Exploration in Linear MDPs

Sham Kakade and Wen Sun

CS 6789: Foundations of Reinforcement Learning

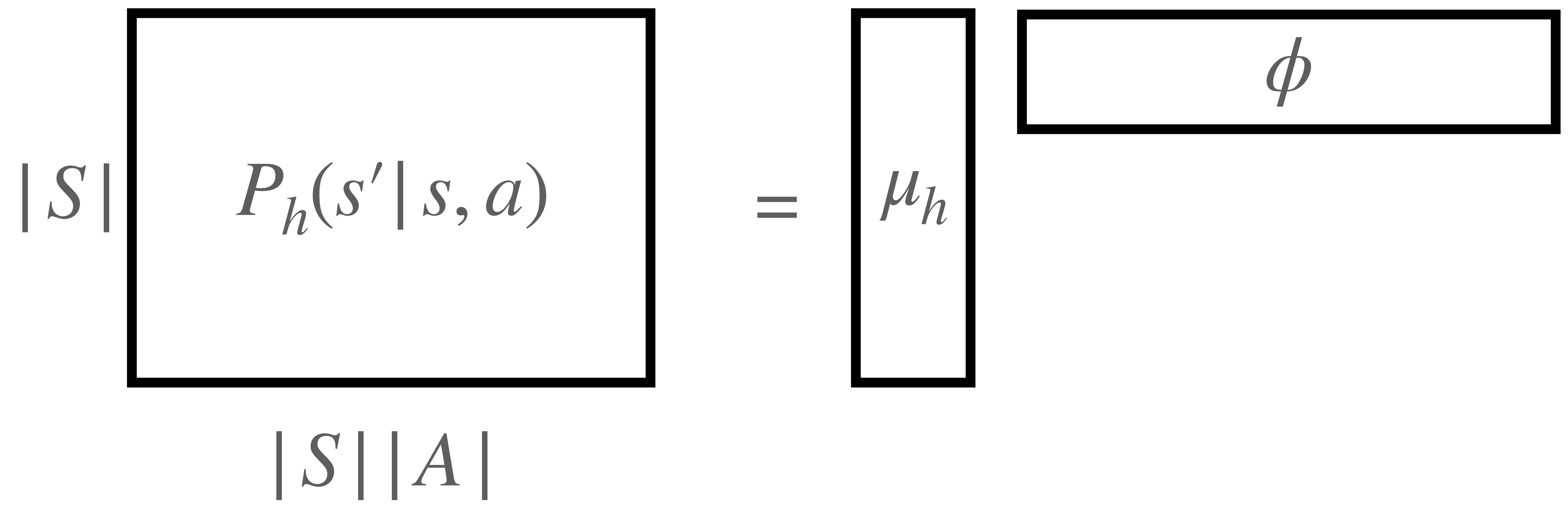
Recap: linear MDP definition

Finite horizon time-dependent episodic MDP $\mathscr{M} = \{S, A, H, \{r\}_h, \{P\}_h, s_0\}$

Recap: linear MDP definition

Finite horizon time-dependent episodic MDP $\mathscr{M} = \{S, A, H, \{r\}_h, \{P\}_h, s_0\}$

Low-Rank Decomposition:



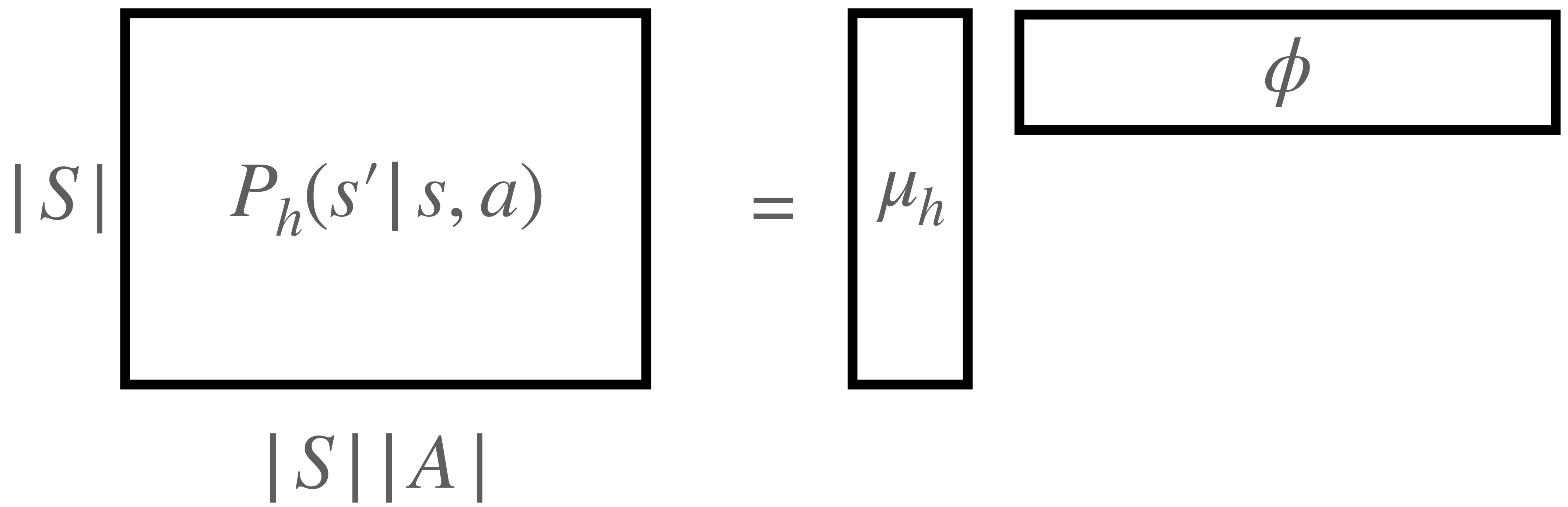
Recap: linear MDP definition

Finite horizon time-dependent episodic MDP $\mathscr{M} = \{S, A, H, \{r\}_h, \{P\}_h, s_0\}$

$$P_h(s' \mid s, a) = \mu_h^\star(s') \cdot \phi(s, a), \quad \mu_h^\star \in S \mapsto \mathbb{R}^d, \phi \in S \times A \mapsto \mathbb{R}^d$$

$$r(s, a) = \theta_h^\star \cdot \phi(s, a), \quad \theta_h^\star \in \mathbb{R}^d$$

Low-Rank Decomposition:



Recap:

1. Linear MDP's $Q_h^\star$ & Q_h^π are all linear functions wrt ϕ

Recap:

1. Linear MDP's $Q_h^\star$ & Q_h^π are all linear functions wrt ϕ

2. Since $P_h(\cdot \mid s, a) = \mu_h^\star \phi(s, a)$, we can estimate $\mu_h^\star$ by via regression

Recap:

1. Linear MDP's $Q_h^\star$ & Q_h^π are all linear functions wrt ϕ

2. Since $P_h(\cdot \mid s, a) = \mu_h^\star \phi(s, a)$, we can estimate $\mu_h^\star$ by via regression

Since given $s, a, s' \sim P_h^\star(\cdot \mid s, a)$, we know $\mathbb{E}_{s' \sim P_h^\star(\cdot \mid s, a)} \delta(s') = P_h^\star(\cdot \mid s, a)$,

Recap:

1. Linear MDP's $Q_h^\star$ & Q_h^π are all linear functions wrt ϕ

2. Since $P_h(\cdot \mid s, a) = \mu_h^\star \phi(s, a)$, we can estimate $\mu_h^\star$ by via regression

Since given $s, a, s' \sim P_h^\star(\cdot \mid s, a)$, we know $\mathbb{E}_{s' \sim P_h^\star(\cdot|s,a)} \delta(s') = P_h^\star(\cdot \mid s, a)$,

Given dataset $\{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}$

Ridge Linear Regression:

$$\min_{\mu} \sum_{i=1}^{n-1} \|\mu \phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\|_2^2 + \lambda \|\mu\|_F^2$$

Recap:

1. Linear MDP's $Q_h^\star$ & Q_h^π are all linear functions wrt ϕ

2. Since $P_h(\cdot \mid s, a) = \mu_h^\star \phi(s, a)$, we can estimate $\mu_h^\star$ by via regression

Since given $s, a, s' \sim P_h^\star(\cdot \mid s, a)$, we know $\mathbb{E}_{s' \sim P_h^\star(\cdot|s,a)} \delta(s') = P_h^\star(\cdot \mid s, a)$,

Given dataset $\{s_h^i, a_h^i, s_{h+1}^i\}_{i=1}^{n-1}$

Ridge Linear Regression:

$$\min_\mu \sum_{i=1}^{n-1} \|\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\|_2^2 + \lambda\|\mu\|_F^2$$

$$\widehat{\mu}_h^n = \sum_{i=1}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top(\Lambda_h^n)^{-1}; \quad \widehat{P}_h^n(\cdot \mid s, a) = \widehat{\mu}_h^n \phi(s, a)$$

Today:

Regret bound for the UCBVI algorithm for Linear MDP and its proof sketch

Outline:

1. Model fitting and its guarantee $(\hat{P}(\,.\,|\,s,a) - P(\,.\,|\,s,a))^\top V,$ for some fixed V

2. Covering argument to bound $(\hat{P}(\,.\,|\,s,a) - P(\,.\,|\,s,a))^\top V$, for ALL $V \in \mathcal{F}$

3. UCBVI revisit and its guarantee

1. Model Learning in Linear MDPs

$$\widehat{P}_h^n(\cdot \mid s, a) = \widehat{\mu}_h^n \phi(s, a)$$

1. Model Learning in Linear MDPs

$$\widehat{P}_h^n(\cdot \mid s, a) = \widehat{\mu}_h^n \phi(s, a)$$

Cannot directly bound $\| \widehat{P}^n(\cdot \mid s, a) - P(\cdot \mid s, a) \|_1$ since $P(\,.\mid s, a)$ is in $\mathbb{R}^{|S|}$

1. Model Learning in Linear MDPs

$$\widehat{P}_h^n(\cdot \mid s, a) = \widehat{\mu}_h^n \phi(s, a)$$

Cannot directly bound $\|\widehat{P}^n(\cdot \mid s, a) - P(\cdot \mid s, a)\|_1$ since $P(\,.\mid s, a)$ is in $\mathbb{R}^{|S|}$

As in tabular-UCBVI and Generative Model, we care **average model error:**

1. Model Learning in Linear MDPs

$$\widehat{P}_h^n(\cdot \mid s, a) = \widehat{\mu}_h^n \phi(s, a)$$

Cannot directly bound $\| \widehat{P}^n(\cdot \mid s, a) - P(\cdot \mid s, a) \|_1$ since $P(\,.\mid s, a)$ is in $\mathbb{R}^{|S|}$

As in tabular-UCBVI and Generative Model, we care **average model error:**

Consider a fixed function $V : S \mapsto [0,H]$, we can bound:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right) \cdot V \right|$$

1. Model Learning in Linear MDPs

Ridge Linear Regression: $$\min_{\mu} \sum_{i=1}^{n-1} \|\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\|_2^2 + \|\mu\|_F^2$$

$$\widehat{\mu}_h^n = \sum_{i=1}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}, \quad \Lambda_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$$

$$\widehat{P}_h^n(\cdot \mid s, a) = \widehat{\mu}_h^n \phi(s, a)$$

1. Model Learning in Linear MDPs

Ridge Linear Regression: $$\min_{\mu} \sum_{i=1}^{n-1} \|\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\|_2^2 + \|\mu\|_F^2$$

$$\widehat{\mu}_h^n = \sum_{i=1}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}, \quad \Lambda_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$$

$$\widehat{P}_h^n(\cdot \mid s, a) = \widehat{\mu}_h^n \phi(s, a)$$

Lemma [Model Average Error under a fixed V]:

Consider a fixed $V : S \to [0,H]$. With probability at least $1 - \delta$, for any s, a, h, n, we have:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right) \cdot V \right| \leq \|\phi(s, a)\|_{(\Lambda_h^n)^{-1}} \times \left(2H\sqrt{\ln \frac{H \det(\Lambda_h^n)^{1/2} \det(\lambda I)^{-1/2}}{\delta}} + H\sqrt{\lambda d} \right)$$

1. Model Learning in Linear MDPs: proof

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

1. Model Learning in Linear MDPs: proof

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\widehat{\mu}_h^n = \sum_{i=1}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}, \quad \Lambda_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$$

1. Model Learning in Linear MDPs: proof

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\widehat{\mu}_h^n = \sum_{i=1}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}, \quad \Lambda_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$$

$$\widehat{\mu}_h^n = \sum_{i=1}^{n-1} \left(P_h(\cdot \mid s_h^i, a_h^i) + \epsilon_h^i\right) \phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}$$

1. Model Learning in Linear MDPs: proof

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\widehat{\mu}_h^n = \sum_{i=1}^{n-1} \delta(s_{h+1}^i) \phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}, \quad \Lambda_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i) \phi(s_h^i, a_h^i)^\top + \lambda I$$

$$\widehat{\mu}_h^n = \sum_{i=1}^{n-1} \left(P_h(\cdot \mid s_h^i, a_h^i) + \epsilon_h^i\right) \phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1} = \sum_{i=1}^{n-1} \left(\mu_h^\star \phi(s_h^i, a_h^i) + \epsilon_h^i\right) \phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}$$

1. Model Learning in Linear MDPs: proof

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\widehat{\mu}_h^n = \sum_{i=1}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}, \quad \Lambda_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$$

$$\widehat{\mu}_h^n = \sum_{i=1}^{n-1} \left(P_h(\cdot \mid s_h^i, a_h^i) + \epsilon_h^i\right) \phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1} = \sum_{i=1}^{n-1} \left(\mu_h^\star \phi(s_h^i, a_h^i) + \epsilon_h^i\right) \phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}$$

$$= \mu_h^\star \left(\sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top \right) (\Lambda_h^n)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}$$

1. Model Learning in Linear MDPs: proof

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\widehat{\mu}_h^n = \sum_{i=1}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}, \quad \Lambda_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$$

$$\widehat{\mu}_h^n = \sum_{i=1}^{n-1} \left(P_h(\cdot \mid s_h^i, a_h^i) + \epsilon_h^i\right) \phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1} = \sum_{i=1}^{n-1} \left(\mu_h^\star \phi(s_h^i, a_h^i) + \epsilon_h^i\right) \phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}$$

$$= \mu_h^\star \left(\sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top \right) (\Lambda_h^n)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}$$

$$= \mu_h^\star - \lambda \mu_h^\star (\Lambda_h^n)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}$$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left((\widehat{\mu}_h^n - \mu_h^\star)\phi(s,a)\right)^\top V$$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left((\widehat{\mu}_h^n - \mu_h^\star)\phi(s,a)\right)^\top V$$

$$= -\lambda \phi(s,a)^\top \left(\Lambda_h^n\right)^{-1} \left(\mu_h^\star\right)^\top V + \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V$$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left((\widehat{\mu}_h^n - \mu_h^\star)\phi(s,a)\right)^\top V$$

$$= -\lambda \phi(s,a)^\top \left(\Lambda_h^n\right)^{-1} \left(\mu_h^\star\right)^\top V + \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V$$

$$\left| \lambda \phi(s,a)^\top \left(\Lambda_h^n\right)^{-1} \left(\mu_h^\star\right)^\top V \right| \leq \lambda \| \phi(s,a)^\top (\Lambda_h^n)^{-1/2} \|_2 \| (\Lambda_h^n)^{-1/2} (\mu_h^\star)^\top V \|_2$$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left((\widehat{\mu}_h^n - \mu_h^\star)\phi(s,a)\right)^\top V$$

$$= -\lambda \phi(s,a)^\top \left(\Lambda_h^n\right)^{-1} \left(\mu_h^\star\right)^\top V + \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V$$

$$\left| \lambda \phi(s,a)^\top \left(\Lambda_h^n\right)^{-1} \left(\mu_h^\star\right)^\top V \right| \leq \lambda \| \phi(s,a)^\top (\Lambda_h^n)^{-1/2} \|_2 \| (\Lambda_h^n)^{-1/2} (\mu_h^\star)^\top V \|_2$$

$$= \lambda \| \phi(s,a) \|_{(\Lambda_h^n)^{-1}} \| (\Lambda_h^n)^{-1/2} (\mu_h^\star)^\top V \|_2$$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left((\widehat{\mu}_h^n - \mu_h^\star)\phi(s,a)\right)^\top V$$

$$= -\lambda \phi(s,a)^\top \left(\Lambda_h^n\right)^{-1} \left(\mu_h^\star\right)^\top V + \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V$$

$$\left| \lambda \phi(s,a)^\top \left(\Lambda_h^n\right)^{-1} \left(\mu_h^\star\right)^\top V \right| \leq \lambda \|\phi(s,a)^\top (\Lambda_h^n)^{-1/2}\|_2 \|(\Lambda_h^n)^{-1/2} (\mu_h^\star)^\top V\|_2$$

$$= \lambda \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \|(\Lambda_h^n)^{-1/2} (\mu_h^\star)^\top V\|_2$$

$$\leq \lambda \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \|(\Lambda_h^n)^{-1/2}\|_2 \|(\mu_h^\star)^\top V\|_2$$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left((\widehat{\mu}_h^n - \mu_h^\star)\phi(s,a)\right)^\top V$$

$$= -\lambda \phi(s,a)^\top \left(\Lambda_h^n\right)^{-1} \left(\mu_h^\star\right)^\top V + \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V$$

$$\left|\lambda \phi(s,a)^\top \left(\Lambda_h^n\right)^{-1} \left(\mu_h^\star\right)^\top V\right| \leq \lambda \|\phi(s,a)^\top (\Lambda_h^n)^{-1/2}\|_2 \|(\Lambda_h^n)^{-1/2}(\mu_h^\star)^\top V\|_2$$

$$= \lambda \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \|(\Lambda_h^n)^{-1/2}(\mu_h^\star)^\top V\|_2$$

$$\leq \lambda \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \|(\Lambda_h^n)^{-1/2}\|_2 \|(\mu_h^\star)^\top V\|_2$$

Normalization assumption on $\mu_h^\star$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left((\widehat{\mu}_h^n - \mu_h^\star)\phi(s,a)\right)^\top V$$

$$= -\lambda \phi(s,a)^\top \left(\Lambda_h^n\right)^{-1} \left(\mu_h^\star\right)^\top V + \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V$$

$$\left| \lambda \phi(s,a)^\top \left(\Lambda_h^n\right)^{-1} \left(\mu_h^\star\right)^\top V \right| \leq \lambda \|\phi(s,a)^\top (\Lambda_h^n)^{-1/2}\|_2 \|(\Lambda_h^n)^{-1/2} (\mu_h^\star)^\top V\|_2$$

$$= \lambda \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \|(\Lambda_h^n)^{-1/2} (\mu_h^\star)^\top V\|_2$$

$$\leq \lambda \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \|(\Lambda_h^n)^{-1/2}\|_2 \|(\mu_h^\star)^\top V\|_2 \leq \lambda \|\phi(s,a)^\top\|_{(\Lambda_h^n)^{-1}} \frac{H\sqrt{d}}{\sqrt{\lambda}}$$

Normalization assumption on $\mu_h^\star$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left| \left((\widehat{\mu}_h^n - \mu_h^\star)\phi(s,a) \right)^\top V \right| \leq H\sqrt{\lambda d} \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} + \left| \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right|$$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left| \left((\widehat{\mu}_h^n - \mu_h^\star)\phi(s,a) \right)^\top V \right| \leq H\sqrt{\lambda d} \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} + \left| \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right|$$

$$\left| \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right| \leq \|\phi(s,a)^\top (\Lambda_h^n)^{-1/2}\|_2 \left\| \sum_{i=1}^{n-1} (\Lambda_h^n)^{-1/2} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right\|_2$$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left| \left((\widehat{\mu}_h^n - \mu_h^\star)\phi(s,a) \right)^\top V \right| \leq H\sqrt{\lambda d} \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} + \left| \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right|$$

$$\left| \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right| \leq \|\phi(s,a)^\top (\Lambda_h^n)^{-1/2}\|_2 \left\| \sum_{i=1}^{n-1} (\Lambda_h^n)^{-1/2} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right\|_2$$

$$= \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \left\| \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\left((\epsilon_h^i)^\top V\right) \right\|_{(\Lambda_h^n)^{-1}}$$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left| \left((\widehat{\mu}_h^n - \mu_h^\star) \phi(s,a) \right)^\top V \right| \leq H\sqrt{\lambda d} \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} + \left| \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i) (\epsilon_h^i)^\top V \right|$$

$$\left| \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i) (\epsilon_h^i)^\top V \right| \leq \|\phi(s,a)^\top (\Lambda_h^n)^{-1/2}\|_2 \left\| \sum_{i=1}^{n-1} (\Lambda_h^n)^{-1/2} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right\|_2$$

$$= \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \left\| \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i) \left((\epsilon_h^i)^\top V \right) \right\|_{(\Lambda_h^n)^{-1}}$$

$$\mathbb{E}\left[(\epsilon_h^i)^\top V \,|\, \mathscr{H}_{i,h} \right] = 0, \quad |(\epsilon_h^i)^\top V| \leq \|\epsilon_h^i\|_1 \|V\|_\infty \leq 2H$$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left| \left((\widehat{\mu}_h^n - \mu_h^\star) \phi(s,a) \right)^\top V \right| \leq H\sqrt{\lambda d} \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} + \left| \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i) (\epsilon_h^i)^\top V \right|$$

$$\left| \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right| \leq \|\phi(s,a)^\top (\Lambda_h^n)^{-1/2}\|_2 \left\| \sum_{i=1}^{n-1} (\Lambda_h^n)^{-1/2} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right\|_2$$

$$= \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \left\| \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i) \left((\epsilon_h^i)^\top V \right) \right\|_{(\Lambda_h^n)^{-1}}$$

$$\leq \|\phi(s,a)^\top\|_{(\Lambda_h^n)^{-1}} \times 2H \sqrt{\ln \frac{\det(\Lambda_h^n)^{1/2} \det(\lambda I)^{-1/2}}{\delta}}$$

$$\mathbb{E}\left[(\epsilon_h^i)^\top V \mid \mathscr{H}_{i,h} \right] = 0, \quad |(\epsilon_h^i)^\top V| \leq \|\epsilon_h^i\|_1 \|V\|_\infty \leq 2H$$

1. Model Learning in Linear MDPs

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}$$

$$\left| \left((\widehat{\mu}_h^n - \mu_h^\star) \phi(s,a) \right)^\top V \right| \leq H\sqrt{\lambda d} \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} + \left| \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right|$$

$$\left| \sum_{i=1}^{n-1} \phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right| \leq \|\phi(s,a)^\top (\Lambda_h^n)^{-1/2}\|_2 \left\| \sum_{i=1}^{n-1} (\Lambda_h^n)^{-1/2} \phi(s_h^i, a_h^i)(\epsilon_h^i)^\top V \right\|_2$$

$$= \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \left\| \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\left((\epsilon_h^i)^\top V\right) \right\|_{(\Lambda_h^n)^{-1}} \qquad \leq \|\phi(s,a)^\top\|_{(\Lambda_h^n)^{-1}} \times 2H \sqrt{\ln \frac{\det(\Lambda_h^n)^{1/2} \det(\lambda I)^{-1/2}}{\delta}}$$

With prob $1-\delta, \ \forall n$

$$\mathbb{E}\left[(\epsilon_h^i)^\top V \,|\, \mathscr{H}_{i,h}\right] = 0, \quad |(\epsilon_h^i)^\top V| \leq \|\epsilon_h^i\|_1 \|V\|_\infty \leq 2H$$

1. Model Learning in Linear MDPs

Lemma [Model Average Error under a fixed V]:

Consider a fixed $V : S \to [0,H]$. With probability at least $1-\delta$, for all s, a, n, h, we have:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V \right| \leq \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \times \left(2H\sqrt{\ln \frac{H\det(\Lambda_h^n)^{1/2}\det(\lambda I)^{-1/2}}{\delta}} + H\sqrt{\lambda d} \right)$$

1. Model Learning in Linear MDPs

Lemma [Model Average Error under a fixed V]:

Consider a fixed $V : S \to [0,H]$. With probability at least $1-\delta$, for all s, a, n, h, we have:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V \right| \leq \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \times \left(2H\sqrt{\ln \frac{H \det(\Lambda_h^n)^{1/2} \det(\lambda I)^{-1/2}}{\delta}} + H\sqrt{\lambda d} \right)$$

$$\det(\Lambda_h^n)^{1/2}\det(\lambda I)^{-1/2} \leq (n+\lambda)^{d/2}\lambda^{-d/2} = (N/\lambda + 1)^{d/2}$$

1. Model Learning in Linear MDPs

Lemma [Model Average Error under a fixed V]:

Consider a fixed $V : S \to [0,H]$. With probability at least $1-\delta$, for all s, a, n, h, we have:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V \right| \leq \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \times \left(2H\sqrt{\ln \frac{H \det(\Lambda_h^n)^{1/2} \det(\lambda I)^{-1/2}}{\delta}} + H\sqrt{\lambda d} \right)$$

$$\det(\Lambda_h^n)^{1/2}\det(\lambda I)^{-1/2} \leq (n+\lambda)^{d/2}\lambda^{-d/2} = (N/\lambda + 1)^{d/2}$$

$$\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V \right| \leq \| \phi(s,a) \|_{(\Lambda_h^n)^{-1}} \left(2H\sqrt{d \ln\left(\frac{NH}{\lambda} + 1\right) + \ln\left(\frac{1}{\delta}\right)} + H\sqrt{\lambda d} \right)$$

1. Model Learning in Linear MDPs

Lemma [Model Average Error under a fixed V]:

Consider a fixed $V : S \to [0,H]$. With probability at least $1-\delta$, for all s, a, n, h, we have:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V \right| \leq \| \phi(s,a) \|_{(\Lambda_h^n)^{-1}} \times \left(2H \sqrt{\ln \frac{H \det(\Lambda_h^n)^{1/2} \det(\lambda I)^{-1/2}}{\delta}} + H\sqrt{\lambda d} \right)$$

$$\det(\Lambda_h^n)^{1/2} \det(\lambda I)^{-1/2} \leq (n+\lambda)^{d/2} \lambda^{-d/2} = (N/\lambda + 1)^{d/2}$$

$$\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V \right| \leq \| \phi(s,a) \|_{(\Lambda_h^n)^{-1}} \left(2H \sqrt{d \ln \left(\frac{NH}{\lambda} + 1 \right) + \ln \left(\frac{1}{\delta} \right)} + H\sqrt{\lambda d} \right)$$

$$= \widetilde{O} \left(H\sqrt{d} + H\sqrt{\ln(1/\delta)} \right) \| \phi(s,a) \|_{(\Lambda_h^n)^{-1}}$$

1. Model learning: summary

Lemma [Model Average Error under a fixed V]:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right) \cdot V \right| = \widetilde{O}\left(H\sqrt{d} + H\sqrt{\ln(1/\delta)} \right) \| \phi(s, a) \|_{(\Lambda_h^n)^{-1}}$$

1. Model learning: summary

Lemma [Model Average Error under a fixed V]:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V \right| = \widetilde{O}\left(H\sqrt{d} + H\sqrt{\ln(1/\delta)} \right) \| \phi(s,a) \|_{(\Lambda_h^n)^{-1}}$$

Q: Can we get a uniform convergence argument for a function class $\mathscr{F}$?

Outline:

1. Model fitting and its guarantee $(\hat{P}(\,.\,|s,a) - P(\,.\,|s,a))^\top V,$ for some fixed V

2. Covering argument to bound $(\hat{P}(\,.\,|s,a) - P(\,.\,|s,a))^\top V$, for ALL $V \in \mathscr{F}$

3. UCBVI revisit and its guarantee

Detour: Covering Number

Consider the ball $\Theta = \{\theta : \theta \in \mathbb{R}^d, \|\theta\|_2 \leq R\}$.

Denote ϵ**-Net** as a subset $\mathcal{N}_\epsilon \subseteq \Theta$, such that $\forall \theta \in \Theta$:

$\exists \theta' \in \mathcal{N}_\epsilon,$ s.t. $\|\theta' - \theta\|_2 \leq \epsilon$.

Denote ϵ**-cover** as the smallest $\mathcal{N}_\epsilon$

Detour: Covering Number

Consider the ball $\Theta = \{\theta : \theta \in \mathbb{R}^d, \|\theta\|_2 \leq R\}$.

Denote ϵ**-Net** as a subset $\mathcal{N}_\epsilon \subseteq \Theta$, such that $\forall \theta \in \Theta$:

$\exists \theta' \in \mathcal{N}_\epsilon,$ s.t. $\|\theta' - \theta\|_2 \leq \epsilon$.

Denote ϵ**-cover** as the smallest $\mathcal{N}_\epsilon$

Lemma [Covering of Θ] We have $|\mathcal{N}_\epsilon| \leq (1 + 2R/\epsilon)^d$, and $\ln(|\mathcal{N}_\epsilon|) \leq d \ln(1 + 2R/\epsilon)$

Detour: Covering Number

Consider the ball $\Theta = \{\theta : \theta \in \mathbb{R}^d, \|\theta\|_2 \leq R\}$.

Denote ϵ-**Net** as a subset $\mathcal{N}_\epsilon \subseteq \Theta$, such that $\forall \theta \in \Theta$:

$\exists \theta' \in \mathcal{N}_\epsilon,$ s.t. $\|\theta' - \theta\|_2 \leq \epsilon$.

Denote ϵ-**cover** as the smallest $\mathcal{N}_\epsilon$

Lemma [Covering of Θ] We have $|\mathcal{N}_\epsilon| \leq (1 + 2R/\epsilon)^d$, and $\ln(|\mathcal{N}_\epsilon|) \leq d \ln(1 + 2R/\epsilon)$

Now consider a function class $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$,

and for any $f_{\theta_1}, f_{\theta_2} \in \mathcal{F}$, $\|f_{\theta_1} - f_{\theta_2}\|_\infty \leq L\|\theta_1 - \theta_2\|_2$

Detour: Covering Number

Consider the ball $\Theta = \{\theta : \theta \in \mathbb{R}^d, \|\theta\|_2 \leq R\}$.

Denote ϵ**-Net** as a subset $\mathcal{N}_\epsilon \subseteq \Theta$, such that $\forall \theta \in \Theta$:

$\exists \theta' \in \mathcal{N}_\epsilon,$ s.t. $\|\theta' - \theta\|_2 \leq \epsilon$.

Denote ϵ**-cover** as the smallest $\mathcal{N}_\epsilon$

Lemma [Covering of Θ] We have $|\mathcal{N}_\epsilon| \leq (1 + 2R/\epsilon)^d$, and $\ln(|\mathcal{N}_\epsilon|) \leq d \ln(1 + 2R/\epsilon)$

Now consider a function class $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$,

and for any $f_{\theta_1}, f_{\theta_2} \in \mathcal{F}$, $\|f_{\theta_1} - f_{\theta_2}\|_\infty \leq L\|\theta_1 - \theta_2\|_2$

Then $(\epsilon/L)-$Net on Θ gives us an $\epsilon-$Net on $\mathcal{F}$ with $d(f_{\theta_1}, f_{\theta_2}) := \|f_{\theta_1} - f_{\theta_2}\|_\infty$

Detour: Covering Number and An Example

Consider a specific parameterization $\theta = (w, \beta, \Lambda)$,

$$\Theta = \{(w, \beta, \Lambda) : \|w\|_2 \leq L, \beta \in [0,B], \sigma_{\min}(\Lambda) \geq \lambda\}$$

Detour: Covering Number and An Example

Consider a specific parameterization $\theta = (w, \beta, \Lambda)$,

$$\Theta = \{(w, \beta, \Lambda) : \|w\|_2 \leq L, \beta \in [0,B], \sigma_{\min}(\Lambda) \geq \lambda\}$$

Define the function $f_{w,\beta,\Lambda} : S \to [0,H]$

$$f_{w,\beta,\Lambda}(s) := \min\left\{ \max_a \left(w^\top \phi(s,a) + \beta \sqrt{\phi(s,a)^\top \Lambda^{-1} \phi(s,a)} \right), H \right\}$$

Detour: Covering Number and An Example

Consider a specific parameterization $\theta = (w, \beta, \Lambda)$,
$\Theta = \{(w, \beta, \Lambda) : \|w\|_2 \leq L, \beta \in [0,B], \sigma_{\min}(\Lambda) \geq \lambda\}$

Define the function $f_{w,\beta,\Lambda} : S \to [0,H]$

$$f_{w,\beta,\Lambda}(s) := \min \left\{ \max_a \left(w^\top \phi(s,a) + \beta \sqrt{\phi(s,a)^\top \Lambda^{-1} \phi(s,a)} \right), H \right\}$$

Denote $\mathcal{F} = \{f_{w,\beta,\Lambda} : \|w\|_2 \leq L, \beta \in [0,B], \sigma_{\min}(\Lambda) \geq \lambda\}$, **what's the covering number of $\mathcal{F}$ under ℓ_∞**

Detour: Covering Number

$$f_{w,\beta,\Lambda} :, f_{w,\beta,\Lambda}(s) := \min\left\{ \max_a \left(w^\top \phi(s,a) + \beta\sqrt{\phi(s,a)^\top \Lambda^{-1} \phi(s,a)} \right), H \right\}$$

Detour: Covering Number

$$f_{w,\beta,\Lambda} :, f_{w,\beta,\Lambda}(s) := \min \left\{ \max_a \left(w^\top \phi(s,a) + \beta\sqrt{\phi(s,a)^\top \Lambda^{-1} \phi(s,a)} \right), H \right\}$$

Lemma: Denote $\mathcal{F} = \{f_{w,\beta,\Lambda} : \|w\|_2 \leq L, \beta \in [0,B], \sigma_{\min}(\Lambda) \geq \lambda\}$,

under ℓ_∞ we have: $\ln|\mathcal{N}_\epsilon| \leq d\ln(1 + 6L/\epsilon) + 2d^2 \ln(1 + 18B^2\sqrt{d}/(\lambda\epsilon^2)) = \widetilde{O}\left(d^2\right)$

Detour: Covering Number

$$f_{w,\beta,\Lambda} :, f_{w,\beta,\Lambda}(s) := \min \left\{ \max_a \left(w^\top \phi(s,a) + \beta \sqrt{\phi(s,a)^\top \Lambda^{-1} \phi(s,a)} \right), H \right\}$$

Lemma: Denote $\mathcal{F} = \{f_{w,\beta,\Lambda} : \|w\|_2 \leq L, \beta \in [0,B], \sigma_{\min}(\Lambda) \geq \lambda\}$,

under ℓ_∞ we have: $\ln |\mathcal{N}_\epsilon| \leq d \ln(1 + 6L/\epsilon) + 2d^2 \ln(1 + 18B^2\sqrt{d}/(\lambda\epsilon^2)) = \widetilde{O}\left(d^2\right)$

Key step in the proof:

$$\left| f_\theta(s) - f_{\hat{\theta}}(s) \right| \leq \|w - \hat{w}\|_2 + |\beta - \hat{\beta}|/\sqrt{\lambda} + B\sqrt{\|\Lambda^{-1} - \hat{\Lambda}^{-1}\|_F}$$

Detour: Uniform Convergence

$$f_{w,\beta,\Lambda} :, f_{w,\beta,\Lambda}(s) := \min\left\{ \max_a \left(w^\top \phi(s,a) + \beta\sqrt{\phi(s,a)^\top \Lambda^{-1} \phi(s,a)} \right), H \right\}$$

Lemma: Denote $\mathscr{F} = \{f_{w,\beta,\Lambda} : \|w\|_2 \leq L, \beta \in [0,B], \sigma_{\min}(\Lambda) \geq \lambda\}$,
under ℓ_∞ we have: $\ln|\mathcal{N}_\epsilon| \leq d\ln(1+6L/\epsilon) + 2d^2\ln(1+18B^2\sqrt{d}/(\lambda\epsilon^2)) = \widetilde{O}\left(d^2\right)$

Lemma [uniform convergence]: With probability at least $1-\delta$, for all s, a, h, n, **and ALL** $f \in \mathscr{F}$:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot f \right| = \widetilde{O}(Hd) \cdot \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}}$$

Detour: Uniform Convergence

$$f_{w,\beta,\Lambda} :, f_{w,\beta,\Lambda}(s) := \min \left\{ \max_a \left(w^\top \phi(s,a) + \beta \sqrt{\phi(s,a)^\top \Lambda^{-1} \phi(s,a)} \right), H \right\}$$

Lemma: Denote $\mathscr{F} = \{f_{w,\beta,\Lambda} : \|w\|_2 \leq L, \beta \in [0,B], \sigma_{\min}(\Lambda) \geq \lambda\}$,

under ℓ_∞ we have: $\ln|\mathcal{N}_\epsilon| \leq d\ln(1 + 6L/\epsilon) + 2d^2\ln(1 + 18B^2\sqrt{d}/(\lambda\epsilon^2)) = \widetilde{O}\left(d^2\right)$

Lemma [uniform convergence]: With probability at least $1-\delta$, for all s, a, h, n, **and ALL** $f \in \mathscr{F}$:

$$\left| \left(\widehat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a) \right) \cdot f \right| = \widetilde{O}(Hd) \cdot \| \phi(s,a) \|_{(\Lambda_h^n)^{-1}}$$

Proof sketch: the model error we had for a fixed V + ϵ-Net argument
(Same high level steps as the ones we used for HW1's last question)

Detour: Uniform Convergence

$$f_{w,\beta,\Lambda} :, f_{w,\beta,\Lambda}(s) := \min \left\{ \max_a \left(w^\top \phi(s,a) + \beta \sqrt{\phi(s,a)^\top \Lambda^{-1} \phi(s,a)} \right), H \right\}$$

Lemma: Denote $\mathscr{F} = \{f_{w,\beta,\Lambda} : \|w\|_2 \leq L, \beta \in [0,B], \sigma_{\min}(\Lambda) \geq \lambda\}$,
under ℓ_∞ we have: $\ln |\mathcal{N}_\epsilon| \leq d \ln(1 + 6L/\epsilon) + 2d^2 \ln(1 + 18B^2\sqrt{d}/(\lambda\epsilon^2)) = \widetilde{O}\left(d^2\right)$

Lemma [uniform convergence]: With probability at least $1-\delta$, for all s, a, h, n, **and ALL** $f \in \mathscr{F}$:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot f \right| = \widetilde{O}(Hd) \cdot \| \phi(s,a) \|_{(\Lambda_h^n)^{-1}}$$

This will be our bonus term

Proof sketch: the model error we had for a fixed V + ϵ-Net argument
(Same high level steps as the ones we used for HW1's last question)

Summary of Covering Argument

Covering allows us to build a uniform convergence result (i.e., $\forall f \in \mathcal{F}$)
over a infinite hypothesis class
(Intuitively, log of covering number scales w.r.t to the # of parameters)

Outline:

1. Model fitting and its guarantee $(\hat{P}(\,.\,|s,a) - P(\,.\,|s,a))^\top V$, for some fixed V

2. Covering argument to bound $(\hat{P}(\,.\,|s,a) - P(\,.\,|s,a))^\top V$, for ALL $V \in \mathcal{F}$

3. UCBVI revisit and its guarantee

Algorithm: UCBVI in Linear MDPs

At the beginning of iteration n:

1. Learn transition model $\{\widehat{P}_h^n\}_{h=0}^{H-1}$ from all previous data via Ridge linear regression

comment: $\min_{\mu} \sum_{i=1}^{n-1} \|\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\|_2^2 + \lambda\|\mu\|_F^2$

Algorithm: UCBVI in Linear MDPs

At the beginning of iteration n:

1. Learn transition model $\{\widehat{P}_h^n\}_{h=0}^{H-1}$ from all previous data via Ridge linear regression

comment: $\min_{\mu} \sum_{i=1}^{n-1} \|\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\|_2^2 + \lambda\|\mu\|_F^2$

2. Design reward bonus $b_h^n(s,a) = \beta\sqrt{\phi(s,a)(\Lambda_h^n)^{-1}\phi(s,a)}$

Algorithm: UCBVI in Linear MDPs

At the beginning of iteration n:

1. Learn transition model $\{\widehat{P}_h^n\}_{h=0}^{H-1}$ from all previous data via Ridge linear regression

comment: $\min_{\mu} \sum_{i=1}^{n-1} \|\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\|_2^2 + \lambda\|\mu\|_F^2$

2. Design reward bonus $b_h^n(s,a) = \beta\sqrt{\phi(s,a)(\Lambda_h^n)^{-1}\phi(s,a)}$

#Comment: $\beta = \widetilde{O}(Hd)$, reward bonus upper bounds $\left|(\widehat{P}_h^n(\cdot\,|\,s,a) - P_h(\cdot\,|\,s,a))\cdot f\right|, \forall f \in \mathcal{F}$

Algorithm: UCBVI in Linear MDPs

At the beginning of iteration n:

1. Learn transition model $\{\widehat{P}_h^n\}_{h=0}^{H-1}$ from all previous data via Ridge linear regression

comment: $\min_{\mu} \sum_{i=1}^{n-1} \|\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\|_2^2 + \lambda\|\mu\|_F^2$

2. Design reward bonus $b_h^n(s,a) = \beta\sqrt{\phi(s,a)(\Lambda_h^n)^{-1}\phi(s,a)}$

#Comment: $\beta = \widetilde{O}(Hd)$, reward bonus upper bounds $\left|(\widehat{P}_h^n(\cdot\,|\,s,a) - P_h(\cdot\,|\,s,a))\cdot f\right|, \forall f \in \mathcal{F}$

3. Plan: π^{n+1} = Value-Iter $\left(\{\widehat{P}^n\}_h, \{r_h + b_h^n\}\right)$

Algorithm: UCBVI in Linear MDPs

At the beginning of iteration n:

1. Learn transition model $\{\widehat{P}_h^n\}_{h=0}^{H-1}$ from all previous data via Ridge linear regression

comment: $\min_{\mu} \sum_{i=1}^{n-1} \|\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)\|_2^2 + \lambda\|\mu\|_F^2$

2. Design reward bonus $b_h^n(s, a) = \beta\sqrt{\phi(s, a)(\Lambda_h^n)^{-1}\phi(s, a)}$

#Comment: $\beta = \widetilde{O}(Hd)$, reward bonus upper bounds $\left|(\widehat{P}_h^n(\cdot \mid s, a) - P_h(\cdot \mid s, a)) \cdot f\right|, \forall f \in \mathcal{F}$

3. Plan: π^{n+1} = Value-Iter $\left(\{\widehat{P}^n\}_h, \{r_h + b_h^n\}\right)$

4. Execute π^{n+1} for H steps

Regret bound for UCBVI in linear MDP

$$\mathbb{E}\left[\sum_{n=1}^{N}\left(V^{\star} - V^{\pi^n}\right)\right] \leq \widetilde{O}\left(H^2 d^{1.5}\sqrt{N}\right)$$

No S or A polynomial dependence!

2. Value Iteration in the Learned Model w/ Reward Bonus

$$\pi^{n+1} = \text{Value-Iter}\left(\{ \widehat{P}^{\,n} \}_h, \{ r_h + b_h^n \} \right)$$

2. Value Iteration in the Learned Model w/ Reward Bonus

$$\pi^{n+1} = \text{Value-Iter}\left(\{ \widehat{P}^{n} \}_h, \{ r_h + b_h^n \} \right)$$

$\widehat{V}^{n}_{H}(s) = 0, \forall s,$

2. Value Iteration in the Learned Model w/ Reward Bonus

$$\pi^{n+1} = \text{Value-Iter}\left(\{ \widehat{P}^n \}_h, \{ r_h + b_h^n \} \right)$$

$\widehat{V}_H^n(s) = 0, \forall s,$

$\widehat{Q}_h^n(s, a) = r_h(s, a) + b_h^n(s, a) + \widehat{P}_h^n(\cdot \mid s, a) \cdot \widehat{V}_{h+1}^n$

2. Value Iteration in the Learned Model w/ Reward Bonus

$$\pi^{n+1} = \text{Value-Iter}\left(\{ \widehat{P}^n \}_h, \{ r_h + b_h^n \} \right)$$

$$\widehat{V}_H^n(s) = 0, \forall s,$$

$$\widehat{Q}_h^n(s,a) = r_h(s,a) + b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n$$

$$= \theta_h^\star \cdot \phi(s,a) + \beta \sqrt{\phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s,a)} + \left(\widehat{\mu}_h^n \phi(s,a) \right)^\top \widehat{V}_{h+1}^n$$

2. Value Iteration in the Learned Model w/ Reward Bonus

$$\pi^{n+1} = \text{Value-Iter}\left(\{ \widehat{P}^n \}_h, \{ r_h + b_h^n \} \right)$$

$\widehat{V}^n_H(s) = 0, \forall s,$

$$\widehat{Q}^n_h(s,a) = r_h(s,a) + b_h^n(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1}$$

$$= \theta_h^\star \cdot \phi(s,a) + \beta\sqrt{\phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s,a)} + \left(\widehat{\mu}_h^n \phi(s,a) \right)^\top \widehat{V}^n_{h+1}$$

$$= \beta\sqrt{\phi(s,a)^\top (\Lambda_h^n)^{-1} \phi(s,a)} + \left(\theta_h^\star + (\widehat{\mu}_h^n)^\top \widehat{V}^n_{h+1}(s') \right)^\top \phi(s,a)$$

2. Value Iteration in the Learned Model w/ Reward Bonus

$$\pi^{n+1} = \text{Value-Iter}\left(\{\widehat{P}^n\}_h, \{r_h + b_h^n\}\right)$$

$$\widehat{V}_H^n(s) = 0, \forall s,$$

$$\widehat{Q}_h^n(s,a) = r_h(s,a) + b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n$$

$$= \theta_h^\star \cdot \phi(s,a) + \beta\sqrt{\phi(s,a)^\top(\Lambda_h^n)^{-1}\phi(s,a)} + \left(\widehat{\mu}_h^n\phi(s,a)\right)^\top \widehat{V}_{h+1}^n$$

$$= \beta\sqrt{\phi(s,a)^\top(\Lambda_h^n)^{-1}\phi(s,a)} + \left(\theta_h^\star + (\widehat{\mu}_h^n)^\top \widehat{V}_{h+1}^n(s')\right)^\top \phi(s,a)$$

$$= \beta\sqrt{\phi(s,a)^\top(\Lambda_h^n)^{-1}\phi(s,a)} + \phi(s,a)^\top \widehat{w}_h^n$$

2. Value Iteration in the Learned Model w/ Reward Bonus

$$\pi^{n+1} = \text{Value-Iter}\left(\{\widehat{P}^n\}_h, \{r_h + b_h^n\}\right)$$

$$\widehat{V}_H^n(s) = 0, \forall s,$$

$$\widehat{Q}_h^n(s,a) = r_h(s,a) + b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n$$

$$= \theta_h^\star \cdot \phi(s,a) + \beta\sqrt{\phi(s,a)^\top (\Lambda_h^n)^{-1}\phi(s,a)} + \left(\widehat{\mu}_h^n \phi(s,a)\right)^\top \widehat{V}_{h+1}^n$$

$$= \beta\sqrt{\phi(s,a)^\top (\Lambda_h^n)^{-1}\phi(s,a)} + \left(\theta_h^\star + (\widehat{\mu}_h^n)^\top \widehat{V}_{h+1}^n(s')\right)^\top \phi(s,a)$$

$$= \beta\sqrt{\phi(s,a)^\top (\Lambda_h^n)^{-1}\phi(s,a)} + \phi(s,a)^\top \widehat{w}_h^n$$

$$\widehat{V}_h^n(s) = \min\left\{\max_a \left(\phi(s,a)^\top \widehat{w}_h^n + \beta\sqrt{\phi(s,a)^\top (\Lambda_h^n)^{-1}\phi(s,a)}\right), H\right\}, \quad \pi_h^n(s) = \arg\max_a \widehat{Q}_h^n(s,a)$$

3. Prove Optimism

Lemma [Optimism]: with high probability, for all n, h, s:

$$\widehat{V}_h^n(s) \geq V_h^\star(s)$$

3. Prove Optimism

Lemma [Optimism]: with high probability, for all n, h, s:

$$\widehat{V}_h^n(s) \geq V_h^\star(s)$$

Proof Sketch: let's do induction here with Inductive hypothesis: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \forall s$

3. Prove Optimism

Lemma [Optimism]: with high probability, for all n, h, s:

$$\widehat{V}_h^n(s) \geq V_h^\star(s)$$

Proof Sketch: let's do induction here with Inductive hypothesis: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \forall s$

$$\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) = b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - P_h(\cdot \mid s,a) \cdot V_{h+1}^\star$$

3. Prove Optimism

Lemma [Optimism]: with high probability, for all n, h, s:
$$\widehat{V}_h^n(s) \geq V_h^\star(s)$$

Proof Sketch: let's do induction here with Inductive hypothesis: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \forall s$

$$\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) = b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - P_h(\cdot \mid s,a) \cdot V_{h+1}^\star$$

$$\geq b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - P_h(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n$$

3. Prove Optimism

Lemma [Optimism]: with high probability, for all n, h, s:

$$\widehat{V}_h^n(s) \geq V_h^\star(s)$$

Proof Sketch: let's do induction here with Inductive hypothesis: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \forall s$

$$\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) = b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - P_h(\cdot \mid s,a) \cdot V_{h+1}^\star$$

$$\geq b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - P_h(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n$$

NOTE this is different from what we did in tabular MDP!!!

3. Prove Optimism

Lemma [Optimism]: with high probability, for all n, h, s:

$$\widehat{V}_h^n(s) \geq V_h^\star(s)$$

Proof Sketch: let's do induction here with Inductive hypothesis: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \forall s$

$$\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) = b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - P_h(\cdot \mid s,a) \cdot V_{h+1}^\star$$

$$\geq b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - P_h(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n$$

NOTE this is different from what we did in tabular MDP!!!

$$\geq b_h^n(s,a) - \left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \widehat{V}_{h+1}^n \right|$$

3. Prove Optimism

Lemma [Optimism]: with high probability, for all n, h, s:

$$\widehat{V}_h^n(s) \geq V_h^\star(s)$$

Proof Sketch: let's do induction here with Inductive hypothesis: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \forall s$

$$\widehat{Q}_h^n(s,a) - Q_h^\star(s,a) = b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - P_h(\cdot \mid s,a) \cdot V_{h+1}^\star$$

$$\geq b_h^n(s,a) + \widehat{P}_h^n(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n - P_h(\cdot \mid s,a) \cdot \widehat{V}_{h+1}^n$$

NOTE this is different from what we did in tabular MDP!!!

$$\geq b_h^n(s,a) - \left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \widehat{V}_{h+1}^n \right|$$

$$\geq 0$$

4. Regret Decomposition

Conditioned on history up to the end of episode n-1:

$$V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$$

4. Regret Decomposition

Conditioned on history up to the end of episode n-1:

$$V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$$

(apply Simulation Lemma here)

4. Regret Decomposition

Conditioned on history up to the end of episode n-1:

$$V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$$

(apply Simulation Lemma here)

$$\leq \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi^n}} \left[b_h^n(s_h, a_h) + \left(\widehat{P}_h^n(\cdot \mid s_h, a_h) - P_h(\cdot \mid s_h, a_h) \right) \cdot \widehat{V}_{h+1}^n \right]$$

4. Regret Decomposition

Conditioned on history up to the end of episode n-1:

$$V_0^\star(s_0) - V_0^{\pi^n}(s_0) \le \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$$

(apply Simulation Lemma here)

$$\le \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi^n}} \left[b_h^n(s_h, a_h) + \left(\widehat{P}_h^n(\cdot \mid s_h, a_h) - P_h(\cdot \mid s_h, a_h) \right) \cdot \widehat{V}_{h+1}^n \right]$$

$$\lesssim \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi^n}} \left[b_h^n(s_h, a_h) \right]$$

4. Regret Decomposition

Conditioned on history up to the end of episode n-1:

$$V_0^\star(s_0) - V_0^{\pi^n}(s_0) \le \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$$

(apply Simulation Lemma here)

$$\le \sum_{h=0}^{H-1} \mathbb{E}_{s_h,a_h\sim d_h^{\pi^n}} \left[b_h^n(s_h,a_h) + \left(\widehat{P}_h^n(\cdot \mid s_h,a_h) - P_h(\cdot \mid s_h,a_h) \right) \cdot \widehat{V}_{h+1}^n \right]$$

$$\lesssim \sum_{h=0}^{H-1} \mathbb{E}_{s_h,a_h\sim d_h^{\pi^n}} \left[b_h^n(s_h,a_h) \right]$$

$$= \sum_{h=0}^{H-1} \mathbb{E}_{s_h,a_h\sim d_h^{\pi^n}} \left[\beta \sqrt{\phi(s_h,a_h)^\top (\Lambda_h^n)^{-1} \phi(s_h,a_h)} \right]$$

4. Concluding the Regret Computation

$$\mathbb{E}\left[\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right] = \mathbb{E}\left[\mathbf{1}[\text{good event holds}]\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{E}\left[\mathbf{1}[\text{good event doesn't hold}]\sum_{n=1}^{N}\left(V_0^\star(s_0) - V_0^{\pi^n}(s_0)\right)\right]$$

4. Concluding the Regret Computation

$$\mathbb{E}\left[\sum_{n=1}^{N}\left(V_0^{\star}(s_0) - V_0^{\pi^n}(s_0)\right)\right] = \mathbb{E}\left[\mathbf{1}[\text{good event holds}]\sum_{n=1}^{N}\left(V_0^{\star}(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{E}\left[\mathbf{1}[\text{good event doesn't hold}]\sum_{n=1}^{N}\left(V_0^{\star}(s_0) - V_0^{\pi^n}(s_0)\right)\right]$$

$$\lesssim \beta\mathbb{E}\left[\sum_{n=1}^{N}\sum_{h=0}^{H-1}\sqrt{\phi(s_h^n, a_h^n)^{\top}(\Lambda_h^n)^{-1}\phi(s_h^n, a_h^n)}\right] + \delta NH$$

4. Concluding the Regret Computation

$$\mathbb{E}\left[\sum_{n=1}^{N}\left(V_0^{\star}(s_0) - V_0^{\pi^n}(s_0)\right)\right] = \mathbb{E}\left[\mathbf{1}[\text{good event holds}]\sum_{n=1}^{N}\left(V_0^{\star}(s_0) - V_0^{\pi^n}(s_0)\right)\right] + \mathbb{E}\left[\mathbf{1}[\text{good event doesn't hold}]\sum_{n=1}^{N}\left(V_0^{\star}(s_0) - V_0^{\pi^n}(s_0)\right)\right]$$

$$\lesssim \beta\mathbb{E}\left[\sum_{n=1}^{N}\sum_{h=0}^{H-1}\sqrt{\phi(s_h^n, a_h^n)^{\top}(\Lambda_h^n)^{-1}\phi(s_h^n, a_h^n)}\right] + \delta NH$$

$$\lesssim \beta\mathbb{E}\left[\sum_{h=0}^{H-1}\sqrt{N}\sqrt{\sum_{n=1}^{N}\phi(s_h^n, a_h^n)^{\top}(\Lambda_h^n)^{-1}\phi(s_h^n, a_h^n)}\right] + \delta NH$$

4. Concluding the Regret Computation

$$\mathbb{E}\left[\sum_{n=1}^{N}\left(V_0^\star(s_0)-V_0^{\pi^n}(s_0)\right)\right]=\mathbb{E}\left[\mathbf{1}[\text{good event holds}]\sum_{n=1}^{N}\left(V_0^\star(s_0)-V_0^{\pi^n}(s_0)\right)\right]+\mathbb{E}\left[\mathbf{1}[\text{good event doesn't hold}]\sum_{n=1}^{N}\left(V_0^\star(s_0)-V_0^{\pi^n}(s_0)\right)\right]$$

$$\lesssim \beta\mathbb{E}\left[\sum_{n=1}^{N}\sum_{h=0}^{H-1}\sqrt{\phi(s_h^n,a_h^n)^\top(\Lambda_h^n)^{-1}\phi(s_h^n,a_h^n)}\right]+\delta NH$$

$$\lesssim \beta\mathbb{E}\left[\sum_{h=0}^{H-1}\sqrt{N}\sqrt{\sum_{n=1}^{N}\phi(s_h^n,a_h^n)^\top(\Lambda_h^n)^{-1}\phi(s_h^n,a_h^n)}\right]+\delta NH$$

$$\lesssim \widetilde{O}(H^2d^{1.5}\sqrt{N})$$

Summary

W/ optimism (thanks to bonus which offsets errors in estimated models):

$$V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \lesssim \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi^n}} \left[b_h^n(s_h, a_h) \right]$$

Summary

W/ optimism (thanks to bonus which offsets errors in estimated models):

$$V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \lesssim \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi^n}} \left[b_h^n(s_h, a_h) \right]$$

Thus, when our policy is suboptimal, i.e., $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \geq \epsilon$

we have: $\sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi^n}} \left[b_h^n(s_h, a_h) \right] \geq \epsilon$

Summary

W/ optimism (thanks to bonus which offsets errors in estimated models):

$$V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \lesssim \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi^n}} \left[b_h^n(s_h, a_h) \right]$$

Thus, when our policy is suboptimal, i.e., $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \geq \epsilon$

we have: $\sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi^n}} \left[b_h^n(s_h, a_h) \right] \geq \epsilon$

i.e., π^n will visit places where the current bonus is non-trivial, namely explore!

Summary

W/ optimism (thanks to bonus which offsets errors in estimated models):

$$V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0) \lesssim \sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi^n}} \left[b_h^n(s_h, a_h) \right]$$

Thus, when our policy is suboptimal, i.e., $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \geq \epsilon$

we have: $\sum_{h=0}^{H-1} \mathbb{E}_{s_h, a_h \sim d_h^{\pi^n}} \left[b_h^n(s_h, a_h) \right] \geq \epsilon$

i.e., π^n will visit places where the current bonus is non-trivial, namely explore!

Since we live in a d-dim space, eventually we will explore all possible directions.