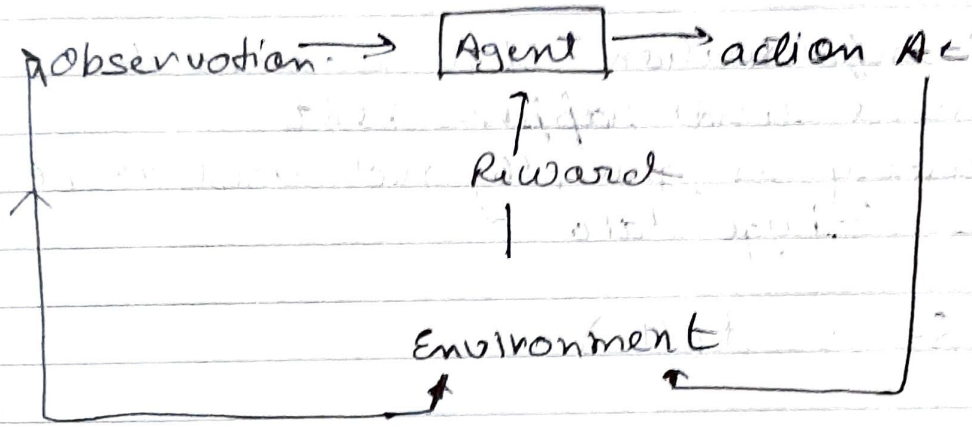
David Silver

Lecture 1 →

→ Goal of RL → Select actions to maximize total future Rewards
↳ series of action to get maximum total result/reward.



Agent Executes Action At
Recieves observation Ot
" scalar reward Rt

Rewards can be scalar but ultimately we want to compare the rewards we can get upon taking different actions so it is prefered to be scalar.

———— × ———— × ———— × ————

→ History → History is the sequence of actions, rewards and observations seen by the agent so far.

$H_t = A_1, O_1, R_1, A_2 \ldots\ldots A_t, O_t, R_t$

↳ all observable variables upto time t
↳ what happens next depens on history

→ State → Information that is used to determine what happens next
↳ History is generally not used as it has huge data.

$$S_t = f(H_t)$$

This information can be anything → 3 types :-

↳ Environment State → The information that is used within the environment. The agent can not see it. It only observes the what happens due to or by this state. For eg. Intermolecular collisions - We dont see them but we know there is water or air or solid on the table.

↳ Agent State → The information in the agent's internal representation. The useful observations by the agent which can be used in the algorithm.

~~the Rt~~

Agent state depends on the function 'f' is

$S_t^a = f(H_t)$ which is decided by us.

↳ Information State → a.k.a Markov State contains all useful information from the history.

Markov Stat S_t if an only if

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

the probability of the next state = The probability of the next state given current state given all the states visited so far.

future is independent of the past given the present.

——— ∝ ——— ∝ ———

→ Environment.

↳ Fully observable Environment → Fully able to see what going on in the environment or directly observes the environment.

Agent state = Environment state = Information State

$O_t = S_t^a = S_t^e$

Formally this is a MDP (Markov Decision Process)

↳ Partially observable Environment: Agent indirectly observes the environment.

↳ Ex → A blind person trying to know the shape of an ~~ato~~ object. It cannot see it so it indirectly gathers information by touching

agent state ≠ environment state

→ Partially observable MDP.

→ Policy(π) → How the agent picks an action
→ Value function → How good is each state/or action
→ Model → ~~Agents~~ Agents "understanding" of the environment.

↳ map from state to action → given a state, it will generate an action.
can be deterministic or stocastic

$$\pi(a|s) = P(A=a|S=s)$$

Value Function → It is the prediction of the expected total future reward.

$$V_\pi(s) = E_\pi \left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} \cdots \Big| S_t = s \right]$$

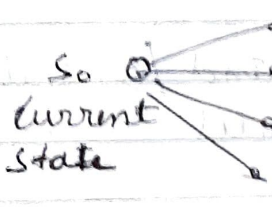
↑

↓

Value function for a policy

→ model → A model predicts what environment will do next. ~~Its uncertainty~~

→ transitions → P predicts the next state



each state has some p of occuring after S_0. transitions calulate thes p.

→ Rewards → R predicts the imidiate reward
It is stocastic in nature

$$P^a_{ss'} = P(s' = s' / s = s, A = a)$$

$$R^a_s = E\left[R / s = s, A = a\right]$$

Tyres of RL agents

→ Value Based → No Policy
→ Value function
: obseves the values for each action and taks action

→ Policy Based → Policy
→ No value Function
: takes action according to the policy

→ Model Free → Policy and/or Value function
→ No model

We only look at the policy and/or the value function for the particular state. We do not try to understand the working of the environment of make any sort of dynamics of the envt.

→ Model Based → Policy and/or Value Function
→ No model

We make an elaborate model of the environment based on the data and try to understand its dynamics