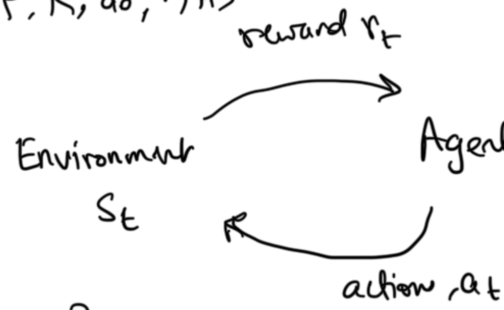
Reinforcement Learning

MDP $(S, A, P, R, d_0, \gamma/H)$ $\qquad r_t \leftarrow s_t, a_t$



$s_1 \sim d_0 \qquad s_{t+1} \xleftarrow{P} s_t, a_t$

Goal: find policy $\pi : S \to a$ $\quad E\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid \pi, s\right]$

Planning problem: Given state transition + reward models, find optimal policy

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

Bellman optimality equations, optimal Q^*, V^*

$$V^*(s) = \max_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s') \right] \equiv T_v V^*$$

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

Challenge: Nonlinear equations (bcoz of max)

① Dynamic Programming / Iterative solutions $\quad$ finite S, A

Policy Iteration

Initialize π_0

1) Policy evaluation $Q_{\pi_{k-1}}$ $\quad$ (linear Bellman eq's)

2) Policy improvement $\pi_k = \pi_{Q_{\pi_{k-1}}}$ $\quad \leftarrow \Rightarrow A_{\pi_k}(s, \pi_{k+1}) \geq 0$

Guaranteed to improve montonically, hence terminate $Q_{\pi_k} = Q_{\pi_{k-1}}$.

$\rightarrow$ # policies $= |A|^{|S|}$

Policy Improvement Theorem: In policy iteration, $V_{\pi_k}(s) \geq V_{\pi_{k-1}}(s)$.

$\forall k \geq 1,\ s \in S,\ a \in A$

& strictly positive in atleast 1 state until π^* is found.

Proof: Advantage of action a in state s over policy π is defined

as $A_\pi(s,a) = Q_\pi(s,a) - Q_\pi(s,\pi(s)) = Q_\pi(s,a) - V_\pi(s)$

Advantage of policy π' over π $\quad A_\pi(s,\pi') := A_\pi(s,\pi'(s))$

Performance Difference Lemma For any π, π' & $s \in S$

$$V_{\pi'}(s) - V_\pi(s) = \frac{1}{1-\gamma} E_{\underbrace{s' \sim d_{\pi',s}}_{}} [A_\pi(s',\pi')]$$

↳ normalized discounted occupancy induced by π' starting in state s.

Proof: Consider a seq^n of policies $\{\pi_i\}_{i=0}^{\infty}$

$\pi_0 = \pi$

$\pi_\infty = \pi'$

π_i follow π' for first i steps, then switch to π.

$$\begin{aligned}
V_{\pi'}(s) - V_\pi(s) &= V_{\pi_\infty}(s) - V_{\pi_0}(s) \\
&= \sum_{i=0}^{\infty} \left(V_{\pi_{i+1}}(s) - V_{\pi_i}(s) \right) \\
&= \sum_{i=0}^{\infty} \Big(E\Big[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\Big|\, \pi_{i+1}, s_1 = s\Big] \\
&\qquad - E\Big[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\Big|\, \pi_i, s_1 = s\Big]\Big)
\end{aligned}$$

Note: π_{i+1} & π_i only deviate at step $i+1$.

Same roll-in policy π' for first i steps

$\Rightarrow$ defines roll-in distribution $P(s_{i+1} \mid s_1 = s, \pi')$

Same roll out policy π starting at $i+2$

$\Rightarrow$ conditioned on $s_{i+1} = s$, $a_{i+1} = a$ total expected discounted reward picked up in rest of trajectory is

$$\gamma^i Q_\pi(s,a)$$

$$\Rightarrow \quad V_{\pi'}(s) - V_\pi(s) = \sum_{i=0} \gamma^i \sum_{s' \in S} P(s_{i+1} = s' \mid s_1 = s, \pi') \underbrace{\left(Q_\pi(s', \pi'(s')) - Q_\pi(s', \pi(s')) \right)}_{A_\pi(s', \pi')}$$

$$= \sum_{s' \in S} P(s_{i+1} = s' \mid s_1 = s, \pi') \, A_\pi(s', \pi') \cdot \frac{1}{1-\gamma}$$

$$= \mathbb{E}_{s' \sim d_{\pi', s}} [A_\pi(s', \pi')] \frac{1}{1-\gamma} .$$

Proof of Policy improvement thm

Invoke policy diff^n lemma $\pi' = \pi_{k+1}$ & $\pi = \pi_k$.

$\Rightarrow V_{\pi_{k+1}} \geq V_{\pi_k}$ $\because$ $A_{\pi_k}(s, \pi_{k+1}) \geq 0 \quad \forall s.$

due to policy improvement step

∎

Worst case $\underbrace{|A|^{|S|}}$ but in practice works better.

Often approximate sol^n suffices .

Thm: Policy iteration converges exponentially in sup-norm

$$\| Q_{\pi_{k+1}} - Q^* \|_\infty \leq \gamma \| Q_{\pi_k} - Q^* \|_\infty$$

Contraction property of T_Q :

T_Q is a γ-contraction in sup. norm

$$\| TQ - TQ' \|_\infty \leq \gamma \| Q - Q' \|_\infty \qquad \forall Q, Q'$$

Proof: $\| TQ - TQ' \|_\infty$ with $TQ \to (s,a)$, $TQ' \to (s,a)$:

$$\| TQ - TQ' \|_\infty = \gamma \max_{s,a} \Big| \mathbb{E}_{s' \sim P(\cdot|s,a)} [\underbrace{\max_{a'} Q(s', a')}_{V(s')}] - \mathbb{E}_{s' \sim P(\cdot|s,a)} [\max_{a'} Q'(s', a')] \Big|$$

$|\mathbb{E}| \leq \mathbb{E}|\ |$

$$\leq \gamma \max_{s,a} \mathbb{E}_{s' \sim P(\cdot|s,a)} \left| \max_{a'} Q(s', a') - \max_{a'} Q'(s', a') \right|$$

$|\max| \leq \max|\ |$

$$\leq \gamma \max_{s,a} \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} | Q(s', a') - Q'(s', a') |$$

$$\leq \gamma \|Q - Q'\|_\infty \quad \boxed{}$$

Proof of exp conv of policy iteration.

$$Q_{\pi_{k+1}} \geq TQ_{\pi_k} \quad **$$

Using this, $\quad Q^* - Q_{\pi_{k+1}} \leq TQ^* - TQ_{\pi_k}$

$$\Rightarrow \|Q^* - Q_{\pi_{k+1}}\| \leq \gamma \|Q^* - Q_{\pi_k}\|_\infty \quad \text{contraction of } T$$

Now lets prove $**$

$$Q_{\pi_{k+1}}(s,a) = R(s,a) + \gamma\, E_{s'\sim P(\cdot|s,a)} V_{\pi_{k+1}}(s')$$

$$\geq R(s,a) + \gamma\, E_{s'\sim P(\cdot|s,a)} V_{\pi_k}(s') \quad \text{Policy improvement thm}$$

$$\geq R(s,a) + \gamma\, E_{s'\sim P(\cdot|s,a)} \max_{a'} Q_{\pi_k}(s',a') = TQ_{\pi_k} \quad \boxed{}$$

Value Iteration

Approx Q/V directly using a seq of Q function (without going between Q & π)

Initialize Q_0

$$Q_t = TQ_{t-1} \qquad Q_{\pi_{k+1}} \geq TQ_{\pi_k} \text{ (policy iteration)}$$

Contraction property implies Q convgs to Q^* exponentially.

Translate to policy value error: $\quad \pi_Q(s) = a = \arg\max_{a'} Q(s,a')$

$$V^*(s) - V_{\pi_Q}(s) = Q^*(s,\pi^*(s)) - Q_{\pi_Q}(s, \pi_Q(s)) \qquad \text{Note: } Q_{\pi_Q} \neq Q$$

$$= \underbrace{Q^*(s,\pi^*(s)) - Q(s,a)}_{1} + \underbrace{Q(s,a) - Q^*(s,a)}_{2} + \underbrace{Q^*(s,a) - Q_{\pi_Q}(s,a)}_{3}$$

$$1 \quad \leq Q^*(s,\pi^*(s)) - Q(s,\pi^*(s)) \qquad \because a = \arg\max_{a'} Q(s,a')$$

$$\leq \|Q - Q^*\|_\infty$$

$$2 \quad \leq \|Q - Q^*\|_\infty$$

$$3 \quad \leq \gamma\, E_{s'\sim P(\cdot|s,a)} [V^*(s') - V_{\pi_Q}(s')]$$

$$\leq 2\|Q^* - Q\|_\infty + \gamma\, E_{s'\sim P(\cdot|s,a)} [V^*(s') - V_{\pi_Q}(s')]$$

$$\Rightarrow \underset{Q \to Q_t}{\cancel{(\ldots)}} \| V^*(s) - V_{\pi_Q}(s) \|_\infty \le \frac{2\|Q^* - Q\|_\infty}{1-\gamma} \le \frac{2\gamma^t R_{max}}{1-\gamma}$$

■

② Linear Programing

$$\min_V \quad d_0^T V$$

$$\text{s.t.} \quad V \geq TV \qquad\qquad V = [V(s)]_{s \in S}$$

S ~~equations~~ constraints (nonlinear)

$\max_a$

$\downarrow$

$S \times A$ constraints (linear)