Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [Robust Adversarial Model-Based Offline RL (RAMBO-RL)]
Date: [21-08-2024]

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 The problem the paper is trying to address:

The paper addresses the challenge of offline reinforcement learning (RL), which involves learning effective policies from pre-recorded datasets without further interaction with the environment. A significant issue in offline RL is the distributional shift between the data in the dataset and the state-action pairs that the learned policy might encounter, which can lead to poor performance.

To tackle this, the paper introduces a novel approach called Robust Adversarial Model-Based Offline RL (RAMBO-RL). The key idea is to formulate the problem as a two-player zero-sum game against an adversarial environment model. This adversarial model is trained to minimize the value function while accurately predicting transitions in the dataset. This forces the policy to act conservatively in areas not well-covered by the dataset, addressing the challenge of distributional shift and leading to more reliable offline RL performance.

2 Key contribution in the paper

Robust Adversarial Reinforcement Learning: RARL addresses the problem of finding a robust agent policy, π, in the online RL setting by posing the problem as a two-player zero-sum game against an adversary policy, $\tilde{\pi}$:

$$\pi = \arg\max_{\pi \in \Pi} \min_{\tilde{\pi} \in \tilde{\Pi}} V_M^{\pi, \tilde{\pi}} \tag{1}$$

where $V_M^{\pi,\tilde{\pi}}$ is the expected value from executing π and $\tilde{\pi}$ in environment M. Different approaches define the action space for $\tilde{\pi}$ in different ways.

Model Gradient Let ϕ denote the parameters of a parametric MDP model $\hat{T}_\phi$, and let V_ϕ^π denote the value function for policy π in $\hat{T}_\phi$. Then:

$$\nabla_\phi V_\phi^\pi = \mathbb{E}_{s,a,s',r,\pi,\hat{T}_\phi} \left[\left(r + \gamma V_\phi^\pi(s') - Q_\phi^\pi(s,a) \right) \cdot \nabla_\phi \log \hat{T}_\phi(s' \mid r, s, a) \right] \tag{2}$$

The equation defines the **gradient of the value function** V_ϕ^π with respect to the model parameters ϕ in the RAMBO-RL framework. This gradient is used to **update the adversarial transition model** $\hat{T}_\phi$, which generates perturbations that minimize the value function. The goal is to train the policy π to be robust against worst-case scenarios modeled by $\hat{T}_\phi$.

Objective: Minimize the value function while ensuring the model remains close to the MLE through a trade-off parameter

TV Distance: Used to measure the difference between the learned model and the MLE.

Final Loss: Combines value optimization with a regularization term that penalizes deviations from the MLE, ensuring that the model is not only accurate but also aligned with observed data.

1. Constrained Optimization:

$$\min_{\hat{T}_\phi} V_\phi^\pi, \quad \text{s.t. } \mathbb{E}_{\mathcal{D}}\left[\text{TV}\left(\hat{T}_{\text{MLE}}(\cdot \mid s,a), \hat{T}_\phi(\cdot \mid s,a)\right)^2\right] \leq \epsilon$$

where TV denotes the total variation distance.

2. Lagrangian Relaxation:

$$\max_{\lambda \geq 0} \min_{\hat{T}_\phi} \left(V_\phi^\pi + \lambda\left(\mathbb{E}_{\mathcal{D}}\left[\text{TV}\left(\hat{T}_{\text{MLE}}(\cdot \mid s,a), \hat{T}_\phi(\cdot \mid s,a)\right)^2\right] - \epsilon\right)\right)$$

where λ is the Lagrange multiplier.

3. Simplified Objective:

$$\min_{\hat{T}_\phi} \left(\lambda V_\phi^\pi + \mathbb{E}_{\mathcal{D}}\left[\text{TV}\left(\hat{T}_{\text{MLE}}(\cdot \mid s,a), \hat{T}_\phi(\cdot \mid s,a)\right)^2\right]\right)$$

4. Final Loss Function:

$$\mathcal{L}_\phi = V_\phi^\pi + \mathbb{E}_{s,a,s',r\sim\mathcal{D}}\left[\log \hat{T}_\phi(s' \mid r,s,a)\right]$$

where $\mathcal{L}_\phi$ incorporates both the value function term and the adversarial loss term based on the MLE.

3 Proposed Algorithm

Algorithm 1 RAMBO-RL

Require: Normalized dataset, D

1: $\hat{T}_\star \leftarrow$ MLE dynamics model.
2: **for** $i = 1, 2, \ldots, n_{\text{iter}}$ **do**
3: Generate synthetic k-step rollouts. Add transition data to $D_{\hat{T}_\star}$.
4: **Agent update**: Update π and Q_θ^π with an actor-critic algorithm, using samples from $D \cup D_{\hat{T}_\star}$.
5: **Adversarial model update**: Update $\hat{T}_\phi$ according to Eq. 9 using samples from D for the MLE component, and the current critic Q_θ^π and synthetic data sampled from π and $\hat{T}_\phi$ for the adversarial component.
6: **end for**

4 How the algorithm addressed the described problem:

The proposed RAMBO-RL algorithm addresses the challenge of distributional shift in offline RL through the following mechanisms:

- **Adversarial Modeling**: The adversarial model $\hat{T}_\phi$ generates worst-case scenarios to force the policy π to be robust against unseen states.
- **Synthetic Data Generation**: By generating synthetic rollouts using the learned dynamics model $\hat{T}_\star$, the algorithm expands the state-action distribution, reducing overfitting to the offline dataset.
- **Agent Update**: The policy π and critic Q_θ^π are updated using both real and adversarially generated data, improving generalization.