Theory Project of Offline Reinforcement Learning with Implicit Q-Learning

Yu-Wei Yang
Institute of Data Science and Engineering
National Yang Ming Chiao Tung University
ray17606.cs10@nycu.edu.tw

1 Introduction

Offline reinforcement learning(RL) is the task learns from the fixed dataset that collected from the environment. However, offline RL should consider two related but conflicted issues: on one hand, it needs to improving the learned policy over the behavior policy that collected the dataset, on the other hand, value errors may occur if the agent encounter out-of-distribution samples with the learned policy too deviate from the behavior policy. To have a good trade-off between the policy improvement and distribution shifts is a quite challenging task, and many prior works tried to solve this issue by constraining distance between the learned policy and behavior policy(Kumar et al. [2019], Wu et al. [2019], Fujimoto et al. [2018], Wang et al. [2020], Fujimoto and Gu [2021]), or lowering the values of out-of-distribution actions(Kumar et al. [2020], Kidambi et al. [2020], Kostrikov et al. [2021]).

In this work, however, the authors have a straightforward thought: is it possible to train the policy only by in-distribution data, and not querying the value of unseen actions? Therefore, authors propose a method called **Implicit Q-learning**(IQL), and the primary idea of this work is to approximate the value of the upper expectile of the distribution with respect to the actions for each state in the dataset distribution. They construct the value function with the expectile regression, and use it to get the Q-function by modifying the loss function with SARSA-style temporal difference(TD), then extract the corresponding policy by using advantage weighted regression.

The first contribution of this work is to avoid estimating the unseen actions values which may cause distribution shift, by making a little change to the loss function with the SARSA-style TD backup. Also, compared to other prior works, IQL not only has better performances on the tasks of D4RL, but the whole process does not involve explicit constraining policy or out-of-distribution actions values regularization as they did.

2 Problem Formulation

In this work, a Markov decision process (MDP) is given, which is defined as $(\mathcal{S}, \mathcal{A}, p_0(s), p(s'|s,a), r(s,a), \gamma)$, where $\mathcal{S}, \mathcal{A}$ represent state and action spaces, $p_0(s)$ represents the distribution of initial states, $p(s'|s,a)$ is the environment dynamics, $r(s,a)$ is the reward function, and γ represents the discount factor. The agent uses policy $\pi(a|s)$ to interact with the MDP. The optimal policy can be defined as:

$$\pi^* = \arg\max_{\theta} \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 \sim p_0(\cdot), a_t \sim \pi(\cdot|s_t), s_{t+1} \sim p(\cdot|s_t, a_t)]$$

Unlike online RL methods, offline RL only uses data collected by the behavior policy π_β to do training without any further environment interaction. Many offline RL works try to minimizes the TD error by approximate dynamic programming methods, according to the following loss function:

$$L_{TD}(\theta) = \mathbb{E}_{(s,a,s') \sim D}[(r(s,a) + \gamma \max_{a'} Q_{\hat{\theta}}(s', a') - Q_\theta(s,a))^2] \tag{1}$$

where D is the training dataset, $Q_\theta(s,a)$ is the parameterized Q-function, $Q_{\hat{\theta}}(s,a)$ is the target network, and the policy is defined as $\pi(s) = \arg\max_a Q_\theta(s,a)$. However, here is an issue that the loss function (1) needs to estimate the unseen action values, the policy may be overestimated since $Q_{\hat{\theta}}(s,a)$ may be wrong values. To avoid estimating out-of-distribution actions, the loss function could be modified with a SARSA-style as:

$$L(\theta) = \mathbb{E}_{(s,a,s',a')\sim D}[(r(s,a) + \gamma Q_{\hat{\theta}}(s',a') - Q_\theta(s,a))^2] \tag{2}$$

Many prior works use this objective function to learn the value of π_β and get the corresponded policy. This method can get good results in simple tasks, like MuJoCo locomotion tasks, but for complex tasks with multi-step dynamic programming, this method doesn't work well.

Therefore, in this work, authors focus on estimating the max Q-value over actions in the dataset, and they want to modify the equation (2) as the following equation:

$$L_{TD}(\theta) = \mathbb{E}_{(s,a,s')\sim D}[(r(s,a) + \gamma \max_{\substack{a'\in A,\\ s.t.\pi_\beta(a'|s')>0}} Q_{\hat{\theta}}(s',a') - Q_\theta(s,a))^2] \tag{3}$$

3 Algorithm

3.1 Expectile Regression

In order to turn the equation (2) into equation (3), authors use expectile regression(Spiegel et al. [2021]). The expectile of a random variable X is defined as:

$$\arg\min_{m_\tau} \mathbb{E}_{x\sim X}[L_2^\tau(x - m_\tau)], \quad L_2^\tau(u) = \begin{cases} (1-\tau)u^2 & \text{if } u \le 0 \\ \tau u^2 & \text{otherwise} \end{cases}$$

where $\tau \in (0,1)$. When $\tau > 0.5$, if m_τ is larger than x, this asymmetric function will decrease the importance of x, and will increase the importance of x if m_τ is smaller than x.

By adding the expectile regression into equation (2), it can have the upper expectile regression of TD targets, and the equation can be defined as:

$$L(\theta) = \mathbb{E}_{(s,a,s',a')\sim D}[L_2^\tau(r(s,a) + \gamma Q_{\hat{\theta}}(s',a') - Q_\theta(s,a))] \tag{4}$$

This objective equation can leads equation (2) to predict an approximation of maximum $r(s,a) + \gamma Q_{\hat{\theta}}(s',a')$ from actions in the dataset, as mentioned in equation (3).

3.2 Value function and Q-function

But there is still an issue that equation (4) involves the stochasticity of environment dynamics $p(\cdot|s,a)$, which may have a chance to make s' happened to be a good state by the state transition and have a nice target value. Such kind of target value may not reflect the values of the actions appropriately.

Thus, authors first take the value function updated by equation (5) that approximates an expectile only with respect to the state action distribution. Then, the Q-function can be updated by using loss equation (6) with the estimation of value function obtained in first step. These two separate steps can average over the stochasticity of s' environment dynamics and prevent the issue mentioned above.

$$L_V(\psi) = \mathbb{E}_{(s,a)\sim D}[L_2^\tau(Q_{\hat{\theta}}(s,a) - V_\psi(s))] \tag{5}$$

$$L_Q(\theta) = \mathbb{E}_{(s,a,s')\sim D}[(r(s,a) + \gamma V_\psi(s') - Q_\theta(s,a))^2] \tag{6}$$

3.3 Complete Algorithm

The optimal Q-function can be obtained by the modified TD learning above, and the next is to use advantaged weighted regression(Peters and Schaal [2007], Wang et al. [2018]) to extract the corresponding policy.

$$L_\pi(\phi) = \mathbb{E}_{(s,a)\sim D}[\exp(\beta(Q_{\hat{\theta}}(s,a) - V_\psi(s))) \log \pi_\phi(a|s)] \tag{7}$$

where $\beta \in [0,\infty)$. When β is large, this function try to perform maximum of the Q-function, and when it get small value, this function will attempts to perform behavioral-cloning.

The complete algorithm, has two parts, the first part uses equation(5) and equation (6) to do gradient updates to fit the value function and Q function, and the second part use equation (7) to train the policy by stochastic gradient decent. In the algorithm, both parts use the clipped double Q-learning taking a minimum of two Q-functions for V-function and policy updates.

4 Theoretical Analysis

After showing IQL can avoid querying unseen actions, in this section, the following analyses will prove that this method can reach optimal value function.

Lemma 1. Let X be a real-valued random variable with a bounded support and supremum of the support is x^*, then:

$$\lim_{\tau \to 1} m_\tau = x^*$$

Proof. By the defintion of expectile in 3.1 section, assume there are two expectiles of X, τ_1 and τ_2, and they share the same supremum x^*. For any $\tau_1 < \tau_2$, we can get $m_{\tau_1} \le m_{\tau_2}$ easily. By following the properties of bounded monotonically non-decreasing, the equation above can be proved.

Next,let $\mathbb{E}^{\tau}_{x \sim X}[x]$ be the τ^{th} expectile of the random variable X, if $\tau_1 < \tau_2$, and we can get $\mathbb{E}^{\tau_1}_{x \sim X}[x] \le \mathbb{E}^{\tau_2}_{x \sim X}[x]$. Then define the optimal solutions of equation (5), (6) as $V_\tau(s)$ and $Q_\tau(s,a)$ respectively, which can be defined as:

$$V_\tau(s) = \mathbb{E}^{\tau}_{a \sim \mu(\cdot|s)}[Q_\tau(s,a)]$$
$$Q_\tau(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)}[V_\tau(s')]$$

Lemma 2. For all state s, if $\tau_1 < \tau_2$, then:

$$V_{\tau_1}(s) \le V_{\tau_2}(s)$$

Proof.

$$\begin{aligned}
V_{\tau_1}(s) &= \mathbb{E}^{\tau_1}_{a \sim \mu(\cdot|s)}[Q_\tau(s,a)] \\
&= \mathbb{E}^{\tau_1}_{a \sim \mu(\cdot|s)}[r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} V_{\tau_1}(s')]] \\
&\le \mathbb{E}^{\tau_2}_{a \sim \mu(\cdot|s)}[r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)}[V_{\tau_1}(s')]] \\
&= \mathbb{E}^{\tau_2}_{a \sim \mu(\cdot|s)}[r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} \mathbb{E}^{\tau_1}_{a' \sim \mu(\cdot|s')}[r(s',a') + \gamma \mathbb{E}_{s'' \sim p(\cdot|s',a')}[V_{\tau_1}(s'')]]] \\
&\le \mathbb{E}^{\tau_2}_{a \sim \mu(\cdot|s)}[r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} \mathbb{E}^{\tau_2}_{a' \sim \mu(\cdot|s')}[r(s',a') + \gamma \mathbb{E}_{s'' \sim p(\cdot|s',a')}[V_{\tau_1}(s'')]]] \\
&= \mathbb{E}^{\tau_2}_{a \sim \mu(\cdot|s)}[r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} \mathbb{E}^{\tau_2}_{a' \sim \mu(\cdot|s')}[r(s',a') + \gamma \mathbb{E}_{s'' \sim p(\cdot|s',a')} \mathbb{E}^{\tau_1}_{a'' \sim \mu(\cdot|s'')}[r(s'',a'') + ...]]] \\
&\vdots \\
&\le V_{\tau_2}(s)
\end{aligned}$$

Corollary 2.1 For any τ and state s, $V_\tau(s)$ defined above, and $Q^*(s,a)$ defined as the optimal Q-function with dataset constraint, we can get:

$$V_\tau(s) \le \max_{\substack{a' \in A, \\ s.t. \pi_\beta(a|s) > 0}} Q^*(s,a)$$

Proof. $V_\tau(s)$ can be viewed as a weighted sum of $Q_\tau(s,a)$, and the function above can be proved by the fact that the convex combination is smaller than maximum.

Theorem 3.

$$\lim_{\tau \to 1} V_\tau(s) = \max_{\substack{a' \in A, \\ s.t. \pi_\beta(a|s) > 0}} Q^*(s,a)$$

Proof. It can be proved by combining Lemma 1 and 2 together.

By the lemmas and theorem that proved above, we can see that IQL achieves the optimal value function, and the larger $\tau < 1$ is, the IQL reach a better approximation of max value function.

5 Conclusion

In this work, I think the idea of avoiding querying of-out-distribution actions in offline RL is quite interesting, since it is a straightforward thought to prevent error caused by distribution shift. I believe that there were many works trying to use similar method to solve the problem, but the result may not be outperformed than other works in the end. Authors of this work use solid theories and clever methods to conquer the technical barriers, which is surprising.

In the theoretical analysis, authors show that as τ goes limit to 1, the optimal value function can be obtained, which implies that τ should be set as 0.9 or 0.99 in the experiment. However, in the code that provided by authors, the τ of MuJoCo and kitchen tasks are both 0.7, which are not close to 1 enough. Authors admitted on the paper review that the implementation has a gap with the theory and sometimes it is difficult to estimate larger expectiles of a distribution. But still, larger τ performs better result.

To sum up, I think this paper provide a promising method to offline RL, and maybe this contribution can combine other policy extraction methods in the future to create more impressive works.

References

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *CoRR*, abs/1906.00949, 2019. URL http://arxiv.org/abs/1906.00949.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *CoRR*, abs/1911.11361, 2019. URL http://arxiv.org/abs/1911.11361.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. *CoRR*, abs/1812.02900, 2018. URL http://arxiv.org/abs/1812.02900.

Ziyu Wang, Alexander Novikov, Konrad Zolna, Jost Tobias Springenberg, Scott E. Reed, Bobak Shahriari, Noah Y. Siegel, Josh Merel, Çaglar Gülçehre, Nicolas Heess, and Nando de Freitas. Critic regularized regression. *CoRR*, abs/2006.15134, 2020. URL https://arxiv.org/abs/2006.15134.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *CoRR*, abs/2106.06860, 2021. URL https://arxiv.org/abs/2106.06860.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *CoRR*, abs/2006.04779, 2020. URL https://arxiv.org/abs/2006.04779.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel : Model-based offline reinforcement learning. *CoRR*, abs/2005.05951, 2020. URL https://arxiv.org/abs/2005.05951.

Ilya Kostrikov, Jonathan Tompson, Rob Fergus, and Ofir Nachum. Offline reinforcement learning with fisher divergence critic regularization. *CoRR*, abs/2103.08050, 2021. URL https://arxiv.org/abs/2103.08050.

Elmar Spiegel, Thomas Kneib, Petra von Gablenz, and Fabian Otto-Sobotka. Generalized expectile regression with flexible response function. *Biometrical Journal*, 63(5):1028–1051, 2021. doi: https://doi.org/10.1002/bimj.202000203. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202000203.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 745–750, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937933. doi: 10.1145/1273496.1273590. URL https://doi.org/10.1145/1273496.1273590.

Qing Wang, Jiechao Xiong, Lei Han, peng sun, Han Liu, and Tong Zhang. Exponentially weighted imitation learning for batched historical data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/4aec1b3435c52abbdf8334ea0e7141e0-Paper.pdf.