A Dataset Perspective on Offline Reinforcement Learning

Kajetan Schweighofer[#,1], Andreas Radler[#,1], Marius-Constantin Dinu[#,1,3],
Markus Hofmarcher[1], Vihang Patil[1],
Angela Bitto-Nemling[1,2], Hamid Eghbal-zadeh[1], Sepp Hochreiter[1,2]

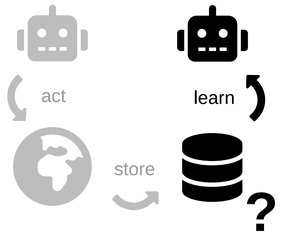
1) ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning, Johannes Kepler University Linz, Austria
2) Institute of Advanced Research in Artificial Intelligence (IARAI)
3) Dynatrace Research
# Authors contributed equally

The application of Reinforcement Learning (RL) in real world environments can be expensive or risky due to sub-optimal policies during training. In Offline RL, this problem is avoided since interactions with an environment are prohibited. Policies are learned from a given dataset, which solely determines their performance. Despite this fact, how dataset characteristics influence Offline RL algorithms is still hardly investigated. The dataset characteristics are determined by the behavioral policy that samples this dataset. Therefore, we define characteristics of behavioral policies as exploratory for yielding high expected information in their interaction with the Markov Decision Process (MDP) and as exploitative for having high expected return. We implement two corresponding empirical measures for the datasets sampled by the behavioral policy in deterministic MDPs. The first empirical measure SACo is defined by the normalized unique state-action pairs and captures exploration. The second empirical measure TQ is defined by the normalized average trajectory return and captures exploitation. Empirical evaluations show the effectiveness of TQ and SACo. In large-scale experiments using our proposed measures, we show that the unconstrained off-policy Deep Q-Network family requires datasets with high SACo to find a good policy. Furthermore, experiments show that policy constraint algorithms perform well on datasets with high TQ and SACo. Finally, the experiments show, that purely dataset-constrained Behavioral Cloning performs competitively to the best Offline RL algorithms for datasets with high TQ.

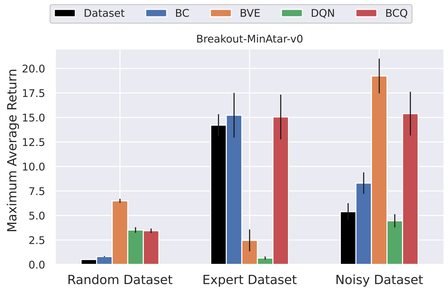
Main contributions

Our main contributions are aligned along the question, how algorithms in Offline RL are influenced by the characteristics of the dataset in finding a good policy:

- We derive theoretical measures that capture exploration and exploitation under a policy
- We provide an effective method to characterize datasets through the empirical measures TQ and SACo
- We conduct an extensive empirical evaluation on how dataset characteristics influence algorithms in Offline RL



Paper (arxiv) **Code**



Comparison of multiple algorithms in an Offline RL setting, using different datasets sampled from the same environment.

Characterizing RL Datasets

Exploitation: Performance under behavioral policy (expected return)

$$g_\pi = \mathbb{E}_\pi\left[\sum_{t=0}^{T}\gamma^t R_{t+1}\right] \qquad \bar{g}(\mathcal{D}) = \frac{1}{B}\sum_{b=0}^{B}\sum_{t=0}^{T_b}\gamma^t r_{b,t}$$

$$TQ(\mathcal{D}) := \frac{\bar{g}(\mathcal{D}) - \bar{g}(\mathcal{D}_{\min})}{\bar{g}(\mathcal{D}_{\text{expert}}) - \bar{g}(\mathcal{D}_{\min})}$$

Exploration: Entropy of transition probability under behavioral policy (transition-entropy)

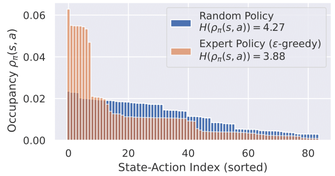
$$\begin{aligned} &H(p_\pi(s,a,r,s')) \\ &= -\sum_{s,a,r,s'} p_\pi(s,a,r,s')\log(p_\pi(s,a,r,s')) \\ &= \sum_{s,a}\rho_\pi(s,a)\, H(p(r,s' \mid s,a)) + H(\rho_\pi(s,a)) \end{aligned}$$

Left term vanishes for deterministic MDPs and the transition-entropy simplifies to occupancy-entropy.

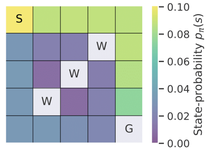
Unique state-action pairs upper bound perplexity estimator

$$e^{\hat{H}(\mathcal{D})} \leq u_{s,a}(\mathcal{D})$$

$$SACo(\mathcal{D}) := \frac{u_{s,a}(\mathcal{D})}{u_{s,a}(\mathcal{D}_{\text{ref}})}$$



Random Policy

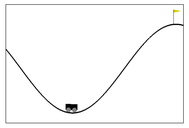


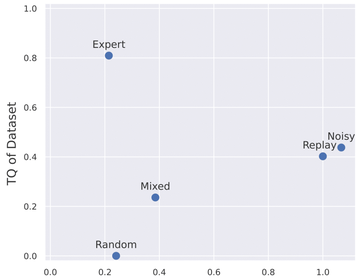
Expert Policy (ε-greedy)

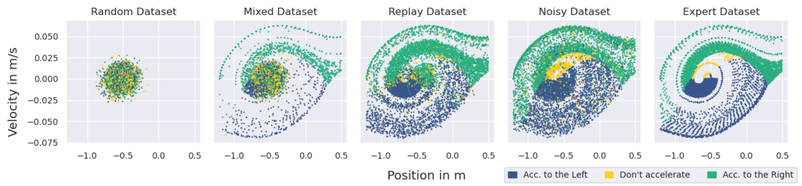
Dataset generation

Five different schemes to generate datasets are used, mostly based on the online policy used for the normalization of dataset measures:

- Random
- Expert
- Mixed
- Replay
- Noisy







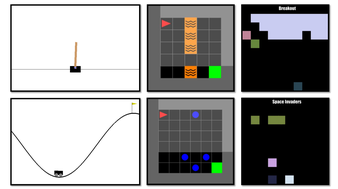
Sample datasets for each scheme, where the state is represented by the position and the action by the color of single points.

Algorithms and Environments

Baseline algorithms
- Behavioral Cloning (BC)
- Behavior Value Estimation (BVE)
- Monte-Carlo Estimation (MCE)

Dataset-constrained off-policy algorithms
- Batch-Constrained Q-learning (BCQ)
- Conservative Q-learning (CQL)
- Critic Regularized Regression (CRR)

Unconstrained off-policy algorithms
- Deep Q-Network (DQN)
- Quantile Regression DQN (QR-DQN)
- Random Ensemble Mixture (REM)



Experimental Results

Baseline algorithms
- BC best on datasets with high TQ
- BVE & MCE tend to work better for datasets with moderate SACo

Unconstrained off-policy algorithms
- Best on datasets with high SACo

Dataset-constrained off-policy algorithms
- Good on datasets with high SACo or high TQ or moderate TQ & SACo