n+1} = \alpha_X \beta (\lambda - V_{\text{tot}})$$

ROBERT RESCORLA



ALLAN R. WAGNER

REINFORCEMENT LEARNING

What makes reinforcement learning different from other machine learning paradigms?

- There is no supervisor, only a **reward** signal
- **Feedback** is delayed, not instantaneous
- Time really matters
- Agent’s actions affect the subsequent data it receives

REINFORCEMENT LEARNING

REINFORCEMENT LEARNING FEATURES

Distinguishing features of RL

Trial and Error

Delayed Reward

RL tries to maximize a reward instead of finding a hidden structure (i.e. unsupervised)

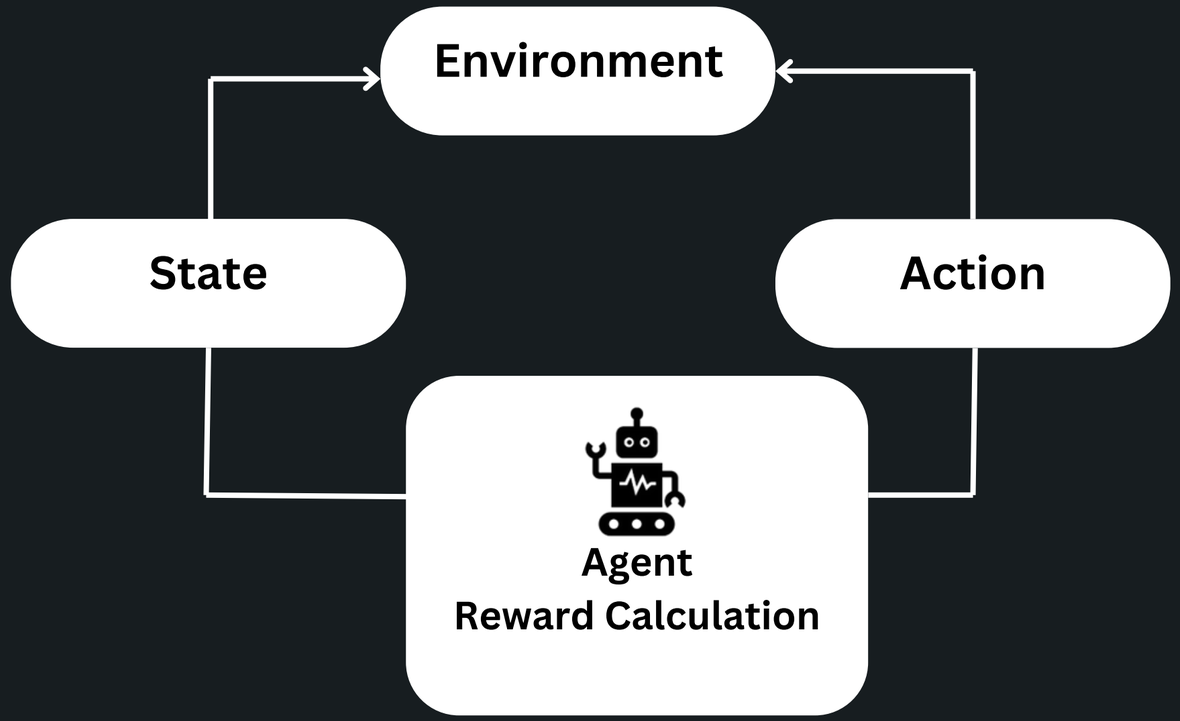
REINFORCEMENT LEARNING

REINFORCEMENT LEARNING BASICS

An agent learns by interacting with an environment to maximize its reward.

Self-learning agent:

- Observe the state of the environment
- Perform an action
- Observe the new state of the environment
- Calculate the reward



REINFORCEMENT LEARNING

REINFORCEMENT LEARNING EXAMPLES

Self-Driving cars

Goal: Drive without crashing for the longest time possible

State:

- Sensor data (left,right,forward, backwork)
- Position in trail

Action:

- Move forward, backward
- Steer left, right

Reward:

- r = timeElapsed + distanceTraveled
- r = timeElapsed + distanceTraveled + distanceFromWalls

REINFORCEMENT LEARNING

REINFORCEMENT LEARNING EXAMPLES

Self-Driving cars

REINFORCEMENT LEARNING

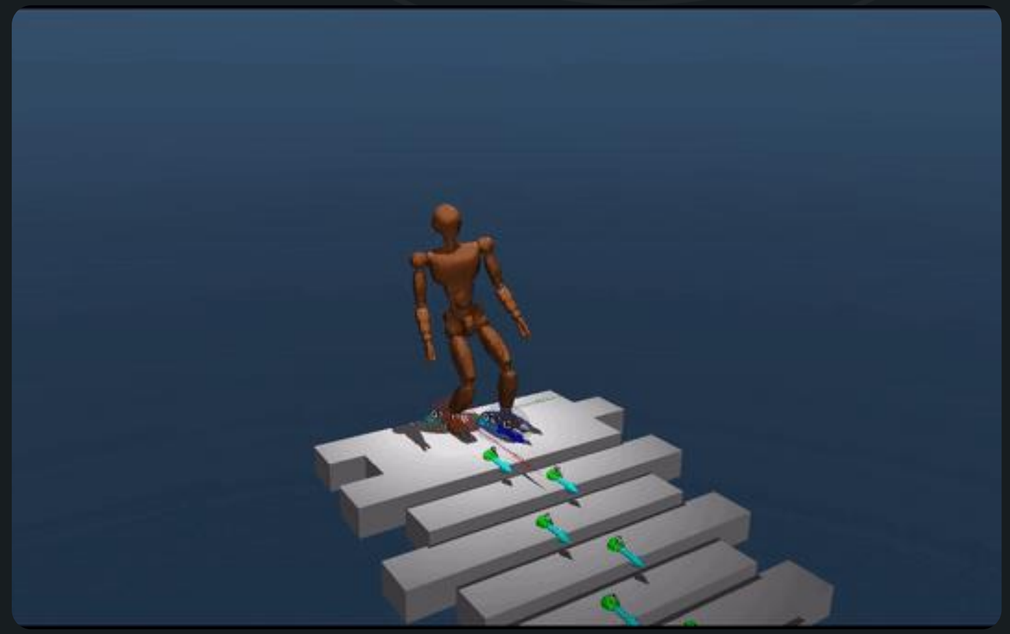
REINFORCEMENT LEARNING EXAMPLES

Make Humanoid Robot Walk

Goal: Walk till reaching the finish line

State:

- Position
- Joints angles
- Head height



Actions:

- Rotate joints

Rewards:

- r = - timeElapsed + position/totalDistance + headHeight/defaultHeadHeight

REINFORCEMENT LEARNING

REINFORCEMENT LEARNING EXAMPLES

Make Humanoid Robot Walk

REINFORCEMENT LEARNING

SUBELEMENTS OF REINFORCEMENT LEARNING

Policy (π)

- Defines agent's behavior
- Maps states to actions
- Probabilities for each action
- Could be:
  - Lookup table
  - Neural network

Example Policy

	Action 1	Action 2
State 1	0.6	0.4
State 2	0.2	0.8
State 3	0.25	0.75

REINFORCEMENT LEARNING

SUBELEMENTS OF REINFORCEMENT LEARNING

Action Choice

Which action should we pick at a given state?

- **Highest probability?**
- **Random?**

	Action 1	Action 2
State 1	0.6	0.4
State 2	0.2	0.8
State 3	0.25	0.75

REINFORCEMENT LEARNING

SUBELEMENTS OF REINFORCEMENT LEARNING

Exploration v.s. Exploitation

Exploitation: Selecting actions of the highest probability.

Exploration: Exploring different actions.

ε-greedy methods:

- Explore with probability ε
- Exploit with probability 1-ε
- ε is in the range of [0,1]

REINFORCEMENT LEARNING

SUBELEMENTS OF REINFORCEMENT LEARNING

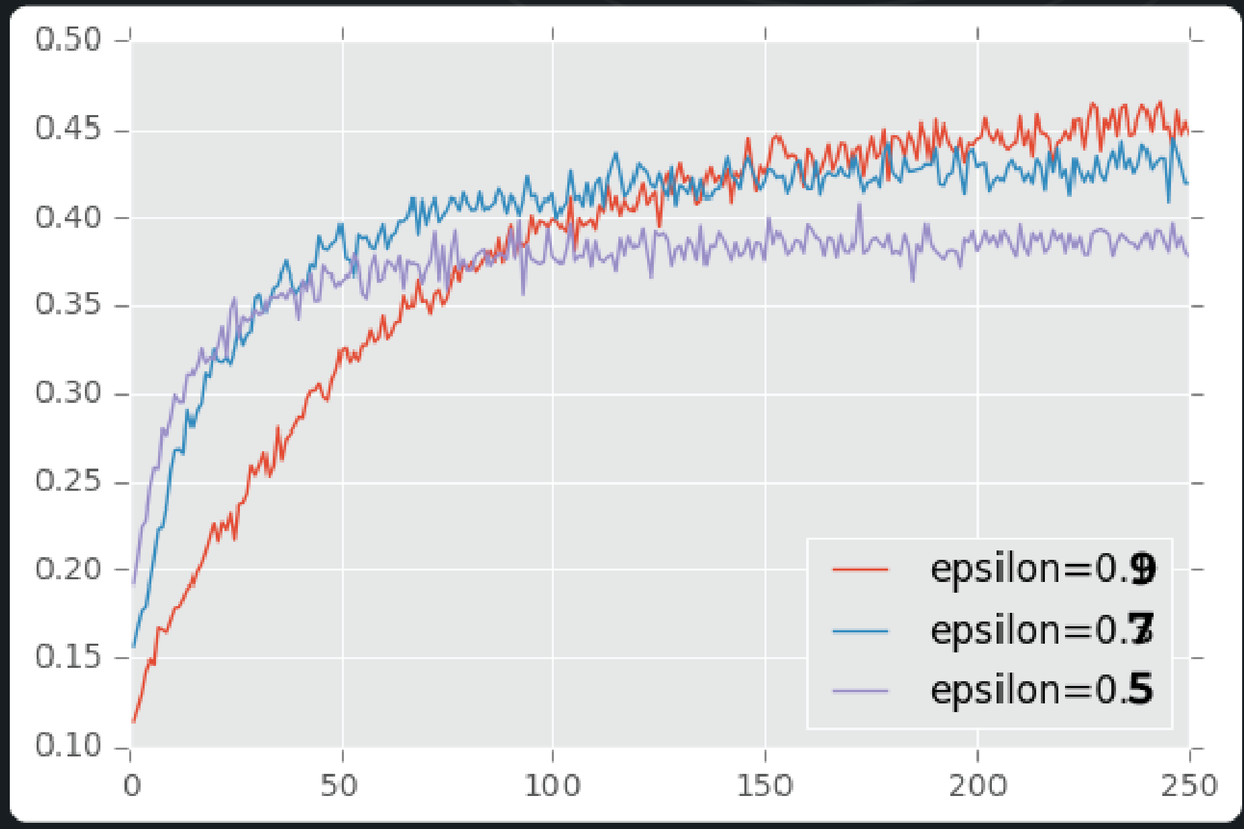
Exploration v.s. Exploitation

If ε is low:

- Fast convergence.
- May not reach the maximum reward value.

If ε is high:

- Slow convergence.
- May achieve high reward values, but not necessarily the maximum.



Reward v.s. time step

REINFORCEMENT LEARNING

SUBELEMENTS OF REINFORCEMENT LEARNING

Value Function (Vπ)

- Defines **states‘ values**
- States with higher values are better
- Similar to pain-pleasure
- More farsighted judgment than immediate rewards
- Could be:
  - Lookup table
  - Neural network

Example Value Function

	Value
State 1	3
State 2	4
State 3	2.5
State 4	5

REINFORCEMENT LEARNING

EXAMPLE

Start and reach the goal while acquiring maximum reward.

The value of each cell is the cost of reaching that state.

Reward = cost

r(1,1) = -1

r(1,2) = -4

State	1	2	3
1	-1	-4	Goal (-1)
2	-2	-1	-1
3	-1	-2	-1
4	Start	-1	-3

REINFORCEMENT LEARNING

EXAMPLE

V(3,2) = ??

V(3,2) = r(2,2) + r(1,2) + r(1,3) = -6

At the end of training, what are the probabilities of right and up actions at V(3,2) and V(2,2)?

Is there a better way to calculate the value function?

State	1	2	3
1	-1	-4	Goal (-1)
2	-2	-1	-1
3	-1	-2	-1
4	Start	-1	-3

REINFORCEMENT LEARNING

EXAMPLE

Consider training a car on a road

Case 1

Case 2

REINFORCEMENT LEARNING

EXAMPLE

Consider training a car on a road

Case 1

Straight road

The car either:

- **Speed up**
- **Speed down**

No critical decisions are to be made

Short-term rewards help me reach the end faster

REINFORCEMENT LEARNING

EXAMPLE

Consider training a car on a road

Crossroad

The car either:

- Steer left
- Steer right
- Keep moving forward

The decision is critical and highly impacts the future outcomes

Case 2

REINFORCEMENT LEARNING

SUBELEMENTS OF REINFORCEMENT LEARNING

Cumulative Reward

$$\bar{r} = r_t + \gamma * r_{t+1} + \gamma^2 * r_{t+2} + \cdots + \gamma^T * r_T$$

$\bar{r}$: return

r_t : reward obtained at time t

γ : discount factor

T: end of an episode or window of interest

γ is high => we favor long term rewards

γ is low => we favor short term rewards

REINFORCEMENT LEARNING

EXAMPLE

V(4,1) = r(3,1) + r(2,1) + r(2,2) + r(2,3) + r(1,3) = -6

State	1	2	3
1	-1	-4	Goal (-1)
2	-2	-1	-1
3	-1	-2	-1
4	Start	-1	-3

REINFORCEMENT LEARNING

EXAMPLE

V(4,1) = r(3,1) + r(2,1) + r(2,2) + r(2,3) + r(1,3) = -7

State	1	2	3
1	-1	-4	Goal (-1)
2	-2	-1	-1
3	-1	-2	-1
4	Start	-1	-3

REINFORCEMENT LEARNING

EXAMPLE

$\gamma = 0.1$

$V(4,1) = r(3,1) + 0.1*r(2,1) + 0.01*r(2,2) + 0.001*r(2,3) + 0.0001*r(1,3) = -1.2111$

$\gamma = 0.9$

$V(4,1) = r(3,1) + 0.9*r(2,1) + 0.81*r(2,2) + 0.729*r(2,3) + 0.65*r(1,3) = -4.989$

State	1	2	3
1	-1	-4	Goal (-1)
2	-2	-1	-1
3	-1	-2	-1
4	Start	-1	-3

REINFORCEMENT LEARNING

EXAMPLE

$\gamma = 0.1$

$V(4,1) = r(3,1) + 0.1*r(2,1) + 0.01*r(2,2) + 0.001*r(2,3) + 0.0001*r(1,3) = -1.3111$

$\gamma = 0.9$

$V(4,1) = r(3,1) + 0.9*r(2,1) + 0.81*r(2,2) + 0.729*r(2,3) + 0.65*r(1,3) = -5.889$

State	1	2	3
1	-1	-4	Goal (-1)
2	-2	-1	-1
3	-1	-2	-1
4	Start	-1	-3

THANK YOU EVERYONE !

Ibrahim Sammour | isammour@outlook.com