Reinforcement Learning without Intractable Oracles?

Ankur Moitra (MIT)

April 7th, ONR Program Meeting

OUTLINE

Part I: Introduction

- Models and Problems
- Hardness and Beyond Worst-Case Analysis
- Our Results

Part II: Planning

Part III: Learning

- Approximate MDPs via Barycentric Spanners

OUTLINE

Part I: Introduction

- **Models and Problems**
- Hardness and Beyond Worst-Case Analysis
- Our Results

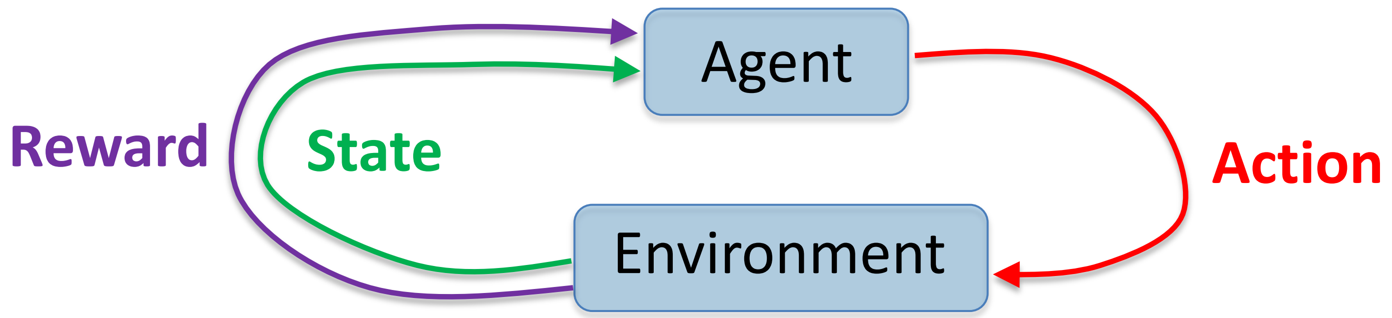
Part II: Planning

Part III: Learning

- Approximate MDPs via Barycentric Spanners

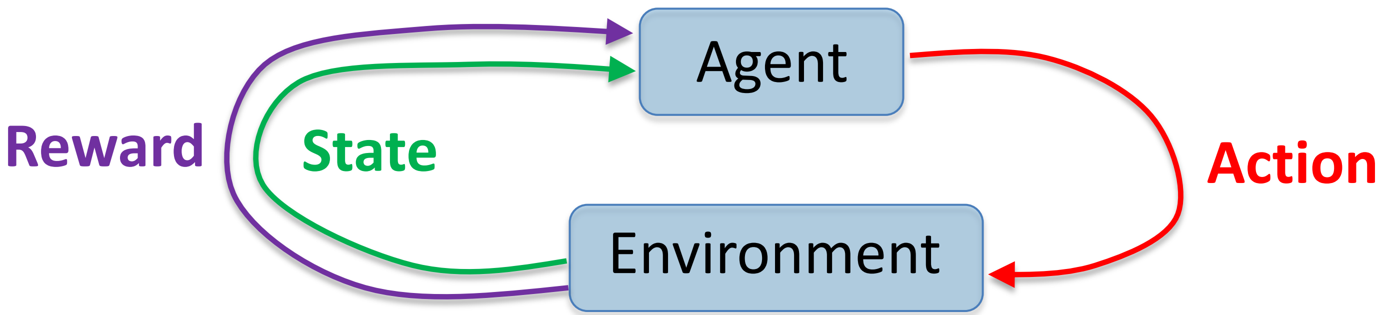
REINFORCEMENT LEARNING (RL)

Goal: Agent learns by interacting with the environment



REINFORCEMENT LEARNING (RL)

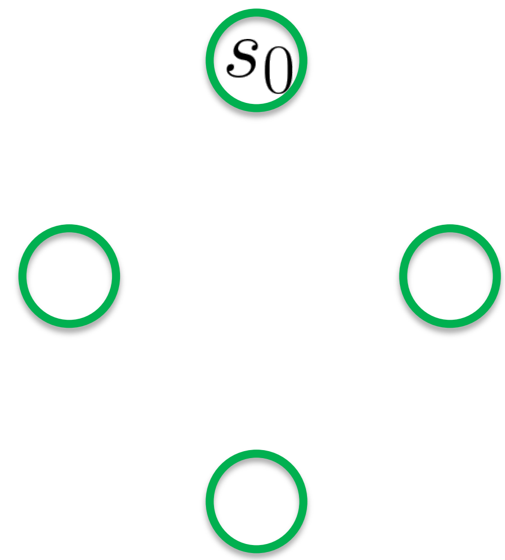
Goal: Agent learns by interacting with the environment



Standard model is a **Markov Decision Process**

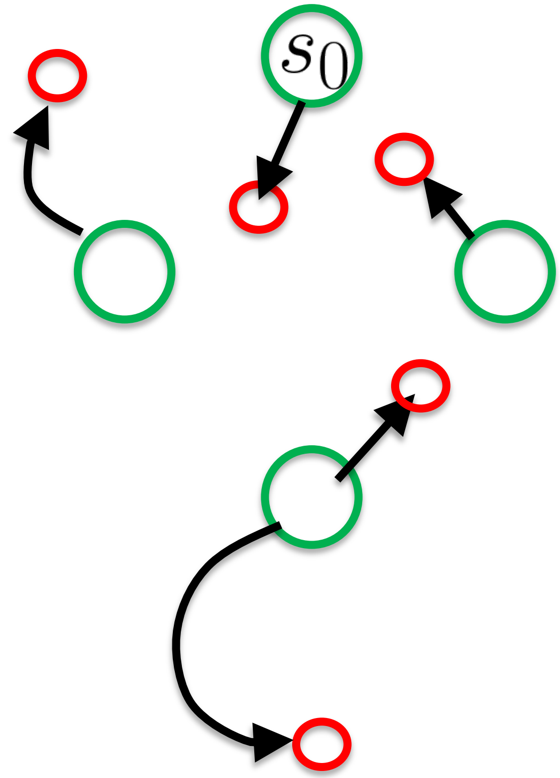
MARKOV DECISION PROCESSES

- State Space $\mathcal{S}$, start at s_0



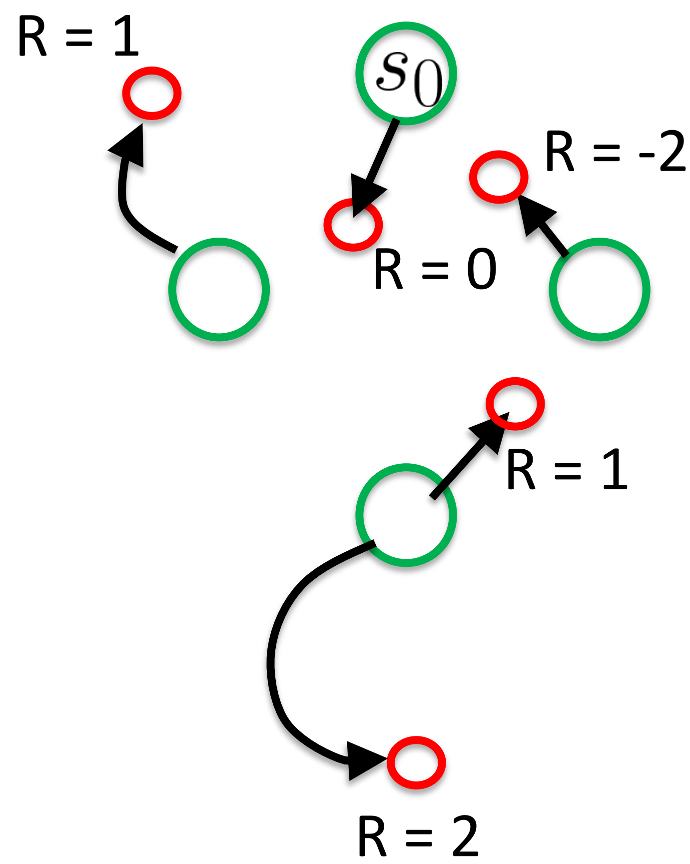
MARKOV DECISION PROCESSES

- State Space $\mathcal{S}$, start at s_0
- Action Space $\mathcal{A}$



MARKOV DECISION PROCESSES

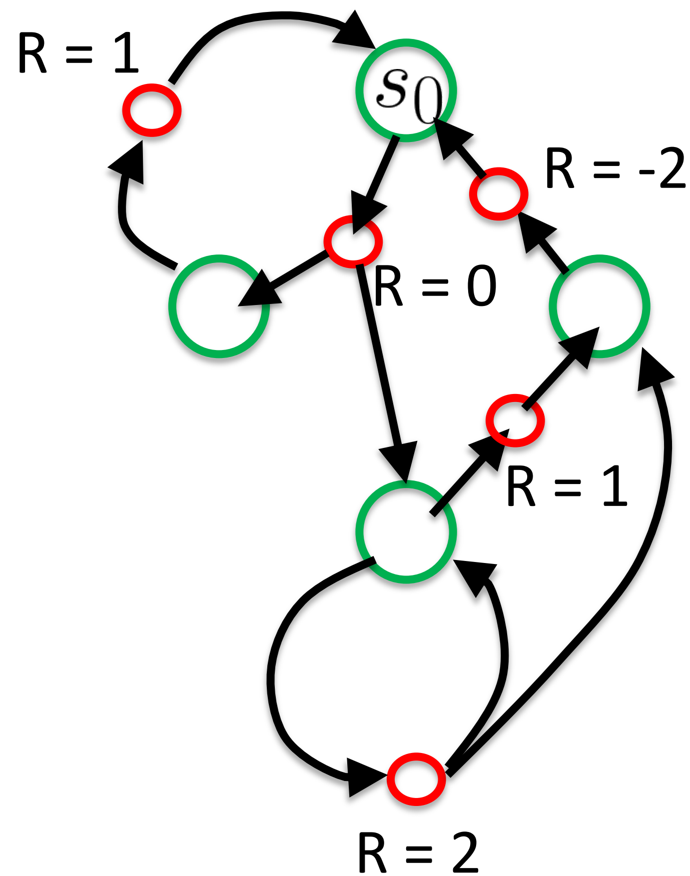
- State Space $\mathcal{S}$, start at s_0
- Action Space $\mathcal{A}$
- Rewards $R_h(s, a)$



MARKOV DECISION PROCESSES

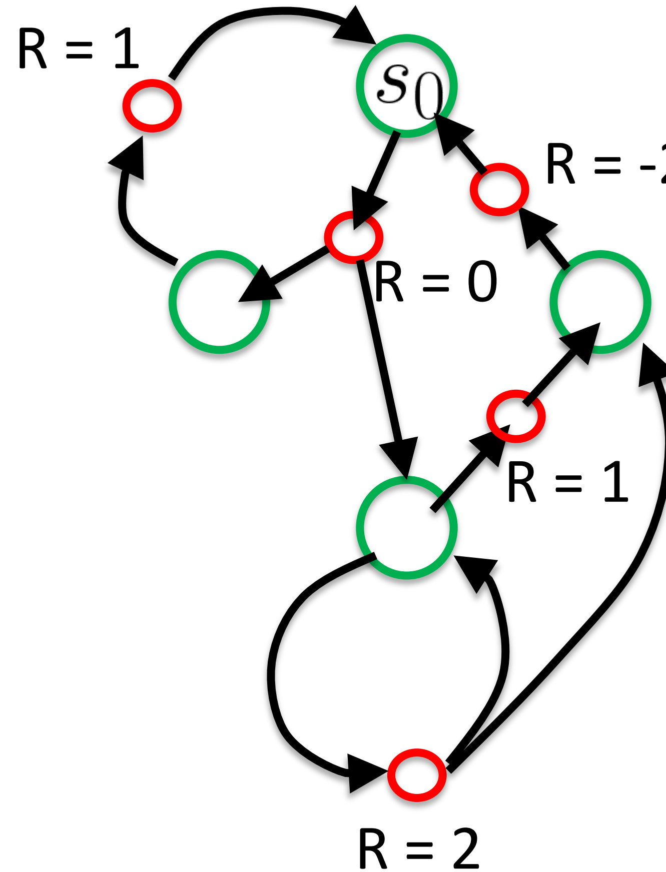
- State Space $\mathcal{S}$, start at s_0
- Action Space $\mathcal{A}$
- Rewards $R_h(s,a)$
- Transition Probabilities

$$\mathbb{T}_h(s'|s,a)$$



MARKOV DECISION PROCESSES

- State Space $\mathcal{S}$, start at s_0
- Action Space $\mathcal{A}$
- Rewards $R_h(s,a)$
- Transition Probabilities $\mathbb{T}_h(s'|s,a)$
- Horizon H

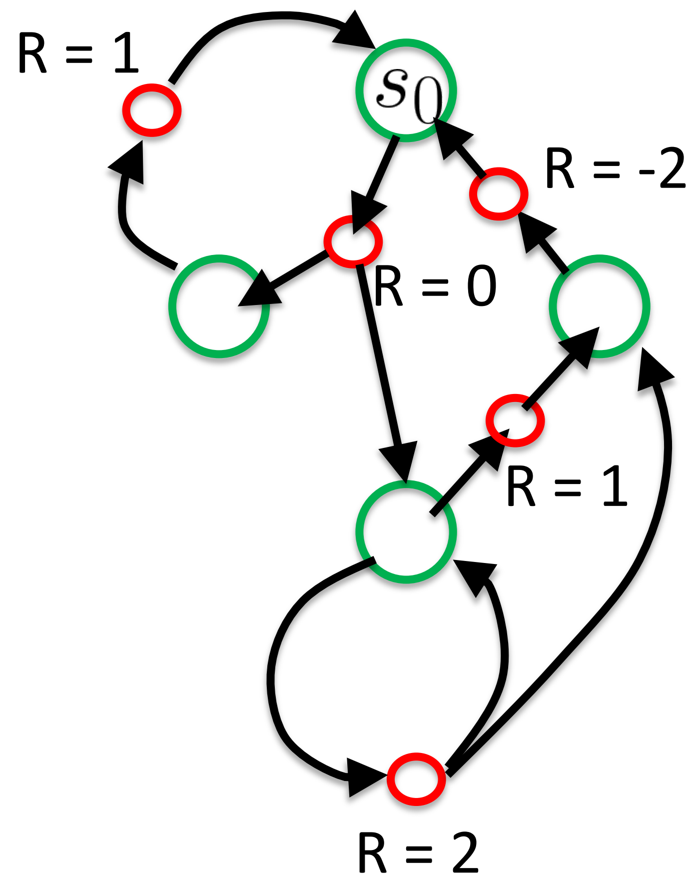


MARKOV DECISION PROCESSES

- State Space $\mathcal{S}$, start at s_0
- Action Space $\mathcal{A}$
- Rewards $R_h(s,a)$
- Transition Probabilities

$$\mathbb{T}_h(s'|s,a)$$

- Horizon H



Goal: Find a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes expected reward

MARKOV DECISION PROCESSES

Main problems:

(1) **Planning:** Given a full description of the MDP, **compute** an optimal policy

MARKOV DECISION PROCESSES

Main problems:

(1) **Planning:** Given a full description of the MDP, **compute** an optimal policy

e.g. value iteration, policy iteration, linear programming

MARKOV DECISION PROCESSES

Main problems:

(1) **Planning:** Given a full description of the MDP, **compute** an optimal policy

e.g. value iteration, policy iteration, linear programming

(2) **Learning:** Given budget of iterations with the environment (e.g. simulator, episodic), **learn** an optimal policy

MARKOV DECISION PROCESSES

Main problems:

(1) **Planning:** Given a full description of the MDP, **compute** an optimal policy

e.g. value iteration, policy iteration, linear programming

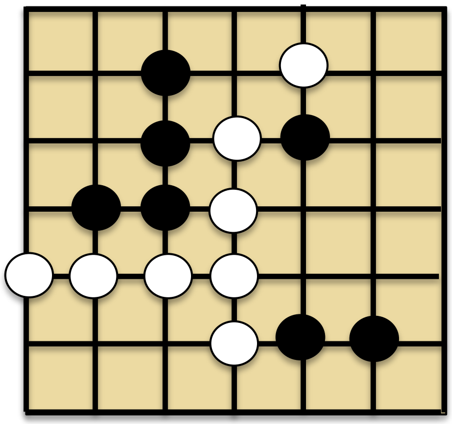
(2) **Learning:** Given budget of iterations with the environment (e.g. simulator, episodic), **learn** an optimal policy

e.g. model based, q-learning, actor-critic, policy gradient

And yet, for many applications tabular MDPs are **insufficient**

And yet, for many applications tabular MDPs are **insufficient**

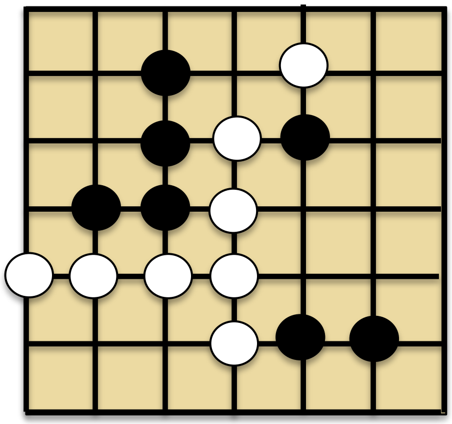
Too many states to write down or visit?



And yet, for many applications tabular MDPs are **insufficient**

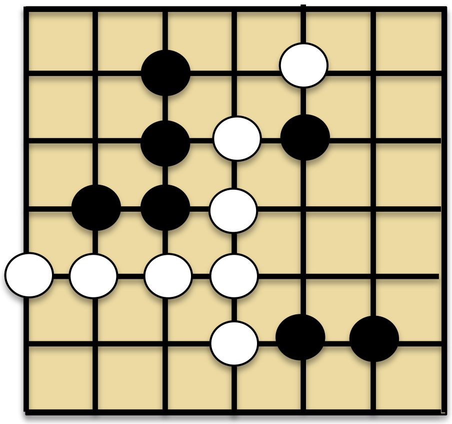
Too many states to write down or visit?

function approximation, block MDPs, etc

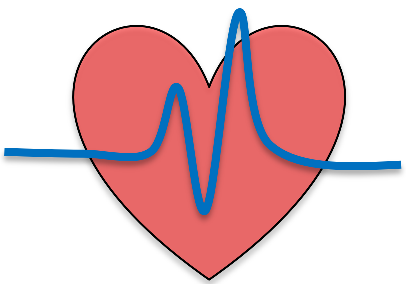


And yet, for many applications tabular MDPs are **insufficient**

Too many states to write down or visit?



function approximation, block MDPs, etc

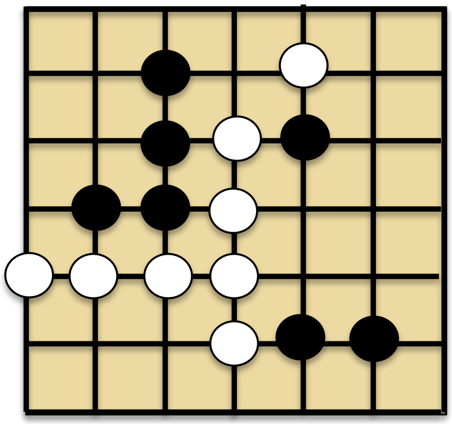


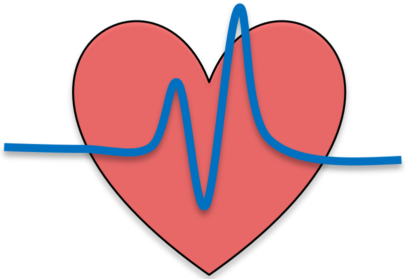
Cannot directly observe the full state?

And yet, for many applications tabular MDPs are **insufficient**

Too many states to write down or visit?

function approximation, block MDPs, etc





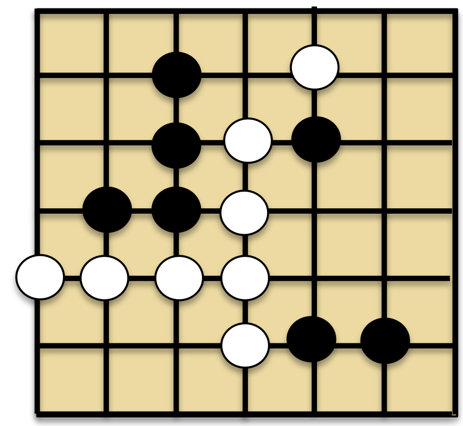
Cannot directly observe the full state?

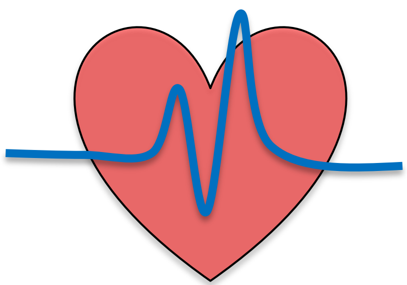
Partially observable MDPs (POMDPs)

And yet, for many applications tabular MDPs are **insufficient**

Too many states to write down or visit?

function approximation, block MDPs, etc





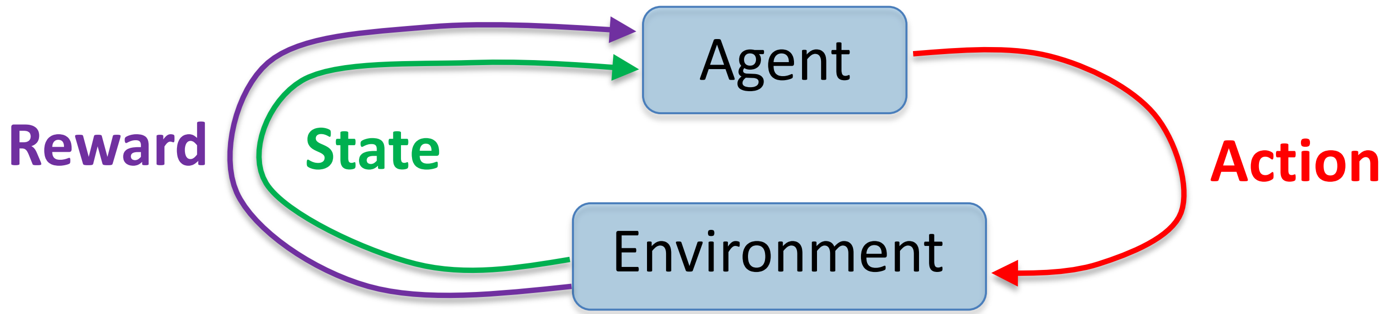
Cannot directly observe the full state?

Partially observable MDPs (POMDPs)

There is a rich understanding of how to augment the model, and still be able to bound **sample complexity**

WHAT ABOUT COMPUTATIONAL COMPLEXITY?

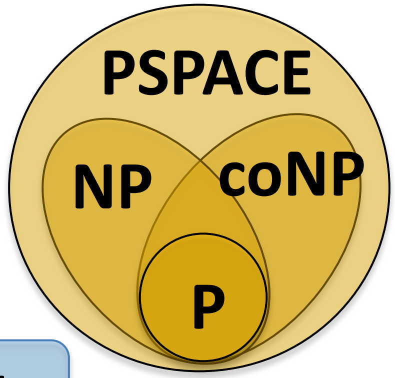
Returning to our earlier picture

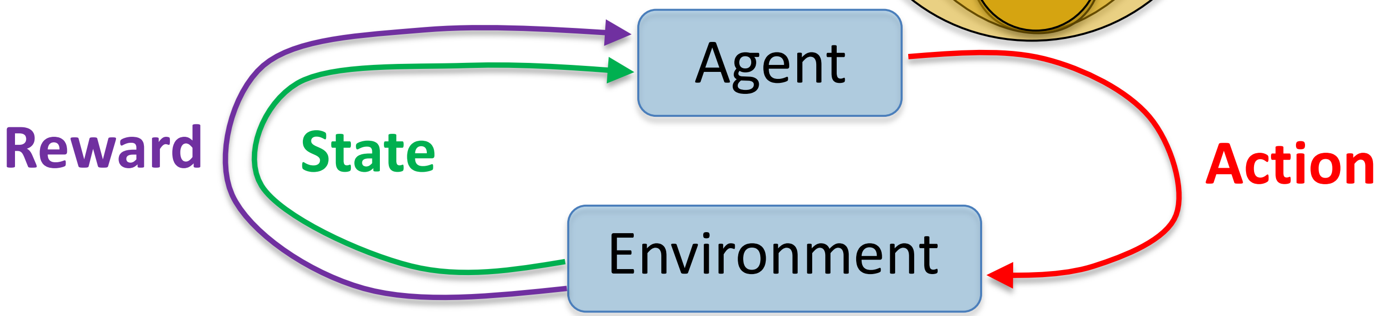


Modern RL is generally built on computationally intractable oracles

WHAT ABOUT COMPUTATIONAL COMPLEXITY?

Returning to our earlier picture

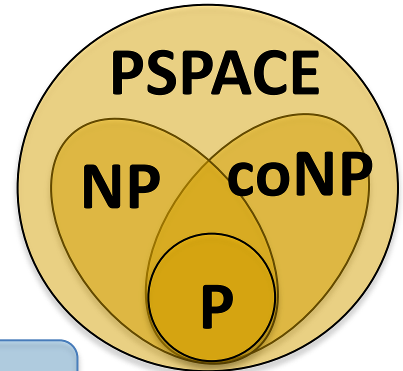


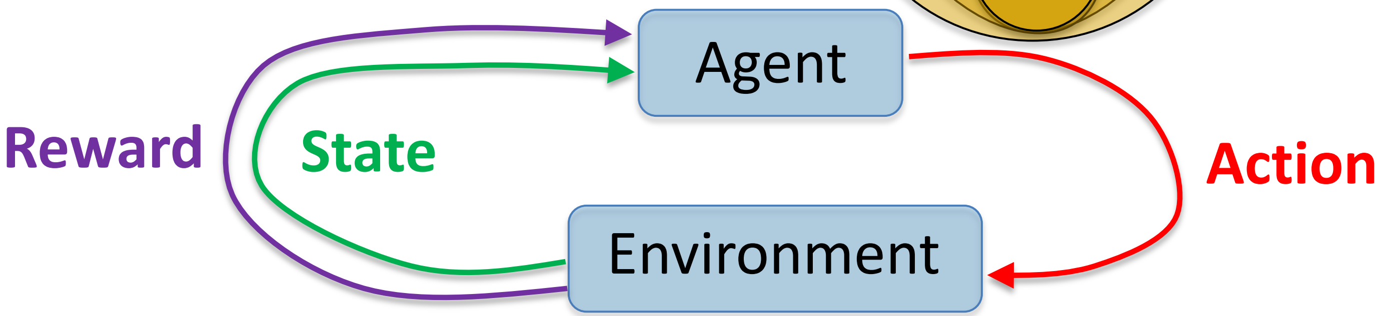


Modern RL is generally built on computationally intractable oracles

WHAT ABOUT COMPUTATIONAL COMPLEXITY?

Returning to our earlier picture



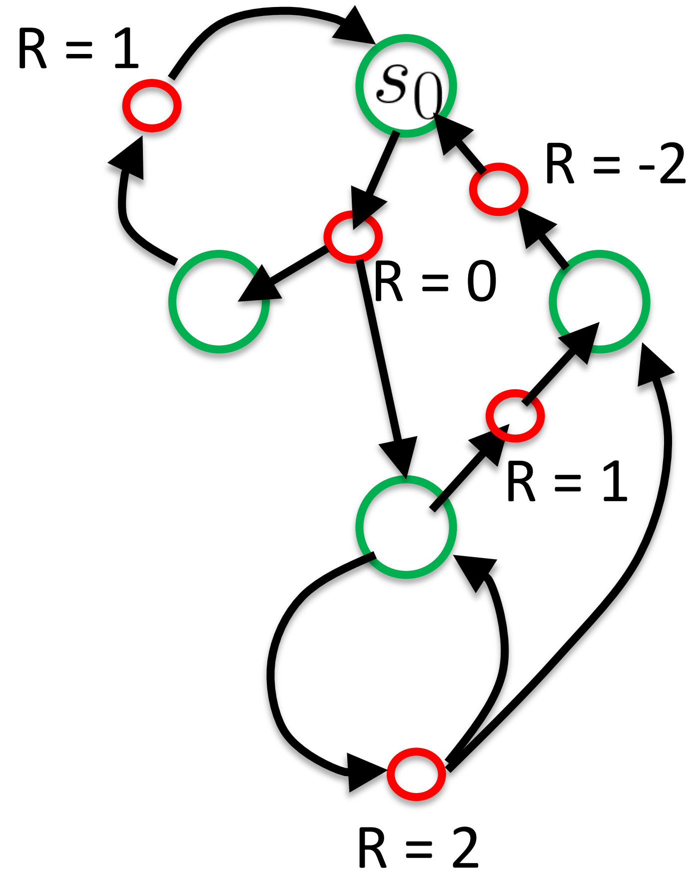


Modern RL is generally built on computationally intractable oracles

Are there computationally efficient algorithms with strong end-to-end provable guarantees?

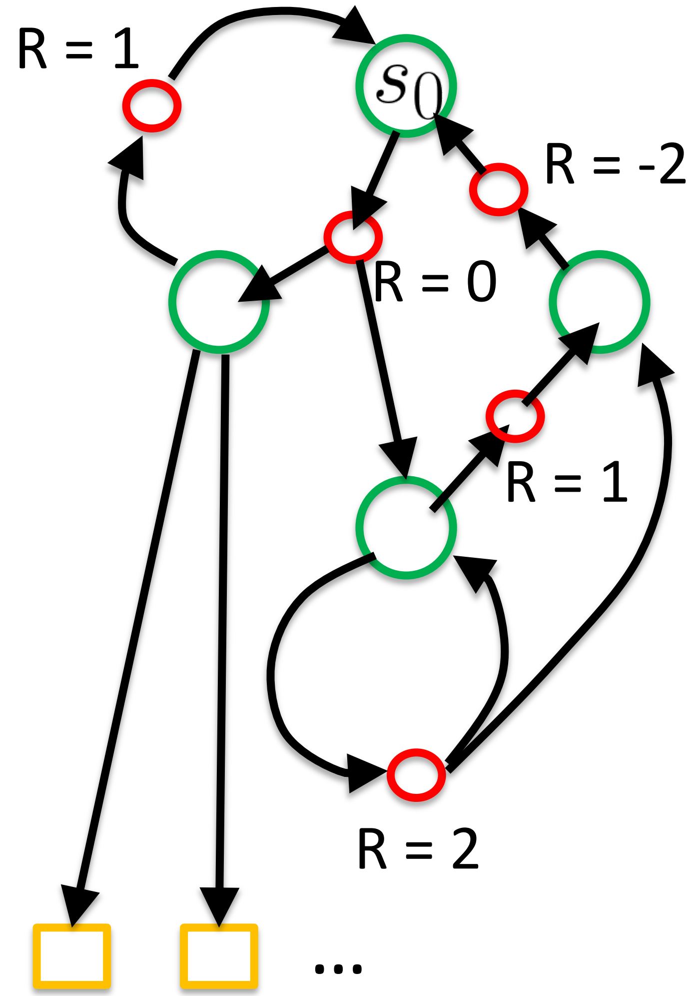
MARKOV DECISION PROCESSES

- State Space $\mathcal{S}$, start at s_0
- Action Space $\mathcal{A}$
- Rewards $R_h(s, a)$
- Transition Probabilities $\mathbb{T}_h(s'|s, a)$
- Horizon H



PARTIALLY OBSERVABLE MDPS (POMDPS)

- State Space $\mathcal{S}$, start at s_0
- Action Space $\mathcal{A}$
- Rewards $R_h(s,a)$
- Transition Probabilities $\mathbb{T}_h(s'|s,a)$
- Horizon H
- Observation Space $\mathcal{O}$ and probabilities $\mathbb{O}_h(o|s)$



OUTLINE

Part I: Introduction

- Models and Problems
- Hardness and Beyond Worst-Case Analysis
- Our Results

Part II: Planning

Part III: Learning

- Approximate MDPs via Barycentric Spanners

OUTLINE

Part I: Introduction

- Models and Problems
- **Hardness and Beyond Worst-Case Analysis**
- Our Results

Part II: Planning

Part III: Learning

- Approximate MDPs via Barycentric Spanners

PLANNING IS HARD

Classic lower bound:

Theorem [Papadimitriou, Tsitsiklis]: Optimal planning in a POMDP is PSPACE hard

THE CURSE OF HISTORY

Can you succinctly represent an optimal policy?

THE CURSE OF HISTORY

Can you succinctly represent an optimal policy?

MDPs	POMDPs
Optimal action only depends on current state $\pi : \mathcal{S} \rightarrow \mathcal{A}$	

THE CURSE OF HISTORY

Can you succinctly represent an optimal policy?

MDPs	POMDPs
Optimal action only depends on current state	**Optimal action depends on action/observation history**
$\pi : \mathcal{S} \rightarrow \mathcal{A}$	$\pi : \mathcal{A} \times \mathcal{O} \cdots \times \mathcal{O} \rightarrow \mathcal{A}$

THE CURSE OF HISTORY

Can you succinctly represent an optimal policy?

MDPs	POMDPs
Optimal action only depends on current state	**Optimal action depends on action/observation history**
$\pi : \mathcal{S} \to \mathcal{A}$	$\pi : \mathcal{A} \times \mathcal{O} \cdots \times \mathcal{O} \to \mathcal{A}$
	Alternatively, it depends on the current belief
	$\pi : \Delta^{\mathcal{S}} \to \mathcal{A}$

THE CURSE OF HISTORY

Can you succinctly represent an optimal policy?

MDPs	POMDPs
Optimal action only depends on current state	**Optimal action depends on action/observation history**
$\pi : \mathcal{S} \rightarrow \mathcal{A}$	$\pi : \mathcal{A} \times \mathcal{O} \cdots \times \mathcal{O} \rightarrow \mathcal{A}$
	Alternatively, it depends on the current belief
	$\pi : \Delta^{\mathcal{S}} \rightarrow \mathcal{A}$

Natural approaches use exponential space $(|\mathcal{A}||\mathcal{O}|)^H$ or $C^{|\mathcal{S}|}$

PLANNING IS EVEN HARDER

Even worse news:

Theorem [Golowich, Moitra, Rohatgi]: Unless the exponential time hierarchy collapses, there is no polynomial sized description of an approximately optimal policy

PLANNING IS EVEN HARDER

Even worse news:

Theorem [Golowich, Moitra, Rohatgi]: Unless the exponential time hierarchy collapses, there is no polynomial sized description of an approximately optimal policy

Why should real-world POMDPs have succinct descriptions of good policies?

BEYOND WORST-CASE ANALYSIS

The hard instances have a curious feature:

“The observations don’t tell you anything about the state”

BEYOND WORST-CASE ANALYSIS

The hard instances have a curious feature:

"The observations don't tell you anything about the state"

But what if they are at least somewhat informative?

"The observations leak some information about the state"

BEYOND WORST-CASE ANALYSIS

The hard instances have a curious feature:

“The observations don’t tell you anything about the state”

But what if they are at least somewhat informative?

“The observations leak some information about the state”

Could this enable tractable planning/learning?

BEYOND WORST-CASE ANALYSIS

Definition: We say the POMDP is γ-observable if for all h and all distributions b,b' on states we have

$$\|\mathbb{O}_h b - \mathbb{O}_h b'\|_1 \geq \gamma \|b - b'\|_1$$

i.e. well-separated distributions on states lead to well-separated distributions on observations

BEYOND WORST-CASE ANALYSIS

Definition: We say the POMDP is γ-observable if for all h and all distributions b,b' on states we have

$$\|\mathbb{O}_h b - \mathbb{O}_h b'\|_1 \geq \gamma \|b - b'\|_1$$

i.e. well-separated distributions on states lead to well-separated distributions on observations

Introduced by **[Even-Dar, Kakade, Mansour]** for understanding stability of beliefs in HMMs under misspecification

BEYOND WORST-CASE ANALYSIS

Definition: We say the POMDP is γ-observable if for all h and all distributions b, b' on states we have

$$\|\mathbb{O}_h b - \mathbb{O}_h b'\|_1 \geq \gamma \|b - b'\|_1$$

i.e. well-separated distributions on states lead to well-separated distributions on observations

Introduced by **[Even-Dar, Kakade, Mansour]** for understanding stability of beliefs in HMMs under misspecification

Key Point: No assumption on transition dynamics like e.g. **deterministic transitions** or **mixing (under every possible policy)**

OUTLINE

Part I: Introduction

- Models and Problems
- Hardness and Beyond Worst-Case Analysis
- Our Results

Part II: Planning

Part III: Learning

- Approximate MDPs via Barycentric Spanners

OUTLINE

Part I: Introduction

- Models and Problems
- Hardness and Beyond Worst-Case Analysis
- **Our Results**

Part II: Planning

Part III: Learning

- Approximate MDPs via Barycentric Spanners

MAIN RESULTS (PLANNING)

There is a quasi-polynomial time algorithm for planning under observability:

Theorem [Golowich, Moitra, Rohatgi]: Given the description of a γ-observable POMDP there is an algorithm running in time

$$H(|\mathcal{O}||\mathcal{A}|)^{C\log(|\mathcal{S}|H/\epsilon)/\gamma^4}$$

that outputs an ϵ-suboptimal policy

MAIN RESULTS (PLANNING)

There is a quasi-polynomial time algorithm for planning under observability:

Theorem [Golowich, Moitra, Rohatgi]: Given the description of a γ-observable POMDP there is an algorithm running in time

$$H(|\mathcal{O}||\mathcal{A}|)^{C \log(|\mathcal{S}|H/\epsilon)/\gamma^4}$$

that outputs an ϵ-suboptimal policy

Key Idea: The Bayes filter is **exponentially** stable

compute posterior on states, given actions/observations

MAIN RESULTS (PLANNING), CONTINUED

Moreover these results are tight

Theorem [Golowich, Moitra, Rohatgi]: Under the Exponential Time Hypothesis, there is no algorithm running in time

$$(|\mathcal{S}||\mathcal{A}|H|\mathcal{O}|)^{o(\log(|\mathcal{S}||\mathcal{A}|H|\mathcal{O}|/\epsilon)/\gamma)}$$

for finding an ϵ-suboptimal policy in a γ-observable POMDP

MAIN RESULTS (PLANNING), CONTINUED

Moreover these results are tight

Theorem [Golowich, Moitra, Rohatgi]: Under the Exponential Time Hypothesis, there is no algorithm running in time

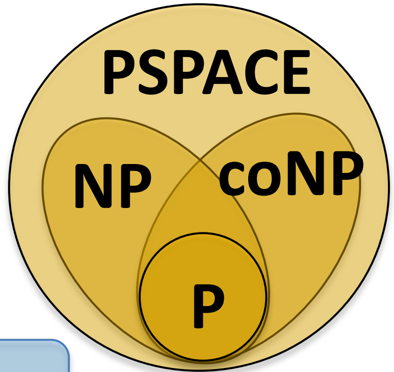
$$(|\mathcal{S}||\mathcal{A}|H|\mathcal{O}|)^{o(\log(|\mathcal{S}||\mathcal{A}|H|\mathcal{O}|/\epsilon)/\gamma)}$$

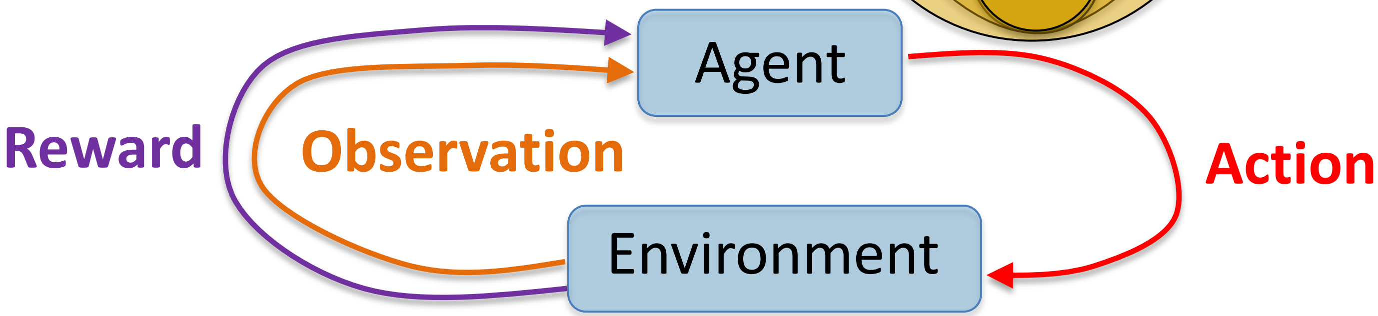
for finding an ϵ-suboptimal policy in a γ-observable POMDP

It's hard even in the **lossy case**, where you observe the state with probability γ independently at each step

WHAT ABOUT LEARNING?

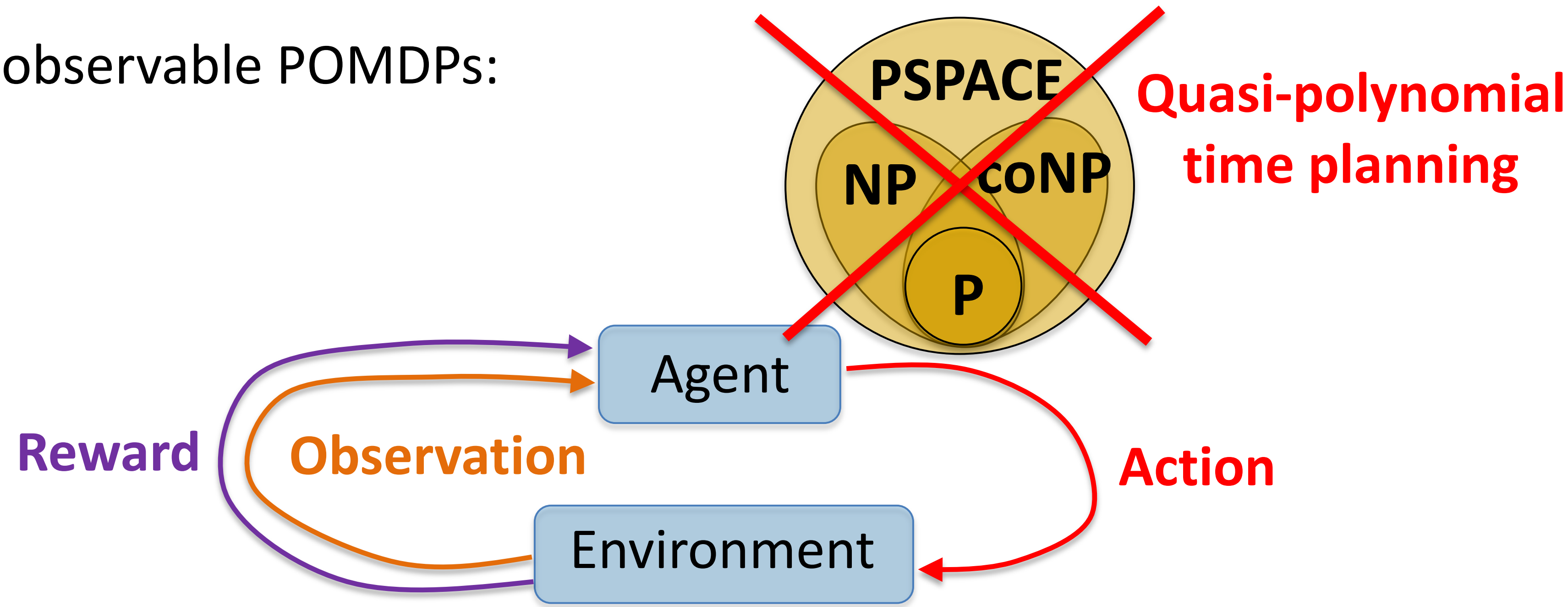
In observable POMDPs:





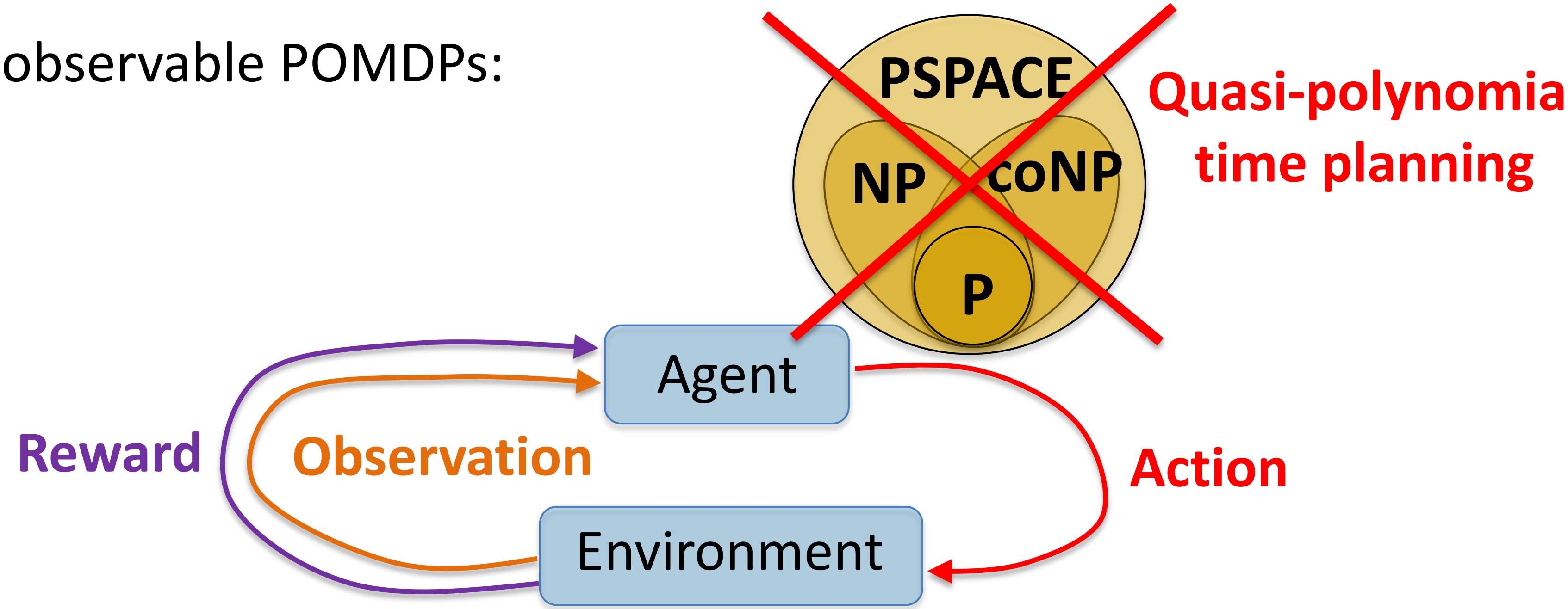
WHAT ABOUT LEARNING?

In observable POMDPs:



WHAT ABOUT LEARNING?

In observable POMDPs:



Can we build a learning algorithm on top of this primitive?

SAMPLE EFFICIENT LEARNING?

Assumption 1: The POMDP is undercomplete, i.e. $|\mathcal{S}| \leq |\mathcal{O}|$

And moreover $\sigma_{min}(\mathbb{O}_h) \geq \alpha$ for all h

SAMPLE EFFICIENT LEARNING?

Assumption 1: The POMDP is undercomplete, i.e. $|\mathcal{S}| \leq |\mathcal{O}|$
And moreover $\sigma_{min}(\mathbb{O}_h) \geq \alpha$ for all h

Theorem [Jin, Kakade, Krishnamurthy, Liu]: Given access to an **optimistic planning oracle**, there is an algorithm that uses

$$\mathrm{poly}(|\mathcal{S}|, |\mathcal{A}|, H, |\mathcal{O}|, 1/\alpha)$$

samples and finds an ϵ-suboptimal policy under **Assumption 1**

SAMPLE EFFICIENT LEARNING?

Assumption 1: The POMDP is undercomplete, i.e. $|\mathcal{S}| \leq |\mathcal{O}|$
And moreover $\sigma_{min}(\mathbb{O}_h) \geq \alpha$ for all h

Theorem [Jin, Kakade, Krishnamurthy, Liu]: Given access to an **optimistic planning oracle**, there is an algorithm that uses

$$\text{poly}(|\mathcal{S}|, |\mathcal{A}|, H, |\mathcal{O}|, 1/\alpha)$$

samples and finds an ϵ-suboptimal policy under **Assumption 1**

i.e. given a constrained, non-convex set of POMDPs, find the maximum value achievable by any policy in the set

SAMPLE EFFICIENT LEARNING?

Assumption 1: The POMDP is undercomplete, i.e. $|\mathcal{S}| \leq |\mathcal{O}|$
And moreover $\sigma_{min}(\mathbb{O}_h) \geq \alpha$ for all h

Theorem [Jin, Kakade, Krishnamurthy, Liu]: Given access to an **optimistic planning oracle**, there is an algorithm that uses

$$\mathrm{poly}(|\mathcal{S}|, |\mathcal{A}|, H, |\mathcal{O}|, 1/\alpha)$$

samples and finds an ϵ-suboptimal policy under **Assumption 1**

i.e. given a constrained, non-convex set of POMDPs, find the maximum value achievable by any policy in the set

But optimism is very hard!

MAIN RESULTS (LEARNING)

We show how to solve learning by using barycentric spanners to construct a policy cover. As a result:

Theorem [Golowich, Moitra, Rohatgi]: There is an algorithm with running time and sample complexity

$$(|\mathcal{O}||\mathcal{A}|)^{C \log(H|\mathcal{S}||\mathcal{O}|/\epsilon\gamma)/\gamma^4}$$

that outputs an ϵ-suboptimal policy in a γ-observable POMDP

MAIN RESULTS (LEARNING)

We show how to solve learning by using barycentric spanners to construct a policy cover. As a result:

Theorem [Golowich, Moitra, Rohatgi]: There is an algorithm with running time and sample complexity

$$(|\mathcal{O}||\mathcal{A}|)^{C \log(H|\mathcal{S}||\mathcal{O}|/\epsilon\gamma)/\gamma^4}$$

that outputs an ϵ-suboptimal policy in a γ-observable POMDP

These are the first end-to-end algorithmic guarantees for learning POMDPs, without oracles or strong assumptions about the model dynamics

OUTLINE

Part I: Introduction

- Models and Problems
- Hardness and Beyond Worst-Case Analysis
- Our Results

Part II: Planning

Part III: Learning

- Approximate MDPs via Barycentric Spanners

OUTLINE

Part I: Introduction

- Models and Problems
- Hardness and Beyond Worst-Case Analysis
- Our Results

Part II: Planning

Part III: Learning

- Approximate MDPs via Barycentric Spanners

BELIEF CONTRACTION

Theorem: Fix any γ-observable POMDP and policy π . Then

$$\mathbb{E}_\tau[\|b_t - b'_t\|_1] \leq (1-\gamma^4)^t |\mathcal{S}|$$

posterior, starting from arbitrary belief state

posterior, starting from uniform belief state

where τ is the trajectory from the POMDP by playing π

BELIEF CONTRACTION

Theorem: Fix any γ-observable POMDP and policy π . Then

$$\mathbb{E}_\tau[\|b_t - b'_t\|_1] \leq (1-\gamma^4)^t |\mathcal{S}|$$

posterior, starting from arbitrary belief state

posterior, starting from uniform belief state

where τ is the trajectory from the POMDP by playing π

Parallels well-known stability results for Kalman filtering

BELLMAN UPDATES FOR POMDPS

Can find an optimal policy through:

$$\text{Value}(x) = E\Big[\max_{\text{actions } a} \text{Reward}(a) + \text{Value}(x')\Big]$$

current action/obs. sequence

new action/obs. sequence

latent state sampled from current belief

TRUNCATED BELLMAN UPDATES

Belief contraction allows us to **truncate**

TRUNCATED BELLMAN UPDATES

Belief contraction allows us to **truncate**

$$\text{Value}(x) = \tilde{E}\Big[\max_{\text{actions } a} \text{Reward}(a) + \text{Value}(x')\Big]$$

length t window

length t window

latent state sampled from truncated belief, with uniform prior

TRUNCATED BELLMAN UPDATES

Belief contraction allows us to **truncate**

$$\text{Value}(x) = \tilde{E}\left[\max_{\text{actions } a} \text{Reward}(a) + \text{Value}(x')\right]$$

length t window

length t window

latent state sampled from truncated belief, with uniform prior

We only need a quasi-polynomial number of belief states

OUTLINE

Part I: Introduction

- Models and Problems
- Hardness and Beyond Worst-Case Analysis
- Our Results

Part II: Planning

Part III: Learning

- Approximate MDPs via Barycentric Spanners

OUTLINE

Part I: Introduction

- Models and Problems
- Hardness and Beyond Worst-Case Analysis
- Our Results

Part II: Planning

Part III: Learning

- **Approximate MDPs via Barycentric Spanners**

APPROXIMATION BY MDPS

Corollary: Any γ-observable POMDP P can be approximated by an MDP M with a quasi-polynomial number of states

APPROXIMATION BY MDPS

Corollary: Any γ-observable POMDP P can be approximated by an MDP M with a quasi-polynomial number of states

(1) P can be thought of as an MDP on belief states

APPROXIMATION BY MDPS

Corollary: Any γ-observable POMDP P can be approximated by an MDP M with a quasi-polynomial number of states

(1) P can be thought of as an MDP on belief states

(2) Construct M as follows:

states = length L sequences of actions/observations

APPROXIMATION BY MDPS

Corollary: Any γ-observable POMDP P can be approximated by an MDP M with a quasi-polynomial number of states

(1) P can be thought of as an MDP on belief states

(2) Construct M as follows:

states = length L sequences of actions/observations

transitions = shift in/out the newest/oldest actions/obs.

APPROXIMATION BY MDPS

Corollary: Any γ-observable POMDP P can be approximated by an MDP M with a quasi-polynomial number of states

(1) P can be thought of as an MDP on belief states

(2) Construct M as follows:

states = length L sequences of actions/observations

transitions = shift in/out the newest/oldest actions/obs.

(3) States in M can mapped to beliefs (using a uniform prior).

By belief contraction, M and P approximate each other

APPROXIMATION BY MDPS

Corollary: Any γ-observable POMDP P can be approximated by an MDP M with a quasi-polynomial number of states

Can we learn M efficiently?

APPROXIMATION BY MDPS

Corollary: Any γ-observable POMDP P can be approximated by an MDP M with a quasi-polynomial number of states

Can we learn M efficiently?

Simplification: For any latent state x in P, and any timestep h, there is some policy π that visits x at h with nonnegligible probability

APPROXIMATION BY MDPS

Corollary: Any γ-observable POMDP P can be approximated by an MDP M with a quasi-polynomial number of states

Can we learn M efficiently?

Simplification: For any latent state x in P, and any timestep h, there is some policy π that visits x at h with nonnegligible probability

How can we find a mixture of policies that visits all latent states?

BARYCENTRIC SPANNERS

Definition: Given a set $\mathcal{X} \subseteq \mathbb{R}^d$, a λ**-approximate barycentric spanner** is a set $\mathcal{C} \subseteq \mathcal{X}$ of size d such that every point in $\mathcal{X}$ can be expressed as a linear combination of points in $\mathcal{C}$ with coefficients in the range $[-\lambda, \lambda]$

BARYCENTRIC SPANNERS

Definition: Given a set $\mathcal{X} \subseteq \mathbb{R}^d$, a λ**-approximate barycentric spanner** is a set $\mathcal{C} \subseteq \mathcal{X}$ of size d such that every point in $\mathcal{X}$ can be expressed as a linear combination of points in $\mathcal{C}$ with coefficients in the range $[-\lambda, \lambda]$

Theorem [Awerbuch, Kleinberg '04]: Given an oracle for optimizing linear functions over $\mathcal{X}$, there is a polynomial time algorithm for constructing a λ-approximate barycentric spanner with

$$O(d^2 \log_\lambda d)$$

calls to the optimization oracle (assuming $\mathcal{X}$ is compact)

POLICY COVERS

Now let

$$\mathcal{X} = \text{set of all distributions on observations at step h that can be obtained by a policy}$$

POLICY COVERS

Now let

$$\mathcal{X} = \text{set of all distributions on observations at step h that can be obtained by a policy}$$

Claim: By observability, if we can construct policies

$$\pi_1, \pi_2, \ldots, \pi_{|\mathcal{O}|}$$

whose induced distributions on observations at step h are an approximate barycentric spanner

POLICY COVERS

Now let

$$\mathcal{X} = \text{set of all distributions on observations at step h that can be obtained by a policy}$$

Claim: By observability, if we can construct policies

$$\pi_1, \pi_2, \ldots, \pi_{|\mathcal{O}|}$$

whose induced distributions on observations at step h are an approximate barycentric spanner, **we must visit each latent state with nonnegligible probability**

ITERATIVE EXPLORATION

Our approach is:

$$M_h$$

MDP that approximates P up to step h

ITERATIVE EXPLORATION

Our approach is:

$$M_h \longrightarrow X_h$$

MDP that approximates P up to step h

Barycentric spanner for observation distributions at step h

ITERATIVE EXPLORATION

Our approach is:

$$M_h \longrightarrow X_h$$

MDP that approximates P up to step h

Barycentric spanner for observation distributions at step h

$\downarrow$ **"explorability"**

Reaches all latent states

ITERATIVE EXPLORATION

Our approach is:

$M_h \longrightarrow X_h$

MDP that approximates P up to step h

Barycentric spanner for observation distributions at step h

"explorability"

Reaches all latent states

Estimate next layer of transitions

ITERATIVE EXPLORATION

Our approach is:

M_{h+1} → X_h

MDP that approximates P up to step h+1

Barycentric spanner for observation distributions at step h

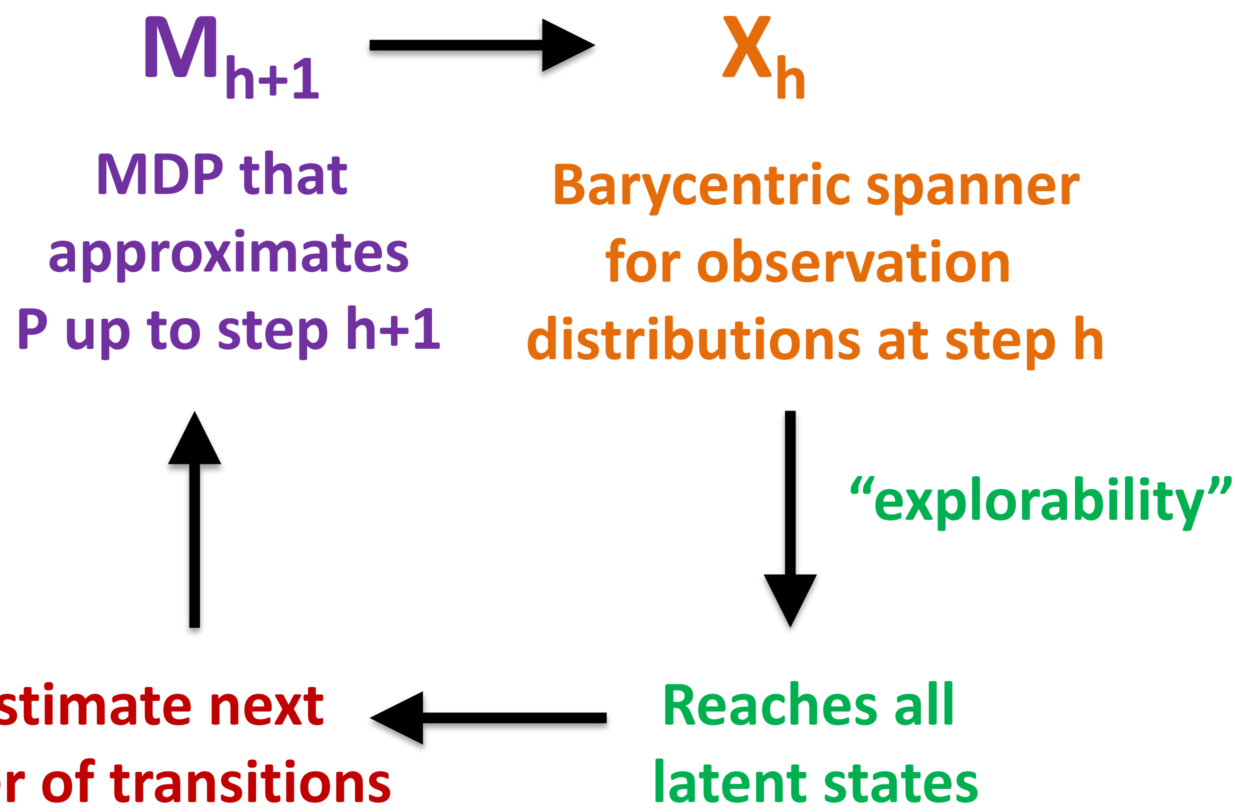
"explorability"

Reaches all latent states

Estimate next layer of transitions

ITERATIVE EXPLORATION

Our approach is:



Without explorability, need more complex measure of progress

LOOKING FORWARD

To get end-to-end algorithmic guarantees, we need to explore new assumptions and frameworks

LOOKING FORWARD

To get end-to-end algorithmic guarantees, we need to explore new assumptions and frameworks

In **[Golowich, Moitra]**, we took a learning-augmented algorithms approach:

"Can you improve Q-learning with advice?"

LOOKING FORWARD

To get end-to-end algorithmic guarantees, we need to explore new assumptions and frameworks

In **[Golowich, Moitra]**, we took a learning-augmented algorithms approach:

"Can you improve Q-learning with advice?"

Takeaway: Improved regret bounds, where you only need to explore state-action pairs with substantially inaccurate predictions, **even without knowing which ones are accurate in advance**

Summary:

- Modern RL is built on computationally intractable oracles. **Are there end-to-end guarantees?**
- Quasi-polynomial time algorithm for planning in **observable** POMDPs, no assumption on dynamics
- New framework for learning without optimism

Summary:

- Modern RL is built on computationally intractable oracles. **Are there end-to-end guarantees?**
- Quasi-polynomial time algorithm for planning in **observable** POMDPs, no assumption on dynamics
- New framework for learning without optimism

Thanks! Any Questions?