Exploration in Linear MDPs (Part II)

Recap: Linear MDP Definition

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

S & A could be large or even continuous, hence poly(S,A) is not acceptable

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

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

$$P_h(\cdot|s,a) = \mu^* \cdot \phi(s,a) = \begin{bmatrix} | & | & & | \\ \mu_1 & \mu_2 & \cdots & \mu_d \\ | & | & & | \end{bmatrix} \phi(s,a)$$

Feature map ϕ is known to the learner!

(We assume reward is known, i.e., $\theta^\star$ is known)

Recap: Model Learning in Linear MDPs

$$\mathcal{D}_h^n = \left\{ s_h^i, a_h^i, s_{h+1}^i \right\}_{i=1}^{n-1}$$

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

Recap: Model Learning in Linear MDPs

$$\mathscr{D}_h^n = \left\{ s_h^i, a_h^i, s_{h+1}^i \right\}_{i=1}^{n-1} \qquad \Lambda_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$$

Ridge Linear Regression:

one-hot vector

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

Recap: Model Learning in Linear MDPs

$$\mathscr{D}_h^n = \left\{s_h^i, a_h^i, s_{h+1}^i\right\}_{i=1}^{n-1} \qquad \Lambda_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$$

Ridge Linear Regression:

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

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

Recap: Model Learning in Linear MDPs

$$\mathscr{D}_h^n = \left\{ s_h^i, a_h^i, s_{h+1}^i \right\}_{i=1}^{n-1} \qquad \Lambda_h^n = \sum_{i=1}^{n-1} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top + \lambda I$$

Ridge Linear Regression:

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

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

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

1

$$E\left[\epsilon_h^i \mid s_h^i, a_h^i\right] = 0 \qquad \|\epsilon_h^i\|_1 \le 2$$

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:

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

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:

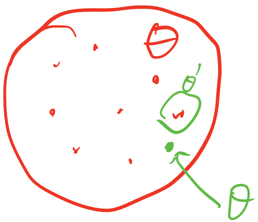
$$\left| \left(\widehat{P}_h^n(\cdot \mid s, a) - P_h(\cdot \mid s, a) \right) \cdot V \right| \lesssim \| \phi(s,a) \|_{(\Lambda_h^n)^{-1}} \cdot H \left(\sqrt{\ln \frac{H}{\delta}} + \sqrt{d \ln \left(1 + \frac{N}{\lambda} \right)} \right)$$

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

$|F| = \infty$

Covering

Detour: Covering Number



$\| \theta' - \theta \|_2 \le \epsilon$

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

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. $\|\theta' - \theta\|_2 \le \epsilon$.

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 $\mathscr{F} = \{f_\theta : \theta \in \Theta\}$,

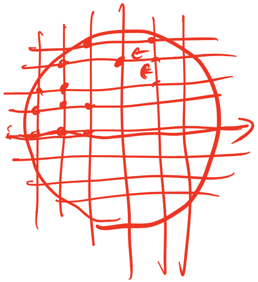
and for any $f_{\theta_1}, f_{\theta_2} \in \mathscr{F}$, $\|f_{\theta_1} - f_{\theta_2}\|_\infty \leq \underline{L\|\theta_1 - \theta_2\|_2}$

$f_{\theta_1}(x) = x^\top \cdot \theta_1$

$f_{\theta_2}(x) = x^\top \cdot \theta_2$

$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$

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 $\mathcal{N}_\epsilon$

$\left(\frac{1}{\epsilon}\right)^2$

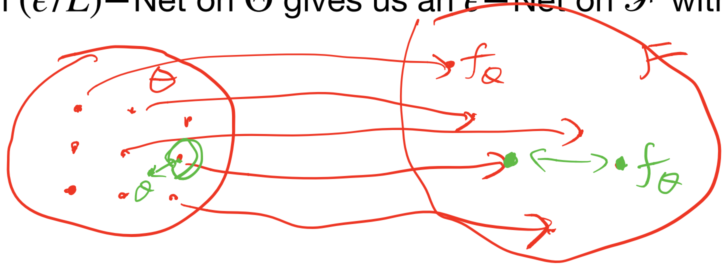
Lemma [Covering of Θ] We have $|\mathcal{N}_\epsilon| \leq (1 + 2R/\epsilon)^d$, and $\ln(|\mathcal{N}_\epsilon|) \leq d \ln(1 + 2R/\epsilon)$

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 $d(f_{\theta_1}, f_{\theta_2}) := \|f_{\theta_1} - f_{\theta_2}\|_\infty$

$\frac{\epsilon}{L}$



Lipchitz

Detour: Covering Number and An Example

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

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

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

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

$\theta = (w, \beta, \Lambda)$

$f_{w,\beta,\Lambda} : S \mapsto \mathbb{R}^+$

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

Detour: Covering Number and Example

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

Detour: Covering Number and Example

$\theta \in R^d$

$|N_\epsilon| \approx \left(\frac{1}{\epsilon}\right)^d$

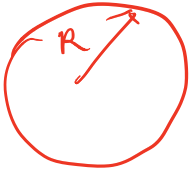
$$\underbrace{f_{w,\beta,\Lambda}}_{d \;\; 1 \;\; d^2} :, f_{w,\beta,\Lambda}(s) := \min \left\{ \max_a \left(w^\top \phi(s,a) + \beta\sqrt{\phi(s,a)^\top \Lambda^{-1} \phi(s,a)} \right), H \right\}$$

$\Lambda \in R^{d \times d}$

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

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

Detour: Covering Number and Example



$|N_\epsilon| = \left(1 + \frac{R}{\epsilon}\right)^d$

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

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

Key step in the proof:

$\forall s \in S$

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

(w, β, Λ) $(\hat w, \hat\beta, \hat\Lambda)$

$\frac{\epsilon}{3}$-Net $\sqrt{\lambda}\epsilon$ $\frac{\epsilon^2}{B^2}$-Net $\underbrace{\{A \in \mathbb{R}^{d\times d} : \sigma_{\max}(A)}_{} \leq \frac{1}{\lambda}\}$

Detour: Uniform Convergence using Covering

Define $f_{w,\beta,\Lambda} : S \to [0,H]$, $f_{w,\beta,\Lambda}(s) := \min \left\{ \max_a \left(w^\top \phi(s,a) + \beta \sqrt{\phi(s,a)^\top \Lambda^{-1} \phi(s,a)} \right), H \right\}$

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

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

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

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

$|F| = \infty$

$\| \widehat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a) \|_1 \approx \sqrt{\frac{S \ln(1/\delta)}{\cdots}}$

Detour: Uniform Convergence using Covering

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

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

Detour: Uniform Convergence using Covering

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

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

Proof Sketch: let's start with $\underbrace{\mathcal{N}_\epsilon \text{ over } \mathcal{F}}$

$\ln |\mathcal{N}_\epsilon| \approx d^2$

Detour: Uniform Convergence using Covering

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

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

Proof Sketch: let's start with $\mathcal{N}_\epsilon$ over $\mathcal{F}$

1. For a fixed $\hat{f} \in \mathcal{N}_\epsilon$: $\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \hat{f} \right| \lesssim \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \cdot H \left(\sqrt{\ln \frac{H}{\delta}} + \sqrt{d \ln\left(1 + \frac{N}{\lambda}\right)} \right)$

Detour: Uniform Convergence using Covering

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

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

$\ln |N_\epsilon| \leq d^2$

Proof Sketch: let's start with $\mathcal{N}_\epsilon$ over $\mathcal{F}$

1. For a fixed $\hat{f} \in \mathcal{N}_\epsilon$: $\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \hat{f} \right| \lesssim \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \cdot H \left(\sqrt{\ln \frac{H}{\delta}} + \sqrt{d \ln\left(1 + \frac{N}{\lambda}\right)} \right)$

2. union bound: $\forall \hat{f} \in \mathcal{N}_\epsilon$: $\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V \right| \lesssim \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \cdot H \left(\sqrt{\ln \frac{H|\mathcal{N}_\epsilon|}{\delta}} + \sqrt{d \ln\left(1 + \frac{N}{\lambda}\right)} \right)$

failure $\frac{\delta}{|N_\epsilon|}$

$\sqrt{\ln \frac{|N_\epsilon|}{\delta}} = \sqrt{d^2 + \ln\frac{1}{\delta}}$

$\approx O(d)$

Detour: Uniform Convergence using Covering

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

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

Proof Sketch: let's start with $\mathscr{N}_\epsilon$ over $\mathscr{F}$

1. For a fixed $\hat{f} \in \mathscr{N}_\epsilon$: $\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \hat{f} \right| \lesssim \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \cdot H \left(\sqrt{\ln \frac{H}{\delta}} + \sqrt{d \ln\left(1 + \frac{N}{\lambda}\right)} \right)$

2. union bound: $\forall \hat{f} \in \mathscr{N}_\epsilon$: $\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V \right| \lesssim \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \cdot H \left(\sqrt{\ln \frac{H|\mathscr{N}_\epsilon|}{\delta}} + \sqrt{d \ln\left(1 + \frac{N}{\lambda}\right)} \right)$

$\mathcal{F}$

$f \overset{\epsilon}{\longleftrightarrow} \hat{f}$

3. Consider any $f \in \mathscr{F}$:

$$\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \underset{1}{f} \right| \le \underbrace{\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \underset{1}{\hat{f}} \right|}_{\text{Step 2}} + \underbrace{\left| \left(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \underline{(f - \hat{f})} \right|}_{\le \epsilon}$$

$\hat{f} \in \mathcal{N}_\epsilon$

Detour: Uniform Convergence using Covering

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

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

Proof Sketch: let's start with $\mathscr{N}_\epsilon$ over $\mathscr{F}$

1. For a fixed $\hat{f} \in \mathscr{N}_\epsilon$: $\left| \left(\widehat{P}^n_h(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \hat{f} \right| \lesssim \|\phi(s,a)\|_{(\Lambda^n_h)^{-1}} \cdot H \left(\sqrt{\ln \frac{H}{\delta}} + \sqrt{d \ln\left(1 + \frac{N}{\lambda}\right)} \right)$

2. union bound: $\forall \hat{f} \in \mathscr{N}_\epsilon$: $\left| \left(\widehat{P}^n_h(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot V \right| \lesssim \|\phi(s,a)\|_{(\Lambda^n_h)^{-1}} \cdot H \left(\sqrt{\ln \frac{H|\mathscr{N}_\epsilon|}{\delta}} + \sqrt{d \ln\left(1 + \frac{N}{\lambda}\right)} \right)$

3. Consider any $f \in \mathscr{F}$:

$$\left| \left(\widehat{P}^n_h(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot f \right| \leq \underbrace{\left| \left(\widehat{P}^n_h(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \hat{f} \right|}_{Hd\sqrt{\ln(1/\epsilon)}} + \left| \left(\widehat{P}^n_h(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot (f - \hat{f}) \right|$$

4. Tune parameter ϵ

Summary of Covering Argument

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

$$\ln(|N_\epsilon|) \approx \tilde{O}(d^2)$$

$$f_{w,\beta,A}$$

$w \uparrow d$, $\beta \uparrow 1$, $A \uparrow d \times d$

Summary of Covering Argument

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

Let's get back to Linear MDPs again!

$\hat{P}_h^n(\cdot|s,a) = \hat{\mu}_h^n \phi(s,a)$ $\qquad$ $\hat{P}_h^n(\cdot|s,a) \notin \Delta(S)$

$\hat{P}_h^n(s'|s,a) = \dfrac{N(s,a,s')}{N(s,a)}$

Algorithm: UCBVI in Linear MDPs

Algorithm: UCBVI in Linear MDPs

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

Algorithm: UCBVI in Linear MDPs

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

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

A

Algorithm: UCBVI in Linear MDPs

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

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

Covering

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

Algorithm: UCBVI in Linear MDPs

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

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

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

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

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

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

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

h=0, ... H-1,

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

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

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

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

$\mathbb{E}_{s' \sim \widehat{P}(\cdot|s,a)} \widehat{V}^n_{h+1}(s')$

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

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

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

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

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

$\hat{\mu}_h^n \phi(s,a)$

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

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

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

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

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

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

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

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

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

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

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

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

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

$\in \mathbb{R}^d$

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

$:=> \widehat{w}^{\,n}_h$

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

1 2

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

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

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

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

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

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

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

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

$\mathcal{F}_{w,\beta,\Lambda}$

$= \min\{\max_a w^\top\phi + \beta\sqrt{\cdots}, H\}$

$\widehat{V}^n_h \in \mathcal{F}$

Δ

$\left(\widehat{P}^n_h(\cdot|s,a) - P(\cdot|s,a)\right)\cdot \widehat{V}^n_h$

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: $\widehat{V}_{h+1}^n(s) \geq V_{h+1}^\star(s), \forall s$

$\cancel{r+}\, b + \hat{P} \cdot \hat{V}_{h+1}$

$(\cancel{r+}\, P(\cdot|s,a) \cdot V^\star_{h+1})$

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

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:

$$\widehat{V}^n_h(s) \geq V^\star_h(s)$$

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

→ valid distribution

$$\widehat{Q}^n_h(s,a) - Q^\star_h(s,a) = b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1} - P_h(\cdot \mid s,a) \cdot V^\star_{h+1}$$

$$\cancel{\geq b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1} - P_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1}}$$

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

$$\geq b^n_h(s,a) + \underbrace{\widehat{P}^n_h(\cdot|s,a)}_{\uparrow \text{ Valid Distribution}} \cdot V^\star_{h+1} - P_h(\cdot|s,a) V^\star_{h+1} \quad \checkmark$$

$$\| \widehat{P}(\cdot|s,a) - P(\cdot|s,a) \|_1 \leq \sqrt{\frac{S \ln(1/\delta)}{N(s,a)}}$$

3. Prove Optimism

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

$$\widehat{V}^n_h(s) \geq V^\star_h(s)$$

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

$$\widehat{Q}^n_h(s,a) - Q^\star_h(s,a) = b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1} - P_h(\cdot \mid s,a) \cdot V^\star_{h+1}$$

$$\geq b^n_h(s,a) + \widehat{P}^n_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1} - P_h(\cdot \mid s,a) \cdot \widehat{V}^n_{h+1}$$

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

$$\geq b^n_h(s,a) - \left| \underbrace{\left(\widehat{P}^n_h(\cdot \mid s,a) - P_h(\cdot \mid s,a) \right) \cdot \widehat{V}^n_{h+1}}_{\approx (Hd)\|\phi(s,a)\|_{(\Lambda^n_h)^{-1}}} \right|$$

$$\underbrace{\hat{V}^n_{h+1}(s) = \min\left\{\max_a w^\top\phi + \beta\sqrt{\phi\Lambda^{-1}\phi},\ H\right\}}$$

$$(Hd)\cdot\|\phi(s,a)\|_{(\Lambda^n_h)^{-1}}$$

$= 0$

covering argument

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:

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

Regret at ep-n

π^n under $(\widehat{P}^n, r + b^n)$

π^n under (P, r)

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:

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

$d_h^\pi(s,a)$

(apply Simulation Lemma here)

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

Tabular: Holder $\|\hat{P}-P\|_1 \cdot \|\hat{V}\|_\infty$

$\le H\cdot\sqrt{\frac{S\ln(1/\delta)}{N(s,a)}}$

$$= \text{①}\ \underbrace{(\hat{P}-P)\cdot V^\star}_{\le\sqrt{\frac{\ln(1/\delta)}{N(s,a)}}} + \underbrace{(\hat{P}-P)(\hat{V}-V^\star)}_{\substack{\hat{P}\to P\\ \hat{V}\to V^\star}}$$

$\to \frac{1}{N(s,a)}$

$\sum \frac{1}{N(s)} \to \ln(N)$

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)

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

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

4. Regret Decomposition

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

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

(apply Simulation Lemma here)

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

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

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

$$\left| \left(\hat{P}^n_h(\cdot|s,a) - P_h(\cdot|s,a) \right) \cdot f \right| \leq H d \|\phi(s,a)\|_{(\Lambda^n_h)^{-1}} \quad \forall s,a,n,h,\ f \in \mathcal{F}$$

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

(ltd)

4. Concluding the Regret Computation

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

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

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

CH-inequality

4. Concluding the Regret Computation

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

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

$\lesssim$

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

Set $\delta = \frac{1}{NH}$

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

Tabular:

$\phi(s,a) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{SA}$ $d = SA$ $H^2 (SA)^{1.5}\sqrt{N}$

Linear Bandits $H = 1$

$d^{1.5}\sqrt{N}$

Linear Bandit $\to d\sqrt{N}$

$\hat{P}(\cdot|s,a) \Leftarrow$ linear Regression

$\left| \left(\hat{P}(\cdot|s,a) - P(\cdot|s,a) \right)^\top f \right|$ ← pre-fixed f ($*$)

$\mathcal{F}$, $|\mathcal{F}| = \infty$

① Covering ϵ-Net for $\mathcal{F}$

② ($*$) + union Bound on ϵ-Net

③ Def ϵ-Net + Triangle Inequality

---

① optimism

② simulation lemma

③ $\sum_{n=1}^{N} \phi(s_h^n, a_h^n)^\top (\Sigma_h^n)^{-1} \phi(s_h^n, a_h^n)$

$\leq d \cdot \ln(N)$

$$|(\hat{\theta} - \theta^*) \cdot x| \leq \underbrace{\|x\|_{\Lambda^{-1}}} \cdot \sqrt{d}$$

$$\Lambda = \sum_{i=1}^{n} x_i x_i^T$$

$$\Lambda = U \Sigma U^T$$

$$\phi(s, a)$$