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

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

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

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

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

Recap: Model Learning in Linear MDPs

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

Ridge Linear Regression:

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

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:

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

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

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

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 $\mathscr{F}$?

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

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

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

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

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

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

Detour: Covering Number and Example

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

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

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

Key step in the proof:

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

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

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 using Covering

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

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

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 $\mathcal{F}$

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

Detour: Uniform Convergence using Covering

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

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

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

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

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

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

Detour: Uniform Convergence using Covering

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

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

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

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

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

4. Tune parameter ϵ

Summary of Covering Argument

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

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)

Let's get back to Linear MDPs again!

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 $b_h^n(s,a) = \beta\sqrt{\phi(s,a)(\Lambda_h^n)^{-1}\phi(s,a)}$

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 $\left|(\widehat{P}_h^n(\cdot\mid s,a) - P_h(\cdot\mid s,a))\cdot f\right|, \forall f \in \mathscr{F}$

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 $\left|(\widehat{P}_h^n(\cdot \mid s,a) - P_h(\cdot \mid s,a)) \cdot f\right|, \forall f \in \mathscr{F}$

3. Plan: π^{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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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:

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

(apply Simulation Lemma here)

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

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

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

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

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

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

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