Reinforcement Learning

reward r_t

Environment $\xrightarrow{\text{reward } r_t}$ Agent

state s_t $\xleftarrow{\text{action } a_t}$

$r_t = r(\underbrace{s_t, a_t})$

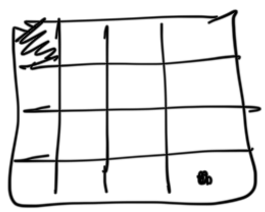
or

$r(\underbrace{s_1 \ldots s_t}_{\tilde{s}_t}, \underbrace{a_1 \ldots a_t}_{\tilde{a}_t})$

$\Big[\; s_{t+1} \leftarrow s_t, a_t$

Environment Model.

eg. grid world



s_t = location / neighborhood of agent

a_t = up / down / left / right

r_t = 1

Markov Decision Process

state space S

action space A

initial state distribution $d_0 \in \Delta(S)$

state transition probability $P: S \times A \rightarrow \Delta(S)$ $\quad P(s' \mid s, a)$

reward function $R: S \times A \longrightarrow [0, R_{max}=1]$ / $\Delta[0,1]$

discount factor $\gamma \in [0,1]$ how much future rewards are discounted

or

time horizon H

MDP $(S, A, d_0, P, R, \gamma/H)$

Interaction protocol

Env. state $s_1 \sim d_0$

For $t = 1, 2, \ldots$

agent takes action $a_t \in A$

attains rewards $r_t = R(s_t, a_t)$

observes next state of env $s_{t+1} \sim P(s_t, a_t)$

(1) Planning : Given state model, find desired policy

(2) learning : Given trajectories observation, learn desired policy.

deterministic policy $\pi : S \to A$ $\quad a_t = \pi(s_t)$

stochastic " $\pi : S \to \Delta(A)$ $\quad a_t \sim \pi(s_t)$

<u>Goal</u>: Choose policy π to max expected discounted sum of rewards starting at state s_1

$$E\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid \pi, s_1\right] \quad \text{or} \quad E\left[\sum_{t=1}^{H} r_t \mid \pi, s_1\right]$$

<u>Note</u>: $$0 < \sum_{t=1}^{\infty} \gamma^{t-1} r_t \leq R_{max} \sum_{t=1}^{\infty} \gamma^{t-1} = \frac{R_{max}}{1-\gamma} \approx R_{max} H$$

effective horizon $\frac{1}{1-\gamma} \approx H$

<u>Value function</u> $$V_\pi(s) = E\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid \pi, s_1 = s\right] \quad \forall s \in S.$$

Value of policy π starting at state s

<u>Action-value or Q function</u>

$$Q_\pi(s,a) = E\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid \pi, s_1 = s, a_1 = a\right]$$

<u>Bellman equations for policy evaluation</u>

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

$$Q_\pi(s,a) = R(s,a) + \underbrace{E\left[\sum_{t=2}^{\infty} \gamma^{t-1} r_t \mid \pi, s_1 = s, a_1 = a\right]}$$

$$= \gamma\, E_{s' \sim P(\cdot|s,a)}\left[E\left[\sum_{t=2}^{\infty} \gamma^{t-2} r_t \mid \pi, s_2 = s'\right]\right]$$

$$= \gamma\, E_{s' \sim P(\cdot|s,a)}\left[\underbrace{E\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid \pi, s_1 = s'\right]}_{V_\pi(s') = Q_\pi(s', \pi(s'))}\right]$$

If S, A finite, write as linear equations in matrix-vector form

$$\underset{|S\times A|\times 1}{Q_\pi} = \underset{|S\times A|\times 1}{R} + \gamma \underset{|S\times A|\times|S|}{\underline{P}} \underset{|S|\times 1}{V_\pi} \qquad P = P(s', a' | s)$$

Bellman eq^s

$$\underset{|S\times A|\times 1}{Q_\pi} = R + \gamma \underset{|S\times A|\times|S\times A|}{P_\pi} \underset{|S\times A|\times 1}{Q_\pi}$$

$$(P_\pi)_{(s,a),(s',a')} = P(s', a' | s, a) = P(s' | s, a) \cdot \pi(a' | s').$$

set of $|S\times A|$ linear equations

$$Q_\pi - \gamma P_\pi Q_\pi = R$$

Bellman sol^n

$$\underline{\underline{Q_\pi}} = \underbrace{(I - \gamma P_\pi)^{-1}}R =$$

$$x^T (I - \gamma P_\pi) x = x^T x - \gamma x^T P_\pi x \geq (1-\gamma) x^T x > 0$$

→ State-action value function is linear in R.

→ rows of $(I - \gamma P_\pi)^{-1}$ - exp no. of times policy π visits each state, action pair

state. action occupancy distribution $\underline{\underline{d^{\pi, s}}} = \underline{(I - \gamma P_\pi)^{-1}} (1-\gamma)$

$$1^T (I - \gamma P_\pi)^{-1} = 1^T \sum_{t=1}^{\infty} \gamma^{t-1} (P_\pi)^{t-1} = \sum_{t=1}^{\infty} \gamma^{t-1} \underbrace{1^T (P_\pi)^{t-1}}_{1}$$

$$= \frac{1}{\underline{1-\gamma}} \underline{1}$$

<u>Bellman Optimality</u> $\pi^* = \arg\max_\pi V_\pi(s)$ ←

<u>Thm</u>: There always exists a stationary and deterministic policy π^* that simultaneously maximizes $V_\pi(s)$ $\forall s \in S$ & $Q_\pi(s,a)$ $\forall s \in S, a \in A$.

$$V^*(s) = \max_{a \in A} \underbrace{Q^*(s,a)}$$

<u>Bellman Optimality Eq^ns</u>

$$V^*(s) = \underline{\underline{\max_{a\in A}}} \underbrace{[R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s')]} \quad ✓$$

nonlinear eq^s. $\qquad V^* = T_V V^*$

If we know V^*/Q^* can find π^*

$$\Longrightarrow \quad \pi^*(s) = \underset{a \in A}{\operatorname{argmax}} \; R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s')$$

$$Q^*(s,a) = \underbrace{R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \max_{a \in A} Q^*(s',a)} \quad \checkmark$$

$$\to \quad \underline{Q^* = T_Q Q^*} \qquad \text{nonlinear eq}^n\text{s}$$

Thm: Q is optimal iff satisfies $Q = TQ$

Proof: Sufficiency - by construction of T.

necessity - if $Q = TQ$ then $Q = Q^*$

$$\to \quad \underline{Q = TQ} = R + \gamma P_\pi Q \qquad \text{where } \pi = \underset{a}{\operatorname{argmax}} \, Q(s,a) \;\; \forall s$$

$$=: \pi_Q$$

$$= (I - \gamma P_\pi)^{-1} R$$

$$\to \quad [P_\pi Q - P_{\pi'} Q]_{s,a} = E_{s' \sim P(\cdot|s,a)} [Q(s', \pi(s')) - Q(s', \pi'(s'))] \geq 0$$

$$\because \pi = \pi_Q.$$

$$Q - Q_{\pi'} = (I - \gamma P_\pi)^{-1} R - (I - \gamma P_{\pi'})^{-1} R$$

$$= (I - \gamma P_{\pi'})^{-1} [(I - \gamma P_{\pi'}) - (I - \gamma P_\pi)] \underset{(I-\gamma P_\pi)^{-1} R}{\overset{\shortparallel}{Q}}$$

$$= \underset{\geq 0}{\gamma} \; \underbrace{(I - \gamma P_{\pi'})^{-1}}_{\geq 0} \; \underbrace{(P_\pi - P_{\pi'}) Q}_{\geq 0}$$

$$Q \geq Q_{\pi'} \quad \forall \pi'$$

Non-stationary (Time dependent) MDP . finite horizon.

$$MDP(S, A, \{P_h\}_{h \in 1..H}, \{R_h\}_{h \in 1..H}, H)$$

$P_h : S \times A \to \Delta(S)$ $\qquad P_h(s'|s,a)$ at each time step h.

$R_h \qquad V_h^\pi(s) \qquad Q_h^\pi(s,a) \qquad \pi_h^* = \underset{a}{\operatorname{argmax}} \, Q_h^{*\pi}(s,a)$