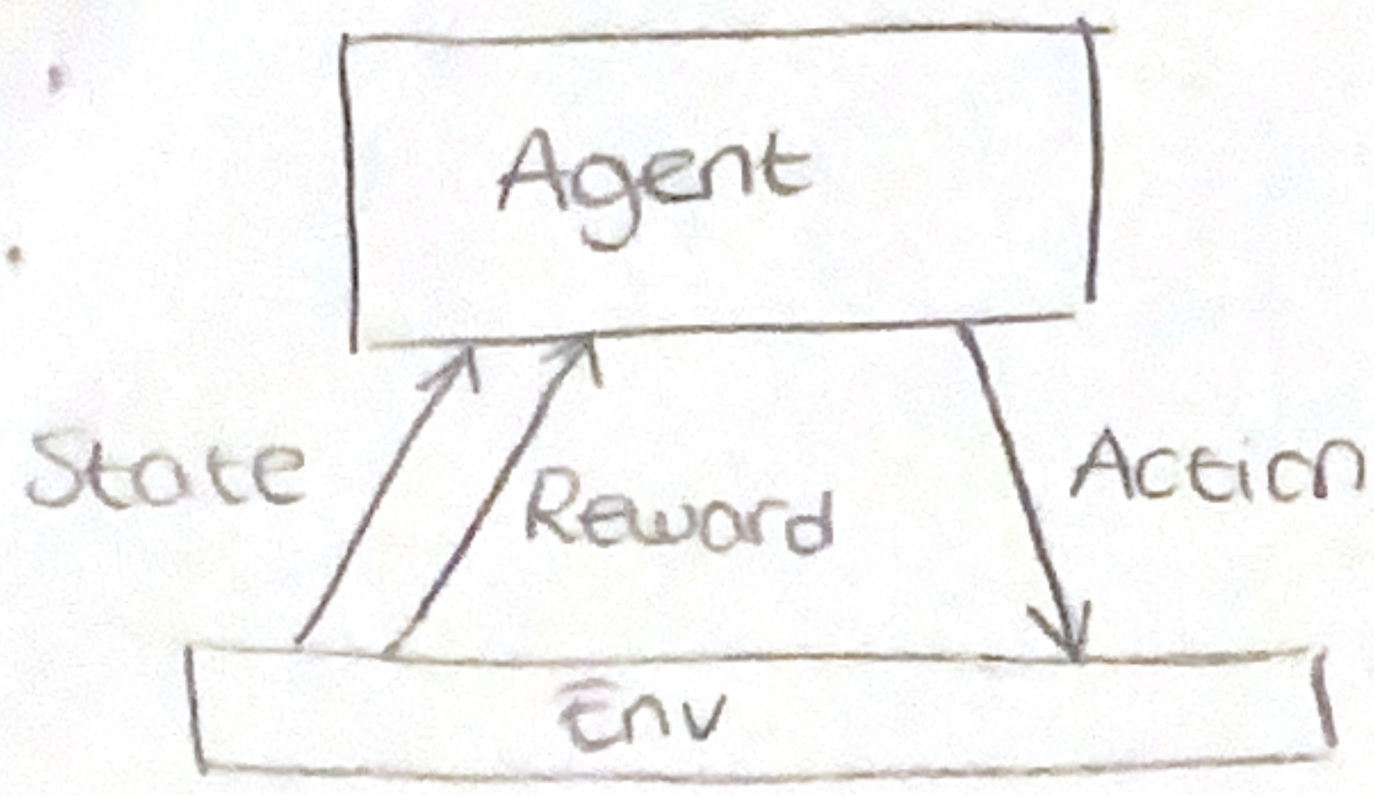
Reinforcement Learning



$s_0 \xrightarrow[r_0]{a_0} s_1 \xrightarrow[r_1]{a_1} \dots$

Agent chooses action a_i in state s_i and gets reward$_i$. But goal is to maximize:

$\underbrace{r_0 + \gamma r_1 + \gamma^2 r_2 + \dots}_{\text{Reward on the long term}}$ $0 \leq \gamma < 1$

γ: discount factor
(a parameter on how much we care about immediate & future rewards)

- Learn a control policy
  $\underset{\text{set of states}}{\pi: S} \rightarrow \underset{\text{set of actions}}{A}$
- Take a from A given current state s from S
- Problems
  - Delayed Reward & Temporal Credit Assignment
    Determine which of the actions in its sequence are to be credited for eventual rewards.
  - Exploration vs Exploitation
    Trade-off in choosing exploration of unknown states & actions (more info) or exploitation of states & actions that are known to yield reward.

The Learning Task

Markov Decision Process

$S \rightarrow$ set of states
$A \rightarrow$ set of actions

At each t, in s_t, perform a_t, get $r(s_t, a_t)$, produce $s_{t+1} = \delta(s_t, a_t)$

These only depend on current state.

- $\pi: S \rightarrow A$
  learn this
- $V^{\pi}(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots = \sum^{\infty} \gamma^i r_{t+i}$
  cumulative reward
  $0 \leq \gamma < 1$
  - When $\gamma = 0 \rightarrow$ only immediate reward is considered
  - When $\gamma \simeq 1 \rightarrow$ Future rewards are more important

Optimal policy is:

$\pi \equiv \operatorname{argmax} V^{\pi}(s)$

pick the policy with the most value