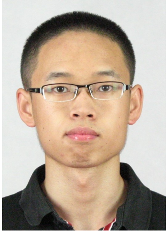
Minimax-optimal reward-agnostic exploration in reinforcement learning

Yuxin Chen

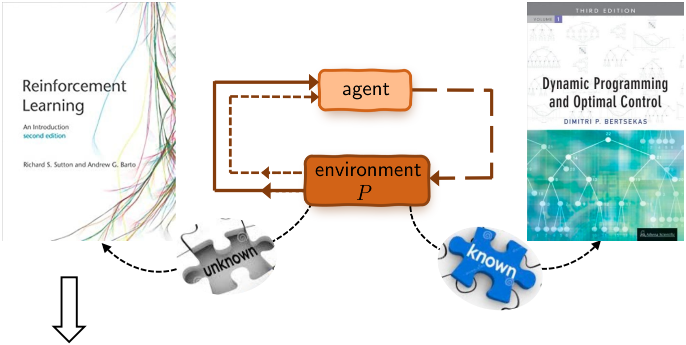
Statistics & ESE, UPenn



Gen Li
CUHK

Yuling Yan
MIT

Jianqing Fan
Princeton



In RL, we need to collect data to learn unknown environments

1. simulator (Li, Wei, Chi, Chen '24, Operations Research)
2. **online RL** **(Zhang, Chen, Lee, Du '24, COLT)**
3. **offline RL** **(Li, Shi, Chen, Chi, Wei '24, Annals. Stats)**

offline RL

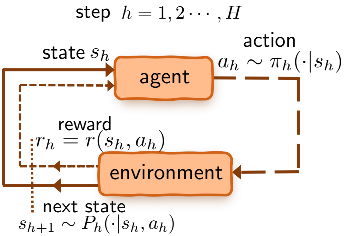
online RL

Key takeaway of this talk: insights from offline RL can inspire online RL algorithms

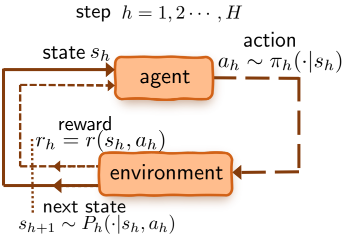
offline RL

reward-agnostic
online RL

Key takeaway of this talk: insights from offline RL can inspire (reward-agnostic) online RL algorithms

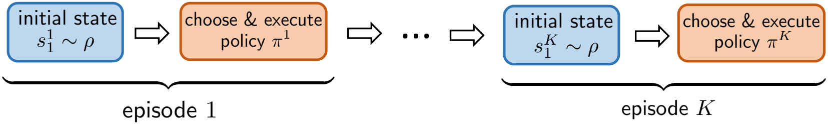


- H: horizon length (large)
- $\mathcal{S} = \{1, \ldots, S\}$: state space (large)
- $\mathcal{A} = \{1, \ldots, A\}$: action space (large)



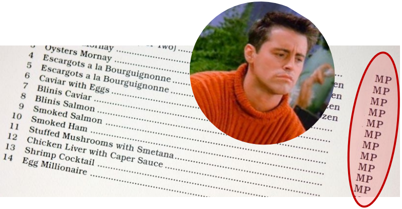
- H: horizon length (large)
- $\mathcal{S} = \{1, \dots, S\}$: state space (large)
- $\mathcal{A} = \{1, \dots, A\}$: action space (large)

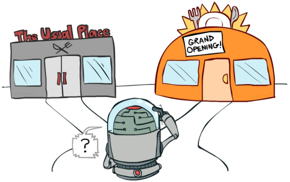
sequentially execute MDP for K episodes, each containing H steps



Reward-agnostic exploration?

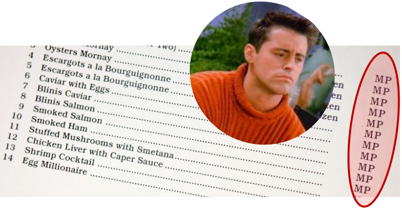
The learner is unaware of the rewards during exploration ...

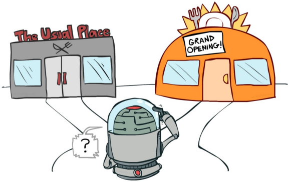




Reward-agnostic exploration?

The learner is unaware of the rewards during exploration ...



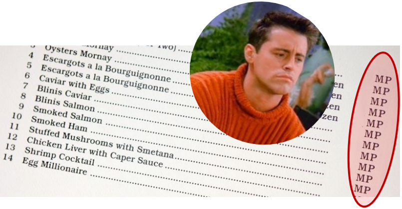


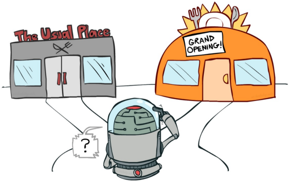
Motivation

- (significantly) delayed feedback
- reward functions keep changing
- many reward functions of interest

Reward-agnostic exploration?

The learner is unaware of the rewards during exploration . . .





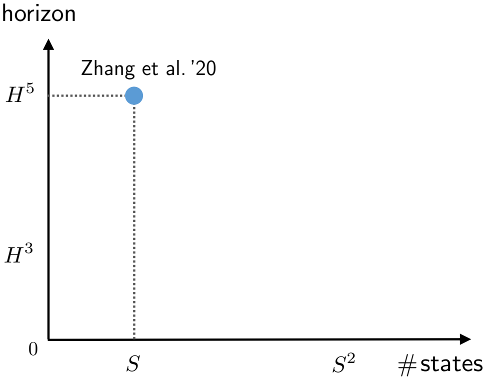
Motivation

- (significantly) delayed feedback
- reward functions keep changing
- many reward functions of interest

Question: can we perform pure exploration just once but achieve efficiency for many unseen reward functions at once?

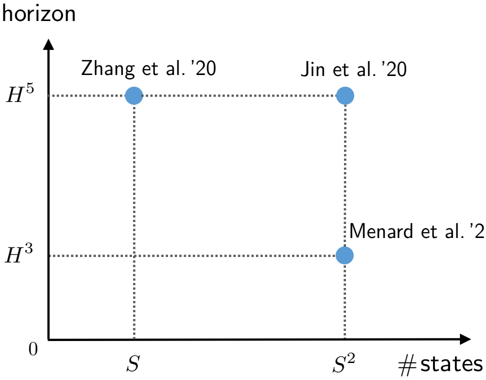
Prior art: sample complexity upper bounds

Suppose there is one fixed (but unseen) reward function of interest . . .



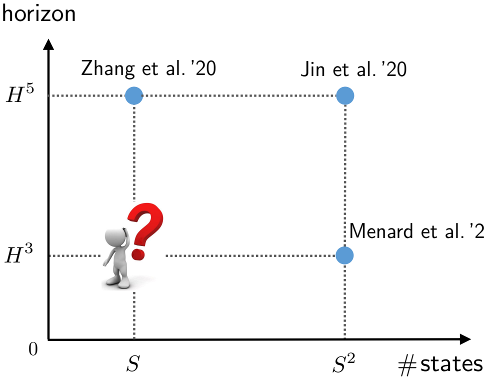
Prior art: sample complexity upper bounds

Suppose there is one fixed (but unseen) reward function of interest ...



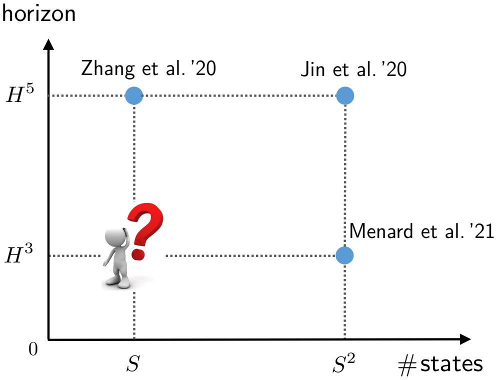
Prior art: sample complexity upper bounds

Suppose there is one fixed (but unseen) reward function of interest ...

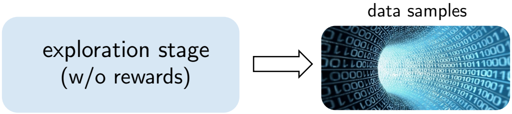


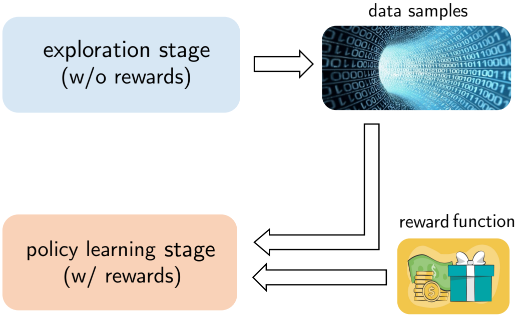
Prior art: sample complexity upper bounds

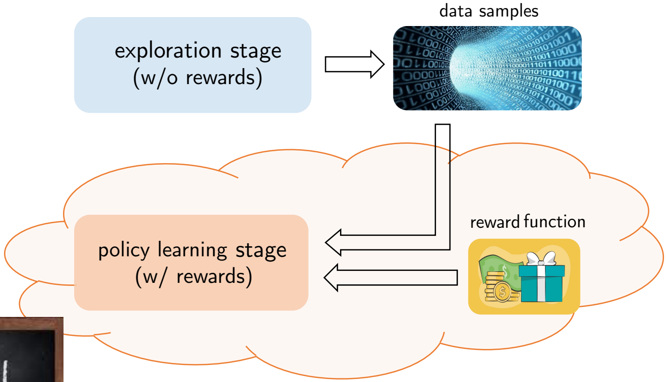
Suppose there is one fixed (but unseen) reward function of interest . . .



Question: can we simultaneously optimize dependency on S & H?


exploration stage
(w/o rewards)
data samples

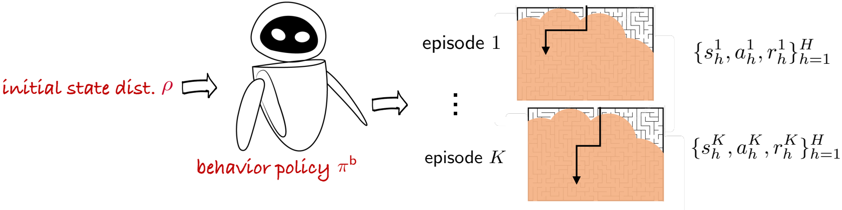

exploration stage
(w/o rewards)
data samples
policy learning stage
(w/ rewards)
reward function


exploration stage
(w/o rewards)
data samples
policy learning stage
(w/ rewards)
reward function


LESSON 1
Offline RL

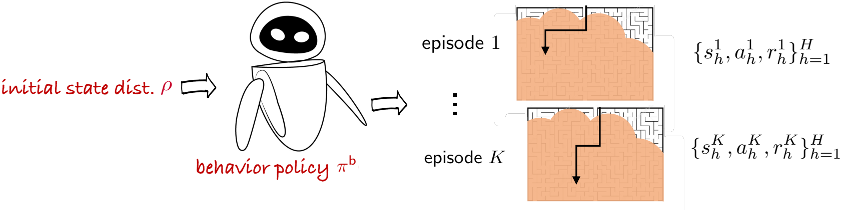
A mathematical model for offline RL

A historical dataset $\mathcal{D}$ containing K episodes generated by π^{b}:

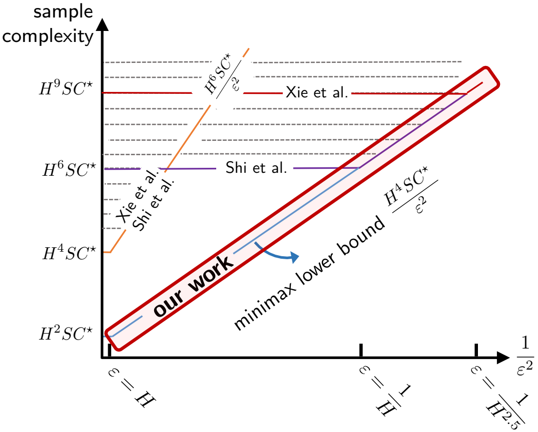


A mathematical model for offline RL

A historical dataset $\mathcal{D}$ containing K episodes generated by π^{b}:



- single-policy concentrability coefficient: $C^\star := \left\| \frac{d^{\pi^\star}}{d^{\pi^{\mathsf{b}}}} \right\|_\infty$



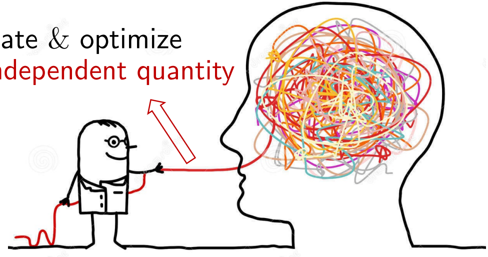
Theorem 1 (Li, Shi, Chen, Chi, Wei '24)

For any $0 < \varepsilon \le H$, we can design a pessimistic model-based algorithm that achieves $V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \le \varepsilon$ with

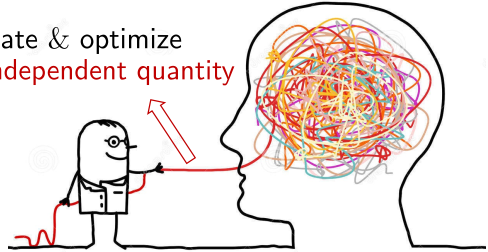
$$\widetilde{O}\left(\frac{H^3SC^\star}{\varepsilon^2}\right) \text{ episodes} \quad \text{or} \quad \widetilde{O}\left(\frac{H^4SC^\star}{\varepsilon^2}\right) \text{ samples}$$

isolate & optimize

reward-independent quantity



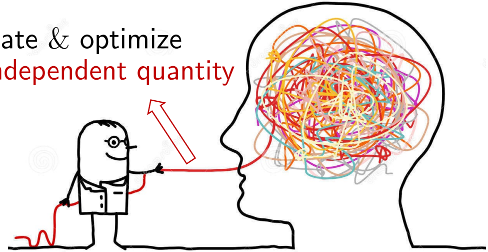
isolate & optimize
reward-independent quantity



lessons learned from offline RL: offline model-based alg. gives

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \lesssim \frac{1}{\sqrt{K}} \sum_{h,s,a} d_h^{\pi^\star}(s,a) \min\left\{ \sqrt{\frac{\mathsf{Var}_{h,s,a}(V_{h+1}^\star)}{d_h^{\mathsf{behavior}}(s,a)}},\ H \right\}$$

isolate & optimize
reward-independent quantity

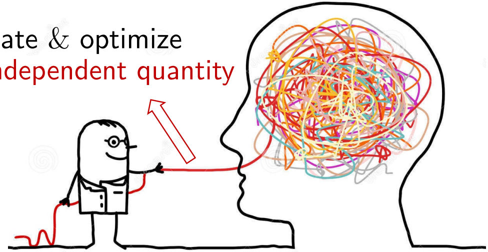


lessons learned from offline RL: offline model-based alg. gives

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \lesssim \frac{1}{\sqrt{K}} \sum_{h,s,a} d_h^{\pi^\star}(s,a) \min\left\{ \sqrt{\frac{\mathsf{Var}_{h,s,a}(V_{h+1}^\star)}{d_h^{\mathsf{behavior}}(s,a)}},\ H \right\}$$

$$\lesssim \frac{1}{\sqrt{K}} \underbrace{\left(\max_\pi \sum_{h,s,a} \frac{d_h^\pi(s,a)}{\frac{1}{KH} + d_h^{\mathsf{behavior}}(s,a)} \right)^{\frac{1}{2}}}_{\text{reward-independent}} \underbrace{\left(\sum_{h,s,a} d_h^{\pi^\star}(s,a) \mathsf{Var}_{h,s,a}(V_{h+1}^\star) + H \right)^{\frac{1}{2}}}_{\text{reward-dependent}}$$

isolate & optimize
reward-independent quantity



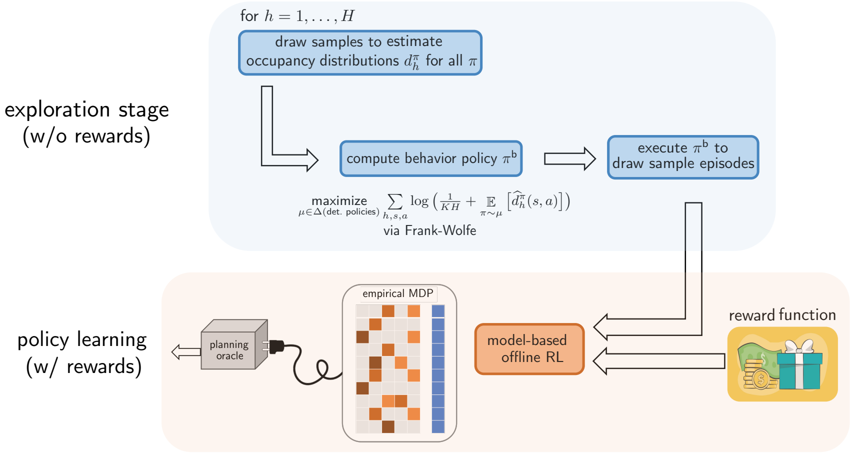
lessons learned from offline RL: offline model-based alg. gives

$$V_1^\star(\rho) - V_1^{\widehat{\pi}}(\rho) \lesssim \frac{1}{\sqrt{K}} \sum_{h,s,a} d_h^{\pi^\star}(s,a) \min \left\{ \sqrt{\frac{\mathsf{Var}_{h,s,a}(V_{h+1}^\star)}{d_h^{\mathsf{behavior}}(s,a)}},\, H \right\}$$

$$\lesssim \frac{1}{\sqrt{K}} \Bigg(\underbrace{\max_\pi \sum_{h,s,a} \frac{d_h^\pi(s,a)}{\frac{1}{KH} + d_h^{\mathsf{behavior}}(s,a)}}_{\text{reward-independent}} \Bigg)^{\frac{1}{2}} \Bigg(\underbrace{\sum_{h,s,a} d_h^{\pi^\star}(s,a)\mathsf{Var}_{h,s,a}(V_{h+1}^\star) + H}_{\text{reward-dependent}} \Bigg)^{\frac{1}{2}}$$

key: find behavior policy to optimize reward-independent quantity

Our algorithm



Main results

Theorem 2 (Li, Yan, Chen, Fan '23)

Suppose there are N fixed reward functions of interest, and suppose ε is small enough. Using the same batch of samples w/

$$\widetilde{O}\left(\frac{H^3 S A \log N}{\varepsilon^2}\right) \text{ episodes,}$$

our algorithm can find, for each reward function, a policy $\widehat{\pi}$ obeying

$$V_1^{\star}(\rho) - V_1^{\widehat{\pi}}(\rho) \le \varepsilon$$

Main results

Theorem 2 (Li, Yan, Chen, Fan '23)

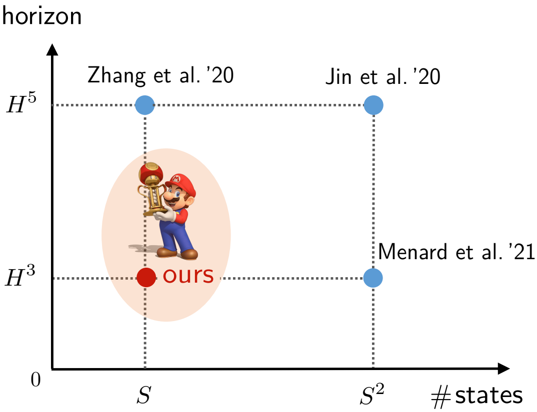
Suppose there are N fixed reward functions of interest, and suppose ε is small enough. Using the same batch of samples w/

$$\widetilde{O}\left(\frac{H^3 S A \log N}{\varepsilon^2}\right) \text{ episodes,}$$

our algorithm can find, for each reward function, a policy $\widehat{\pi}$ obeying

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

- optimal sample complexity
- collect data once $\longrightarrow$ work for poly(H, S, A) reward functions



The studies of offline RL inspire optimal reward-agnostic exploration!

Concluding remarks

Theoretical studies of offline RL shed light on data-efficient algorithm designs for other RL scenarios:

- online exploration
- hybrid RL
- ...

"Minimax-optimal reward-agnostic exploration in reinforcement learning," G. Li, Y. Yan, Y. Chen, J. Fan, *COLT* 2024

"Settling the sample complexity of model-based offline reinforcement learning," G. Li, L. Shi, Y. Chen, Y. Chi, Y. Wei, *Annals of Statistics*, 2024

"Reward-agnostic fine-tuning: provable statistical benefits of hybrid reinforcement learning," G. Li, W. Zhan, J. Lee, Y. Chi, Y. Chen, *NeurIPS* 2023