
Generative Adversarial Imitation from Observation

Faraz Torabi¹ Garrett Warnell² Peter Stone¹

Abstract

Imitation from observation (IfO) is the problem of learning directly from state-only demonstrations without having access to the demonstrator’s actions. The lack of action information both distinguishes *IfO* from most of the literature in imitation learning, and also sets it apart as a method that may enable agents to learn from a large set of previously inapplicable resources such as internet videos. In this paper, we propose both a general framework for *IfO* approaches and also a new *IfO* approach based on generative adversarial networks called *generative adversarial imitation from observation (GAIfo)*. We conduct experiments in two different settings: (1) when demonstrations consist of low-dimensional, manually-defined state features, and (2) when demonstrations consist of high-dimensional, raw visual data. We demonstrate that our approach performs comparably to classical imitation learning approaches (which have access to the demonstrator’s actions) and significantly outperforms existing imitation from observation methods in high-dimensional simulation environments.

1. Introduction

One well-known way in which artificially-intelligent agents are able to learn to perform tasks is via *reinforcement learning (RL)* (Sutton & Barto, 1998) techniques. Using these techniques, if agents are able to interact with the world and receive feedback (known as *reward*) based on how well they are performing with respect to a particular task, they are able to use their own experience to improve their future behavior. However, designing a proper feedback mechanism for complex tasks can sometimes prove to be extremely difficult for system designers. Moreover, learning based solely on one’s own experience can be exceedingly slow.

^{*}Equal contribution ¹University of Texas at Austin, Austin, USA ²U.S. Army Research Laboratory. Correspondence to: Faraz Torabi <faraztrb@cs.utexas.edu>.

Concerns such as the ones above have given rise to the study of *imitation learning* (Schaal, 1997; Billard et al., 2008; Argall et al., 2009), where agents instead attempt to learn a task by observing another, more expert agent perform that task. Because the information about how to perform the task is communicated to the imitating agent via a demonstration, this paradigm does not require the explicit design of a reward function. Moreover, because the demonstrations directly provide rich information regarding how to perform the task correctly, imitation learning is typically faster than *RL*. While there are multiple ways that this problem can be formulated, one general approach is referred to as *inverse reinforcement learning (IRL)* (Russell, 1998). *IRL*-based techniques aim to first infer the expert agent’s reward function, and then learn imitating behavior using *RL* techniques that utilize the inferred function.

Importantly, most of the imitation learning literature has thus far concentrated only on situations in which the imitator not only has the ability to observe the demonstrating agent’s *states* (e.g., observable quantities such as spatial location), but also the ability to observe the demonstrator’s *actions* (e.g., internal control signals such as motor commands). While this extra information can make the imitation learning problem easier, requiring it is also limiting. In particular, requiring action observations makes a large number of valuable learning resources – e.g., vast quantities of online videos of people performing different tasks (Zhou et al., 2017) – useless. For the demonstrations present in such resources, the actions of the expert are unknown. This limitation has recently motivated work in the area of *imitation from observation (IfO)* (Liu et al., 2017), in which agents seek to perform imitation learning using state-only demonstrations.

Broadly speaking, the *IfO* problem consists of two major subproblems: (1) perception of the demonstrations, i.e., extracting useful features from raw visual data, and (2) learning a control policy using the extracted features. Most *IfO* work thus far (Liu et al., 2017; Sermanet et al., 2017) has focused on perception and not on control. While powerful methods for perceiving the demonstrations have been developed, the control problem is solved via relatively simple means, i.e., reinforcement learning over a pre-defined reward function. Depending on the defined reward function, this approach could be restrictive, as discussed further in

the next section. Therefore, we seek a more sophisticated control algorithm that is able to learn the task automatically from the demonstrations without explicitly defining a reward function.

In this paper, we propose a general framework for the control aspect of *Ifo* in which we characterize the cost as a function of state transitions only. Under this framework, the *Ifo* problem becomes one of trying to recover the state-transition cost function of the expert. Inspired by the work of Ho & Ermon (2016), we introduce a novel, model-free algorithm called *generative adversarial imitation from observation* (*GAIfo*) and prove that it is a specific version of the general framework proposed for *Ifo*. We then experimentally evaluate *GAIfo* in high-dimensional simulation environments in two different settings: (1) demonstrations and states of the imitator are manually-defined features, and (2) demonstrations and states of the imitator come exclusively from raw visual observation. We show that the proposed method compares favorably to other recently-developed methods for *Ifo* and also that it performs comparably to state-of-the-art conventional imitation learning methods that *do* have access to the demonstrator’s actions.

The rest of this paper is organized as follows. In Section 2, we cover related work in imitation learning and review existing research in imitation from observation. Then, we present the notation and background needed in Section 3. In Section 4, we introduce our proposed general framework for *Ifo* problems and, in Sections 5 and 6, we discuss our *Ifo* algorithm, *GAIfo*. Finally, we describe and discuss our experiments in Sections 7 and 8, respectively.

2. Related Work

Because our work is related to imitation learning (Schaal et al., 2003), we first discuss here different approaches and recent advancements in this area. In general, existing work in imitation learning can be split into two categories: (1) behavioral cloning (*BC*) (Bain & Sammut, 1995; Pomerleau, 1989), and (2) inverse reinforcement learning (*IRL*) (Ng et al., 2000; Abbeel & Ng, 2004; Ziebart et al., 2008; Fu et al., 2017).

Behavioral cloning methods use supervised learning as a means by which to find a direct mapping from states to actions. *BC* approaches have been used to successfully learn many different tasks such as navigation for quadrotors (Giusti et al., 2016) or autonomous ground vehicles (Bojarski et al., 2016). Inverse reinforcement learning (*IRL*) techniques, on the other hand, seek to learn the demonstrator’s cost function and then use this learned cost function in order to learn an imitation policy through *RL* techniques. *IRL* methods have been used for interesting tasks such as dish placement and pouring (Finn et al., 2016). To the best

of our knowledge, the current state of the art in imitation learning is an *IRL*-based technique called generative adversarial imitation learning (*GAIL*) (Ho & Ermon, 2016). *GAIL* uses generative adversarial networks (*GANs*) (Goodfellow et al., 2014) as a means by which to bring the distribution of state and action pairs of the imitator and the demonstrator closer together.

Most existing imitation learning approaches require demonstrations that include the expert actions. However, these actions are not always observable, and sometimes it is more practical to be able to imitate state-only demonstrations. One step towards this goal is the work of Finn et al. (2017) where a meta-learning imitation learning method is proposed that enables a robot to reuse past experience and learn new skills from a single demonstration. In particular, raw pixel videos are used as the source of demonstration information. However, it is still assumed that the expert actions are available during meta-training; the requirement for actions is only lifted at test time when learning the new task.

One way to approach the aforementioned problem is to “learn to imitate” (as opposed to imitation learning), i.e., by doing some preprocessing, enable the agent to follow a single demonstration exactly. Two such approaches are proposed by Nair et al. (2017) and Pathak et al. (2018). These methods first learn an inverse dynamics model through self-supervised exploration, and then use it to infer the demonstrator’s action at each step and perform that in the environment. These approaches mimic the one demonstration that they are exposed to exactly (as opposed to learning and generalizing a task from multiple different demonstrations).

A second approach to imitation from action-free demonstrations is *behavioral cloning from observation* (*BCO*) (Torabi et al., 2018). This method also learns an inverse dynamics model through self-supervised exploration which is then used to infer actions from demonstrations. The problem is then treated as a regular imitation learning problem, and behavioral cloning is used to learn an imitation policy that maps states to the inferred actions. Therefore, this method is able to learn and generalize from different demonstrations but, since it is based on behavioral cloning, it may suffer from the well-studied compounding error caused by covariate shift (Ross & Bagnell, 2010; Ross et al., 2011; Laskey et al., 2016).

A third class of techniques that is able to perform imitation learning without requiring knowledge of actions includes those that first focus on learning a representation of the task and then use an *RL* method with a predefined surrogate reward over that representation. For example, Gupta et al. (2017) have proposed an invariant feature space to transfer skills between agents with different embodiments, Liu et al. (2017) have presented a network architecture

which is capable of handling differences in viewpoints and contexts between the imitator and the demonstrator, and [Sermanet et al. \(2017\)](#) have proposed a time-contrastive network which is invariant to both different embodiments and viewpoints. While these techniques represent significant advances in representation learning, each of them uses the same surrogate reward function, i.e., the proximity of the imitator’s and demonstrator’s encoded representation at each time step. One of the downsides of this reward function is that each provided demonstration needs to be time-aligned, i.e., at every time step, each demonstration needs to have advanced to the same point of the task. Another approaches developed by [Merel et al.](#) and [Henderson et al.](#) aim to imitate the state distribution of the expert. However, the state distribution does not represent the demonstrator policy and the learned policy may fail in tasks such as the cyclic ones. Moreover, these approaches have thus far focused mostly on experimentation and less on the theoretical underpinnings of the control problem. In our work, we propose a new algorithm to remove the constraints mentioned above, and also provide theoretical analysis of this approach.

3. Preliminaries

Notation We consider agents within the framework of Markov decision processes (MDPs). In this framework, \mathcal{S} and \mathcal{A} are the state and action spaces, respectively. An agent at a particular state $s \in \mathcal{S}$, chooses an action $a \in \mathcal{A}$, based on a policy $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ and transitions to state s' with probability of $P(s'|s, a)$ that is predefined by the environment transition dynamics. In this process, the agent gets feedback $c(s, a)$ which is coming from a cost function $c : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. In this paper, $\overline{\mathbb{R}}$ means the extended real numbers $\mathbb{R} \cup \{+\infty\}$ and expectation over a policy means the expectation over all the trajectories that it generates.

Inverse Reinforcement Learning (IRL) As described earlier, one general approach to imitation learning is based on IRL. The first step of this approach is to learn a cost function based on the given state-action demonstrations. This cost function is learned such that it is minimal for the trajectories demonstrated by the expert and maximal for every other policy ([Abbeel & Ng, 2004](#)). However, since the problem is underconstrained — many policies can lead to the same (demonstrated) trajectories — another constraint is usually assigned as well which chooses the policy that has the maximum entropy. This method is called maximum entropy inverse reinforcement learning (*MaxEnt IRL*) ([Ziebart et al., 2008](#)). A very general form of this framework can be described as

$$IRL_{\psi}(\pi_E) = \arg \max_{c \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}} - \psi(c) + \left(\min_{\pi \in \Pi} - \lambda_H H(\pi) + \mathbb{E}_{\pi} [c(s, a)] - \mathbb{E}_{\pi_E} [c(s, a)] \right), \quad (1)$$

where $\psi(c) : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \rightarrow \overline{\mathbb{R}}$ is a convex cost function regularizer, π_E is the expert policy, Π is the space of all the possible policies, and $H(\pi)$ and λ_H are the entropy function of the policy π and its weighting parameter respectively. The output here is the desired cost function. The second step of this framework is to input the learned cost function into a standard reinforcement learning problem. An entropy-regularized version of *RL* can be described as

$$RL(c) = \arg \min_{\pi \in \Pi} - \lambda_H H(\pi) + \mathbb{E}_{\pi} [c(s, a)], \quad (2)$$

which aims to find a policy that minimizes the cost function and maximizes the entropy.

Generative Adversarial Imitation Learning (GAIL)

Recently, [Ho & Ermon](#) have shown that by considering a specific function as the cost regularizer $\psi(c)$, the described pipeline ((1) and (2)) can be solved instead as

$$\min_{\pi \in \Pi} \max_{D \in (0,1)^{\mathcal{S} \times \mathcal{A}}} - \lambda_H H(\pi) + \mathbb{E}_{\pi} [\log(D(s, a))] + \mathbb{E}_{\pi_E} [\log(1 - D(s, a))], \quad (3)$$

where $D : \mathcal{S} \times \mathcal{A} \rightarrow (0, 1)$ is a classifier trained to discriminate between the state-action pairs that arise from the demonstrator and the imitator. Excluding the entropy term, the loss function in (3) is similar to the loss of generative adversarial networks ([Goodfellow et al., 2014](#)). Instead of first learning the cost function and then learning the policy on top of that, this method directly learns the optimal policy by bringing the distribution of the state-action pairs of the imitator as close as possible to that of the demonstrator.

4. A General Framework for Imitation from Observation

In *IRL*, both states and actions are available and the goal is to find a cost function that on average has a smaller value for the trajectories generated by the expert policy compared to the ones generated by any other policy. In the case of imitation from observation, however, the demonstrations that the agent receives are limited to the expert’s state-only trajectories. In the context of the *IRL*-based approaches to imitation learning discussed above, this lack of action information makes it impossible to calculate the $\mathbb{E}_{\pi_E} [c(s, a)]$ term in (1). Consequently, none of the approaches described in Section 3 is directly applicable in this setting.

In imitation from observation, the goal is for the imitator to perform similarly to the expert in the environment, i.e., for the actions of the demonstrator and imitator to have the same effect on the environment (performing the task), rather than taking exactly the same actions. Therefore, instead of characterizing the cost signal as a function of states and actions $c : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, we define them as a function of the

state transitions $c : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$. Based on this characterization, we formulate inverse reinforcement learning from observation as

$$\begin{aligned} \text{IRLfo}_\psi(\pi_E) = & \arg \max_{c \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}} -\psi(c) + \left(\min_{\pi \in \Pi} \mathbb{E}_\pi [c(s, s')] \right) \\ & - \mathbb{E}_{\pi_E} [c(s, s')], \end{aligned} \quad (4)$$

which outputs \tilde{c} . Note that in (4) we ignore the entropy term so as to simplify the theoretical analysis presented in Section 5. Evidence from Ho & Ermon suggests that doing so is fine from an empirical perspective (they set $\lambda_H = 0$ in more than 80% of their successful experiments). We leave detailed analysis of the effect of this term to future work. From a high-level perspective, in imitation from observation, the goal is to enable the agent to extract what the task is by observing some state sequences. Intuitively, this extraction is possible because we expect the beneficial state transitions for any given task to form a low-dimensional manifold within the $\mathcal{S} \times \mathcal{S}$ space. Thus, the intuition behind our definition of the cost function is to penalize based on how close each transition is to that manifold.

Now using an *RL* algorithm for \tilde{c} amounts to solving:

$$\text{RL}(\tilde{c}) = \arg \min_{\pi \in \Pi} \mathbb{E}_\pi [\tilde{c}(s, s')] \quad (5)$$

where the output, $\tilde{\pi}$, is the imitation policy.

5. Generative Adversarial Imitation from Observation

Having developed the general framework in (4), we now propose a specific algorithm, generative adversarial imitation from observation (*GAIfo*). To this end, we first define the state-transition occupancy measure, $\rho_\pi^s : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ as

$$\rho_\pi^s(s_i, s_j) = \sum_a P(s_j | s_i, a) \pi(a | s_i) \sum_{t=0}^{\infty} \gamma^t P(s_t = s_i | \pi). \quad (6)$$

This occupancy measure corresponds to the distribution of state transitions that an agent encounters when using policy π . We define the set of valid state-transition occupancy measures as $\mathcal{P}^s \triangleq \{\rho_\pi^s : \pi \in \Pi\}$.

We now introduce a proposition which is the foundation of our algorithm. In the following proposition we use the convex conjugate concept which is defined as follows: for a function $f : X \rightarrow \mathbb{R}$, the convex conjugate $f^* : X^* \rightarrow \mathbb{R}$ is defined as $f^*(x^*) \triangleq \sup_{x \in X} \langle x^*, x \rangle - f(x)$.

Proposition 5.1. *$\text{RL} \circ \text{IRLfo}_\psi(\pi_E)$ and $\arg \min_{\pi \in \Pi} \psi^*(\rho_\pi^s - \rho_{\pi_E}^s)$ induce policies that have the same state-transition occupancy measure, $\rho_{\tilde{\pi}}^s$.*

In the rest of this section, we prove this proposition and then by choosing a specific regularizer, we present our algorithm.

At the end we propose a practical implementation of the algorithm. To prove the proposition, we first define another problem, $\overline{\text{RL}} \circ \overline{\text{IRLfo}}_\psi(\pi_E)$, and argue that it outputs a state-transition occupancy measure which is the same as ρ_π^s induced by $\text{RL} \circ \text{IRLfo}_\psi(\pi_E)$. We define

$$\begin{aligned} \overline{\text{IRLfo}}_\psi(\pi_E) = & \arg \max_{c \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}} \left(\min_{\rho_\pi^s \in \mathcal{P}^s} \sum_{s, s'} \rho_\pi^s(s, s') c(s, s') \right) \\ & - \sum_{s, s'} \rho_{\pi_E}^s(s, s') c(s, s') - \psi(c), \end{aligned} \quad (7)$$

where, the output is a cost function \bar{c} . Note that, $\mathbb{E}_\pi [c(s, s')] = \sum_{s, s'} \rho_\pi^s(s, s') c(s, s')$ so (4) and (7) are similar except that the former is optimized over $\pi \in \Pi$ and the latter over $\rho_\pi^s \in \mathcal{P}^s$. If we consider using an *RL* method to find a state-transition occupancy measure under \bar{c} , (5) can be rewritten as

$$\overline{\text{RL}}(\bar{c}) = \min_{\rho_\pi^s \in \mathcal{P}^s} \sum_{s, s'} \rho_\pi^s(s, s') \bar{c}(s, s'), \quad (8)$$

which would now output the desired state-transition occupancy measure $\bar{\rho}_\pi^s$.

Lemma 5.1. *$\overline{\text{RL}} \circ \overline{\text{IRLfo}}_\psi(\pi_E)$ outputs a state-transition occupancy measure, $\bar{\rho}_\pi^s$, which is the same as ρ_π^s induced by $\text{RL} \circ \text{IRLfo}_\psi(\pi_E)$.*

Proof. From the definition of \mathcal{P}^s , the mapping from Π to \mathcal{P}^s is surjective, i.e., for every $\rho_\pi^s \in \mathcal{P}^s$, there exists at least one $\pi \in \Pi$. Therefore, we can say $\bar{\rho}_\pi^s = \rho_\pi^s$ (where $\tilde{\pi}$ and $\bar{\rho}_\pi^s$, as already defined, are the outputs of (5) and (8), and ρ_π^s is the state-transition occupancy measure that corresponds to $\tilde{\pi}$). Therefore, solving $\text{RL} \circ \text{IRLfo}_\psi(\pi_E)$ results in the same ρ_π^s as applying $\overline{\text{RL}}$ using the cost function returned by $\overline{\text{IRLfo}}$ in (7). \square

Note that, in this lemma, the returned policies from these two problems are not necessarily the same. The reason is that the mapping from Π to \mathcal{P}^s is not injective, i.e., there could be one or multiple $\pi \in \Pi$ that corresponds to the same $\rho_\pi^s \in \mathcal{P}^s$. Consequently, it is not necessarily the case that a policy that gives rise to $\bar{\rho}_\pi^s$ is the same as $\tilde{\pi}$. However, as we discussed in the previous section, in imitation from observation, we are primarily concerned with the effect of the policy on the environment so this situation is acceptable.

Now we introduce another lemma that helps us in the proof of Proposition 5.1.

Lemma 5.2. *$\overline{\text{RL}} \circ \overline{\text{IRLfo}}_\psi(\pi_E) = \arg \min_{\rho_\pi^s \in \mathcal{P}^s} \psi^*(\rho_\pi^s - \rho_{\pi_E}^s)$*

This lemma is proven in the appendix ¹ using the min-max principle (Millar, 1983). Thus far, by combin-

¹The appendix is anonymously presented <https://tinyurl.com/ybkn8v7n> <https://tinyurl.com/ybkn8v7n>

ing Lemmas 5.1 and 5.2, we can conclude that ρ_π^s induced by $\text{RL} \circ \text{IRLfo}_\psi(\pi_E)$ is the same as the output of $\arg \min_{\rho_\pi^s \in \mathcal{P}^s} \psi^*(\rho_\pi^s - \rho_{\pi_E}^s)$. Now, we only need one more step to prove Proposition 5.1:

Lemma 5.3. $\arg \min_{\pi \in \Pi} \psi^*(\rho_\pi^s - \rho_{\pi_E}^s)$ is a policy that has a state-transition occupancy measure that is the same as the output of $\arg \min_{\rho_\pi^s \in \mathcal{P}^s} \psi^*(\rho_\pi^s - \rho_{\pi_E}^s)$.

The proof of Lemma 5.3 is similar to that of Lemma 5.1. Now based on Lemmas 5.1, 5.2, and 5.3 we can conclude that Proposition 5.1 holds.

Having proved this proposition, we can solve $\arg \min_{\pi \in \Pi} \psi^*(\rho_\pi^s - \rho_{\pi_E}^s)$ instead of $\text{RL} \circ \text{IRLfo}_\psi(\pi_E)$. To this end, we consider the generative adversarial regularizer

$$\psi_{GA}(c) \triangleq \begin{cases} \mathbb{E}_{\pi_E}[g(c(s, s'))] & \text{if } c < 0 \\ +\infty & \text{otherwise} \end{cases} \quad (9)$$

where

$$g(x) = \begin{cases} -x - \log(1 - e^x) & \text{if } x < 0 \\ +\infty & \text{otherwise} \end{cases} \quad (10)$$

which is a closed, proper, convex function and has convex conjugate

$$\begin{aligned} \psi_{GA}^*(\rho_\pi^s - \rho_{\pi_E}^s) &= \max_{D \in (0,1)^{\mathcal{S} \times \mathcal{S}}} \sum_{s,s'} \rho_\pi^s(s, s') \log(D(s, s')) + \\ &\quad \rho_{\pi_E}^s(s, s') \log(1 - D(s, s')), \end{aligned} \quad (11)$$

where $D : \mathcal{S} \times \mathcal{S} \rightarrow (0, 1)$ is a discriminative classifier. A similar convex conjugate is derived in Ho & Ermon; However, for the sake of completeness, we prove the properties claimed for (9) and show that (11) is its convex conjugate in the appendix.²

Using the above, the imitation from observation problem can be solved as:

$$\begin{aligned} \min_{\pi \in \Pi} \psi_{GA}^*(\rho_\pi^s - \rho_{\pi_E}^s) &= \min_{\pi \in \Pi} \max_{D \in (0,1)^{\mathcal{S} \times \mathcal{S}}} \mathbb{E}_\pi[\log(D(s, s'))] + \\ &\quad \mathbb{E}_{\pi_E}[\log(1 - D(s, s'))] \end{aligned} \quad (12)$$

We can see that the loss function in (12) is similar to the generative adversarial loss. We can connect this to general GANs if we interpret the expert's demonstrations as the real data, and the data coming from the imitator as the generated

²This proof closely follows the proofs of Proposition A.1. and Corollary A.1.1. of Ho & Ermon and it is included here for the sake of completeness. The only substantive difference is that in our case we consider state-transition occupancy measure (s, s') instead of (s, a) .

Algorithm 1 *GAIfo*

- 1: Initialize parametric policy π_ϕ with random ϕ
 - 2: Initialize parametric discriminator D_θ with random θ
 - 3: Obtain state-only expert demonstration trajectories $\tau_E = \{(s, s')\}$
 - 4: **while** Policy Improves **do**
 - 5: Execute π_ϕ and store the resulting state transitions $\tau = \{(s, s')\}$
 - 6: Update D_θ using loss
$$-\left(\mathbb{E}_\tau[\log(D_\theta(s, s'))] + \mathbb{E}_{\tau_E}[\log(1 - D_\theta(s, s'))]\right)$$
 - 7: Update π_ϕ by performing *TRPO* updates with reward function
$$-\left(\mathbb{E}_\tau[\log(D_\theta(s, s'))]\right)$$
 - 8: **end while**
-

data. The discriminator seeks to distinguish the source of the data, and the imitator policy (i.e., the generator) seeks to fool the discriminator to make it look like the state transitions it generates are coming from the expert. The entire process can be interpreted as bringing the distribution of the imitator's state transitions closer to that of the expert. We call this process Generative Adversarial Imitation from Observation (*GAIfo*).

6. Practical Implementation

Based on the preceding analysis, we now specify our practical implementation of the *GAIfo* algorithm. We represent the discriminator, D , using a multi-layer perceptron with parameters θ that takes as input a state transition and outputs a value between 0 and 1. We represent the policy, π , using a multi-layer perceptron with parameters ϕ that takes as input a state and outputs an action. We begin by randomly initializing each of these networks, after which the imitator selects an action according to π_ϕ and executes that action. This action leads to a new state, and we feed both this state transition and the entire set of expert state transitions to the discriminator. The discriminator is updated using the Adam optimization algorithm (Kingma & Ba, 2014), with cross-entropy loss that seeks to push the output for expert state transitions closer to 0 and the imitator's state transitions closer to 1. After the discriminator update, we perform trust region policy optimization (*TRPO*) (Schulman et al., 2015) to improve the policy using a reward function that encourages state transitions that yield small outputs from the discriminator (i.e., those that appear to be from the demonstrator). This process continues until convergence. The algorithm is shown in Algorithm 1 and the framework is summarized in Figure 1.

The implementation described above is only effective

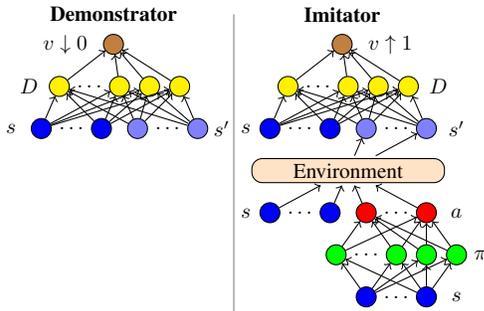


Figure 1. A diagrammatic representation of *GAIfo*. On the left, s (dark blue) and s' (light blue) are the state and next-state features in a demonstration transition, respectively. On the right, dark blue neurons represent the imitator’s states. Based on policy π (green), it performs action a (red) in the environment, and encounters the next-state s' (light blue). We aim to find a policy that generates state-transitions close to the demonstrations. To this end, we iteratively train the discriminator and the policy. The discriminator is trained in a way to output values v (brown) close to zero for the data coming from the expert (left) and close to one for the data coming from the imitator (right). The policy is trained to generate state-transitions close to the demonstrations so that the discriminator is not able to distinguish them from the demonstrations.

for cases in which the demonstration consists of low-dimensional state representations. In particular, the imitation policy maps a single state to the imitating action and the reward function operates on a single state transition. This approach is feasible for cases in which (a) the states can be assumed to be fully-observable, and (b) the system is strictly Markovian. However, when considering visual state representations, neither of these assumptions is necessarily valid. Therefore, agents operating in such state spaces are typically provided instead a recent state history. This is useful because, for example, having knowledge about the velocity of the agent at each time step is important in order to select the correct action, and velocity information is not available when considering a single image. Therefore, we propose here a second implementation of *GAIfo* that enables imitation from visual demonstration data. It modifies the implementation used for low-dimensional state representations by adding convolutional layers and using images from multiple time steps as the input to the generator and discriminator. This implementation is summarized in Figure 2.

7. Experimental Setup and Implementation Details

We evaluate our algorithm in domains from OpenAI Gym (Brockman et al., 2016) based on the Pybullet simulator (Coumans & Bai, 2016-2017). In each of the domains, we used trust region policy optimization (*TRPO*) (Schulman

et al., 2015) to train the expert agents, and we recorded the demonstrations using the resulting policy.

The results shown in the figures are the average over ten independent trials. We compare our algorithm against three baselines:

- **Behavioral Cloning from Observation (BCO)**(Torabi et al., 2018): *BCO* first learns an inverse dynamics model through self-supervised exploration, and then uses that model to infer the missing actions from state-only demonstrated trajectories. *BCO* then uses the inferred actions to learn an imitation policy using conventional behavioral cloning.
- **Time Contrastive Networks (TCN)**(Sermanet et al., 2017): *TCNs* use a triplet loss to train a neural network to learn an encoded form of the task at each time step. This loss function brings the states that occur in a small time-window closer together in the embedding space and pushes the ones from distant time-steps far apart. A reward function is then defined as the Euclidean distance between the embedded demonstration and the embedded agent’s state at each time step. The imitation policy is learned using *RL* techniques that seek to optimize this reward function.
- **Generative Adversarial Imitation Learning (GAIL)** (Ho & Ermon, 2016): This method is as specified in Section 3. Note that, this method has access to the demonstrator’s actions while the others do not.

8. Results and Discussion

In this section, we present the results of the two sets of experiments described above.

8.1. Low-dimensional State Representations

Figure 3 illustrate the comparative performance of *GAIfo* in our experimental domains using the low-dimensional state representations. We can see that, for the domains considered here, *GAIfo* (a) performs very well compared to other IFO techniques, and (b) is surprisingly comparable to *GAIL* even though *GAIfo* lacks access to explicit action information.

Figure 3 compares the final performance of the imitation policies learned by different algorithms. We can clearly see that *GAIfo* outperforms the other imitation from observation algorithms by a large margin in most of the experiments. For the InvertedDoublePendulum domain, we can see that the *TCN* method does not perform well at all. We hypothesize that this is the case because *TCN* relies on time synchronization in order to find the imitating policy, i.e., it learns what the state should be at each time step. However, successfully

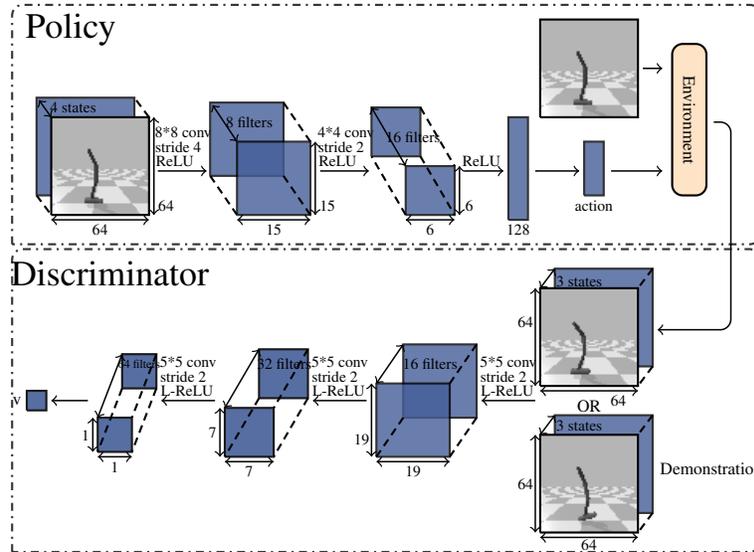


Figure 2. A diagrammatic representation of our *GAIfo* implementation for processing visual state representations. A stack of 4 grayscale images from $t - 3$ to t (t being the current time-step) enters the policy CNN (top left). The policy outputs an action that the agent takes in the environment and goes to the next state in time $t + 1$ (top right). A stack of 3 grayscale images from $t - 1$ to $t + 1$ of the agent is prepared along with a stack of 3 consecutive state images (grayscale) of the demonstrator (bottom right). When data from the imitation policy is provided, the stack from the imitator enters the discriminator and outputs the reward for taking that action (bottom left). This reward value is then used to both update the policy using *TRPO* and also update the discriminator using supervised learning (to drive the value closer to zero). When data from the demonstrator is provided, the stack from the demonstrator enters the discriminator and outputs a value which is then used to update the discriminator (to drive the value closer to one).

performing the InvertedDoublePendulum task requires the agent to simply keep the pendulum upright, and requiring it to time synchronize with the demonstrator may be too restrictive a requirement. *BCO*, on the other hand, performs very well in this domain, which demonstrates that, here, the inverse dynamics model learned by *BCO* is accurate and that the compounding error problem is negligible. We can see that *GAIfo* also performs very well here, achieving performance similar to that of the expert, which shows that the algorithm has been able to extract the goal of the task and find a reasonable cost function from which to learn the policy.

For the InvertedPendulumSwingup domain, we can see that *TCN* again does not perform well, perhaps because the goal of the task is not well-represented in the encoding-learning phase. *BCO* also does not perform well. We hypothesize that this is the case because of the compounding error problem since performing this task successfully is contingent on taking several specific actions consecutively – deviation from those actions would cause the pendulum to drop down and not reach the goal. *GAIfo* and *GAIL*, on the other hand, perform as well as the expert, which reveals that these algorithms have successfully extracted the goal and learned the task.

For both the Hopper and Walker2D domains, it can be seen that, again, *TCN* does not work well. We posit that this

might be due to the fact that these tasks require behavior that is cyclic in nature, i.e., the expert demonstrations contain repeated states. Because *TCN* learns a time-dependent representation of the task, it cannot appropriately handle this periodicity and, therefore, the learned representations are not sufficient. *GAIfo*, however, learns a distribution of the state transitions that is not time-dependent; therefore, periodicity does not affect its performance. *BCO* also does not perform well in either of these two domains, perhaps again due to the compounding error problem. Learning in these domains has two steps: first, the agent needs to learn to stand, and then the agent needs to learn to walk or hop. With *BCO*, it would seem that the imitating agent begins to deviate from the expert early in the task, and this early deviation ultimately leads to the imitating agent being unable to learn the secondary walking and hopping behaviors. *GAIfo*, on the other hand, does not suffer from this issue because it learns by executing its own policy in the environment (on-policy learning) and is therefore able to address deviation from the expert during the learning process.

8.2. Visual State Representations

In this section, we discuss the results of the experiments performed on the cases where the states are represented using the raw visual data. Figure 4 illustrates the comparison be-

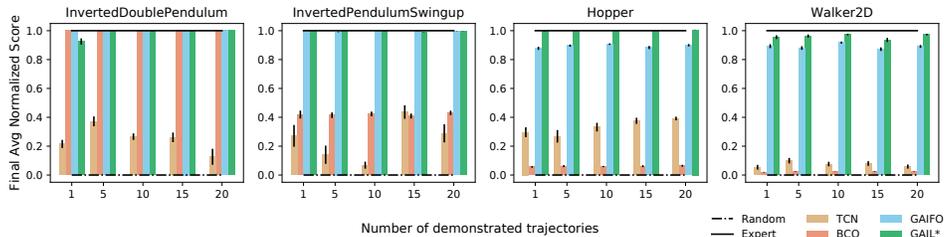


Figure 3. Performance of algorithms in low-dimensional experiments with respect to the number of demonstration trajectories. Rectangular bars and error bars represent mean return and standard deviations, respectively. For comparison purposes, we have scaled all the performances such that a random and the expert policy score 0.0 and 1.0, respectively. **GAIL* has access to action information.

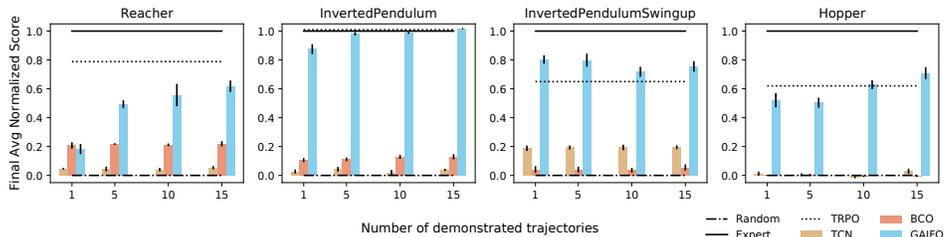


Figure 4. Performance of algorithms in visual experiments with respect to the number of demonstration trajectories. Rectangular bars and error bars represent mean return and standard deviations, respectively. For comparison purposes, we have scaled all the performances such that a random and the expert policy score 0.0 and 1.0, respectively.

tween the performance of *GAIfO*, *BCO* and *TCN*.³ In these experiments, like the ones done using the lower-dimensional state representations, the expert is trained with *TRPO* using low-level state features, and the quantities 0 and 1 represent the performance of a random agent and the expert, respectively. The demonstrations, though, consist of visual recordings using the trained policy. Accordingly, for a more-representative baseline, we also learn a policy with *TRPO* using visual states only (as opposed to the low-dimensional state observations) and represent the performance of that agent using a black dotted line on the plots. This line is important in our comparison because it shows (everything being similar to *IfO* methods) what would have been the resulting performance if the agent had access to the reward. Figure 4 shows that *GAIfO* outperforms other approaches by a large margin.

It is interesting to notice that, even though *GAIfO* (like the other *IfO* techniques) does not achieve the performance of the expert agent (solid line), it *does* achieve the performance of the *TRPO*-trained agent that used visual state representations. This suggests that, in these cases, the drop in imitation performance is perhaps due to a fundamental limitation of learning the task from visual data (i.e., partial state observability).

³Here, we do not compare against *GAIL* because doing so would require a drastic change to the structure of its discriminator in order to process raw visual data, i.e., the discriminator would need to be altered to appropriately mix action and visual data.

Finally, it can be seen that *BCO* does not perform well in any of the domains, perhaps due to (a) the complexity of learning dynamics models over visual states, and (b) compounding error. *TCN* also does not work well, perhaps due to the demonstrations not being time-synchronized.

9. Conclusion and Future Work

In this paper, we presented a general framework for imitation from observation ($RL \circ IRLfO_{\psi}(\pi_E)$) and then proposed a specific algorithm (*GAIfO*) for doing so. *GAIfO* removes the need for several restrictive assumptions that are required for some other *IfO* techniques, including the need for multiple demonstrations to be time-synchronized. Moreover, the on-policy nature of *GAIfO* allows it to avoid the compounding error problem experienced by more brittle imitation techniques. The result is an approach that is able to find better imitation policies without the need for action information, and is also able to find imitation policies that perform very close to those found by techniques that do have access to this information.

Regarding future work, note that, in our analysis, we did not consider policy entropy terms in either the *IRL* step, nor in the *RL* step. Therefore, it would be interesting to include entropy in these equations – as has been shown to be beneficial in some cases (Haarnoja et al., 2017; 2018) – and investigate its effects on the overall problem and results as has been shown to be beneficial in some cases.

Acknowledgements

This work has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by grants from the National Science Foundation (IIS-1637736, IIS-1651089, IIS-1724157), the Office of Naval Research (N00014-18-2243), Future of Life Institute (RFP2-000), Army Research Lab, DARPA, Intel, Raytheon, and Lockheed Martin. Peter Stone serves on the Board of Directors of Cogitai, Inc. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

References

- Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first International Conference on Machine Learning*, pp. 1. ACM, 2004.
- Argall, B. D., Chernova, S., Veloso, M., and Browning, B. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.
- Bain, M. and Sammut, C. A framework for behavioral cloning. *Machine Intelligence 14*, 1995.
- Billard, A., Calinon, S., Dillmann, R., and Schaal, S. Robot programming by demonstration. In *Springer handbook of robotics*, pp. 1371–1394. Springer, 2008.
- Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.
- Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym, 2016.
- Coumans, E. and Bai, Y. Pybullet, a python module for physics simulation for games, robotics and machine learning. 2016-2017. URL <http://pybullet.org/>.
- Finn, C., Levine, S., and Abbeel, P. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*, pp. 49–58, 2016.
- Finn, C., Yu, T., Zhang, T., Abbeel, P., and Levine, S. One-shot visual imitation learning via meta-learning. In *Conference on Robot Learning*, pp. 357–368, 2017.
- Fu, J., Luo, K., and Levine, S. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.
- Giusti, A., Guzzi, J., Cireşan, D. C., He, F.-L., Rodríguez, J. P., Fontana, F., Faessler, M., Forster, C., Schmidhuber, J., Di Caro, G., et al. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robotics and Automation Letters*, 1(2):661–667, 2016.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.
- Gupta, A., Devin, C., Liu, Y., Abbeel, P., and Levine, S. Learning invariant feature spaces to transfer skills with reinforcement learning. *arXiv preprint arXiv:1703.02949*, 2017.
- Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. *arXiv preprint arXiv:1702.08165*, 2017.
- Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.
- Henderson, P., Chang, W.-D., Bacon, P.-L., Meger, D., Pineau, J., and Precup, D. Optiongan: Learning joint reward-policy options using generative adversarial inverse reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- Ho, J. and Ermon, S. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pp. 4565–4573, 2016.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Laskey, M., Staszak, S., Hsieh, W. Y.-S., Mahler, J., Pokorny, F. T., Dragan, A. D., and Goldberg, K. Shiv: Reducing supervisor burden in dagger using support vectors for efficient learning from demonstrations in high dimensional state spaces. In *International Conference on Robotics and Automation (ICRA)*, pp. 462–469. IEEE, 2016.
- Liu, Y., Gupta, A., Abbeel, P., and Levine, S. Imitation from observation: Learning to imitate behaviors from raw video via context translation. *arXiv preprint arXiv:1707.03374*, 2017.
- Merel, J., Tassa, Y., Srinivasan, S., Lemmon, J., Wang, Z., Wayne, G., and Heess, N. Learning human behaviors from motion capture by adversarial imitation. *arXiv preprint arXiv:1707.02201*, 2017.
- Millar, P. W. The minimax principle in asymptotic statistical theory. In *Ecole d’Eté de Probabilités de Saint-Flour XI-1981*, pp. 75–265. Springer, 1983.

- Nair, A., Chen, D., Agrawal, P., Isola, P., Abbeel, P., Malik, J., and Levine, S. Combining self-supervised learning and imitation for vision-based rope manipulation. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pp. 2146–2153. IEEE, 2017.
- Ng, A. Y., Russell, S. J., et al. Algorithms for inverse reinforcement learning. In *ICML*, pp. 663–670, 2000.
- Pathak, D., Mahmoudieh, P., Luo, G., Agrawal, P., Chen, D., Shentu, Y., Shelhamer, E., Malik, J., Efros, A. A., and Darrell, T. Zero-shot visual imitation. *arXiv preprint arXiv:1804.08606*, 2018.
- Pomerleau, D. A. Alvin: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems*, pp. 305–313, 1989.
- 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, 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, 2011.
- Russell, S. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on computational learning theory*, pp. 101–103. ACM, 1998.
- Schaal, S. Learning from demonstration. In *Advances in Neural Information Processing Systems*, pp. 1040–1046, 1997.
- Schaal, S., Ijspeert, A., and Billard, A. Computational approaches to motor learning by imitation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 358(1431):537–547, 2003.
- Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015.
- Sermanet, P., Lynch, C., Hsu, J., and Levine, S. Time-contrastive networks: Self-supervised learning from multi-view observation. *arXiv preprint arXiv:1704.06888*, 2017.
- Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.
- Torabi, F., Warrnell, G., and Stone, P. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018.
- Zhou, L., Xu, C., and Corso, J. J. Towards automatic learning of procedures from web instructional videos. *arXiv preprint arXiv:1703.09788*, 2017.
- Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.