<u>RL</u> $MDP(\overset{d_0,}{S}, A, P, R, H/\gamma)$

$\pi^* = \arg\max_{\pi} E[V(\pi)]$ $\qquad V(\pi) = \sum_{t=1}^{T} \sum_{h=1}^{H} r_h^t$

<u>Planning</u>: Given $\underline{\underline{P}}, R, d_0$ $\qquad d_0 : S \to \mathbb{R}$

find π^* $\qquad R : S \times A \to \mathbb{R}$

$P : S \times A \to S$ (drives complexity of learning)

Dynamic Prog

/ → Policy iteration $\qquad \pi$ random

Q_π evaluate policy

π_{Q_π} improvement policy

✓ → Value iteration $\qquad Q_0$ random

$\pi_k = TQ_{k-1}$ improve policy

$\pi_k \leftarrow$ greedy Q_{π_k}

exp conv. $\qquad \|Q_k - Q^*\| \le \gamma^k \|Q_0 - Q^*\|$

value $\qquad \|V_k - V^*\| \;\; \simeq \;\; \|Q_k - Q^*\| \;\; \le \gamma^k \|Q_0 - Q^*\|$

$Q_{\pi_k} \neq Q_k$

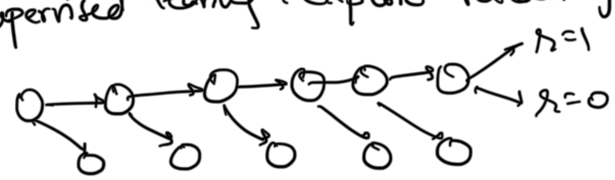
✓ Linear Prog.

<u>Learning Policy</u>

Don't know state transitions $\underline{\underline{P}}$ (rewards R, initial state distⁿ d_0)

Need to "explore" to learn $\underline{\underline{P}}, R, d_0$.

Why not use supervised learning (explore randomly?



$\underline{\underline{2^H}}$ episodes

Episodic MDP $\qquad M = (S, A, P, R, H)$ $\qquad$ starting state ↑ reward $\qquad$ deterministic ↑ known $\quad s_1$

Episode t generates trajectory $\tau_t = \{s_1^t, a_1^t, r_1^t, \ldots, s_H^t, a_H^t, r_H^t\}$

$$\text{Regret}(T) = T\,\mathbb{E}[V_{\pi^*}(s)] - \mathbb{E}\Big[\sum_{t=1}^{T}\sum_{h=1}^{H} r_h^t\Big]$$
$$= \sum_{t=1}^{T}\big(\mathbb{E}[V_{\pi^*}(s)] - \mathbb{E}[V_{\pi_t}(s)]\big)$$

policy π_t deploy at time t.

(1) Natural idea

1) estimate $\hat{P}_t$ using trajectories in episode t
2) plan $\hat{\pi}$ using $\hat{P}$
3) deploy $\hat{\pi}$ next episode

greedy approach $\equiv$ ε-greedy

MAB
$$\mu(a^*) - \mu(a_t) \le \underbrace{\mu(a^*) - \hat{\mu}_t(a^*)} + \underbrace{\overset{\le 0}{\hat{\mu}_t(a^*) - \hat{\mu}_t(a_t)}}_{\text{greedy}} + \underbrace{\hat{\mu}_t(a_t) - \mu(a_t)}$$

$a_t = \arg\max_a \hat{\mu}_t(a)$

$$\le \sigma_t(a^*) - \sigma_t(a_t) \quad ||$$

may not shrink $\qquad$ shrinks

(2) UCB-VI (Upper Confidence Bound - Value Iteration)

MAB
$$\underbrace{\mu(a^*) - \mu(a_t)} \le \underbrace{\hat{\mu}_t(a^*) + \sigma_t(a^*)} - \hat{\mu}_t(a_t) + \sigma_t(a_t)$$
$$\underset{\text{UCB}}{\le} \cancel{\hat{\mu}_t(a_t)} + \sigma_t(a_t) - \cancel{\hat{\mu}_t(a_t)} + \sigma_t(a_t)$$
$$= 2\sigma_t(a_t) \quad || \qquad \sigma_t(a) \sim \sqrt{\frac{\log T}{n_t(a)}}$$

shrinks

Goal: $\rightarrow$ $\underbrace{\mathbb{E}[V_{\pi^*}(s)] - \mathbb{E}[V_{\pi}(s)]} \le$ confidence (shrink)

compute using data

Optimistic Regret Decomposition

$$\mathbb{E}[V_{\pi^*}(s)] - \mathbb{E}[V_{\bar{\pi}}(s)] \le \sum_{h=1}^{H} \mathbb{E}_{(s_h, a_h)\sim d_h^{\bar{\pi}}}[\text{conf}_h(s_h, a_h)]$$

$\rightarrow$ if $\bar{\pi}$ is greedy policy corresponding to $\bar{Q}$: $\bar{\pi} = \arg\max_a \bar{Q}(s,a)$

when $\quad Q_h^*(s,a) \le \bar{Q}_h(s,a) \le T\bar{Q}_{h+1}(s,a) + \text{conf}_h(s,a)$

$\forall h, s, a$ Optimistic nearly Bellman consistent

Proof: Assuming $\bar{Q}$ exists with these properties,

$$E[V_{\pi^*}(s)] - E[V_{\bar{\pi}}(s)] = V_1^*(s_1) - V_1^{\bar{\pi}}(s_1) \quad s_1 \text{ deterministic}$$

$$= Q_1^*(s_1, \pi^*(s_1)) - Q_1^{\bar{\pi}}(s_1, \bar{\pi}(s_1))$$

optimistic $\bar{Q}$ $\leq \bar{Q}_1(s_1, \pi^*(s_1)) - Q_1^{\bar{\pi}}(s_1, \bar{\pi}(s_1))$

bcos $\bar{\pi}$ is greedy policy for $\bar{Q}$ $\leq \bar{Q}_1(s_1, \bar{\pi}(s_1)) - Q_1^{\bar{\pi}}(s_1, \bar{\pi}(s_1))$ — (a)

nearly Bellman consistency of $\bar{Q}$ $\leq T\bar{Q}_2(s_1, \bar{\pi}(s_1)) + \text{conf}_1(s_1, \bar{\pi}(s_1)) - Q_1^{\bar{\pi}}(s_1, \bar{\pi}(s_1))$

$$= \cancel{R(s_1, \bar{\pi}(s_1))} + \underbrace{E_{s_2 \sim P(\cdot|s_1, \bar{\pi}(s_1))}[\bar{Q}_2(s_2, \bar{\pi}(s_2))]}_{\text{Bellman operator when } \bar{\pi} \text{ greedy for } \bar{Q}} + \text{conf}_1(s_1, \bar{\pi}(s_1))$$

$$\underbrace{- \cancel{R(s_1, \bar{\pi}(s_1))} - E_{s_2 \sim P(\cdot|s_1, \bar{\pi}(s_1))}[Q_2^{\bar{\pi}}(s_2, \bar{\pi}(s_2))]}_{\text{linear Bellman eq}^n}$$

(b) $$= E_{s_2 \sim P(\cdot|s_1, \bar{\pi}(s_1))}[\bar{Q}_2(s_2, \bar{\pi}(s_2)) - Q_2^{\bar{\pi}}(s_2, \bar{\pi}(s_2))] + E_{(s_1, a_1) \sim d_1^{\bar{\pi}}}[\text{conf}_1(s_1, a_1)]$$

deterministic

Notice that (a) - (b) form a recursion

$$E[V_{\pi^*}(s)] - E[V_{\bar{\pi}}(s)] \leq \sum_{h=1}^{H} E_{(s_h, a_h) \sim d_h^{\bar{\pi}}}[\text{conf}_h(s_h, a_h)]$$

$\because \bar{Q}_{H+1} = 0$ & $Q_{H+1}^{\bar{\pi}} = 0$. $\blacksquare$

How to construct $\bar{Q}$?

Optimistic planning via bonus

UCB-VI

$N_{t-1}(s, a, s')$ - # times see (s, a, s') in past $t-1$ trajectories

$N_{t-1}(s, a)$ - # " " (s, a) " " " "

$$P^{t-1}(s'|s, a) = \frac{N_{t-1}(s, a, s')}{N_{t-1}(s, a)}$$

$Q_H^{t-1}(s,a) = R(s,a)$

optimistic VI update $\rightarrow$ $Q_h^{t-1}(s,a) = \boxed{R(s,a) + \sum_{s'} P^{t-1}(s'|s,a) \max_{a'} Q_{h+1}^{t-1}(s',a') + \underbrace{b^{t-1}(s,a)}_{\text{reward bonus}}}$ Bellman eq^n.

$\min(H, \square)$

$\rightarrow$ Reward bonus, $b^{t-1}(s,a) = H\sqrt{\dfrac{S\Delta}{N^{t-1}(s,a)}}$ $\quad \Delta = \log \dfrac{SAHT}{\delta}$

$\equiv$ VI with MDP $(S, A, P^{t-1}, R + b^{t-1}, d_0, H)$

Lemma: w.p. $\geq 1-\delta$

$\forall t, h, s, a \quad Q_h^*(s,a) \leq Q_h^{t-1}(s,a) \leq T Q_{h+1}^{t-1}(s,a) + \underbrace{2b^{t-1}(s,a)}_{\text{conf.}}$

Proof: By martingale version of Bernstein inequality

$$\forall s,a,t \quad |P^{t-1}(s'|s,a) - P(s'|s,a)| \lesssim \sqrt{\frac{P(s'|s,a)\log \frac{SAHT}{\delta}}{N^{t-1}(s,a)}} + O\left(\frac{1}{N^{t-1}(s,a)}\right)$$

$$\Rightarrow \|P^{t-1}(s,a) - P(s,a)\|_{TV} = \sum_{s'} 1 \cdot |P^{t-1}(s'|s,a) - P(s'|s,a)|$$

$$\leq \sqrt{\sum_{s'} 1^2 \cdot \sum_{s'} |P^{t-1}(s'|s,a) - P(s'|s,a)|^2}$$

$$\leq \sqrt{\frac{S\Delta}{N^{t-1}(s,a)} \underbrace{\sum_{s'} P(s'|s,a)}_{1}} =: \frac{b^{t-1}(s,a)}{H}$$

Induction to show optimistic Q_h^{t-1}

$$Q_{h+1}^{t-1}(s,a) \geq Q_{h+1}^*(s,a) \quad \forall s, a$$

$$Q_h^{t-1}(s,a) = R(s,a) + b^{t-1}(s,a) + \sum_{s'} P^{t-1}(s'|s,a) \max_{a'} Q_{h+1}^{t-1}(s',a')$$

$$\geq \text{"} + \text{"} + \sum_{s'} P^{t-1}(s'|s,a) V_{h+1}^*(s') \quad \text{optimism of } h+1$$

$$= \text{"} + \text{"} + \sum_{s'} \left(P^{t-1}(s'|s,a) - P(s'|s,a)\right) V_{h+1}^*(s')$$

$$+ \sum_{s'} P(s'|s,a) V_{h+1}^*(s')$$

$$\geq \text{"} + \text{"} - \|P^{t-1}(s,a) - P(s,a)\|_{TV} \cdot H + \text{"}$$

$$\geq R(s,a) + \sum_{s'} P(s'|s,a) V_{h+1}^*(s') = Q_h^*(s,a)$$

$$Q_h^{t-1}(s,a) = R(s,a) + b^{t-1}(s,a) + \sum_{s'} P(s'|s,a) \max_{a'} Q_{h+1}^{t-1}(s',a')$$

$$\leq b^{t-1}(s,a) + \underbrace{T Q_{h+1}^{t-1}(s,a)} + \sum_{s'} \underbrace{\left(P^{t-1}(s'|s,a) - P(s'|s,a)\right)}_{b^{t-1}/H} \cdot \underbrace{\max_{a'} Q_{h+1}^{t-1}(s',a')}_{H}$$

$$\leq \underbrace{2 b^{t-1}(s,a)}_{\text{conf}} + T Q_{h+1}^{t-1}(s,a)$$

Regret of UCB-VI

Optimistic Regret Decomp.

$\bar{Q}_h \equiv Q_h^{t-1}$ optimistic & nearly Bellman consistent

$$\text{Regret}(T) = \sum_{t=1}^{T} E\left[V_{\pi^*}(s) - V_{\pi_t}(s)\right] \leq \sum_{t=1}^{T}\sum_{h=1}^{H} 2\, E_{s,a \sim d_h^{\pi_t}}\left[b^{t-1}(s,a)\right]$$

π_t: greedy for Q_h^{t-1}

$$\sum_{t=1}^{T}\sum_{h=1}^{H} H\sqrt{\frac{S\Delta}{N^{t-1}(s_h,a_h)}} = H\sqrt{S\Delta} \sum_{t,h} \frac{1}{\sqrt{N^{t-1}(s_h,a_h)}}$$

$$= H\sqrt{S\Delta} \sum_h \sum_{s,a} \sum_{i=1}^{N_h^T(s,a)} \frac{1}{\sqrt{i}}$$

$$\sum_{i=1}^{n} \frac{1}{\sqrt{i}} \leq 2\sqrt{n}$$

$$\lesssim H\sqrt{S\Delta} \sum_h \sum_{s,a} 1 \cdot \sqrt{N_h^T(s,a)}$$

$$\leq H\sqrt{S\Delta} \sum_h \sqrt{\underbrace{\sum_{s,a} 1^2}_{SA} \cdot \underbrace{\sum_{s,a} N_h^T(s,a)}_{T}}$$

$$\mathcal{E} = H^2 S \sqrt{AT\Delta}$$

$\blacksquare$

$T \asymp S^2 A$ $\qquad$ $\underline{P} : S \times A \to S$

Can be improved:

$\Rightarrow$ $T \asymp SA$ $\qquad$ Optimizing Regret / Find optimal policy $\}$ easier than learn model.