Is Pessimism Provably Efficient for Offline RL?

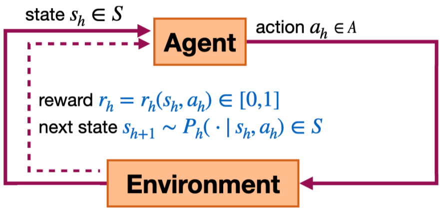
Ying Jin [1] Zhuoran Yang [2] Zhaoran Wang [3]

[1]Stanford University

[2]Princeton University

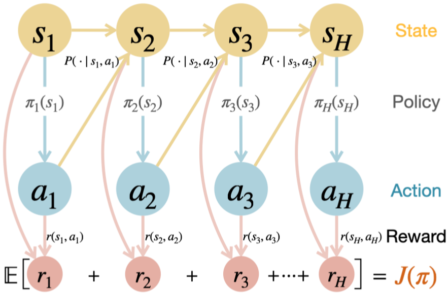
[3]Northwestern University

Episodic MDP



- $\mathcal{S}$: infinite state space. $\mathcal{A}$: finite action space.
- Unknown reward function $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$.
- Unknown transition kernel $\mathbb{P}_h(\cdot \mid x, a) \in \Delta(\mathcal{S})$.
- Finite horizon H: terminate when $h = H$.

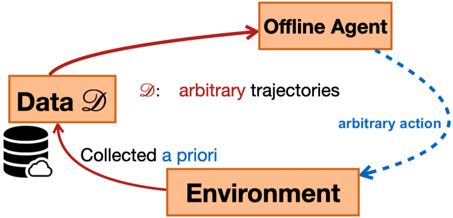
Episodic MDP



- Policy: $\pi = \{\pi_h\}_{h\in[H]} : \mathcal{S} \to \Delta(\mathcal{A})$, $a_h \sim \pi_h(s_h)$.
- Expected total reward: $J(\pi, x) = \mathbb{E}_\pi[\sum_{h=1}^{H} r_h \mid s_1 = x] \in [0, H]$.
- Optimal policy: $\pi^\star(\cdot) = \mathrm{argmax}_\pi\, J(\pi, \cdot)$.

Offline Policy Learning

Learn from Given Datasets



- Offline Data: collected a priori.
- Arbitrary trajectories: actions a_h by an offline agent (unknown rule).
- No further interactions with MDP.
- Learning objective: performance of the learned policy

$$\text{SubOpt}(\widehat{\pi}, x) = J(\pi^\star, x) - J(\widehat{\pi}, x),$$

where $\widehat{\pi} = \text{OfflineRL}(\mathcal{D}, \mathcal{F})$, $x \in \mathcal{S}$.

Why May Greedy Value Iterations Fail?

Epistemic Uncertainty

- Some policy $\widetilde{\pi}$ might be insufficiently covered by dataset $\mathcal{D}$
  $\Rightarrow$ Large uncertainty in our knowledge about a policy $\widetilde{\pi}$.
- Epistemic Uncertainty spuriously correlates with decision-making,

$$J(\widehat{\pi}) = J\Big(\operatorname*{argmax}_{\pi} \widehat{J}(\pi)\Big).$$

  $\widehat{J}$ might be far from J for some π.

- **Ruined if a bad π with large uncertainty appears to be good!**
- No further interactions with MDP $\Rightarrow$ unable to reduce uncertainty.

Question

Is it possible to design a provably efficient algorithm for offline RL under minimal assumptions on the dataset?

- Our solution by **Pessimism**: penalize large epistemic uncertainties.

General Algorithm: Pessimistic Value Iteration

Algorithm: Pessimistic Value Iterations (General Form)

- Estimate: $\overline{Q}_h \leftarrow \texttt{Regress}(\mathbb{B}_h \widehat{Q}_{h+1}, \mathcal{D}, \mathcal{F})$.
- Uncertainty quantification (UQ): w.h.p.

$$\left|\overline{Q}_h - (\mathbb{B}_h \widehat{Q}_{h+1})\right| \leq \Gamma_h, \quad \forall h \in [H].$$

- Construct pessimistic value function

$$\widehat{Q}_h(x,a) = \underbrace{\overline{Q}_h(x,a)}_{\text{VI}} \underbrace{-\Gamma_h(x,a)}_{\text{penalty}}$$

- Optimize: $\widehat{\pi}_h(x) = \operatorname{argmax}_{a \in \mathcal{A}} \widehat{Q}_h(x,a)$.

Why Pessimism Helps?

Suboptimality Upper Bound [1]

- A clean suboptimality bound

$$\text{SubOpt}(\widehat{\pi}; x) \le 2 \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\Gamma_h(s_h, a_h) \,\middle|\, s_1 = x\right]$$

  - Only depends on the trajectory of π^*
  - Pessimism eliminates spurious correlation.

[1] Adapted from Theorem 4.2 in (JYW'20)

Why Pessimism Helps?

Suboptimality Upper Bound [1]

- A clean suboptimality bound

$$\mathrm{SubOpt}(\widehat{\pi}; x) \leq 2 \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\Gamma_h(s_h, a_h) \,\middle|\, s_1 = x\right]$$

  - Only depends on the trajectory of π^*
  - Pessimism eliminates spurious correlation.

Question

How to construct the uncertainty quantifier Γ_h?

[1] Adapted from Theorem 4.2 in (JYW'20)

Instantiation of PEVI

Linear MDP

Definition (Linear MDP)

We say an episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is a linear MDP with a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if there exist d unknown (signed) measures $\mu_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$ such that

$$\mathbb{P}_h(x' \mid x, a) = \langle \phi(x, a), \mu_h(x') \rangle,$$
$$\mathbb{E}\big[r_h(s_h, a_h) \,\big|\, s_h = x, a_h = a\big] = \langle \phi(x, a), \theta_h \rangle$$

for all $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ at each step $h \in [H]$. Here we assume $\|\phi(x, a)\| \le 1$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$ and $\max\{\|\mu_h(\mathcal{S})\|, \|\theta_h\|\} \le \sqrt{d}$ at each step $h \in [H]$, where $\|\mu_h(\mathcal{S})\| = \int_{\mathcal{S}} \|\mu_h(x)\| \,\mathrm{d}x$.

- Linearity of Bellman update: $\mathbb{B}_h \widehat{Q}_{h+1} = \phi^\top \widehat{\theta}_h$ for some $\widehat{\theta}_h \in \mathbb{R}^d$.
- Linear function approximation $\mathcal{F} = \{f_\theta(x, a) = \phi(x, a)^\top \theta,\ \theta \in \mathbb{R}^d\}$.

Instantiation of PEVI

Linear MDP

Algorithm: PEVI for Linear MDP

- Estimate: $\overline{Q}_h(x,a) = \phi(x,a)^\top \widehat{\theta}_h$ via ridge regression.
- Uncertainty quantification

$$\Gamma_h(x,a) \asymp dH \cdot \big(\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)\big)^{1/2},$$

  where Λ_h is the augmented sample covariance matrix of $\phi(s_h, a_h)$.

- Pessimistic value function

$$\widehat{Q}_h(x,a) = \phi(x,a)^\top \widehat{\theta}_h - c \cdot dH \cdot \big(\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)\big)^{1/2}$$

- Optimize: $\widehat{\pi}_h(x) = \mathrm{argmax}_{a \in \mathcal{A}} \widehat{Q}_h(x,a)$.

Instantiation of PEVI - Linear MDP

Compliance Assumption

Assumption: Compliance

Let $\mathbb{P}_{\mathcal{D}}$ be the joint distribution of the dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$. We say $\mathcal{D}$ is compliant with an MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ if

$$\mathbb{P}_{\mathcal{D}}\big(r_h^\tau = r', x_{h+1}^\tau = x' \,\big|\, \{(x_h^j, a_h^j)\}_{j=1}^{\tau}, \{(r_h^j, x_{h+1}^j)\}_{j=1}^{\tau-1}\big)$$
$$= \mathbb{P}\big(r_h = r', s_{h+1} = x' \,\big|\, s_h = x_h^\tau, a_h = a_h^\tau\big)$$

for all $r' \in [0,1]$, $x' \in \mathcal{S}$, $h \in [H]$, $\tau \in [K]$. Here $\mathbb{P}$ is taken with respect to the underlying MDP.

- Only require that $\mathcal{D}$ evolves according to the MDP.
- Minimal assumptions on actions a_h^τ: allow for arbitrarily collected data.
  - i.i.d. trajectories from a behavior policy ✓
  - sequentially adjusted actions $a_h^\tau \in \sigma(\{x_{h+1}^j, r_h^j\}_{j<\tau})$ ✓

Instantiation of PEVI - Linear MDP

Suboptimality Upper Bound

Theorem 4.4 (JYW'20)

If $\mathcal{D}$ is compliant with the underlying MDP, then w.h.p,

$$\text{SubOpt}\big(\widehat{\pi}; x\big) \leq c \cdot dH \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \Big| s_1 = x\Big].$$

up to logarithm factors of d, H, K.

- Minimal-assumption guarantee: only require compliance of $\mathcal{D}$.
- Oracle property: only depends on how well $\pi^\star$ is covered - no requirement on coverage of all trajectories.
- Data-dependent upper bound: (offline) data is what it is.

Question

Is coverage of optimal $\pi^\star$ the essential information in $\mathcal{D}$?

Minimax Optimality of Pessimism: Linear MDP

- Answer: Coverage of optimal $\pi^\star$ is the essential information in $\mathcal{D}$.
- Pessimism is (nearly) minimax optimal in linear setting.

Minimax Optimality of Pessimism: Linear MDP

- Answer: Coverage of optimal $\pi^\star$ is the essential information in $\mathcal{D}$.
- Pessimism is (nearly) minimax optimal in linear setting.

Minimax Optimality in Linear MDP

- Upper bound: pessimistic policy $\widehat{\pi}$ and compliant $\mathcal{D} \sim \mathcal{M}$,

$$\mathsf{SubOpt}\big(\mathcal{M}, \widehat{\pi}; x\big) \le c \cdot dH \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x\Big].$$

Minimax Optimality of Pessimism: Linear MDP

- Answer: Coverage of optimal $\pi^\star$ is the essential information in $\mathcal{D}$.
- Pessimism is (nearly) minimax optimal in linear setting.

Minimax Optimality in Linear MDP

- Upper bound: pessimistic policy $\widehat{\pi}$ and compliant $\mathcal{D} \sim \mathcal{M}$,

$$\mathsf{SubOpt}\big(\mathcal{M}, \widehat{\pi}; x\big) \leq c \cdot dH \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x\Big].$$

- Lower bound: for any offline learning algorithm $\texttt{Algo}(\cdot)$,

$$\sup_{\mathcal{M}, \mathcal{D}} \mathbb{E}_{\mathcal{D}}\left[\frac{\mathsf{SubOpt}\big(\mathcal{M}, \texttt{Algo}(\mathcal{D}); x\big)}{\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x\Big]}\right] \geq c.$$

  - Dependence on true MDP $\mathcal{M}$ and its optimal policy $\pi^\star$.
  - Essential Hardness in $\mathcal{D}$: how well (sample covariance) Λ_h covers $\pi^\star$.