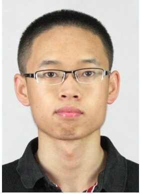
Pessimistic Q-Learning for Offline Reinforcement Learning: Towards Optimal Sample Complexity

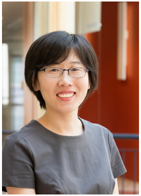
Laixi Shi
CMU



Gen Li
Uppen

Yuting Wei
Uppen

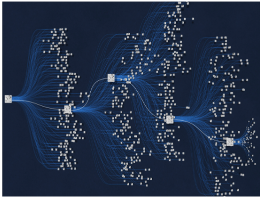
Yuxin Chen
Uppen



Yuejie Chi
CMU

Reinforcement learning (RL) and its challenges

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



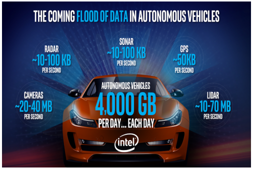
Challenges:

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

Offline/Batch RL motivation: sample efficiency

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

medical records



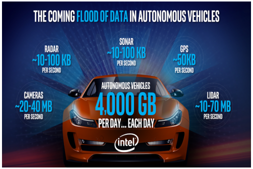
data of self-driving

clicking times of ads

Offline/Batch RL motivation: sample efficiency

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

medical records



data of self-driving

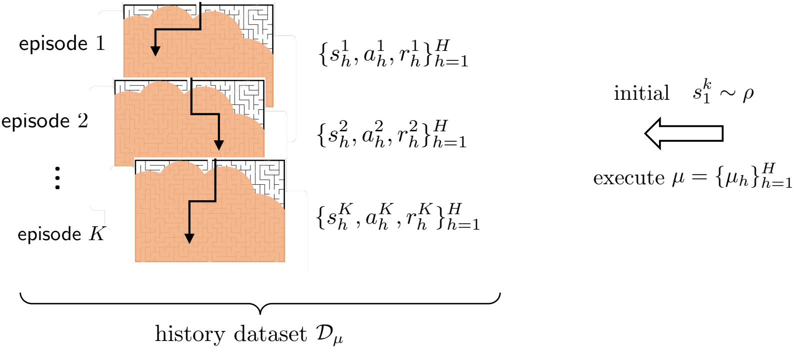
clicking times of ads

Can we design sample-efficient algorithms based on only history data?

Offline/Batch RL: no interaction with environments

Given a history dataset $\mathcal{D}_\mu$ of K episodes, each consisting of H steps:

$$\mathcal{D}_\mu := \left\{ \left(s_1^k, a_1^k, r_1^k, \cdots, s_H^k, a_H^k, r_H^k \right) \right\}_{k=1}^{K}$$



Offline RL: find an ϵ-optimal policy

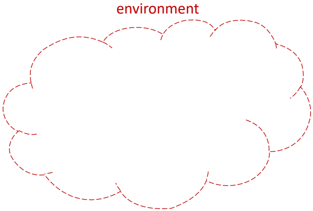
Performance metric: Given initial state distribution ρ and any accuracy level ϵ. An ϵ-optimal policy $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^{H}$ obeys

$$V_1^{\star}(\rho) - V_1^{\widehat{\pi}}(\rho) \leq \epsilon$$

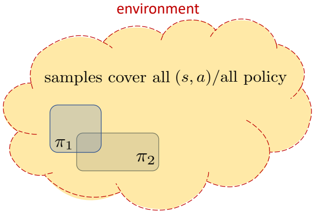
Goal: find an ϵ-optimal policy using only history dataset

— in a sample-efficient manner

Challenges of offline RL: partial coverage

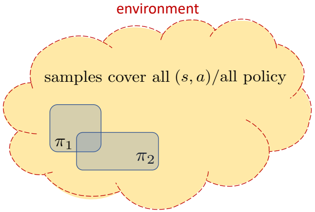


Challenges of offline RL: partial coverage

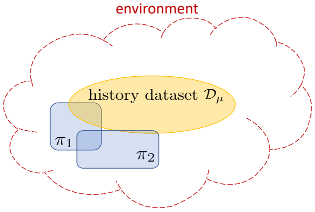


Online/Vanilla Offline: Uniform coverage

Challenges of offline RL: partial coverage

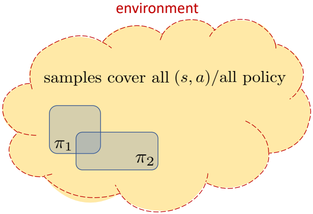


Online/Vanilla Offline: Uniform coverage

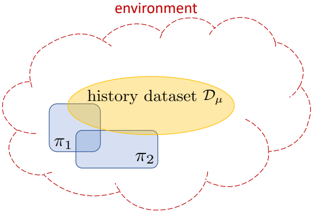


Offline RL with partial coverage

Challenges of offline RL: partial coverage



Online/Vanilla Offline: Uniform coverage

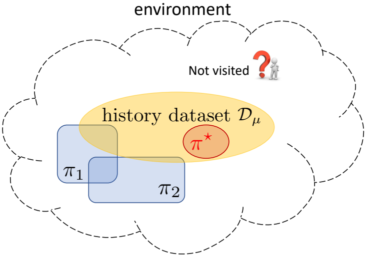


Offline RL with partial coverage

Assumption on $\mathcal{D}_\mu$: finite single-policy concentrability

Key idea: pessimism/conservatism

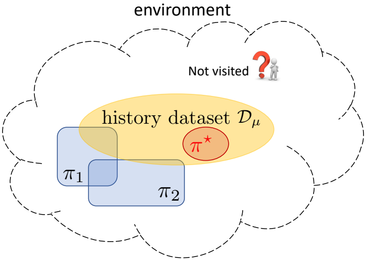
— *Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21*



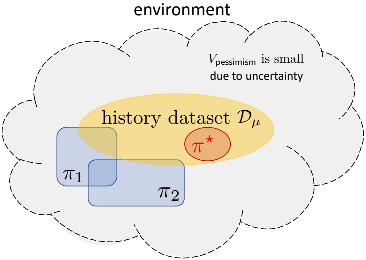
Challenge: partial coverage

Key idea: pessimism/conservatism

— *Jin et al. '20, Rashidinejad et al. '21, Xie et al. '21*



Challenge: partial coverage



Pessimism helps: $V_{\text{pessimism}} < V^\star$
$V_{\text{pessimism}}$ near $V^\star$

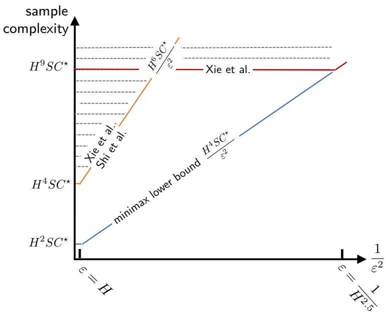
Pessimism in Offline RL:

add (s, a)-dependent penalties to reduce uncertainty damage.

Prior art: Xie et al. '21

Sample complexity $T = KH = |\mathcal{D}_\mu|$

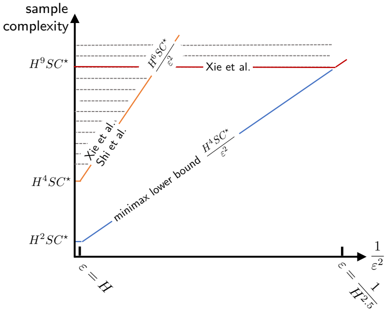
Algorithm	Type	Sample complexity
VI-LCB (Xie et al., 2021)	model-based	$H^6 SC^\star/\epsilon^2$
PEVI-Adv (Xie et al., 2021)	model-based	$H^4 SC^\star/\epsilon^2$
lower bound (Xie et al., 2021)	n/a	$H^4 SC^\star/\epsilon^2$



Prior art: Xie et al. '21

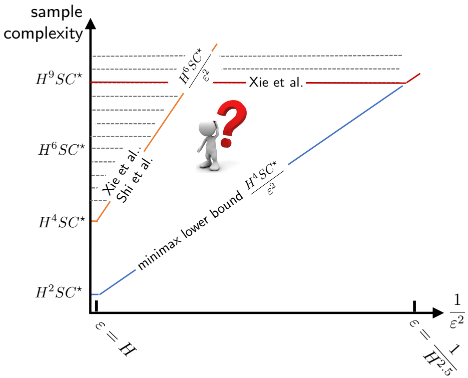
Sample complexity $T = KH = |\mathcal{D}_\mu|$

Algorithm	Type	Sample complexity
VI-LCB (Xie et al., 2021)	model-based	$H^6 SC^\star/\epsilon^2$
PEVI-Adv (Xie et al., 2021)	model-based	$H^4 SC^\star/\epsilon^2$
lower bound (Xie et al., 2021)	n/a	$H^4 SC^\star/\epsilon^2$

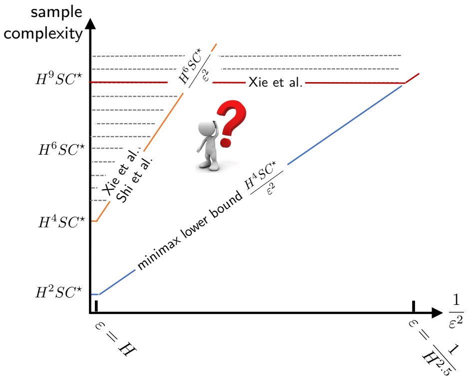


Model-based RL achieve optimal sample complexity when the accuracy level is small enough ($\epsilon \leq \frac{1}{H^{2.5}}$)

No model-free offline RL analysis

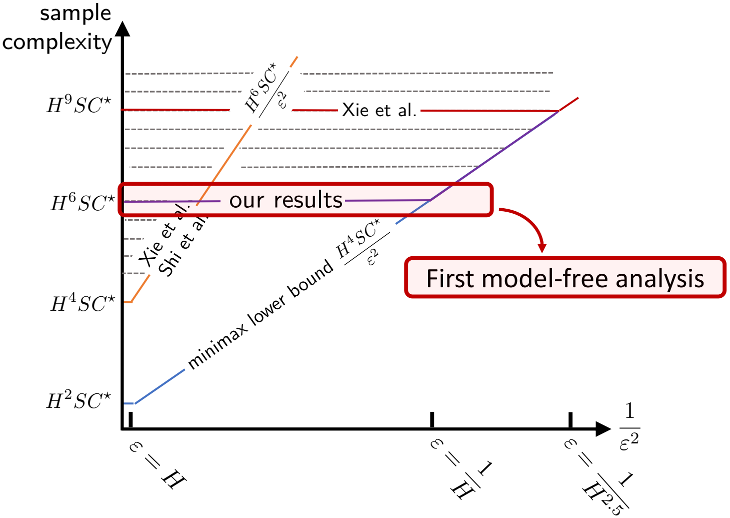


No model-free offline RL analysis



Will flexible model-free RL work?
Can we enlarge the range of accuracy level ϵ?

This work



sample complexity
$H^9SC^\star$
$H^6SC^\star$
$H^4SC^\star$
$H^2SC^\star$
$\frac{H^6SC^\star}{\varepsilon^2}$
Xie et al.
our results
Xie et al.
Shi et al.
minimax lower bound $\frac{H^4SC^\star}{\varepsilon^2}$
First model-free analysis
$\frac{1}{\varepsilon^2}$
$\varepsilon = H$
$\varepsilon = \frac{1}{H}$
$\varepsilon = \frac{1}{H^{2.5}}$

Our algorithm: LCB-Q-Advantage

Theorem (Shi, Li, Wei, Chen, Chi, 2022)

With high prob., for $\epsilon \in \left(0, \frac{1}{H}\right]$, LCB-Q-Advantage *can find an* ϵ*-optimal policy* $\widehat{\pi}$ *as long as (up to log factor)*

$$T \gtrsim O\left(\frac{H^4 SC^\star}{\epsilon^2}\right).$$

Our algorithm: LCB-Q-Advantage

Theorem (Shi, Li, Wei, Chen, Chi, 2022)

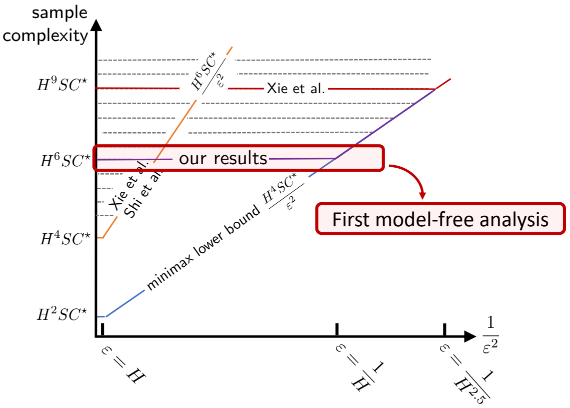
With high prob., for $\epsilon \in \left(0, \frac{1}{H}\right]$, LCB-Q-Advantage *can find an* ϵ*-optimal policy* $\widehat{\pi}$ *as long as (up to log factor)*

$$T \gtrsim O\left(\frac{H^4 S C^\star}{\epsilon^2}\right).$$

- model-free RL achieves optimal sample complexity for certain accuracy ϵ
- optimal in a larger accuracy range (improved by a factor of $H^{1.5}$)

$$\underbrace{\epsilon \leq \left(0, H^{-1}\right]}_{\text{(Our LCB-Q-Advantage)}} \qquad \text{vs.} \qquad \underbrace{\epsilon \leq \left(0, H^{-2.5}\right]}_{\text{(PEVI-Adv in [Xie et al., 2021])}}$$

Concluding remarks



Model-free RL matches the minimax-optimal sample complexity for model-based ones!

— *in a much larger range of the accuracy level*