Paper Critique

Anirud N, CE21B014

Course: DA7400, Fall 2024, IITM
Paper:COptiDICE: Offline Constrained Reinforcement Learning via Stationary Distribution Correction Estimation
Date: 27 September 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

The paper addresses the problem of solving offline RL while being constrained. The agent needs to learn from an offline dataset, and it also needs to follow and respect the constraints that are set in the environment for safety. Offline RL suffers from distribution shifts and overestimation of values for state actions that may not be present in the dataset. Making the policy trained on an offline dataset to satisfy the cost constraints is an added challenge because - off-policy evaluation involves estimation errors. So it is not so easy to ensure that the policy estimated from a finite dataset would adhere to the constraints in the environment. Including the cost constraints mean that we need to optimize over the Lagrange multipliers. Actor critic-based constrained RL might be very unstable in practice.

2 Key Contributions

The paper introduces the idea of optimizing the state action stationary distribution directly, instead of the Q functions or the policy. This bypasses need the multiple estimators for values and policy, giving a single objective function to solve. The paper also introduces a new method to compute a policy more robust in constraint violation, computing the upper bound using a off policy confidence interval estimation. This provides a more computationally efficient way to estimate an upper bound on cost value compared to traditional bootstrapping methods.

3 Proposed Algorithm/Framework

The two constraints involved are - cost constraints and bellman flow constraints. FLow constraint is give below:

$$d_\pi(s) = (1-\gamma)d_0(s) + \gamma \sum_{s',a'} T(s \mid s', a') d_\pi(s', a')$$

We formulate the constrained optimization problem as a Lagrangian:

$$\min_{\lambda \geq 0, \nu} \max_{d \geq 0} \mathbb{E}_{(s,a)\sim d}[R(s,a)] - \alpha D_f(d||d^D) - \sum_{k=1}^{K} \lambda_k \big(\mathbb{E}_{(s,a)\sim d}[C_k(s,a)] - \hat{c}_k\big) - \sum_{s'} \nu(s') \Big[\sum_{a'} d(s',a') - (1-\gamma)p_0(s') - \gamma \sum_{s,a} d(s,a) T(s'|s,a) \Big] \tag{1}$$

introducing new optimization variables to make the optimization tractable, the paper introduces $w(s,a) = d(s,a)/d^D(s,a)$,and formulates into a convex minimisation problem:

$$\min_{\lambda \geq 0, \nu} L(\lambda, \nu) = \mathbb{E}_{(s,a)\sim d^D}\left[w^*_{\lambda,\nu}(s,a)e_{\lambda,\nu}(s,a) - \alpha f(w^*_{\lambda,\nu}(s,a))\right] + (1-\gamma)\mathbb{E}_{s_0\sim p_0}[\nu(s_0)] + \lambda^\top \hat{c} \tag{2}$$

This is formulated in a Joint optimization problem as follows:

$$\begin{aligned} &\nu \leftarrow argmin_\nu L(\lambda, \nu) && \text{(OptiDICE for } R - \lambda^\top C) \quad (3)\\ &\lambda \leftarrow argmin_{\lambda \geq 0} \lambda^\top \big(\hat{c} - \underbrace{_{(s,a)\sim d^D}[w^*_{\lambda,\nu}(s,a)C(s,a)]}_{\approx V_C(\pi)}\big) && \text{(Cost Lagrange multiplier)} \end{aligned}$$

For performing a conservative cost-constrained policy optimization, the paper does it in an adversarial fashion - adversarially optimized the distribution over data samples $\tilde{p}$, to overestimate the cost value, and it also intrudes constraints to ensure that this data sample distribution does not deviate by a lot from the dataset distribution. This is given by the equation:

$$\max_{\tilde{p}\in\Delta(X)} \mathbb{E}_{(s_0,s,a,s')\sim\tilde{p}}[w(s,a)C_k(s,a)] \tag{4}$$

$$\text{s.t. } D_{\mathrm{KL}}(\tilde{p}(s_0,s,a,s')||d^D(s_0,s,a,s')) \leq \epsilon \tag{5}$$

$$\sum_{a'} \tilde{p}(s',a')w(s',a') = (1-\gamma)\tilde{p}_0(s') + \gamma \sum_{s,a} \tilde{p}(s,a)w(s,a)\tilde{p}(s'|s,a) \quad \forall s' \tag{6}$$

to extract the policy: for finite CMDP : $\pi^*(a|s) = \frac{d^{\pi^*}(s,a)}{\sum_{a'} d^{\pi^*}(s,a')} = \frac{d^D(s,a)w^*(s,a)}{\sum_{a'} d^D(s,a)w^*(s,a)}$. But incase of continuous CMDPs, the policy is extracted in a importance weighted behavior cloning.

$$\max_\pi \mathbb{E}_{(s,a)\sim d^{\pi^*}}[\log \pi(a|s)] =_{(s,a)\sim d^D} [w^*(s,a)\log \pi(a|s)] \tag{7}$$

A lot of approximations were involved, during implementation, as given in the paper.

4 Conclusions and Results

- COptiDICE manages to balance between reward performance and strict cost constraint satisfaction by constraining the upper bound of the cost value. This method is particularly effective in both high-data and low-data scenarios.
- The baselines struggled in smaller datasets and they very often violated the constraints.
- Pessimistic approach in other offline RL algorithms, tried to stay close to the dataset distribution.
- In settings where the data-collection policy is constraint-violating, only COptiDICE ensures constraint adherence, making it a superior algorithm for such situations.