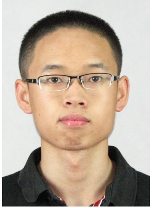
The Cost of Distributional Robustness in Reinforcement Learning

— minimax-optimal sample efficiency

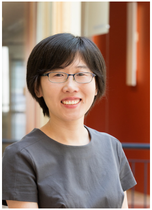
Laixi Shi

Computing & Mathematical Sciences
California Institute of Technology

WORDS 2023
The Fuqua School of Business, Duke University





Gen Li
CUHK

Yuting Wei
UPenn

Yuxin Chen
UPenn

Matthieu Geist
Google Brain

Yuejie Chi
CMU

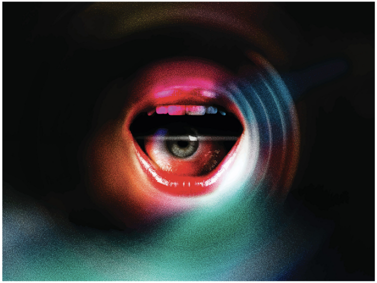
Artificial intelligence (AI): an amazing future

The New ChatGPT Can 'See' and 'Talk.' Here's What It's Like.

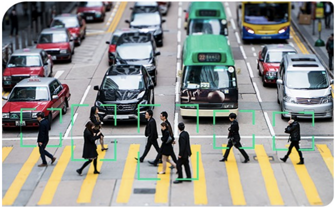
The image-recognition feature could have many uses, and the voice feature is even more intriguing.

Published Sept. 27, 2023

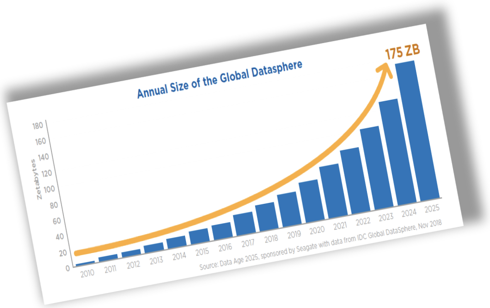
The New York Times



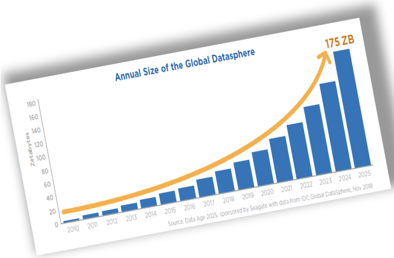
Artificial intelligence (AI): an amazing future

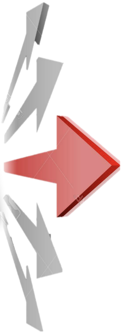


Data is the key of AI

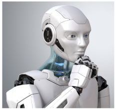


Data is the key of AI

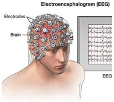





Radio Astronomy


Robotics

Decision-making



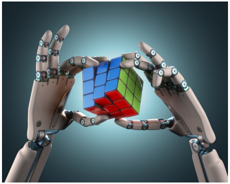
Biology

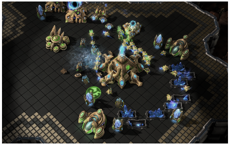

Sports

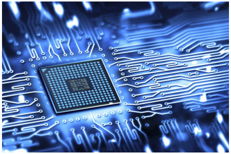

Healthcare

Creating AI for diverse applications using data science.

Decision-making AI: RL is promising

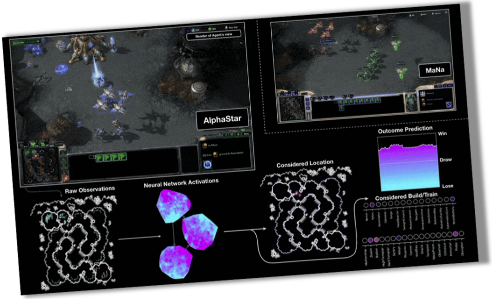






RL holds great promise in the next era of artificial intelligence.

RL: pretty data-starved



30 millions of moves

200 years of StarCraft video play

The agent need to explore a lot for difficult/complicated tasks.

Sample efficiency

A pressing need of sample efficiency:

- Enormous state/action space of the unknown environment
- Data collection can be costly, time-consuming, or high-stakes

clinical trials

autonomous driving

Chat robot

A pressing need of sample efficiency:

- Enormous state/action space of the unknown environment
- Data collection can be costly, time-consuming, or high-stakes

clinical trials

autonomous driving

Chat robot

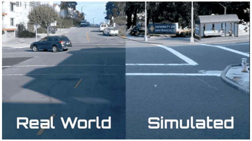
Calls for design of sample-efficient RL algorithms!

Robustness

Robustness is a cornerstone of tackling with

- Uncertainty and noise of the environment
- Simulation-to-reality gaps and generalization requirements

Uncertainty



Sim-to-real gaps

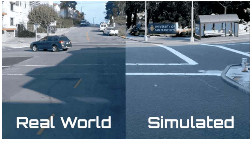
Generalization

Robustness

Robustness is a cornerstone of tackling with

- Uncertainty and noise of the environment
- Simulation-to-reality gaps and generalization requirements

Uncertainty



Sim-to-real gaps

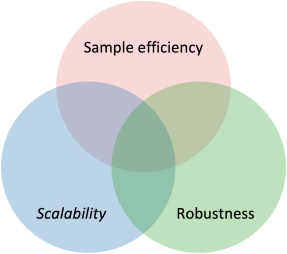
Generalization

Calls for design of robust RL algorithms!

Overview

Understand and design RL algorithms in the face of sample efficiency, scalability, and robustness.

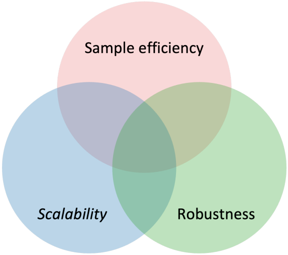
Theory	**Robust RL:** *[Shi et al. ‘23], [Shi and Chi. '22]*
	Online RL: *[Li et al. '21]*
	Offline RL: *[Shi et al. '22], [Li et al. '22]*
Practice	**Robust RL:** *[Ding et al. '23]*
	Offline RL: *[Shi et al. ‘23], [Wang et al. '23]*
	Curriculum RL: *[Huang et al. '22]*



Overview

Understand and design RL algorithms in the face of sample efficiency, scalability, and robustness.

Theory	**Robust RL:** *[Shi et al. '23], [Shi and Chi. '22]* **Online RL:** *[Li et al. '21]* **Offline RL:** *[Shi et al. '22], [Li et al. '22]*
Practice	**Robust RL:** *[Ding et al. '23]* **Offline RL:** *[Shi et al. '23], [Wang et al. '23]* **Curriculum RL:** *[Huang et al. '22]*



Outline of this talk: robust RL

Background: Markov decision processes (MDPs)

Outline of this talk: robust RL

Background: Markov decision processes (MDPs)

Problem formulation: distributionally robust RL

Outline of this talk: robust RL

Background: Markov decision processes (MDPs)

Problem formulation: distributionally robust RL

I: The cost of distributional robustness in RL

Standard RL: Learn the optimal policy for a fixed environment?

Robust RL: Learn the optimal policy with additional robustness to environment shift

Do robust RL need more samples

Outline of this talk: robust RL

Background: Markov decision processes (MDPs)

Problem formulation: distributionally robust RL

I: The cost of distributional robustness in RL

Standard RL: Learn the optimal policy for a fixed environment?

Robust RL: Learn the optimal policy with additional robustness to environment shift

Do robust RL need more samples → **This work:** solving robust RL may need less samples

Outline of this talk: robust RL

Background: Markov decision processes (MDPs)

Problem formulation: distributionally robust RL

I: The cost of distributional robustness in RL

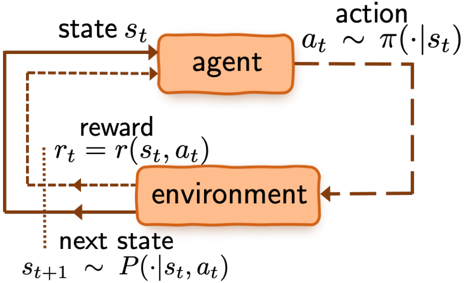
Will solving robust RL be inherently harder than standard RL in terms of sample requirements?

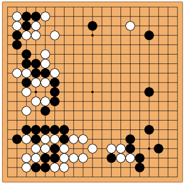
II: Design sample efficient offline robust RL algorithm

Can we design a near-optimal algorithm that can learn under simultaneous model uncertainty and limited historical datasets?

Background: Markov decision process

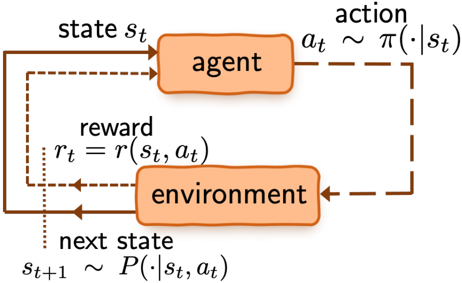
Markov decision processes

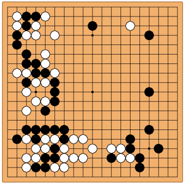




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

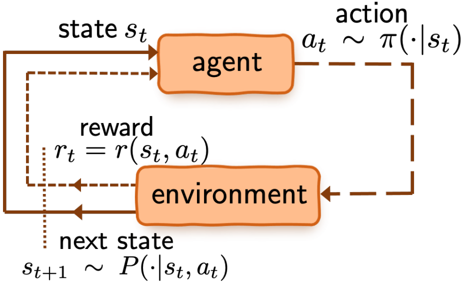
Markov decision processes

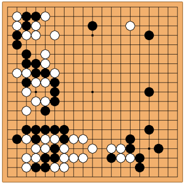




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward

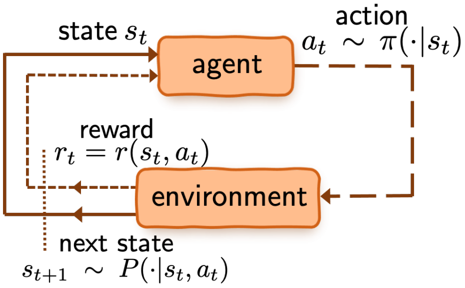
Markov decision processes

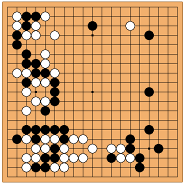




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

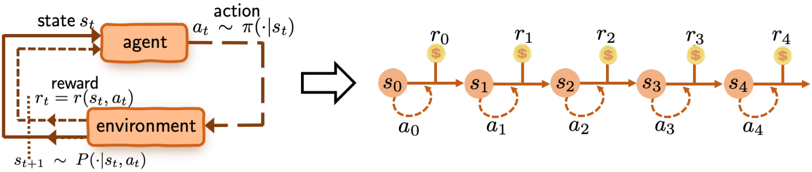
Markov decision processes





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

Value function

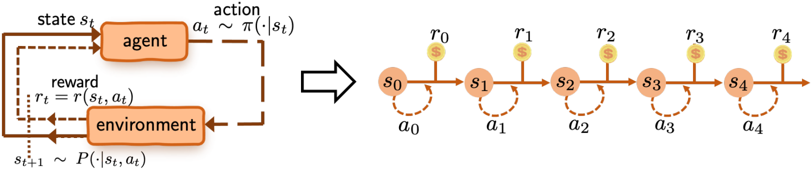


Value/Q-function function of policy π:

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

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

Value function



Value/Q-function function of policy π:

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

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

- $\gamma \in [0,1)$ is the discount factor; $\frac{1}{1-\gamma}$ is effective horizon
- Expectation is w.r.t. the sampled trajectory under π over P

Problem formulation: robust RL

Motivation: safety and robustness in RL

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

Training environment $\neq$ Test environment

(Sim-to-real gaps / generalization requirements / random noise)

Motivation: safety and robustness in RL

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

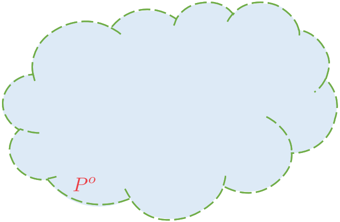
Training environment $\neq$ Test environment

(Sim-to-real gaps / generalization requirements / random noise)

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

Uncertainty set of the nominal transition kernel P^o:

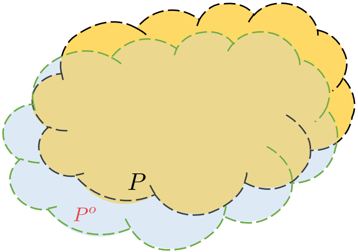
$$\mathcal{U}^\sigma(P^o) = \left\{P : \rho\big(P, P^o\big) \leq \sigma\right\}$$

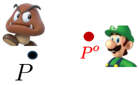


Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

$$\mathcal{U}^\sigma(P^o) = \left\{P : \ \rho\big(P, P^o\big) \le \sigma\right\}$$

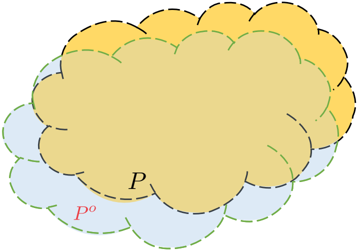


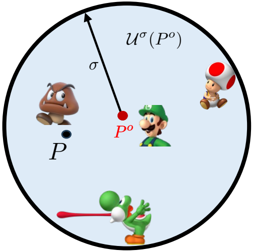


Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

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

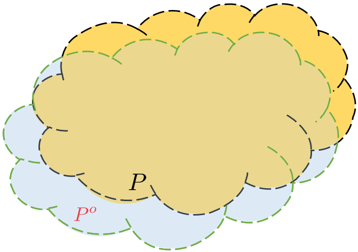


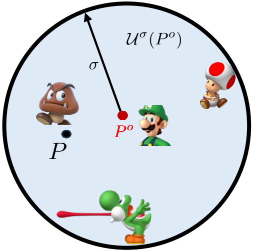


Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

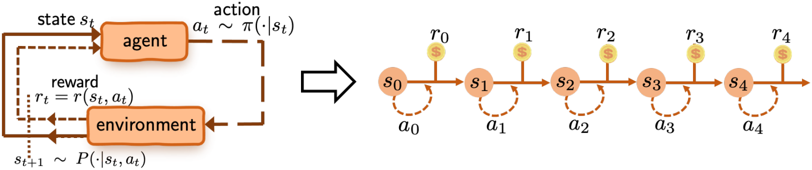
$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \le \sigma\}$$





- Examples of ρ: f-divergence (TV, χ^2, KL...), Wasserstein distance
- Under (s,a)-rectangularity: $P_{s,a} \in \mathcal{U}^\sigma(P^o_{s,a})$

Robust value/Q function



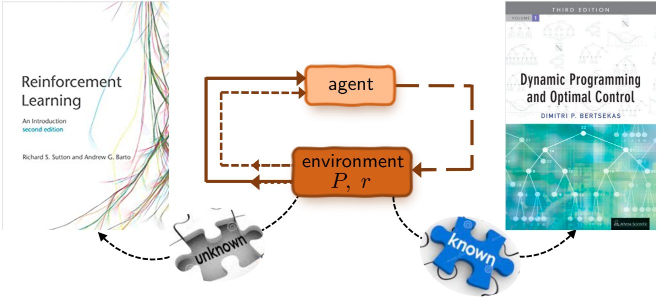
Robust value/Q function of policy π:

$$\forall s \in \mathcal{S}: \quad V^{\pi,\sigma}(s) := \inf_{P \in \mathcal{U}^{\sigma}(P^{o})} V^{\pi,P}(s)$$

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q^{\pi,\sigma}(s,a) := \inf_{P \in \mathcal{U}^{\sigma}(P^{o})} Q^{\pi,P}(s,a)$$

Measures the worst-case performance of the policy when the transition kernel $P \in$ uncertainty set $\mathcal{U}^{\sigma}(P^{o})$.

Distributionally robust MDP

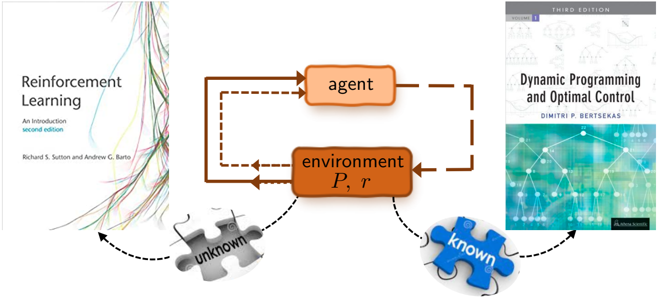


Robust MDP

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

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

Distributionally robust MDP



Robust MDP

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

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

- optimal robust value / Q function: $V^{\star,\sigma} := V^{\pi^\star,\sigma}$, $Q^{\star,\sigma} := Q^{\pi^\star,\sigma}$
- optimal robust policy $\pi^\star(s) = \mathrm{argmax}_{a\in\mathcal{A}}\, Q^{\star,\sigma}(s,a)$

Distributionally robust Bellman's optimality equation

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

Robust Bellman's optimality equation: the optimal robust policy $\pi^\star$ satisfies

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

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

Distributionally robust Bellman's optimality equation

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

Robust Bellman's optimality equation: the optimal robust policy $\pi^\star$ satisfies

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

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

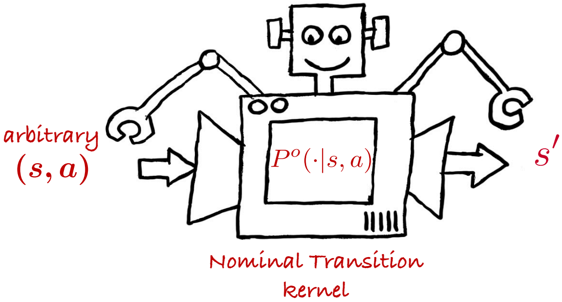
$$Q(s,a) \leftarrow r(s,a) + \gamma \inf_{P_{s,a}\in\mathcal{U}^\sigma\left(P^o_{s,a}\right)} \langle P_{s,a}, V\rangle,$$

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

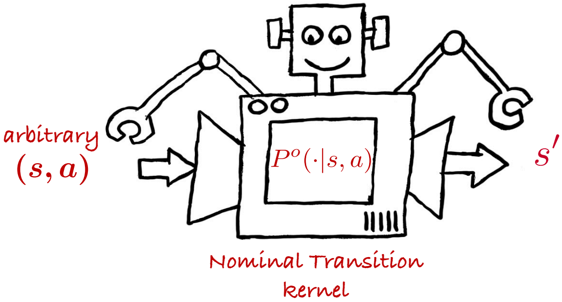
I: The curious sample complexity price of solving distributionally robustness RL

— Benchmark with standard RL

Distributionally robust RL with a generative model



Distributionally robust RL with a generative model



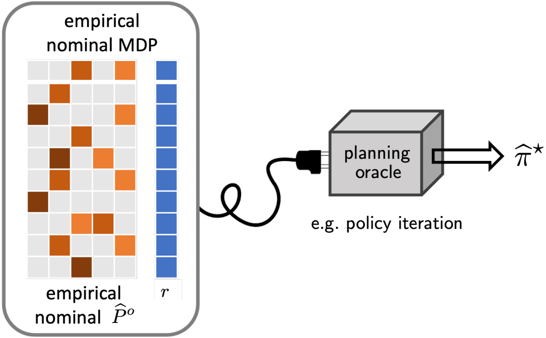
Goal of robust RL: given $\mathcal{D} := \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ from the *nominal* environment P^o, find an ϵ-optimal robust policy $\widehat{\pi}$ obeying

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

— *in a sample-efficient manner*

Model-based RL: empirical MDP + planning

— Azar et al., 2013, Agarwal et al., 2019



Find policy based on the empirical MDP

$$\underbrace{\text{Find policy}}_{\text{using, e.g., policy iteration}} \text{ based on the } \underbrace{\text{empirical MDP}}_{(\widehat{P}^o, r)}$$

Distributionally robust Bellman's optimality equation

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

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

Distributionally robust Bellman's optimality equation

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

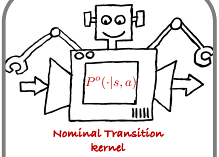
Planning by **distributionally robust value iteration (DRVI)**:

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

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

Involves an additional inner optimization problem $(\inf_{P_{s,a} \in \mathcal{U}^{\sigma}(\widehat{P}^{o}_{s,a})})$ compared to standard RL

A curious open question: robust RL v.s. standard RL



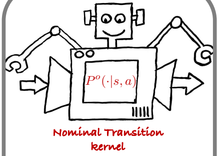
with dataset $\mathcal{D}$ from nominal P^o

Standard RL: Learn the optimal policy of the nominal MDP?

Robust RL: Learn the **robust** policy around the nominal MDP?

Which one need more samples

A curious open question: robust RL v.s. standard RL



with dataset $\mathcal{D}$ from nominal P^o

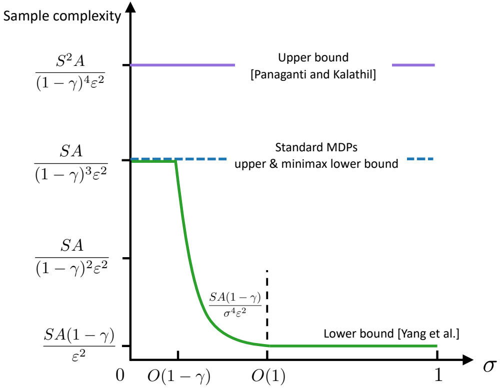
Standard RL: Learn the optimal policy of the nominal MDP?

Robust RL: Learn the **robust** policy around the nominal MDP?

Which one need more samples

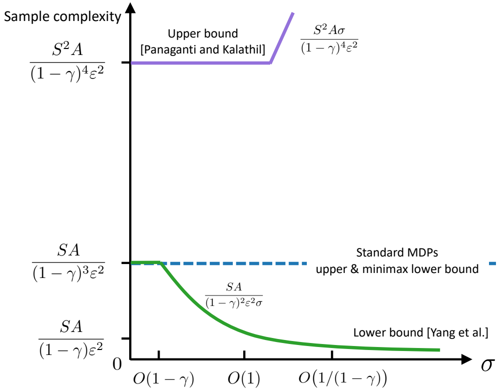
Robustness-statistical trade-off? Is there a statistical premium that one needs to pay in quest of additional robustness?

Prior art: robust RL with TV uncertainty



- Large gaps between existing upper and lower bounds
- Unclear benchmarking with standard MDP

Prior art: robust RL with χ^2 uncertainty



- Large gaps between existing upper and lower bounds
- Unclear benchmarking with standard MDP

Our theorems 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 $\epsilon > 0$, DRVI outputs a policy $\widehat{\pi}$ that satisfies $V^{\star,\sigma} - V^{\widehat{\pi},\sigma} \leq \epsilon$ with sample complexity at most

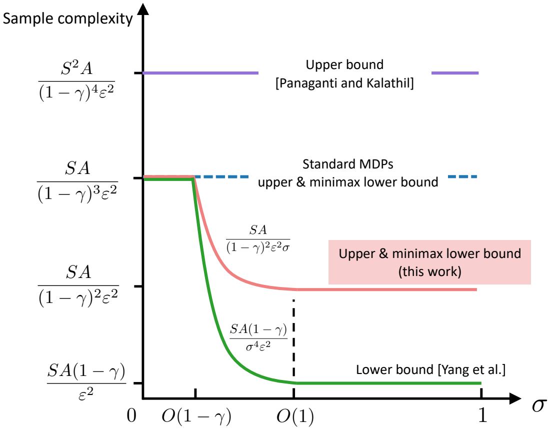
$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\}\epsilon^2}\right)$$

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

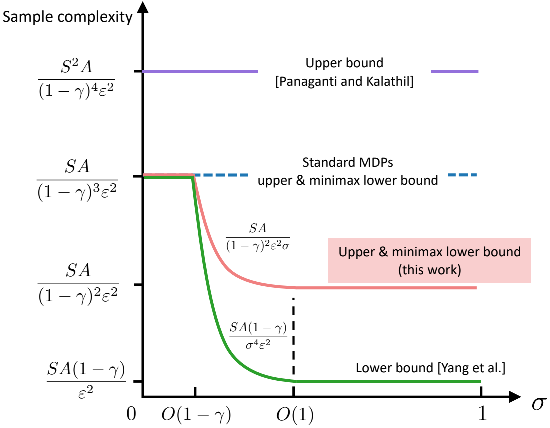
$$\widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\}\epsilon^2}\right).$$

- 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.

Our theorems under χ^2 uncertainty

Theorem (Upper bound, Shi et al., 2023)

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

$$\widetilde{O}\left(\frac{SA(1+\sigma)}{(1-\gamma)^4\epsilon^2}\right)$$

ignoring logarithmic factors.

Our theorems under χ^2 uncertainty

Theorem (Upper bound, Shi et al., 2023)

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

$$\widetilde{O}\left(\frac{SA(1+\sigma)}{(1-\gamma)^4\epsilon^2}\right)$$

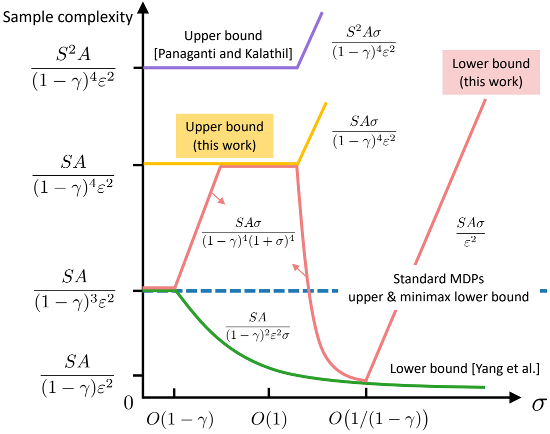
ignoring logarithmic factors.

Theorem (Lower bound, Shi et al., 2023)

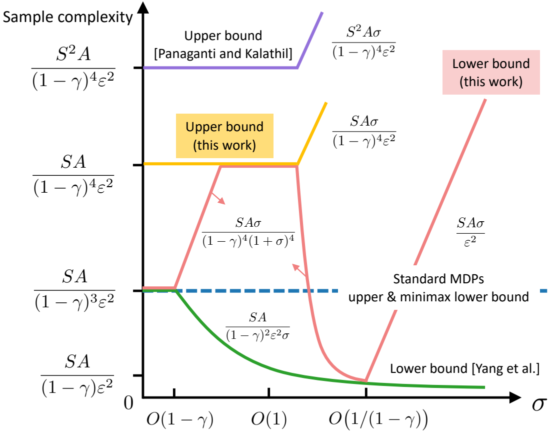
In addition, no algorithm succeeds when the sample size is below

$$\begin{cases} \widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^3\epsilon^2}\right) & \text{if } \sigma \lesssim 1-\gamma \\ \widetilde{\Omega}\left(\frac{\sigma SA}{\min\{1,(1-\gamma)^4(1+\sigma)^4\}\epsilon^2}\right) & \text{otherwise} \end{cases}$$

When the uncertainty set is χ^2 divergence



When the uncertainty set is χ^2 divergence



RMDPs can be much **harder** to learn than standard MDPs.

Why robust RL is easier/harder than standard RL?

Technical challenge: robust RL v.s. standard RL

- Control the error terms based on estimate $\widehat{P}^o$:

Standard RL: $$\delta_{\mathsf{RL}} = \underbrace{\left| P^o \widehat{V} - \widehat{P}^0 \widehat{V} \right|}_{\text{linear w.r.t. } P^o - \widehat{P}^0}$$

Robust RL: $$\delta_{\mathsf{rob}} = \underbrace{\left| \inf_{\mathcal{P} \in \mathcal{U}^\sigma_\rho(P^o)} \mathcal{P} \widehat{V}^\sigma_{\mathsf{rob}} - \inf_{\mathcal{P} \in \mathcal{U}^\sigma_\rho(\widehat{P}^0)} \mathcal{P} \widehat{V}^\sigma_{\mathsf{rob}} \right|}_{\text{complex form w.r.t. } P^o - \widehat{P}^0 \text{ due to inner problem over } \mathcal{U}^\sigma_\rho(\cdot)}$$

Technical challenge: robust RL v.s. standard RL

- Control the error terms based on estimate $\widehat{P}^o$:

Standard RL: $$\delta_{\mathsf{RL}} = \underbrace{\left| P^o \widehat{V} - \widehat{P}^0 \widehat{V} \right|}_{\text{linear w.r.t. } P^o - \widehat{P}^0}$$

Robust RL: $$\delta_{\mathsf{rob}} = \underbrace{\left| \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(P^o)} \mathcal{P} \widehat{V}_{\mathsf{rob}}^\sigma - \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(\widehat{P}^0)} \mathcal{P} \widehat{V}_{\mathsf{rob}}^\sigma \right|}_{\text{complex form w.r.t. } P^o - \widehat{P}^0 \text{ due to inner problem over } \mathcal{U}_\rho^\sigma(\cdot)}$$

- Main factors:
    - the error function (δ_{RL} or δ_{rob}) w.r.t. model estimate error $P^o - \widehat{P}^0$;
    - the range of value functions $\widehat{V}$ or $\widehat{V}_{\mathsf{rob}}^\sigma$.

Technical challenge: robust RL v.s. standard RL

- Control the error terms based on estimate $\widehat{P}^o$:

$$\text{Standard RL:} \quad \delta_{\mathsf{RL}} = \underbrace{\left| P^o \widehat{V} - \widehat{P}^0 \widehat{V} \right|}_{\text{linear w.r.t. } P^o - \widehat{P}^0}$$

$$\text{Robust RL:} \quad \delta_{\mathsf{rob}} = \underbrace{\left| \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(P^o)} \mathcal{P} \widehat{V}^\sigma_{\mathsf{rob}} - \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(\widehat{P}^0)} \mathcal{P} \widehat{V}^\sigma_{\mathsf{rob}} \right|}_{\text{complex form w.r.t. } P^o - \widehat{P}^0 \text{ due to inner problem over } \mathcal{U}_\rho^\sigma(\cdot)}$$

- Main factors:
  - the error function (δ_{RL} or δ_{rob}) w.r.t. model estimate error $P^o - \widehat{P}^0$;
  - the range of value functions $\widehat{V}$ or $\widehat{V}^\sigma_{\mathsf{rob}}$.

Using same size of samples (same $\widehat{P}^o$), smaller error $\rightarrow$ easier task

Intuition for tighter bound

- **TV:**
  - linear dependency w.r.t $P^o - \widehat{P}^0$: $\delta_{\mathsf{rob}} = \left| P^o \widehat{V}_{\mathsf{rob}} - \widehat{P}^0 \widehat{V}_{\mathsf{rob}} \right|$
  - the range of $\widehat{V}_{\mathsf{rob}}^{\sigma}$ contracts rapidly as σ grows $\rightarrow$ smaller than the range of $\widehat{V}$ in standard RL

Intuition for tighter bound

- **TV:**
  - linear dependency w.r.t $P^o - \widehat{P}^0$: $\delta_{\mathsf{rob}} = \left| P^o \widehat{V}_{\mathsf{rob}} - \widehat{P}^0 \widehat{V}_{\mathsf{rob}} \right|$
  - the range of $\widehat{V}^{\sigma}_{\mathsf{rob}}$ contracts rapidly as σ grows $\to$ smaller than the range of $\widehat{V}$ in standard RL

smaller range of $\widehat{V}^{\sigma}_{\mathsf{rob}}$ $\to$ RMDPs are easier than standard MDPs

Intuition for tighter bound

- **TV:**
  - linear dependency w.r.t $P^o - \widehat{P}^0$: $\delta_{\mathsf{rob}} = \left| P^o \widehat{V}_{\mathsf{rob}} - \widehat{P}^0 \widehat{V}_{\mathsf{rob}} \right|$
  - the range of $\widehat{V}^\sigma_{\mathsf{rob}}$ contracts rapidly as σ grows $\rightarrow$ smaller than the range of $\widehat{V}$ in standard RL

smaller range of $\widehat{V}^\sigma_{\mathsf{rob}}$ $\rightarrow$ RMDPs are easier than standard MDPs

- χ^2**:**
  - Non-linear and sensitive w.r.t. $P^o - \widehat{P}^0$ $\rightarrow$ even if $P^o - \widehat{P}^0$ is small, the error term δ_{rob} can explode.
  - the range of $\widehat{V}^\sigma_{\mathsf{rob}}$ can be similar to V

Intuition for tighter bound

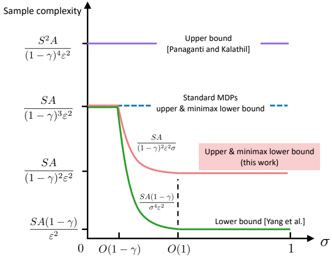
- **TV:**
  - linear dependency w.r.t $P^o - \widehat{P}^0$: $\delta_{\text{rob}} = \left| P^o \widehat{V}_{\text{rob}} - \widehat{P}^0 \widehat{V}_{\text{rob}} \right|$
  - the range of $\widehat{V}_{\text{rob}}^{\sigma}$ contracts rapidly as σ grows $\to$ smaller than the range of $\widehat{V}$ in standard RL

smaller range of $\widehat{V}_{\text{rob}}^{\sigma}$ $\to$ RMDPs are easier than standard MDPs

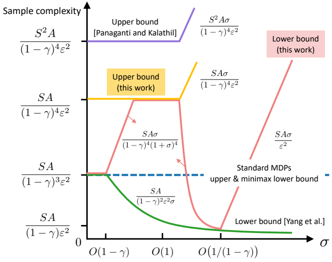
- χ^2:
  - Non-linear and sensitive w.r.t. $P^o - \widehat{P}^0$ $\to$ even if $P^o - \widehat{P}^0$ is small, the error term δ_{rob} can explode.
  - the range of $\widehat{V}_{\text{rob}}^{\sigma}$ can be similar to V

Complicated error terms $\to$ RMDPs are harder than standard MDPs

Takeaway: statistical implications of robustness



TV uncertainty

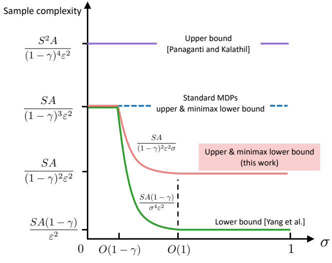


χ^2 uncertainty

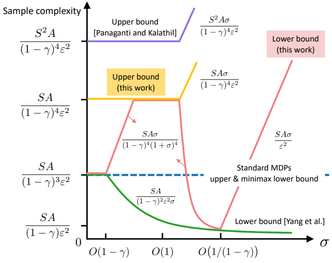
RMDPs are neither necessarily harder nor easier than standard RL in terms of sample requirements.

— depend heavily on the shape and size of the uncertainty set

Takeaway: statistical implications of robustness



TV uncertainty



χ^2 uncertainty

RMDPs are neither necessarily harder nor easier than standard RL in terms of sample requirements.

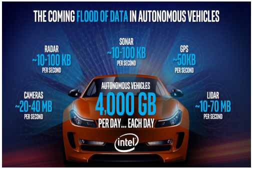
— depend heavily on the shape and size of the uncertainty set

Solving distributionally robust formulation for RL is potentially more sample-efficient

II: Provable sample efficiency in offline robust RL

- Having stored tons of history data
- Collecting new data might be expensive or time-consuming

medical records



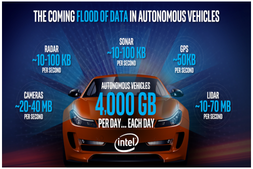
data of self-driving

clicking times of ads

Offline/Batch RL

- Having stored tons of history data
- Collecting new data might be expensive or time-consuming

medical records

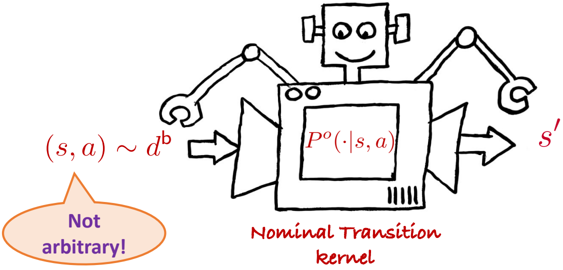


data of self-driving

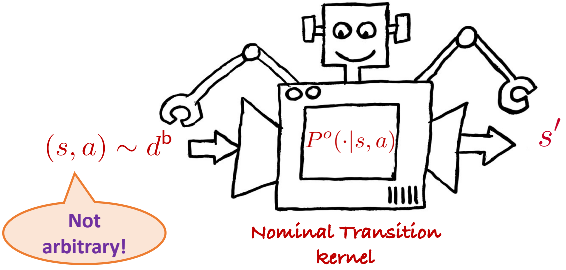
clicking times of ads

Can we design algorithms based on only history data?

Distributionally robust offline RL



Distributionally robust offline RL

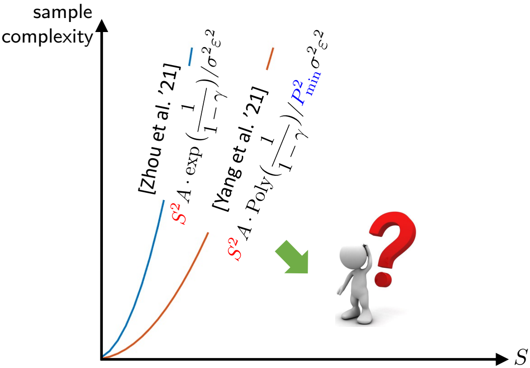


Goal of robust offline RL: given $\mathcal{D} := \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ from the nominal environment P^0, find an ϵ-optimal robust policy $\widehat{\pi}$ obeying

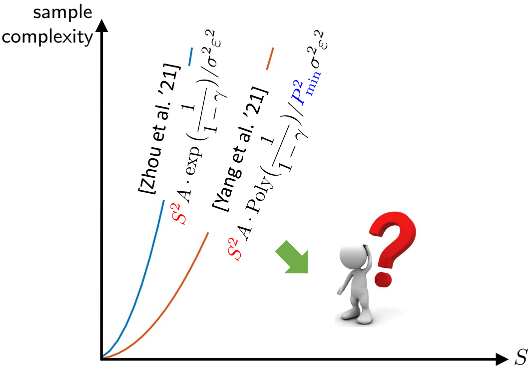
$$V^{\star,\sigma}(\rho) - V^{\widehat{\pi},\sigma}(\rho) \le \epsilon$$

— *in a sample-efficient manner*

Prior art under full coverage: KL uncertainty



Prior art under full coverage: KL uncertainty



Questions: Can we improve the sample efficiency and allow partial coverage?

How to quantify the compounded distribution shift?

Robust single-policy concentrability coefficient

$$C^{\star}_{\mathsf{rob}} := \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^{\sigma}(P^{o})} \frac{\min\{d^{\pi^\star,P}(s,a), \frac{1}{S}\}}{d^{\mathsf{b}}(s,a)}$$
$$= \left\| \frac{\textit{occupancy distribution of } (\pi^\star, \mathcal{U}^{\sigma}(P^{o}))}{\textit{occupancy distribution of } \mathcal{D}} \right\|_{\infty}$$

where $d^{\pi,P}$ is the state-action occupation density of π under P.

How to quantify the compounded distribution shift?

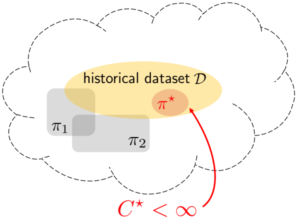
Robust single-policy concentrability coefficient

$$C^{\star}_{\mathsf{rob}} := \max_{(s,a,P)\in\mathcal{S}\times\mathcal{A}\times\mathcal{U}^{\sigma}(P^o)} \frac{\min\{d^{\pi^\star,P}(s,a), \frac{1}{S}\}}{d^{\mathsf{b}}(s,a)}$$

$$= \left\| \frac{\textit{occupancy distribution of } (\pi^\star, \mathcal{U}^{\sigma}(P^o))}{\textit{occupancy distribution of } \mathcal{D}} \right\|_\infty$$

where $d^{\pi,P}$ is the state-action occupation density of π under P.

- captures distributional shift due to behavior policy and environment.
- $C^{\star}_{\mathsf{rob}} \le A$ under full coverage.



DRVI with pessimism

Distributionally robust value iteration (DRVI) with LCB:

$$\widehat{Q}(s,a) \leftarrow \max\Big\{r(s,a) + \gamma \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^{o}_{s,a})} \mathcal{P}\widehat{V} - \underbrace{b(s,a;\widehat{V})}_{\text{uncertainty penalty}}, 0\Big\},$$

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

Key innovation: design the penalty term to capture the uncertainty of both model and the data in robust RL:

$$\underbrace{\left| \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(P^{o}_{s,a})} \mathcal{P}\widehat{V} - \inf_{\mathcal{P} \in \mathcal{U}^{\sigma}(\widehat{P}^{o}_{s,a})} \mathcal{P}\widehat{V} \right|}_{\text{No closed form w.r.t. } P^{o}_{s,a} - \widehat{P}^{o}_{s,a} \text{ due to } \mathcal{U}^{\sigma}(\cdot)}$$

Sample complexity of DRVI-LCB

Theorem (Shi and Chi '22)

For any uncertainty level $\sigma > 0$ *and small enough* ϵ*, DRVI-LCB outputs an* ϵ*-optimal policy with high prob., with sample complexity at most*

$$\widetilde{O}\left(\frac{SC^{\star}_{\mathsf{rob}}}{P^{\star}_{\mathsf{min}}(1-\gamma)^4\sigma^2\epsilon^2}\right),$$

where $P^{\star}_{\mathsf{min}}$ *is the smallest positive state transition probability of the nominal kernel* visited by the optimal robust policy $\pi^{\star}$.

Sample complexity of DRVI-LCB

Theorem (Shi and Chi '22)

For any uncertainty level $\sigma > 0$ *and small enough* ϵ*, DRVI-LCB outputs an* ϵ*-optimal policy with high prob., with sample complexity at most*

$$\widetilde{O}\left(\frac{SC^{\star}_{\mathsf{rob}}}{P^{\star}_{\mathsf{min}}(1-\gamma)^4\sigma^2\epsilon^2}\right),$$

where $P^{\star}_{\mathsf{min}}$ *is the smallest positive state transition probability of the nominal kernel* visited by the optimal robust policy $\pi^{\star}$.

- scales linearly with respect to S
- reflects the impact of distribution shift of offline dataset ($C^{\star}_{\mathsf{rob}}$) and also model shift level (σ)

Minimax lower bound

Theorem (Shi and Chi '22)

Suppose that $\frac{1}{1-\gamma} \geq e^8$, $S \geq \log\left(\frac{1}{1-\gamma}\right)$, $C^\star_{\mathsf{rob}} \geq 8/S$, $\sigma \asymp \log \frac{1}{1-\gamma}$ *and* $\epsilon \lesssim \frac{1}{(1-\gamma)\log\frac{1}{1-\gamma}}$, *there exists some MDP and batch dataset such that no algorithm succeeds if the sample size is below*

$$\widetilde{\Omega}\left(\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^2\sigma^2\epsilon^2}\right).$$

Minimax lower bound

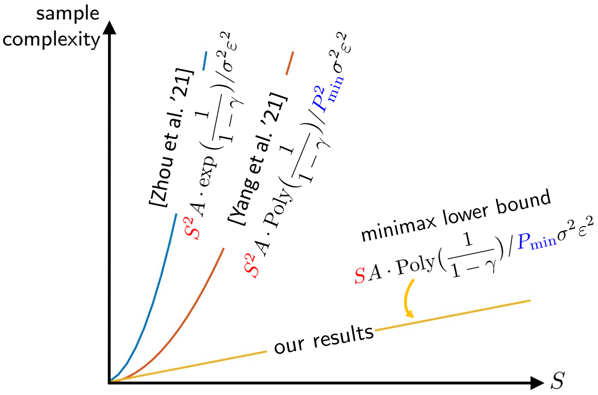
Theorem (Shi and Chi '22)

Suppose that $\frac{1}{1-\gamma} \geq e^8$, $S \geq \log\left(\frac{1}{1-\gamma}\right)$, $C^\star_{\mathsf{rob}} \geq 8/S$, $\sigma \asymp \log\frac{1}{1-\gamma}$ *and* $\epsilon \lesssim \frac{1}{(1-\gamma)\log\frac{1}{1-\gamma}}$, *there exists some MDP and batch dataset such that no algorithm succeeds if the sample size is below*

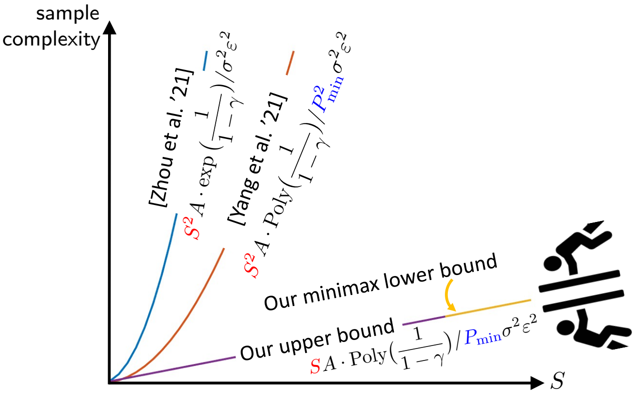
$$\widetilde{\Omega}\left(\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^2\sigma^2\epsilon^2}\right).$$

- the first lower bound for robust MDP with KL divergence
- Establishes the near minimax-optimality of DRVI-LCB up to factors of $1/(1-\gamma)$

Compare to prior art under full coverage



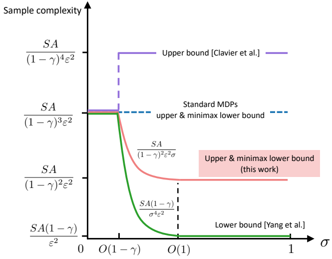
Compare to prior art under full coverage



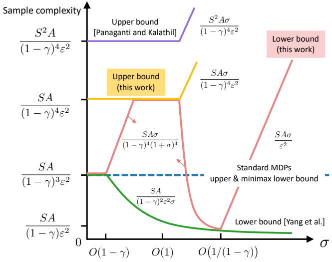
We develop the first minimax lower bound on this.
Our DRVI-LCB method is near minimax-optimal!

Concluding remarks

Statistical implications of distributionally robustness



TV uncertainty

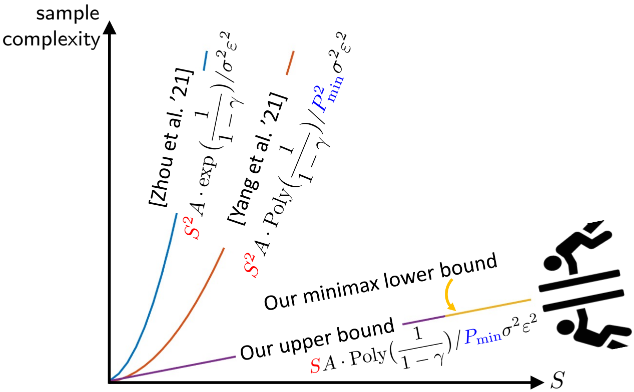


χ^2 uncertainty

RMDPs are neither necessarily harder nor easier than standard RL in terms of sample requirements.

— depend heavily on the shape and size of the uncertainty set

Near-optimal robust offline RL



We develop the first minimax lower bound on this.
Our DRVI-LCB method is near minimax-optimal!

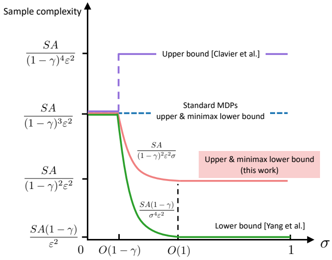
References

- L. Shi, G. Li, Y. Wei, Y. Chen, M. Geist, and Y. Chi, "The curious price of distributional robustness in reinforcement learning with a generative model," *arXiv preprint arXiv:2305.16589*, 2023.
- L. Shi, R. Dadashi, Y. Chi, P. S. Castro, and M. Geist, "Offline reinforcement learning with on-policy Q-function regularization," *European Conference on Machine Learning*, 2023.
- G. Li, L. Shi, Y. Chen, and Y. Chi, "Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning," *Information and Inference: A Journal of the IMA*, vol. 12, no. 2, pp. 969–1043, 2023.
- W. Ding*, L. Shi*, Y. Chi, and D. Zhao, "Seeing is not believing: Robust reinforcement learning against spurious correlation," *In submission. A short version at ICML Workshop on Spurious Correlations, Invariance and Stability*, 2023.
- L. Shi and Y. Chi, "Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity," *arXiv preprint arXiv:2208.05767*, 2022.

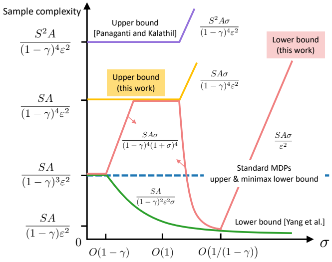
References

- Y. Wang, M. Xu, L. Shi, and Y. Chi, "A trajectory is worth three sentences: Multimodal transformer for offline reinforcement learning," *The Conference on Uncertainty in Artificial Intelligence*, 2023.
- L. Shi, G. Li, Y. Wei, Y. Chen, and Y. Chi, "Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity," in *International Conference on Machine Learning*. PMLR, 2022, pp. 19 967–20 025.
- G. Li, L. Shi, Y. Chen, Y. Chi, and Y. Wei, "Settling the sample complexity of model-based offline reinforcement learning," *arXiv preprint arXiv:2204.05275*, 2022.
- P. Huang, M. Xu, J. Zhu, L. Shi, F. Fang, and D. Zhao, "Curriculum reinforcement learning using optimal transport via gradual domain adaptation," *Advances in Neural Information Processing Systems*, 2022.

Thank you!



TV uncertainty



χ^2 uncertainty

RMDPs are neither necessarily harder nor easier than standard RL in terms of sample requirements.

— depend heavily on the shape and size of the uncertainty set