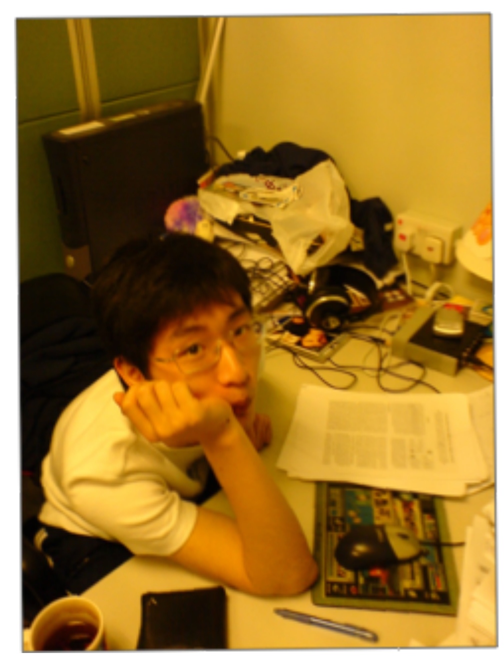
Some Recent Progresses on Partially Observable RL: B-Stability, Sharp Algorithms, and Lower Bounds

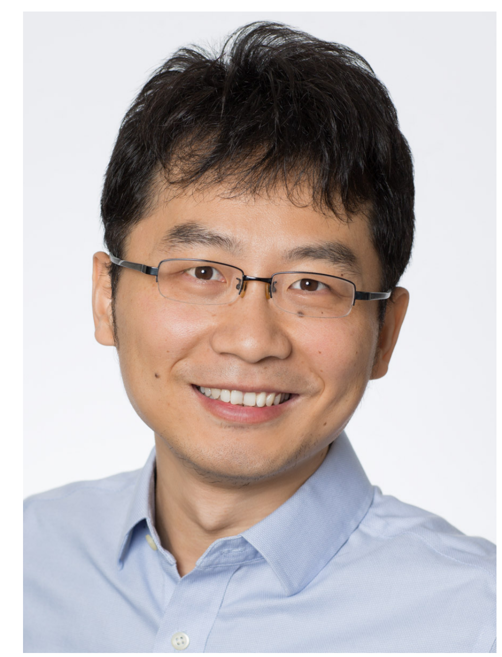
Yu Bai
Salesforce Research

Fan Chen (Peking U -> PhD)

Song Mei (UC Berkeley)

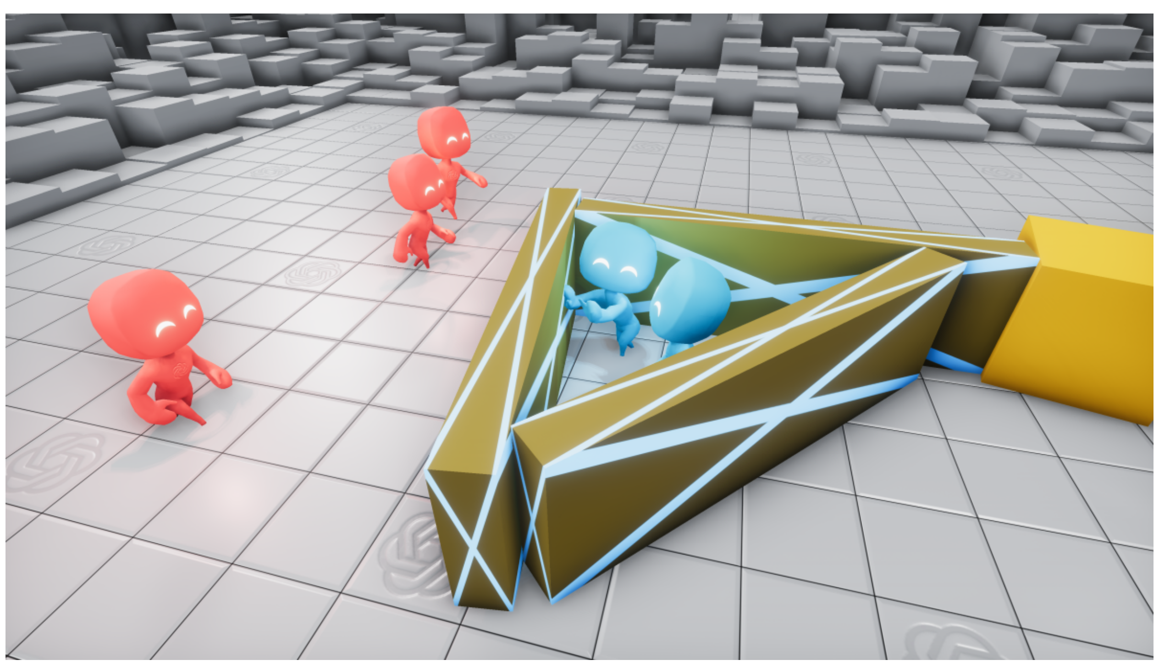


Huan Wang (Salesforce)



Caiming Xiong (Salesforce)

Partial Observability in Reinforcement Learning



Partially Observable Markov Decision Processes (POMDPs)

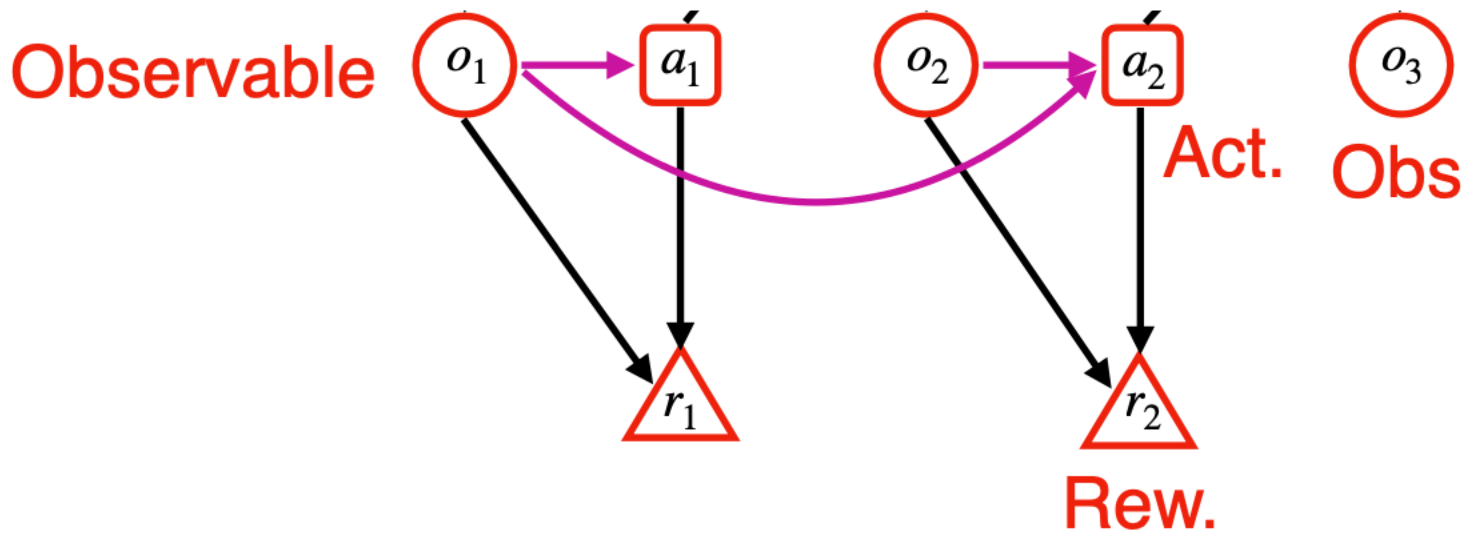


Figure credit: Song Mei

Partially Observable Markov Decision Processes (POMDPs)

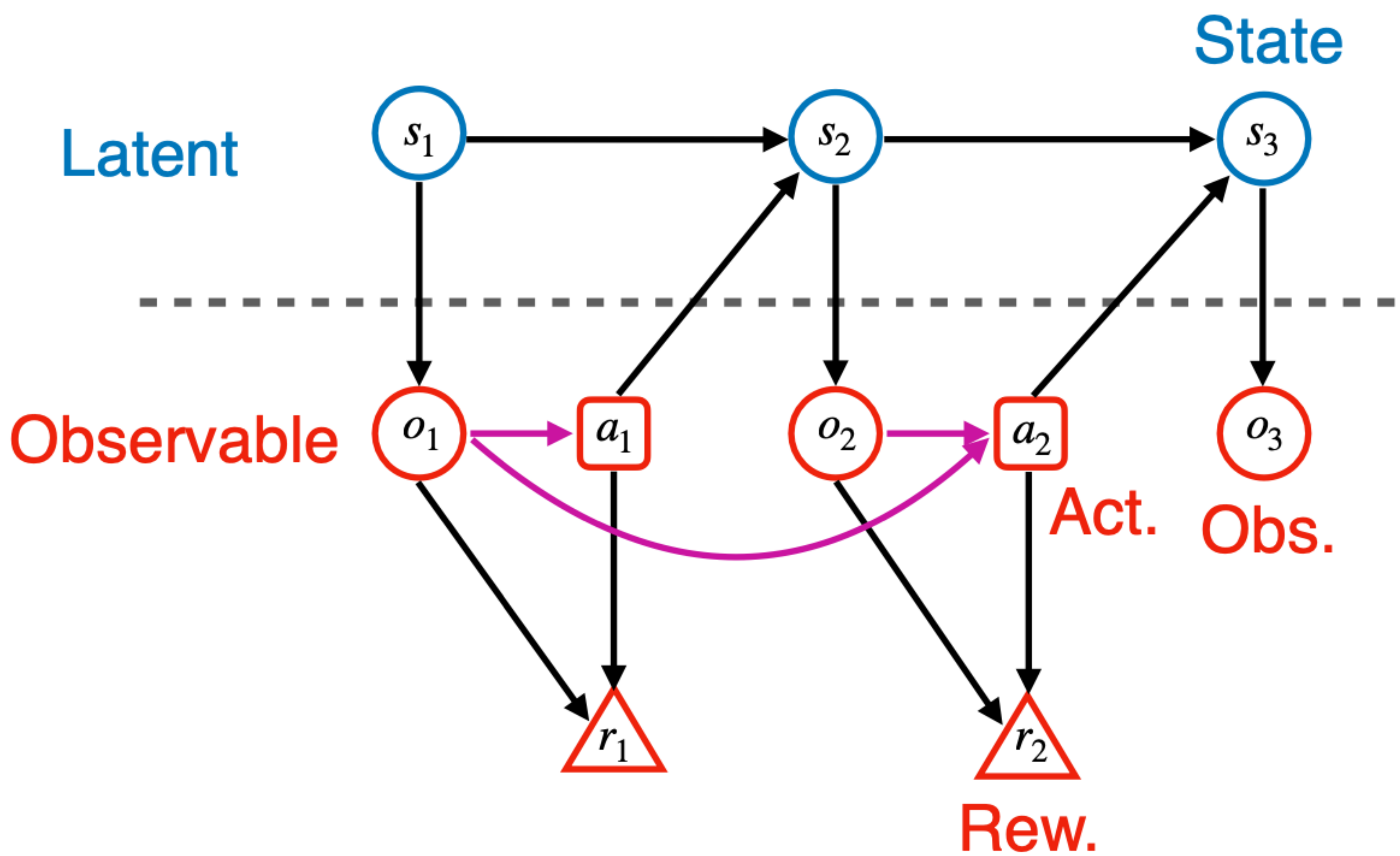
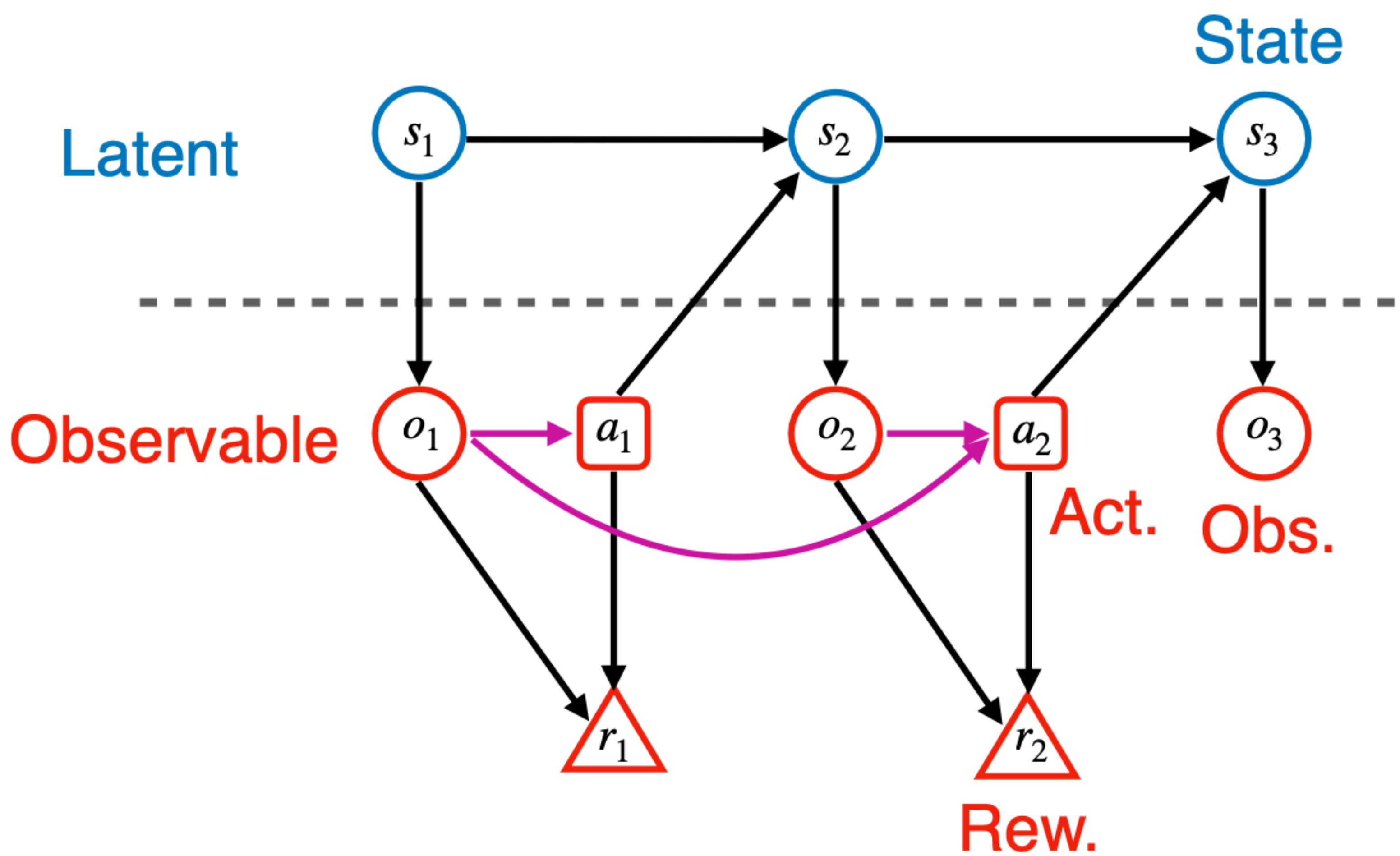


Figure credit: Song Mei

Partially Observable Markov Decision Processes (POMDPs)



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

Figure credit: Song Mei

Challenge for Learning in POMDPs

Challenge for Learning in POMDPs

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

ε-optimal policy can be found in $\mathrm{poly}(H, S, A, 1/\varepsilon)$ *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):

ε-optimal policy can be found in $\text{poly}(H, S, A, 1/\varepsilon)$ *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 [Vlassis 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 [Vlassis et al. '12]

Statistical hardness   (with ∞ 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 [Vlassis et al. '12]

Statistical hardness   (with ∞ 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 [Vlassis et al. '12]

Statistical hardness (with ∞ 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 [Vlassis et al. '12]

Statistical hardness (with ∞ compute):

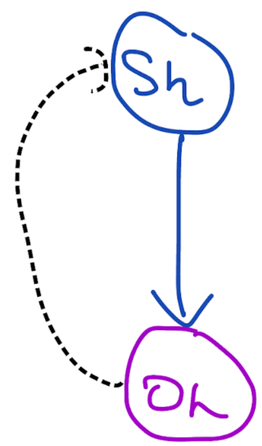
- Learning optimal policy requires $\exp(\Omega(H))$ samples in the worst-case [Krishnamurthy et al. '16]
- Hard instance: "non-revealing combination lock" with "dummy observations"

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

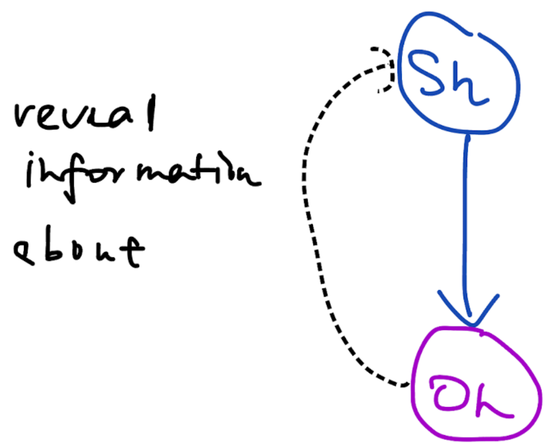
Tractable Subclasses of POMDPs

Example 1: Revealing (Observable) POMDPs

reveal
information
about



Example 1: Revealing (Observable) POMDPs



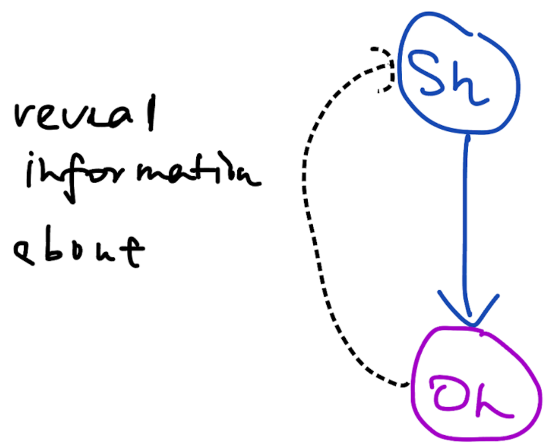
Emission matrix:

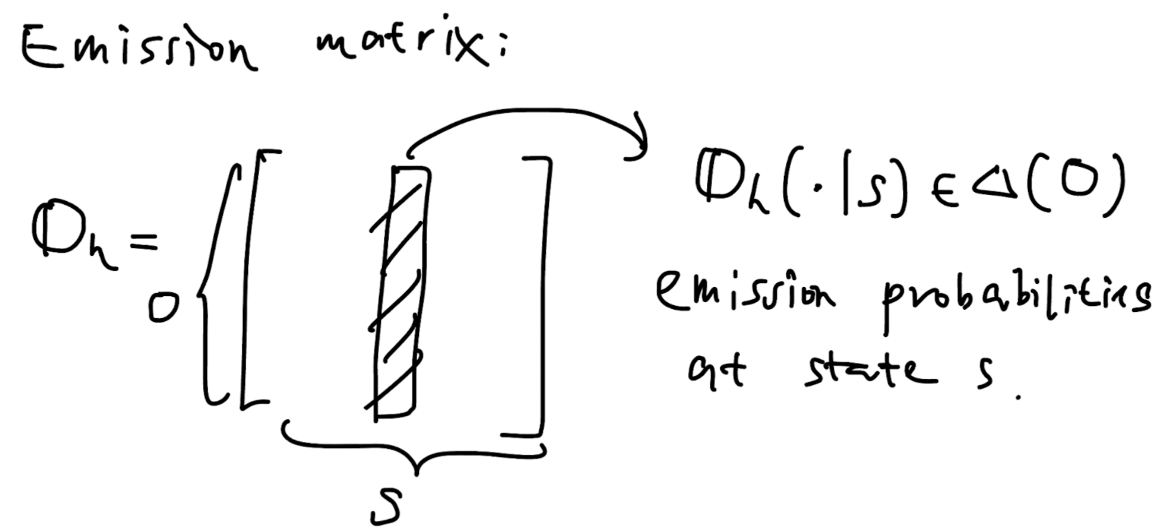
$$\mathbb{O}_h = \underbrace{\left.\begin{bmatrix} & \cdots & \end{bmatrix}\right\}}_{S} O$$

$\mathbb{O}_h(\cdot \mid s) \in \Delta(O)$

emission probabilities at state s.

Example 1: Revealing (Observable) POMDPs

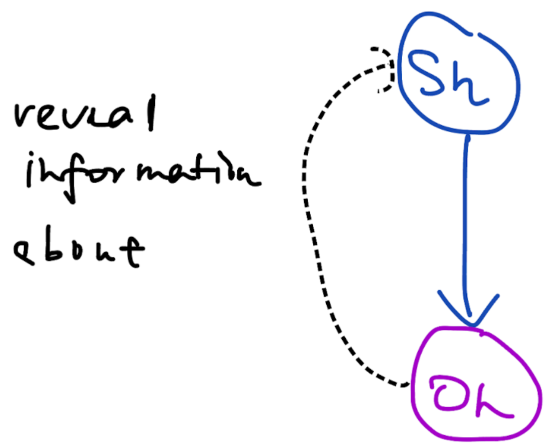




Desire: Emission matrices have full column rank ==> different states are probabilistically distinguishable from their emitted observations

Rules out “uninformative” observations.

Example 1: Revealing (Observable) POMDPs



Emission matrix:

$$\mathbb{O}_h = \begin{array}{c} \\ O \end{array} \left\{ \underbrace{\begin{bmatrix} \quad & \quad & \quad \\ & & \\ & & \end{bmatrix}}_{S} \right.$$

$\mathbb{O}_h(\cdot|s) \in \Delta(O)$

emission probabilities at state s.

Desire: Emission matrices have full column rank ==> different states are probabilistically distinguishable from their emitted observations

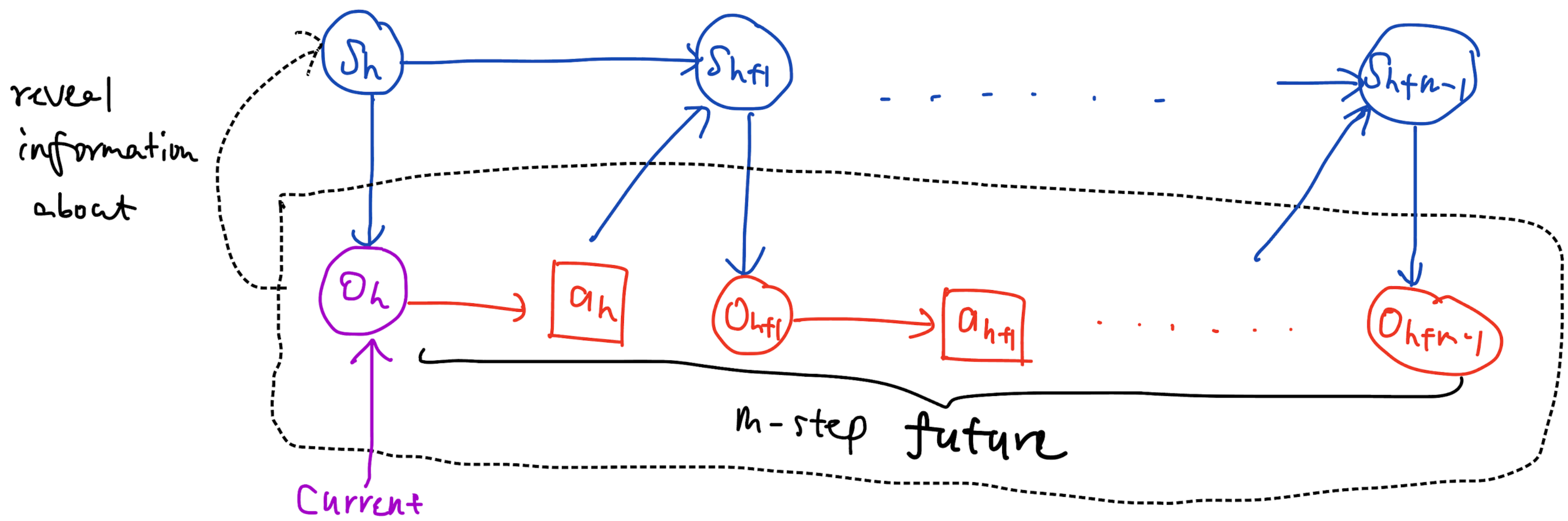
Rules out "uninformative" observations.

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

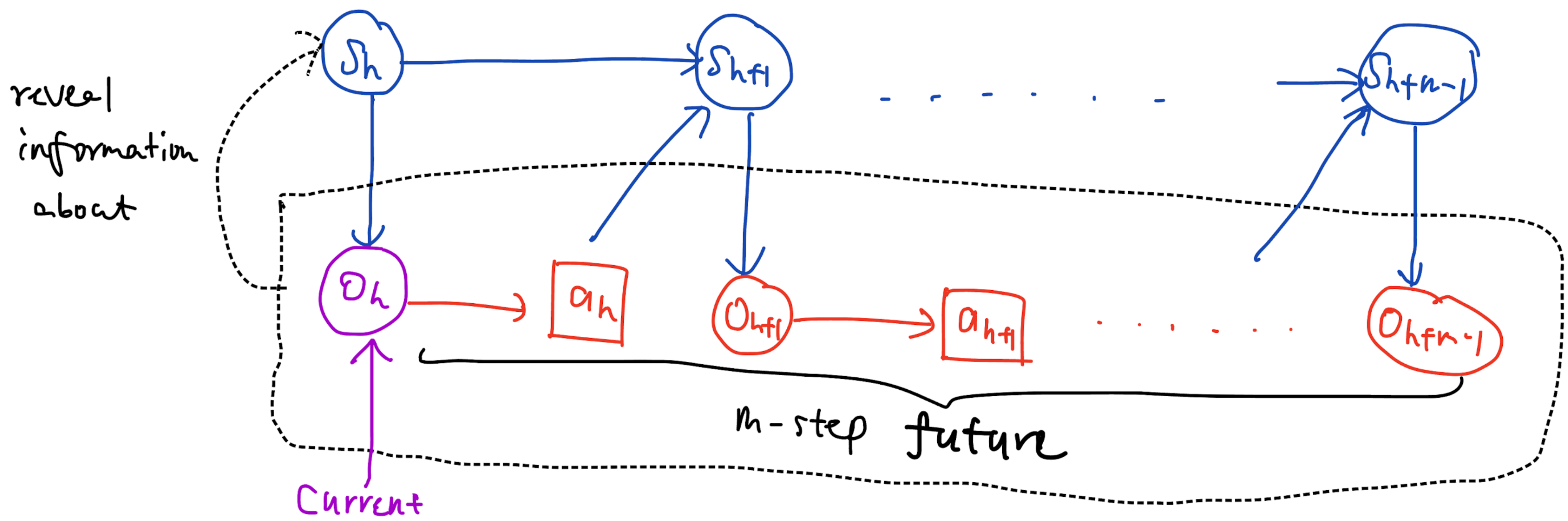
$$\|\mathbb{O}_h^+\| \leq \alpha^{-1},$$

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

Multi-step Revealing POMDPs



Multi-step Revealing POMDPs

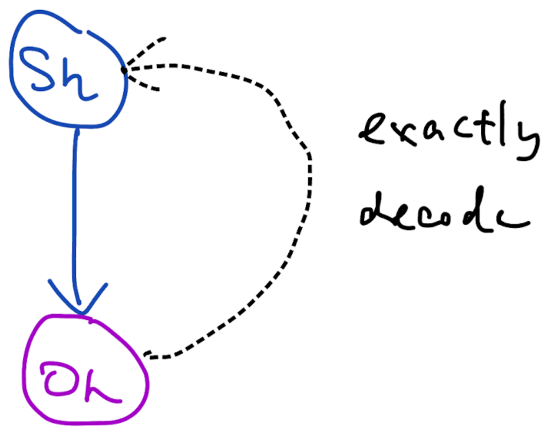


m-step α-revealing POMDPs [Liu et al. '22a]: The m-step emission-action matrices at all step $h \in [H]$ satisfy

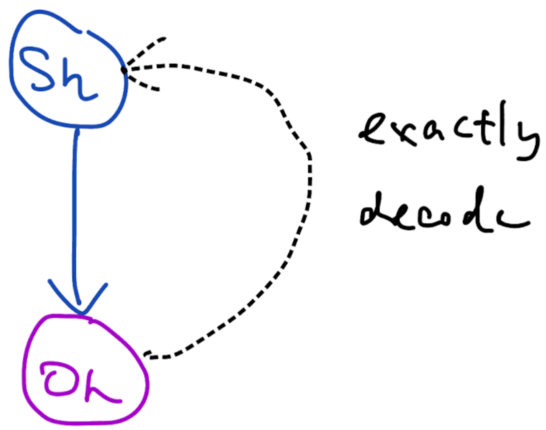
$$\|\mathbb{M}_{h,m}^{+}\| \leq \alpha^{-1},$$

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

Example 2: Decodable POMDPs

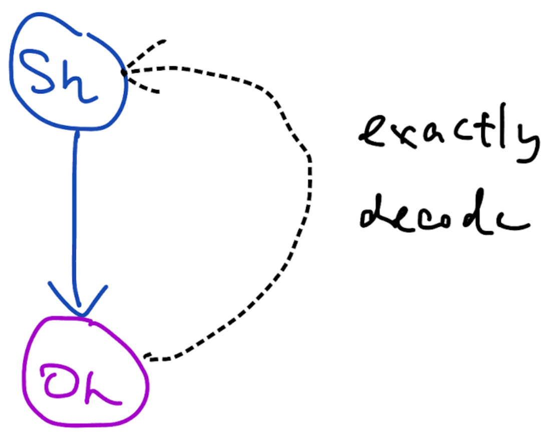


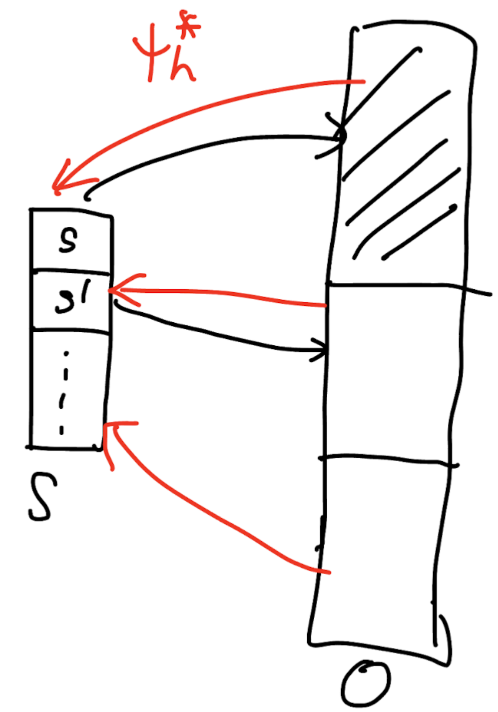
Example 2: Decodable POMDPs



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

Example 2: Decodable POMDPs



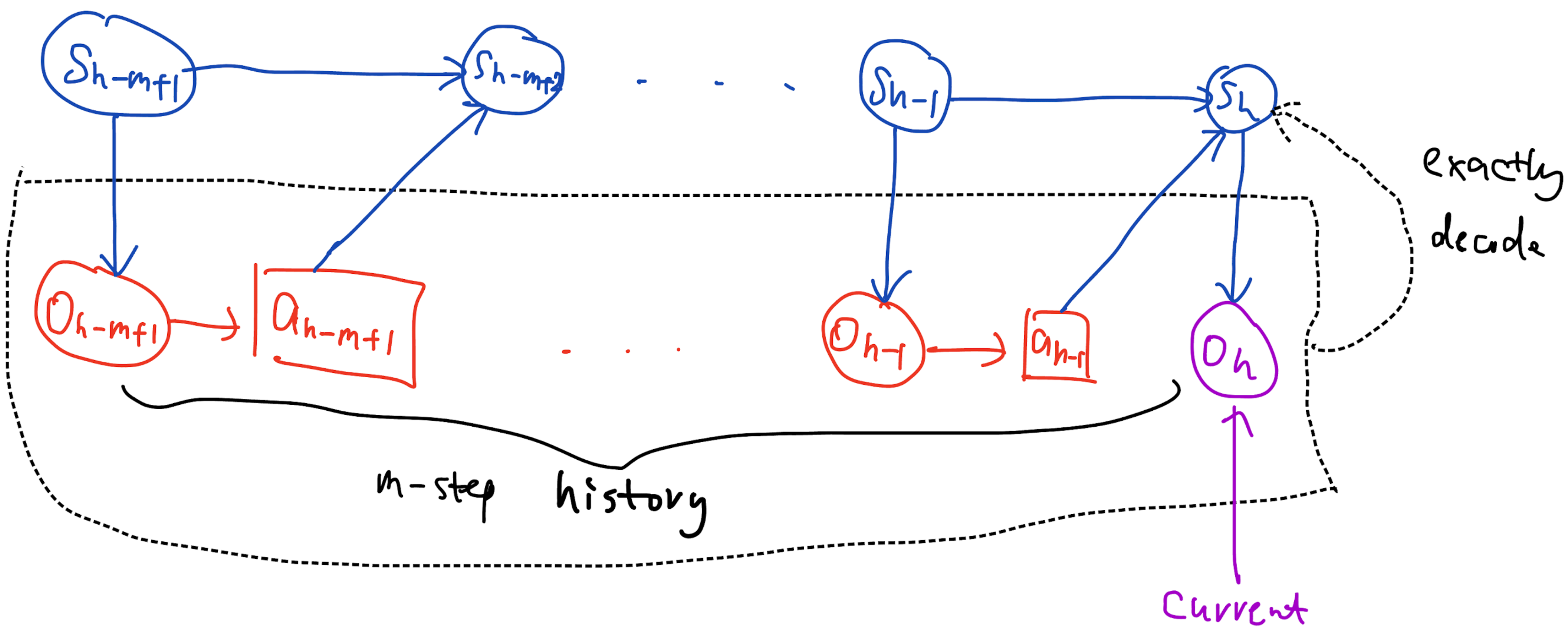


Desire: Latent state can be uniquely determined from the observation.

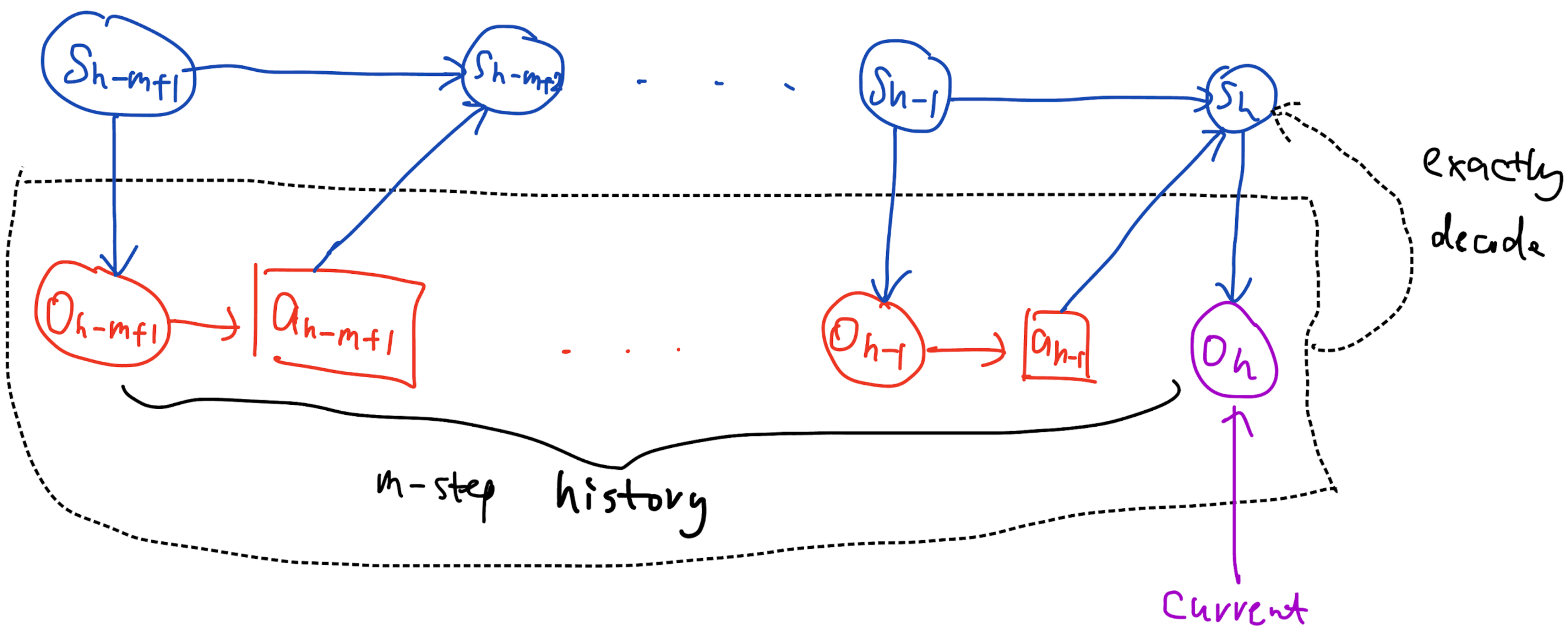
Block MDPs [Du et al. '19]: There exists an (unknown) decoder $\psi_h^\star$ at every step $h \in [H]$ such that

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

Multi-Step Decodable POMDPs



Multi-Step Decodable POMDPs



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

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

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 μ_1 such that for any trajectory τ, policy π,

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

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 μ_1 such that for any trajectory τ, policy π,

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

Above,

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

Example: Single-step revealing POMDPs

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

B-Representation of POMDPs

[Jaeger '00]

B-representation: *Any* set of matrices $\{B_h(o,a)\}_{h,o,a}$ and vector μ_1 such that for any trajectory τ, policy π,

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

Above,

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

Example: Single-step revealing POMDPs

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

OxS emission matrix

B-Representation of POMDPs

[Jaeger '00]

B-representation: *Any* set of matrices $\{B_h(o,a)\}_{h,o,a}$ and vector μ_1 such that for any trajectory τ, policy π,

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

Above,

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

Example: Single-step revealing POMDPs

SxS latent transition matrix

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

OxS emission matrix

B-Representation of POMDPs

[Jaeger '00]

B-representation: *Any* set of matrices $\{B_h(o,a)\}_{h,o,a}$ and vector μ_1 such that for any trajectory τ, policy π,

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

Above,

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

Example: Single-step revealing POMDPs

SxS latent transition matrix

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

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 μ_1 such that for any trajectory τ, policy π,

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

Above,

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

Example: Single-step revealing POMDPs

SxS latent transition matrix

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

OxS emission matrix

SxS diagonal matrix of emission probabilities of o_h

To verify, for any fixed h,

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

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 μ_1 such that for any trajectory τ, policy π,

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

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 μ_1 such that for any trajectory τ, policy π,

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

Above,

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

More Generally,

B-representation <==> 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 μ_1 such that for any trajectory τ, policy π,

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

Above,

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

More Generally,

B-representation <==> 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 $h \in [H]$,

$$\|\mathscr{B}_{H:h}\|_{*\to\Pi} \le \Lambda_B,$$

where operator

$$\mathscr{B}_{H:h} : q \to [B_{H:h}(\tau_{H:h})q]_{\tau_{h:H}} = \left[B_H(o_H, a_H)\ldots B_h(o_h, a_h)q\right]_{(oa)_{h:H}}$$

B-Stability Condition

[Chen, **Bai**, Mei '22]

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

$$\|\mathscr{B}_{H:h}\|_{\underbrace{*\to\Pi}_{\text{A certain operator norm}}} \leq \Lambda_B,$$

where operator

$$\mathscr{B}_{H:h} : q \to [B_{H:h}(\tau_{H:h})q]_{\tau_{h:H}} = \left[B_H(o_H, a_H)\ldots B_h(o_h, a_h)q\right]_{(oa)_{h:H}}$$

B-Stability Condition

[Chen, **Bai**, Mei '22]

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

$$\|\mathscr{B}_{H:h}\|_{\underbrace{*\to\Pi}_{\text{A certain operator norm}}} \leq \Lambda_B,$$

where operator

$$\mathscr{B}_{H:h} : q \to [B_{H:h}(\tau_{H:h})q]_{\tau_{h:H}} = \left[B_H(o_H, a_H)\ldots B_h(o_h, a_h)q\right]_{(oa)_{h:H}}$$

Intuition:

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

B-Stability Condition

[Chen, **Bai**, Mei '22]

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

$$\|\mathscr{B}_{H:h}\|_{\underbrace{*\to\Pi}_{\text{A certain operator norm}}} \leq \Lambda_B,$$

where operator

$$\mathscr{B}_{H:h} : q \to [B_{H:h}(\tau_{H:h})q]_{\tau_{h:H}} = \left[B_H(o_H, a_H)\ldots B_h(o_h, a_h)q\right]_{(oa)_{h:H}}$$

Intuition:

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

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 $h \in [H]$,

$$\|\mathscr{B}_{H:h}\|_{\underline{*\to\Pi}} \leq \Lambda_B,$$

A certain operator norm

where operator

$$\mathscr{B}_{H:h} : q \to [B_{H:h}(\tau_{H:h})q]_{\tau_{h:H}} = \left[B_H(o_H, a_H)\ldots B_h(o_h, a_h)q\right]_{(oa)_{h:H}}$$

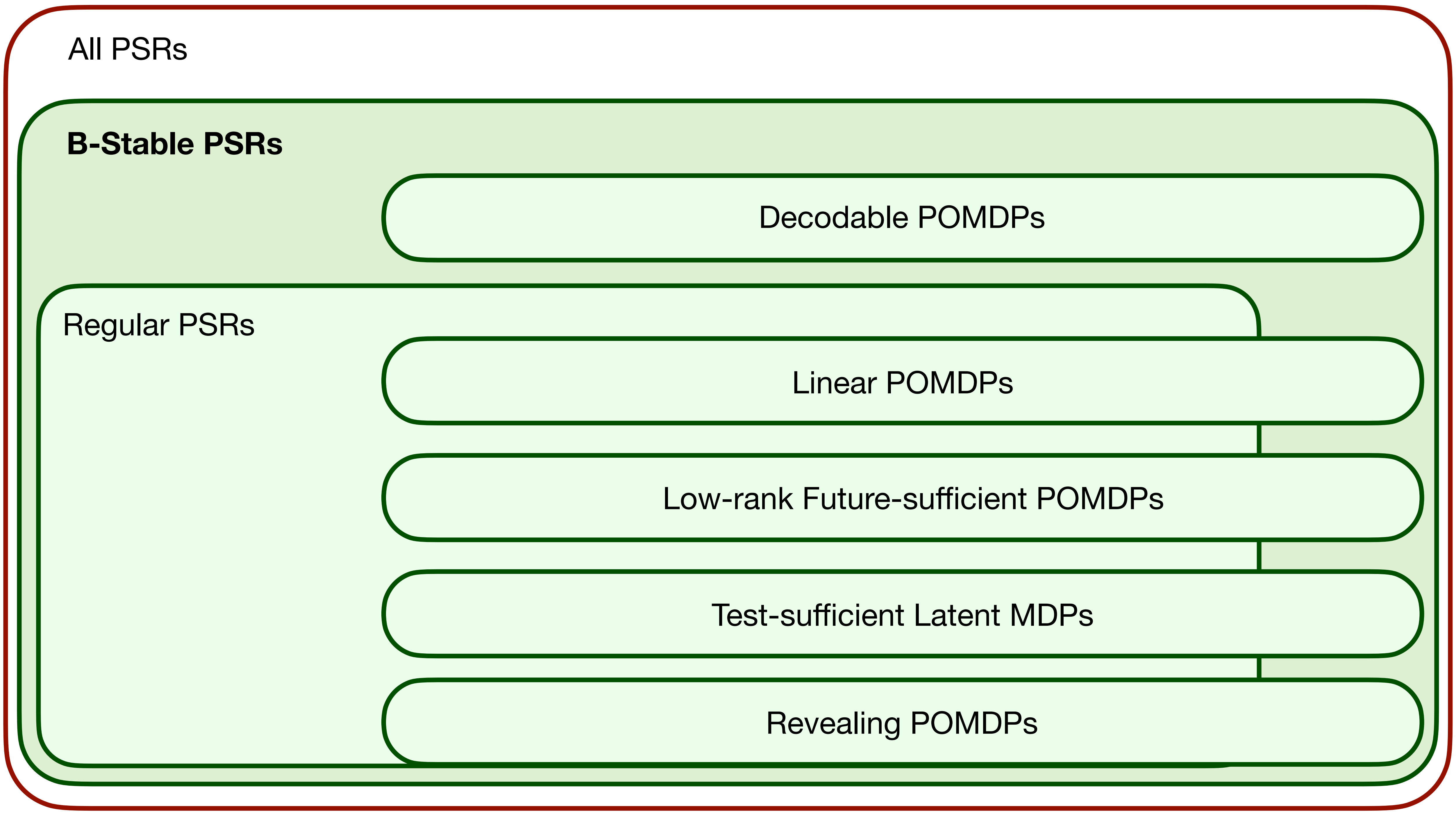
Intuition:

Estimation error of B matrices
(Algorithm can bound)

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

Error from performance difference
(for bounding Regret/PAC)

Landscape of POMDP/PSRs



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

Algorithm 1: OMLE

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

In each iteration k,

1. Set $\pi^k = \arg\max_{\pi} \max_{\theta \in \mathscr{B}^k} V_\theta^\pi$ to be optimistic greedy policy wrt $\mathscr{B}^k$
2. Play corresponding "exploration policies" $\pi^k_{h,\exp} = \Pi_{h,\exp}(\pi^k)$

Algorithm 1: OMLE

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

In each iteration k,

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

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

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* $(p_{\mathrm{exp}}^k, p_{\mathrm{out}}^k)$ 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* $(p^k_{\mathrm{exp}}, p^k_{\mathrm{out}})$ to minimize risk $V^{\mu^k}(\,\cdot\,,\,\cdot\,)$
2. Sample and play exploration policy $\pi^k \sim p^k_{\mathrm{exp}}$, obtain trajectory τ^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* $(p_{\mathrm{exp}}^k, p_{\mathrm{out}}^k)$ to minimize risk $V^{\mu^k}(\,\cdot\,,\cdot\,)$
2. Sample and play exploration policy $\pi^k \sim p_{\mathrm{exp}}^k$, obtain trajectory τ^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 τ^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 τ^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 $(\widetilde{\mathbb{P}}, \Theta_0)$
2: **Initialize:** $\mu^1 = \mathrm{Unif}(\Theta_0)$
3: **for** $t = 1, \ldots, T$ **do**
4: Sample $\theta^t \sim \mu^t$ and $h^t \sim \mathrm{Unif}(\{0, 1, \cdots, H-1\})$.
5: Set $\pi^t = \pi_{\theta^t} \circ_{h^t} \mathrm{Unif}(\mathcal{A}) \circ_{h^t+1} \mathrm{Unif}(\mathcal{U}_{A,h+1})$, execute π^t and observe τ^t.
6: Compute $\mu^{t+1} \in \Delta(\Theta_0)$ by

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

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

Similar as E2D, except for using optimistic posterior.

Main Result for Learning B-Stable POMDP/PSRs

Thm [Chen, **Bai**, Mei '22a]: Algorithms {OMLE, E2D, MOPS} can all learn a Λ_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, **Bai**, Mei '22a]: Algorithms {OMLE, E2D, MOPS} can all learn a Λ_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, **Bai**, Mei '22a]: Algorithms {OMLE, E2D, MOPS} can all learn a Λ_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, **Bai**, Mei '22a]: Algorithms {OMLE, E2D, MOPS} can all learn a Λ_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, **Bai**, Mei '22a]: Algorithms {OMLE, E2D, MOPS} can all learn a Λ_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, **Bai**, Mei '22a]: Algorithms {OMLE, E2D, MOPS} can all learn a Λ_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 Λ_B^2 rate (previous works at least Λ_B^4 on their stability/regularity parameters)

Instantiations to Concrete Subclasses

Table 1: **Comparisons of sample complexities** for learning an ε 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 α_{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[1]. All results are scaled to the setting with total reward in $[0, 1]$.

Problem Class	Current Best	Ours
Λ_{B}-stable PSR	-	$\widetilde{\mathcal{O}}\left(d_{\mathsf{PSR}} A U_A H^2 \log \mathcal{N}_\Theta \cdot \Lambda_{\mathsf{B}}^2/\varepsilon^2\right)$
α_{psr}-regular PSR	$\widetilde{\mathcal{O}}\left(d_{\mathsf{PSR}}^4 A^4 U_A^9 H^6 \log(\mathcal{N}_\Theta O)/(\alpha_{\mathsf{psr}}^6 \varepsilon^2)\right)$	$\widetilde{\mathcal{O}}\left(d_{\mathsf{PSR}} A U_A^2 H^2 \log \mathcal{N}_\Theta/(\alpha_{\mathsf{psr}}^2 \varepsilon^2)\right)$
α_{rev}-revealing tabular POMDP	$\widetilde{\mathcal{O}}\left(S^4 A^{6m-4} H^6 \log \mathcal{N}_\Theta/(\alpha_{\mathsf{rev}}^4 \varepsilon^2)\right)$	$\widetilde{\mathcal{O}}\left(S^2 A^m H^2 \log \mathcal{N}_\Theta/(\alpha_{\mathsf{rev}}^2 \varepsilon^2)\right)$
ν-future-suff. rank-d_{trans} POMDP	$\widetilde{\mathcal{O}}\left(d_{\mathsf{trans}}^4 A^{5m+3l+1} H^2 (\log \mathcal{N}_\Theta)^2 \cdot \nu^4 \gamma^2/\varepsilon^2\right)$	$\widetilde{\mathcal{O}}\left(d_{\mathsf{trans}} A^{2m-1} H^2 \log \mathcal{N}_\Theta \cdot \nu^2/\varepsilon^2\right)$
decodable rank-d_{trans} POMDP	$\widetilde{\mathcal{O}}\left(d_{\mathsf{trans}} A^m H^2 \log \mathcal{N}_{\mathcal{G}}/\varepsilon^2\right)$	$\widetilde{\mathcal{O}}\left(d_{\mathsf{trans}} A^m H^2 \log \mathcal{N}_\Theta/\varepsilon^2\right)$

$\log \mathcal{N}_\Theta$ = 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 ℓ_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 ℓ_2-Eluder argument to bridge step 1 & 2

* Concurrent work [Liu et al. ’22b] shows B-errors <= TV distance in their step 2, and performs ℓ_1-Eluder argument in their step 3, which gives similar result but worse rate.

Lower Bounds

Towards Fine-Grained Studies

Understanding fundamental limits <== 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

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

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

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

- Often provide intuitions / directions for improvement

Case Study: (Tabular) Revealing POMDPs

Our result (current best) for learning m-step α-revealing POMDPs:

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

Case Study: (Tabular) Revealing POMDPs

Our result (current best) for learning m-step α-revealing POMDPs:

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

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 α-revealing POMDPs:

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

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}:

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

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 α-revealing	$\tilde{\mathcal{O}}\left(\frac{S^2OA}{\alpha^2\varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{SO^{1/2}A}{\alpha^2\varepsilon^2}\right)$ (Theorem 4)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2O^2A}{\alpha^2}\cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{SO^{1/2}A}{\alpha^2}\cdot T}\right)$ (Corollary 7)
m-step ($m \geqslant 2$) α-revealing	$\tilde{\mathcal{O}}\left(\frac{S^2OA^m}{\alpha^2\varepsilon^2}\right)$ (Chen et al., 2022a)	$\boxed{\Omega\left(\frac{(S^{3/2}+SA)O^{1/2}A^{m-1}}{\alpha^2\varepsilon^2}\right)}$ (Theorem 5)	$\tilde{\mathcal{O}}\left(\left(\frac{S^2OA^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{SO^{1/2}A^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Theorem 6)

Omitting H, assuming O >= 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 α-revealing	$\tilde{\mathcal{O}}\left(\frac{S^2OA}{\alpha^2\varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{SO^{1/2}A}{\alpha^2\varepsilon^2}\right)$ (Theorem 4)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2O^2A}{\alpha^2}\cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{SO^{1/2}A}{\alpha^2}\cdot T}\right)$ (Corollary 7)
m-step ($m \geqslant 2$) α-revealing	$\tilde{\mathcal{O}}\left(\frac{S^2OA^m}{\alpha^2\varepsilon^2}\right)$ (Chen et al., 2022a)	$\boxed{\Omega\left(\frac{(S^{3/2}+SA)O^{1/2}A^{m-1}}{\alpha^2\varepsilon^2}\right)}$ (Theorem 5)	$\tilde{\mathcal{O}}\left(\left(\frac{S^2OA^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{SO^{1/2}A^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Theorem 6)

Omitting H, assuming O >= 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 α-revealing	$\tilde{\mathcal{O}}\left(\frac{S^2OA}{\alpha^2\varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{SO^{1/2}A}{\alpha^2\varepsilon^2}\right)$ (Theorem 4)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2O^2A}{\alpha^2}\cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{SO^{1/2}A}{\alpha^2}\cdot T}\right)$ (Corollary 7)
m-step $(m \geqslant 2)$ α-revealing	$\tilde{\mathcal{O}}\left(\frac{S^2OA^m}{\alpha^2\varepsilon^2}\right)$ (Chen et al., 2022a)	$\boxed{\Omega\left(\frac{(S^{3/2}+SA)O^{1/2}A^{m-1}}{\alpha^2\varepsilon^2}\right)}$ (Theorem 5)	$\tilde{\mathcal{O}}\left(\left(\frac{S^2OA^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{SO^{1/2}A^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Theorem 6)

Omitting H, assuming O >= SA in upper bounds

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

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 α-revealing	$\tilde{\mathcal{O}}\left(\frac{S^2OA}{\alpha^2\varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{SO^{1/2}A}{\alpha^2\varepsilon^2}\right)$ (Theorem 4)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2O^2A}{\alpha^2}\cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{SO^{1/2}A}{\alpha^2}\cdot T}\right)$ (Corollary 7)
m-step ($m \geqslant 2$) α-revealing	$\tilde{\mathcal{O}}\left(\frac{S^2OA^m}{\alpha^2\varepsilon^2}\right)$ (Chen et al., 2022a)	$\boxed{\Omega\left(\frac{(S^{3/2}+SA)O^{1/2}A^{m-1}}{\alpha^2\varepsilon^2}\right)}$ (Theorem 5)	$\tilde{\mathcal{O}}\left(\left(\frac{S^2OA^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{SO^{1/2}A^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Theorem 6)

Omitting H, assuming O >= SA in upper bounds

- First multiplicative dependence on $S, O, A, 1/\alpha, 1/\varepsilon$
- Suggests our $1/\alpha^2$ is sharp dependence on α
- 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 α-revealing	$\widetilde{\mathcal{O}}\left(\frac{S^2OA}{\alpha^2\varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{SO^{1/2}A}{\alpha^2\varepsilon^2}\right)$ (Theorem 4)	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{S^2O^2A}{\alpha^2}\cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{SO^{1/2}A}{\alpha^2}\cdot T}\right)$ (Corollary 7)
m-step ($m \geqslant 2$) α-revealing	$\widetilde{\mathcal{O}}\left(\frac{S^2OA^m}{\alpha^2\varepsilon^2}\right)$ (Chen et al., 2022a)	$\boxed{\Omega\left(\frac{(S^{3/2}+SA)O^{1/2}A^{m-1}}{\alpha^2\varepsilon^2}\right)}$ (Theorem 5)	$\widetilde{\mathcal{O}}\left(\left(\frac{S^2OA^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{SO^{1/2}A^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Theorem 6)

Omitting H, assuming O >= SA in upper bounds

- First multiplicative dependence on $S, O, A, 1/\alpha, 1/\varepsilon$
- Suggests our $1/\alpha^2$ is sharp dependence on α
- First joint dependence on O and $1/(\alpha\varepsilon)$
- Regret is $\Omega(T^{2/3})$ for m-step revealing, whereas $\widetilde{O}(\sqrt{T})$ 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 α-revealing	$\tilde{\mathcal{O}}\left(\frac{S^2OA}{\alpha^2\varepsilon^2}\right)$ (Chen et al., 2022a)	$\Omega\left(\frac{SO^{1/2}A}{\alpha^2\varepsilon^2}\right)$ (Theorem 4)	$\tilde{\mathcal{O}}\left(\sqrt{\frac{S^2O^2A}{\alpha^2}\cdot T}\right)$ (Theorem 8)	$\Omega\left(\sqrt{\frac{SO^{1/2}A}{\alpha^2}\cdot T}\right)$ (Corollary 7)
m-step ($m \geqslant 2$) α-revealing	$\tilde{\mathcal{O}}\left(\frac{S^2OA^m}{\alpha^2\varepsilon^2}\right)$ (Chen et al., 2022a)	$\boxed{\Omega\left(\frac{(S^{3/2}+SA)O^{1/2}A^{m-1}}{\alpha^2\varepsilon^2}\right)}$ (Theorem 5)	$\tilde{\mathcal{O}}\left(\left(\frac{S^2OA^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Chen et al., 2022a)*	$\Omega\left(\left(\frac{SO^{1/2}A^m}{\alpha^2}\right)^{1/3}T^{2/3}\right)$ (Theorem 6)

Omitting H, assuming O >= SA in upper bounds

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

Lower Bounds for Revealing POMDPs

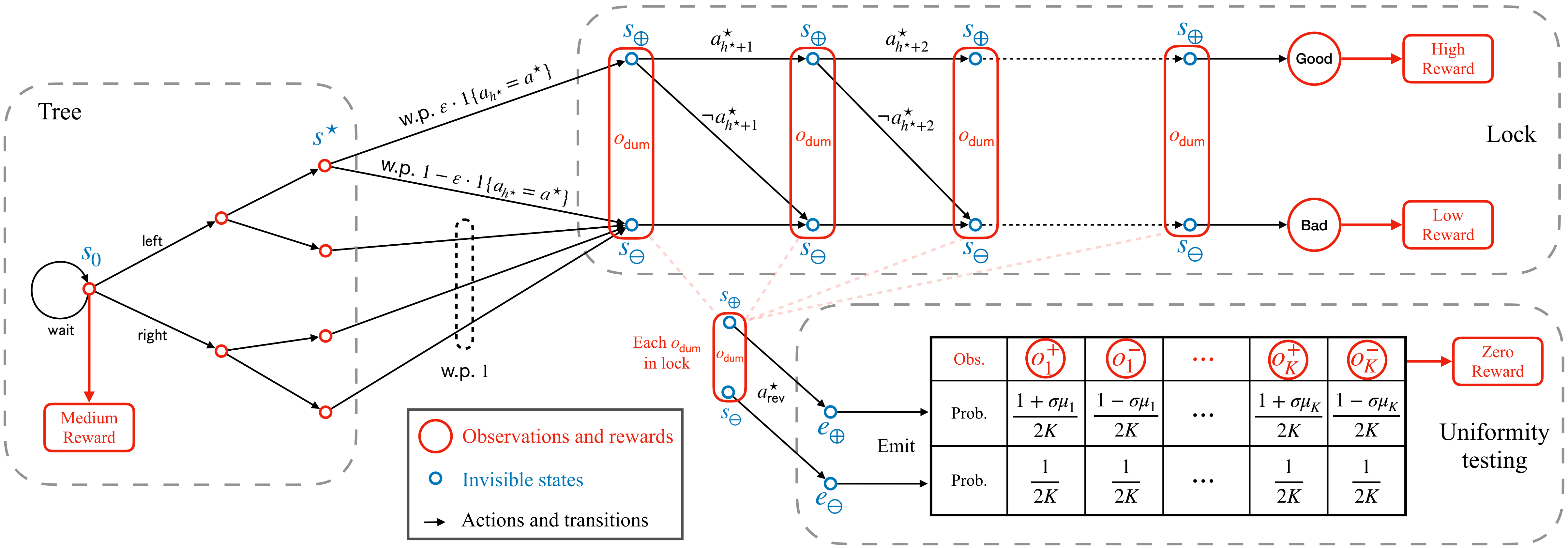
[Chen, Wang, Xiong, Mei, **Bai** '23]

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

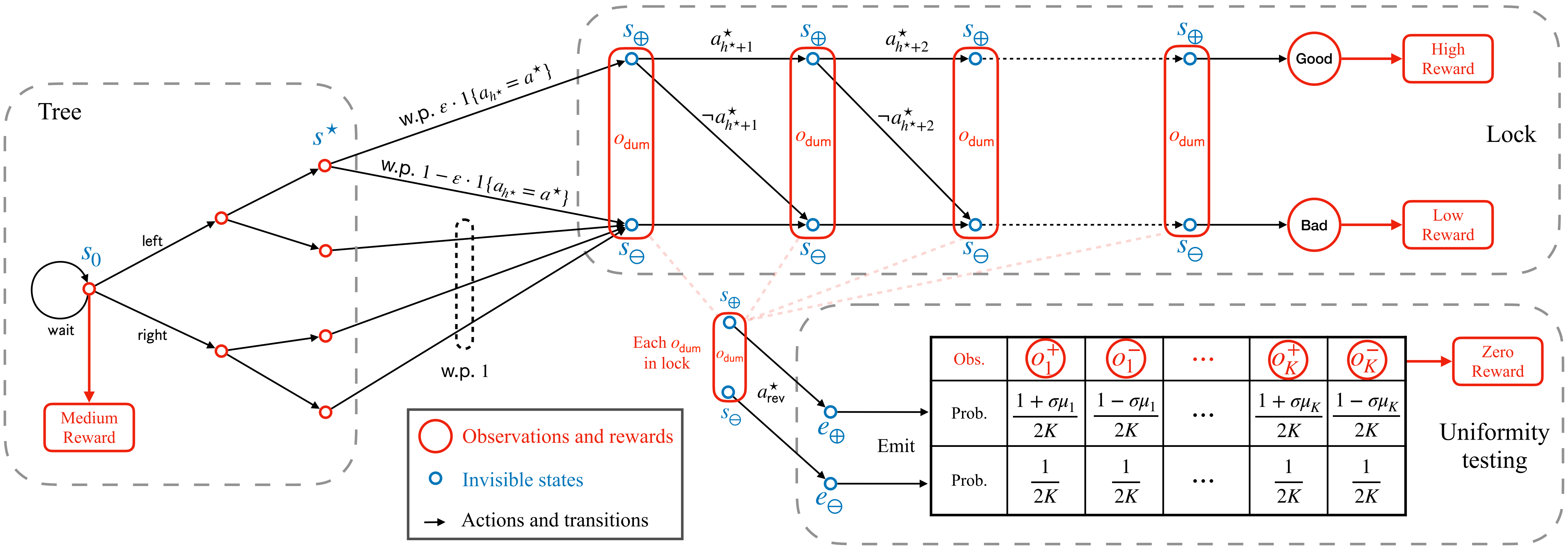
Omitting H, assuming O >= SA in upper bounds

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

Hard instance construction (2-step case, simplified)

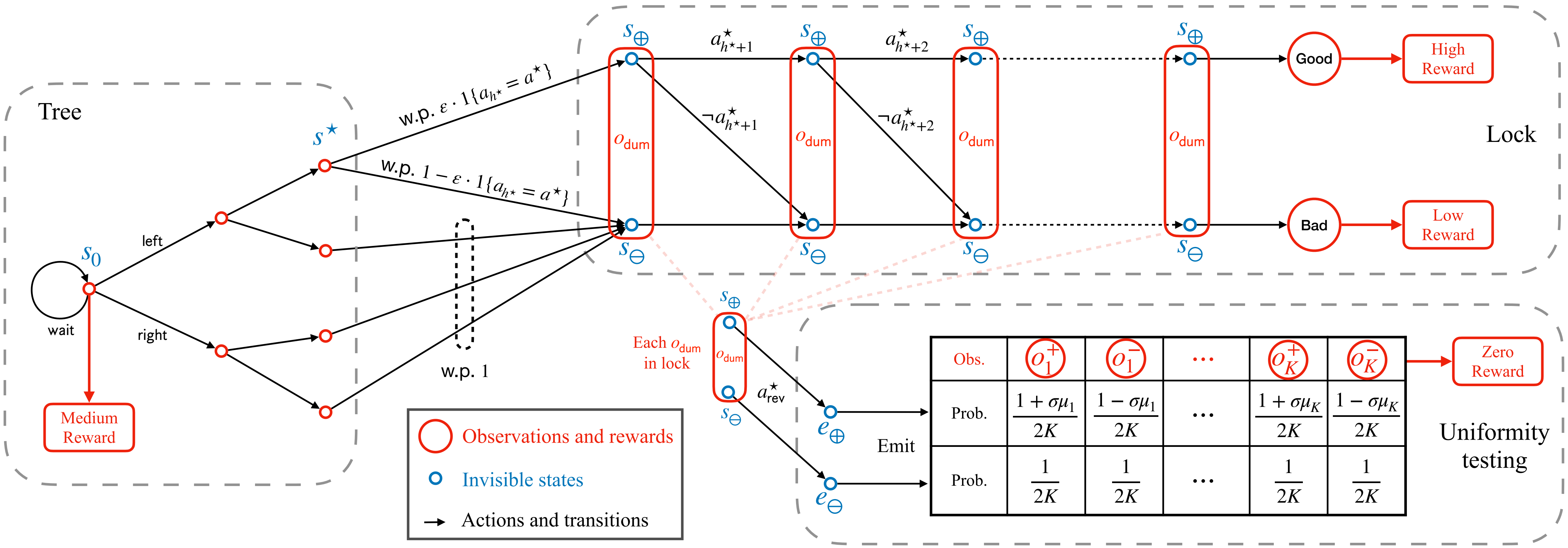


Hard instance construction (2-step case, simplified)



Building blocks

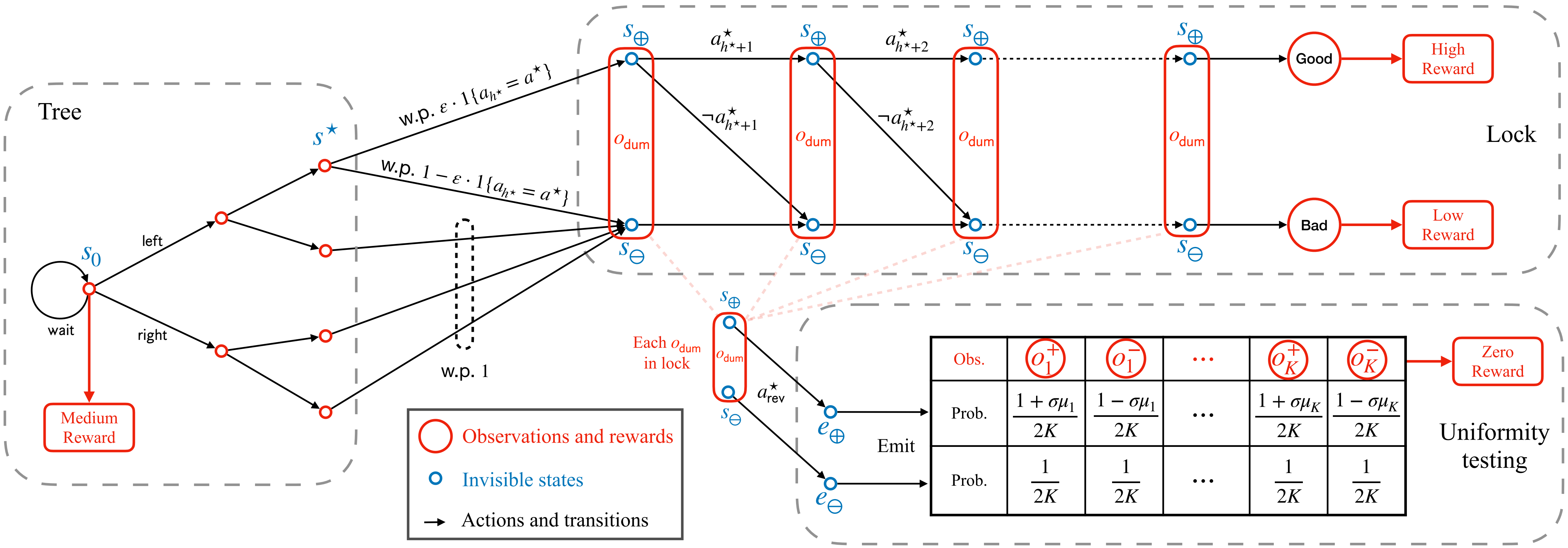
Hard instance construction (2-step case, simplified)



Building blocks

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

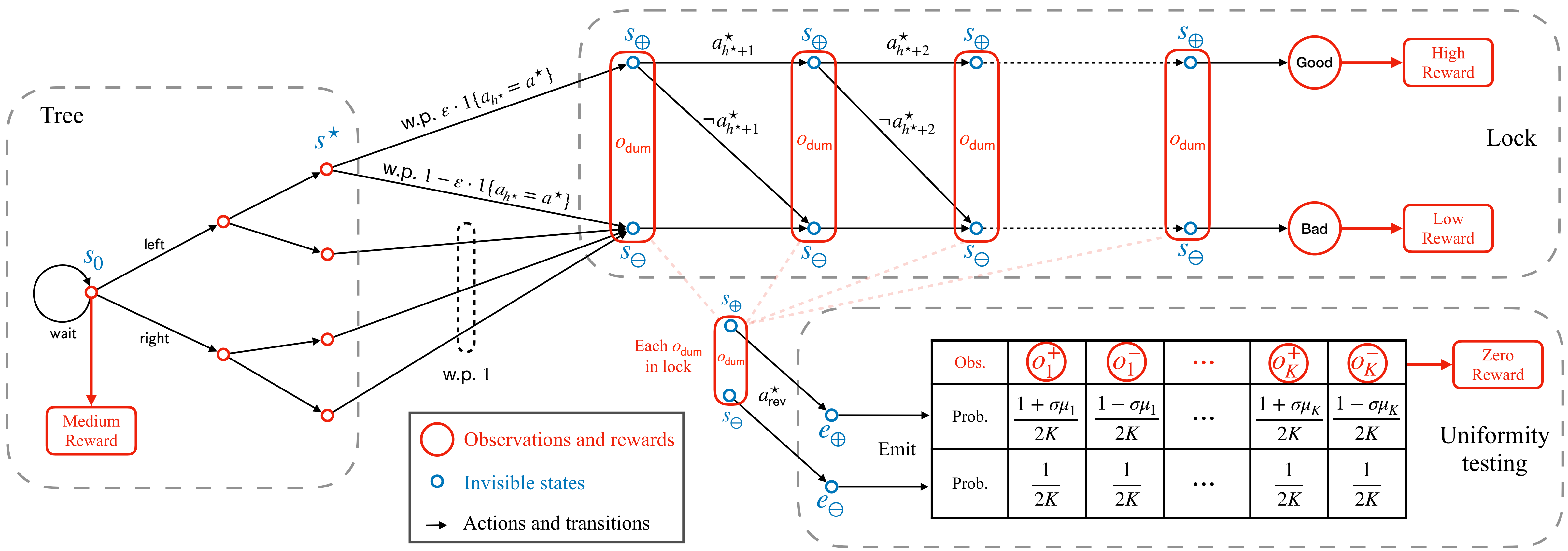
Hard instance construction (2-step case, simplified)



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 $\sqrt{O}/(\alpha^2\varepsilon^2)$ 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

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

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 Π-*norm*: For a vector $\mathbf{b}$ indexed by $\tau_{h:H} \in (\mathcal{O}\times\mathcal{A})^{H-h+1}$, we define

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

where the maximization is over all policies $\bar{\pi}$ starting from step h (ignoring the history τ_{h-1}) and $\bar{\pi}(\tau_{h:H}) = \prod_{h\leqslant h'\leqslant H} \bar{\pi}_{h'}(a_{h'}|o_{h'}, \tau_{h:h'-1})$. 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 Π'-norm[5]:

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

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

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

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 Π-norm.

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

$$\sup_{h\in[H]} \max_{\|\mathbf{q}\|_*=1} \|\mathcal{B}_{H:h}\mathbf{q}\|_{\Pi} \leqslant \Lambda_{\mathsf{B}}. \tag{6}$$

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 $h \leqslant H - m$, we consider $t_h = (o_h, a_h, \cdots, o_{h+m-1}) \in \mathcal{U}_h$, $t_{h+1} = (o'_{h+1}, a'_{h+1}, \cdots, o'_{h+m}) \in \mathcal{U}_{h+1}$, and define

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

where ϕ_{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

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

for any reachable (τ_{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. [18] Then, for $h \in [H]$, we can take

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

where $\mathbb{1}((o, a) \to t_h)$ is 1 if t_h starts with (o, a) and 0 otherwise[19].

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 α-revealing POMDP M is a α^{-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 $h \leqslant H - m$, set

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

and when $h > H - m$, take

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

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.