Nonlinear function approx^n in RL

Structural conditions:

local optimism

$Q^*_h(s,a) \le Q_h(s,a)$ $\forall h, s, a$

Q^* realizable $\qquad Q^* \in \mathcal{F} / \mathcal{Q}$

linear MDP - low rank enough & LVI-UCB works

P/R are linear

$\Downarrow$

$Q \leftarrow$ $\underbrace{\text{linear + bonus}}_{\text{linear}}$

(linear) Bellman completeness.

$Q \in \mathcal{F} \Rightarrow TQ \in \mathcal{F}$

How to measure complexity of $\mathcal{F}$ for RL?

Finite-horizon MDP $\quad M = (S, A, P, R, \mu, H)$

Realizability $\quad Q^*_h(s,a) = \langle \theta^*_h, \phi(s,a) \rangle$ for linear $\quad \in \mathcal{F}$ for nonlinear

Bellman completeness (linear) $\quad$ For any θ, $\exists\, \bar\theta$ s.t. $\qquad \theta \equiv Q$

$$(T\theta)(s,a) = \langle \bar\theta, \phi(s,a) \rangle$$

Global optimistic regret decomposition $\quad$ If $Q_1 \ldots Q_H$ s.t.

$\Rightarrow E_{s_1} \max_a Q_1(s_1,a) \ge E_{s_1} \max_a Q^*_1(s_1,a)$

(earlier $Q \ge Q^*$ optimistic $\forall s,a,h$

$Q \le TQ + \text{conf}$)

and π greedy w.r.t. Q_1 then

$$\underset{E[V^*(s_1) - V_\pi(s_1)]}{J(\pi^*) - J(\pi)} \le \sum_{h=1}^{H} E_{(s,a)\sim d_h^\pi}\big[\underbrace{Q_h(s,a) - (TQ_{h+1})(s,a)}_{\text{conf}_h(s,a)}\big]$$

Proof: $\quad J(\pi^*) - J(\pi) = E_{s_1}\big[\underbrace{Q^*(s_1, \pi^*(s_1))}_{\max_a Q^*(s_1,a)} - Q_1^\pi(s_1, \pi(s_1))\big]$

by global opt $\quad \le E_{s_1}\big(\max_a Q_1(s_1,a) - Q_1^\pi(s_1, \pi(s_1))\big]$ (1)

π is greedy wrt Q_1 $\quad \le E_{(s_1,a_1)\sim d_0^\pi}\big[Q_1(s_1, \underset{\pi(s_1)}{a_1}) - Q_1^\pi(s_1, \pi(s_1))\big]$

$$= E_{(s_1,a_1)\sim d_0^\pi}\big[\underbrace{Q_1(s_1,a_1) - (TQ_2)(s_1,a_1)} + \underbrace{(TQ_2)}(s_1,a_1) - \underbrace{Q_1^\pi(s_1,\pi(s_1))}\big]$$

$$= \underbrace{\text{first term}}_{Q_1 - TQ_2} + \cancel{R(s_1,a_1)} + E_{s_2}[\max_a Q_2(s_2,a_2)] \underbrace{- \cancel{R(s_1,a_1)} - E_{s_2}[Q_2^\pi(s_2,\pi(s_2))]}_{B}$$ (2a)

$$= \sum_{h=1}^{H} E[Q_h - TQ_{h+1}]$$

Main challenge: Show global optimism only assuming Bellman completeness.

[Linear] Bellman completeness: For any θ, $\exists \bar\theta$ s.t.

$$\forall (s,a) \quad (T\theta)(s,a) = \langle \bar\theta, \phi(s,a)\rangle$$

$$\| \quad E[R(s,a) + \max_{a'} Q(s',a')]$$

$$\| \quad \langle \theta, \phi(s',a')\rangle$$

$\Rightarrow$ no misspecification in regression if take Q to be linear.

(1) Regression error $R_h^{t-1}(\theta, \tilde\theta_{h+1}) = \sum_{i=1}^{t-1} \Big(\langle \phi(s_h^i, a_h^i), \theta\rangle - \underbrace{r_h^i - \underbrace{\max_a \langle \phi(s_{h+1}^i, a), \tilde\theta_{h+1}\rangle}_{V_{h+1}^i(s)}}_{Y_i}\Big)^2$

$$\equiv (Q - T\theta)^2$$

earlier Q = linear + bonus
now Q = linear
no optimism!

earlier $EY_i \neq$ linear.
now EY_i = linear acc to Bellman completeness.

Build in optimism using nested confidence balls

(2) $\mathrm{BALL}^t := \{(\theta_1 \ldots \theta_H) : \theta_H = 0, \ \forall h \ \ R_h^{t-1}(\theta_h, \theta_{h+1}) \le \min_\theta R_h^{t-1}(\theta, \theta_{h+1}) + \beta^2\}$

$\nexists$ optimism

$\max_{\theta \in C_t} x^T\theta$
$\theta^* \in C_t$

Note that θ^* : $R_h^{t-1}(\theta_h^*, \theta_{h+1}^*) = 0$

$$\theta^* = TQ^*$$

$$\theta_h^* = R(s_h, a_h) + E[V_{h+1}^*(s')]$$

$$\{Q_h^*\}_{h \in H} \equiv \theta_1^* \ldots \theta_H^*$$

(3) $\Rightarrow (\theta_1^t \ldots \theta_H^t) \leftarrow \arg\max_{(\theta_1 \ldots \theta_H) \in \mathrm{BALL}^t} E_{s_1} \max_a \langle \phi(s_1,a_1), \theta_1\rangle$

$$(\theta_1^* \ldots \theta_H^*) \in \mathrm{BALL}^t$$

... $E_{s_1} \max \langle \phi(s_1,a_1), \theta_1^*\rangle$

$\Rightarrow$ Global optimism holds w.r.t. a

$\leq E_{s_1} \max_a \langle \phi(s_1,a_1), \hat{\theta}_1 \rangle$

(4) greedy policy wrt θ^t to collect another episode.

Regret analysis:

$$\text{Regret} \leq E\Big[\sum_t \sum_h Q_h^t(s_h^t, a_h^t) - (T Q_{h+1}^t)(s_h^t, a_h^t)\Big] \equiv E\Big[\sum_t \sum_h \text{conf}_h^t\Big]$$

$$\leq \sum_t \sum_h Q_h^t(s_h^t, a_h^t) - (T Q_{h+1}^t)(s_h^t, a_h^t) + \tilde{O}(H\sqrt{T})$$

depend on linear assumption

$$\leq \sum_t \sum_h \langle \phi(s_h^t, a_h^t), \theta_h^t - \bar{\theta}_h^t \rangle + \tilde{O}(H\sqrt{T})$$

$$\leq \underbrace{\sum_t \sum_h \|\phi(s_h^t, a_h^t)\|_{\Lambda_{h,t-1}^{-1}}}_{\text{elliptic potential lemma}} \cdot \underbrace{\|\theta_h^t - \bar{\theta}_h^t\|_{\Lambda_{h,t-1}}}_{\beta = O(H\sqrt{d})} + \tilde{O}(H\sqrt{T})$$

saving $\sqrt{d}$ factor using global opt.

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

$$= \tilde{O}(H^2 d\sqrt{T})$$ ∎

Generalization to nonlinear functions.

- only place linear assumption needed is to bound $\|\theta_h^t - \bar{\theta}_h^t\|$ d

Notion for nonlinear complexity

Bellman rank: Given $\mathcal{F}$ ($\mathcal{Q}$), let Π be induced policy class

$\Pi = \{\pi_f : f \in \mathcal{F}\}$. For each h, $\exists$ embedding function

$w_h : \Pi \to \mathbb{R}^d$ and $v_h : \mathcal{F} \to \mathbb{R}^d$ s.t.

Bellman error $\mathcal{E}_h(\Pi, F) = \langle w_h(\pi), v_h(f) \rangle$

$Q - TQ$

where $\mathcal{E}_h(\pi, F) = E[Q_h(s_h, a_h) - r_h - \max_{a'} Q_{h+1}(s_{h+1}, a')$

$s_h \sim d_h^\pi,\ a_h = \pi_Q(s_h)]$

not π as for linear

d - Bellman rank

[Note: $Q \neq$ linear]

Regret: $J(\pi^*) - J(\pi) = \sum_t \sum_h \langle w_h(\pi^t), v_h(f^t) \rangle$ (earlier $\sum_t (Q - TQ)$)

$\leq \sum_t \sum_h \|w_h(\pi^t)\|_{\Sigma^{-1}_{t-1,h}} \|v_h(f^t)\|_{\Sigma_{t-1,h}}$

$\uparrow \lambda I + \sum v v^T$

assuming $\|w_h(\pi)\|_2 \leq W$, $\|v_h(f)\|_2 \leq V$

can show $\|v_h(f^t)\|_{\Sigma_{t-1,h}} \leq \sqrt{\lambda V^2 + 4\beta^2}$ $= \|\theta - \hat{\theta}\|$

$\|w_h(f^t)\|_{\Sigma^{-1}_{t-1,h}} \leq \sqrt{\frac{2Hd}{T} \log(\ )}$) elliptic potential lemma

$\Rightarrow$ Regret bound that depends on Bellman rank d.