Off-Policy Deep Reinforcement Learning without Exploration

Yang Chang
Department of Computer Science
National Yang Ming Chiao Tung University
nyms0390.cs10@nycu.edu.tw

May 2022

1 Introduction

This paper focuses on batch reinforcement learning, which means training an agent on a fixed dataset without further exploration. The authors first show that standard off-policy methods like DQN and DDPG are not able to learn well without data correlated with current policy. By conducting a preliminary experiment, they show that the off-policy agent performs much worse than the behavioral agent trained by the DDPG algorithm due to the extrapolation errors. To deal with the fixed dataset setting, the authors introduce an algorithm called "Batch-Constrained Reinforcement Learning". In the proposed method, they try to force the policy to select actions which are closer to those in the pre-collected batch. By doing so, the agent can somehow avoid large extrapolation errors.

2 Problem Formulation

In this paper, they analyze batch-constrained policies in a finite Markov decision process (MDP) setting $(\mathcal{S}, \mathcal{A}, p_M, r, \gamma)$, with state space S, action space A, transition dynamics $p_M(s'|s,a)$, reward function r and discount factor γ. For a given policy π, the value function can be calculated using the Bellman operator $\mathcal{T}^\pi$:

$$\mathcal{T}^\pi Q(s,a) = \mathbb{E}_{s'}[r + \gamma Q(s', \pi(s'))] \tag{1}$$

2.1 Extrapolation Error

Extrapolation error is introduced by the mismatch between the dataset and the true state-action visitation of the current policy. In other words, the policy may select an action a' and enter state s', while the state-action pair (s', a') is not

contained in the collected dataset. The cause of this phenomenon can be listed below.

Absent Data

The pre-collected dataset might not have enough exploration in the real state distribution. If any state-action pair is not available in the batch, the estimated Q-value may be extremely bad.

Model Bias

For a stochastic MDP without infinite state-action visitation, the result Q-estimator will be biased.

$$\mathcal{T}^{\pi}Q(s,a) \approx \mathbb{E}_{s'}[r + \gamma Q(s', \pi(s'))] \tag{2}$$

Training Mismatch

Even if we have sufficient visitation, the distribution of the data collected by the current policy could still be different from those in the batch. The mismatch in training may also lead to a poor estimate of the value function.

2.2 Batch-Constrained Reinforcement Learning

For the proposed method, the policy should select actions with respect to three objectives:

1. Minimize the distance between the selected actions and the data in the batch.
2. Lead to states where familiar data can be observed.
3. Maximize the value function.

They found that by inducing a data distribution that is contained entirely within the batch, the extrapolation error can be eliminated entirely for deterministic MDPs, which fulfill objective 1. They denoted policies that satisfy this notion as batch-constrained.

3 Theoretical Analysis

The analysis is divided into two different parts. First, they show that the batch-constrained variant of Q-learning will also converge to an optimal policy. Then, they prove that for a deterministic MDP, batch-constrained Q-learning is guaranteed to match, or outperform, the behavioral policy when starting from any state contained in the batch.

Theorem 1. *Performing Q-learning by sampling from a batch $\mathcal{B}$ converges to the optimal value function under MDP $M_{\mathcal{B}}$*

The tabular extrapolation error ϵ_{MDP} can be computed by the following:

$$\epsilon_{\text{MDP}}(s,a) = Q^{\pi}(s,a) - Q^{\pi}_{\mathcal{B}}(s,a) \tag{3}$$

which can be further extracted as:

$$\begin{aligned}\epsilon_{\text{MDP}}(s,a) = \sum_{s'}&(p_M(s'|s,a) - p_{\mathcal{B}}(s'|s,a)) \\ &(r(s,a,s') + \gamma \sum_{a'} \pi(a'|s') Q^{\pi}_{\mathcal{B}}(s',a')) \\ &+ p_M(s'|s,a)\gamma \sum_{a'} \pi(a'|s') \epsilon_{\text{MDP}}(s',a')\end{aligned} \tag{4}$$

Proof.

$$\begin{aligned}\epsilon_{\text{MDP}}(s,a) &= Q^{\pi}(s,a) - Q^{\pi}_{\mathcal{B}}(s,a) \\ &= \sum_{s'} p_M(s'|s,a)\Big(r(s,a,s') + \gamma \sum_{a'} \pi(a'|s') Q^{\pi}(s',a')\Big) \\ &- \Big(\sum_{s'} p_{\mathcal{B}}(s'|s,a)\Big(r(s,a,s') + \gamma \sum_{a'} \pi(a'|s') Q_{\mathcal{B}^{\pi}}(s',a')\Big)\Big) \\ &= \sum_{s'} (p_M(s'|s,a) - p_{\mathcal{B}}(s'|s,a)) r(s,a,s') \\ &+ p_M(s'|s,a)\gamma \sum_{a'} \pi(a'|s')(Q^{\pi}_{\mathcal{B}}(s',a') + \epsilon_{\text{MDP}}(s',a')) \\ &- p_{\mathcal{B}}(s'|s,a)\gamma \sum_{a'} \pi(a'|s') Q^{\pi}_{\mathcal{B}}(s',a') \\ &= \sum_{s'} (p_M(s'|s,a) - p_{\mathcal{B}}(s'|s,a)) r(s,a,s') \\ &+ p_M(s'|s,a)\gamma \sum_{a'} \pi(a'|s')(Q^{\pi}_{\mathcal{B}}(s',a') + \epsilon_{\text{MDP}}(s',a')) \\ &+ p_M(s'|s,a)\gamma \sum_{a'} \pi(a'|s')(\epsilon_{\text{MDP}}(s',a') - \epsilon_{\text{MDP}}(s',a')) \\ &- p_{\mathcal{B}}(s'|s,a)\gamma \sum_{a'} \pi(a'|s') Q^{\pi}_{\mathcal{B}}(s',a') \\ &= \text{equation(4)}\end{aligned} \tag{5}$$

□

This means that the extrapolation error is a function of divergence in the transition distributions, weighted by value, along with the error in the succeeding states. If the transition distribution of our carefully chosen policy is similar

to the batch transition distribution, we can minimize the error. For simplicity in notation, they denote

$$\epsilon^{\pi}_{\text{MDP}} = \sum_{s} \mu_{\pi}(s) \sum_{a} \pi(a|s) |\epsilon_{\text{MDP}}(s,a)| \tag{6}$$

To evaluate a policy π exactly at relevant state-action pairs, only $\epsilon^{\pi}_{\text{MDP}} = 0$ is required.

Lemma 1. *For all reward functions, $\epsilon^{\pi}_{\text{MDP}} = 0$ if and only if $p_{\mathcal{B}}(s'|s,a) = p_M(s'|s,a)$ for all $s' \in S$ and (s,a) such that $\mu_{\pi}(s) > 0$ and $\pi(a|s) > 0$*

Lemma 1 states that if $M_{\mathcal{B}}$ and M exhibit the same transition probabilities in the regions of relevance, the policy can be accurately evaluated. In other words, a policy that only traverses transitions contained in the batch can eliminate the extrapolation error entirely.
Let us proceed to prove the existence of such a batch-constrained policy. More formally, the authors define a policy $\pi \in \Pi_{\mathcal{B}}$ as *batch-constrained* if for all (s,a) where $\mu_{\pi}(s) > 0$ and $\pi(a|s) > 0$ then $(s,a) \in \mathcal{B}$. Furthermore, they define a batch $\mathcal{B}$ as *coherent* if for all $(s,a) \in \mathcal{B}$ then the following $s' \in \mathcal{B}$ unless s' is a terminal state.

Theorem 2. *For a deterministic MDP and all reward functions, $\epsilon^{\pi}_{\text{MDP}} = 0$ if and only if the policy π is batch-constrained. Furthermore, if $\mathcal{B}$ is coherent, then such a policy must exist if the start state $s_0 \in \mathcal{B}$*

Proof. By Lemma 1, we know that for the policy π, if $(s,a) \in \mathcal{B}$, then $p_{\mathcal{B}}(s'|s,a) = p_M(s'|s,a)$ for all $s' \in S$. We can construct the batch-constrained policy by selecting a in the state $s \in \mathcal{B}$, such that $(s,a) \in \mathcal{B}$. Since the MDP is deterministic and the batch is coherent, when starting from s_0, we must be able to follow at least one trajectory until termination. □

The next step is to combine batch-constrained policy with Q-learning to form batch-constrained Q-learning (BCQL). BCQL can be updated just like the standard form of Q-learning:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a' \text{s.t.}(s',a') \in \mathcal{B}} Q(s',a')) \tag{7}$$

Theorem 3. *Given the Robbins-Monro stochastic convergence conditions on the learning rate α and standard environmental sampling requirements, BCQL converges to the optimal value function Q^* .*

Theorem 4. *Given a deterministic MDP and coherent batch $\mathcal{B}$, along with the Robbins-Monro stochastic convergence conditions on the learning rate α and standard sampling requirements on the batch $\mathcal{B}$, BCQL converges to $Q^{\pi}_{\mathcal{B}}(s,a)$ where $\pi^*(s) = \arg\max_{a s.t.(s,a) \in \mathcal{B}} Q^{\pi}_{\mathcal{B}}(s,a)$ is the optimal batch-constrained policy.*

Theorems 3 and 4 show that BCQL is guaranteed to outperform any behavioral policy when starting from any state contained in the batch.

4 Conclusion

In this work, they point out the problem of using standard off-policy learning method with batch setting. They introduced the idea of batch-constrained policy to deal with the problem. Although they theoretically prove that the batch-constrained policy is guaranteed to outperform the behavioral policy, something we can argue about is that the behavioral policy might not be optimal. If we use a suboptimal policy to collect our batch data, outperforming such policy means nothing. The latest result of [2]model-based offline policy optimization (Tianhe Yu et al., 2020) provides another way to trade off between the risk of uncertainty and the value of exploration.

References

[1] Fujimoto, S., Meger, D., Precup, D. Off-Policy Deep Reinforcement Learning without Exploration. PMLR 97:2052-2062, 2019.

[2] Tianhe, Y., Garrett, T., Lantao, Y., Stefano, E., James, Z., Sergey, L., Chelsea, F., Tengyu, M. MOPO: Model-based Offline Policy Optimization. arXiv:2005.13239, 2020.