Algorithmic Aspects of Reinforcement Learning I

Ankur Moitra (MIT)

February 24th, ITA+ALT Tutorial

TUTORIAL GOALS

(1) Overview of theoretical foundations of RL

TUTORIAL GOALS

(1) Overview of theoretical foundations of RL

(2) Gaps in our *algorithmic* understanding

TUTORIAL GOALS

(1) Overview of theoretical foundations of RL

(2) Gaps in our *algorithmic* understanding

(3) Deep dive into some success stories, emphasizing connections to other areas

INTRODUCTION

Success stories of reinforcement learning:

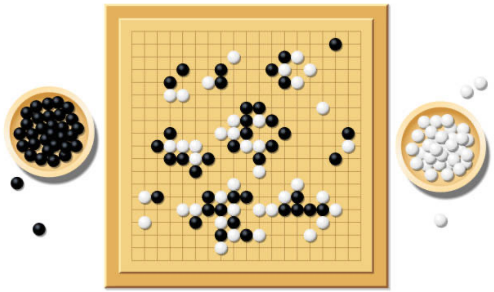
robotic manipulation

INTRODUCTION

Success stories of reinforcement learning:

robotic manipulation

playing strategic games

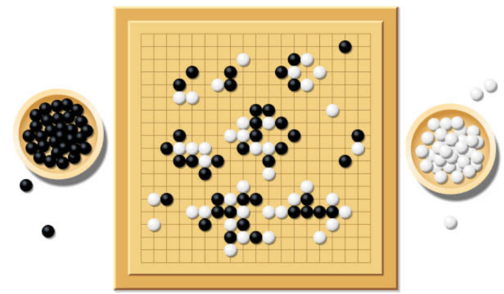


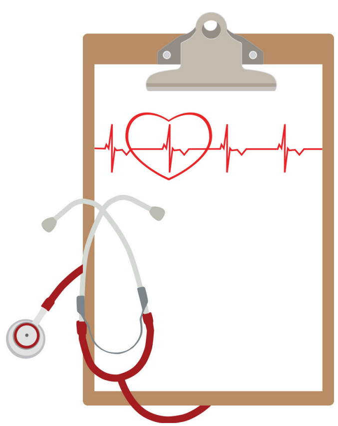
INTRODUCTION

Success stories of reinforcement learning:

robotic manipulation

playing strategic games

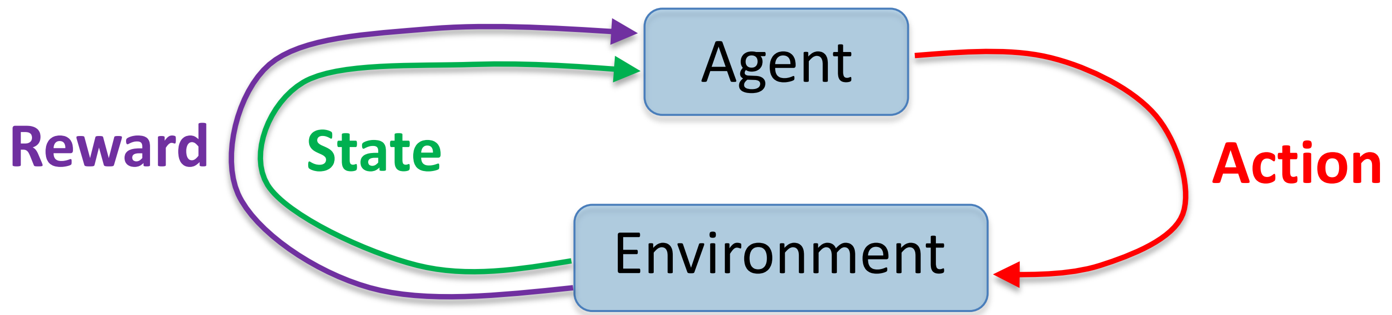




personalized treatment in medicine

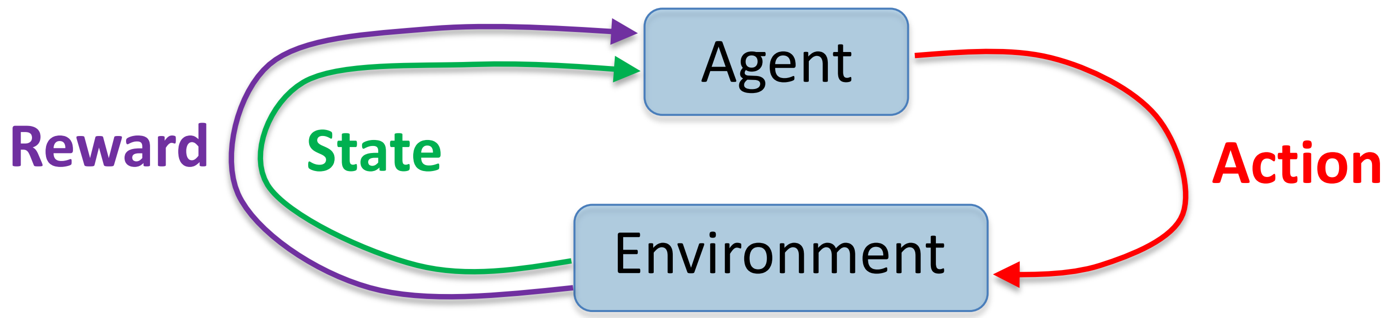
INTRODUCTION

Goal: Agent learns by interacting with the environment



INTRODUCTION

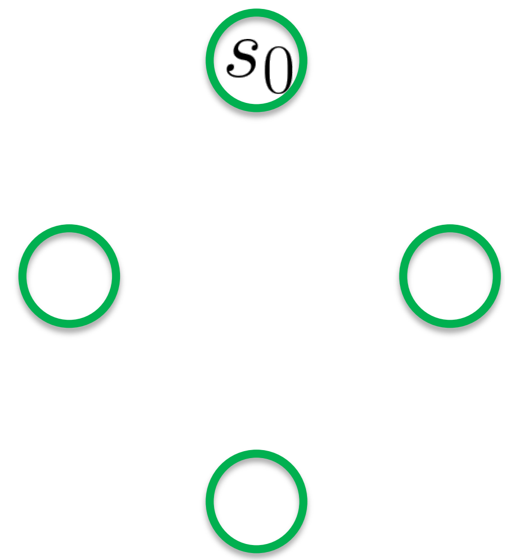
Goal: Agent learns by interacting with the environment



The basic model is a **Markov Decision Process**

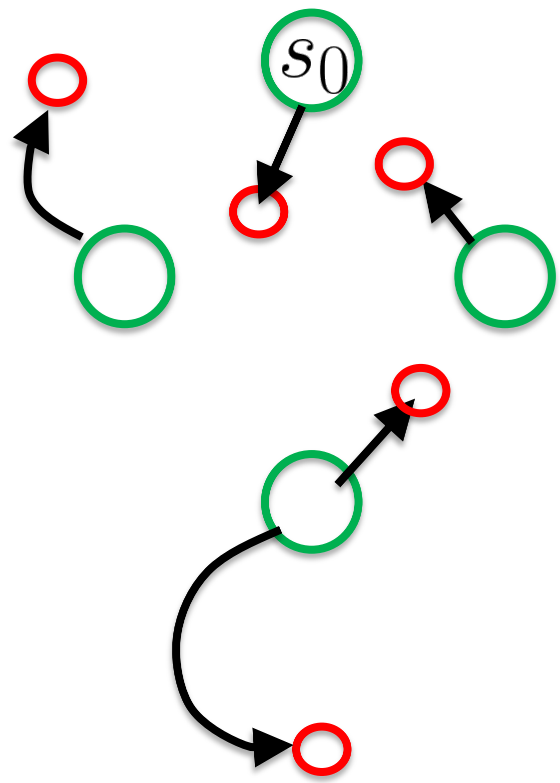
MARKOV DECISION PROCESSES

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



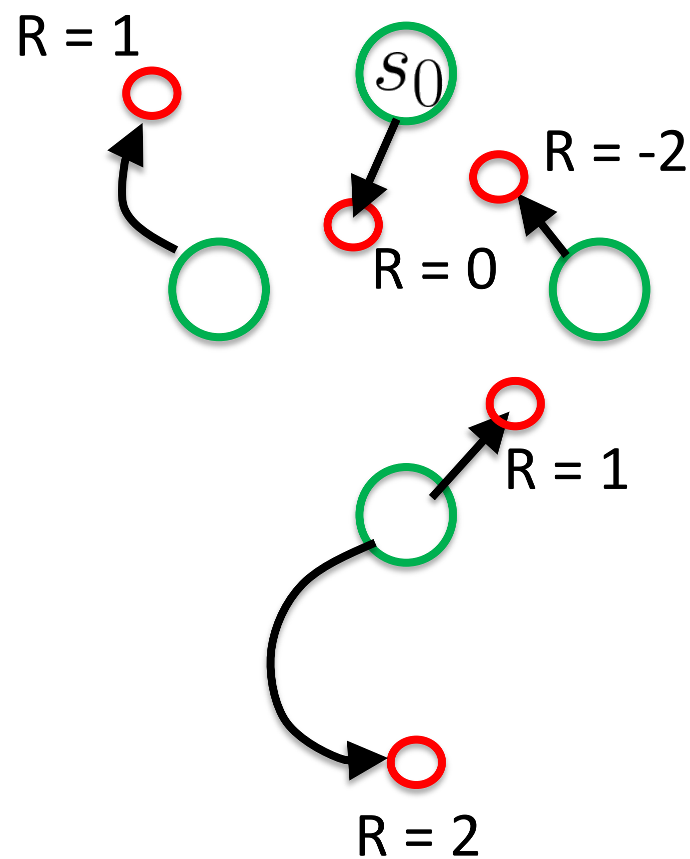
MARKOV DECISION PROCESSES

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



MARKOV DECISION PROCESSES

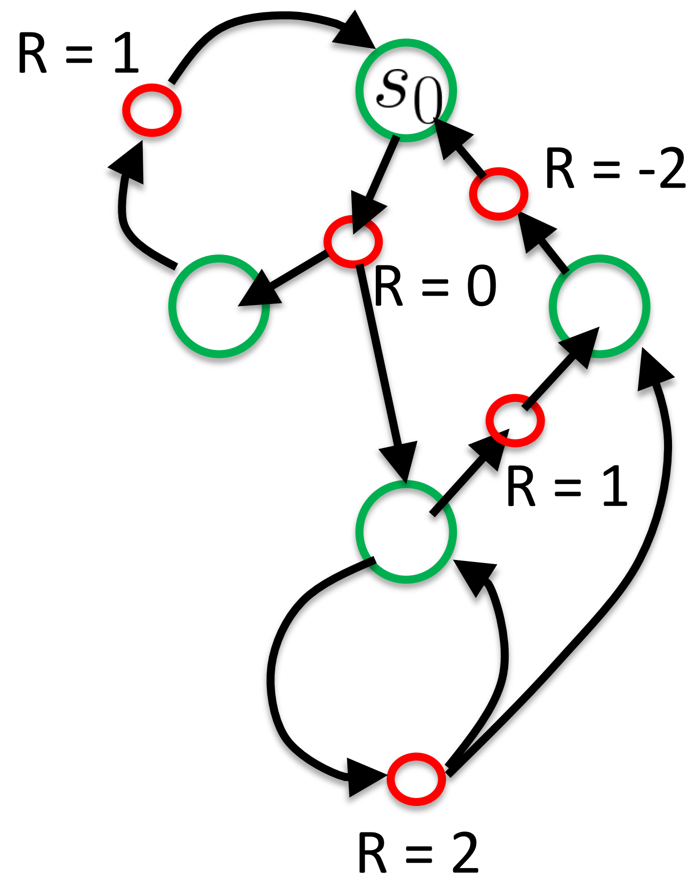
- State Space $\mathcal{S}$, start at s_0
- Action Space $\mathcal{A}$
- Rewards $R_h(s, a)$



MARKOV DECISION PROCESSES

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

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

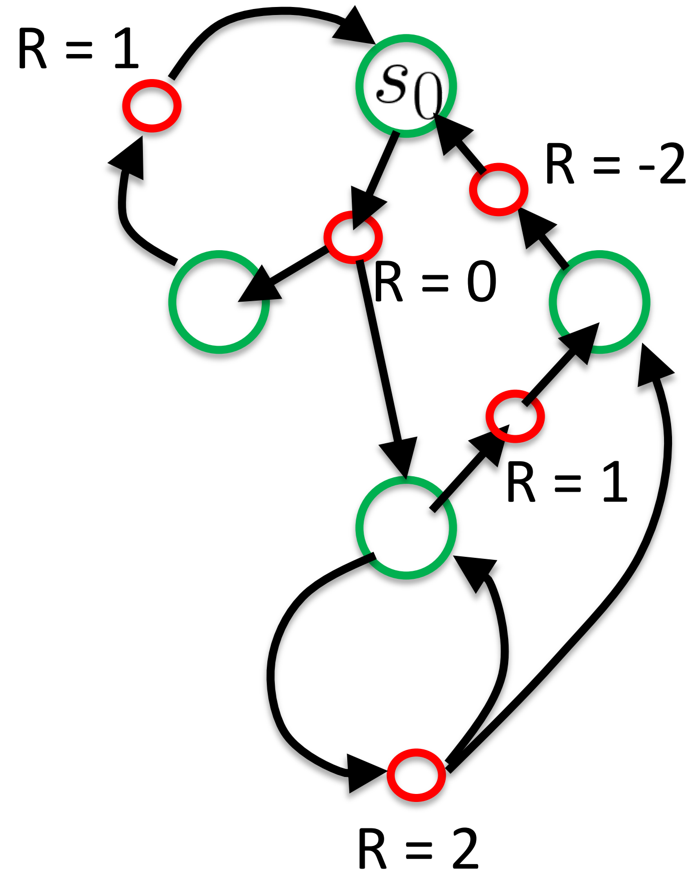


MARKOV DECISION PROCESSES

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

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

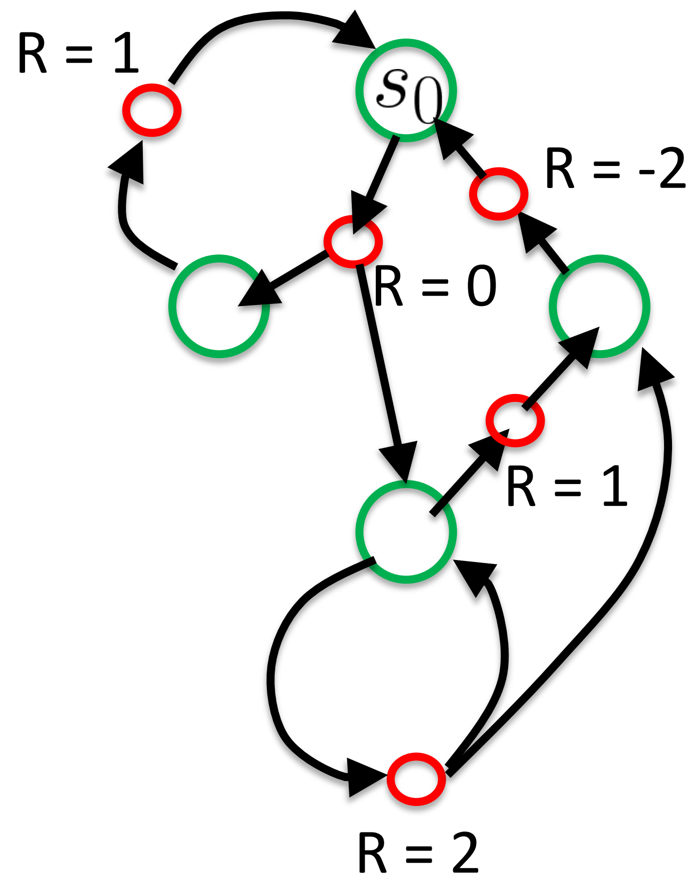
- Horizon H



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



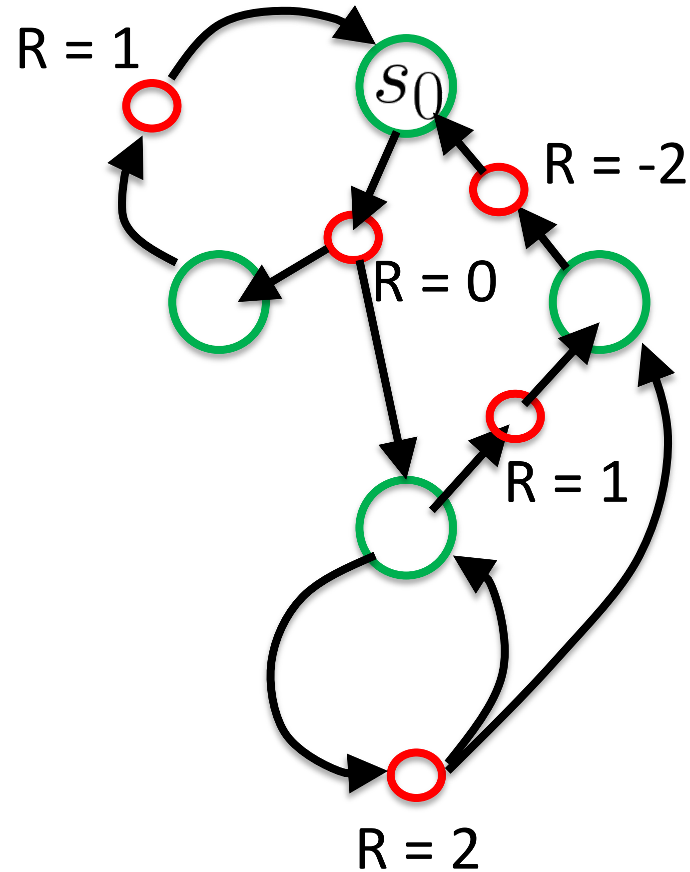
Goal: Find a strategy that maximizes expected reward

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 ~~strategy~~ that maximizes expected reward

policy $\pi : \mathcal{S} \to \mathcal{A}$ i.e. choice of action doesn't depend on past

OUTLINE

Part I: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

Part III: Planning and Belief Contraction

- Polynomial vs. Exponential Rates

Part IV: Learning

OUTLINE

Part I: Tabular Markov Decision Processes

- **Planning vs. Learning**

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

Part III: Planning and Belief Contraction

- Polynomial vs. Exponential Rates

Part IV: Learning

MAIN PROBLEMS

Planning (computational)	
Given full description of the MDP, compute an optimal policy	

MAIN PROBLEMS

Planning
(computational)

Given full description of the MDP, compute an optimal policy

Learning
(statistical)

Given budget of interactions with the environment, learn an optimal policy

MAIN PROBLEMS

Planning (computational)	**Learning (statistical)**
Given full description of the MDP, compute an optimal policy	Given budget of interactions with the environment, learn an optimal policy
	Uses planning as a subroutine

VALUE FUNCTIONS

Key: Keep track of how well you will do in the future

VALUE FUNCTIONS

Definition: The **value function** of a policy is

$$V_h^\pi(s) = \mathbb{E}[R_h(s_h, a_h) + \cdots + R_H(s_H, a_H) | s_h = s]$$

i.e. it gives the expected future reward starting from state s at timestep h

VALUE FUNCTIONS

Definition: The **value function** of a policy is

$$V_h^{\pi}(s) = \mathbb{E}[R_h(s_h, a_h) + \cdots + R_H(s_H, a_H) | s_h = s]$$

i.e. it gives the expected future reward starting from state s at timestep h

Bellman Optimality: An optimal policy π* must satisfy

$$V_h^{\pi^*}(s) = \max_a R_h(s, a) + \mathbb{E}_{s'}[V_{h+1}^{\pi^*}(s')]$$

for every state s, i.e. value function must be consistent

VALUE ITERATION

This gives an efficient algorithm for planning:

Initialize V = 0 (assuming no rewards at step H)

Repeat until convergence

Scan through states, update any violated V constraint

VALUE ITERATION

This gives an efficient algorithm for planning:

Initialize V = 0 (assuming no rewards at step H)

Repeat until convergence

Scan through states, update any violated V constraint

Of course, this is just **dynamic programming**

VALUE ITERATION

This gives an efficient algorithm for planning:

Initialize V = 0 (assuming no rewards at step H)

Repeat until convergence

Scan through states, update any violated V constraint

Of course, this is just **dynamic programming**

Moreover can find the optimal policy from the V^{π^*} values

MAIN PROBLEMS

Planning (computational)	**Learning (statistical)**
Given full description of the MDP, compute an optimal policy	Given budget of interactions with the environment, learn an optimal policy

MAIN PROBLEMS

Planning (computational)	**Learning (statistical)**
Given full description of the MDP, compute an optimal policy	Given budget of interactions with the environment, learn an optimal policy
e.g. value iteration, policy iteration, linear programming	

EXPLORATION VS EXPLOITATION

First non-asymptotic result:

Theorem [Kearns, Singh '02]: There is an algorithm that has polynomial running time and sample complexity that outputs an ϵ-suboptimal policy in tabular MDPs

EXPLORATION VS EXPLOITATION

First non-asymptotic result:

Theorem [Kearns, Singh '02]: There is an algorithm that has polynomial running time and sample complexity that outputs an ϵ-suboptimal policy in tabular MDPs

(1) Build a partial model on known states

(2) Trade off playing the optimal policy in current model vs. discovering new states

EXPLORATION VS EXPLOITATION

First non-asymptotic result:

Theorem [Kearns, Singh '02]: There is an algorithm that has polynomial running time and sample complexity that outputs an ϵ-suboptimal policy in tabular MDPs

(1) Build a partial model on known states

(2) Trade off playing the optimal policy in current model vs. discovering new states

Tight regret bounds given by **[Azar, Osband, Munos '17]**

POLICY GRADIENTS

Suppose we parameterize the class of policies by θ --- i.e. we want to maximize

$$J(\theta) = \mathbb{E}_{\pi_\theta}[R(\tau)]$$

random trajectory under π_θ

POLICY GRADIENTS

Suppose we parameterize the class of policies by θ --- i.e. we want to maximize

$$J(\theta) = \mathbb{E}_{\pi_\theta}[R(\tau)]$$

random trajectory under π_θ

How can we compute the gradient without full knowledge of the environment?

POLICY GRADIENTS

Suppose we parameterize the class of policies by θ --- i.e. we want to maximize

$$J(\theta) = \mathbb{E}_{\pi_\theta}[R(\tau)]$$

random trajectory under π_θ

How can we compute the gradient without full knowledge of the environment?

Policy Gradient Theorem: In fact

$$\nabla J(\theta) = \mathbb{E}_{\pi_\theta}[R(\tau) \nabla \log \pi_\theta(\tau)]$$

probability of trajectory under π_θ

POLICY GRADIENTS

Thus we can approximate the gradient through samples

POLICY GRADIENTS

Thus we can approximate the gradient through samples

Theorem: With softmax parameterization, there are **no spurrious critical points**

POLICY GRADIENTS

Thus we can approximate the gradient through samples

Theorem: With softmax parameterization, there are **no spurrious critical points**

Many challenges both in theory and practice, e.g. delayed feedback can cause gradients to be extremely small

e.g. see **[Agarwal, Kakade, Lee, Mahajan ‘19]**

MAIN PROBLEMS

Planning (computational)	**Learning (statistical)**
Given full description of the MDP, compute an optimal policy	Given budget of interactions with the environment, learn an optimal policy
e.g. value iteration, policy iteration, linear programming	

MAIN PROBLEMS

Planning (computational)	**Learning (statistical)**
Given full description of the MDP, compute an optimal policy	Given budget of interactions with the environment, learn an optimal policy
e.g. value iteration, policy iteration, linear programming	**e.g. model based, Q-learning, actor-critic policy gradient**

OUTLINE

Part I: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

Part III: Planning and Belief Contraction

- Polynomial vs. Exponential Rates

Part IV: Learning

OUTLINE

Part I: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

Part III: Planning and Belief Contraction

- Polynomial vs. Exponential Rates

Part IV: Learning

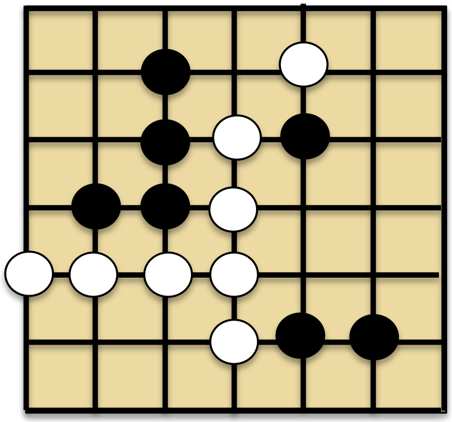
BEYOND TABULAR?

The trouble is most applications are not tabular, e.g.

BEYOND TABULAR?

The trouble is most applications are not tabular, e.g.

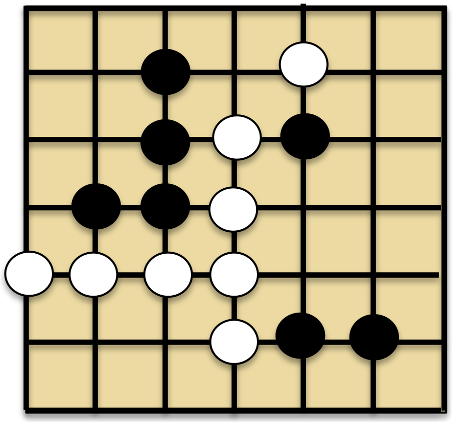
(1) Too many states to write down or visit



BEYOND TABULAR?

The trouble is most applications are not tabular, e.g.

(1) Too many states to write down or visit

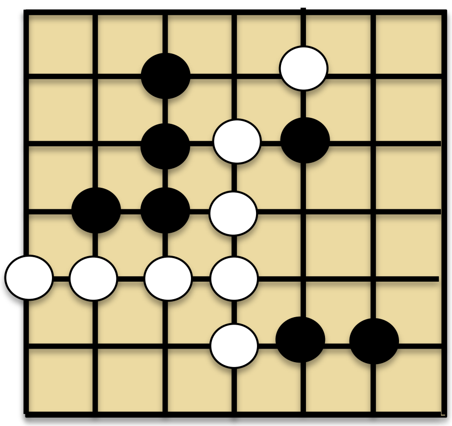


function approximation, block MDPs, etc

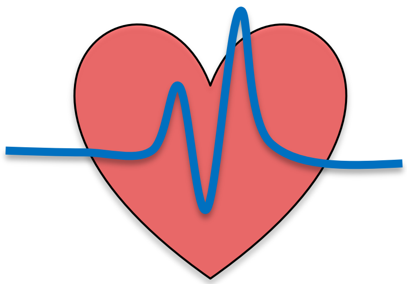
BEYOND TABULAR?

The trouble is most applications are not tabular, e.g.

(1) Too many states to write down or visit



function approximation, block MDPs, etc

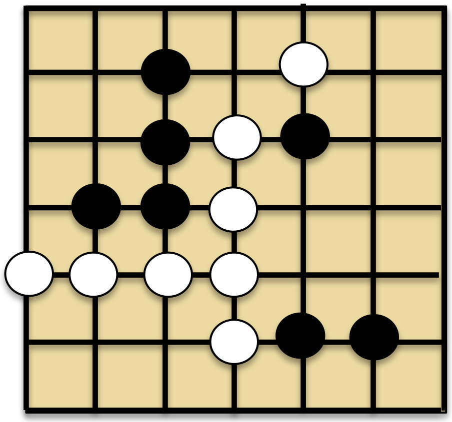


(2) Cannot observe full state

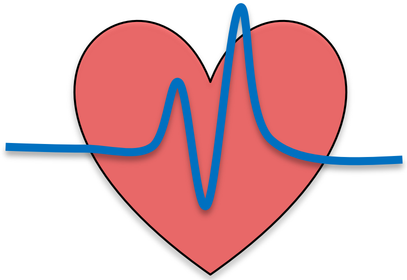
BEYOND TABULAR?

The trouble is most applications are not tabular, e.g.

(1) Too many states to write down or visit



function approximation, block MDPs, etc



(2) Cannot observe full state

Partially observable MDPs (POMDPs)

BEYOND TABULAR?

What do we want from our theoretical models?

BEYOND TABULAR?

What do we want from our theoretical models?

(1) Allow for very large, or even infinitely many states

BEYOND TABULAR?

What do we want from our theoretical models?

(1) Allow for very large, or even infinitely many states

(2) Be able to learn a near optimal policy from a small number of interactions

BEYOND TABULAR?

What do we want from our theoretical models?

(1) **Allow for very large, or even infinitely many states**

(2) **Be able to learn a near optimal policy from a small number of interactions**

(3) **Have computationally efficient algorithms**

BEYOND TABULAR?

What do we want from our theoretical models?

(1) Allow for very large, or even infinitely many states

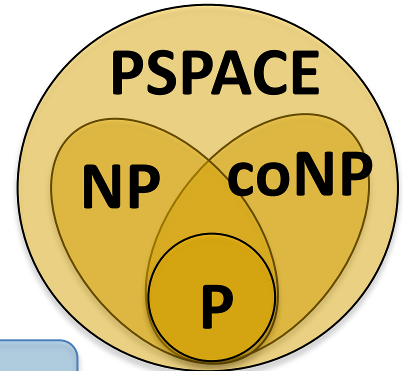
(2) Be able to learn a near optimal policy from a small number of interactions

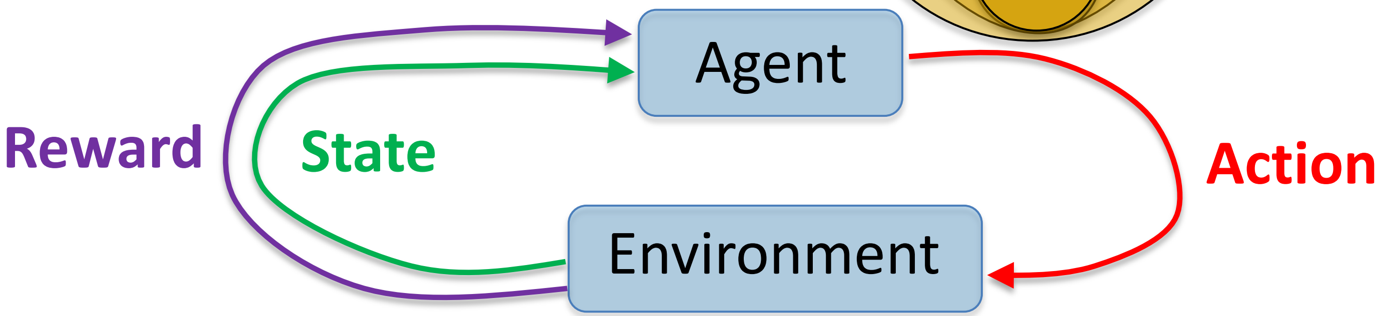
(3) Have computationally efficient algorithms

Existing theory is built around **(1)** and **(2)** but what do we miss out on by ignoring **(3)**?

WHAT ABOUT COMPUTATIONAL COMPLEXITY?

Returning to our earlier picture





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

OUTLINE

Part I: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

Part III: Planning and Belief Contraction

- Polynomial vs. Exponential Rates

Part IV: Learning

OUTLINE

Part I: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

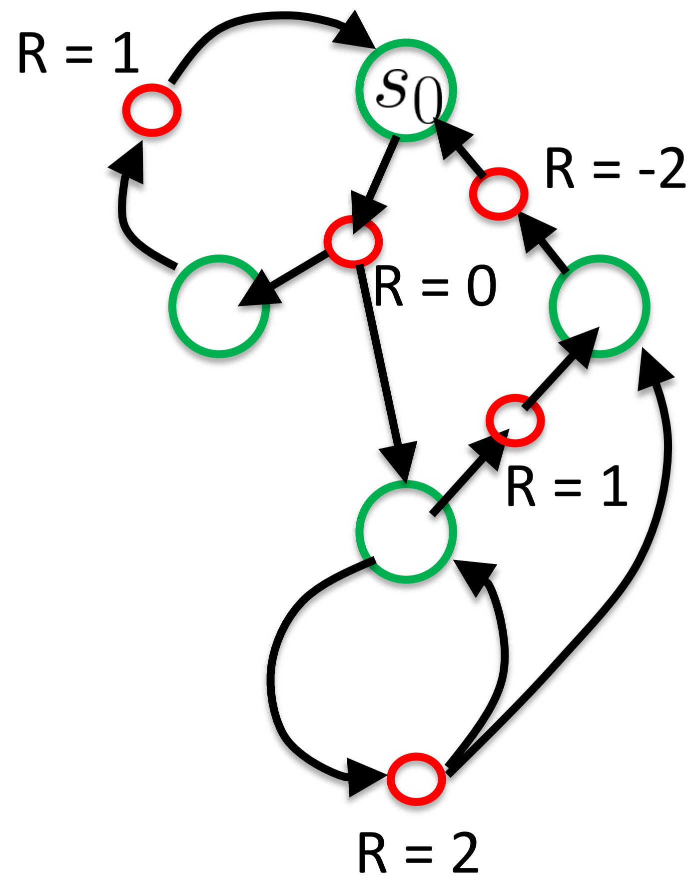
Part III: Planning and Belief Contraction

- Polynomial vs. Exponential Rates

Part IV: Learning

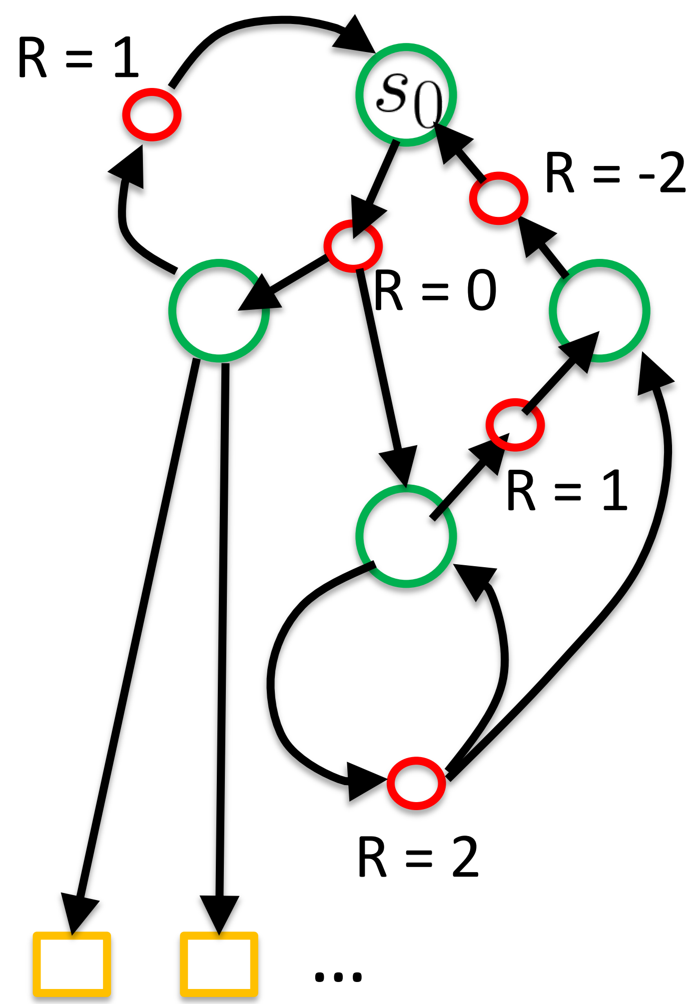
MARKOV DECISION PROCESSES

- 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 $\mathbb{O}_h(o|s)$



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**
$\pi : \mathcal{S} \to \mathcal{A}$	$\pi : \mathcal{A} \times \mathcal{O} \cdots \times \mathcal{O} \to \mathcal{A}$

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**
$\pi : \mathcal{S} \to \mathcal{A}$	$\pi : \mathcal{A} \times \mathcal{O} \cdots \times \mathcal{O} \to \mathcal{A}$
	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**
$\pi : \mathcal{S} \to \mathcal{A}$	$\pi : \mathcal{A} \times \mathcal{O} \cdots \times \mathcal{O} \to \mathcal{A}$
	Alternatively, it depends on the current belief
	$\pi : \Delta^{\mathcal{S}} \to \mathcal{A}$

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 '23]: 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 '23]: 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?

OUTLINE

Part I: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

Part III: Planning and Belief Contraction

- Polynomial vs. Exponential Rates

Part IV: Learning

OUTLINE

Part I: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- **Beyond Worst-Case Analysis**

Part III: Planning and Belief Contraction

- Polynomial vs. Exponential Rates

Part IV: Learning

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)**

PLANNING VIA STABILITY

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

Theorem [Golowich, Moitra, Rohatgi '23]: Given 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

PLANNING VIA STABILITY

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

Theorem [Golowich, Moitra, Rohatgi '23]: Given 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

PLANNING VIA STABILITY

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

Theorem [Golowich, Moitra, Rohatgi '23]: Given 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

Parallels well-known stability results for Kalman filtering

LOWER BOUNDS

Moreover these results are tight

Theorem [Golowich, Moitra, Rohatgi ‘23]: 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

LOWER BOUNDS

Moreover these results are tight

Theorem [Golowich, Moitra, Rohatgi '23]: 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

OUTLINE

Part I: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

Part III: Planning and Belief Contraction

- Polynomial vs. Exponential Rates

Part IV: Learning

OUTLINE

Part I: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

Part III: Planning and Belief Contraction

- Polynomial vs. Exponential Rates

Part IV: Learning

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

Thus, we could ignore all but the most recent history

BELIEF UPDATES

Definition: The **Bayes operator**, given an observation

$$B_h : \Delta^{\mathcal{S}} \times \mathcal{O} \to \Delta^{\mathcal{S}}$$

initial belief **observation** **revised belief**

updates the posterior on states

BELIEF UPDATES

Definition: The **Bayes operator**, given an observation

$$B_h : \Delta^{\mathcal{S}} \times \mathcal{O} \to \Delta^{\mathcal{S}}$$

initial belief **observation** **revised belief**

updates the posterior on states, and is defined as

$$B_h(b, y)(x) = \frac{\mathbb{O}_h(y|x) b(x)}{\sum_{z \in \mathcal{S}} \mathbb{O}_h(y|z) b(z)}$$

BELIEF UPDATES

Definition: And the **update operator,** given both an action and observation

$$U_h : \Delta^{\mathcal{S}} \times \mathcal{A} \times \mathcal{O} \rightarrow \Delta^{\mathcal{S}}$$

initial belief

observation

revised belief

based on the chosen action, takes a step

updates the posterior

BELIEF UPDATES

Definition: And the **update operator,** given both an action and observation

$$U_h : \Delta^{\mathcal{S}} \times \mathcal{A} \times \mathcal{O} \to \Delta^{\mathcal{S}}$$

initial belief **observation** **revised belief**

based on the chosen action, takes a step

updates the posterior, and is defined as

$$U_h(b, a, y) = B_h(\mathbb{T}_h(a)b, y)$$

OUTLINE

Part I: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

Part III: Planning and Belief Contraction

- Polynomial vs. Exponential Rates

Part IV: Learning

OUTLINE

Part I: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

Part III: Planning and Belief Contraction

- **Polynomial vs. Exponential Rates**

Part IV: Learning

TOWARDS A WEAKER BOUND

From the **data processing inequality**, we have that for any action

$$KL(\mathbb{T}_h(a)b||\mathbb{T}_h(a)b') \leq KL(b||b')$$

TOWARDS A WEAKER BOUND

From the **data processing inequality**, we have that for any action

$$KL(\mathbb{T}_h(a)b||\mathbb{T}_h(a)b') \leq KL(b||b')$$

But for some observations, the Bayes operator can **increase** the KL-divergence

TOWARDS A WEAKER BOUND

From the **data processing inequality**, we have that for any action

$$KL(\mathbb{T}_h(a)b||\mathbb{T}_h(a)b') \leq KL(b||b')$$

But for some observations, the Bayes operator can **increase** the KL-divergence

Do we make progress in expectation?

Lemma: Given beliefs b, b'

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] = KL(b||b') - KL(\mathbb{O}_h b||\mathbb{O}_h b')$$

Lemma: Given beliefs b, b'

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] = KL(b||b') - KL(\mathbb{O}_h b||\mathbb{O}_h b')$$

Notation: Let $P_X = b$, $Q_X = b'$ and $P_{Y|X=x} = \mathbb{O}_h(\cdot|x)$.

Lemma: Given beliefs b, b'

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] = KL(b||b') - KL(\mathbb{O}_h b||\mathbb{O}_h b')$$

Notation: Let $P_X = b$, $Q_X = b'$ and $P_{Y|X=x} = \mathbb{O}_h(\cdot|x)$.
$P_{X,Y} = P_{Y|X}P_X$ and $Q_{X,Y} = P_{Y|X}Q_X$

Lemma: Given beliefs b, b'

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] = KL(b||b') - KL(\mathbb{O}_h b||\mathbb{O}_h b')$$

Notation: Let $P_X = b$, $Q_X = b'$ and $P_{Y|X=x} = \mathbb{O}_h(\cdot|x)$.
$P_{X,Y} = P_{Y|X} P_X$ and $Q_{X,Y} = P_{Y|X} Q_X$

Proof: $KL(b||b') = KL(P_X||Q_X)$

Lemma: Given beliefs b, b'

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] = KL(b||b') - KL(\mathbb{O}_h b||\mathbb{O}_h b')$$

Notation: Let $P_X = b,\ Q_X = b'$ and $P_{Y|X=x} = \mathbb{O}_h(\cdot|x)$.
$P_{X,Y} = P_{Y|X}P_X$ and $Q_{X,Y} = P_{Y|X}Q_X$

Proof: $$KL(b||b') = KL(P_X||Q_X)$$

$$= KL(P_{X,Y}||Q_{X,Y}) + \mathop{\mathbb{E}}_{x \sim P_X}[KL(P_{Y|X=x}||Q_{Y|X=x})]$$

..using the chain rule

Lemma: Given beliefs b, b'

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] = KL(b||b') - KL(\mathbb{O}_h b||\mathbb{O}_h b')$$

Notation: Let $P_X = b$, $Q_X = b'$ and $P_{Y|X=x} = \mathbb{O}_h(\cdot|x)$.

$P_{X,Y} = P_{Y|X} P_X$ and $Q_{X,Y} = P_{Y|X} Q_X$

Proof: $KL(b||b') = KL(P_X||Q_X)$

zero

$$= KL(P_{X,Y}||Q_{X,Y}) - \mathop{\mathbb{E}}_{x \sim P_X}[KL(P_{Y|X=x}||Q_{Y|X=x})]$$

..using the chain rule, and the fact that $Q_{Y|X=x} = P_{Y|X=x}$

Lemma: Given beliefs b, b'

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] = KL(b||b') - KL(\mathbb{O}_h b||\mathbb{O}_h b')$$

Notation: Let $P_X = b,\ Q_X = b'$ and $P_{Y|X=x} = \mathbb{O}_h(\cdot|x)$.
$P_{X,Y} = P_{Y|X}P_X$ and $Q_{X,Y} = P_{Y|X}Q_X$

Proof: $KL(b||b') = KL(P_X||Q_X)$

$$= KL(P_{X,Y}||Q_{X,Y})$$

Lemma: Given beliefs b, b'

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] = KL(b||b') - KL(\mathbb{O}_h b||\mathbb{O}_h b')$$

Notation: Let $P_X = b$, $Q_X = b'$ and $P_{Y|X=x} = \mathbb{O}_h(\cdot|x)$.
$P_{X,Y} = P_{Y|X} P_X$ and $Q_{X,Y} = P_{Y|X} Q_X$

Proof:

$$KL(b||b') = KL(P_X||Q_X)$$

$$= KL(P_{X,Y}||Q_{X,Y})$$

$$= KL(P_Y||Q_Y) + \mathop{\mathbb{E}}_{y \sim P_Y}[KL(P_{X|Y=y}||Q_{X|Y=y})]$$

... this time using the chain rule in opposite order

Lemma: Given beliefs b, b'

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] = KL(b||b') - KL(\mathbb{O}_h b||\mathbb{O}_h b')$$

Notation: Let $P_X = b$, $Q_X = b'$ and $P_{Y|X=x} = \mathbb{O}_h(\cdot|x)$.
$P_{X,Y} = P_{Y|X}P_X$ and $Q_{X,Y} = P_{Y|X}Q_X$

Proof:
$$\begin{aligned} KL(b||b') &= KL(P_X||Q_X) \\ &= KL(P_{X,Y}||Q_{X,Y}) \\ &= KL(P_Y||Q_Y) + \mathop{\mathbb{E}}_{y \sim P_Y}\underbrace{[KL(P_{X|Y=y}||Q_{X|Y=y})]}_{KL(B_h(b,y)||B_h(b',y))} \end{aligned}$$

Lemma: Given beliefs b, b'

$$\mathop{\mathbb{E}}_{y\sim\mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] = KL(b||b') - KL(\mathbb{O}_h b||\mathbb{O}_h b')$$

Notation: Let $P_X = b,\ Q_X = b'$ and $P_{Y|X=x} = \mathbb{O}_h(\cdot|x)$.

$P_{X,Y} = P_{Y|X}P_X$ and $Q_{X,Y} = P_{Y|X}Q_X$

Proof:

$$KL(b||b') = KL(P_X||Q_X)$$

$$= KL(P_{X,Y}||Q_{X,Y})$$

$$= KL(P_Y||Q_Y) + \mathop{\mathbb{E}}_{y\sim P_Y}[\underbrace{KL(P_{X|Y=y}||Q_{X|Y=y})}_{KL(B_h(b,y)||B_h(b',y))}] \qquad \blacksquare$$

Lemma: Given beliefs b, b'

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] = KL(b||b') - KL(\mathbb{O}_h b||\mathbb{O}_h b')$$

Lemma: Given beliefs b, b'

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] = KL(b||b') - KL(\mathbb{O}_h b||\mathbb{O}_h b')$$

Does this imply fast enough convergence?

Using Pinsker's inequality **(1)** and observability **(2)**, we have

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] \overset{(1)}{\leq} KL(b||b') - \frac{1}{2}\|\mathbb{O}_h b - \mathbb{O}_h b'\|_1^2$$

Using Pinsker's inequality **(1)** and observability **(2)**, we have

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] \overset{(1)}{\leq} KL(b||b') - \frac{1}{2}\|\mathbb{O}_h b - \mathbb{O}_h b'\|_1^2$$

$$\overset{(2)}{\leq} KL(b||b') - \frac{\gamma^2}{2}\|b - b'\|_1^2$$

Using Pinsker's inequality **(1)** and observability **(2)**, we have

$$\underset{y \sim \mathbb{O}_h b}{\mathbb{E}}[KL(B_h(b,y)||B_h(b',y))] \overset{(1)}{\leq} KL(b||b') - \frac{1}{2}\|\mathbb{O}_h b - \mathbb{O}_h b'\|_1^2$$

$$\overset{(2)}{\leq} KL(b||b') - \frac{\gamma^2}{2}\|b - b'\|_1^2$$

Using reverse Pinsker's inequality, we get

$$\underset{y \sim \mathbb{O}_h b}{\mathbb{E}}[KL(B_h(b,y)||B_h(b',y))] \leq KL(b||b') - c\gamma^2 KL(b||b')^2$$

provided that $||b/b'||_\infty$ is bounded

Theorem [Even-Dar et al.]: Fix any γ-observable POMDP and policy π . Then we have

$$\mathbb{E}_{\tau}[KL(b_t||b'_t)] \leq \epsilon$$

provided that $t \geq 1/(\gamma^2\epsilon)$

Theorem [Even-Dar et al.]: Fix any γ-observable POMDP and policy π. Then we have

$$\mathbb{E}_{\mathcal{T}}[KL(b_t||b'_t)] \leq \epsilon$$

provided that $t \geq 1/(\gamma^2 \epsilon)$

Inverse polynomial, rather than exponential convergence :(

Theorem [Even-Dar et al.]: Fix any γ-observable POMDP and policy π. Then we have

$$\mathbb{E}_{\mathcal{T}}[KL(b_t||b_t')] \leq \epsilon$$

provided that $t \geq 1/(\gamma^2 \epsilon)$

Inverse polynomial, rather than exponential convergence :(

Unfortunately there are cases where progress can be slow, but…

A WIN-WIN ARGUMENT

We show that either

(1) A **stronger reverse Pinsker** holds, i.e.

$$\underset{y \sim \mathbb{O}_h b}{\mathbb{E}}[KL(B_h(b,y)||B_h(b',y))] \leq KL(b||b') - \frac{\gamma^2}{32}\min(KL(b||b'),1)$$

A WIN-WIN ARGUMENT

We show that either

(1) A **stronger reverse Pinsker** holds, i.e.

$$\mathbb{E}_{y \sim \mathbb{O}_h b}[KL(B_h(b,y)||B_h(b',y))] \leq KL(b||b') - \frac{\gamma^2}{32}\min(KL(b||b'),1)$$

or instead

(2) Progress is **anti-concentrated,** i.e. for some event $\mathcal{E} \subset \mathcal{O}$

$$\mathbb{E}_{y \sim \mathbb{O}_h b}[(KL(B_h(b,y)||B_h(b',y)) - KL(b||b'))\mathbf{1}_{y\in\mathcal{E}}] \leq -\frac{\gamma}{8}KL(b||b')$$

A WIN-WIN ARGUMENT

As a result, we get:

Corollary: For any γ-observable POMDP

$$\underset{y \sim \mathbb{O}_h b}{\mathbb{E}}[\sqrt{KL(B_h(b,y)||B_h(b',y))}]$$

$$\leq \left(1 - \Omega\left(\frac{\gamma^2}{\max(1, KL(b||b'))}\right)\right)\sqrt{KL(b||b')}$$

A WIN-WIN ARGUMENT

As a result, we get:

Corollary: For any γ-observable POMDP

$$\underset{y \sim \mathbb{O}_h b}{\mathbb{E}}[\sqrt{KL(B_h(b,y)||B_h(b',y))}]$$
$$\leq \left(1 - \Omega\left(\frac{\gamma^2}{\max(1, KL(b||b'))}\right)\right)\sqrt{KL(b||b')}$$

Variations on this argument lead to different rates of contraction

A WIN-WIN ARGUMENT

As a result, we get:

Corollary: For any γ-observable POMDP

$$\mathop{\mathbb{E}}_{y \sim \mathbb{O}_h b}[\sqrt{KL(B_h(b,y)||B_h(b',y))}] \leq \left(1 - \Omega\left(\frac{\gamma^2}{\max(1, KL(b||b'))}\right)\right)\sqrt{KL(b||b')}$$

Variations on this argument lead to different rates of contraction

Open: Prove sharp rates that match Chernoff bounds

BELLMAN UPDATES

How does this lead to better algorithms for planning?

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

current action/obs. sequence

new action/obs. sequence

latent state sampled from current belief, stochastic transition based on chosen action

BELLMAN UPDATES

Belief contraction allows us to **truncate**

TRUNCATED BELLMAN UPDATES

Belief contraction allows us to **truncate**

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

length t window

length t window

latent state sampled from truncated belief, with uniform prior

TRUNCATED BELLMAN UPDATES

Belief contraction allows us to **truncate**

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

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: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

Part III: Planning and Belief Contraction

- Polynomial vs. Exponential Rates

Part IV: Learning

OUTLINE

Part I: Tabular Markov Decision Processes

- Planning vs. Learning

Interlude: Can We Make RL Algorithmically Tractable?

Part II: Partial Observations and the Curse of History

- Beyond Worst-Case Analysis

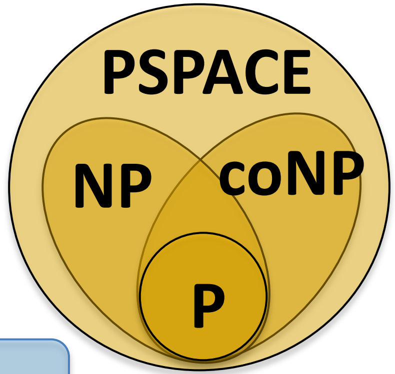
Part III: Planning and Belief Contraction

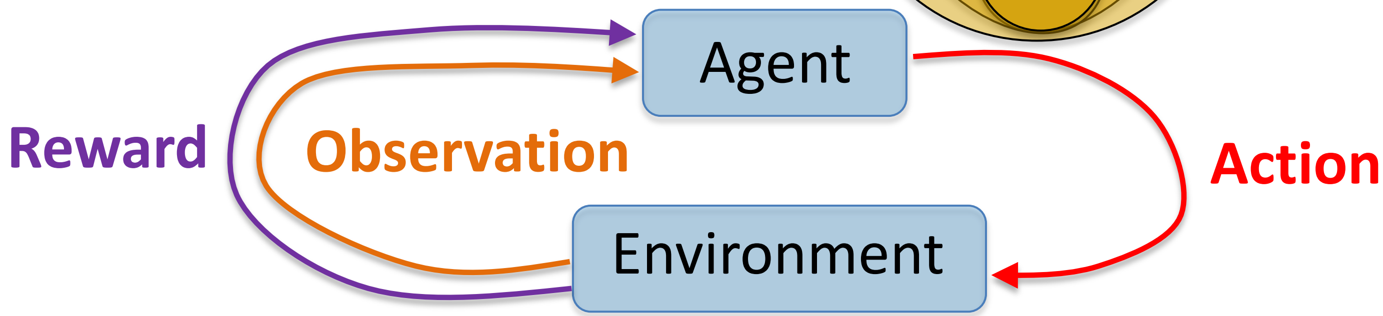
- Polynomial vs. Exponential Rates

Part IV: Learning

WHAT ABOUT LEARNING?

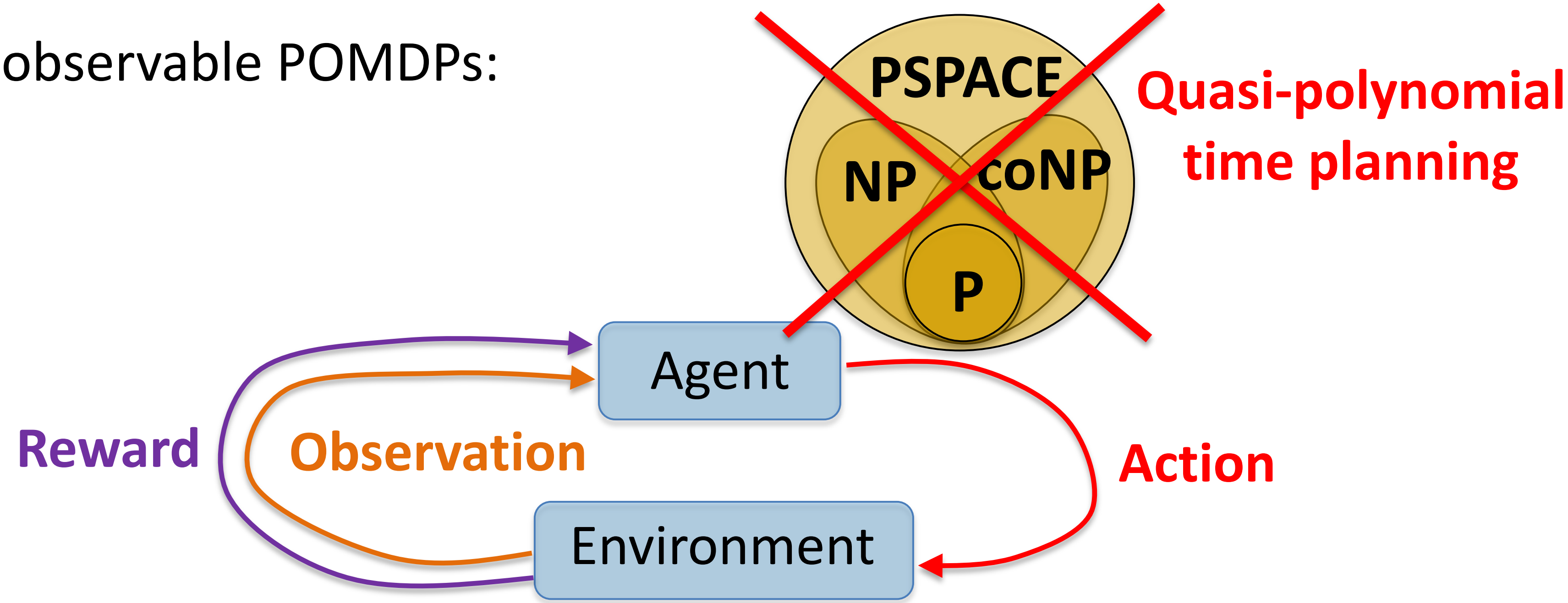
In observable POMDPs:





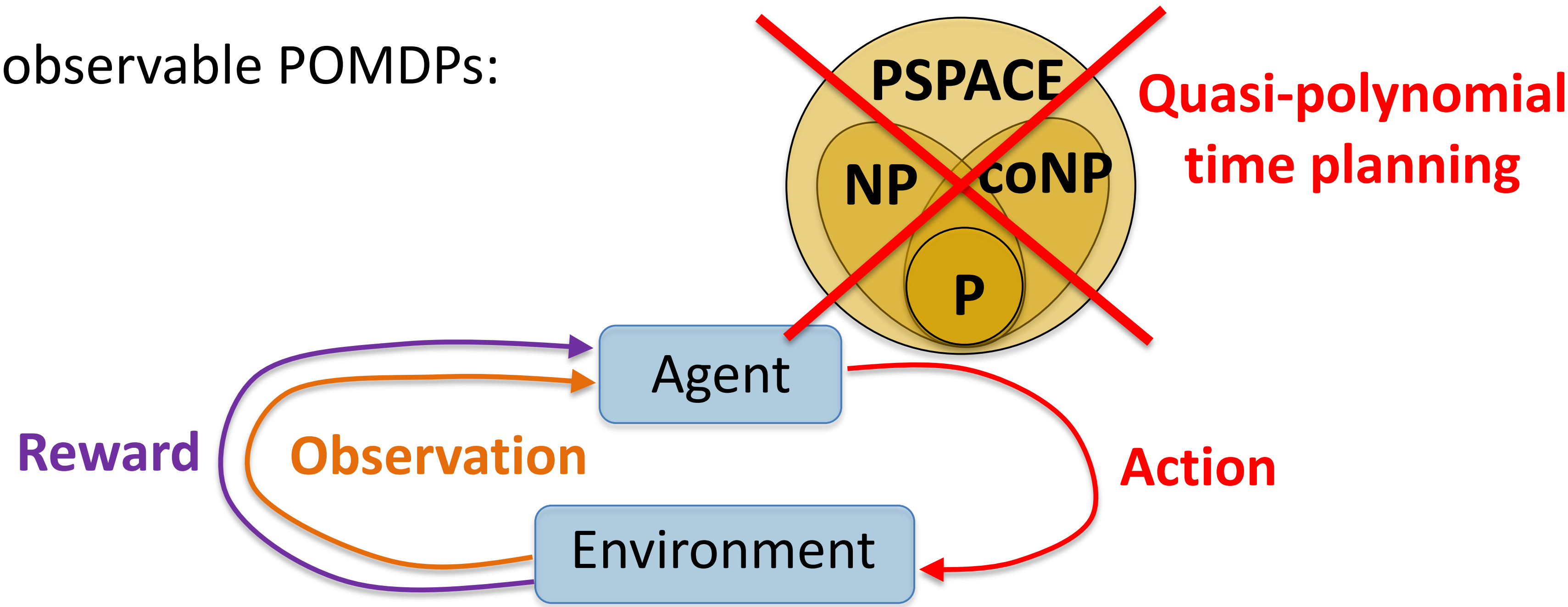
WHAT ABOUT LEARNING?

In observable POMDPs:



WHAT ABOUT LEARNING?

In observable POMDPs:



Can belief contraction be used for learning too?

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 '20]: 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 '20]: 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

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 '20]: 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!

Alternatively:

[Lin, Chung, Szepesvari, Jin ‘23] gave a framework based on **optimistic maximum likelihood estimation**

Alternatively:

[Lin, Chung, Szepesvari, Jin '23] gave a framework based on **optimistic maximum likelihood estimation**

i.e. given sample trajectories, find a POMDP that gets maximum value conditioned on approximately maximizing the likelihood

Alternatively:

[Lin, Chung, Szepesvari, Jin ‘23] gave a framework based on **optimistic maximum likelihood estimation**

i.e. given sample trajectories, find a POMDP that gets maximum value conditioned on approximately maximizing the likelihood

Can we circumvent optimism?

COMPUTATIONALLY EFFICIENT LEARNING

We show:

Theorem [Golowich, Moitra, Rohatgi '23]: 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

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?

Reachability: 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?

Reachability: 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:

$$\mathbf{M_h}$$

MDP that approximates P up to step h

ITERATIVE EXPLORATION

Our approach is:

$$\mathbf{M_h} \longrightarrow \mathbf{X_h}$$

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 → 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}

MDP that approximates P up to step h+1

X_h

Barycentric spanner for observation distributions at step h

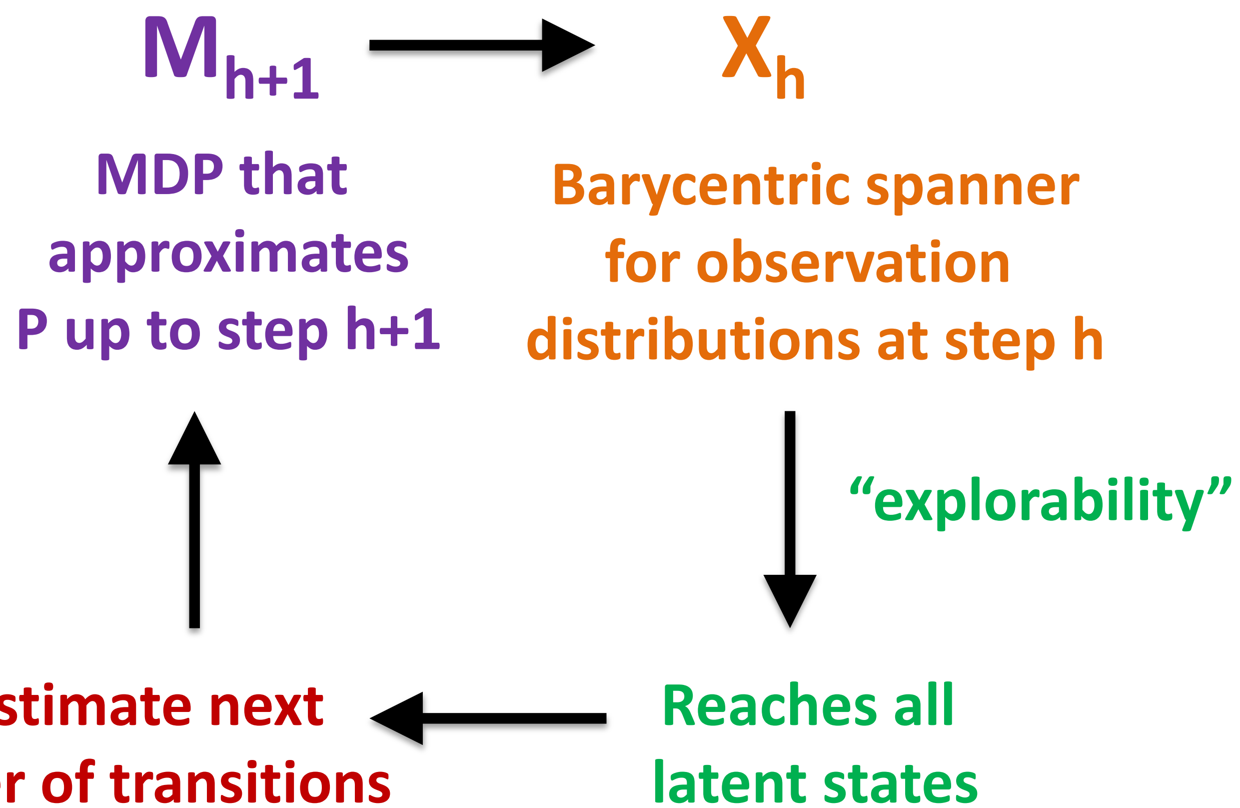
"explorability"

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

e.g. in **[Golowich, Moitra ‘22]**, 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

e.g. in **[Golowich, Moitra '22]**, 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?