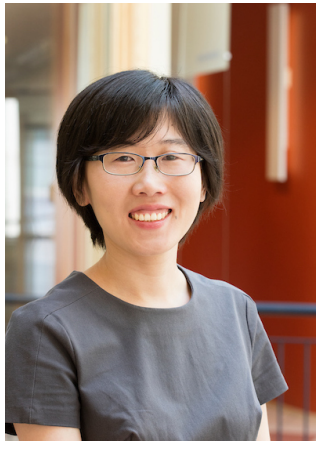
Statistical and algorithmic foundations of reinforcement learning



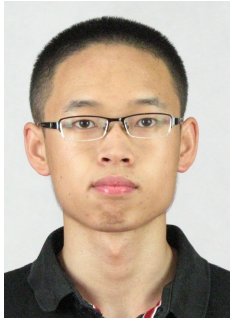
Yuejie Chi
Yale

Yuxin Chen
UPenn

Yuting Wei
UPenn

Tutorial, INFORMS 2025

Our wonderful collaborators



Gen Li

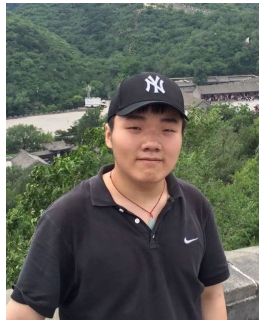
CUHK

Zihan Zhang

HKUST

Laixi Shi

JHU

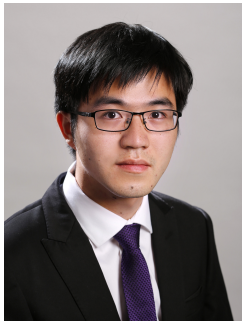


Yuling Yan

UWM

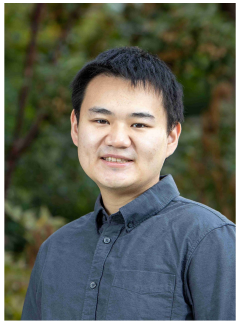
Shicong Cen

XTY



Changxiao Cai

UMich



Simon Du

UW

Jianqing Fan

Princeton

Matthieu Geist

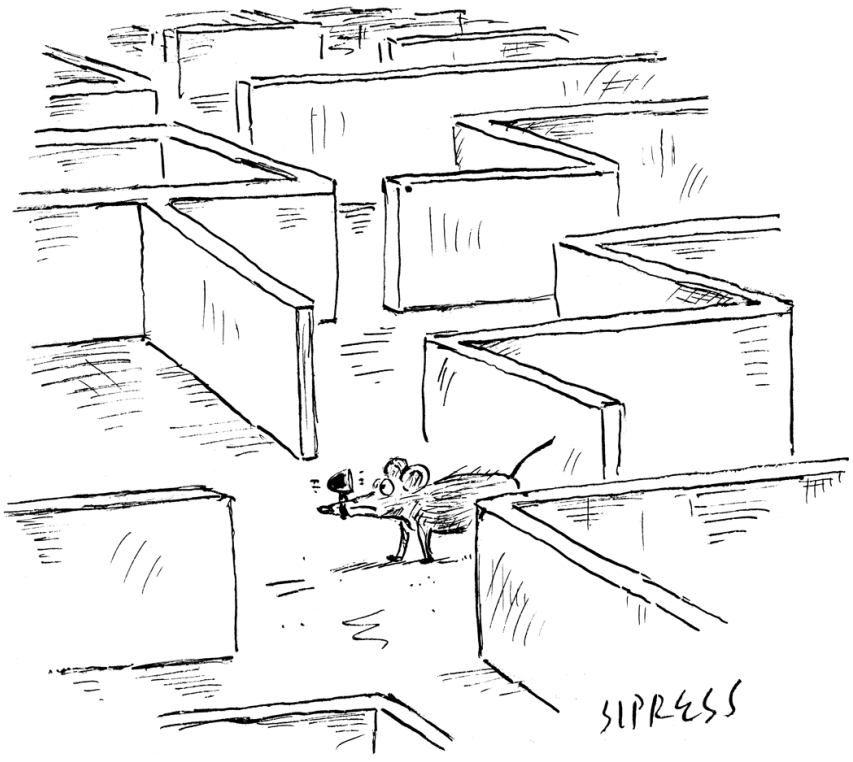
ESL

Jason Lee

Berkeley

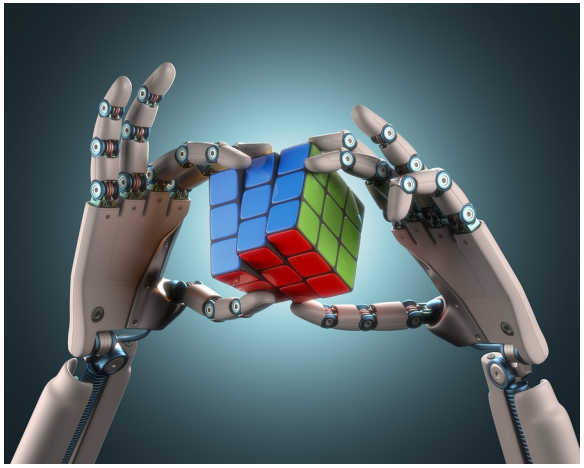
Reinforcement learning (RL)

In RL, an agent learns by interacting with an *unknown* environment through trial-and-error to maximize long-term total reward.

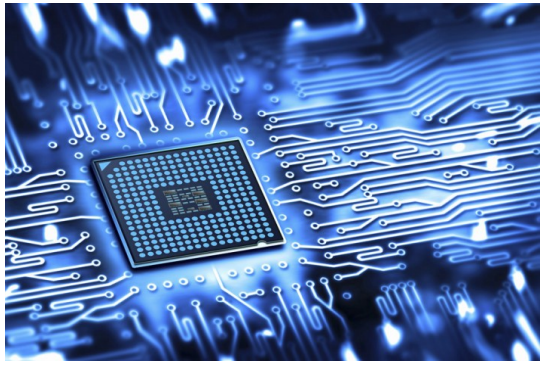


"Recalculating ... recalculating ..."

More successes of RL since AlphaGo



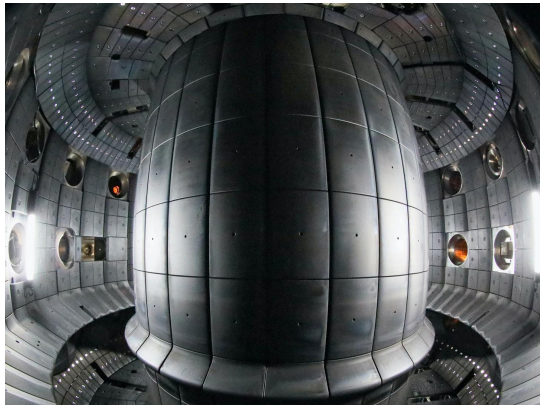
robotics



chip designs

resource management

strategic games



nuclear plant control

UAV and drones

One more: RL for foundation models

ChatGPT

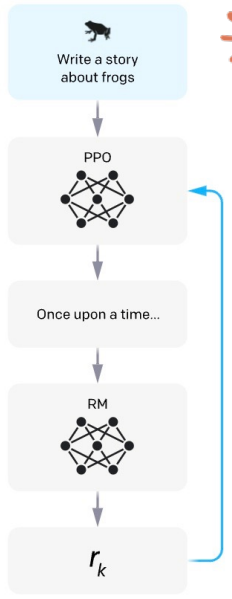
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Gemini

Claude

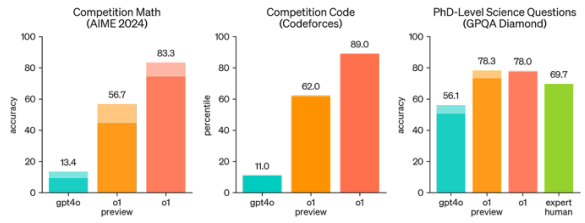
Meta AI

deepseek

What can I help with?

Ask anything

Search Reason



Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a-\sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a-\sqrt{a+x}} = x$, let's start by squaring both $\cdots$

$\left(\sqrt{a-\sqrt{a+x}}\right)^2 = x^2 \implies a-\sqrt{a+x} = x^2$.

Rearrange to isolate the inner square root term:

$(a-x^2)^2 = a+x \implies a^2-2ax^2+(x^2)^2 = a+x \implies x^4-2ax^2-x+(a^2-a) = 0$

$\ldots$

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$\sqrt{a-\sqrt{a+x}} = x$

First, let's square both sides:

$a-\sqrt{a+x} = x^2 \implies \sqrt{a+x} = a-x^2$

Next, I could square both sides again, treating the equation: $\cdots$

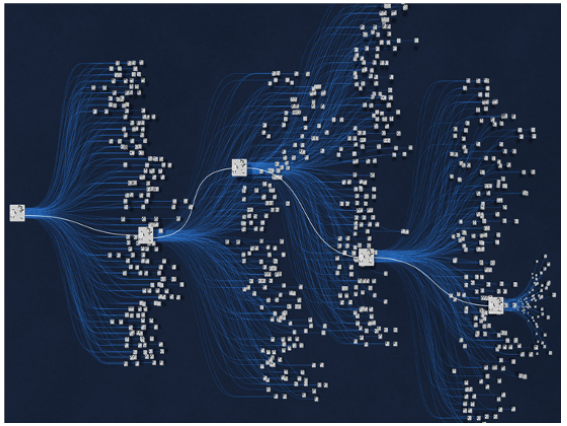
$\ldots$

Alignment: safety, human value..

Reasoning: math, coding...

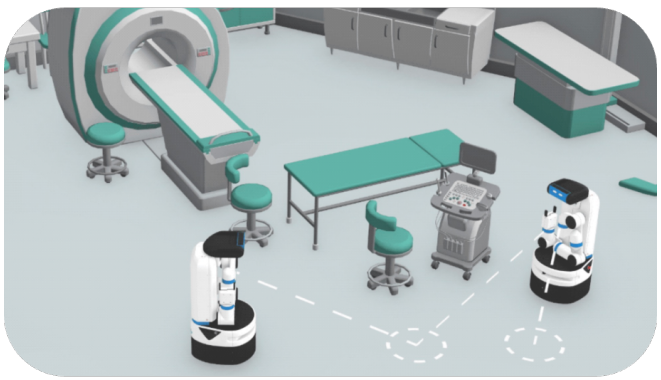
Challenges of RL

- explore or exploit: unknown or changing environments
- credit assignment problem: delayed rewards or feedback
- enormous state and action space



Data efficiency

Data collection might be expensive, time-consuming, or high-stakes



clinical trials

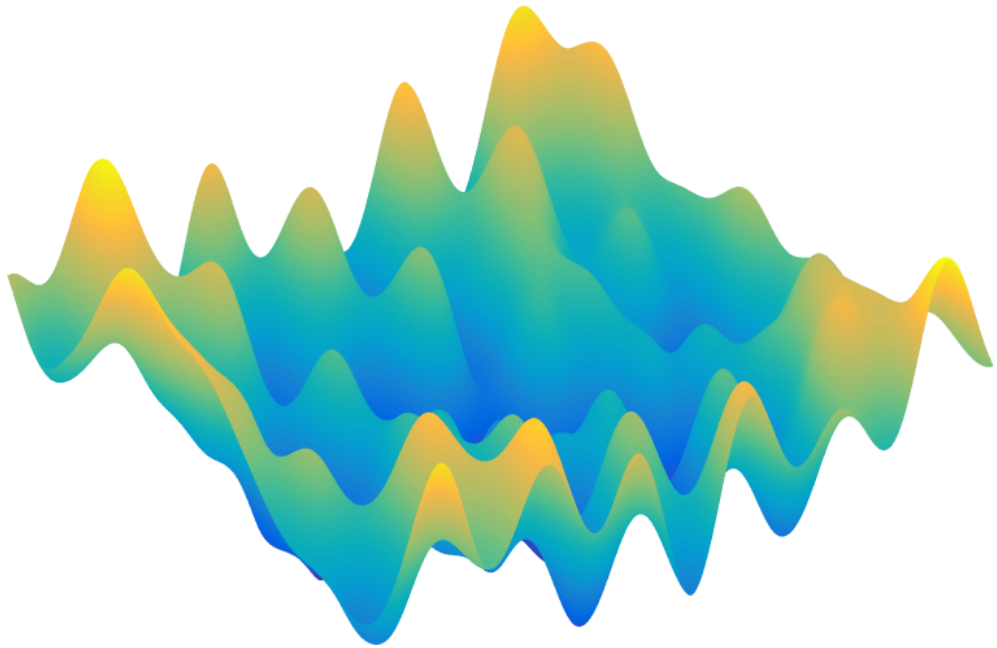
self-driving cars

Calls for design of sample-efficient RL algorithms!

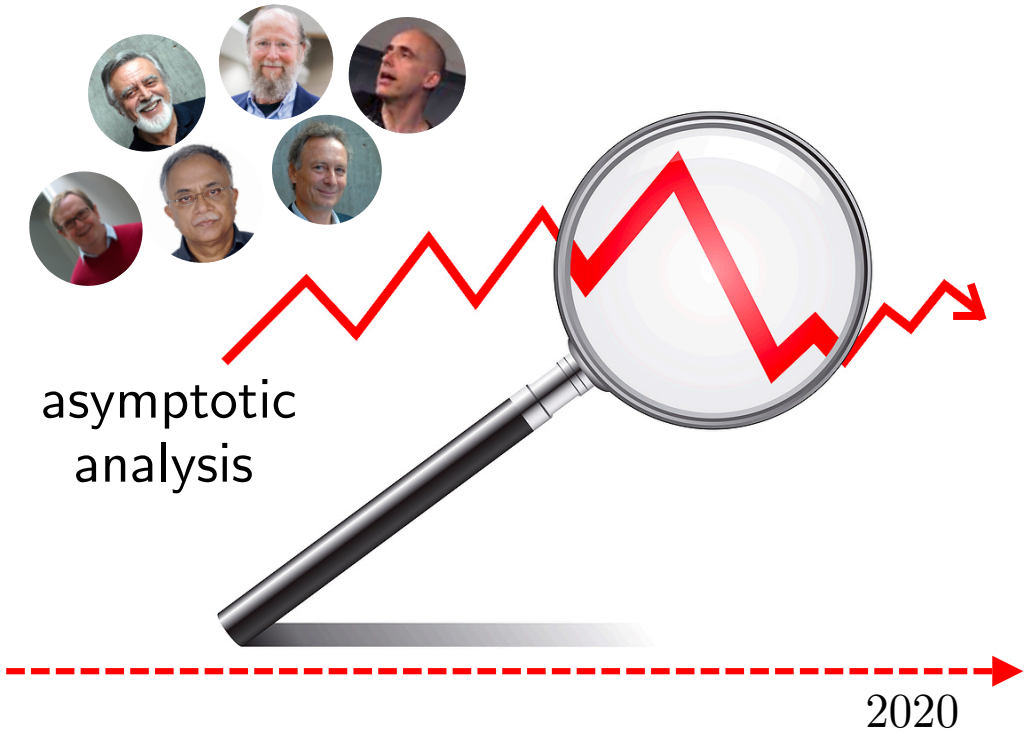
Computational efficiency

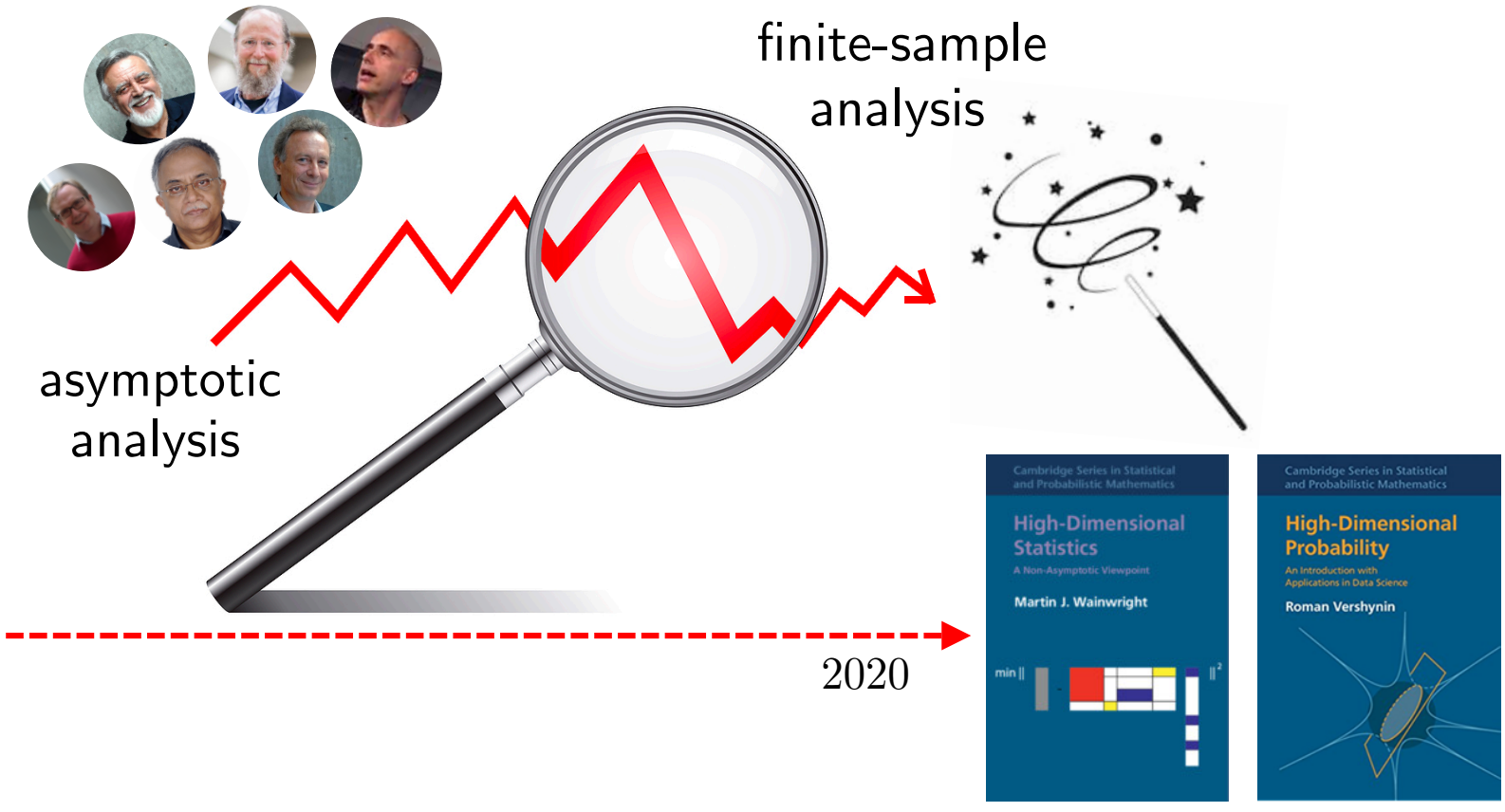
Running RL algorithms might take a long time ...

- enormous state-action space
- nonconvexity



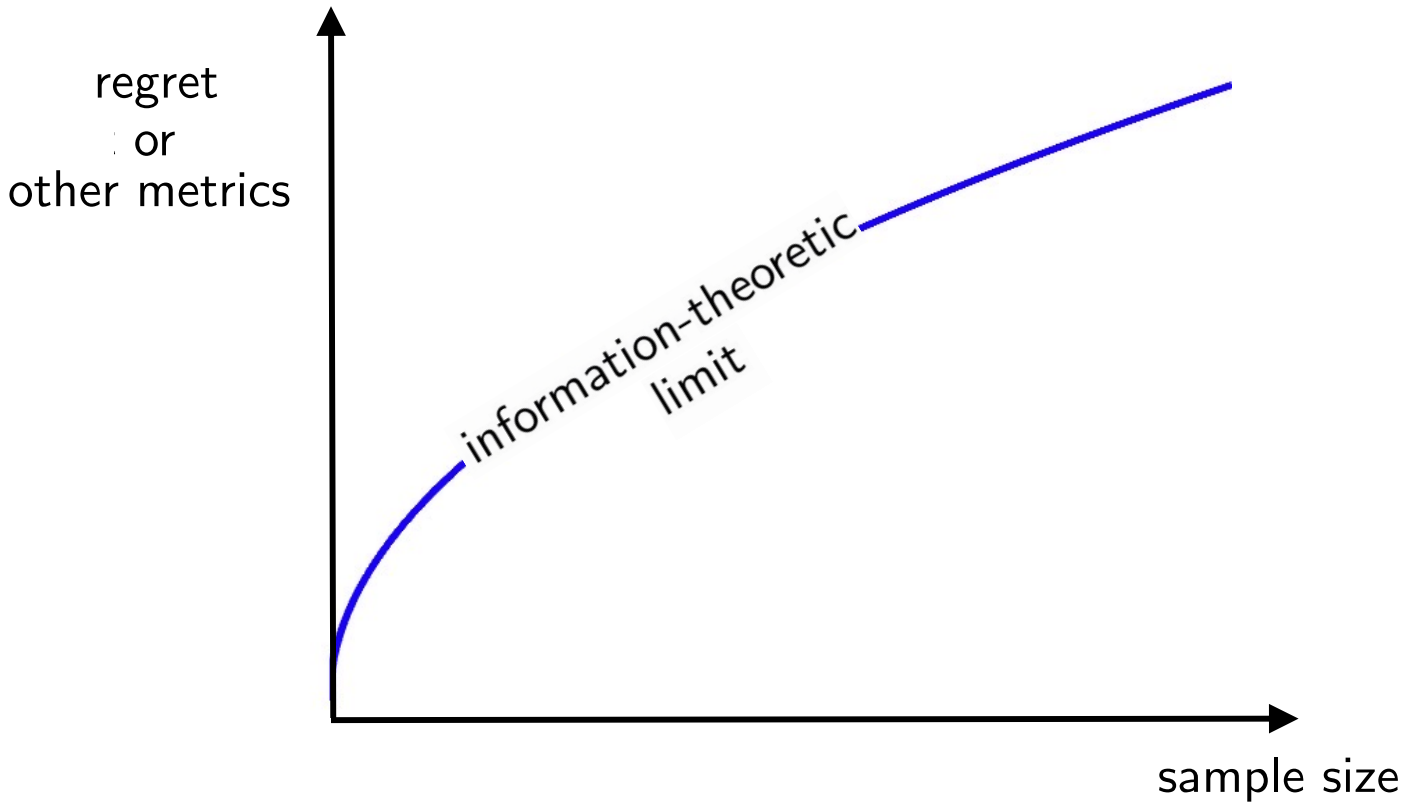
Calls for computationally efficient RL algorithms!


asymptotic
analysis
2020

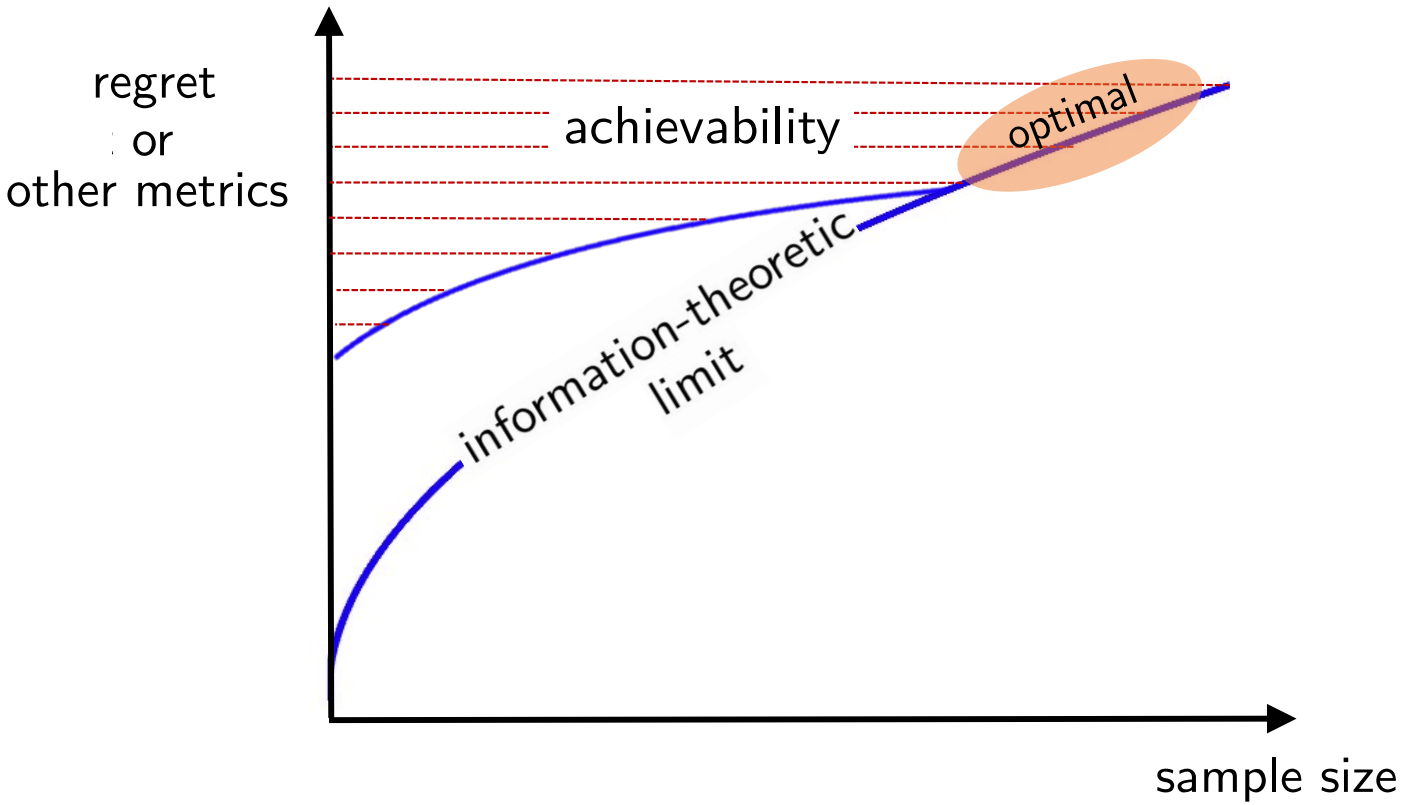


Understanding efficiency of contemporary RL requires a modern suite of non-asymptotic analysis

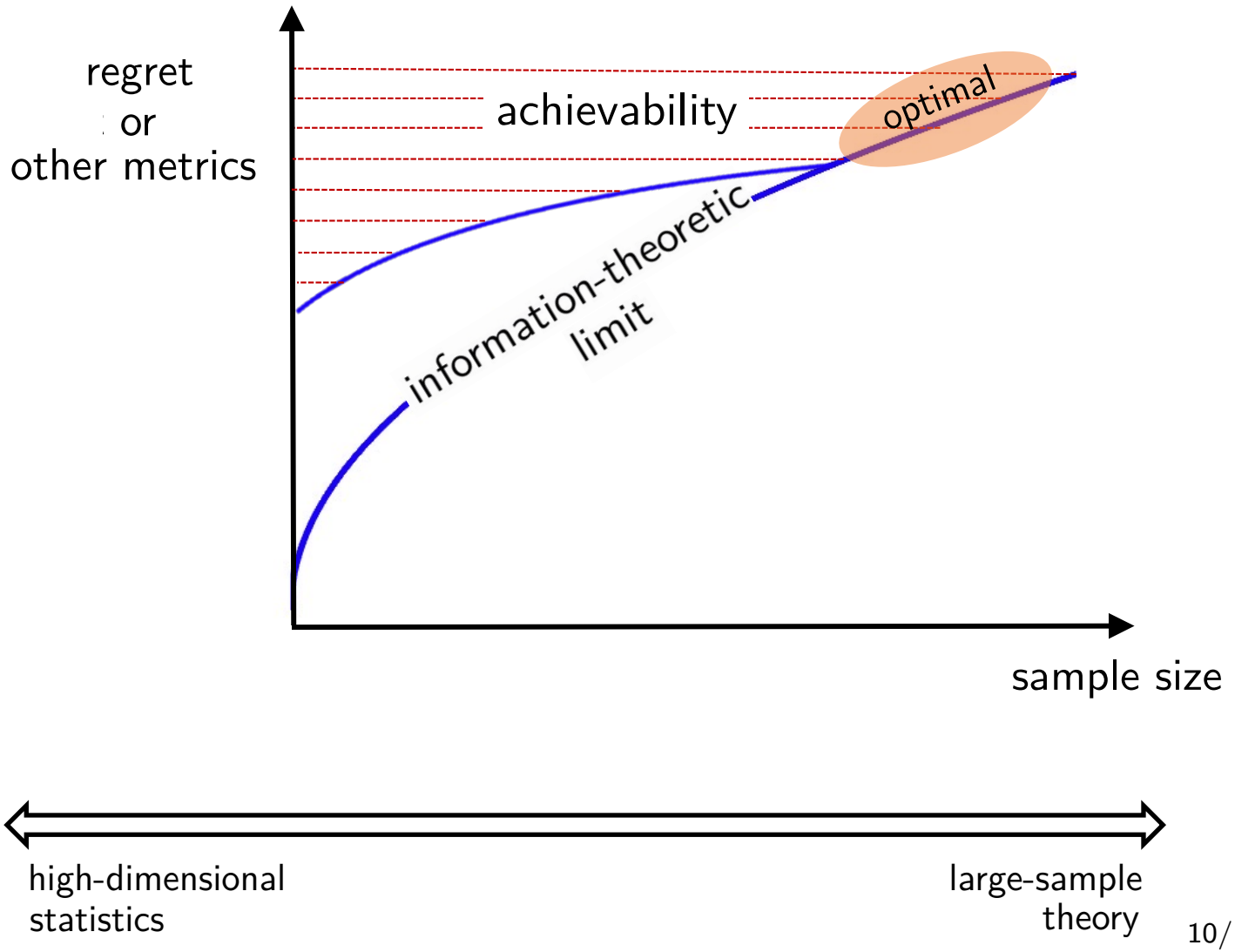
Sample complexity issues that permeate state-of-the-art RL theory



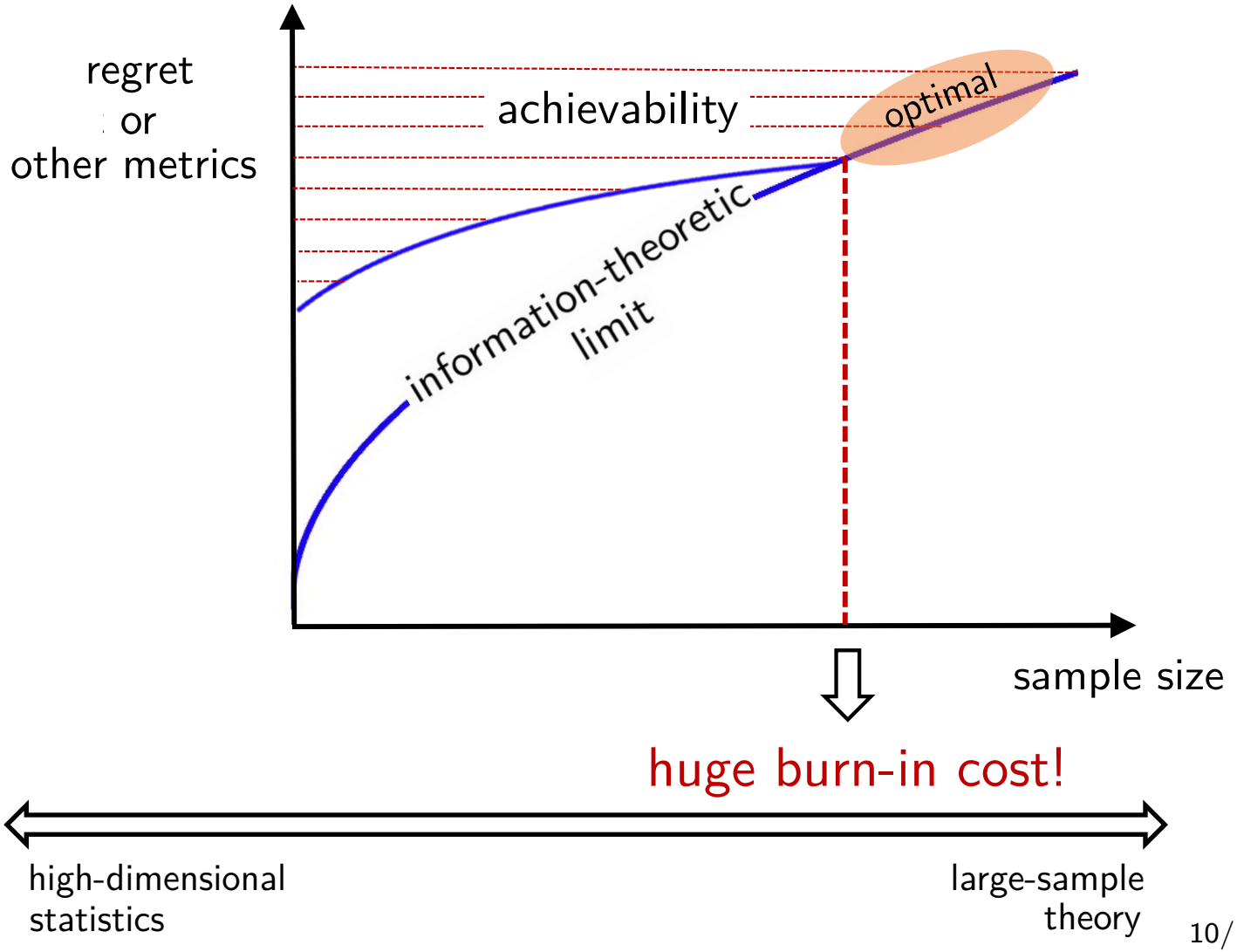
Sample complexity issues that permeate state-of-the-art RL theory



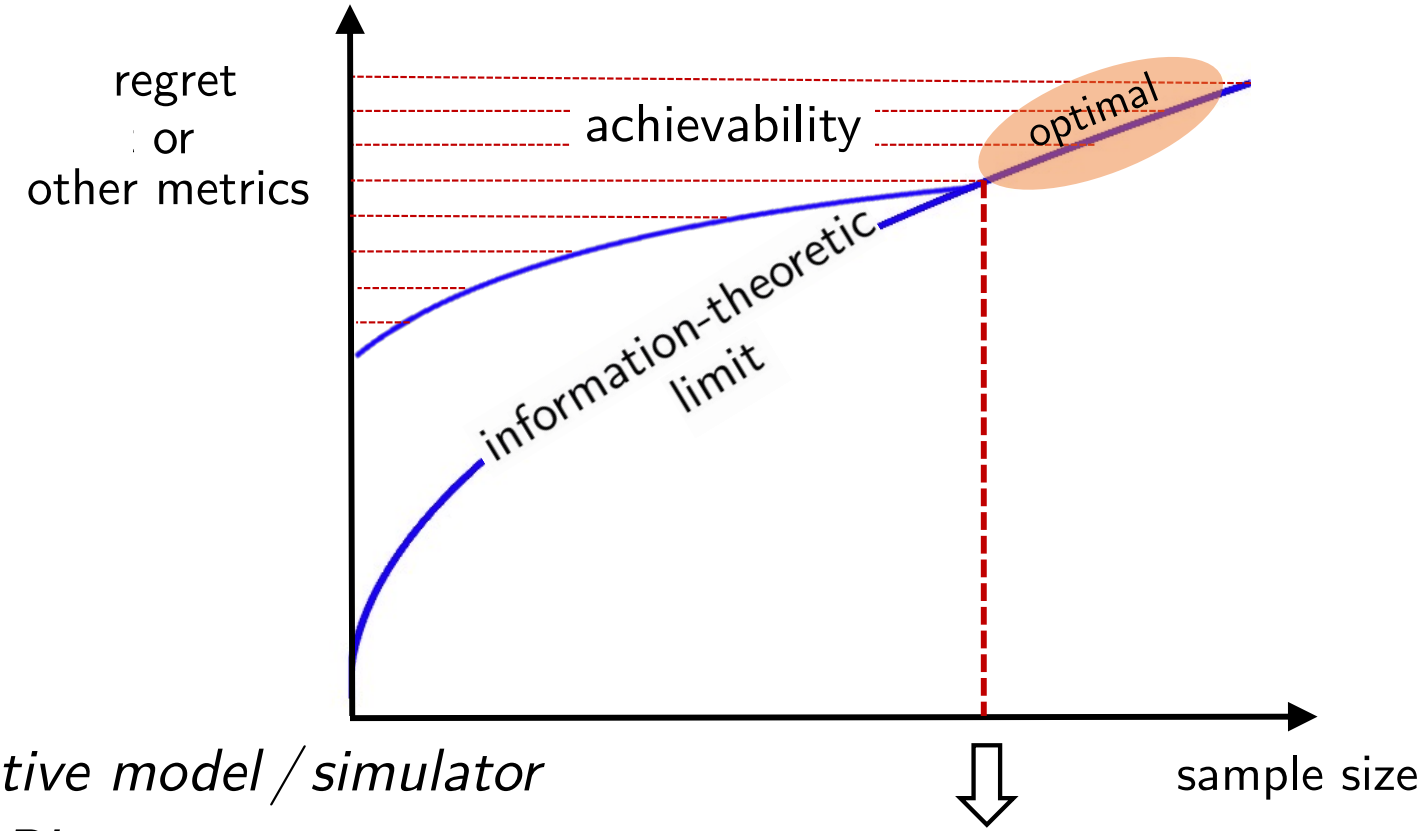
Sample complexity issues that permeate state-of-the-art RL theory



Sample complexity issues that permeate state-of-the-art RL theory

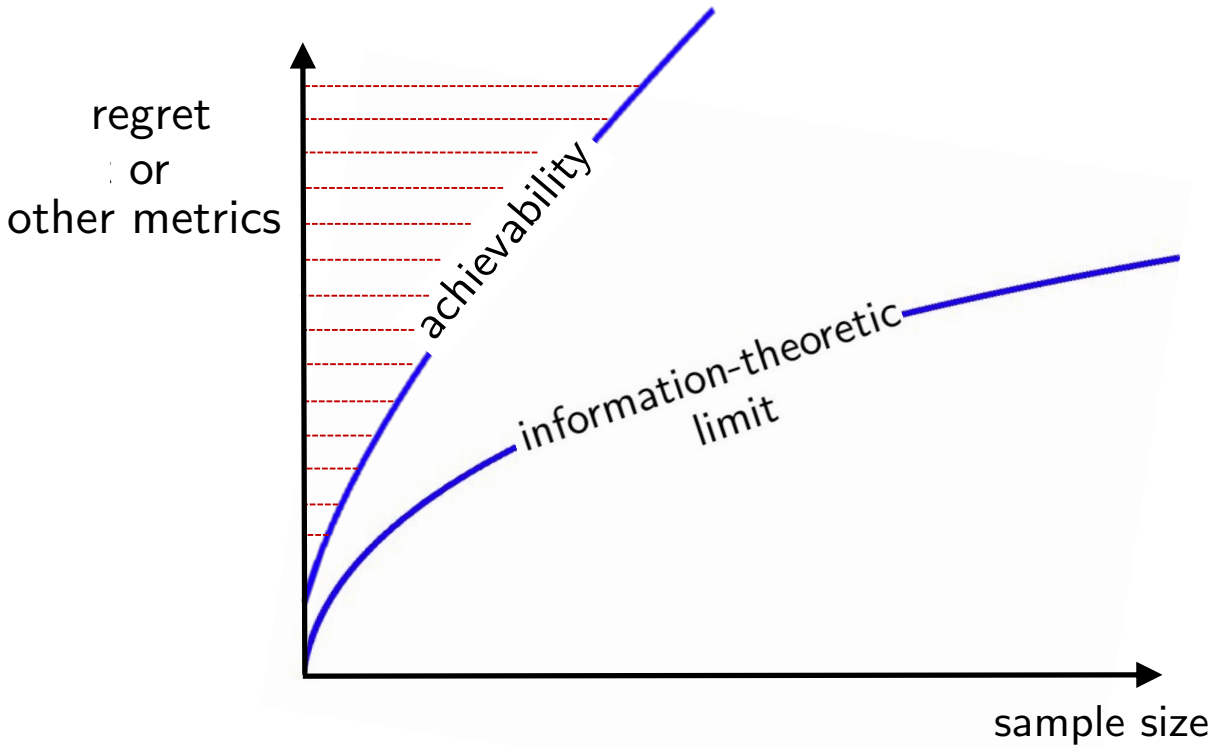


Sample complexity issues that permeate state-of-the-art RL theory



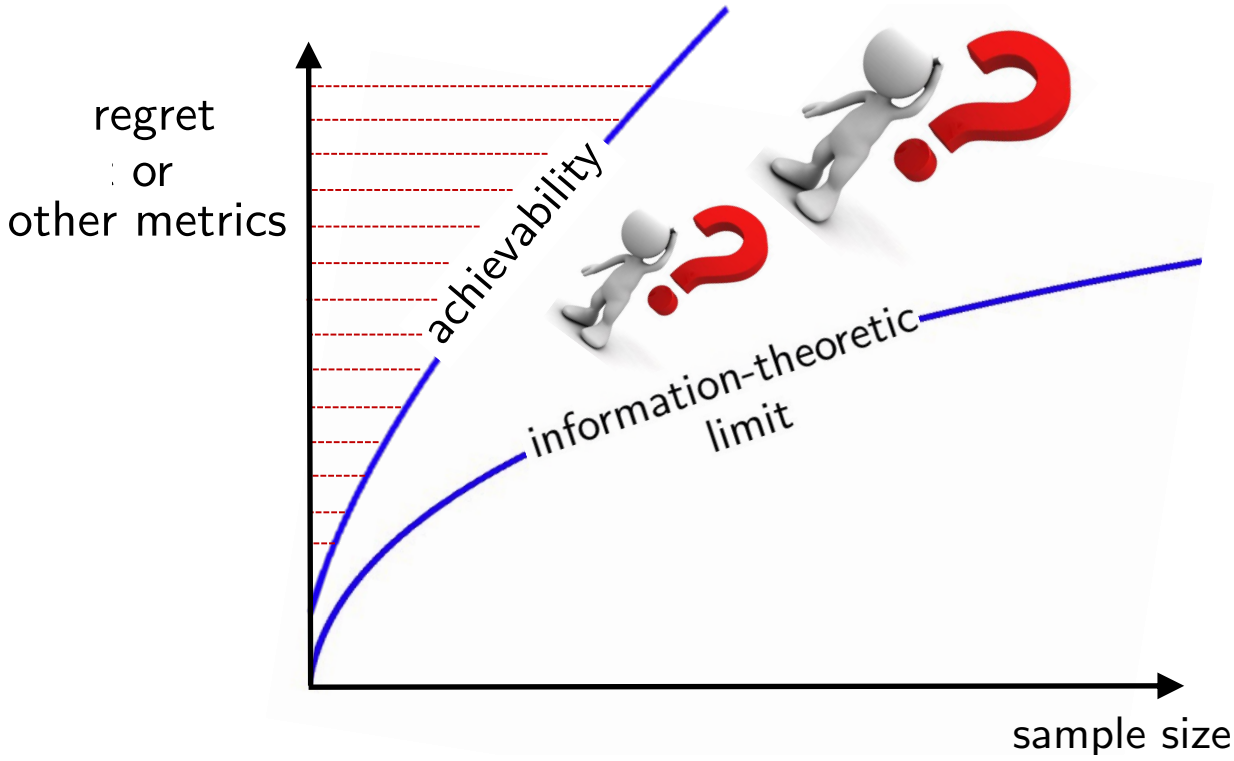
- *generative model / simulator*
- *online RL*
- *offline RL*
- ...

Sample complexity issues that permeate state-of-the-art RL theory



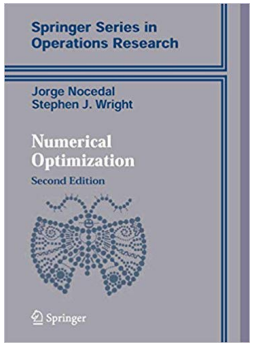
- *robust RL*
- *multi-agent RL*
- ...

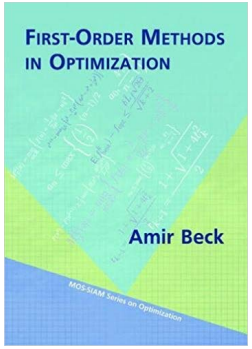
Sample complexity issues that permeate state-of-the-art RL theory



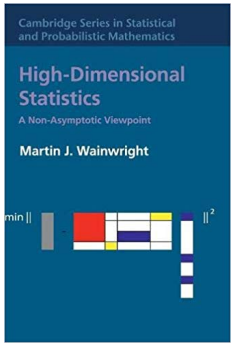
- *robust RL*
- *multi-agent RL*
- ...

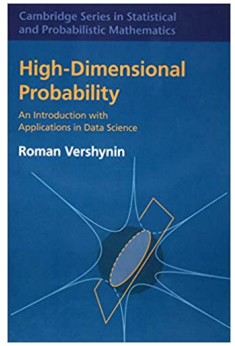
This tutorial





(large-scale) optimization

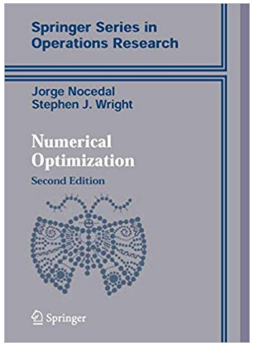


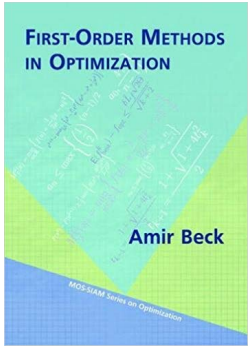


(high-dimensional) statistics

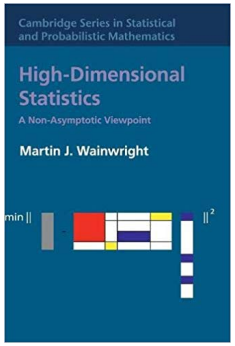
A taste of recent advances in understanding and designing sample- and computationally-efficient RL algorithms

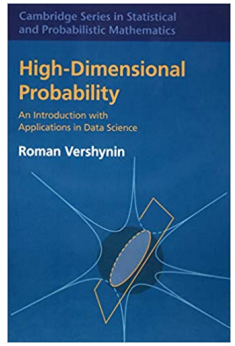
This tutorial





(large-scale) optimization

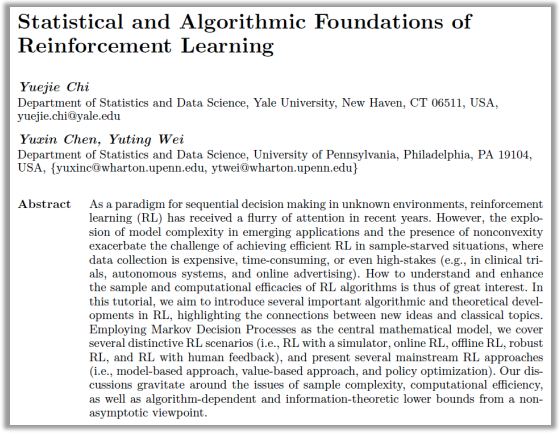




(high-dimensional) statistics

A taste of recent advances in understanding and designing sample- and computationally-efficient RL algorithms

1. Sample complexity of Q-Learning
2. Offline RL
3. Robust RL
4. RLHF (time permitting)



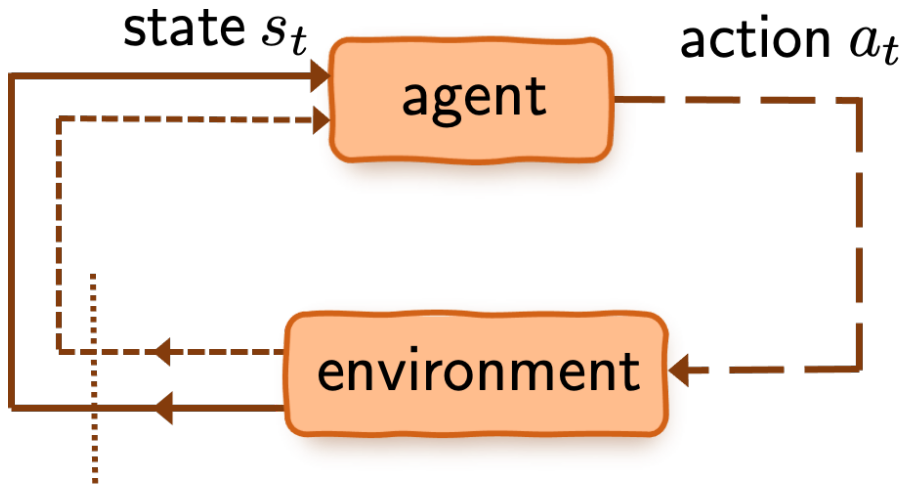
Statistical and Algorithmic Foundations of Reinforcement Learning

Yuejie Chi
Department of Statistics and Data Science, Yale University, New Haven, CT 06511, USA, yuejie.chi@yale.edu

Yuxin Chen, Yuting Wei
Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104, USA, {yuxinc@wharton.upenn.edu, ytwei@wharton.upenn.edu}

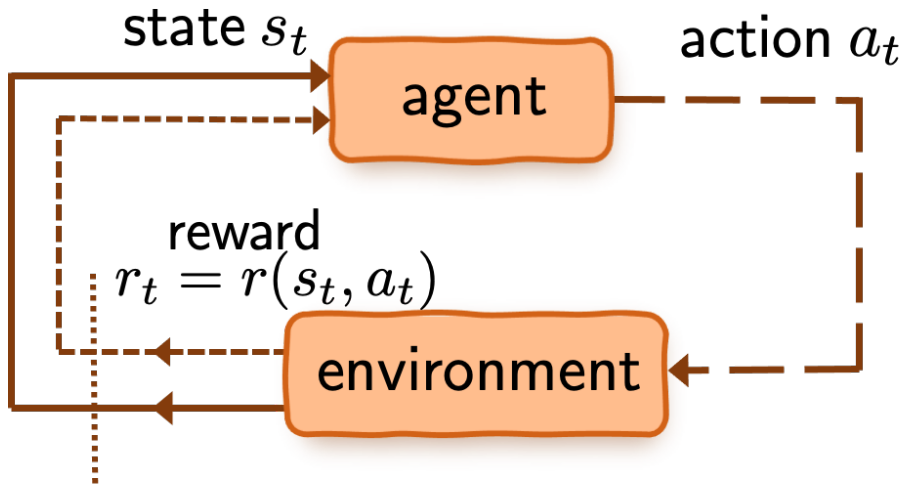
Abstract As a paradigm for sequential decision making in unknown environments, reinforcement learning (RL) has received a flurry of attention in recent years. However, the explosion of model complexity in emerging applications and the presence of nonconvexity exacerbate the challenge of achieving efficient RL in sample-starved situations, where data collection is expensive, time-consuming, or even high-stakes (e.g., in clinical trials, autonomous systems, and online advertising). How to understand and enhance the sample and computational efficacies of RL algorithms is thus of great interest. In this tutorial, we aim to introduce several important algorithmic and theoretical developments in RL, highlighting the connections between new ideas and classical topics. Employing Markov Decision Processes as the central mathematical model, we cover several distinctive RL scenarios (i.e., RL with a simulator, online RL, offline RL, robust RL, and RL with human feedback), and present several mainstream RL approaches (i.e., model-based approach, value-based approach, and policy optimization). Our discussions gravitate around the issues of sample complexity, computational efficiency, as well as algorithm-dependent and information-theoretic lower bounds from a non-asymptotic viewpoint.

Markov decision process (MDP)



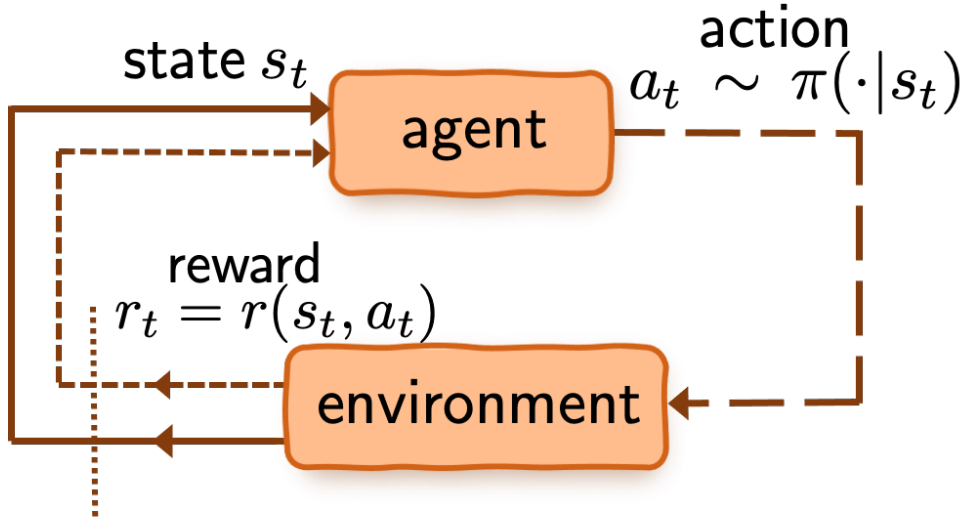
- $\mathcal{S} = \{1, \ldots, S\}$: state space (containing S states)
- $\mathcal{A} = \{1, \ldots, A\}$: action space (containing A actions)

Markov decision process (MDP)



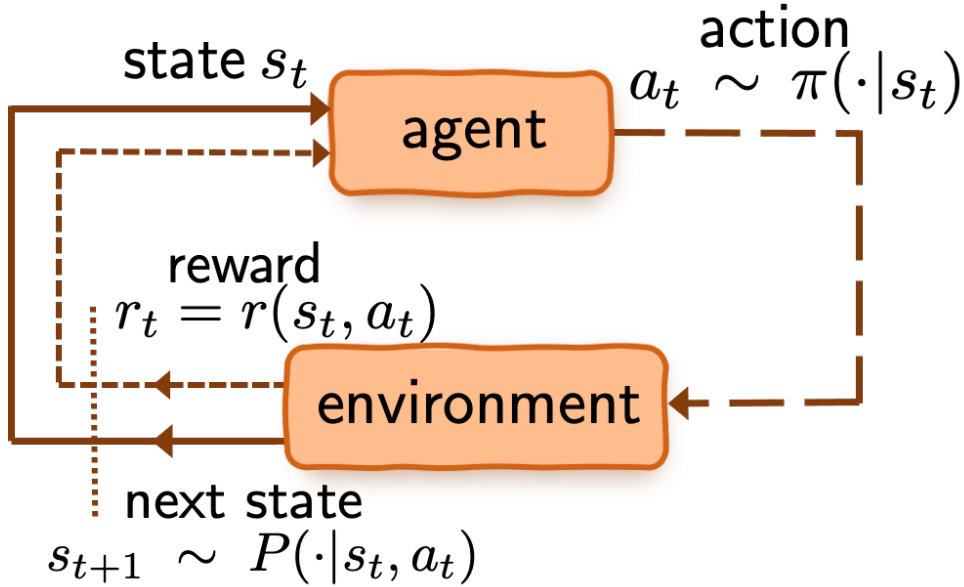
- $\mathcal{S} = \{1, \ldots, S\}$: state space (containing S states)
- $\mathcal{A} = \{1, \ldots, A\}$: action space (containing A actions)
- $r(s, a) \in [0, 1]$: immediate reward

Discounted infinite-horizon MDPs



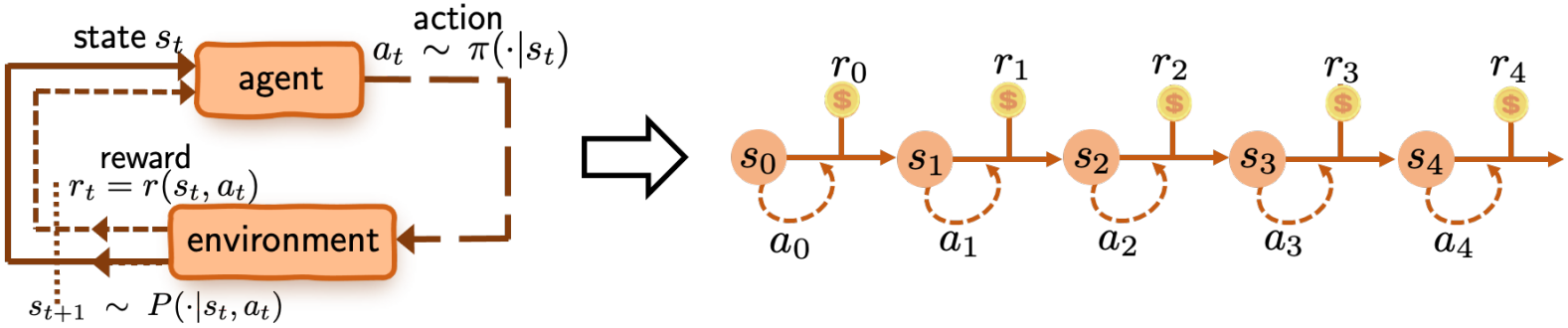
- $\mathcal{S} = \{1, \ldots, S\}$: state space (containing S states)
- $\mathcal{A} = \{1, \ldots, A\}$: action space (containing A actions)
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

Discounted infinite-horizon MDPs



- $\mathcal{S} = \{1, \ldots, S\}$: state space (containing S states)
- $\mathcal{A} = \{1, \ldots, A\}$: action space (containing A actions)
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)
- $P(\cdot|s, a)$: unknown transition probabilities

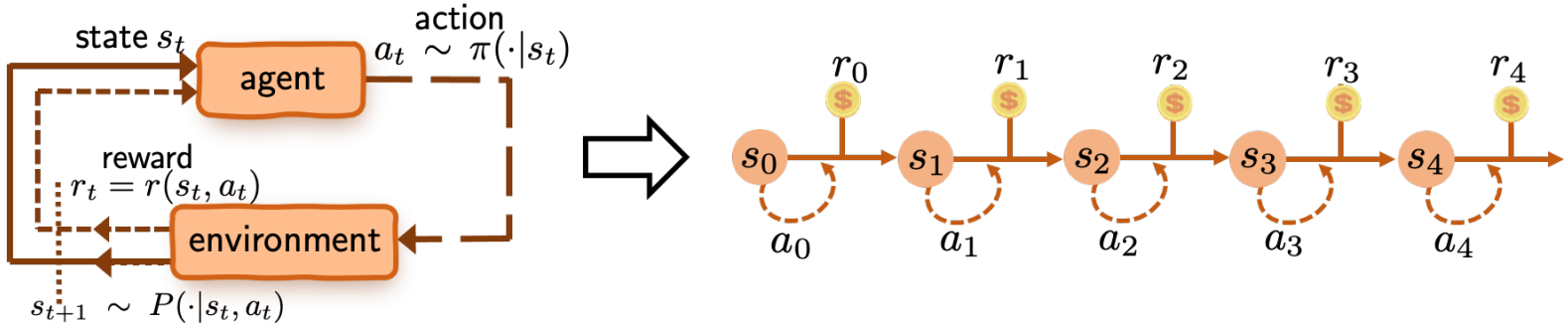
Value function



Value of policy π: cumulative discounted reward

$$\forall s \in \mathcal{S}: \quad V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right]$$

Value function

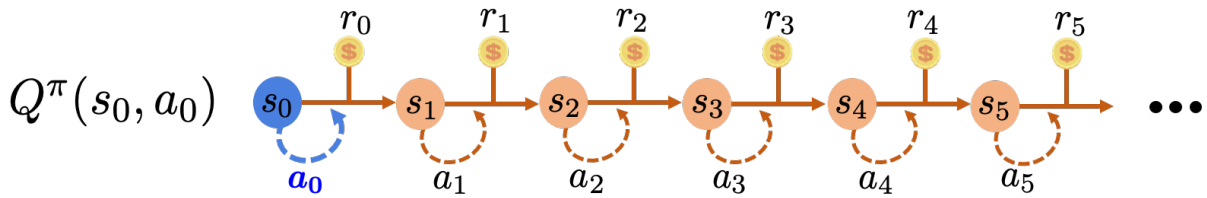


Value of policy π: cumulative discounted reward

$$\forall s \in \mathcal{S}: \quad V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right]$$

- $\gamma \in [0, 1)$: discount factor
  - take $\gamma \to 1$ to approximate long-horizon MDPs
  - **effective horizon**: $\frac{1}{1-\gamma}$

Q-function (action-value function)

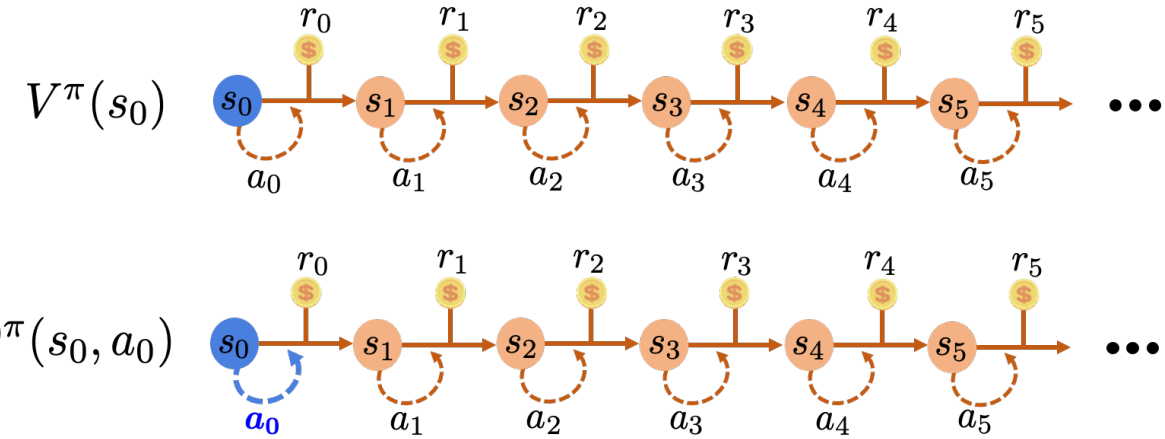


Q-function of policy π:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q^\pi(s,a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_0 = a\right]$$

- $(\cancel{a_0}, s_1, a_1, s_2, a_2, \cdots)$: induced by policy π

Q-function (action-value function)

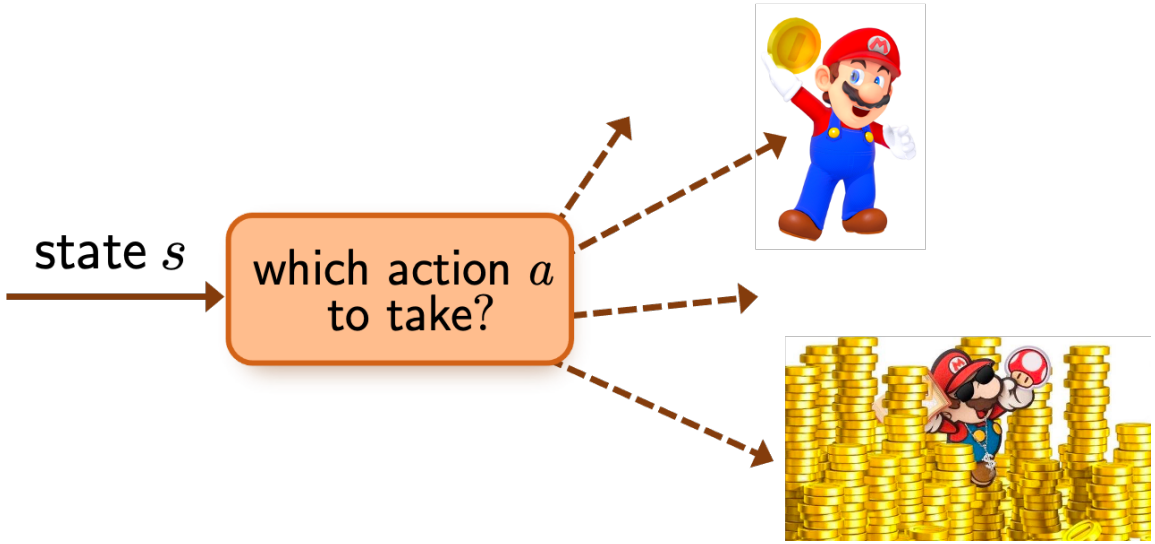


Q-function of policy π:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q^\pi(s,a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_0 = a\right]$$

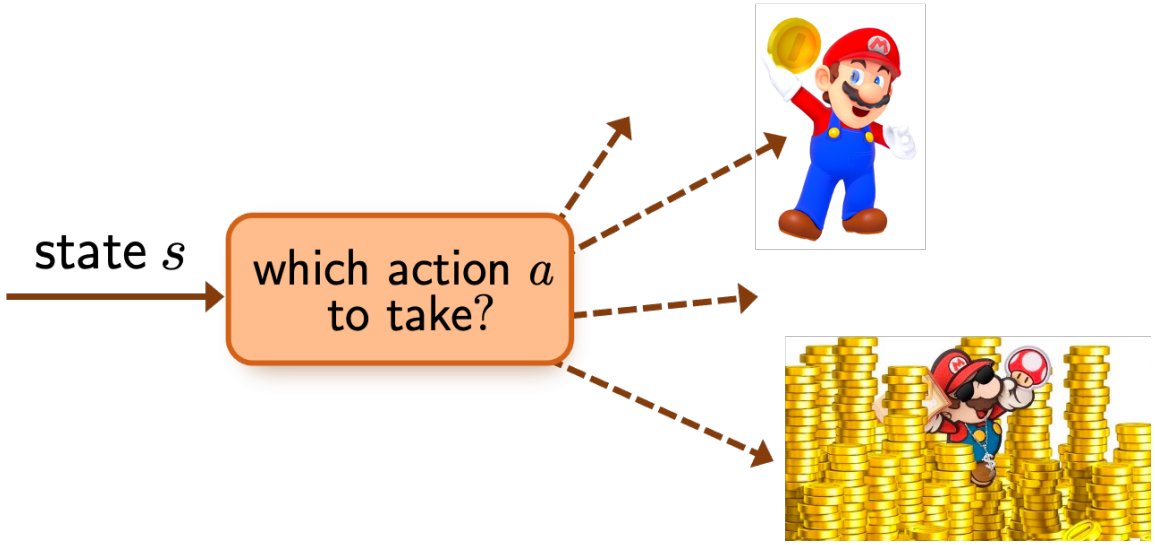
- $(\cancel{a_0}, s_1, a_1, s_2, a_2, \cdots)$: induced by policy π

Optimal policy and optimal value



- **optimal policy** $\pi^\star$: maximizing value function $\max_\pi V^\pi$

Optimal policy and optimal value



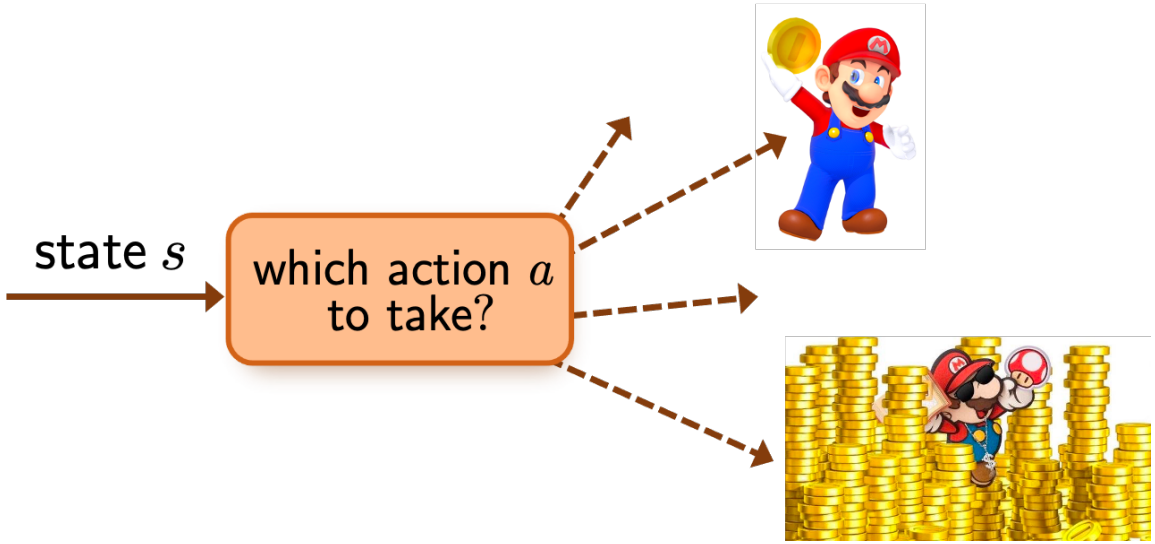
- **optimal policy** $\pi^\star$: maximizing value function $\max_\pi V^\pi$

Theorem (Puterman'94)

For infinite horizon discounted MDP, there always exists a deterministic policy $\pi^\star$, such that

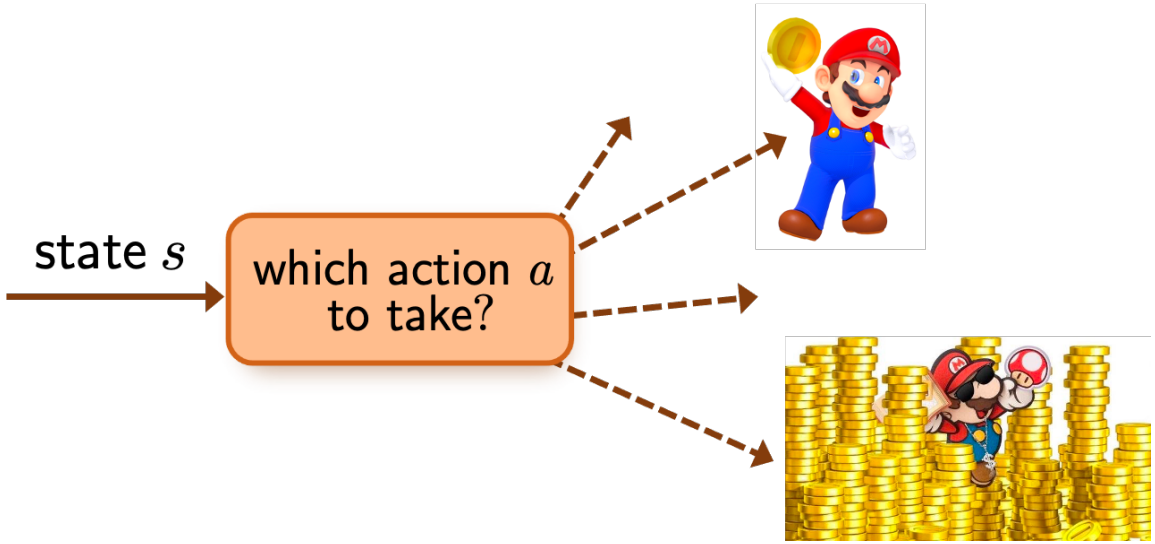
$$V^{\pi^\star}(s) \geq V^\pi(s), \quad \forall s, \text{ and } \pi.$$

Optimal policy and optimal value



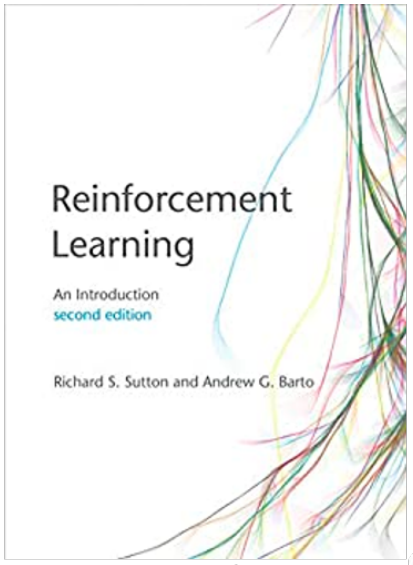
- **optimal policy** $\pi^\star$: maximizing value function $\max_\pi V^\pi$
- **optimal value / Q function**: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$

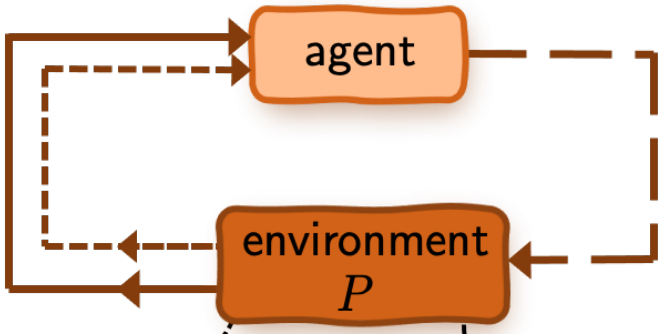
Optimal policy and optimal value

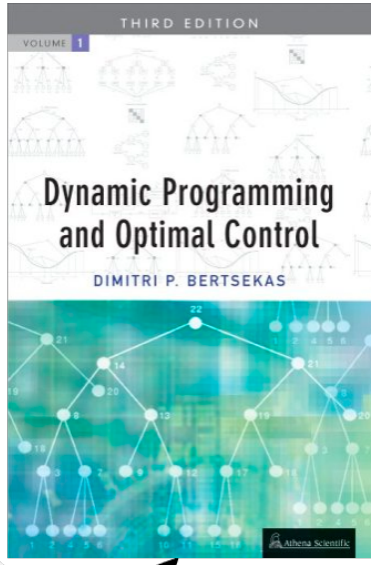


- **optimal policy** $\pi^\star$: maximizing value function $\max_\pi V^\pi$
- **optimal value / Q function**: $V^\star := V^{\pi^\star}$, $Q^\star := Q^{\pi^\star}$
- A question to keep in mind: *how to find optimal* $\pi^\star$*?*

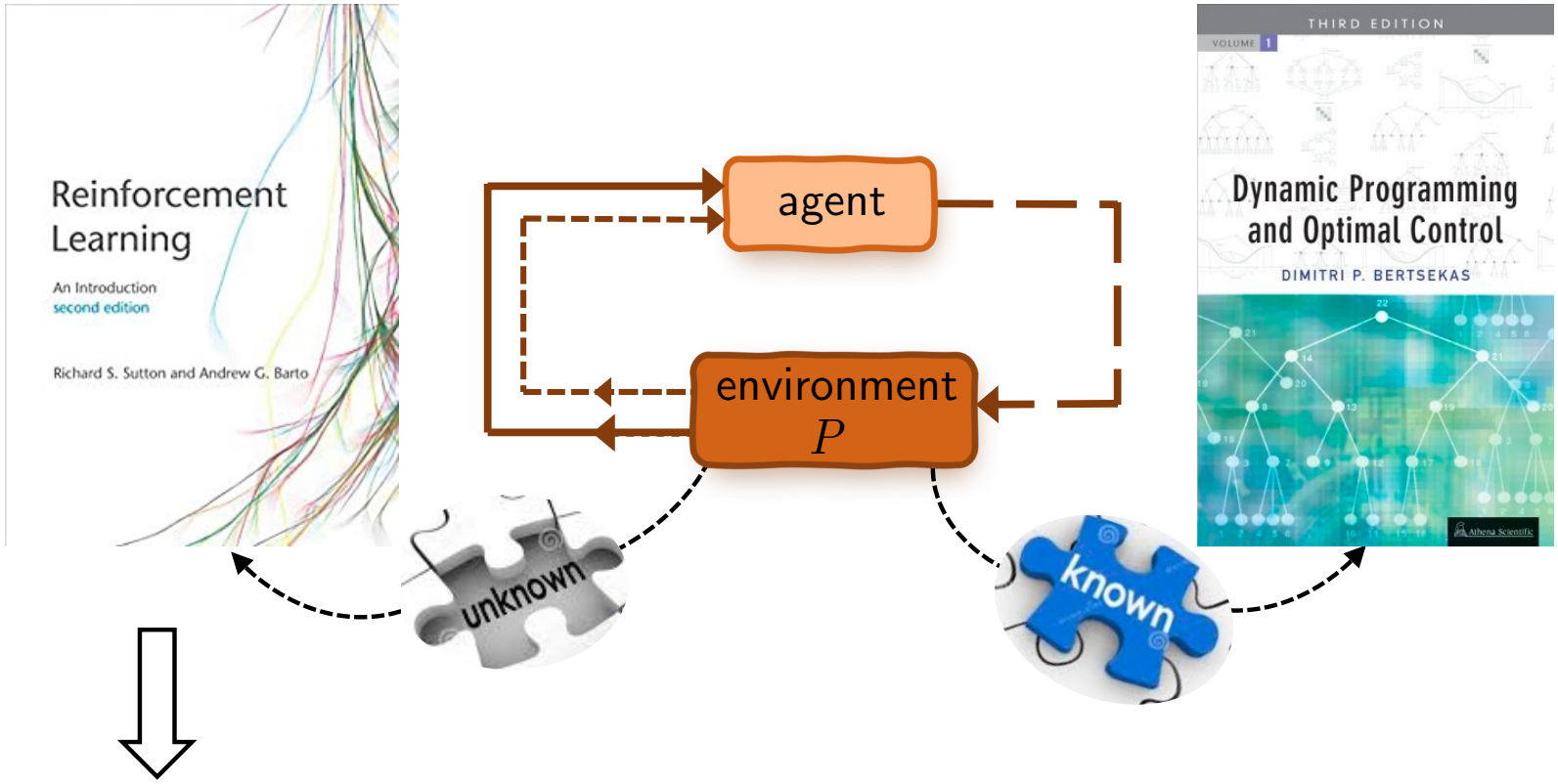
RL: when the model is unknown . . .





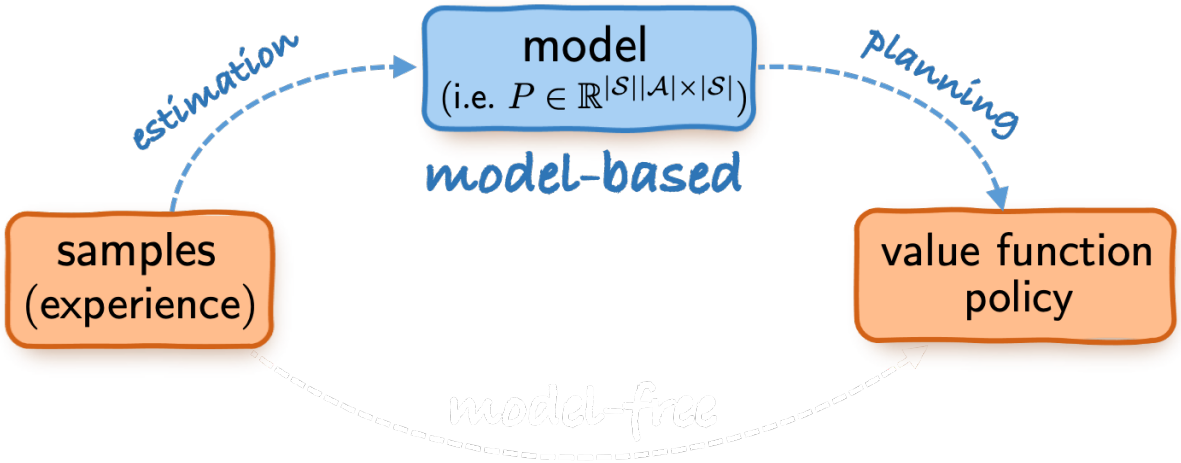


RL: when the model is unknown . . .



Need to learn optimal policy from samples w/o model specification

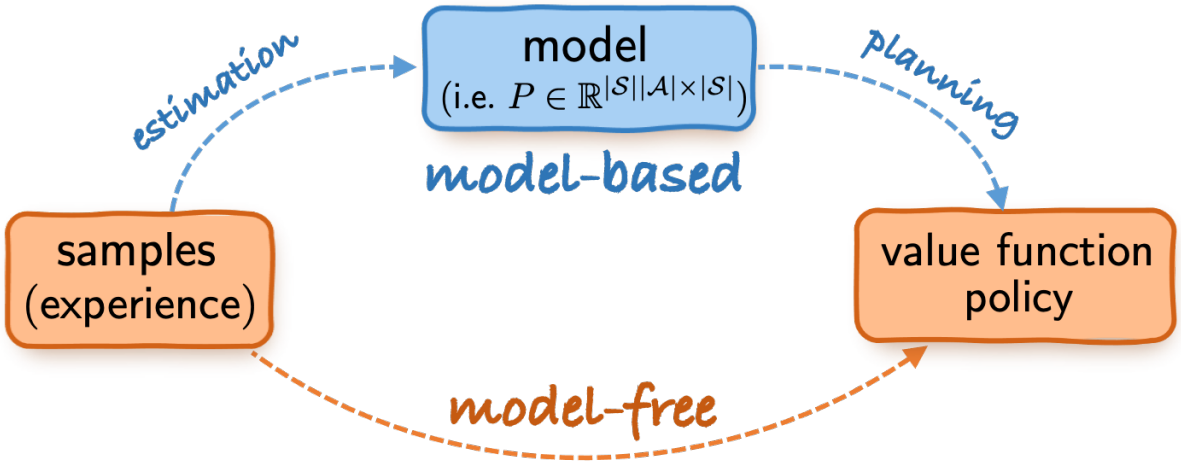
Two approaches



Model-based approach ("plug-in")

1. build an empirical estimate $\widehat{P}$ for P
2. planning based on the empirical $\widehat{P}$

Two approaches



Model-based approach ("plug-in")

1. build an empirical estimate $\widehat{P}$ for P
2. planning based on the empirical $\widehat{P}$

Model-free approach

— learning w/o estimating the model explicitly

Sampling mechanisms

1. RL w/ a generative model (a.k.a. simulator)
    - can query arbitrary state-action pairs to draw samples

Sampling mechanisms

1. RL w/ a generative model (a.k.a. simulator)
    - can query arbitrary state-action pairs to draw samples
2. online RL
    - execute MDP in real time to obtain sample trajectories

Sampling mechanisms

1. RL w/ a generative model (a.k.a. simulator)
    - can query arbitrary state-action pairs to draw samples
2. online RL
    - execute MDP in real time to obtain sample trajectories
3. offline RL
    - use pre-collected historical data

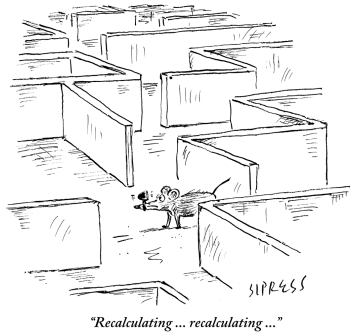
Sampling mechanisms

1. RL w/ a generative model (a.k.a. simulator)
   - can query arbitrary state-action pairs to draw samples
2. online RL
   - execute MDP in real time to obtain sample trajectories
3. offline RL
   - use pre-collected historical data

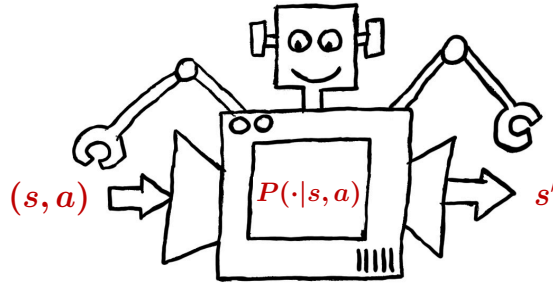
Question: *how many samples are sufficient to learn an* $\underbrace{\varepsilon\text{-optimal policy}}_{V^{\widehat{\pi}} \geq V^{\star} - \varepsilon}$*?*

Exploration vs exploitation

offline RL

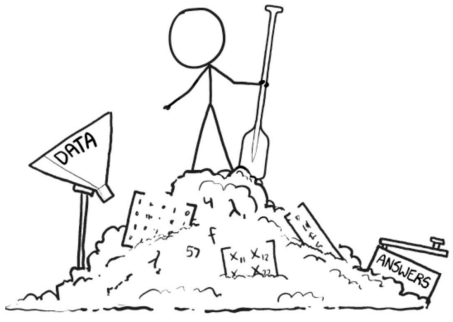


online RL

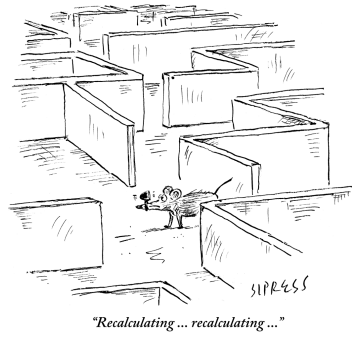


generative model

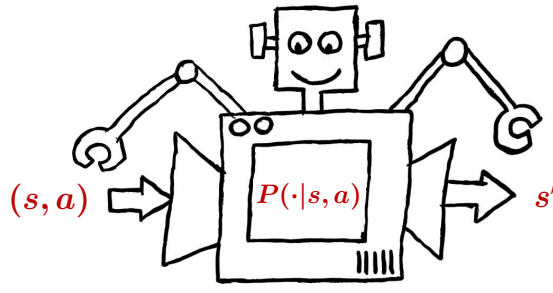
Exploration vs exploitation



offline RL



online RL



generative model

Varying levels of trade-offs between exploration and exploitation.

Sample Complexity of Q-Learning

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

Stochastic approximation for solving the **Bellman equation**

Robbins & Monro, 1951

$$\mathcal{T}(Q) - Q = 0$$

where

$$\mathcal{T}(Q)(s,a) := \underbrace{r(s,a)}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \underbrace{\left[\max_{a' \in \mathcal{A}} Q(s',a')\right]}_{\text{next state's value}}.$$

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

Stochastic approximation for solving Bellman equation $\mathcal{T}(Q) - Q = 0$

$$\underbrace{Q_{t+1}(s,a) = Q_t(s,a) + \eta_t\big(\mathcal{T}_t(Q_t)(s,a) - Q_t(s,a)\big)}_{\text{sample transition } (s,a,s')}, \quad t \geq 0$$

Q-learning: a stochastic approximation algorithm

Chris Watkins

Peter Dayan

Stochastic approximation for solving Bellman equation $\mathcal{T}(Q) - Q = 0$

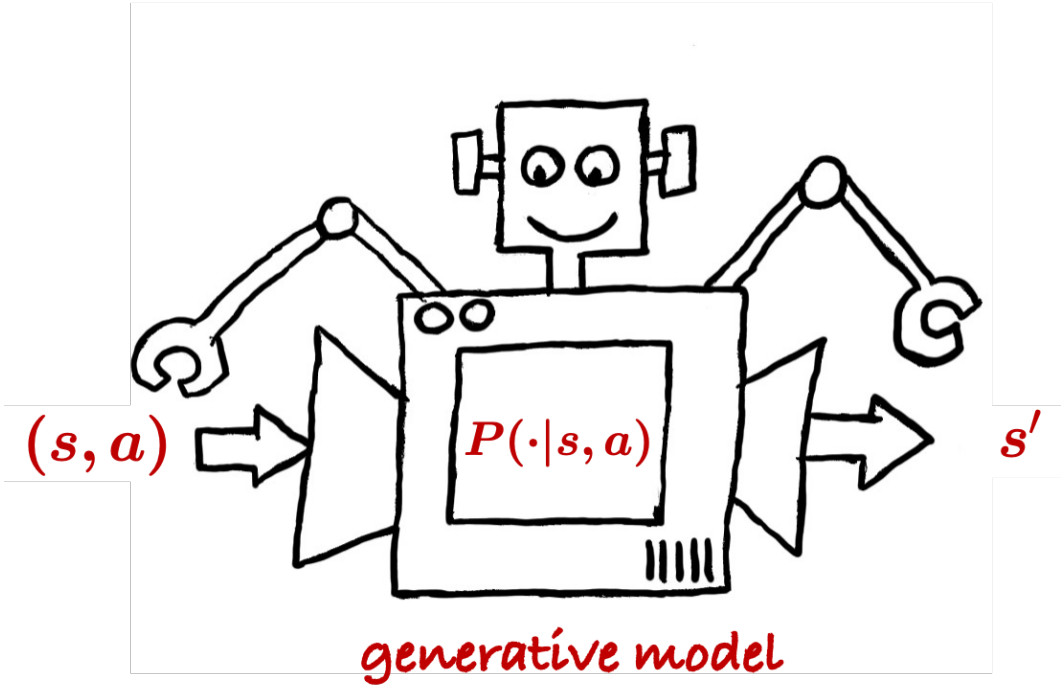
$$\underbrace{Q_{t+1}(s,a) = Q_t(s,a) + \eta_t\big(\mathcal{T}_t(Q_t)(s,a) - Q_t(s,a)\big)}_{\text{sample transition } (s,a,s')}, \quad t \geq 0$$

$$\mathcal{T}_t(Q)(s,a) = r(s,a) + \gamma \max_{a'} Q(s',a')$$

$$\mathcal{T}(Q)(s,a) = r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \Big[\max_{a'} Q(s',a')\Big]$$

A generative model / simulator

— *Kearns, Singh, 1999*



Each iteration, draw an independent sample (s, a, s') for given (s, a)

Synchronous Q-learning

Chris Watkins

Peter Dayan

for $t = 0, 1, \ldots, T$

  for each $(s, a) \in \mathcal{S} \times \mathcal{A}$

    draw a sample (s, a, s'), run

$$Q_{t+1}(s,a) = (1-\eta_t)Q_t(s,a) + \eta_t\Big\{r(s,a) + \gamma \max_{a'} Q_t(s',a')\Big\}$$

synchronous: all state-action pairs are updated simultaneously

- total sample size: TSA

ℓ_∞-**sample complexity:** how many samples are required to learn an $\underbrace{\varepsilon\text{-optimal policy}}_{\forall s:\ V^{\widehat{\pi}}(s) \geq V^\star(s) - \varepsilon}$?

Minimax lower bound

Theorem (minimax lower bound; Azar et al., 2013)

For all $\varepsilon \in [0, \frac{1}{1-\gamma})$, there exists some MDP such that the total number of samples need to be at least

$$\widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right)$$

to achieve $V^\star - V^{\widehat{\pi}} \leq \varepsilon$, where $\widehat{\pi}$ is the output of any RL algorithm.

Minimax lower bound

Theorem (minimax lower bound; Azar et al., 2013)

For all $\varepsilon \in [0, \frac{1}{1-\gamma})$, there exists some MDP such that the total number of samples need to be at least

$$\widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\right)$$

to achieve $V^\star - V^{\widehat{\pi}} \leq \varepsilon$, where $\widehat{\pi}$ is the output of any RL algorithm.

- holds for both finding the optimal Q-function and the optimal policy over the entire range of ε
- much smaller than the model dimension S^2A

Sample complexity of synchronous Q-learning

Theorem (Li, Cai, Chen, Wei, Chi '21, OR'24)

For any $0 < \varepsilon \leq 1$, *synchronous Q-learning yields* $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ *with high prob. and* $\mathbb{E}[\|\widehat{Q} - Q^\star\|_\infty] \leq \varepsilon$, *with sample size at most*

$$\begin{cases} \widetilde{O}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right) & \text{if } A \geq 2 \\ \widetilde{O}\left(\frac{S}{(1-\gamma)^3\varepsilon^2}\right) & \text{if } A = 1 \qquad (\textit{TD learning}) \end{cases}$$

Sample complexity of synchronous Q-learning

Theorem (Li, Cai, Chen, Wei, Chi '21, OR'24)

For any $0 < \varepsilon \leq 1$, *synchronous Q-learning yields* $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ *with high prob. and* $\mathbb{E}[\|\widehat{Q} - Q^\star\|_\infty] \leq \varepsilon$, *with sample size at most*

$$\begin{cases} \widetilde{O}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right) & \text{if } A \geq 2 \\ \widetilde{O}\left(\frac{S}{(1-\gamma)^3\varepsilon^2}\right) & \text{if } A = 1 \qquad (\textit{TD learning}) \end{cases}$$

- Covers both *constant* and *rescaled linear* learning rates:

$$\eta_t \equiv \frac{1}{1 + \frac{c_1(1-\gamma)T}{\log^2 T}} \quad \text{or} \quad \eta_t = \frac{1}{1 + \frac{c_2(1-\gamma)t}{\log^2 T}}$$

Sample complexity of synchronous Q-learning

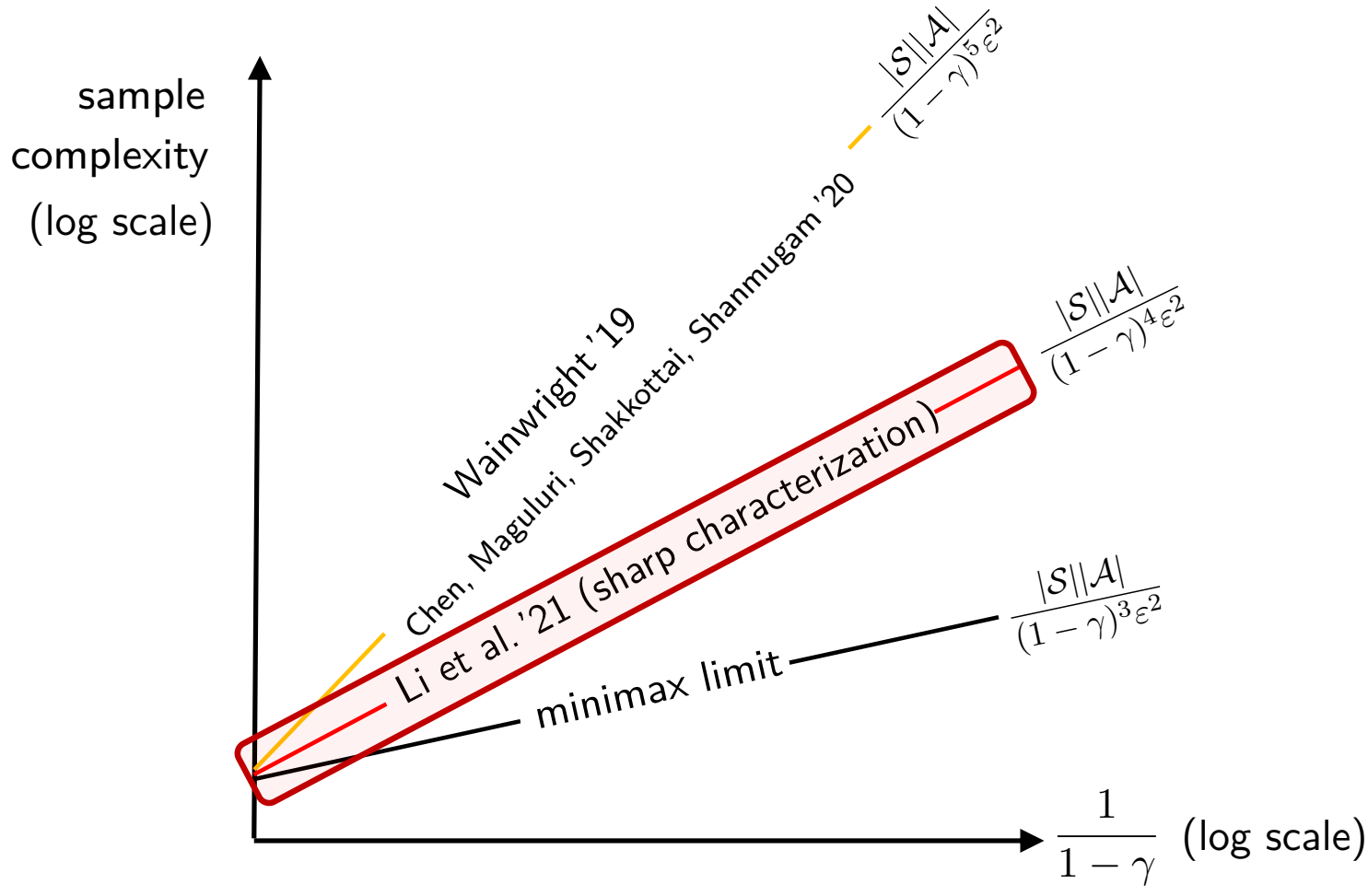
Theorem (Li, Cai, Chen, Wei, Chi '21, OR'24)

For any $0 < \varepsilon \leq 1$, synchronous Q-learning yields $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$ with high prob. and $\mathbb{E}[\|\widehat{Q} - Q^\star\|_\infty] \leq \varepsilon$, with sample size at most

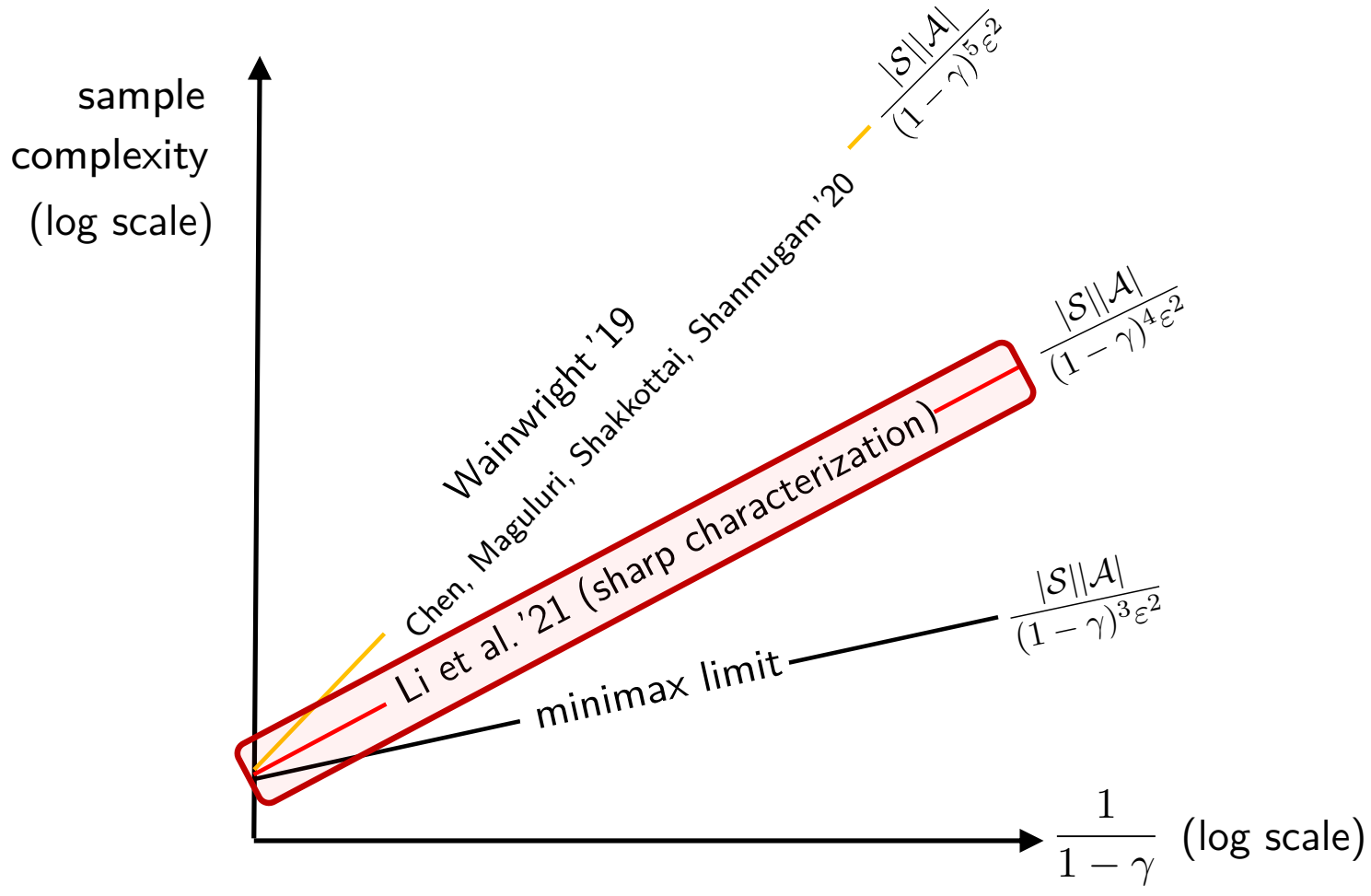
$$\begin{cases} \widetilde{O}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right) & \text{if } A \geq 2 \qquad (?) \\ \widetilde{O}\left(\frac{S}{(1-\gamma)^3\varepsilon^2}\right) & \text{if } A = 1 \qquad (\text{minimax optimal}) \end{cases}$$

other papers	sample complexity
Even-Dar & Mansour, 2003	$2^{\frac{1}{1-\gamma}} \frac{SA}{(1-\gamma)^4\varepsilon^2}$
Beck, Srikant, 2012	$\frac{S^2A^2}{(1-\gamma)^5\varepsilon^2}$
Wainwright, 2019	$\frac{SA}{(1-\gamma)^5\varepsilon^2}$
Chen, Maguluri, Shakkottai, Shanmugam, 2020	$\frac{SA}{(1-\gamma)^5\varepsilon^2}$

All this requires sample size at least $\frac{SA}{(1-\gamma)^4\varepsilon^2}$ $(A \geq 2)$ ...



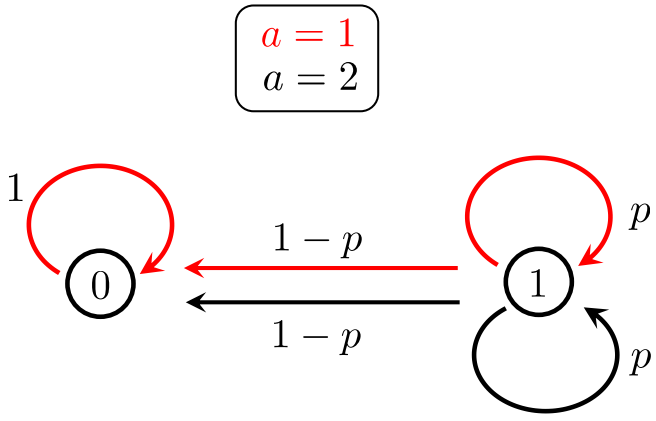
All this requires sample size at least $\frac{SA}{(1-\gamma)^4\varepsilon^2}$ $(A \geq 2)$ ...



Question: *Is Q-learning sub-optimal, or is it an analysis artifact?*

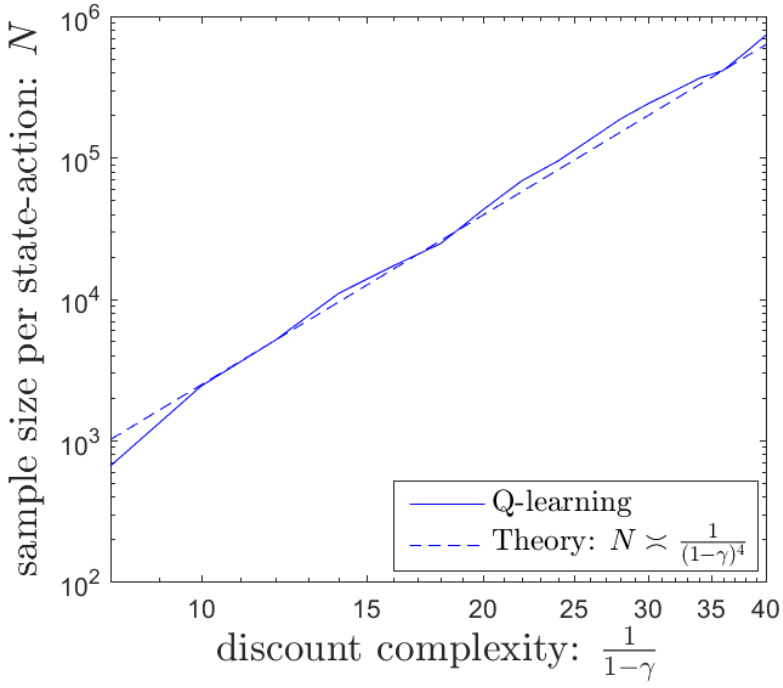
A numerical example: $\frac{SA}{(1-\gamma)^4\varepsilon^2}$ samples seem necessary . . .

— observed in Wainwright '19



$$p = \frac{4\gamma - 1}{3\gamma}$$

$$r(0,1) = 0, \quad r(1,1) = r(1,2) = 1$$



Q-learning is NOT minimax optimal

Theorem (Li, Cai, Chen, Wei, Chi '21, OR'24)

For any $0 < \varepsilon \leq 1$, there exists an MDP with $A \geq 2$ such that to achieve $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$, synchronous Q-learning needs at least

$$\widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right) \quad \textit{samples}$$

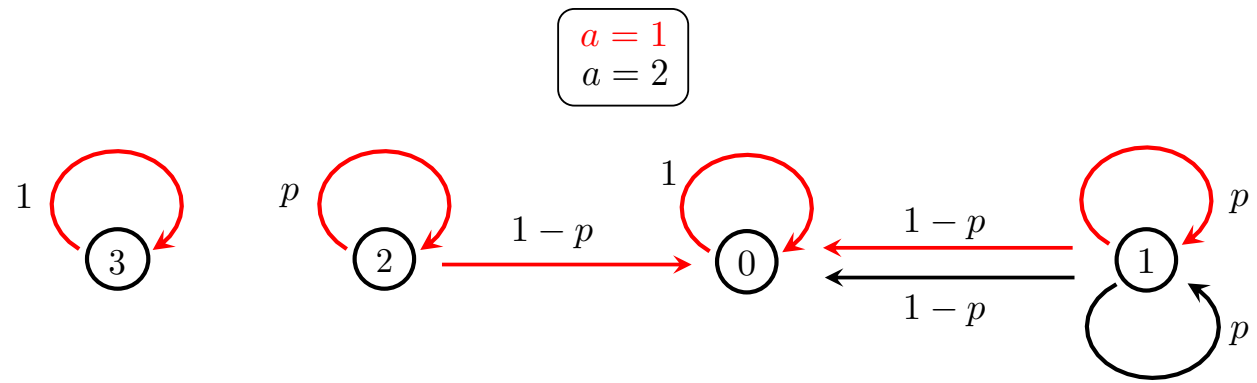
Q-learning is NOT minimax optimal

Theorem (Li, Cai, Chen, Wei, Chi '21, OR'24)

For any $0 < \varepsilon \leq 1$, there exists an MDP with $A \geq 2$ such that to achieve $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$, synchronous Q-learning needs at least

$$\widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right) \quad \textit{samples}$$

- Tight **algorithm-dependent** lower bound
- Holds for both constant and rescaled linear learning rates

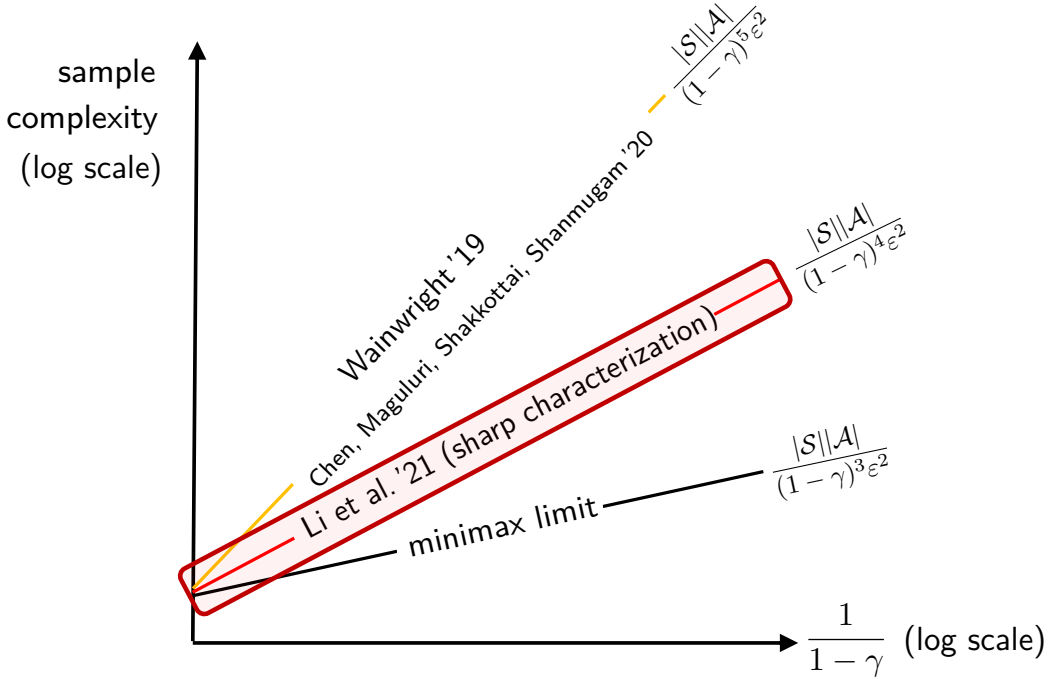


Q-learning is NOT minimax optimal

Theorem (Li, Cai, Chen, Wei, Chi '21, OR'24)

For any $0 < \varepsilon \leq 1$, there exists an MDP with $A \geq 2$ such that to achieve $\|\widehat{Q} - Q^\star\|_\infty \leq \varepsilon$, synchronous Q-learning needs at least

$$\widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^4\varepsilon^2}\right) \quad \textit{samples}$$



Why is Q-learning sub-optimal?

Over-estimation of Q-functions (Thrun & Schwartz '93; Hasselt '10)

- $\max_{a \in \mathcal{A}} \mathbb{E}[X(a)]$ tends to be over-estimated (high positive bias) when $\mathbb{E}[X(a)]$ is replaced by its empirical estimates using a small sample size
- often gets worse with a large number of actions (Hasselt, Guez, Silver '15)

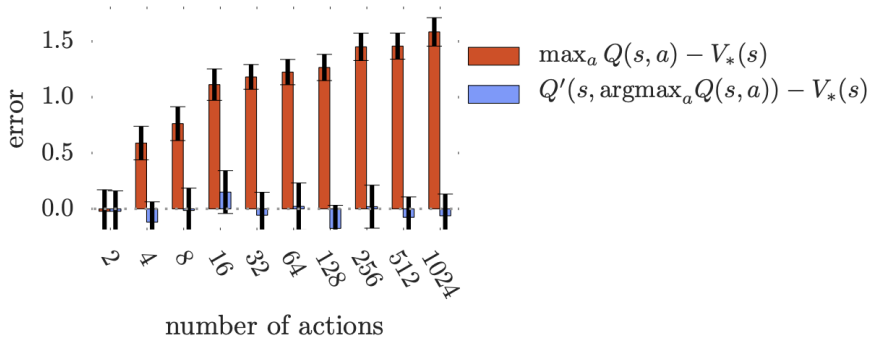


Figure 1: The orange bars show the bias in a single Q-learning update when the action values are $Q(s,a) = V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values Q', used for the blue bars, was generated identically and independently. All bars are the average of 100 repetitions.

Why is Q-learning sub-optimal?

Over-estimation of Q-functions (Thrun & Schwartz '93; Hasselt '10)

- $\max_{a \in \mathcal{A}} \mathbb{E}[X(a)]$ tends to be over-estimated (high positive bias) when $\mathbb{E}[X(a)]$ is replaced by its empirical estimates using a small sample size
- often gets worse with a large number of actions (Hasselt, Guez, Silver '15)

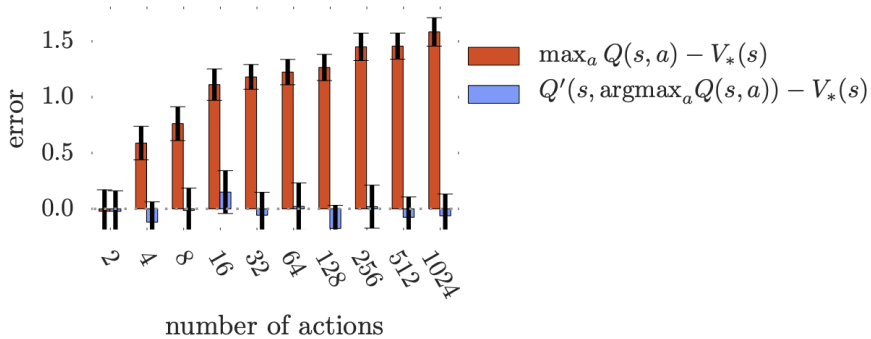


Figure 1: The orange bars show the bias in a single Q-learning update when the action values are $Q(s,a) = V_*(s) + \epsilon_a$ and the errors $\{\epsilon_a\}_{a=1}^m$ are independent standard normal random variables. The second set of action values Q', used for the blue bars, was generated identically and independently. All bars are the average of 100 repetitions.

A provable improvement: Q-learning with variance reduction

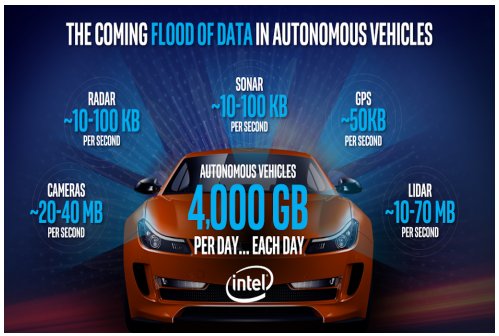
(Wainwright 2019)

Offline RL

Offline/batch RL

- Collecting new data might be costly, unsafe, unethical, or time-consuming

medical records



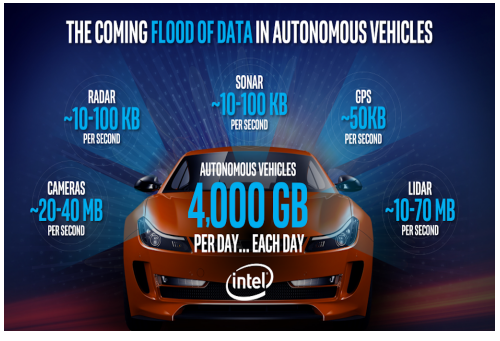
data of self-driving

clicking times of ads

Offline/batch RL

- Collecting new data might be costly, unsafe, unethical, or time-consuming
- But we have already stored tons of historical data

medical records



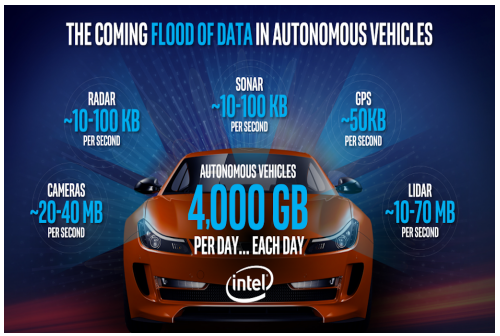
data of self-driving

clicking times of ads

Offline/batch RL

- Collecting new data might be costly, unsafe, unethical, or time-consuming
- But we have already stored tons of historical data

medical records

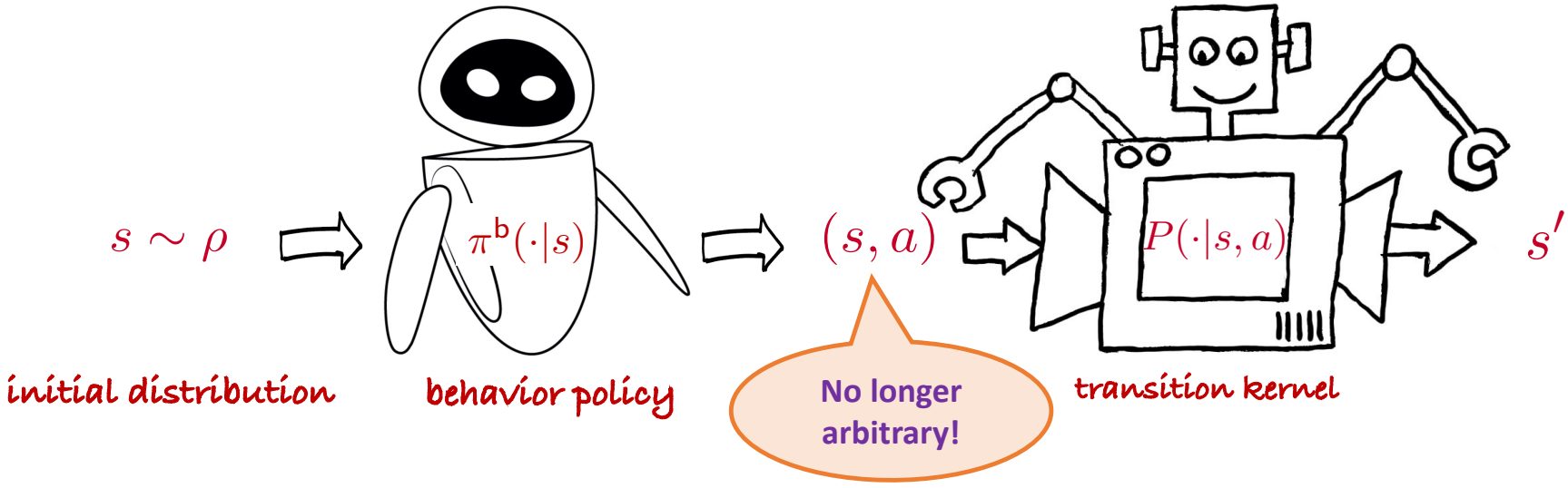


data of self-driving

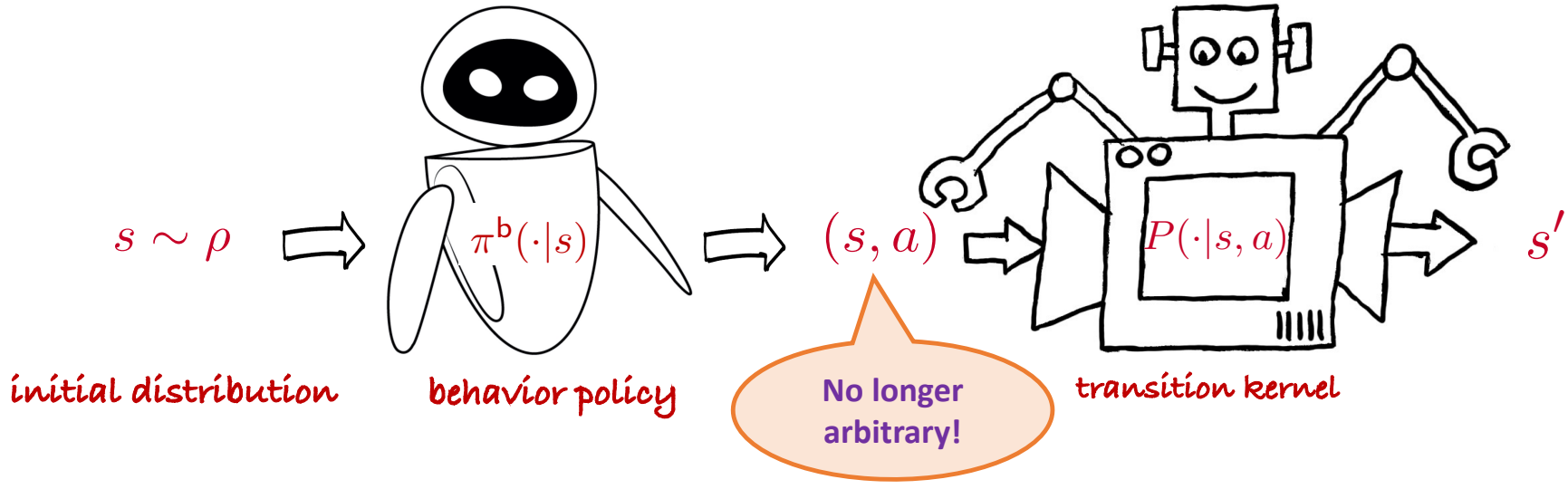
clicking times of ads

Question: can we learn based solely on historical data w/o active exploration?

A mathematical model of offline data



A mathematical model of offline data

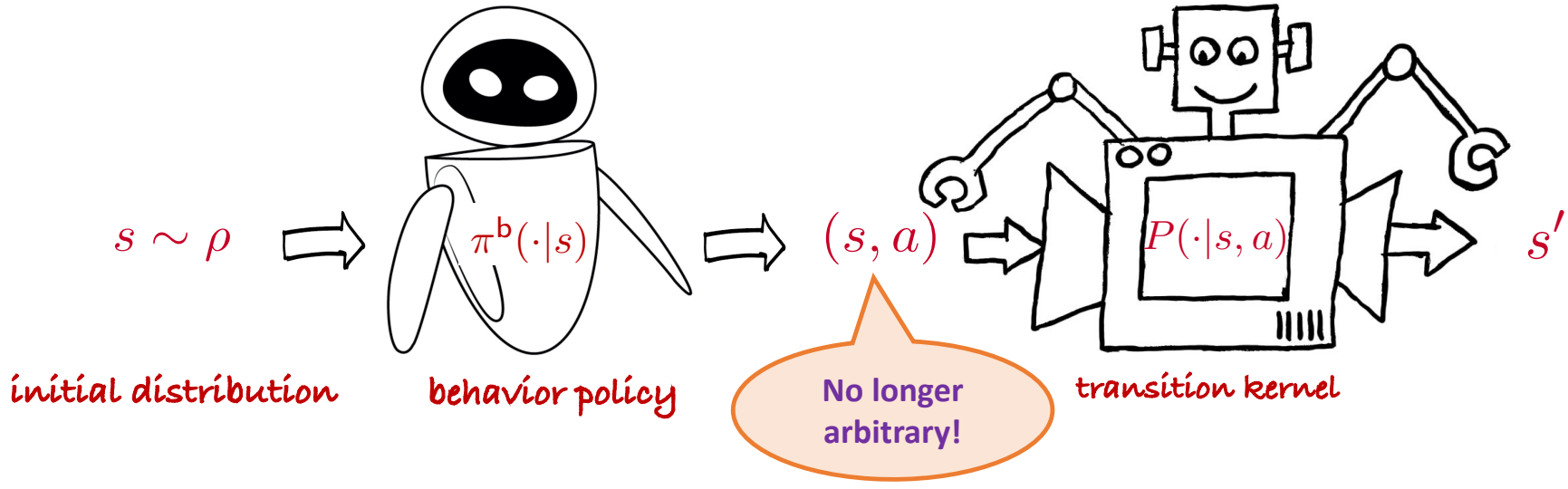


historical dataset $\mathcal{D} = \{(s^{(i)}, a^{(i)}, s'^{(i)})\}$**:** N independent copies of

$$s \sim \rho, \qquad a \sim \pi^{\mathsf{b}}(\cdot \,|\, s), \qquad s' \sim P(\cdot \,|\, s, a)$$

- ρ: initial state distribution; π^{b}: behavior policy

A mathematical model of offline data



Goal: given a target accuracy level $\varepsilon \in (0, \frac{1}{1-\gamma}]$, find $\widehat{\pi}$ s.t.

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) := \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\star}(s)\right] - \mathop{\mathbb{E}}_{s\sim\rho}\left[V^{\widehat{\pi}}(s)\right] \leq \varepsilon$$

— *in a sample-efficient manner*

Challenges of offline RL

- **Distribution shift**:

$$\text{distribution}(\mathcal{D}) \;\neq\; \text{target distribution under optimal } \pi^\star$$

Challenges of offline RL

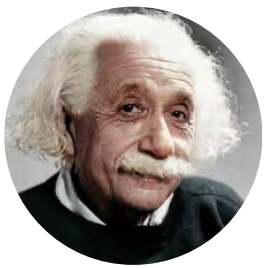
- **Distribution shift**:

$$\text{distribution}(\mathcal{D}) \neq \text{target distribution under optimal } \pi^\star$$

easier

harder

distance$(\pi^{\mathsf{b}}, \pi^\star)$



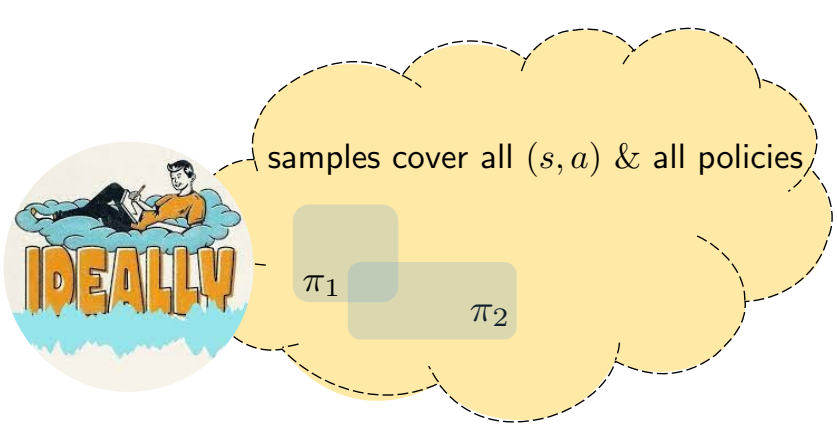
expert data

Challenges of offline RL

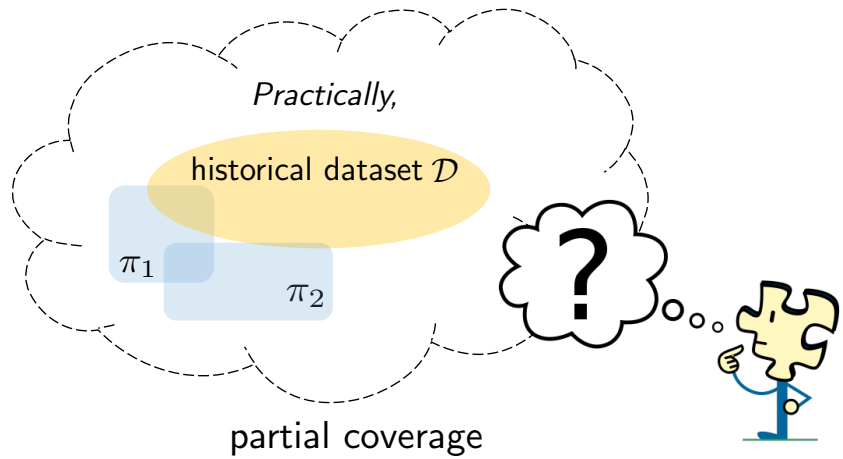
- **Distribution shift**:

$$\text{distribution}(\mathcal{D}) \;\neq\; \text{target distribution under optimal } \pi^\star$$

- **Partial coverage of state-action space**:



uniform coverage over entire space
(sufficiently explored)



partial coverage
(inadequately explored)

How to quantify quality of historical dataset $\mathcal{D}$ (induced by π^{b})?

How to quantify quality of historical dataset $\mathcal{D}$ (induced by π^{b})?

Single-policy concentrability coefficient (Rashidineiad et al. '21)

$$C^\star := \max_{s,a} \frac{d^{\pi^\star}(s,a)}{d^{\pi^{\mathsf{b}}}(s,a)} = \left\| \frac{\textit{occupancy distribution of } \pi^\star}{\textit{occupancy distribution of } \pi^{\mathsf{b}}} \right\|_\infty \geq 1$$

How to quantify quality of historical dataset $\mathcal{D}$ (induced by π^{b})?

Single-policy concentrability coefficient (Rashidineiad et al. '21)

$$C^\star := \max_{s,a} \frac{d^{\pi^\star}(s,a)}{d^{\pi^{\mathsf{b}}}(s,a)} = \left\| \frac{\textit{occupancy distribution of } \pi^\star}{\textit{occupancy distribution of } \pi^{\mathsf{b}}} \right\|_\infty \geq 1$$

- captures distributional shift

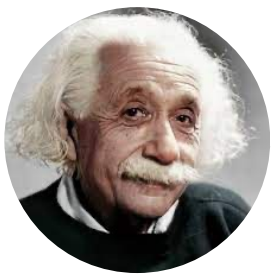
How to quantify quality of historical dataset $\mathcal{D}$ (induced by π^{b})?

Single-policy concentrability coefficient (Rashidineiad et al. '21)

$$C^\star := \max_{s,a} \frac{d^{\pi^\star}(s,a)}{d^{\pi^{\mathsf{b}}}(s,a)} = \left\| \frac{\textit{occupancy distribution of } \pi^\star}{\textit{occupancy distribution of } \pi^{\mathsf{b}}} \right\|_\infty \geq 1$$

- captures distributional shift

$C^\star = O(1)$ → large $C^\star$



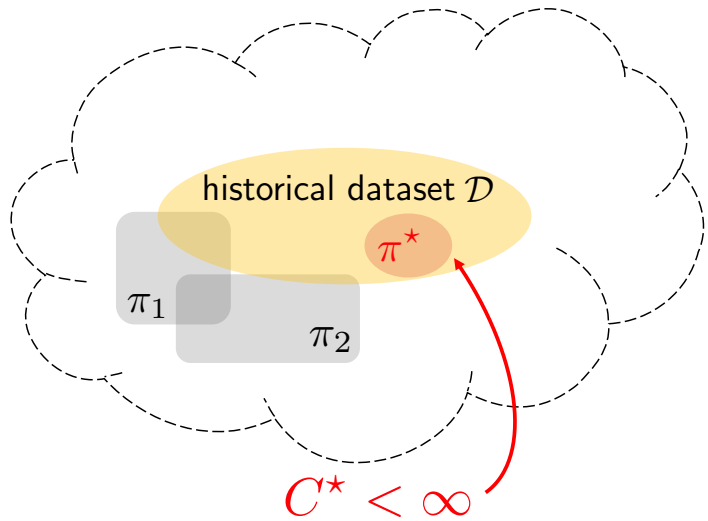
expert data

How to quantify quality of historical dataset $\mathcal{D}$ (induced by π^{b})?

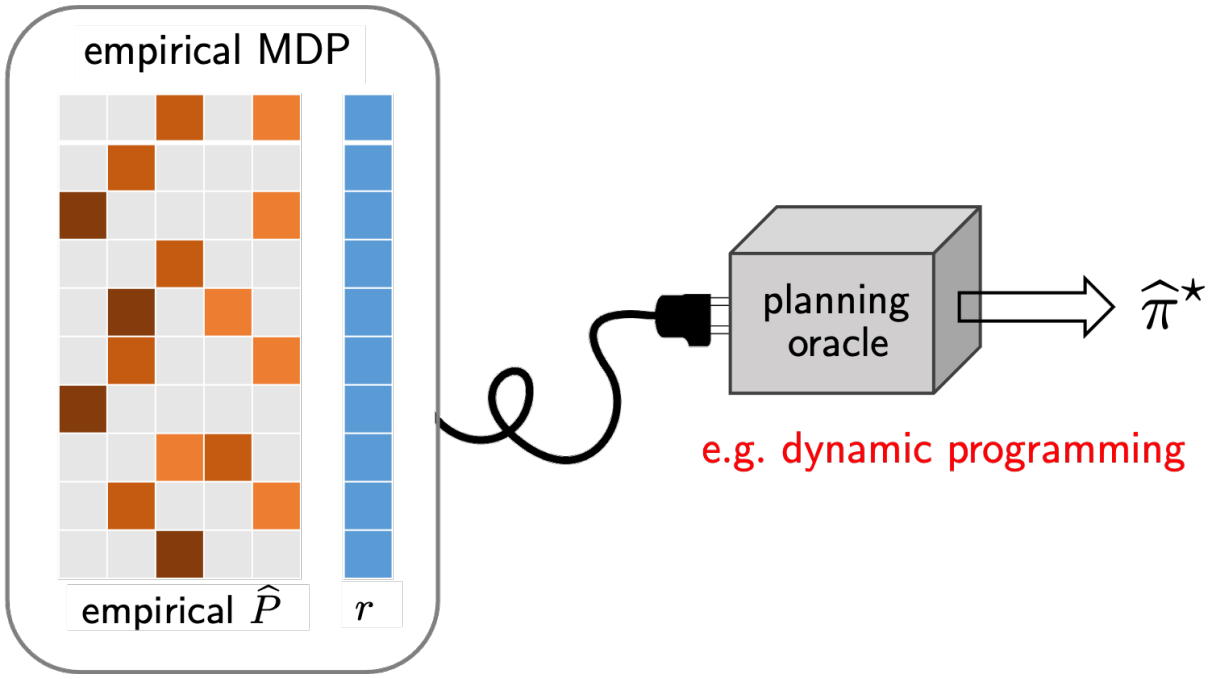
Single-policy concentrability coefficient (Rashidineiad et al. '21)

$$C^\star := \max_{s,a} \frac{d^{\pi^\star}(s,a)}{d^{\pi^{\mathsf{b}}}(s,a)} = \left\| \frac{\textit{occupancy distribution of } \pi^\star}{\textit{occupancy distribution of } \pi^{\mathsf{b}}} \right\|_\infty \geq 1$$

- captures distributional shift
- allows for partial coverage
  - as long as it covers the part reachable by $\pi^\star$

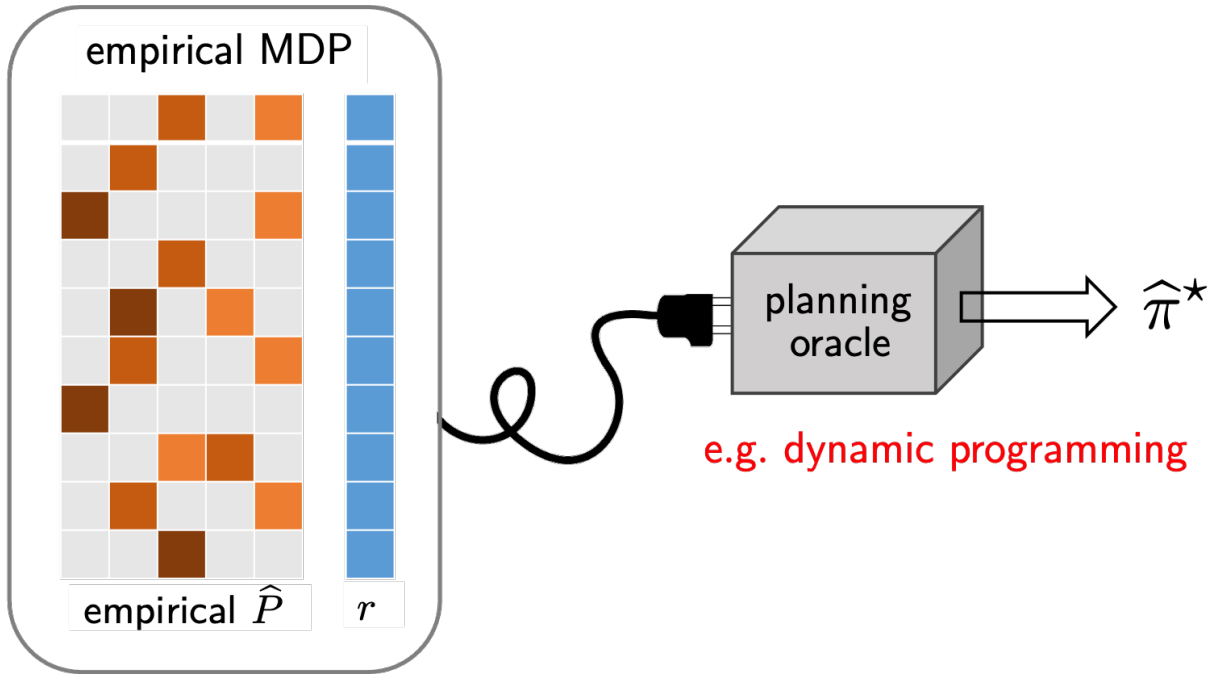


Model-based ("plug-in") approach?



1. construct empirical model $\widehat{P}$:

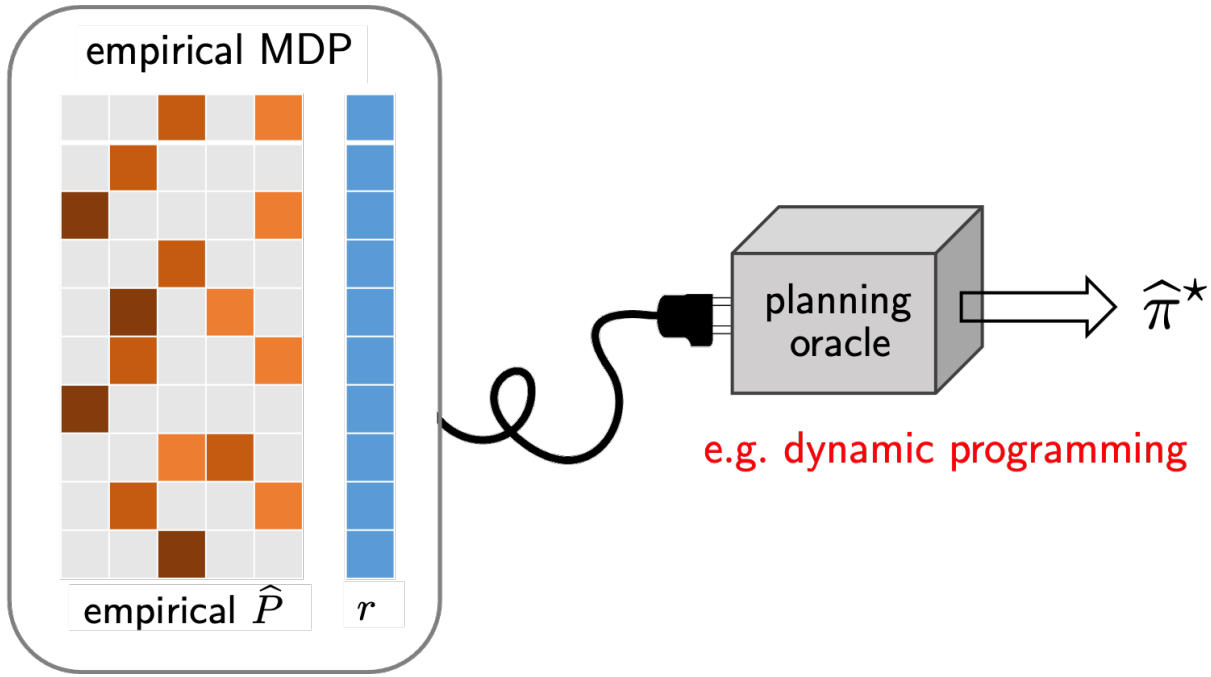
Model-based ("plug-in") approach?



1. construct empirical model $\widehat{P}$:

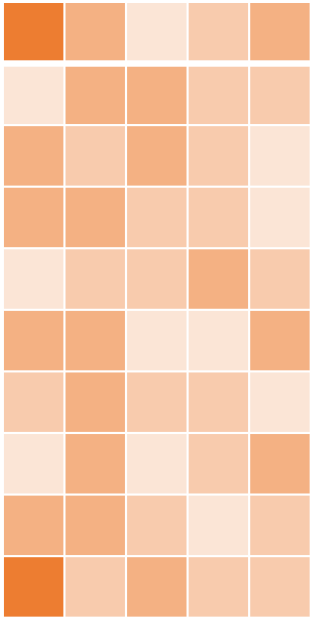
$$\widehat{P}(s' \mid s, a) = \underbrace{\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{s'^{(i)} = s'\}}_{\text{empirical frequency}}$$

Model-based ("plug-in") approach?

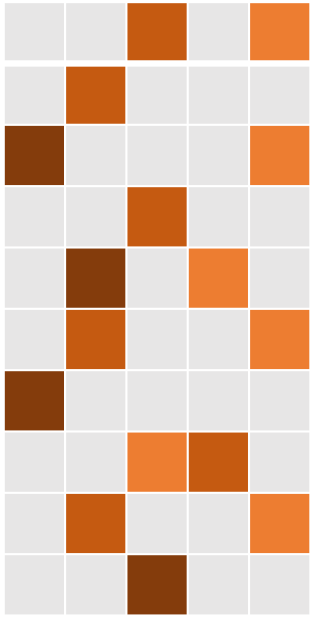


1. construct empirical model $\widehat{P}$:
2. planning (e.g. value iteration) based on empirical MDP

Issues & challenges in the sample-starved regime



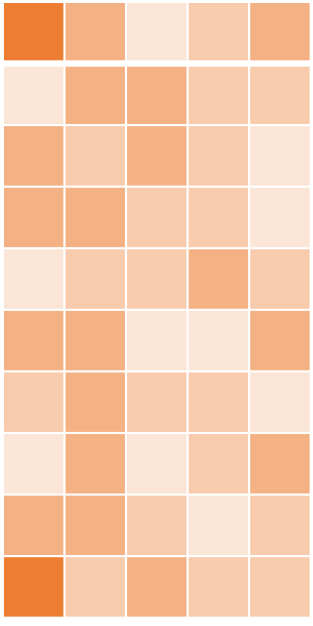
truth: $P \in \mathbb{R}^{SA \times S}$



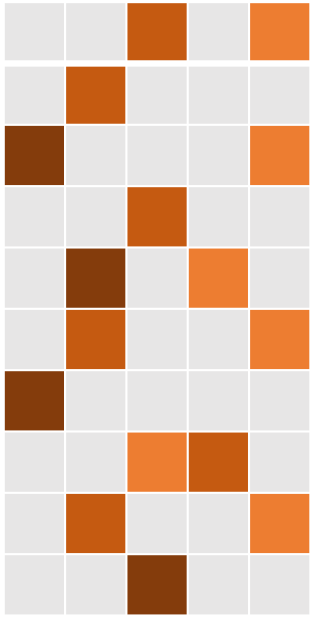
empirical $\widehat{P}$ (simulator)

- can't recover P faithfully if sample size $\ll S^2 A$

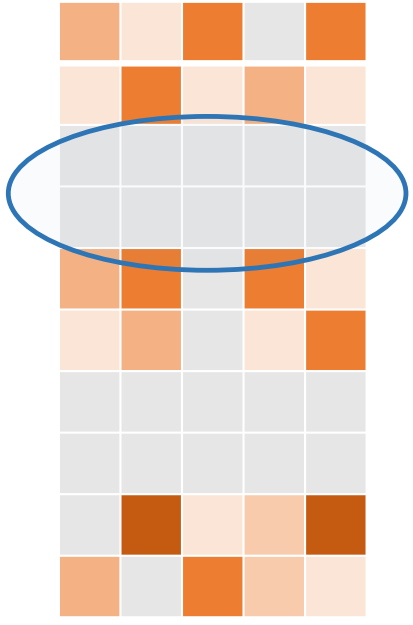
Issues & challenges in the sample-starved regime



truth: $P \in \mathbb{R}^{SA \times S}$



empirical $\widehat{P}$ (simulator)



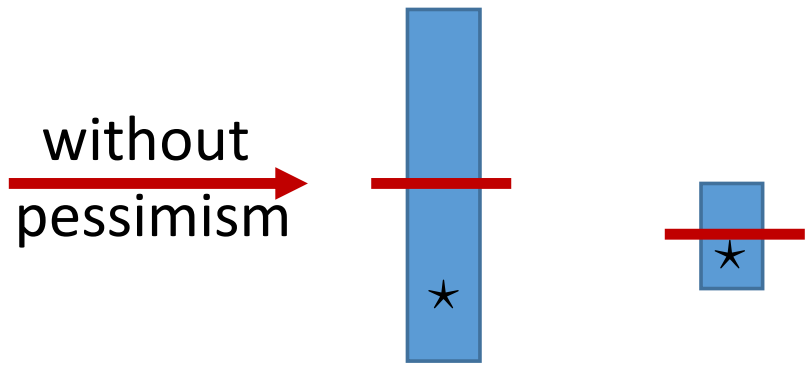
empirical $\widehat{P}$ (offline)

- can't recover P faithfully if sample size $\ll S^2A$
- (possibly) insufficient coverage under offline data

Pessimism in the face of uncertainty

Penalize value estimate of (s, a) pairs that were poorly visited

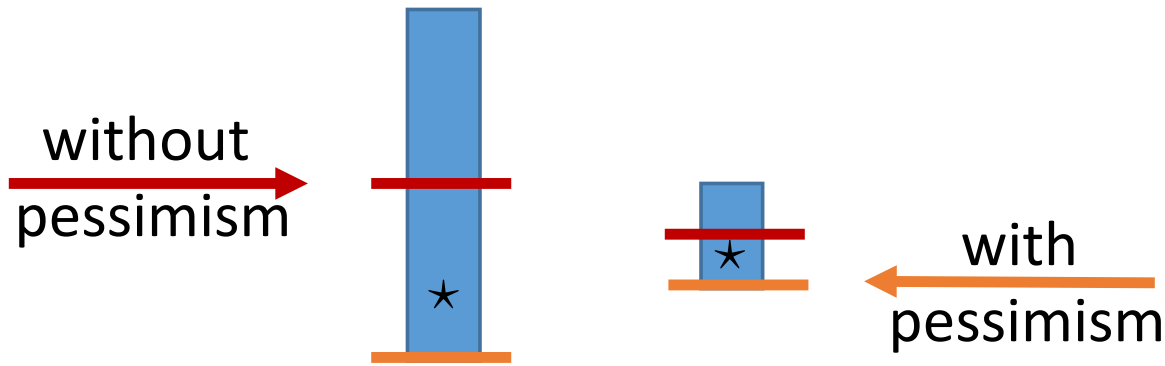
— (Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21)



Pessimism in the face of uncertainty

Penalize value estimate of (s, a) pairs that were poorly visited

— (Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21)



Value iteration with lower confidence bound (VI-LCB):

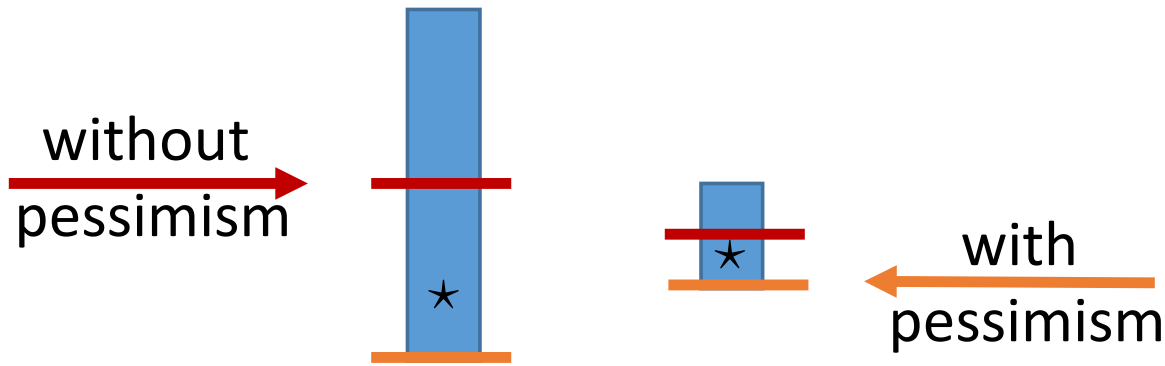
$$\widehat{Q}(s,a) \;\leftarrow\; \max\Big\{r(s,a) + \gamma\big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}\big\rangle$$

where $\widehat{V}(s) = \max_a \widehat{Q}(s,a)$.

Pessimism in the face of uncertainty

Penalize value estimate of (s, a) pairs that were poorly visited

— (Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21)



Value iteration with lower confidence bound (VI-LCB):

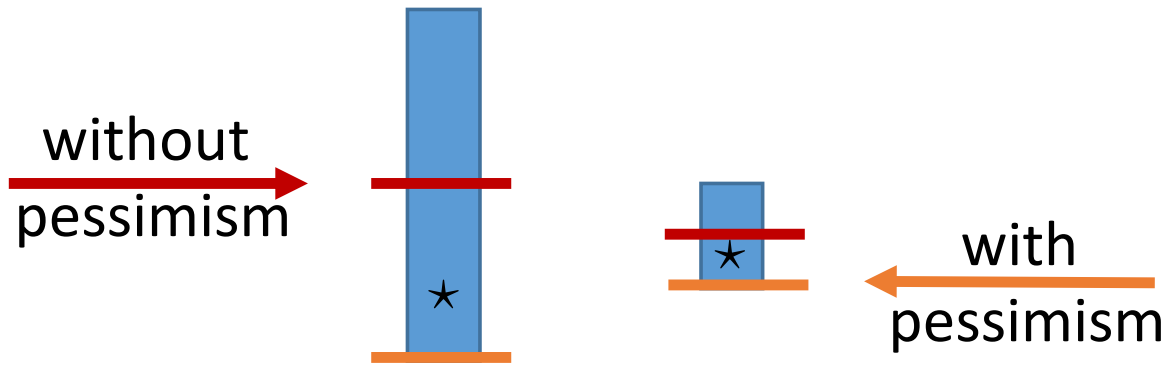
$$\widehat{Q}(s,a) \;\leftarrow\; \max\Big\{r(s,a) + \gamma\big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}\big\rangle,\; 0\Big\}$$

where $\widehat{V}(s) = \max_a \widehat{Q}(s,a)$.

Pessimism in the face of uncertainty

Penalize value estimate of (s, a) pairs that were poorly visited

— (Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21)



Value iteration with lower confidence bound (VI-LCB):

$$\widehat{Q}(s,a) \;\leftarrow\; \max\Big\{r(s,a) + \gamma\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}\rangle - \underbrace{b(s,a;\widehat{V})}_{\text{uncertainty penalty}},\; 0\Big\}$$

where $\widehat{V}(s) = \max_a \widehat{Q}(s,a)$.

Sample complexity of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '24)

For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *the policy* $\widehat{\pi}$ *returned by VI-LCB using a Bernstein-style penalty term achieves*

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{SC^{\star}}{(1-\gamma)^3\varepsilon^2}\right)$$

Sample complexity of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '24)

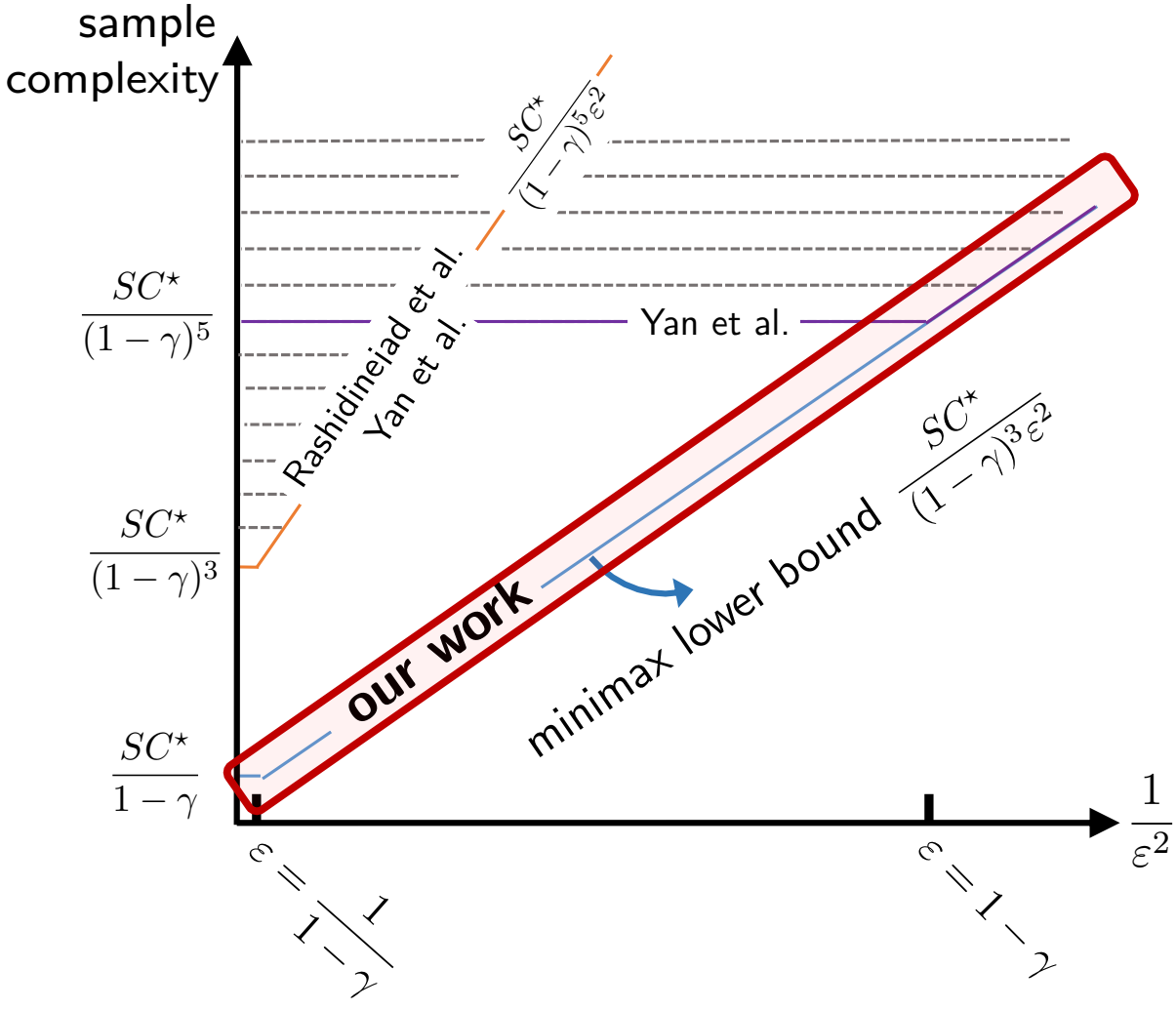
For any $0 < \varepsilon \leq \frac{1}{1-\gamma}$, *the policy* $\widehat{\pi}$ *returned by VI-LCB using a Bernstein-style penalty term achieves*

$$V^{\star}(\rho) - V^{\widehat{\pi}}(\rho) \leq \varepsilon$$

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{SC^{\star}}{(1-\gamma)^3\varepsilon^2}\right)$$

- depends on distribution shift (as reflected by $C^{\star}$)
- achieves minimax optimality
- full ε-range (no burn-in cost)



Model-based offline RL is minimax optimal with no burn-in cost!

Robust RL

Safety and robustness in RL

(Zhou et al., 2021; Panaganti and Kalathil, 2022; Yang et al., 2022;)

Training environment $\neq$

Test environment

Safety and robustness in RL

(Zhou et al., 2021; Panaganti and Kalathil, 2022; Yang et al., 2022;)

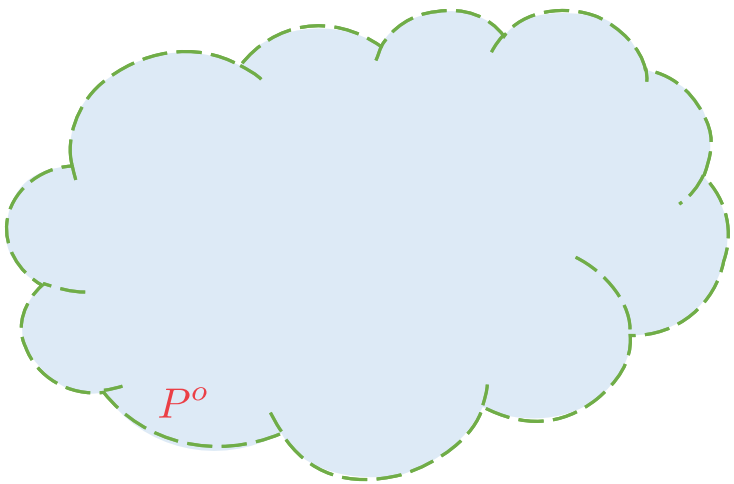
Training environment $\neq$ Test environment

Sim2Real Gap: Can we learn optimal policies that are robust to model perturbations?

Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

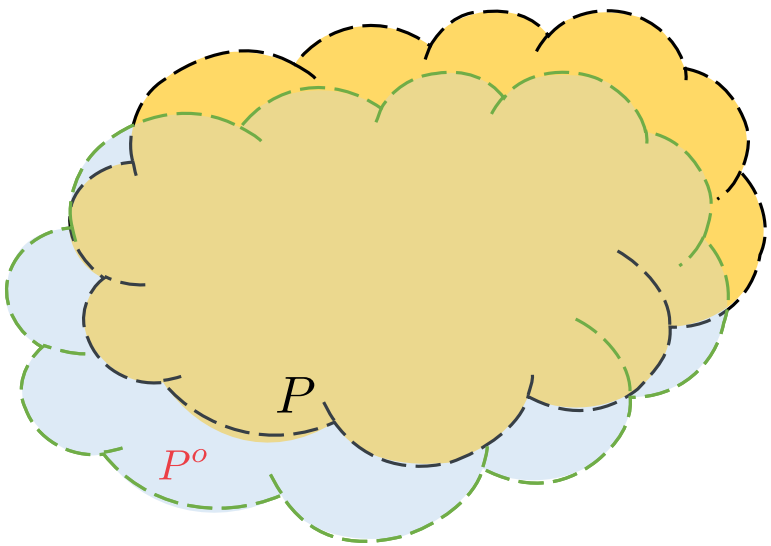
$$\mathcal{U}^\sigma(P^o) = \{P : \ \rho(P, P^o) \leq \sigma\}$$

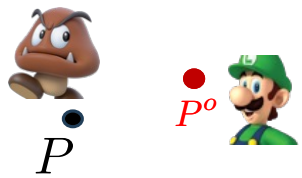


Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$

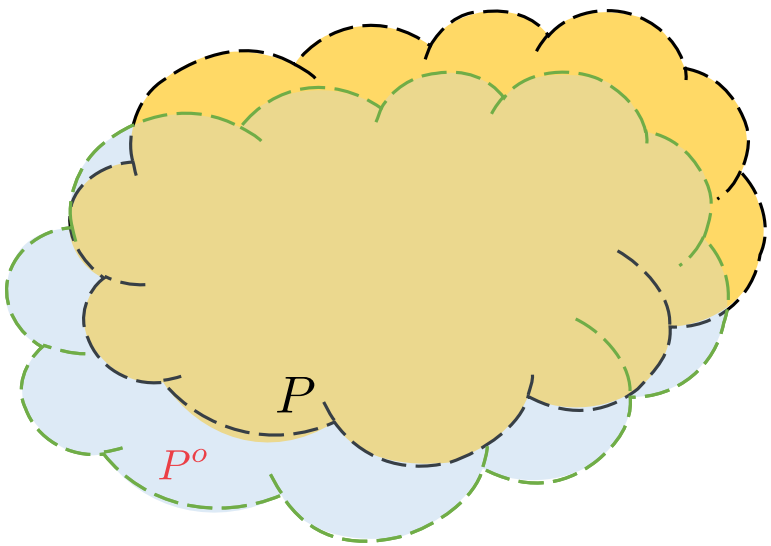


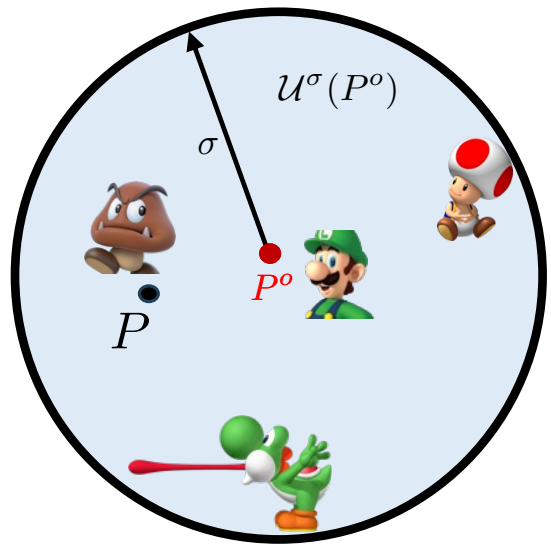


Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$

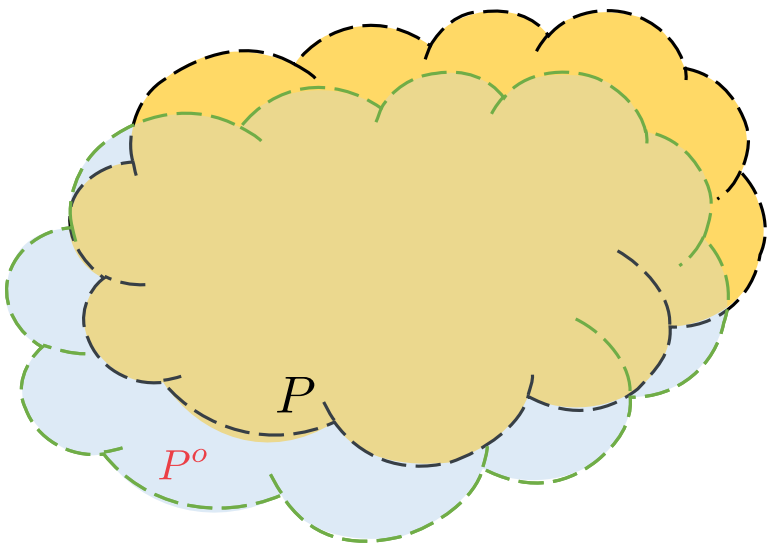


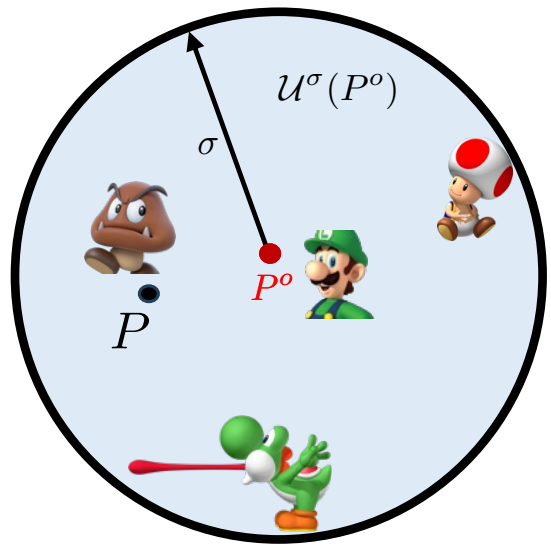


Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

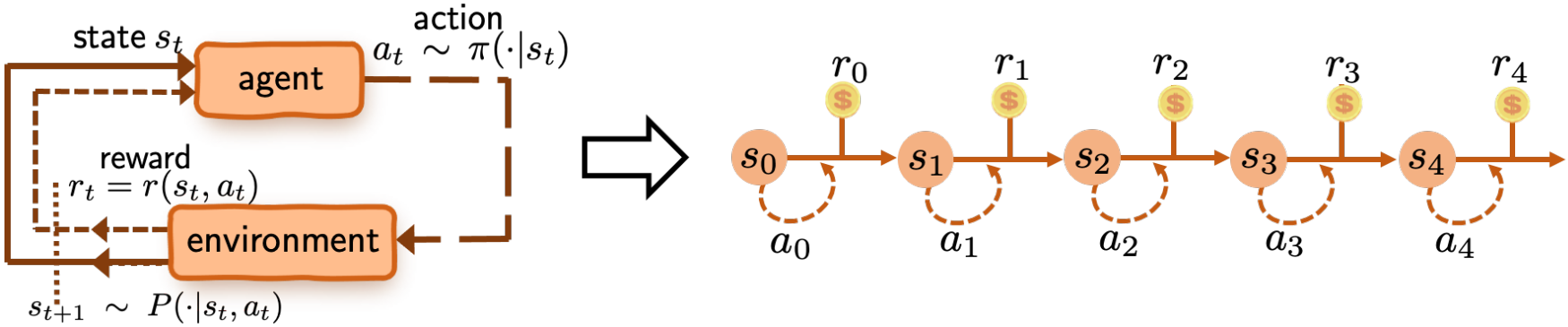
$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$





- Examples of ρ: f-divergence (TV, χ^2, KL...)

Robust value/Q function



Robust value/Q function of policy π:

$$\forall s \in \mathcal{S}: \qquad V^{\pi,\sigma}(s) := \inf_{P \in \mathcal{U}^\sigma(P^o)} \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s\right]$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \qquad Q^{\pi,\sigma}(s,a) := \inf_{P \in \mathcal{U}^\sigma(P^o)} \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_0 = a\right]$$

Measures the worst-case performance of the policy in the uncertainty set.

Distributionally robust MDP

Robust MDP

Find the policy $\pi^\star$ that maximizes $V^{\pi,\sigma}$

(Iyengar. '05, Nilim and El Ghaoui. '05)

Distributionally robust MDP

Robust MDP

Find the policy $\pi^\star$ that maximizes $V^{\pi,\sigma}$

(Iyengar. '05, Nilim and El Ghaoui. '05)

Robust Bellman's optimality equation: the optimal robust policy $\pi^\star$ and optimal robust value $V^{\star,\sigma} := V^{\pi^\star,\sigma}$ satisfy

$$Q^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma\left(P^o_{s,a}\right)} \left\langle P_{s,a}, V^{\star,\sigma} \right\rangle,$$

$$V^{\star,\sigma}(s) = \max_a Q^{\star,\sigma}(s,a)$$

Distributionally robust MDP

Robust MDP

Find the policy $\pi^\star$ that maximizes $V^{\pi,\sigma}$

(Iyengar. '05, Nilim and El Ghaoui. '05)

Robust Bellman's optimality equation: the optimal robust policy $\pi^\star$ and optimal robust value $V^{\star,\sigma} := V^{\pi^\star,\sigma}$ satisfy

$$Q^{\star,\sigma}(s,a) = r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma\left(P^o_{s,a}\right)} \langle P_{s,a}, V^{\star,\sigma} \rangle \,,$$

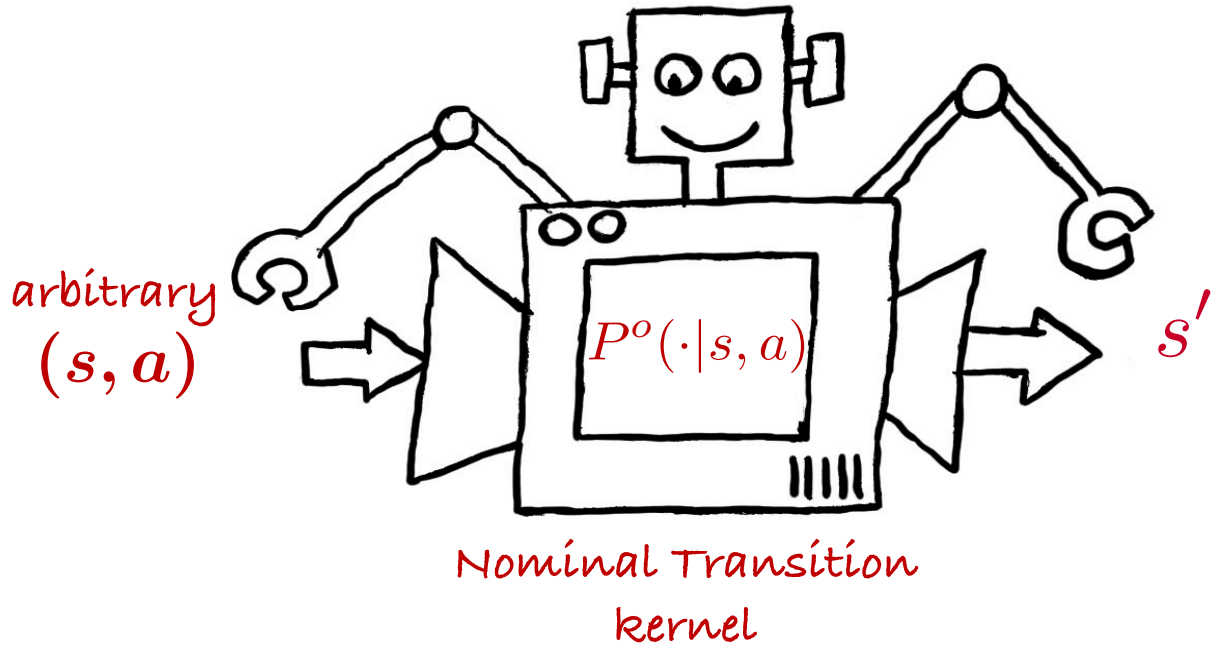
$$V^{\star,\sigma}(s) = \max_a \, Q^{\star,\sigma}(s,a)$$

Distributionally robust value iteration (DRVI):

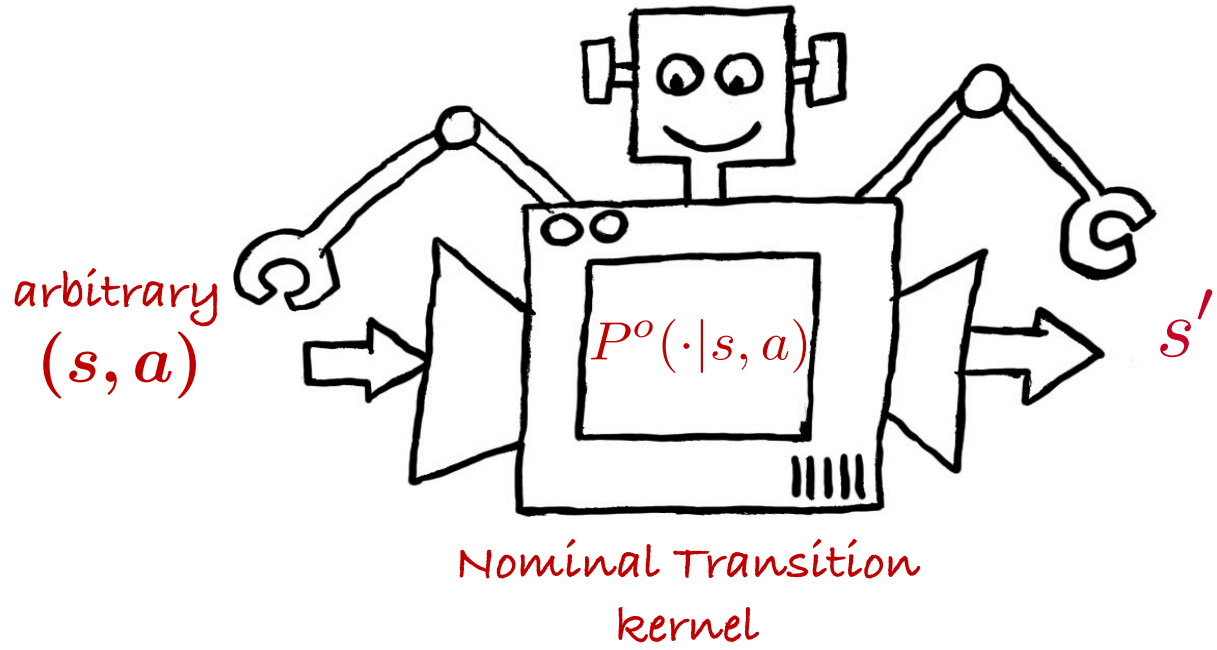
$$Q(s,a) \leftarrow r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma\left(P^o_{s,a}\right)} \langle P_{s,a}, V \rangle,$$

where $V(s) = \max_a \, Q(s,a)$.

Learning distributionally robust MDPs



Learning distributionally robust MDPs

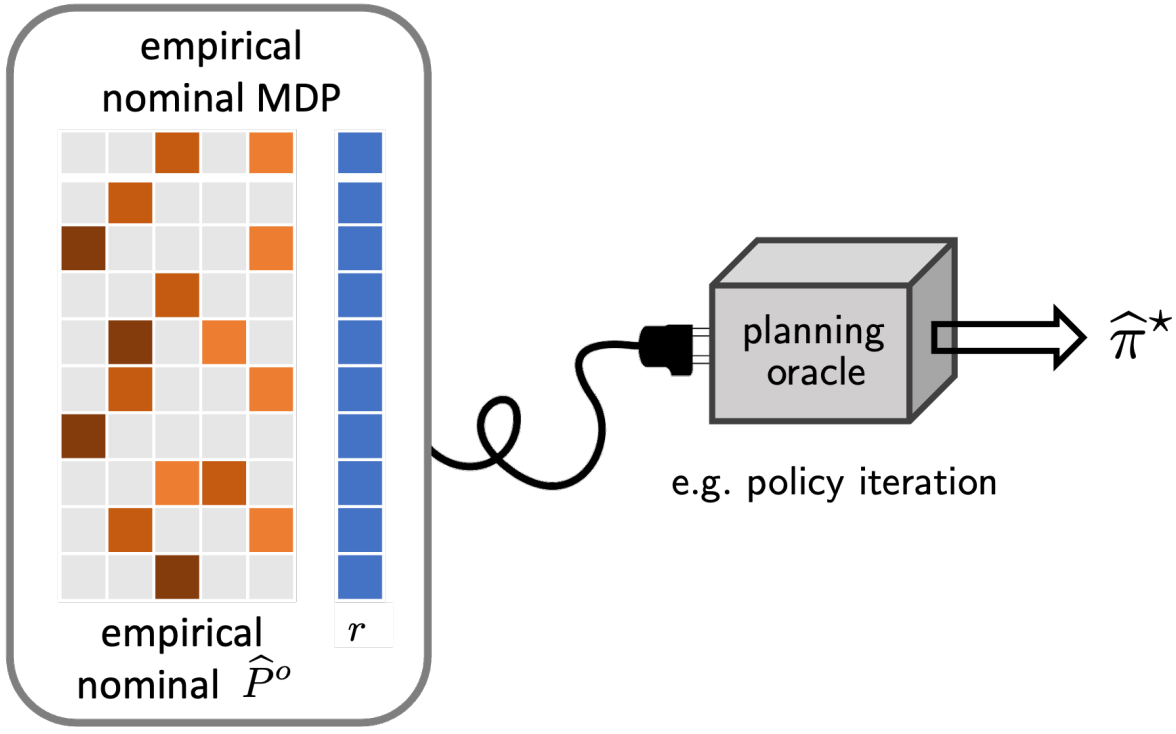


Goal of robust RL: given $\mathcal{D} := \{(s_i, a_i, s_i')\}_{i=1}^N$ from the *nominal* environment P^0, find an ε-optimal robust policy $\widehat{\pi}$ obeying

$$V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \leq \varepsilon$$

— *in a sample-efficient manner*

Model-based RL: empirical MDP + planning



Planning by **distributionally robust value iteration (DRVI)**:

$$\widehat{Q}(s,a) \leftarrow r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma\left(\widehat{P}^o_{s,a}\right)} \langle P_{s,a}, \widehat{V} \rangle,$$

where $\widehat{V}(s) = \max_a \widehat{Q}(s,a)$.

Duality for scalability

Dual problem can be solved efficiently (w.r.t. a scalar)

(Iyengar. '05, Nilim and El Ghaoui. '05)

TV uncertainty: divergence function $\rho =$ total variation

$$\widehat{Q}(s,a) \leftarrow r(s,a) + \gamma \max_{\lambda \in \left[\min_s \widehat{V}(s), \max_s \widehat{V}(s)\right]} \left\{ \widehat{P}^o_{s,a}\left[\widehat{V}\right]_\lambda - \sigma\left(\lambda - \min_{s'} \left[\widehat{V}\right]_\lambda (s')\right)\right\},$$

where $[\widehat{V}]_\lambda(s) := \lambda$ if $\widehat{V}(s) > \lambda$, otherwise $[\widehat{V}]_\lambda(s) = \widehat{V}(s)$.

Duality for scalability

Dual problem can be solved efficiently (w.r.t. a scalar)

(Iyengar. '05, Nilim and El Ghaoui. '05)

TV uncertainty: divergence function $\rho =$ total variation

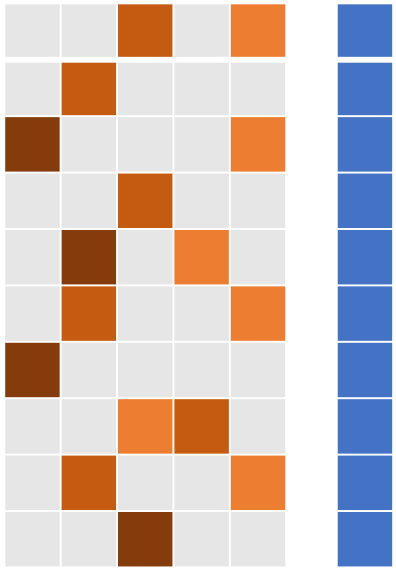
$$\widehat{Q}(s,a) \leftarrow r(s,a) + \gamma \max_{\lambda \in \left[\min_s \widehat{V}(s), \max_s \widehat{V}(s)\right]} \left\{ \widehat{P}^o_{s,a}\left[\widehat{V}\right]_\lambda - \sigma\left(\lambda - \min_{s'} \left[\widehat{V}\right]_\lambda (s')\right)\right\},$$

where $[\widehat{V}]_\lambda(s) := \lambda$ if $\widehat{V}(s) > \lambda$, otherwise $[\widehat{V}]_\lambda(s) = \widehat{V}(s)$.

χ^2 uncertainty: divergence function $\rho = \chi^2$

$$\widehat{Q}(s,a) \leftarrow r(s,a) + \gamma \max_{\lambda \in \left[\min_s \widehat{V}(s), \max_s \widehat{V}(s)\right]} \left\{ \widehat{P}^o_{s,a}\left[\widehat{V}\right]_\lambda - \sqrt{\lambda \mathsf{Var}_{\widehat{P}^o_{s,a}}\left(\left[\widehat{V}\right]_\lambda\right)}\right\}.$$

A curious question

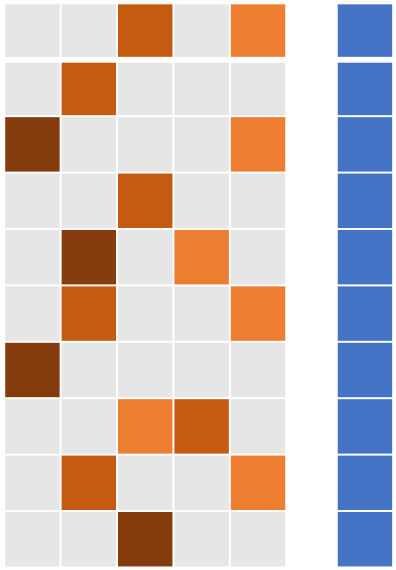


empirical MDP

Learn the optimal policy of the nominal MDP?

Learn the **robust** policy around the nominal MDP?

A curious question



empirical MDP

Learn the optimal policy of the nominal MDP?

Learn the **robust** policy around the nominal MDP?

Robustness-statistical trade-off? Is there a statistical premium that one needs to pay in quest of additional robustness?

Sample complexity under TV uncertainty

Theorem (Shi et al., 2023)

Assume the uncertainty set is measured via the TV distance with radius $\sigma \in [0, 1)$. *For sufficiently small* $\varepsilon > 0$, *DRVI outputs a policy* $\widehat{\pi}$ *that satisfies* $V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \leq \varepsilon$ *with sample complexity at most*

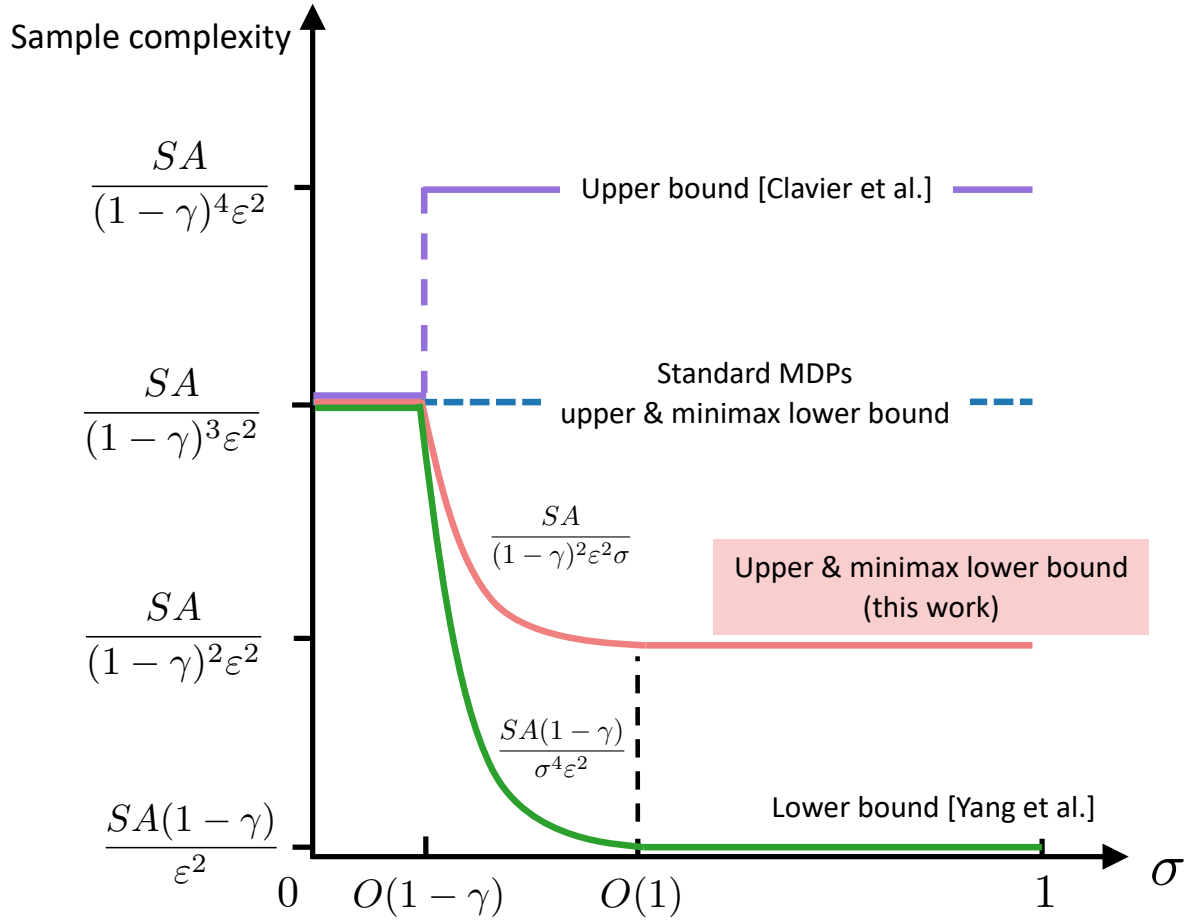
$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\}\varepsilon^2}\right)$$

ignoring logarithmic factors. In addition, no algorithm can succeed if the sample size is below

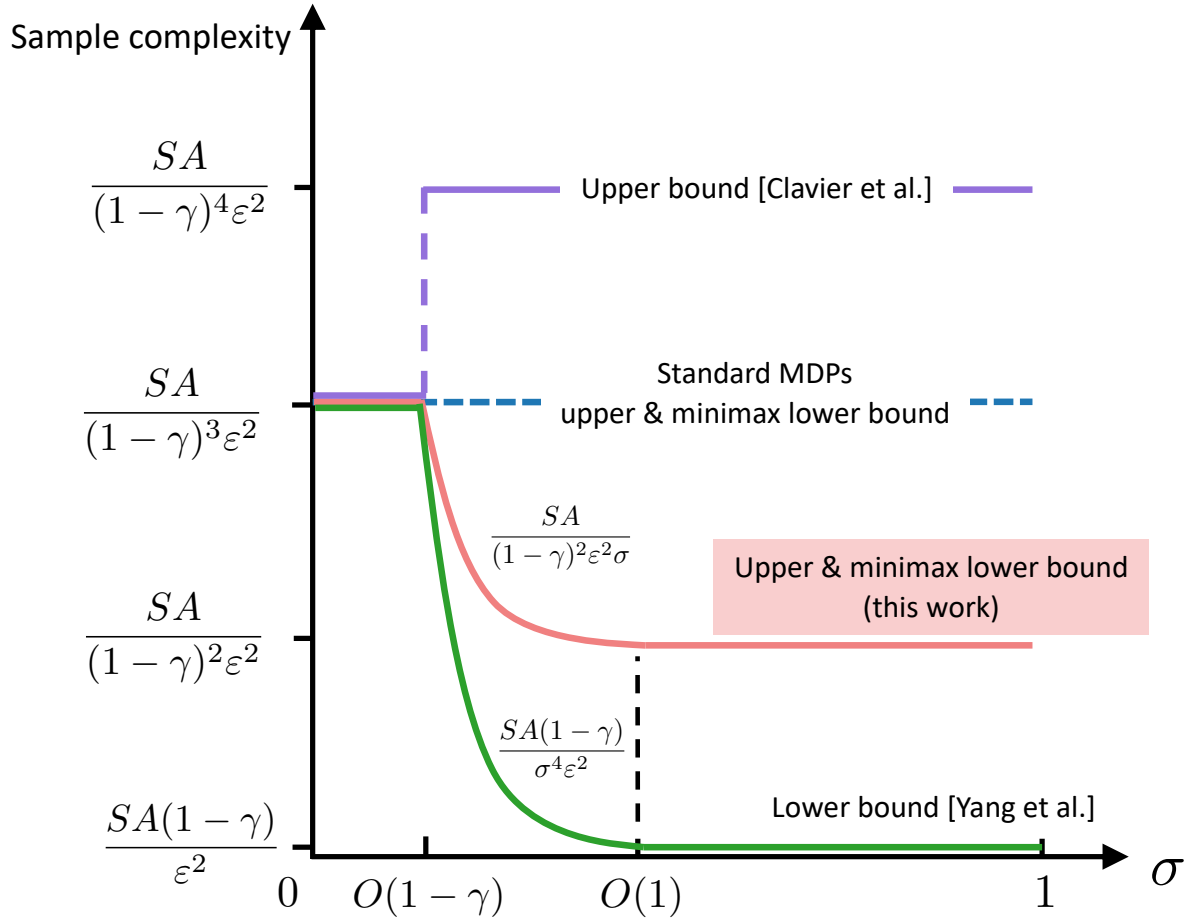
$$\widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\}\varepsilon^2}\right).$$

- Establish the minimax optimality of DRVI for RMDP under the TV uncertainty set over the full range of σ.

When the uncertainty set is TV



When the uncertainty set is TV



RMDPs are **easier** to learn than standard MDPs.

Sample complexity theorem under χ^2 uncertainty

Theorem (Shi et al., 2025)

Assume the uncertainty set is measured via the χ^2 divergence with radius $\sigma \in [0, \infty)$. For sufficiently small $\varepsilon > 0$, DRVI outputs a policy $\widehat{\pi}$ that satisfies $V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \leq \varepsilon$ with sample complexity at most

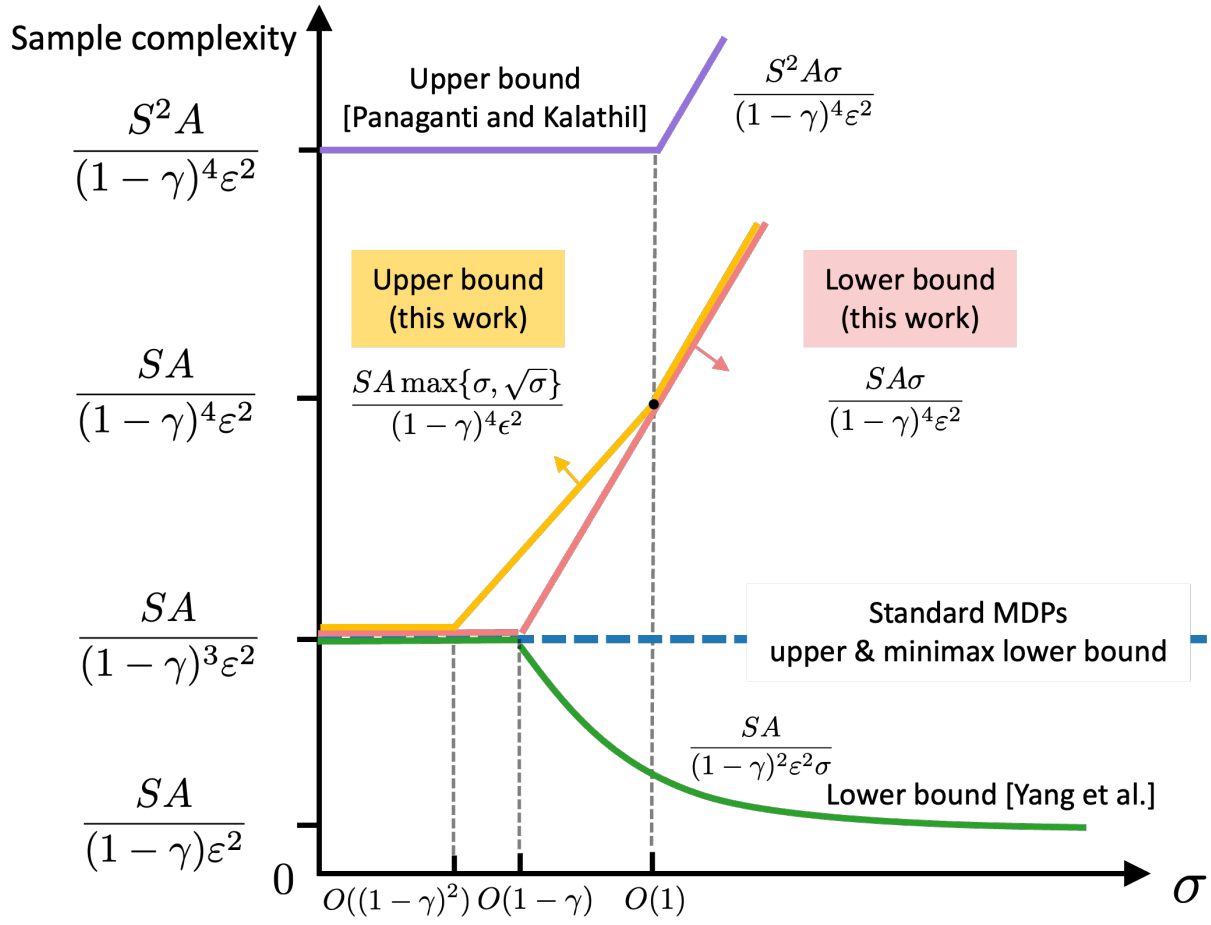
$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\left(1+\frac{\sqrt{\sigma}+\sigma}{1-\gamma}\right)\right)$$

ignoring logarithmic factors. In addition, no algorithm can succeed if the sample size is below

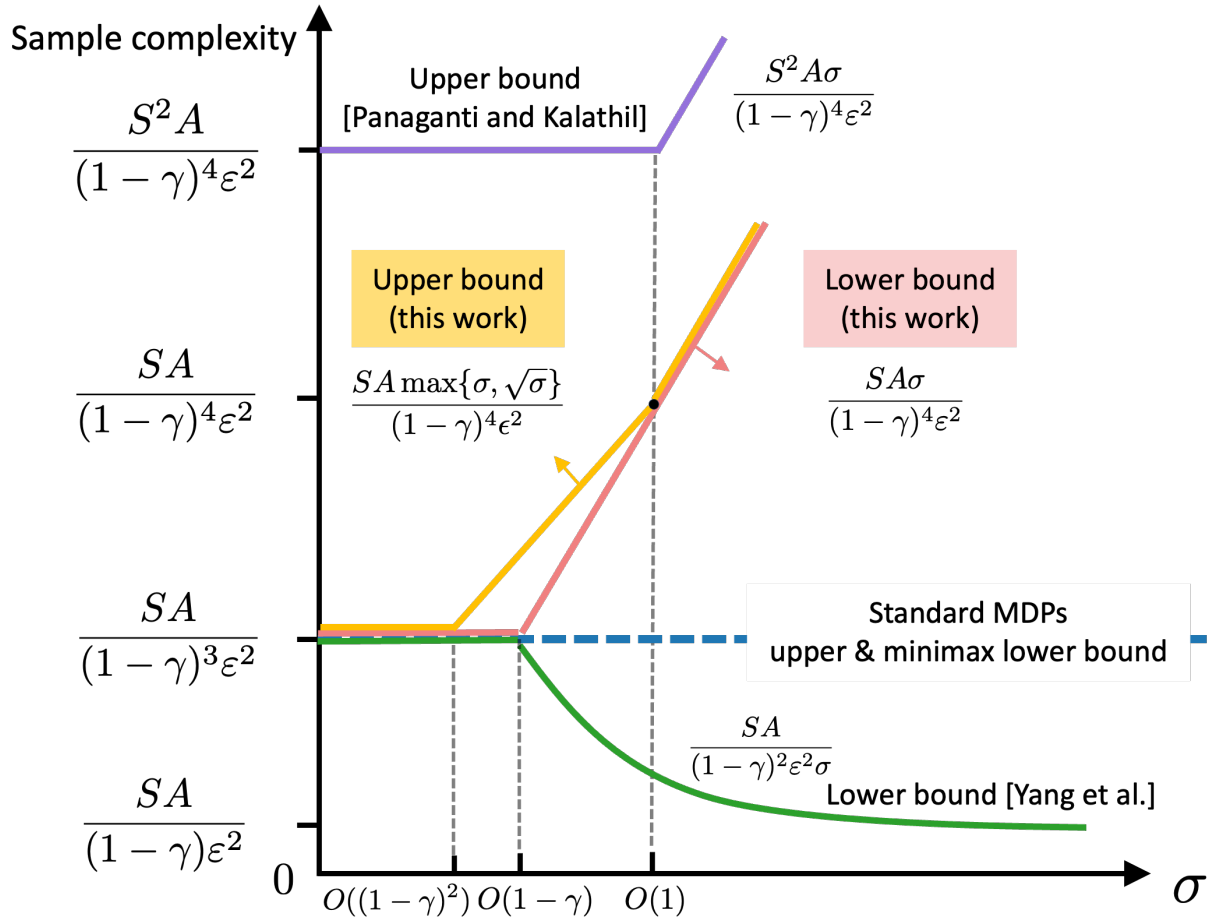
$$\widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\left(1+\frac{\sigma}{1-\gamma}\right)\right)$$

- Establish the minimax optimality of DRVI for RMDP under the TV uncertainty set over the full range of σ.

When the uncertainty set is χ^2 divergence



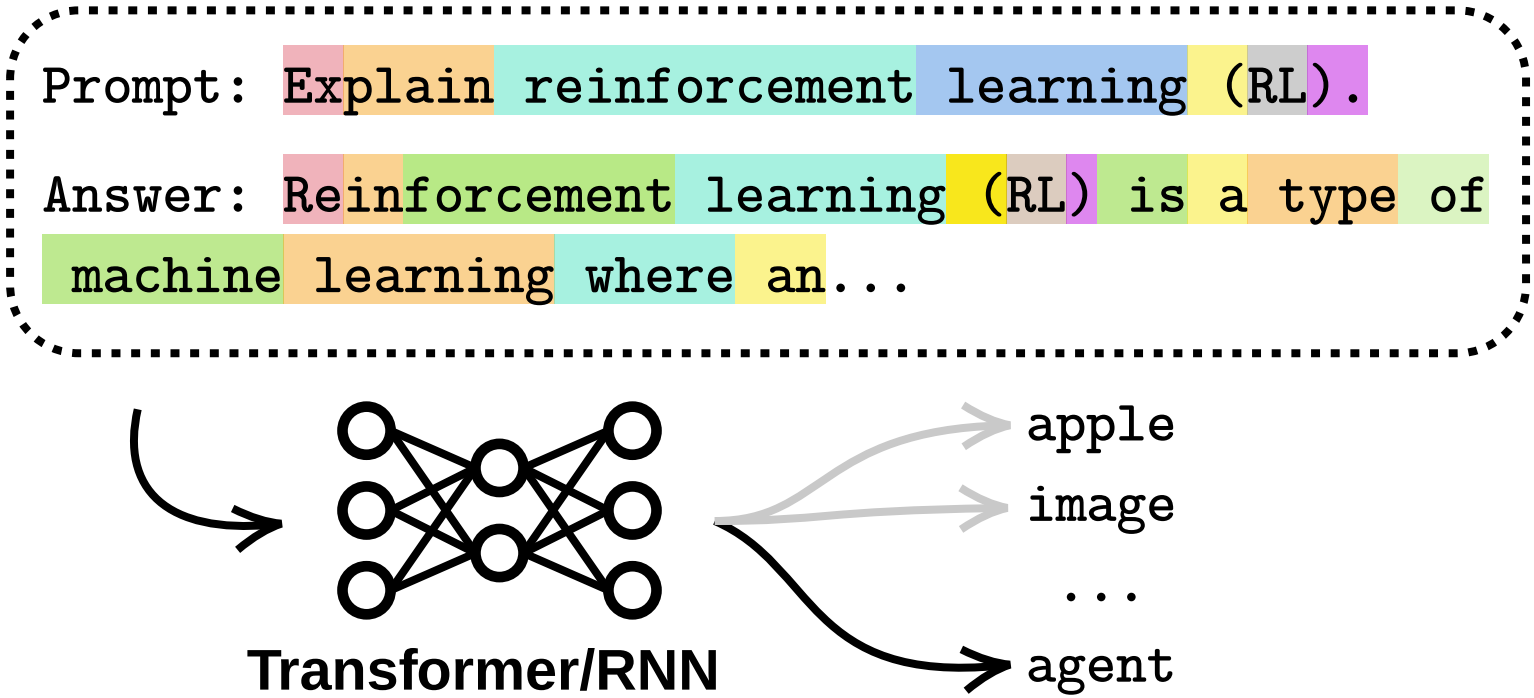
When the uncertainty set is χ^2 divergence



RMDPs are **harder** to learn than standard MDPs.

A crumb of RLHF

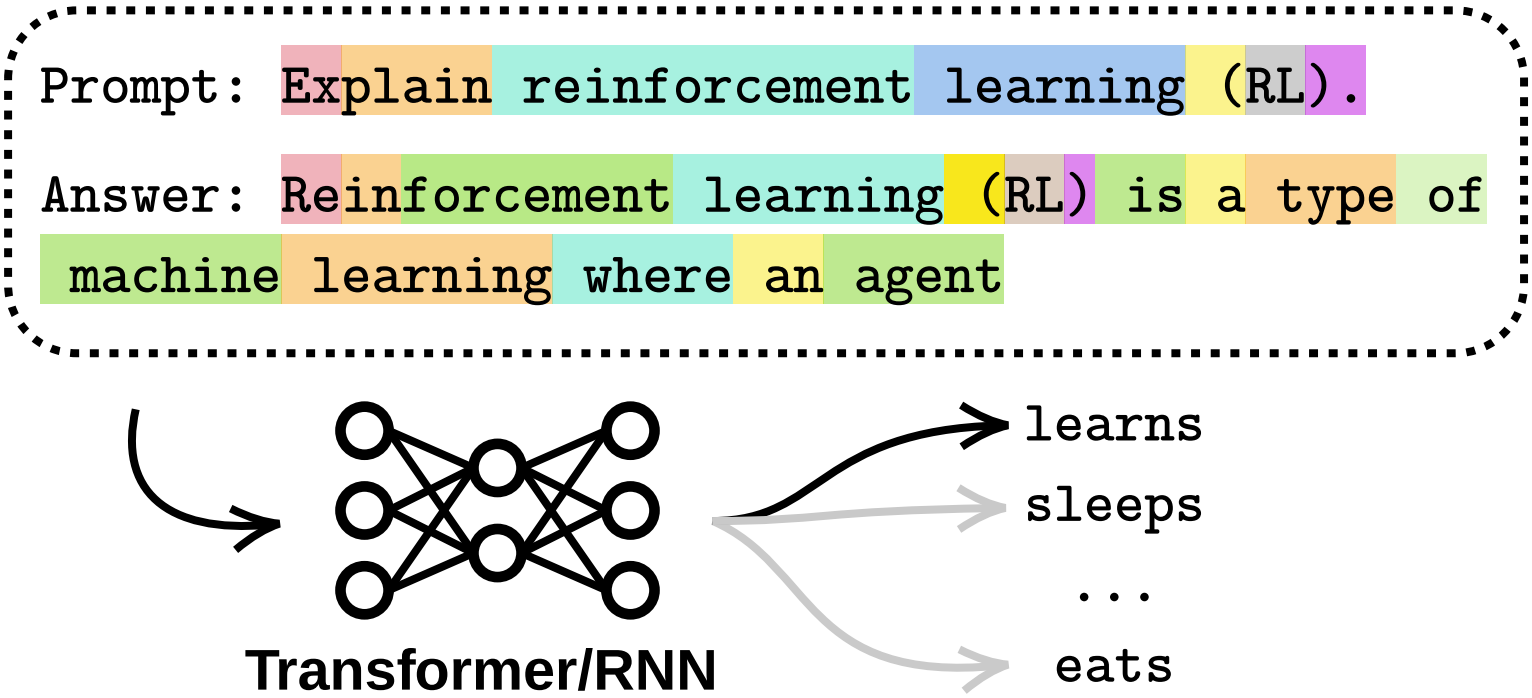
Language models as policies



Given prompt $x \in \mathcal{X}$, a language model generates an answer:

$$y \sim \underbrace{\pi(\cdot|x)}_{\text{parameterized by LLM}}$$

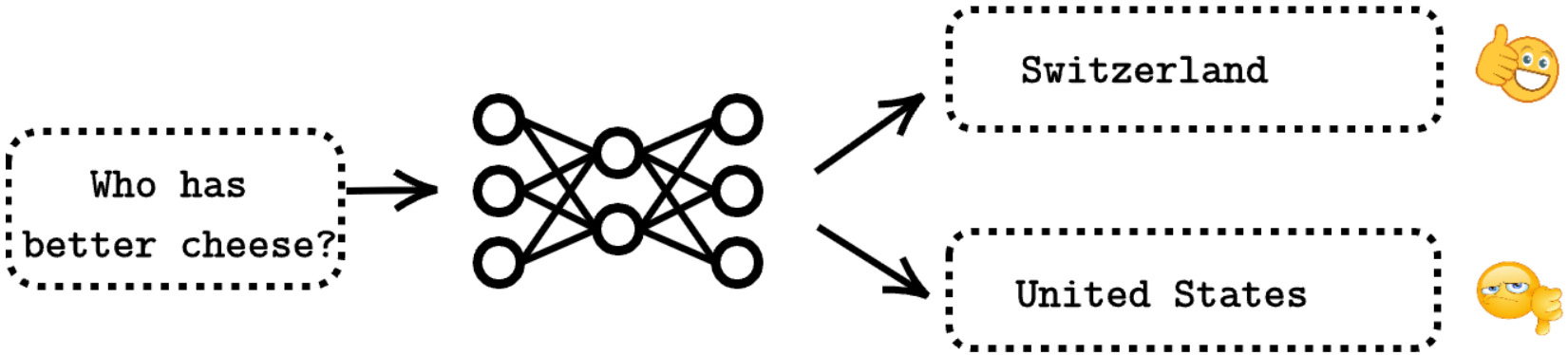
Language models as policies



Given prompt $x \in \mathcal{X}$, a language model generates an answer:

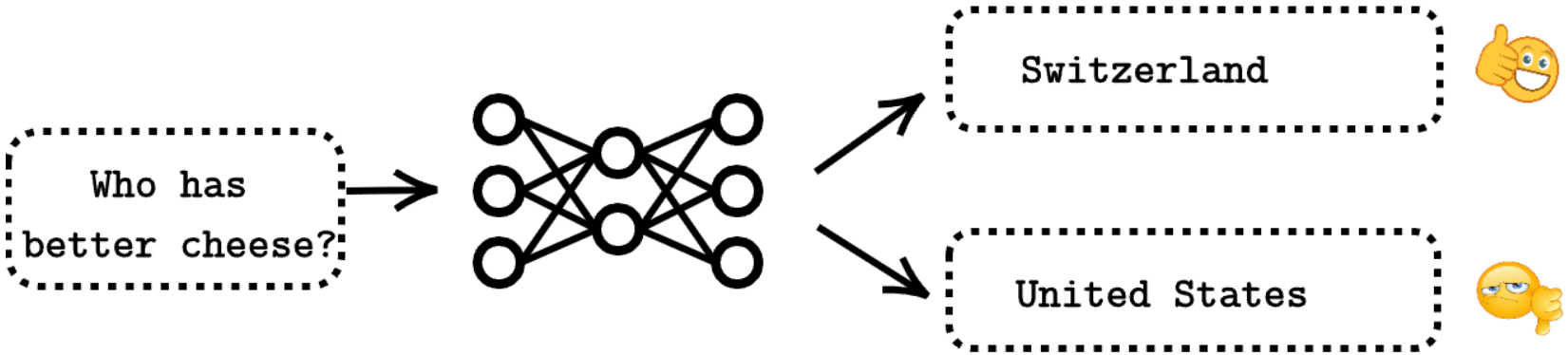
$$y \sim \underbrace{\pi(\cdot|x)}_{\text{parameterized by LLM}}$$

RL with human feedback (RLHF)



Goal: finetune the LLM to align with human preference

RL with human feedback (RLHF)



Goal: finetune the LLM to align with human preference

Prototypical pipeline:

- *Reward learning*: learn a reward model from preference data;
- *Policy optimization*: optimize the LLM to maximize the reward.

RLHF: reward learning

Bradly-Terry model

The probability of pairwise comparison $i \succ j$ is modeled by

$$\mathbb{P}(i \succ j) = \frac{\exp(r_i^\star)}{\exp(r_i^\star) + \exp(r_j^\star)} = \sigma(r_i^\star - r_j^\star),$$

where $r_i^\star \in \mathbb{R}$ is the score associated with item i.

RLHF: reward learning

Bradly-Terry model

The probability of pairwise comparison $i \succ j$ is modeled by

$$\mathbb{P}(i \succ j) = \frac{\exp(r_i^\star)}{\exp(r_i^\star) + \exp(r_j^\star)} = \sigma(r_i^\star - r_j^\star),$$

where $r_i^\star \in \mathbb{R}$ is the score associated with item i.

- **Reward model:** $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, evaluating the quality of a prompt-answer pair (x, y) that aligns with human preference;

RLHF: reward learning

Bradly-Terry model

The probability of pairwise comparison $i \succ j$ is modeled by

$$\mathbb{P}(i \succ j) = \frac{\exp(r_i^\star)}{\exp(r_i^\star) + \exp(r_j^\star)} = \sigma(r_i^\star - r_j^\star),$$

where $r_i^\star \in \mathbb{R}$ is the score associated with item i.

- **Reward model:** $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, evaluating the quality of a prompt-answer pair (x, y) that aligns with human preference;
- **Reward learning:** Given comparison data $\mathcal{D} = \{(x^i, y_+^i, y_-^i)\}_{i=1}^N$, the MLE of the reward function is given by

$$r_{\mathsf{MLE}} = \mathsf{argmin}_r \ell(r, \mathcal{D}),$$

where $\ell(r, \mathcal{D}) = -\sum_{i=1}^N \log \sigma(r(x^i, y_+^i) - r(x^i, y_-^i))$.

RLHF: policy optimization

Policy optimization via reward maximization

Find π that (approximately) maximizes the objective w.r.t. r

$$J(r, \pi) = \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y \sim \pi(\cdot|x)}} [r(x, y)] - \beta \mathop{\mathbb{E}}_{x \sim \rho} \left[\mathsf{KL}\big(\pi(\cdot|x) \,\|\, \pi_{\mathrm{ref}}(\cdot|x)\big) \right]$$

- $\beta > 0$: KL regularization parameter;
- π_{ref}: a reference policy, typically the model after SFT;
- $\rho \in \Delta(\mathcal{X})$: prompt distribution.

Direct Preference Optimization (DPO)

— (Rafailov et al., 2023)

1. Reward learning: $\widehat{r} \leftarrow \text{argmin}_r \ell(r, \mathcal{D})$
2. Policy learning: $\widehat{\pi} \leftarrow \text{argmax}_\pi J(\widehat{r}, \pi)$

Direct Preference Optimization (DPO)

— (Rafailov et al., 2023)

Observation: the optimal π w.r.t. r admits a closed-form solution

$$\pi_r = \text{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

Direct Preference Optimization (DPO)

— (Rafailov et al., 2023)

Observation: the optimal π w.r.t. r admits a closed-form solution

$$\pi_r = \text{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x,y)/\beta)}{Z(r,x)}.$$

- The reward function r in terms of its optimal π_r is

$$r(x,y) = \underbrace{\beta\big(\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x) + \log Z(r,x)\big)}_{=:r(\pi)}.$$

Direct Preference Optimization (DPO)

— (Rafailov et al., 2023)

- The reward function r in terms of its optimal π_r is

$$r(x,y) = \underbrace{\beta\big(\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x) + \log Z(r,x)\big)}_{=:r(\pi)}.$$

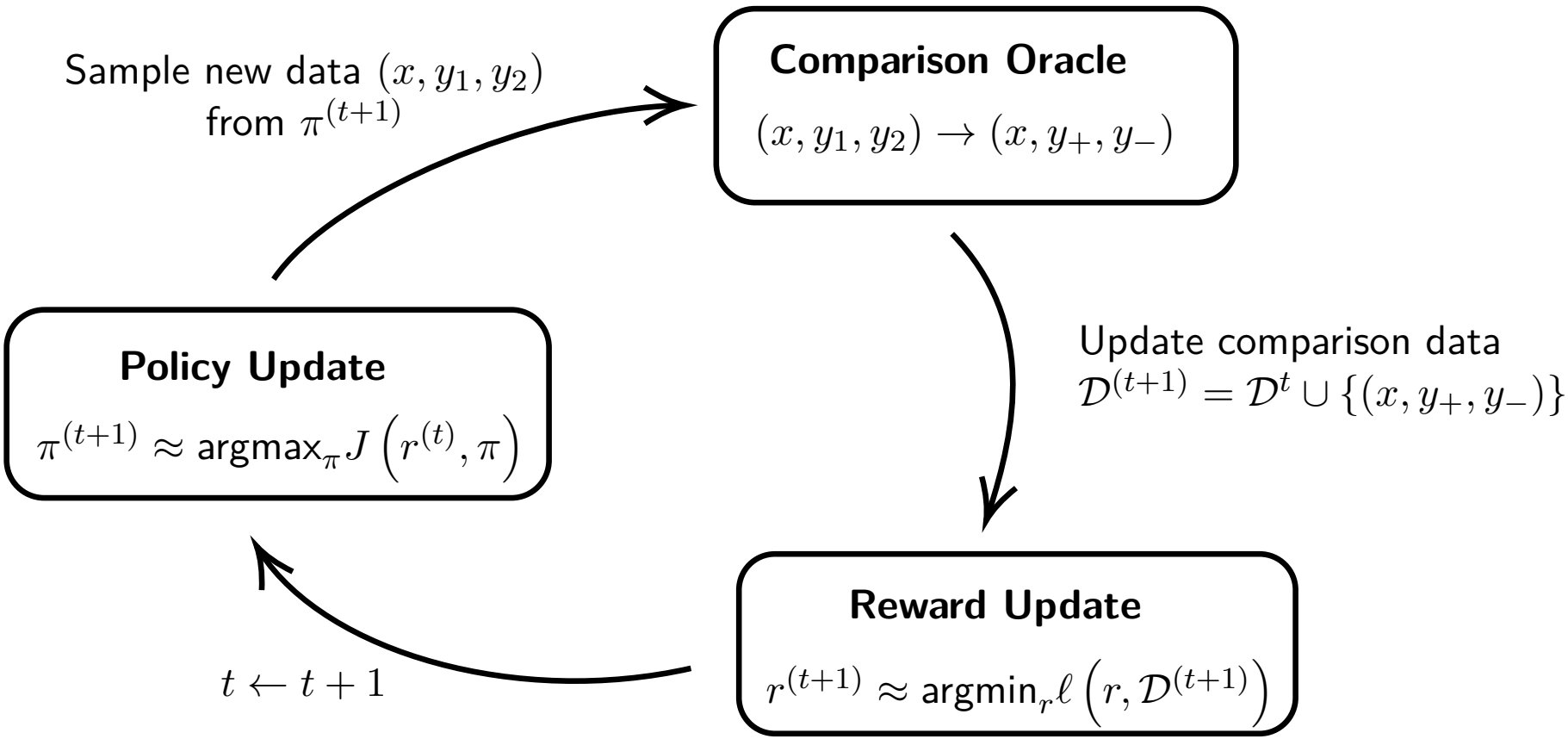
- The two-step procedure is equivalent to

$$\begin{aligned}\widehat{\pi} &\leftarrow \mathsf{argmin}_\pi \ell(r(\pi), \mathcal{D}) \\ &= -\sum_{i=1}^{N} \log \sigma\left(\beta \log \frac{\pi(y_+^i|x^i)}{\pi_{\text{ref}}(y_+^i|x^i)} - \beta \log \frac{\pi(y_-^i|x^i)}{\pi_{\text{ref}}(y_-^i|x^i)}\right).\end{aligned}$$

Single-step and **policy-only**! Very popular in practice.

Online RLHF

Leverage online data collection to improve data coverage - how do we perform **exploration** in the policy space directly?



Exploration via optimistic MLE

Optimistic MLE: Bias the estimate towards the models with higher optimal objective $J^\star(r) = \max_\pi J(r, \pi)$ by

$$r^{(t+1)} \leftarrow \mathsf{argmin}_r \{\ell(r, \mathcal{D}^{(t)}) - \alpha J^\star(r)\}.$$

Exploration via optimistic MLE

Optimistic MLE: Bias the estimate towards the models with higher optimal objective $J^\star(r) = \max_\pi J(r, \pi)$ by

$$r^{(t+1)} \leftarrow \mathsf{argmin}_r \{\ell(r, \mathcal{D}^{(t)}) - \alpha J^\star(r)\}.$$

- *Small caveat:* the update is not well-defined, since the BT model cannot distinguish between r and $r + c \cdot \mathbf{1}$, while

$$J^\star(r + c \cdot \mathbf{1}) = J^\star(r) + c.$$

Exploration via optimistic MLE

Optimistic MLE: Bias the estimate towards the models with higher optimal objective $J^\star(r) = \max_\pi J(r, \pi)$ by

$$r^{(t+1)} \leftarrow \mathsf{argmin}_{r \in \mathcal{R}} \{\ell(r, \mathcal{D}^{(t)}) - \alpha J^\star(r)\}.$$

- *Small caveat:* the update is not well-defined, since the BT model cannot distinguish between r and $r + c \cdot \mathbf{1}$, while

$$J^\star(r + c \cdot \mathbf{1}) = J^\star(r) + c.$$

- We can resolve the shift ambiguity by focusing on the following equivalent class of reward functions:

$$\mathcal{R} = \left\{ r : \mathop{\mathbb{E}}_{x \sim \rho, y \sim \pi_{\mathrm{cal}}(\cdot|x)} [r(x, y)] = 0 \right\}.$$

Exploration via optimistic MLE

Optimistic MLE: Bias the estimate towards the models with higher optimal objective $J^\star(r) = \max_\pi J(r, \pi)$ by

$$r^{(t+1)} \leftarrow \mathsf{argmin}_{r \in \mathcal{R}} \{\ell(r, \mathcal{D}^{(t)}) - \alpha J^\star(r)\}.$$

- *Small caveat:* the update is not well-defined, since the BT model cannot distinguish between r and $r + c \cdot \mathbf{1}$, while

$$J^\star(r + c \cdot \mathbf{1}) = J^\star(r) + c.$$

- We can resolve the shift ambiguity by focusing on the following equivalent class of reward functions:

$$\mathcal{R} = \left\{ r : \mathop{\mathbb{E}}_{x \sim \rho, y \sim \pi_{\mathrm{cal}}(\cdot|x)} [r(x, y)] = 0 \right\}.$$

Can we avoid solving a bilevel optimization problem?

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \mathsf{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \text{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

- We can write the $J^\star(r)$ as

$$\begin{aligned} J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[\log Z(r, x) \right] \end{aligned}$$

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \text{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

- We can write the $J^\star(r)$ as

$$\begin{aligned} J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} [\log Z(r, x)] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log Z(r, x)] \end{aligned}$$

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \text{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

- We can write the $J^\star(r)$ as

$$\begin{aligned} J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} [\log Z(r, x)] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log Z(r, x)] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \end{aligned}$$

Exploiting the structure of $J(r, \pi)$

- The optimal policy admits the following closed-form solution:

$$\pi_r = \mathsf{argmax}_\pi J(r, \pi) \iff \pi_r(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}.$$

- We can write the $J^\star(r)$ as

$$\begin{aligned} J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[r(x, y) - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} \left[\log Z(r, x) \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} \left[\log Z(r, x) \right] \\ &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} \left[\cancel{r(x, y)} - \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} \right] \end{aligned}$$

Value-incentivized preference optimization (VPO)

$$\pi^{(t+1)} \leftarrow \mathsf{argmin}_\pi\{\ell(r(\pi), \mathcal{D}^{(t)}) - \alpha J^\star(r(\pi))\}.$$

- The negative log-likelihood term reformulates into DPO loss:

$$\ell(r(\pi), \mathcal{D}^{(t)})$$
$$= -\sum_{(x,y_+,y_-)\in\mathcal{D}^{(t)}} \log\sigma\bigg(\beta\Big(\log\frac{\pi(y_+|x)}{\pi_{\mathrm{ref}}(y_+|x)} - \log\frac{\pi(y_-|x)}{\pi_{\mathrm{ref}}(y_-|x)}\Big)\bigg).$$

- The reward bias term can be written as:

$$J^\star(r(\pi)) = -\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\mathrm{cal}}(\cdot|x)} \left[\log\pi(y|x) - \log\pi_{\mathrm{ref}}(y|x)\right],$$

  which is essentially becomes a reverse-KL regularization that maximizes $\mathsf{KL}\big(\pi_{\mathrm{cal}}(\cdot|x) \,\|\, \pi(\cdot|x)\big)$.

Main results - online VPO

Theorem (Cen et al., ICLR 2025)

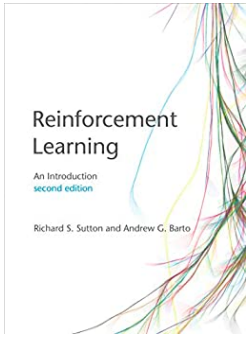
Assume that reward estimates $\|r^{(t)}\|_\infty \le B$ *and* $\|r^\star\|_\infty \le B$ *for some* $B > 0$. *With high probability we have*

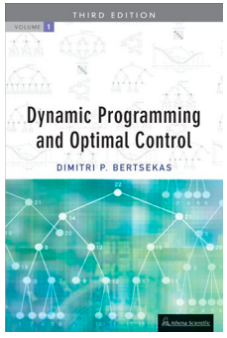
$$\sum_{t=1}^{T} \left[J^\star(r^\star) - J(r^\star, \pi^{(t)})\right] \le \widetilde{O}(\sqrt{T}).$$

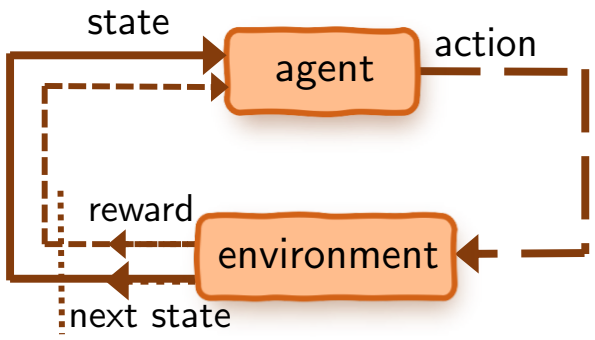
- We can obtain similar regret bounds under general function approximation of the reward model.
- Consistent with the $\widetilde{O}(\sqrt{T})$ regret for online RL with UCB bonus.
- Offline RL: flipping the sign of α leads to a pessimistic algorithm.

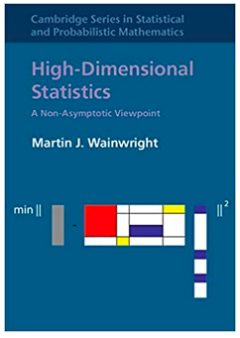
Concluding Remarks

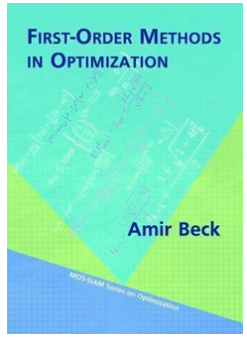
Concluding remarks











Understanding non-asymptotic performances of RL algorithms is a fruitful playground!

Promising directions:

- function approximation
- multi-agent RL
- RL for foundation models
- many more...

Thank you!

Statistical and Algorithmic Foundations of Reinforcement Learning

Yuejie Chi
Department of Statistics and Data Science, Yale University, New Haven, CT 06511, USA, yuejie.chi@yale.edu

Yuxin Chen, Yuting Wei
Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104, USA, {yuxinc@wharton.upenn.edu, ytwei@wharton.upenn.edu}

Abstract As a paradigm for sequential decision making in unknown environments, reinforcement learning (RL) has received a flurry of attention in recent years. However, the explosion of model complexity in emerging applications and the presence of nonconvexity exacerbate the challenge of achieving efficient RL in sample-starved situations, where data collection is expensive, time-consuming, or even high-stakes (e.g., in clinical trials, autonomous systems, and online advertising). How to understand and enhance the sample and computational efficacies of RL algorithms is thus of great interest. In this tutorial, we aim to introduce several important algorithmic and theoretical developments in RL, highlighting the connections between new ideas and classical topics. Employing Markov Decision Processes as the central mathematical model, we cover several distinctive RL scenarios (i.e., RL with a simulator, online RL, offline RL, robust RL, and RL with human feedback), and present several mainstream RL approaches (i.e., model-based approach, value-based approach, and policy optimization). Our discussions gravitate around the issues of sample complexity, computational efficiency, as well as algorithm-dependent and information-theoretic lower bounds from a non-asymptotic viewpoint.