Offline Reinforcement Learning: Towards Optimal Sample Complexity and Distributional Robustness

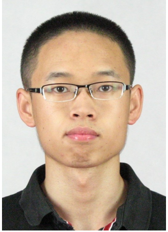
Yuejie Chi

Carnegie Mellon University

University of Virginia
March 2023

My wonderful collaborators

Laixi Shi
CMU



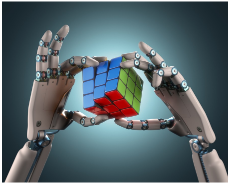
Gen Li
UPenn

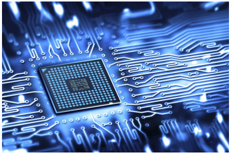
Yuxin Chen
UPenn

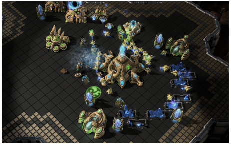
Yuting Wei
UPenn

Recent successes in reinforcement learning (RL)

In RL, an agent learns by interacting with an environment.







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

Sample efficiency

Collecting data samples might be expensive or time-consuming

clinical trials

autonomous driving

online ads

Collecting data samples might be expensive or time-consuming

clinical trials

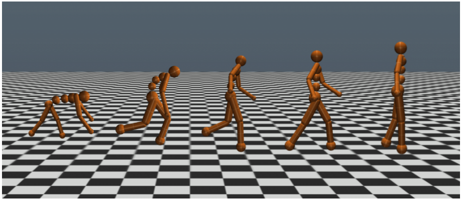
autonomous driving

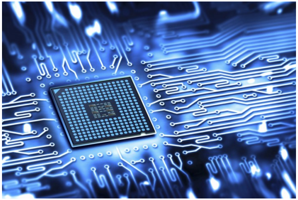
online ads

Calls for design of sample-efficient RL algorithms!

Computational efficiency

Running RL algorithms might take a long time and space

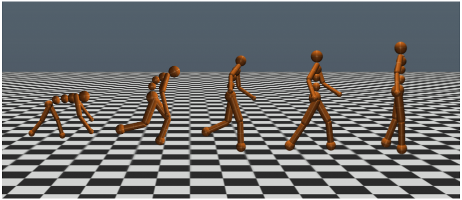


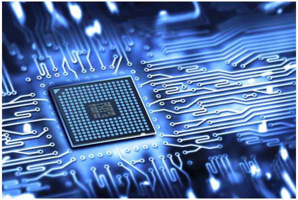


many CPUs / GPUs / TPUs + computing hours

Computational efficiency

Running RL algorithms might take a long time and space





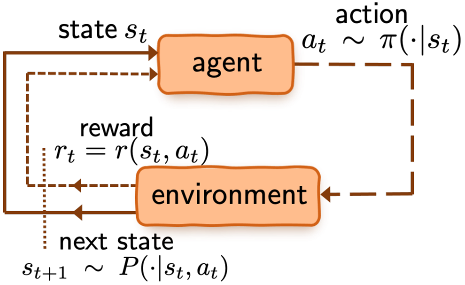
many CPUs / GPUs / TPUs + computing hours

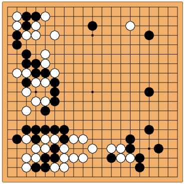
Calls for computationally efficient RL algorithms!

Recent advances in statistical RL

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

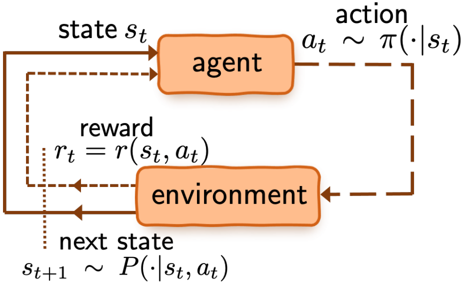
Markov decision processes

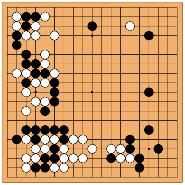




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space

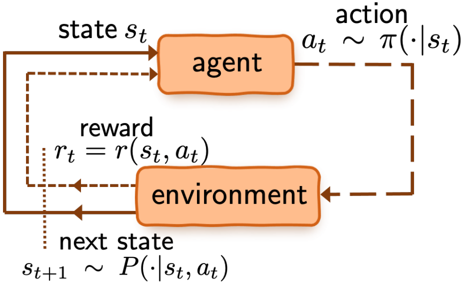
Markov decision processes

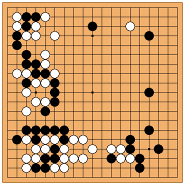




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s,a) \in [0,1]$: immediate reward

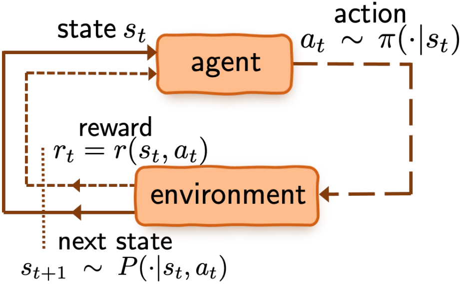
Markov decision processes

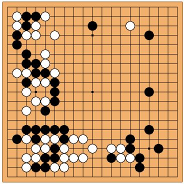




- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule)

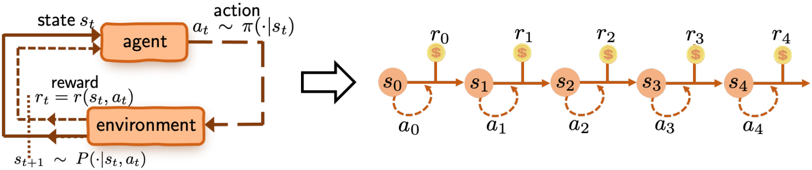
Markov decision processes





- $\mathcal{S}$: 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

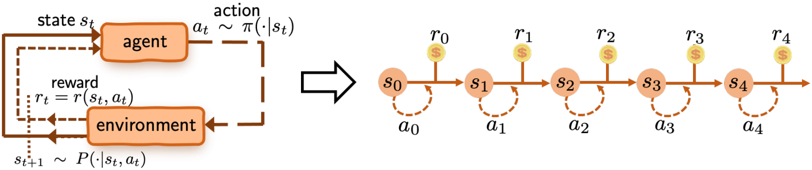


Value/Q-function function of policy π:

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

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

Value function



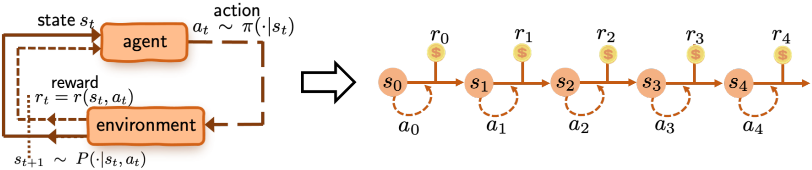
Value/Q-function function of policy π:

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

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

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

Value function



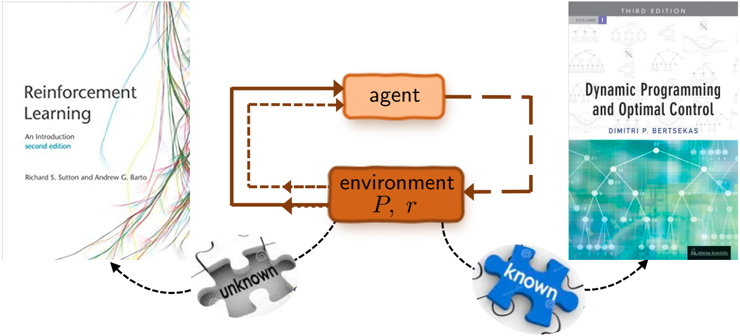
Value/Q-function function of policy π:

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

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

- $\gamma \in [0,1)$ is the discount factor; $\frac{1}{1-\gamma}$ is effective horizon
- Expectation is w.r.t. the sampled trajectory under π
- Given initial state distribution ρ, let $V^\pi(\rho) = \mathbb{E}_{s\sim\rho} V^\pi(s)$.

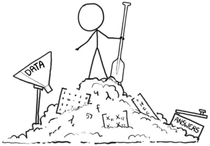
Searching for the optimal policy



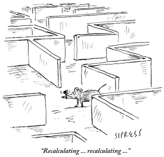
Goal: find the optimal policy $\pi^\star$ that maximize $V^\pi(\rho)$

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

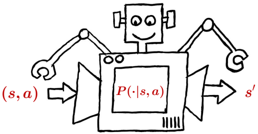
Data source in RL



offline RL



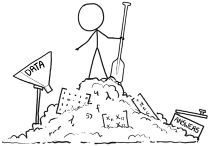
online RL



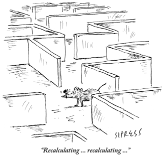
generative model

Data source in RL

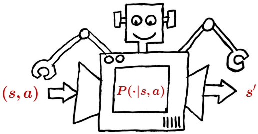
Exploration



offline RL



online RL



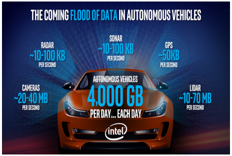
generative model

Our focus: offline RL without exploration

Offline RL / Batch RL

- Sometimes we can not explore or generate new data
- But we have already stored tons of historical data

medical records



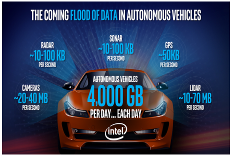
data of self-driving

clicking times of ads

Offline RL / Batch RL

- Sometimes we can not explore or generate new data
- But we have already stored tons of historical data

medical records



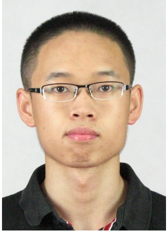
data of self-driving

clicking times of ads

Can we learn a good policy based solely on historical data without active exploration?

Model-based offline RL is nearly minimax optimal

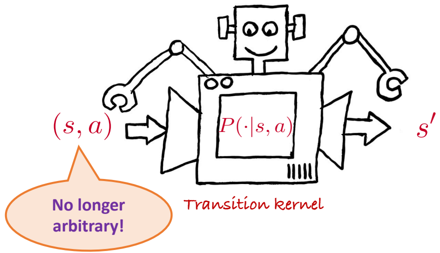

Laixi Shi
CMU


Gen Li
UPenn

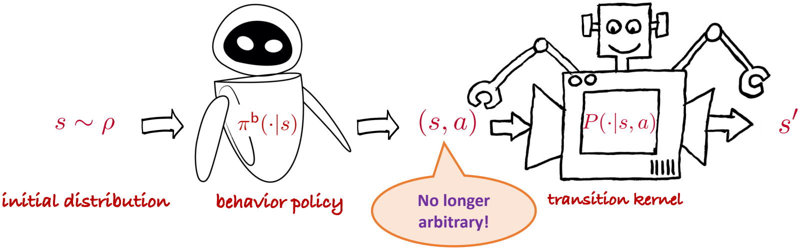

Yuxin Chen
UPenn


Yuting Wei
UPenn

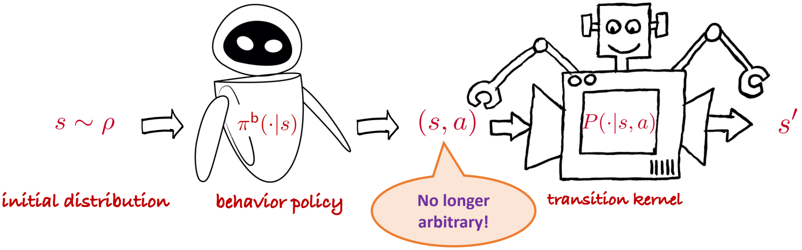
A simplified model of history data from behavior policy



A simplified model of history data from behavior policy



A simplified model of history data from behavior policy



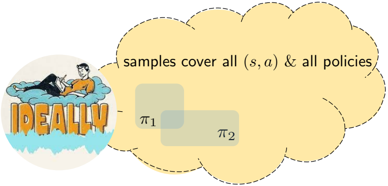
Goal of offline RL: given history data $\mathcal{D} := \{(s_i, a_i, s_i')\}_{i=1}^N$, find an ϵ-optimal policy $\widehat{\pi}$ obeying

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

— *in a sample-efficient manner*

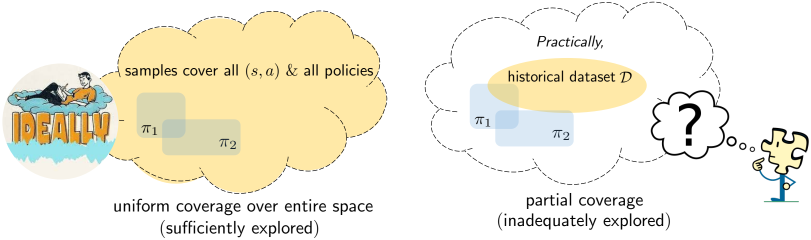
Challenges of offline RL

Partial coverage of state-action space:



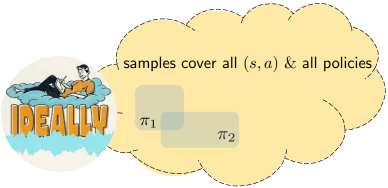
uniform coverage over entire space
(sufficiently explored)

Partial coverage of state-action space:

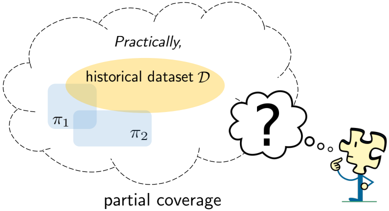


Challenges of offline RL

Partial coverage of state-action space:



uniform coverage over entire space
(sufficiently explored)



partial coverage
(inadequately explored)

Distribution shift:

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

How to quantify the distribution shift?

Single-policy concentrability coefficient (Rashidineiad et al.)

$$C^\star := \max_{s,a} \frac{d^{\pi^\star}(s,a)}{d^{\pi^\mathsf{b}}(s,a)} \geq 1$$

where $d^\pi(s,a)$ is the state-action occupation density of policy π.

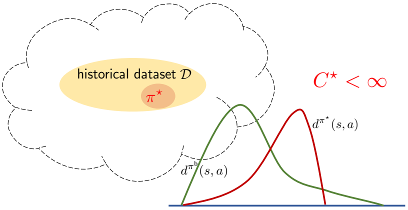
How to quantify the distribution shift?

Single-policy concentrability coefficient (Rashidineiad et al.)

$$C^\star := \max_{s,a} \frac{d^{\pi^\star}(s,a)}{d^{\pi^\mathsf{b}}(s,a)} \geq 1$$

where $d^\pi(s,a)$ is the state-action occupation density of policy π.

- captures distribution shift
- allows for partial coverage



How to quantify the distribution shift? — a refinement

Single-policy clipped concentrability coefficient (Li et al., '22)

$$C^{\star}_{\mathsf{clipped}} := \max_{s,a} \frac{\min\{d^{\pi^\star}(s,a), 1/S\}}{d^{\pi^{\mathsf{b}}}(s,a)} \geq 1/S$$

where $d^{\pi}(s,a)$ *is the state-action occupation density of policy* π.

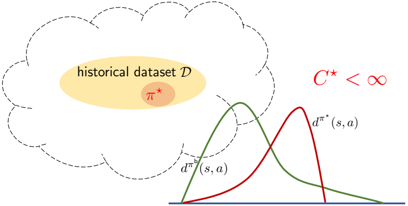
How to quantify the distribution shift? — a refinement

Single-policy clipped concentrability coefficient (Li et al., '22)

$$C^{\star}_{\text{clipped}} := \max_{s,a} \frac{\min\{d^{\pi^\star}(s,a), 1/S\}}{d^{\pi^{\mathsf{b}}}(s,a)} \geq 1/S$$

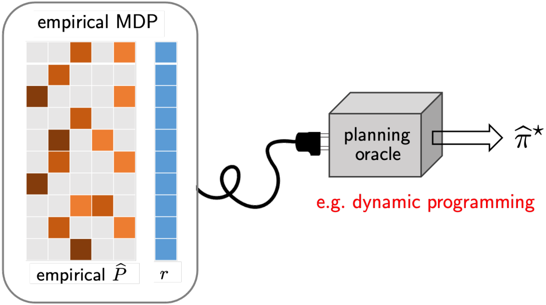
where $d^{\pi}(s,a)$ is the state-action occupation density of policy π.

- captures distribution shift
- allows for partial coverage
- $C^{\star}_{\text{clipped}} \leq C^{\star}$
- $C^{\star}_{\text{clipped}} \leq A$ (while $C^{\star} \leq SA$) under full coverage.



A "plug-in" model-based approach

— (Azar et al. '13, Agarwal et al. '19, Li et al. '20)

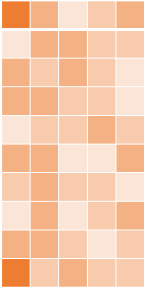


Empirical estimates: estimate $\widehat{P}(s'|s,a)$ by $\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{s'_{(i)} = s'\}}_{\text{empirical frequency}}$

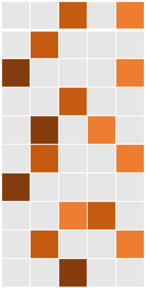
Planning (e.g., value iteration) based on $\widehat{P}$:

$$\widehat{Q}(s,a) \;\leftarrow\; r(s,a) + \gamma\big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}\big\rangle, \quad \widehat{V}(s) = \max_a \widehat{Q}(s,a).$$

Challenges in the sample-starved regime



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



empirical estimate:
$\widehat{P}$

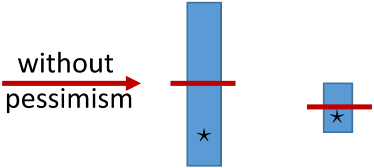
- Can't recover P faithfully if sample size $\ll |\mathcal{S}|^2|\mathcal{A}|$!

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

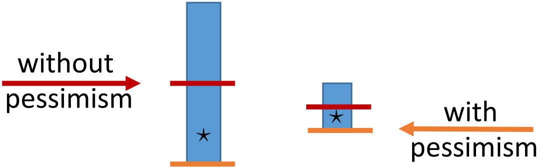
— (Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21)



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)

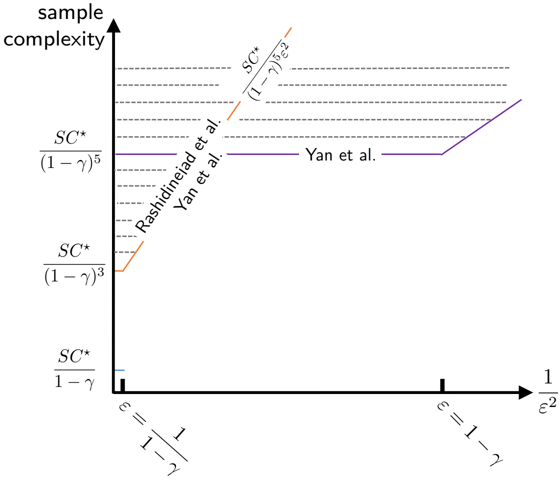


Value iteration with lower confidence bound (VI-LCB):

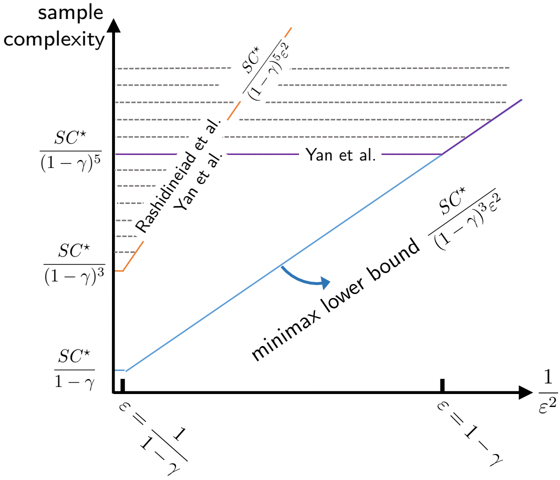
$$\widehat{Q}(s,a) \leftarrow \max\Big\{ r(s,a) + \gamma \big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V} \big\rangle - \underbrace{b(s,a;\widehat{V})}_{\text{uncertainty penalty}}, 0 \Big\},$$

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

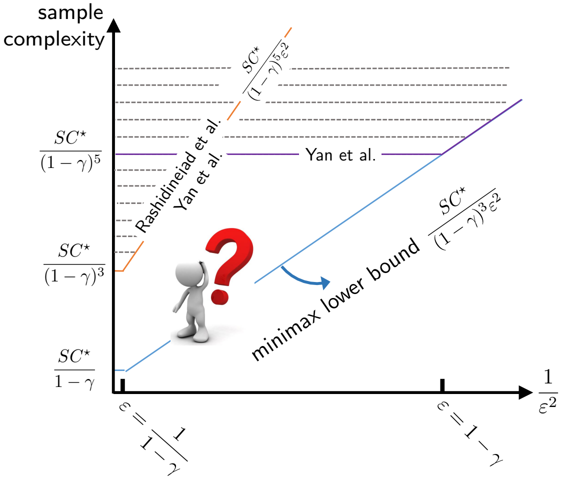
A benchmark of prior arts



A benchmark of prior arts



A benchmark of prior arts



Can we close the gap with the minimax lower bound?

Sample complexity of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '22)

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

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

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{SC^{\star}_{\text{clipped}}}{(1-\gamma)^3\epsilon^2}\right).$$

Sample complexity of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '22)

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

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

with high prob., with sample complexity at most

$$\widetilde{O}\left(\frac{SC^{\star}_{\text{clipped}}}{(1-\gamma)^3\epsilon^2}\right).$$

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

Minimax optimality of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '22)

For any $\gamma \in [2/3, 1)$, $S \geq 2$, $C^{\star}_{\text{clipped}} \geq 8\gamma/S$, *and* $0 < \epsilon \leq \frac{1}{42(1-\gamma)}$, *there exists some MDP and batch dataset such that no algorithm succeeds if the sample size is below*

$$\widetilde{\Omega}\left(\frac{SC^{\star}_{\text{clipped}}}{(1-\gamma)^3\epsilon^2}\right).$$

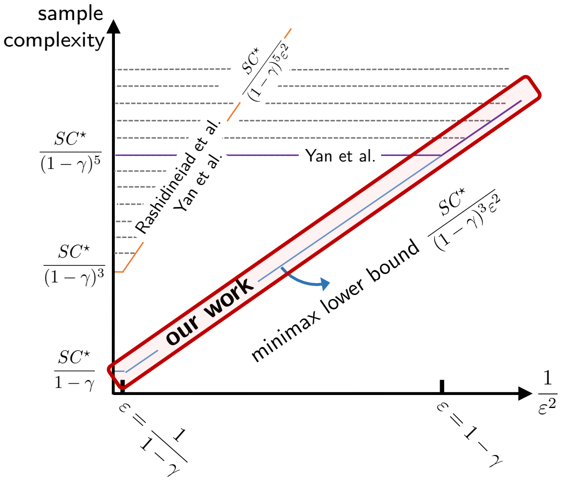
Minimax optimality of model-based offline RL

Theorem (Li, Shi, Chen, Chi, Wei '22)

For any $\gamma \in [2/3, 1)$, $S \geq 2$, $C^{\star}_{\text{clipped}} \geq 8\gamma/S$, *and* $0 < \epsilon \leq \frac{1}{42(1-\gamma)}$, *there exists some MDP and batch dataset such that no algorithm succeeds if the sample size is below*

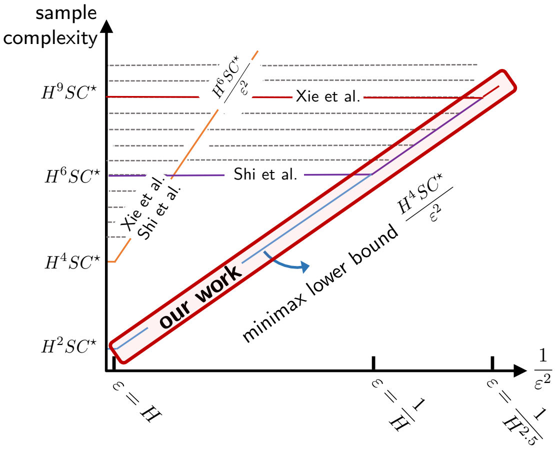
$$\widetilde{\Omega}\left(\frac{SC^{\star}_{\text{clipped}}}{(1-\gamma)^3\epsilon^2}\right).$$

- verifies the near-minimax optimality of the pessimistic model-based algorithm
- improves upon prior results by allowing $C^{\star}_{\text{clipped}} \asymp 1/S$.



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

The finite-horizon case



Offline RL meets distributional robustness

Laixi Shi
CMU

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;)

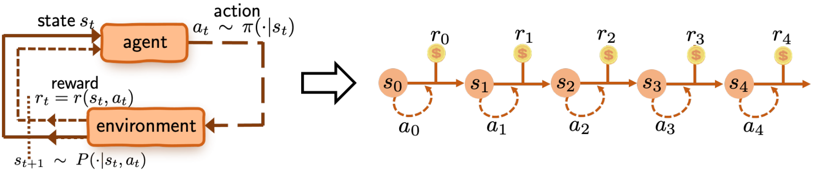
Training environment

$\neq$

Test environment

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

Distributionally robust MDP



Uncertainty set of the normal transition kernel P^o:

$$\mathcal{U}^\sigma(P^o) = \left\{P : \quad \mathsf{KL}\left(P \parallel P^o\right) \le \sigma\right\}$$

Robust value/Q function of policy π:

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

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

The optimal robust policy $\pi^\star$ maximizes $V^{\pi,\sigma}(\rho)$

Distributionally robust Bellman's optimality equation

(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 Bellman's optimality equation

(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

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

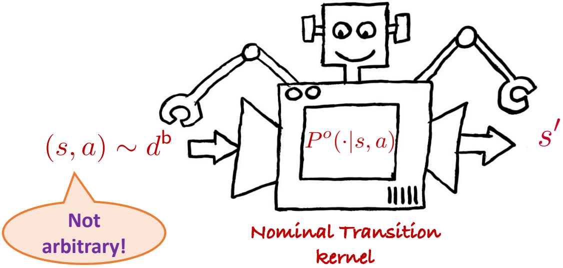
$$V^{\star,\sigma}(s) = \max_a Q^{\star,\sigma}(s,a)$$

Robust value iteration:

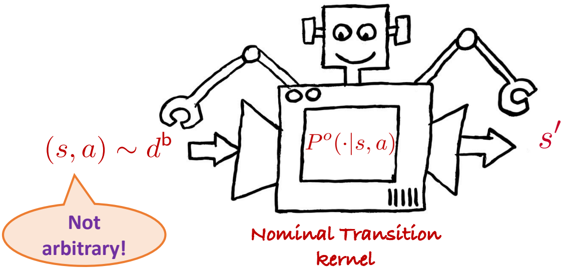
$$Q(s,a) \leftarrow r(s,a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma\left(P^o_{s,a}\right)} \langle P_{s,a}, V \rangle,$$

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

Distributionally robust offline RL



Distributionally robust offline RL

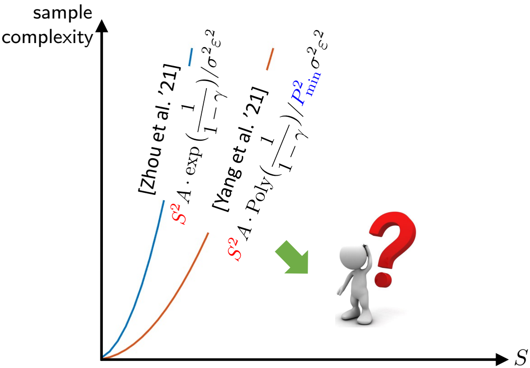


Goal of robust offline RL: given $\mathcal{D} := \{(s_i, a_i, s_i')\}_{i=1}^N$ from the *nominal* environment P^0, find an ϵ-optimal robust policy $\widehat{\pi}$ obeying

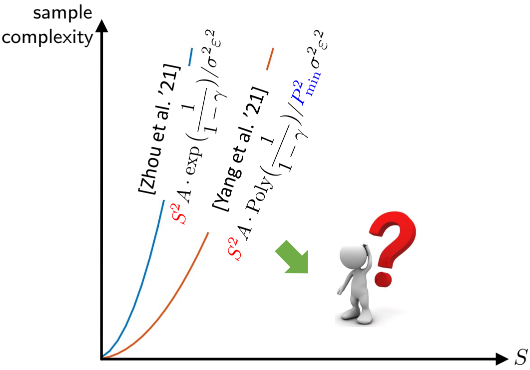
$$V^{\star,\sigma}(\rho) - V^{\widehat{\pi},\sigma}(\rho) \le \epsilon$$

— *in a sample-efficient manner*

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

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

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

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

How to quantify the compounded distribution shift?

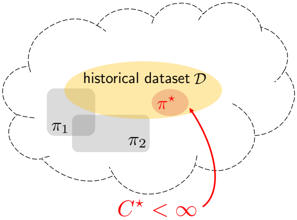
Robust single-policy concentrability coefficient

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

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

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

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



Distributionally robust value iteration 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 variability 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

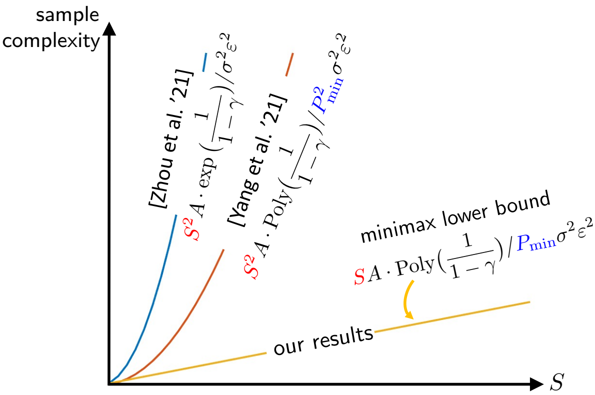
Theorem (Shi and Chi '22)

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*

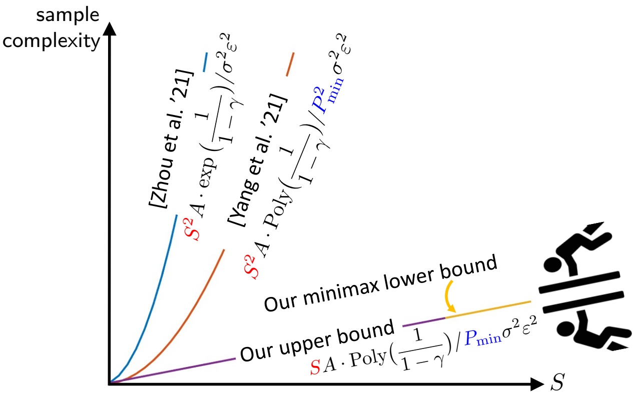
$$\widetilde{\Omega}\left(\frac{SC^\star_{\mathsf{rob}}}{P^\star_{\mathsf{min}}(1-\gamma)^2\sigma^2\epsilon^2}\right).$$

- 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

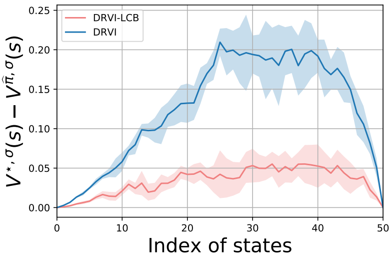


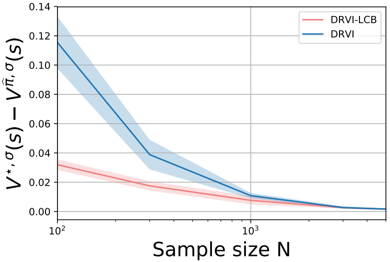
Compare to prior art under full coverage



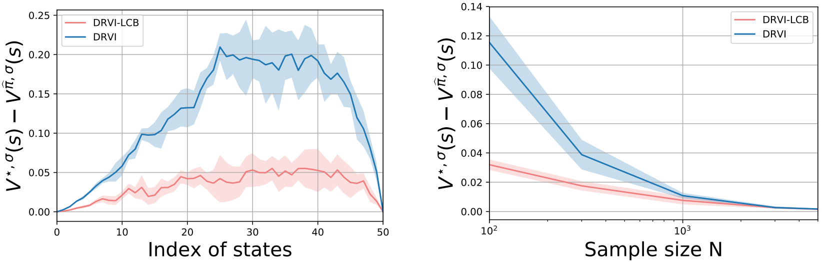
Our DRVI-LCB method is near minimax-optimal!

Numerical experiments





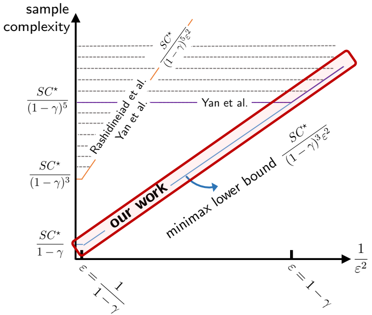
Numerical experiments

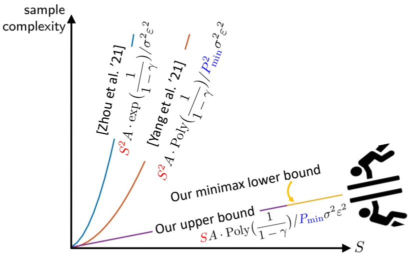


Pessimism improves the sample efficiency in robust offline RL!

Concluding remarks

Concluding remarks





Model-based offline RL algorithms with pessimism are near minimax-optimal in both nominal MDP and robust MDP!

Thank you!

- Settling the sample complexity of model-based offline reinforcement learning, arXiv:2204.05275.
- Pessimistic Q-Learning for Offline Reinforcement Learning: Towards Optimal Sample Complexity, ICML 2022.
- Distributionally Robust Model-Based Offline Reinforcement Learning with Near-Optimal Sample Complexity, arXiv:2208.05767.

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