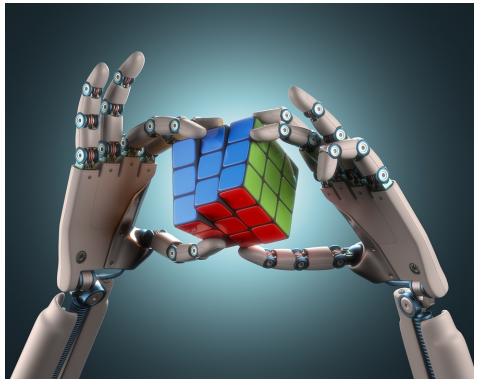
Taming the Sim-to-Real Gap in Reinforcement Learning

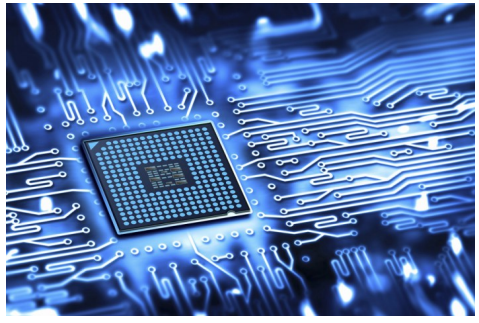
Yuejie Chi

Carnegie Mellon University

Fields Institute
April 2024

Recent successes in RL





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

Sample efficiency

Collecting data samples might be expensive or time-consuming due to the enormous state and action space

clinical trials

autonomous driving

online ads

Sample efficiency

Collecting data samples might be expensive or time-consuming due to the enormous state and action space

clinical trials

autonomous driving

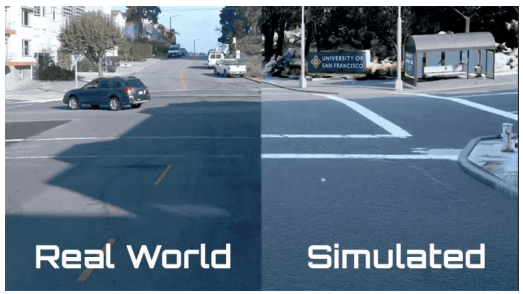
online ads

Calls for design of sample-efficient RL algorithms!

Robustness to sim-to-real gap

The experienced environment can be perturbed from the training one due to sim-to-real gaps, noise, and generalization.

Uncertainty



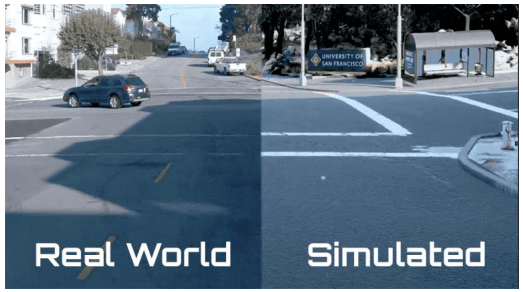
Sim-to-real gaps

Generalization

Robustness to sim-to-real gap

The experienced environment can be perturbed from the training one due to sim-to-real gaps, noise, and generalization.

Uncertainty



Sim-to-real gaps

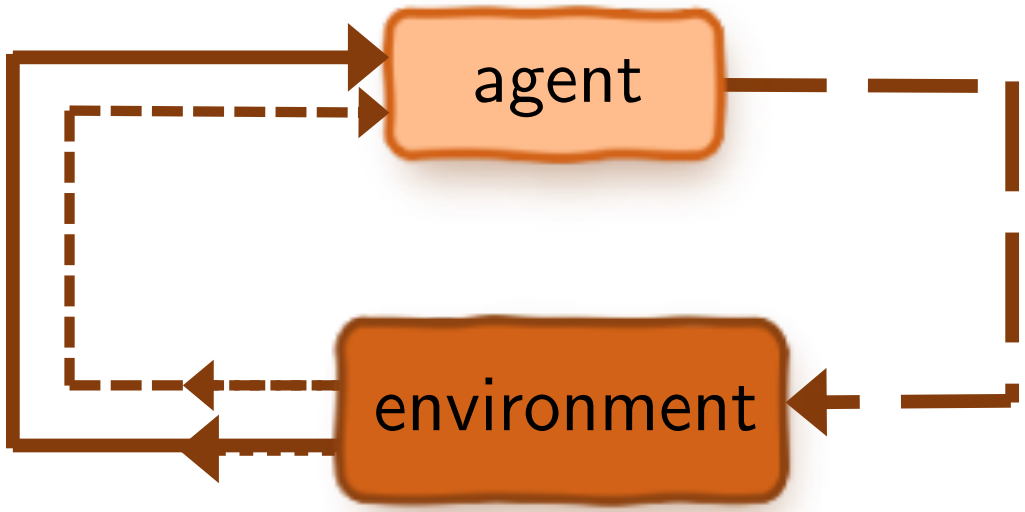
Generalization

Calls for robust RL algorithms!

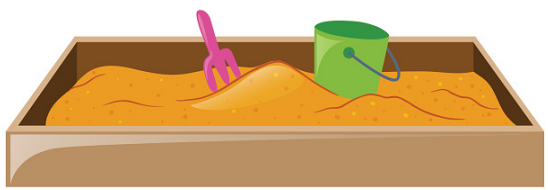
Statistical thinking in RL: non-asymptotic analysis

Non-asymptotic analyses are key to understand statistical efficiency in modern RL.

Recent advances in statistical RL

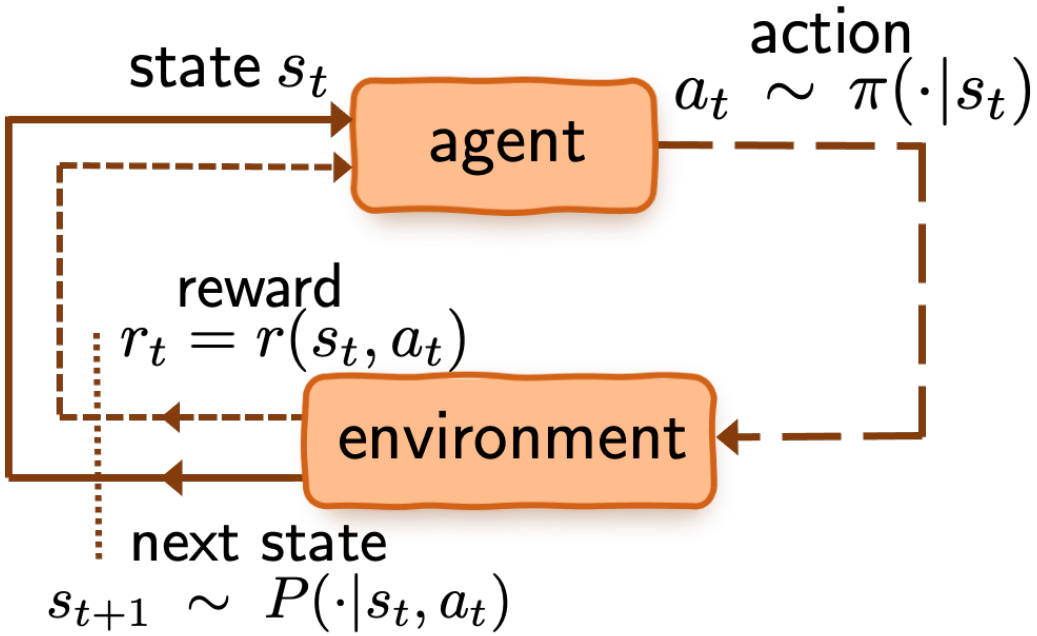


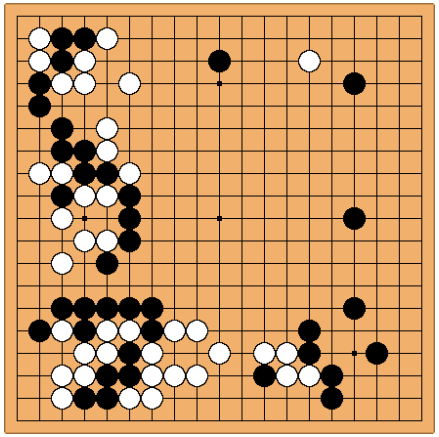
The playground: Markov decision processes



Backgrounds: Markov decision processes

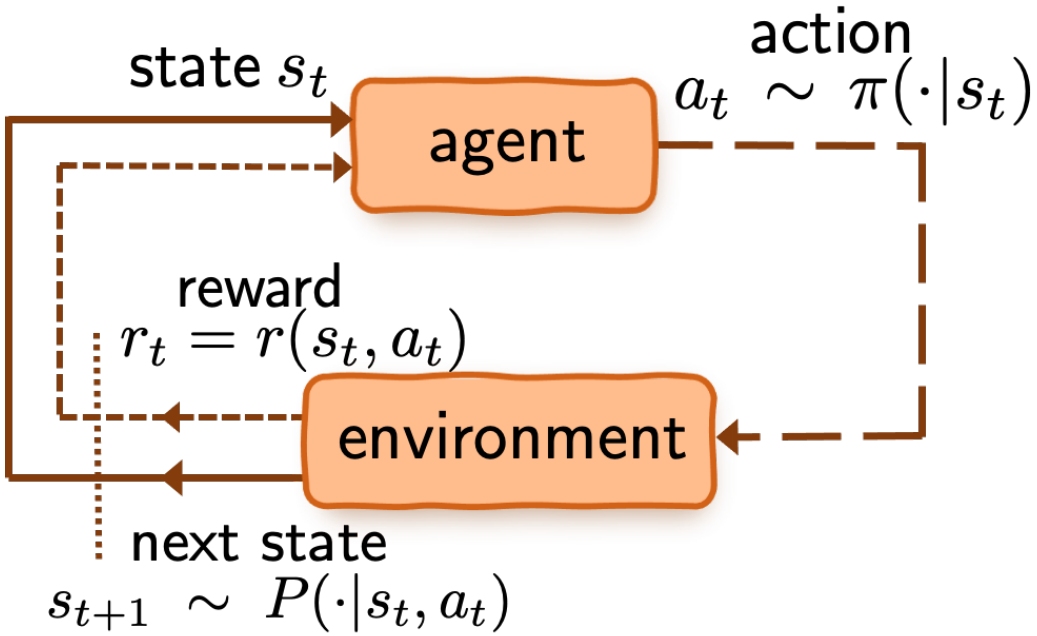
Markov decision process (MDP)

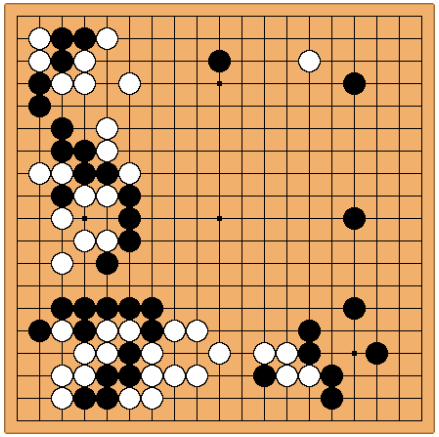




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

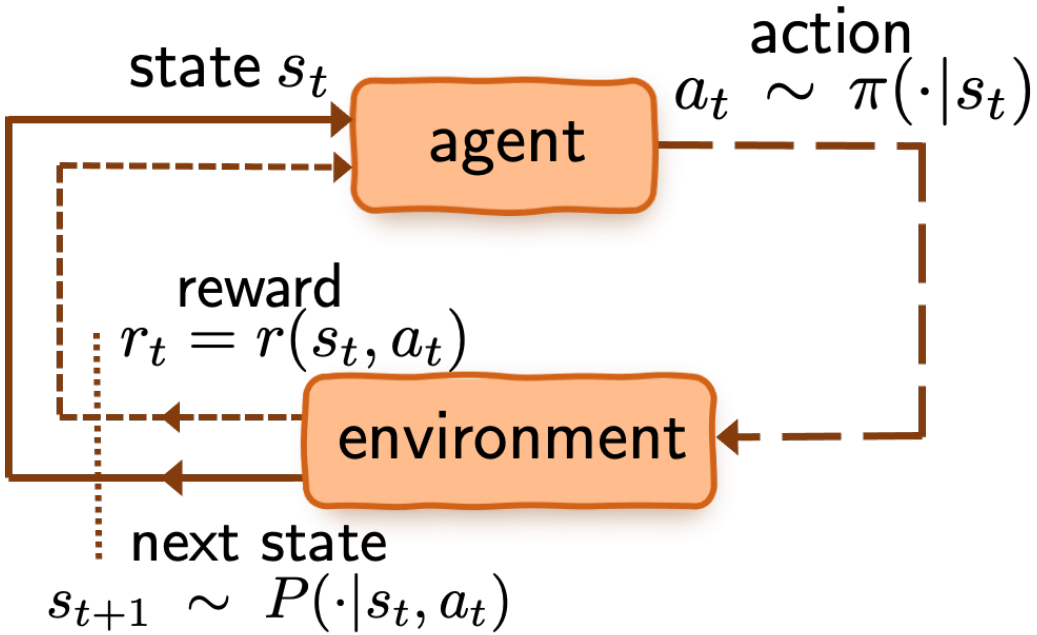
Markov decision process (MDP)

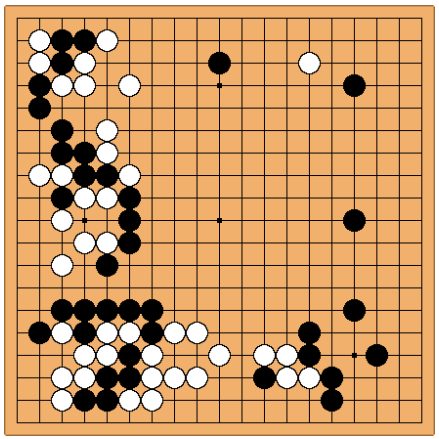




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward

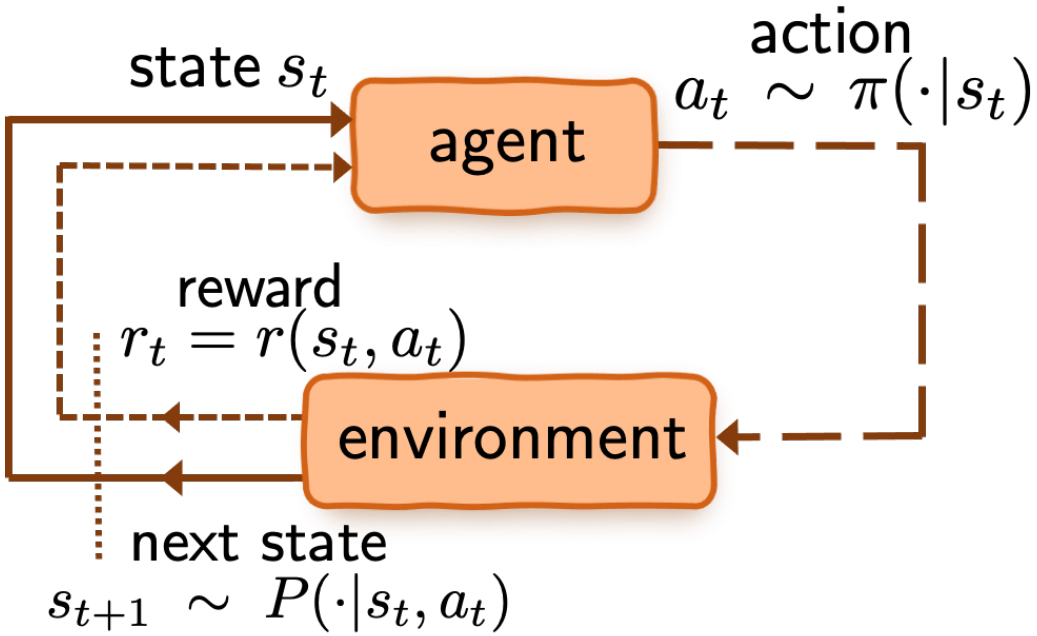
Markov decision process (MDP)

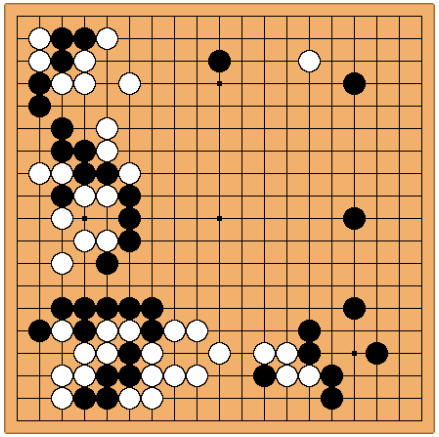




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

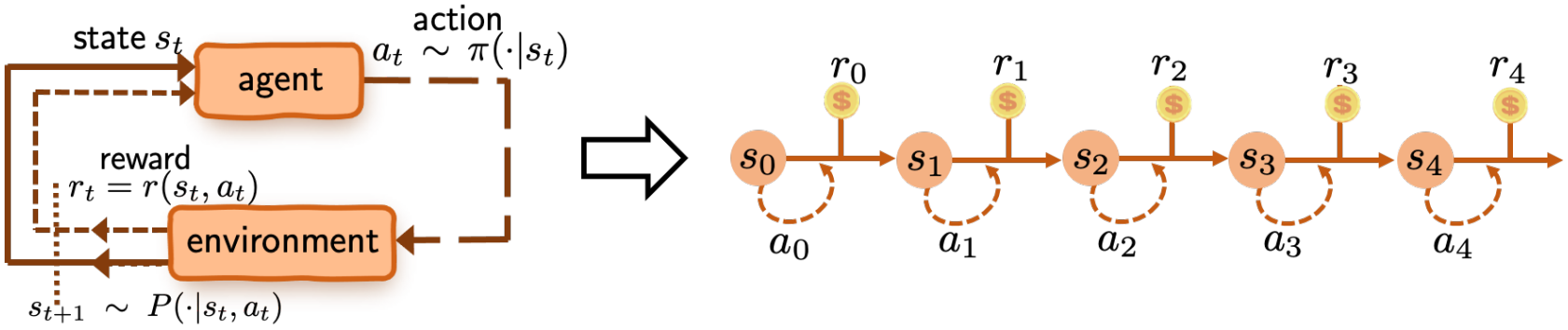
Markov decision process (MDP)





- $\mathcal{S}$: 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



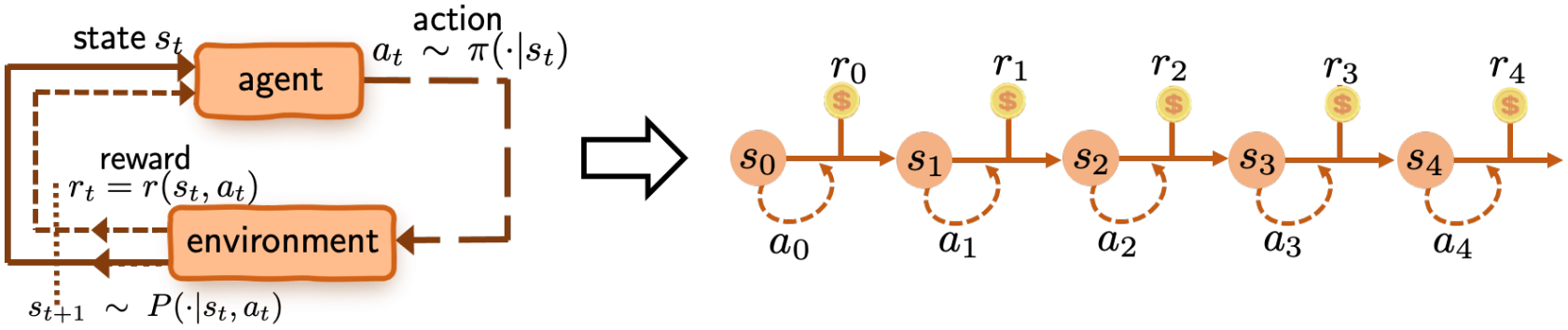
Value function of policy π:

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

Q-function of policy π:

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

Value function



Value function of policy π:

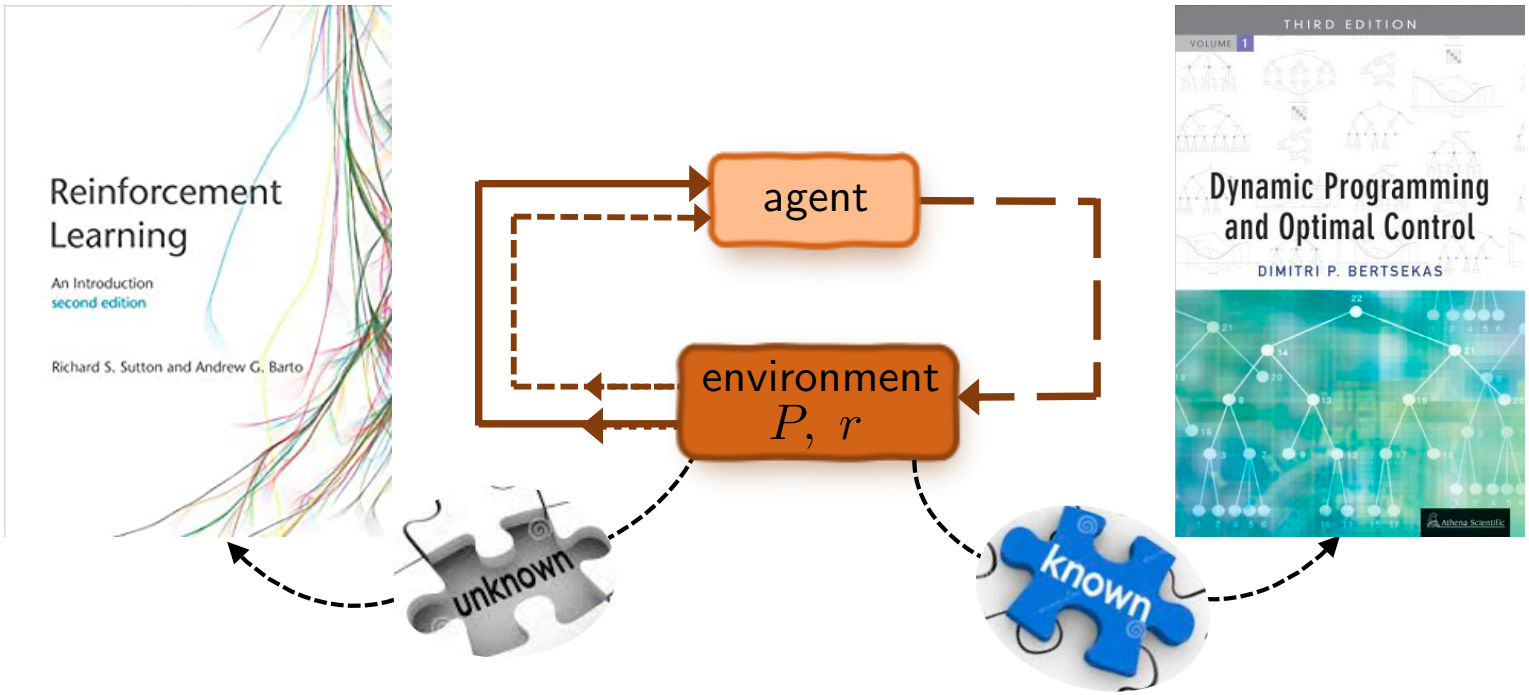
$$\forall s \in \mathcal{S}: \qquad V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s\right]$$

Q-function of policy π:

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

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

Searching for the optimal policy



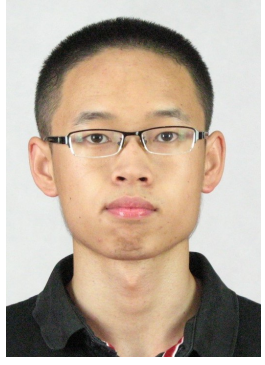
Goal: find the optimal policy $\pi^\star$ that maximize $V^\pi(s)$

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

RL meets distributional robustness: towards minimax-optimal sample complexity

Laixi Shi

Caltech



Gen Li

CUHK

Yuxin Chen

UPenn

Yuting Wei

UPenn

Matthieu Geist

Cohere

"The Curious Price of Distributional Robustness in Reinforcement Learning with a Generative Model," arXiv:2305.16589. Short version at NeurIPS 2023.

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 $\neq$

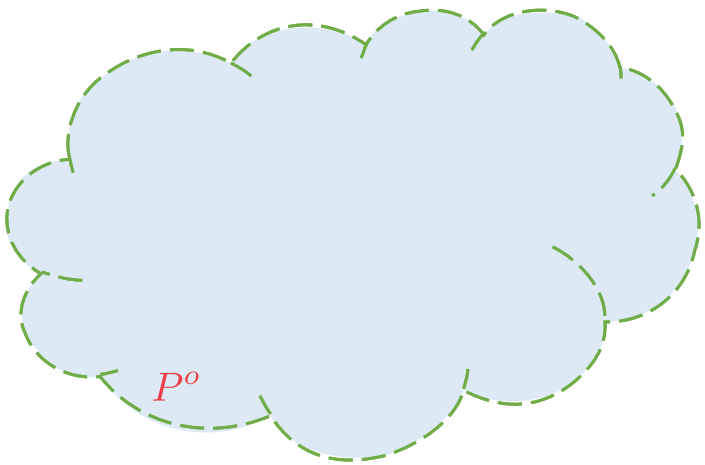
Test environment

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:

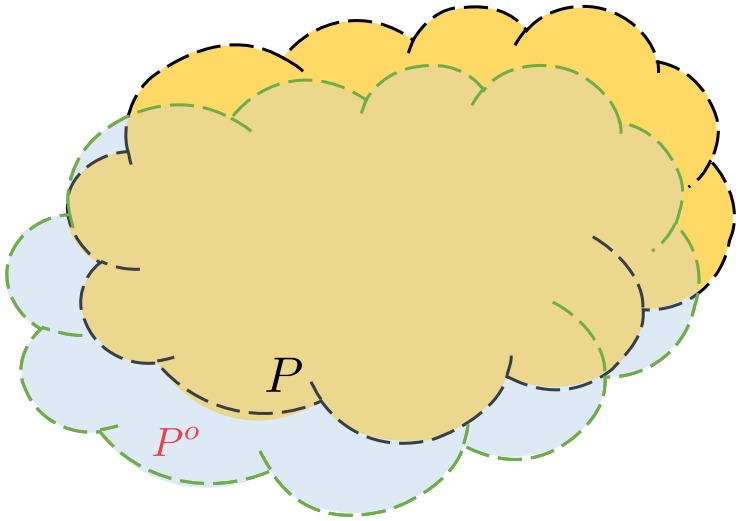
$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$

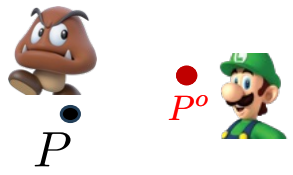


Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

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

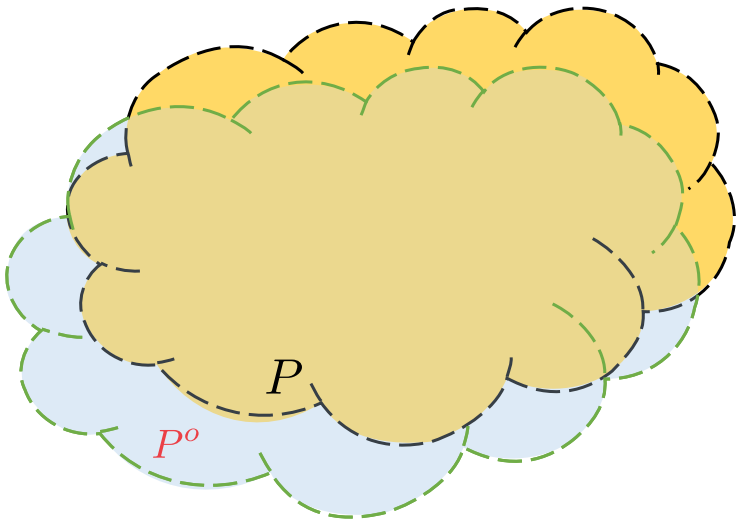


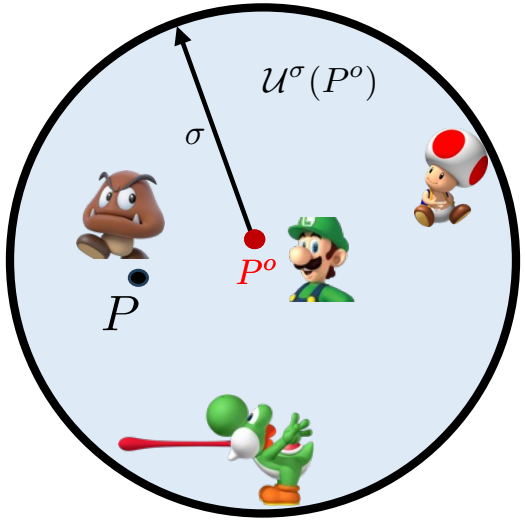


Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

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

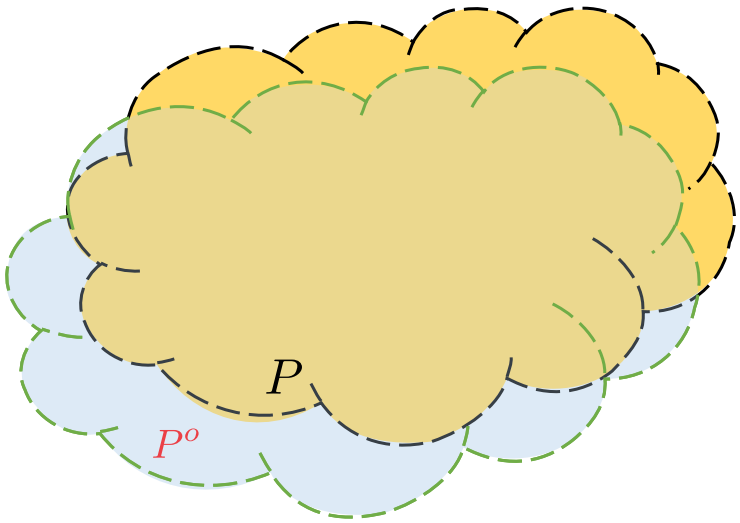


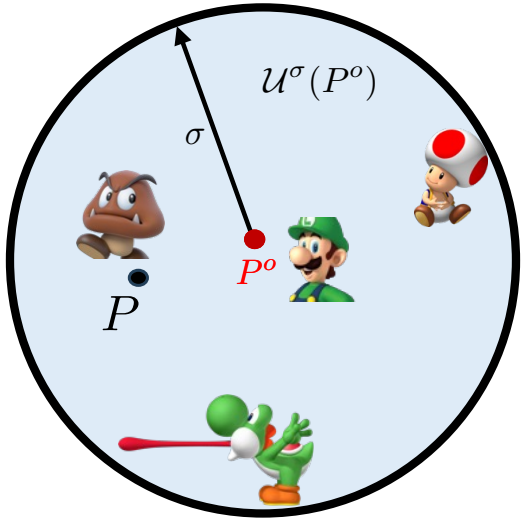


Modeling environment uncertainty

Uncertainty set of the nominal transition kernel P^o:

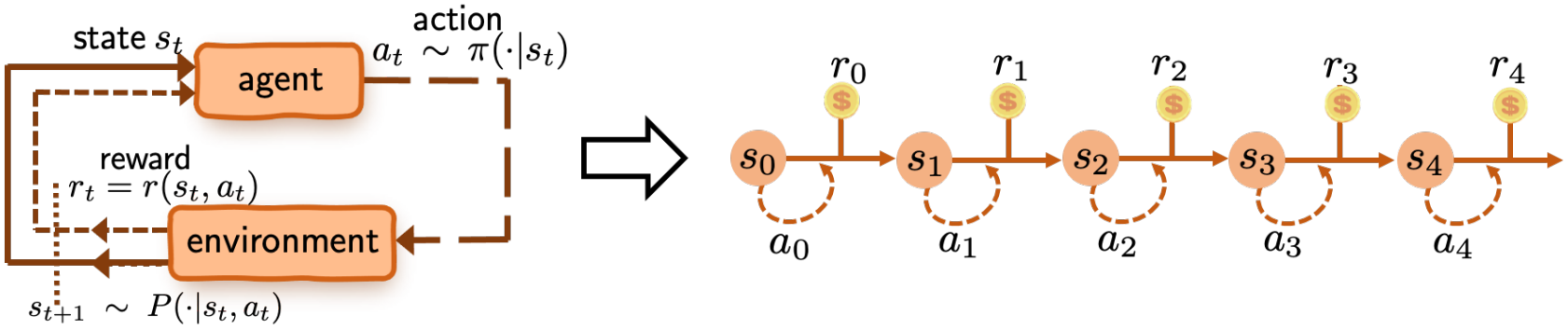
$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$





- 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

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

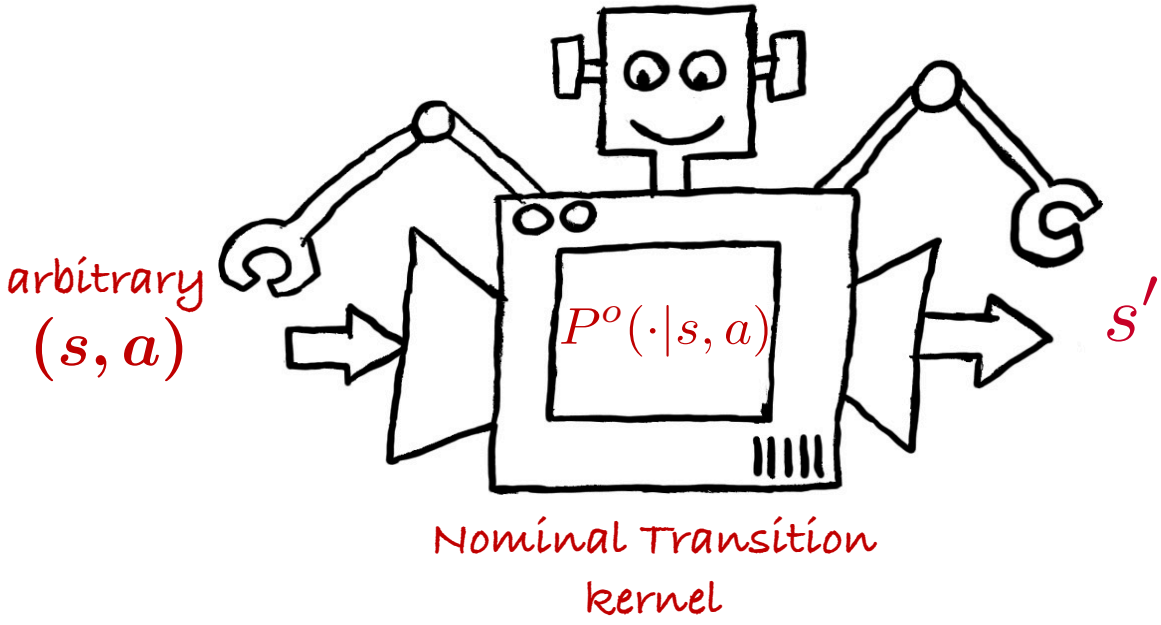
$$V^{\star,\sigma}(s) = \max_a Q^{\star,\sigma}(s,a)$$

Distributionally robust value iteration (DRVI):

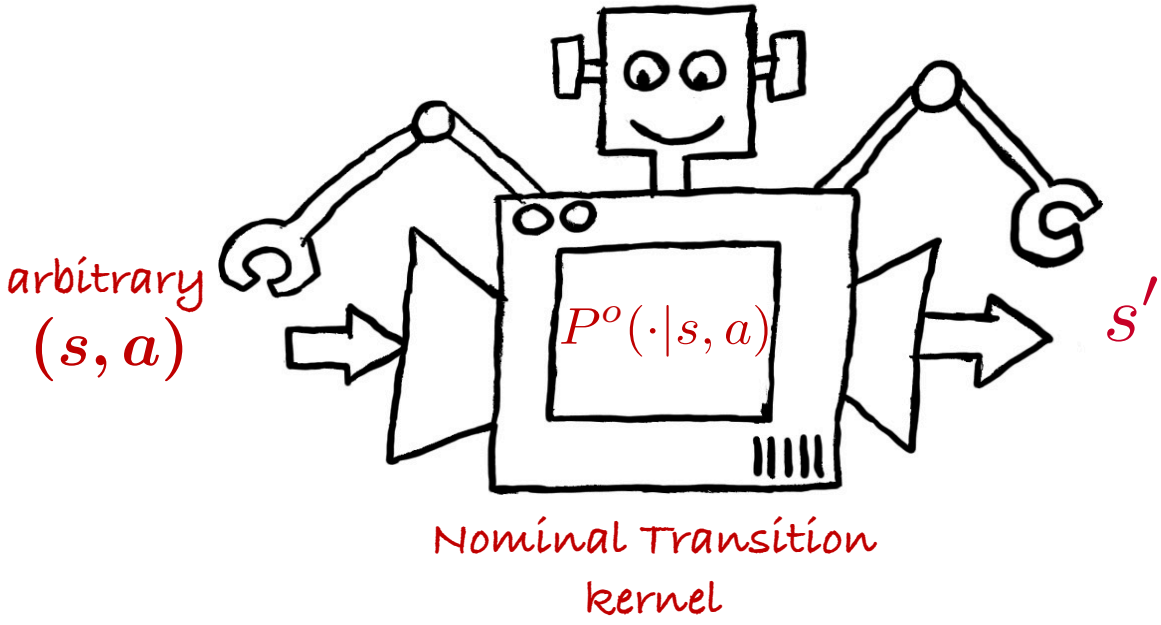
$$Q(s,a) \leftarrow r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma\left(P^o_{s,a}\right)} \langle P_{s,a}, V \rangle,$$

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

Learning distributionally robust MDPs



Learning distributionally robust MDPs



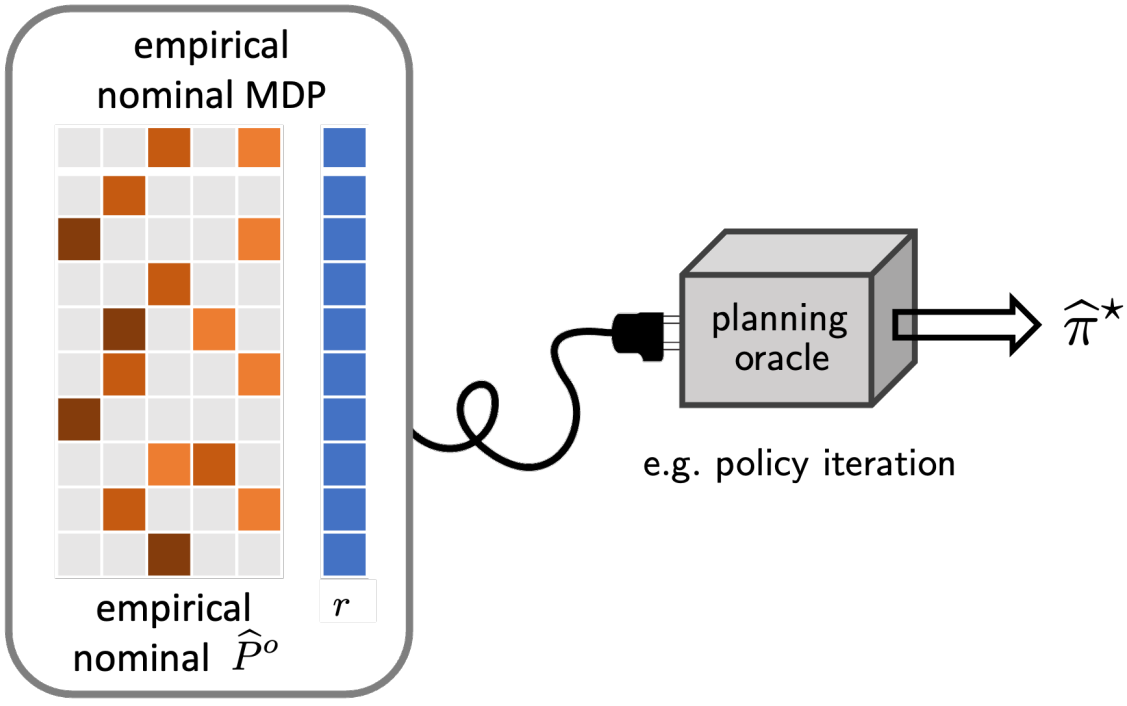
Goal of robust RL: given $\mathcal{D} := \{(s_i, a_i, s_i')\}_{i=1}^N$ from the *nominal* environment P^0, find an ϵ-optimal robust policy $\widehat{\pi}$ obeying

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

— *in a sample-efficient manner*

Model-based RL: empirical MDP + planning

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



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

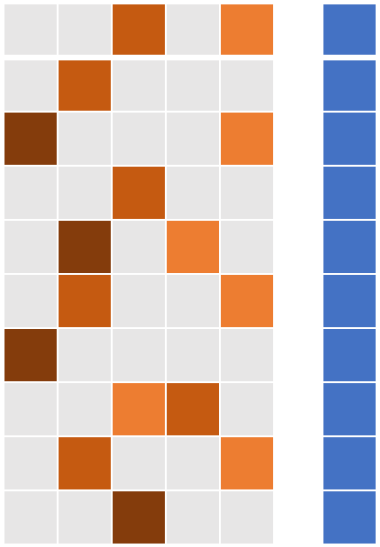
$$\widehat{Q}(s,a) \leftarrow r(s,a) + \gamma \max_{\lambda \in \left[\min_s \widehat{V}(s), \max_s \widehat{V}(s)\right]} \left\{ \widehat{P}^o_{s,a}\left[\widehat{V}\right]_\lambda - \sigma\left(\lambda - \min_{s'} \left[\widehat{V}\right]_\lambda (s')\right)\right\},$$

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

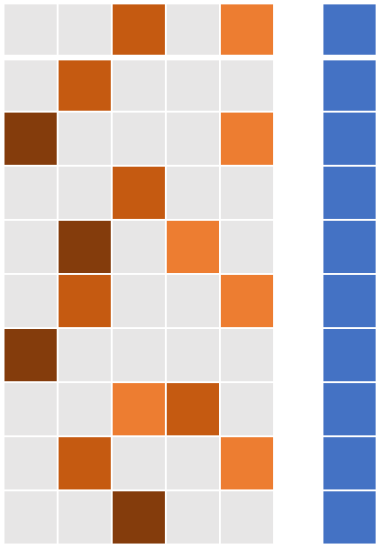


empirical MDP

Learn the optimal policy of the nominal MDP?

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

A curious question



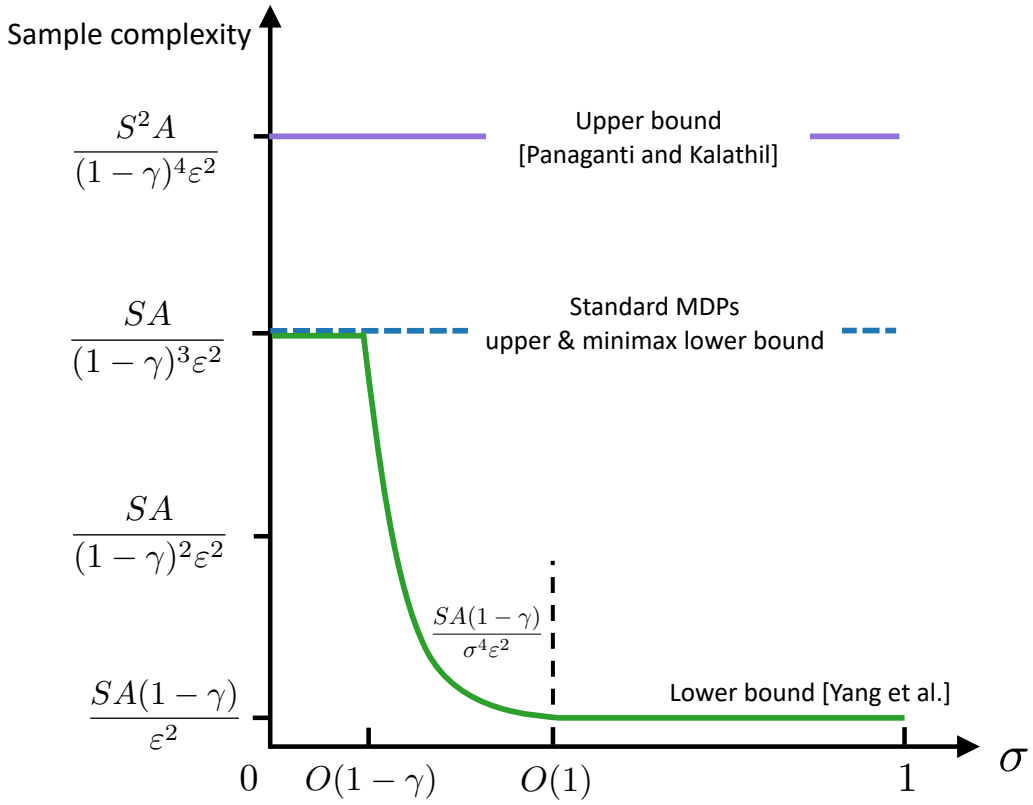
empirical MDP

Learn the optimal policy of the nominal MDP?

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

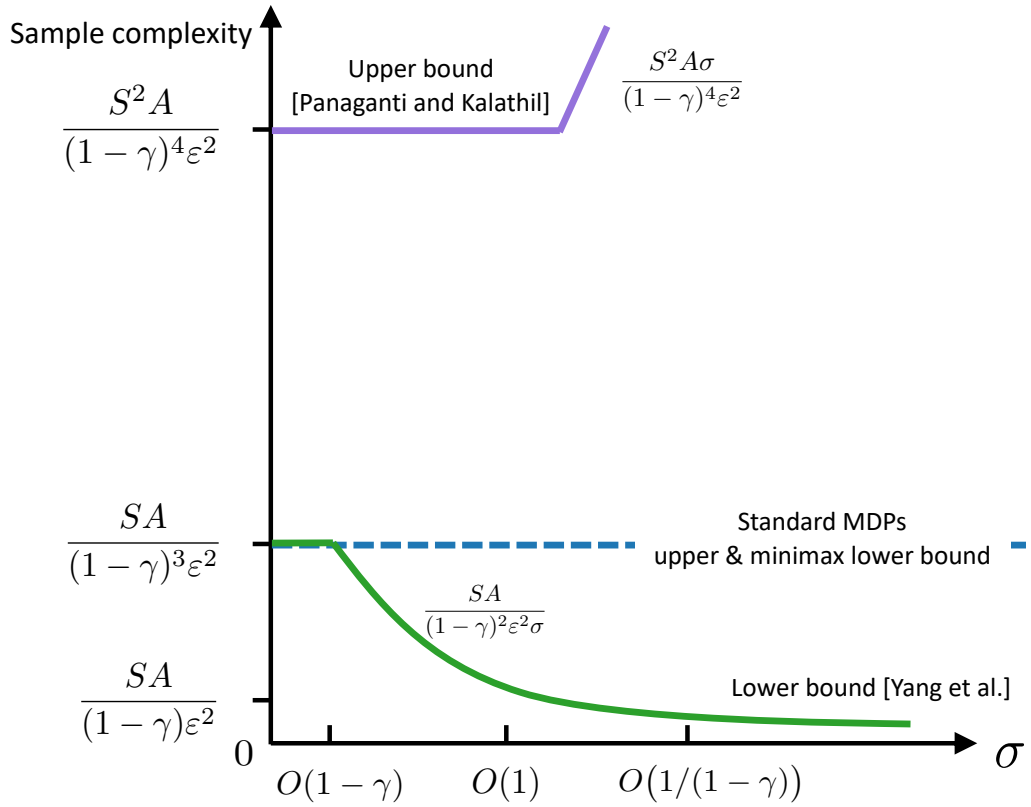
Robustness-statistical trade-off? Is there a statistical premium that one needs to pay in quest of additional robustness?

Prior art: TV uncertainty



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

Prior art: χ^2 uncertainty



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

Our theorem 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*

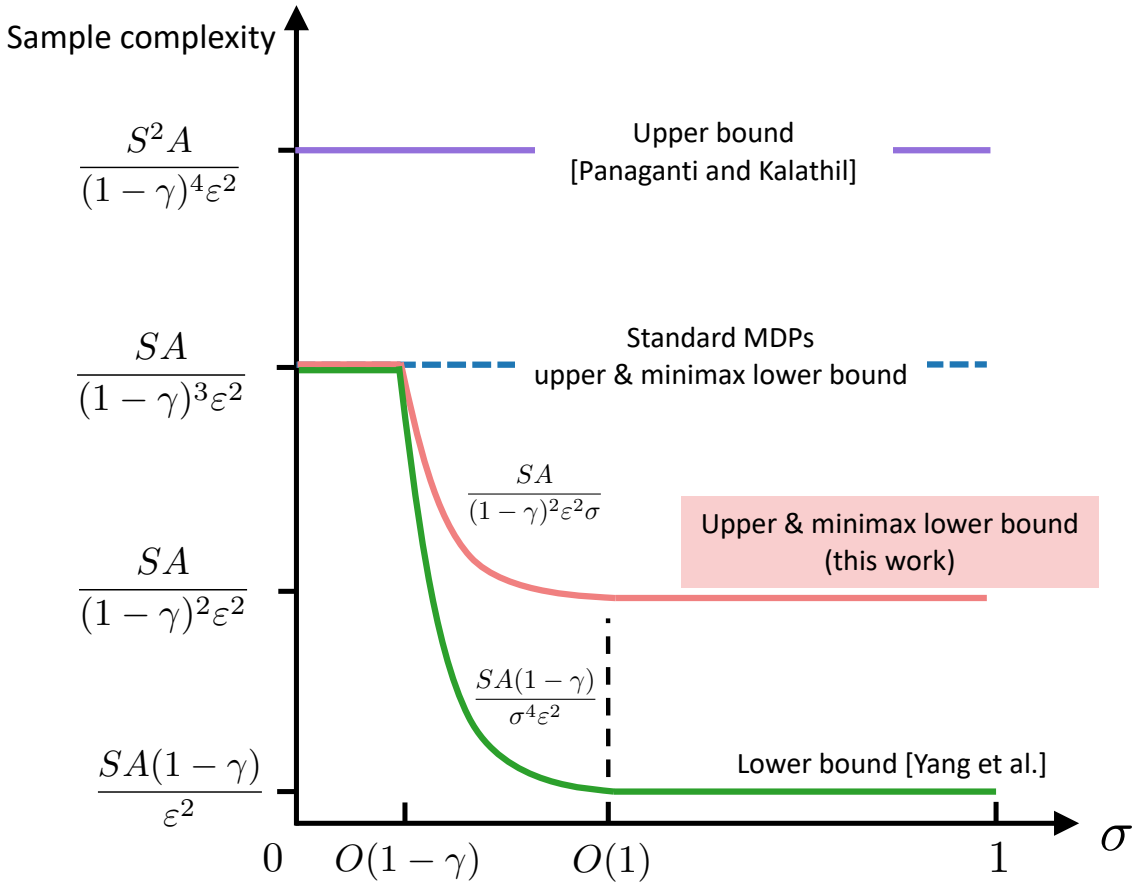
$$\widetilde{O}\left(\frac{SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\}\epsilon^2}\right)$$

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

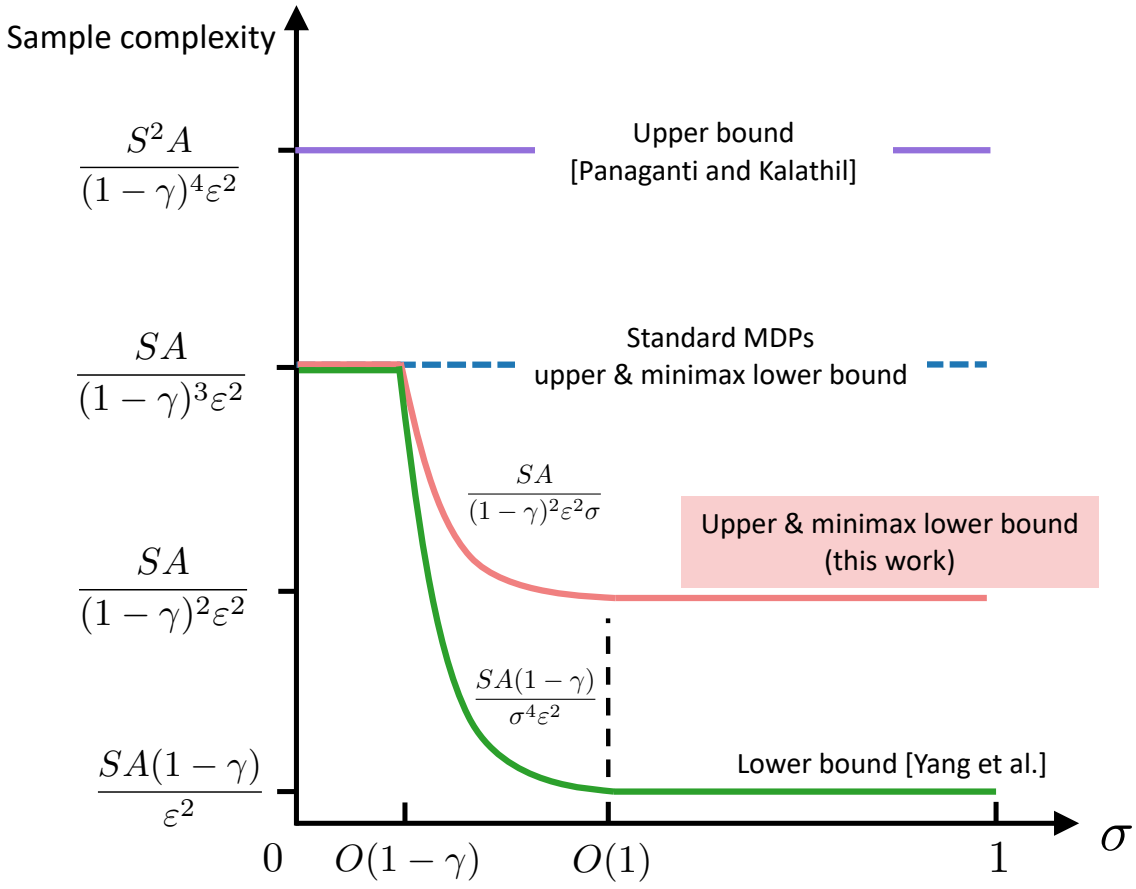
$$\widetilde{\Omega}\left(\frac{SA}{(1-\gamma)^2 \max\{1-\gamma, \sigma\}\epsilon^2}\right).$$

- 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 theorem 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 theorem 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

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

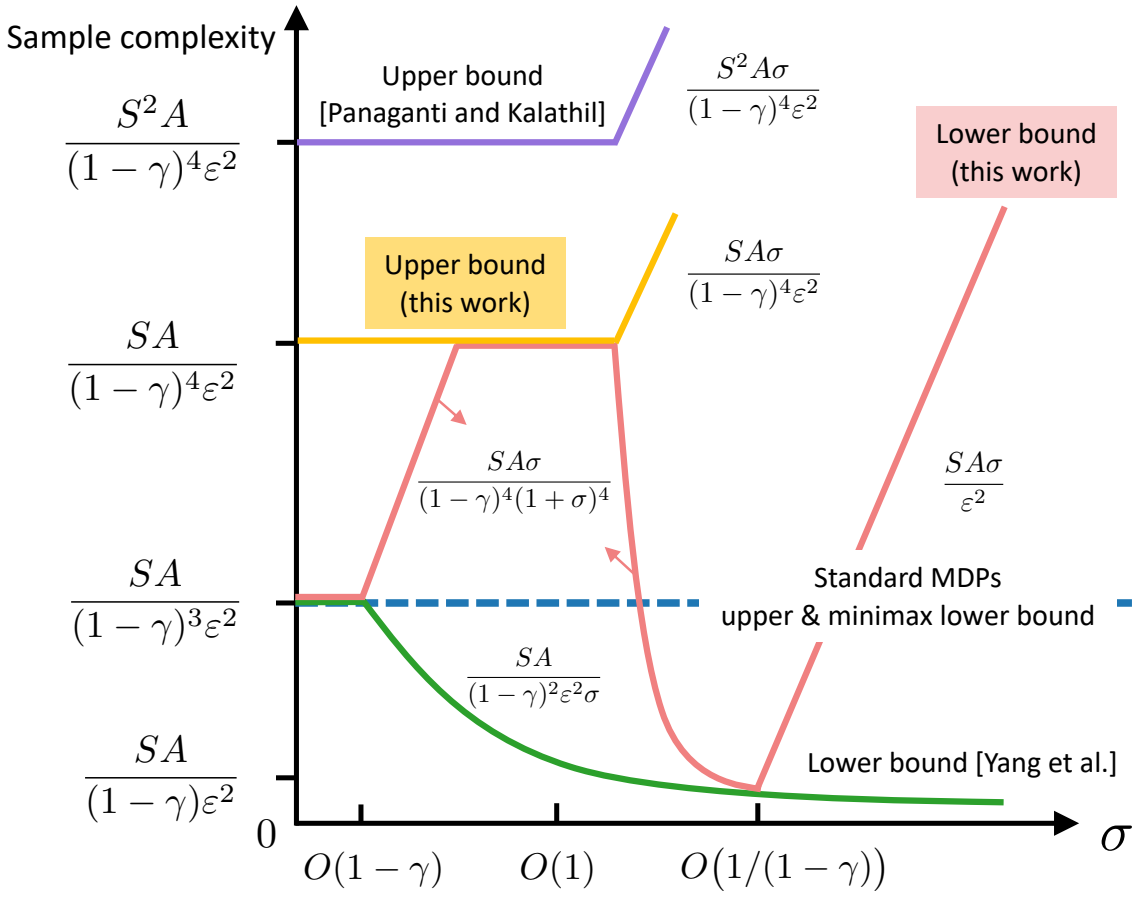
ignoring logarithmic factors.

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

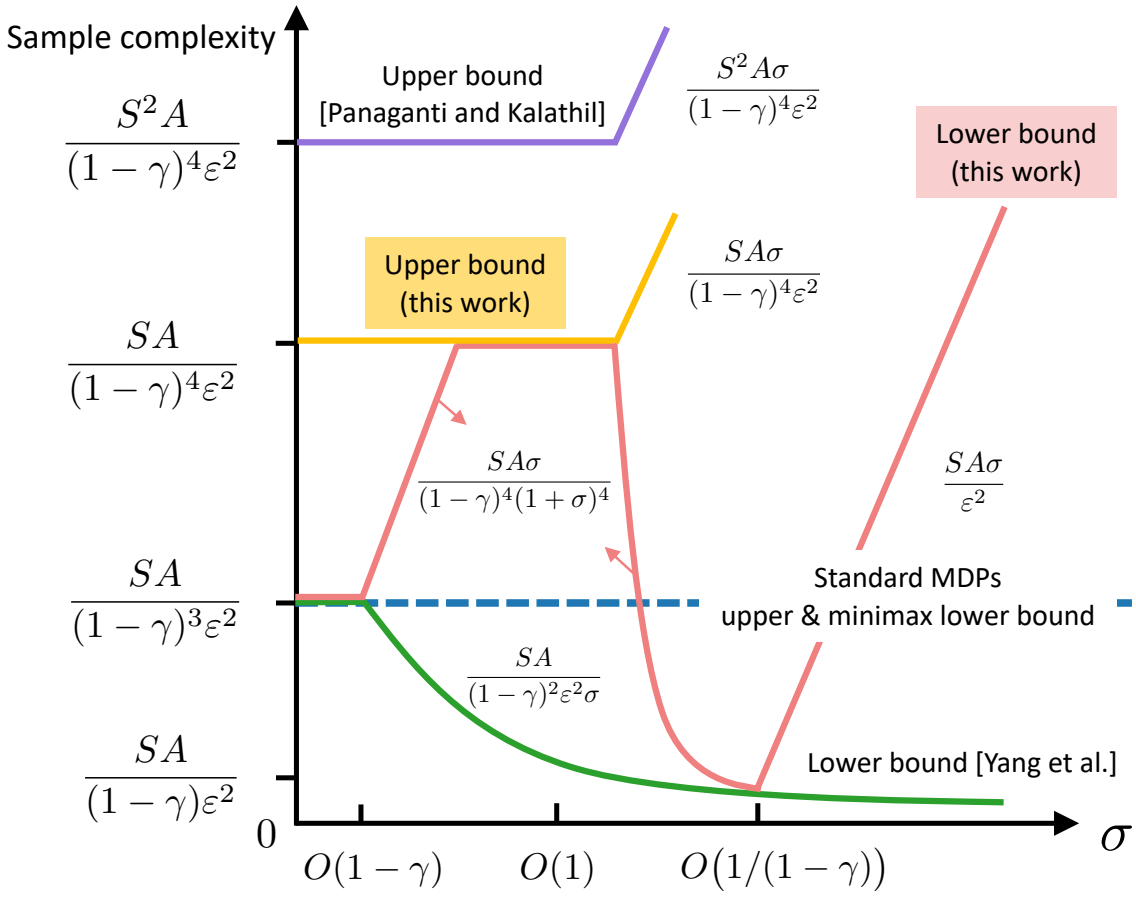
In addition, no algorithm succeeds when the sample size is below

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

When the uncertainty set is χ^2 divergence

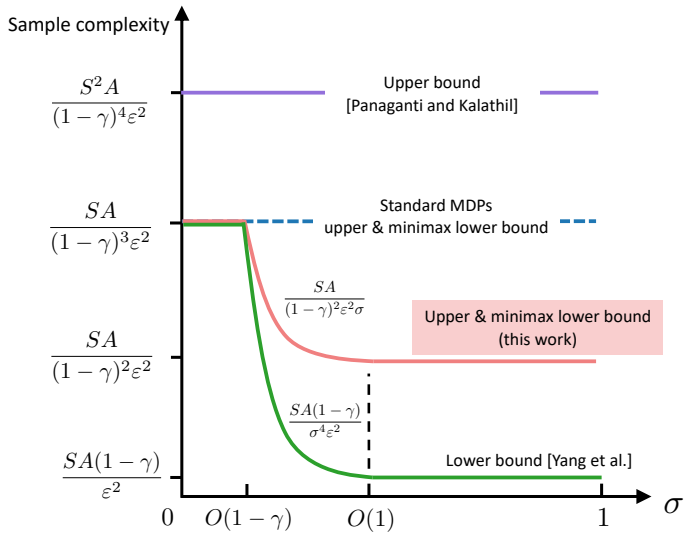


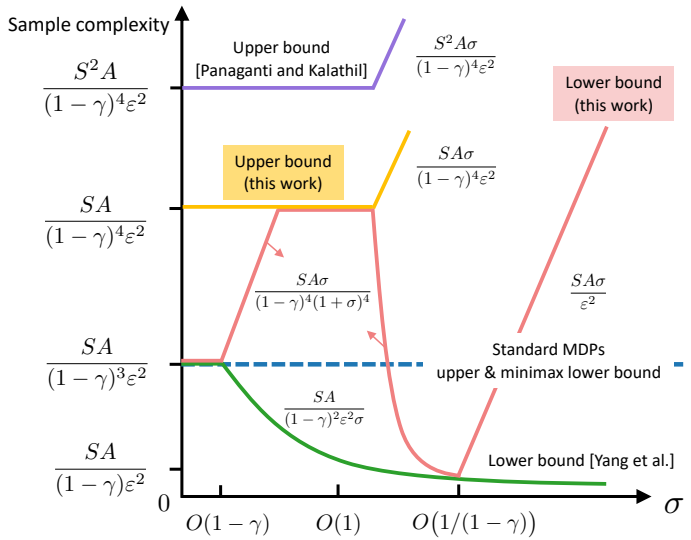
When the uncertainty set is χ^2 divergence



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

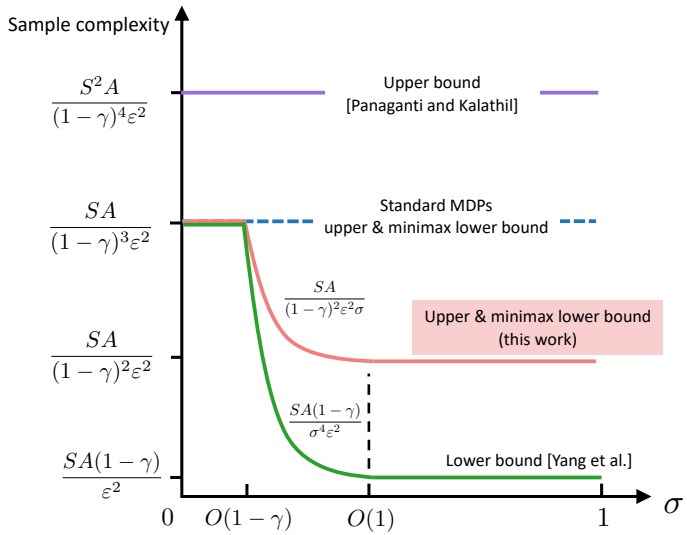
Summary

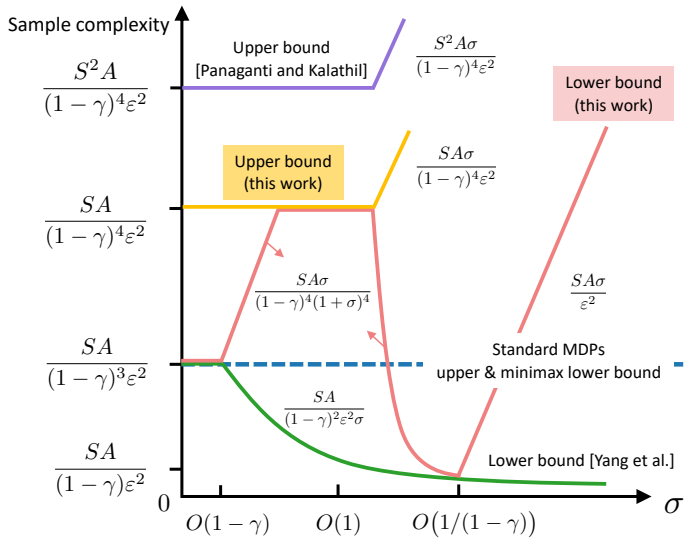




The statistical price of robustness varies: the choice of uncertainty sets matters.

Summary





The statistical price of robustness varies: the choice of uncertainty sets matters.

Future work:

- Function approximation and multi-agent settings.

Distributional robustness meets offline RL

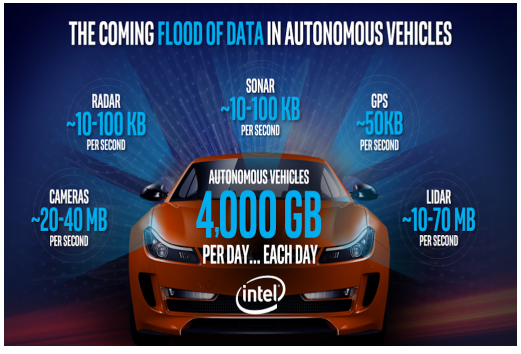
Laixi Shi

CMU→Caltech

Offline/Batch RL

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

medical records



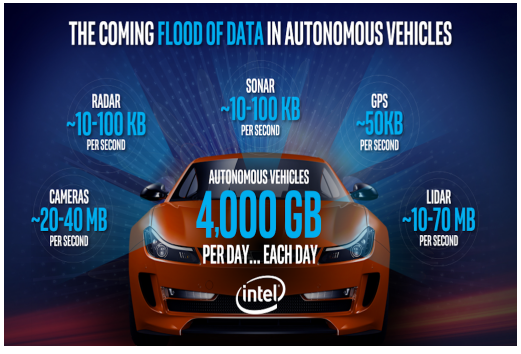
data of self-driving

clicking times of ads

Offline/Batch RL

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

medical records

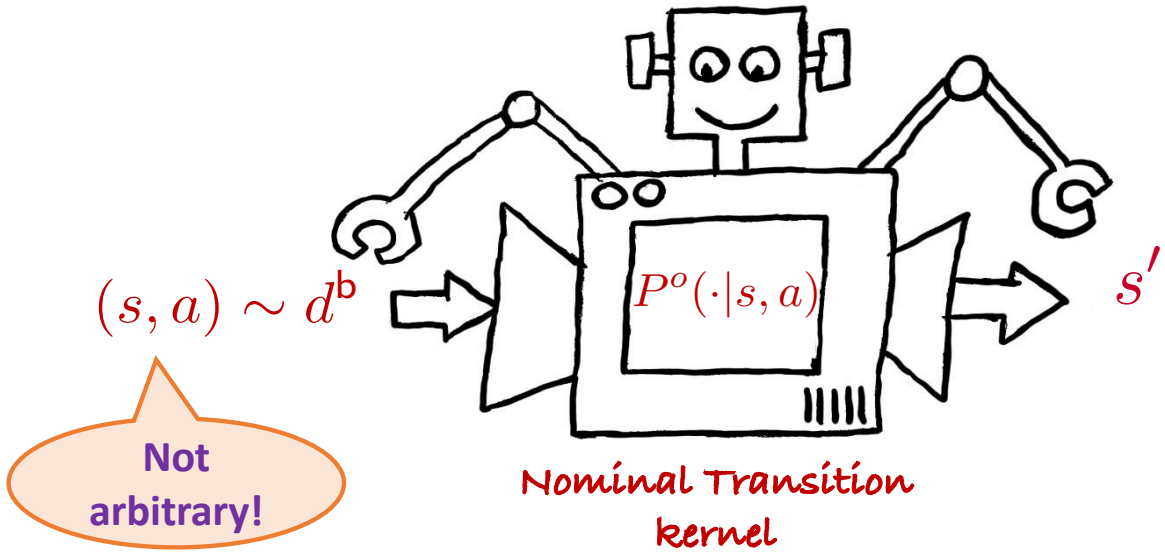


data of self-driving

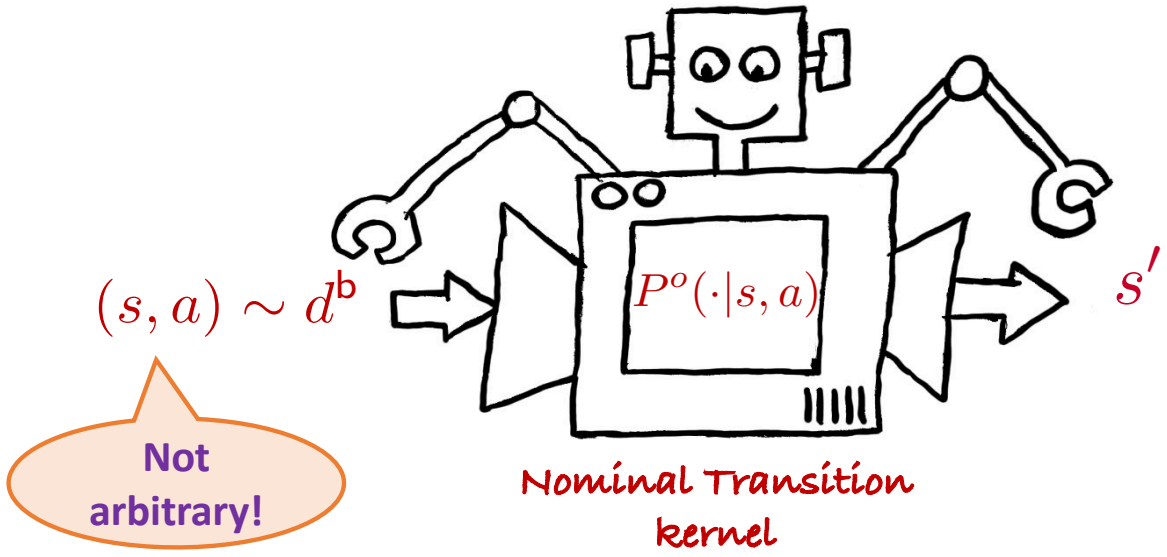
clicking times of ads

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

Distributionally robust offline RL



Distributionally robust offline RL



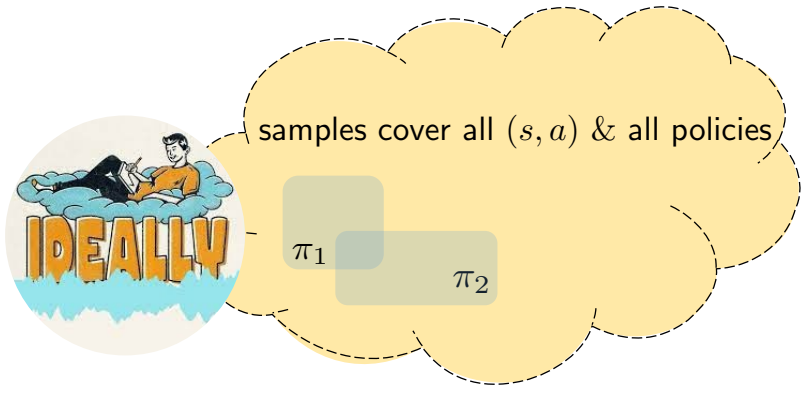
Goal of robust offline RL: given $\mathcal{D} := \{(s_i, a_i, s_i')\}_{i=1}^N$ from the *nominal* environment P^0, find an ϵ-optimal robust policy $\widehat{\pi}$ obeying

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

— *in a sample-efficient manner*

Challenges of offline RL

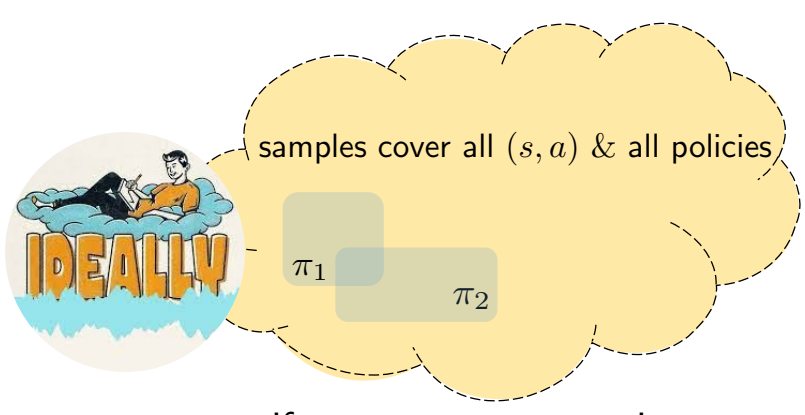
Partial coverage of state-action space:



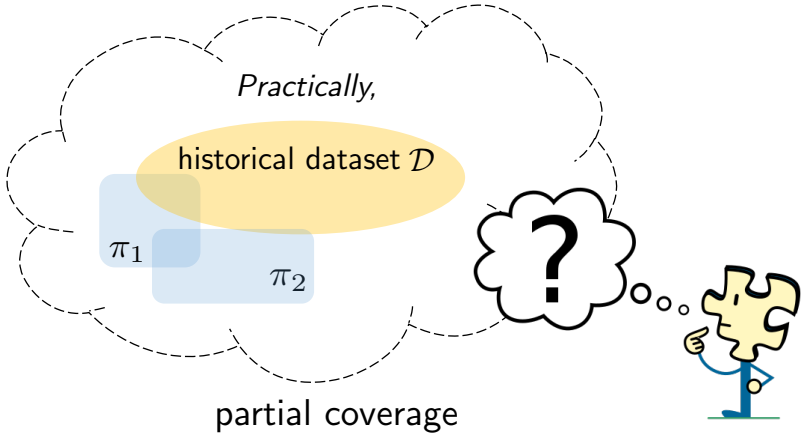
uniform coverage over entire space
(sufficiently explored)

Challenges of offline RL

Partial coverage of state-action space:



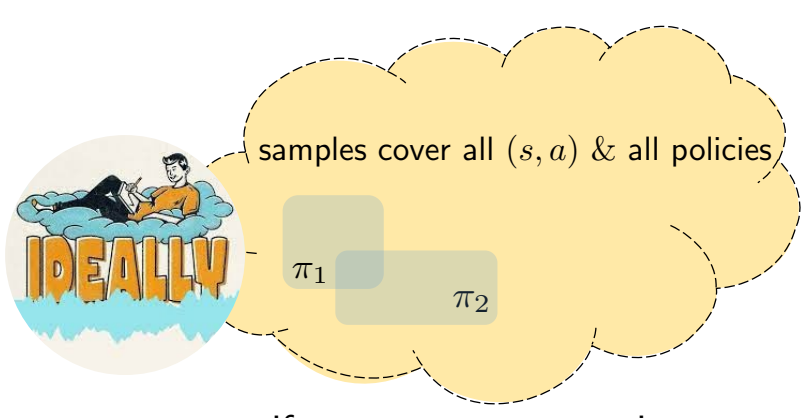
uniform coverage over entire space
(sufficiently explored)



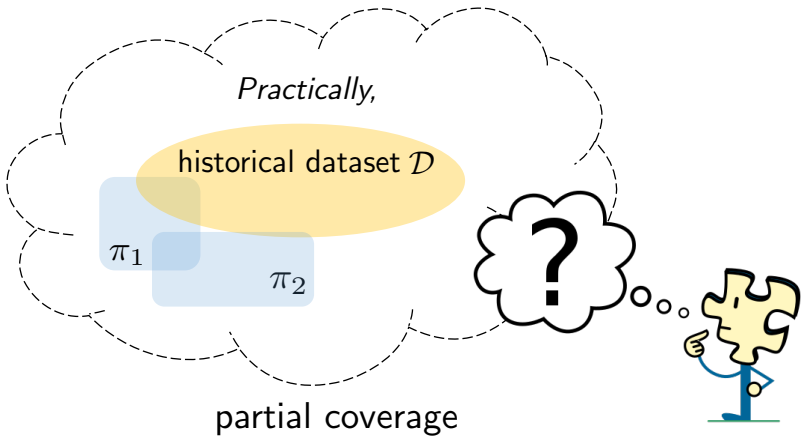
partial coverage
(inadequately explored)

Challenges of offline RL

Partial coverage of state-action space:



uniform coverage over entire space
(sufficiently explored)

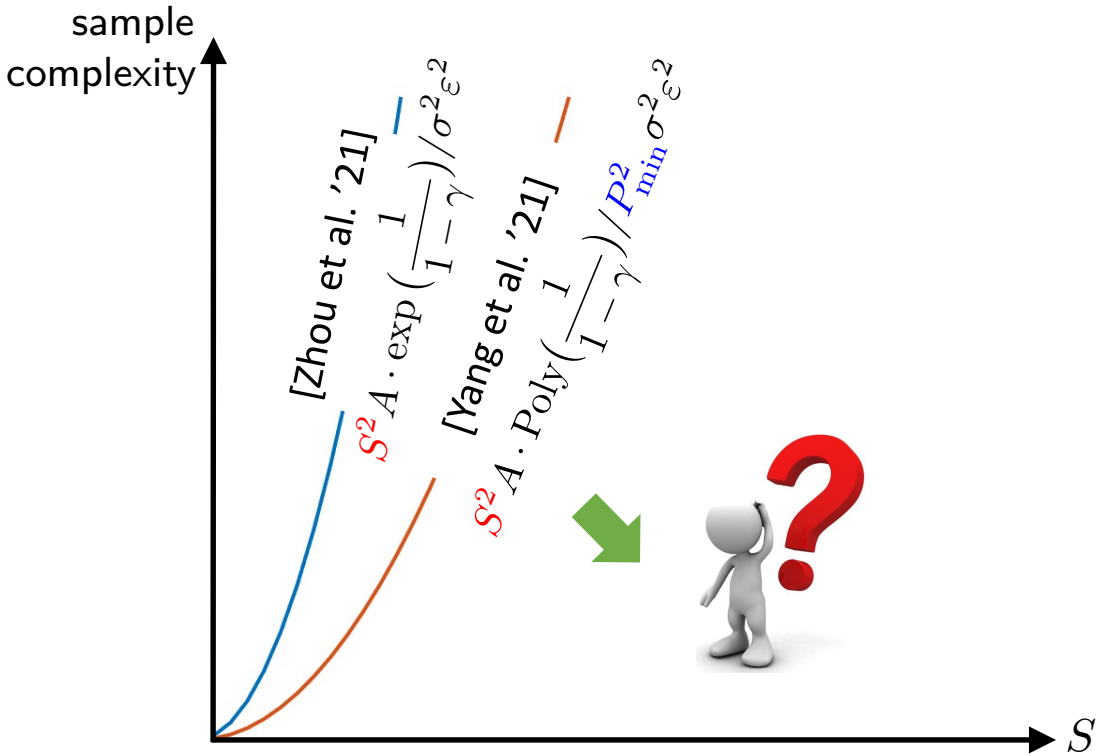


partial coverage
(inadequately explored)

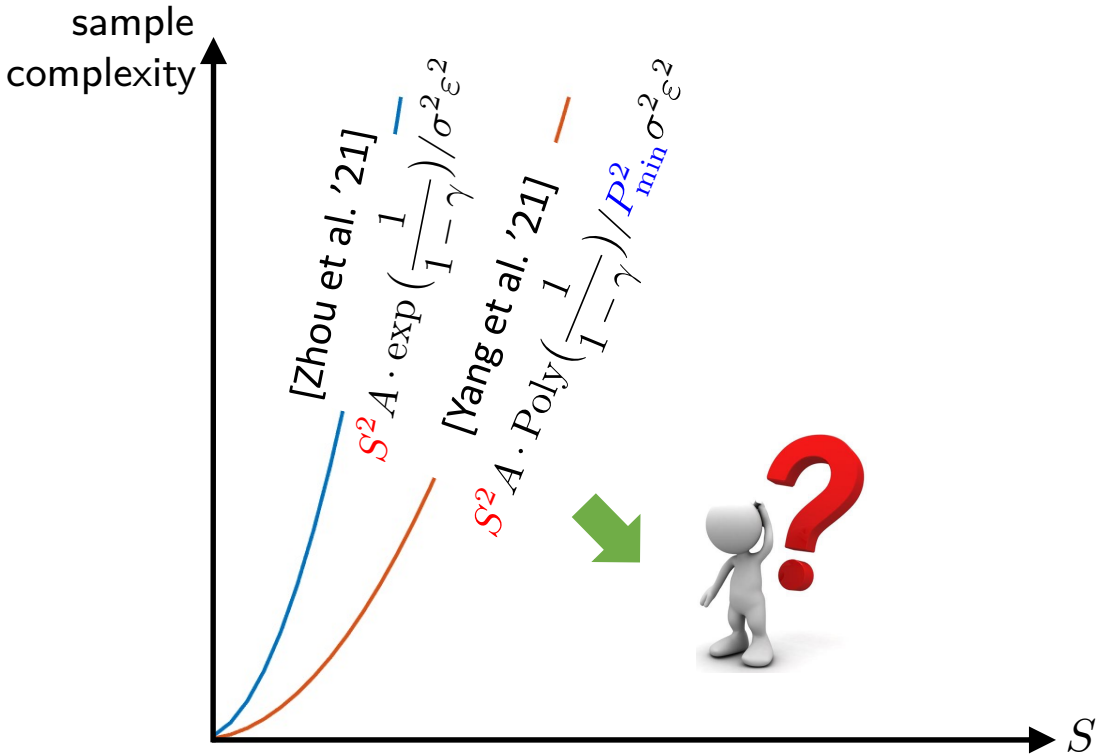
Distribution shift:

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

Prior art under full coverage



Prior art under full coverage



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

How to quantify the compounded distribution shift?

Robust single-policy concentrability coefficient

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

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

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

How to quantify the compounded distribution shift?

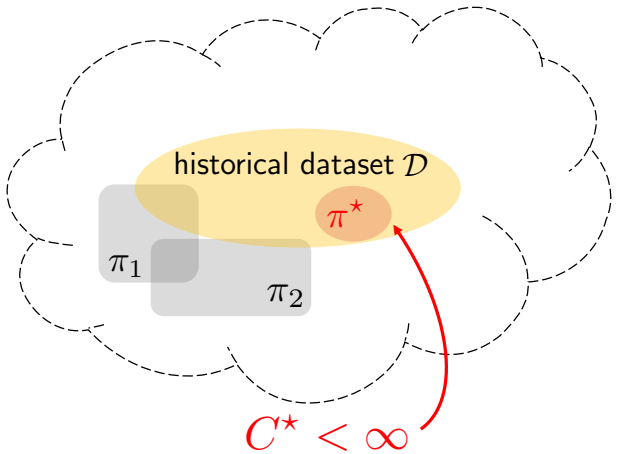
Robust single-policy concentrability coefficient

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

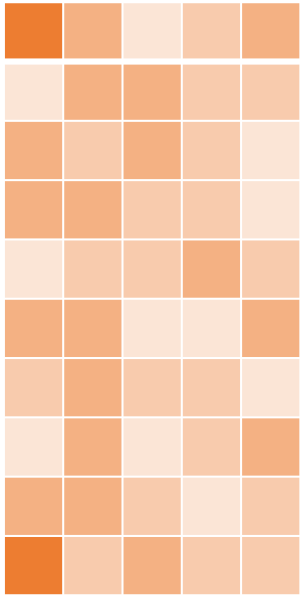
$$= \left\| \frac{\textit{occupancy distribution of } (\pi^\star, P\in\mathcal{U}(P^o))}{\textit{occupancy distribution of } \mathcal{D}} \right\|_\infty$$

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

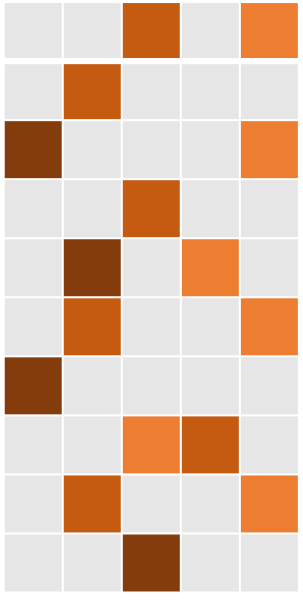
- captures distributional shift due to behavior policy and environment.
- $C^\star_{\text{rob}} \le A$ under full coverage.



Challenges in the sample-starved regime



truth:
$P \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$



empirical estimate: $\widehat{P}$

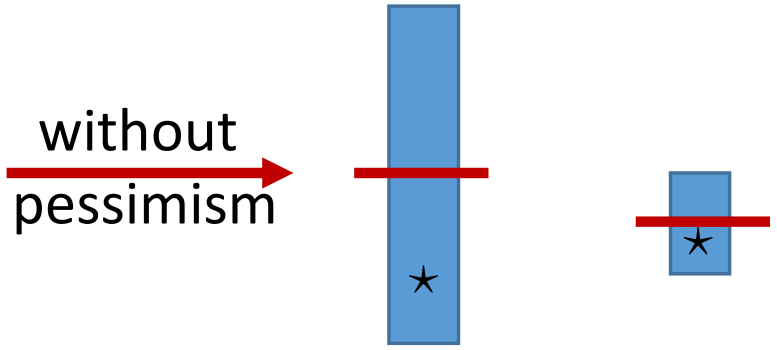
- Can't recover P faithfully if sample size $\ll |\mathcal{S}|^2|\mathcal{A}|$!

Issue: poor value estimates under partial and poor coverage.

Pessimism in the face of uncertainty

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

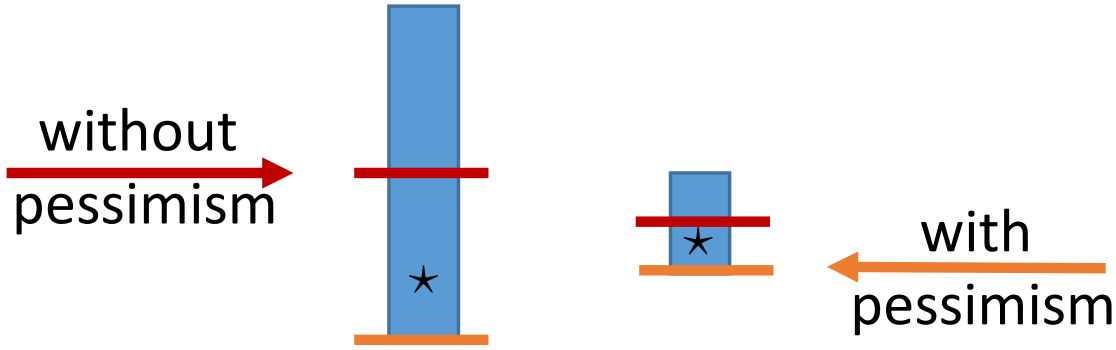
— (Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21)



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)



Distributionally robust value iteration (DRVI) with lower confidence bound (LCB):

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

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

Key novelty: design the penalty term to capture the variability in robust RL.

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

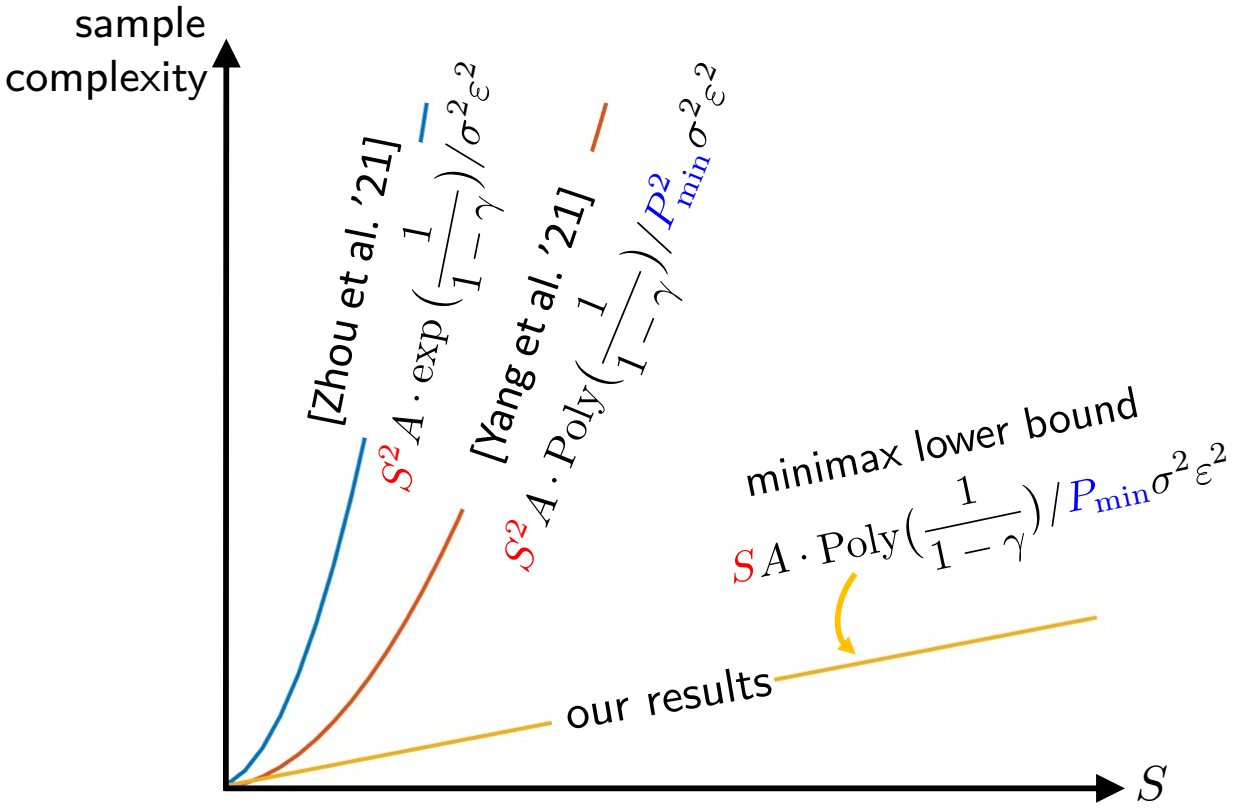
Theorem (Shi and Chi '22)

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*

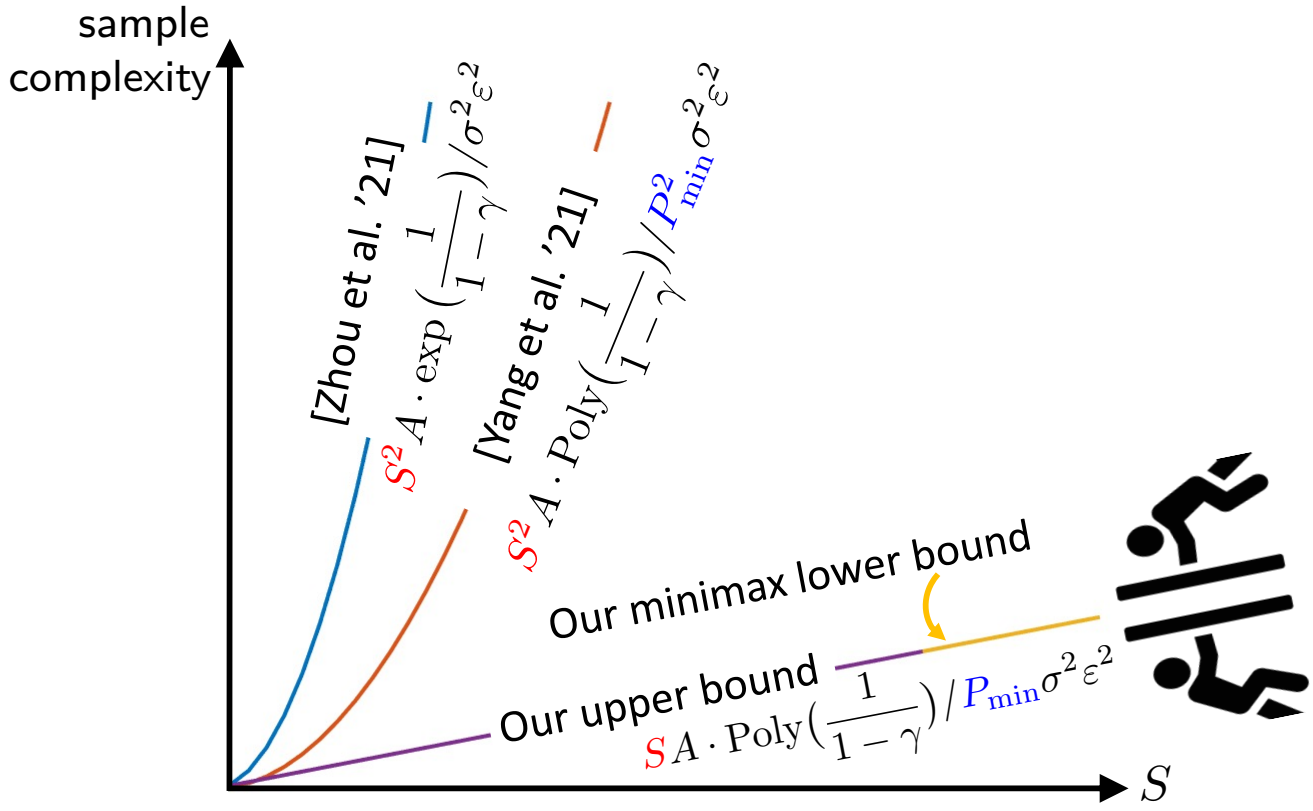
$$\widetilde{\Omega}\left(\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^2\sigma^2\epsilon^2}\right).$$

- 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



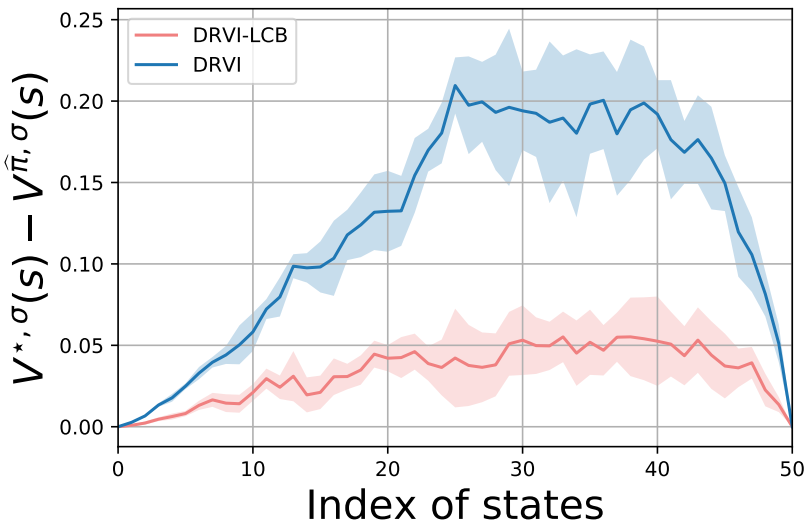
Compare to prior art under full coverage

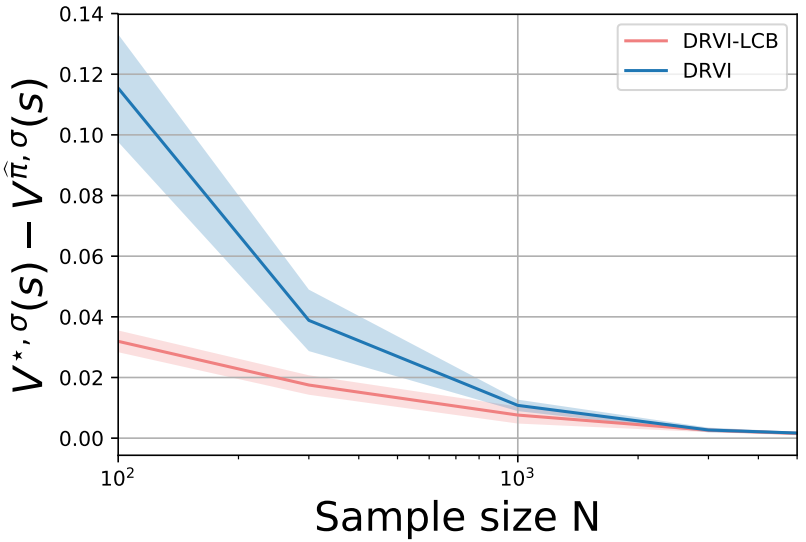


Our DRVI-LCB method is near minimax-optimal!

Numerical experiments

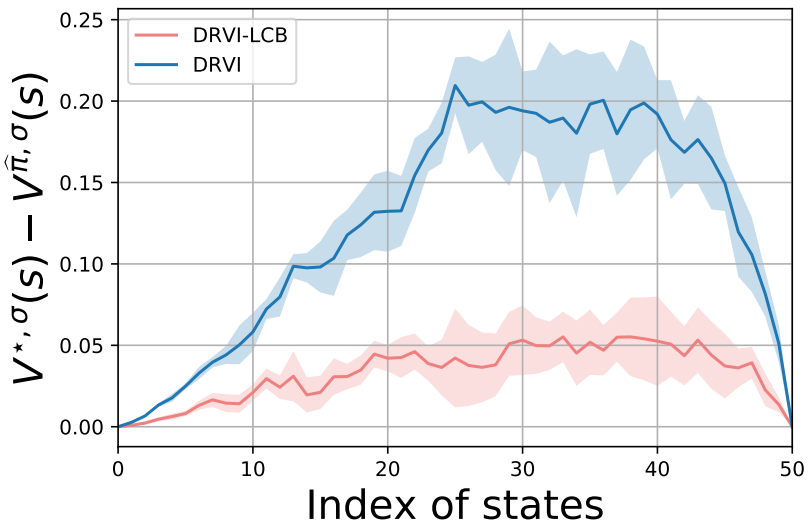
Gambler's problem: a gambler bets on a sequence of coin flips, winning the stake with heads and losing with tails. Starting from some initial balance, the game ends when the gambler's balance either reaches 50 or 0, or the total number of bets H is hit.

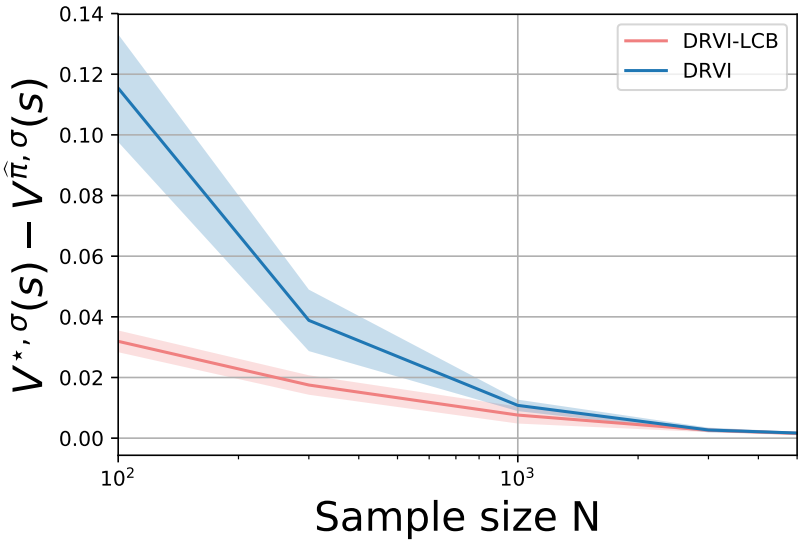




Numerical experiments

Gambler's problem: a gambler bets on a sequence of coin flips, winning the stake with heads and losing with tails. Starting from some initial balance, the game ends when the gambler's balance either reaches 50 or 0, or the total number of bets H is hit.

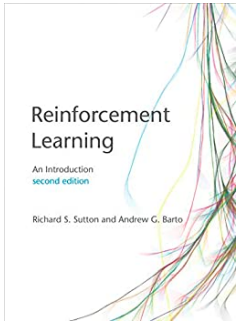


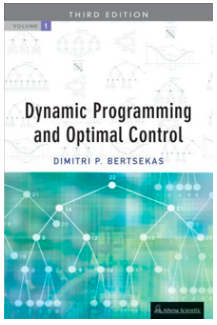


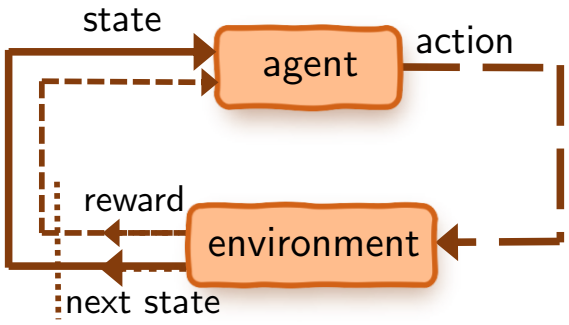
Pessimism improves the sample efficiency in robust offline RL!

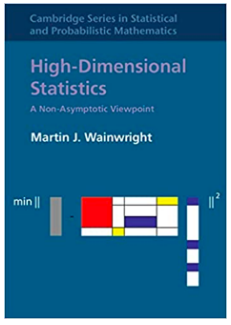
Concluding remarks

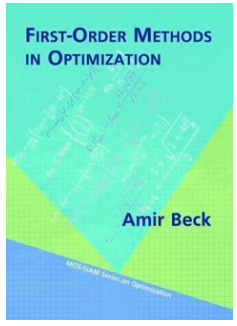
Concluding remarks











Understanding non-asymptotic performances of robust RL algorithms sheds light to their empirical successes (and failures)!

Thanks!

- The Curious Price of Distributional Robustness in Reinforcement Learning with a Generative Model, arXiv:2305.16589; short version at NeurIPS 2023.
- Distributionally Robust Model-Based Offline Reinforcement Learning with Near-Optimal Sample Complexity, arXiv:2208.05767.

https://users.ece.cmu.edu/~yuejiec/