Adversarially Trained Actor Critic for Offline Reinforcement Learning

Presentor: Tian Xu

xut@lamda.nju.edu.cn

Nanjing University, Nanjing, China

Based on :
Ching-An Cheng, Tengyang Xie, Nan Jiang, Alekh Agarwal. "Adversarially Trained Actor Critic for Offline Reinforcement Learning." ICML (Outstanding Paper), 2022.

November 8, 2022

Motivation

Two desiderata for an offline RL algorithm.

- Safe Policy Improvement: with a proper hyperparameter, the learned policy should outperform the behavioral policy. Robust Policy Improvement (RPI): safe policy improvement holds across large hyperparameter choices.
  - RPI is important for some risk-sensitive tasks like healthcare.
  - RPI makes policy finetuning with additional online interactions possible.
- Learning Optimality: the learned policy should outperform the policy whose state-action distribution is well covered by the dataset.

Existing theoretical and empirical works fail to satisfy both two properties, especially the RPI property.

Main Contributions

- The authors proposed a Stackelberg game formulation based on the relative pessimism. Built upon this Stackelberg game, they developed the method adversarially trained actor critic (ATAC).
- Theoretically, they proved that ATAC satisfies the properties of both robust policy improvement and learning optimality.
- Empirically, ATAC has a scalable implementation, which is verified to hold the mentioned two desiderata on D4RL benchmarks.

Background

- Infinite-horizon discounted MDP: $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma, s_0)$.
- Policy value and Q-function:
$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | \pi\right], Q^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | (s_0, a_0) = (s, a), \pi\right].$$
- For a policy π, the Bellman operator $\mathcal{T}^\pi$: $(\mathcal{T}^\pi f)(s, a) := R(s, a) + \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[f(s', \pi)\right]$, where $f(s', \pi) = \mathbb{E}_{a' \sim \pi(\cdot|s')}\left[f(s', a')\right]$.
- State-action distribution: $d^\pi(s, a) = (1 - \gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{I}\left(s_t = s, a_t = a\right) \mid \pi\right]$.
- The data-weighted squared ℓ_2-norm: $\|f\|_{2,\mu}^2 = \mathbb{E}_{(s,a)\sim\mu}\left[f(s, a)^2\right]$.

Offline RL and Function Approximation

In the offline setting, we only have access to a pre-collected dataset $\mathcal{D} = \{(s, a, r, s')\}$, where $(s, a) \sim \mu = d^\mu, r = R(s, a), s' \sim P(\cdot|s, a)$ and μ is the behavioral policy.

They consider the function approximation setting. The learner have access to a value function class $\mathcal{F} \subseteq (\mathcal{S} \times \mathcal{A} \to [0, V_{\max}])$ and a policy class $\Pi \subseteq (\mathcal{S} \to \Delta(\mathcal{A}))$.

- Assumption 1 (Realizability): $\forall \pi \in \Pi, Q^\pi \in \mathcal{F}$.
- Assumption 2 (Bellman Completeness): $\forall f \in \mathcal{F}, \forall \pi \in \Pi, \mathcal{T}^\pi f \in \mathcal{F}$.
- The theoretical results also hold when the above assumptions are satisfied approximately.
- These two assumptions are necessary for efficient learning with function approximation in the offline setting. [Foster et al., 2022, Wang et al., 2021]

Key Idea: Relative Pessimism

Optimize for the best worst-case performance compared with the behavior policy.

$$\widehat{\pi}^* \in \arg\max_{\pi \in \Pi} \underbrace{\text{lower bound of}}_{\text{Pessimism}} \underbrace{J(\pi) - J(\mu)}_{\text{Relative}}.$$

Stackelberg Game Formulation

$$\widehat{\pi}^* \in \underset{\pi \in \Pi}{\operatorname{argmax}} \mathcal{L}_\mu\left(\pi, f^\pi\right)$$
$$\text{s.t.} \quad f^\pi \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \mathcal{L}_\mu(\pi, f) + \beta \mathcal{E}_\mu(\pi, f)$$

where $\mathcal{L}_\mu(\pi, f) := \mathbb{E}_{(s,a)\sim\mu}[f(s, \pi) - f(s, a)]$, $\mathcal{E}_\mu(\pi, f) := \mathbb{E}_\mu\left[\left(\left(f - \mathcal{T}^\pi f\right)(s, a)\right)^2\right]$ is the Bellman error and β is the hyperparameter.

- $\widehat{\pi}^*$ maximizes the value function on $a \sim \pi(\cdot|s)$ predicted by f^π, and minimizes the value function on $a \sim \mu(\cdot|s)$ predicted by f^π.
- f^π performs the relatively pessimistic policy evaluation of π compared with μ (we will prove $\mathcal{L}_\mu\left(\pi, f^\pi\right) \precsim J(\pi) - J(\mu)$). On the one hand, the Bellman error $\mathcal{E}_\mu(\pi, f)$ ensures f^π's Bellman consistency on data. On the other hand, $\mathcal{L}_\mu(\pi, f)$ ensures f^π's relative pessimism.

Robust Policy Improvement

Proposition 1 (Robust Policy Improvement).

Suppose that the realizability assumption holds and $\mu \in \Pi$, for any $\beta \geq 0$, we have that

$$\mathcal{L}_\mu\left(\pi, f^\pi\right) \leq (1-\gamma)(J(\pi)-J(\mu)), \ \forall \pi \in \Pi.$$

Furthermore, it holds that $J\left(\widehat{\pi}^*\right) \geq J(\mu)$.

- Relative pessimism: The first claim suggests that $\widehat{\pi}^\star$ optimizes the lower bound of $J(\pi)-J(\mu)$ up to constants. This lower bound is tight when π is the behavior policy μ.
- The second claim shows that the policy that solves this Stackelberg game holds the robust policy improvement property. The RPI property is mainly due to the relative pessimism.

Analysis

Claim I: For any $\pi \in \Pi$, $Q^\pi \in \mathcal{F}$ and $f^\pi \in \underset{f \in \mathcal{F}}{\mathrm{argmin}} \mathcal{L}_\mu(\pi, f) + \beta \mathcal{E}_\mu(\pi, f)$, we have

$$\begin{aligned}
\mathcal{L}_\mu(\pi, f^\pi) &\le \mathcal{L}_\mu(\pi, f^\pi) + \beta \mathcal{E}_\mu(\pi, f^\pi) \\
&\le \mathcal{L}_\mu(\pi, Q^\pi) + \beta \mathcal{E}_\mu(\pi, Q^\pi) \\
&= \mathcal{L}_\mu(\pi, Q^\pi) && \mathcal{E}_\mu(\pi, Q^\pi) = 0 \\
&= \mathbb{E}_{(s,a)\sim d^\mu} \left[Q^\pi(s, \pi) - Q^\pi(s, a) \right] \\
&= \mathbb{E}_{(s,a)\sim d^\mu} \left[-A^\pi(s, a) \right] \\
&= (1-\gamma)\left(J(\pi) - J(\mu)\right). && \text{policy difference lemma}
\end{aligned}$$

$$\begin{aligned}
\text{Claim II: } (1-\gamma)\left(J(\widehat{\pi}^*) - J(\mu)\right) &\ge \mathcal{L}_\mu(\widehat{\pi}^*, f^{\widehat{\pi}^*}) + \beta \mathcal{E}_\mu(\widehat{\pi}^*, f^{\widehat{\pi}^*}) \\
&\ge \mathcal{L}_\mu(\widehat{\pi}^*, f^{\widehat{\pi}^*}) && \mathcal{E}_{\widehat{\pi}^*}(\widehat{\pi}^*, f^{\widehat{\pi}^*}) \ge 0 \\
&\ge \mathcal{L}_\mu(\mu, f^\mu) && \mu \in \Pi \\
&= 0.
\end{aligned}$$

Useful fact: $\mathcal{L}_\mu(\pi, f^\pi) \le \mathcal{L}_\mu(\pi, Q^\pi) = (1-\gamma)(J(\pi) - J(\mu))$.

Why Relative Pessimism Leads to Robust Improvement

Relative Pessimism:

$$\widehat{\pi}^* \in \underset{\pi\in\Pi}{\mathrm{argmax}} \mathcal{L}_\mu\left(\pi, f^\pi\right)$$

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

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

Absolute Pessimism (a variant of [Xie et al., 2021]):

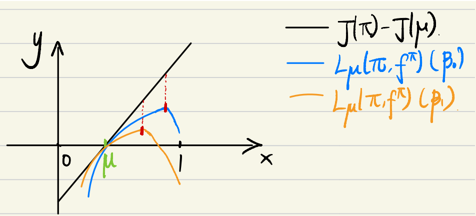
$$\widetilde{\pi}^* \in \underset{\pi\in\Pi}{\mathrm{argmax}} f^\pi(s_0, \pi)$$

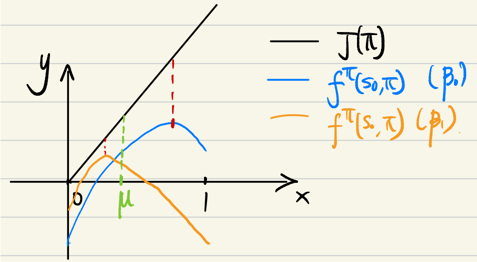
$$\text{s.t.} \quad f \in \underset{f\in\mathcal{F}}{\mathrm{argmin}} f^\pi(s_0, \pi) + \beta\mathcal{E}_\mu(\pi, f)$$

$\forall\pi$, $f^\pi(s_0,\pi) \le f^\pi(s_0,\pi) + \beta\mathcal{E}_\mu(\pi, f) \le Q^\pi(s_0,\pi) + \beta\mathcal{E}_\mu(\pi, Q^\pi) = Q^\pi(s_0,\pi)$.

The Perspective of Lower Bound Maximization

- For illustration, we consider the bandit problem with two actions $(a^1, a^2, r(a^1) > r(a^2))$. The policy π: $\pi(a^1) = x, \pi(a^2) = 1 - x$. The behavior policy μ: $\mu(a^1) = \mu, \mu(a^2) = 1 - \mu$, where $\mu \in (0, 1)$.
- For relative pessimism, we can ensure that the constructed lower bound is sharp at $\pi = \mu$ for any $\beta \geq 0$. $\mathcal{L}_\mu(\mu, f^\mu; \beta) = (1 - \gamma)\,(J(\mu) - J(\mu)) = 0$ for any $\beta \geq 0$.
- For absolute pessimism, the constructed lower bound could be loose at $\pi = \mu$ for some $\beta \geq 0$. For example, when $\beta = 0$, $f^\pi(s_0, \pi) = -\infty, \forall \pi \in \Pi$ if $\mathcal{F}$ contains all state-action functions.





The Perspective of Imitation Learning

When $\beta = 0$:

$$\widehat{\pi}^* \in \underset{\pi \in \Pi}{\operatorname{argmax}} \mathcal{L}_\mu \left(\pi, f^\pi\right)$$

$$\text{s.t.} \quad f^\pi \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \mathcal{L}_\mu(\pi, f).$$

We re-formulate it as

$$\begin{aligned}\max_{\pi \in \Pi} \underset{f \in \mathcal{F}}{\operatorname{argmin}} \mathcal{L}_\mu(\pi, f) &= \max_{\pi \in \Pi} \underset{f \in \mathcal{F}}{\operatorname{argmin}} \mathbb{E}_{(s,a) \sim d^\mu} \left[f(s, \pi) - f(s, a)\right] \\ &\overset{(a)}{=} \max_{\pi \in \Pi} -\mathbb{E}_{s \sim d^\mu} \left[\mathsf{IPM}(\pi(\cdot|s), \mu(\cdot|s))\right].\end{aligned}$$

In (a), it resembles BC which minimizes the policy difference in terms of IPM distance on states induced by d^μ. Here f can be interpreted as the test/discriminator function to distinguish the actions sampled from $\pi(\cdot|s)$ and $\mu(\cdot|s)$.

Offline RL + Relative Pessimism = Imitation + Bellman Regularization.

Adversarially Trained Actor Critic

How to utilize this Stackelberg game formulation to design algorithm?

- $\mathcal{L}_\mu(\pi, f) = \mathbb{E}_{(s,a)\sim\mu}[f(s,\pi) - f(s,a)]$ and $\mathcal{E}_\mu(\pi, f) = \mathbb{E}_\mu\left[\left(\left(f - \mathcal{T}^\pi f\right)(s,a)\right)^2\right]$ is defined with μ and $\mathcal{T}^\pi$, which is unknown to the learner.
- Empirical estimates: $\mathcal{L}_\mathcal{D}(f,\pi) := \mathbb{E}_\mathcal{D}[f(s,\pi) - f(s,a)]$.

$$\mathbb{E}_\mu\left[\left(f(s,a) - \mathcal{T}^\pi f(s,a)\right)^2\right] = \mathbb{E}_{\mu\times(P,R)}\left[\left(f(s,a) - r - \gamma f(s',\pi)\right)^2\right]$$
$$- \mathbb{E}_\mu\left[\mathrm{Var}_{(R,P)}\left[r + \gamma f(s',\pi)|s,a\right]\right]$$
$$\mathcal{E}_\mathcal{D}(f,\pi) := \mathbb{E}_\mathcal{D}\left[\left(f(s,a) - r - \gamma f\left(s',\pi\right)\right)^2\right]$$
$$- \min_{f'\in\mathcal{F}} \mathbb{E}_\mathcal{D}\left[\left(f'(s,a) - r - \gamma f\left(s',\pi\right)\right)^2\right].$$

- We directly know that $\mathcal{L}_\mathcal{D}(f,\pi)$ can approximate $\mathcal{L}_\mu(f,\pi)$ well. Later, we will show that $\mathcal{E}_\mathcal{D}(f,\pi)$ is a good estimate of the true Bellman error $\mathcal{E}_\mu(\pi, f)$ [Antos et al., 2008].

Solving the Stackelberg game

$$\widehat{\pi}^* \in \underset{\pi \in \Pi}{\operatorname{argmax}} \mathcal{L}_{\mathcal{D}}\left(\pi, f^{\pi}\right)$$

$$\text{s.t.} \quad f^{\pi} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \mathcal{L}_{\mathcal{D}}(\pi, f) + \beta \mathcal{E}_{\mathcal{D}}(\pi, f)$$

Best Response: $f_k \in \arg\min_{f \in \mathcal{F}} \mathcal{L}_{\mathcal{D}}\left(\pi_k, f\right) + \beta \mathcal{E}_{\mathcal{D}}\left(\pi_k, f\right)$

Repeat for K iterations

No-regret Learning: $\pi_{k+1} = \operatorname{NoRegret}(\pi_k, f_k)$

Output: $\overline{\pi} = \textbf{Unif}\left(\{\pi_1, \cdots, \pi_K\}\right)$

No-regret Learning Oracle

Definition 1 (No-regret policy optimization oracle).

An algorithm PO is called a no-regret policy optimization oracle if for any sequence of value functions $f_1, \cdots, f_K$ where each $f_k : \mathcal{S} \times \mathcal{A} \to [0, V_{\mathsf{max}}]$, PO outputs a sequence of policies $\pi_1, \cdots, \pi_K$ output by PO satisfy, for any comparator $\pi \in \Pi$:

$$\varepsilon_{\mathrm{opt}}^{\pi} := \frac{1}{1-\gamma} \sum_{k=1}^{K} \mathbb{E}_{\pi} \left[f_k(s, \pi) - f_k\left(s, \pi_k\right) \right] = o(K).$$

- In some scenarios, we can apply natural policy gradient based on multiplicative weights updates (an instance of mirror descent) [Shalev-Shwartz, 2012].

$$\pi_{k+1}(a \mid s) \propto \pi_k(a \mid s) \exp\left(\eta f_k(s, a)\right), \text{with step size } \eta.$$

Theoretical Guarantees: RPI

Theorem 2 (Robust Policy Improvement).

Assume that $\mathcal{F}$ satisfies the realizability and $\mu \in \Pi$, consider that $\bar{\pi}$ is the learned policy, with probability at least $1-\delta$,

$$J(\mu) - J(\bar{\pi}) \leq \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^2 N^{1/2}} + \frac{\beta}{(1-\gamma)^3 N}\right) + \frac{\varepsilon_{\mathit{opt}}^{\mu}}{K}.$$

- RPI: As long as $\beta = o(N)$, $\bar{\pi}$ is guaranteed to outperform μ when N is large and the number of iteration K is large. This robust policy improvement property is due to the relative pessimism principle.
- When $\beta = 0$, ATAC based on relative pessimism is still guaranteed to learn a policy which is no worse than the behavior policy μ. On the other hand, the methods based on absolute pessimism degenerates when the Bellman error loss is removed.

Proof Sketch: Error Decomposition

As $\bar{\pi}$ is the mixture policy between $\{\pi_1, \cdots, \pi_K\}$, $J(\mu) - J(\bar{\pi}) = \frac{1}{K}\sum_{k=1}^{K} J(\mu) - J(\pi^k)$,

$$
\begin{aligned}
J(\mu) - J(\pi^k) &= \mathbb{E}_{s\sim d^\mu}\left[Q^{\pi^k}(s,\mu) - Q^{\pi^k}(s,\pi^k)\right] && \text{Policy difference lemma} \\
&= -\mathcal{L}_\mu(\pi_k, Q^{\pi_k}) \\
&= \mathcal{L}_\mu\left(\pi_k, f_k\right) - \mathcal{L}_\mu(\pi_k, Q^{\pi_k}) - \mathcal{L}_\mu\left(\pi_k, f_k\right) \\
&= \mathcal{L}_\mu\left(\pi_k, f_k\right) - \mathcal{L}_\mu(\pi_k, Q^{\pi_k}) + \underbrace{\mathbb{E}_{s\sim d^\mu}\left[f_k(s,\mu) - f_k(s,\pi_k)\right]}_{\text{opt}_k}.
\end{aligned}
$$

Recall $f_k = \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{L}_\mathcal{D}(\pi_k, f) + \beta\mathcal{E}_\mathcal{D}(\pi_k, f)$ Intuition: ignoring the statistical error ($f_k = \operatorname{argmin}_{f\in\mathcal{F}} \mathcal{L}_\mu(\pi_k, f) + \beta\mathcal{E}_\mu(\pi_k, f)$), we directly have $\mathcal{L}_\mu\left(\pi_k, f_k\right) - \mathcal{L}_\mu(\pi_k, Q^{\pi_k}) \le 0$ due to the lower bound argument.

Proof Sketch: Error Decomposition

We only need to address the statistical error. Construct the empirical objective $\mathcal{L}_{\mathcal{D}}\left(\pi_k, f_k\right)+\beta \varepsilon_{\mathcal{D}}\left(\pi_k, f_k\right)$.

$$
\begin{aligned}
& \mathcal{L}_\mu\left(\pi_k, f_k\right)-\mathcal{L}_\mu(\pi_k, Q^{\pi_k}) \\
& =\mathcal{L}_{\mathcal{D}}\left(\pi_k, f_k\right)-\mathcal{L}_{\mathcal{D}}(\pi_k, Q^{\pi_k})+\underbrace{\left|\mathcal{L}_\mu\left(\pi_k, f_k\right)-\mathcal{L}_{\mathcal{D}}\left(\pi_k, f_k\right)\right|+\left|\mathcal{L}_\mu(\pi_k, Q^{\pi_k})-\mathcal{L}_{\mathcal{D}}(\pi_k, Q^{\pi_k})\right|}_{\varepsilon_{\text{stat}}^1} \\
& \leq \mathcal{L}_{\mathcal{D}}\left(\pi_k, f_k\right)+\beta \varepsilon_{\mathcal{D}}\left(\pi_k, f_k\right)-\mathcal{L}_{\mathcal{D}}(\pi_k, Q^{\pi_k})+\varepsilon_{\text{stat}}^1 \\
& =\mathcal{L}_{\mathcal{D}}\left(\pi_k, f_k\right)+\beta \varepsilon_{\mathcal{D}}\left(\pi_k, f_k\right)-\mathcal{L}_{\mathcal{D}}(\pi_k, Q^{\pi_k})-\beta \varepsilon_{\mathcal{D}}(\pi_k, Q^{\pi_k})+\beta \varepsilon_{\mathcal{D}}(\pi_k, Q^{\pi_k})+\varepsilon_{\text{stat}}^1 \\
& \leq \beta \varepsilon_{\mathcal{D}}(\pi_k, Q^{\pi_k})+\varepsilon_{\text{stat}}^1 \\
& \leq \beta \varepsilon_{\text{stat}}^2+\varepsilon_{\text{stat}}^1 .
\end{aligned}
$$

In the last inequality, $\varepsilon_{\mathcal{D}}(\pi_k, Q^{\pi_k})=|\varepsilon_{\mathcal{D}}(\pi_k, Q^{\pi_k})-\varepsilon_\mu(\pi_k, Q^{\pi_k})| \leq \varepsilon_{\text{stat}}^2$. $\varepsilon_{\text{stat}}^1$ can be directly upper bounded by Hoeffding's inequality. We will upper bound $\varepsilon_{\text{stat}}^2$.

The Bellman Error Proxy

Theorem 3.

For any $\delta \in (0,1)$, *w.p.*$\geq 1-\delta$, $\forall \pi \in \Pi$,

$$\mathcal{E}_{\mathcal{D}}(\pi, Q^\pi) = \mathcal{E}_{\mathcal{D}}(\pi, Q^\pi) - \|Q^\pi - \mathcal{T}^\pi Q^\pi\|_{2,\mu}^2 \leq \frac{23 V_{\max}^2 \log\left(2|\mathcal{F}||\Pi|/\delta\right)}{n}.$$

The Bellman Error Proxy

$\mathcal{E}_{\mathcal{D}}(f,\pi) := \mathbb{E}_{\mathcal{D}}\left[\left(f(s,a)-r-\gamma f\left(s',\pi\right)\right)^2\right] - \min_{f'\in\mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\left(f'(s,a)-r-\gamma f\left(s',\pi\right)\right)^2\right].$

$\mathcal{E}_{\mathcal{D}}(f,\pi)$ is a proxy of the Bellman error $\|f-\mathcal{T}^{\pi}f\|_{2,\mu}^2$. To understand this claim, the key observation builds upon the classical bias-variance decomposition. Let $g \in \operatorname{argmin}_{f'\in\mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\left(f'(s,a)-r-\gamma f\left(s',\pi\right)\right)^2\right]$ be the empirical minimizer of the regression problem.

$$\begin{aligned}
\mathcal{E}_{\mathcal{D}}(f,\pi) &\approx \mathbb{E}_{\mu\times(P,R)}\left[\left(f(s,a)-r-\gamma f(s',\pi)\right)^2\right] - \mathbb{E}_{\mu\times(P,R)}\left[\left(g(s,a)-r-\gamma f(s',\pi)\right)^2\right] \\
&= \left(\mathbb{E}_{\mu}\left[\left(f(s,a)-\mathcal{T}^{\pi}f(s,a)\right)^2\right] + \mathbb{E}_{\mu}\left[\operatorname{Var}_{(R,P)}\left[r+\gamma f(s',\pi)|s,a\right]\right]\right) \\
&- \left(\mathbb{E}_{\mu}\left[\left(g(s,a)-\mathcal{T}^{\pi}f(s,a)\right)^2\right] + \mathbb{E}_{\mu}\left[\operatorname{Var}_{(R,P)}\left[r+\gamma f(s',\pi)|s,a\right]\right]\right) \\
&= \underbrace{\mathbb{E}_{\mu}\left[\left(f(s,a)-\mathcal{T}^{\pi}f(s,a)\right)^2\right]}_{\|f-\mathcal{T}^{\pi}f\|_{2,\mu}^2} - \underbrace{\mathbb{E}_{\mu}\left[\left(g(s,a)-\mathcal{T}^{\pi}f(s,a)\right)^2\right]}_{\|g-\mathcal{T}^{\pi}f\|_{2,\mu}^2}.
\end{aligned}$$

$\|g-\mathcal{T}^{\pi}f\|_{2,\mu}^2$ measures the difference between the empirical minimizer g and the population minimizer (Bayes-optimal minimizer) $\mathcal{T}^{\pi}f$, which diminishes as n increases.

Proof Sketch: The Bellman Error Proxy

Lemma 4: $$\left|\mathcal{E}_{\mathcal{D}}(f,\pi) - \left(\|f - \mathcal{T}^{\pi} f\|_{2,\mu}^2 - \|g - \mathcal{T}^{\pi} f\|_{2,\mu}^2\right)\right| \leqslant \tilde{O}\left(\frac{\|g - f\|_{2,\mu}}{\sqrt{N}}\right)$$

Lemma 5: $$\|g - \mathcal{T}^{\pi} f\|_{2,\mu}^2 \leq \tilde{O}\left(\frac{1}{N}\right)$$

$\downarrow$

$$\left|\mathcal{E}_{\mathcal{D}}(f,\pi) - \|f - \mathcal{T}^{\pi} f\|_{2,\mu}^2\right| \leq \tilde{O}\left(\frac{\|g - f\|_{2,\mu}}{\sqrt{N}}\right)$$

$\downarrow$ Set $f = Q^{\pi}$

Theorem 3: $$\mathcal{E}_{\mathcal{D}}(Q^{\pi},\pi) \leq \tilde{O}\left(\frac{\|g - Q^{\pi}\|_{2,\mu}}{\sqrt{N}}\right) = \tilde{O}\left(\frac{\|g - \mathcal{T}^{\pi} Q^{\pi}\|_{2,\mu}}{\sqrt{N}}\right) = \tilde{O}(\frac{1}{N})$$

Proof Sketch: The Bellman Error Proxy

Lemma 4.

For any $\delta \in (0,1)$, w.p. $\geq 1-\delta$, for any $f, g_1, g_2 \in \mathcal{F}$ and $\pi \in \Pi$,

$$\Bigg| \|g_1 - \mathcal{T}^\pi f\|_{2,\mu}^2 - \|g_2 - \mathcal{T}^\pi f\|_{2,\mu}^2 - \left(\mathbb{E}_{\mathcal{D}}\left[\left(g_1(s,a) - r - \gamma f\left(s', \pi\right)\right)^2 \right] - \mathbb{E}_{\mathcal{D}}\left[\left(g_2(s,a) - r - \gamma f\left(s', \pi\right)\right)^2 \right] \right) \Bigg| \leq \|g_1 - g_2\|_{2,\mu} \sqrt{\frac{24 V_{\max}^2 \log(2|\mathcal{F}||\Pi|/\delta)}{n}} + \frac{V_{\max}^2 \log(2|\mathcal{F}||\Pi|/\delta)}{n}.$$

The proof builds upon the bias-variance decomposition. To establish a sharp bound, we need to use Bernstein's inequality, which introduces the variance term related to $\|g_1 - g_2\|_{2,\mu}$.

Lemma 5.

Consider a real-valued regression problem with feature space $\mathcal{X}$ and label space $\mathcal{Y} \in [0, V_{\max}]$. Let $\{(X_i, Y_i))\}_{i=1}^n$ be the i.i.d. data where $X_i \sim P(\cdot)$ and $Y_i \sim Q(\cdot|X_i)$. Let $\mathcal{F} \subset \mathcal{X} \to \mathcal{Y}$ be the function class, which is assumed to be finite but exponentially large. Let $f^*(X) = \mathbb{E}[Y|X]$ be the Bayes-optimal minimizer. We assume that $\mathcal{F}$ satisfies the realizability, i.e., $f^* \in \mathcal{F}$. Let $\widehat{f}^*$ be the empirical risk minimizer on $\{(X_i, Y_i))\}_{i=1}^n$.

$$\widehat{f}^* = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2 .$$

Then w.p. $\geq 1 - \delta$,

$$\mathbb{E}_{x \sim P(\cdot)} \left[\left(\hat{f}^*(x) - f^*(x) \right)^2 \right] \leq \frac{8 V_{\max}^2 \log 2 \left(|\mathcal{F}| / \delta \right)}{n} .$$

In offline RL, the realizability corresponds to the Bellman completeness, i.e., $\forall \pi \in \Pi$, $\forall f \in \mathcal{F}$, $\mathcal{T}^\pi f \in \mathcal{F}$.

Theoretical Guarantees: Learning Optimality

Theorem 6 (Learning Optimality).

Assume that $\mathcal{F}$ satisfies the realizability and completeness. We define $\mathscr{C}(\nu;\mu,\mathcal{F},\pi) := \max_{f\in\mathcal{F}} \frac{\mathcal{E}_\nu(\pi,f)}{\mathcal{E}_\mu(\pi,f)}$. For any policy $\pi \in \Pi$, suppose that $\max_{k\in[K]} \mathscr{C}(d^\pi;\mu,\mathcal{F},\pi_k) \le C$, with $\beta = \Theta\left(\sqrt[3]{V_{\max} N^2}\right)$, with high probability,

$$J(\pi) - J(\bar{\pi}) \le \mathcal{O}\left(\frac{\sqrt{C}}{(1-\gamma)^2 N^{1/3}}\right) + \frac{\varepsilon_{\text{opt}}^\pi}{K}.$$

- $\mathscr{C}(\nu;\mu,\mathcal{F},\pi)$ measures how well the distribution ν is covered by the data distribution μ w.r.t π and $\mathcal{F}$. This is a sharper measure compared with the concentrability coefficient ($\mathscr{C}(\nu;\mu,\mathcal{F},\pi) \le \max_{s,a} \nu(s,a)/\mu(s,a), \forall\pi, \forall\mathcal{F}$).
- This result shows that the learned policy can compete with any policy whose state-action distribution is covered by the dataset, when N is large and the number of iterations is large.

A Practical Implementation of ATAC

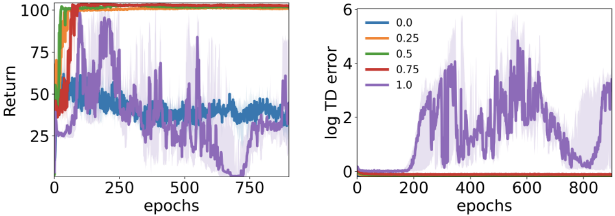
- In practice, the above process is approximated by a two-timescale gradient-based method.
- The critic is updated with a much faster rate η_{fast} than the actor with η_{slow}.
- The critic update with η_{fast} can mimic the best response procedure and the actor update with η_{slow} can mimic the no-regret learning procedure.

Practical Designs of Critic Update

- Projection: We parameterize $\mathcal{F}$ as neural networks with ℓ_2 bounded weights. The projection is crucial to ensure stable learning across all β-values.
- A surrogate loss of the Bellman error $\mathcal{E}_{\mathcal{D}}(f,\pi)$:

$$\mathcal{E}_{\mathcal{D}}^{w}(f,\pi) := (1-w)\mathcal{E}_{\mathcal{D}}^{\mathrm{td}}(f,f,\pi) + w\mathcal{E}_{\mathcal{D}}^{\mathrm{td}}\left(f,\bar{f}_{\min},\pi\right),$$

  where $w \in (0,1)$, $\mathcal{E}_{\mathcal{D}}^{\mathrm{td}}(f,f',\pi) := \mathbb{E}_{\mathcal{D}}\left[\left(f(s,a) - r - \gamma f'(s',\pi)\right)^2\right]$ and $\bar{f}_{\min}(s,a) := \min_{i=1,2} \bar{f}_i(s,a)$. Using this surrogate loss significantly improves the optimization stability.



Practical Designs of Actor Update

- Actor loss with a single critic: While the critic optimization uses the double Q networks for numerical stability, the actor loss only uses one of the critics (f_1). This is different from SAC which takes $\min_{i=1,2} f_i(s,a)$ as the objective.
- This design choice is critical to enable ATAC's IL behavior when β is low.

Experimental Evaluations

- Target: test the effectiveness of ATAC in terms of performance and robust policy improvement.
- Baseline: CQL, COMBO, TD3+BC, IQL, BC.
- Variants of ATAC:
  - An absolute pessimism version of ATAC (denoted ATAC_0): $\mathcal{L}_{\mathcal{D}}(f, \pi)$ in $\mathcal{L}_{\text{critic}}$ is replaced with $f(s_0, \pi)$.
  - Since ATAC does not have guarantees on last-iterate convergence, they report also the results of both the last iterate (denoted as ATAC and ATAC_0) and the best checkpoint (denoted as $\mathrm{ATAC}^{\star}$ and ATAC_0) selected among 9 checkpoints.

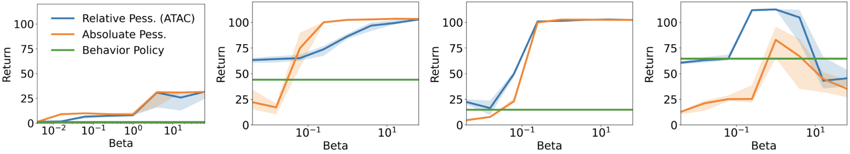
Performance

	Behavior	ATAC*	ATAC	ATAC_0^*	ATAC_0	CQL	COMBO	TD3+BC	IQL	BC
halfcheetah-rand	-0.1	4.8	3.9	2.3	2.3	35.4	**38.8**	10.2	-	2.1
walker2d-rand	0.0	**8.0**	6.8	7.6	5.7	7.0	7.0	1.4	-	1.6
hopper-rand	1.2	**31.8**	17.5	31.6	18.2	10.8	17.9	11.0	-	9.8
halfcheetah-med	40.6	**54.3**	**53.3**	43.9	36.8	44.4	**54.2**	42.8	47.4	36.1
walker2d-med	62.0	**91.0**	**89.6**	**90.5**	**89.6**	74.5	75.5	79.7	78.3	6.6
hopper-med	44.2	**102.8**	85.6	**103.5**	94.8	86.6	94.9	**99.5**	66.3	29.0
halfcheetah-med-replay	27.1	49.5	48.0	49.2	47.2	46.2	**55.1**	43.3	44.2	38.4
walker2d-med-replay	14.8	**94.1**	92.5	**94.2**	89.8	32.6	56.0	25.2	73.9	11.3
hopper-med-replay	14.9	**102.8**	**102.5**	**102.7**	**102.1**	48.6	73.1	31.4	94.7	11.8
halfcheetah-med-exp	64.3	**95.5**	**94.8**	41.6	39.7	62.4	90.0	**97.9**	86.7	35.8
walker2d-med-exp	82.6	**116.3**	**114.2**	**114.5**	104.9	98.7	96.1	101.1	**109.6**	6.4
hopper-med-exp	64.7	**112.6**	**111.9**	83.0	46.5	**111.0**	**111.1**	**112.2**	91.5	**111.9**
pen-human	207.8	79.3	53.1	**106.1**	61.7	37.5	-	-	71.5	34.4
hammer-human	25.4	**6.7**	1.5	3.8	1.2	**4.4**	-	-	1.4	1.5
door-human	28.6	8.7	2.5	**12.2**	7.4	**9.9**	-	-	4.3	0.5
relocate-human	86.1	**0.3**	**0.1**	**0.5**	**0.1**	**0.2**	-	-	**0.1**	**0.0**
pen-cloned	107.7	73.9	43.7	**104.9**	68.9	39.2	-	-	37.3	56.9
hammer-cloned	8.1	2.3	1.1	**3.2**	0.4	2.1	-	-	2.1	0.8
door-cloned	12.1	**8.2**	3.7	6.0	0.0	0.4	-	-	1.6	-0.1
relocate-cloned	28.7	**0.8**	**0.2**	**0.3**	**0.0**	**-0.1**	-	-	**-0.2**	**-0.1**
pen-exp	105.7	**159.5**	136.2	**154.4**	97.7	107.0	-	-	-	85.1
hammer-exp	96.3	**128.4**	**126.9**	118.3	99.2	86.7	-	-	-	**125.6**
door-exp	100.5	**105.5**	**99.3**	**103.6**	48.3	**101.5**	-	-	-	34.9
relocate-exp	101.6	**106.5**	**99.4**	**104.0**	74.3	95.0	-	-	-	**101.3**

Table 1. Evaluation on the D4RL dataset. Algorithms with score within ϵ from the best on each domain are marked in bold, where $\epsilon = 0.1|J(\mu)|$. Baseline results are from the respective papers. For ATAC variants, we take the median score over 10 seeds.

ATAC and $\text{ATAC}^\star$ outperforms other model-free baselines consistently and model-based method COMBO mostly..

Robust Policy Improvement



(a) hopper-random (b) hopper-medium (c) hopper-medium-replay (d) hopper-medium-expert

- ATAC based on relative pessimism improves from behavior policies over a wide range of hyperparameters β. On the other hand, offline RL based on absolute pessimism has safe policy improvement only for well-tuned hyperparameters β.
- This empirical results validate the role of relative pessimism on robust policy improvement.

References I

A. Antos, C. Szepesvári, and R. Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. Machine Learning, 71(1): 89–129, 2008.

D. J. Foster, A. Krishnamurthy, D. Simchi-Levi, and Y. Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In Proceedings of the 35th Annual Conference on Learning Theory, page 3489, 2022.

S. Shalev-Shwartz. Online learning and online convex optimization. Foundations and Trends in Machine Learning, 4(2):107–194, 2012.

R. Wang, D. P. Foster, and S. M. Kakade. What are the statistical limits of offline RL with linear function approximation? In Proceedings of the 9th International Conference on Learning Representations, 2021.

References II

T. Xie, C. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offline reinforcement learning. arXiv, 2106.06926, 2021.