Sample-efficient Reinforcement Learning of Undercomplete POMDPs

based on joint work with Chi Jin, Sham Kakade and Akshay Krishnamurthy

Qinghua Liu
Princeton University

RL Theory Seminar, February 23, 2021

Table of contents

Introduction

Background

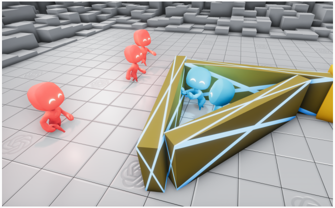
- Partial observability is a common feature in real world.

Texas Hold'em Poker

Robotics

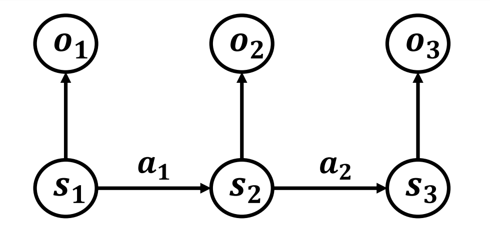
Starcraft

Hide-and-seek



Background

- POMDP is a classic model for modeling partial observability.



POMDP = hidden Markov model + input control.

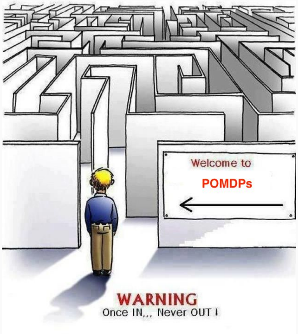
POMDPs is challenging

- Cannot observe the current state

  ⇒ cannot determine if a new state is reached

- The current hidden state depends on the **entire history**

  ⇒ **exponential possibilities!**



Computational hardness

Planning is Hard! When the parameters are *known*,

- PSPACE-complete to compute the optimal policy [PT87]
- NP-hard to compute the optimal memoryless policy [VLB12]

Computational hardness

Planning is Hard! When the parameters are *known*,

- PSPACE-complete to compute the optimal policy [PT87]
- NP-hard to compute the optimal memoryless policy [VLB12]

Q. Can we obtain any positive result for POMDPs?

Computational hardness

Planning is Hard! When the parameters are *known*,

- PSPACE-complete to compute the optimal policy [PT87]
- NP-hard to compute the optimal memoryless policy [VLB12]

Q. Can we obtain any positive result for POMDPs?

A. Yes! A rich class of POMDPs is sample-efficiently learnable!

Existing works

- [EDKM05], [RCdP08], [PV08]: without sample complexity guarantee.
- [GDB16, ALA16]: assume all latent states can be reached by random actions or given polices.

Existing works

- [EDKM05], [RCdP08], [PV08]: without sample complexity guarantee.
- [GDB16, ALA16]: assume all latent states can be reached by random actions or given polices.

Existing works do NOT address the EXPLORATION challenge.

Existing works

- [EDKM05], [RCdP08], [PV08]: without sample complexity guarantee.
- [GDB16, ALA16]: assume all latent states can be reached by random actions or given polices.

Existing works do NOT address the EXPLORATION challenge.

This work: attack EXPLORATION directly.

Settings and lower bounds

Definition of POMDPs

Formally, a **POMDP** is specified by

- state set $\mathcal{S}$, observation set $\mathcal{O}$, action set $\mathcal{A}$.
- H: length of horizon.

Definition of POMDPs

Formally, a **POMDP** is specified by

- state set $\mathcal{S}$, observation set $\mathcal{O}$, action set $\mathcal{A}$.
- H: length of horizon.

- $\mathbb{T}_h(s' \mid s, a)$: transition measure.
- $\mathbb{O}_h(o \mid s)$: emission measure.
- μ_1: distribution of s_1.

Definition of POMDPs

Formally, a **POMDP** is specified by

- state set $\mathcal{S}$, observation set $\mathcal{O}$, action set $\mathcal{A}$.
- H: length of horizon.

- $\mathbb{T}_h(s' \mid s, a)$: transition measure.
- $\mathbb{O}_h(o \mid s)$: emission measure.
- μ_1: distribution of s_1.

- $r : (\mathcal{O} \times \mathcal{A})^H \to [0, H]$: reward function.

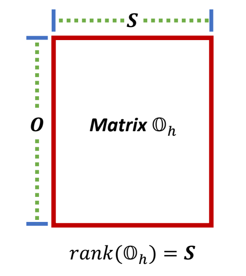
Settings

Assumption

(a) The POMDP is undercomplete, i.e. $S \leq O$

(b) $\sigma_{\min}(\mathbb{O}_h) \geq \alpha > 0$ for all h

(a)+(b) is a robust version of $\mathrm{rank}(\mathbb{O}_h) = S$



Theorem (Lower bound)

Without either (a) or (b), learning a 1/4-optimal policy needs at least $\Omega(A^{H-1})$ samples in general.

Observable operator models

Definition of OOMs

Definition (A philosophical one)

probability of *observable* sequence = product of *operators*.

An operator view of POMDPs

$$\mathbb{P}(o_{1:H} \mid a_{1:H-1}) = \mathbf{e}_{o_H}^{\mathrm{T}} \cdot \mathbf{B}(a_{H-1}, o_{H-1}) \cdots \mathbf{B}(a_1, o_1) \cdot \mathbf{b}_0$$

where $\mathbf{B}(a, o) = \mathbb{O}\mathbb{T}(a)\mathrm{diag}(\mathbb{O}(o \mid \cdot))\mathbb{O}^{\dagger}$ and $\mathbf{b}_0 = \mathbb{O}\mu_1$.

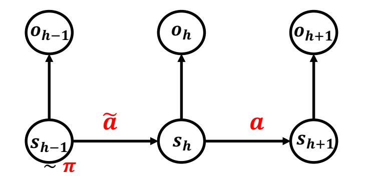
Benefits of the operator view

- No need to recover model parameters: **learning operators suffices**.
- Operators are indexed by observations and actions, not by unobservable underlying hidden states.
- Most importantly, the operators satisfy certain **moment constraints**!

The moment constraint

Given *arbitrary*

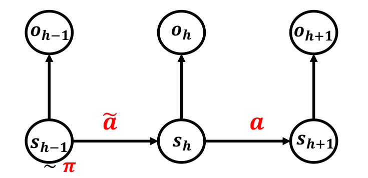
- actions a and $\tilde{a}$
- policy π



The moment constraint

Given *arbitrary*

- actions a and $\tilde{a}$
- policy π



Let $\mathbf{N}_h(a, \tilde{a}), \mathbf{M}_h(o, a, \tilde{a}) \in \mathbb{R}^{O \times O}$ be the probability matrices s.t.

$$\mathbf{N}_h(a, \tilde{a}) = \mathbb{P}(o_h = \cdot, o_{h-1} = \cdot)$$
$$\mathbf{M}_h(o, a, \tilde{a}) = \mathbb{P}(o_{h+1} = \cdot, o_h = o, o_{h-1} = \cdot)$$

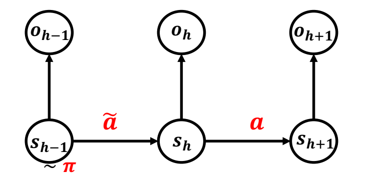
Then

$$\mathbf{B}(a, o)\mathbf{N}_h(a, \tilde{a}) = \mathbf{M}_h(o, a, \tilde{a}) \qquad (*)$$

The moment constraint

Given *arbitrary*

- actions a and $\tilde{a}$
- policy π



Let $\mathbf{N}_h(a, \tilde{a}), \mathbf{M}_h(o, a, \tilde{a}) \in \mathbb{R}^{O \times O}$ be the probability matrices s.t.

$$\mathbf{N}_h(a, \tilde{a}) = \mathbb{P}(o_h = \cdot, o_{h-1} = \cdot)$$
$$\mathbf{M}_h(o, a, \tilde{a}) = \mathbb{P}(o_{h+1} = \cdot, o_h = o, o_{h-1} = \cdot)$$

Then

$$\mathbf{B}(a, o)\mathbf{N}_h(a, \tilde{a}) = \mathbf{M}_h(o, a, \tilde{a}) \qquad (*)$$

Moreover, if $\mathrm{rank}(\mathbf{N}) = S$, then $\mathbf{B}$ is identified by $(*)$.

Algorithm OOM-UCB

Algorithm OOM-UCB

For $k = 1, \ldots, K$

1. Optimistic planning

$$\pi_k \leftarrow \arg\max_{\pi} \max_{\hat{\theta} \in \Theta_k} V_1^{\pi}(\hat{\theta}).$$

2. Collect data using π_k.
3. Construct the confidence set Θ_k.

Output π_k sampled u.a.r. from $\{\pi_k\}_{k=1}^{K}$.

Local Confidence Set + Global Optimism

Data collection

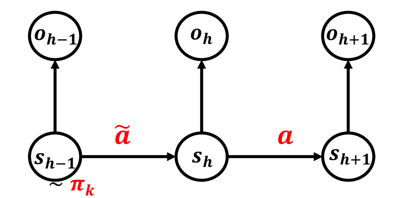
2. Collect data using π_k

For all $(h, a, \tilde{a})$ do:

(1) execute π_k for step 1 to $h-2$

(2) take action $\tilde{a}$ and a at step $h-1$ and h, respectively

(3) add 1 to the $(o_h, o_{h-1})^{\text{th}}$ entry of $\widehat{\mathbf{N}}_h(a, \tilde{a})$
and the $(o_{h+1}, o_{h-1})^{\text{th}}$ entry of $\widehat{\mathbf{M}}_h(o, a, \tilde{a})$



Construct the confidence set

3. **Construct the confidence set Θ_k**

- Replace $\mathbf{N}_h$ and $\mathbf{M}_h$ by empirical estimate $\widehat{\mathbf{N}}_h$ and $\widehat{\mathbf{M}}_h$.

Construct the confidence set

3. **Construct the confidence set Θ_k**

- Replace $\mathbf{N}_h$ and $\mathbf{M}_h$ by empirical estimate $\widehat{\mathbf{N}}_h$ and $\widehat{\mathbf{M}}_h$.
- Construct the confidence set for each $o, a, \tilde{a}, h$

$$\mathfrak{B}_h(o, a, \tilde{a}) \triangleq \left\{\hat{\theta} : \|\mathbf{B}(a, o; \hat{\theta})\widehat{\mathbf{N}}_h(a, \tilde{a}) - \widehat{\mathbf{M}}_h(o, a, \tilde{a})\| \leq \gamma\right\}.$$

Construct the confidence set

3. Construct the confidence set Θ_k

- Replace $\mathbf{N}_h$ and $\mathbf{M}_h$ by empirical estimate $\widehat{\mathbf{N}}_h$ and $\widehat{\mathbf{M}}_h$.

- Construct the confidence set for each $o, a, \tilde{a}, h$

$$\mathfrak{B}_h(o, a, \tilde{a}) \triangleq \left\{\hat{\theta} : \|\mathbf{B}(a, o; \hat{\theta})\widehat{\mathbf{N}}_h(a, \tilde{a}) - \widehat{\mathbf{M}}_h(o, a, \tilde{a})\| \leq \gamma\right\}.$$

- Take the intersection of all confidence sets

$$\Theta \triangleq \left[\cap_{(o,a,\tilde{a},h)} \mathfrak{B}_h(o, a, \tilde{a})\right] \cap \{\hat{\theta} : \sigma_{\min}(\hat{\mathbb{O}}) \geq \alpha\}.$$

Remark. The confidence set for $\mathbf{b}_0$ is simple; we neglect it here.

Main theorem

Assumption

(a) The POMDP is undercomplete, i.e. $S \leq O$.

(b) $\sigma_{\min}(\mathbb{O}_h) \geq \alpha > 0$ for all h.

Theorem

Under the assumption above, OOM-UCB outputs an ϵ-optimal policy within $\mathrm{poly}(H, S, A, O, \alpha^{-1})/\epsilon^2$ iterations with probability at least $2/3$.

The first polynomial sample complexity guarantee for RL of POMDPs in the exploration-setting.

Proof Sketch (1/2)

- Martingale concentration $\Rightarrow \theta^\star \in \Theta^k$
- Optimistic planning: $(\pi_k, \theta_k) \leftarrow \arg\max_\pi \max_{\hat{\theta} \in \Theta_k} V_1^\pi(\hat{\theta})$

$$\Rightarrow \sum_{k=1}^{K} \underbrace{[V^\star(\theta^\star) - V^{\pi_k}(\theta^\star)]}_{\text{suboptimality gap}} \leq \sum_{k=1}^{K} \underbrace{[V^{\pi_k}(\theta_k) - V^{\pi_k}(\theta^\star)]}_{\text{same policy, different models}}$$

Proof Sketch (2/2)

$$\sum_{k=1}^{K} \underbrace{\left[V^{\pi_k}(\theta_k) - V^{\pi_k}(\theta^\star)\right]}_{\text{same policy, different models}}$$

$$\lesssim \sum_{k=1}^{K} \sum_{h,a,\tilde{a},o,s} \underbrace{\left\| \left[\mathbf{B}(a,o;\theta_k) - \mathbf{B}(a,o;\theta^\star)\right] \mathbb{OT}(\tilde{a}) \mathbf{e}_s \right\|_1}_{\text{operator error of } \theta_k \text{ on } s\text{-direction}} \cdot \underbrace{\mathbb{P}^{\pi_k}_{\theta^\star}(s_{h-1} = s)}_{\text{prob. of visiting } s \text{ by } \pi_k}$$

NO need to recover B.

Being accurate in the directions of frequently visited states suffices.

Future directions

- Over-complete POMDPs.
- Markov games with partial observations.
- Function approximation.
- Stronger assumptions for computational efficiency.

  ...

Thank You!

Reference

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar.
Reinforcement learning of pomdps using spectral methods.
29th Annual Conference on Learning Theory, 2016.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour.
Reinforcement learning in pomdps without resets.
2005.

Zhaohan Daniel Guo, Shayan Doroudi, and Emma Brunskill.
A pac rl algorithm for episodic pomdps.
In *Artificial Intelligence and Statistics*, pages 510–518, 2016.

Christos H Papadimitriou and John N Tsitsiklis.
The complexity of markov decision processes.
Mathematics of operations research, 12(3):441–450, 1987.

Pascal Poupart and Nikos Vlassis.
Model-based bayesian reinforcement learning in partially observable domains.

In *Proc Int. Symp. on Artificial Intelligence and Mathematics,*, pages 1–2, 2008.

Stephane Ross, Brahim Chaib-draa, and Joelle Pineau.
Bayes-adaptive pomdps.
In *Advances in neural information processing systems*, pages 1225–1232, 2008.

Nikos Vlassis, Michael L Littman, and David Barber.
On the computational complexity of stochastic controller optimization in pomdps.
ACM Transactions on Computation Theory (TOCT), 4(4):1–8, 2012.