- HW2 (due 2 weeks)
- presentation: critique form
  - please send me your slides / notes after your presentation.

Linear MDPs: $M = (r_h, P_h)_{h=0}^{H-1}$, feature map $\phi: S \times A \to \mathbb{R}^d$.

$\exists\, (\alpha_h^*)_{h=0}^{H-1} \subseteq \mathbb{R}^d$,

$(\mu_h^*)_{h=0}^{H-1} \subseteq \mathbb{R}^{S \times d}$, such that:

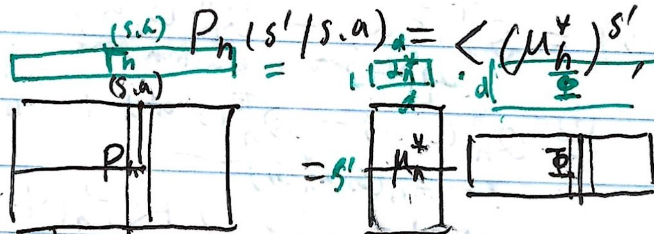
(1) $\forall h, \forall s, a, s'$: $r_h(s,a) = \langle \alpha_h^*, \phi(s,a) \rangle$

$P_h(s'|s,a) = \langle (\mu_h^*)^{s'}, \phi(s,a) \rangle$

(2) $\max_{s,a} \|\phi(s,a)\|_2 \leq$

(3) $\forall V: \|V\|_\infty \leq 1$ $\Rightarrow \|V^T \mu_h^*\|_2 \leq \sqrt{d}$

~~(4) $\forall h, \|\alpha_h^*\|_2 \leq \sqrt{}$~~



Recall (exercise): linear MDPs satisfies that: $\forall V: S \to \mathbb{R}$.

$\mathcal{L}_h: \mathbb{R}^S \to \mathbb{R}^{SA}$.

[$\mathcal{L}_h$ links V_{h+1}^π to Q_h^π, V_{h+1}^* to Q_h^*]

$(\mathcal{L}_h V)(s,a) = r_h(s,a) + \langle P_h(\cdot|s,a), V \rangle$

$= \langle \alpha_h^*, \phi(s,a) \rangle + \langle \mu_h^* \phi(s,a), V \rangle$

$= (\alpha_h^*)^T \phi(s,a) + V^T \mu_h^* \phi(s,a)$ Important notation.

$= \langle \alpha_h^* + (\mu_h^*)^T V, \phi(s,a) \rangle = \langle \theta_h^*(V), \phi(s,a) \rangle$ linear in $\phi(s,a)$.

How to design an online RL algorithm for linear MDPs?

Idea: combine LSVI with optimism. (useful in online RL)

For episodes $k = 0, 1, \cdots K$:

// Data: $D_h = \{(s_h^i, a_h^i, r_h^i, s_{h+1}^i)\}_{i=0}^{k-1}$, $h = 0, 1, \cdots H-1$.

// Idea: construct optimistic estimates of Q_h^*, V_h^*

$V_H^k \leftarrow 0$.

For $h = H-1, H-2, \cdots 0$:

$\hat{Q}_h^k \geq Q_h^*$

// given $\hat{V}_{h+1}^k \geq V_{h+1}^*$. want to construct $\hat{Q}_h^k \geq Q_h^*$.

write on another part of board

Recall: $Q_h^*(s,a) = (\mathcal{L}_h V_{h+1}^*)(s,a)$

~~can we build estimate of $\mathcal{L}_h \hat{V}_{h+1}^k$~~ ~~$\hat{Q}_h^k \geq \mathcal{L}_h \hat{V}_{h+1}^k$~~

although we don't have V_{h+1}^*. we have $\hat{V}_{h+1}^k$. as an upper bound of it.

$\mathcal{L}_h \hat{V}_{h+1}^k \geq \mathcal{L}_h V_{h+1}^*$, (exercise).

it suffices to build tight upper bound estimates of $\mathcal{L}_h \hat{V}_{h+1}^k$.

Idea: $\mathcal{L}_h \hat{V}_{h+1}^k = \langle \theta_h^*(\hat{V}_{h+1}^k), \phi(s,a) \rangle$. so it suffices to estimate $\theta_h(\hat{V}_{h+1}^k)$; specifically.

Similar to UCB. we will use a confidence set (confidence interval) to estimate it.

$(\phi(s_h^i, a_h^i), r_h^i + \hat{V}_{h+1}^k(s_{h+1}^i))_{i=0}^{k-1}$ ~~generalizing~~

- Define. $\hat{\theta}_h^k = \arg\min_\theta \left\{ \lambda \|\theta\|^2 + \sum_{i=0}^{k-1} (\langle \theta, \phi(s_h^i, a_h^i) \rangle - (r_h^i + \hat{V}_{h+1}^k(s_{h+1}^i)))^2 \right\}$ ridge regression

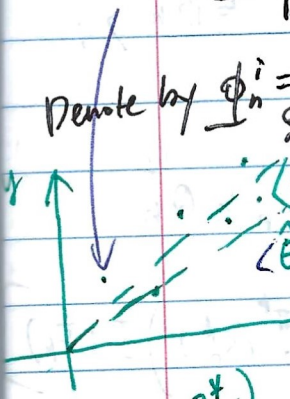
Denote by $\phi_h^i = \phi(s_h^i, a_h^i)$.

$= (\Lambda_h^k)^{-1} \left(\sum_{i=0}^{k-1} \phi_h^i (r_h^i + \hat{V}_{h+1}^k(s_{h+1}^i)) \right)$

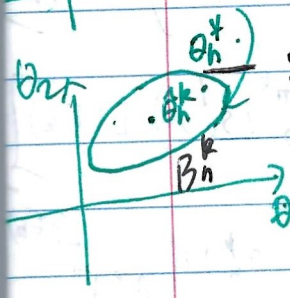
where $\Lambda_h^k = \lambda I + \sum_{i=0}^{k-1} \phi_h^i (\phi_h^i)^T$.

~~set $B_h^k = \{ (\theta_h^*(\hat{V}_{h+1}^k), \phi(s,a)) = (\mathcal{L}_h \hat{V}_{h+1}^k)(s,a) \}$~~

$\therefore \langle \hat{\theta}_h^k, \phi(s,a) \rangle$



set $B_h^k = \{ \theta : \|\theta - \hat{\theta}_h^k\|_{\Lambda_h^k} \leq \beta \}$, where



(with prob $\geq 1 - \frac{1}{k}$: $\theta_h^*(\hat{V}_{h+1}^k) \in B_h^k$).

$\beta = \text{const} \cdot H \cdot d \sqrt{\ln \frac{d(k+1)}{\delta}}$.

(so w.h.p. $\geq \langle \theta_h^*(\hat{V}_{h+1}^k), \phi(s,a) \rangle = (\mathcal{L}_h \hat{V}_{h+1}^k)(s,a)$)

- Define $\hat{Q}_h^k(s,a) = \max_{\theta \in B_h^k} \langle \theta, \phi(s,a) \rangle$

$= \max_{\|\Delta\theta\|_{\Lambda_h^k} \leq \beta} \langle \hat{\theta}_h^k, \phi(s,a) \rangle + \langle \Delta\theta, \phi(s,a) \rangle$

(see textbook for another "model-based" explanation of the same algorithm)

$M_h^k = (\hat{P}_h^k, \hat{r}_h^k)_{h=0}^{H-1}$.

$= \langle \hat{\theta}_h^k, \phi(s,a) \rangle + \beta \cdot \| \phi(s,a) \|_{(\Lambda_h^k)^{-1}}$.

— Define $\hat{V}_h^k(s) = \min\left(\max_{a \in A} \hat{Q}_h^k(s,a), H\right)$.

— Define $\hat{\pi}_h^k(s) = \arg\max_{a \in A} \hat{Q}_h^k(s,a)$.

// execute policy $\hat{\pi}^k$.

For $h = 0, 1, \cdots H-1$:

see s_h^k. take action $a_h^k = \hat{\pi}_h^k(s_h^k)$,

receive $r_h^k = r_h(s_h^k, a_h^k)$, transition to s_{h+1}^k.

Analysis: LSVI-UCB algorithm.

Theorem: LSVI-UCB has regret

$$\mathrm{Reg}(K) \le \tilde{O}\left(H^2 \sqrt{d^3 K}\right).$$

$\tilde{O}(\cdot)$ ignore log factors

Pf structure: ① ~~optimism: $V_h^k \ge V_h^*$, $Q_h^k \ge Q_h^*$ whp.~~

concentration: $\forall k$, $E_k := \{ \forall h,\ \theta_h^*(\hat{V}_{h+1}^k) \in B_h^k \}$;

② $P(E_k) \ge 1 - \frac{1}{k}$.

③ optimism: $\forall k$ on E_k, $\hat{V}_h^k \ge V_h^*$, $\hat{Q}_h^k \ge Q_h^*$.

$$\mathrm{Reg}(K) \lesssim \mathbb{E}\left[\sum_{k=0}^{K-1} \left(V_0^*(s_0^k) - V_0^{\pi^k}(s_0^k)\right) \mathbb{I}(E_k)\right]$$

$$\overset{\text{optimism}}{\le} \mathbb{E}\left[\sum_{k=0}^{K-1} \left(\hat{V}_0^k(s_0^k) - V_0^{\pi^k}(s_0^k)\right) \mathbb{I}(E_k)\right]$$

$$\lesssim \mathbb{E}\left[\sum_{k=0}^{K-1} \sum_{h=0}^{H-1} 2\beta \cdot \| \phi(s_h^k, a_h^k) \|_{(\Lambda_h^k)^{-1}}^{\mathbb{I}(E_k)}\right]$$

bonus / uncertainty of experienced state-action pairs

④ $\sum_{k=0}^{K-1} \sum_{h=0}^{H-1} \| \phi(s_h^k, a_h^k) \|_{(\Lambda_h^k)^{-1}} \le \tilde{O}\left(H\sqrt{dK}\right)$

using elliptic potential lemma (will introduce next)

In summary, ① – ④:

$\Rightarrow Reg(K) \le \tilde{O}(\beta \cdot H\sqrt{dK}) = \tilde{O}(H^2\sqrt{d^3K})$. ($\beta = \tilde{O}(Hd)$)

pf of ①: Let $f = \hat{V}_{h+1}^k$.

$(\mathcal{L}_h f)(s_h^i, a_h^i) = \langle \theta_h^*(f), \phi(s_h^i, a_h^i) \rangle \leftarrow r_h^i + f(s_{h+1}^i)$ (Expectation.)

$= \langle \theta_h^*(f), \phi_h^i \rangle + \xi_h^i(f)$ wt)

$\hat{\theta}_h^k(f) = (\Lambda_h^k)^{-1}\left(\sum_{i=1}^{k-1} \phi_h^i (\phi_h^{iT} \cdot \theta_h^*(f) + \xi_h^i(f))\right)$

$\Rightarrow \hat{\theta}_h^k(f) - \theta_h^*(f)$

$= (\Lambda_h^k)^{-1}\left[\left((\Lambda_h^k) - \lambda I\right)\theta_h^*(f) + \sum_{i=0}^{k-1} \phi_h^i \xi_h^i(f)\right] - \theta_h^*(f)$

$= -\lambda \cdot (\Lambda_h^k)^{-1} \theta_h^*(f) + (\Lambda_h^k)^{-1}\left(\sum_{i=0}^{k-1} \phi_h^i \xi_h^i(f)\right)$

$\Rightarrow \| \hat{\theta}_h^k(f) - \theta_h^*(f) \|_{\Lambda_h^k}$

$\le \sqrt{\lambda} \cdot \| \theta_h(f) \| + \| \sum_{i=0}^{k-1} \phi_h^i \xi_h^i(f) \|_{(\Lambda_h^k)^{-1}}$.

$\hat{\theta}_h^k(f)$ is the ridge regression soln of

$\left(\phi(s_h^i, a_h^i), \underbrace{(\mathcal{L}_h f)(s_h^i, a_h^i)}_{\| \langle \theta_h^*(f), \phi(s_h^i, a_h^i) \rangle} + \xi_i(f) \right)_{i=0}^{k-1}$

regression guarantee

$\| \hat{\theta}_h^k(f) - \theta_h^*(f) \|_{(\Lambda_h^k)^{-1}} \le \| \theta_h^*(f) \|_2 + H\sqrt{d + \ln\frac{1}{\delta}}$

$\le \tilde{O}(H\sqrt{d + \ln\frac{1}{\delta}})$ (*)

However: $f = \hat{V}^k_{h+1}$ depends on previous data collected.
so conditioned on $\mathcal{D}^k_{h+1} (s^i_h, a^i_h)_{i=0}^{k-1}$, $(s^i_{h+1})_{i=0}^{k-1}$ may not be independent
so (*) may not be true.

Fix: key observation: $\mathcal{D}^k_{h+1}$ come from a "simple" fn class
argue that the concentration holds for all $f \in F_{L,B,\lambda}$

$$f \in F_{L,B} = \{ f_{w,\beta,\lambda} : \|w\|_2 \le L, \beta \le B, \lambda \ge I \}$$

with $L = 2HK$
$B = \tilde{O}(Hd)$.

where $f_{w,\beta,\lambda} = \min\left(H, \max_a \left(\langle w, \phi(s,a)\rangle + \beta \|\phi(s,a)\|_{\Lambda^{-1}}\right)\right)$

we show w.p. $\ge 1-\delta$. $\forall f \in F_{L,B,\lambda}$.

$$\|\hat{\theta}^k_h(f) - \theta^*_h(f)\|_{(\Lambda^k_h)^{-1}} \le \tilde{O}\left(H\sqrt{d^2 + \ln\tfrac{1}{\delta}}\right).$$

via a covering argument. (AJKS. Sect. 8.4)

Pf of ②: Backward induction. $\forall k$. suppose E_k happens.

① $\forall h, s$ $\hat{V}^k_h(s) = \min(H, \max_a \hat{Q}^k_h(s,a))$

② $\forall h, s, a$. $\hat{Q}^k_h(s,a) = \max_{\theta \in \mathcal{B}^k_h} \langle \theta, \phi(s,a)\rangle$

$\ge \langle \theta^*_h(\hat{V}^k_{h+1}), \phi(s,a)\rangle$

$= (\mathcal{L}_h \hat{V}^k_{h+1})(s,a)$

$h = H$. trivial.

suppose $\hat{Q}^k_{h+1} \ge Q^*_{h+1}$. $\hat{V}^k_{h+1} \ge V^*_{h+1}$. (inductive hypothesis)

as $\forall s,a$ $\hat{Q}^k_h(s,a) = \max_{\theta \in \mathcal{B}^k_h} \langle \theta, \phi(s,a)\rangle$

$\ge \langle \theta^*_h(\hat{V}^k_{h+1}), \phi(s,a)\rangle$

$= (\mathcal{L}_h \hat{V}^k_{h+1})(s,a)$

$\ge (\mathcal{L}_h V^*_{h+1})(s,a) = Q^*_h(s,a)$.

$\Rightarrow \hat{Q}^k_h \ge Q^*_h$

vs. $\max_a \hat{Q}^k_h(s,a) \ge \max_a Q^*_h(s,a) = V^*_h(s)$ and $H \ge V^*_h(s)$.

$\Rightarrow \hat{V}^k_h(s) \ge V^*_h(s)$.

leave this as an exercise

① ② $\Rightarrow$ strong optimism:

strong optimism $\Rightarrow$ optimism (Exercise)

(ψ_k almost satisfy Bellman opt. eqn w/ equality replaced w/ inequality)

pf of ③ : $\hat{V}_0^k(s_0^k) - V_0^{\pi^k}(s_0^k)$

let $a_0^k = \pi_0^k(s_0^k)$

$= \min(\hat{Q}_0^k(s_0^k, a_0^k), H) - Q_0^{\pi^k}(s_0^k, a_0^k)$

$\leq \hat{Q}_0^k(s_0^k, a_0^k) - Q_0^{\pi^k}(s_0^k, a_0^k)$.

Generalized Simulation Lemma:

given episodic MDP. M. $\hat{Q} = (\hat{Q}_h)_{h=0}^{H-1} : S \times A \to \mathbb{R}$, $\pi = (\pi_h)_{h=0}^{H-1}$ history independent policy. then

$\hat{Q}_h(s_h, a_h) - Q_h^\pi(s_h, a_h)$

$= \mathbb{E}\left[\sum_{i=h}^{H-1} Bell_h^\pi(s_h, a_h) \;\middle|\; s_h, a_h, M, \pi \right]$, where

$Bell_h^\pi(s,a) = \hat{Q}_h(s,a) - (J_h^\pi \hat{Q}_{h+1})(s,a)$ is the Bellman error of $\hat{Q}$ at step h (how much does $\hat{Q}_h$ violate Bellman consistency equation of π)

$(J_h^\pi f)(s,a) = r_h(s,a) + \sum_{s'} P_h(s'|s,a) f(s, \pi_{h+1}(s))$.

① it generalizes simulation Lemma b/c $(\hat{Q}_h)_{h=0}^{H-1}$ may not be value fn of some policy.

② intuitively. if Bellman error of $\hat{Q}$ is small at all steps, $\hat{Q} \approx Q^\pi$.

Applying to our setting: at episode k.

$Bell_h^{\pi^k}(s,a) = \hat{Q}_h^k(s,a) - \left(r_h(s,a) + \sum_{s'} P_h(s'|s,a) \hat{Q}_{h+1}^k(s, \pi_{h+1}^k(s)) \right)$

$\max_a \hat{Q}_{h+1}^k(s,a) = \hat{V}_{h+1}^k(s)$

$\leq \hat{Q}_h^k(s,a) - (\mathcal{L}_h \hat{V}_{h+1}^k)(s,a)$

$= \max_{\theta \in B_h^k} \langle \theta, \phi(s,a) \rangle - \langle \theta_h^*(\hat{V}_{h+1}^k), \phi(s,a) \rangle$

$\theta, \theta' \in B_h^k \cdot \|\theta - \theta'\|_{\Lambda_h^k} \leq 2\beta$.

$\leq 2\beta \cdot \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}$.

in other words.

$$\left(\hat{Q}_0^k(s_0^k, a_0^k) - Q_0^{\pi^k}(s_0^k, a_0^k) \right) I(E_k)$$

$$\le \mathbb{E}\left[\sum_{h=0}^{H-1} 2\beta \cdot \| \phi(s_h^k, a_h^k) \|_{(\Lambda_h^k)^{-1}} \,\middle|\, \mathcal{H}_{k-1} \right] I(E_k)$$

$$\mathcal{H}_{k-1} = \left\{ (s_h^{k'}, a_h^{k'})_{h=0, k'=0}^{H-1, k-1}, s_0^k, a_0^k \right\}.$$

Pf of ④: $\sum_{k=0}^{K-1} \sum_{h=0}^{H-1} \| \underbrace{\phi(s_h^k, a_h^k)}_{\phi_h^k} \|_{(\Lambda_h^k)^{-1}}$

$$\Lambda_h^k = I + \sum_{k'=0}^{k-1} \phi_h^{k'} (\phi_h^{k'})^T.$$

so we just need to bound

$$\sum_{k=0}^{K-1} \| \phi_h^k \|_{(\Lambda_h^k)^{-1}} \quad \text{for each } h \text{ separately}$$

(*)

The Elliptic potential Lemma: (EPL)

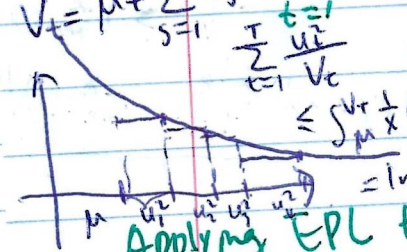
suppose $u_1 \cdots u_T \in \mathbb{R}^d$. $V_t = \mu I + \sum_{s=1}^{t} u_s u_s^T$. then

$$\sum_{t=1}^{T} \| u_t \|^2_{V_t^{-1}} \le \ln \frac{\det(V_T)}{\det(V_0)}$$

in 1-d:

$V_t = \mu + \sum_{s=1}^{t} u_s^2$

$\sum_{t=1}^{T} \frac{u_t^2}{V_t} \le \int_{\mu}^{V_T} \frac{1}{x} dx = \ln \frac{V_T}{\mu}.$



If all $\|u_t\| \le D$.

$\le d \ln \left(1 + \frac{D^2 T}{d\mu} \right) \simeq \tilde{O}(d)$

Applying EPL to (*):

$$(*) \le \sqrt{K \cdot \sum_{k=0}^{K-1} \| \underbrace{\phi_h^k}_{u_t} \|^2_{\underbrace{(\Lambda_h^k)^{-1}}_{V_{t-1}^{-1}}}}, \text{ not } V_t^{-1}!$$

$$\Lambda_h^{k+1} = \Lambda_h^k + \phi_h^k (\phi_h^k)^T$$
$$\preceq \Lambda_h^k + I$$
$$\preceq 2\Lambda_h^k \quad \text{(Exercise)}$$

$A \succeq B \Rightarrow A^{-1} \preceq B^{-1}$

$$\Rightarrow (\Lambda_h^k)^{-1} \preceq 2(\Lambda_h^{k+1})^{-1}$$

$$\leq \sqrt{K \sum_{k=0}^{K-1} \| \phi_h^k \|^2_{(\Lambda_h^{k+1})^{-1}}}$$

$$\leq \tilde{O}(\sqrt{kd}).$$

Pf of generalized simulation Lemma:

h, s_h, a_h: $\hat{Q}_h(s_h, a_h) - Q_h^\pi(s_h, a_h)$

Introduce Bellman error

$$= \hat{Q}_h(s_h, a_h) - (\mathcal{T}_h^\pi \hat{Q}_{h+1})(s_h, a_h) + (\mathcal{T}_h^\pi \hat{Q}_{h+1})(s_h, a_h) - (\mathcal{T}_h^\pi Q_{h+1}^\pi)(s_h, a_h)$$

$$= \text{Bell}_h^\pi(s_h, a_h) + \left[r_h(s_h, a_h) + \langle \cancel{P_h(\cdot|s_h, a_h)}, \cancel{\hat{Q}_{h+1}} \rangle \right.$$
$$\left. \sum_{s'} P_h(s'|s_h, a_h) \cdot \hat{Q}_{h+1}(s', \pi_{h+1}(s')) \right]$$

$$- \left[r_h(s_h, a_h) + \sum_{s'} P_h(s'|s_h, a_h) \cdot Q_{h+1}^\pi(s', \pi_{h+1}(s')) \right]$$

$$= \text{Bell}_h^\pi(s_h, a_h) + \mathbb{E}\left[+ \hat{Q}_{h+1}(s_{h+1}, a_{h+1}) - Q_{h+1}^\pi(s_{h+1}, a_{h+1}) \,\middle|\, s_h, a_h, \pi, M \right]$$

Keep expanding.