Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [Mutual Information Regularized Offline Reinforcement Learning]
Date: [21-08-2024]

Make sure your critique Address 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 The problem the paper is trying to address

In offline RL, the learning process is based on previously collected data rather than live interactions with the environment. A key issue arises when the learning policy queries out-of-distribution (OOD) actions—actions that were not part of the dataset used for training. This can lead to extrapolation errors, where the value functions (Q-values) are inaccurately estimated for these unseen actions, causing the policy improvement direction to be biased and potentially leading to catastrophic failures, such as the collapse of the learned policy.

The paper proposes a novel framework called MISA (Mutual Information between States and Actions) to address this problem by constraining the direction of policy improvement. MISA works by estimating and maximizing the mutual information between states and actions in the offline dataset. Doing so ensures that policy improvements stay within the data manifold, thereby reducing the impact of distribution shift and improving the robustness and performance of the offline RL algorithm.

2 Key contributions of the paper

Mutual Information State-Action is a framework introduced for improving offline reinforcement learning (RL) by addressing the issue of distribution shift, which occurs when the policy encounters states or actions not well-represented in the training data. MISA regularizes the learning process by constraining policy updates to remain within the distribution of the offline dataset. This is achieved by maximizing the mutual information between states and actions within the data, ensuring the learned policy stays close to the data distribution, thereby reducing the risk of extrapolation errors and improving policy performance.

MISA with f-divergence (MISA-f):

$$I_f(S;A) = \mathbb{E}_{\mu(s)}\left[D_f\left(\pi(a|s) \parallel p(a)\right)\right] \tag{1}$$

MISA with Donsker-Varadhan (MISA-DV):

$$I_{\mathrm{DV}}(S;A) = \sup_{f:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathbb{E}_{\mu(s,a)}\left[f(s,a)\right] - \log \mathbb{E}_{\mu(s)\mu(a)}\left[e^{f(s,a)}\right] \tag{2}$$

2.1 Policy Evaluation and Improvement

$$\min_{\phi} \quad \mathbb{E}_{s,a,r,s'\sim D}\left[\frac{1}{2}\left(Q_\phi(s,a) - \mathcal{B}^\pi Q_\theta(s,a)\right)^2\right] - \alpha_1 \hat{\mathcal{I}}_2(D(\theta,\phi)), \quad \text{(Policy Evaluation)} \tag{3}$$

$$\max_{\theta} \quad \mathbb{E}_{s,a\sim D,a'\sim\pi_\theta(s')}\left[Q_\theta(s,a) + \alpha_2 \hat{\mathcal{I}}_2(D(\theta,\phi))\right], \quad \text{(Policy Improvement)} \tag{4}$$

3 Proposed algorithm/framework

Algorithm 1 Mutual Information Regularized Offline RL

```
Require: Initialize Q network $Q_\phi$, policy network $\pi_\theta$, dataset $\mathcal{D}$, hyperparameters $\alpha_1$ and $\alpha_2$.
for $t \in \{1, \ldots, \text{MAX\_STEP}\}$ do
    Train the Q network by gradient descent with objective $J_Q(\phi)$
    $\phi := \phi - \eta_Q \nabla_\phi J_Q(\phi)$
    Improve policy network by gradient ascent with objective $J_\pi(\theta)$
    $\theta := \theta + \eta_\pi \nabla_\theta \mathbb{E}_{s,a\sim\mathcal{D},a'\sim\pi_\theta(s')}[Q_\phi(s,a)] + \alpha_2 \nabla_\theta I_{\text{MISA}}$
end for
Ensure: The well-trained $\pi_\theta$.
```

The losses are defined as

$$J_Q(\phi) = J_Q^{\mathcal{B}}(\phi) - \gamma_1 \mathbb{E}_{s,a\sim\mathcal{D}}\left[Q_\phi(s,a)\right] - \gamma_1 \mathbb{E}_{s,a\sim\mathcal{D}}\left[\log \mathbb{E}_{\pi_\theta(a|s)}\left[e^{Q_\phi(s,a)}\right]\right], \quad (12) \tag{5}$$

$$J_Q^{\mathcal{B}}(\phi) = \mathbb{E}_{s,a,r,s'\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_\phi(s,a) - \mathcal{B}^\pi Q_\theta(s,a)\right)^2\right] \quad \text{represents the TD error.} \tag{6}$$

For policy improvement, note that the entropy term $H(a)$ can be omitted as it is a constant given dataset $\mathcal{D}$. Thus, we have the below objective to maximize:

$$J_\pi(\theta) = \mathbb{E}_{s,a\sim\mathcal{D},a'\sim\pi_\theta(s')}\left[Q_\phi(s,a)\right] + \gamma_2 \mathbb{E}_{s,a\sim\mathcal{D}}\left[\log \pi_\theta(a|s)\right] - \gamma_2 \mathbb{E}_{s,a\sim\mathcal{D}}\left[\log \mathbb{E}_{\pi_\theta(a|s)}\left[e^{Q_\phi(s,a)}\right]\right] \tag{7}$$

Intuitive Explanation on the Mutual Information Regularizer. By rearranging the terms in Eqn. red10, MISA can be written as:

$$I_{\text{MISA}} = \mathbb{E}_{s,a\sim\mathcal{D}}\left[\log \frac{\pi_\theta(a|s)e^{Q_\phi(s,a)}}{\mathbb{E}_{\pi_\theta(a|s')}\left[e^{Q_\phi(s',a')}\right]}\right] \tag{8}$$

4 How the proposed algorithm addressed the described problem

The proposed algorithm, **MISA (Mutual Information State-Action)**, addresses distribution shift in offline reinforcement learning by:

- **Mutual Information Regularization:** It adds a regularization term based on mutual information between states and actions, keeping the policy close to the data distribution and reducing the risk of selecting out-of-distribution (OOD) actions.

- **Conservative Policy Improvement:** The algorithm penalizes deviations from the data distribution, ensuring more reliable and conservative policy updates.