DISCRIMINATOR-WEIGHTED OFFLINE IMITATION LEARNING FROM SUBOPTIMAL DEMONSTRATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the problem of offline Imitation Learning (IL) where an agent aims to learn an optimal expert behavior policy without additional online environment interactions. Instead, the agent is provided with a static offline dataset of state-action-next state transition triples from both optimal and non-optimal expert behaviors. This strictly offline imitation learning problem arises in many real-world problems, where environment interactions and expert annotations are costly. Prior works that address the problem either require that expert data occupies the majority proportion of the offline dataset, or need to learn a reward function and perform offline reinforcement learning (RL) based on the learned reward function. In this paper, we propose an imitation learning algorithm to address the problem without additional steps of reward learning and offline RL training for the case when demonstrations containing large proportion of suboptimal data. Built upon behavioral cloning (BC), we introduce an additional discriminator to distinguish expert and non-expert data, we propose a cooperation strategy to boost the performance of both tasks, this will result in a new policy learning objective and surprisingly, we find its equivalence to a generalized BC objective, where the outputs of discriminator serve as the weights of the BC loss function. Experimental results show that the proposed algorithm can learn behavior policies that are much closer to the optimal policies than policies learned by baseline algorithms.

1 INTRODUCTION

The recent success of reinforcement learning (RL) in many domains showcases the great potential of applying this family of learning methods to real-world applications. A key prerequisite for RL is to design a reward function that specifies what kind of agent behavior is preferred. However, in many real-world applications, designing a reward function is prohibitively difficult (Ng et al., 1999; Irpan, 2018). By contrast, imitation learning (IL) provides a much easier way to leverage the reward function implicitly from the collected demonstrations and has achieved great success in many sequential decision making problems (Pomerleau, 1989; Ng et al., 2000; Ho & Ermon, 2016).

However, popular IL methods such as behavioral cloning (BC) and generative adversarial imitation learning (GAIL) (Ho & Ermon, 2016), assume the expert demonstration is optimal. Unfortunately, it is often difficult to obtain optimal demonstrations for many real-world tasks, because human experts often make mistakes due to various reasons, such as the difficulty of the task, partial observability of the environment, or the presence of distraction. Given such noisy expert demonstrations, which contain records of both optimal and non-optimal behaviors, BC and GAIL all fail to imitate the optimal policy (Wu et al., 2019b; Ma, 2020). Current methods that deal with suboptimal demonstrations either require additional labels, which can be done explicitly by annotating each demonstration with confidence scores by human experts (Wu et al., 2019b), or implicitly by ranking noisy demonstrations according to their relative performance through interacting with the environment (Brown et al., 2019; 2020). However, human annotation and environment interaction are laborious and expensive in real-world settings, such as in medicine, healthcare, and industrial processes.

In this work, we investigate a pure offline learning setting where the agent has access to neither the expert nor the environment for additional information. The agent, instead, has only access to a small pre-collected dataset of state-action-next state transition triples sampled from the expert and a large batch offline dataset sampled from one or multiple behavior policies that could be highly sub-optimal.

Prior works that address the problem are based on variants of BC or inverse RL, Sasaki & Yamashina (2021) reuse another policy learned by BC as the weight of original BC objective, however, this requires that expert data occupy the majority proportion of the offline dataset, otherwise the policy will be misguided to imitate the suboptimal data. Zolna et al. (2020) first learns a reward function that prioritizes expert data over others and then performs offline RL based on this reward function. This algorithm is extremely expensive to run, requiring solving offline RL in an inner loop, which itself is a challenging problem and prone to training instability (Kumar et al., 2019) and hyperparameter sensitivity (Wu et al., 2019a).

In this paper, we propose an offline imitation learning algorithm to learn from demonstrations that (perhaps) contain a large proportion of suboptimal data without additional steps of reward learning and offline RL training. Built upon the task of behavioral cloning (BC), we introduce an additional task to learn a discriminator to distinguish expert and non-expert data, we propose a cooperation strategy to boost the performance of both tasks. This results in a new policy learning objective and surprisingly, we find its equivalence to a generalized BC objective, where the outputs of the discriminator serve as the weights of the BC loss function. We thus term our resulting algorithm Discriminator-Weighted Behavioral Cloning (DWBC). Experimental results show that DWBC can learn policies that are much closer to the optimal policy than the ones learned by other baselines under different scenarios.

2 PRELIMINARY

2.1 PROBLEM SETTING

We consider the standard fully observed Markov Decision Process (MDP) setting (Sutton et al., 1998), $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma, d_0\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the MDP's transition probability, r is the reward function, $\gamma \in [0, 1)$ is the discount factor for future reward and d_0 is the initial distribution. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps from state to distribution over actions. We denote $d^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$ as the discounted state-action distribution of π under transition kernel P, that is, $d^\pi = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t d_t^\pi$, where $d_t^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$ is the distribution of $\left(s^{(t)}, a^{(t)}\right)$ under π at step t. Following the standard IL setting, the ground truth reward function r is unknown. Instead, we have the demonstrations by the expert specified by $\pi_e : \mathcal{S} \to \Delta(\mathcal{A})$ (potentially stochastic and not necessarily optimal). Concretely, we have an expert dataset in the form of i.i.d tuples $\mathcal{D}_e = \{s_i, a_i, s_i'\}_{i=1}^{n_e}$ where (s, a) is sampled from distribution d^{π_e} and s' is sampled from $P(s, a)$.

In our problem setting, we also have an offline static dataset consisting of i.i.d tuples $\mathcal{D}_o = \{s_i, a_i, s_i'\}_{i=1}^{n_o}$ s.t. $(s, a) \sim \rho(s, a), s' \sim P(s, a)$, where $\rho \in \Delta(\mathcal{S} \times \mathcal{A})$ is an offline state-action distribution resulting from some other behavior policies. Note that these behavior policies could be much worse than the expert π_e. Our goal is to only leverage the offline batch data $\mathcal{D}_b = \mathcal{D}_e \cup \mathcal{D}_o$ to learn an optimal policy π with regard to optimizing the ground truth reward r, without any interaction with the environment or expert.

2.2 A GENERALIZED BEHAVIORAL CLONING OBJECTIVE

In order to discard low-quality demonstrations and only clone the best behavior available, previous works (Sasaki & Yamashina, 2021; Zolna et al., 2020) consider a generalized behavioral cloning objective to imitate demonstrations unequally, that is,

$$\min_{\pi} \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_b} \left[-\log \pi(a|s) \cdot f(s, a)\right] \tag{1}$$

where $f : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes an arbitrary weight function.

- If $f(s, a) = 1$ for $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, the objective (1) corresponds to the vanilla BC objective.
- If $f(s, a) = \pi'(a|s)$, where π' is an old policy which was previously optimized with $\mathcal{D}_b$, the objective (1) corresponds to the objective of Behavioral Cloning from Noisy Demonstrations (Sasaki & Yamashina, 2021). Since $\sum_a \pi'(a|s) = 1$ for $\forall s \in \mathcal{S}$ is satisfied, $\pi'(a|s)$ can be interpreted as the weights for the weighted action sampling.

- If $f(s,a) = \mathbb{1}\left[A^{\pi}(s,a)\right]$, where $\mathbb{1}$ is the indicator function which creates a boolean mask that eliminates samples which are thought to be worse than the current policy, the objective (1) corresponds to the objective of Offline Reinforced Imitation Learning (Zolna et al., 2020).

The objective (1) can also be deemed as the objective of Soft Q Imitation Learning (Reddy et al., 2019) with $f(s,a) = 1$ for $(s,a) \in \mathcal{D}_e$ and $f(s,a) = 0$ for $(s,a) \in \mathcal{D}_o$ in online IL literature; or the objective of off-policy actor-critic (Off-PAC) algorithm (Degris et al., 2012) with $f(s,a) = Q^{\pi}(s,a) \cdot \pi(a|s)/\pi_b(a|s)$ in online RL literature.

3 Discriminator-Weighted Behavioral Cloning

We now continue to describe our approach for offline imitation learning from demonstrations that (perhaps) contain large-proportional suboptimal data without additional steps of reward learning and offline RL training. Built upon the task of BC, we introduce an additional task to learn a discriminator to distinguish expert and non-expert data, we propose a cooperation strategy to boost the performance of both tasks, this will result in a new generalized BC objective. We then provide the interpretation of weights in our generalized BC objective, this gives the intuition about why our method can work.

3.1 Learn the policy and discriminator separately

It is obvious that we can avoid the negative impact of suboptimal demonstrations presented in $\mathcal{D}_o$ by only imitating $\mathcal{D}_e$, which can be written as

$$\min_{\pi} \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e} \left[-\log \pi(a|s)\right] \tag{2}$$

We call the task of learning a policy using objective (2) as the BC task. The drawback of BC task is that the learned policy may not be able to generalize due to the potential limited size and state coverage of $\mathcal{D}_e$. It does not fully utilize the information from $\mathcal{D}_o$. If we can select those high-reward transitions from $\mathcal{D}_o$ and combine them with $\mathcal{D}_e$, we are expected to get a better policy.

Now let's consider another different task, which aims to learn a discriminator by contrasting expert and non-expert transitions, given by

$$\min_{d} \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e} \left[-\log d(s,a)\right] + \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_o} \left[-\log(1 - d(s,a))\right] \tag{3}$$

We call this the discriminating task, which is similar to how the discriminator is trained in GAIL (Ho & Ermon, 2016) and GAN (Goodfellow et al., 2014), except that the second term is sampled from a fixed dataset instead of new samples obtained from interacting with the environment.

However, optimizing objective (3) will make the discriminator learned to assign 1 to all transitions from $\mathcal{D}_e$ and 0 to all transitions from $\mathcal{D}_o$. This limiting behavior is unsatisfactory because $\mathcal{D}_o$ can contain some successful (high-reward) transitions. This bears similarity to the positive-unlabeled classification problem (Elkan & Noto, 2008), where both positive and negative samples exist in the unlabeled data.

To solve this problem, we adopt the approach from positive-unlabeled (PU) learning (Du Plessis et al., 2015). The main idea is to re-weight different losses for positive and unlabeled data, in order to obtain an estimate of model loss on negative samples that is not directly available. Applying PU learning to objective (3) yields the following objective:

$$\min_{d} \eta \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e} \left[-\log d(s,a)\right] + \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_o} \left[-\log(1 - d(s,a))\right] - \eta \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e} \left[-\log(1 - d(s,a))\right] \tag{4}$$

where η is a hyperparameter, corresponds to the proportion of positive samples to unlabeled samples. Intuitively, the second term in (4) could make $d(s,a)$ of state-action pairs from $\mathcal{D}_e$ become 0 if similar state-action pairs are included in $\mathcal{D}_o$, and the third term in (4) balances the impact of the second term, i.e., avoids $d(s,a)$ of state-action pairs from $\mathcal{D}_e$ becoming 0.

To summarize, the BC task aims to imitate the expert behavior from $\mathcal{D}_e$, but ignores the valuable information in $\mathcal{D}_o$; the discriminating task aims to contrast expert and non-expert transitions from $\mathcal{D}_e$ and $\mathcal{D}_o$, by only using state-action information as input. Both tasks lack enough information to

improve their own performance, which, however, can be obtained from the other task. It then seems natural to find a scheme to incorporate the policy into the training of the discriminator and effectively use the discriminator to help the training of the policy.

3.2 LEARN THE POLICY AND DISCRIMINATOR COOPERATIVELY

We propose a cooperation strategy to boost the performance of both tasks. To boost the performance of the discriminating task, we add the imitation information from π to the input of the discriminator, yielding the following discriminator learning objective

$$\begin{aligned}\min_d \eta \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e}[-\log d(s,a,\log\pi(a|s))] + \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_o}[-\log(1-d(s,a,\log\pi(a|s)))] \\ -\eta \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e}[-\log(1-d(s,a,\log\pi(a|s)))]\,.\end{aligned} \tag{5}$$

Supposed π is learned to be optimal, i.e., assigns large probabilities to expert actions in expert states, the discriminator will receive additional learning signal. It will be easier for the discriminator to contrast expert and non-expert transitions in $\mathcal{D}_o$, as $\log\pi(a|s)$ will be large if (s,a) are from expert behaviors and small if (s,a) are from non-expert behaviors. Without this information from π, the discriminator is much hard to learn by only using information from (s,a).

Now let's dive deeper into objective (5), which we denote it as $\min_d \mathcal{L}_d$. As π now appears in the input of the discriminator d, hence both d and its loss $\mathcal{L}_d$ become functionals of π. We are interested to see how the variation of π affects $\mathcal{L}_d$. More specifically, we fix the parameter of d to exclude its influence, in such case, we can define the functional $J(\pi)$ for $\mathcal{L}_d$ in the following integral form:

$$J(\pi) = \int\int \frac{\partial \mathcal{L}_d(s,a,d,\log\pi)}{\partial d(s,a,\log\pi)}\, ds\, da = \int\int F(s,a,\pi,\pi')\, ds\, da, \tag{6}$$

where $F = \partial\mathcal{L}_d/\partial d$ is assumed to be continuously differentiable with respect to s, a, π and π' (derivative of π). To more robustly learn the discriminator d, inspired by the idea of adversarial learning (Lowd & Meek, 2005), we make π challenge the discriminator by providing as little information in $\log\pi$ as possible such that minimizing $\mathcal{L}_d$ becomes harder for the discriminator. This is equivalent to maximize $\mathcal{L}_d$ for π, so we let the functional $J(\pi)$ attains its maxima with respect to π. According to the calculus of variations (Gelfand et al., 2000), the extrema (maxima or minima) of functional $J(\pi)$ can be obtained by finding a function π such that the functional derivative of $J(\pi)$ is equal to zero. We can show with theoretical derivation (see Appendix B) that ensuring the functional derivative of $J(\pi)$ equal to zero requires $\partial F/\partial\theta = 0$, with θ being the parameters of π. We have

$$\begin{aligned}\frac{\partial F}{\partial\theta} = \frac{\partial F}{\partial\log\pi}\cdot\nabla_\theta\log\pi = -\mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e}\left[\frac{\eta}{d}\cdot\nabla_\theta\log\pi(a|s)\right] + \\ \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_o}\left[\frac{1}{1-d}\cdot\nabla_\theta\log\pi(a|s)\right] - \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e}\left[\frac{\eta}{1-d}\cdot\nabla_\theta\log\pi(a|s)\right],\end{aligned} \tag{7}$$

where write $d(s,a,\log\pi(a|s))$ as d for simplicity by slightly abusing notations. Notice that above derviative can be equivalently perceived as the gradient of a new loss term $\mathcal{L}_w$ of the policy π ($\partial\mathcal{L}_w/\partial\theta = -\partial F/\partial\theta$) with following form:

$$\mathcal{L}_w = \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e}\left[\log\pi(a|s)\cdot\frac{\eta}{d}\right] - \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_o}\left[\log\pi(a|s)\cdot\frac{1}{1-d}\right] + \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e}\left[\log\pi(a|s)\cdot\frac{\eta}{1-d}\right].$$

Hence minimizing $\mathcal{L}_w$ with respect to π (make $\partial\mathcal{L}_w/\partial\theta = 0$) will drive the functional derivative of $J(\pi)$ to zero, which also lead to the maxima of $J(\pi)$ (update with gradient direction $-\partial F/\partial\theta$). Adding this loss term $\mathcal{L}_w$ to the BC task, we get the following new learning objective of π as:

$$\min_\pi \alpha \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e}[-\log\pi(a|s)] - \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e}\left[-\log\pi(a|s)\cdot\frac{\eta}{d\,(1-d)}\right] + \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_o}\left[-\log\pi(a|s)\cdot\frac{1}{1-d}\right], \tag{8}$$

where α is the weight factor ($\alpha \geq 1$). This new objective essentially transforms the original BC task into a cost-sensitive learning problem (Ling & Sheng, 2008) by imposing the following weight on imitating each state-action transition as

$$\text{Behavioral cloning weights} = \begin{cases}\alpha - \eta/d(1-d), & (s,a)\in\mathcal{D}_e \\ 1/\left(1-d\right), & (s,a)\in\mathcal{D}_o\end{cases}. \tag{9}$$

Note that the derivation of the above behavioral cloning weights requires uniform continuity to be satisfied in F and its derivative (details see Appendix B). The involvement of $1/d$ and $1/(1-d)$ may violate the continuity assumption. We thus clip the value d to the range of $[0.1, 0.9]$.

Above behavioral cloning weights induce different behaviors on the imitation of transitions from $\mathcal{D}_e$ and $\mathcal{D}_o$. Supposed d is learned to be optimal, i.e., assigns large values (close to 1) to expert transitions and small values to non-expert transitions (close to 0). The weight of those expert transitions in $\mathcal{D}_o$ will become large while the weight of those non-expert transitions will become small. For transitions in $\mathcal{D}_e$, their weights can be adjusted by tuning the parameter α. Note that even if the discriminator is learned to be totally wrong (i.e., assign small values to expert transitions and large values to non-expert transitions), which may occur at the very beginning of training, the behavior cloning weights $\alpha - \eta/d(1-d)$ ($\alpha \geq 1, \eta < 1$) will not be drastically changed under value clipping. This means that the policy can still learn from the expert dataset $\mathcal{D}_e$. Even though the weight for $\mathcal{D}_o$ is temporarily incorrect, it will be corrected as the discriminator becomes better and better.

Eq. (8) implies that our approach is also a variant of generalized behavioral cloning objective, as discussed in section 2.2, but uses a different form of weights. Unlike Offline Reinforced Imitation Learning (Zolna et al., 2020), which uses the discriminator as the reward and learns a value function as the weight, our approach uses the discriminator outputs directly as the weight. This can greatly reduce the training time and avoid the overestimation issue in estimating the value function offline (Kumar et al., 2019). We thus term our algorithm Discriminator-Weighted Behavioral Cloning (DWBC).

A keen reader may find the similarity of our approach to the adversarial learning in GAN (Goodfellow et al., 2014), which makes the generator (in our case is the policy π) and the discriminator d learn against each other, thus improves the performance of both tasks. However, our proposed cooperation strategy is much easier to learn than GAN. GAN needs to solve a min-max optimization problem and is known to suffer from training instability and issues such as mode collapse (Arjovsky et al., 2017). In DWBC, we solve the min-max optimization problem implicitly by transforming the max optimization problem to a new learning loss imposed on the policy.

4 Related Work

4.1 Offline Imitation Learning

Offline IL, which has not received considerable attention, is a promising area because it makes IL more practical to satisfy critical safety desiderata. Offline IL methods can be typically folded into two paradigms: Behavioral Cloning (BC) and Offline Inverse Reinforcement Learning (Offline IRL).

BC (Pomerleau, 1989) is the simplest IL method that can be used in the offline setting, it considers the policy as a conditional distribution $\pi(\cdot|s)$ over actions, recent work (Florence et al., 2021) enhances BC by using energy-based models (LeCun et al., 2006). BC has shown to have no inferior performance compared to popular IL algorithms such as GAIL (Ho & Ermon, 2016) when clean expert demonstrations are available (Ma, 2020). Unlike BC, offline IRL considers matching the state-action distribution induced by the expert policy, this can be achieved implicitly by adversarial training or explicitly by learning a reward function. Offline IRL algorithms based on adversarial training (Kostrikov et al., 2019; Sun et al., 2021; Swamy et al., 2021; Jarboui & Perchet, 2021) use Intergral Probability Metrics (Sriperumbudur et al., 2009) (IPMs) as a distance measure to solve the dual problem. They introduce a discriminator and aim to find the saddle point of a min-max optimization problem, like GAN (Goodfellow et al., 2014). Jarrett et al. (2020) avoids the need of min-max problem by fixing the policy to be energy-based models, in such case the KL divergence from the demonstrator's state-action distribution to that of the policy can be computed in closed form. However, recent work finds several fundamental mathematical misconceptions in their proposed approach and we refer the reader to Swamy et al. (2021) for more details.

The common problem of these works is that they imitate equally to all demonstrations, this will hinder the performance if the demonstrations contain suboptimal data. To solve this, Sasaki & Yamashina (2021) reuses another policy learned by BC as the weight of original BC objective, however, this requires that expert data occupies the majority proportion of the offline dataset, otherwise the policy will be misguided to imitate the suboptimal data. Zolna et al. (2020) and Konyushkova et al. (2020) first construct a reward function that discriminates expert and exploratory trajectories, then use it to

solve an offline RL problem. However, offline IRL based on reward learning is extremely expensive to run, requiring solving offline RL in an inner loop, which itself is a challenging problem and prone to training instability (Kumar et al., 2019) and hyperparameter sensitivity (Wu et al., 2019a). Our algorithm can be seen as a combination of these two algorithms in that we use a generalized BC objective to imitate demonstrations selectively and we train a discriminator to distinguish expert and non-expert data and use the output of the discriminator as the weight of the generalized BC objective.

4.2 Offline Reinforcement Learning

One research area highly related to offline IL is offline RL (Lange et al., 2012; Levine et al., 2020), which considers performing effective RL by utilizing arbitrary given, static offline datasets, without any further environment interactions. Note that in offline RL, the training dataset is allowed to have non-optimal trajectories and the reward for each state-action-next state transition triple is known.

Our algorithm draws connection to a branch of methods in offline RL literature that performs "filtered" behavioral cloning explicitly or implicitly. More specifically, Peng et al. (2019), Nair et al. (2020) and Wang et al. (2020) estimate an advantage function, which represents the change in expected return when taking action a instead of following the current policy, and perform weighted regression based on the advantage function, defined as $\mathcal{L}_\pi = \mathbb{E}_{(s,a)\sim\mathcal{D}_b}\left[-\log \pi(a|s) \cdot f\left(A^\pi(s,a)\right)\right]$. The advantage A^π can be estimated by Monte-Carlo methods (Schulman et al., 2017; Peng et al., 2019) or Q-value based methods (Schulman et al., 2015; Nair et al., 2020). The filter function f can be a binary filter (Wang et al., 2020) or an exponential filter (Peng et al., 2019; Nair et al., 2020).

While Chen et al. (2021) and Janner et al. (2021) perform filtered behavioral cloning more implicitly. They abstract offline RL as a sequence modeling problem and use Transformer architecture (Vaswani et al., 2017) to perform credit assignments directly via self-attention. Owing to the memorization power of Transformer in capturing long-term dependencies across timesteps, these methods are able to discard low-quality transitions, do behavior cloning only on high-reward transitions across different trajectories and stitch them together to obtain an optimal trajectory.

5 Experiments

We present empirical evaluations of DWBC in a variety of settings. We start with describing our experimental setup, considered datasets and baseline algorithms. Then we evaluate DWBC against other baseline algorithms on a range of robotic locomotion tasks with different types of datasets. Finally, we analyze the property of the discriminator. We show that a well-trained discriminator can be used to perform offline policy selection (Fu et al., 2020b), which is of independent interest.

5.1 Settings

We conducted experiments on four MuJoCo continuous control tasks: `Ant-v2`, `Hopper-v2`, `HalfCheetah-v2` and `Walker2d-v2`. These tasks imitate scenarios encountered by robots in real life. The robot is composed of multiple joints, at each step the agent selects the amount of torque to apply to each joint.

In practice, the dataset $\mathcal{D}_e$ would be collected by an expert and $\mathcal{D}_o$ is meant to be built based on the logged interactions of the agent from the past. Similar to Zolna et al. (2020). In this paper, we build these datasets based on existing offline RL benchmark datasets. Given a logged dataset consists of multiple trajectories, we extract a small subset of well-performing trajectories to constitute demonstrations $\mathcal{D}_e$ and use the rest to constitute demonstrations $\mathcal{D}_o$. We use ground truth rewards only to perform this data split and discard the rewards afterward.

Table 1: **Datasets statistics.** The number of data transitions in $\mathcal{D}_e$ and $\mathcal{D}_o$ per dataset.

Dataset-X	$\#\boldsymbol{\mathcal{D}_e}$	$\#\boldsymbol{\mathcal{D}_o}$
Ant_mixed-2	46,646	254,869
Ant_mixed-5	19,209	282,306
Ant_mixed-10	9,866	29,1649
Hopper_mixed-2	96,222	303,737
Hopper_mixed-5	39,590	360,369
Hopper_mixed-10	19,176	380,783
Halfcheetah_mixed-2	19,980	181,818
Halfcheetah_mixed-5	7,992	193,806
Halfcheetah_mixed-10	3,996	197,802
Walker2d_mixed-2	74,857	226,050
Walker2d_mixed-5	31,010	269,897
Walker2d_mixed-10	15,569	285,338

We use the mixed datasets in the open D4RL benchmark datasets (Fu et al., 2020a). The mixed datasets are generated by recording the replay buffer of a policy trained up to the performance of a predefined threshold (usually 1/3 of the performance of an online SAC agent (Haarnoja et al., 2018)). We sort from high to low of all trajectories based on the total reward summed over the entire trajectory. We define a trajectory as well-performing if it is among the top 20% of all trajectories. We then sample every X^{th} trajectory from the well-performing trajectories to constitute $\mathcal{D}_e$ and use the remaining trajectories in the dataset to constitute $\mathcal{D}_o$. Note that with X becomes larger, $\mathcal{D}_o$ will contain more proportion of well-performing data.

Table 1 lists all datasets used in this paper and the number of transitions in $\mathcal{D}_e$ and $\mathcal{D}_o$, different X is labeled after the dataset name.

5.2 Baseline and ablated algorithms

We compare DWBC with the following baseline algorithms:

BC-pos: Behavioral cloning on positive (expert) data only. This baseline is trained only on $\mathcal{D}_e$, hence it is not exposed to low-performing transitions in $\mathcal{D}_o$. On the other hand, it may not be able to generalize due to the limited size of $\mathcal{D}_e$.

BC-all: Behavioral cloning on all data. BC-all may generalize better than BC-pos due to access to a much larger dataset, but its performance may be negatively impacted by the low-quality data in $\mathcal{D}_o$.

BCND-all: Behavioral Cloning from Noisy Demonstrations (Sasaki & Yamashina, 2021) on all data. BCND-all reuses another policy learned by BC as the weight of BC objective, but its performance will be worse if the suboptimal data occupies the major part of the offline dataset.

ORIL: Offline Reinforced Imitation Learning (Zolna et al., 2020) on $\mathcal{D}_e$ and $\mathcal{D}_o$. ORIL first learns a discriminator as the reward function, then uses it to solve an offline RL problem. It suffers from large computational costs and the difficulty of performing offline RL under distributional shift.

5.3 Comparative Evaluations

It can be shown from Figure 1 that DWBC outperforms the baseline algorithms on almost all the tasks (DWBC only falls behind in `ant-mixed-2`), showing that DWBC is well suited to make effective use of the mixed quality dataset $\mathcal{D}_o$.

As expected, the performance of BC-pos declines as X becomes larger. This is because that a larger X means the number of well-performing transitions is smaller. In some datasets (e.g., `halfcheetah-mixed-5`, `ant-mixed-10`, `walker2d-mixed-10`), there is no clear winner between BC-pos and BC-all, which suggests that the quality of $\mathcal{D}_o$ for the considered tasks varies. BCND performs poorly compared to other methods due to the majority of low-quality data in the mixed datasets. It usually scores below BC-all. ORIL struggles to learn in some tasks (especially in the ant datasets), which suggests their learned reward function does not accurately contrast expert and non-expert data. We also find that the performance of ORIL and DWBC tends to decrease in some tasks (e.g., `hopper-mixed-5`, `halfcheetah-mixed-5` and `walker-mixed-2`), this "overfitting" phenomenon also occurs in the experiments of offline RL papers (Kumar et al., 2019; Wu et al., 2019a). This is perhaps due to the limited data size and model generalization bottleneck (Neyshabur, 2017). This is a challenging open problem and one may need to use techniques from supervised learning (Lee et al., 2020; Kumar et al., 2020) to solve it.

5.4 Additional Experiments

Offline policy selection by the discriminator. Offline policy selection (OPS) (Paine et al., 2020; Yang et al., 2020; Dereventsov et al., 2021) considers the problem of choosing the best policy from a set of policies given only offline data. This problem is critical in the offline settings (i.e., offline RL and offline IL) because the online execution is often very costly and safety-aware, deploying a problematic policy may damage the real-world systems (Tang & Wiens, 2021). Note that existing offline RL/IL methods break the offline assumption by evaluating different policies corresponding to their rewards in online environment interactions. However, this online evaluation is often infeasible and hence undermines the initial assumption of offline RL/IL.

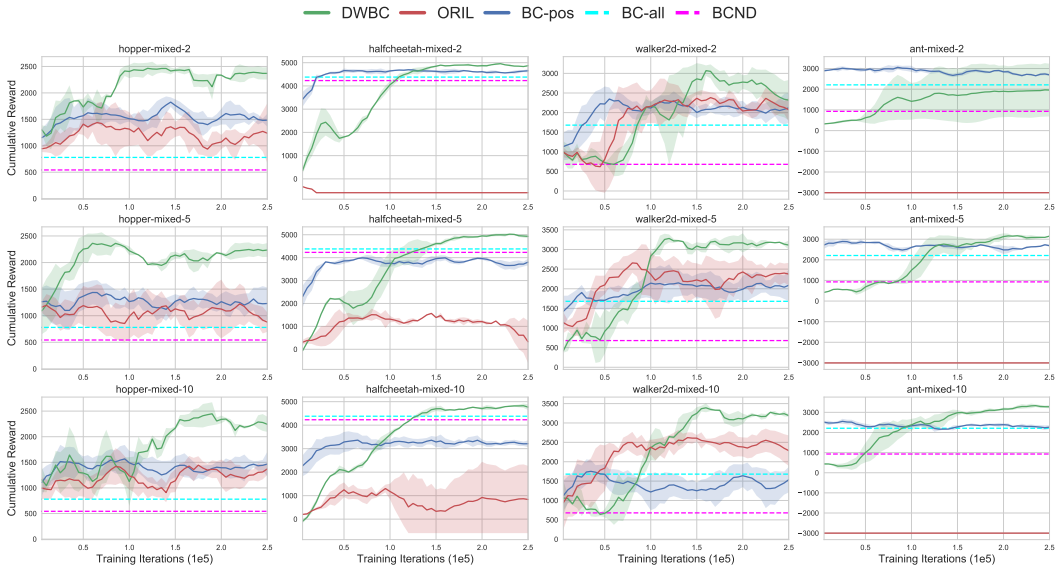

Figure 1: Learning curves comparing the performance of DWBC against other baseline algorithms in different datasets. Curves are averaged over 5 seeds, with the shaded area representing the standard deviation across seeds. DWBC achieves the highest results in almost every task.

In this paper, we propose a novel offline policy selection method, simply by using our learned discriminator d. More specifically, as the discriminator gets state s, action a and the log probability $\log \pi(a|s)$ as input, we can use expert state-action pairs from $\mathcal{D}_e$ and different policy π as input. The discriminator learned by our method will assign large values (close to 1) when the evaluated policy is close to the policy learned by DWBC, this means that the policy is closer to the optimal policy.

To validate our proposed idea, we conduct experiments in the `Hopper-v2` environment, we use the three datasets in previous environment, namely `Hopper-mixed-2`, `Hopper-mixed-5` and `Hopper-mixed-10`. We compare three algorithms (BC-pos, ORIL and DWBC) trained in these datasets, total of 9 policies. We use the datasets $\mathcal{D}_e$ and the discriminator learned by DWBC to compute the average output values of the discriminator for these 9 policies and plot their actual performance versus average discriminator output values. As shown in Figure 2, the discriminator output values can well reflect the rank between almost every two policies. This means that we can first train a DWBC policy and use the trained discriminator d to compare with other policies (trained by other algorithms).

Comparision of run time. We also evaluate the run time of training DWBC and other baseline algorithms for 250 thousand training steps. All run time experiments were executed on NVIDIA V100 GPUs. For a fair comparison, we use the same policy network size in BC, BCND, ORIL and DWBC. The discriminator network size is also kept the same in ORIL and DWBC. The results are reported in Figure 2. Unsurprisingly, we find the run time of our approach is only slightly more than BC, while other baselines (ORIL, BCND) are over 7 times more costly than BC. The reason that ORIL is costly to run is due to the additional effort to solve an offline RL problem. The high computation cost of BCND is due to its inner iterations of training K policy ensembles ($K = 5$ in our experiment), which is also mentioned in their paper (Sasaki & Yamashina, 2021). This demonstrates the effectiveness of DWBC by only adds a limited cost to the original BC algorithm while providing substantially improved performance.

6 Conclusion and Limitation

In this paper, we propose a new offline imitation learning algorithm that can learn from suboptimal demonstrations without environment interactions or expert annotations. Experimental results show

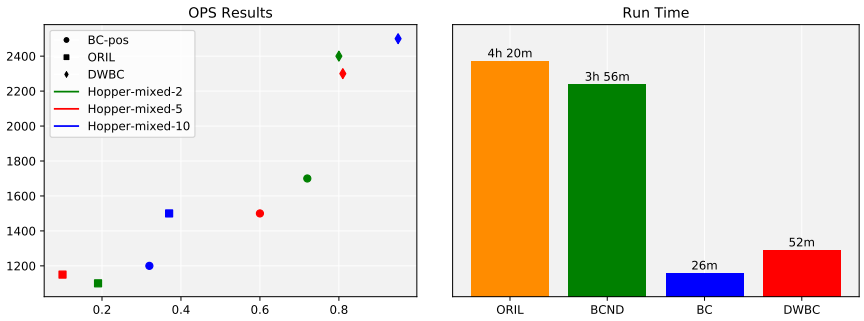

Figure 2: Additional experiments. Left: the results of offline policy selection by the discriminator learned by DWBC. Right: run time comparison of training each offline IL algorithm (does not include evaluation run time cost).

that our algorithm can learn behavior policies that are much closer to the true expert policies than ones learned by baseline algorithms, along with less wall-clock training time. Since our algorithm copes well with the noisy expert demonstrations and performs high-fidelity imitation in an offline, data-driven manner, our algorithm is more applicable to real-world problems than prior works. One limitation of our work is that since our algorithm is based on weighted BC, the covariate shift problem of BC (Ross et al., 2011) will be inherited. That is, there is no way for the policy to learn how to recover if it deviates from the behavior policy to a new state not seen in the demonstrations. In future work, we will consider modifying the main task from action matching to state-action distribution matching, which is known to be more robust to distributional shift (Kostrikov et al., 2019).

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pp. 783–792. PMLR, 2019.

Daniel S Brown, Wonjoon Goo, and Scott Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on robot learning*, pp. 330–359. PMLR, 2020.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *arXiv preprint arXiv:2106.01345*, 2021.

Thomas Degris, Martha White, and Richard S Sutton. Off-policy actor-critic. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 179–186, 2012.

Anton Dereventsov, Joseph D Daws Jr, and Clayton Webster. Offline policy comparison under limited historical agent-environment interactions. *arXiv preprint arXiv:2106.03934*, 2021.

Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pp. 1386–1394. PMLR, 2015.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–220, 2008.

Pete Florence, Corey Lynch, Andy Zeng, Oscar Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. *arXiv preprint arXiv:2109.00137*, 2021.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020a.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Alexander Novikov, Mengjiao Yang, Michael R Zhang, Yutian Chen, Aviral Kumar, Cosmin Paduraru, et al. Benchmarks for deep off-policy evaluation. In *International Conference on Learning Representations*, 2020b.

Izrail Moiseevitch Gelfand, Richard A Silverman, et al. *Calculus of variations*. Courier Corporation, 2000.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870, 2018.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29:4565–4573, 2016.

Alex Irpan. Deep reinforcement learning doesn't work yet. `https://www.alexirpan.com/2018/02/14/rl-hard.html`, 2018.

Michael Janner, Qiyang Li, and Sergey Levine. Reinforcement learning as one big sequence modeling problem. *arXiv preprint arXiv:2106.02039*, 2021.

Firas Jarboui and Vianney Perchet. Offline inverse reinforcement learning. *arXiv preprint arXiv:2106.05068*, 2021.

Daniel Jarrett, Ioana Bica, and Mihaela van der Schaar. Strictly batch imitation learning by energy-based distribution matching. *arXiv preprint arXiv:2006.14154*, 2020.

Ksenia Konyushkova, Konrad Zolna, Yusuf Aytar, Alexander Novikov, Scott Reed, Serkan Cabi, and Nando de Freitas. Semi-supervised reward learning for offline reinforcement learning. *arXiv preprint arXiv:2012.06899*, 2020.

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. In *International Conference on Learning Representations*, 2019.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pp. 11761–11771, 2019.

Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, and Sergey Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *International Conference on Learning Representations*, 2020.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Byungjun Lee, Jongmin Lee, Peter Vrancx, Dongho Kim, and Kee-Eung Kim. Batch reinforcement learning with hyperparameter gradients. In *International Conference on Machine Learning*, pp. 5725–5735. PMLR, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Charles X Ling and Victor S Sheng. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning*, 2011:231–235, 2008.

Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 641–647, 2005.

Yecheng Jason Ma. *From Adversarial Imitation Learning to Robust Batch Imitation Learning*. PhD thesis, 2020.

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

Behnam Neyshabur. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287, 1999.

Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Tom Le Paine, Cosmin Paduraru, Andrea Michi, Caglar Gulcehre, Konrad Zolna, Alexander Novikov, Ziyu Wang, and Nando de Freitas. Hyperparameter selection for offline reinforcement learning. *arXiv preprint arXiv:2007.09055*, 2020.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Advances in neural information processing systems*, pp. 305–313, 1989.

Siddharth Reddy, Anca D Dragan, and Sergey Levine. Sqil: Imitation learning via reinforcement learning with sparse rewards. *arXiv preprint arXiv:1905.11108*, 2019.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635, 2011.

Fumihiro Sasaki and Ryota Yamashina. Behavioral cloning from noisy demonstrations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=zrT3HcsWSAt`.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Mingfei Sun, Anuj Mahajan, Katja Hofmann, and Shimon Whiteson. Softdice for imitation learning: Rethinking off-policy distribution matching. *arXiv preprint arXiv:2106.03155*, 2021.

Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

Gokul Swamy, Sanjiban Choudhury, Zhiwei Steven Wu, and J Andrew Bagnell. Of moments and matching: Trade-offs and treatments in imitation learning. *arXiv preprint arXiv:2103.03236*, 2021.

Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. *arXiv preprint arXiv:2107.11003*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Ziyu Wang, Alexander Novikov, Konrad Zolna, Josh S Merel, Jost Tobias Springenberg, Scott E Reed, Bobak Shahriari, Noah Siegel, Caglar Gulcehre, Nicolas Heess, et al. Critic regularized regression. *Advances in Neural Information Processing Systems*, 33, 2020.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019a.

Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, and Masashi Sugiyama. Imitation learning from imperfect demonstration. In *International Conference on Machine Learning*, pp. 6818–6827. PMLR, 2019b.

Mengjiao Yang, Bo Dai, Ofir Nachum, George Tucker, and Dale Schuurmans. Offline policy selection under uncertainty. *arXiv preprint arXiv:2012.06919*, 2020.

Konrad Zolna, Alexander Novikov, Ksenia Konyushkova, Caglar Gulcehre, Ziyu Wang, Yusuf Aytar, Misha Denil, Nando de Freitas, and Scott Reed. Offline learning from demonstrations and unlabeled experience. *arXiv preprint arXiv:2011.13885*, 2020.

A IMPLEMENTATION DETAILS

In this paper, all experiments are implemented with Tensorflow and executed on NVIDIA V100 GPUs. For all function approximators, we use fully connected neural networks with RELU activations. For policy networks, we use tanh (Gaussian) on outputs. We use Adam for all optimizers. The batch size is 256 and γ is 0.99. We rescale the reward to $[0, 1]$ as $r' = (r - r_{\min}) \,/\, (r_{\max} - r_{\min})$, where $r_{\max}$ and $r_{\min}$ is the maximum and the minimum reward in the dataset. Note that any affine transformation of the reward function does not change the optimal policy of the MDP. The policy and discriminator network are all 3-layer MLP with 256 hidden units in each layer. The learning rate for the policy is $1e-5$ and the learning rate for the discriminator network is $1e-4$. We search α in $\{1, 2, 5, 10\}$ for best model performance. We clip the output of d to $[0.1, 0.9]$. We set η to 0.5 across all tasks, which is the same as the ORIL paper (Zolna et al., 2020).

B DERIVATION DETAILS OF SECTION 3.2

In this section, we provide the detailed theoretical derivation of the impact of BC policy π on the functional $J(\pi)$ defined in Eq.(6).

Proposition 1. *With the parameters of the discriminator d fixed, considering the functional $J(\pi)$ for the discriminator loss $\mathcal{L}_d$ with respect to π defined in the following integral form:*

$$J(\pi) = \int\int \frac{\partial \mathcal{L}_d(s, a, d, \log \pi)}{\partial d(s, a, \log \pi)} \, ds \, da = \int\int F(s, a, \pi, \pi') \, ds \, da,$$

where $F = \partial \mathcal{L}_d / \partial d$ is assumed to be a continuously differentiable function with respect to s, a, π and π'. $J(\pi)$ attains its extrema when $\partial F(s, a, \pi, \pi')/\partial \theta = 0$, and θ is the parameters of π.

Proof. We derive the finite difference approximation of the derivative of $J(\pi)$ by Taylor expansion,

$$J(\pi) = J(\pi_{\theta_0}) + J'(\pi_{\theta_0})\delta\pi + o(\delta\pi), \tag{10}$$

where θ is the parameter of π and $\delta\pi$ is the variation of π. As $o(\delta\pi)$ is a higher order infinitesmall term when $\theta \to \theta_0$, which can be ignored. Plug Eq.(10) into the form of $J(\pi)$, we obtain the finite difference approximation of $J(\pi)$ as follows:

$$\Delta J(\pi) = J(\pi) - J(\pi_{\theta_0}) = J'(\pi_{\theta_0})\delta\pi = \frac{d \int\int F(s, a, \pi, \pi') \, ds \, da}{d\theta} \delta\pi. \tag{11}$$

If the uniform continuity of both F and its derivative is satisfied, we can swap the order integration and differentiation as well as $ds \, da$ and $\delta\pi$. The above equation can be simplified as:

$$\Delta J(\pi) = \frac{\int\int \partial F(s, a, \pi, \pi') \, ds \, da}{\partial \theta} \delta\pi = \int\int \frac{\partial F(s, a, \pi, \pi')}{\partial \theta} \delta\pi \, ds \, da. \tag{12}$$

According to the calculus of variations (Gelfand et al., 2000), the extrema of functional $J(\pi)$ can be obtained by finding a function π where the functional derivative is equal to zero. As π can be any function in some function space $\mathcal{F}$. Ensuring the functional derivative of $J(\pi)$ equal to 0 also requires $\partial F(s, a, \pi, \pi')/\partial \theta = 0$. □

In the follows, we derive the detailed form of $\partial F / \partial \theta$ as presented in Eq.(7) of the main article:

$$\begin{aligned}
\frac{\partial F(s, a, \pi, \pi')}{\partial \theta} &= \frac{\partial F(s, a, \pi, \pi')}{\partial \log \pi} \cdot \frac{\partial \log \pi}{\partial \pi} \cdot \frac{\partial \pi}{\partial \theta} = \frac{\partial F(s, a, \pi, \pi')}{\partial \log \pi} \cdot \nabla_\theta \log \pi \\
&= \frac{\partial \mathcal{L}_d(s, a, \log \pi)}{\partial d} \cdot \frac{\partial d}{\partial \log \pi} \cdot \nabla_\theta \log \pi \\
&= -\mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e} \left[\frac{\eta}{d} \cdot \nabla_\theta \log \pi(a|s) \right] + \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_o} \left[\frac{1}{1-d} \cdot \nabla_\theta \log \pi(a|s) \right] \\
&\quad - \mathop{\mathbb{E}}_{(s,a)\sim\mathcal{D}_e} \left[\frac{\eta}{1-d} \cdot \nabla_\theta \log \pi(a|s) \right],
\end{aligned}$$