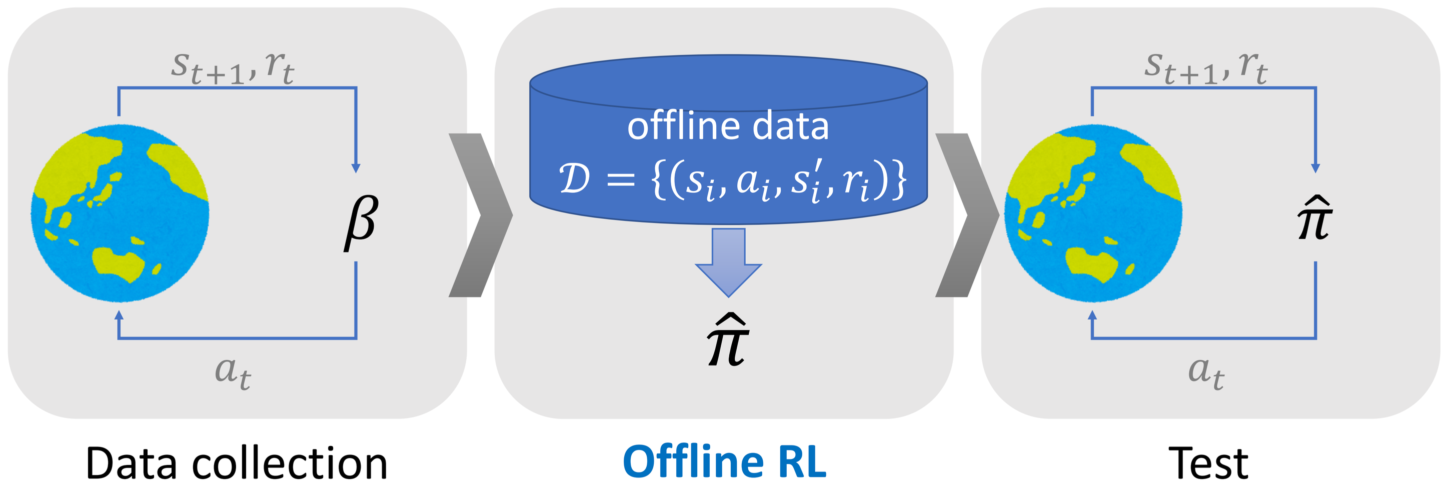
NEURAL INFORMATION PROCESSING SYSTEMS

Worst-Case Offline Reinforcement Learning with Arbitrary Data Support

Kohei Miyaguchi (IBM Research – Tokyo/LY Corporation), NeurIPS2024

TLDR; weakest known sufficient condition for offline RL is updated, specifically in terms of data coverage and sample size.

Background



Q. What is weakest practical condition for successful offline RL?

Our contribution

Previous weakest condition [Zhan et al. 2022]

- model-free realizability
- concentrability (data coverage) → **removed**
- sample size of $O(\epsilon^{-6})$ → **reduced to $O(\epsilon^{-2})$**

Why removing concentrability (CC)?

Requirement:

$$\sup_{s,a} \frac{d^{\hat{\pi}}(s,a)}{p_{\text{data}}(s,a)} < \infty$$

test-time occupancy of $\hat{\pi}$

offline state-action distribution

Issue1: concentrable $\hat{\pi}$ may NOT exist due to

- data fragmentation/censorship
- initial-state distribution shift
- unknown constraints on behavior actions

Issue2: Coefficient of CC is hard to estimate

→ **CC is easily violated and difficult to verify in practice**

Proposal: Worst-case offline RL

New performance metric w/ **built-in pessimism**:

$$\tilde{J}(\pi) := \min_{\mathcal{M} \in \boxed{\mathfrak{U}}} \boxed{J(\pi|\mathcal{M})}$$

policy value under MDP $\mathcal{M}$

uncertainty set under distribution oracle

$$\mathfrak{U} := \{\mathcal{M} = (T, r) : (T, r) = (T^\star, r^\star)|_{\text{supp}(p_{\text{data}})}, 0 \le r \le 1\}$$

Justifications:

1. tractability: can be estimated w/o CC
2. generality: recovers standard metric if CC holds:
$$J(\pi) := J(\pi|\mathcal{M}^\star) = \tilde{J}(\pi), \qquad \forall \pi \in \Pi_{\text{CC}}$$
3. sufficiency: generalized suboptimality dominates standard one:
$$\max_{\pi^* \in \Pi_{\text{CC}}} J(\pi^*) - J(\pi) \le \max_{\tilde{\pi}^* \in \Pi_{\text{all}}} \tilde{J}(\tilde{\pi}^*) - \tilde{J}(\pi)$$

Result 1: Worst-case offline RL is still RL

Def: Worst-case MDP $\widetilde{\mathcal{M}} = (\tilde{T}, \tilde{r})$ is given by

$$\tilde{T}(s,a) = \mathbf{1}_{\{p_{\text{data}}(s,a)>0\}} T^\star(s,a) + \mathbf{1}_{\{p_{\text{data}}(s,a)=0\}} \delta_\perp$$

$$\tilde{r}(s,a) = \mathbf{1}_{\{p_{\text{data}}(s,a)>0\}} r^\star(s,a)$$

where $\perp$ is terminal state.

Thm: $\tilde{J}(\pi) = J(\pi|\widetilde{\mathcal{M}})$ for all π

→ **Standard RL methods are still applicable**

1. solve Bellman equation of $\widetilde{\mathcal{M}}$
2. extract optimal policies from Bellman eq.'s solution

Result 2: Saddle-point characterization

Consider "Lagrangian of offline RL":

$$L(v, f) := \langle (1-\gamma)v + f \cdot (r + Tv - v) \rangle_{\text{data}}$$

Known: Saddle point under $f \ge 0$ is

1. well-defined only if optimal policy π^* is concentrable and
2. solution of standard Bellman eq., i.e., gives optimal value function $v^*(s)$ and optimal occupancy density $f^*(s,a)$.

New: Saddle point under $v \ge 0$ and $f \ge 0$ is

1. well-defined unconditionally and
2. solution of Bellman eq. of $\widetilde{\mathcal{M}}$

Result 3: Algorithm & sample complexity

We propose to minimize

$$\mathcal{L}(f; w, \pi) = \boxed{\mathcal{L}_{\text{SP}}(f)} + \boxed{\mathcal{L}_{\text{PX}}(f; w, \pi)}$$

saddle-point loss   policy-extraction loss

where

$$\mathcal{L}_{\text{SP}}(f) := \max_{v \ge 0} \left\{ -L(v, f) - \frac{1-\gamma}{2} \|v\|^2 \right\}$$

$$\mathcal{L}_{\text{PX}}(f; w, \pi) := \max_{\xi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}} \langle f\xi - w\xi(\cdot, \pi) \rangle_{\text{data}}$$

… achieving SOTA sample complexity bound!

Method	Assumptions: Concentrability	Assumptions: Realizability	Sample complexity bound
Zhan et al. (2022)	π^*	π_n	$\epsilon^{-6}(1-\gamma)^{-4}\ln(\mathcal{N}/\delta)$
Chen and Jiang (2022)	π^*	π^*	$\epsilon^{-2}H^5 C_{\text{gap}}^{-2}\ln(\mathcal{N}/\delta)$
Ozdaglar et al. (2023)	π^*	π^*	$\epsilon^{-2}(1-\gamma)^{-6}C_{\text{gap}}^{-2}\ln(\mathcal{N}/\delta)$
Uehara et al. (2023)	π^*	π^*	$\epsilon^{-2-4/\beta_{\text{gap}}}(1-\gamma)^{-6-4/\beta_{\text{gap}}}\ln(\mathcal{N}/\delta)$
Ours (Corollary 6.3)	—	$\tilde{\pi}^*$	$\epsilon^{-2}(1-\gamma)^{-4}\ln(\mathcal{N}/\delta)$

ϵ: policy subopt; δ: confidence; γ: discount factor; $\mathcal{N}$: hypothesis size; C_{gap}: min action value gap