NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

SCSE21075- Offline Reinforcement Learning

Presented by Chen Kang Ming

Supervised by Asst Prof Jun Zhao

OFFLINE REINFORCEMENT LEARNING

REFERENCES

Kumar, A. (2019, December 5). Data-Driven Deep Reinforcement Learning. The Berkeley Artificial Intelligence Research Blog. https://bair.berkeley.edu/blog/2019/12/05/bear/

Agarwal, R., Schuurmans, D., & Norouzi, M. (2019). An Optimistic Perspective on Offline Reinforcement Learning. ArXiv. https://doi.org/10.48550/ARXIV.1907.04543

Fu, J., Kumar, A., Nachum, O., Tucker, G., & Levine, S. (2020). D4RL: Datasets for Deep Data-Driven Reinforcement Learning (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2004.07219

Presented by: Chen Kang Ming

INTRODUCTION

Offline reinforcement learning, also known as off-policy learning or batch RL, is a variation of conventional reinforcement learning (RL) methods that aims to train agents through conventional RL means but constricted **with a finite data set** (Figure 1).

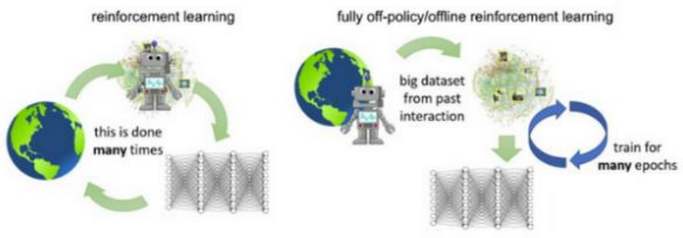
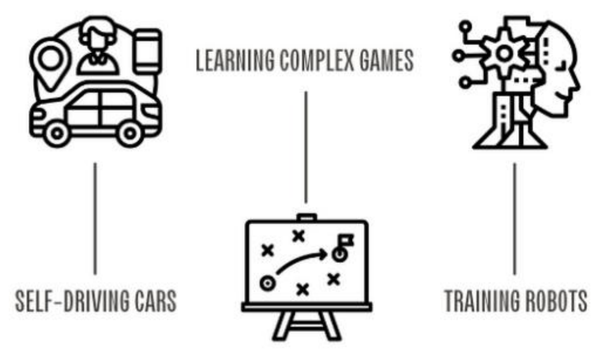


Figure 1: RL (left) vs Offline RL (right) (Kumar, 2019)

USE CASES

Offline RL has huge significance in cases where data collection would be **tedious or potentially unsafe** (autonomous driving and robots, or learning how to play strategy games for example).



PROBLEMS REGARDING OFFLINE RL

Offline RL presents many challenges:

Out-of-Distribution (OOD) actions:
- positive Q-value errors contribute to **bootstrap error accumulation** and policy skewing as the values are continuously backed up (Figure 2).
- Error due to nature of using dynamic programming in solving offline RL, where bad and mediocre policies have to be collected on top of optimal ones.

Distribution Mismatch:
- policy that is in process of learning takes a different method from the data collection policy
- Raises **ambiguity** on reward-giving.

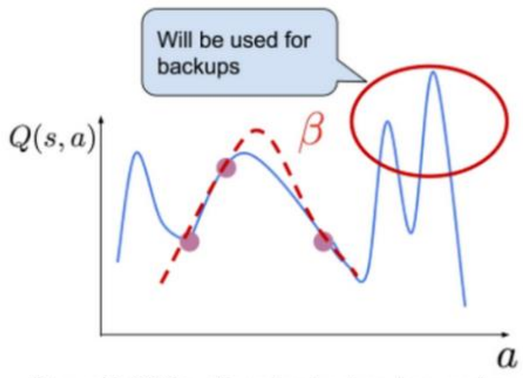


Figure 2: OOD actions backed up, increasing error (Kumar, 2019)

TYPES OF OFFLINE RL ALGORITHMS

2 types will be discussed:

Bootstrapping Error Accumulation Reduction (BEAR) (Figure 3)

Random Ensemble Mixture (REM) (Figure 4)

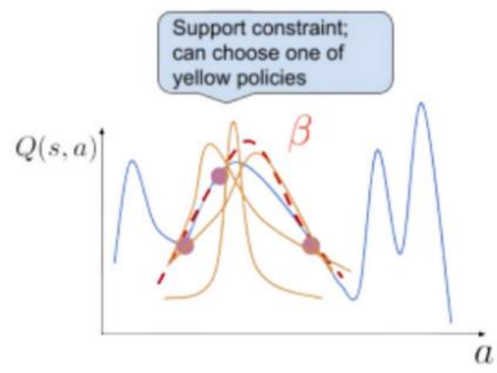


Figure 3: Illustration of BEAR (Kumar, 2019)

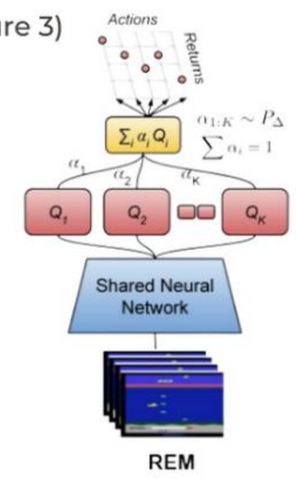


Figure 4: Illustration of REM (Agarwal et al., 2019)

ANALYSIS & ELABORATION

BEAR:

Utilizes sampled **maximum mean discrepancy** to constraint policies to a select few within margin.

Outperforms other algorithms like **Behavioural Cloning and Naïve RL,** in various types of policies (Figure 5).

$$\pi_\phi := \max_{\pi \in \Delta_{|S|}} \mathbb{E}_{s\sim\mathcal{D}} \mathbb{E}_{a\sim\pi(\cdot|s)} [Q_\theta(s,a)] \quad \text{s.t.} \quad \mathbb{E}_{s\sim\mathcal{D}}[\text{MMD}(\beta(\cdot|s), \pi(\cdot|s))] \leq \varepsilon$$

Figure 5: BEAR against other forms of RL policies (Kumar, 2019)

REM:

Based on **Q-Learning:** using combination of **multiple Q-Learning estimates** to give a **new** Q-Learning estimate to plan states/actions.

Outperforms other offline RL methods like **QR-DQN** on tested games for the Atari 2600 (Figure 6).

REM Loss Function:

$$\mathcal{L}(\theta) = \mathbb{E}_{s,a,r,s'\sim\mathcal{D}} \left[\mathbb{E}_{\alpha\sim P_\Delta} \left[\ell_\lambda \left(\Delta_\theta^\alpha(s,a,r,s') \right) \right] \right],$$
$$\Delta_\theta^\alpha = \sum_k \alpha_k Q_\theta^k(s,a) - r - \gamma \max_{a'} \sum_k \alpha_k Q_{\theta'}^k(s',a')$$

Figure 6 REM (orange) against offline QR-DQN (green) in games of Pong (Agarwal, Schuurmans, Norouzi, 2019)

CURRENT AND FUTURE WORK (D4RL)

D4RL is a set of tests and data geared towards offline learning. In particular, two of them will be discussed: *FrankaKitchen* and *CARLA* (Figure 7).

FrankaKitchen:
Aims to control a 9-DoF (degree of freedom) robot that has access to basic kitchen appliances like a microwave and an oven among others to reach a desired **goal configuration.**

In this task, **BEAR was able to perform better** than cREM (continuous REM).

CARLA:
Agents aim to control a throttle, steering and brake pedal, following lanes and navigating in a small town while receiving images as observations

Neither algorithm was able to solve the task, which struggled in the presence of **undirected data.**

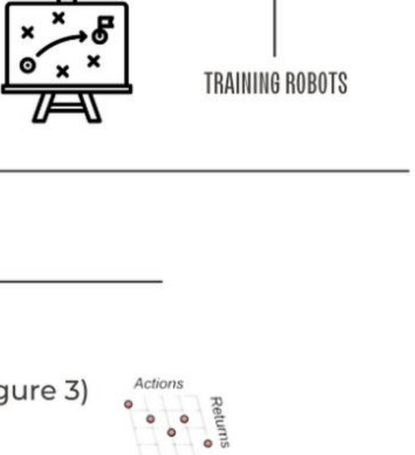


Figure 7: *FrankaKitchen* (top) and *CARLA* (bottom) (Fu et al., 2020)

CONCLUSION

Offline reinforcement learning will become one of the leading topics in the next decade to come, with a paradigm shift likely to occur in the sectors where traditional methods prove difficult or risky. In future, offline RL could have the potential to extend to sectors of NLP (natural language processing) and healthcare, which could ultimately benefit our society even more.