Efficient Inverse Reinforcement Learning without Compounding Errors

Nicolas Espinosa Dice [1] Gokul Swamy [2] Sanjiban Choudhury [1] Wen Sun [1]

Abstract

Inverse reinforcement learning (IRL) is an on-policy approach to imitation learning (IL) that allows the learner to observe the consequences of their actions at train-time. Accordingly, there are two seemingly contradictory desiderata for IRL algorithms: *(a)* preventing the compounding errors that stymie offline approaches like behavioral cloning and *(b)* avoiding the worst-case exploration complexity of reinforcement learning (RL). Prior work has been able to achieve either *(a)* or *(b)* but not both simultaneously. In our work, we first prove a negative result showing that, without further assumptions, there are no efficient IRL algorithms that avoid compounding errors in the worst case. We then provide a positive result: under a novel structural condition we term *reward-agnostic policy completeness*, we prove that efficient IRL algorithms *do* avoid compounding errors, giving us the best of both worlds. We also propose a principled method for using sub-optimal data to further improve the sample-efficiency of efficient IRL algorithms.

1. Introduction

Inverse reinforcement learning (IRL) is an on-policy approach to imitation learning that involves simultaneously learning a reward function from expert demonstrations and learning a policy that optimizes the learned reward (Ziebart et al., 2008a). IRL has been applied to a diverse set of applications, including robotics (Ratliff et al., 2007; Abbeel & Ng, 2008; Ratliff et al., 2009; Silver et al., 2010; Zucker et al., 2011), autonomous driving (Bronstein et al., 2022; Igl et al., 2022; Vinitsky et al., 2022), and route finding (Ziebart et al., 2008a;b; Barnes et al., 2023).

Compared to offline imitation learning methods such as behavior cloning, IRL offers the following advantages. First, IRL is more sample efficient, with respect to expert samples, than behavior cloning (Swamy et al., 2021; 2022). Second, IRL offers better error scaling, with respect to the horizon, than behavior cloning (Ross & Bagnell, 2010; Swamy et al., 2021; 2022). Unlike behavior cloning, IRL is capable of avoiding quadratically compounding errors in the horizon (Ross & Bagnell, 2010; Swamy et al., 2021).

However, the expert sample efficiency of traditional IRL methods comes at the cost of environment interactions. Traditional IRL methods can require an exponential number of environment interactions in the worst case (Swamy et al., 2023). Because the reward function and policy are learned simultaneously, IRL requires policy optimization to be performed repeatedly, making it susceptible to RL's worst-case exploration complexity (Swamy et al., 2023). In order to focus the exploration on useful states, prior work has leveraged the expert's state distribution for learner resets, resulting in an exponential speedup in interaction complexity (Swamy et al., 2023).

Unfortunately, the improvement of efficient IRL's interaction efficiency sacrifices traditional IRL's linear error scaling. Swamy et al. (2023)'s Moment Matching by Dynamic Programming (MMDP) and No-Regret Moment Matching (NRMM) are exponentially faster than traditional IRL algorithms, but they suffer from quadratically compounding errors in the horizon.

Based on the prior work, it seems that two desiderata of IRL – interaction efficiency and avoidance of compounding errors – are contradictory, with algorithms only being able to attain one or the other. Our key insight is that the commonly imposed assumption of *expert realizability* (i.e. the expert policy is within the learner's policy class) is insufficient to address both interaction efficiency and error scaling. In our paper, we introduce a novel structural condition, *reward-agnostic policy completeness*, under which IRL can both be efficient and avoid compounding errors.

More explicitly, our contributions are as follows:

1. We first consider the *agnostic* setting, where no assumptions are made about the MDP's structure, and present a lower bound that shows it is impossible to learn a competitive policy with polynomial environment

[1] Department of Computer Science, Cornell University, Ithaca, NY, USA [2] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. Correspondence to: Nicolas Espinosa Dice <ne229@cornell.edu>.

ICML 2024 Workshop on Models of Human Feedback for AI Alignment, Vienna, Austria. Copyright 2024 by the author(s).

interaction complexity in the worst case. In other words, efficient IRL is not possible without assuming additional structure on the MDP.

2. We define a new structural condition, *reward-agnostic policy completeness*, under which our efficient, reset-based IRL algorithm is capable of avoiding quadratically compounding errors. Importantly, our analysis holds for *approximate* policy completeness, and the optimal (i.e. expert) policy does not have to be in the policy class.

3. We extend our algorithm to incorporate sub-optimal data. We show that the benefits of incorporating sub-optimal data are a function of the quantity of data and how well the sub-optimal data *covers* the expert data. Our theoretical results are aligned in the intuition that suggests the greater the overlap between sub-optimal and expert states, the more beneficial to learning the sub-optimal data is.

2. Related Work

Prior work in reinforcement learning (RL) has examined leveraging exploration distributions to improve learning (Kakade & Langford, 2002; Bagnell et al., 2003; Ross et al., 2011). We adapt the Policy Search via Dynamic Programming (PSDP) algorithm of Bagnell et al. (2003) as our RL solver and leverage its performance guarantees in our analysis. Our policy completeness error is inspired by Agarwal et al. (2019)'s adapted analysis of Kakade & Langford (2002)'s Conservative Policy Iteration (CPI) algorithm. Our paper also builds on work in agnostic RL. Jia et al. (2024) analyze the conditions for which agnostic RL is statistically tractable. We use Jia et al. (2024)'s lower bound on agnostic RL with expert feedback to show why agnostic IRL is hard.

Our work examines the issue of distribution shift due to compounding errors in IRL, which was introduced by Ross & Bagnell (2010). Ross et al. (2011)'s DAgger algorithm is capable of avoiding compounding errors but requires an interactive expert, which we do not assume in our setting.

We incorporate Swamy et al. (2023)'s novel approach of leveraging the expert's state distribution for learner resets. Our algorithm builds upon Swamy et al. (2023)'s MMDP and NRMM algorithms by avoiding quadratically compounding error in the horizon.

Our algorithm and results are not limited to the tabular and linear MDP settings, differentiating from some prior work in efficient imitation learning (Xu et al., 2023; Viano et al., 2024). Our work also relates to (Shani et al., 2022), who propose a mirror descent based no-regret algorithm for online apprenticeship learning (OAL). We similarly use a mirror descent based update to our reward function, but differ from Shani et al. (2022)'s work by leveraging resets to expert and sub-optimal data to improve the interaction efficiency of our algorithm.

Poiani et al. (2024) propose a technique of incorporating sub-optimal experts as a means of addressing the ambiguity in IRL problems, specifically the lack of uniqueness in reward functions that rationalize the observed behavior. Our work contrasts Poiani et al. (2024)'s because we do not use sub-optimal data in learning a reward function, instead using it to improve policy optimization training.

3. Setup and Motivation

3.1. Problem Setup

Markov Decision Process We consider a finite-horizon Markov Decision Process (MDP), $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P_h, r^*, H, \mu \rangle$. $\mathcal{S}$ and $\mathcal{A}$ are the state space and action space, respectively. $P = \{P_h\}_{h=1}^H$ is the time-dependent transition function, where $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$. $r^* : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the ground-truth reward function, which is unknown. Let $\mathcal{R}$ be the class of reward functions, such that $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ for all $r \in \mathcal{R}$. H is the horizon, and $\mu \in \Delta(\mathcal{S})$ is the starting state distribution. Let $\Pi = \{\pi : \mathcal{S} \to \Delta(\mathcal{A})\}$ be the class of stationary policies. Let the class of non-stationary policies be defined by $\Pi^H = \{\pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}_{h=1}^H$. A trajectory is given by $\tau = \{(s_h, a_h, r_h)\}_{h=1}^H$, where $s_h \in \mathcal{S}, a_h \in \mathcal{A}$, and $r_h = f(s_h, a_h)$ for some $f \in \mathcal{R}$. The distribution over trajectories formed by a policy is given by: $a_h \sim \pi(\cdot \mid s_h)$, $r_h = R_h(s_h, a_h)$, and $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$, for $h = 1, \dots, H$. Let $d^\pi_{s_0,h}(s) = \mathbb{P}^\pi[s_h = s \mid s_0]$ and $d^\pi_{s_0}(s) = \frac{1}{H}\sum_{h=1}^H d^\pi_{s_0,h}(s)$. Overloading notation slightly, we have $d^\pi_\mu = \mathbb{E}_{s_0\sim\mu}\, d^\pi_{s_0}$.

We index the value function by the reward function, such that for any $\pi \in \Pi^H$ and $r \in \mathcal{R}$, $V^\pi_{r,h}(s) := \mathbb{E}_{\tau\sim\pi}\left[\sum_{h'=h}^H r_{h'} \mid s_h = s\right]$, and $V^\pi_r = \mathbb{E}_{\tau\sim\pi}\sum_{h=1}^H r(s_h, a_h)$. We do a corresponding indexing for the advantage function. We will overload notation such that a state-action pair can be sampled from the visitation distributions, e.g. $(s,a) \sim d^\pi_\mu$ and $(s,a) \sim \rho_E$, as well as a state, e.g. $s \sim d^\pi_\mu$ and $s \sim \rho_E$. Note that by definition of d^π_μ, $\mathbb{E}_{\tau\sim\pi}\left[\sum_{h=1}^H r(s_t, a_t)\right] = H\,\mathbb{E}_{(s,a)\sim d^\pi_\mu}[r(s,a)]$.

Expert Data There exists an expert policy π_E, of which a sample of its trajectories are known. The dataset of state-action pairs sampled from the expert is $D_E = D_1 \cup D_2 \cup \ldots \cup D_H$, where $D_h = \{s_h, a_h\} \sim d^{\pi_E}_{\mu,h}$ and $|D_E| = N$. Let ρ_h be a uniform distribution over the samples in D_h, and ρ_E be a uniform distribution over the samples in D_E.

Goal of IRL We adopt the formulation of Swamy et al. (2021), casting IRL as a Nash equilibrium problem. The goal is to find a policy π such that

$$\min_{\pi \in \Pi} \max_{r \in \mathcal{R}} J(\pi_E, r) - J(\pi, r),$$

where $J(\pi, r) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{T} r(s_t, a_t)\right]$.

3.2. IRL in the Agnostic Setting

We first consider IRL in the agnostic setting, where no assumptions are made about the MDP's structure, the policy class, or the expert's policy (i.e. we do not assume $\pi_E \in \Pi^H$). We restate Theorem 9 from Jia et al. (2024).

Theorem 3.1 (Lower Bound on Agnostic RL with Expert Feedback (Jia et al., 2024))**.** *For any $H \in \mathbb{N}$ and $C \in [2^H]$, there exists a policy class Π with $|\Pi| = C$, expert policy $\pi_E \notin \Pi$, and a family of MDPs $\mathcal{M}$ with state space $\mathcal{S}$ of size $O(2^H)$, binary action space, and horizon H such that any algorithm that returns a 1/4-optimal policy must either use $\Omega(C)$ queries to a generative model or $\Omega(C)$ queries to the expert oracle $O_{exp} : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$, which returns $Q^{\pi_E}(s, a)$ (i.e. the Q value of expert policy π_E).*

Theorem 3.1 presents a lower bound on agnostic RL with expert feedback. Specifically, it assumes access to the true reward function and an expert oracle, $O_{\text{exp}} : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$, which returns $Q^{\pi_E}(s, a)$ for a given state-action pair (s, a). The lower bound in Theorem 3.1 applies in the case where the expert oracle is replaced with a weaker expert action oracle (i.e. $\pi_E(s) : \mathcal{S} \to \mathcal{A}$) (Amortila et al., 2022; Jia et al., 2024). In agnostic IRL, we consider the even weaker setting of having a dataset of state-action pairs from the expert policy π_E. From Theorem 3.1, we can infer that polynomial sample complexity in the agnostic IRL setting is not possible in the worst case.

It should be noted that the classical importance sampling (IS) algorithm (Kearns et al., 1999) can be employed to find an approximately optimal policy in the agnostic setting, but it requires an exponential number of interactions (Agarwal et al., 2019; Jia et al., 2024).

4. Policy Complete Inverse Reinforcement Learning

Theorem 3.1 establishes a lower bound in the agnostic setting, where no assumptions are made about the MDP or expert policy. It naturally motivates the question,

Under what conditions is it possible for efficient IRL algorithms to avoid quadratically compounding errors?

Expert realizability was assumed by Swamy et al. (2023)'s efficient IRL algorithms but fails to avoid compounding errors.

We introduce *reward-agnostic policy completeness error* to specify the conditions under which compounding errors can be avoided efficiently. Policy completeness error can be thought of as measuring the policy class's ability to approximate the maximum possible advantage over the expert's state distribution under any reward function in the reward class.

Definition 4.1 (Reward-Indexed Policy Completeness Error)**.** Given some expert state distribution ρ_E, MDP $\mathcal{M}$ with policy class Π and reward class $\mathcal{R}$, learned policy π_i, and learned reward function r_i, define the reward-indexed policy completeness error of $\mathcal{M}$ to be

$$\epsilon_\Pi^{\pi_i, r_i} := \mathop{\mathbb{E}}_{s \sim \rho_E}\left[\max_{a \in \mathcal{A}} A_{r_i}^{\pi_i}(s, a)\right] - \max_{\pi' \in \Pi} \mathop{\mathbb{E}}_{s \sim \rho_E} \mathop{\mathbb{E}}_{a \sim \pi'(\cdot|s)} \left[A_{r_i}^{\pi_i}(s, a)\right].$$

We first present the reward-indexed policy completeness error in Definition 4.1, where π_i and r_i represent the learned policy and reward, respectively, from iteration i of a generic IRL algorithm. Our definition of reward-indexed policy completeness error is inspired by one used in Agarwal et al. (2019)'s adapted analysis of CPI, extended to the IRL setting. Notably, our definition is distinct in using the expert's state distribution rather than the learner's.

The reward-indexed policy completeness error measures how well the policy class can approximate the advantage of optimal actions over policy π_i under reward r_i. Because there do not exist strong guarantees on how closely r_i will resemble the true reward r^* during early iterations of an IRL algorithm, the expert policy may not be optimal under r_i. We consider a maximum over all actions to determine the maximum possible advantage over policy π_i, i.e. $\max_{a \in \mathcal{A}} A_{r_i}^{\pi_i}(s, a)$. In the worst case, where the policy class is poorly restricted under the expert's state distribution, then $\epsilon_\Pi = H$, due to the bound on the reward function.

In order to extend the definition to other policies and reward functions learned at separate iterations, we pessimistically consider the worst case over all possible policies and rewards, leading to Definition 4.2. Note that $0 \leq \epsilon_\Pi^{\pi_i, r_i} \leq \epsilon_\Pi \leq H$ for any $\pi_i \in \Pi$, $r_i \in \mathcal{R}$.

Definition 4.2 (Reward-Agnostic Policy Completeness Error)**.** Given some expert state distribution ρ_E and MDP $\mathcal{M}$ with policy class Π and reward class $\mathcal{R}$, define the reward-agnostic policy completeness error of $\mathcal{M}$ to be

$$\begin{aligned}\epsilon_\Pi &:= \max_{\pi \in \Pi, r \in \mathcal{R}} \epsilon_\Pi^{\pi, r} \\ &= \max_{\pi \in \Pi, r \in \mathcal{R}} \left(\mathop{\mathbb{E}}_{s \sim \rho_E}\left[\max_{a \in \mathcal{A}} A_r^\pi(s, a)\right] - \max_{\pi' \in \Pi} \mathop{\mathbb{E}}_{s \sim \rho_E} \mathop{\mathbb{E}}_{a \sim \pi'(\cdot|s)} \left[A_r^\pi(s, a)\right] \right).\end{aligned}$$

Algorithm 1 Policy Search Via Dynamic Programming (Bagnell et al., 2003)

```
Input: $H$, expert state distribution $\rho_E$ and its time indexed components, reward function $r_i$, and policy class $\Pi$
Output: Trained policy $\pi$
for $h = H, H-1, \ldots, 1$ do
  Optimize
    $\pi_h \leftarrow \operatorname{argmax}_{\pi' \in \Pi} \mathbb{E}_{s \sim \rho_{E,h}} \mathbb{E}_{a \sim \pi'(\cdot|s)} A_{r_i}^{\pi_{h+1},\ldots,\pi_H}(s,a)$
end for
Return $\pi = \{\pi_h\}_{h=1}^H$
```

4.1. Efficient IRL Under Approximate Policy Completeness

We present MMDP-SR (Moment Matching by Dynamic Programming: Sub-optimal Reset), an efficient IRL algorithm that can be considered a variant of Swamy et al. (2023)'s MMDP algorithm. MMDP-SR can incorporate sub-optimal data resets, which we describe in Section 5. We analyze its sample complexity in the approximate policy completeness setting.

Following Swamy et al. (2021)'s classification of IRL algorithms, we propose an efficient dual variant algorithm, where the discriminator is updated via a no-regret step, and the policy is updated via a best-response step. We employ online mirror descent for the discriminator update, such that our reward function is updated via

$$r_i \leftarrow \operatorname*{argmax}_{r \in \mathcal{R}} \hat{L}(\pi_{i-1}, r) + \eta^{-1} \Delta_R(r \mid r_{i-1}),$$

where Δ_R is the Bregman divergence with respect to the negative entropy function R. $\hat{L}(\pi, r)$ is the loss, defined by

$$\hat{L}(\pi, r) = \mathop{\mathbb{E}}_{(s,a) \sim \rho_E} r(s,a) - \mathop{\mathbb{E}}_{(s,a) \sim d_\mu^\pi} r(s,a),$$

with respect to the distribution of expert samples, ρ_E. Importantly, for our analysis, we assume that the ground-truth reward function is realizable such that $r^* \in \mathcal{R}$. An interesting direction of future work is extending our analysis to the case of a non-realizable reward.

We employ Bagnell et al. (2003)'s PSDP algorithm, shown in Algorithm 1, for the policy update step. We use the distribution of expert samples, ρ_E, as the distribution for resets. The IRL procedure is outlined in Algorithm 2.

Algorithm 2 MMDP-SR (Moment Matching by Dynamic Programming: Sub-optimal Reset)

```
Input: Expert state visitation distributions $\rho_E$, policy class $\Pi$, reward class $\mathcal{R}$
Output: Trained policy $\pi$
Set $\pi_0 \in \Pi$
for $i = 1$ to $N$ do
  Let
    $\hat{L}(\pi, r) = \mathbb{E}_{(s,a) \sim \rho_E} r(s,a) - \mathbb{E}_{(s,a) \sim d_\mu^\pi} r(s,a)$
  Optimize
    $r_i \leftarrow \operatorname{argmax}_{r \in \mathcal{R}} \hat{L}(\pi_{i-1}, r) + \eta^{-1} \Delta_R(r \mid r_{i-1}).$
  Optimize
    $\pi_i \leftarrow \text{PSDP}(r_i)$
end for
Return $\pi_i$ with lowest validation error
```

4.2. Analysis in the Infinite-Sample Regime

Theorem 4.3 (Sample Complexity of Algorithm 2)**.** *Consider the case of infinite expert data samples. If $\pi_i = (\pi_{i,1}, \pi_{i,2}, \ldots, \pi_{i,H})$ is the policy returned by ϵ-approximate PSDP at iteration $i \in [n]$ of Algorithm 2 and $\rho_E = d_\mu^{\pi_E}$, then*

$$V^{\pi_E} - V^{\overline{\pi}} \leq H^2 \epsilon + H \epsilon_\Pi + H \sqrt{\frac{\ln |\mathcal{R}|}{n}},$$

where H is the horizon, n is the number of outer-loop iterations of the algorithm, and $\overline{\pi}$ is the average of the learned policies, π_i at each iteration $i \in [n]$.

The sample complexity of Algorithm 2 in the infinite expert sample regime is given in Theorem 4.3. The error is comprised of three terms. The first term, $H^2\epsilon$, stems from the policy optimization error of PSDP. It can be mitigated be improving the accuracy parameter ϵ of PSDP. Set to $\epsilon = \frac{1}{H}$, the term is reduced to linear error in the horizon H. This error can be interpreted as representing a tradeoff between environment interactions (i.e. computation) and error.

The second term, $H\epsilon_\Pi$, stems from the richness of the policy class. In the worst case where the policy class cannot approximate the maximum advantage, $\epsilon_\Pi = H$, resulting in quadratically compounding errors. Unlike the policy optimization error, the policy completeness error cannot be reduced with more environment interactions. Instead, it represents a fixed error that is a property of the MDP, the policy class, and the reward class. Under the approximate policy completeness setting, we assume $\epsilon_\Pi = O(1)$, reducing the error to linear in the horizon.

Finally, the last term $H\sqrt{\frac{\ln|\mathcal{R}|}{n}}$ stems from the regret of the online mirror descent update to the reward function. Assuming approximate policy completeness, such that $\epsilon_\Pi =$

$O(1)$, Theorem 4.3 shows that quadratically compounding errors in the horizon can be avoided by setting a small accuracy parameter ϵ in the PSDP procedure. The finite sample analysis of Algorithm 2 is provided in Section 5, where we also incorporate sub-optimal data.

5. Leveraging Sub-Optimal Data in IRL

5.1. Resetting to Sub-Optimal Data

In addition to the expert dataset, we also consider the case where we have an offline dataset $D_{\text{off}} = \{s_i, a_i\}_{i=1}^{M}$, where $(s,a) \sim d_\mu^{\pi_b}$ and π_b is some behavior policy that is not necessarily as a high-quality as the expert π_E. We measure the overlap of π_b to the expert π_E using the standard concentrability coefficient: $C_b = \left\| \frac{d_\mu^{\pi_E}}{d_\mu^{\pi_b}} \right\|_\infty$. We will show that we can gain benefit of using D_{off} as long as $C_b < \infty$ and the number of offline data points M is large.

Let us define $D_{\text{mix}} = D_E \cup D_{\text{off}}$ and ρ_{mix} as the uniform distribution over D_{mix}. We will use ρ_{mix} as the reset distribution for policy optimization. Let

$$\nu = \frac{N}{N+M} d_\mu^{\pi_E} + \frac{M}{N+M} d_\mu^{\pi_b}.$$

We only incorporate sub-optimal data for the policy optimization step. Using sub-optimal data for the reward update may lead to learning a reward function that values sub-optimal behavior as optimal, so the reward update remains the same as (2). Instead, we incorporate the sub-optimal for the policy optimization step, specifically resetting to the mixture of sub-optimal and expert states. Our replacement for policy optimization step (1) becomes

$$\pi_h \leftarrow \operatorname*{argmax}_{\pi' \in \Pi} \mathop{\mathbb{E}}_{s \sim \rho_{\text{mix},h}} \mathop{\mathbb{E}}_{a \sim \pi'(\cdot|s)} A_{r_i}^{\pi_{h+1},\ldots,\pi_H}(s,a).$$

5.2. Analysis in the Finite-Sample Regime

Lemma 5.1 (Advantage Bound)**.** *Suppose that $\epsilon = 0$ and reward function r_i are the input parameters to PSDP, and $\pi_i = (\pi_1^i, \pi_2^i, \ldots, \pi_H^i)$ is the output learned policy. Then, with probability at least $1-\delta$,*

$$\mathop{\mathbb{E}}_{s \sim d^{\pi_E}} \max_{a \in \mathcal{A}} A^{\pi_i}(s,a) \le \min \Bigg\{ \epsilon_\Pi + \epsilon_\Pi \sqrt{\frac{C_0}{N}}, \\ C_b \left(\epsilon_\Pi + \epsilon_\Pi \sqrt{\frac{C_0}{N+M}} \right) \Bigg\}$$

where $C_b = \left\| \frac{d_\mu^{\pi_E}}{d_\mu^{\pi_b}} \right\|_\infty$, H is the horizon, N is the number of expert state-action pairs, M is the number of offline state-action pairs, and $C_0 = 2 \ln \frac{|\Pi||\mathcal{R}|}{\delta}$.

Theorem 5.2 (Sample Complexity of Algorithm 2)**.** *Suppose that PSDP's accuracy parameter is set to $\epsilon = 0$, meaning we assume access to infinite computations of PSDP. Then, upon termination of Algorithm 2 with policy optimization step (5.1), with probability at least $1-\delta$, we have*

$$V^{\pi_E} - V^{\overline{\pi}} \le H \min \Bigg\{ \epsilon_\Pi + \epsilon_\Pi \sqrt{\frac{C_0}{N}}, \\ C_b \left(\epsilon_\Pi + \epsilon_\Pi \sqrt{\frac{C_0}{N+M}} \right) \Bigg\} \\ + H \sqrt{\frac{C}{N}} + H \sqrt{\frac{C_1}{n}},$$

where H is the horizon, N is the number of expert state-action pairs, M is the number of offline state-action pairs, n is the number of reward updates, $C_0 = 2 \ln \frac{|\Pi||\mathcal{R}|}{\delta}$, $C = \ln \frac{2|\mathcal{R}|}{\delta}$, $C_1 = 2 \ln |\mathcal{R}|$, and $C_b = \left\| \frac{d_\mu^{\pi_E}}{d_\mu^{\pi_b}} \right\|_\infty$.

Lemma 5.1 upper bounds the advantage over the distribution induced by the expert policy. Theorem 5.2 upper bounds the sample complexity of Algorithm 2 with policy optimization step (5.1). The error consists of three terms. The first term stems from the policy completeness error. The second term stems from the statistical error of estimating the expert policy's state distribution $d_\mu^{\pi_E}$ with the distribution over samples ρ_E. The third term stems from the regret of the reward update. Unlike Theorem 4.3, which considers ϵ-approximate PSDP, Theorem 5.2 examines the case of infinite computations of PSDP such that $\epsilon = 0$, resulting in a vanishing policy optimization error term. Importantly, the assumption of $\epsilon = 0$ is not necessary but rather convenient in simplifying the analysis. Moreover, the $\epsilon > 0$ case was presented in Theorem 4.3.

From Theorem 5.2, we observe the condition under which sub-optimal data benefits learning is when

$$\epsilon_\Pi + \epsilon_\Pi \sqrt{\frac{C_0}{N}} \le \left\| \frac{d_\mu^{\pi_E}}{d_\mu^{\pi_b}} \right\|_\infty \left(\epsilon_\Pi + \epsilon_\Pi \sqrt{\frac{C_0}{N+M}} \right).$$

When the sub-optimal data covers the expert data well, $C_b = \left\| \frac{d_\mu^{\pi_E}}{d_\mu^{\pi_b}} \right\|_\infty$ is small, so the sub-optimal data may be beneficial. Considering the special case where the "sub-optimal" data is collected from the expert policy π_E, then $C_b = \left\| \frac{d_\mu^{\pi_E}}{d_\mu^{\pi_E}} \right\|_\infty = 1$. The advantage bound becomes equivalent to the case of having $N+M$ number of expert data samples. However, because we only use the expert data for the reward update, rather than the sub-optimal data, the reward error terms remain the same.

6. Discussion

We address the seemingly contradictory goals of preventing compounding errors in IRL and avoiding the worst-case exploration complexity of RL. We introduce a novel structural condition, reward-agnostic policy completeness, under which both compounding errors can be avoided efficiently. We then present a reset-based IRL algorithm and perform a finite-sample analysis. Finally, we identify the conditions under which sub-optimal data can be beneficial to the sample-efficiency of the algorithm.

One direction for future work is extending our analysis to RL solvers beyond PSDP, such as replacing CPI's reset distribution by the expert and sub-optimal data distributions. This can also include generalizing our analysis to abstracted RL procedures. Another approach may be to empirically demonstrate the tradeoff between the coverage and amount of sub-optimal data in terms of IRL performance.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-first International Conference on Machine Learning*. ACM, 2008.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32, 2019.

Amortila, P., Jiang, N., Madeka, D., and Foster, D. P. A few expert queries suffices for sample-efficient rl with resets and linear value approximation. *Advances in Neural Information Processing Systems*, 35:29637–29648, 2022.

Bagnell, J., Kakade, S. M., Schneider, J., and Ng, A. Policy search by dynamic programming. *Advances in neural information processing systems*, 16, 2003.

Barnes, M., Abueg, M., Lange, O. F., Deeds, M., Trader, J., Molitor, D., Wulfmeier, M., and O'Banion, S. Massively scalable inverse reinforcement learning in google maps. *arXiv preprint arXiv:2305.11290*, 2023.

Bronstein, E., Palatucci, M., Notz, D., White, B., Kuefler, A., Lu, Y., Paul, S., Nikdel, P., Mougin, P., Chen, H., et al. Hierarchical model-based imitation learning for planning in autonomous driving. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8652–8659. IEEE, 2022.

Igl, M., Kim, D., Kuefler, A., Mougin, P., Shah, P., Shiarlis, K., Anguelov, D., Palatucci, M., White, B., and Whiteson, S. Symphony: Learning realistic and diverse agents for autonomous driving simulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2445–2451. IEEE, 2022.

Jia, Z., Li, G., Rakhlin, A., Sekhari, A., and Srebro, N. When is agnostic reinforcement learning statistically tractable? *Advances in Neural Information Processing Systems*, 36, 2024.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Kearns, M., Mansour, Y., and Ng, A. Approximate planning in large pomdps via reusable trajectories. *Advances in Neural Information Processing Systems*, 12, 1999.

Poiani, R., Curti, G., Metelli, A. M., and Restelli, M. Inverse reinforcement learning with sub-optimal experts. *arXiv preprint arXiv:2401.03857*, 2024.

Ratliff, N., Bagnell, J., and Zinkevich, M. (semi-) autonomous navigation (san) using the maximum margin planning framework. In *Proceedings of Robotics: Science and Systems*. MIT Press, 2007.

Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A. Maximum margin planning. In *Proceedings of the 23rd International Conference on Machine learning*, pp. 729–736. ACM, 2009.

Ross, S. and Bagnell, D. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

Ross, S., Gordon, G., and Bagnell, D. 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. JMLR Workshop and Conference Proceedings, 2011.

Shani, L., Zahavy, T., and Mannor, S. Online apprenticeship learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8240–8248, 2022.

Silver, D., Bagnell, J. A., and Stentz, A. Learning from demonstration for autonomous navigation in complex unstructured terrain. *The International Journal of Robotics Research*, 29(12):1565–1592, 2010.

Swamy, G., Choudhury, S., Bagnell, J. A., and Wu, S. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pp. 10022–10032. PMLR, 2021.

Swamy, G., Rajaraman, N., Peng, M., Choudhury, S., Bagnell, J., Wu, S. Z., Jiao, J., and Ramchandran, K. Minimax optimal online imitation learning via replay estimation. *Advances in Neural Information Processing Systems*, 35: 7077–7088, 2022.

Swamy, G., Wu, D., Choudhury, S., Bagnell, D., and Wu, S. Inverse reinforcement learning without reinforcement learning. In *International Conference on Machine Learning*, pp. 33299–33318. PMLR, 2023.

Viano, L., Skoulakis, S., and Cevher, V. Imitation learning in discounted linear mdps without exploration assumptions. *arXiv preprint arXiv:2405.02181*, 2024.

Vinitsky, E., Lichtlé, N., Yang, X., Amos, B., and Foerster, J. Nocturne: a scalable driving benchmark for bringing multi-agent learning one step closer to the real world. *Advances in Neural Information Processing Systems*, 35: 3962–3974, 2022.

Xu, T., Li, Z., Yu, Y., and Luo, Z.-Q. Provably efficient adversarial imitation learning with unknown transitions. In *Uncertainty in Artificial Intelligence*, pp. 2367–2378. PMLR, 2023.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008a.

Ziebart, B. D., Maas, A. L., Dey, A. K., and Bagnell, J. A. Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. In *Proceedings of the 10th international conference on Ubiquitous computing*, pp. 322–331, 2008b.

Zucker, M., Ratliff, N., Stolle, M., Chestnutt, J., Bagnell, J. A., Atkeson, C. G., and Kuffner, J. Optimization and learning for rough terrain legged locomotion. *The International Journal of Robotics Research*, 30(2):175–191, 2011.

A. Proofs of Section 4

A.1. Proof of Theorem 4.3

Proof. We consider the imitation gap of the expert and the average of the learned policies $\overline{\pi}$,

$$
\begin{aligned}
V^{\pi_E} - V^{\overline{\pi}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathop{\mathbb{E}}_{\zeta\sim\pi_E} \sum_{h=1}^{H} r^*(s,a) - \mathop{\mathbb{E}}_{\zeta\sim\pi_i} \sum_{h=1}^{H} r^*(s,a) \right) \\
&= H\frac{1}{n}\sum_{i=1}^{n}\left(\mathop{\mathbb{E}}_{(s,a)\sim d_\mu^{\pi_E}} r^*(s,a) - \mathop{\mathbb{E}}_{(s,a)\sim d_\mu^{\pi_i}} r^*(s,a) \right) \\
&= H\frac{1}{n}\sum_{i=1}^{n} L(\pi_i, r^*) \\
&\le H\frac{1}{n}\max_{r\in\mathcal{R}}\sum_{i=1}^{n} L(\pi_i, r) \\
&\le H\frac{1}{n}\max_{r\in\mathcal{R}}\sum_{i=1}^{n} L(\pi_i, r) - L(\pi_i, r_i) + L(\pi_i, r_i) \\
&= H\frac{1}{n} L(\pi_i, r_i) + H\frac{1}{n}\max_{r\in\mathcal{R}}\sum_{i=1}^{n} L(\pi_i, r) - L(\pi_i, r_i)
\end{aligned}
$$

Applying the regret bound of Online Mirror Descent (Theorem C.2), we have

$$
\begin{aligned}
V^{\pi_E} - V^{\overline{\pi}} &\le H\frac{1}{n}\sum_{i=1}^{n} L(\pi_i, r_i) + H\sqrt{\frac{\ln|\mathcal{R}|}{n}} \\
&= H\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{H}\sum_{h=1}^{H} \mathop{\mathbb{E}}_{(s_h,a_h)\sim d_h^{\pi_E}} r_i(s_h,a_h) - \frac{1}{H}\sum_{h=1}^{H} \mathop{\mathbb{E}}_{(s_h,a_h)\sim d_h^{\pi_i}} r_i(s_h,a_h) \right) + H\sqrt{\frac{\ln|\mathcal{R}|}{n}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\mathop{\mathbb{E}}_{s\sim\mu} V_{r_i}^{\pi_E} - \mathop{\mathbb{E}}_{s\sim\mu} V_{r_i}^{\pi_i} \right) + H\sqrt{\frac{\ln|\mathcal{R}|}{n}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{h=0}^{H-1}\left(\mathop{\mathbb{E}}_{(s_h,a_h)\sim d_h^{\pi_E}} A_{r_i,h}^{\pi_i}(s_h,a_h) \right) + H\sqrt{\frac{\ln|\mathcal{R}|}{n}}
\end{aligned}
$$

Focusing on the interior summation, we have

$$
\begin{aligned}
\sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{(s_h,a_h)\sim d_h^{\pi_E}} A_h^{\pi_i}(s_h,a_h) &\le \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{s_h\sim d_h^{\pi_E}} \max_{a\in\mathcal{A}} A_h^{\pi_i}(s_h,a) \\
&= \sum_{h=0}^{H-1} \mathop{\mathbb{E}}_{s_h\sim d_h^{\pi_E}} \max_{a\in\mathcal{A}} A_h^{\pi_i}(s_h,a) - \epsilon_{\Pi,h} + \epsilon_{\Pi,h} \\
&= \sum_{h=0}^{H-1} \max_{\pi'\in\Pi} \mathop{\mathbb{E}}_{s_h\sim d_h^{\pi_E}} \mathop{\mathbb{E}}_{a\sim\pi'(\cdot|s)} A_h^{\pi_i}(s_h,a) + \epsilon_{\Pi,h} \\
&\le H^2\epsilon + H\epsilon_{\Pi,h}
\end{aligned}
$$

where the last line holds by PSDP's performance guarantee (Bagnell et al., 2003).

Applying (A.1) to (A.1), we have

$$
\begin{aligned}
V^{\pi_E} - V^{\overline{\pi}} &\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{h=0}^{H-1}\left(\mathop{\mathbb{E}}_{(s_h,a_h)\sim d_h^{\pi_E}} A_{r_i,h}^{\pi_i}(s_h,a_h)\right) + H\sqrt{\frac{\ln|\mathcal{R}|}{n}} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left(H^2\epsilon + H\epsilon_{\Pi,h}\right) + H\sqrt{\frac{\ln|\mathcal{R}|}{n}} \\
&\leq H^2\epsilon + H\epsilon_{\Pi} + H\sqrt{\frac{\ln|\mathcal{R}|}{n}}
\end{aligned}
$$

which completes the proof. □

B. Proofs of Section 5

B.1. Lemmas of Theorem 5.2

Lemma B.1 (Reward Regret Bound)**.** *Recall that*

$$
\hat{L}(\pi, r) = \mathop{\mathbb{E}}_{(s,a)\sim\rho_E} r(s,a) - \mathop{\mathbb{E}}_{(s,a)\sim d_\mu^\pi} r(s,a).
$$

Suppose that we update the reward via the online mirror descent (ascent) algorithm. Since $0 \leq r(s,a) \leq 1$ *for all* s, a*, then* $\sup_{\pi\in\Pi, r\in\mathcal{R}} \hat{L}(\pi, r) \leq 1$*. Applying Theorem C.2 with* $B = 1$*, the regret is given by*

$$
\begin{aligned}
\mathit{Reg}_n &= \sup_{r\in\mathcal{R}} \frac{1}{n}\sum_{i=1}^{n}\hat{L}(\pi_i, r) - \frac{1}{n}\sum_{i=1}^{n}\hat{L}(\pi_i, r_i) \\
&\leq \sqrt{\frac{2\ln|\mathcal{R}|}{n}} \\
&= \sqrt{\frac{C_1}{n}},
\end{aligned}
$$

where $C_1 = 2\ln|\mathcal{R}|$ *and* n *is the number of updates.*

Lemma B.2 (Statistical Difference of Losses)**.** *With probability at least* $1-\delta$*,*

$$
L(\pi, r) \leq \hat{\mathrm{L}}(\pi, r) + \sqrt{\frac{C}{N}},
$$

where $C = \ln\frac{2|\mathcal{R}|}{\delta}$ *and* N *is the number of state-action pairs from the expert.*

Proof. By definition of L and $\hat{\mathrm{L}}$, for any $\pi \in \Pi$ and $r \in \mathcal{R}$, we have

$$
\begin{aligned}
\left|L(\pi, r) - \hat{\mathrm{L}}(\pi, r)\right| &= \left|\mathop{\mathbb{E}}_{(s,a)\sim d_\mu^{\pi_E}} r(s,a) - \mathop{\mathbb{E}}_{(s,a)\sim d_\mu^{\pi}} r(s,a) - \left(\mathop{\mathbb{E}}_{(s,a)\sim\rho_E} r(s,a) - \mathop{\mathbb{E}}_{(s,a)\sim d_\mu^{\pi}} r(s,a)\right)\right| \\
&= \left|\mathop{\mathbb{E}}_{(s,a)\sim d_\mu^{\pi_E}} r(s,a) - \mathop{\mathbb{E}}_{(s,a)\sim\rho_E} r(s,a)\right| \\
&= \left|\mathop{\mathbb{E}}_{(s,a)\sim d_\mu^{\pi_E}} r(s,a) - \frac{1}{N}\sum_{(s_i,a_i)\in D_E}^{N} r(s_i,a_i)\right| \\
&\leq \sqrt{\frac{1}{2N}\ln\frac{2|\mathcal{R}|}{\delta}} \\
&\leq \sqrt{\frac{C}{N}},
\end{aligned}
$$

where $C = 4\ln\frac{2|\mathcal{R}|}{\delta}$. The fourth line holds by Hoeffding's inequality and a union bound. Specifically, we apply Corollary C.1 with $c = 1$, since all rewards are bounded by 0 and 1. We take a union bound over all reward functions in the reward class $\mathcal{R}$. Note that the terms involving π cancel out, so the union bound only applies to the reward function class $\mathcal{R}$. Rearranging terms gives the desired bound. □

Lemma B.3 (Loss Bound)**.** *Suppose that $\epsilon = 0$ and reward function r_i are the input parameters to PSDP, and $\pi_i = (\pi_1^i, \pi_2^i, \dots, \pi_H^i)$ is the output learned policy. Then, with probability at least $1-\delta$,*

$$\hat{\mathrm{L}}(\pi_i, r_i) \le \min\left\{\epsilon_\Pi + \epsilon_\Pi\sqrt{\frac{C_0}{N}}, C_b\left(\epsilon_\Pi + \epsilon_\Pi\sqrt{\frac{C_0}{N+M}}\right)\right\} + \sqrt{\frac{C}{N}},$$

where $C_b = \left\|\frac{d_\mu^{\pi_E}}{d_\mu^{\pi_b}}\right\|_\infty$, H is the horizon, N is the number of expert state-action pairs, M is the number of offline state-action pairs, $C_0 = 2\ln\frac{|\Pi||\mathcal{R}|}{\delta}$, and $C = \ln\frac{2|\mathcal{R}|}{\delta}$.

Proof. By Lemma B.2, we have

$$\begin{aligned}
\hat{\mathrm{L}}(\pi_i, r_i) &\le L(\pi_i, r_i) + \sqrt{\frac{C}{N}} \\
&= \mathop{\mathbb{E}}_{(s,a)\sim d_\mu^{\pi_E}}[r_i(s,a)] - \mathop{\mathbb{E}}_{(s,a)\sim d_\mu^{\pi_i}}[r_i(s,a)] + \sqrt{\frac{C}{N}} \\
&= \frac{1}{H}\left(V_{r_i}^{\pi_E} - V_{r_i}^{\pi_i}\right) + \sqrt{\frac{C}{N}} \\
&= \frac{1}{H}\left(\sum_{h=1}^{H}\mathop{\mathbb{E}}_{(s_h,a_h)\sim d_h^{\pi_E}} A_{r_i,h}^{\pi_i}(s_h,a_h)\right) + \sqrt{\frac{C}{N}} \\
&\le \frac{1}{H}\left(\sum_{h=1}^{H}\mathop{\mathbb{E}}_{s_h\sim d_h^{\pi_E}} \max_{a\in\mathcal{A}} A_{r_i,h}^{\pi_i}(s_h,a)\right) + \sqrt{\frac{C}{N}} \\
&= \frac{1}{H}\left(H\mathop{\mathbb{E}}_{s\sim d^{\pi_E}} \max_{a\in\mathcal{A}} A_{r_i}^{\pi_i}(s,a)\right) + \sqrt{\frac{C}{N}}
\end{aligned}$$

where $C = \ln\frac{2|\mathcal{R}|}{\delta}$. The second line holds by the definition of $L(\pi_i, r_i)$, and the third line holds by the definition of the reward-indexed value function. The fourth line holds by the Performance Difference Lemma (PDL). Applying Lemma 5.1, we have

$$\hat{\mathrm{L}}(\pi_i, r_i) \le \min\left\{\epsilon_\Pi + \epsilon_\Pi\sqrt{\frac{C_0}{N}}, C_b\left(\epsilon_\Pi + \epsilon_\Pi\sqrt{\frac{C_0}{N+M}}\right)\right\} + \sqrt{\frac{C}{MN}},$$

where $C_b = \left\|\frac{d_\mu^{\pi_E}}{d_\mu^{\pi_b}}\right\|_\infty$, H is the horizon, N is the number of expert state-action pairs, M is the number of offline state-action pairs, $C_0 = 2\ln\frac{|\Pi||\mathcal{R}|}{\delta}$, and $C = \ln\frac{2|\mathcal{R}|}{\delta}$. □

B.2. Proof of Lemma 5.1

Proof. Suppose that $\epsilon = 0$ is the input accuracy parameter to PSDP, and the advantages are computed under reward function r_i. PSDP is guaranteed to terminate and output a policy $\pi_i = (\pi_1^i, \pi_2^i, \dots, \pi_H^i)$, such that

$$H\epsilon \ge \max_{\pi'\in\Pi}\mathop{\mathbb{E}}_{s_h\sim\rho_{mix,h}}\mathop{\mathbb{E}}_{a\sim\pi'(\cdot|s)} A_h^{\pi_i}(s_h,a)$$

for all $h \in [H]$ (Bagnell et al., 2003). Consequently, we have

$$\begin{aligned}
H\epsilon &\ge \max_{\pi'\in\Pi}\mathop{\mathbb{E}}_{s\sim\rho_{\text{mix}}}\mathop{\mathbb{E}}_{a\sim\pi'(\cdot|s)} A^{\pi_i}(s,a) \\
&= \max_{\pi'\in\Pi}\mathop{\mathbb{E}}_{s\sim\rho_{\text{mix}}}\mathop{\mathbb{E}}_{a\sim\pi'(\cdot|s)} A^{\pi_i}(s,a) + \epsilon_{\Pi,r_i} - \epsilon_{\Pi,r_i} \\
&= \mathop{\mathbb{E}}_{s\sim\rho_{\text{mix}}}\max_{a\in\mathcal{A}} A^{\pi_i}(s,a) - \epsilon_{\Pi,r_i}
\end{aligned}$$

By definition, $0 \leq \epsilon_{\Pi,r_i} \leq \epsilon_\Pi$, so for any $x \in \mathbb{R}$, $x - \epsilon_{\Pi,r_i} \geq x - \epsilon_\Pi$, so

$$H\epsilon \geq \mathop{\mathbb{E}}_{s\sim\rho_{\text{mix}}} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a) - \epsilon_\Pi.$$

Rearranging the terms gives us

$$\begin{aligned}\mathop{\mathbb{E}}_{s\sim\rho_{\text{mix}}} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a) &\leq H\epsilon + \epsilon_\Pi \\ &= \epsilon_\Pi,\end{aligned}$$

where the last line holds by our assumption that $\epsilon = 0$.

Case 1: Jettison Offline Data We will first consider the case where offline data is useless, in which case we will focus on the expert data.

Note that $\max_{a\in\mathcal{A}} A^{\pi_i}(s,a) \geq 0$ for all $s \in \mathcal{S}$ and $h \in [H]$. Applying the definition of ρ_{mix},

$$\mathop{\mathbb{E}}_{s\sim\rho_{\text{mix}}} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a) = \mathop{\mathbb{E}}_{s\sim\rho_E} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a) + \mathop{\mathbb{E}}_{s\sim\rho_b} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a).$$

Consequently, we know that

$$\begin{aligned}\epsilon_\Pi &\geq \mathop{\mathbb{E}}_{s\sim\rho_E} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a) \\ &= \frac{1}{N}\sum_{s_i\in D_E}^{N} \max_{a\in\mathcal{A}} A^{\pi_i}(s_i,a)\end{aligned}$$

Because $\max_{a\in\mathcal{A}} A^{\pi_i}(s,a) \geq 0$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we know $\max_{a\in\mathcal{A}} A^{\pi_i}(s_i,a) \leq \epsilon_\Pi$ for all $s_i \in D_E$. We apply Hoeffding's inequality (Corollary C.1) with $c = {\epsilon_\Pi}^2$ to bound the difference between $\mathbb{E}_{s\sim d^{\pi_E}} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a)$ and $\mathbb{E}_{s\sim\rho_E} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a)$. We apply a union bound on the policy and reward function. As stated previously, $\max_{a\in\mathcal{A}} A^{\pi_i}(s,a) \geq 0$ for all $s \in \mathcal{S}$. By Hoeffding's inequality, with probability $1-\delta$, we have

$$\begin{aligned}\left|\mathop{\mathbb{E}}_{s\sim d_\mu^{\pi_E}} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a) - \mathop{\mathbb{E}}_{s\sim\rho_E} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a)\right| &= \left|\mathop{\mathbb{E}}_{s\sim d_\mu^{\pi_E}} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a) - \frac{1}{N}\sum_{s_i\in D_E}^{N} \max_{a\in\mathcal{A}} A^{\pi_i}(s_i,a)\right| \\ &\leq \sqrt{\epsilon_\Pi^2 \frac{1}{2N} \ln\frac{|\Pi||\mathcal{R}|}{\delta}} \\ &\leq \epsilon_\Pi\sqrt{\frac{C_0}{N}},\end{aligned}$$

where $C_0 = 2\ln\frac{|\Pi||\mathcal{R}|}{\delta}$. Note that the cardinality of the set of advantage functions over all possible policies is upper bounded by the cardinalities of the policy and reward classes. Rearranging the terms and applying (B.2) yields

$$\mathop{\mathbb{E}}_{s\sim d_\mu^{\pi_E}} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a) \leq \epsilon_\Pi + \epsilon_\Pi\sqrt{\frac{C_0}{N}}.$$

Case 2: Leverage Offline Data Next, we consider the case where offline data is useful, specifically where there is good coverage of the expert data.

Next, we apply Hoeffding's inequality (Corollary C.1) to bound the difference between $\mathbb{E}_{s\sim\nu} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a)$ and $\mathbb{E}_{s\sim\rho_{\text{mix}}} \max_{a\in\mathcal{A}} A^{\pi_i}(s,a)$. We apply a union bound on the policy and reward function. We use $c = \epsilon_\Pi^2$ for a similar argument to the one used in Case 1. As stated previously, $\max_{a\in\mathcal{A}} A^{\pi_i}(s,a) \geq 0$ for all $s \in \mathcal{S}$. By Hoeffding's inequality, with probability $1-\delta$, we have

$$\begin{aligned}\left|\mathop{\mathbb{E}}_{s\sim\nu}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a)-\mathop{\mathbb{E}}_{s\sim\rho_{\text{mix}}}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a)\right| &= \left|\mathop{\mathbb{E}}_{s\sim\nu}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a)-\frac{1}{N+M}\sum_{s_i\in D_{mix}}^{N+M}\max_{a\in\mathcal{A}}A^{\pi_i}(s_i,a)\right| \\ &\leq \sqrt{\epsilon_\Pi\frac{1}{2(N+M)}\ln\frac{|\Pi||\mathcal{R}|}{\delta}} \\ &\leq \epsilon_\Pi\sqrt{\frac{C_0}{N+M}},\end{aligned}$$

where $C_0 = 2\ln\frac{|\Pi||\mathcal{R}|}{\delta}$. Note that the cardinality of the set of advantage functions over all possible policies is upper bounded by the cardinalities of the policy and reward classes. Rearranging the terms and applying (B.2) yields

$$\mathop{\mathbb{E}}_{s\sim\nu}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a)\leq\epsilon_\Pi+\epsilon_\Pi\sqrt{\frac{C_0}{N+M}}.$$

By linearity of expectation, and using the fact that $1\leq C_b<\infty$, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{s\sim d^{\pi_E}}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a) &= \frac{N}{N+M}\mathop{\mathbb{E}}_{s\sim d^{\pi_E}}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a)+\frac{M}{N+M}\mathop{\mathbb{E}}_{s\sim d^{\pi_E}}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a) \\ &\leq \frac{N}{N+M}\mathop{\mathbb{E}}_{s\sim d^{\pi_E}}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a)+C_b\frac{M}{N+M}\mathop{\mathbb{E}}_{s\sim d^{\pi_b}}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a) \\ &\leq C_b\frac{N}{N+M}\mathop{\mathbb{E}}_{s\sim d^{\pi_E}}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a)+C_b\frac{M}{N+M}\mathop{\mathbb{E}}_{s\sim d^{\pi_b}}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a) \\ &= C_b\left(\frac{N}{N+M}\mathop{\mathbb{E}}_{s\sim d^{\pi_E}}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a)+\frac{M}{N+M}\mathop{\mathbb{E}}_{s\sim d^{\pi_b}}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a)\right) \\ &\leq C_b\mathop{\mathbb{E}}_{s\sim\nu}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a).\end{aligned}$$

Applying (B.2) to (B.2), we have

$$\begin{aligned}\mathop{\mathbb{E}}_{s\sim d^{\pi_E}}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a) &\leq C_b\mathop{\mathbb{E}}_{s\sim\nu}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a) \\ &\leq C_b\left(\epsilon_\Pi+\epsilon_\Pi\sqrt{\frac{C_0}{N+M}}\right)\end{aligned}$$

Final Result Using the bounds from Case 1 and Case 2, we know that

$$\mathop{\mathbb{E}}_{s\sim d^{\pi_E}}\max_{a\in\mathcal{A}}A^{\pi_i}(s,a)\leq\min\left\{\epsilon_\Pi+\epsilon_\Pi\sqrt{\frac{C_0}{N}},C_b\left(\epsilon_\Pi+\epsilon_\Pi\sqrt{\frac{C_0}{N+M}}\right)\right\}$$

where $C_b=\left\|\frac{d_\mu^{\pi_E}}{d_\mu^{\pi_b}}\right\|_\infty$, H is the horizon, N is the number of expert state-action pairs, M is the number of offline state-action pairs, and $C_0=2\ln\frac{|\Pi||\mathcal{R}|}{\delta}$. □

B.3. Proof of Theorem 5.2

Proof. We consider the imitation gap of the expert and the averaged learned policies, $\overline{\pi}$,

$$\begin{aligned}
V^{\pi_E} - V^{\overline{\pi}} &= \frac{1}{n}\sum_{i=0}^{n}\left(\mathop{\mathbb{E}}_{\zeta\sim\pi_E}\left[\sum_{h=1}^{H} r^*(s_h,a_h)\right] - \mathop{\mathbb{E}}_{\zeta\sim\pi_i}\left[\sum_{h=1}^{H} r^*(s_h,a_h)\right]\right) \\
&= \frac{1}{n}H\sum_{i=0}^{n}\left(\mathop{\mathbb{E}}_{(s,a)\sim d_\mu^{\pi_E}}\left[r^*(s,a)\right] - \mathop{\mathbb{E}}_{(s,a)\sim d_\mu^{\pi_i}}\left[r^*(s,a)\right]\right) \\
&= \frac{1}{n}H\sum_{i=0}^{n} L(\pi_i, r^*) \\
&\leq \frac{1}{n}H\max_{r\in\mathcal{R}}\sum_{i=0}^{n} L(\pi_i, r)
\end{aligned}$$

where n is the number of updates to the reward function. The second line holds by definition of d_μ^π. The third line holds by definition of L. Applying the Statistical Difference of Losses (Lemma B.2), we have

$$\begin{aligned}
V^{\pi_E} - V^{\overline{\pi}} &\leq \frac{1}{n}H\max_{r\in\mathcal{R}}\sum_{i=0}^{n}\left(\hat{\mathrm{L}}(\pi_i, r) + \sqrt{\frac{C}{N}}\right) \\
&= \frac{1}{n}H\max_{r\in\mathcal{R}}\sum_{i=0}^{n}\left(\hat{\mathrm{L}}(\pi_i, r) - \hat{\mathrm{L}}(\pi_i, r_i) + \hat{\mathrm{L}}(\pi_i, r_i) + \sqrt{\frac{C}{N}}\right)
\end{aligned}$$

where $C = \ln\frac{2|\mathcal{R}|}{\delta}$ and M is the number of state-action pairs from the expert. Applying the Reward Regret Bound (Lemma B.1), we have

$$V^{\pi_E} - V^{\overline{\pi}} \leq \frac{1}{n}H\sum_{i=0}^{n}\left(\hat{\mathrm{L}}(\pi_i, r_i) + \sqrt{\frac{C}{N}}\right) + H\sqrt{\frac{C_1}{n}}$$

where $C_1 = 2\ln|\mathcal{R}|$. Applying the Loss Bound (Lemma B.3), we have

$$V^{\pi_E} - V^{\overline{\pi}} \leq \frac{1}{n}H\sum_{i=0}^{n}\left(\min\left\{\epsilon_\Pi + \epsilon_\Pi\sqrt{\frac{C_0}{N}}, C_b\left(\epsilon_\Pi + \epsilon_\Pi\sqrt{\frac{C_0}{N+M}}\right)\right\} + \sqrt{\frac{C}{N}},\right) + H\sqrt{\frac{C_1}{n}},$$

which simplifies to

$$V^{\pi_E} - V^{\overline{\pi}} \leq H\min\left\{\epsilon_\Pi + \epsilon_\Pi\sqrt{\frac{C_0}{N}}, C_b\left(\epsilon_\Pi + \epsilon_\Pi\sqrt{\frac{C_0}{N+M}}\right)\right\} + H\sqrt{\frac{C}{N}}, +H\sqrt{\frac{C_1}{n}},$$

where $C_b = \left\|\frac{d_\mu^{\pi_E}}{d_\mu^{\pi_b}}\right\|_\infty$, H is the horizon, N is the number of expert state-action pairs, M is the number of offline state-action pairs, n is the number of reward updates, $C_0 = 2\ln\frac{|\Pi||\mathcal{R}|}{\delta}$, $C = \ln\frac{2|\mathcal{R}|}{\delta}$, and $C_1 = 2\ln|\mathcal{R}|$. □

C. Useful Lemmas

Theorem C.1 (Hoeffding's Inequality)**.** *If $Z_1,\ldots,Z_M$ are independent with $P(a \leq Z_i \leq b) = 1$ and common mean μ, then, with probability at least $1-\delta$,*

$$|\overline{Z}_M - \mu| \leq \sqrt{\frac{c}{2M}\ln\frac{2}{\delta}}$$

where $c = \frac{1}{M}\sum_{i=1}^{M}(b_i - a_i)^2$.

Lemma C.2 (Online Mirror Descent Regret)**.** *Regret is defined as*

$$Reg_N = \frac{1}{N}\sum_{t=1}^{N}\ell(\hat{\mathbf{y}}_t, z_t) - \inf_{\mathbf{f}\in\mathcal{F}}\frac{1}{N}\sum_{t=1}^{N}\ell(\mathbf{f}, z_t).$$

Given $\mathcal{F} = \Delta(\mathcal{F}')$ and $\langle \mathbf{f}, \nabla_t \rangle = \mathbb{E}_{f' \sim \mathbf{f}}[\ell(f', (x_t, y_t))]$, where $\sup_{\nabla \in \mathcal{D}} \|\nabla\|_\infty \leq B$, let R be any 1-strongly convex function. If we use the Mirror descent algorithm with $\eta = \sqrt{\frac{2 \sup_{\mathbf{f} \in \mathcal{F}} R(\mathbf{f})}{NB^2}}$, then,

$$Reg_n \leq \sqrt{\frac{2B^2 \sup_{\mathbf{f} \in \mathcal{F}} R(\mathbf{f})}{N}}.$$

If R is the negative entropy function, then $\sup_{\mathbf{f} \in \mathcal{F}} R(\mathbf{f}) \leq \log |\mathcal{F}'|$.