Paper Critique

Anirud N, CE21B014

Course: DA7400, Fall 2024, IITM
Paper: Offline RL Implicit Q-Learning
Date: 16 August 2024

Make sure your critique addresses the 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 Problem Statement

Offline RL suffers from distributional shift due to unseen actions in the dataset. This paper aims to tackle this problem without further interaction with the environment. It balances improving the behavior policy and minimizing the distributional shift. The paper introduces an algorithm that avoids evaluating outside the dataset while using generalization to improve the behavior policy.

2 Key Contributions

- Implicit policy improvement using upper expectile, which avoids querying the values of unseen actions while still performing dynamic programming updates.
- Computational efficiency and state-of-the-art (SOTA) performance.
- Expectile regression to approximate the best action values.
- Advantage-weighted regression to extract the policy.
- The study offers empirical validation across a wide range of benchmarks, showing that IQL outperforms offline RL algorithms, especially for high-dimensional continuous tasks.

3 Proposed Algorithm/Framework

3.1 Learning Value Function Using Expectile Regression

$$L_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[L_2^\tau\left(Q_{\hat{\theta}}(s,a) - V_\psi(s)\right)\right]. \tag{1}$$

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

$$\underset{m_\tau}{\operatorname{argmin}}\, \mathbb{E}_{x\sim X}\left[L_2^\tau\left(x - m_\tau\right)\right], \tag{3}$$

where

$$L_2^\tau(u) = |\tau - 1(u < 0)|u^2. \tag{4}$$

Equation (1) prevents overestimation due to "lucky samples" by approximating the value function with respect to only the action distribution, thus mitigating the effect of stochasticity. Equation (2) uses mean squared error (MSE) loss to estimate the Q function, averaging over stochastic transitions to avoid issues with lucky samples. Note that no new policies are explored; the algorithm only considers actions present in the dataset.

3.2 Policy Extraction

$$L_\pi(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\exp\left(\beta\left(Q_{\hat{\theta}}(s,a) - V_\psi(s)\right)\right)\log \pi_\phi(a \mid s)\right], \tag{5}$$

where $\beta \in [0, \infty)$ is an inverse temperature parameter. For β values close to zero, the objective behaves similarly to behavioral cloning, while larger values attempt to recover the maximum of the Q function.

Algorithm 1 Implicit Q-Learning

Initialize parameters ψ, θ, $\hat{\theta}$, ϕ.
TD Learning (IQL):
for each gradient step **do**
  $\psi \leftarrow \psi - \lambda_V \nabla_\psi L_V(\psi)$
  $\theta \leftarrow \theta - \lambda_Q \nabla_\theta L_Q(\theta)$
  $\hat{\theta} \leftarrow (1-\alpha)\hat{\theta} + \alpha\theta$
end for
Policy Extraction (AWR):
for each gradient step **do**
  $\phi \leftarrow \phi - \lambda_\pi \nabla_\phi L_\pi(\phi)$
end for

For a larger value of τ we get a better approximation of the maximum but it becomes a computationally challenging problem to optimize.

4 Advantages and Conclusions of the Algorithm

- Avoided out of sample actions during training and used a multi step dp
- Computationally very efficient
- Matches the best prior performance on the MuJoCo tasks and performs the best in the challenging ant maze environment.
- Separating Value function estimation and Q value estimation prevented the effect of "lucky" actions
- The ant maze environment showed the differnce between 1 step policy improvement vs IQL. The value function in IQL closely matches the true optimal. It propagates iterative dynamic programming, and so the values are no longer determined by noise