# Paper Critique

Shuvrajeet Das, DA24D402

**Course:** DA7400, Fall 2024, IITM
**Paper:** [Discriminator-Weighted Offline Imitation Learning from Suboptimal Demonstrations]
**Date:** [16-08-2024]

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.

---

## 1 The problem the paper is trying to address

The paper addresses the problem of learning an optimal policy from offline imitation learning (IL) in situations where the available demonstrations are suboptimal or mixed in quality. In many real-world scenarios, obtaining optimal demonstrations is challenging, expensive, or impractical, leading to datasets that contain a significant proportion of suboptimal examples. Traditional IL methods, such as Behavioral Cloning (BC), assume that the demonstrations are near-optimal, which can lead to poor performance when this assumption does not hold. These methods may overfit to the suboptimal data, resulting in a policy that replicates the mistakes or inefficiencies present in the demonstrations.

## 2 Key contributions of the paper

The key contributions of this paper can be defined as through the following points which is described below:

- Cooperative Framework between Policy and Discriminator: The paper proposes a framework where both the policy $\pi_\theta$ and the discriminator $D_\phi$ are learned in a cooperative manner. The discriminator $D_\phi(s,a)$ is trained to distinguish between optimal and suboptimal actions, while the policy $\pi_\theta$ is trained using a weighted loss function that leverages the output of the discriminator.

  **Discriminator Objective**: The discriminator is trained to classify whether a state-action pair $(s,a)$ is from an optimal or suboptimal policy:

  $$\mathcal{L}_D(\phi) = -\mathbb{E}_{(s,a)\sim\pi^*}[\log D_\phi(s,a)] - \mathbb{E}_{(s,a)\sim\pi_\theta}[\log(1 - D_\phi(s,a))]$$

  where $\pi^*$ represents the expert policy, and $\pi_\theta$ is the current policy being learned.

- Discriminator-Weighted Behavioral Cloning (DWBC):The policy is trained using a modified behavioral cloning objective, where the loss function is weighted by the discriminator's output. This ensures that higher quality demonstrations have a greater influence on the learned policy.

  **Policy Objective (DWBC)**: The policy $\pi_\theta$ is optimized using the following weighted behavioral cloning loss:

  $$\mathcal{L}_{\text{DWBC}}(\theta) = -\mathbb{E}_{(s,a)\sim\mathcal{D}}[D_\phi(s,a)\log\pi_\theta(a|s)]$$

  Here, $D_\phi(s,a)$ acts as a weight that emphasizes the importance of higher-quality demonstrations. $\mathcal{D}$ denotes the dataset of demonstrations.

# 3 Proposed algorithm/framework

**Algorithm 1** Discriminator-Weighted Behavior Cloning (DWBC)

```
Require: Dataset $D_e$ and $D_o$, hyperparameter $\eta$, $\alpha$
Initialize the imitation policy $\pi$ and the discriminator $d$
while training do
    Sample $(s_e, a_e) \sim D_e$ and $(s_o, a_o) \sim D_o$ to form a training batch $B$
    Compute $\log \pi(a|s)$ values for samples in $B$ using the learned policy $\pi$
    Compute discriminator output values $d(s, a, \log \pi(a|s))$ using sampled $(s, a)$ and computed $\log \pi(a|s)$
    Update $d$ by minimizing the learning objective $L_d$ every 100 training steps
    Update $\pi$ by minimizing the learning objective $L_\pi$ every 1 training step
end while
```

# 4 How the proposed algorithm addressed the described problem

The DWBC algorithm addresses the challenge of learning from suboptimal demonstrations in offline imitation learning through the following mechanisms:

- **Discriminator to Identify Demonstration Quality**

  **Problem:** Traditional methods like Behavioral Cloning (BC) assume all demonstrations are near-optimal, leading to poor policy performance when suboptimal behavior is present.

  **Solution:** The algorithm introduces a discriminator $D_\phi(s, a)$ to distinguish between high-quality (optimal) and low-quality (suboptimal) demonstrations. The discriminator assigns higher values to likely optimal actions, guiding the policy to focus on more relevant data.

- **Discriminator-Weighted Policy Learning**

  **Problem:** Standard BC loss treats all demonstrations equally, which can degrade performance in the presence of mixed-quality data.

  **Solution:** DWBC modifies the policy learning process by weighting the BC loss with the discriminator's output and by prioritizing high-quality demonstrations and reducing the influence of suboptimal data.

- **Cooperative Training Between Policy and Discriminator**

  **Problem:** A static weighting mechanism may not adapt well to a changing policy or non-stationary environment.

  **Solution:** DWBC employs a cooperative framework where the policy $\pi_\theta$ and discriminator $D_\phi$ are updated iteratively. The policy informs the discriminator by generating actions, and the discriminator adjusts the weights in the loss function, refining its ability to distinguish between optimal and suboptimal actions over time.