Paper Critique

Anirud N, CE21B014

Course: DA7400, Fall 2024, IITM
Paper: [BAIL: Best-Action Imitation Learning for Batch Deep Reinforcement Learning]
Date: [9 August 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.

---

1 Problem Statement

The paper talks about the challenges in Deep Reinforcement Learning. The current algorithms, such as Q function-based, DDPG, etc, for learning from the given dataset without additional interactions with the environment don't do well and give poor results due to **extrapolation error**. The paper implements BAIL - for simplicity, better performance, and faster computationally.

2 Key Contributions

This paper introduces the BAIL - Batch DRL algorithm. It devised the concept of the "Upper envelope of data." It learns a V function using a neural network to find the upper envelope of the data. It trains the policy network with this upper network using Imitation Learning. The paper therefore, combines V learning with Imitation Learning. The paper experiments these algorithms on different datasets and analyses the performance.

3 Proposed Algorithm/Framework

The paper assumes that the environment is **Deterministic**. However, It does not assume the knowledge of the function for finding s' and r from the state action pair but assumes the existence of the functions. It also assumes that the batch data is generated and ordered episodic. However, We do not have access to the policy that generated the batch B.

- Estimate a V(s) close enough to V*(s)
- Find G(s,a) - discounted Monte Carlo Returns
- Identify the State action pairs that have G(s,a) close to V*(s) using the upper envelope of the data
- Using the selected state action pairs - Perform Imitation Learning.

BAIL uses 2 neural networks.

- Approximate the optimal value function
- Learing policy using Imitation learning

3.1 Upper Envelope of the Data

Batch of data $\mathcal{B} = \{(s_i, a_i, r_i, s_i'),\ i = 1, ..., m\}$. For each data point $i \in \{1, \ldots, m\}$, Monte Carlo discounted return is found G_i as $G_i = \sum_{t=i}^{T} \gamma^{t-i} r_t$ where T denotes the time the episode ends for the episode that contains the ith data point. To find the upper envelope, a neural network is used : $V_\phi(s)$ parameterized by $\phi = (w, b)$ that takes as input a state s and outputs a real number, where w and b denote the weights and bias, respectively. We find $V_\phi(s)$ as we solve a constraint optimisation problem as :

$$\min_{\phi} \sum_{i=1}^{m} [V_\phi(s_i) - G_i]^2 + \lambda \|w\|^2 \qquad s.t. \qquad V_\phi(s_i) \geq G_i, \qquad i = 1, 2, \ldots, m \tag{1}$$

λ ¿ 0 is set- regularisation to prevent overfitting. Eq(1) could be converted to Unconstrained optimization with K¿¿1 as the penalty factor

$$L^K(\phi) = \sum_{i=1}^{m} (V_\phi(s_i) - G_i)^2 \{1_{(V_\phi(s_i) \geq G_i)} + K \cdot 1_{(V_\phi(s_i) < G_i)}\} + \lambda \|w\|^2 \tag{2}$$

3.2 Selecting Best Actions

BAIL ratio For a fixed $x > 0$:

$$G_i > xV(s_i) \tag{3}$$

x is set such that atleast p(=25%) % of the points are selected.

BAIL Difference For a fixed $x > 0$:

$$G_i \geq V(s_i) - x \tag{4}$$

The results from the above 2 methods were similar

4 Advantages and Conclusions of the Algorithm

- BAIL performs better than BCQ and BEAR intuitively because these are policy constraint methods. So they depend on tuned constraints to prevent problems due to out-of-dataset distribution actions.
- Loose constraint - Extrapolation errors. Tight Constraint - Misses some good actions
- Considering the training batches, BAIL is a better algorithm than BCQ, BEAR, BC, and MARWIL, as it wins in 20 out of 22 batch generations of the dataset.
- BAIL is roughly 35 times faster than BCQ and 50 times faster than BEAR