What is RL? → learning what to do in an environment

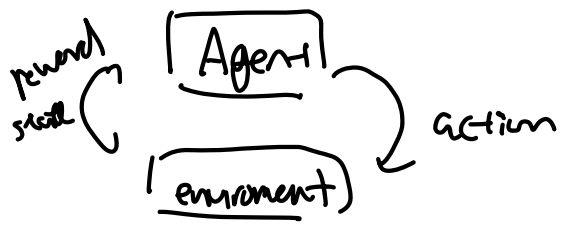
Exploration ✓ Exploitation

Explore: trying out new things

Exploitation: doing things you already know will work well.

how?

→ we want to maximize Reward

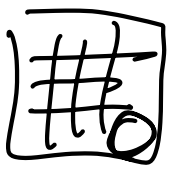


① start at some state S_t ✓ discrete → $t = 0, 1, 2, \ldots$

② pick Action A_t

③ receive S_{t+1}, R_{t+1}

④ repeat



S → states

$A(s)$ ↗ actions

$R \subset \mathbb{R}$ ↗ rewards

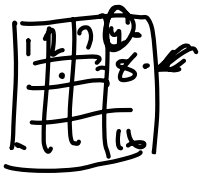
A MDP → markov decision process

- S
- A(s) →
- transition probabilities

$S_0 \to A_0 \to R_1 \to S_1 \to A_1 \to R_2 \to S_2 \dots$ trajectory

R_t & S_t ← random variables → discrete probability distributions

$p(s', r \mid s, a)$: given that I'm at state s, if I take action a, what is the probability that I will end up in s', w/ reward r



$p(8, 50 \mid 4, \downarrow) = 1$

↳ $p(16, 20 \mid 3, \downarrow) = 0$

$\sum_{r \in R} p(s', r \mid s, a)$ = probability of going from state $s \to s'$ taking action a.

transition probability = $p(s' \mid s, a)$

Discounted rewards

$G_t = \sum_{k=0}^{T} \gamma^k R_{t+k+1} = R_{t+1} + \gamma R_{t+1} + \gamma^2 R_{t+2} \dots$

discount factor $0 \le \gamma \le 1$

↑ sum of rewards

Maximize $\mathbb{E}[G_t]$

policies & value function

π : state → action

$V_\pi(s) = \mathbb{E}[G_t | S_t = s] \rightarrow V_{\pi^*} \geq V_\pi$

state - value function

$q_\pi(s,a) = \mathbb{E}[G_t | S_t = s, A_t = a]$ $q_{\pi^*}(s,a) \geq q_\pi(s,a)$

↑ action - value function

$V_\pi(s) = \mathbb{E}[G_t | S_t = s]$

$= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma \cdot V_\pi(s')]$

state - value - Bellman equation

Intuitively

optimize (maximize)

$V_{\pi^*} \geq V_\pi \quad \forall \pi$