Reinforcement Learning

Lecture - 01

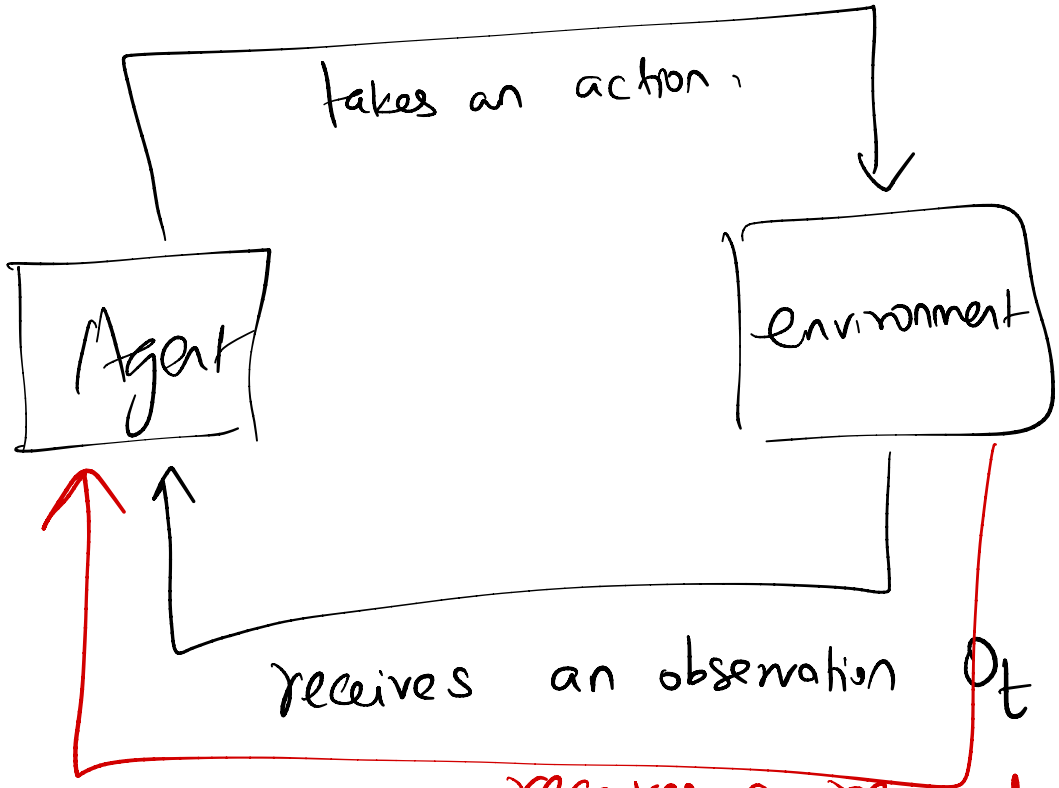
Artificial Intelligence :-

Goal of AI : → to understand the general principles behind human intelligence.

→ to simulate them in machines.

ML : takes the "learning" approach to solve AI.

- Active learning (RL) — learning through interaction.
- Passive learning (SL) — learning from data.



Objective: to learn how to map situations to action — so as to maximize a numerical reward signal.

Example: Trash collecting robot.

+1 — whenever it collects a trash,

0 — rest of the time.

-100 — if it runs out of battery.

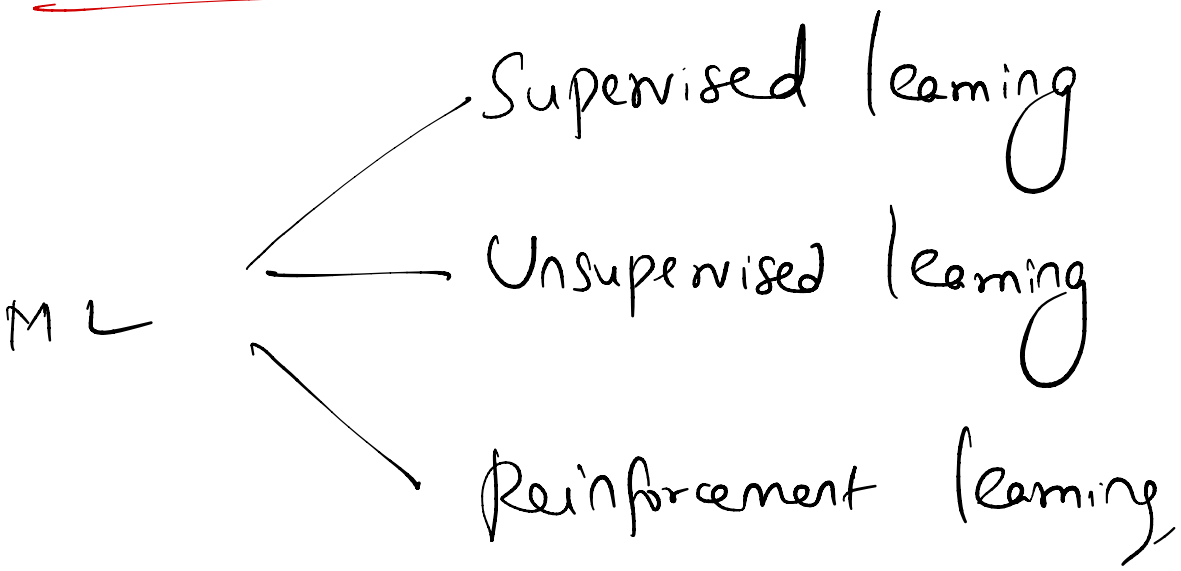
Reward hypothesis:-

That all of what we mean by goals and purposes can be well thought of as max. of the expected value of the cumulative sum of a received scalar signal (reward)

Reward-is-Enough hypothesis (Silver et al. 2021)

Intelligence, and its associated abilities, can be understood as subserving the max. of reward by an agent acting in its environment.

Detour to ML :—



Supervised learning :

$$\mathcal{D}_{Tr} = \left\{ x^{(i)}, y^{(i)} \right\}_{i=1}^{N}$$

$$f : x \rightarrow y$$

new $x \rightarrow$?

x : review

y : $+1$ Positive review

-1 negative rev.

"Prediction"

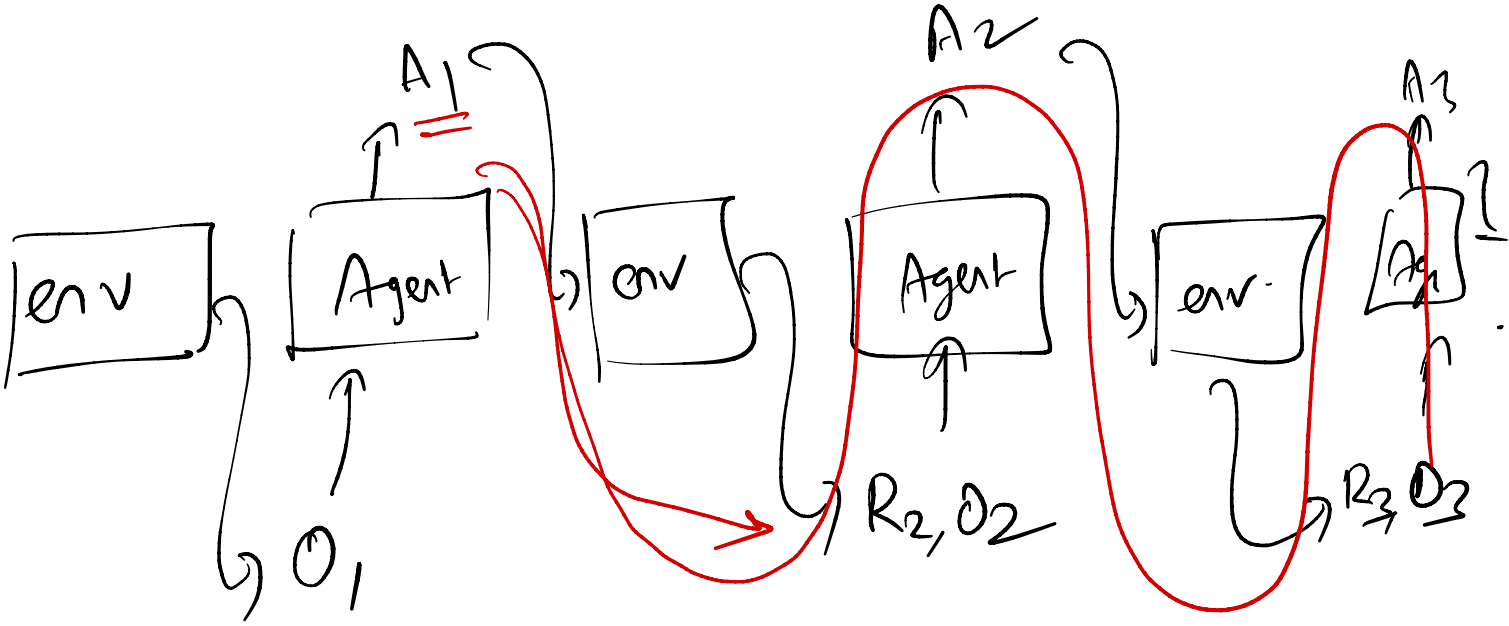
Unsupervised learning:-

$$\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$$

goal: find some structure in the data.

ex: clustering

Reinforcement Learning:

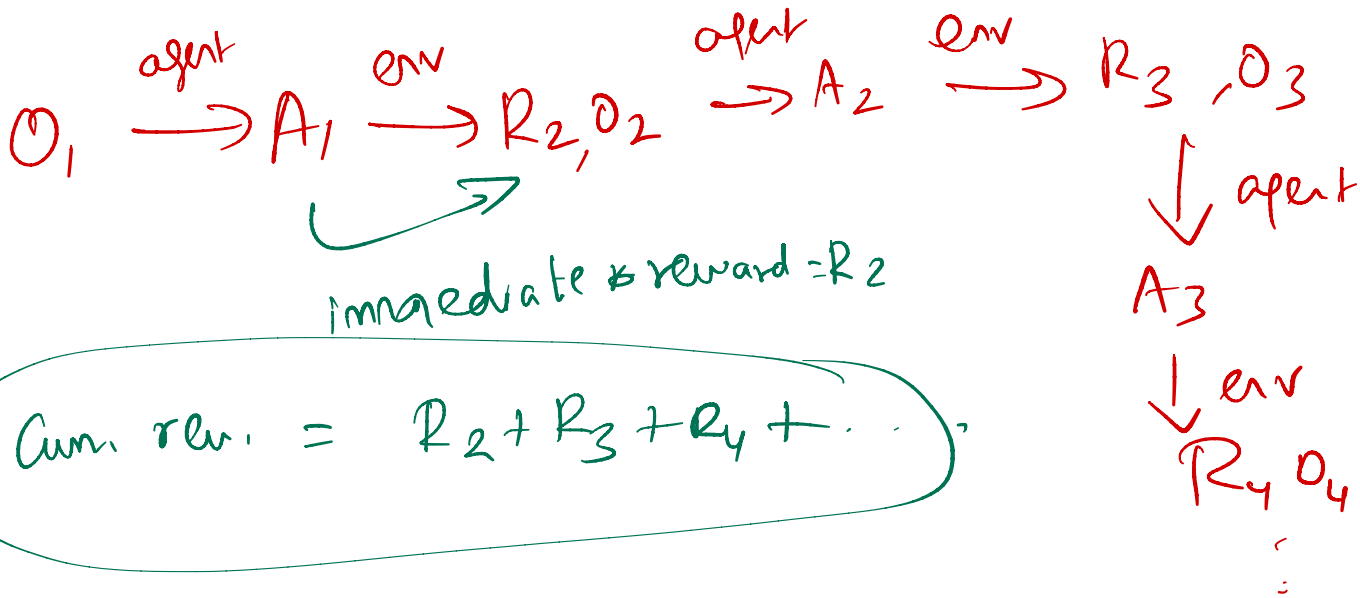


→ Interaction

→ notion of "time"

Implications of closed loop:-

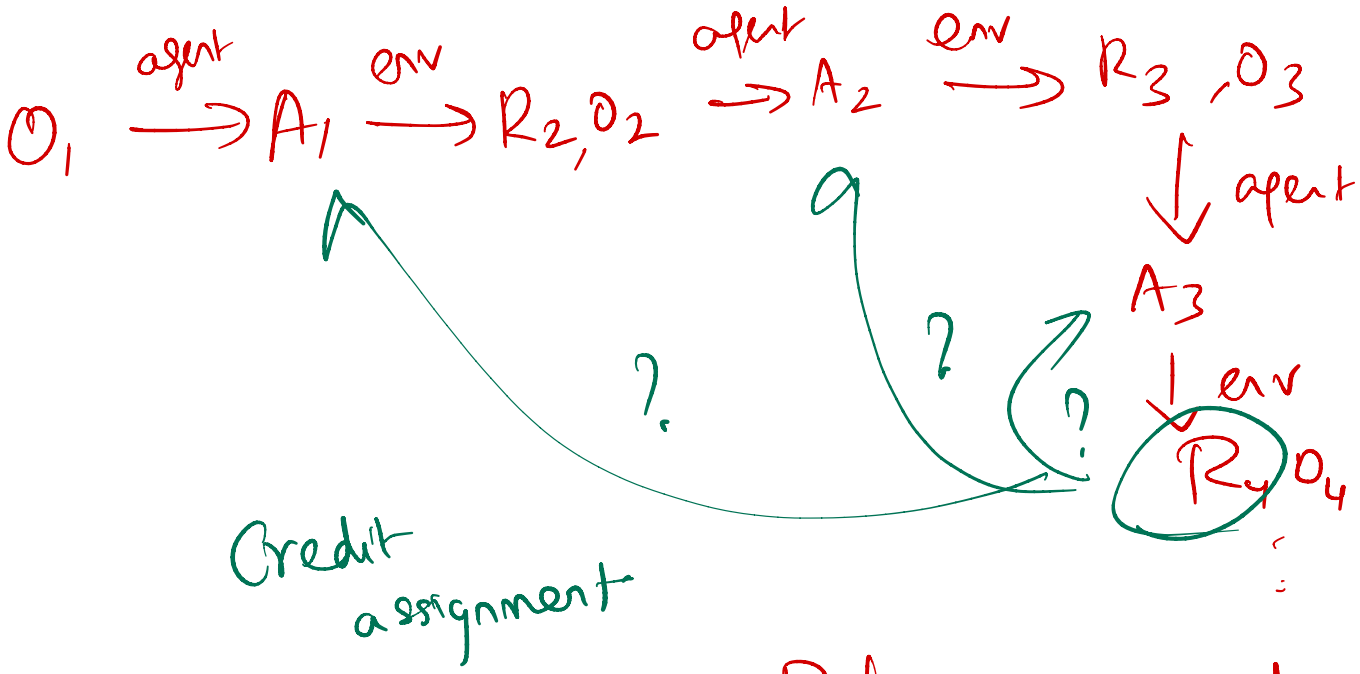
Forward view:-



Cum. rew. $= R_2 + R_3 + R_4 + \ldots$

Goal: to max. cumulative reward.

Backward view:



Delayed rewards.

---

trial & error

labels are given.

RL Vs SL :-

learn from interaction

learn from data.

Predict

Control

Prediction

delayed reward

Cur. RL.

RL vs. VS L :-

Exploration vs. exploitation dilemma ;-

exploit → what it has already exp. in order to obtain reward

explore → in order to make better action selection in the future.

RL - Summary:-

1. Agent view of AI.
2. Sequential decision making process.
3. Delayed rewards & Credit assignment
4. Exploration / Exploitation dilemma.

Immediate RL

Key char. of an RL problem:-

1. learning to act in many situation.
2. Delayed rewards / credit assignment.
3. Exploration / Exploitation dilemma.

types of feedback

- SL → Instructive feedback
- RL → evaluative feedback

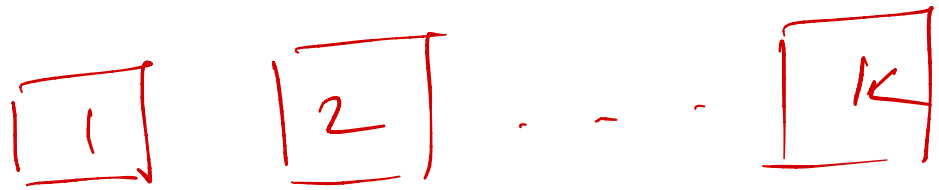
$O_t \rightarrow$ [Agent] $\rightarrow A_t$.

optimal action : a^*

(A_1, A_2, A_3, A_4)

$a^* = A_2$

Immediate RL: / k-armed bandit problem



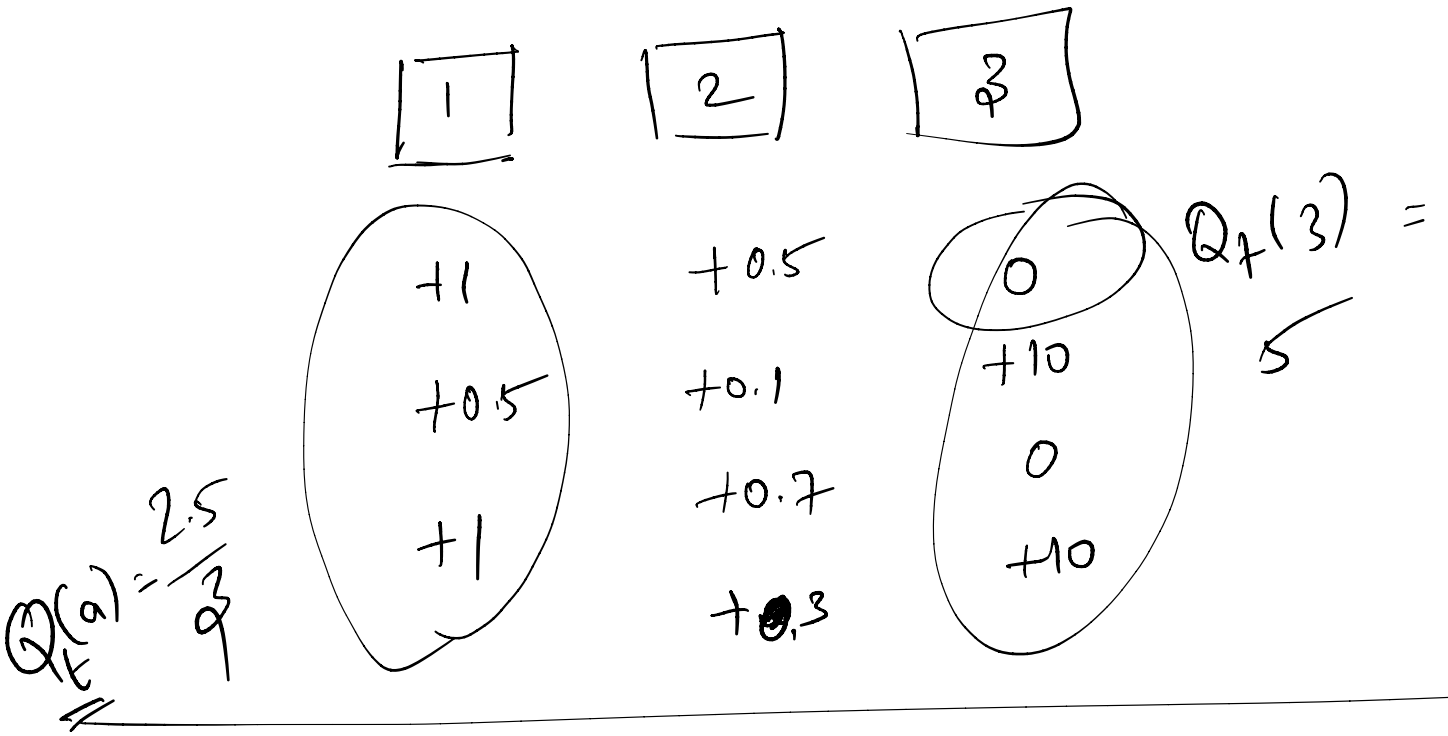
options/actions.

reward → chosen from stationary prob. distrib. that depends on your action.

objective: to maximize the expected total reward over some time period.

Note: 1. Agent sees the same state (same set of actions) all the time.

2. Rewards are immediate



Value of action $a = q_*(a)$

$= \mathbb{E}[R_t \mid A_t = a]$

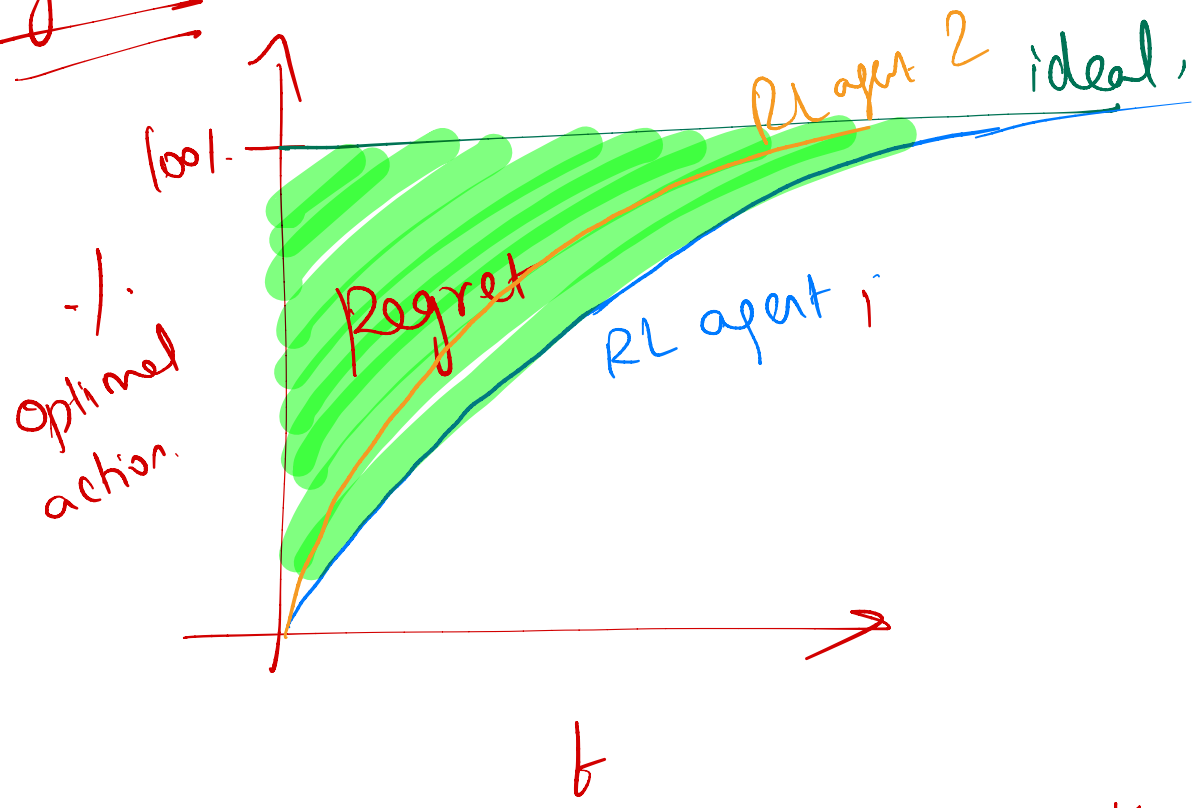
$q_*(a)$ — actual.

$Q_t(a)$ — estimate

$$a^* = \underset{a \in \{a_1 \ldots a_k\}}{\operatorname{argmax}} \; q_*(a)$$

$$A_t = \underset{a \in \{a_1 \ldots a_k\}}{\operatorname{argmax}} \; Q_t(a)$$

regret:



$$\text{Regret} = k\, q_*(a^*) - \sum_{t=1}^{k} R_t$$

Action-value methods:-

$$Q_t(a) = \frac{\text{Sum of rewards when 'a' taken prior to 't'}}{\text{\# of times 'a' taken prior to 't'}}$$

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_i = a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i = a}}$$

$$\mathbb{1}_{A_i = a} = \begin{cases} 1 & \text{if action 'a' was taken in timestep 'i'} \\ 0 & \text{otherwise.} \end{cases}$$

As $t \to \infty$ $\quad Q_t(a) \to q_*(a)$

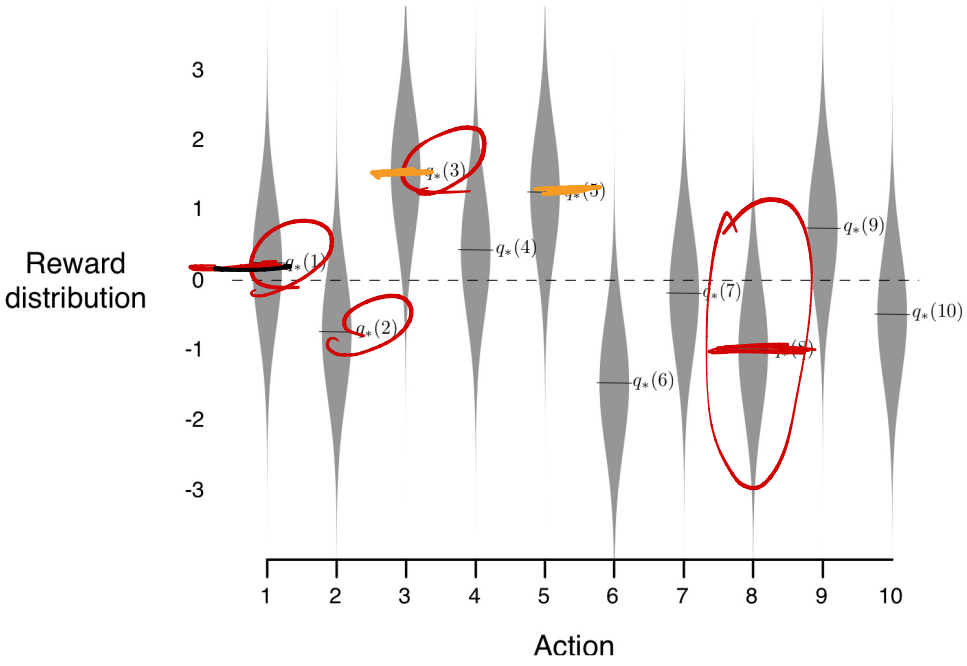
Exploit: $A_t = \underset{a}{\text{argmax}}\ Q_t(a)$

"greedy" action.

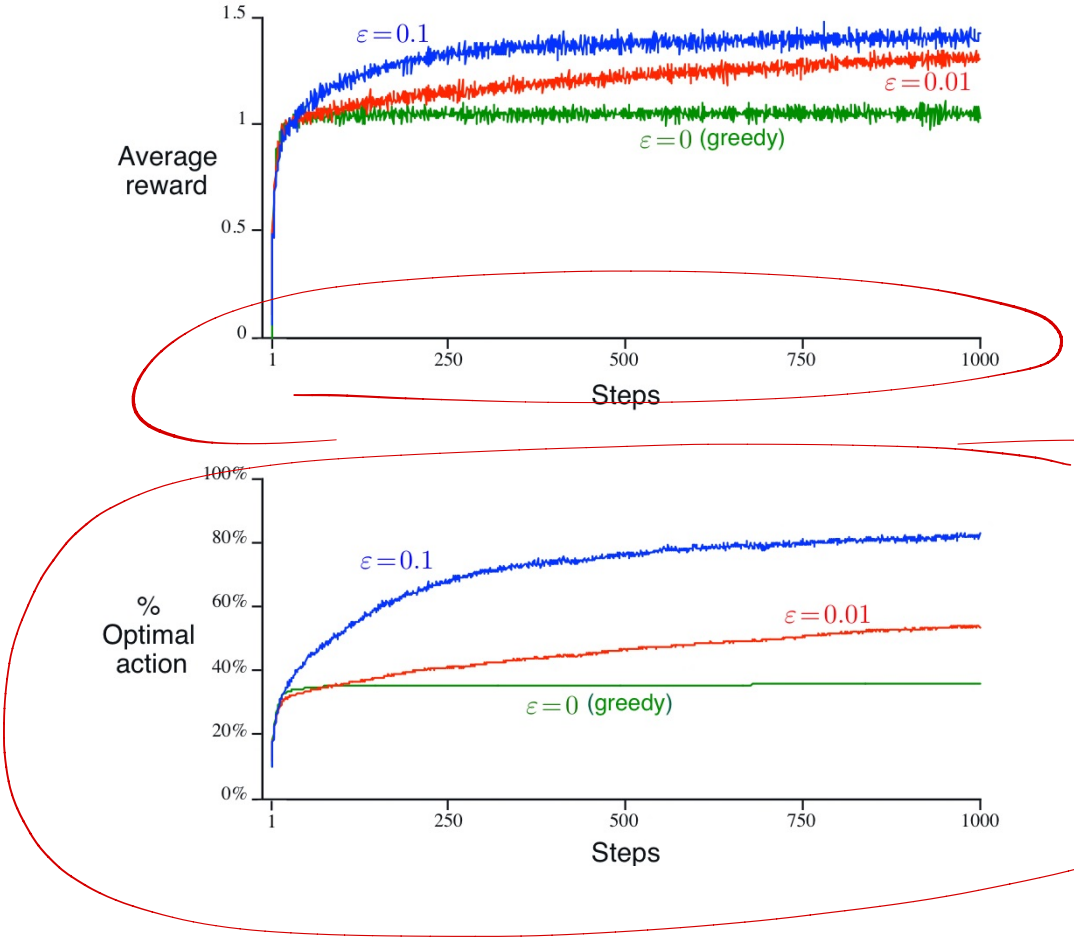
Prob. ϵ 0.1, 0.01, 0.05

$$A_t = \begin{cases} \underset{a}{\text{argmax}}\ Q_t(a) & 1-\epsilon \\ \text{random action} & \epsilon \end{cases}$$

ϵ-greedy



mean of $q_*(i) = \mathcal{N}(0, 1)$


1.5
1
0.5
0
Average reward
ε = 0.1
ε = 0.01
ε = 0 (greedy)
1
250
500
750
1000
Steps
100%
80%
60%
40%
20%
0%
% Optimal action
ε = 0.1
ε = 0.01
ε = 0 (greedy)
1
250
500
750
1000
Steps