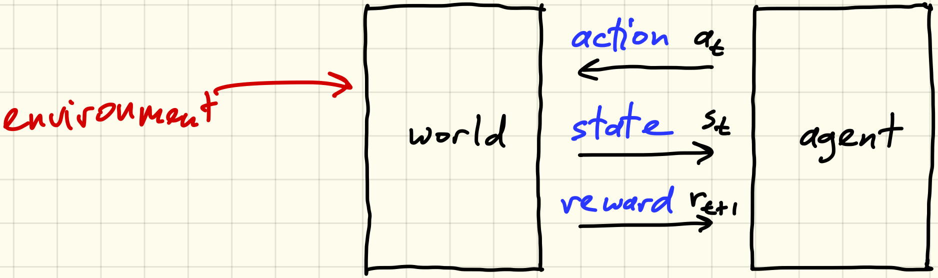
MDP: Markov Decision Process (S&B Ch 3)

- stochastic, discrete state, discrete action, state feedback



t = time

3 system spaces: state $s_t \in S$, action $a_t \in A$, reward $r_t \in \mathbb{R}$

3 system functions: model $p(s_{t+1} \mid s_t, a_t)$ = probability of s_{t+1} given s_t, a_t

policy $\pi(a_t \mid s_t)$ = probability of a_t given s_t

reward $r_{t+1} = R(s_t, a_t)$ ← not stochastic (if so, take $E[\,]$)

Types of ML

supervised learning: given many pairs (x_i, y_i) instruction

learn approximate map $y = f_\theta(x)$

$$y = \theta_0 + \theta_1 x$$

$$\underset{\theta}{\text{minimize}} \sum_i \| y_i - f_\theta(x_i) \|^2$$

unsupervised learning: given many x_i

- clustering, k-means
- autoencoders
- GANs (generative adversarial networks)

self-supervised learning: given sequence $(x_0, x_1, x_2, \ldots, x_n)$

learn model $x_{k+1} = f_\theta(x_k, x_{k-1}, x_{k-2}, \ldots, x_{k-m})$

- OpenAI GPT-3

reinforcement learning : $a_t = \pi_\theta(s_t)$ policy $(s_{t+1}, r_t) = p(s_t, a_t)$

recursive evaluation

$$\underset{\theta}{\text{maximize}} \sum_t r_t$$

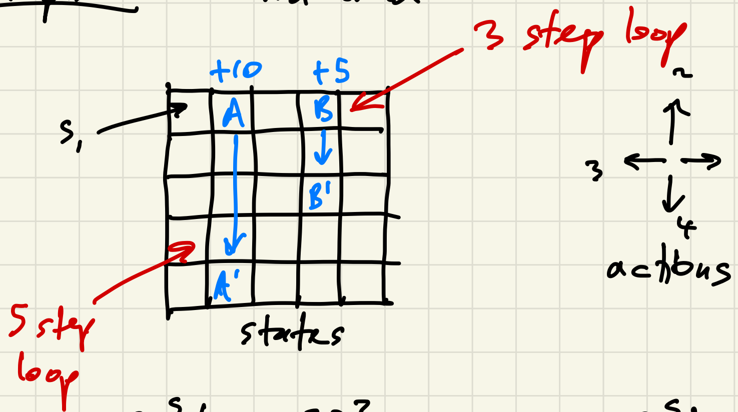
can we call $\nabla p()$

model-based RL : algorithms can directly call $p()$

model-free RL : only have access to $p()$ via trajectories

↳ pros : - can do it on reality
- only need simulator

Example : Gridworld



reward : $r_t = -1$ except at A, B

$s \in \{1, \dots, 25\}$ $\quad$ $a \in \{1, \dots, 4\}$

(0 to 24 in Python)

If the system is time-dependent, we can add time as a state.

Value functions

$$V_\pi(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \;\middle|\; \pi, S_0 = s\right]$$

infinite horizon

policy π

$\leftarrow$ total future discounted reward starting from s under policy π

expected

state value function

discount factor $\gamma < 1$

effective time horizon $\sim \frac{1}{\gamma}$

smaller $\gamma \Rightarrow$ prioritize short-term rewards

typical $\gamma = 0.95,\ 0.99$

(<u>or</u> use finite horizon $V_\pi(s) = E\left[\sum_{t=0}^{T} r_{t+1}\right]$)

finite horizon

$$Q_\pi(s, a) = E\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \;\middle|\; \pi, S_0 = s, A_0 = a\right]$$

state-action value function

Relationships :

det: $V_\pi(s) = Q_\pi(s, \underbrace{\pi(s)}_{a})$

$$V_\pi(s) = E_{a\sim\pi(\cdot|s)}\left[Q_\pi(s,a)\right] = \sum_a \pi(a|s)\, Q_\pi(s,a)$$

$$Q_\pi(s,a) = R(s,a) + \gamma\, E_{s'\sim p(\cdot|s,a)}\left[V_\pi(s')\right]$$

$$= R(s,a) + \gamma \sum_{s'} p(s'|s,a)\, V_\pi(s')$$

det: $Q_\pi(s,a) = R(s,a) + \gamma\, V_\pi(\underbrace{p(s,a)}_{s'})$

recursive relation for V:

$$V_\pi(s) = E_{a\sim\pi(\cdot|s)}\left[R(s,a) + \gamma\, E_{s'\sim p(\cdot|s,a)}\left[V_\pi(s')\right]\right]$$

Optimality:

$$V^*(s) = \max_{\pi} V_\pi(s)$$

$$Q^*(s,a) = \max_{\pi} Q_\pi(s,a)$$

$$\pi^*(a|s) = \begin{cases} 1 & \text{if} \quad a = \operatorname{argmax}_{a'} Q^*(s,a') \\ 0 & \text{otherwise} \end{cases}$$