__/__/____

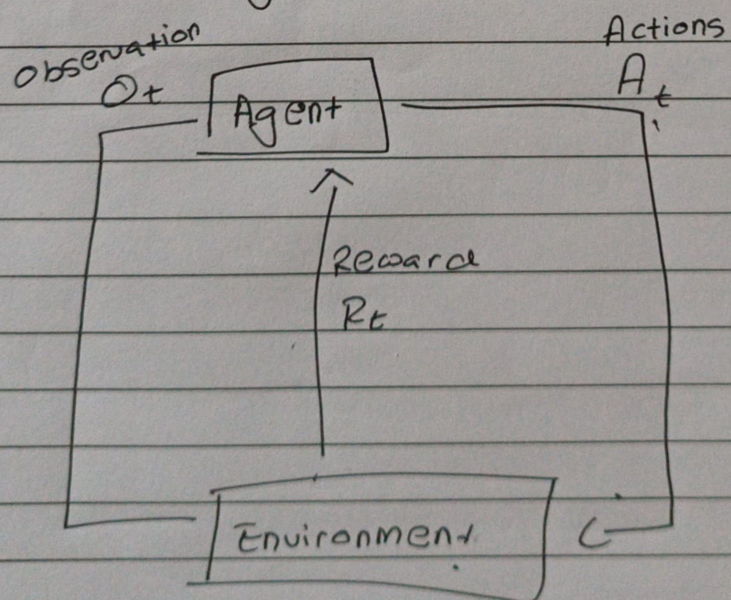
RL course by David Silver
↳ optimal way to make decisions
↳ intersection of many fields
mimics dopamine system as rewards

trial & error paradigm instead of supervisor
not instant rewards (DELAYED FEEDBACK)
dynamic environment (not iid data)
↓
independent & identically distribute

Problem
scalar R_t reward at time step t

Sequential Decision Making
↳ SELECT ACTIONS TO MAXIMIZE FUTURE REWARDS

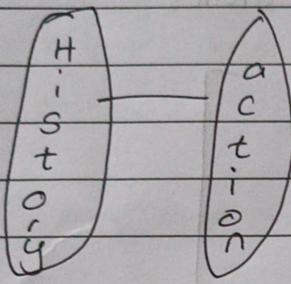
sometimes better to sacrifice immediate rewards to gain more long term rewards



History

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

all observable variables



State : concise summary of history

$S_t = f(H_t)$ only imp things

① Environment state S_t^e : environments ~~state~~ private representation

NOT visible is agent usually

NOT USED
can be irrelevant

② Agent state S_t^a (history of agent)

$S_t^a = f(H_t)$

USED ALOT

An information state (i.e. Markov state) contains all useful info from history

S_t is Markov ONLY IF

$$\mathbb{P}[S_{t+1} \mid S_t] = \mathbb{P}[S_{t+1} \mid S_1, \ldots, S_t]$$

probability

you can throw away all previous states & retain current state still you'll get same characterization of future

THE FUTURE IS INDEPENDENT ~~OF~~ PAST GIVEN PRESENT

__/__/__

S_t is sufficient

Example helicopter ~~a~~ performing stunts

Markov State $\begin{bmatrix} \text{position} \\ \text{velocity} \\ \vdots \\ \text{angular velocity} \\ \text{" " position} \\ \text{wind direction} \end{bmatrix}$

now we dont care of the state 10 mins back. What does it matter?

Imperfect Non-Markov [position] need to now look back on history & maybe calculate velocity/momentum

OUR JOB: defining good state that does best job of prediction

full observability

$O_t = S_t^a = S_t^e$ MDP

partial observability Agent state $\neq$ environment state

↳ Partially observable MDP (POMDP)

↳ remember all histor

↳ built beliefs (probability distribution)

↳ linear combination of state in last time step + latest observation

_/_/__

Agent

main components

↳ Policy: agents behaviour function

↳ Value: how good is each state and/or action

↳ Model: agents understanding of model

① Policy (map state to action)

- deterministic $a = \pi(s)$
- Stochastic $\pi(a|s) = \mathbb{P}[A = a | S = s]$

② Value prediction of future reward of state/action

$$v_\pi(s) = \mathbb{E}_\pi[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots | S_t = s]$$

↑ end is ~~steps~~ given by γ^n when too low

③ Model predictions about environment (OPTIONAL)

- transition (predicting next state)

state transition model

$$P^a_{ss'} = \mathbb{P}[S' = s' \mid S = s, A = a]$$

- reward (predicting immediate reward)

$$R^a_s = \mathbb{E}[R \mid \underbrace{S = s, A = a}_{p.}]$$

Value based RL → only Value function

Policy " " → Policy

Actor critic → Policy + Value

Model Free → no model

Model Based → first craft model

__/__/__

Problems

→ Planning & RL
↓
know everything abt environment before

→ Exploration & Exploitation
(new path) (ur first instinct)

→ Prediction & Control
(evaluate policy) (find best policy)