A Game Theoretic Approach to Offline Reinforcement Learning

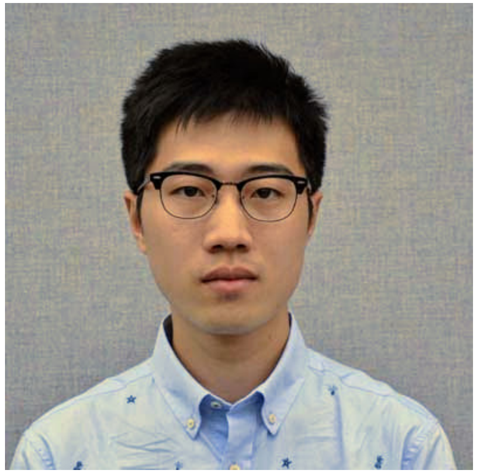
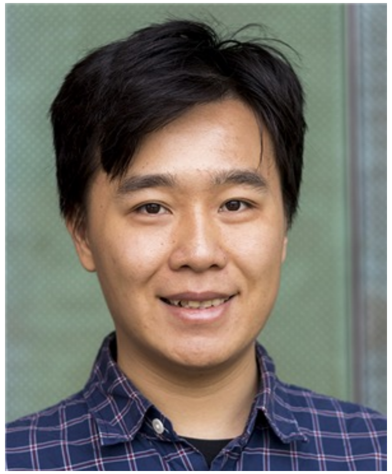
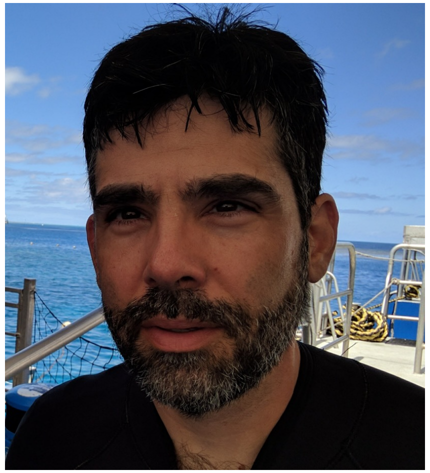
Ching-An Cheng

Robot Learning Group, Redmond

Microsoft

Acknowledgement

Tengyang Xie | Nan Jiang | Paul Mineiro | Alekh Agarwal


UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Microsoft

Google Research

Motivation

- Challenge of real-world decision-making problems

Data collection is costly and risky

How to make decisions under systematic uncertainty caused by missing data coverage?

Collected data lack diversity, despite quantity, as data can only be collected by qualified policies

Offline Reinforcement Learning

- **Goal**: learn good decision policies from non-exploratory datasets.

- **Core challenge:**

  Because of missing data coverage, in general, it's impossible to estimate how well a policy performs.

  How to optimize a policy without being able to estimate how well it performs?

How to understand a driving behavior is unsafe if all the data are safe?

Offline Reinforcement Learning

- **Principle of Pessimism:**

  Optimize performance lower bounds, that is, worst-case performance.

- But there're many ways to define and construct worst-case scenarios.

  How to properly trade off between conservatism and generalization??

How to understand a driving behavior is unsafe if all the data are safe?

A Game Theoretic Approach to Offline RL

Offline RL

$$\max_{\pi \in \Pi} \quad J(\pi)$$
$$s.t. \quad \text{no env interaction}$$

Maximize return in the true environment using data with partial coverage

Two-player Game

Learner Adversary

$$\max_{\pi\in\Pi} \underbrace{\min_{\hat{J}\in\mathcal{J}} \hat{J}(\pi)}_{\text{lower bound}}$$

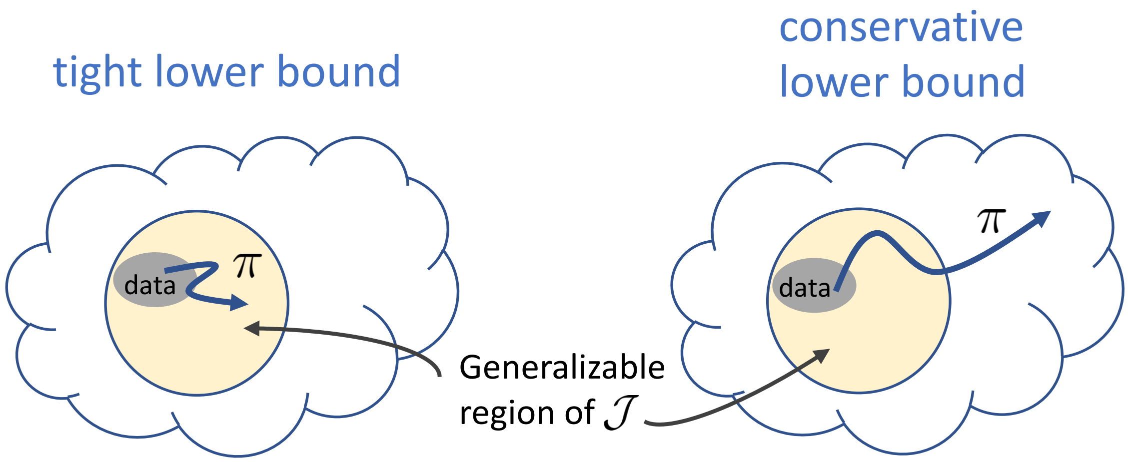
$$\text{s.t. } \hat{J} \text{ is data consistent}$$
$$J \in \mathcal{J}$$

Maximize a **performance lower bound** by a **two-player game**

A Game Theoretic Approach to Offline RL

Two-player game naturally handles the missing data uncertainty according to a prior hypothesis class $\mathcal{J}$.

Thus, the learned policy can generalize well!



Two-player Game

Learner Adversary

$$\max_{\pi\in\Pi} \underbrace{\min_{\hat{J}\in\mathcal{J}} \hat{J}(\pi)}_{\text{lower bound}}$$

$$\text{s.t. } \hat{J} \text{ is data consistent}$$

$$J \in \mathcal{J}$$

Maximize a **performance lower bound** by a **two-player game**

A Game Theoretic Approach to Offline RL

Outline

- A generic game-theoretic framework for designing offline RL algorithms
- Different concepts of pessimism
  - Absolute pessimism
  - Relative pessimism
- Robust policy improvement (RPI)

Two-player Game

Learner Adversary

$$\max_{\pi \in \Pi} \underbrace{\min_{\hat{J} \in \mathcal{J}} \hat{J}(\pi)}_{\text{lower bound}}$$

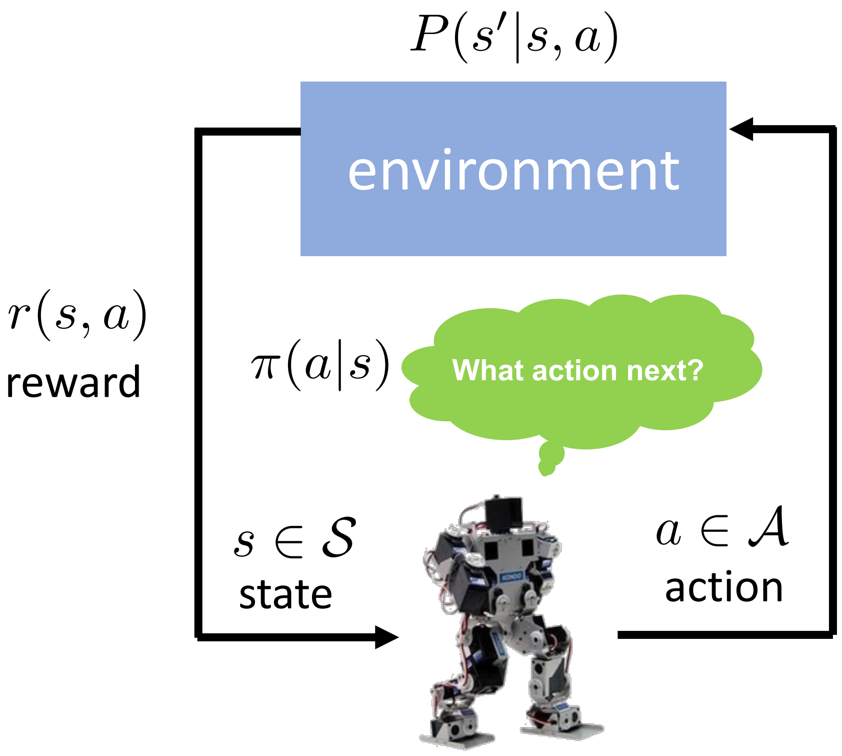
$$\text{s.t. } \hat{J} \text{ is data consistent}$$

$$J \in \mathcal{J}$$

Maximize a **performance lower bound** by a **two-player game**

Problem Setup

Suppose the world is a Markov decision process



$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$$

Offline setting assumption: offline data $\mathcal{D}$, collected by a behavior policy μ starting from s_0. No interaction with environment for learning.

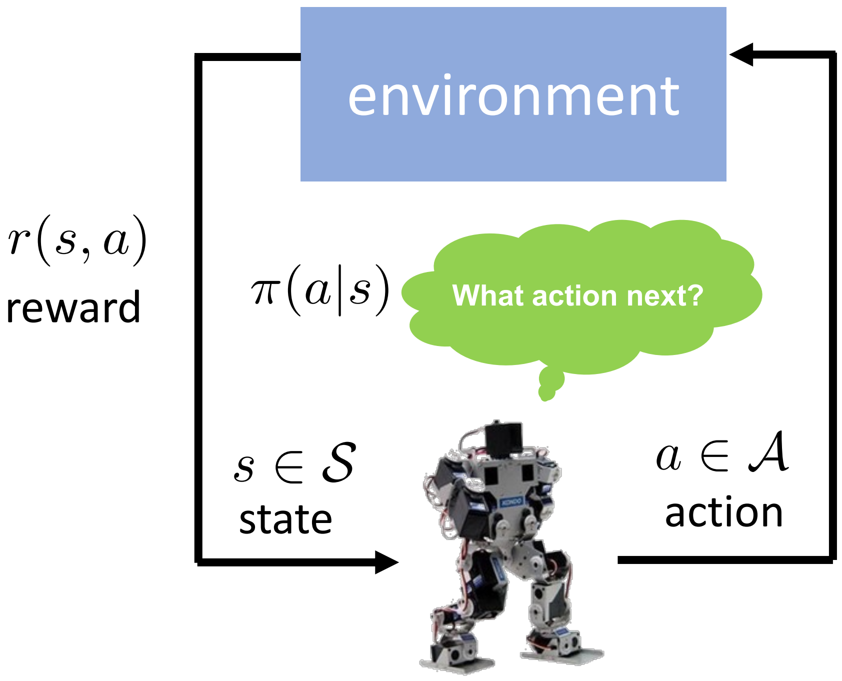
Goal: Find a policy π that has high return starting from s_0.

$$J(\pi) = \mathbb{E}_{d^\pi}[r(s,a)] = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$$

Problem Setup

Suppose the world is a Markov decision process

$P(s'|s,a)$



$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$$

This talk will focus on the **model-free version**.

Bellman operator and Q function

$$(\mathcal{T}^\pi f)(s,a) = r(s,a) + \gamma \mathbb{E}_{s'|s,a}[f(s,\pi)]$$

$$Q^\pi(s,a) = (\mathcal{T}^\pi Q^\pi)(s,a)$$

$$J(\pi) = Q^\pi(s_0,\pi)$$

Assumption: Given a function class $\mathcal{F}$ such that

- Realizability $Q^\pi \in \mathcal{F}$
- Completeness $\mathcal{T}^\pi \mathcal{F} \in \mathcal{F}$

Stackelberg Game for Offline RL

Each game is defined an objective ϕ and a regularization $\mathcal{E}$ to encourage data-consistency.

Learner optimizes policy

$$\hat{\pi}^* \in \arg\max_{\pi \in \Pi} \phi(\pi, f^\pi)$$

Leader

Adversary selects the worst-case hypothesis

$$f^\pi \in \arg\min_{f \in \mathcal{F}} \phi(\pi, f) + \beta \mathcal{E}(\pi, f)$$

Follower

Follower can also use a constrained version.

$$\max_{\pi \in \Pi} \min_{\hat{J} \in \mathcal{J}} \hat{J}(\pi)$$

$$\text{s.t. } \hat{J} \text{ is data consistent}$$

$$J \in \mathcal{J}$$

Maximize a **performance lower bound** by a **two-player game**

Stackelberg Game for Offline RL

Each game is defined an objective ϕ and a regularization $\mathcal{E}$ to encourage data-consistency.

Learner optimizes policy

$$\hat{\pi}^* \in \arg\max_{\pi \in \Pi} \phi(\pi, f^\pi)$$

Leader

Adversary selects the worst-case hypothesis

$$f^\pi \in \arg\min_{f \in \mathcal{F}} \phi(\pi, f) + \beta \mathcal{E}(\pi, f)$$

Follower

Follower can also use a constrained version.

Pessimism Lemma
If $\mathcal{E}(\pi, f) \geq 0$ and $\mathcal{E}(\pi, Q^\pi) = 0$
then $\phi(\pi, f^\pi) \leq \phi(\pi, Q^\pi), \forall \beta \geq 0$

Absolute Pessimism Game

$$\phi(\pi, f) := f(s_0, \pi)$$
$$\mathcal{E}(\pi, f) := \mathbb{E}_\mu[(f - \mathcal{T}^\pi f)^2]$$

Model-free

Bellman-consistent pessimism (Xie and **Cheng**, et al, 2021)

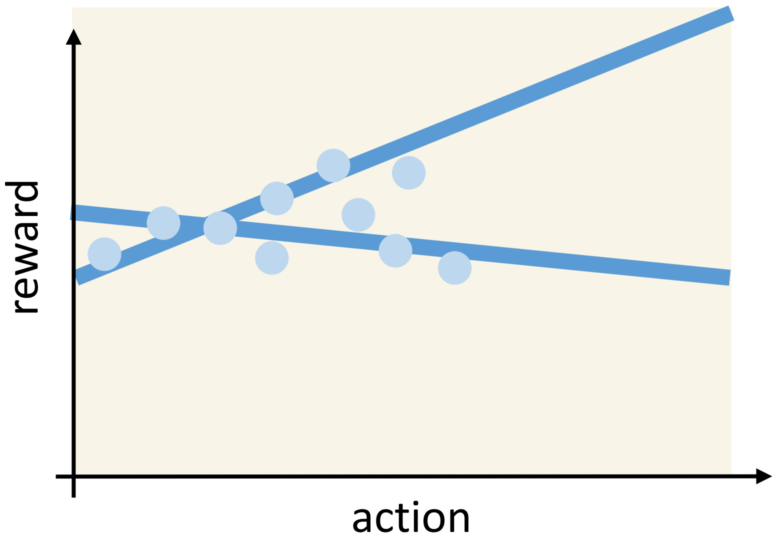
Since $J(\pi) = Q^\pi(s_0, \pi)$ by Pessimism Lemma, learner optimizes a performance LCB

$$\phi(\pi, f^\pi) \leq J(\pi) \quad \forall \beta \geq 0$$

This would imply for any comparator π'

$$J(\pi') - J(\hat{\pi}^*) \leq \underbrace{J(\pi') - \phi(\pi', f^{\pi'})}_{\text{underestimation error measured at the comparator}}$$

An Illustrative Example of Absolute Pessimism Game



data

hypothesis $f(s, \cdot)$ with small $\beta\mathcal{E}(\pi, f)$

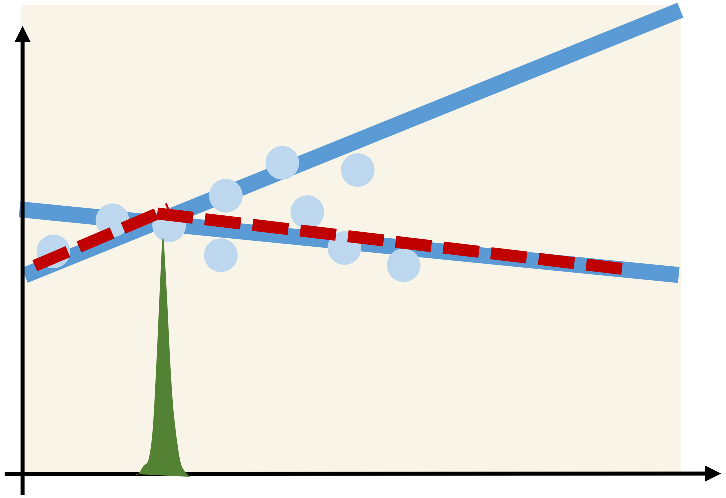
Let's use a toy example to compare

- **Absolute Pessimism Game**

- **Pointwise Pessimism**:
  Algorithms based on bonus/truncations (Kostrikov et al., 2021, Liu et al., 2020, Jin et al. 2021, Kidambi et al., 2020, Yu et al. 2020)

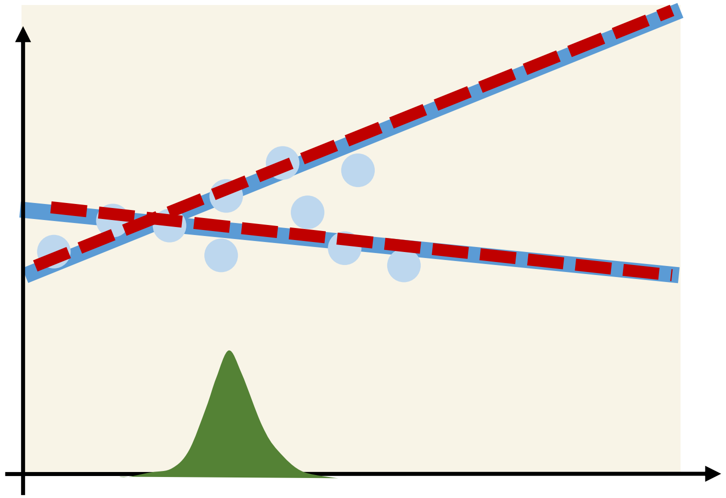
An Illustrative Example of Absolute Pessimism Game

Pointwise Pessimism



Multiple hypotheses are merged into a new hypothesis that may be outside the original hypothesis class

Absolute Pessimism Game



Learner needs to balance multiple hypotheses in the hypothesis class

data

hypothesis $f(s, \cdot)$ with small $\beta\mathcal{E}(\pi, f)$

learned policy

objective(s)

Solving the Stackelberg Game

$$\hat{\pi}^* = \arg\max_{\pi \in \Pi} \phi(\pi, f^\pi)$$

$$f^\pi \in \arg\min_{f \in \mathcal{F}} \phi(\pi, f) + \beta \mathcal{E}(\pi, f)$$

No-Regret + Best Response Scheme

Repeat for K iterations

$$f^k \in \arg\min_{f \in \mathcal{F}} \phi(\pi^k, f) + \beta \mathcal{E}(\pi^k, f)$$

ϕ, ψ are approximated by samples

$$\pi^{k+1} = \mathrm{NoRegret}(\pi^k, f^k)$$

e.g. $\pi^{k+1}(a|s) \propto \exp(\eta \sum_{i=1}^{k} f^i(s,a))$

Output uniform mixture of policies $\hat{\pi}$

For Absolute Pessimistic Game, this algorithm is known as PSPI (Pessimistic Soft Policy Iteration) (Xie and **Cheng**, et al., 2021)

Theory (Absolute Pessimism Game)

Learning Optimality

With a well tuned β, the learned policy can compete with any policy within the data coverage.

Assume $\mathcal{F}$ satisfies realizability and completeness.
Given dataset $\mathcal{D}$ s.t. $|\mathcal{D}| = N$. With $\beta = \sqrt[3]{\frac{V_{\max} N^2}{d^2_{\mathcal{F},\Pi}}}$. Then $\forall \pi \in \Pi$,

$$J(\pi) - J(\hat{\pi}) \leq \mathcal{O}\left(\frac{\sqrt{C} V_{\max}}{(1-\gamma)} \sqrt[3]{\frac{d_{\mathcal{F},\Pi}}{N}} \right) + \frac{\sum_{k=1}^{K} \mathbb{E}_{d^\pi \backslash \nu}[e^k]}{K(1-\gamma)} + \mathcal{O}\left(\frac{V_{\max}}{1-\gamma} \frac{\text{Regret}(K)}{K} \right)$$

where $\nu(s,a)$ is any distribution satisfying $\max_{k \in [K]} \frac{\mathbb{E}_\nu[e_k^2]}{\mathbb{E}_\mu[e_k^2]} \leq C$,

$d^\pi \backslash \nu(s,a) = \max(d^\pi(s,a) - \nu(s,a), 0)$

$e^k = f^k - \mathcal{T}^{\pi^k} f^k \qquad d_{\mathcal{F},\Pi} = \log \frac{|\mathcal{F}||\Pi|}{\delta}$

Theory (Absolute Pessimism Game)

Learning Optimality

With a well tuned β, the learned policy can compete with any policy within the data coverage.

Assume $\mathcal{F}$ satisfies realizability and completeness.
Given dataset $\mathcal{D}$ s.t. $|\mathcal{D}| = N$. With $\beta = \sqrt[3]{\frac{V_{\max} N^2}{d^2_{\mathcal{F},\Pi}}}$. Then $\forall \pi \in \Pi$,

$$J(\pi) - J(\hat{\pi}) \leq \mathcal{O}\left(\frac{\sqrt{C} V_{\max}}{(1-\gamma)} \sqrt[3]{\frac{d_{\mathcal{F},\Pi}}{N}}\right) + \frac{\sum_{k=1}^{K} \mathbb{E}_{d^\pi \setminus \nu}[e^k]}{K(1-\gamma)} + \mathcal{O}\left(\frac{V_{\max}}{1-\gamma} \frac{\mathrm{Regret}(K)}{K}\right)$$

In-Support Error

where $\nu(s,a)$ is any distribution satisfying $\max_{k \in [K]} \frac{\mathbb{E}_\nu[e_k^2]}{\mathbb{E}_\mu[e_k^2]} \leq C$,

$$d^\pi \setminus \nu(s,a) = \max(d^\pi(s,a) - \nu(s,a), 0)$$

$$e^k = f^k - \mathcal{T}^{\pi^k} f^k \qquad d_{\mathcal{F},\Pi} = \log \frac{|\mathcal{F}||\Pi|}{\delta}$$

Theory (Absolute Pessimism Game)

With a well tuned β, the learned policy can compete with any policy within the data coverage.

Learning Optimality

Assume $\mathcal{F}$ satisfies realizability and completeness.
Given dataset $\mathcal{D}$ s.t. $|\mathcal{D}| = N$. With $\beta = \sqrt[3]{\frac{V_{\max} N^2}{d^2_{\mathcal{F},\Pi}}}$. Then $\forall \pi \in \Pi$,

$$J(\pi) - J(\hat{\pi}) \leq \mathcal{O}\left(\frac{\sqrt{C}V_{\max}}{(1-\gamma)}\sqrt[3]{\frac{d_{\mathcal{F},\Pi}}{N}}\right) + \frac{\sum_{k=1}^{K}\mathbb{E}_{d^\pi \setminus \nu}[e^k]}{K(1-\gamma)} + \mathcal{O}\left(\frac{V_{\max}}{1-\gamma}\frac{\mathrm{Regret}(K)}{K}\right)$$

In-Support Error Out-of-Support Error

where $\nu(s,a)$ is any distribution satisfying $\max_{k\in[K]} \frac{\mathbb{E}_\nu[e_k^2]}{\mathbb{E}_\mu[e_k^2]} \leq C$,

$d^\pi \setminus \nu(s,a) = \max(d^\pi(s,a) - \nu(s,a), 0)$

$e^k = f^k - \mathcal{T}^{\pi^k} f^k \qquad d_{\mathcal{F},\Pi} = \log \frac{|\mathcal{F}||\Pi|}{\delta}$

Theory (Absolute Pessimism Game)

With a well tuned β, the learned policy can compete with any policy within the data coverage.

Learning Optimality

Assume $\mathcal{F}$ satisfies realizability and completeness.
Given dataset $\mathcal{D}$ s.t. $|\mathcal{D}| = N$. With $\beta = \sqrt[3]{\frac{V_{\max} N^2}{d^2_{\mathcal{F},\Pi}}}$. Then $\forall \pi \in \Pi$,

$$J(\pi) - J(\hat{\pi}) \leq \underbrace{\mathcal{O}\left(\frac{\sqrt{C} V_{\max}}{(1-\gamma)} \sqrt[3]{\frac{d_{\mathcal{F},\Pi}}{N}}\right)}_{\text{In-Support Error}} + \underbrace{\frac{\sum_{k=1}^{K} \mathbb{E}_{d^\pi \setminus \nu}[e^k]}{K(1-\gamma)}}_{\text{Out-of-Support Error}} + \underbrace{\mathcal{O}\left(\frac{V_{\max}}{1-\gamma} \frac{\mathrm{Regret}(K)}{K}\right)}_{\text{Optimization Error in } o(1)}$$

where $\nu(s,a)$ is any distribution satisfying $\max_{k \in [K]} \frac{\mathbb{E}_\nu[e_k^2]}{\mathbb{E}_\mu[e_k^2]} \leq C$,

$d^\pi \setminus \nu(s,a) = \max(d^\pi(s,a) - \nu(s,a), 0)$

$e^k = f^k - \mathcal{T}^{\pi^k} f^k \qquad d_{\mathcal{F},\Pi} = \log \frac{|\mathcal{F}||\Pi|}{\delta}$

Theory (Absolute Pessimism Game)

- Proof Sketch:

$$J(\pi) - J(\hat{\pi}) = \frac{1}{K} \sum_{k=1}^{K} J(\pi) - J(\pi^k)$$

Theory (Absolute Pessimism Game)

- Proof Sketch:

$$r^k(s,a) := f^k(s,a) - \gamma \mathbb{E}_{s'|s,a}[f^k(s',\pi^k)]$$

$$\begin{aligned}
J(\pi) - J(\hat{\pi}) =& \frac{1}{K}\sum_{k=1}^{K} J(\pi) - J(\pi^k) \\
=& \frac{1}{K}\sum_{k=1}^{K} \big(J(\pi) - \mathbb{E}_{d^\pi}[r^k]\big) + \big(\mathbb{E}_{d^\pi}[r^k] - f^k(s_0,\pi^k)\big) + \big(f^k(s_0,\pi^k) - J(\pi^k)\big)
\end{aligned}$$

Theory (Absolute Pessimism Game)

- Proof Sketch:

$$r^k(s,a) := f^k(s,a) - \gamma \mathbb{E}_{s'|s,a}[f^k(s',\pi^k)]$$

$$J(\pi) - J(\hat{\pi}) = \frac{1}{K}\sum_{k=1}^{K} J(\pi) - J(\pi^k)$$

$$= \frac{1}{K}\sum_{k=1}^{K} \boxed{\left(J(\pi) - \mathbb{E}_{d^\pi}[r^k]\right)} + \boxed{\left(\mathbb{E}_{d^\pi}[r^k] - f^k(s_0,\pi^k)\right)} + \left(f^k(s_0,\pi^k) - J(\pi^k)\right)$$

$$= \frac{1}{K}\sum_{k=1}^{K} \underbrace{\boxed{\mathbb{E}_{d^\pi}[e^k]}}_{\text{Bellman error}} + \underbrace{\boxed{\mathbb{E}_{d^\pi}[f^k(s,\pi) - f^k(s,\pi^k)]}}_{\text{regret}} + \underbrace{\left(f^k(s_0,\pi^k) - J(\pi^k)\right)}_{\leq 0}$$

$$e^k = f^k - \mathcal{T}^{\pi^k} f^k$$

Theory (Absolute Pessimism Game)

- Proof Sketch:

$$r^k(s,a) := f^k(s,a) - \gamma \mathbb{E}_{s'|s,a}[f^k(s',\pi^k)]$$

$$J(\pi) - J(\hat{\pi}) = \frac{1}{K}\sum_{k=1}^{K} J(\pi) - J(\pi^k)$$

$$= \frac{1}{K}\sum_{k=1}^{K} \left(J(\pi) - \mathbb{E}_{d^\pi}[r^k]\right) + \left(\mathbb{E}_{d^\pi}[r^k] - f^k(s_0,\pi^k)\right) + \left(f^k(s_0,\pi^k) - J(\pi^k)\right)$$

$$= \frac{1}{K}\sum_{k=1}^{K} \boxed{\mathbb{E}_{d^\pi}[e^k]} + \boxed{\mathbb{E}_{d^\pi}[f^k(s,\pi) - f^k(s,\pi^k)]} + \underbrace{\left(f^k(s_0,\pi^k) - J(\pi^k)\right)}_{\leq 0}$$

$$e^k = f^k - \mathcal{T}^{\pi^k} f^k$$

$$J(\pi) - J(\hat{\pi}) \leq \mathcal{O}\left(\frac{\sqrt{C}V_{\max}}{(1-\gamma)}\sqrt[3]{\frac{d_{\mathcal{F},\Pi}}{N}}\right) + \frac{\sum_{k=1}^{K}\mathbb{E}_{d^\pi\setminus\nu}[e^k]}{K(1-\gamma)} + \mathcal{O}\left(\frac{V_{\max}}{1-\gamma}\frac{\mathrm{Regret}(K)}{K}\right)$$

In-Support Error | Out-of-Support Error | Optimization Error in $o(1)$

$$\beta = \sqrt[3]{\frac{V_{\max} N^2}{d^2_{\mathcal{F},\Pi}}}$$

What is missing?

- In the offline setting, it is hard to tune hyperparameters, but when β (i.e., the degree of pessimism) is selected incorrectly, we lose the guarantees. When β is wrong, the learned can be even worse than the behavior policy! Same for other LCB-based algorithms.

- *Why?* Recall, by optimizing LCB, we have

$$J(\pi') - J(\hat{\pi}^*) \leq \boxed{J(\pi') - \phi(\pi', f^{\pi'})}$$

But this gap depends on β

Can we design offline RL algorithms that are robust to hyperparameter selection?

Relative Pessimism Game

(**Cheng*** and Xie* et al, 2022)

Absolute Pessimism Game

$$\phi(\pi, f) := f(s_0, \pi)$$
$$\mathcal{E}(\pi, f) := \mathbb{E}_\mu[(f - \mathcal{T}^\pi f)^2]$$

Model-free

Relative Pessimism Game

$$\phi(\pi, f) := \mathbb{E}_{d^\mu}[f(s, \pi) - f(s, a)]$$
$$\mathcal{E}(\pi, f) := \mathbb{E}_\mu[(f - \mathcal{T}^\pi f)^2]$$

Model-free

$J(\pi) = Q^\pi(s_0, \pi)$	**Objective**	$J(\pi) - J(\mu) = \mathbb{E}_{d^\mu}[Q^\pi(s, \pi) - Q^\pi(s, a)]$

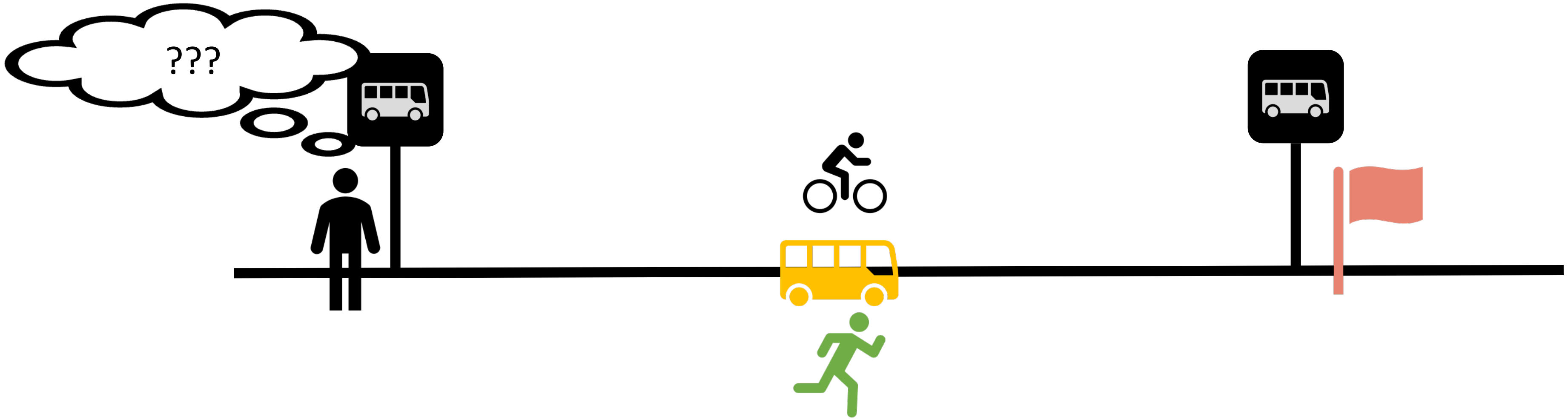
We can solve this Stackelberg Game with the same no-regret + best response scheme. This algorithm is known as **ATAC** (Adversarially Trained Actor Critic) (**Cheng*** and Xie, et al, 2022)

While optimizing the two is the same in online RL, **the results are different in the offline case!** Because the agent cannot explore to reduce the uncertainty due to partial data coverage.

Absolute Pessimism vs Relative Pessimism

Hypothesis class

1. Good traffic: Bus 5 min, Walk 30 min, Bike 20 min
2. Bad traffic: Bus 30 min, Walk 30 min, Bike 30 min



Absolute Pessimism vs Relative Pessimism

Hypothesis class

1. Good traffic: Bus 5 min, Walk 30 min, Bike 20 min
2. Bad traffic: Bus 30 min, Walk 30 min, Bike 30 min

Absolute Time

	Bus	Walk	Bike
Case 1	10	30	20
Case 2	30	30	30

Absolute Pessimism

Either

Absolute Pessimism vs Relative Pessimism

Hypothesis class

1. Good traffic: Bus 5 min, Walk 30 min, Bike 20 min
2. Bad traffic: Bus 30 min, Walk 30 min, Bike 30 min

Relative Time to Bus

	Bus	Walk	Bike
Case 1	0	25	15
Case 2	0	0	0

Absolute Pessimism

Either

Relative Pessimism

Take Bus!

Relative Pessimism Game

(**Cheng*** and Xie* et al, 2022)

Absolute Pessimism Game		Relative Pessimism Game
$\phi(\pi, f) \coloneqq f(s_0, \pi)$ $\mathcal{E}(\pi, f) \coloneqq \mathbb{E}_\mu[(f - \mathcal{T}^\pi f)^2]$ Model-free		$\phi(\pi, f) \coloneqq \mathbb{E}_{d^\mu}[f(s, \pi) - f(s, a)]$ $\mathcal{E}(\pi, f) \coloneqq \mathbb{E}_\mu[(f - \mathcal{T}^\pi f)^2]$ Model-free
$J(\pi) = Q^\pi(s_0, \pi)$	**Objective**	$J(\pi) - J(\mu) = \mathbb{E}_{d^\mu}[Q^\pi(s, \pi) - Q^\pi(s, a)]$

We can solve this Stackelberg Game with the same no-regret + best response scheme. This algorithm is known as **ATAC** (Adversarially Trained Actor Critic) (**Cheng*** and Xie, et al, 2022

While optimizing the two is the same in online RL, **the results are different in the offline case!**
Because the agent cannot explore to reduce the uncertainty due to partial data coverage.

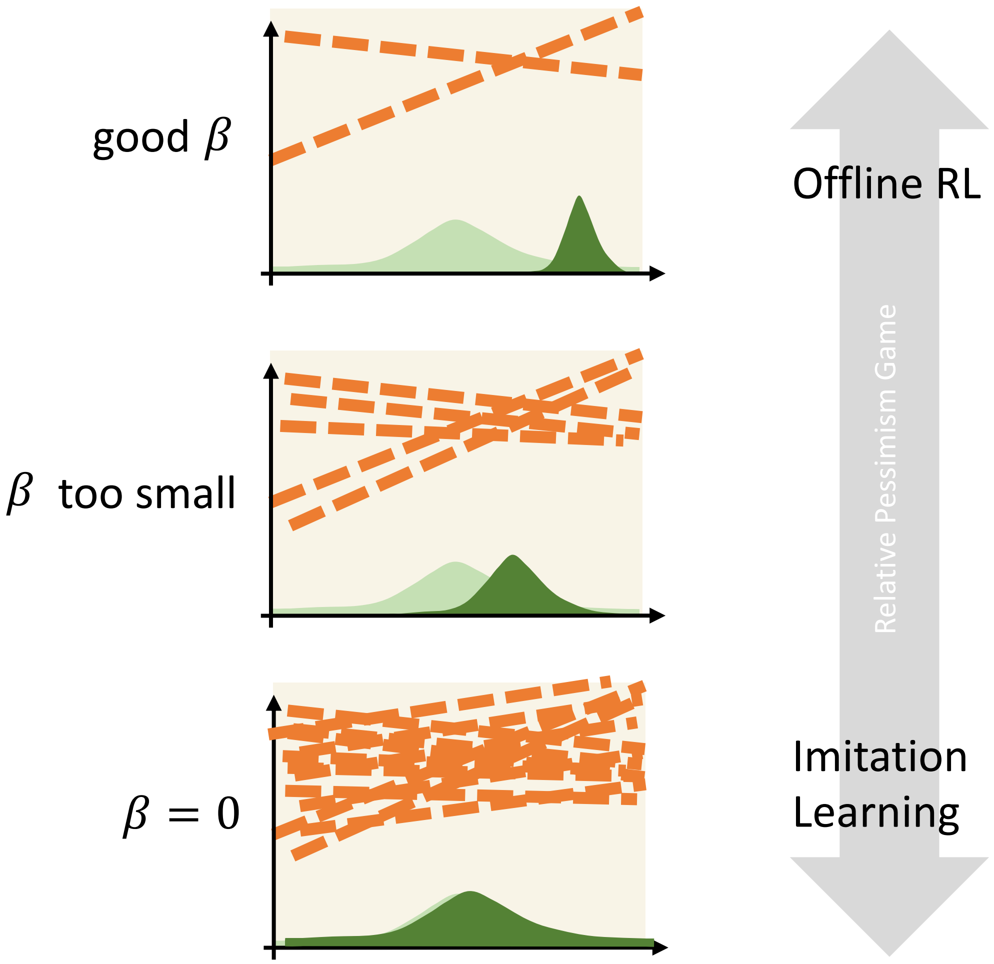
Relative Pessimism Game

(**Cheng*** and Xie* et al, 2022)

Absolute Pessimism Game		Relative Pessimism Game
$\phi(\pi, f) := f(s_0, \pi)$ $\mathcal{E}(\pi, f) := \mathbb{E}_\mu[(f - \mathcal{T}^\pi f)^2]$ **Model-free**		$\phi(\pi, f) := \mathbb{E}_{d^\mu}[f(s, \pi) - f(s, a)]$ $\mathcal{E}(\pi, f) := \mathbb{E}_\mu[(f - \mathcal{T}^\pi f)^2]$ **Model-free**
$J(\pi) = Q^\pi(s_0, \pi)$	**Objective**	$J(\pi) - J(\mu) = \mathbb{E}_{d^\mu}[Q^\pi(s, \pi) - Q^\pi(s, a)]$
	Pessimism Lemma	
$J(\pi) \geq \phi(\pi, f^\pi) \quad \forall \beta \geq 0$	**Lower bound**	$J(\pi) - J(\mu) \geq \phi(\pi, f^\pi) \quad \forall \beta \geq 0$
	Optimizing LCB	
$J(\pi') - J(\hat{\pi}^*) \leq J(\pi') - \phi(\pi', f^{\pi'})$	**Performance**	$J(\pi') - J(\hat{\pi}^*) \leq J(\pi') - J(\mu) - \phi(\pi', f^{\pi'})$

Robust Policy Improvement (RPI) $J(\hat{\pi}^*) - J(\mu) \geq \phi(\hat{\pi}^*, f^{\hat{\pi}^*}) \geq \phi(\mu, f^\mu) = 0 \ \forall \beta \geq 0$

Source of Robust Policy Improvement



Relative Pessimism Game (ATAC)

$$\begin{aligned} &\widehat{\pi}^* \in \arg\max_{\pi \in \Pi} \mathbb{E}_{d^\mu}[f(s,\pi) - f(s,a)] \\ \text{s.t.} \quad & f^\pi \in \arg\min_{f \in \mathcal{F}} \mathbb{E}_{d^\mu}[f(s,\pi) - f(s,a)] + \beta \mathcal{E}(\pi, f) \end{aligned}$$

Actor = Conditional generator
Critic = Discriminator

Relative Pessimism Game provides a bridge between offline RL and imitation learning with IPM via the lens of generative adversarial networks (GAN)

Offline RL + Relative Pessimism
= IL + Bellman Regularization

Theory (Relative Pessimism Game)

With a well tuned β, the learned policy can compete with any policy within the data coverage.

Learning Optimality

Same as Absolute Pessimism!

$$J(\pi) - J(\hat{\pi}) \leq \mathcal{O}\left(\frac{\sqrt{C}V_{\max}}{(1-\gamma)}\sqrt[3]{\frac{d_{\mathcal{F},\Pi}}{N}}\right) + \frac{\sum_{k=1}^{K}\mathbb{E}_{d^\pi\setminus\nu}[e^k]}{K(1-\gamma)} + \mathcal{O}\left(\frac{V_{\max}}{1-\gamma}\frac{\mathrm{Regret}(K)}{K}\right)$$

In-Support Error | Out-of-Support Error | Optimization Error in $o(1)$

where $\nu(s,a)$ is any distribution satisfying $\max_{k\in[K]} \frac{\mathbb{E}_\nu[e_k^2]}{\mathbb{E}_\mu[e_k^2]} \leq C$,

$d^\pi\setminus\nu(s,a) = \max(d^\pi(s,a) - \nu(s,a), 0)$

$e^k = f^k - \mathcal{T}^{\pi^k} f^k \qquad d_{\mathcal{F},\Pi} = \log\frac{|\mathcal{F}||\Pi|}{\delta}$

Theory (Relative Pessimism Game)

Robust Policy Improvement (RPI)

The learned policy always improves the behavior policy so long as $\beta = o(N)$.

Assume $\mathcal{F}$ satisfies realizability, though not necesssarily completeness. Given dataset $\mathcal{D}$ s.t. $|\mathcal{D}| = N$. For any $\beta \geq 0$, if $\mu \in \Pi$, then

$$J(\mu) - J(\hat{\pi}) \leq \mathcal{O}\left(\frac{V_{\max}}{(1-\gamma)} \sqrt{\frac{d_{\mathcal{F},\Pi}}{N}} + \frac{\beta V_{\max}^2 d_{\mathcal{F},\Pi}}{(1-\gamma)N} + \frac{V_{\max}}{1-\gamma} \frac{\mathrm{Regret}(K)}{K} \right)$$

Statistical Error

Optimization Error in $o(1)$

Theory (Relative Pessimism Game)

- Proof Sketch (Robust Policy Improvement):

$$
\begin{aligned}
J(\mu) - J(\hat{\pi}) =& \frac{1}{K} \sum_{k=1}^{K} J(\mu) - J(\pi^k) \\
=& \frac{1}{K} \sum_{k=1}^{K} \big(J(\mu) - J(\pi^k) + \phi(\pi^k, f^k) \big) - \phi(\pi^k, f^k)
\end{aligned}
$$

Theory (Relative Pessimism Game)

- Proof Sketch (Robust Policy Improvement):

$$
\begin{aligned}
J(\mu) - J(\hat{\pi}) &= \frac{1}{K}\sum_{k=1}^{K} J(\mu) - J(\pi^k) \\
&= \frac{1}{K}\sum_{k=1}^{K} \big(J(\mu) - J(\pi^k) + \phi(\pi^k, f^k)\big) - \phi(\pi^k, f^k) \\
&= \frac{1}{K}\sum_{k=1}^{K} \big(J(\mu) - J(\pi^k) + \phi(\pi^k, f^k)\big) + \mathbb{E}_{d^\mu}[f^k(s,\mu) - f^k(s,\pi^k)]
\end{aligned}
$$

Theory (Relative Pessimism Game)

- Proof Sketch (Robust Policy Improvement):

$$J(\mu) - J(\hat{\pi}) = \frac{1}{K}\sum_{k=1}^{K} J(\mu) - J(\pi^k)$$

$$= \frac{1}{K}\sum_{k=1}^{K} \left(J(\mu) - J(\pi^k) + \phi(\pi^k, f^k)\right) - \phi(\pi^k, f^k)$$

$$= \frac{1}{K}\sum_{k=1}^{K} \boxed{\left(J(\mu) - J(\pi^k) + \phi(\pi^k, f^k)\right)} + \boxed{\mathbb{E}_{d^\mu}\left[f^k(s,\mu) - f^k(s,\pi^k)\right]}$$

Ideally ≤ 0 but since ϕ is estimated by finite samples

$$J(\mu) - J(\hat{\pi}^*) \leq \mathcal{O}\left(\boxed{\frac{V_{\max}}{(1-\gamma)}\sqrt{\frac{d_{\mathcal{F},\Pi}}{N}} + \frac{\beta V_{\max}^2 d_{\mathcal{F},\Pi}}{(1-\gamma)N}} + \boxed{\frac{V_{\max}}{1-\gamma}\frac{\mathrm{Regret}(K)}{K}}\right)$$

Statistical Error

Optimization Error in $o(1)$

Comparison of Offline RL Approaches

	Relative Pessimism	**Absolute Pessimism**
Game-Theoretic	Relative Pessimism Game (Cheng et al. 2022) RPI Less Conservative	Absolute Pessimism Game (Xie et al., 2021, Uehara et al., 2021) Less Conservative
Single MDP	Behavior regularization (Fujimoto et al. 2019,2021, Kuma et al., 2019, Laroche et al. 2019) RPI Simple	Algorithms based on bonus/truncations (Kostrikov et al., 2021, Liu et al., 2020, Jin et al. 2021, Kidambi et al., 2020, Yu et al. 2020) Simple

Practical Implementation

- We approximate the No-Regret + Best Response scheme by a two-timescale stochastic gradient update rule.

No-Regret + Best Response Scheme

$$f^k \in \arg\min_{f \in \mathcal{F}} \phi(\pi^k, f) + \beta \mathcal{E}(\pi^k, f)$$

$$\pi^{k+1} = \mathrm{NoRegret}(\pi^k, f^k)$$

We trained NN policies and values on the D4RL benchmark and compare the results with other deep offline RL algorithms (CQL, COMBO, IQL, TD3+BC).

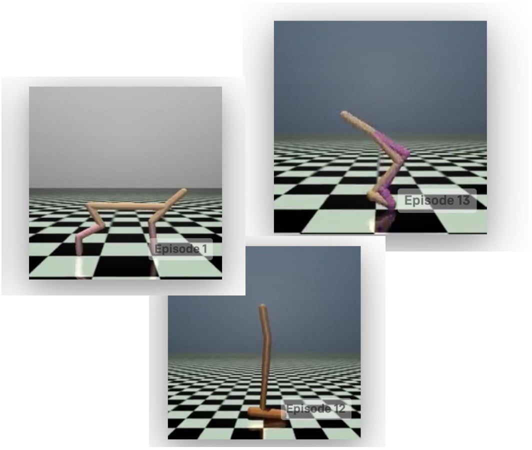
Algorithm

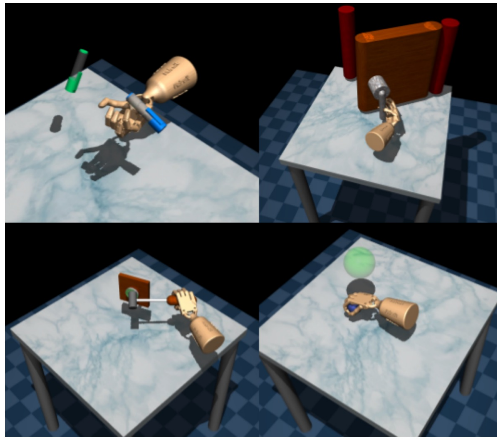
Input: Batch data $\mathcal{D}$, policy π, critics f_1, f_2, constants $\beta \geq 0$, $\tau \in [0,1]$, $w \in [0,1]$

1: Initialize target networks $\bar{f}_1 \leftarrow f_1$, $\bar{f}_2 \leftarrow f_2$
2: **for** $k = 1, 2, \ldots, K$ **do**
3: Sample minibatch $\mathcal{D}_{\text{mini}}$ from dataset $\mathcal{D}$.
4: For $f \in \{f_1, f_2\}$, update critic networks
$l_{\text{critic}} = \phi_{\mathcal{D}_{\text{mini}}}(f, \pi) + \beta \mathcal{E}_{\mathcal{D}_{\text{mini}}}(f, \pi)$
$f \leftarrow \mathrm{Proj}_{\mathcal{F}}(f - \eta_{\text{fast}} \nabla l_{\text{critic}})$
5: Update actor network
$l_{\text{actor}} = -\phi_{\mathcal{D}_{\text{mini}}}(f, \pi)$
$\pi \leftarrow \mathrm{Proj}_{\Pi}(\pi - \eta_{\text{slow}} \nabla l_{\text{actor}})$
6: For $(f, \bar{f}) \in \{(f_i, \bar{f}_i)\}_{i=1,2}$, update target networks
$\bar{f} \leftarrow (1-\tau)\bar{f} + \tau f$.
7: **end for**

Experimental Results

PSPI (Absolute Pessimism) outperforms baseline algorithms in 17/24 datasets



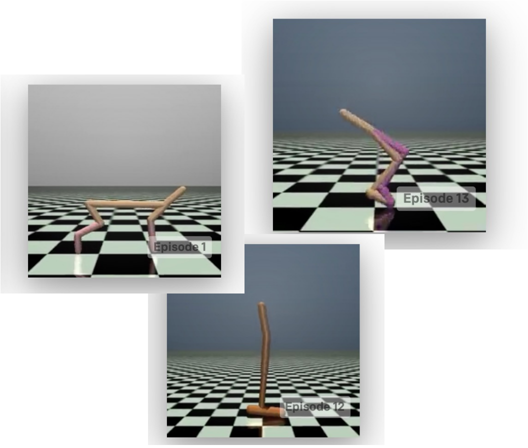


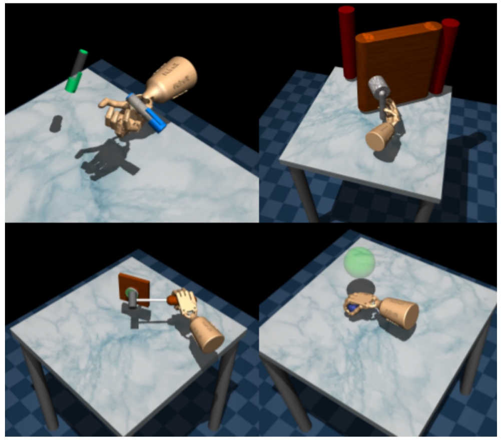
	Behavior	ATAC	PSPI	CQL	COMBO	TD3BC	IQL	BC
halfcheetah-rand	-0.1	4.8	2.3	35.4	**38.8**	10.2	-	2.1
walker2d-rand	0.0	**8.0**	7.6	7.0	7.0	1.4	-	1.6
hopper-rand	1.2	**31.8**	31.6	10.8	17.9	11.0	-	9.8
halfcheetah-med	40.6	**54.3**	43.9	44.4	**54.2**	42.8	47.4	36.1
walker2d-med	62.0	**91.0**	**90.5**	74.5	75.5	79.7	78.3	6.6
hopper-med	44.2	**102.8**	**103.5**	86.6	94.9	**99.5**	66.3	29.0
halfcheetah-med-replay	27.1	49.5	49.2	46.2	**55.1**	43.3	44.2	38.4
walker2d-med-replay	14.8	**94.1**	**94.2**	32.6	56.0	25.2	73.9	11.3
hopper-med-replay	14.9	**102.8**	**102.7**	48.6	73.1	31.4	94.7	11.8
halfcheetah-med-exp	64.3	**95.5**	41.6	62.4	90.0	**97.9**	86.7	35.8
walker2d-med-exp	82.6	**116.3**	**114.5**	98.7	96.1	101.1	**109.6**	6.4
hopper-med-exp	64.7	**112.6**	83.0	**111.0**	**111.1**	**112.2**	91.5	**111.9**
pen-human	207.8	79.3	**106.1**	37.5	-	-	71.5	34.4
hammer-human	25.4	**6.7**	3.8	**4.4**	-	-	1.4	1.5
door-human	28.6	8.7	**12.2**	**9.9**	-	-	4.3	0.5
relocate-human	86.1	**0.3**	**0.5**	**0.2**	-	-	**0.1**	**0.0**
pen-cloned	107.7	73.9	**104.9**	39.2	-	-	37.3	56.9
hammer-cloned	8.1	2.3	**3.2**	2.1	-	-	2.1	0.8
door-cloned	12.1	**8.2**	6.0	0.4	-	-	1.6	-0.1
relocate-cloned	28.7	**0.8**	**0.3**	**-0.1**	-	-	**-0.2**	**-0.1**
pen-exp	105.7	**159.5**	**154.4**	107.0	-	-	-	85.1
hammer-exp	96.3	**128.4**	118.3	86.7	-	-	-	**125.6**
door-exp	100.5	**105.5**	**103.6**	**101.5**	-	-	-	34.9
relocate-exp	101.6	**106.5**	**104.0**	95.0	-	-	-	**101.3**

Experimental Results

ATAC (Relative Pessimism) achieves SOTA performance, outperforming baseline algorithms in 21/24 datasets

9% improvement (median) compared with the best baseline algorithm.



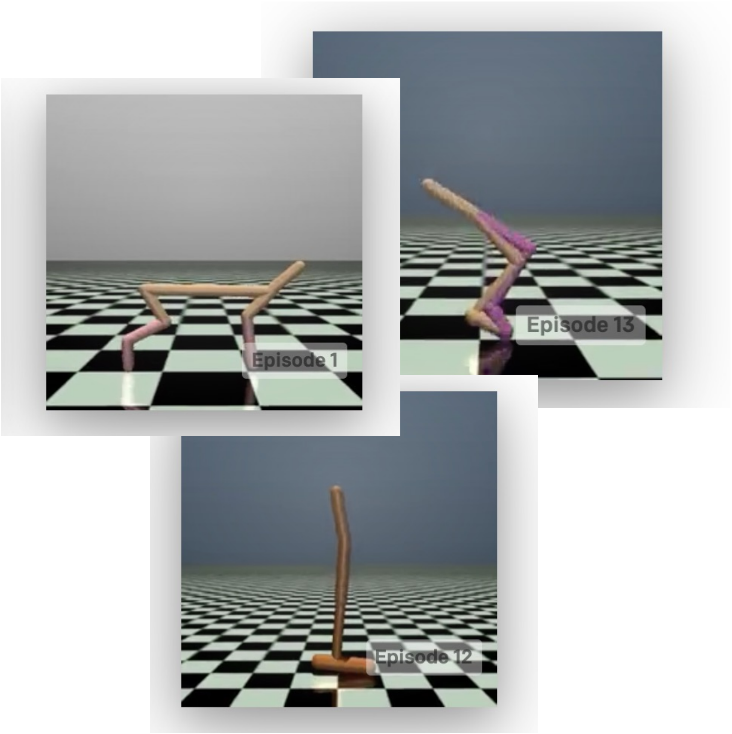


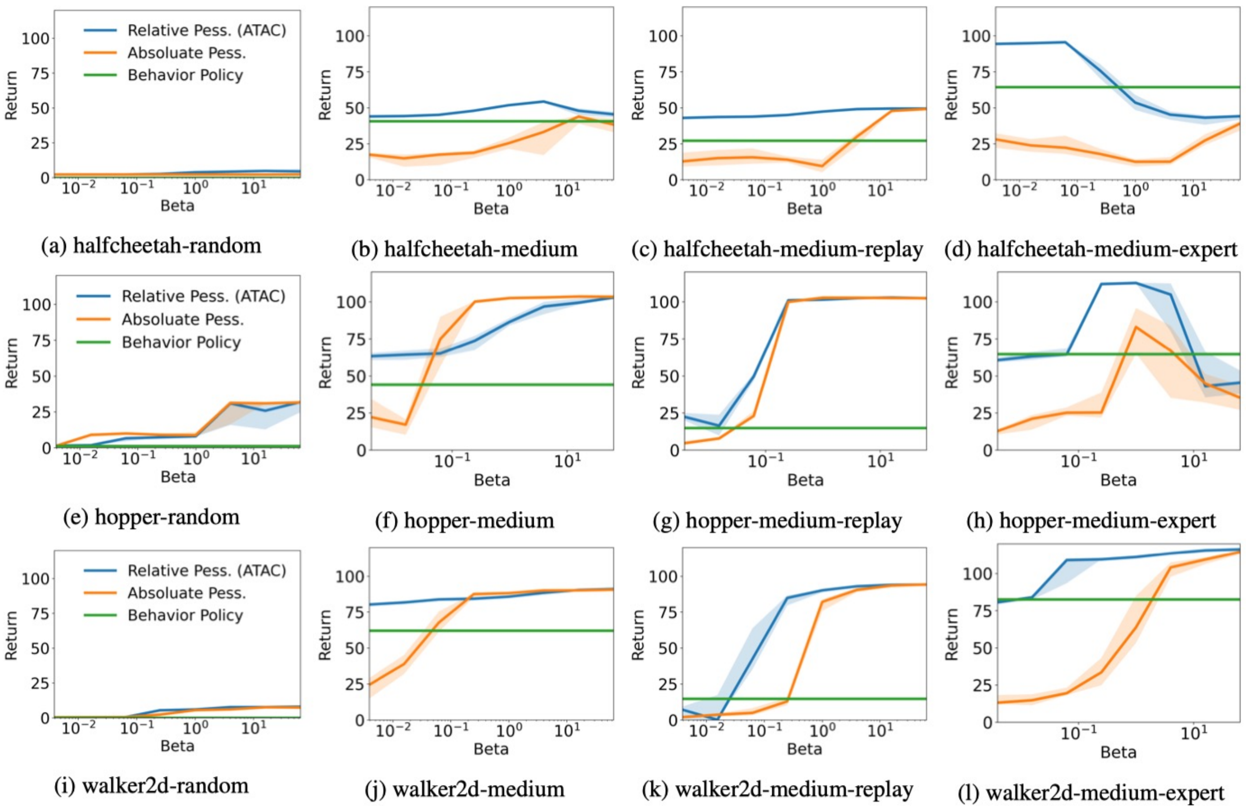
	Behavior	ATAC	PSPI	CQL	COMBO	TD3BC	IQL	BC
halfcheetah-rand	-0.1	4.8	2.3	35.4	**38.8**	10.2	-	2.1
walker2d-rand	0.0	**8.0**	7.6	7.0	7.0	1.4	-	1.6
hopper-rand	1.2	**31.8**	31.6	10.8	17.9	11.0	-	9.8
halfcheetah-med	40.6	**54.3**	43.9	44.4	**54.2**	42.8	47.4	36.1
walker2d-med	62.0	**91.0**	**90.5**	74.5	75.5	79.7	78.3	6.6
hopper-med	44.2	**102.8**	**103.5**	86.6	94.9	**99.5**	66.3	29.0
halfcheetah-med-replay	27.1	49.5	49.2	46.2	**55.1**	43.3	44.2	38.4
walker2d-med-replay	14.8	**94.1**	**94.2**	32.6	56.0	25.2	73.9	11.3
hopper-med-replay	14.9	**102.8**	**102.7**	48.6	73.1	31.4	94.7	11.8
halfcheetah-med-exp	64.3	**95.5**	41.6	62.4	90.0	**97.9**	86.7	35.8
walker2d-med-exp	82.6	**116.3**	**114.5**	98.7	96.1	101.1	**109.6**	6.4
hopper-med-exp	64.7	**112.6**	83.0	**111.0**	**111.1**	**112.2**	91.5	**111.9**
pen-human	207.8	79.3	**106.1**	37.5	-	-	71.5	34.4
hammer-human	25.4	**6.7**	3.8	**4.4**	-	-	1.4	1.5
door-human	28.6	8.7	**12.2**	**9.9**	-	-	4.3	0.5
relocate-human	86.1	**0.3**	**0.5**	**0.2**	-	-	**0.1**	**0.0**
pen-cloned	107.7	73.9	**104.9**	39.2	-	-	37.3	56.9
hammer-cloned	8.1	2.3	**3.2**	2.1	-	-	2.1	0.8
door-cloned	12.1	**8.2**	6.0	0.4	-	-	1.6	-0.1
relocate-cloned	28.7	**0.8**	**0.3**	**-0.1**	-	-	**-0.2**	**-0.1**
pen-exp	105.7	**159.5**	**154.4**	107.0	-	-	-	85.1
hammer-exp	96.3	**128.4**	118.3	86.7	-	-	-	**125.6**
door-exp	100.5	**105.5**	**103.6**	**101.5**	-	-	-	34.9
relocate-exp	101.6	**106.5**	**104.0**	95.0	-	-	-	**101.3**

Experimental Results

Robust Policy Improvement

RPI is also verified empirically, This property can be used for online HP selection: we can gradually increase β to tune its performance without breaking the baseline performance.





Summary

We propose a game theoretic approach to offline RL

Learner optimizes policy

$$\hat{\pi}^* \in \arg\max_{\pi \in \Pi} \phi(\pi, f^\pi)$$

Leader

Adversary selects the worst-case hypothesis

$$f^\pi \in \arg\min_{f \in \mathcal{F}} \phi(\pi, f) + \beta \mathcal{E}(\pi, f)$$

Follower

Follower can also use a constrained version.

Offline RL

$$\max_{\pi \in \Pi} \quad J(\pi)$$
$$s.t. \quad \text{no env interaction}$$

Maximize return in the true environment using data with partial coverage

Summary

Papers

Github Code

$$\hat{\pi}^* = \arg\max_{\pi \in \Pi} \phi(\pi, f^\pi)$$

$$f^\pi \in \arg\min_{f \in \mathcal{F}} \phi(\pi, f) + \beta \mathcal{E}(\pi, f)$$

Absolute Pessimism

Relative Pessimism

Robust Policy Improvement (RPI)

Model-free

$$f(s, a)$$

$$\mathcal{E}(\pi, f) := \mathbb{E}_\mu[(f - \mathcal{T}^\pi f)^2]$$

$$\phi(\pi, f) := f(s_0, \pi)$$

(Xie and Cheng et al., 2021)

$$\phi(\pi, f) := \mathbb{E}_{d^\mu}[f(s, \pi) - f(s, a)]$$

(Cheng* and Xie* et al., 2022)

SoTA Empirical Results

Learning Optimality

Learn the best policy that the data can afford despite missing coverage

Useful for online HP tuning and applications where decisions can lead to risky consequences

Robust Policy Improvement

Learn a policy better than the data collection policy, regardless of hyperparameters.

Summary

Papers

Github Code

$$\hat{\pi}^* = \arg\max_{\pi \in \Pi} \phi(\pi, f^\pi)$$

$$f^\pi \in \arg\min_{f \in \mathcal{F}} \phi(\pi, f) + \beta \mathcal{E}(\pi, f)$$

	Absolute Pessimism	**Relative Pessimism** Robust Policy Improvement (RPI)
Model-free $f(s,a)$ $\mathcal{E}(\pi, f) := \mathbb{E}_\mu[(f - \mathcal{T}^\pi f)^2]$	$\phi(\pi, f) := f(s_0, \pi)$ (Xie and Cheng et al., 2021)	$\phi(\pi, f) := \mathbb{E}_{d^\mu}[f(s,\pi) - f(s,a)]$ (Cheng* and Xie* et al., 2022)
Model-based $\hat{M} = (\hat{r}, \hat{P})$ $\mathcal{E}(\pi, M) := \mathbb{E}_{d^\mu}[(\hat{r} - r)^2 - \log P]$	$\phi(\pi, M) := J_{\hat{M}}(\pi)$ (Uehara and Sun., 2021)	$\phi(\pi, M) := J_{\hat{M}}(\pi) - J_{\hat{M}}(\mu)$ (Xie and Bhardwaj et al., 2022)

Many more choices to explore in the future…