Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [Model-Based Offline Reinforcement Learning (MOReL)]
Date: [09-08-2024]

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.

Offline Reinforcement learning where the main aim is to get a very high rewarding policy based on the dataset that is created using a behavioural policy π_β, the behavioural policy could be anything like from a random policy to a highly sophistic heuristic based approach policy etc. This gives us a historical interactions based dataset where we train our model. Offline RL also termed as batch RL suffers hugely with unique difficulties especially from the distribution shift. The distribution shift can be described as the state visitation of the a policy that we are choosing deviates from the data distribution that was collected by the behavioral policy. Following an example of Q-learning the scheme the target value can be

$$y(s,a) = r(s,a) + \mathbb{E}_{a' \sim \pi_{new}}[Q(s', a']$$

where π_{new} denotes the policy to be evaluated, thus in teh later part the update function will tend to the objective as shown below:

$$min_Q \mathbb{E}_{(s,a)\sim\pi_\beta}[(Q(s,a) - y(s,a))^2]$$

Here in this setting we are forced to use the samples form $\pi_\beta(a|s)$. Apart from learning from the Q-function itself, the policy is generated by

$$\pi_{new}(a|s) = argmax_\pi \mathbb{E}_{a\sim\pi(a|s)}[Q(s,a)]$$

Thus now the states are produced to yield maximum value of the Q function.

Key contribution of the paper on Model-Based Reinforcement learning or MORel are the paper provides a 2 step approach to tackle the issue of distribution shift and using of a model based approach. The two step approach can be defined as:

- **Learning a Pessimistic MDP (P-MDP)**: This step approximated the learning of a dynamic model $\hat{P}(.|s,a)$ and then using it to construct a P-MDP that divides the states into "Known" and "Unknown" based on the usage of total variance formulation, $D_{TV}(\hat{P}(.|s,a), P(.|s,a))$ between the approximate model and the transition kernel, heavily penalizing the policies that venture into the unknown regions.
- **Learning a near optimal policy**: The performance of the policy in the true environment is lower bounded by the performance in the P-MDP ensuring robustness learning of the learning of the policies against model accuracies.

The use of behavioral cloning is treated as optional choice to estimate the behavior policy π_β

Algorithm 1 MOReL: Model Based Offline Reinforcement Learning

Require: Dataset $\mathbb{D}$

1: Learn approximate dynamics model $\hat{P} : \mathbb{S} \times \mathbb{A} \to \mathcal{S}$ using $\mathbb{D}$.
2: Construct α-USAD, $U^\alpha : \mathbb{S} \times \mathbb{A} \to \{\text{TRUE, FALSE}\}$ using $\mathbb{D}$
3: Construct the pessimistic MDP $\mathbb{M}_p = \{\mathbb{S} \cup \text{HALT}, \mathbb{A}, r_p, \hat{P}_p, \rho_0, \gamma\}$
4: (OPTIONAL) Use a behavior cloning approach to estimate the behavior policy $\hat{\pi}_b$.
5: $\pi_{\text{out}} \leftarrow \text{PLANNER}(\mathbb{M}_p, \pi_{\text{init}} = \hat{\pi}_b)$
6: **return** π_{out}

The paper dealt with the core problem address the use of the Pessimistic Markov Decision based approach to tackle the issue of distribution shift and by penalizing the action that make it least visitful towards the states that mostly swades it away from the dataset that it trained on by introducing a larger negative reward. It effectively forces to stay in the known region so that model's prediction becomes reliable and making the policy operates in the region where dataset provides sufficient information to deal with.

The other was using a model based approach, model exploitation occurs when the learned policy does inaccuracies in the model to achieve high rewards. This MOReL approach gurantees that the policy is lower bounded by the performance of the MDP, ensuring there is no over estimation. By incorporating uncertainty estimates into the P-MDP, MOReL ensures that the learned policy does not exploit model inaccuracies. Instead, it favors actions that are supported by the data, leading to more reliable and robust policies.