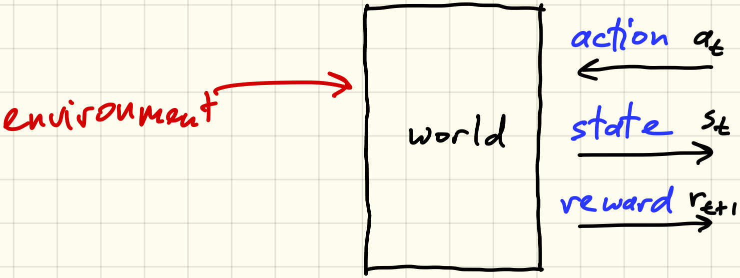
MDP: Markov Decision Process (S&B Ch 3)

- 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 $p(s_{t+1} | s_t, a_t)$ = probability of s_{t+1} given s_t, a_t

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

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

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

recursive evaluation

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

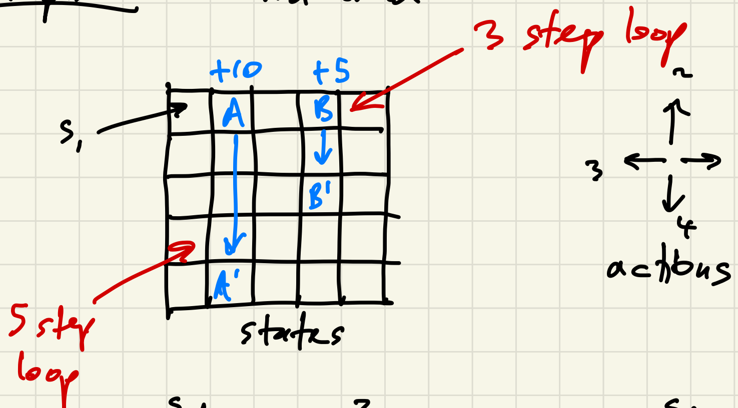
can we call $\nabla p()$

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

$s \in \{1, \dots, 25\}$

(0 to 24 in Python)

$a \in \{1, \dots, 4\}$

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

Value functions

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

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

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

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

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

$\pi(s)$ if π is deterministic

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

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

Deterministic policy

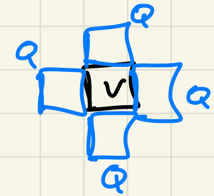
$a = \pi(s)$

$$\pi(a|s) = \begin{cases} 1 & \text{if } a = a' \\ 0 & \text{otherwise} \end{cases}$$

$V_\pi(s) = Q(s, \pi(s))$

$a \quad \propto$

$$V^*(s) = \max_a Q^*(s,a)$$



$$V^*(s) = \max_a \left(R(s,a) + \gamma \, \mathbb{E}_{s' \sim p(\cdot|s,a)} \left[V^*(s') \right] \right)$$

Bellman optimality equation

$$Q^*(s,a) = R(s,a) + \gamma \, \mathbb{E}_{s' \sim p(\cdot|s,a)} \left[\max_{a'} Q^*(s',a') \right]$$

Policy evaluation (S&B 4.1): prediction $\quad \pi \to V \quad$ (know p)

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

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

$s \in \{1, \dots, 25\}$

$a \in \{1, \dots, 4\}$

$$\pi = \begin{array}{cc} \begin{bmatrix} x & x & x & x \\ & \pi(16,2) & & \\ & \bullet & & \\ & & & \\ & & & \end{bmatrix} & 25 \\ 4 & \end{array} \qquad V = \begin{array}{cc} \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix} & 25 \\ 1 & \end{array}$$

$$V_\pi(s) - \gamma \sum_{s'} \underbrace{\left(\sum_a \pi(a|s)\, p(s'|s,a) \right)}_{A(s,s') \,\leftarrow\, 25 \times 25} V_\pi(s') = \underbrace{\sum_a \pi(a|s) R(s,a)}_{b(s) \,\leftarrow\, 25 \times 1}$$

$$(\underbrace{I - \gamma A}_{25 \times 25}) V = \underbrace{b}_{25 \times 1}$$

$$(I - A)x = b$$

$$\underbrace{\begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}\begin{bmatrix} x \end{bmatrix}} - \begin{bmatrix} & A & \end{bmatrix}\begin{bmatrix} x \end{bmatrix} = \begin{bmatrix} b \end{bmatrix} \leftarrow i\text{th}$$

$$x_i - \sum_j A_{ij} x_j = b_i$$

$$V(s) - \gamma \sum_{s'} A(s, s') V(s') = b(s)$$

$$y = Ax$$

$$i\begin{bmatrix} y \end{bmatrix} = i\begin{bmatrix} & \overset{j}{A} & \end{bmatrix}\begin{bmatrix} x \end{bmatrix} j$$

Iterative solution: $v_k(s) \longmapsto v_{k+1}(s)$

k: iteration number

$$v_{k+1}(s) = E_{a\sim\pi(\cdot|s)}\left[R(s,a) + \gamma\, E_{s'\sim p(\cdot|a,s)}\left[v_k(s')\right]\right] \quad \forall s$$

$$v_{k+1} = b + \gamma A v_k$$

(same as $(I - \gamma A)v = b$)

Can prove this converges if $\gamma < 1$

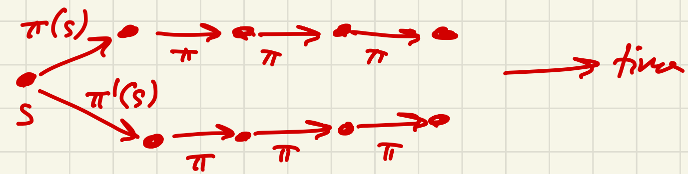
Policy improvement (S&B 4.2) $\pi \to \pi'$

Policy improvement theorem Given π, π' deterministic policies

If $Q_\pi(s, \pi'(s)) \geq V_\pi(s) \quad \forall s$ Better 1-step decisions

then $V_{\pi'}(s) \geq V_\pi(s) \quad \forall s$ $\Rightarrow$ better overall performance

Recall $Q_\pi(s, \pi(s)) = V_\pi(s)$



Pf Unroll:

$$
\begin{aligned}
V_\pi(s) &\leq Q_\pi(\underbrace{s}_{S_t}, \underbrace{\pi'(s)}_{A_t}) \\
&= E[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = \pi'(s)] \\
&= E_{\pi'}[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s] \\
&\leq E_{\pi'}[R_{t+1} + \gamma Q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s] \\
&= E_{\pi'}[R_{t+1} + \gamma E[R_{t+2} + \gamma V_\pi(S_{t+2}) \mid A_{t+1} = \pi'(S_{t+1})] \mid S_t = s] \\
&= E_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 V_\pi(S_{t+2}) \mid S_t = s] = \dots = V_{\pi'}(s)
\end{aligned}
$$

Greedy policy: $\pi'(s) = \arg\max_a Q_\pi(s,a)$

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

Observe that $Q_\pi(s, \pi'(s)) \geq Q_\pi(s,a) \quad \forall a$

$\implies \quad Q_\pi(s, \pi'(s)) \geq V_\pi(s)$ (by thm $\Rightarrow V_{\pi'}(s) \geq V_\pi(s)$ for all s)

What if $\pi'(s) = \pi(s) \quad \forall s$?

$\implies \quad V_{\pi'}(s) = \max_a E_{s' \sim p(\cdot|s,a)}[R(s,a) + \gamma V_\pi(s')]$ ← Bellman

Thus if greedy policy makes no improvement, π must already be optimal.

Policy iteration (S & B 4.3)

$$\pi_0 \xrightarrow{E} V_{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} V_{\pi_1} \xrightarrow{I} \pi_2 \rightarrow \dots \rightarrow \pi^* \rightarrow V^*$$

evaluation alg. (iterative)

improvement

π_1 is the 1-step greedy policy from V_{π_0}

Q: how many iterations of evaluation should we do before improving the policy?

Value iteration (S&B 4.4)

Key idea: do policy improvement but with a single evaluation sweep

$$V_{k+1}(s) = \max_a \left(R(s,a) + \gamma \, \mathbb{E}_{s' \sim p(\cdot|s,a)} [V_k(s')] \right) \quad \forall s$$

$$= \max_a \sum_{s'} \left(R(s,a) + \gamma \, p(s'|s,a) \, V_k(s') \right)$$