Foundations of Reinforcement Learning

Offline RL: the pessimism principle

Yuejie Chi

Department of Electrical and Computer Engineering

Carnegie Mellon University

Spring 2023

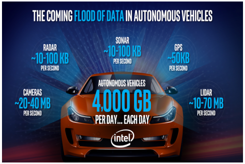
Ack: some materials of this lecture are borrowed/adapted from Cong Ma (Chicago).

Outline

Offline RL / Batch RL

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

medical records



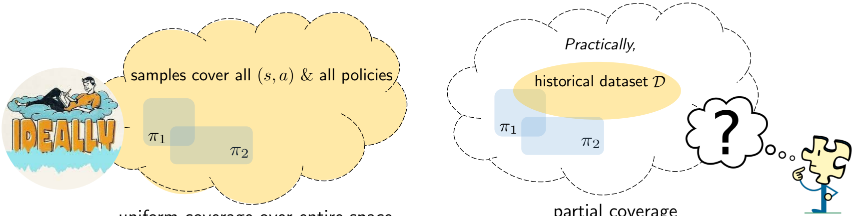
data of self-driving

clicking times of ads

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

Challenges of offline RL

Partial coverage of state-action space:



uniform coverage over entire space
(sufficiently explored)

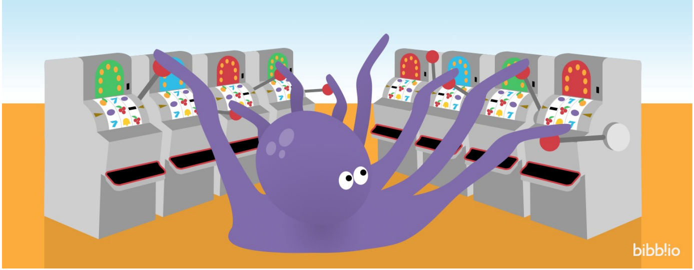
partial coverage
(inadequately explored)

Distribution shift:

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

Offline multi-arm bandits

Multi-arm bandit



- Action space: $\mathcal{A} = \{1, 2, \ldots, A\}$
- Reward distributions: $R(\cdot \mid a)$ with mean $r(a)$

— *correspond to MDP with single state and* $\gamma = 0$

Offline learning in multi-arm bandit

Batch dataset: $\mathcal{D} = \{(a_i, r_i)\}_{1 \le i \le N}$, where

$$a_i \sim \mu, \qquad r_i \sim R(\cdot \mid a_i)$$

are collected in an *i.i.d.* manner, where $\mu \in \Delta(\mathcal{A})$ is the **behavior** policy.

Goal: minimize expected sub-optimality based on collected data

$$\mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\widehat{a})],$$

where $a^\star = \arg\max_a r(a)$ is the optimal arm with the highest mean reward.

How to capture the distribution shift?

Single-policy concentrability coefficient [Rashidinejad et al., 2021]

$$C^\star := \max_a \frac{\pi^\star(a)}{\mu(a)} = \frac{1}{\mu(a^\star)}.$$

μ: behavior policy

$\pi^\star$: optimal policy

- When $C^\star = 1$: expert data
- When $C^\star > 1$: behavior policy deviates from the optimal policy
- When μ is uniform (random exploration), $C^\star = A$.
- Partial coverage: $C^\star$ is finite as long as $\mu(a^\star) > 0$.

A natural idea: empirical best arm

A natural idea is to pick the empirical best arm

$$\hat{a} := \arg\max_{a} \hat{r}(a),$$

where $\hat{r}(a)$ is the empirical mean reward of arm a.

Theorem 1 ([Rashidinejad et al., 2021])

For any $\epsilon < 0.05$, $N \geq 500$, there exists a bandit problem with two arms such that for $\hat{a} = \mathrm{argmax}_a \hat{r}(a)$, one has

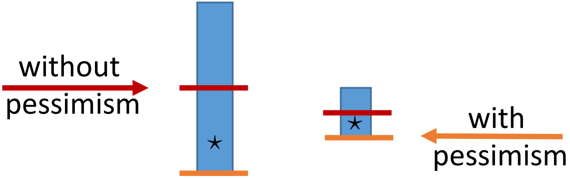
$$\mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\hat{a})] \geq \epsilon.$$

- Empirical best arm is sensitive to arms with few observations
- This happens even when $C^\star$ is close to 1

Pessimism via lower confidence bound

Lessons learned from failure of empirical best arm

- Should not treat arms equally
- Need to be **pessimistic** about arms with few observations



Lower confidence bound (LCB) for bandit: fix some $L > 0$, return

$$\hat{a} := \arg\max_{a} \quad \hat{r}(a) - \frac{L}{\sqrt{\max\{N(a), 1\}}}$$

$N(a)$: *number of times arm a is seen*

A closer look at LCB

Lower confidence bound for bandit: fix some $L > 0$, return

$$\hat{a} := \arg\max_{a} \quad \hat{r}(a) - \frac{L}{\sqrt{\max\{N(a), 1\}}}$$

$N(a)$: *number of times arm a is seen*

- View $\hat{r}(a) - \frac{L}{\sqrt{\max\{N(a),1\}}}$ as lower confidence bound of $r(a)$
- $\frac{L}{\sqrt{\max\{N(a),1\}}}$ arises from Hoeffding concentration inequality
- $\frac{L}{\sqrt{\max\{N(a),1\}}}$ is large when $N(a)$ is small: discount empirical mean with few observations

Performance guarantees

Theorem 2 ([Rashidinejad et al., 2021])

Set $L \asymp \sqrt{\log(AN)}$. Policy $\hat{a}$ returned by LCB algorithm obeys

$$\mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\hat{a})] \lesssim \sqrt{\frac{C^\star}{N}}$$

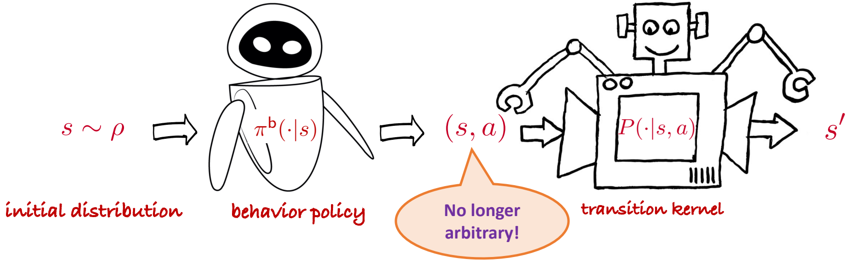
- LCB beats empirical best arm
- To achieve ϵ-optimality, the sample size needs to scale as

$$N \gtrsim \frac{C^\star}{\epsilon^2}.$$

- Performance of LCB degrades gracefully w.r.t. $C^\star$.

Offline RL: mathematical setup

A 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

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

— *in a sample-efficient manner*

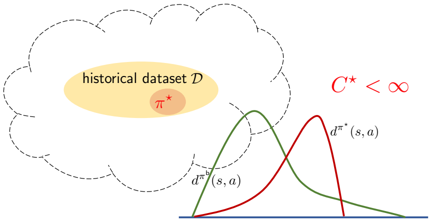
How to capture the distribution shift?

Single-policy concentrability coefficient [Rashidinejad et al., 2021]

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

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

- allows for partial coverage
- Behavior cloning $C^\star = 1$
- Generative model $C^\star = SA$



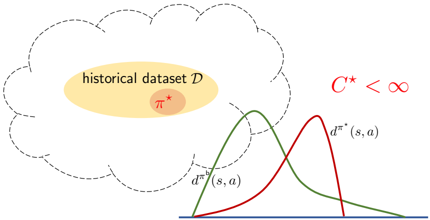
How to capture the distribution shift? a refinement

Clipped single-policy concentrability coefficient [Li et al., 2022]

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

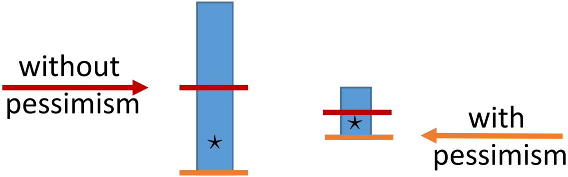
where $d^{\pi}(s,a)$ is the state-action occupation density of policy π.

- allows for partial coverage
- $C^{\star}_{\text{clipped}} \leq C^{\star}$
- Generative model $C^{\star}_{\text{clipped}} = A$



Model-free offline RL: pessimistic Q-learning

LCB-Q: Q-learning with LCB penalty



— [Shi et al., 2022, Yan et al., 2022]

$$Q_{t+1}(s_t, a_t) \leftarrow \underbrace{(1-\eta_t)Q_t(s_t, a_t) + \eta_t \mathcal{T}_t\,(Q_t)\,(s_t, a_t)}_{\text{classical Q-learning}} - \eta_t \underbrace{b_t(s_t, a_t)}_{\text{LCB penalty}}$$

- $b_t(s, a)$: Hoeffding-style confidence bound
- pessimism in the face of uncertainty

sample size: $\widetilde{O}\Big(\frac{SC^\star}{(1-\gamma)^5\varepsilon^2}\Big)$ $\Longrightarrow$ sub-optimal by a factor of $\frac{1}{(1-\gamma)^2}$

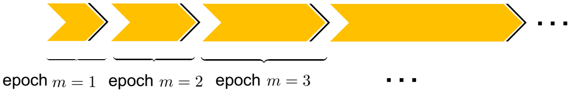
Issue: *large variability in stochastic update rules*

Q-learning with LCB and variance reduction

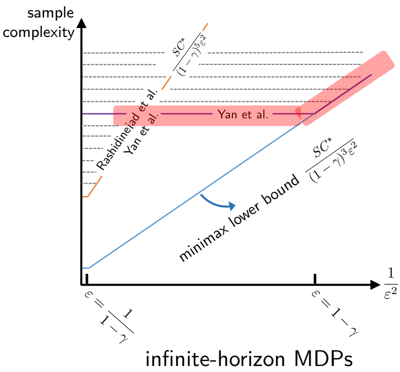
— [Shi et al., 2022, Yan et al., 2022]

$$Q_{t+1}(s_t,a_t) \leftarrow (1-\eta_t)Q_t(s_t,a_t) - \eta_t \underbrace{b_t(s_t,a_t)}_{\text{LCB penalty}}$$
$$+\,\eta_t\Big(\underbrace{\mathcal{T}_t(Q_t) - \mathcal{T}_t(\overline{Q})}_{\text{advantage}} + \underbrace{\widehat{\mathcal{T}}(\overline{Q})}_{\text{reference}}\Big)(s_t,a_t)$$

- incorporates variance reduction into LCB-Q



optimal sample size: $\widetilde{O}\Big(\frac{SC^\star}{(1-\gamma)^3\varepsilon^2}\Big)$ for $\varepsilon \in (0, 1-\gamma]$



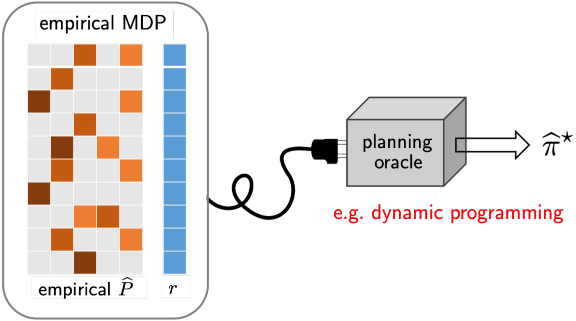
Model-free offline RL attains sample optimality too!

— *with some burn-in cost though . . .*

Model-based offline RL: pessimistic value iteration

A "plug-in" model-based approach

— [Azar et al., 2013]



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

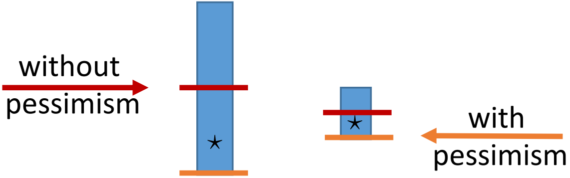
$$\widehat{Q}(s,a) \;\leftarrow\; r(s,a) + \gamma\big\langle \widehat{P}(\cdot \,|\, s,a), \widehat{V}\big\rangle, \quad \widehat{V}(s) = \max_a \widehat{Q}(s,a).$$

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., 2021, Rashidinejad et al., 2021, Li et al., 2022]

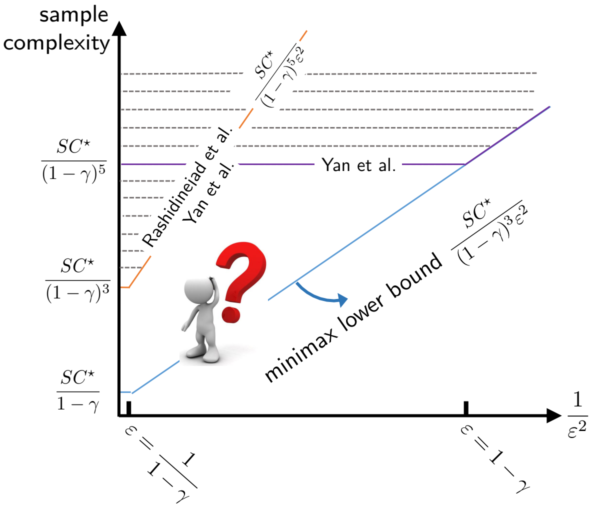


Value iteration with lower confidence bound (VI-LCB):

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

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

A benchmark of prior arts



Can we close the gap with the minimax lower bound?

Sample complexity of model-based offline RL

Theorem 3 ([Li et al., 2022])

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

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

with high prob., with sample complexity at most

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

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

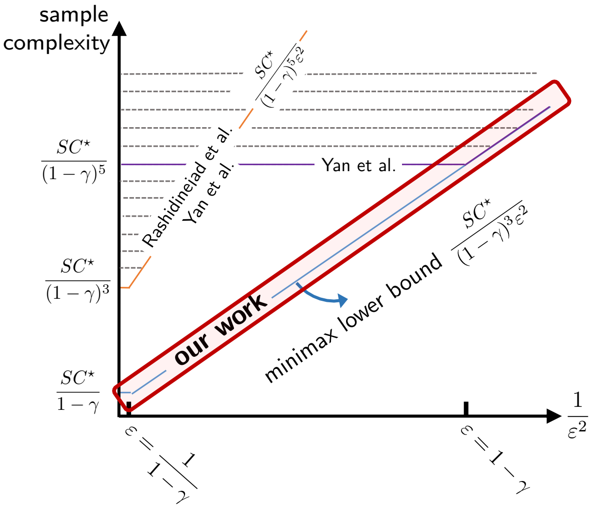
Minimax optimality of model-based offline RL

Theorem 4 ([Li et al., 2022])

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

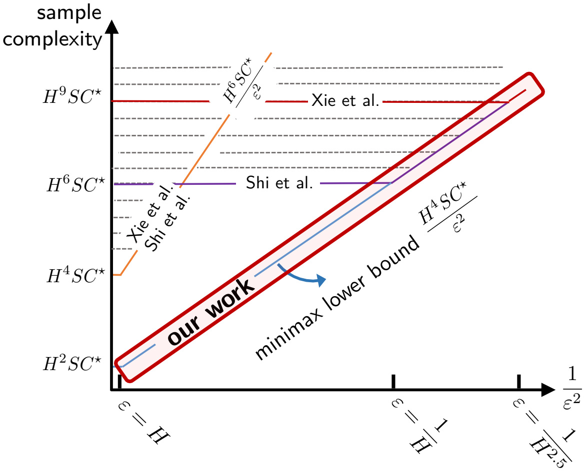
$$\widetilde{\Omega}\left(\frac{SC^{\star}_{\text{clipped}}}{(1-\gamma)^3\varepsilon^2}\right).$$

- 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



References I

Azar, M. G., Munos, R., and Kappen, H. J. (2013).
Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model.
Machine learning, 91(3):325–349.

Jin, Y., Yang, Z., and Wang, Z. (2021).
Is pessimism provably efficient for offline RL?
In *International Conference on Machine Learning*, pages 5084–5096.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. (2022).
Settling the sample complexity of model-based offline reinforcement learning.
arXiv preprint arXiv:2204.05275.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021).
Bridging offline reinforcement learning and imitation learning: A tale of pessimism.
Neural Information Processing Systems (NeurIPS).

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022).
Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity.
In *International Conference on Machine Learning*, pages 19967–20025. PMLR.

Yan, Y., Li, G., Chen, Y., and Fan, J. (2022).
The efficacy of pessimism in asynchronous Q-learning.
arXiv preprint arXiv:2203.07368.