南京航空航天大學
Nanjing University of Aeronautics and Astronautics

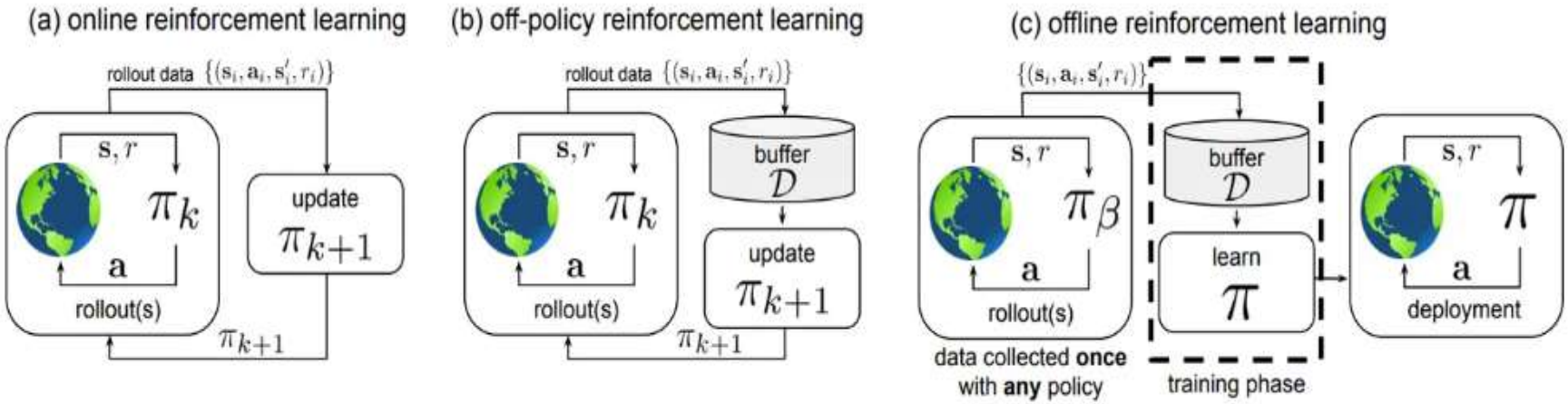
模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Beyond Uniform Sampling: Offline Reinforcement Learning with Imbalanced Datasets

NIPS | 2023

Background

- **Reinfocement learning**



- **Offline Reinfocement learning**

$$J(\pi) = \mathbb{E}_{\tau \sim p_\pi(\tau)} \left[\sum_{t=0}^{H} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t) \right]. \xrightarrow{\text{constraint}} \max_{\pi_\theta} J(\pi_\theta) - \alpha \mathbb{E}_{(s,a)\sim\mathcal{D}} \left[\mathcal{C}(s,a) \right],$$

Motivation

- **Problem Statement**

多数offlineRL模型趋向于选择数据集中出现的动作，当数据集本身不平衡或者次优数据较多，结果会很差。

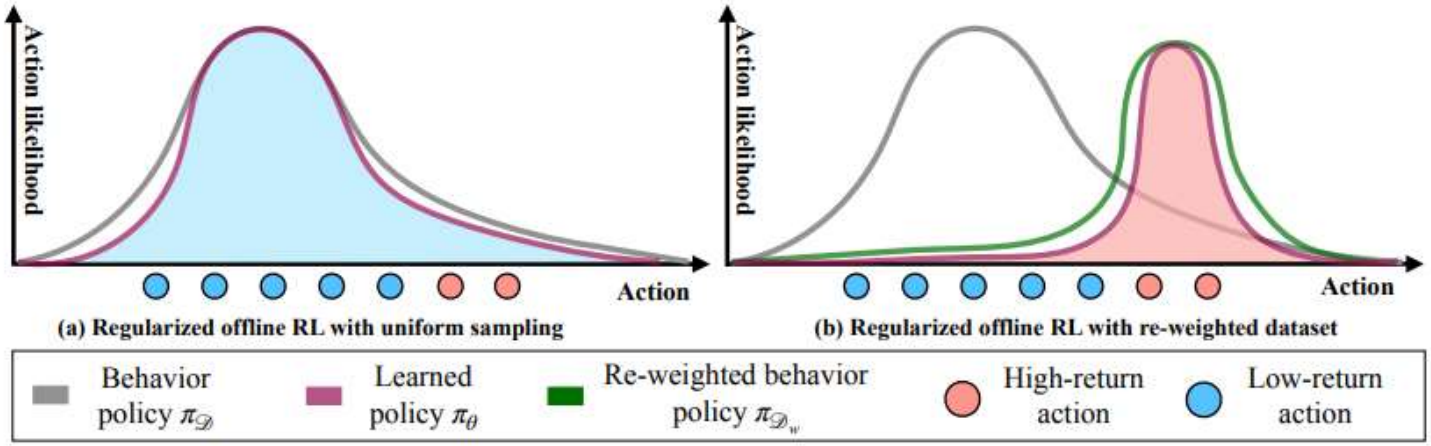


Figure 1: The dots represent actions in the dataset, where imbalanced datasets have more low-return actions. **(a)** Regularized offline RL algorithms [22, 7, 18] equally regularize the policy π_θ on each action, leading to imitation of low-return actions and a low-performing π_θ. The color under the curves shows the policy's performance $J(\pi_\theta)$, with red indicating higher performance and blue indicating lower performance. **(b)** Re-weighting the dataset based on actions' returns allows the algorithm to only regularize on actions with high returns, enabling the policy π_θ to imitate high-return actions while ignoring low-return actions.

Background

Definition 3.1 (Dataset imbalance). RPSV of a dataset, $\mathbb{V}_+[G(\tau_i)]$, corresponds to the second-order moment of the positive component of the difference between trajectory return: $G(\tau_i) := \sum_{t=0}^{T_i-1} \gamma^t r(s_t^i, a_t^i)$ and its expectation, where τ_i denote trajectory in the dataset:

$$\mathbb{V}_+[G(\tau_i)] \doteq \mathbb{E}_{\tau_i \sim \mathcal{D}} \left[(G(\tau_i) - \mathbb{E}_{\tau_i \sim \mathcal{D}}[G(\tau_i)])_+^2 \right] \quad \text{with} \quad x_+ = \max\{x, 0\}, \tag{3}$$

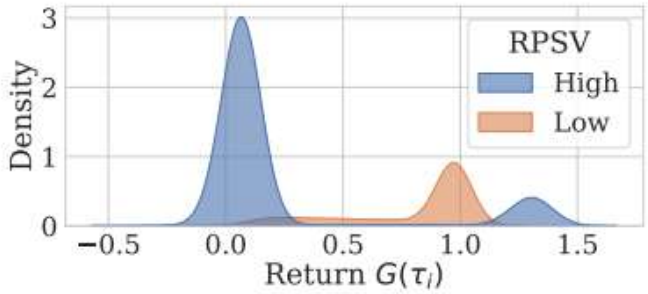


Figure 2: Return distribution of datasets with high and low RPSV. Low RPSV datasets have returns centered at the mean, while high RPSV datasets have a wider distribution extending towards higher returns. See Appendix A.4 for details.

- **unnecessarily conservative**

Imbalance dataset + constraint

Methodology

- **Mitigating Unnecessary Conservativeness By Weighting Samples**

Adding more experiences from high-performing policies

↓ Importance sampling

$$w(s,a) = \frac{\mathcal{D}_w(s,a)}{\mathcal{D}(s,a)}, \qquad \text{(Dw unknown)}$$

↓

$$\max_{\pi} \widehat{J}^{\gamma}_{\mathcal{D}}(\pi) - \alpha \mathbb{E}_{(s_t,a_t)\sim\mathcal{D}}\left[w(s,a)\mathcal{C}(s_t,a_t)\right]$$

$$\Updownarrow$$

$$\max_{\pi} \widehat{J}^{\gamma}_{\mathcal{D}}(\pi) - \alpha \mathbb{E}_{(s_t,a_t)\sim\mathcal{D}_w}\left[\mathcal{C}(s_t,a_t)\right].$$

Difficulty:
The key challenge is determining the weights w since the state-action distribution of the better policy πDw is initially unknown

↓

off-policy evaluation techniques(optidice)

to determine if a weighting function corresponds to a high-return policy

Methodology

- OptiDICE

provides a framework for improving policy optimization using offline datasets in reinforcement learning

$$\max_{d \geq 0} \min_{\nu} \mathbb{E}_{(s,a)\sim d}[R(s,a)] - \alpha D_f(d||d^D) \qquad (5)$$

$$+ \sum_s \nu(s)\Big((1-\gamma)p_0(s) + \gamma(\mathcal{T}_* d)(s) - (\mathcal{B}_* d)(s)\Big),$$

$$= (1-\gamma)\mathbb{E}_{s\sim p_0}[\nu(s)] + \mathbb{E}_{(s,a)\sim d^D}\left[-\alpha f\big(w(s,a)\big)\right]$$

$$+ \mathbb{E}_{(s,a)\sim d^D}\left[w(s,a)\big(e_\nu(s,a)\big)\right] =: L(w,\nu). \qquad (7)$$

$$D_f(d^\pi||d^D) := \mathbb{E}_{(s,a)\sim d^D}\left[f\left(\frac{d^\pi(s,a)}{d^D(s,a)}\right)\right]$$

F = xlogx时为KL散度

DICE简单来说可以理解为初始状态转移分布和状态转移分布，在结合重要度采样所构造的优化目标

$$(\mathcal{B}_* d)(s) := \sum_{\bar{a}} d(s,\bar{a})$$

$$(\mathcal{T}_* d)(s) := \sum_{\bar{s},\bar{a}} T(s|\bar{s},\bar{a})d(\bar{s},\bar{a})$$

Methodology： Density ratio Weighting (DW)

Our goal is to discover a weighting function w that can emulate drawing state-action samples from a better dataset Dw that is collected by an alternative behavior policy πDw with higher return than the behavior policy πD that collected the original dataset D

- **Optimizing the Weightings**

$$J^{\gamma}(\pi_{\mathcal{D}_w}) \approx \mathbb{E}_{(s,a)\sim\mathcal{D}_w}\left[r(s,a)\right] = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[w(s,a)r(s,a)\right].$$

when the the dataset Dw represents a stationary state-action distribution:

$$\mathcal{D}_w(s') = (1-\gamma)\rho_0(s') + \gamma\sum_{s,a}\mathcal{T}(s'|s,a)\mathcal{D}_w(s,a)\;\forall s'\in\mathcal{S}, \quad \mathcal{D}_w(s') := \sum_{a'\in\mathcal{A}}\mathcal{D}_w(s',a') \quad (7)$$

$$\downarrow$$

$$\mathcal{D}_w(s') = \sum_{s,a}\mathcal{T}(s'|s,a)\mathcal{D}_w(s,a)\;\forall s'\in\mathcal{S}.$$

Methodology： Density ratio Weighting (DW)

- 将数据集Dw写成w(s,a)的形式，则有下式的约束成立：

$$\mathcal{D}(s')w(s') = \sum_{s,a} \mathcal{T}(s'|s,a)w(s,a)\ \forall s' \in \mathcal{S}, \qquad w(s) := \sum_{a\in\mathcal{A}} \frac{\mathcal{D}_w(s,a)}{\mathcal{D}(s,a)} \tag{9}$$

$$J^{\gamma}(\pi_{\mathcal{D}_w}) \approx \mathbb{E}_{(s,a)\sim\mathcal{D}_w}\left[r(s,a)\right] = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[w(s,a)r(s,a)\right]. \tag{6}$$

将（6）式的无折扣版本和（9）式的约束合并，则有以下带约束的优化目标：

$$\max_{w} J(\pi_{\mathcal{D}_w}) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[w(s,a)r(s,a)\right] \tag{10}$$

$$\text{subject to } \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[w(s') - w(s,a) \mid s'\right] = 0 \ \ \forall s' \in \mathcal{S}.$$

Implementation

进行参数化，并给出状态边际分布的计算方式

$$w(s,a) \doteq \frac{\mathcal{D}_w(s,a)}{\mathcal{D}(s,a)} = \frac{\mathcal{D}_w(s)\pi_{\mathcal{D}_w}(a|s)}{\mathcal{D}(s)\pi_{\mathcal{D}}(a|s)} = \frac{\mathcal{D}_w(s)}{\mathcal{D}(s)} \times \frac{\pi_{\mathcal{D}_w}(a|s)}{\pi_{\mathcal{D}}(a|s)}.$$

$$w_{\phi,\psi}(s,a) = \exp\phi(s)\exp\psi(s,a), \qquad w_\phi(s) = \exp\phi(s) \tag{12}$$

φ and ψ are neural networks
(two-layer Multilayer Perceptron (MLP) with 256 neurons and ReLU activation in each layer.)

$$\max_{\phi,\psi} \mathbb{E}_{(s,a,s')\sim\mathcal{D}} \left[\underbrace{w_{\phi,\psi}(s,a)r(s,a)}_{\text{Return}} - \lambda_F \underbrace{(w_\phi(s') - w_{\phi,\psi}(s,a))^2}_{\text{Bellman flow conservation penalty}} \right] - \lambda_K \underbrace{D_{KL}(\mathcal{D}_w||\mathcal{D})}_{\text{KL regularization}}, \tag{13}$$

最终优化目标

加快收敛 提高稳定性

数据集覆盖范围有限

Methodology：DW & IQL

Algorithm 1 Implicit Q-learning

Initialize parameters ψ, θ, $\hat{\theta}$, ϕ.

TD learning (IQL):

for each gradient step **do**

$\psi \leftarrow \psi - \lambda_V \nabla_\psi L_V(\psi)$ → $L_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}[L_2^\tau(Q_{\hat{\theta}}(s,a) - V_\psi(s))]$.

$\theta \leftarrow \theta - \lambda_Q \nabla_\theta L_Q(\theta)$ → $L_Q(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}[(r(s,a) + \gamma V_\psi(s') - Q_\theta(s,a))^2]$

$\hat{\theta} \leftarrow (1-\alpha)\hat{\theta} + \alpha\theta$

end for

Policy extraction (AWR):

for each gradient step **do**

$\phi \leftarrow \phi - \lambda_\pi \nabla_\phi L_\pi(\phi)$ → $L_\pi(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}[\exp(\beta(Q_{\hat{\theta}}(s,a) - V_\psi(s))) \log \pi_\phi(a|s)]$.

end for

Methodology： DW & IQL

IQL. We reweight state-value function (V), state-action value function (i.e,. Q-function, Q), and policy (π) in IQL as follows:

$$\min_V \mathbb{E}_{(s,a)\sim\mathcal{D}} \left[w_\phi(s) L_2^\tau(Q(s,a) - V(s))\right] \tag{20}$$

$$\min_Q \mathbb{E}_{(s,a,s')\sim\mathcal{D}} \left[w_{\phi,\psi}(s,a) \left(r(s,a) + \gamma V(s') - V(s)\right)^2\right] \tag{21}$$

$$\max_\pi \mathbb{E}_{(s,a)\sim\mathcal{D}} \left[w_{\phi,\psi}(s,a) \exp\left(\beta \left(Q(s,a) - V(s)\right)\right) \log \pi(a|s)\right], \tag{22}$$

where L_2^τ denotes the upper expectile loss [18] and β denotes the temperature parameter for IQL. Our implementation is adapted from the official implementation[4] for implicit Q-learning (IQL) [18].

Algorithm

Algorithm 1 Density-ratio weighting with generic offline RL algorithms (details in Appendix A.3)

Input: Dataset $\mathcal{D}$
Initialize policy π and weighting function $w_{\phi,\psi}$
while not converged **do**
    Sample a batch $\mathcal{B}$ of tuples of states, actions, rewards, and next states (s, a, r, s') from $\mathcal{D}$
    Update $w_{\phi,\psi}$ with batch $\mathcal{B}$ using Equation 13
    Update policy π and value function with an offline RL training objective with weights $w_{\phi,\psi}(s, a)$ and $\mathcal{B}$ (e.g., [22, 7, 18])
end while

Summary

目标：本文希望优化目标不是聚焦于整个数据集，而是仅聚焦于数据集中好的部分

实施：

1. 通过重要度采样将原数据集D和未知数据集Dw联系起来
2. 通过off-policy evaluation方法，将数据集分布Dw和策略πw的累积回报联系起来
3. 通过对w进行参数化，并加入KL散度，最终形成优化目标。
4. 与现有离线强化学习算法结合。

Experiments

- **Two stages**

  (i) Trajectories with similar initial states
  (ii) Trajectories with diverse initial states

- **interquartile mean（IQM）**

  discards the bottom and top 25% of the runs and calculates the mean score of the remaining 50% runs (=N M/2c for N runs each on M tasks)

  表示统计资料中各变量分散情形，但四分差更多为一种稳健统计

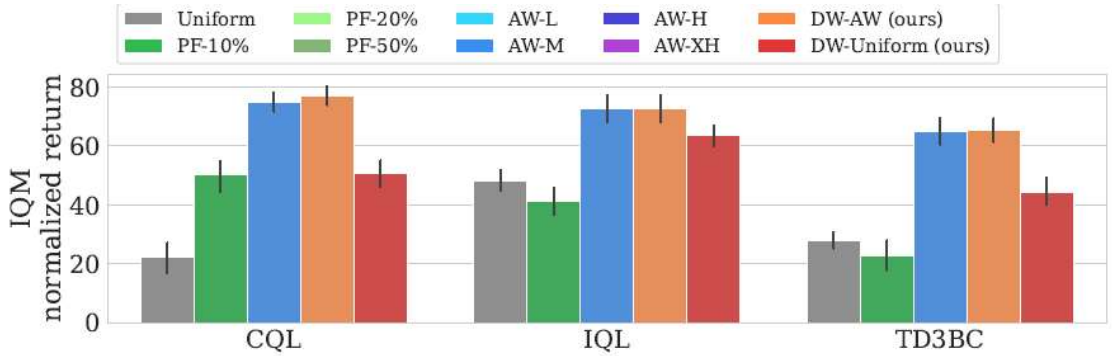
- **Baseline**

  advantage-weighting (AW) (provide a better initial sampling distribution to train the weightingfunction in DW)
  percentage-filtering (PF)

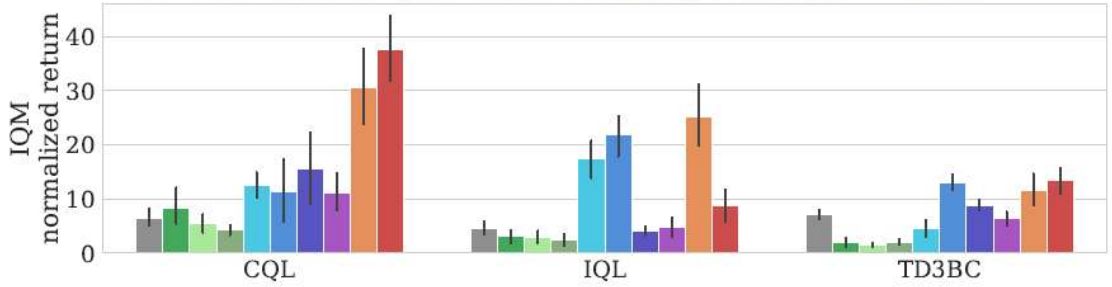
$$\mathcal{P}_{\mathrm{AW}}(s_t^i, a_t^i) \propto \exp((G(\tau_i) - V_0(s_0^i))/\eta) \qquad \text{(Advantage-weighting)} \qquad (14)$$

$$\mathcal{P}_{\mathrm{PF}}(s_t^i, a_t^i) \propto \mathbb{1}\left[G(\tau_i) \geq G_{K\%}\right] \qquad \text{(Percentage-filtering)}, \qquad (15)$$

Experiments



(a) Results on imbalanced datasets of trajectories with similar initial states (Section 5.1).



(b) Results on smaller version of datasets used in Figure 3a.

Figure 3: **(a)** Our methods, DW-AW and DW-Uniform, achieve higher return than Uniform, indicating that DW can enhance the performance of offline RL algorithms on imbalanced datasets. Note that our methods in IQL, although not surpassing AW and PF-10% in performance, ours can be applied to offline RL dataset that are not curated with trajectories. **(b)** Our methods outperform Uniform in CQL, IQL, and TD3BC, indicating no significant overfitting in smaller datasets. DW-AW demonstrates superior returns compared to AW and PF, particularly in CQL, indicating our method effectively leverages limited data. IQL shows limited gains likely due to its difficulties in utilizing data from the rest of low-return trajectories in the dataset (see Section 5.2).

Experiments

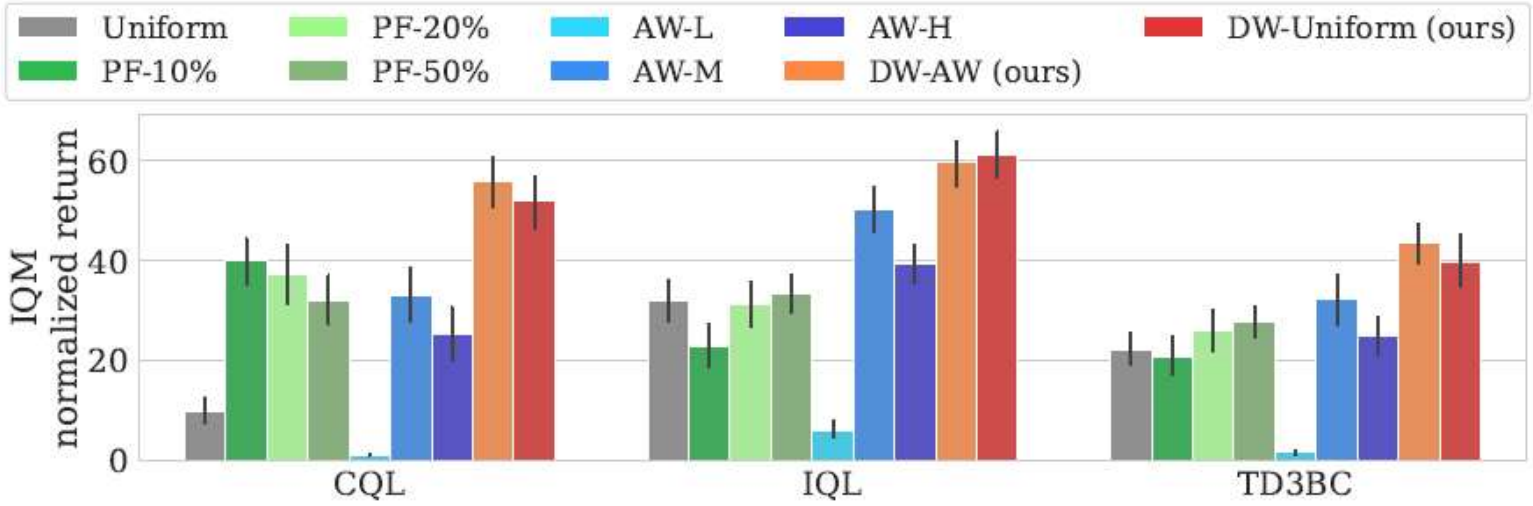


Figure 4: Results on imbalanced datasets with trajectories starting from diverse initial states (Section 5.3). Compared to Figure 3a, the performance of uniform sampling and AW decrease, showing that diverse initial states exacerbate the issue of imbalance. Our methods, DW-AW and DW-Uniform, achieve higher return than all the baselines, which suggests DW is advantageous in broader types of imbalanced datasets.