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 $S$, 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** $R_h(s, a)$

![Diagram showing transitions and rewards](image-url)
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 $\pi : \mathcal{S} \rightarrow \mathcal{A}$ 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)

Given full description of the MDP, compute an optimal policy

Learning (statistical)

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

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

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

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

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

Bellman Optimality: An optimal policy \( \pi^* \) must satisfy

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

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 $V_{\pi^*}$ values
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)

Given full description of the MDP, compute an optimal policy

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

Learning (statistical)

Given budget of interactions with the environment, learn an optimal policy
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 $\epsilon$-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 $\epsilon$-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 \( \epsilon \)-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 $\theta$ --- i.e. we want to maximize

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

random trajectory under $\pi_\theta$
POLICY GRADIENTS

Suppose we parameterize the class of policies by $\theta$ --- i.e. we want to maximize

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

random trajectory under $\pi_\theta$

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

Suppose we parameterize the class of policies by $\theta$ --- i.e. we want to maximize

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

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

**Policy Gradient Theorem: In fact**

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

random trajectory under $\pi_\theta$

probability of trajectory under $\pi_\theta$
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)

Given full description of the MDP, compute an optimal policy

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

Learning (statistical)

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

Planning (computational)

Given full description of the MDP, compute an optimal policy

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

Learning (statistical)

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

- 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 $H$

Diagram:

- $s_0$ with $R = 1$
- $s'$ with $R = 0$
- $s''$ with $R = -2$
- $s'''$ with $R = 1$
- $s''''$ with $R = 2$
PARTIALLY OBSERVABLE MDPS (POMDPS)

- State Space $\mathcal{S}$, start at $s_0$
- Action Space $\mathcal{A}$
- Rewards $R_h(s, a)$
- Transition Probabilities $T_h(s' | s, a)$
- Horizon $H$
- Observation Space $\mathcal{O}$ and probabilities $\mathcal{O}_h(o | s)$

Diagram:
R = 1
R = 0
R = -2
R = 1
R = 2
...
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?

<table>
<thead>
<tr>
<th>MDPs</th>
<th>POMDPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal action only depends on current state</td>
<td></td>
</tr>
</tbody>
</table>

\[ \pi : S \rightarrow A \]
THE CURSE OF HISTORY

Can you succinctly represent an optimal policy?

<table>
<thead>
<tr>
<th>MDPs</th>
<th>POMDPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal action only depends on current state</td>
<td>Optimal action depends on action/observation history</td>
</tr>
<tr>
<td>( \pi : S \rightarrow A )</td>
<td>( \pi : A \times \mathcal{O} \cdots \times \mathcal{O} \rightarrow A )</td>
</tr>
</tbody>
</table>
THE CURSE OF HISTORY

Can you succinctly represent an optimal policy?

<table>
<thead>
<tr>
<th>MDPs</th>
<th>POMDPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal action only</td>
<td>Optimal action depends on</td>
</tr>
<tr>
<td>depends on current state</td>
<td>action/observation history</td>
</tr>
<tr>
<td>$\pi : S \rightarrow A$</td>
<td>$\pi : A \times O \times \cdots \times O \rightarrow A$</td>
</tr>
<tr>
<td></td>
<td>Alternatively, it depends</td>
</tr>
<tr>
<td></td>
<td>on the current belief</td>
</tr>
<tr>
<td></td>
<td>$\pi : \Delta^S \rightarrow A$</td>
</tr>
</tbody>
</table>
THE CURSE OF HISTORY

Can you succinctly represent an optimal policy?

<table>
<thead>
<tr>
<th>MDPs</th>
<th>POMDPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal action only depends on current state</td>
<td>Optimal action depends on action/observation history</td>
</tr>
<tr>
<td>( \pi : S \rightarrow A )</td>
<td>( \pi : A \times O \times \cdots \times O \rightarrow A )</td>
</tr>
</tbody>
</table>

Alternatively, it depends on the current belief

\[ \pi : \Delta^S \rightarrow A \]

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

“\textbf{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 $\gamma$-observable if for all $\mu$ and all distributions $b, b'$ on states we have

$$\| \Theta_h b - \Theta_h b' \|_1 \geq \gamma \| b - b' \|_1$$

i.e. well-separated distributions on states lead to well-separated distributions on observations.
BEYOND WORST-CASE ANALYSIS

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

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

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 $\gamma$-observable if for all $h$ and all distributions $b, b'$ on states we have

$$\|\mathcal{O}_hb - \mathcal{O}_hb'\|_1 \geq \gamma\|b - b'\|_1$$

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 $\gamma$-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 $\epsilon$-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 $\gamma$-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 $\epsilon$-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 $\gamma$-observable POMDP there is an algorithm running in time

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

that outputs an $\epsilon$-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

\[ O\left(\frac{\log(|S| |A| |H| |O|)}{|\epsilon|/\gamma}\right) \]

for finding an \( \epsilon \)-suboptimal policy in a \( \gamma \)-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

\[
\left(|S| |A| H |O|\right)^{o\left(\log\left(|S| |A| H |O| / \epsilon\right)/ \gamma\right)}
\]

for finding an $\epsilon$-suboptimal policy in a $\gamma$-observable POMDP

It’s hard even in the **lossy case**, where you observe the state with probability $\gamma$ 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 $\gamma$-observable POMDP and policy $\pi$. Then

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

**posterior, starting from arbitrary belief state**

**posterior, starting from uniform belief state**

where $\tau$ is the trajectory from the POMDP by playing $\pi$
BELIEF CONTRACTION

**Theorem:** Fix any $\gamma$-observable POMDP and policy $\pi$. 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 $\tau$ is the trajectory from the POMDP by playing $\pi$

Thus, we could ignore all but the most recent history
BELIEF UPDATES

Definition: The Bayes operator, given an observation

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

initial belief \hspace{2cm} revised belief

observation

updates the posterior on states
BELIEF UPDATES

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

\[ B_h : \Delta^S \times \mathcal{O} \rightarrow \Delta^S \]

updates the posterior on states, and is defined as

\[
B_h(b, y)(x) = \frac{\mathcal{O}_h(y|x)b(x)}{\sum_{z \in \mathcal{S}} \mathcal{O}_h(y|z)b(z)}
\]
BELIEF UPDATES

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

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

- **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

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

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

\[ KL(T_h(a)b || T_h(a)b') \leq KL(b || b') \]
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

\[
KL(T_h(a)b || T_h(a)b') \leq KL(b || b')
\]

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

**Do we make progress in expectation?**
Lemma: Given beliefs $b, b'$

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

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

Notation: Let $P_X = b$, $Q_X = b'$ and $P_Y|X=x = \mathcal{O}_h(\cdot|x)$. 
Lemma: Given beliefs $b, b'$

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

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 $b, b'$

$$\mathbb{E}[KL(B_h(b, y) || B_h(b', y))] = KL(b || b') - KL(\bigcirc_h b || \bigcirc_h b')$$

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

Proof: $KL(b || b') = KL(P_X || Q_X)$
Lemma: Given beliefs $b, b'$

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

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: $KL(b \mid \mid b') = KL(P_X \mid \mid Q_X)$

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

..using the chain rule
**Lemma:** Given beliefs $b, b'$

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

**Notation:** Let $P_X = b$, $Q_X = b'$ and $P_{Y \mid X=x} = \Theta_h (\cdot \mid x)$.

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

**Proof:**

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

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

..using the chain rule, and the fact that $Q_{Y \mid X=x} = P_{Y \mid X=x}$
Lemma: Given beliefs $b, b'$

\[ \mathbb{E}[KL(B_h(b, y) \mid B_h(b', y))] = KL(b \mid b') - KL(\mathbb{Q}_h b \mid \mathbb{Q}_h b') \]

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

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

Proof: \[ KL(b \mid b') = KL(P_X \mid Q_X) \]

\[ = KL(P_{X,Y} \mid Q_{X,Y}) \]
Lemma: Given beliefs \( b, b' \)

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

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: \( KL(b || b') = KL(P_X || Q_X) \)

\[
= KL(P_{X,Y} || Q_{X,Y})
\]

\[
= KL(P_Y || Q_Y) + \mathbb{E}_{y \sim P_Y}[KL(P_X|_{Y=y} || Q_X|_{Y=y})]
\]

... this time using the chain rule in opposite order
**Lemma:** Given beliefs $b, b'$

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

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

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

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

$$= KL(P_Y \| Q_Y) + \mathbb{E}_{y \sim P_Y} [KL(P_X | Y = y \| Q_X | Y = y)]$$

$$KL(B_h(b, y) \| B_h(b', y))$$
**Lemma:** Given beliefs $b, b'$

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

**Notation:** Let $P_X = b$, $Q_X = b'$ and $P_{Y \mid X = x} = \mathbb{Q}_h(\cdot \mid x)$. Then $P_{X,Y} = P_Y \mid X P_X$ and $Q_{X,Y} = P_Y \mid X Q_X$

**Proof:**

$$
KL(b \mid b') = KL(P_X \mid Q_X)
$$

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

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

$$
KL(B_h(b, y) \mid B_h(b', y))
$$
**Lemma:** Given beliefs $b, b'$

\[
\mathbb{E}_{y \sim \mathcal{D}_b} [KL(B_h(b, y) \| B_h(b', y))] = KL(b \| b') - KL(\mathcal{D}_b \| \mathcal{D}_{b'})
\]
**Lemma:** Given beliefs \( b, b' \)

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

Does this imply fast enough convergence?
Using Pinsker’s inequality (1) and observability (2), we have

\[
\mathbb{E}[KL(B_h(b, y) \| B_h(b', y))] \leq KL(b \| b') - \frac{1}{2} \| \mathbb{O}_h b - \mathbb{O}_h b' \|_1^2
\]
Using Pinsker’s inequality (1) and observability (2), we have

\[
\mathbb{E}[KL(B_h(b, y) \| B_h(b', y))] \leq KL(b \| b') - \frac{1}{2} \| \Theta_h b - \Theta_h b' \|_1^2
\]

\[
\leq KL(b \| b') - \gamma^2 \frac{1}{2} \| b - b' \|_1^2
\]
Using Pinsker’s inequality (1) and observability (2), we have

\[ \mathbb{E}[KL(B_h(b, y) \mid B_h(b', y))] \leq KL(b \mid b') - \frac{1}{2}\|\Omega_h b - \Omega_h b'\|^2_1 \]

\[ \leq KL(b \mid b') - \frac{\gamma^2}{2}\|b - b'\|^2_1 \]

Using reverse Pinsker’s inequality, we get

\[ \mathbb{E}[KL(B_h(b, y) \mid B_h(b', y))] \leq KL(b \mid b') - c\gamma^2 KL(b \mid b')^2 \]

provided that \( \|b/b'\|_\infty \) is bounded
Theorem [Even-Dar et al.]: Fix any $\gamma$-observable POMDP and policy $\pi$. Then we have

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

provided that $t \geq 1/(\gamma^2 \epsilon)$
Theorem [Even-Dar et al.]: Fix any $\gamma$-observable POMDP and policy $\pi$. Then we have
\[
\mathbb{E}_\pi[KL(b_t || b'_t)] \leq \epsilon
\]
provided that $t \geq 1/(\gamma^2 \epsilon)$

Inverse polynomial, rather than exponential convergence :(
Theorem [Even-Dar et al.]: Fix any $\gamma$-observable POMDP and policy $\pi$. Then we have
\[
\mathbb{E}_\pi[KL(b_t || b'_t)] \leq \epsilon
\]
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.

\[
\mathbb{E}_{y \sim \mathbb{O}_h^b}\left[KL(B_h(b, y) \mid\mid B_h(b', y))\right] \leq KL(b \mid\mid b') - \frac{\gamma^2}{32} \min(KL(b \mid\mid b'), 1)
\]
A WIN-WIN ARGUMENT

We show that either

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

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

or instead

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

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

As a result, we get:

**Corollary:** For any \( \gamma \)-observable POMDP

\[
\mathbb{E}
\left[
\sqrt{KL(B_h(b, y) \mid \mid B_h(b', y))}
\right]
\leq \left(1 - \Omega\left(\frac{\gamma^2}{\max(1, KL(b \mid \mid b'))}\right)\right) \sqrt{KL(b \mid \mid b')}
\]
A WIN-WIN ARGUMENT

As a result, we get:

**Corollary:** For any $\gamma$-observable POMDP

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

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

Variations on this argument lead to different rates of contraction.
A WIN-WIN ARGUMENT

As a result, we get:

**Corollary:** For any $\gamma$-observable POMDP

$$\mathbb{E} \left[ \sqrt{KL(B_h(b, y) || B_h(b', y))} \mid y \sim \mathcal{O}_h^b \right]$$

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

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?

\[ \text{Value}(x) = \max_a \mathbb{E}[\text{Reward}(a) + \text{Value}(x')] \]

- 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
Belief contraction allows us to truncate

\[
\text{Value}(x) = \max_{\text{actions } a} \mathbb{E}[\text{Reward}(a) + \text{Value}(x')] \\
\text{latent state sampled from truncated belief, with uniform prior}
\]
TRUNCATED BELLMAN UPDATES

Belief contraction allows us to truncate

\[
Value(x) = \max_a \mathbb{E}[\text{Reward}(a) + Value(x')] 
\]

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:

- Agent
- Environment
- Action
- Reward
- Observation

Complexity Classes:
- P
- NP
- coNP
- PSPACE
WHAT ABOUT LEARNING?

In observable POMDPs:

- Agent
- Environment
- Action
- Reward
- Observation

Complexity classes:
- $P$
- $NP$
- $coNP$
- $PSPACE$

Quasi-polynomial time planning
WHAT ABOUT LEARNING?

In observable POMDPs:

Can belief contraction be used for learning too?

Quasi-polynomial time planning

P
NP
coNP
PSPACE
SAMPLE EFFICIENT LEARNING

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

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

Theorem [Jin, Kakade, Krishnamurthy, Liu ‘20]: Given access to an optimistic planning oracle, there is an algorithm that uses
$$\text{poly}(|\mathcal{S}|, |\mathcal{A}|, H, |\mathcal{O}|, 1/\alpha)$$
samples and finds an $\varepsilon$-suboptimal policy under Assumption 1
SAMPLE EFFICIENT LEARNING

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

**Theorem [Jin, Kakade, Krishnamurthy, Liu ‘20]**: Given access to an optimistic planning oracle, there is an algorithm that uses

$$\text{poly}(|S|, |A|, H, |O|, 1/\alpha)$$
samples and finds an $\varepsilon$-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. $|S| \leq |O|$
And moreover $\sigma_{\min}(O_h) \geq \alpha$ for all $h$

Theorem [Jin, Kakade, Krishnamurthy, Liu ‘20]: Given access to an optimistic planning oracle, there is an algorithm that uses

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

samples and finds an $\epsilon$-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

\[ \checkmark \]

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

\[ \text{Can we circumvent optimism?} \]
We show:

**Theorem [Golowich, Moitra, Rohatgi ’23]:** There is an algorithm with running time and sample complexity

\[
(|O| |A|)^C \log(H |S| |O| / \epsilon \gamma) / \gamma^4
\]

that outputs an \(\epsilon\)-suboptimal policy in a \(\gamma\)-observable POMDP.
**Corollary:** Any $\gamma$-observable POMDP $P$ can be approximated by an MDP $M$ with a quasi-polynomial number of states
Corollary: Any $\gamma$-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
Corollary: Any $\gamma$-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.
**Corollary:** Any $\gamma$-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.
Corollary: Any $\gamma$-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.
Corollary: Any $\gamma$-observable POMDP $P$ can be approximated by an MDP $M$ with a quasi-polynomial number of states.

Can we learn $M$ efficiently?
**Corollary:** Any $\gamma$-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 $\pi$ that visits $x$ at $h$ with nonnegligible probability.
APPROXIMATION BY MDPS

**Corollary:** Any $\gamma$-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 $\pi$ 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 $\lambda$-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 \( \lambda \)-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 \( \lambda \)-approximate barycentric spanner with

\[
O(d^2 \log \lambda \ d)
\]
calls to the optimization oracle (assuming \( \mathcal{X} \) is compact)
POLICY COVERS

Now let

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

Now let

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

**Claim:** By observability, if we can construct policies

\[ \pi_1, \pi_2, \ldots, \pi_{|O|} \]

whose induced distributions on observations at step \( h \) are an approximate barycentric spanner
POLICY COVERS

Now let

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

**Claim:** By observability, if we can construct policies

\[ \pi_1, \pi_2, \ldots, \pi_{|O|} \]

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:

\( M_h \)  \( \rightarrow \)  \( X_h \)

MDP that approximates \( P \) up to step \( h \)

Barycentric spanner for observation distributions at step \( h \)
Our approach is:

\[ M_h \quad \rightarrow \quad X_h \]

- \( M_h \): MDP that approximates \( P \) up to step \( h \)
- \( X_h \): Barycentric spanner for observation distributions at step \( h \)

Reaches all latent states

“explorability”
ITERATIVE EXPLORATION

Our approach is:

\( M_h \)  \( \rightarrow \)  \( X_h \)

- **MDP that approximates** \( P \) **up to step** \( h \)
- **Barycentric spanner** for observation distributions at step \( h \)

- “explorability”
- Estimate next layer of transitions
- Reaches all latent states
ITERATIVE EXPLORATION

Our approach is:

\[ M_{h+1} \rightarrow X_h \]

MDP that approximates \( P \) up to step \( h+1 \)

Barycentric spanner for observation distributions at step \( h \)

Estimate next layer of transitions

Reaches all latent states

“explorability”
ITERATIVE EXPLORATION

Our approach is:

\[ M_{h+1} \rightarrow X_h \]

- MDP that approximates \( P \) up to step \( h+1 \)
- Barycentric spanner for observation distributions at step \( h \)

Estimate next layer of transitions
Reaches all latent states

“explorability”

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?