Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [Offline Reinforcement learning with implicit Q-learning]
Date: [16-08-2024]

Make sure your critique Address these following points:
1. The problem the paper is trying to address
2. Key contributions of the paper
3. Proposed algorithm/framework
4. How the proposed algorithm addressed the described problem
Note: Be concise with your explanations. Unnecessary verbosity will be penalized. Please don't exceed 2 pages.

---

1 The problem the paper is trying to address

In offline RL, the goal is to learn effective policies from previously collected data without further interaction with the environment. However, improving the policy usually requires evaluating actions that were not seen in the dataset, which can lead to errors due to distributional shift. Existing methods often tackle this by either constraining the policy to remain close to the behavior policy or by regularizing the value functions to limit errors for out-of-distribution actions.

2 Key contributions of the paper

The RL problem is formulated in the context of a Markov decision process (MDP) $(S, \mathcal{A}, p_0(s), p(s'|s,a), r(s,a), \gamma)$, where S is a state space, $\mathcal{A}$ is an action space, $p_0(s)$ is a distribution of initial states, $p(s'|s,a)$ is the environment dynamics, $r(s,a)$ is a reward function, and γ is a discount factor. The agent interacts with the MDP according to a policy $\pi(a|s)$. The goal is to obtain a policy that maximizes the cumulative discounted returns:

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

Their work builds on approximate dynamic programming methods that minimize temporal difference error, according to the following loss:

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

where $\mathcal{D}$ is the dataset, $Q_\theta(s,a)$ is a parameterized Q-function, $Q_{\theta'}(s,a)$ is a target network (e.g., with soft parameters updates defined via Polyak averaging), and the policy is defined as $\pi(s) = \arg\max_a Q_\theta(s,a)$.

2.1 Implicit Q-Learning

The paper introduces a novel algorithm, IQL, that enables policy improvement in offline reinforcement learning without the need to evaluate or select out-of-distribution actions. This method differs from previous approaches by avoiding the issues related to distributional shift that often arise when the policy deviates from the behavior policy.

$$L(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}} \left[\left(r(s,a) + \gamma \max_{a' \notin \mathcal{A},\ \text{s.t. } \pi(a'|s')>0} Q_{\theta'}(s',a') - Q_\theta(s,a) \right)^2 \right].$$

2.2 Exceptile Regression

IQL leverages expectile regression to estimate the value of actions indirectly. By treating the state value function as a random variable and focusing on the expectile that corresponds to the best possible actions, IQL can improve the policy without requiring explicit evaluation of actions that were not seen in the dataset.

$$\hat{V}(s) = \arg\min_{V} \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}} \left[\rho_\tau \left(r + \gamma \max_{a'} Q(s', a') - V(s) \right) \right]$$

Thus $\rho_\tau(u) = |\tau - \mathbb{I}(u < 0)| \cdot u^2$ is the expectile loss function. Also, τ is the expectile parameter that controls the focus on higher or lower quantiles.

3 Proposed algorithm/framework

Algorithm 1 Implicit Q-Learning (IQL) Algorithm

```
1. Estimate the Q-function:
    Start with the Q-function estimate based on the data available in the offline dataset.
2. Decompose the Q-function:
    Decompose the Q-function into the state value function $V(s)$ and the advantage function $A(s, a)$.
3. Perform Expectile Regression:
    Use expectile regression to estimate the state value function $V(s)$, focusing on the best possible actions within the observed data.
4. Update the Policy:
    Update the policy through advantage-weighted regression, ensuring that actions with higher advantages are more likely to be selected by the policy.
5. Iterate:
    Repeat the above steps iteratively, refining the policy and the Q-function estimates until convergence.
```

4 How the proposed algorithm addressed the described problem

The algorithm addresses the problem of distributional shift in offline reinforcement learning by avoiding the direct evaluation of out-of-distribution actions. Instead of estimating Q-values for all actions, IQL uses expectile regression to focus on the best actions observed in the data, ensuring more reliable value estimates. Additionally, it implicitly improves the policy through advantage-weighted regression, which emphasizes actions with higher advantages within the observed data, leading to stable and effective policy learning without the risks associated with unseen actions.