Exploration in Linear MDPs

Wen Sun

CS 6789: Foundations of Reinforcement Learning

Recap:

Recap:

Stochastic Linear Bandits

$\mathscr{D} \subset \mathbb{R}^d \qquad r(x) = \theta^\star \cdot x, \forall x$

Recap:

Stochastic Linear Bandits

$$\mathscr{D} \subset \mathbb{R}^d \qquad r(x) = \theta^\star \cdot x, \forall x$$

Every episode n, learner picks $x_n \in \mathscr{D}$

Recap:

Stochastic Linear Bandits

$$\mathscr{D} \subset \mathbb{R}^d \qquad r(x) = \theta^\star \cdot x, \forall x$$

Every episode n, learner picks $x_n \in \mathscr{D}$

Learner receives a scalar $r_n = \theta^\star \cdot x_n + \epsilon_n$, $\mathbb{E}[\epsilon_n] = 0, |\epsilon_n| < \alpha$

Recap:

Stochastic Linear Bandits

$$\mathscr{D} \subset \mathbb{R}^d \qquad r(x) = \theta^\star \cdot x, \forall x$$

Every episode n, learner picks $x_n \in \mathscr{D}$

Learner receives a scalar $r_n = \theta^\star \cdot x_n + \epsilon_n$, $\mathbb{E}[\epsilon_n] = 0, |\epsilon_n| < \alpha$

$$\text{Regret} = \mathbb{E}\left[\sum_{n=1}^{N} \theta^\star \cdot x^\star - \sum_{n=1}^{N} \theta^\star x_n\right]$$

Important Lemma:

Lemma [Self Normalized Bound for Vector-Valued Martingales] Suppose $\{\epsilon_n\}_{n=1}^{\infty}$ are mean zero random variables with $|\epsilon_n| \leq \alpha$, for all n; Let $\{x_i \in \mathbb{R}^d\}_{n=1}^{\infty}$ be some stochastic random process; Define $\Lambda^n = \lambda I + \sum_{i=1}^{n} x_i x_i^\top$, then with probability at least $1-\delta$, for all $n \geq 1$:

$$\left\| \sum_{i=1}^{n} x_i \epsilon_i \right\|_{(\Lambda^n)^{-1}}^2 \leq 2\sigma^2 \ln\left(\frac{\det(\Lambda^n)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)$$

Important Lemma:

Lemma [Self Normalized Bound for Vector-Valued Martingales] Suppose $\{\epsilon_n\}_{n=1}^{\infty}$ are mean zero random variables with $|\epsilon_n| \leq \alpha$, for all n; Let $\{x_i \in \mathbb{R}^d\}_{n=1}^{\infty}$ be some stochastic random process; Define $\Lambda^n = \lambda I + \sum_{i=1}^{n} x_i x_i^\top$, then with probability at least $1-\delta$, for all $n \geq 1$:

$$\left\| \sum_{i=1}^{n} x_i \epsilon_i \right\|_{(\Lambda^n)^{-1}}^2 \leq 2\sigma^2 \ln\left(\frac{\det(\Lambda^n)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)$$

$$2\sigma^2 \ln\left(det(\Lambda^n)^{1/2} det(\lambda I)^{-1/2} / \delta \right) \leq \sigma^2 \left(d \ln(1 + n/\lambda) + 2\ln(1/\delta) \right)$$

Today's question

We extended MAB to linear bandit so that we can deal w/ infinitely many actions…

Can we extend discrete MDPs to some kind linear MDPs?

Outline for this lecture:

1. Introduction of low-rank MDP

2. Planning in low-rank MDP (i.e., DP) and UCBVI algorithm

3.Non-parametric model learning in linear MDPs

Notations and Useful Inequalities

For real-value matrix A:

$$\|A\|_F^2 = \sum_{i,j} A_{i,j}^2 \qquad \|A\|_2 = \sup_{x:\|x\|_2 \le 1} \|Ax\|_2$$

Notations and Useful Inequalities

For real-value matrix A:

$$\|A\|_F^2 = \sum_{i,j} A_{i,j}^2 \qquad \|A\|_2 = \sup_{x:\|x\|_2 \le 1} \|Ax\|_2$$

$$\|Ax\|_2 \le \|A\|_2 \|x\|_2$$

Notations and Useful Inequalities

For real-value matrix A:

$$\|A\|_F^2 = \sum_{i,j} A_{i,j}^2 \qquad \|A\|_2 = \sup_{x:\|x\|_2 \le 1} \|Ax\|_2$$

$$\|Ax\|_2 \le \|A\|_2 \|x\|_2$$

For Positive Definite matrix Λ:

Notations and Useful Inequalities

For real-value matrix A:

$$\|A\|_F^2 = \sum_{i,j} A_{i,j}^2 \qquad \|A\|_2 = \sup_{x:\|x\|_2 \leq 1} \|Ax\|_2$$

$$\|Ax\|_2 \leq \|A\|_2 \|x\|_2$$

For Positive Definite matrix Λ:

$\sigma_{\min}(\Lambda), \sigma_{\max}(\Lambda)$

Notations and Useful Inequalities

For real-value matrix A:

$$\|A\|_F^2 = \sum_{i,j} A_{i,j}^2 \qquad \|A\|_2 = \sup_{x:\|x\|_2 \le 1} \|Ax\|_2$$

$$\|Ax\|_2 \le \|A\|_2 \|x\|_2$$

For Positive Definite matrix Λ:

$$\sigma_{\min}(\Lambda), \sigma_{\max}(\Lambda) \qquad \det(\Lambda) = \prod_{i=1}^{d} \sigma_i$$

Notations and Useful Inequalities

For real-value matrix A:

$$\|A\|_F^2 = \sum_{i,j} A_{i,j}^2 \qquad \|A\|_2 = \sup_{x:\|x\|_2 \le 1} \|Ax\|_2$$

$$\|Ax\|_2 \le \|A\|_2 \|x\|_2$$

For Positive Definite matrix Λ:

$$\sigma_{\min}(\Lambda), \sigma_{\max}(\Lambda) \qquad \det(\Lambda) = \prod_{i=1}^{d} \sigma_i$$

$$\|x\|_\Lambda^2 = x^\top \Lambda x$$

Notations and Useful Inequalities

For real-value matrix A:

$$\|A\|_F^2 = \sum_{i,j} A_{i,j}^2 \qquad \|A\|_2 = \sup_{x:\|x\|_2 \le 1} \|Ax\|_2$$

$$\|Ax\|_2 \le \|A\|_2 \|x\|_2$$

For Positive Definite matrix Λ:

$$\sigma_{\min}(\Lambda), \sigma_{\max}(\Lambda) \quad \det(\Lambda) = \prod_{i=1}^{d} \sigma_i$$

$$\|x\|_\Lambda^2 = x^\top \Lambda x$$

$$\mathbb{E}_{s' \sim P_h(\cdot|s,a)} f(s') = P_h(\cdot \mid s, a) \cdot f$$

Low-Rank MDP Definition

Low-Rank MDP Definition

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

Low-Rank MDP Definition

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

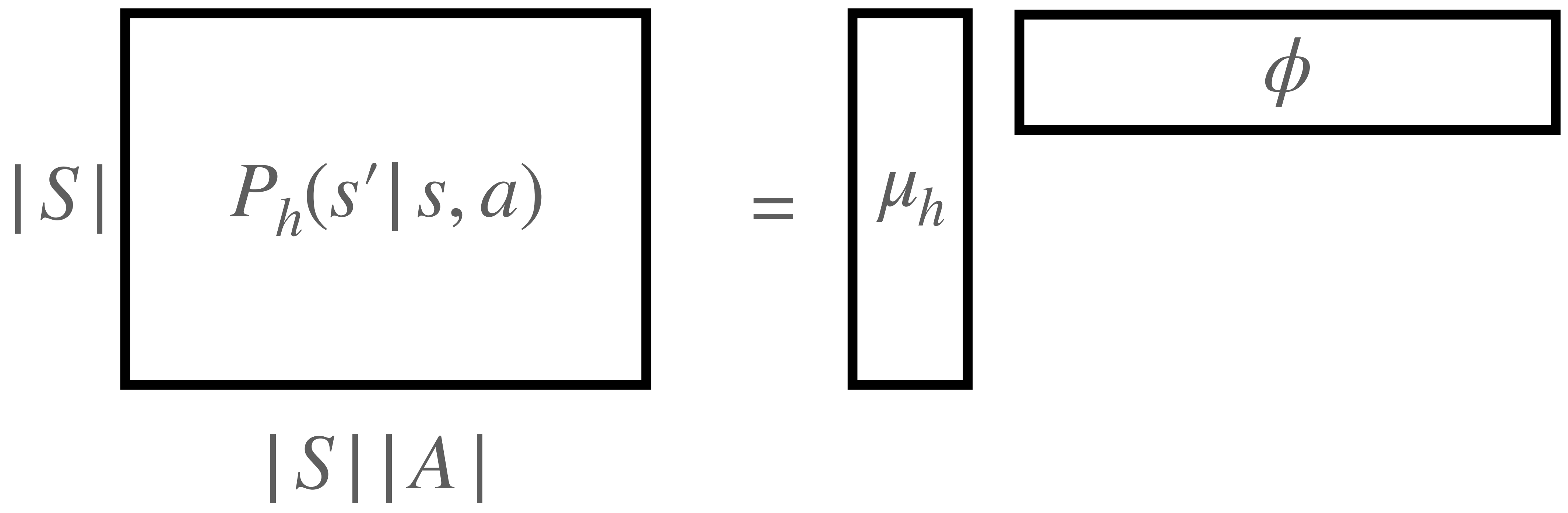
S & A could be large or even continuous, hence poly(S,A) is not acceptable

Low-Rank 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

Low-Rank Decomposition:



Low-Rank MDP Definition

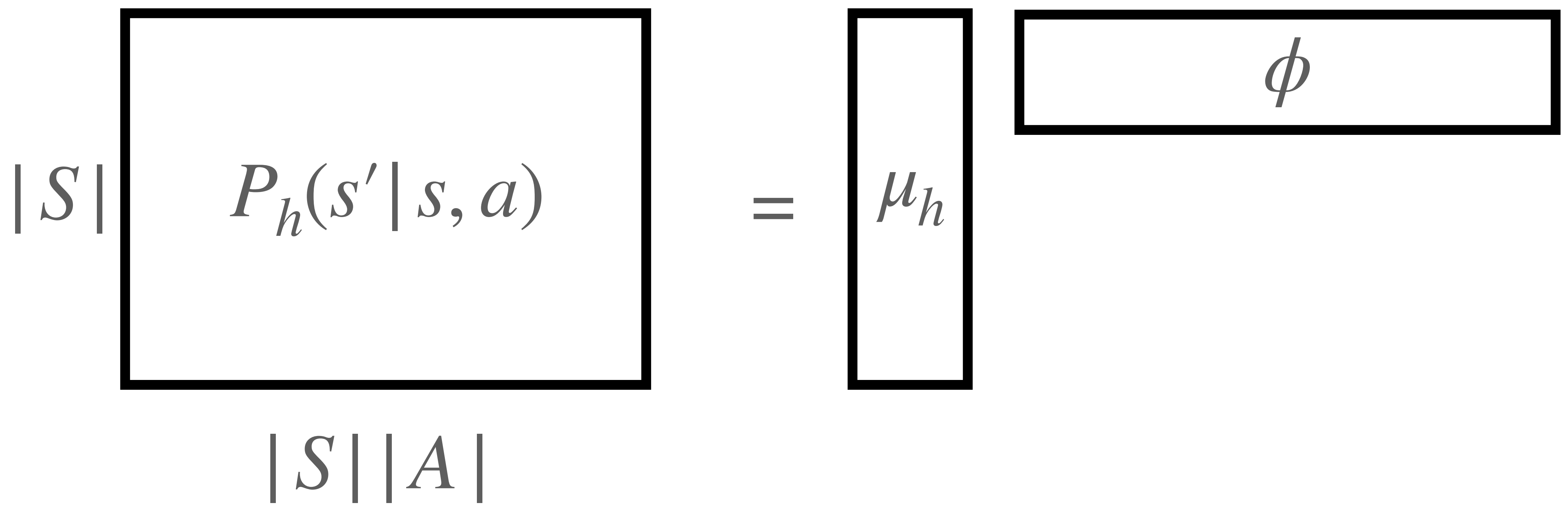
Finite horizon time-dependent episodic MDP $\mathscr{M} = \{S, A, H, \{r\}_h, \{P\}_h, s_0\}$

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

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

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

Low-Rank Decomposition:



Low-Rank MDP Definition

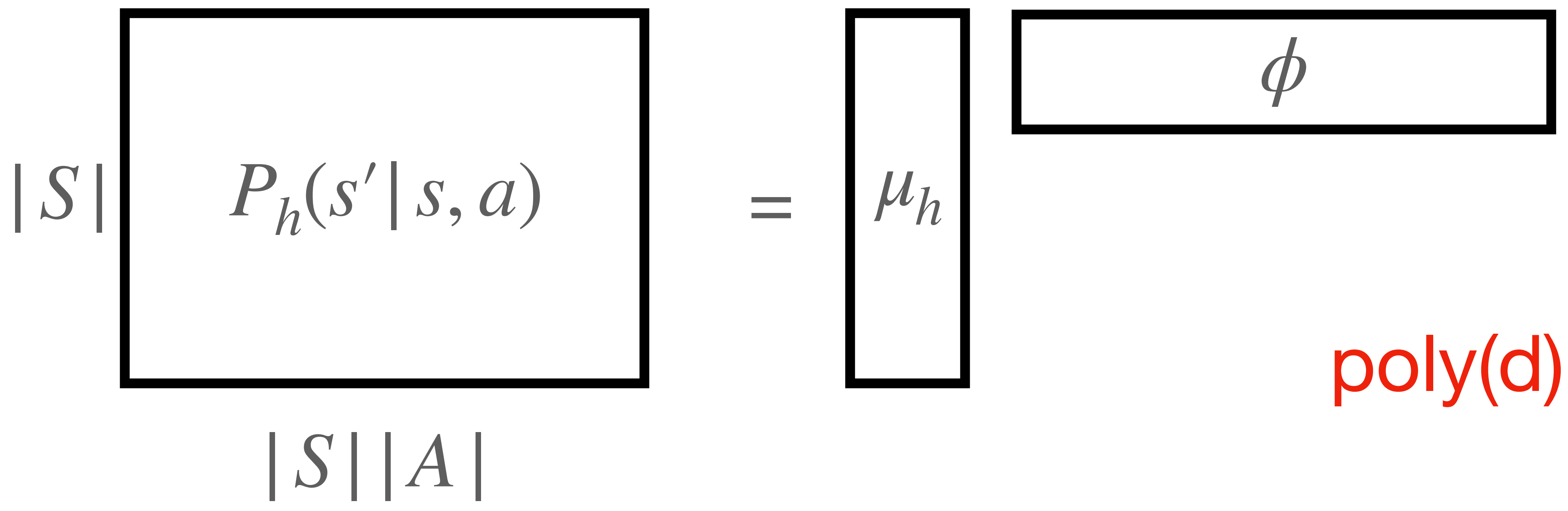
Finite horizon time-dependent episodic MDP $\mathscr{M} = \{S, A, H, \{r\}_h, \{P\}_h, s_0\}$

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

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

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

Low-Rank Decomposition:



poly(d) rather than poly(SA)

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

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

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

Feature map ϕ is known to the learner!

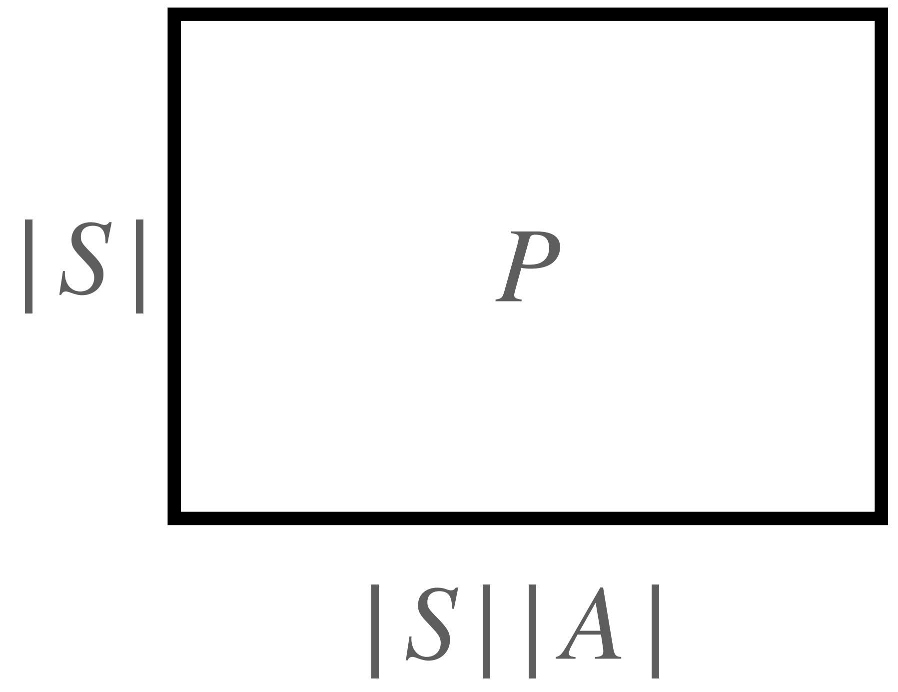
(We assume reward is known, i.e., $\theta^\star$ is known)

Linear MDP Example

It generalizes tabular MDPs: $\phi(s, a)$ one-hot vector

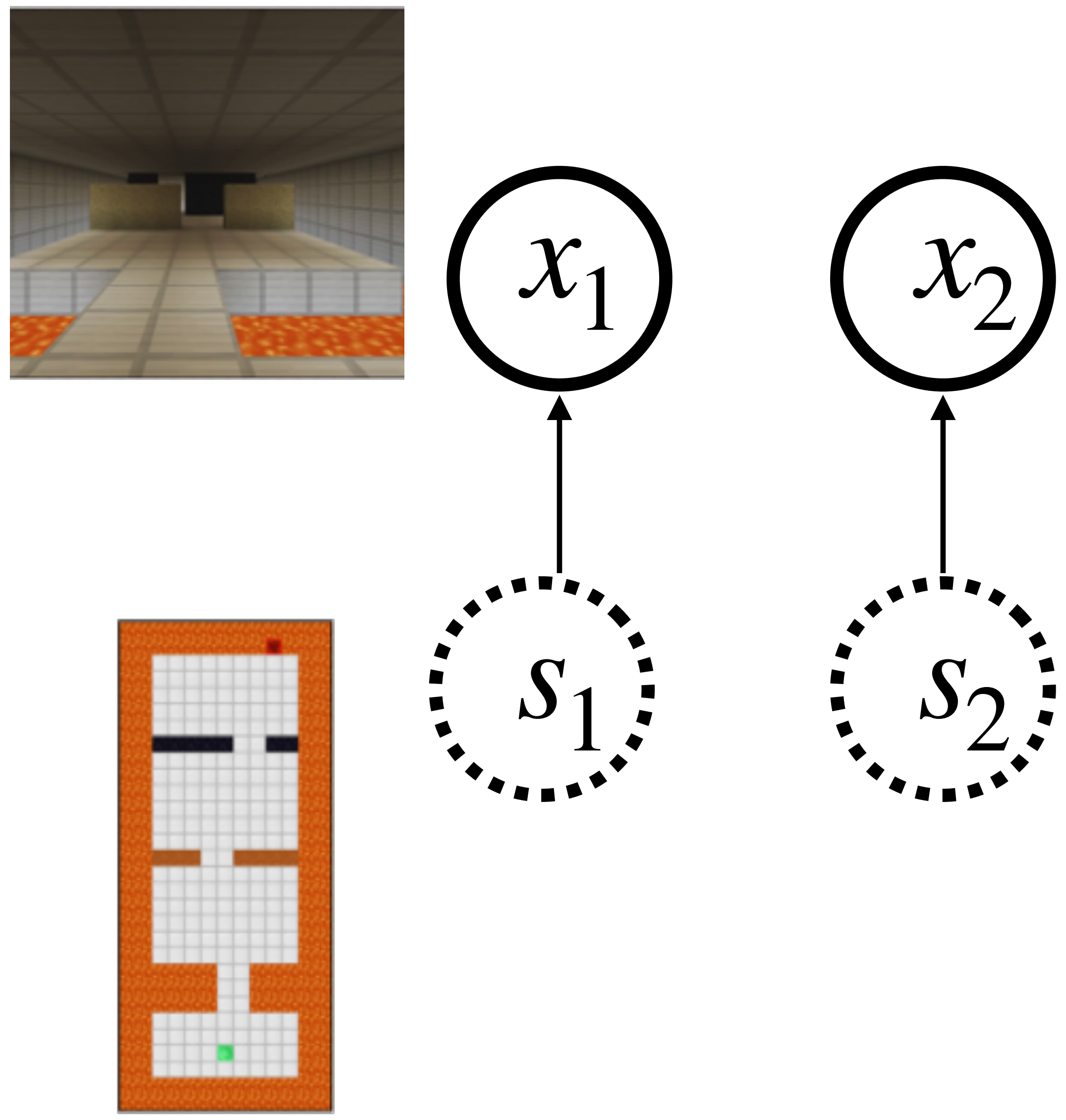
$$P(\cdot \mid s, a) = P\phi(s, a)$$

where $P \in \mathbb{R}^{|S| \times |SA|}$ is the transition matrix



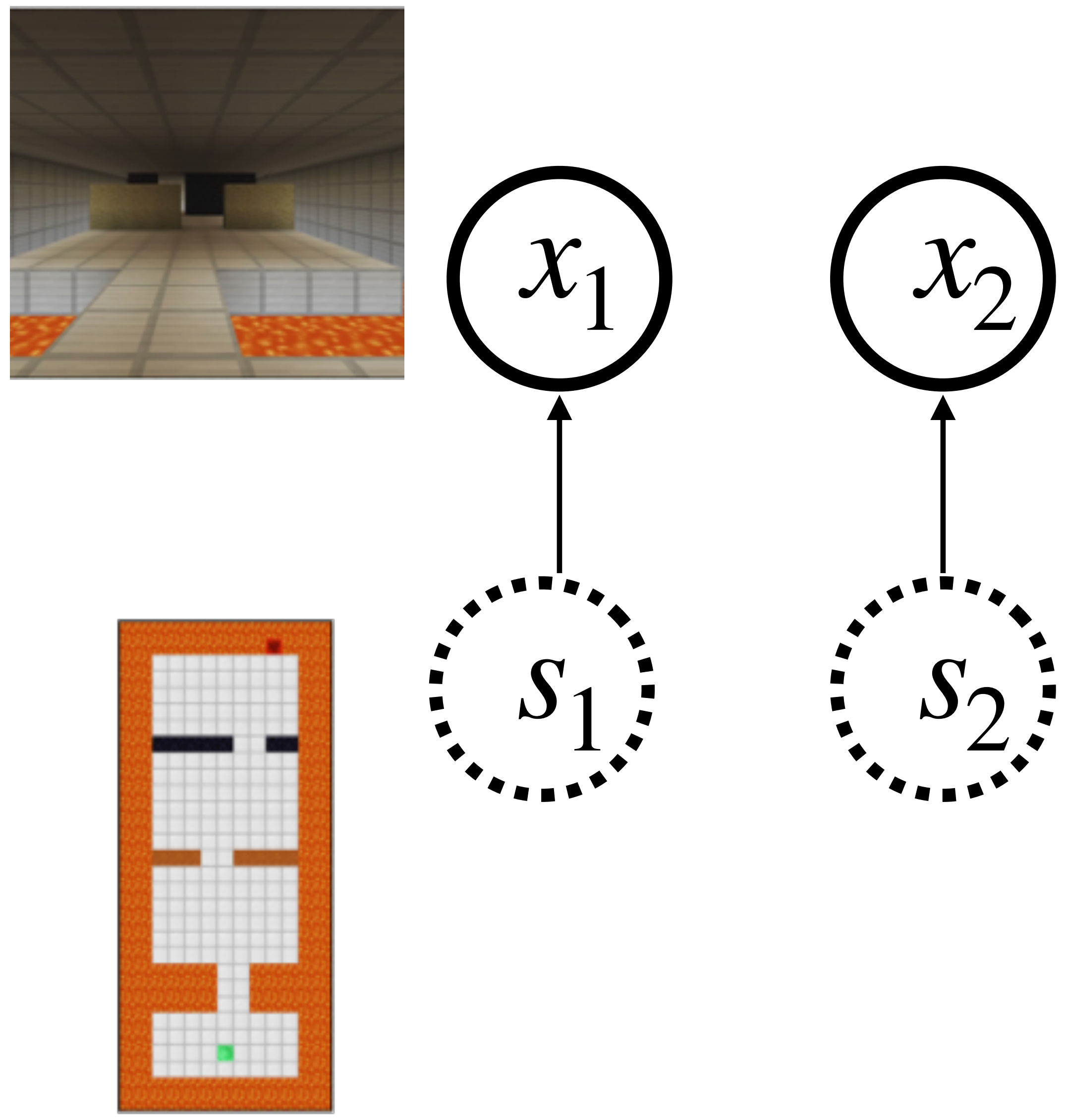
Low-Rank Example

Can encode latent variables: block-MDPs



Low-Rank Example

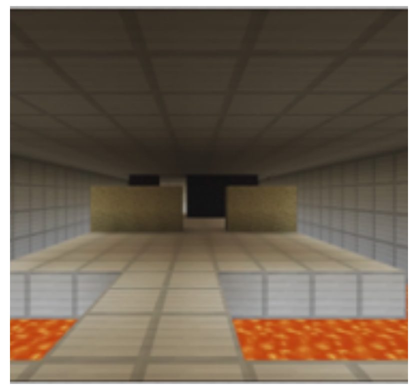
Can encode latent variables: block-MDPs

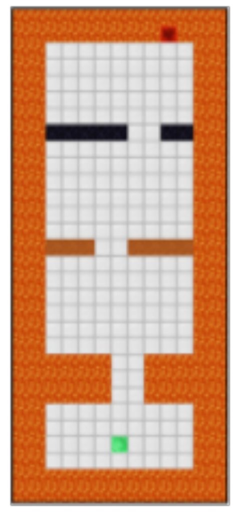


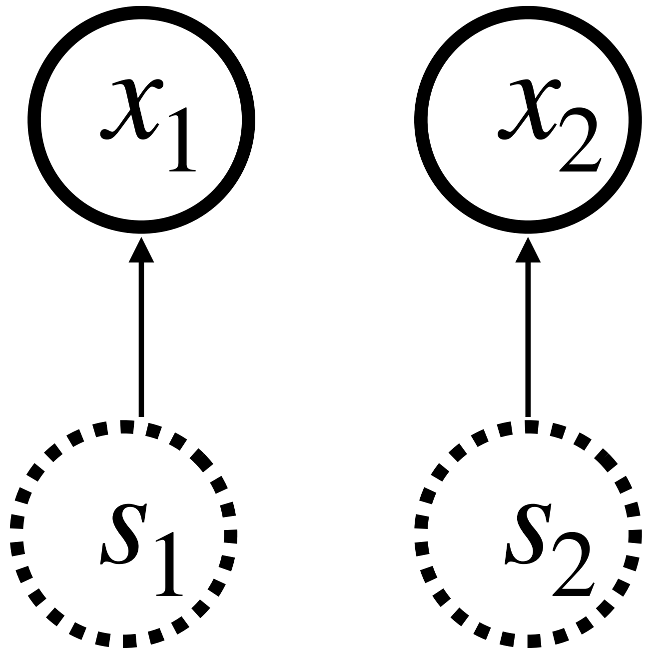
Discrete latent state space S: $|S|$ is small, transition $T : S \times A \mapsto S$

Low-Rank Example

Can encode latent variables: block-MDPs





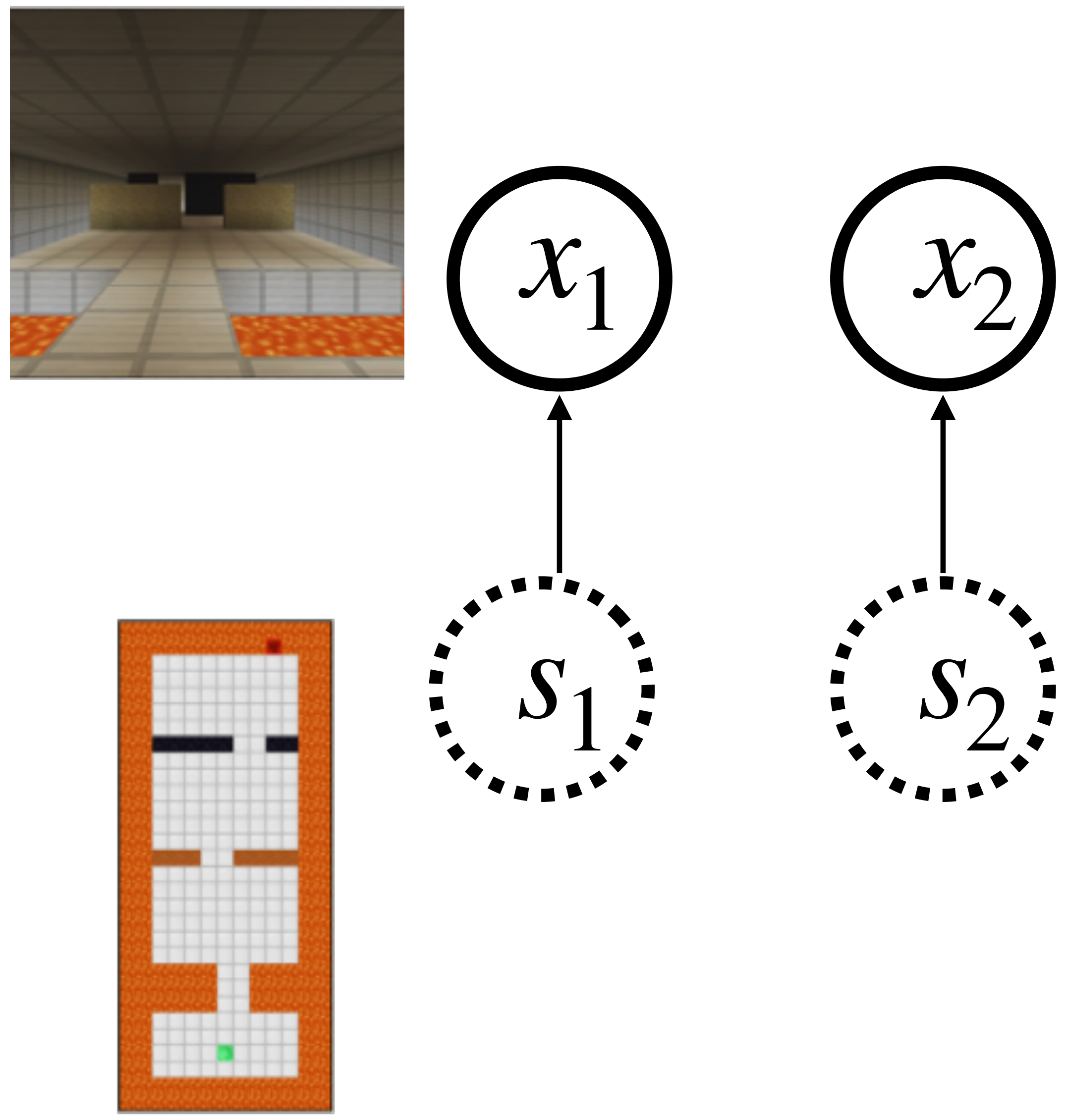


Discrete latent state space S: $|S|$ is small, transition $T: S \times A \mapsto S$

Large observation space X (hence any poly dependency on $|X|$ is bad)

Low-Rank Example

Can encode latent variables: block-MDPs



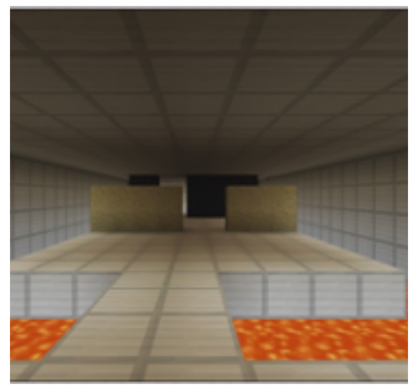
Discrete latent state space S: $|S|$ is small, transition $T: S \times A \mapsto S$

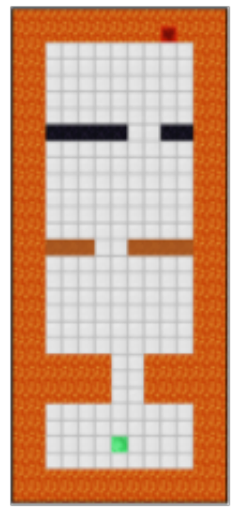
Large observation space X (hence any poly dependency on $|X|$ is bad)

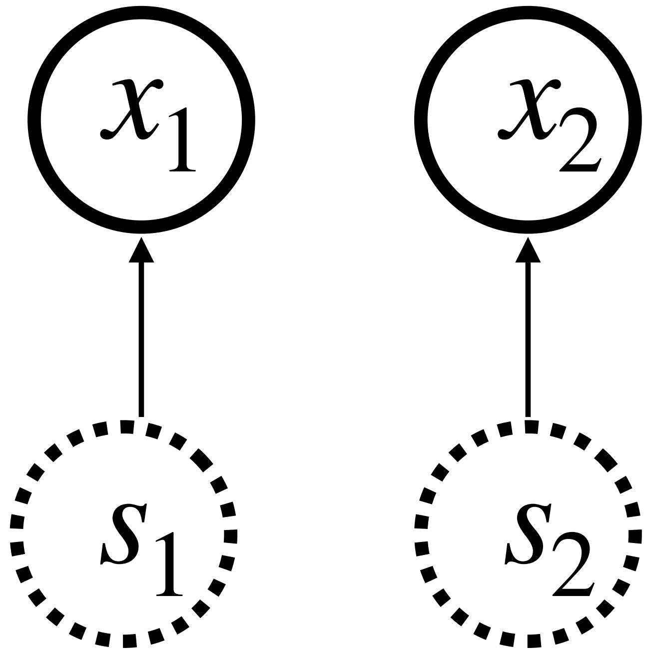
Each state s has an emission distribution $\mu_s \in \Delta(X)$, also μ_s and $\mu_{s'}$ have **disjoint support** for any $s \neq s'$
(i.e, latent state is decodable)

Low-Rank Example

Can encode latent variables: block-MDPs







Discrete latent state space S: $|S|$ is small, transition $T: S \times A \mapsto S$

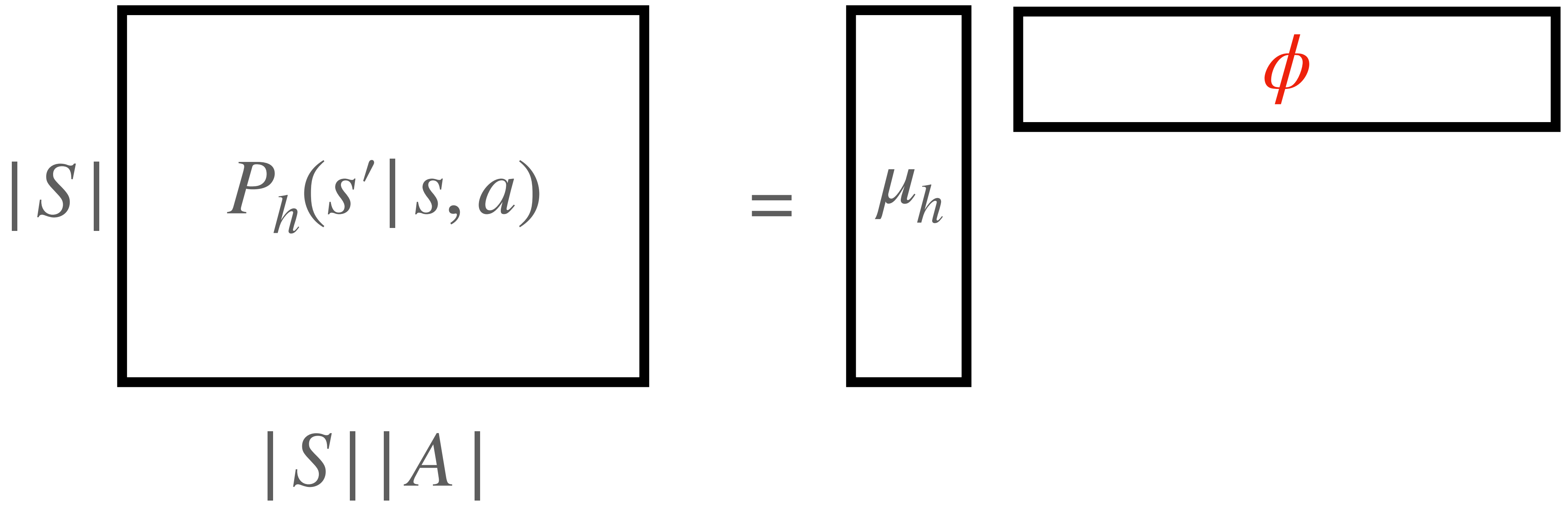
Large observation space X (hence any poly dependency on $|X|$ is bad)

Each state s has an emission distribution $\mu_s \in \Delta(X)$, also μ_s and $\mu_{s'}$ have **disjoint support** for any $s \neq s'$ (i.e, latent state is decodable)

$$P(x'|x,a) = \sum_{s' \in \{s_1,s_2,s_3\}} T(s'|\omega(x),a)\mu_{s'}(x') = [\mu_{s_1}(x'), \mu_{s_2}(x'), \mu_{s_3}(x')] \begin{bmatrix} T(s_1|\omega(x),a) \\ T(s_2|\omega(x),a) \\ T(s_3|\omega(x),a) \end{bmatrix}$$

We only study Linear MDPs here (i..e, low-rank + known ϕ).
Learning in Low-rank MDP is much harder (coming later!)

Low-Rank Decomposition:



Outline for this lecture:

1. Introduction of low-rank MDP

2. Planning in low-rank MDP (i.e., DP) and the UCBVI algorithm

3. Non-parametric model learning in linear MDPs

Planning in Linear MDP: Value Iteration

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

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

Planning in Linear MDP: Value Iteration

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

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

$V_H^\star(s) = 0, \forall s,$

Planning in Linear MDP: Value Iteration

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

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

$V_H^\star(s) = 0, \forall s,$

$Q_h^\star(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} V_{h+1}^\star(s')$

Planning in Linear MDP: Value Iteration

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

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

$V_H^\star(s) = 0, \forall s,$

$$Q_h^\star(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} V_{h+1}^\star(s')$$

$$= \theta_h^\star \cdot \phi(s, a) + \left(\mu_h^\star \phi(s, a)\right)^\top V_{h+1}^\star$$

Planning in Linear MDP: Value Iteration

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

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

$V_H^\star(s) = 0, \forall s,$

$$Q_h^\star(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)} V_{h+1}^\star(s')$$

$$= \theta_h^\star \cdot \phi(s,a) + \left(\mu_h^\star \phi(s,a)\right)^\top V_{h+1}^\star$$

$$= \phi(s,a)^\top \left(\theta_h^\star + (\mu_h^\star)^\top V_{h+1}^\star\right)$$

Planning in Linear MDP: Value Iteration

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

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

$$V_H^\star(s) = 0, \forall s,$$

$$Q_h^\star(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)} V_{h+1}^\star(s')$$

$$= \theta_h^\star \cdot \phi(s,a) + \left(\mu_h^\star \phi(s,a)\right)^\top V_{h+1}^\star$$

$$= \phi(s,a)^\top \left(\theta_h^\star + (\mu_h^\star)^\top V_{h+1}^\star\right)$$

$$= \phi(s,a)^\top w_h$$

Planning in Linear MDP: Value Iteration

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

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

$$V_H^\star(s) = 0, \forall s,$$

$$Q_h^\star(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)} V_{h+1}^\star(s')$$

$$= \theta_h^\star \cdot \phi(s,a) + \left(\mu_h^\star \phi(s,a)\right)^\top V_{h+1}^\star$$

$$= \phi(s,a)^\top \left(\theta_h^\star + (\mu_h^\star)^\top V_{h+1}^\star\right)$$

$$= \phi(s,a)^\top w_h$$

$$V_h^\star(s) = \max_a \phi(s,a)^\top w_h, \quad \pi_h^\star(s) = \arg\max_a \phi(s,a)^\top w_h$$

Planning in Linear MDP: Value Iteration

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

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

$$V_H^\star(s) = 0, \forall s,$$

$$Q_h^\star(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)} V_{h+1}^\star(s')$$

$$= \theta_h^\star \cdot \phi(s,a) + \left(\mu_h^\star \phi(s,a)\right)^\top V_{h+1}^\star$$

$$= \phi(s,a)^\top \left(\theta_h^\star + (\mu_h^\star)^\top V_{h+1}^\star\right)$$

$$= \phi(s,a)^\top w_h$$

$$V_h^\star(s) = \max_a \phi(s,a)^\top w_h, \quad \pi_h^\star(s) = \arg\max_a \phi(s,a)^\top w_h$$

Indeed we can show that $Q_h^\pi(\cdot,\cdot)$

Is linear with respect to ϕ as well, for any π, h

UCBVI in Linear MDPs

At the beginning of iteration n:

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 $\{s_h^i, a_h^i, s_{h+1}^i\}_{i=0}^{n-1}$

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 $\{s_h^i, a_h^i, s_{h+1}^i\}_{i=0}^{n-1}$

2. Design reward bonus $b_h^n(s, a), \forall s, a$

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 $\{s_h^i, a_h^i, s_{h+1}^i\}_{i=0}^{n-1}$

2. Design reward bonus $b_h^n(s, a), \forall s, a$

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

Outline for this lecture:

1. Introduction of low-rank MDP

2. Planning in low-rank MDP (i.e., DP) and the UCBVI algorithm

3. Non-parametric model learning in linear MDPs

Additional Assumptions in Linear MDPs to permit linear regression analysis

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

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

Additional Assumptions in Linear MDPs to permit linear regression analysis

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

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

Norm bounds:

$$\sup_{s,a} \|\phi(s,a)\|_2 \le 1, \ \|\theta_h^\star\|_2 \le W, \quad \|v^\top \mu_h^\star\|_2 \le \sqrt{d}, \forall v \text{ s.t. } \|v\|_\infty \le 1$$

1. Model Learning in Linear MDPs

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

1. Model Learning in Linear MDPs

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

Denote $\delta(s) \in \mathbb{R}^{|S|}$ with zero everywhere except the entry corresponding to s

1. Model Learning in Linear MDPs

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

Denote $\delta(s) \in \mathbb{R}^{|S|}$ with zero everywhere except the entry corresponding to s

Given s, a, note that $\mathbb{E}_{s' \sim P_h(\cdot|s,a)} \left[\delta(s') \right] = P_h(\cdot \mid s, a) = \mu_h^\star \phi(s, a)$

1. Model Learning in Linear MDPs

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

Denote $\delta(s) \in \mathbb{R}^{|S|}$ with zero everywhere except the entry corresponding to s

Given s, a, note that $\mathbb{E}_{s' \sim P_h(\cdot|s,a)} \left[\delta(s') \right] = P_h(\cdot \mid s, a) = \mu_h^\star \phi(s, a)$

Denote $\epsilon_{s,a} = \delta(s') - P_h(\cdot \mid s, a)$, we have $\mathbb{E}_{s' \sim P_h(\cdot|s,a)}[\epsilon_{s,a}] = 0$, and $\|\epsilon_{s,a}\|_1 \leq 2$

1. Model Learning in Linear MDPs

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

Denote $\delta(s) \in \mathbb{R}^{|S|}$ with zero everywhere except the entry corresponding to s

Given s, a, note that $\mathbb{E}_{s' \sim P_h(\cdot|s,a)} \left[\delta(s') \right] = P_h(\cdot \mid s, a) = \mu_h^\star \phi(s, a)$

Denote $\epsilon_{s,a} = \delta(s') - P_h(\cdot \mid s, a)$, we have $\mathbb{E}_{s' \sim P_h(\cdot|s,a)}[\epsilon_{s,a}] = 0$, and $\|\epsilon_{s,a}\|_1 \leq 2$

Ridge Linear Regression:

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

1. Model Learning in Linear MDPs

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

Denote $\delta(s) \in \mathbb{R}^{|S|}$ with zero everywhere except the entry corresponding to s

Given s, a, note that $\mathbb{E}_{s' \sim P_h(\cdot|s,a)} \left[\delta(s') \right] = P_h(\cdot \mid s, a) = \mu_h^\star \phi(s, a)$

Denote $\epsilon_{s,a} = \delta(s') - P_h(\cdot \mid s, a)$, we have $\mathbb{E}_{s' \sim P_h(\cdot|s,a)}[\epsilon_{s,a}] = 0$, and $\|\epsilon_{s,a}\|_1 \leq 2$

Ridge Linear Regression:

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

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

1. Model Learning in Linear MDPs

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

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

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

1. Model Learning in Linear MDPs

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

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

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

Can we bound the ℓ_1 error on distributions, i.e., $\|\widehat{P}^n(\cdot \mid s, a) - P(\cdot \mid s, a)\|_1$?

1. Model Learning in Linear MDPs

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

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

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

Can we bound the ℓ_1 error on distributions, i.e., $\| \widehat{P}^n(\cdot \mid s, a) - P(\cdot \mid s, a)\|_1$?

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

1. Model Learning in Linear MDPs

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

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

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

Can we bound the ℓ_1 error on distributions, i.e., $\|\widehat{P}^n(\cdot \mid s, a) - P(\cdot \mid s, a)\|_1$?

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

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

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

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

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

1. Model Learning in Linear MDPs

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

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

1. Model Learning in Linear MDPs

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

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

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

1. Model Learning in Linear MDPs

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

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

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

1. Model Learning in Linear MDPs

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

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

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

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

1. Model Learning in Linear MDPs

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1. Model Learning in Linear MDPs

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

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

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

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

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:

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

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:

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

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

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:

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

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

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

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:

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

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

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

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

2. Reward Bonus Design

Lemma [Model Average Error under a fixed V]:

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

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

Next lecture: reward bonus design + regret bound

Summary for today:

1. Introduction of low-rank / Linear MDPs (linear $Q^\star, Q^\pi$ in feature ϕ)

2. Model-fitting in low-rank MDP (non-parametric regression)

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

3. Average model-error (over a fixed function V):

Summary for today:

1. Introduction of low-rank / Linear MDPs (linear $Q^\star, Q^\pi$ in feature ϕ)

2. Model-fitting in low-rank MDP (non-parametric regression)

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

3. Average model-error (over a fixed function V):

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