Finite Markov Decision Process
(MDPs)

evaluative feedback

associative aspect (choose different actions in different situations)

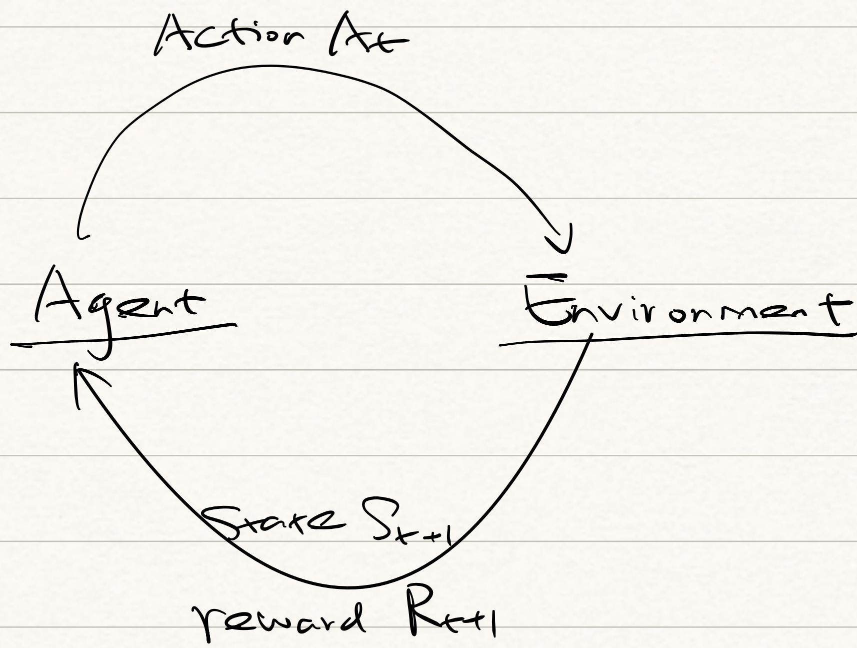
MDPs : sequential decision making

bandit : $q_*(a)$ w.r.t. action a

MDPs : $q_*(s, a)$ w.r.t. $\begin{cases} \text{action } a \\ \text{state } s \end{cases}$

$V_*(s)$ w.r.t. state s given optimal action selections

Agent-Environment Interface



Sequence / trajectory :

$$\boxed{S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots}$$

$$p : S \times R \times S \times A \to [0, 1]$$

$$\boxed{p(s', r \mid s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\}}$$

$$\sum_{s' \in S} \sum_{r \in R} p(s', r \mid s, a) = 1$$

for all $s \in S$, $a \in A(s)$

"Markov Property"

"state-transition"

probabilities

$p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0,1]$

$$p(s'|s,a) \doteq \Pr\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\}$$

$$= \sum_{r \in \mathcal{R}} p(s', r \mid s, a)$$

Expected rewards for state-action pairs

$r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$r(s,a) \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a]$$

$$= \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r \mid s, a)$$

Expected rewards for state-action-next-state triples

$r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$

$$r(s,a,s') \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s']$$

$$= \sum_{r \in \mathcal{R}} r \frac{p(s', r \mid s, a)}{p(s' \mid s, a)}$$

Reward

$R_t \in \mathbb{R}$

Goal : maximize cumulative reward in the long run

Expected return G_t

$$G_t \doteq R_{t+1} + R_{t+2} + \cdots + R_T$$

subsequences $\longrightarrow$ episodes

$\mathcal{S}$

Starting state $\longrightarrow$ terminal state

$\mathcal{S}^+$

$\updownarrow$

episodic task

Discounted Return :

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

$$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

γ : discount rate $\quad \gamma \in [0, 1]$

$\gamma = 0$: Agent "myopic" "immediate rewards"

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots$$

$$= R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \cdots)$$

$$= R_{t+1} + \gamma G_{t+1}$$

$$G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$$

defined w.r.t. particular ways of acting

called Policy

Value Functions

policy: a mapping from states to probabilities of
π
selecting each possible action

$\pi(a|s)$: the probability that $A_t = a$ if $S_t = s$

Value function of a state s under a policy π

state-value function for policy π

$$v_\pi(s) \doteq \mathbb{E}[G_t \mid S_t = s]$$
$$= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right]$$

action-value function for policy π

$$q_\pi(s,a) \doteq \mathbb{E}[G_t \mid S_t = s, A_t = a]$$
$$= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

$v_\pi(s)$

$$\doteq \mathbb{E}_\pi[G_t \mid S_t = s]$$

$$= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

$$= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r \mid s, a)\left[r + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s']\right]$$

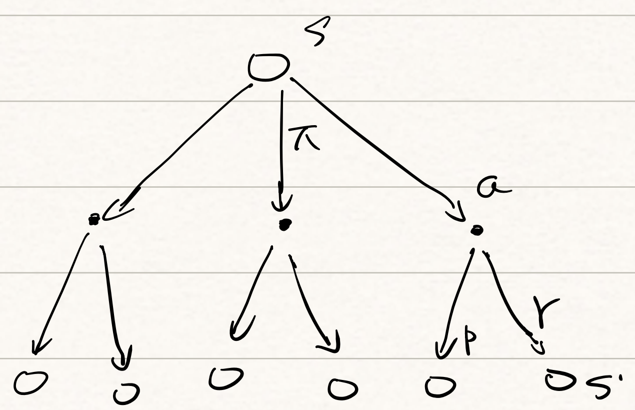
$$= \sum_a \pi(a|s) \sum_{s', r} p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right]$$

Bellman equation for v_π

$$\boxed{v_\pi(s) = \sum_a \pi(a|s) \sum_{s', r} p(s', r \mid s, a)\left[r + \gamma v_\pi(s')\right]}$$

transfer value information back to a state (or a state-action pair) from its successor states (or state-action pairs) "Bootstrap"

Backup diagram for v_π



Optimal Policies / Optimal Value Functions

optimal policy: π_*

optimal state-value function v_*

$$v_*(s) \doteq \max_{\pi} v_\pi(s)$$

optimal action-value function q_*

$$q_*(s,a) \doteq \max_{\pi} q_\pi(s,a)$$

$$= \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right]$$

Bellman Optimality Equation

$$v_*(s) = \max_{a \in A(s)} q_{\pi_*}(s,a)$$

$$= \max_a \mathbb{E}_{\pi_*}\left[G_t \mid S_t = s, A_t = a\right]$$

$$= \max_a \mathbb{E}_{\pi_*}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right]$$

$$\boxed{v_*(s) = \max_a \mathbb{E}\left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a\right]}$$

$$v_*(s) = \max_a \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma v_*(s')\right]$$

Bellman optimality equation for v_*

$$q_*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \,\middle|\, S_t = s, A_t = a\right]$$

$$= \sum_{s', r} p(s', r \mid s, a)\left[r + \gamma \max_{a'} q_*(s', a')\right]$$

Bellman optimality equation for q_*