

# Some Recent Progresses on Partially Observable RL: B-Stability, Sharp Algorithms, and Lower Bounds

Yu Bai

Salesforce Research



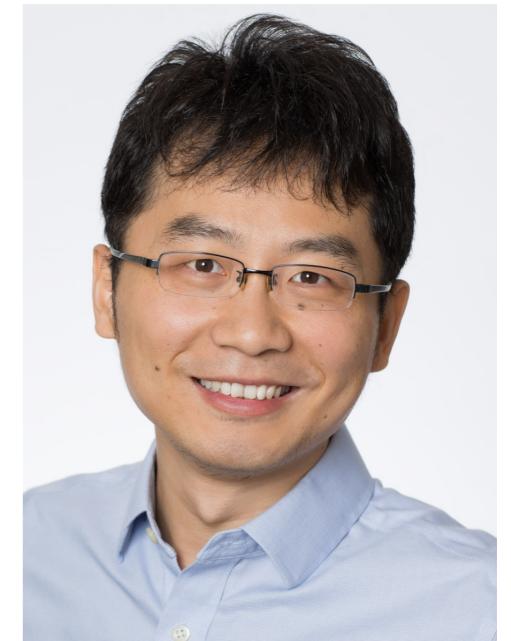
Fan Chen (Peking U -> PhD)



Song Mei (UC Berkeley)

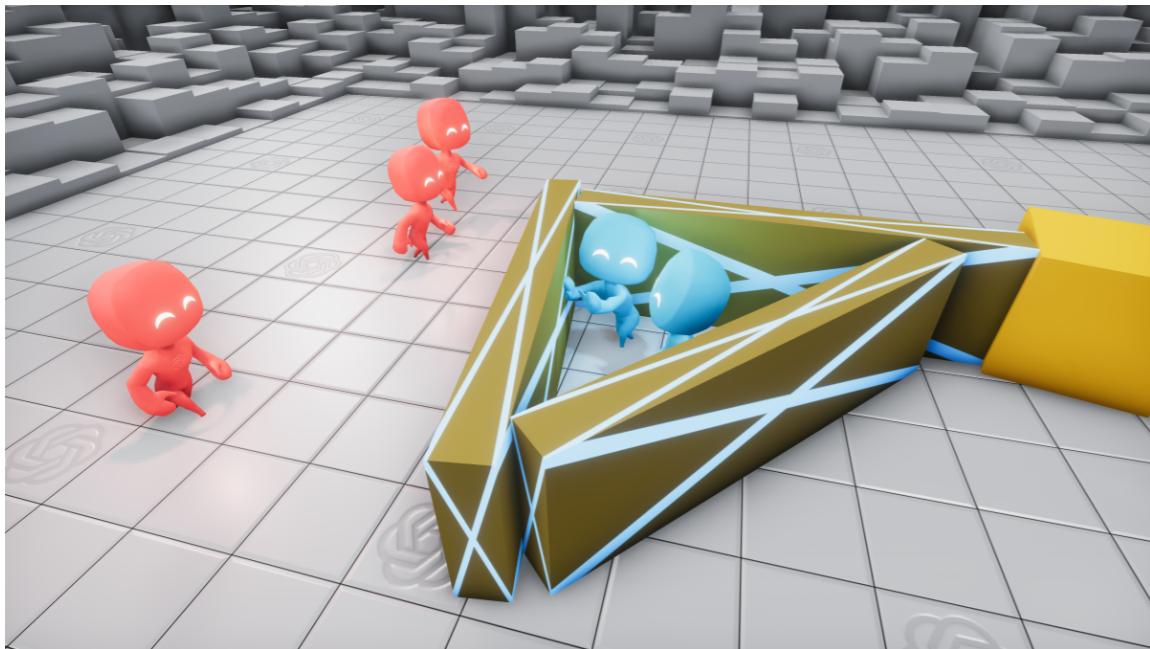


Huan Wang (Salesforce)

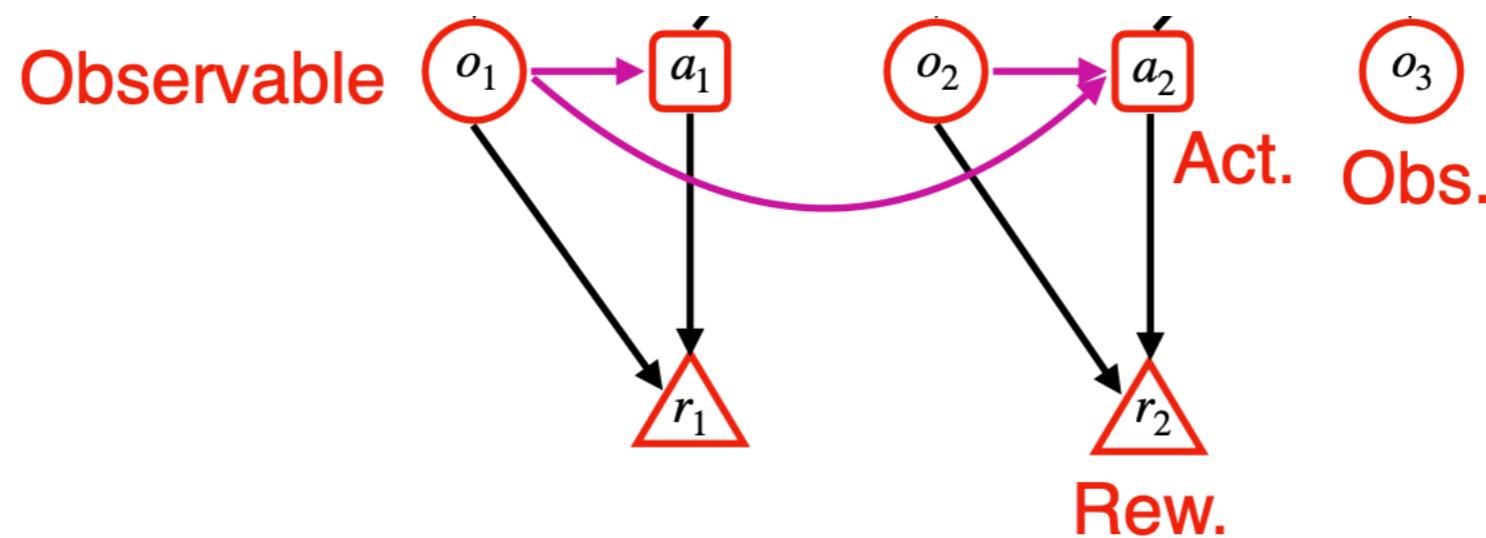


Caiming Xiong (Salesforce)

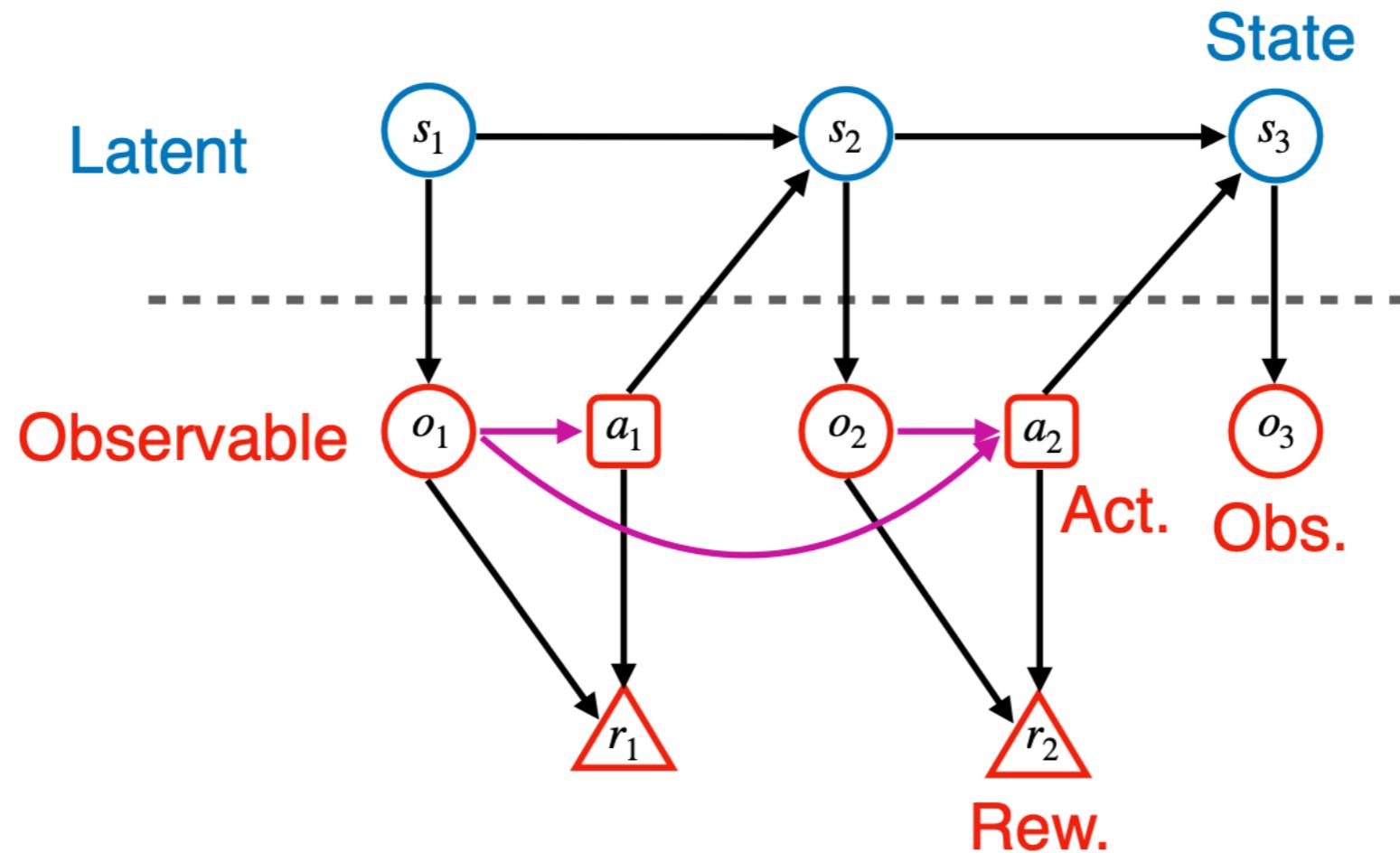
# Partial Observability in Reinforcement Learning



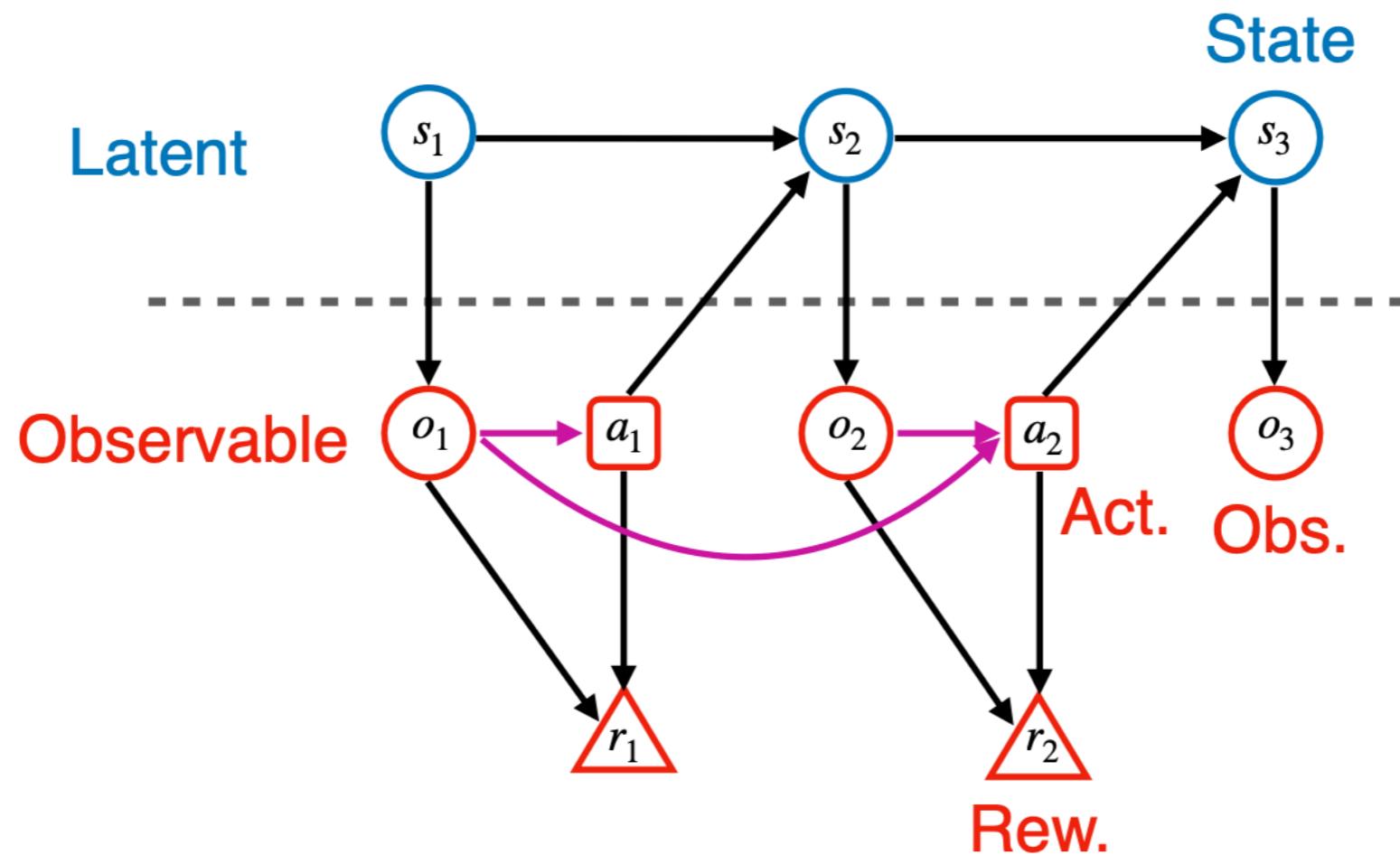
# Partially Observable Markov Decision Processes (POMDPs)



# Partially Observable Markov Decision Processes (POMDPs)



# Partially Observable Markov Decision Processes (POMDPs)



**POMDPs** = MDPs + Observations = HMMs (Hidden Markov Models) + Actions

# Challenge for Learning in POMDPs

# Challenge for Learning in POMDPs

Tabular MDPs ( $S$  states,  $A$  actions,  $H$  steps):

$\epsilon$ -optimal policy can be found in  $\text{poly}(H, S, A, 1/\epsilon)$  time and samples

[Bellman '57, Howard '60, Bertsekas '87, Kearns & Singh '02, Azar et al. '17, Sidford et al. '18, Jin et al. '18...]

# Challenge for Learning in POMDPs

Tabular MDPs ( $S$  states,  $A$  actions,  $H$  steps):

$\epsilon$ -optimal policy can be found in  $\text{poly}(H, S, A, 1/\epsilon)$  time and samples

[Bellman '57, Howard '60, Bertsekas '87, Kearns & Singh '02, Azar et al. '17, Sidford et al. '18, Jin et al. '18...]

Tabular POMDPs ( $S$  *latent* states,  $O$  observations,  $A$  actions,  $H$  steps):

- Reason about belief over states
- Policies are history-dependent in general, requires  $2^{\Omega(H)}$  memory to store
- All while exploring the environment

# Computational and Statistical Hardness

## Computational hardness

*Planning* is already hard:

# Computational and Statistical Hardness

## Computational hardness

*Planning* is already hard:

- Computing optimal policy is PSPACE-complete [Papadimitrou & Tsitsiklis '87]

# Computational and Statistical Hardness

## Computational hardness

*Planning* is already hard:

- Computing optimal policy is PSPACE-complete [Papadimitrou & Tsitsiklis '87]
- Learning optimal *memoryless* policy is NP-hard [Massis et al. '12]

# Computational and Statistical Hardness

## Computational hardness

*Planning* is already hard:

- Computing optimal policy is PSPACE-complete [Papadimitrou & Tsitsiklis '87]
- Learning optimal *memoryless* policy is NP-hard [Massis et al. '12]

## Statistical hardness

(with  $\infty$  compute):

# Computational and Statistical Hardness

## Computational hardness

*Planning* is already hard:

- Computing optimal policy is PSPACE-complete [Papadimitrou & Tsitsiklis '87]
- Learning optimal *memoryless* policy is NP-hard [Massis et al. '12]

## Statistical hardness

(with  $\infty$  compute):

- Learning optimal policy requires  $\exp(\Omega(H))$  samples in the worst-case [Krishnamurthy et al. '16]

# Computational and Statistical Hardness

## Computational hardness

*Planning* is already hard:

- Computing optimal policy is PSPACE-complete [Papadimitrou & Tsitsiklis '87]
- Learning optimal *memoryless* policy is NP-hard [Massis et al. '12]

## Statistical hardness

(with  $\infty$  compute):

- Learning optimal policy requires  $\exp(\Omega(H))$  samples in the worst-case [Krishnamurthy et al. '16]
- Hard instance: “non-revealing combination lock” with “dummy observations”

# Computational and Statistical Hardness

## Computational hardness

*Planning* is already hard:

- Computing optimal policy is PSPACE-complete [Papadimitrou & Tsitsiklis '87]
- Learning optimal *memoryless* policy is NP-hard [Massis et al. '12]

## Statistical hardness

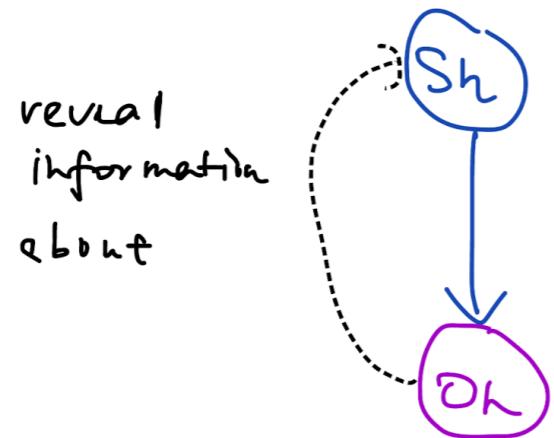
(with  $\infty$  compute):

- Learning optimal policy requires  $\exp(\Omega(H))$  samples in the worst-case [Krishnamurthy et al. '16]
- Hard instance: “non-revealing combination lock” with “dummy observations”

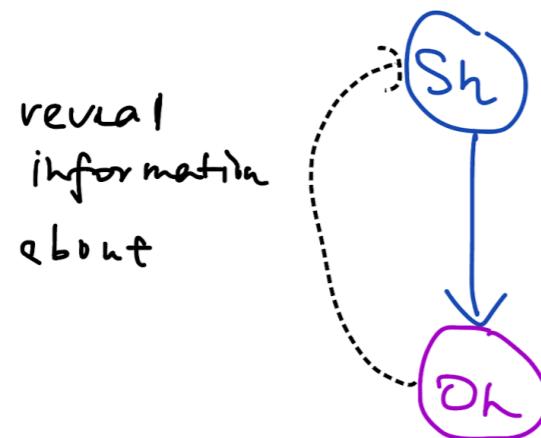
**Question:** What are “**tractable**” **subclasses** of POMDPs that can be learned with poly samples, **how sharply**, and with **what algorithms**?

# Tractable Subclasses of POMDPs

# Example 1: Revealing (Observable) POMDPs



# Example 1: Revealing (Observable) POMDPs



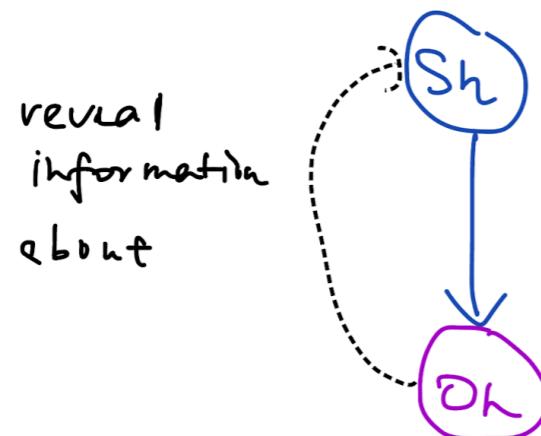
Emission matrix:

$$D_h = \begin{bmatrix} & & \end{bmatrix} \quad \Phi_h(\cdot | s) \in \Delta(O)$$

emission probabilities  
at state  $s$ .

The matrix  $D_h$  is shown with three columns. The middle column is highlighted with diagonal hatching. Braces below the matrix indicate the columns correspond to the states  $s$ .

# Example 1: Revealing (Observable) POMDPs



Emission matrix:

$$\mathbb{D}_h = \mathbb{D}_0 \left[ \begin{bmatrix} & & \end{bmatrix} \right] \quad \mathbb{D}_h(\cdot | s) \in \Delta(O)$$

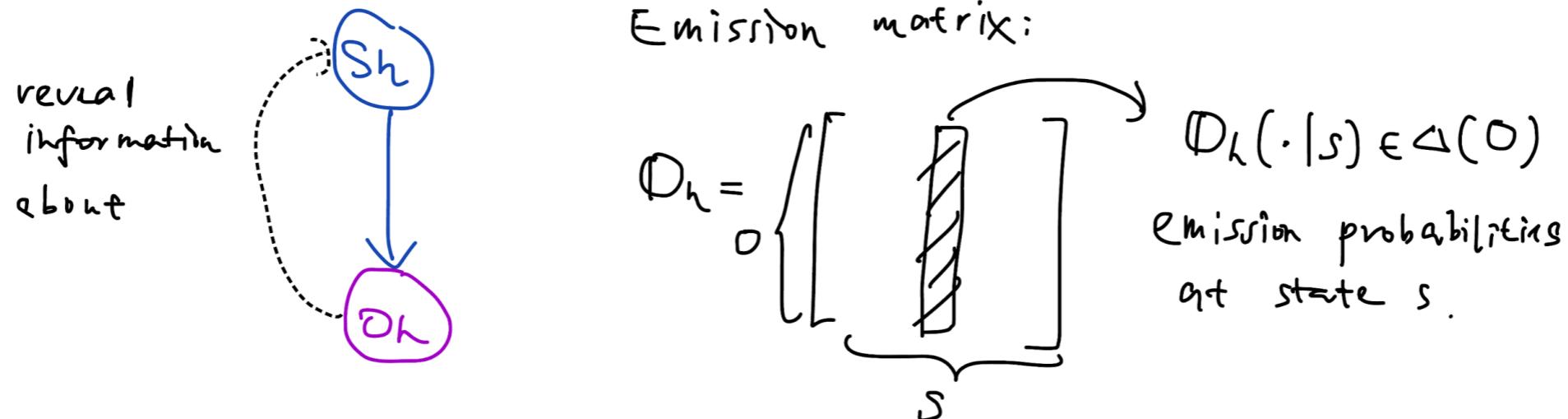
emission probabilities  
at state  $s$ .

The diagram shows a matrix  $\mathbb{D}_h$  with a column highlighted with diagonal lines, representing the emission probabilities for observation  $h$  at state  $s$ .

Desire: Emission matrices have full column rank  $\Rightarrow$  different states are probabilistically distinguishable from their emitted observations

Rules out “uninformative” observations.

# Example 1: Revealing (Observable) POMDPs



Desire: Emission matrices have full column rank  $\Rightarrow$  different states are probabilistically distinguishable from their emitted observations

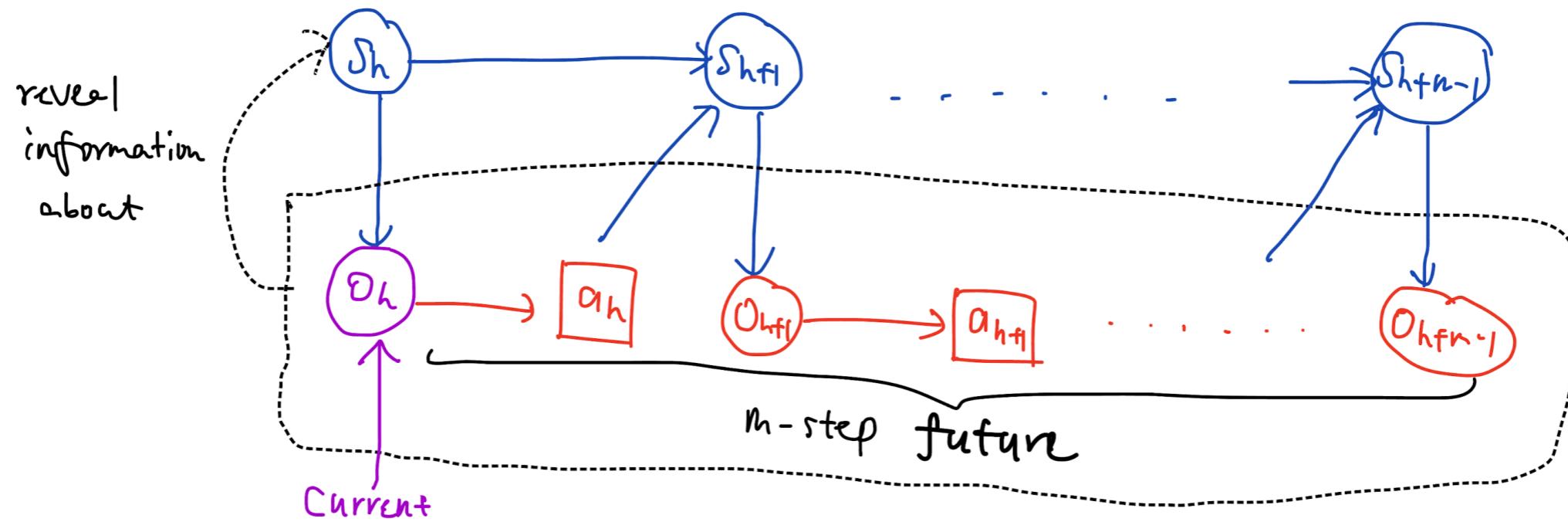
Rules out “uninformative” observations.

Single-step  $\alpha$ -revealing POMDPs [Jin et al. '20]: The emission matrices at all step  $h \in [H]$  satisfy

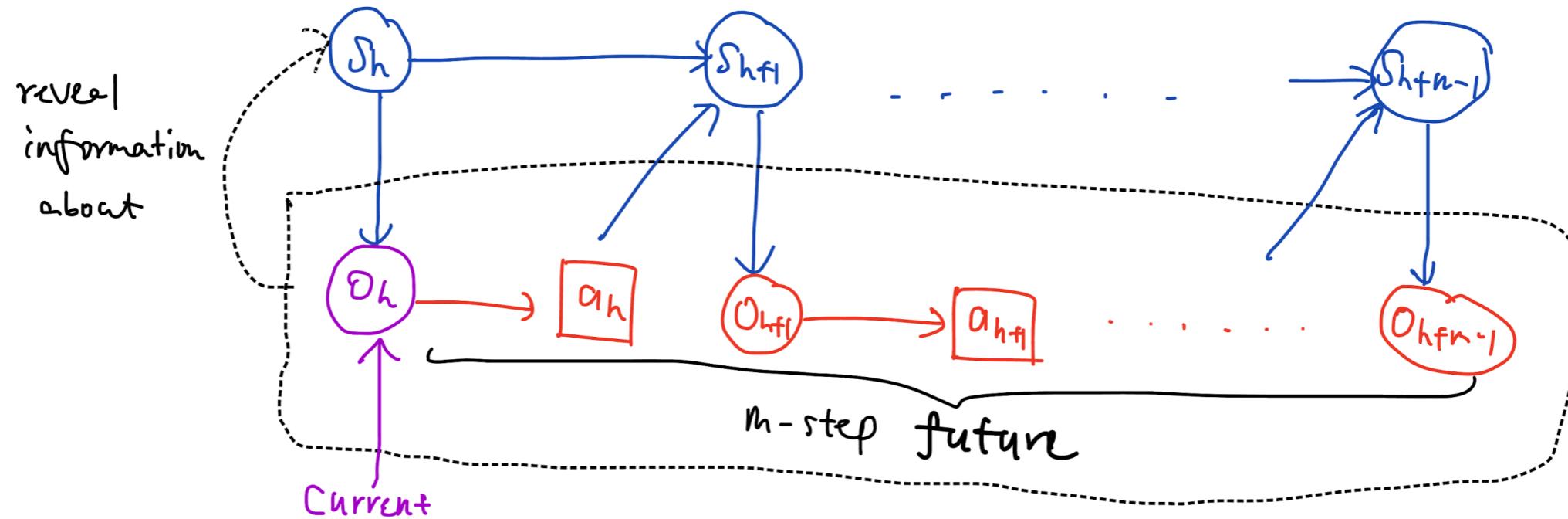
$$\|O_h^+\| \leq \alpha^{-1},$$

( $\|\cdot\|$  is some operator norm, and  $A^+$  is any *left inverse* of matrix  $A$ )

# Multi-step Revealing POMDPs



# Multi-step Revealing POMDPs

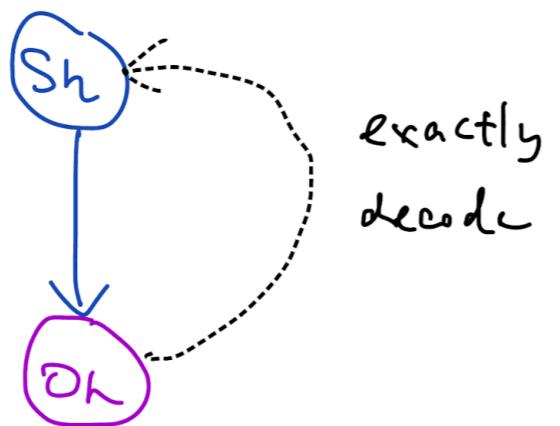


**$m$ -step  $\alpha$ -revealing POMDPs** [Liu et al. '22a]: The  $m$ -step emission-action matrices at all step  $h \in [H]$  satisfy

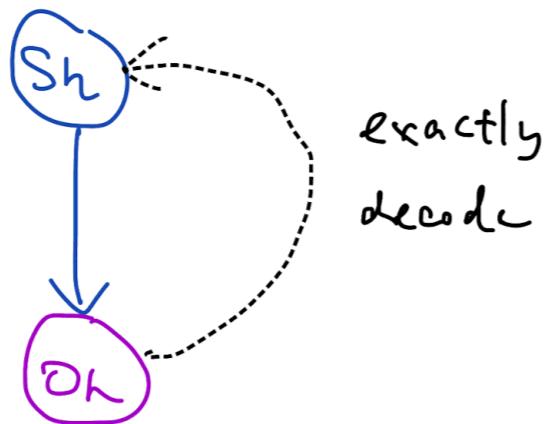
$$\|\mathbb{M}_{h,m}^+\| \leq \alpha^{-1},$$

( $\|\cdot\|$  is some operator norm, and  $A^+$  is any *left inverse* of matrix  $A$ )

## Example 2: Decodable POMDPs

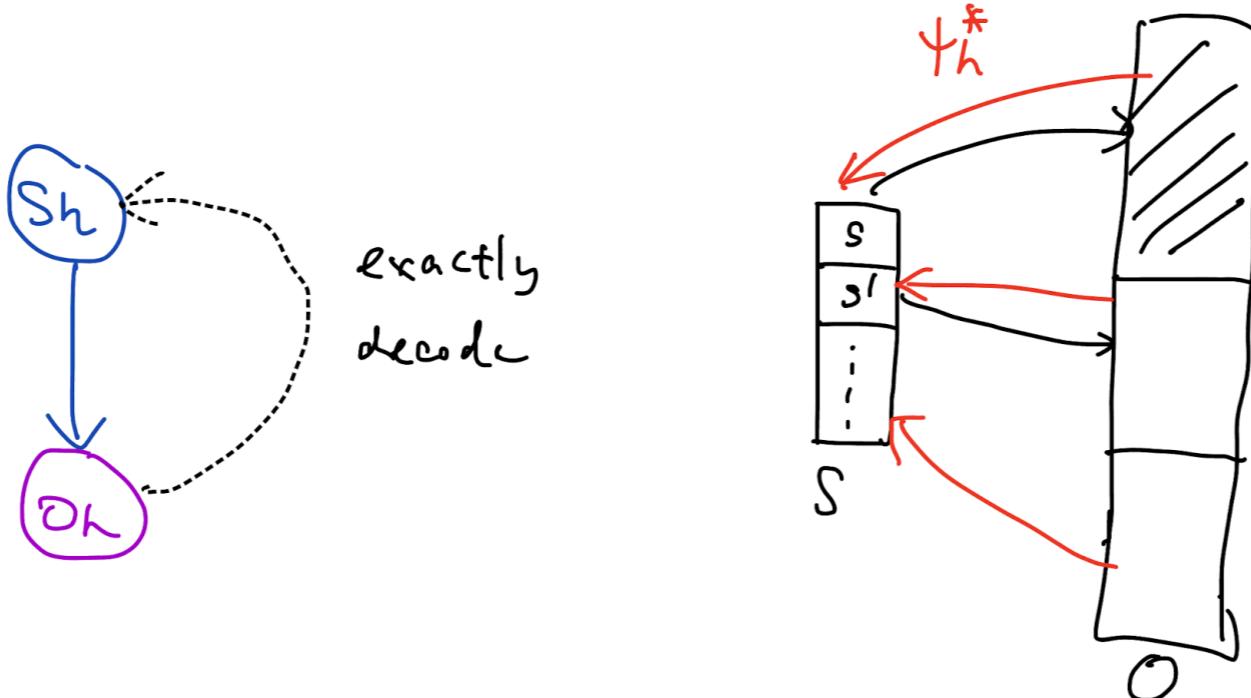


## Example 2: Decodable POMDPs



Desire: Latent state can be uniquely determined from the observation.

## Example 2: Decodable POMDPs

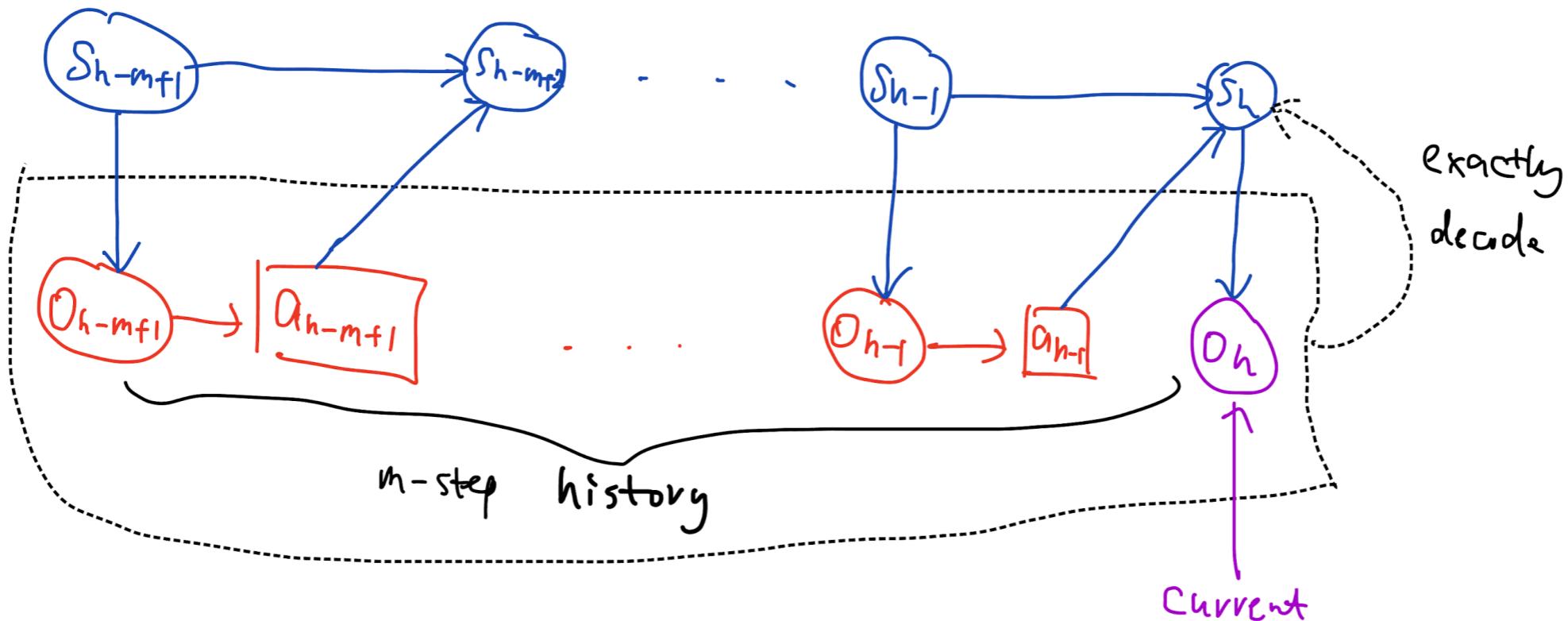


Desire: Latent state can be uniquely determined from the observation.

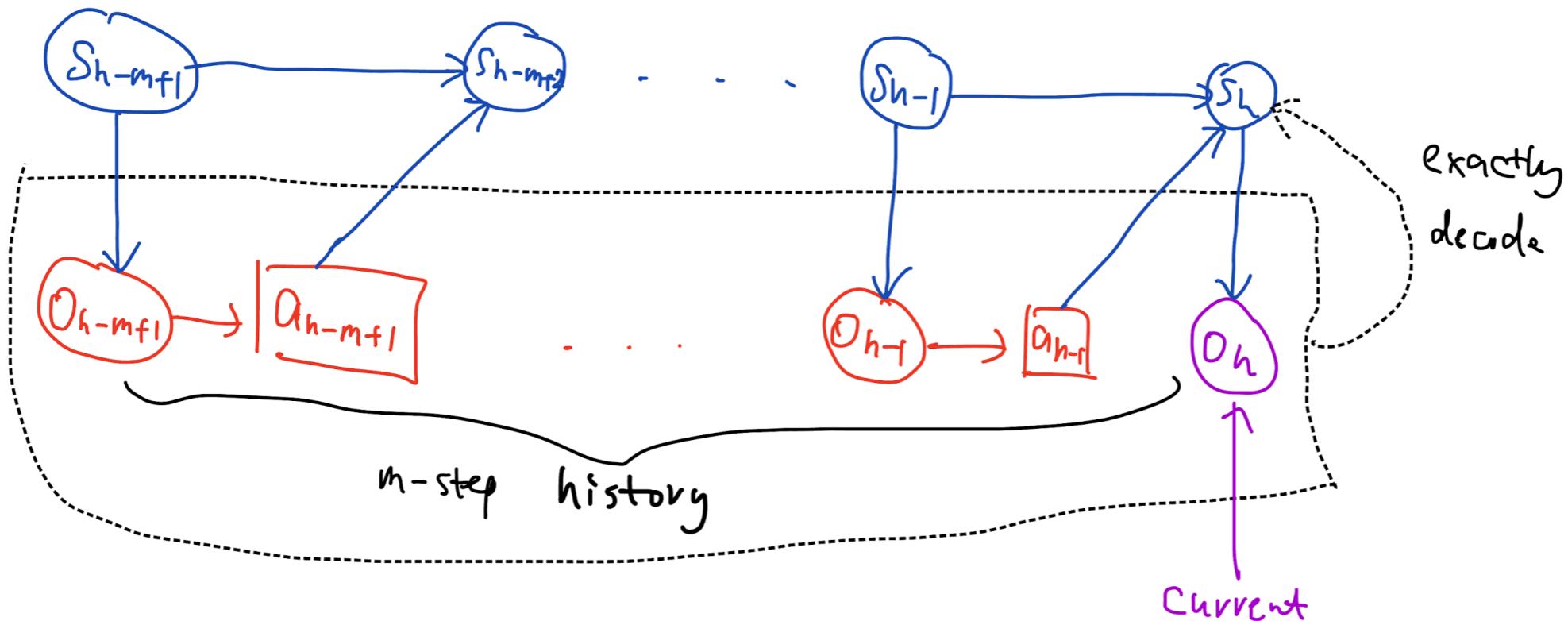
Block MDPs [Du et al. '19]: There exists an (unknown) decoder  $\psi_h^*$  at every step  $h \in [H]$  such that

$$s_h = \psi_h^*(o_h).$$

# Multi-Step Decodable POMDPs



# Multi-Step Decodable POMDPs



**m-step decodable MDPs** [Efroni et al. '22]: There exists an (unknown) decoder  $\psi_h^*$  at every step  $h \in [H]$  such that

$$s_h = \psi_h^*(o_{h-m+1}, a_{h-m+1}, \dots, o_{h-1}, a_{h-1}, o_h).$$

# Existing Work

Learning with polynomial samples has been shown to be possible within revealing POMDPs, decodable POMDPs, etc.

# Existing Work

Learning with polynomial samples has been shown to be possible within revealing POMDPs, decodable POMDPs, etc.

However:

# Existing Work

Learning with polynomial samples has been shown to be possible within revealing POMDPs, decodable POMDPs, etc.

However:

Different Algorithms for each class of POMDPs

# Existing Work

Learning with polynomial samples has been shown to be possible within revealing POMDPs, decodable POMDPs, etc.

However:

Different Algorithms for each class of POMDPs

Case-by-case analysis and no unification of the proof techniques

# Existing Work

Learning with polynomial samples has been shown to be possible within revealing POMDPs, decodable POMDPs, etc.

However:

Different Algorithms for each class of POMDPs

Case-by-case analysis and no unification of the proof techniques

Other tractable classes & tasks:

- Reactive POMDPs [Jiang et al. '17]
- Latent MDPs [Kwon et al. '21, Zhou et al. '22]
- Future-sufficient low-rank POMDPs [Wang et al. '22]
- Linear POMDPs [Cai et al. '22]
- Learning short-memory policies [Uehara et al. '22]
- ...

A partial unification: Regular PSRs [Zhan et al. '22]

# **Unified Condition: B-Stability**

# B-Representation of POMDPs

[Jaeger '00]

# B-Representation of POMDPs

[Jaeger '00]

B-representation: Any set of matrices  $\{B_h(o, a)\}_{h, o, a}$  and vector  $\mu_1$  such that for any trajectory  $\tau$ , policy  $\pi$ ,

$$\mathbb{P}_h^\pi(\tau) = \pi(\tau) \times [B_H(o_H, a_H)B_{H-1}(o_{H-1}, a_{H-1}) \cdots B_1(o_1, a_1)\mu_1],$$

Above,

$$\pi(\tau) = \prod_{h=1}^H \pi(a_h | \tau_{1:h-1}, o_h)$$

# B-Representation of POMDPs

[Jaeger '00]

B-representation: Any set of matrices  $\{B_h(o, a)\}_{h, o, a}$  and vector  $\mu_1$  such that for any trajectory  $\tau$ , policy  $\pi$ ,

$$\mathbb{P}_h^\pi(\tau) = \pi(\tau) \times [B_H(o_H, a_H)B_{H-1}(o_{H-1}, a_{H-1}) \cdots B_1(o_1, a_1)\mu_1],$$

Above,

$$\pi(\tau) = \prod_{h=1}^H \pi(a_h | \tau_{1:h-1}, o_h)$$

Example: Single-step revealing POMDPs

$$B_h(o_h, a_h) = \mathbb{O}_{h+1} \mathbb{T}_{h, a_h} \text{diag}(\mathbb{O}_h(o_h | \cdot)) \mathbb{O}_h^+$$

# B-Representation of POMDPs

[Jaeger '00]

B-representation: Any set of matrices  $\{B_h(o, a)\}_{h, o, a}$  and vector  $\mu_1$  such that for any trajectory  $\tau$ , policy  $\pi$ ,

$$\mathbb{P}_h^\pi(\tau) = \pi(\tau) \times [B_H(o_H, a_H) B_{H-1}(o_{H-1}, a_{H-1}) \cdots B_1(o_1, a_1) \mu_1],$$

Above,

$$\pi(\tau) = \prod_{h=1}^H \pi(a_h | \tau_{1:h-1}, o_h)$$

Example: Single-step revealing POMDPs

$$B_h(o_h, a_h) = \mathbb{O}_{h+1} \mathbb{T}_{h, a_h} \text{diag}(\mathbb{O}_h(o_h | \cdot)) \mathbb{O}_h^+$$

OxS emission matrix

# B-Representation of POMDPs

[Jaeger '00]

B-representation: Any set of matrices  $\{B_h(o, a)\}_{h,o,a}$  and vector  $\mu_1$  such that for any trajectory  $\tau$ , policy  $\pi$ ,

$$\mathbb{P}_h^\pi(\tau) = \pi(\tau) \times [B_H(o_H, a_H)B_{H-1}(o_{H-1}, a_{H-1}) \cdots B_1(o_1, a_1)\mu_1],$$

Above,

$$\pi(\tau) = \prod_{h=1}^H \pi(a_h | \tau_{1:h-1}, o_h)$$

Example: Single-step revealing POMDPs

$$B_h(o_h, a_h) = \mathbb{O}_{h+1} \mathbb{T}_{h,a_h} \text{diag}(\mathbb{O}_h(o_h | \cdot)) \mathbb{O}_h^+$$

SxS latent transition matrix

OxS emission matrix

# B-Representation of POMDPs

[Jaeger '00]

B-representation: Any set of matrices  $\{B_h(o, a)\}_{h, o, a}$  and vector  $\mu_1$  such that for any trajectory  $\tau$ , policy  $\pi$ ,

$$\mathbb{P}_h^\pi(\tau) = \pi(\tau) \times [B_H(o_H, a_H) B_{H-1}(o_{H-1}, a_{H-1}) \cdots B_1(o_1, a_1) \mu_1],$$

Above,

$$\pi(\tau) = \prod_{h=1}^H \pi(a_h | \tau_{1:h-1}, o_h)$$

Example: Single-step revealing POMDPs

$$B_h(o_h, a_h) = \mathbb{O}_{h+1} \mathbb{T}_{h, a_h} \text{diag}(\mathbb{O}_h(o_h | \cdot)) \mathbb{O}_h^+$$

SxS latent transition matrix

OxS emission matrix

SxS diagonal matrix of emission probabilities of  $o_h$

# B-Representation of POMDPs

[Jaeger '00]

B-representation: Any set of matrices  $\{B_h(o, a)\}_{h,o,a}$  and vector  $\mu_1$  such that for any trajectory  $\tau$ , policy  $\pi$ ,

$$\mathbb{P}_h^\pi(\tau) = \pi(\tau) \times [B_H(o_H, a_H) B_{H-1}(o_{H-1}, a_{H-1}) \cdots B_1(o_1, a_1) \mu_1],$$

Above,

$$\pi(\tau) = \prod_{h=1}^H \pi(a_h | \tau_{1:h-1}, o_h)$$

Example: Single-step revealing POMDPs

$$B_h(o_h, a_h) = \mathbb{O}_{h+1} \mathbb{T}_{h,a_h} \text{diag}(\mathbb{O}_h(o_h | \cdot)) \mathbb{O}_h^+$$

SxS latent transition matrix

OxS emission matrix

SxS diagonal matrix of emission probabilities of  $o_h$

To verify, for any fixed  $h$ ,

$$B_{h:1}(\tau_h) = \mathbb{O}_{h+1} \mathbb{T}_{h,a_h} \text{diag}(\mathbb{O}_h(o_h | \cdot)) \mathbb{T}_{h-1,a_{h-1}} \text{diag}(\mathbb{O}_h(o_{h-1} | \cdot)) \cdots \mathbb{T}_{1,a_1} \text{diag}(\mathbb{O}_h(o_1 | \cdot)) \mu_1$$

indeed yields emission probabilities

# B-Representations of POMDPs

[Jaeger '00]

B-representation: Any set of matrices  $\{B_h(o, a)\}_{h, o, a}$  and vector  $\mu_1$  such that for any trajectory  $\tau$ , policy  $\pi$ ,

$$\mathbb{P}_h^\pi(\tau) = \pi(\tau) \times [B_H(o_H, a_H)B_{H-1}(o_{H-1}, a_{H-1}) \cdots B_1(o_1, a_1)\mu_1],$$

Above,

$$\pi(\tau) = \prod_{h=1}^H \pi(a_h | \tau_{1:h-1}, o_h)$$

# B-Representations of POMDPs

[Jaeger '00]

B-representation: Any set of matrices  $\{B_h(o, a)\}_{h, o, a}$  and vector  $\mu_1$  such that for any trajectory  $\tau$ , policy  $\pi$ ,

$$\mathbb{P}_h^\pi(\tau) = \pi(\tau) \times [B_H(o_H, a_H)B_{H-1}(o_{H-1}, a_{H-1}) \cdots B_1(o_1, a_1)\mu_1],$$

Above,

$$\pi(\tau) = \prod_{h=1}^H \pi(a_h | \tau_{1:h-1}, o_h)$$

More Generally,

B-representation  $\iff$  full dynamics of the observables

# B-Representations of POMDPs

[Jaeger '00]

B-representation: Any set of matrices  $\{B_h(o, a)\}_{h, o, a}$  and vector  $\mu_1$  such that for any trajectory  $\tau$ , policy  $\pi$ ,

$$\mathbb{P}_h^\pi(\tau) = \pi(\tau) \times [B_H(o_H, a_H)B_{H-1}(o_{H-1}, a_{H-1}) \cdots B_1(o_1, a_1)\mu_1],$$

Above,

$$\pi(\tau) = \prod_{h=1}^H \pi(a_h | \tau_{1:h-1}, o_h)$$

More Generally,

B-representation  $\iff$  full dynamics of the observables

## Predictive State Representations (PSRs) [Littman & Sutton '01]

- Any Sequential Decision Process (SDP) that admits a B-representation
- Any Sequential Decision Process (SDP) that admits core test sets  
(two equivalent definitions)

A generalization of POMDPs.

# B-Stability Condition

[Chen, Bai, Mei '22]

A POMDP/PSR is called B-Stable with parameter  $\Lambda_B > 0$ , if for all  $h \in [H]$ ,

$$\|\mathcal{B}_{H:h}\|_{* \rightarrow \Pi} \leq \Lambda_B,$$

where operator

$$\mathcal{B}_{H:h} : q \rightarrow [B_{H:h}(\tau_{H:h})q]_{\tau_{h:H}} = [B_H(o_H, a_H) \dots B_h(o_h, a_h)q]_{(oa)_{h:H}}$$

# B-Stability Condition

[Chen, Bai, Mei '22]

A POMDP/PSR is called B-Stable with parameter  $\Lambda_B > 0$ , if for all  $h \in [H]$ ,

$$\|\mathcal{B}_{H:h}\|_{\overline{\star} \rightarrow \Pi} \leq \Lambda_B,$$

A certain operator norm

where operator

$$\mathcal{B}_{H:h} : q \rightarrow [B_{H:h}(\tau_{H:h})q]_{\tau_{h:H}} = [B_H(o_H, a_H) \dots B_h(o_h, a_h)q]_{(oa)_{h:H}}$$

# B-Stability Condition

[Chen, Bai, Mei '22]

A POMDP/PSR is called B-Stable with parameter  $\Lambda_B > 0$ , if for all  $h \in [H]$ ,

$$\|\mathcal{B}_{H:h}\|_{* \rightarrow \overline{\Pi}} \leq \Lambda_B,$$

A certain operator norm

where operator

$$\mathcal{B}_{H:h} : q \rightarrow [B_{H:h}(\tau_{H:h})q]_{\tau_{h:H}} = [B_H(o_H, a_H) \dots B_h(o_h, a_h)q]_{(oa)_{h:H}}$$

Intuition:

$$\|\mathcal{B}_{H:h}^{\theta^*}(B_{h-1:1}^\theta \mu_1 - B_{h-1:1}^{\theta^*} \mu_1)\|_{\Pi} \leq \Lambda_B \|(B_{h-1:1}^\theta - B_{h-1:1}^{\theta^*})\mu_1\|_*$$

# B-Stability Condition

[Chen, Bai, Mei '22]

A POMDP/PSR is called B-Stable with parameter  $\Lambda_B > 0$ , if for all  $h \in [H]$ ,

$$\|\mathcal{B}_{H:h}\|_{* \rightarrow \Pi} \leq \Lambda_B,$$

A certain operator norm

where operator

$$\mathcal{B}_{H:h} : q \rightarrow [B_{H:h}(\tau_{H:h})q]_{\tau_{h:H}} = [B_H(o_H, a_H) \dots B_h(o_h, a_h)q]_{(oa)_{h:H}}$$

Intuition:

$$\|\mathcal{B}_{H:h}^{\theta^*}(B_{h-1:1}^\theta \mu_1 - B_{h-1:1}^{\theta^*} \mu_1)\|_\Pi \leq \Lambda_B \|(B_{h-1:1}^\theta - B_{h-1:1}^{\theta^*})\mu_1\|_*$$

Error from performance difference  
(for bounding Regret/PAC)

# B-Stability Condition

[Chen, Bai, Mei '22]

A POMDP/PSR is called B-Stable with parameter  $\Lambda_B > 0$ , if for all  $h \in [H]$ ,

$$\|\mathcal{B}_{H:h}\|_{* \rightarrow \Pi} \leq \Lambda_B,$$

A certain operator norm

where operator

$$\mathcal{B}_{H:h} : q \rightarrow [B_{H:h}(\tau_{H:h})q]_{\tau_{h:H}} = [B_H(o_H, a_H) \dots B_h(o_h, a_h)q]_{(oa)_{h:H}}$$

Intuition:

$$\|\mathcal{B}_{H:h}^{\theta^*}(B_{h-1:1}^\theta \mu_1 - B_{h-1:1}^{\theta^*} \mu_1)\|_\Pi \leq \Lambda_B \|(B_{h-1:1}^\theta - B_{h-1:1}^{\theta^*})\mu_1\|_*$$

Error from performance difference  
(for bounding Regret/PAC)

Estimation error of B matrices  
(Algorithm can bound)

# Landscape of POMDP/PSRs

All PSRs

**B-Stable PSRs**

Decodable POMDPs

Regular PSRs

Linear POMDPs

Low-rank Future-sufficient POMDPs

Test-sufficient Latent MDPs

Revealing POMDPs

# Algorithms and Guarantees

# Algorithms for B-Stable POMDP/PSRs

Three **model-based** algorithms with **similar principles**:

# Algorithms for B-Stable POMDP/PSRs

Three **model-based** algorithms with **similar principles**:

1. Use policy derived from belief (**optimism / posterior sampling**) to collect data

# Algorithms for B-Stable POMDP/PSRs

Three **model-based** algorithms with **similar principles**:

1. Use policy derived from belief (**optimism / posterior sampling**) to collect data
2. Update **belief about true model**, such as confidence set or posterior

# Algorithms for B-Stable POMDP/PSRs

Three **model-based** algorithms with **similar principles**:

1. Use policy derived from belief (**optimism / posterior sampling**) to collect data
2. Update **belief about true model**, such as confidence set or posterior

Note the principle is general, not limited to POMDP/PSRs.

\* For details on the connections/differences between the 3 algorithms, see our related paper [Chen, Mei, **Bai** '22b]

# Algorithm 1: OMLE

OMLE (Optimistic Maximum Likelihood Estimation) [Liu et al. '22a]

# Algorithm 1: OMLE

OMLE (Optimistic Maximum Likelihood Estimation) [Liu et al. '22a]

In each iteration  $k$ ,

# Algorithm 1: OMLE

OMLE (Optimistic Maximum Likelihood Estimation) [Liu et al. '22a]

In each iteration  $k$ ,

1. Set  $\pi^k = \arg \max_{\pi} \max_{\theta \in \mathcal{B}^k} V_{\theta}^{\pi}$  to be optimistic greedy policy wrt  $\mathcal{B}^k$

# Algorithm 1: OMLE

OMLE (Optimistic Maximum Likelihood Estimation) [Liu et al. '22a]

In each iteration  $k$ ,

1. Set  $\pi^k = \arg \max_{\pi} \max_{\theta \in \mathcal{B}^k} V_{\theta}^{\pi}$  to be **optimistic greedy policy** wrt  $\mathcal{B}^k$
2. Play corresponding “exploration policies”  $\pi_{h,\text{exp}}^k = \Pi_{h,\text{exp}}(\pi^k)$

# Algorithm 1: OMLE

OMLE (Optimistic Maximum Likelihood Estimation) [Liu et al. '22a]

In each iteration  $k$ ,

1. Set  $\pi^k = \arg \max_{\pi} \max_{\theta \in \mathcal{B}^k} V_{\theta}^{\pi}$  to be **optimistic greedy policy** wrt  $\mathcal{B}^k$
2. Play corresponding “exploration policies”  $\pi_{h,\text{exp}}^k = \Pi_{h,\text{exp}}(\pi^k)$
3. Update confidence set  $\mathcal{B}^{k+1}$  given data

$$\mathcal{B}^{k+1} = \left\{ \theta : \sum_{(\pi, \tau) \in \mathcal{D}^{k+1}} \log \mathbb{P}_{\theta}^{\pi}(\tau) \geq \max_{\theta'} \sum_{(\pi, \tau) \in \mathcal{D}^{k+1}} \log \mathbb{P}_{\theta'}^{\pi}(\tau) - \beta \right\}$$

## Algorithm 2: E2D, Based on Decision-Estimation Coefficients (DECs)

E2D (Estimation-To-Decisions) [Chen et al. '22b, Foster et al. '21]

## Algorithm 2: E2D, Based on Decision-Estimation Coefficients (DECs)

E2D (Estimation-To-Decisions) [Chen et al. '22b, Foster et al. '21]

In each iteration  $k$ ,

## Algorithm 2: E2D, Based on Decision-Estimation Coefficients (DECs)

E2D (Estimation-To-Decisions) [Chen et al. '22b, Foster et al. '21]

In each iteration  $k$ ,

1. Set policy *distributions*  $(p_{\text{exp}}^k, p_{\text{out}}^k)$  to minimize risk  $V^{\mu^k}(\cdot, \cdot)$

## Algorithm 2: E2D, Based on Decision-Estimation Coefficients (DECs)

E2D (Estimation-To-Decisions) [Chen et al. '22b, Foster et al. '21]

In each iteration  $k$ ,

1. Set policy *distributions*  $(p_{\text{exp}}^k, p_{\text{out}}^k)$  to minimize risk  $V^{\mu^k}(\cdot, \cdot)$
2. Sample and play exploration policy  $\pi^k \sim p_{\text{exp}}^k$ , obtain trajectory  $\tau^k$

## Algorithm 2: E2D, Based on Decision-Estimation Coefficients (DECs)

E2D (Estimation-To-Decisions) [Chen et al. '22b, Foster et al. '21]

In each iteration  $k$ ,

1. Set policy *distributions*  $(p_{\text{exp}}^k, p_{\text{out}}^k)$  to minimize risk  $V^{\mu^k}(\cdot, \cdot)$
2. Sample and play exploration policy  $\pi^k \sim p_{\text{exp}}^k$ , obtain trajectory  $\tau^k$
3. Update “tempered posterior” of model:

$$\mu^{k+1}(\theta) \propto_{\theta} \mu^k(\theta) \cdot \exp \left( \eta \cdot \log \mathbb{P}_{\theta}^{\pi^k}(\tau^k) \right)$$

## Algorithm 2: E2D, Based on Decision-Estimation Coefficients (DECs)

E2D (Estimation-To-Decisions) [Chen et al. '22b, Foster et al. '21]

In each iteration  $k$ ,

1. Set policy *distributions*  $(p_{\text{exp}}^k, p_{\text{out}}^k)$  to minimize risk  $V^{\mu^k}(\cdot, \cdot)$
2. Sample and play exploration policy  $\pi^k \sim p_{\text{exp}}^k$ , obtain trajectory  $\tau^k$
3. Update “tempered posterior” of model:

$$\mu^{k+1}(\theta) \propto_{\theta} \mu^k(\theta) \cdot \exp\left(\eta \cdot \log \mathbb{P}_{\theta}^{\pi^k}(\tau^k)\right)$$

Output policy  $p_{\text{out}} = \frac{1}{K} \sum_{k=1}^K p_{\text{out}}^k$

## Algorithm 2: E2D, Based on Decision-Estimation Coefficients (DECs)

### E2D (Estimation-To-Decisions) [Chen et al. '22b, Foster et al. '21]

In each iteration  $k$ ,

1. Set policy *distributions*  $(p_{\text{exp}}^k, p_{\text{out}}^k)$  to minimize risk  $V^{\mu^k}(\cdot, \cdot)$
2. Sample and play exploration policy  $\pi^k \sim p_{\text{exp}}^k$ , obtain trajectory  $\tau^k$
3. Update “tempered posterior” of model:

$$\mu^{k+1}(\theta) \propto_{\theta} \mu^k(\theta) \cdot \exp\left(\eta \cdot \log \mathbb{P}_{\theta}^{\pi^k}(\tau^k)\right)$$

Output policy  $p_{\text{out}} = \frac{1}{K} \sum_{k=1}^K p_{\text{out}}^k$

Risk functional determined by the Explorative DEC:

$$V^{\mu^k}(p_{\text{exp}}, p_{\text{out}}) = \mathbb{E}_{\pi \sim p_{\text{out}}}[V_{\theta}^{\pi_{\theta}} - V_{\theta}^{\pi}] - \gamma \mathbb{E}_{\pi \sim p_{\text{exp}}} \mathbb{E}_{\theta^k \sim \mu^k} [D_H^2(\mathbb{P}_{\theta}^{\pi}, \mathbb{P}_{\theta^k}^{\pi})]$$

## Algorithm 3: MOPS

MOPS (Model-based Optimistic Posterior Sampling) [Agarwal & Zhang '22]

---

### Algorithm 4 MODEL-BASED OPTIMISTIC POSTERIOR SAMPLING (Agarwal and Zhang, 2022)

---

- 1: **Input:** Parameters  $\gamma > 0$ ,  $\eta \in (0, 1/2)$ . An  $1/T$ -optimistic cover  $(\tilde{\mathbb{P}}, \Theta_0)$
- 2: **Initialize:**  $\mu^1 = \text{Unif}(\Theta_0)$
- 3: **for**  $t = 1, \dots, T$  **do**
- 4:     Sample  $\theta^t \sim \mu^t$  and  $h^t \sim \text{Unif}(\{0, 1, \dots, H - 1\})$ .
- 5:     Set  $\pi^t = \pi_{\theta^t} \circ_{h^t} \text{Unif}(\mathcal{A}) \circ_{h^t+1} \text{Unif}(\mathcal{U}_{A, h+1})$ , execute  $\pi^t$  and observe  $\tau^t$ .
- 6:     Compute  $\mu^{t+1} \in \Delta(\Theta_0)$  by

$$\mu^{t+1}(\theta) \propto_{\theta} \mu^1(\theta) \exp \left( \sum_{s=1}^t \left( \gamma^{-1} V_{\theta}(\pi_{\theta}) + \eta \log \tilde{\mathbb{P}}_{\theta}^{\pi^s}(\tau^s) \right) \right).$$

**Output:** Policy  $\hat{\pi}_{\text{out}} := \frac{1}{T} \sum_{t=1}^T p_{\text{out}}(\mu^t)$ , where  $p_{\text{out}}(\cdot)$  is defined in (46).

---

Similar as E2D, except for using optimistic posterior.

# Main Result for Learning B-Stable POMDP/PSRs

**Thm** [Chen, Mei, **Bai** '22a]: Algorithms {OMLE, E2D, MOPS} can all learn a  $\Lambda_B$ -stable POMDP/PSR within

$$K = \widetilde{O}(dAU_A\Lambda_B^2/\varepsilon^2)$$

episodes of play.

# Main Result for Learning B-Stable POMDP/PSRs

**Thm** [Chen, Mei, **Bai** '22a]: Algorithms {OMLE, E2D, MOPS} can all learn a  $\Lambda_B$ -stable POMDP/PSR within

$$K = \widetilde{O}(dAU_A\Lambda_B^2/\varepsilon^2)$$

episodes of play.

Above,

# Main Result for Learning B-Stable POMDP/PSRs

**Thm** [Chen, Mei, **Bai** '22a]: Algorithms {OMLE, E2D, MOPS} can all learn a  $\Lambda_B$ -stable POMDP/PSR within

$$K = \widetilde{O}(dAU_A\Lambda_B^2/\varepsilon^2)$$

episodes of play.

Above,

- $d$ : PSR rank ( $d \leq S$  for POMDPs)

# Main Result for Learning B-Stable POMDP/PSRs

**Thm** [Chen, Mei, **Bai** '22a]: Algorithms {OMLE, E2D, MOPS} can all learn a  $\Lambda_B$ -stable POMDP/PSR within

$$K = \widetilde{O}(dAU_A\Lambda_B^2/\varepsilon^2)$$

episodes of play.

Above,

- $d$ : PSR rank ( $d \leq S$  for POMDPs)
- $A$ : number of actions

# Main Result for Learning B-Stable POMDP/PSRs

**Thm** [Chen, Mei, **Bai** '22a]: Algorithms {OMLE, E2D, MOPS} can all learn a  $\Lambda_B$ -stable POMDP/PSR within

$$K = \widetilde{O}(dAU_A\Lambda_B^2/\varepsilon^2)$$

episodes of play.

Above,

- $d$ : PSR rank ( $d \leq S$  for POMDPs)
- $A$ : number of actions
- $U_A$ : number of *core actions* (equals  $A^{m-1}$  for m-step revealing/decodable)

# Main Result for Learning B-Stable POMDP/PSRs

**Thm** [Chen, Mei, **Bai** '22a]: Algorithms {OMLE, E2D, MOPS} can all learn a  $\Lambda_B$ -stable POMDP/PSR within

$$K = \widetilde{O}(dAU_A\Lambda_B^2/\varepsilon^2)$$

episodes of play.

Above,

- $d$ : PSR rank ( $d \leq S$  for POMDPs)
- $A$ : number of actions
- $U_A$ : number of *core actions* (equals  $A^{m-1}$  for m-step revealing/decodable)

First  $\Lambda_B^2$  rate (previous works at least  $\Lambda_B^4$  on their stability/regularity parameters)

# Instantiations to Concrete Subclasses

Table 1: **Comparisons of sample complexities** for learning an  $\varepsilon$  near-optimal policy in POMDPs and PSRs. Definitions of the problem parameters can be found in Section 3.2. The last three rows refer to the  $m$ -step versions of the problem classes (e.g. the third row considers  $m$ -step  $\alpha_{\text{rev}}$ -revealing POMDPs). The current best results within the last four rows are due to [Zhan et al. \(2022\)](#); [Liu et al. \(2022a\)](#); [Wang et al. \(2022\)](#); [Efroni et al. \(2022\)](#) respectively<sup>1</sup>. All results are scaled to the setting with total reward in  $[0, 1]$ .

Problem Class	Current Best	Ours
$\Lambda_B$ -stable PSR	-	$\tilde{\mathcal{O}}(d_{\text{PSR}}AU_AH^2 \log \mathcal{N}_\Theta \cdot \Lambda_B^2/\varepsilon^2)$
$\alpha_{\text{psr}}$ -regular PSR	$\tilde{\mathcal{O}}(d_{\text{PSR}}^4 A^4 U_A^9 H^6 \log(\mathcal{N}_\Theta O)/(\alpha_{\text{psr}}^6 \varepsilon^2))$	$\tilde{\mathcal{O}}(d_{\text{PSR}}AU_A^2 H^2 \log \mathcal{N}_\Theta /(\alpha_{\text{psr}}^2 \varepsilon^2))$
$\alpha_{\text{rev}}$ -revealing tabular POMDP	$\tilde{\mathcal{O}}(S^4 A^{6m-4} H^6 \log \mathcal{N}_\Theta /(\alpha_{\text{rev}}^4 \varepsilon^2))$	$\tilde{\mathcal{O}}(S^2 A^m H^2 \log \mathcal{N}_\Theta /(\alpha_{\text{rev}}^2 \varepsilon^2))$
$\nu$ -future-suff. rank- $d_{\text{trans}}$ POMDP	$\tilde{\mathcal{O}}(d_{\text{trans}}^4 A^{5m+3l+1} H^2 (\log \mathcal{N}_\Theta)^2 \cdot \nu^4 \gamma^2 / \varepsilon^2)$	$\tilde{\mathcal{O}}(d_{\text{trans}} A^{2m-1} H^2 \log \mathcal{N}_\Theta \cdot \nu^2 / \varepsilon^2)$
decodable rank- $d_{\text{trans}}$ POMDP	$\tilde{\mathcal{O}}(d_{\text{trans}} A^m H^2 \log \mathcal{N}_G / \varepsilon^2)$	$\tilde{\mathcal{O}}(d_{\text{trans}} A^m H^2 \log \mathcal{N}_\Theta / \varepsilon^2)$

$\log \mathcal{N}_\Theta$  = log-covering number of model class

Significantly sharper rates on revealing POMDPs, decodable POMDPs, ...

# Overview of Techniques

# Overview of Techniques

1. Performance decomposition into B-errors  
Relate regret/PAC learning objective to estimation error in “B operators”

# Overview of Techniques

1. Performance decomposition into B-errors  
Relate regret/PAC learning objective to estimation error in “B operators”
2. Bounding squared B-errors by squared Hellinger distance, using B-stability.  
All 3 algorithms control this squared Hellinger distance by algorithm design.

# Overview of Techniques

1. Performance decomposition into B-errors  
Relate regret/PAC learning objective to estimation error in “B operators”
2. Bounding squared B-errors by squared Hellinger distance, using B-stability.  
All 3 algorithms control this squared Hellinger distance by algorithm design.
3. A sharp generalized  $\ell_2$ -Eluder argument to bridge step 1 & 2

# Overview of Techniques

1. Performance decomposition into B-errors  
Relate regret/PAC learning objective to estimation error in “B operators”
2. Bounding squared B-errors by squared Hellinger distance, using B-stability.  
All 3 algorithms control this squared Hellinger distance by algorithm design.
3. A sharp generalized  $\ell_2$ -Eluder argument to bridge step 1 & 2

\* Concurrent work [Liu et al. '22b] shows B-errors  $\leq$  TV distance in their step 2, and performs  $\ell_1$ -Eluder argument in their step 3, which gives similar result but worse rate.

# Lower Bounds

# Towards Fine-Grained Studies

Understanding fundamental limits  $\leq$  studying lower bounds

# Towards Fine-Grained Studies

Understanding fundamental limits <== studying **lower bounds**

- In MDPs, lower bounds [Jaksch et al. '10, Azar et al. '13] *predated* the matching upper bounds [Azar et al. '17, Sidford et al. '18] for suggesting the minimax PAC sample complexity

$$\widetilde{\Theta} (H^3SA/\varepsilon^2)$$

# Towards Fine-Grained Studies

Understanding fundamental limits <== studying **lower bounds**

- In MDPs, lower bounds [Jaksch et al. '10, Azar et al. '13] *predated* the matching upper bounds [Azar et al. '17, Sidford et al. '18] for suggesting the minimax PAC sample complexity
$$\widetilde{\Theta} (H^3SA/\varepsilon^2)$$
- Often provide intuitions / directions for improvement

# Case Study: (Tabular) Revealing POMDPs

Our result (current best) for learning  $m$ -step  $\alpha$ -revealing POMDPs:

$$\tilde{O}\left(\frac{\text{poly}(H) \cdot S^2 O A^m}{\alpha^2 \varepsilon^2}\right)$$

# Case Study: (Tabular) Revealing POMDPs

Our result (current best) for learning  $m$ -step  $\alpha$ -revealing POMDPs:

$$\tilde{O}\left(\frac{\text{poly}(H) \cdot S^2 O A^m}{\alpha^2 \varepsilon^2}\right)$$

Lower bounds quite scarce and preliminary...

1. Preliminary lower bound by [Liu et al. '22]:

$$\Omega\left(\min\{1/(\alpha H), A^{H-1}\} + A^{m-1}\right)$$

# Case Study: (Tabular) Revealing POMDPs

Our result (current best) for learning  $m$ -step  $\alpha$ -revealing POMDPs:

$$\tilde{\mathcal{O}}\left(\frac{\text{poly}(H) \cdot S^2 O A^m}{\alpha^2 \varepsilon^2}\right)$$

Lower bounds quite scarce and preliminary...

1. Preliminary lower bound by [Liu et al. '22]:

$$\Omega\left(\min\{1/(\alpha H), A^{H-1}\} + A^{m-1}\right)$$

2. By embedding {MDPs, contextual bandits}:

$$\Omega\left(\frac{H \min\{S, O\} A + O A}{\varepsilon^2}\right)$$

# Lower Bounds for Revealing POMDPs

[Chen, Wang, Xiong, Mei, Bai '23]

Problem	PAC sample complexity		Regret	
	Upper bound	Lower bound	Upper bound	Lower bound
1-step $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{S O^{1/2} A}{\alpha^2 \varepsilon^2}\right)$ (Theorem 4)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2 O^2 A}{\alpha^2} \cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{S O^{1/2} A}{\alpha^2} \cdot T}\right)$ (Corollary 7)
$m$ -step ( $m \geq 2$ ) $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A^m}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{(S^{3/2} + S A) O^{1/2} A^{m-1}}{\alpha^2 \varepsilon^2}\right)$ (Theorem 5)	$\tilde{\mathcal{O}}\left(\left(\frac{S^2 O A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{S O^{1/2} A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Theorem 6)

Omitting H, assuming  $O \geq SA$  in upper bounds

# Lower Bounds for Revealing POMDPs

[Chen, Wang, Xiong, Mei, Bai '23]

Problem	PAC sample complexity		Regret	
	Upper bound	Lower bound	Upper bound	Lower bound
1-step $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{S O^{1/2} A}{\alpha^2 \varepsilon^2}\right)$ (Theorem 4)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2 O^2 A}{\alpha^2} \cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{S O^{1/2} A}{\alpha^2} \cdot T}\right)$ (Corollary 7)
$m$ -step ( $m \geq 2$ ) $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A^m}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{(S^{3/2} + S A) O^{1/2} A^{m-1}}{\alpha^2 \varepsilon^2}\right)$ (Theorem 5)	$\tilde{\mathcal{O}}\left(\left(\frac{S^2 O A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{S O^{1/2} A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Theorem 6)

Omitting H, assuming  $O \geq SA$  in upper bounds

- First multiplicative dependence on  $S, O, A, 1/\alpha, 1/\varepsilon$

# Lower Bounds for Revealing POMDPs

[Chen, Wang, Xiong, Mei, Bai '23]

Problem	PAC sample complexity		Regret	
	Upper bound	Lower bound	Upper bound	Lower bound
1-step $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{S O^{1/2} A}{\alpha^2 \varepsilon^2}\right)$ (Theorem 4)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2 O^2 A}{\alpha^2} \cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{S O^{1/2} A}{\alpha^2} \cdot T}\right)$ (Corollary 7)
$m$ -step ( $m \geq 2$ ) $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A^m}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{(S^{3/2} + S A) O^{1/2} A^{m-1}}{\alpha^2 \varepsilon^2}\right)$ (Theorem 5)	$\tilde{\mathcal{O}}\left(\left(\frac{S^2 O A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{S O^{1/2} A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Theorem 6)

Omitting H, assuming  $O \geq SA$  in upper bounds

- First multiplicative dependence on  $S, O, A, 1/\alpha, 1/\varepsilon$
- Suggests our  $1/\alpha^2$  is sharp dependence on  $\alpha$

# Lower Bounds for Revealing POMDPs

[Chen, Wang, Xiong, Mei, Bai '23]

Problem	PAC sample complexity		Regret	
	Upper bound	Lower bound	Upper bound	Lower bound
1-step $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{S O^{1/2} A}{\alpha^2 \varepsilon^2}\right)$ (Theorem 4)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2 O^2 A}{\alpha^2} \cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{S O^{1/2} A}{\alpha^2} \cdot T}\right)$ (Corollary 7)
$m$ -step ( $m \geq 2$ ) $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A^m}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{(S^{3/2} + S A) O^{1/2} A^{m-1}}{\alpha^2 \varepsilon^2}\right)$ (Theorem 5)	$\tilde{\mathcal{O}}\left(\left(\frac{S^2 O A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{S O^{1/2} A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Theorem 6)

Omitting H, assuming  $O \geq SA$  in upper bounds

- First multiplicative dependence on  $S, O, A, 1/\alpha, 1/\varepsilon$
- Suggests our  $1/\alpha^2$  is sharp dependence on  $\alpha$
- First joint dependence on  $O$  and  $1/(\alpha\varepsilon)$

# Lower Bounds for Revealing POMDPs

[Chen, Wang, Xiong, Mei, Bai '23]

Problem	PAC sample complexity		Regret	
	Upper bound	Lower bound	Upper bound	Lower bound
1-step $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{S O^{1/2} A}{\alpha^2 \varepsilon^2}\right)$ (Theorem 4)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2 O^2 A}{\alpha^2} \cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{S O^{1/2} A}{\alpha^2} \cdot T}\right)$ (Corollary 7)
$m$ -step ( $m \geq 2$ ) $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A^m}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{(S^{3/2} + S A) O^{1/2} A^{m-1}}{\alpha^2 \varepsilon^2}\right)$ (Theorem 5)	$\tilde{\mathcal{O}}\left(\left(\frac{S^2 O A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{S O^{1/2} A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Theorem 6)

Omitting H, assuming  $O \geq SA$  in upper bounds

- First multiplicative dependence on  $S, O, A, 1/\alpha, 1/\varepsilon$
- Suggests our  $1/\alpha^2$  is sharp dependence on  $\alpha$
- First joint dependence on  $O$  and  $1/(\alpha\varepsilon)$
- Regret is  $\Omega(T^{2/3})$  for  $m$ -step revealing, whereas  $\tilde{\mathcal{O}}(\sqrt{T})$  for 1-step

# Lower Bounds for Revealing POMDPs

[Chen, Wang, Xiong, Mei, Bai '23]

Problem	PAC sample complexity		Regret	
	Upper bound	Lower bound	Upper bound	Lower bound
1-step $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{S O^{1/2} A}{\alpha^2 \varepsilon^2}\right)$ (Theorem 4)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2 O^2 A}{\alpha^2} \cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{S O^{1/2} A}{\alpha^2} \cdot T}\right)$ (Corollary 7)
$m$ -step ( $m \geq 2$ ) $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A^m}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{(S^{3/2} + S A) O^{1/2} A^{m-1}}{\alpha^2 \varepsilon^2}\right)$ (Theorem 5)	$\tilde{\mathcal{O}}\left(\left(\frac{S^2 O A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{S O^{1/2} A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Theorem 6)

Omitting H, assuming  $O \geq SA$  in upper bounds

- First multiplicative dependence on  $S, O, A, 1/\alpha, 1/\varepsilon$
- Suggests our  $1/\alpha^2$  is sharp dependence on  $\alpha$
- First joint dependence on  $O$  and  $1/(\alpha\varepsilon)$
- Regret is  $\underline{\Omega}(T^{2/3})$  for  $m$ -step revealing, whereas  $\tilde{\mathcal{O}}(\sqrt{T})$  for 1-step
- ...

# Lower Bounds for Revealing POMDPs

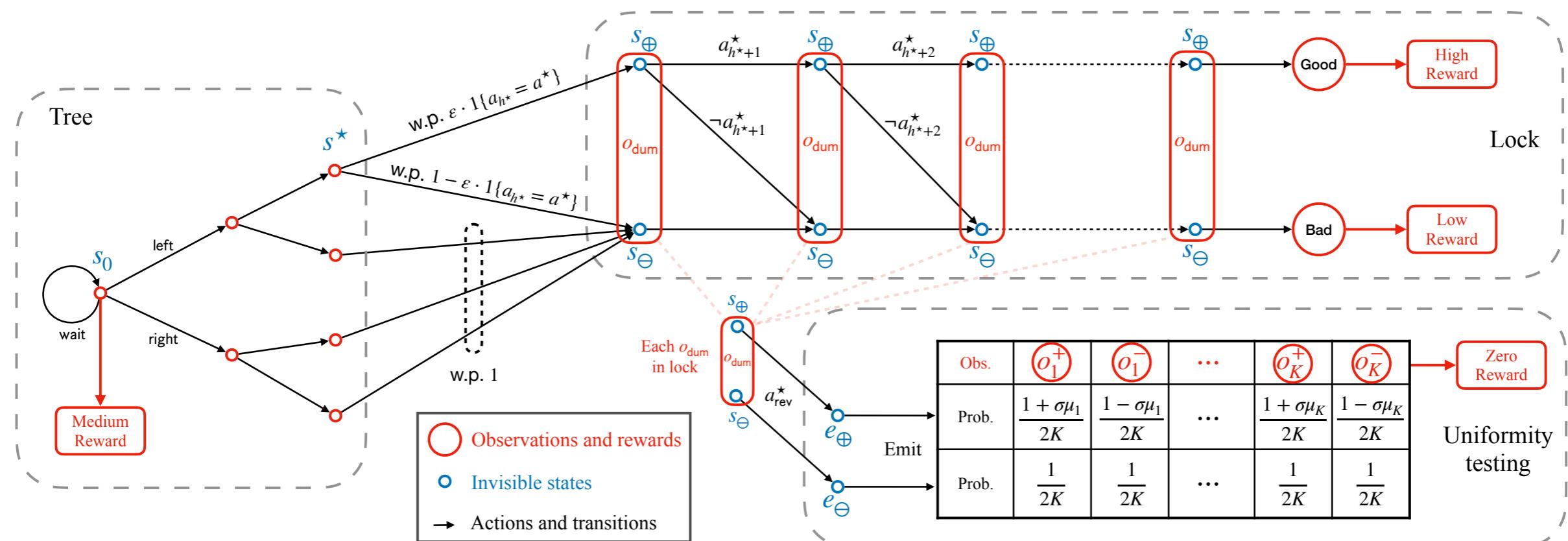
[Chen, Wang, Xiong, Mei, Bai '23]

Problem	PAC sample complexity		Regret	
	Upper bound	Lower bound	Upper bound	Lower bound
1-step $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{S O^{1/2} A}{\alpha^2 \varepsilon^2}\right)$ (Theorem 4)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2 O^2 A}{\alpha^2} \cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{S O^{1/2} A}{\alpha^2} \cdot T}\right)$ (Corollary 7)
$m$ -step ( $m \geq 2$ ) $\alpha$ -revealing	$\tilde{\mathcal{O}}\left(\frac{S^2 O A^m}{\alpha^2 \varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{(S^{3/2} + S A) O^{1/2} A^{m-1}}{\alpha^2 \varepsilon^2}\right)$ (Theorem 5)	$\tilde{\mathcal{O}}\left(\left(\frac{S^2 O A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{S O^{1/2} A^m}{\alpha^2}\right)^{1/3} T^{2/3}\right)$ (Theorem 6)

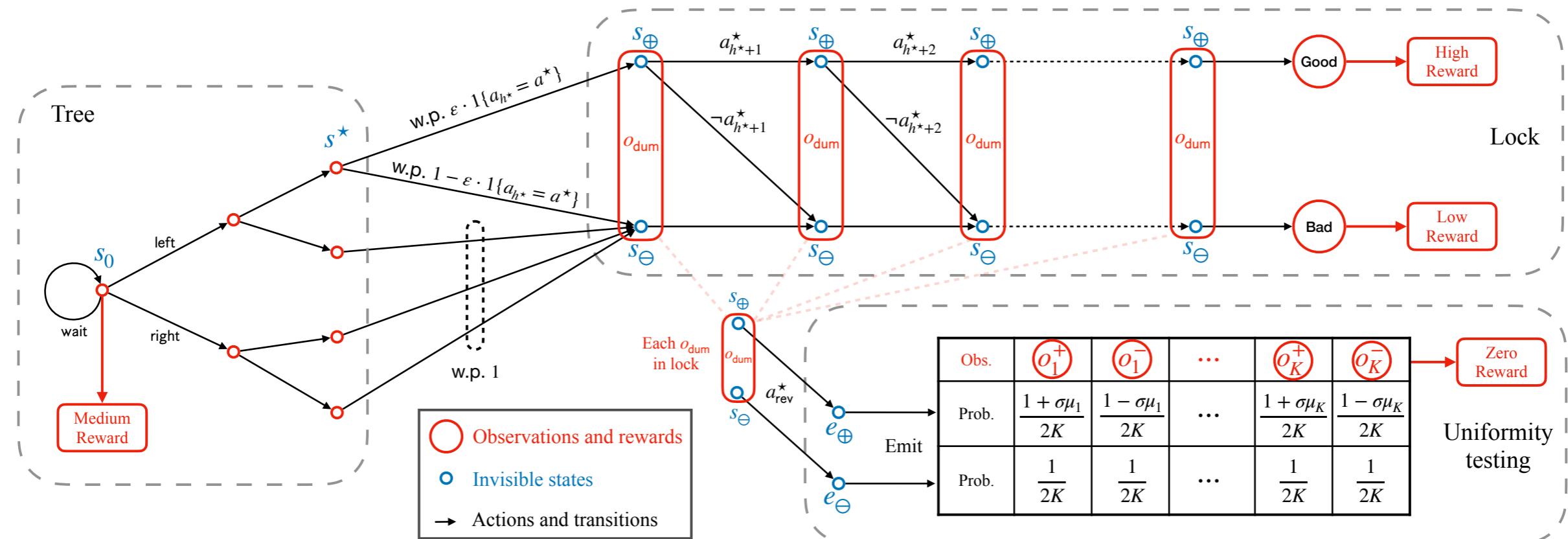
Omitting H, assuming  $O \geq SA$  in upper bounds

- First multiplicative dependence on  $S, O, A, 1/\alpha, 1/\varepsilon$
- Suggests our  $1/\alpha^2$  is sharp dependence on  $\alpha$
- First joint dependence on  $O$  and  $1/(\alpha\varepsilon)$
- Regret is  $\underline{\Omega}(T^{2/3})$  for  $m$ -step revealing, whereas  $\tilde{\mathcal{O}}(\sqrt{T})$  for 1-step
- ...
- Gap is only  $\sqrt{SO}$  in  $(S, O)$  for  $m$ -step revealing

# Hard instance construction (2-step case, simplified)

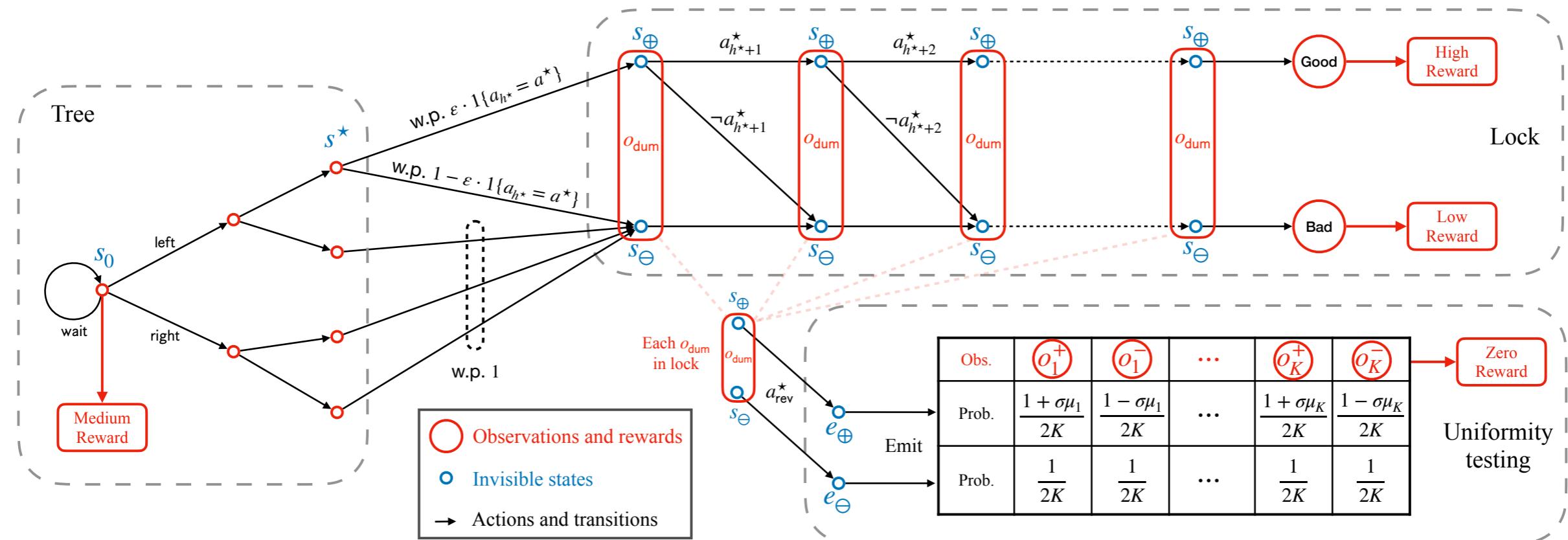


# Hard instance construction (2-step case, simplified)



Building blocks

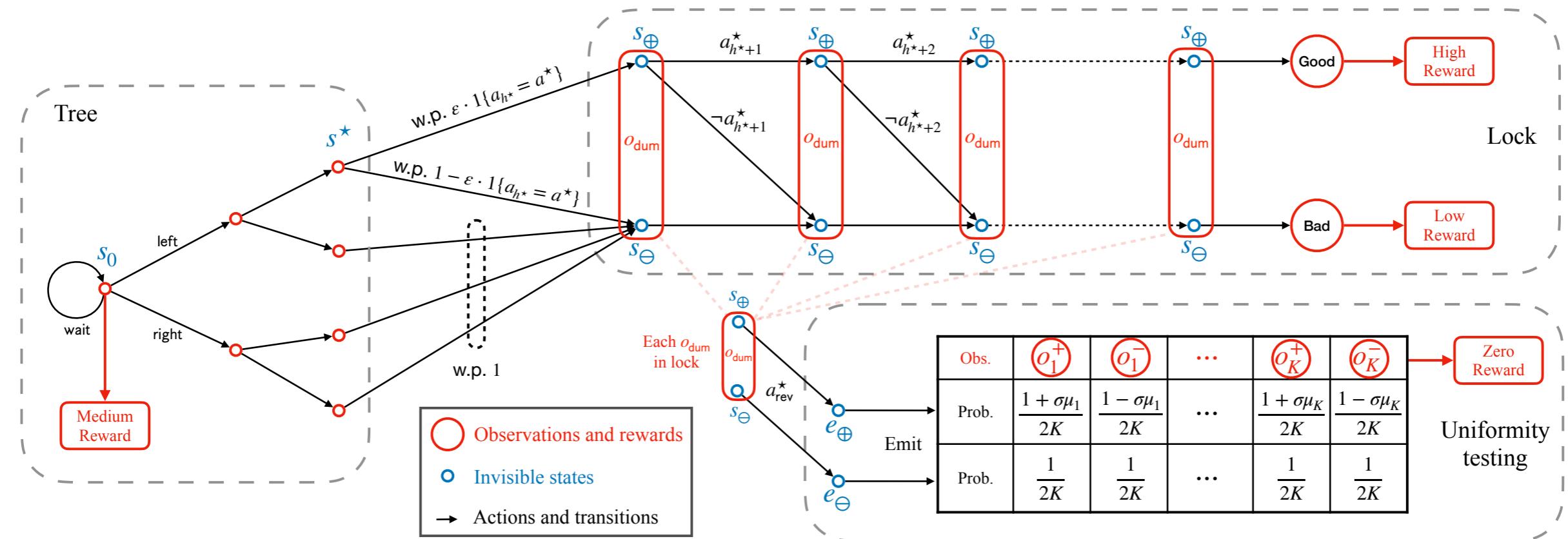
# Hard instance construction (2-step case, simplified)



## Building blocks

- Tree-MDP to obtain *HSA* factor [Domingues et al. '21]

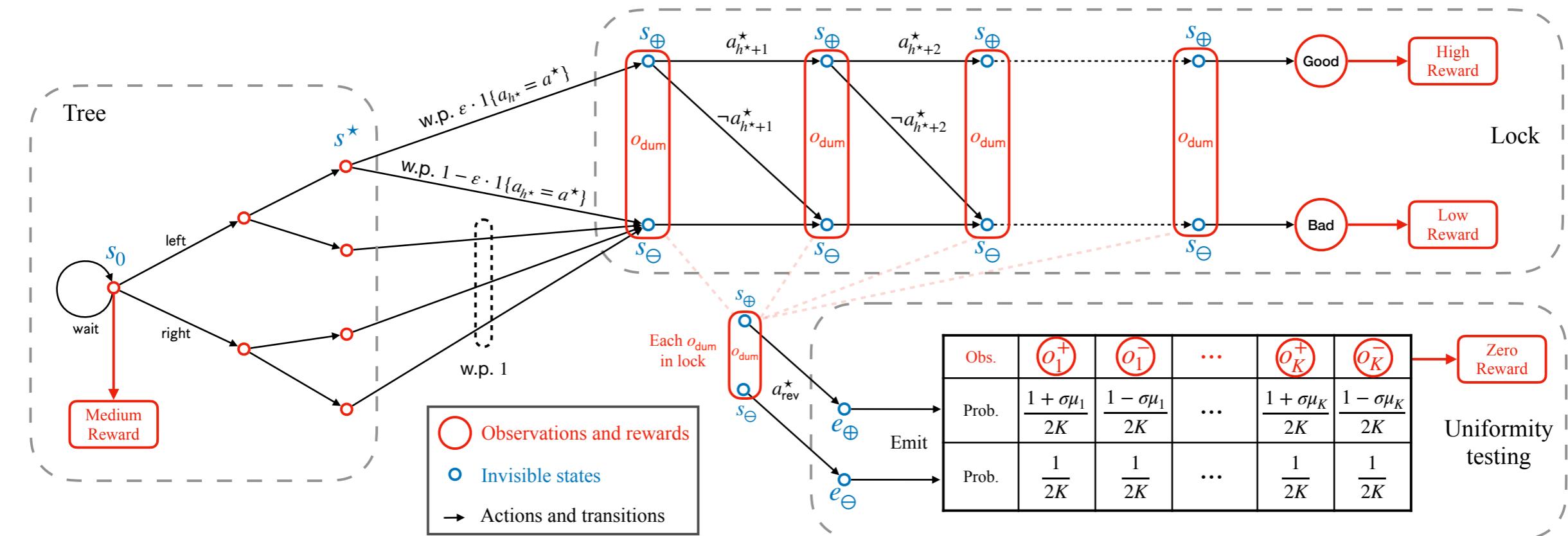
# Hard instance construction (2-step case, simplified)



## Building blocks

- Tree-MDP to obtain *HSA* factor [Domingues et al. '21]
- 2-step revealing combination lock to force exploration with revealing mechanism

# Hard instance construction (2-step case, simplified)



## Building blocks

- Tree-MDP to obtain  $HSA$  factor [Domingues et al. '21]
- 2-step revealing combination lock to force exploration with revealing mechanism
- Uniformity testing for constructing hard-to-distinguish distributions over  $[O]$ , and obtain  $\sqrt{O}/(\alpha^2\epsilon^2)$  factor in lower bound [Paninski '08, Diakonikolas et al. '14, ...]

# Summary

# Summary

We provide

- New unified condition (B-stability) for tractable learning in POMDP/PSRs
- 3 algorithms (OMLE, E2D, posterior sampling)
- Sharp rates via unified analysis (B-stability + L2 Eluder argument)
- Lower bounds for revealing POMDPs

# Summary

We provide

- New unified condition (B-stability) for tractable learning in POMDP/PSRs
- 3 algorithms (OMLE, E2D, posterior sampling)
- Sharp rates via unified analysis (B-stability + L2 Eluder argument)
- Lower bounds for revealing POMDPs

## Future directions

# Summary

We provide

- New unified condition (B-stability) for tractable learning in POMDP/PSRs
- 3 algorithms (OMLE, E2D, posterior sampling)
- Sharp rates via unified analysis (B-stability + L2 Eluder argument)
- Lower bounds for revealing POMDPs

## Future directions

- Alternative algorithms (value-based?)

# Summary

We provide

- New unified condition (B-stability) for tractable learning in POMDP/PSRs
- 3 algorithms (OMLE, E2D, posterior sampling)
- Sharp rates via unified analysis (B-stability + L2 Eluder argument)
- Lower bounds for revealing POMDPs

## Future directions

- Alternative algorithms (value-based?)
- Sharper rates for tabular revealing POMDPs

# Summary

We provide

- New unified condition (B-stability) for tractable learning in POMDP/PSRs
- 3 algorithms (OMLE, E2D, posterior sampling)
- Sharp rates via unified analysis (B-stability + L2 Eluder argument)
- Lower bounds for revealing POMDPs

## Future directions

- Alternative algorithms (value-based?)
- Sharper rates for tabular revealing POMDPs
- Other tractable subclasses beyond revealing/decodable?

# Summary

We provide

- New unified condition (B-stability) for tractable learning in POMDP/PSRs
- 3 algorithms (OMLE, E2D, posterior sampling)
- Sharp rates via unified analysis (B-stability + L2 Eluder argument)
- Lower bounds for revealing POMDPs

## Future directions

- Alternative algorithms (value-based?)
- Sharper rates for tabular revealing POMDPs
- Other tractable subclasses beyond revealing/decodable?

## Thank you!

*Partially Observable RL with B-Stability: Unified Structural Condition and Sharp Sample-Efficient Algorithms.*  
Fan Chen, Yu Bai, Song Mei. ICLR 2023 (spotlight). <https://arxiv.org/abs/2209.14990>

*Lower Bounds for Learning in Revealing POMDPs.*

Fan Chen, Huan Wang, Caiming Xiong, Song Mei, Yu Bai, 2023. <https://arxiv.org/abs/2302.01333>

# Backup Slides

# B-Stability

For any PSR with an associated B-representation, we define its  $\mathcal{B}$ -operators  $\{\mathcal{B}_{H:h}\}_{h \in [H]}$  as

$$\mathcal{B}_{H:h} : \mathbb{R}^{\mathcal{U}_h} \rightarrow \mathbb{R}^{(\mathcal{O} \times \mathcal{A})^{H-h+1}}, \quad \mathbf{q} \mapsto [\mathbf{B}_{H:h}(\tau_{h:H}) \cdot \mathbf{q}]_{\tau_{h:H} \in (\mathcal{O} \times \mathcal{A})^{H-h+1}}.$$

Operator  $\mathcal{B}_{H:h}$  maps any predictive state  $\mathbf{q} = \mathbf{q}(\tau_{h-1})$  at step  $h$  to the vector  $\mathcal{B}_{H:h}\mathbf{q} = (\mathbb{P}(\tau_{h:H} | \tau_{h-1}))_{\tau_{h:H}}$  which governs the probability of transitioning to all possible futures, by properties of the B-representation (cf. (18) & Corollary B.2). For each  $h \in [H]$ , we equip the image space of  $\mathcal{B}_{H:h}$  with the  $\Pi$ -norm: For a vector  $\mathbf{b}$  indexed by  $\tau_{h:H} \in (\mathcal{O} \times \mathcal{A})^{H-h+1}$ , we define

$$\|\mathbf{b}\|_{\Pi} := \max_{\bar{\pi}} \sum_{\tau_{h:H} \in (\mathcal{O} \times \mathcal{A})^{H-h+1}} \bar{\pi}(\tau_{h:H}) \mathbf{b}(\tau_{h:H}), \quad (3)$$

where the maximization is over all policies  $\bar{\pi}$  starting from step  $h$  (ignoring the history  $\tau_{h-1}$ ) and  $\bar{\pi}(\tau_{h:H}) = \prod_{h' \leq h' \leq H} \bar{\pi}_{h'}(a_{h'} | o_{h'}, \tau_{h:h'-1})$ . We further equip the domain  $\mathbb{R}^{\mathcal{U}_h}$  with a *fused-norm*  $\|\cdot\|_*$ , which is defined as the maximum of  $(1, 2)$ -norm and  $\Pi'$ -norm<sup>5</sup>:

$$\|\mathbf{q}\|_* := \max\{\|\mathbf{q}\|_{1,2}, \|\mathbf{q}\|_{\Pi'}\}, \quad (4)$$

$$\|\mathbf{q}\|_{1,2} := \left( \sum_{\mathbf{a} \in \mathcal{U}_{A,h}} \left( \sum_{\mathbf{o} : (\mathbf{o}, \mathbf{a}) \in \mathcal{U}_h} |\mathbf{q}(\mathbf{o}, \mathbf{a})| \right)^2 \right)^{1/2}, \quad \|\mathbf{q}\|_{\Pi'} := \max_{\bar{\pi}} \sum_{t \in \bar{\mathcal{U}}_h} \bar{\pi}(t) |\mathbf{q}(t)|, \quad (5)$$

where  $\bar{\mathcal{U}}_h := \{t \in \mathcal{U}_h : \nexists t' \in \mathcal{U}_h \text{ such that } t \text{ is a prefix of } t'\}$ .

We now define the B-stability condition, which simply requires the  $\mathcal{B}$ -operators  $\{\mathcal{B}_{H:h}\}_{h \in [H]}$  to have bounded operator norms from the fused-norm to the  $\Pi$ -norm.

**Definition 4** (B-stability). *A PSR is B-stable with parameter  $\Lambda_B \geq 1$  (henceforth also  $\Lambda_B$ -stable) if it admits a B-representation with associated  $\mathcal{B}$ -operators  $\{\mathcal{B}_{H:h}\}_{h \in [H]}$  such that*

$$\sup_{h \in [H]} \max_{\|\mathbf{q}\|_* = 1} \|\mathcal{B}_{H:h}\mathbf{q}\|_{\Pi} \leq \Lambda_B. \quad (6)$$

# B-representation for Decodable POMDPs

## B.3.5 Decodable POMDPs

To construct a B-representation for the decodable POMDP, we introduce the following notation. For  $h \leq H - m$ , we consider  $t_h = (o_h, a_h, \dots, o_{h+m-1}) \in \mathcal{U}_h$ ,  $t_{h+1} = (o'_{h+1}, a'_{h+1}, \dots, o'_{h+m}) \in \mathcal{U}_{h+1}$ , and define

$$\mathbb{P}_h(t_{h+1}|t_h) = \begin{cases} \mathbb{P}(o_{h+m} = o'_{h+m} | s_{h+m-1} = \phi_{h+m-1}(t_h), a_{h+m-1}), & \text{if } o_{h+1:h+m-1} = o'_{h+1:h+m-1} \\ & \text{and } a_{h+1:h+m-2} = a'_{h+1:h+m-2}, \\ 0, & \text{otherwise,} \end{cases} \quad (27)$$

where  $\phi_{h+m-1}$  is the decoder function that maps  $t_h$  to a latent state  $s_{h+m-1}$ . Similarly, for  $h > H - m$ ,  $t_h \in \mathcal{U}_h$ ,  $t_{h+1} \in \mathcal{U}_{h+1}$ , we let  $\mathbb{P}_h(t_{h+1}|t_h)$  be 1 if  $t_h$  ends with  $t_{h+1}$ , and 0 otherwise.

Under such definition, for all  $h \in [H]$ ,  $t_h \in \mathcal{U}_h$ ,  $t_{h+1} \in \mathcal{U}_{h+1}$ , it is clear that

$$\mathbb{P}_h(t_{h+1}|t_h) = \mathbb{P}(t_{h+1}|t_h, \tau_{h-1}) \quad (28)$$

for any reachable  $(\tau_{h-1}, t_h)$ , because of decodability. Hence, we can interpret  $\mathbb{P}_h(t_{h+1}|t_h)$  as the probability of observing  $t_{h+1}$  conditional on observing  $t_h$  on step  $h$ . <sup>18</sup> Then, for  $h \in [H]$ , we can take

$$\mathbf{B}_h(o, a) = [\mathbb{1}((o, a) \rightarrow t_h) \mathbb{P}_h(t_{h+1}|t_h)]_{(t_{h+1}, t_h) \in \mathcal{U}_{h+1} \times \mathcal{U}_h}, \quad (29)$$

where  $\mathbb{1}((o, a) \rightarrow t_h)$  is 1 if  $t_h$  starts with  $(o, a)$  and 0 otherwise<sup>19</sup>.

We verify that (29) indeed gives a B-representation for decodable POMDPs:

# B-representation for Revealing POMDPs

*Proof of Proposition C.2.* Chen et al. (2022a, Appendix B.3.3) showed that any  $m$ -step  $\alpha$ -revealing POMDP  $M$  is a  $\alpha^{-1}$ -stable PSR with core test set  $\mathcal{U}_h = (\mathcal{O} \times \mathcal{A})^{\min\{m-1, H-h\}} \times \mathcal{O}$ , and explicitly constructed the following B-representation for it: when  $h \leq H - m$ , set

$$\mathbf{B}_h(o, a) = \mathbb{M}_{h+1} \mathbb{T}_{h,a} \text{diag}(\mathbb{O}_h(o|\cdot)) \mathbb{M}_h^+, \quad h \in [H-m], \quad (12)$$

and when  $h > H - m$ , take

$$\mathbf{B}_h(o_h, a_h) = [\mathbb{1}(t_h = (o_h, a_h, t_{h+1}))]_{(t_{h+1}, t_h) \in \mathcal{U}_{h+1} \times \mathcal{U}_h} \in \mathbb{R}^{\mathcal{U}_{h+1} \times \mathcal{U}_h}, \quad (13)$$

where  $\mathbb{1}(t_h = (o_h, a_h, t_{h+1}))$  is 1 if  $t_h$  equals to  $(o_h, a_h, t_{h+1})$ , and 0 otherwise.