Linear MDP

$$\phi : S \times A \rightarrow \mathbb{R}^d$$

Q^* is linearly realizable $\quad Q^*(s,a) = \langle \phi(s,a), \theta^* \rangle$ ✓

don't explicitly estimate P^{t-1}, dynamic programming for Q directly.

Recap Tabular UCB-VI

$$Q_h^t(s,a) = \min\{H,\ \underbrace{R(s,a) + \sum_{s'} P^{t-1}(s'|s,a) \max_{a'} Q_h^t(s',a')}\ + \underbrace{b^t(s,a)}_{\text{bonus exploration parameter}}\}$$

For bandits, $\quad R(s,a) + b^t(s,a)$

For RL (without exp) $\quad Q = R + \sum_{s'} P(s'|s,a) \max_{a'} Q(s',a')$

Bellman optimality.

UCB-VI update ensures Q is optimistic & nearly Bellman optimal

Linear UCB-VI $\quad$ LSVI-UCB

$$V_H^t = 0$$

$$\forall h < H \qquad \theta_h^t \leftarrow \arg\min_{\theta} \sum_{i=1}^{t-1} (\overbrace{\langle \phi(s_h^i, a_h^i), \theta \rangle}^{\langle X_i, \beta \rangle} - \overbrace{r_h^i - V_{h+1}^t(s_{h+1}^i)}^{Y_i})^2 + \lambda \|\theta\|_2^2$$

$$\Rightarrow Q_h^t(s,a) = \langle \phi(s,a), \theta_h^t \rangle + b_h^t(s,a) \rightarrow \min(H, \ldots)$$

$$V_h^t(s) = \max_{a'} Q_h^t(s,a)$$

→ greedy policy wrt Q_h^t, collect another episode

bonus $\quad b_h^{t-1}(s,a) = \beta \|\phi(s,a)\|_{\Lambda_{h,t-1}^{-1}}, \quad \Lambda_{h,t} = \sum_{i=1}^{t-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^T + \lambda I$

If $H=1$, LinUCB recovered.

Complication: $\quad E[Y_i] = \langle X_i, \beta^* \rangle$

$$E[r_h^i + V_{h+1}^t(s_{h+1}^i)] \rightarrow \text{not linear}$$

Least square fit is not guaranteed to be well-specified.

Need more assumptions beyond realizable Q^*

Linear / Low-rank MDP if $\exists\, \phi : S \times A \to \mathbb{R}^d$, $w^* \in \mathbb{R}^d$, $\mu : S \to \mathbb{R}^d$

s.t. $R(s,a) = \langle \phi(s,a), w^* \rangle$, $P(s'|s,a) = \langle \phi(s,a), \mu(s') \rangle$

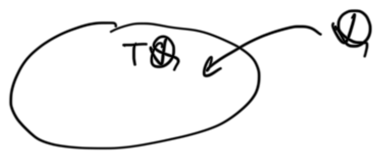
These assumption imply any Q that satisfies Bellman equation is linear

1) $R(s,a)$ is linear

2) $\mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')]$ is linear

$$= \int_{s'} P(s'|s,a) V(s') = \int_{s'} \langle \phi(s,a), \mu(s') \rangle V(s')$$

$$= \langle \phi(s,a), \underbrace{\int_{s'} \mu(s') V(s')}_{= \bar{\theta}_V} \rangle$$



Upper bound: $\forall p \geq 1 - \delta$ LSVI-UCB has $\mathrm{Reg}(T) = \tilde{O}(H^2 \sqrt{d^3 T})$.

$|S|, |A|$ infinite

d-dim.

Proof: By induction show optimistic regret decomposition for tabular holds.

① Assume $Q^*_{h+1}(s,a) \leq Q^t_{h+1}(s,a)$ $\forall s,a$

$$Q^*_h(s,a) = R(s,a) + \int_{s'} P(s'|s,a) V^*_{h+1}(s')$$

$$= \langle \phi(s,a), w^* \rangle + \int_{s'} P(s'|s,a) \max_{a'} Q^*_{h+1}(s',a')$$

$$\leq \langle \phi(s,a), w^* \rangle + \int_{s'} P(s'|s,a) \max_{a'} Q^t_{h+1}(s,a) \quad \star$$

$$= \langle \phi(s,a), w^* \rangle + \langle \phi(s,a), \bar{\theta}^t_h \rangle$$

$$= \langle \phi(s,a), \tilde{\theta}^t_h \rangle \quad \star$$

$$= \langle \phi(s,a), \theta^t_h \rangle + \underbrace{\langle \phi(s,a), \theta^t_h - \tilde{\theta}^t_h \rangle}$$

$$\leq \|\phi(s,a)\|_{\Lambda^{-1}_{h,t-1}} \underbrace{\|\theta^t_h - \tilde{\theta}^t_h\|_{\Lambda_{h,t-1}}}_{= \tilde{O}(H^2 d^2) =: \beta^2}$$

$$= \langle \phi(s,a), \theta^t_h \rangle + b^t_h(s,a)$$

$$= Q^t_h(s,a)$$

For LS linear fit $\|\hat{\theta}_{LS} - \theta^*\|^2_{\Sigma^{-1}} = \sigma^2 d \rightarrow H^2 d$

$\sigma \rightarrow H$ + extra d bcoz $\tilde{\theta}_h^t$ is not fixed as compared to θ^*.

$Y_i = \langle x_i, \beta \rangle + \varepsilon_i$
$\varepsilon_i \hookrightarrow \sigma^2$

(2) $Q_h^t(s,a) \leq T Q_{h+1}^t(s,a) + \text{conf}_h^t(s,a)$ (nearly Bellman opt)

[opt Reg Decomp : if Q s.t. (1) $Q^* \leq Q_h^t$
(2) $Q_h^t \leq T Q_{h-1}^t + \text{conf}_h^t$]
then $\text{Reg} \leq \sum_{t,h} \text{conf}_h^t$

$$Q_h^t(s,a) = \langle \phi(s,a), \theta_h^t \rangle + \beta \|\phi(s,a)\|_{\Lambda^{-1}_{h,t-1}}$$
$$\leq \langle \phi(s,a), \tilde{\theta}_h^t \rangle + \underbrace{\langle \phi(s,a), \theta_h^t - \tilde{\theta}_h^t \rangle} + \beta \|\phi(s,a)\|_{\Lambda^{-1}_{h,t-1}}$$
$$\leq \langle \phi(s,a), \tilde{\theta}_h^t \rangle + 2\beta \|\phi(s,a)\|_{\Lambda^{-1}_{h,t-1}}$$
$$= T Q_{h-1}^t(s,a) + \text{conf}_h^t$$

$$\text{Regret} \leq E\Big[\sum_{t,h} \text{conf}_h^t\Big] \leq 2\sum_{t,h} \beta \|\phi(s,a)\|_{\Lambda^{-1}_{h,t-1}} + \tilde{O}(H\sqrt{T})$$

$\downarrow$ using elliptical potential lemma (from linear bandit)

$\tilde{O}(\beta H \sqrt{dT})$

$= \tilde{O}(H^2 d \sqrt{dT})$

$|S|, |A|$ infinite

Weaker assumptions.

Q^* linear realizable

LSVI-UCB $\quad \{\max_a \langle \phi(s,a), \theta \rangle + \beta \|\phi(s,a)\|_{M^{-1}}\}$ $\quad V$

only requires value functions of ↑ form to have $\quad M \succeq 0, \beta > 0$ $\quad TV$

linear Bellman updates.

Bellman completeness A function class $\mathcal{F}: S \times A \to \mathbb{R}$ and MDP M satisfy Bellman completeness if $\forall f \in \mathcal{F}$ we have $\mathcal{T}f \in \mathcal{F}$.

LSVI-UCB does not work under Bellman completeness.

generalized linear model. $\sigma(\langle \phi(s,a), \theta \rangle)$ included under #

$\uparrow$ nonlinear

LQR (Linear Quadratic Regulator) reward - quadratic

transitions - linear

not included under #

Need algorithm that works under weaker Bellman completeness for extension to nonlinear settings in general.