UNDERSTANDING THE RELATION BETWEEN MAXIMUM-ENTROPY INVERSE REINFORCEMENT LEARNING AND BEHAVIOUR CLONING

Anonymous authors

Paper under double-blind review

ABSTRACT

In many settings, it is desirable to learn decision-making and control policies through learning or bootstrapping from expert demonstrations. The most common approaches under this framework are Behaviour Cloning (BC), and Inverse Reinforcement Learning (IRL). Recent methods for IRL have demonstrated the capacity to learn effective policies with access to a very limited set of demonstrations, a scenario in which BC methods often fail. Unfortunately, directly comparing the algorithms for these methods does not provide adequate intuition for understanding this difference in performance. This is the motivating factor for our work. We begin by presenting f-MAX, a generalization of AIRL (Fu et al., 2017), a state-of-the-art IRL method. f-MAX provides grounds for more directly comparing the objectives for LfD objectives. We empirically evaluate the factors of difference between these objectives in the continuous control domain.

1 INTRODUCTION

Modern advances in reinforcement learning aim to alleviate the need for hand-engineered decision-making and control algorithms by designing general purpose methods that learn to optimize provided reward functions. In many cases however, it is either too challenging to optimize a given reward (e.g. due to sparsity of signal), or it is simply impossible to design a reward function that captures the intricate details of desired outcomes. One approach to overcoming such hurdles is Learning from Demonstrations (LfD), where algorithms are provided with expert demonstrations of how to accomplish desired tasks.

The most common approaches in the LfD framework are Behaviour Cloning (BC) and Inverse Reinforcement Learning (IRL) (Russell, 1998; Ng et al., 2000). In standard BC, learning from demonstrations is treated as a supervised learning problem and policies are trained to regress expert actions from a dataset of expert demonstrations. Other forms of Behaviour Cloning, such as DAgger (Ross et al., 2011), consider how to make use of an expert in a more optimal fashion. On the other hand, in IRL the aim is to infer the reward function of the expert, and subsequently train a policy to optimize this reward. The motivation for IRL stems from the intuition that the reward function is the most concise and portable representation of a task (Ng et al., 2000; Abbeel & Ng, 2004).

Unfortunately, the standard IRL formulation (Ng et al., 2000) faces degeneracy issues [1]. A successful framework for overcoming such challenges is the Maximum-Entropy (Max-Ent) IRL method (Ziebart et al., 2008; Ziebart, 2010). A line of research stemming from the Max-Ent IRL framework has lead to recent "adversarial" methods (Ho & Ermon, 2016; Finn et al., 2016a; Fu et al., 2017). These approaches aim to directly recover the expert policy ensuing from the Max-Ent IRL process without explicitly modelling the reward function, and have shown tremendous success. In benchmarks for continuous control (Brockman et al., 2016), these methods outperform Behaviour Cloning by a wide margin, particularly in the low data regime where a very limited number of expert trajectories are available.

[1] for example, any policy is optimal for the constant reward function $r(s, a) = 0$

However, it is not immediately clear why adversarial Max-Ent IRL methods would outperform BC, since at optimality both methods exactly recover the expert policy. This questions motivates the work presented here. Drawing upon the literature on f-divergences (Lin, 1991; Nowozin et al., 2016), we begin by presenting f-MAX, an algorithm for Max-Ent IRL. We demonstrate how f-MAX generalizes AIRL (Fu et al., 2017), and provides new intuition for what this algorithm accomplishes. Specifically we demonstrate that the objective in AIRL is equivalent to minimizing the reverse KL divergence between the joint state-action marginal distribution of the policy and that of the expert. From these findings, we generate hypotheses for why direct Max-Ent IRL methods outperform BC, and empirically evaluate them in continuous control benchmarks. To tease apart the differences between standard BC and AIRL, we also address the degeneracy of f-MAX in a special case, and provide a one-line modification of AIRL, named FAIRL, which minimizes the *forward* KL divergence between the joint state-action marginal of the expert and the policy. We discuss how our findings may relate to common observations (Bishop, 2006) regarding the mode-covering/mode-seeking behaviour of the different KL divergence directions.

2 Background

2.1 Maximum Entropy Inverse Reinforcement Learning

Consider a Markov Decision Process (MDP) represented as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho_0, \gamma)$ with state-space $\mathcal{S}$, action-space $\mathcal{A}$, dynamics $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, reward function $r(s, a)$, initial state distribution ρ_0, and discount factor $\gamma \in (0, 1)$. In Maximum Entropy (Max-Ent) reinforcement learning (Todorov, 2008; Toussaint, 2009; Rawlik et al., 2013; Fox et al., 2015; Haarnoja et al., 2017; 2018), the goal is to find a policy π such that trajectories sampled using this policy follow the distribution

$$p(\tau) = \frac{1}{Z} \exp(R(\tau)) \tag{1}$$

where $\tau = (s_0, a_0, s_1, a_1, ...)$ denotes a trajectory, and $R(\tau) = \sum_t r(s_t, a_t)$ and Z is the partition function. Hence, trajectories that accumulate more reward are exponentially more likely to be sampled.

Converse to the standard RL setting, in Max-Ent IRL (Ziebart et al., 2008; Ziebart, 2010) we are instead presented with an optimal policy π^{exp}, or more realistically, sample trajectories from such a policy, and we seek to find a reward function r that maximizes the likelihood of the trajectories sampled from π^{exp}. Formally, our objective is:

$$\max_r \mathbb{E}_{\tau \sim \pi^{exp}} [R(\tau) - \log Z] \tag{2}$$

Being an energy-based modelling objective, the difficulty in performing this optimization arises from estimating the partition function Z. Initial methods addressed this problem using dynamic programming (Ziebart et al., 2008; Ziebart, 2010), and recent approaches present methods aimed at intractable domains with unknown dynamics (Finn et al., 2016b; Ho & Ermon, 2016; Finn et al., 2016a; Fu et al., 2017; Kostrikov et al., 2018).

2.2 Direct Methods for Max-Ent IRL

Instead of recovering the expert's reward function and policy, recent successful methods in Max-Ent IRL aim to directly recover the policy that would result from the full process. Since such methods only recover the policy, it would be more accurate to refer to them as Imitation Learning algorithms. However, to avoid confusion with Behaviour Cloning methods, in this work we will refer to them as *direct* methods for Max-Ent IRL.

GAIL (Ho & Ermon, 2016) presents a class of algorithms for direct Max-Ent IRL. A successful special case of this method, dubbed Generative Adversarial Imitation Learning (GAIL), is as follows:

Let $\rho^{\text{exp}}(s,a), \rho^{\pi}(s,a)$ denote the state-action marginal distributions of the expert and student policy (the policy being trained) respectively. Let $D(s,a) : \mathcal{S} \times \mathcal{A} \to [0,1]$ be a binary classifier - often referred to as the discriminator - for identifying positive samples from $\rho^{\text{exp}}(s,a)$ from negative samples from $\rho^{\pi}(s,a)$. Using RL, the student policy is trained to maximize $\mathbb{E}_{\tau\sim\pi}\left[\log D(s,a)\right] - \lambda\mathcal{H}^{\text{causal}}(\pi)$, where λ is a hyperparameter and $\mathcal{H}^{\text{causal}}(\pi)$ is the causal entropy (Ziebart, 2010) of the policy. It is shown (Ho & Ermon, 2016) that this training procedure minimizes the Jensen-Shannon divergence between $\rho^{\text{exp}}(s,a)$ and $\rho^{\pi}(s,a)$. Further connections between this work and ours are discussed in section 3.3.

AIRL: Adversarial Inverse Reinforcement Learning Subsequent to the advent of GAIL (Ho & Ermon, 2016), Finn et al. (2016a) present a theoretical discussion of an adversarial training method for Max-Ent IRL that could recover the Max-Ent reward function from expert demonstrations and simultaneously train the Max-Ent policy corresponding to that reward. Fu et al. (2017) present a practical implementation of this method, dubbed Adversarial Inverse Reinforcement Learning (AIRL). They also present additional contributions that are not directly relevant to our work. The direct version of AIRL is as follows:

As before, let $\rho^{\text{exp}}(s,a), \rho^{\pi}(s,a)$ denote the state-action marginal distributions of the expert and student policy respectively and let $D(s,a) : \mathcal{S} \times \mathcal{A} \to [0,1]$ be the discriminator. AIRL defines the reward function, $r(s,a) := \log D(s,a) - \log\left(1 - D(s,a)\right)$, and sets the objective for the student policy to be the RL objective, $\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_t r(s_t,a_t)\right]$. This leads to an iterative optimization process alternating between optimizing the discriminator and the policy, and it is shown that this procedure implicitly optimizes the Max-Ent IRL objective.

Performance With Respect to BC Methods such as GAIL and AIRL have demonstrated significant performance gains compared to Behaviour Cloning. In particular, in standard Mujoco benchmarks (Todorov et al., 2012; Brockman et al., 2016), adversarial methods for Max-Ent IRL achieve strong performance using a *very limited* amount of demonstrations from an expert policy, an important failure scenario for standard Behaviour Cloning.

2.3 f-DIVERGENCES

Ho & Ermon (2016) demonstrate that Max-Ent IRL is the dual problem of matching $\rho^{\pi}(s,a)$ to $\rho^{\text{exp}}(s,a)$; indeed as noted above, GAIL (Ho & Ermon, 2016) optimizes the Jensen-Shannon divergence between the two distributions. In section 3 we present f-MAX, a method for matching $\rho^{\pi}(s,a)$ to $\rho^{\text{exp}}(s,a)$ using arbitrary f-divergences (Lin, 1991). Hence, in this section we recall this class of statistical divergences as well as methods for using them for training generative models.

Let P, Q be two distributions with density functions p, q. For any convex, lower-semicontinuous function $f : \mathbb{R}^+ \to \mathbb{R}$ a statistical divergence can be defined as: $D_f(P||Q) = \int_{\chi} q(x) f\left(\frac{p(x)}{q(x)}\right)$. Divergences derived in this manner are called *f-divergences* and amongst many interesting divergences, include the forward and reverse KL.

Nguyen et al. (2010) present a variational estimation method for f-divergences between arbitrary distributions P, Q. Using the notation of Nowozin et al. (2016) we can write,

$$D_f(P||Q) \geq \sup_{T_\omega \in \mathcal{T}} \left(\mathbb{E}_{x\sim P}\left[T_\omega(x)\right] - \mathbb{E}_{x\sim Q}\left[f^*(T_\omega(x))\right]\right) \tag{3}$$

where $\mathcal{T}$ is an arbitrary class of functions $T_\omega : X \to \mathbb{R}$, and f^* is the convex conjugate of f. Under mild conditions (Nguyen et al., 2010) equality holds between the two sides. Motivated by this variational approximation as well as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), Nowozin et al. (2016) present an iterative optimization scheme for matching an implicit distribution[2] Q to a fixed distribution P using any f-divergence. For a given f-divergence, the corresponding minimax optimization is,

[2]We use the term "implicit distributions" to refer to distributions we can efficiently sample from, e.g. GAN (Goodfellow et al., 2016) generators

$$\min_{Q} \max_{T_\omega} F(\theta, \omega) = \mathbb{E}_{x\sim P}\left[T_\omega(x)\right] - \mathbb{E}_{x\sim Q}\left[f^*(T_\omega(x))\right] \tag{4}$$

Nowozin et al. (2016) discuss practical parameterizations of T_ω, but to avoid notational clutter we will use the form above.

3 f-MAX: f-DIVERGENCE MAX-ENT IRL

We begin by presenting f-MAX, a generalization of AIRL (Fu et al., 2017) which provides a more intuitive interpretation of what similar algorithms accomplish.

Imagine, for some f, we aim to train a policy by optimizing the f-divergence $D_f\left(\rho^{\text{exp}}(s,a)||\rho^\pi(s,a)\right)$. To do so, we propose the following iterative optimization procedure,

$$\max_{T_\omega} \mathbb{E}_{(s,a)\sim\rho^{\text{exp}}(s,a)}\left[T_\omega(s,a)\right] - \mathbb{E}_{(s,a)\sim\rho^{\pi}(s,a)}\left[f^*(T_\omega(s,a))\right] \tag{5}$$

$$\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_t f^*(T_\omega(s_t,a_t))\right] \tag{6}$$

where f^* and T_ω are as defined in section 2.3. Equation 5 is the same as the inner maximization of the f-GAN objective in equation 4; this objective optimizes T_ω so that equation 5 best approximates $D_f\left(\rho^{\text{exp}}(s,a)||\rho^\pi(s,a)\right)$. On the other hand, as we will show below, the objective for the policy is equivalent to minimizing equation 5. Hence, if in each iteration the optimal T_ω is found, the entire procedure is equivalent to iteratively computing $D_f\left(\rho^{\text{exp}}(s,a)||\rho^\pi(s,a)\right)$ and optiminizing the policy to minimize it. This is the intuition underlying the forthcoming proof.

3.1 PROOF OF VALIDITY

Throughout this work, in our proofs we assume environments with finite horizon T. All proofs can be converted in a straight-forward manner to the infinite horizon case with fixed termination probability by using the infinite horizon version of the identity in appendix A.

For any policy π, let T^π_ω denote the maximizer of equation 5. With the simple identity shown in appendix A we have[3],

$$\frac{1}{T}\mathbb{E}_{\tau\sim\pi}\left[\sum_t f^*(T^\pi_\omega(s_t,a_t))\right] = \mathbb{E}_{(s,a)\sim\rho^{\pi}(s,a)}\left[f^*(T^\pi_\omega(s,a))\right] \tag{7}$$

Letting θ denote the parameters of our policy, we have,

$$\nabla_\theta \frac{1}{T}\mathbb{E}_{\tau\sim\pi}\left[\sum_t f^*(T^\pi_\omega(s_t,a_t))\right] = \nabla_\theta \mathbb{E}_{(s,a)\sim\rho^{\pi}(s,a)}\left[f^*(T^\pi_\omega(s,a))\right] \tag{8}$$

$$= \nabla_\theta \mathbb{E}_{(s,a)\sim\rho^{\pi}(s,a)}\left[f^*(T^\pi_\omega(s,a))\right] - \nabla_\theta \mathbb{E}_{(s,a)\sim\rho^{\text{exp}}(s,a)}\left[T^\pi_\omega(s,a)\right] \tag{9}$$

$$= -\nabla_\theta D_f(\rho^{\text{exp}}(s,a)||\rho^\pi(s,a)) \tag{10}$$

Note that in equation 9, as in GANs (Goodfellow et al., 2014), we ignore the effect of a change in π on the change in T^π_ω which makes the second gradient term 0. Our derivation demonstrates that if in each iteration T^π_ω is found, the policy will be trained to minimize $D_f\left(\rho^{\text{exp}}(s,a)||\rho^\pi(s,a)\right)$.

[3]Note that the T in $\frac{1}{T}$ is the fixed finite length of episodes, not the T_ω function

3.2 COROLLARY: A SIMPLE DERIVATION AND INTUITION FOR AIRL

Choosing $f(u) := -\log u$ leads to $D_f(\rho^{\text{exp}}(s,a)||\rho^{\pi}(s,a)) = \text{KL}\left(\rho^{\pi}(s,a)||\rho^{\text{exp}}(s,a)\right)$. This divergence is commonly referred to as the "reverse" KL divergence. In this setting we have, $f^*(t) = -1 - \log(-t)$, and $T^{\pi}_{\omega}(s,a) = -\frac{\rho^{\pi}(s,a)}{\rho^{\text{exp}}(s,a)}$ (Nowozin et al., 2016). Hence, given T^{π}_{ω}, the policy objective in equation 6 takes the form,

$$\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_t f^*(T^{\pi}_{\omega}(s_t,a_t))\right] = \max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_t \log \rho^{\text{exp}}(s_t,a_t) - \log \rho^{\pi}(s_t,a_t) - 1\right] \quad (11)$$

On the other hand, plugging the optimal discriminator $D^{\pi}(s,a) = \frac{\rho^{\text{exp}}(s,a)}{\rho^{\text{exp}}(s,a)+\rho^{\pi}(s,a)}$ (Goodfellow et al., 2014) into the AIRL (Fu et al., 2017) policy objective, we get,

$$\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_t \log D^{\pi}(s_t,a_t) - \log\left(1 - D^{\pi}(s_t,a_t)\right)\right] = \mathbb{E}_{\tau\sim\pi}\left[\sum_t \log \rho^{\text{exp}}(s_t,a_t) - \log \rho^{\pi}(s_t,a_t)\right] \quad (12)$$

As can be seen, the right hand side of equation 12 matches that of equation 11 up to a constant [4], meaning that AIRL is solving the Max-Ent IRL problem by minimizing the reverse KL divergence, $\text{KL}\left(\rho^{\pi}(s,a)||\rho^{\text{exp}}(s,a)\right)$!

3.3 RELATION TO GAIL

Ho & Ermon (2016) present a class of methods for Max-Ent IRL that directly retrieve the expert policy without explicitly finding the reward function of the expert. Using an interesting connection between surrogate cost functions for binary classification and f-divergences (Nguyen et al., 2009), Ho & Ermon (2016) derive a special case of their method for minimizing any *symmetric*[5] f-divergence between $\rho^{\text{exp}}(s,a)$ and $\rho^{\pi}(s,a)$. Choosing the symmetric f-divergence to be the Jensen-Shannon divergence leads to the successful special case dubbed Generative Adversarial Imitation Learning (GAIL).

4 UNDERSTANDING THE RELATIONS AMONG LEARNING-FROM-DEMONSTRATION ALGORITHMS

Given results derived in the prior section, we can now begin to populate table 1, writing various Imitation Learning algorithms in a common form, as the minimization of some statistical divergence between $\rho^{\text{exp}}(s,a)$ and $\rho^{\pi}(s,a)$. In Behaviour Cloning we minimize $\mathbb{E}_{\rho^{\text{exp}}(s)}\left[\text{KL}\left(\rho^{\text{exp}}(a|s)||\rho^{\pi}(a|s)\right)\right]$[6]. On the other hand, the corollary in section 3.2 demonstrates that AIRL (Fu et al., 2017) minimizes $\text{KL}\left(\rho^{\pi}(s,a)||\rho^{\text{exp}}(s,a)\right)$, while GAIL (Ho & Ermon, 2016) optimizes $D_{JS}(\rho^{\text{exp}}(s,a)||\rho^{\pi}(s,a)) - \lambda\mathcal{H}^{\text{causal}}(\pi)$. Hence, there are two ways in which the direct IRL methods differ from BC. First, in standard BC the policy is optimized to match the conditional distribution $\rho^{\text{exp}}(a|s)$, whereas in the other two the policy is explicitly encouraged to match the marginal state distributions as well. Second, in BC we make use of the forward KL divergence, whereas AIRL and GAIL use divergences that exhibit more mode-seeking behaviour. These observations allow us to generate the following two hypotheses about why direct IRL methods outperform BC, particularly in the low-data regime,

[4] In both settings of fixed finite horizon, and infinite horizon with constant probability of termination, the additional term resulting from the -1 is a constant.

[5] We call an f-divergence symmetric if for any P,Q we have $D_f(P||Q) = D_f(Q,P)$

[6] Since it is equal to minimizing $-\mathbb{E}_{\rho^{\text{exp}}(s,a)}\left[\log \rho^{\pi}(a|s)\right] - \mathcal{H}^{\text{exp}}(s,a)$ and $\mathcal{H}^{\text{exp}}(s,a)$ is constant w.r.t. the policy ($\mathcal{H}^{\text{exp}}(s,a)$ is the entropy of $\rho^{\text{exp}}(s,a)$)

Method	Optimized Objective (Minimization)
Standard Behaviour Cloning	$\mathbb{E}_{\rho^{\text{exp}}(s)}\left[\text{KL}\left(\rho^{\text{exp}}(a\|s)\|\|\rho^{\pi}(a\|s)\right)\right] = -\mathbb{E}_{\rho^{\text{exp}}(s,a)}\left[\log \rho^{\pi}(a\|s)\right]$
DAgger (Ross et al., 2011)	$\mathbb{E}_{\rho^{\text{agg}_{1:n}}(s)}\left[\text{KL}\left(\rho^{\text{exp}}(a\|s)\|\|\rho^{\pi}(a\|s)\right)\right]$ at iteration $n+1$
FAIRL (this work, section 5)	$\text{KL}(\rho^{\text{exp}}(s,a)\|\|\rho^{\pi}(s,a)) = -\mathbb{E}_{\rho^{\text{exp}}(s,a)}\left[\log \rho^{\pi}(s,a)\right] - \mathcal{H}(\rho^{\text{exp}}(s,a))$
AIRL (Fu et al., 2017)	$\text{KL}(\rho^{\pi}(s,a)\|\|\rho^{\text{exp}}(s,a)) = -\mathbb{E}_{\rho^{\pi}(s,a)}\left[\log \rho^{\text{exp}}(s,a)\right] - \mathcal{H}(\rho^{\pi}(s,a))$
GAIL (Ho & Ermon, 2016)	$D_{\text{JS}}(\rho^{\text{exp}}(s,a)\|\|\rho^{\pi}(s,a)) - \lambda\mathcal{H}^{\text{causal}}(\pi)$
Ho & Ermon (2016) symm. f-div	$D_{f\text{-symm}}(\rho^{\pi}(s,a)\|\|\rho^{\text{exp}}(s,a)) - \lambda\mathcal{H}^{\text{causal}}(\pi)$
f-MAX (this work, section 3)	$D_f(\rho^{\pi}(s,a)\|\|\rho^{\text{exp}}(s,a))$

Table 1: The objective function for various imitation learning algorithms, written in a common form as the minimization of statistical divergences. $\mathcal{H}(\cdot)$ denotes entropy, $\mathcal{H}^{\text{causal}}(\pi)$ denotes the causal entropy of the policy (Ziebart, 2010; Ho & Ermon, 2016), and λ is a hyperparameter. JS denotes the Jensen-Shannon divergence and D_f indicates any f-divergence. For DAgger, we are showing the objective for the simplest form of the algorithm, where $\pi^{(i)}$ is the policy obtained at iteration i, $\pi^{(1)}$ is the expert, and $\rho^{\text{agg}_{1:n}}(s) = \frac{1}{n}\sum_{i=1}^{n}\rho^{\pi^{(i)}}(s)$.

Hypothesis 1 *In common MDPs of interest, the reward function depends more on the state than the action. Hence it is plausible that matching state marginals is more useful than matching action conditional marginals.*

Hypothesis 2 *It is known that optimization using the forward KL divergence results in distributions with a mode-covering behaviour, whereas using the reverse KL results in mode-seeking behaviour (Bishop, 2006). Therefore, since in Reinforcement Learning we care about the "quality of trajectories", being mode-seeking is more beneficial than mode-covering, particularly in the low-data regime.*

In what follows, we seek to experimentally evaluate our hypotheses. To tease apart the differences between the direct Max-Ent IRL methods and BC, we present an algorithm that optimizes $\text{KL}\left(\rho^{\text{exp}}(s,a)\|\|\rho^{\pi}(s,a)\right)$. We then compare its performance to Behaviour Cloning and the standard AIRL algorithm using varying amounts of expert demonstrations.

5 Deriving Forward KL

While f-MAX is a general algorithm, useful for most choices of f, it unfortunately cannot be used for the special case of forward KL, i.e. $\text{KL}\left(\rho^{\text{exp}}(s,a)\|\|\rho^{\pi}(s,a)\right)$. In the following sections we identify the problem and present a separate direct Max-Ent IRL method that optimizes this divergence.

5.1 Deriving Forward KL from f-MAX

For the case of forward KL, using notation from section 3.1 and drawing upon equations from Nowozin et al. (2016) we have,

$$u := \frac{\rho^{\text{exp}}(s,a)}{\rho^{\pi}(s,a)} \qquad f(u) := u\log u \qquad f^*(t) = \exp(t-1) \qquad T^{\pi}_{\omega} = 1 + \log\frac{\rho^{\text{exp}}(s,a)}{\rho^{\pi}(s,a)} \tag{13}$$

Given this, the objective for the policy (equation 6) under optimal T^{π}_{ω} becomes,

$$\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_t f^*(T^{\pi}_{\omega}(s_t,a_t))\right] \propto \mathbb{E}_{(s,a)\sim\rho^{\pi}(s,a)}\left[f^*(T^{\pi}_{\omega}(s,a))\right] \tag{14}$$

$$= \mathbb{E}_{(s,a)\sim\rho^{\pi}(s,a)}\left[\exp\left(\left(1+\log\frac{\rho^{\text{exp}}(s,a)}{\rho^{\pi}(s,a)}\right)-1\right)\right] \tag{15}$$

$$= \mathbb{E}_{(s,a)\sim\rho^{\pi}(s,a)}\left[\frac{\rho^{\text{exp}}(s,a)}{\rho^{\pi}(s,a)}\right] \tag{16}$$

$$= 1 \tag{17}$$

Hence, there is no signal to train the policy! [7]

5.2 An Alternative Method for Forward KL

In this section we derive an algorithm for optimizing KL $(\rho^{\text{exp}}(s,a)||\rho^{\pi}(s,a))$. Similar to AIRL (Fu et al., 2017), let us have a discriminator, $D(s,a)$ whose objective is to discriminate between expert and policy state-action pairs,

$$\max_{D} \mathbb{E}_{(s,a)\sim\rho^{\text{exp}}(s,a)}\left[\log D(s,a)\right] + \mathbb{E}_{(s,a)\sim\rho^{\pi}(s,a)}\left[\log\left(1-D(s,a)\right)\right] \tag{18}$$

We now define the objective for the policy to be,

$$h(s,a) := \log D(s,a) - \log\left(1-D(s,a)\right) \tag{19}$$

$$r(s,a) := \exp(h(s,a))\cdot(-h(s,a)) \tag{20}$$

$$\max_{\pi} \mathbb{E}_{\tau\sim\pi_{\theta}}\left[\sum_t r(s_t,a_t)\right] \tag{21}$$

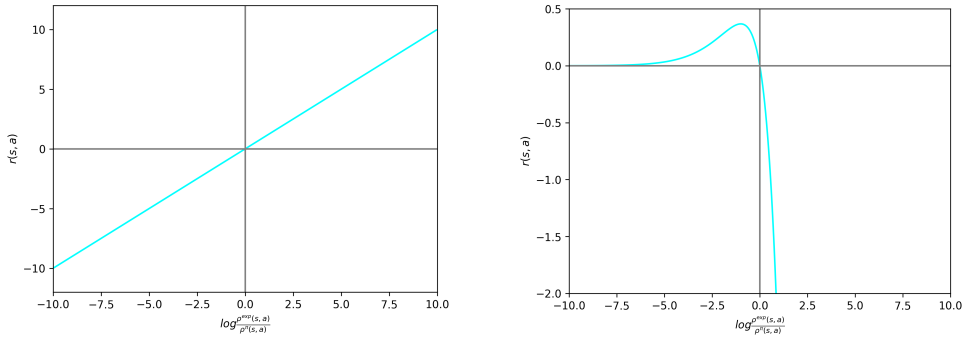
In appendix B we show,

$$\mathbb{E}_{\tau\sim\pi}\left[\sum_t r(s_t,a_t)\right] = -\text{KL}(\rho^{\text{exp}}(s,a)||\rho^{\pi}(s,a)) \tag{22}$$

This is a refreshing result since it demonstrates that we can convert the AIRL algorithm (Fu et al., 2017) into its forward KL counterpart by simply modifying the reward function used; in AIRL (reverse KL) the reward is defined as $r(s,a) := \log D(s,a) - \log\left(1-D(s,a)\right)$, whereas for forward KL it is defined as $r(s,a) := \frac{D(s,a)}{1-D(s,a)}\cdot\log\frac{1-D(s,a)}{D(s,a)}$. We refer to this forward KL version of AIRL as FAIRL.

If we parameterize the discriminator as $D(s,a) := \sigma(\ell(s,a))$, where σ represents the sigmoid activation function, the logit of the discriminator, $\ell(s,a)$, is equal to $\log D(s,a) - \log\left(1-D(s,a)\right)$. Hence, for an optimal discriminator, D^{π}, we have $\ell^{\pi}(s,a) = \log\frac{\rho^{\text{exp}}(s,a)}{\rho^{\pi}(s,a)}$. It is instructive to plot the reward functions under the two different settings as a function of $\ell^{\pi}(s,a)$; figure 1 presents these plots. As can be seen, in the forward KL version of AIRL, if for a state-action pair the expert puts more probability mass than the policy, the policy is *severely* punished. However, if for some state-action pairs the policy places a lot more mass than the expert, it almost does not matter. As a result, the policy spreads its mass. On the other hand, in the original AIRL formulation (reverse KL), the policy is *always* encouraged to put less mass than the expert. These observations are in line with standard intuitions about the mode-covering/mode-seeking behaviours of the two KL divergences (Bishop, 2006).

[7] A similar results holds for the standard f-GAN formulation (Nowozin et al., 2016).



(a) Reverse: KL $(\rho^{\pi}(s,a)||\rho^{\exp}(s,a))$

(b) Forward: KL $(\rho^{\exp}(s,a)||\rho^{\pi}(s,a))$

Figure 1: $r(s,a)$ as the function of the logits of the optimal discriminator, $\ell^{\pi}(s,a) = \log \frac{\rho^{\exp}(s,a)}{\rho^{\pi}(s,a)}$.

6 Experiments

In this section we provide empirical comparisons between AIRL, FAIRL, and standard BC in the Ant and Halfcheetah environments found in Open-AI Gym (Brockman et al., 2016).

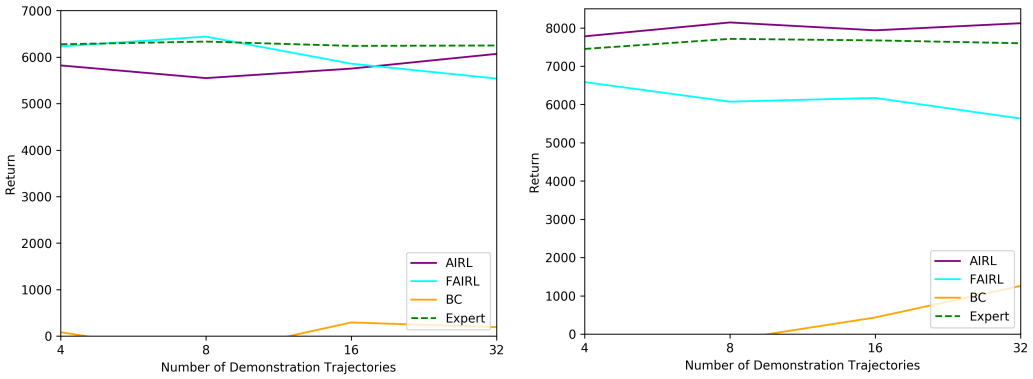
6.1 Setup

Expert Policy To simulate access to expert demonstrations we train an expert policy using Soft-Actor-Critic (SAC) (Haarnoja et al., 2018), a state-of-the-art reinforcement learning algorithm for continuous control. The expert policy consists of a 2-layer MLP with 256-dim layers, ReLU activations, and two output streams for the mean and the diagonal covariance of a `Tanh(Normal(`μ, σ`))` distribution [8]. We use the default hyperparameter settings for training the expert.

Evaluation Setup Using a trained expert policy, we generated 4 sets of expert demonstrations of that contain $\{4, 8, 16, 32\}$ trajectories. Starting from a random offset, each trajectory is subsampled by a factor of 20. This is standard protocol employed in prior direct methods for Max-Ent IRL (Ho & Ermon, 2016; Fu et al., 2017). Also note that when generating demonstrations we *sample* from the expert's action distribution rather than taking the mode. This way, since the expert was trained using Soft-Actor-Critic, the expert should correspond to the Max-Ent optimal policy for the reward function $\frac{1}{\tau} r_g(s,a)$, where τ is the SAC temperature used and $r_g(s,a)$ is the ground-truth reward function. To compare the various learning-from-demonstration algorithms we train each method at each amount of expert demonstrations using 2 random seeds. For each seed, we checkpoint the model at its best validation loss[9] throughout training. At the end of training, the resulting checkpoints are evaluated on 50 test episodes.

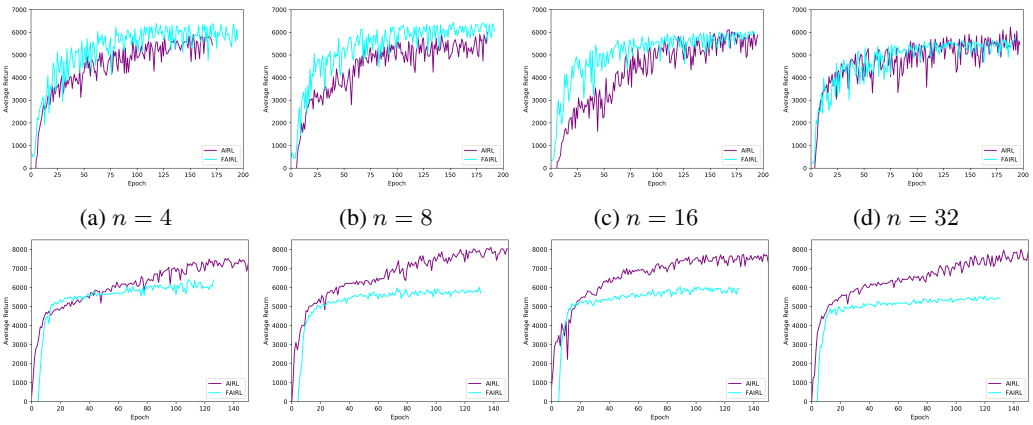
Details for AIRL & FAIRL For AIRL and FAIRL, the student policy has an identical architecture to that of the expert, and the discriminator is a 2-layer MLP with 256-dim layers and Tanh activations. We normalize the observations from the environment by computing the mean and standard deviations of the expert demonstrations. The RL algorithm used for the student policies is SAC (Haarnoja et al., 2018), and the temperature parameter is tuned separately for AIRL & FAIRL.

Details for BC For BC, we use an identical architecture as the expert. The model was fit using Maximum Likelihood Estimation[10]. As before, the observations from the environment are normalized using the mean and standard deviation of the expert demonstrations.



(a) Ant (b) HalfCheetah

Figure 2: Average return on 50 evaluation trajectories as a function of number of expert demonstrations (higher is better). Models evaluated *deterministically*. As we ran two seeds per experiment, we do not present standard deviations. While FAIRL performs comparably to AIRL, Behaviour cloning lags behind quite significantly. Considering the form of their objectives (table 1), this demonstrates that the advantage of direct Max-Ent IRL methods over BC is a result of the additional aspect of their objectives explicitly matching marginal state distributions.



(a) $n = 4$ (b) $n = 8$ (c) $n = 16$ (d) $n = 32$

(e) $n = 4$ (f) $n = 8$ (g) $n = 16$ (h) $n = 32$

Figure 3: Training curves averaged across random seeds. Top row Ant, Bottom row Halfcheetah. n represents the number of expert demonstrations provided. While in the Ant environment FAIRL has slight performance gains, in Halfcheetah AIRL performs noticebaly better. Further experiments are necessary to compare these two methods.

6.2 Results & Interpretations

To match state-action marginals, the optimal student policy must *sample* actions from the state-conditional distribution, $\pi(a|s)$. On the other hand, when we deploy a trained policy it is reasonable to instead choose the mode of this distribution, which we call the deterministic setting. Here, we present evaluation results under the latter, and defer the results for the stochastic setting to the appendix.

Figure 4 demonstrates that both AIRL and FAIRL outperform BC by a large margin, especially in the low data regime. Specifically, the fact that FAIRL outperforms BC supports hypothesis 1 that

[8]This is the architecture presented in SAC (Haarnoja et al., 2018)

[9]Average return on 10 test episodes

[10]Recall that given a state, the output of the policy is a `Tanh(Normal(`μ, σ`))` distribution

the performance gain of Max-Ent IRL is not necessarily due to the direction of KL divergence used, but is the result of explicitly encouraging the policy to match the marginal state distribution of the expert in addition to the matching of conditional action distribution.

To compare AIRL and FAIRL, in figure 5 we plot the validation curves throughout training. While FAIRL has some gains over AIRL in the Ant environment, it does worse in Halfcheetah. However in the stochastic evaluation setting AIRL seems to outperform FAIRL (appendix C). Another trend that we do consistently observe is that, as we increase the amount of demonstrations, the performance of FAIRL slightly diminishes. Such observations provide initial positive support for hypothesis 2; as more expert demonstrations are provided, the policy trained with FAIRL broadens its distribution to cover the data-distribution, resulting in trajectories accumulating less reward in expectation. We note however that more detailed experiments are necessary for adequately comparing the two methods.

7 Conclusion & Future Work

The motivation for this work stemmed from the superior performance of recent direct Max-Ent IRL methods (Ho & Ermon, 2016; Fu et al., 2017) compared to BC in the low-data regime, and the desire to understand the relation between various approaches for Learning from Demonstrations. We first presented f-MAX, a generalization of AIRL (Fu et al., 2017), which allowed us to interpret AIRL as optimizing for $\mathrm{KL}\left(\rho^{\pi}(s,a)||\rho^{\exp}(s,a)\right)$. Comparing to the standard BC objective, $\mathbb{E}_{\rho^{\exp}(s)}\left[\mathrm{KL}\left(\rho^{\exp}(a|s)||\rho^{\pi}(a|s)\right)\right]$, we hypothesized two reasons for the superior performance of AIRL: 1) the additional terms in the objective encouraging the matching of marginal state distributions, and 2) the direction of the KL divergence being optimized. Setting out to empirically evaluate these claims we presented FAIRL, a one-line modification of the AIRL algorithm that optimizes $\mathrm{KL}\left(\rho^{\exp}(s,a)||\rho^{\pi}(s,a)\right)$. FAIRL outperformed BC in a similar fashion to AIRL, which allowed us to conclude the key factor being the matching of state marginals. Additional comparisons between FAIRL and AIRL provided initial understanding about the role of the direction of the KL being optimized.

In future work we aim to produce results on a more diverse set of more challenging environments. Additionally, evaluating other choices of f-divergence beyond forward and reverse KL may present interesting avenues for improvement (Wang et al., 2018). Lastly, but importantly, we would like to understand whether the mode-covering behaviour of FAIRL could result in more robust policies (Rajeswaran et al., 2017).

REFERENCES

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1. ACM, 2004.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*, 2016a.

Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*, pp. 49–58, 2016b.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. *arXiv preprint arXiv:1512.08562*, 2015.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *CoRR*, abs/1710.11248, 2017. URL http://arxiv.org/abs/1710.11248.

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.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1352–1361. JMLR. org, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *CoRR*, abs/1606.03476, 2016. URL http://arxiv.org/abs/1606.03476.

Ilya Kostrikov, Kumar Krishna Agrawal, Sergey Levine, and Jonathan Tompson. Addressing sample inefficiency and reward bias in inverse reinforcement learning. *arXiv preprint arXiv:1809.02925*, 2018.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

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

XuanLong Nguyen, Martin J Wainwright, Michael I Jordan, et al. On surrogate loss functions and f-divergences. *The Annals of Statistics*, 37(2):876–904, 2009.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pp. 271–279, 2016.

Aravind Rajeswaran, Kendall Lowrey, Emanuel V Todorov, and Sham M Kakade. Towards generalization and simplicity in continuous control. In *Advances in Neural Information Processing Systems*, pp. 6550–6561, 2017.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

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.

Stuart J Russell. Learning agents for uncertain environments. In *COLT*, volume 98, pp. 101–103, 1998.

Emanuel Todorov. General duality between optimal control and estimation. In *2008 47th IEEE Conference on Decision and Control*, pp. 4286–4292. IEEE, 2008.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

Marc Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th annual international conference on machine learning*, pp. 1049–1056. ACM, 2009.

Dilin Wang, Hao Liu, and Qiang Liu. Variational inference with tail-adaptive f-divergence. In *Advances in Neural Information Processing Systems*, pp. 5742–5752, 2018.

Brian D Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. 2010.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A SOME USEFUL IDENTITIES

Let $h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be an arbitrary function. If all episodes have the same length T, we have,

$$\mathbb{E}_{\tau\sim\pi}\left[\sum_t h(s_t, a_t)\right] = \sum_t \mathbb{E}_{(s_t,a_t)\sim\rho^\pi(s_t,a_t)}\left[h(s_t, a_t)\right] \tag{23}$$

$$= \sum_t \int_{S,A} \rho^\pi(s_t, a_t) h(s_t, a_t) \tag{24}$$

$$= \int_{S,A} \left[\sum_t \rho^\pi(s_t, a_t)\right] h(s, a) \tag{25}$$

$$= T \cdot \int_{S,A} \rho^\pi(s, a) h(s, a) \tag{26}$$

$$= T \cdot \mathbb{E}_{(s,a)\sim\rho^\pi(s,a)}\left[h(s, a)\right] \tag{27}$$

In a somewhat similar fashion, in the infinite horizon case with fixed probability $\gamma \in (0, 1)$ of transitioning to a terminal state, for the discounted sum below we have,

$$\mathbb{E}_{\tau\sim\pi}\left[\sum_t \gamma^t h(s_t, a_t)\right] = \sum_t \mathbb{E}_{(s_t,a_t)\sim\rho^\pi(s_t,a_t)}\left[\gamma^t h(s_t, a_t)\right] \tag{28}$$

$$= \sum_t \int_{S,A} \gamma^t \rho^\pi(s_t, a_t) h(s_t, a_t) \tag{29}$$

$$= \int_{S,A} \left[\sum_t \gamma^t \rho^\pi(s_t, a_t)\right] h(s, a) \tag{30}$$

$$= \Gamma \cdot \int_{S,A} \rho^\pi(s, a) h(s, a) \tag{31}$$

$$= \Gamma \cdot \mathbb{E}_{(s,a)\sim\rho^\pi(s,a)}\left[h(s, a)\right] \tag{32}$$

where $\Gamma := \frac{1}{1-\gamma}$ is the normalizer of the sum $\sum_t \gamma^t$. Since the integral of an infinite series is not always equal to the infinite series of integrals, some analytic considerations must be made to go from equation 29 to 30. But, one simple case in which it holds is when the ranges of h and all $\rho^\pi(s_t, a_t)$ are bounded.

B DERIVATION FOR FAIRL

Below we present the derivation for equation 22. Recalling definitions,

$$h(s, a) := \log D(s, a) - \log\left(1 - D(s, a)\right) \tag{33}$$

$$r(s, a) := \exp(h(s, a)) \cdot (-h(s, a)) \tag{34}$$

and assuming the discriminator is optimal[11], we have,

$$\mathbb{E}_{\tau\sim\pi}\left[\sum_t r(s_t,a_t)\right] = \mathbb{E}_{\tau\sim\pi}\left[\sum_t \exp(h(s_t,a_t))\cdot(-h(s_t,a_t))\right] \tag{35}$$

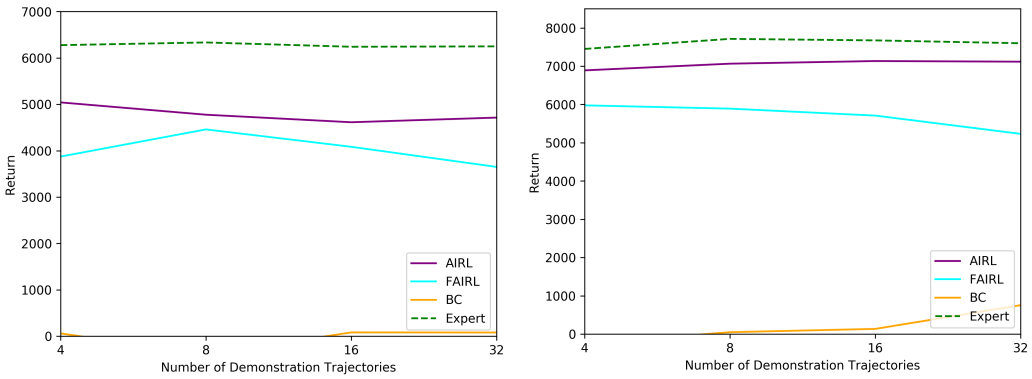
$$= \mathbb{E}_{\tau\sim\pi}\left[\sum_t \frac{\rho^{\exp}(s_t,a_t)}{\rho^{\pi}(s_t,a_t)}\cdot\log\frac{\rho^{\pi}(s_t,a_t)}{\rho^{\exp}(s_t,a_t)}\right] \tag{36}$$

$$\propto \mathbb{E}_{(s,a)\sim\rho^{\pi}(s,a)}\left[\frac{\rho^{\exp}(s_t,a_t)}{\rho^{\pi}(s_t,a_t)}\cdot\log\frac{\rho^{\pi}(s_t,a_t)}{\rho^{\exp}(s_t,a_t)}\right] \tag{37}$$

$$= \mathbb{E}_{(s,a)\sim\rho^{\exp}(s,a)}\left[\log\frac{\rho^{\pi}(s_t,a_t)}{\rho^{\exp}(s_t,a_t)}\right] \tag{38}$$

$$= -\mathrm{KL}(\rho^{\exp}(s,a)||\rho^{\pi}(s,a)) \tag{39}$$

C STOCHASTIC EVALUATION

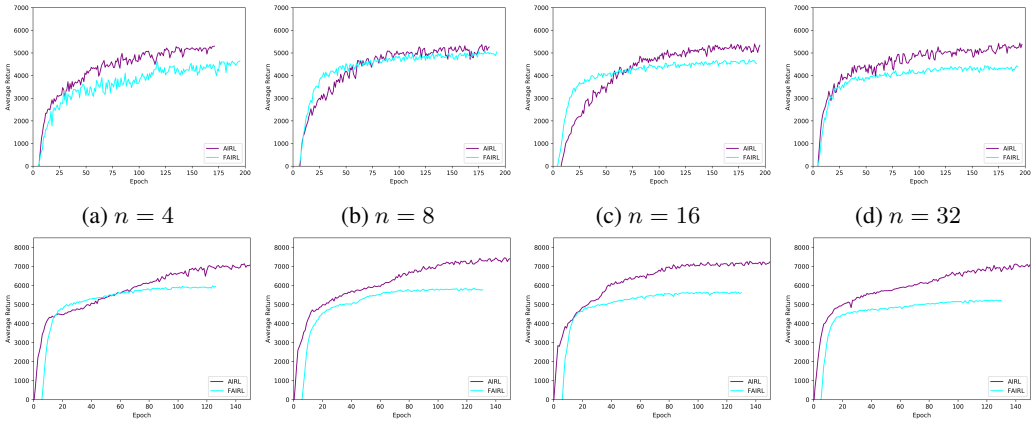


(a) Ant

(b) HalfCheetah

Figure 4: Average return on 50 evaluation trajectories as a function of number of expert demonstrations (higher is better). Models evaluated *stochastically*. As we ran two seeds per experiment, we do not present standard deviations. While FAIRL performs comparably to AIRL, Behaviour cloning lags behind quite significantly. Considering the form of their objectives (table 1), this demonstrates that the advantage of direct Max-Ent IRL methods over BC is a result of the additional aspect of their objectives explicitly matching marginal state distributions.

[11] As a reminder, the optimal discriminator has the form, $D(s,a) = \frac{\rho^{\exp}(s,a)}{\rho^{\exp}(s,a)+\rho^{\pi}(s,a)}$. A simple proof of which can be found in Goodfellow et al. (2014).



(a) $n = 4$ (b) $n = 8$ (c) $n = 16$ (d) $n = 32$

(e) $n = 4$ (f) $n = 8$ (g) $n = 16$ (h) $n = 32$

Figure 5: Training curves averaged across random seeds. Top row Ant, Bottom row Halfcheetah. n represents the number of expert demonstrations provided. AIRL appears to perform better than FAIRL.