Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [COptiDICE: Offline Constrained Reinforcement Learning via Stationary Distribution Correction Estimation]
Date: [27-09-2024]

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

This problem arises in the Offline Constrained Reinforcement Learning setting, where the goal is to learn an optimal policy from a fixed, pre-collected dataset without further interactions with the environment, while satisfying predefined safety or cost constraints. The challenge comes from the distribution shift between the offline data and the learned policy, which makes it hard to guarantee constraint satisfaction. The paper introduces the COptiDICE algorithm to estimate stationary distribution corrections and solve this problem effectively.

Maximize the expected reward:

$$\mathbb{E}_{(s,a)\sim d_\pi}[R(s,a)]$$

Subject to cost constraints:

$$\mathbb{E}_{(s,a)\sim d_\pi}[C_k(s,a)] \leq \hat{c}_k, \quad \forall k \in \{1,\ldots,K\}$$

2 Key contributions of the paper

- The paper introduces a new algorithm, **COptiDICE**, for offline constrained reinforcement learning by estimating the stationary distribution correction.
- **COptiDICE** avoids the nested optimization problem of prior methods by constraining the upper bound of the cost value, ensuring better constraint satisfaction.
- It leverages **DICE**-family methods to estimate the correction between the policy's and dataset's stationary distributions, leading to improved reward maximization and constraint satisfaction.
- Empirical results demonstrate that **COptiDICE** achieves a superior trade-off between reward maximization and constraint satisfaction compared to baseline algorithms.

3 Proposed algorithm/framework

Algorithm 1 COptiDICE

Input: An offline dataset $D = \{(s_0, s, a, r, c, s')_i^N\}_{i=1}$, a learning rate η.
Initialize parameter vectors θ, ϕ, λ, τ, ψ.
for each gradient step **do**
    Sample mini-batches from D.
    Compute gradients and perform SGD update:
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{J}_\theta(\psi)$
    $\tau \leftarrow \tau - [\tau - \eta \nabla_\tau \mathcal{J}_\tau(\tau, \phi)]_+$
    $\phi \leftarrow \phi - \eta \nabla_\phi \mathcal{J}_\lambda(\tau, \phi)$
    $\lambda \leftarrow [\lambda - \eta \nabla_\lambda \mathcal{J}_\lambda]_+$
    $\psi \leftarrow \psi - \eta \nabla_\psi \mathcal{J}_\theta(\psi)$
end for

4 How the proposed algorithm addressed the problem

Objective Function

$$\max_\pi \mathbb{E}_{(s,a)\sim d_\pi}[R(s,a)]$$

subject to:

$$\mathbb{E}_{(s,a)\sim d_\pi}[C_k(s,a)] \leq \hat{c}^k, \quad \forall k \in \{1, \ldots, K\}$$

f-Divergence Penalized Optimization

$$\max_d \mathbb{E}_{(s,a)\sim d}[R(s,a)] - \alpha D_f(d \| d_D)$$

subject to:

$$\mathbb{E}_{(s,a)\sim d}[C_k(s,a)] \leq \hat{c}^k, \quad \forall k \in \{1, \ldots, K\}$$

Bellman Flow Constraint

$$\sum_{a'} d(s', a') = (1-\gamma) p_0(s') + \gamma \sum_{s,a} d(s,a) T(s'|s,a)$$

Lagrangian Formulation

$$L = \mathbb{E}_{(s,a)\sim d}[R(s,a)] - \alpha D_f(d \| d_D) - \sum_{k=1}^{K} \lambda_k \left(\mathbb{E}_{(s,a)\sim d}[C_k(s,a)] - \hat{c}^k \right)$$

Stationary Distribution Correction

$$w(s,a) = \frac{d_\pi(s,a)}{d_D(s,a)}$$

Min-Max Optimization Problem

$$\min_{\lambda \geq 0, \nu} \max_{w \geq 0} \mathbb{E}_{(s,a)\sim d_D} \left[w(s,a) e^{\lambda, \nu(s,a)} - \alpha f(w(s,a)) \right]$$