Adversarially Trained Actor Critic for Offline Reinforcement Learning

Ching-An Cheng*, Tengyang Xie*, Nan Jiang, Alekh Agarwal

(*equal contribution)

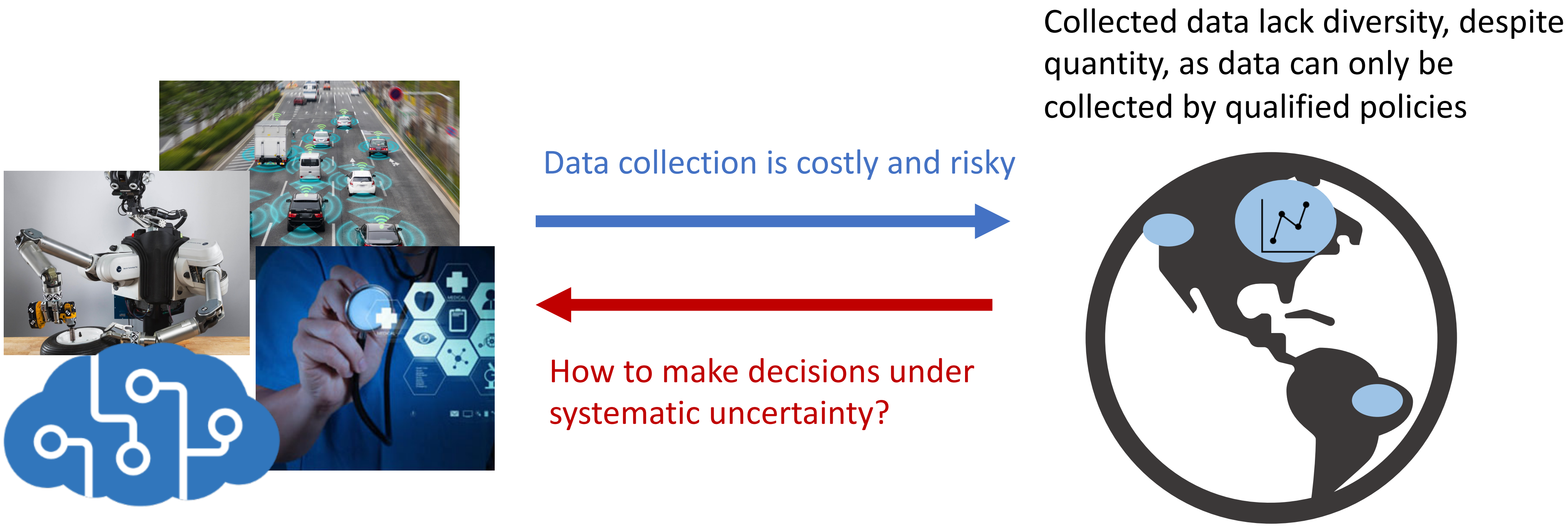
Microsoft

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Google Research

Motivation

- Challenge of real-world decision-making problems



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

Data collection is costly and risky

How to make decisions under systematic uncertainty?

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:**

  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?

Contribution

- **Adversarially Trained Actor Critic (ATAC)**

  A provably correct & scalable framework based on relative pessimism for offline RL with nonlinear function approximators

Learning Optimality

Learn the optimal policy as long as the data covers the optimal policy well

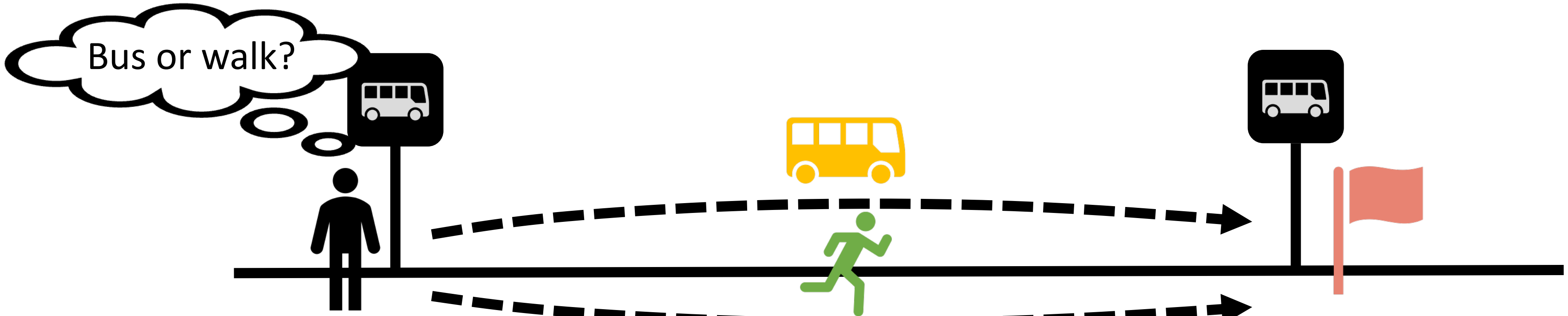
Robust Policy Improvement

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

Why Relative Pessimism?

Hypothesis class

1. Bus arrives on time: Bus 5 min, Walk 30 min
2. Bus is delayed: Bus 30 min, Walk 30 min

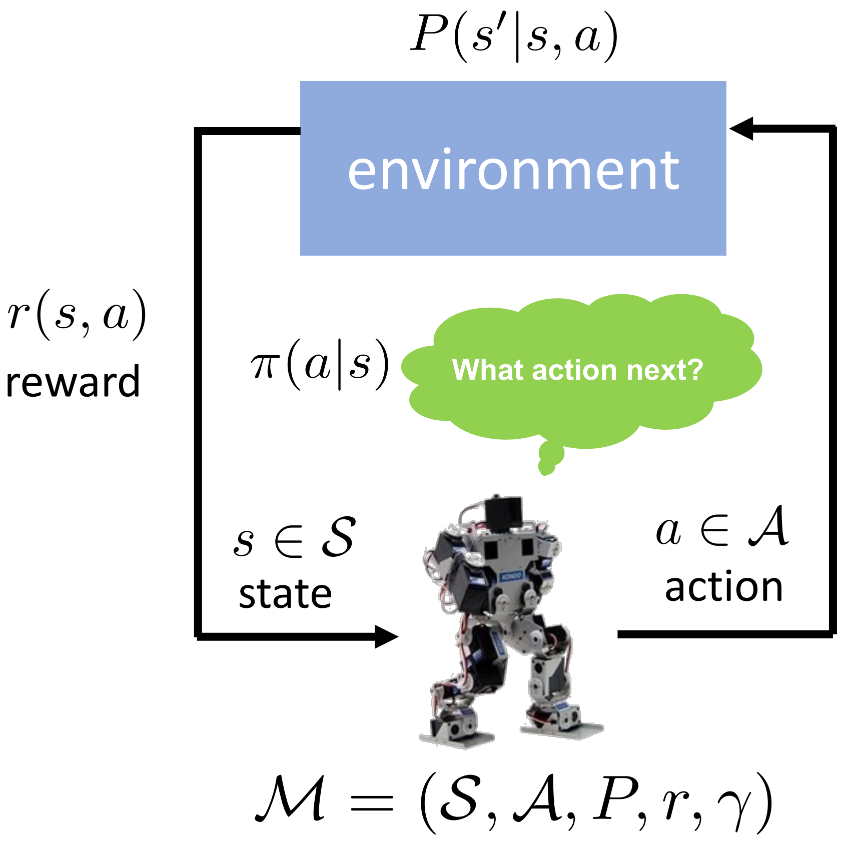


Relative Pessimism

Intuitively we want to take the bus compared with flipping a coin.

Problem Setup

- Suppose the world is a Markov decision process (MDP)



Offline setting assumption: offline data $\mathcal{D}$, collected by a behavior policy μ starting from d_0.

Goal: Find a policy π that has high return

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

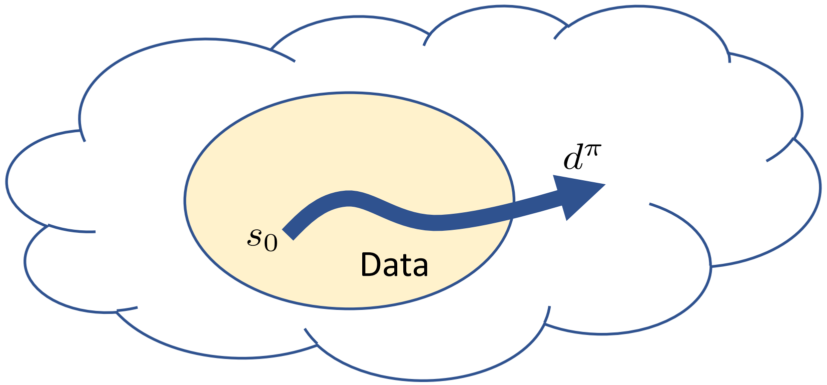
No interaction with environment for learning.

Key Idea: Relative Pessimism

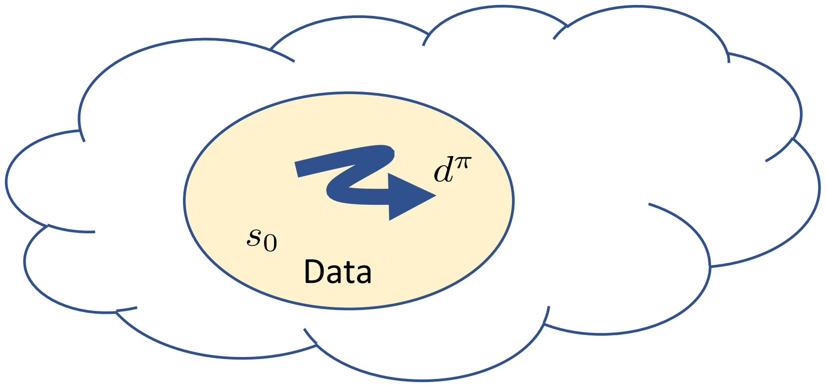
- Optimize for the **worst-case** performance compared with the **behavior policy $\boldsymbol{\mu}$.**

$$\widehat{\pi}^* \in \arg\max_{\pi \in \Pi} \boxed{\text{Lower bound of } J(\pi) - J(\mu)}$$

Lower bound $< J(\pi) - J(\mu)$



Lower bound $\approx J(\pi) - J(\mu)$



Key Idea: Relative Pessimism

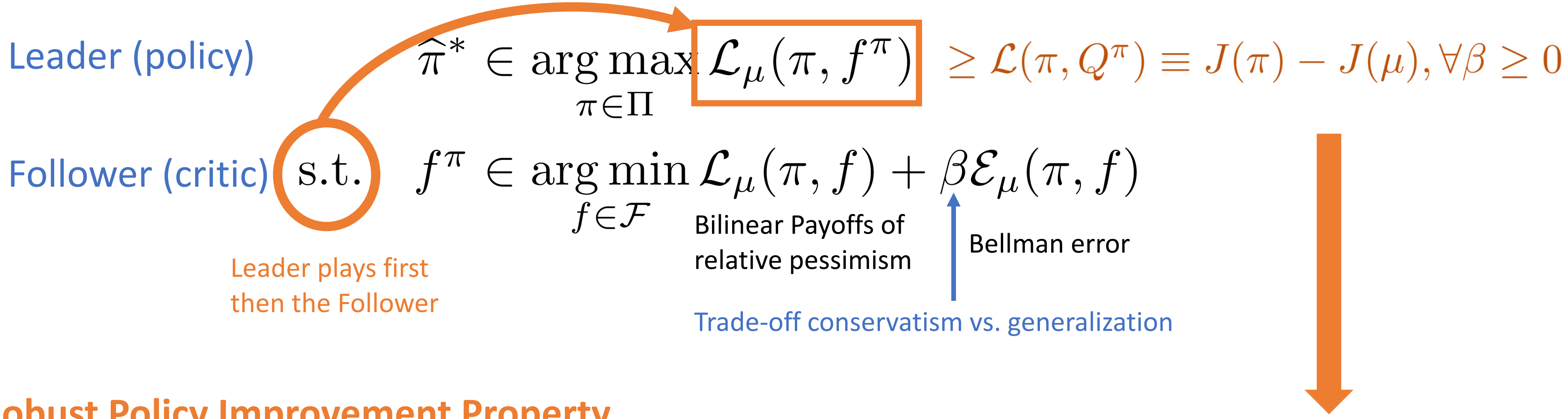
- Optimize for the **best worst-case** performance compared with the **behavior policy $\boldsymbol{\mu}$.**

$$\widehat{\pi}^* \in \underset{\pi \in \Pi}{\arg\max} \boxed{\text{Lower bound of } J(\pi) - J(\mu)}$$

ATAC frames this problem as a Stackelberg game (i.e., bilevel optimization)

A Stackelberg Game for Offline RL

- ATAC optimizes for relative pessimism via solving a Stackelberg game



Leader (policy)

$$\widehat{\pi}^* \in \arg\max_{\pi\in\Pi} \boxed{\mathcal{L}_\mu(\pi, f^\pi)} \geq \mathcal{L}(\pi, Q^\pi) \equiv J(\pi) - J(\mu), \forall\beta \geq 0$$

Follower (critic) s.t.

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

Bilinear Payoffs of relative pessimism

Bellman error

Leader plays first then the Follower

Trade-off conservatism vs. generalization

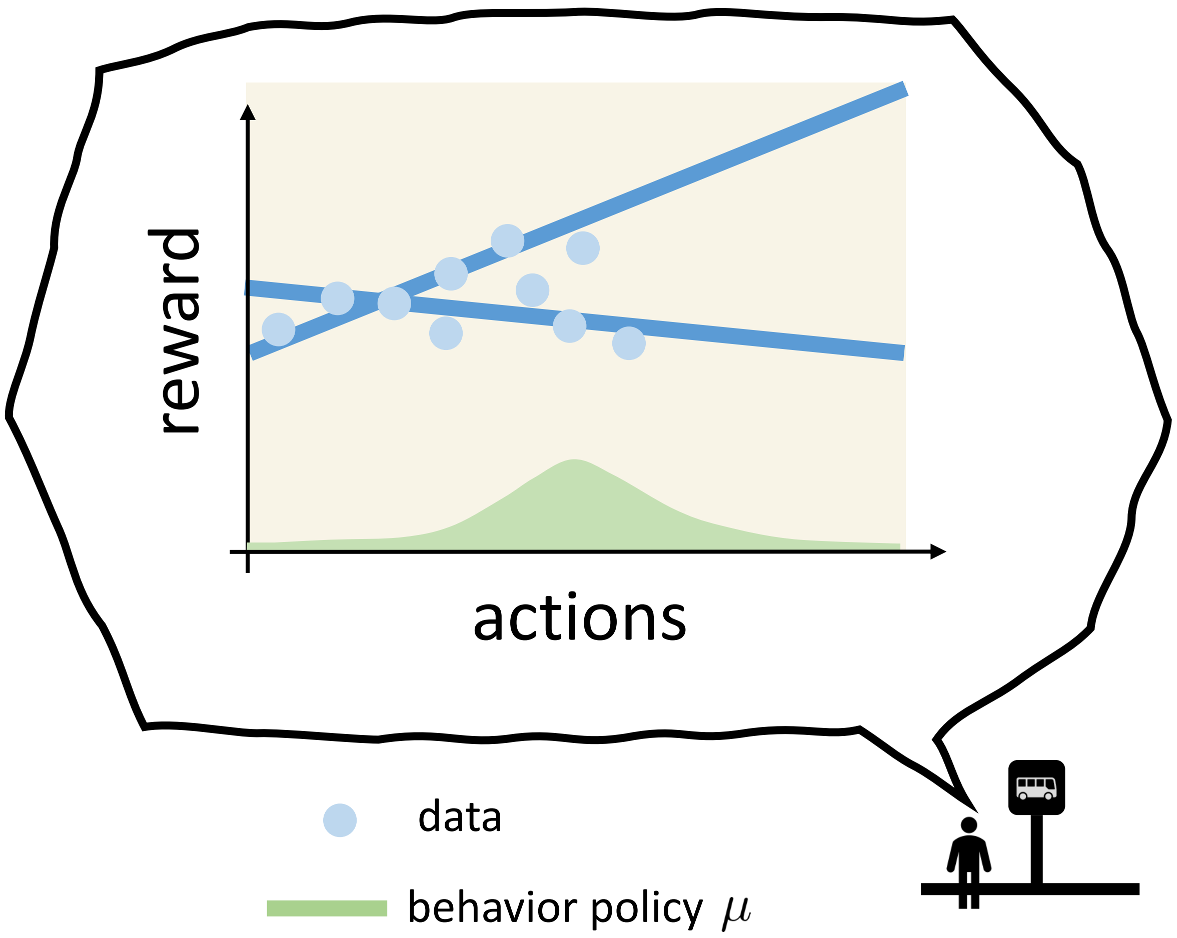
Robust Policy Improvement Property

For all $\beta \geq 0$, the ATAC policy is always no worse than the behavior policy that collected the data.

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

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

A Stackelberg Game for Offline RL



ATAC

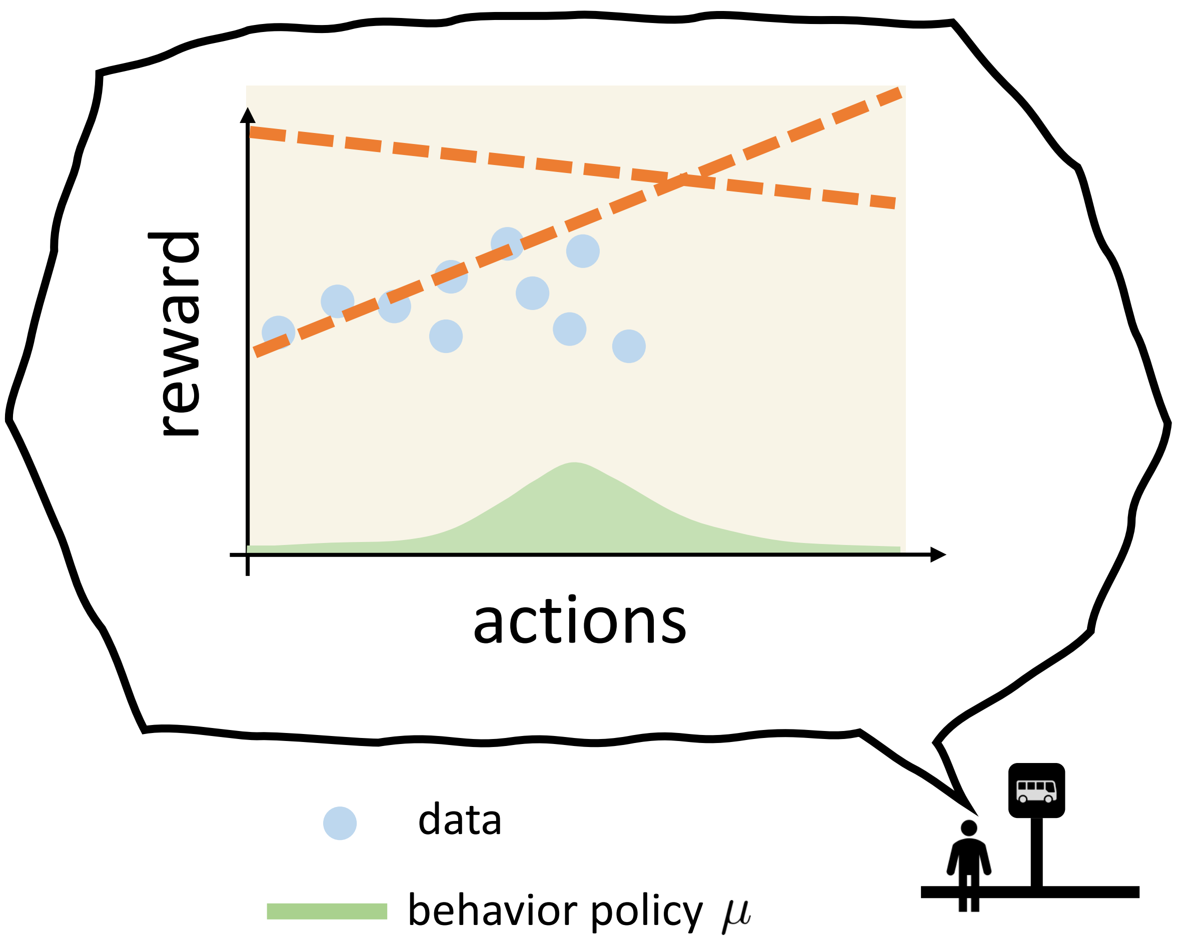
$$\widehat{\pi}^* \in \arg\max_{\pi\in\Pi} \mathcal{L}_\mu(\pi, f^\pi)$$

$$\text{s.t.} \quad \boxed{f^\pi} \in \arg\min_{f\in\mathcal{F}} \mathcal{L}_\mu(\pi, f) + \beta\mathcal{E}_\mu(\pi, f)$$

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

Functions that are consistent with the reward and the dynamics on the behavior data

A Stackelberg Game for Offline RL



ATAC

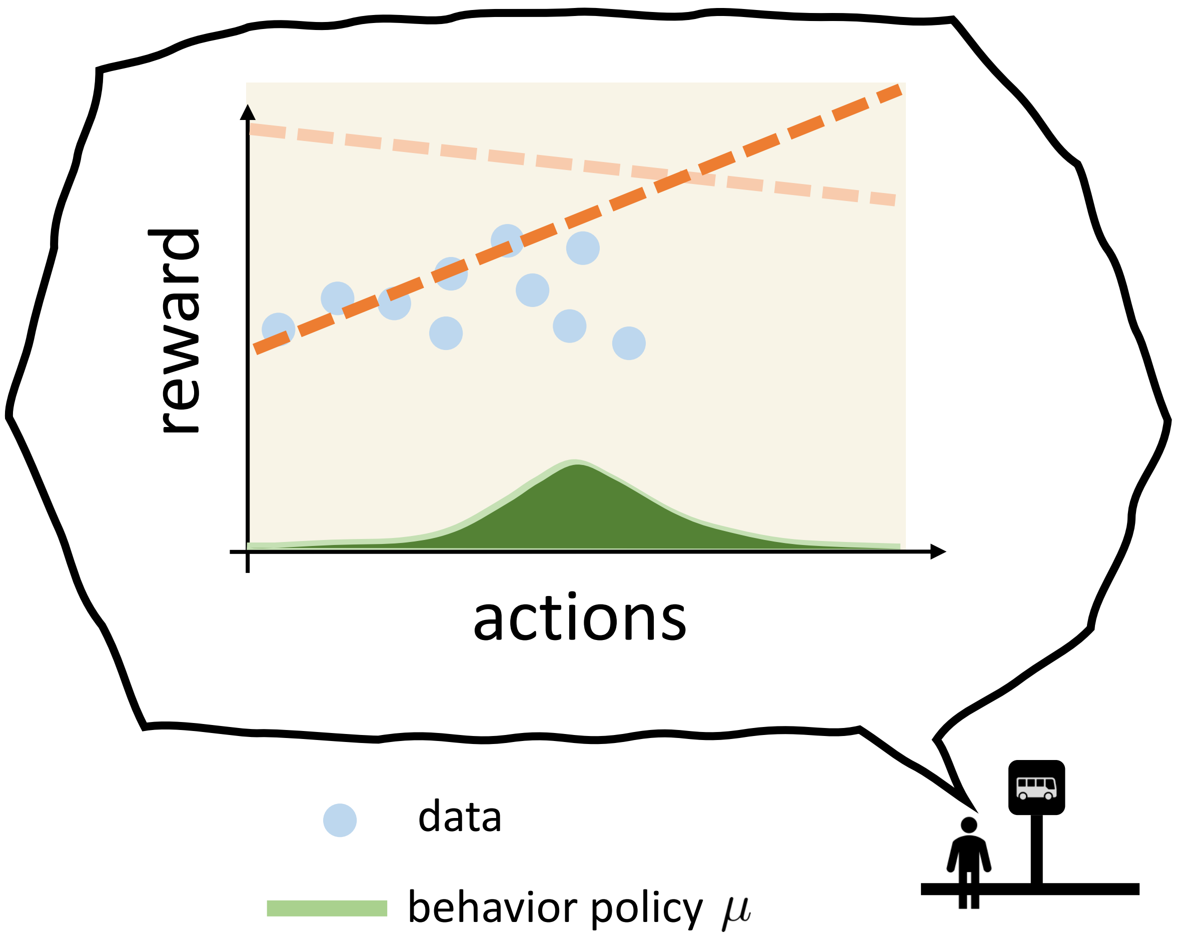
$$
\begin{aligned}
&\widehat{\pi}^* \in \arg\max_{\pi\in\Pi} \boxed{\mathcal{L}_\mu(\pi, f^\pi)} \\
\text{s.t.} \quad & f^\pi \in \arg\min_{f\in\mathcal{F}} \mathcal{L}_\mu(\pi, f) + \beta\mathcal{E}_\mu(\pi, f)
\end{aligned}
$$

– – value difference hypothesis $f(s,\cdot) - f(s,\mu)$

Functions shifted from the original hypotheses

What is the solution to the Stackelberg game?

A Stackelberg Game for Offline RL



ATAC

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

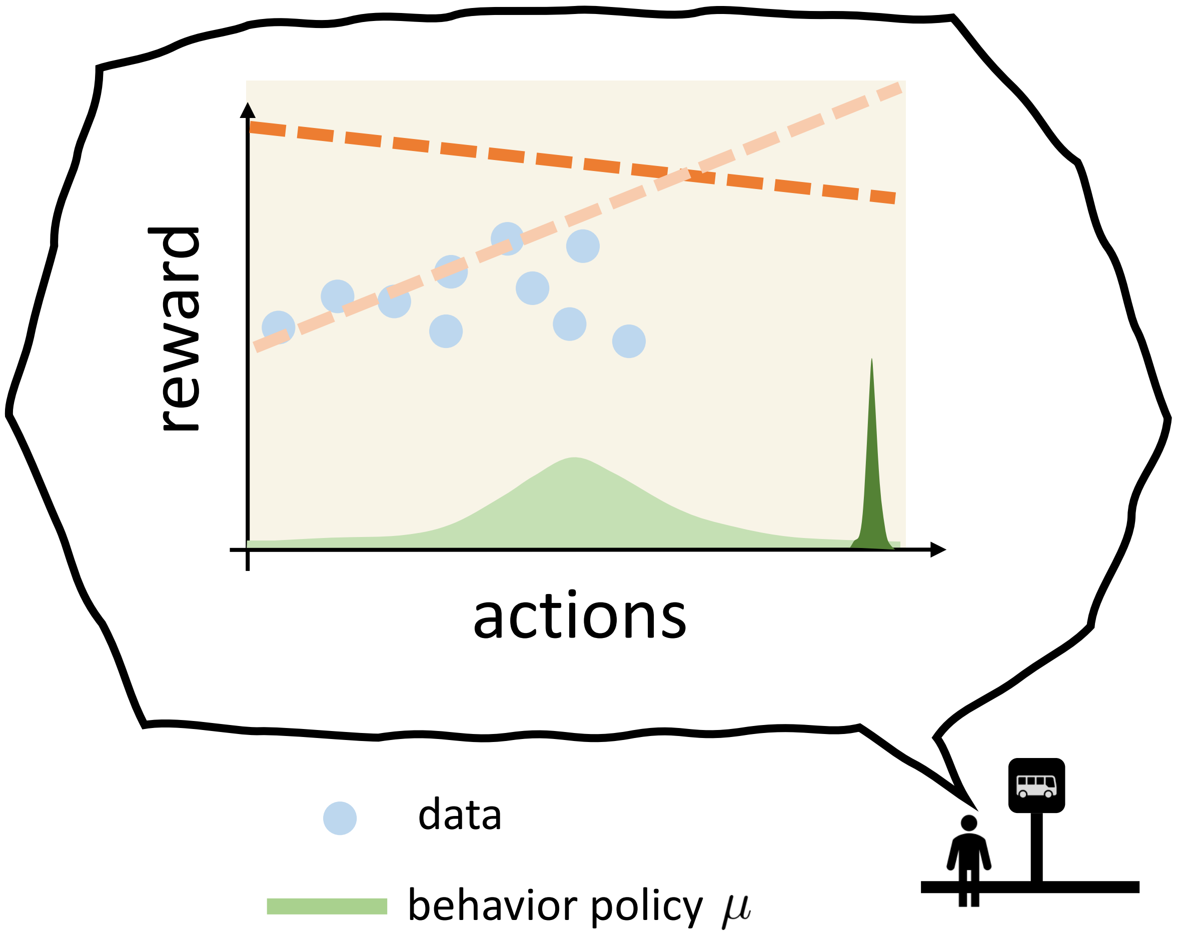
▬ ▬ value difference hypothesis $f(s,\cdot) - f(s,\mu)$

▬ ▬ inactive value difference hypothesis

▬ decision policy π

Not the behavior policy in this case…

A Stackelberg Game for Offline RL



ATAC

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

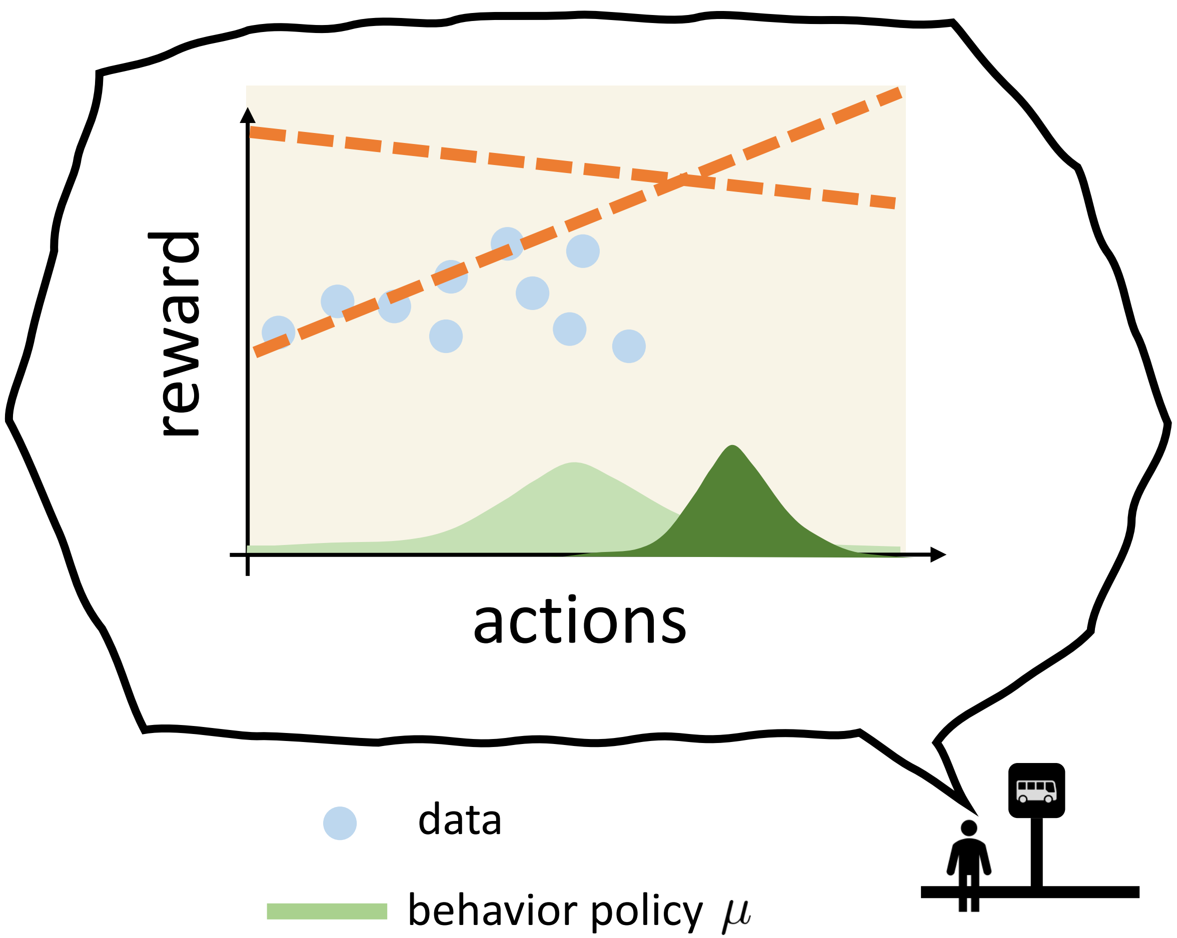
value difference hypothesis $f(s,\cdot) - f(s,\mu)$

inactive value difference hypothesis

decision policy π

Not the policy that maximizes a single hypothesis...

A Stackelberg Game for Offline RL



ATAC

$$\boxed{\widehat{\pi}^*} \in \arg\max_{\pi\in\Pi} \mathcal{L}_\mu(\pi, f^\pi)$$

$$\text{s.t.} \quad f^\pi \in \arg\min_{f\in\mathcal{F}} \mathcal{L}_\mu(\pi, f) + \beta\mathcal{E}_\mu(\pi, f)$$

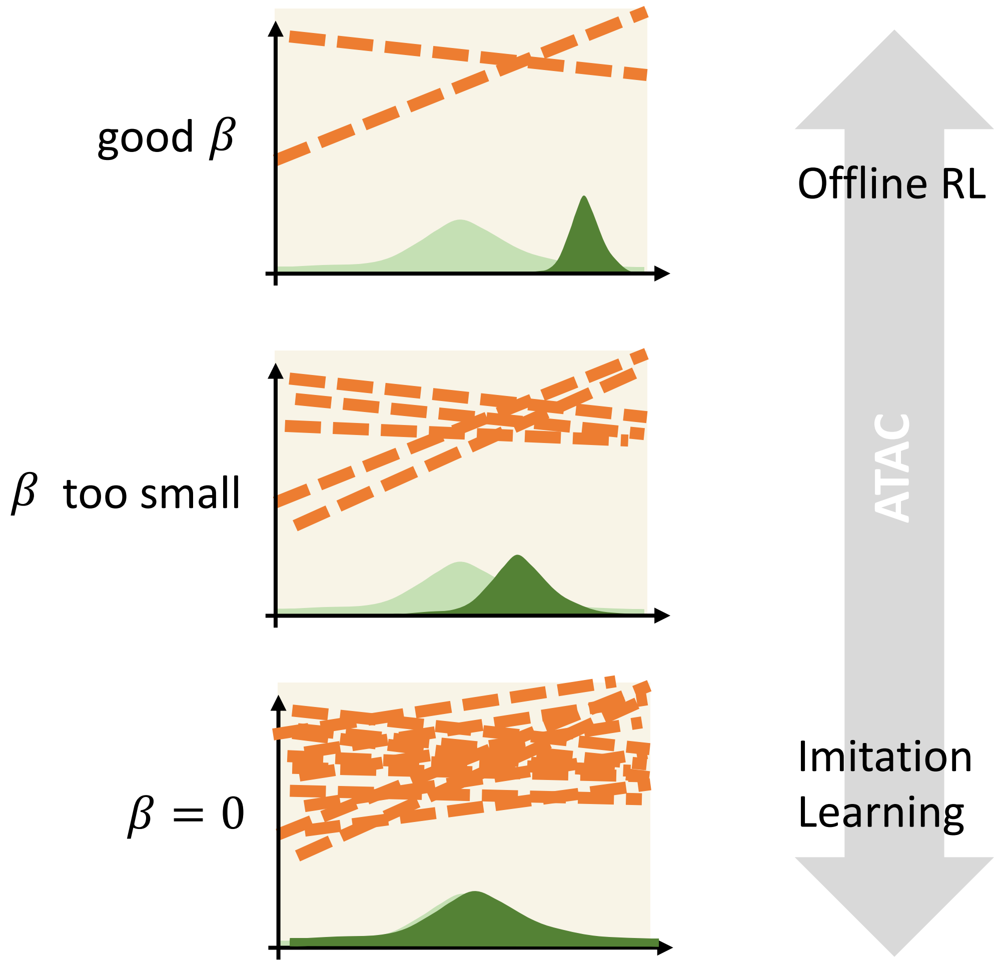
value difference hypothesis $f(s,\cdot) - f(s,\mu)$

inactive value difference hypothesis

decision policy π

The optimal decision balances multiple hypotheses

A Stackelberg Game for Offline RL



Leader = Actor = Conditional generator

Follower = Critic = Discriminator

ATAC

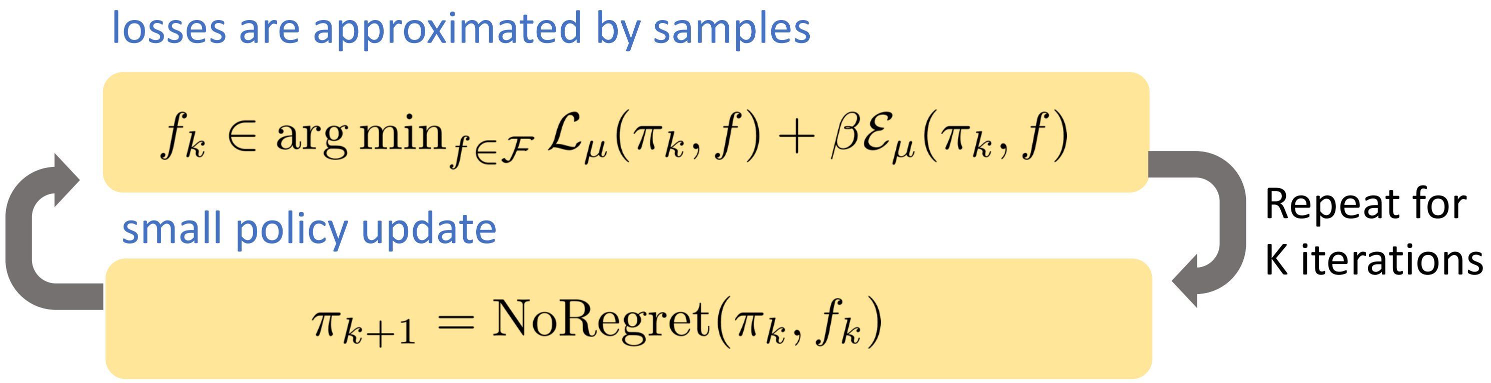
$$\begin{aligned} & \widehat{\pi}^* \in \arg\max_{\pi\in\Pi} \mathcal{L}_\mu(\pi, f^\pi) \\ \text{s.t.} \quad & f^\pi \in \arg\min_{f\in\mathcal{F}} \mathcal{L}_\mu(\pi, f) + \beta\mathcal{E}_\mu(\pi, f) \end{aligned}$$

ATAC 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

Solving the Stackelberg Game

No-Regret + Best Response Scheme



losses are approximated by samples

$$f_k \in \arg\min_{f\in\mathcal{F}} \mathcal{L}_\mu(\pi_k, f) + \beta\mathcal{E}_\mu(\pi_k, f)$$

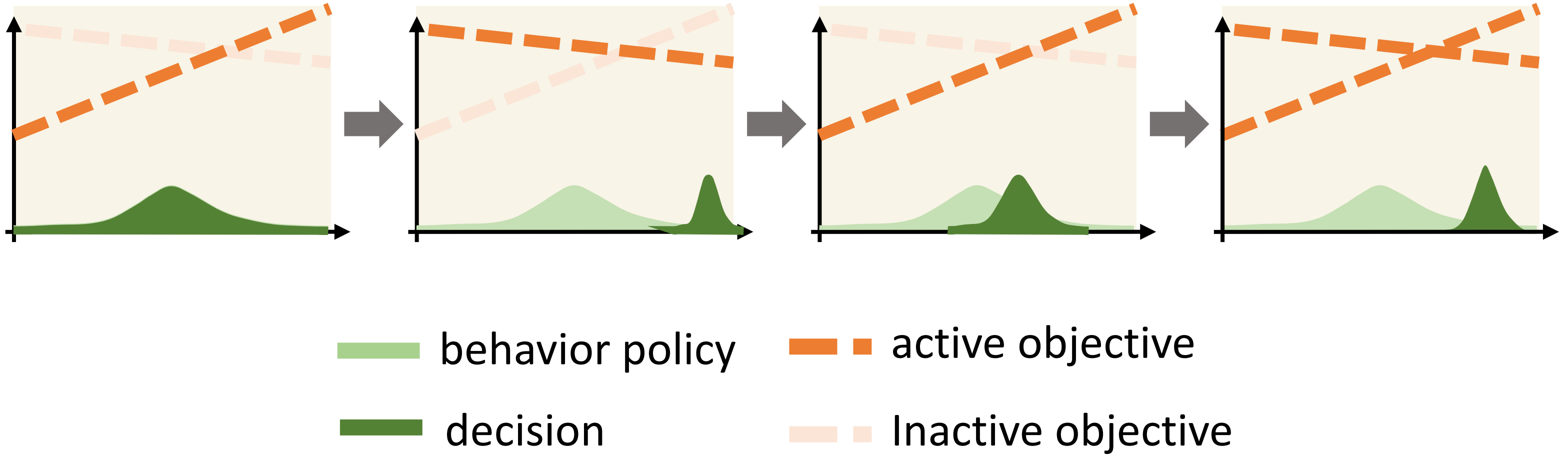
small policy update

$$\pi_{k+1} = \mathrm{NoRegret}(\pi_k, f_k)$$

Repeat for
K iterations

Output uniform mixture of policies (theory) or the last policy (practice)

In practice, the above is implemented by two-timescale SGD updates



ATAC Theory (Informal)

Learning Optimality

Assume $\mathcal{F}$ satisfies realizability and completeness.

Given dataset $\mathcal{D}$ s.t. $|\mathcal{D}| = N$. With $\beta = \Theta(N^{2/3})$. Then $\forall \pi \in \Pi$,

$$J(\pi) - J(\hat{\pi}) \leq \mathcal{O}\left(\frac{1}{(1-\gamma)N^{1/3}}\right) + \epsilon_{\text{generalization}}(\mathcal{F}, \pi, \mathcal{D})$$

average Bellman error of f_t on the distribution of π

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

ATAC Theory (Informal)

Robust Policy Improvement

Assume $\mathcal{F}$ satisfies **realizability** without the need of completeness.

$$\text{If } \mu \in \Pi, \text{ then } \; J(\mu) - J(\bar{\pi}) \le \mathcal{O}\left(\frac{1}{(1-\gamma)N^{1/2}} + \frac{\beta}{(1-\gamma)N}\right)$$

faster rate

ATAC always improves over the behavior policy so long as β = $o(N)$.

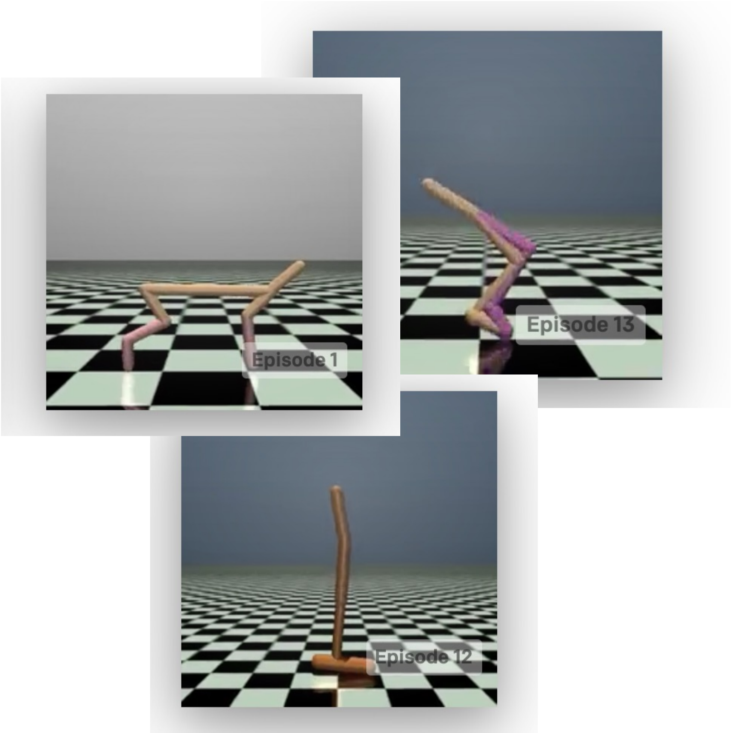
ATAC and Other Offline RL Techniques

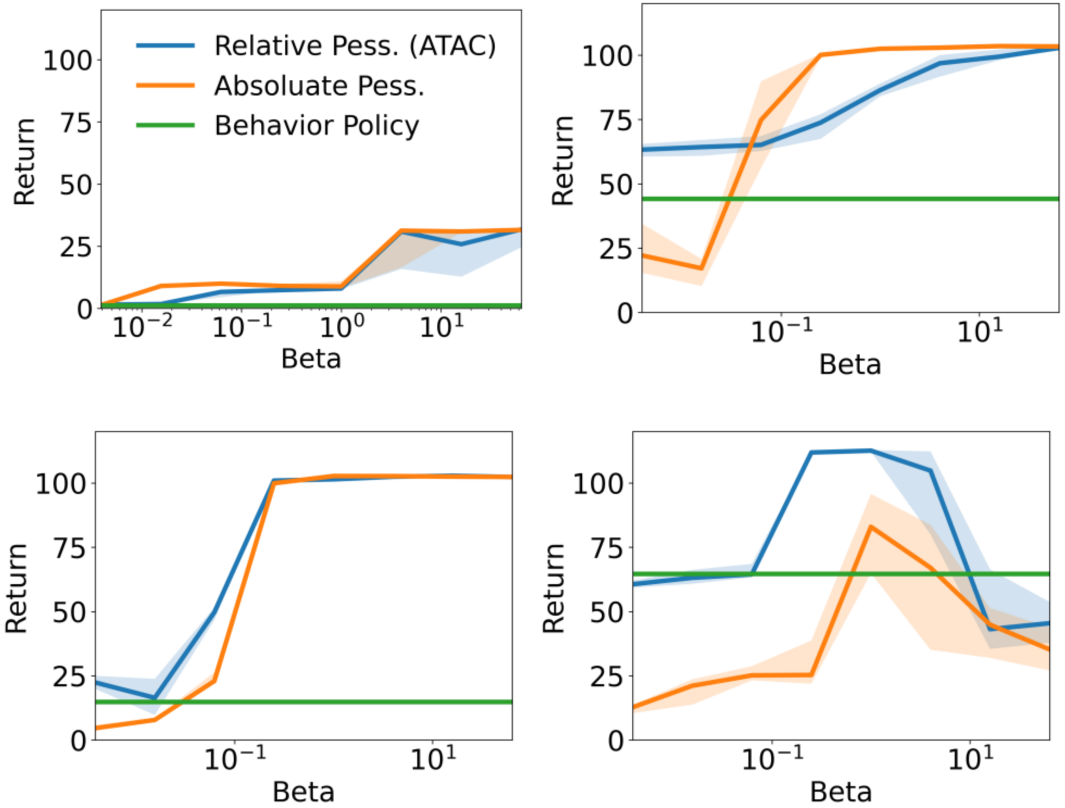
	Relative Pessimism	**Absolute Pessimism**
Two-player Game	ATAC: Maximin problem of relative performance Robust Policy Improvement Less Conservative	Maximin problem of absolute performance (Xie et al., 2021, Uehara et al., 2021) Less Conservative
Single MDP	Behavior regularization (Fujimoto et al. 2019,2021, Kuma et al., 2019, Laroche et al. 2019) Robust Policy Improvement 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

Experimental Results

Robust Policy Improvement

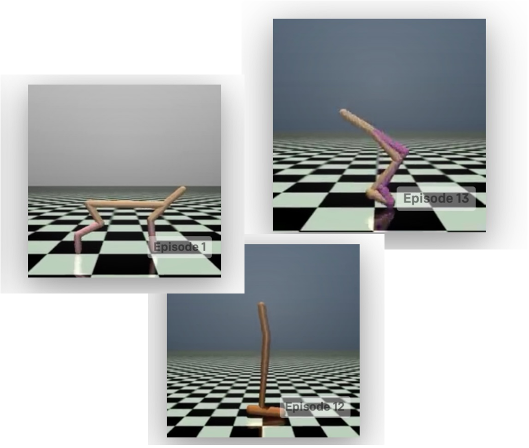
ATAC's robustness property enables online HP selection. We can gradually increase β to tune its performance without breaking the baseline performance.

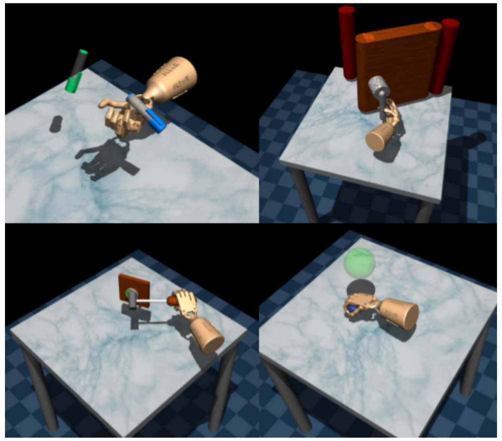


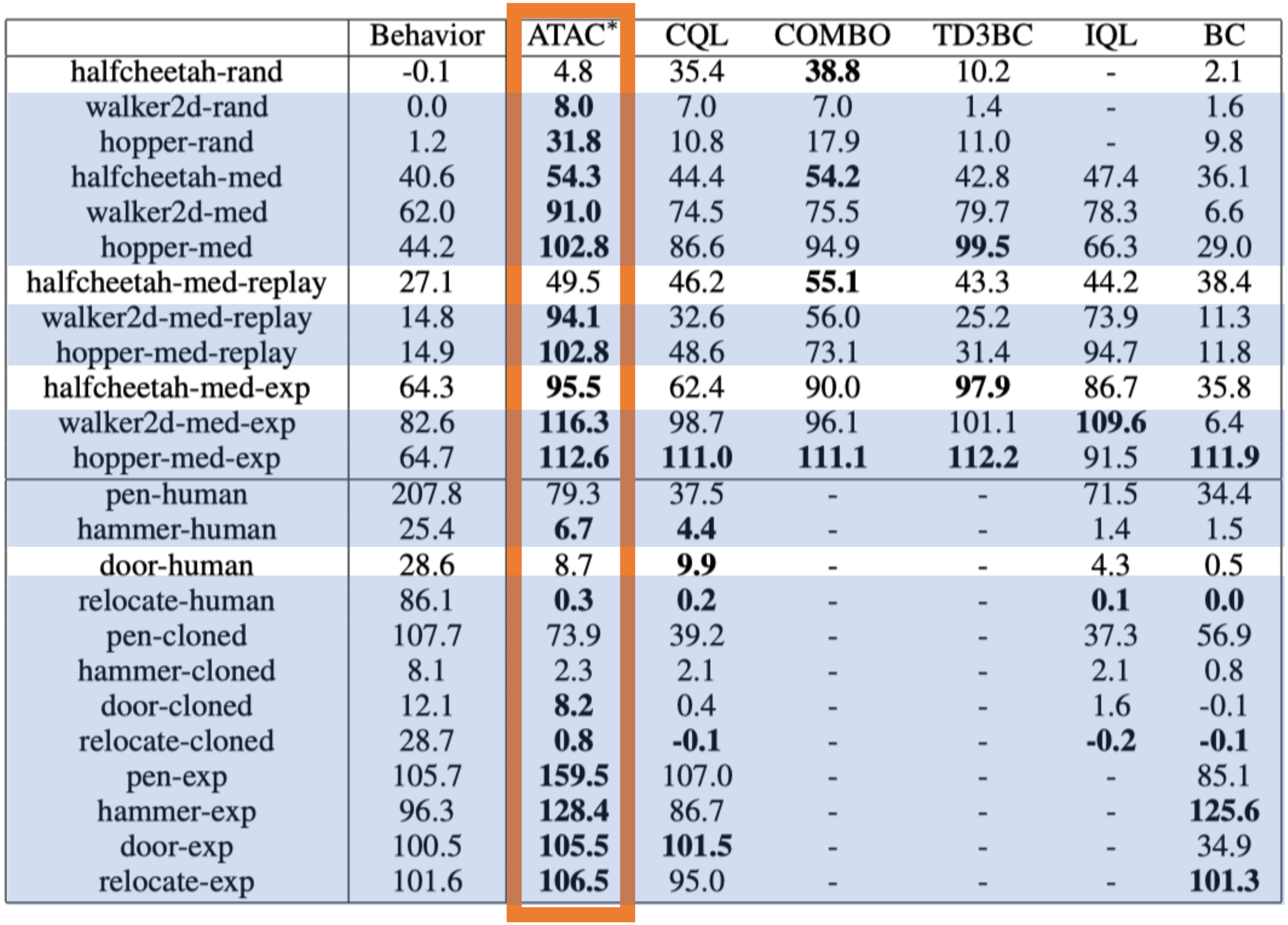


Experimental Results

ATAC achieves SOTA performance, outperforming baseline algorithms in most datasets







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

Datasets where ATAC is the best performing algorithm, with 9% improvement (median) compared with the best baseline algorithm.

Summary

Paper

Github

- **Adversarially Trained Actor Critic (ATAC)**

A provably correct & scalable framework based on relative pessimism for offline RL with nonlinear function approximators

Learning Optaimlity

Learn the optimal policy as long as the data covers the optimal policy well

SoTA Empirical Results

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.