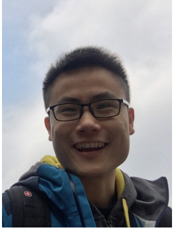
When Is Partially Observable Reinforcement Learning Not Scary?

Chi Jin

Princeton University.

Collaborators



Qinghua Liu
Princeton

Alan Chung
Princeton

Sham Kakade
Harvard

Akshay Krishnamurthy
MSR, NY

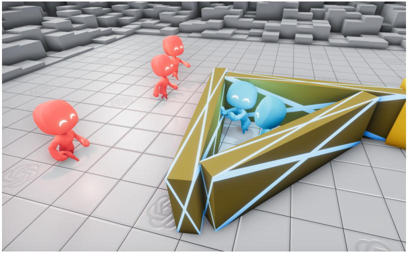
Praneeth Netrapalli
MSR, India

Csaba Szepesvari
U of Alberta

Partial Observability

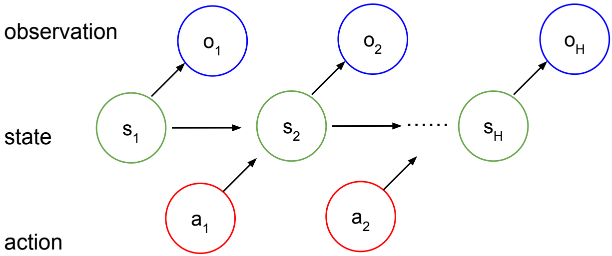
Partial observability is ubiquitous in modern RL.



Mathematical Model

Partially Observable Markov Decision Process (POMDP)



POMDP = MDP + emission OR hidden Markov model + control.

Unique Challenges

Tabular MDP (known states)

- finite size of states/actions/horizon S, A, H.
- computation and sample complexity: poly(S, A, H) [AOM17, JABJ18, ...].

Tabular POMDP (unknown states)

- reason about beliefs over the states.
- maintain memory: $2^{\Omega(H)}$ possible histories!
- all while exploring the environment

- **Challenging!** many hardness results.

Hardness

Computational hardness: **planning alone is already hard**.

*planning: compute optimal policy based on known model or values.

- optimal policy: PSPACE-complete [PT87].
- optimal *memoryless* policy: NP-hard [VLB12].

Statistical hardness (even if allowing infinite computation):

- learning POMDP requires $\Omega(A^H)$ samples!

We will address the statistical efficiency today!

Main Questions

1. Problem Structure:

   Can we identify a rich sub-class of POMDPs that is statistically tractable?

2. Algorithm:

   Can we design simple algorithms that efficiently learn this rich class?

An Overview of Our Results

1. Problem Structure:

   - A new rich sub-class of POMDPs—weakly revealing POMDPs.
   - ruling out the pathological POMDPs with uninformative observations.

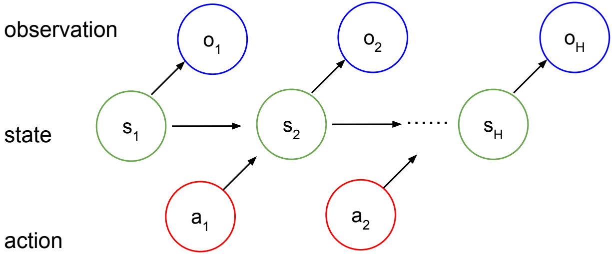
2. Algorithm:

   - A new simple generic algorithm—OMLE.
     Optimism + Maximum Likelihood Estimation (MLE)
   - solve weakly revealing POMDPs in polynomial samples.

First line of sample-efficient results for learning from interactions in large classes of POMDPs.

Formulation and Objectives

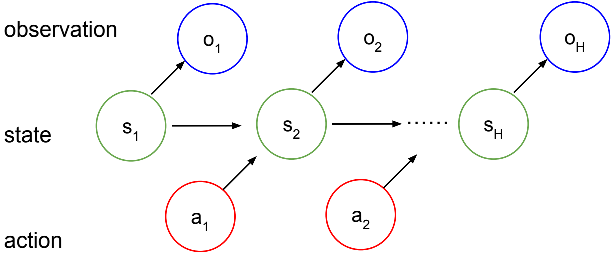
Formal Setup



Partially Observable MDP (POMDP) $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathbb{T}, \mathbb{O}, \mu_1, r, H)$:

- finite state set $\mathcal{S}$, finite action set $\mathcal{A}$, finite observation set $\mathcal{O}$.
- transition $\mathbb{T}_h(s' \mid s, a)$; emission $\mathbb{O}_h(o \mid s)$; initial distribution $\mu_1(s)$
- reward $r_h : \mathcal{O} \to [0, 1]$; horizon length H.

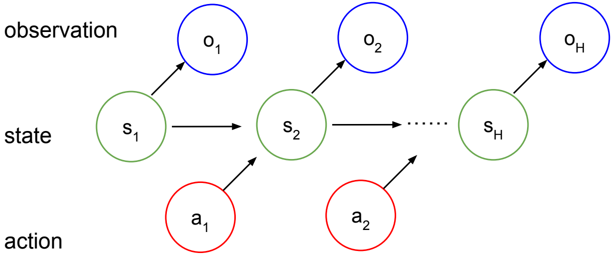
Policy and Values



- **Policy** π: a collection of maps $\{\pi_h\}_{h=1}^{H}$ with $\pi_h : \mathcal{T}_h \to \mathcal{A}$.
  $\mathcal{T}_h = \{(o_1, a_1, \ldots, o_h)\}$ is the set of all possible *h*-step histories.
- **Value** V^π: the total expected reward received under π.

Objective: find the optimal $\pi^\star$ that maximize V^π.

Learning from Interactions



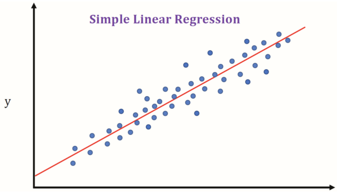
Agent learns by online interaction with POMDPs: at each episode k

- agent picks a policy π to execute.
- receive a trajectory $(o_1, a_1, \ldots, o_{H-1}, a_{H-1}, o_H)$.

Components of Learning

Planning: compute optimal policy based on known model or values.

Estimation: estimate model/values based on collected data samples.



Exploration: strategically collect informed data samples.

Prior Works

Guo et al. (2016); Azizzadenesheli et al. (2016); Xiong et al. (2021)

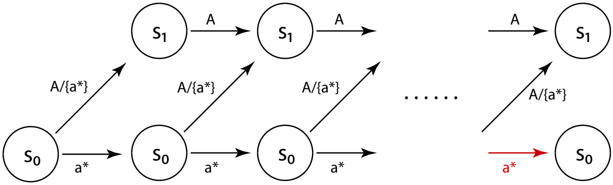
- Made several various strong assumptions about the POMDPs (e.g., reachability, invertibility of the transition matrix, or ergodicity).
- Does not address exploration.

A Rich Class of Tractable POMDPs

A Hard Instance

POMDP "without observations":

- combinatorial lock (as underlying MDP) + dummy observation.



$\Leftrightarrow$ enter a passcode of length H, requires $\Omega(A^H)$ samples.

POMDPs are hard if two (mixtures of) states lead to the same distribution over observations.

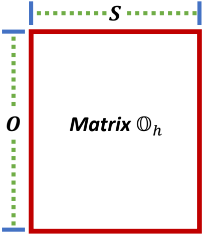
Weakly Revealing POMDPs

Rule out the pathological instances that prevent efficient learning!

- Emission matrix $[\mathbb{O}_h]_{o,s} = \mathbb{O}_h(o|s)$.
- Different mixtures of states induce different observation distributions.

$\Leftrightarrow \quad \mu_1 \neq \mu_2 \Rightarrow \mathbb{O}_h\mu_1 \neq \mathbb{O}_h\mu_2.$

$\Leftrightarrow \quad \mathrm{rank}(\mathbb{O}_h) = S$



$rank(\mathbb{O}_h) = S$

α-weakly revealing condition [JKKL20]

$$\sigma_S(\mathbb{O}_h) \geq \alpha > 0$$

Overcomplete Settings

- $\sigma_S(\mathbb{O}_h) > 0$ is only possible in undercomplete POMDPs ($S \le O$)
- **Overcomplete case**: use multistep observations to distinguish states.

m-step emission-action matrix $\mathbb{M}_h \in \mathbb{R}^{A^{m-1}O^m \times S}$:

$$[\mathbb{M}_h]_{(\mathbf{a},\mathbf{o}),s} = \mathbb{P}(o_{h:h+m-1} = \mathbf{o} \mid s_h = s, a_{h:h+m-2} = \mathbf{a})$$

m-step α-weakly revealing condition [LCSJ22]

$$\sigma_S(\mathbb{M}_h) \ge \alpha > 0$$

Sample-Efficient Algorithms

High-level Ideas

Prior algorithms on tabular MDP: value-based approach

- in POMDP, value depends on entire history $\rightarrow$ exponential size.
- Use model-based approach! POMDPs have polynomial model sizes.

[JKKL20]: design an spectral-based algorithm to estimate model.

[LCSJ22]: why not simply do

Maximum Likelihood Estimation (MLE) + Optimism!

OMLE Algorithm

- $\theta = \{\mathbb{T}_h, \mathbb{O}_h\}_{h=1}^{H}$ are model parameters.
- $V^\pi(\theta)$: value of policy π under model θ.

Optimistic MLE [LCSJ22]

for $k = 1, \ldots, K$

1. **optimistic planning**

   compute $(\theta^k, \pi^k) = \mathrm{argmax}_{\theta \in \mathcal{B}, \pi} V^\pi(\theta)$.

2. **data collection**

   execute π^k to collect a trajectory $\tau^k = (o_1, a_1, \ldots, o_H, a_H)$.

3. **update confidence set $\mathcal{B}$.**

output π^{out} sampled uniformly from $\{\pi^k\}_{k=1}^{K}$.

OMLE Algorithm II

Confidence set $\mathcal{B}$:

$$\mathcal{B} = \left\{ \theta \in \Theta : \underbrace{\sum_{(\pi,\tau)\in\mathcal{D}} \log \mathbb{P}_\theta^\pi(\tau)}_{\text{likelihood of } \theta} \geq \underbrace{\max_{\theta'\in\Theta} \sum_{(\pi,\tau)\in\mathcal{D}} \log \mathbb{P}_{\theta'}^\pi(\tau)}_{\text{MLE}} - \underbrace{\beta}_{\textit{tolerance}} \right\}$$

- $\mathbb{P}_\theta^\pi(\tau)$: the probability of observe trajectory τ if following policy π under model θ.
- Θ: set of model parameters such that $\sigma_S(\mathbb{O}_h) \geq \alpha$.

Theoretical Guarantees

Theorem (undercomplete case)

For α-weakly revealing POMDPs, **OMLE** outputs an $\mathcal{O}(\epsilon)$-optimal policy in $\mathrm{poly}(H, S, A, O, \epsilon^{-1}, \alpha^{-1})$ episodes.

Theorem (overcomplete case)

For m-step α-weakly revealing POMDPs, an m-step version of **OMLE** outputs an $\mathcal{O}(\epsilon)$-optimal policy in $\mathrm{poly}(H, S, A^m, O, \epsilon^{-1}, \alpha^{-1})$ episodes.

First line of sample-efficient results for learning from interactions in rich classes of POMDPs.

Lower Bounds

Are $\mathrm{poly}(\alpha^{-1})$ or $A^{\Omega(m)}$ necessary? Yes.

- $\exists$ undercomplete α-weakly revealing POMDPs such that any algorithm requires $\Omega(\min\{(\alpha H)^{-1}, A^{H-1}\})$ samples to learn $\mathcal{O}(1)$-optimal policy.
- $\exists$ m-step α-weakly revealing POMDPs such that any algorithm requires $\Omega(A^{m-1})$ samples to learn $\mathcal{O}(1)$-optimal policy.

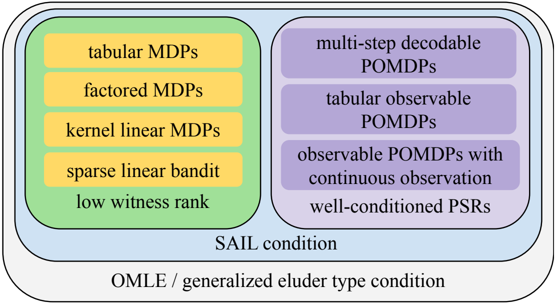
Beyond POMDPs

Multiagent RL under Partial Observability

Partially observable Markov games [LSJ2022]:

- each player has local observation o_i.
- covers imperfect information extensive-form games (IIEFGs).
- joint observations of all players $(o_1, \dots, o_m)$ weakly reveals the states.
- OMLE-Eq learns various equilibria in polynomial samples.

Continuous Observation and Beyond



Generic partially observable sequential decision making [LNSJ2022].

- observable POMDPs with continuous observation
- well-conditioned predictive state representations
- any RL problems satisfying the SAIL condition, which covers a majority of known tractable model-based RL problems.

Conclusion

Summary

First line of sample-efficient algorithms for learning from interactions in large classes of partially observable problems.

- Simple optimism + MLE suffices.
- (multi-step) weakly revealing POMDPs.
- Weakly-revealing POMGs.
- Continuous observations and beyond.

Future directions:

- Richer classes of tractable partially observable problems.
- Computational efficiency.

Thank you!