Machine Learning (2023) 112:2263–2296
https://doi.org/10.1007/s10994-022-06273-x

Model-free inverse reinforcement learning with multi-intention, unlabeled, and overlapping demonstrations

Ariyan Bighashdel[1] · **Pavol Jancura**[1] · **Gijs Dubbelman**[1]

Received: 4 November 2021 / Revised: 18 July 2022 / Accepted: 20 October 2022 /
Published online: 30 November 2022

Abstract
In this paper, we define a novel inverse reinforcement learning (IRL) problem where the demonstrations are multi-intention, i.e., collected from multi-intention experts, unlabeled, i.e., without intention labels, and partially overlapping, i.e., shared between multiple intentions. In the presence of overlapping demonstrations, current IRL methods, developed to handle multi-intention and unlabeled demonstrations, cannot successfully learn the underlying reward functions. To solve this limitation, we propose a novel clustering-based approach to disentangle the observed demonstrations and experimentally validate its advantages. Traditional clustering-based approaches to multi-intention IRL, which are developed on the basis of model-based Reinforcement Learning (RL), formulate the problem using parametric density estimation. However, in high-dimensional environments and unknown system dynamics, i.e., model-free RL, the solution of parametric density estimation is only tractable up to the density normalization constant. To solve this, we formulate the problem as a mixture of logistic regressions to directly handle the unnormalized density. To research the challenges faced by overlapping demonstrations, we introduce the concepts of *shared pair*, which is a state-action pair that is shared in more than one intention, and *separability*, which resembles how well the multiple intentions can be separated in the joint state-action space. We provide theoretical analyses under the global optimality condition and the existence of shared pairs. Furthermore, we conduct extensive experiments on four simulated robotics tasks, extended to accept different intentions with specific levels of separability, and a synthetic driver task developed to directly control the separability. We evaluate the existing baselines on our defined problem and demonstrate, theoretically and experimentally, the advantages of our clustering-based solution, especially when the separability of the demonstrations decreases.

Keywords Inverse reinforcement learning · Multi-intention · Model-free reinforcement learning · Unlabeled demonstrations · Overlapping demonstrations · Mixture of logistic regressions

Communicated by Bo Han, Tongliang Liu, Quanming Yao, Mingming Gong, Gang Niu, Ivor W. Tsang, Masashi Sugiyama.

Extended author information available on the last page of the article

1 Introduction

In the last few decades, there has been a surge of interest in the task of learning from demonstrations (LfD) in various domains (Belogolovsky et al., 2021; Chen et al., 2020; Kangasrääsiö & Kaski, 2018; Neu & Szepesvári, 2009). In an LfD task, the agent learns a mapping, called *policy*, from the world states to actions, solely by observing the experts' demonstrations, which are various sequences of state-action pairs. Even though the policy can be directly learned from demonstrations in a supervised learning fashion, the recovery of the *reward function* via Inverse Reinforcement Learning (IRL) has shown to provide a much more compact and generalizable description of behaviors (Ng et al., 2000).

In many IRL tasks, it can be the case that the demonstrations are collected from experts with multiple and inherently different intentions. In this situation, a single reward function is inadequate to cover all the demonstrations, and multiple reward functions are required to clearly express the differences between the multiple experts' intentions. To be able to recover multiple rewards functions from the experts' demonstrations, each demonstration should be accompanied by an extra piece of information that indicates to which intention it belongs. However, such information may not always be available in many real-world scenarios or is too expensive to be manually added. This leads to the relatively new problem of multi-intention and unlabeled demonstrations, i.e., the demonstrations are without intention labels. Even though the IRL problem with Multi-intention and Unlabeled Demonstrations, referred to as IRL-MUD, is potentially relevant to many real-world applications of IRL, a search of the literature reveals that this domain has received relatively little attention from the community.

It is commonplace to distinguish the IRL-MUD studies based on the dimensionality of the joint state-action space and the knowledge about the system dynamics. The studies either address the problems with low-dimensional spaces and known dynamics models, referred to as model-based IRL-MUD, e.g., Babes et al. (2011); Bighashdel et al. (2021), or they consider high dimensionality in the joint state-action space with unknown dynamics model, referred to as model-free IRL-MUD, e.g., Li et al. (2017); Hsiao et al. (2019). Model-based approaches normally solve the IRL-MUD problem using methods from parameter density estimation, which involves selecting a family of distributions and employing a clustering algorithm, e.g., Expectation Maximization (EM), to estimate both the reward parameters and the intention labels of the demonstrations (Babes et al., 2011; Rajasekaran et al., 2017). Despite the straightforwardness in model-based approaches, this clustering-based technique is more difficult to apply in a model-free setting, i.e., high-dimensional spaces and unknown dynamics models, which is the focus of this work, due to the two following challenges:

Estimation of the partition function. Model-based approaches assume that the experts sample the demonstrations according to a Boltzmann distribution, parameterized up to the normalization factor, i.e., partition function. Given the knowledge of system dynamics and an efficient Reinforcement Learning (RL) solver like dynamic programming, the partition function can be analytically computed in the inner loop of the reward learning algorithm (Ng et al., 2000; Ziebart et al., 2008). In model-free settings, the partition function cannot be expressed analytically, and it is computationally expensive to completely solve the RL via approximate methods in each iteration during reward learning. A practical approach, however, is to interleave the reward learning with a policy optimization procedure and adapt the sampling distribution in the policy optimization to estimate the partition function with importance weights (Finn et al., 2016). Unfortunately, this approach is shown to have

high variance in the cases where the sampling distribution fails to cover the trajectories with high rewards. When the demonstrations have only one intention, this coverage problem can be addressed by mixing the generated samples with some samples from demonstrations with an estimated distribution, e.g., a Gaussian distribution (Finn et al., 2016). However, when it comes to model-free IRL-MUD, the distribution of multi-intention demonstrations is multi-modal and, therefore, more challenging to be estimated. We later show (in Table 3) that using a fixed family of distributions like the Gaussian Mixtures Model (GMM) shows poor performance in estimating the distribution of the demonstrations.

Estimation of the posterior distribution. Since in our IRL-MUD problem the demonstrations are without intention labels, the posterior probabilities of the latent intentions, given the model parameters, need to be estimated. This estimation is normally done via Bayes' rule. In model-free IRL-MUD, the true likelihood function is not known and, at best, can be estimated with the sampling distribution of the policy optimizer. However, even with a known likelihood function, the high dimensionality of the state-action space makes the Bayes' rule impracticable to analytically compute the posterior probabilities for all demonstrations and intentions. Therefore, alternative approximate methods are required to efficiently estimate the posterior distribution of the latent intentions.

Given the two key challenges discussed above, applying a standard clustering-based technique like EM to model-free IRL-MUD remains an unsolved problem, which we aim to tackle in this work. Due to the challenges, a greater focus in the literature has been placed upon non-clustering approaches to solve the problem of model-free IRL-MUD by directly inferring the structures of the latent intention. Two non-clustering methods, namely InfoGAIL (Li et al., 2017), and IntentionGAIL (Hausman et al., 2017), have shown relative success in various illustrated experiments compared to the methods with single-intention assumptions. However, one critical question remains about their performances. Although both models propose a model-free IRL-MUD solution, it is not clear to what extent the multi-intention demonstrations need to be separable in the state-action space for these methods to work successfully. Therefore, in this work, we focus on the setting when the multi-intention and unlabeled demonstrations can be partially overlapping. We refer to this setting as the problem of model-free IRL-MUD with Overlapping demonstrations (IRL-MUD-O).

For the sake of clarity, we discuss the following scenario that is depicted in Fig. 1. Here, the demonstrations, the sequences of state-action pairs, are the paths derived by an expert

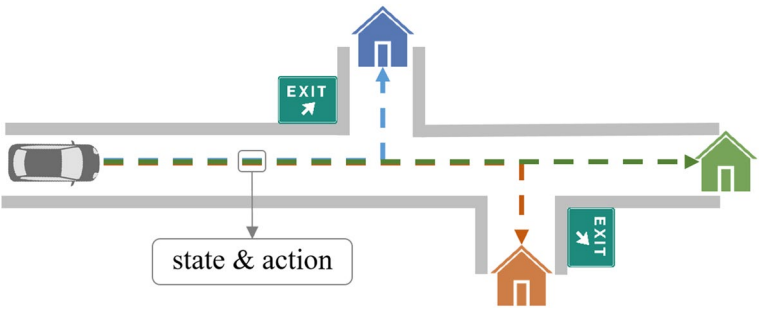


Fig. 1 An example of partially overlapping demonstrations. The three demonstrations, i.e., sequences of state-action pairs, are the paths derived by an expert driver to one of the three colored destinations: blue, green, and brown, each of which is considered as one intention (Color figure online)

driver to one of the three colored destinations: blue, green, and brown, each of which is considered as one intention. What can be clearly seen in this figure is that regardless of different intentions, the demonstrations partially overlap as the paths to different destinations go over the same road. From the state-action point of view, all the state-action pairs from the start point to the first exit are shared between the three destinations. Although the methods like InfoGAIL allocate an intention-specific policy for each intention, we later show in Sect. 4.2 that due to their specific reward structures in the policy optimization phase, a shared, perfectly imitated state-action pair receives a lower reward. This is while reward must be proportional to imitation quality rather than the degree to which the demonstrations overlap with the imitated behavior.

Previous studies have failed to demonstrate (any) convincing evidence that the non-clustering solutions, e.g., InfoGAIL and IntentionGAIL, for the problem of model-free IRL-MUD are also valid when the demonstrations overlap. We claim and experimentally validate that a suitable clustering-based approach performs better in these situations.

In summary, the primary goal of this work is twofold:

- We propose a practical clustering-based approach via EM to the problem of model-free IRL where the demonstrations are multi-intention, unlabeled, and partially overlapping;
- We demonstrate, theoretically and proposed experimentally, the benefits of our clustering-based approach to non-clustering ones, e.g., InfoGAIL and IntentionGAIL.

To accomplish our goals, we first provide suitable definitions for the problems of model-free IRL-MUD and model-free IRL-MUD-O. The latter is done by introducing the concept of the *shared pair* (Sect. 3). Then, we propose a solution for the problem of model-free IRL-MUD in Sect. 4.1 by addressing the aforementioned challenges. To address the challenge regarding the estimation of the partition function, we propose a *mixture of logistic regressions*, where the partition function is considered as a model parameter. As to the impracticality of estimating the posterior distributions, we propose a network, called *posterior network*, which directly outputs an estimate of the posterior probabilities.

In Sect. 4.2 we use our proposed clustering-based solution for the problem of model-free IRL-MUD-O and provide theoretical analysis to prove the correctness of our solution. Furthermore, we give experimental comparisons with non-clustering solutions to further demonstrate the advantages of our proposed solution (Sect. 5). This is done by (1) introducing the metric of *separability* to measure the level of overlap in the demonstrations, (2) extending well-known simulated robotics environments to cover multiple intentions with different separability levels, and (3) developing a synthetic driver environment where the separability can be directly controlled. The experimental results show that our clustering-based approach outperforms the state-of-the-art methods.

2 Related works

The research in IRL can be categorized into three frameworks: (1) Single-intention IRL, (2) Multi-intention IRL, and (3) Meta-IRL. Figure 2 illustrates an overview of each framework. Although all of these research directions target the LfD task, they fundamentally differ in their problem definition. In the following subsections, we clearly define the objectives of the three aforementioned IRL frameworks and provide an overview of the proposed methods in each research direction.

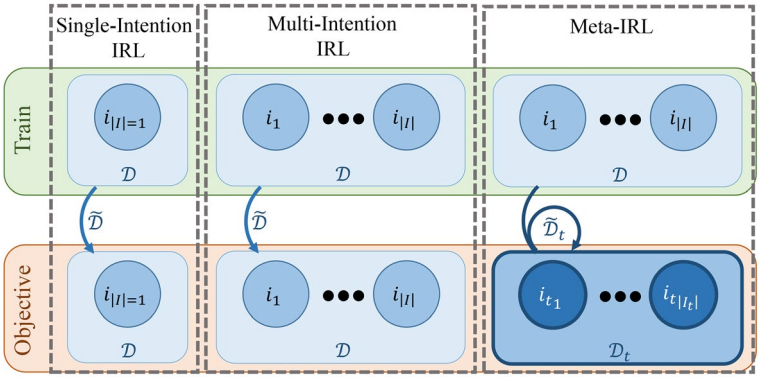


Fig. 2 The schematics of problem definition in three research directions: single-intention IRL, multi-intention IRL, and meta-IRL. Given the expert's demonstration set $\mathcal{D}$ and intention set $\mathcal{I}$, the single-intention IRL methods assume $|\mathcal{I}| = 1$ and try to minimize the distance function $d(\mathcal{D}, \tilde{\mathcal{D}})$, where $\tilde{\mathcal{D}}$ is the imitated behaviors. While minimizing the same distance function, the multi-intention IRL methods assume that the expert has multiple intentions, $|\mathcal{I}| > 1$, and the expert's demonstration set $\mathcal{D}$ typically doesn't include the intention labels. The meta-IRL methods address rapid adaptation to unknown environments that have, so far, unseen sets of target behavior $\mathcal{D}_t$ and intention $\mathcal{I}_t$. The meta-IRL methods minimize $d(\mathcal{D}_t, \tilde{\mathcal{D}}_t)$, where $\mathcal{D}_t, |\mathcal{D}_t| \ll |\mathcal{D}|$, is the new target demonstration set with the new target intention set $\mathcal{I}_t$, $i \neq i_t$ for all $i \in \mathcal{I}$ and $i_t \in \mathcal{I}_t$, and $\tilde{\mathcal{D}}_t$ is the imitated target behaviors (Color figure online)

2.1 Single-intention IRL

Given the expert's demonstration set $\mathcal{D}$ and intention set $\mathcal{I}$, the single-intention IRL methods assumes that the expert has only one intention, i.e., $|\mathcal{I}| = 1$ (see Fig. 2). The objective of the single-intention IRL methods is to perfectly imitate the expert's demonstration set $\mathcal{D}$. If the imitated behavior set can be shown by $\tilde{\mathcal{D}}$, the single-intention IRL methods try to minimize the distance function $d(\mathcal{D}, \tilde{\mathcal{D}})$, which is normally done through recovering the reward function that justifies the expert's behaviors.

A great deal of previous research into IRL task has focused on single-intention IRL. This line of research has addressed various aspects including, reward ambiguity and the problem of degeneracy (Ratliff et al., 2006; Ng et al., 2000), constraints on the demonstrations distribution (Ziebart et al., 2008), nonlinearity of the reward function (Wulfmeier et al., 2015), high dimensionality of state-action spaces (Finn et al., 2016; Fu et al., 2018; Ho & Ermon, 2016), outperforming the demonstrators (Yu et al., 2020), imperfect and noisy demonstrations (Tangkaratt et al., 2020, 2021; Wu et al., 2019), etc. Various survey articles have also extensively reviewed the challenges and future research trends in this domain (Fang et al., 2019; Hussein et al., 2017; Zheng et al., 2021).

2.2 Multi-intention IRL

In the multi-intention IRL approach, which is the focus of this work, it is assumed that the expert has multiple intentions, i.e., $|\mathcal{I}| > 1$ and the expert's demonstration set $\mathcal{D}$ typically doesn't include the intention labels, i.e., Inverse Reinforcement Learning with Multi-intention and Unlabeled Demonstrations (IRL-MUD). As in the case of single-intention IRL, the objective of multi-intention IRL methods is the perfect imitation of the expert's

demonstration set $\mathcal{D}$ (see Fig. 2). As the demonstrations are unlabeled, the multi-intention IRL methods try to infer the intention labels and the corresponding rewards functions in $\mathcal{D}$ in order to minimize the distance function $d(\mathcal{D}, \tilde{\mathcal{D}})$ where $\tilde{\mathcal{D}}$ is the imitated behaviors.

Although it is natural in this research direction to assume that the demonstrations are unlabeled, several studies have simplified the problem by considering the intention label as an available piece of information (Chen et al., 2020; Likmeta et al., 2021; Lin & Zhang, 2018; Nikolaidis et al., 2015; Ramponi et al., 2020). However, providing the intention labels can be too expensive or practically impossible in many real-world applications, and researchers try to address the problem of unlabeled demonstrations, i.e., IRL-MUD. In an early work, Babes et al. (2011) turned the problem into a parametric density estimation of the experts' demonstrations. Using EM, they tried to iteratively cluster the observed demonstrations and estimate the parameters of the rewards functions per cluster. The idea of density estimation has been applied by several authors to the problem of IRL-MUD (Almingol et al., 2013; Bighashdel et al., 2021; Choi & Kim, 2012; Michini et al., 2015; Michini & How, 2012; Rajasekaran et al., 2017; Ranchod et al., 2015). Despite the promising results, these methods are developed to handle low-dimensional problems with the knowledge of system dynamics, i.e., model-based setting.

There are a number of works that address the problem of model-free IRL-MUD (Hsiao et al., 2019; Hausman et al., 2017; Lin & Zhang, 2018; Li et al., 2017; Morton & Kochenderfer, 2017; Wang et al., 2017). Except ACGAIL (Lin & Zhang, 2018) and Goal-GAIL (Ding et al., 2019), where the demonstrations are labeled, all of the proposed methods have focused on a direct inference of the latent intentions in an unsupervised manner by employing generative models, namely variational auto-encoders (Hsiao et al., 2019; Morton & Kochenderfer, 2017; Wang et al., 2017) and Generative Adversarial Networks (GAN) (Hausman et al., 2017; Li et al., 2017). Hausman et al. (2017) and Li et al. (2017) in parallel studies, proposed two equivalent models, referred to as IntentionGAIL and Info-GAIL, respectively, as the extensions of Generative Adversarial Imitation Learning (GAIL) (Ho & Ermon, 2016) for the problem of model-free multi-intention IRL. Inspired by Chen et al. (2016), they tried to infer the latent structures by optimizing the mutual information between the intentions and the demonstrations in the GAN training. In comparison, we propose a clustering-based approach and demonstrate the benefits of our method in imitating the experts' behaviors when the multi-intention demonstrations are overlapping (see Sect. 5).

2.3 Meta-IRL

Another line of research in IRL is associated with the problem of generalization in the LfD task. When applied to unseen environments, satisfactory performance of the above-discussed IRL methods requires training from scratch. To avoid this, researchers have developed IRL methods in the framework of meta-learning, where the main goal is a cross-domain generalization. The meta-IRL methods seek to exploit the structural similarity among a distribution of environments and optimize for rapid adaptation to unknown environments with a limited amount of data. Given the expert's demonstration set $\mathcal{D}$ and intention set $\mathcal{I}$, the meta-IRL methods try to perfectly imitate the so far unseen target demonstration set $\mathcal{D}_t$ with the target intention set $\mathcal{I}_t$, where $|\mathcal{D}_t| \ll |\mathcal{D}|$ and $i \neq i_t$ for all $i \in \mathcal{I}$ and $i_t \in \mathcal{I}_t$ (see Fig. 2). Therefore, the objective of the meta-IRL methods is to minimize the distance function $d(\mathcal{D}_t, \tilde{\mathcal{D}}_t)$, where $\tilde{\mathcal{D}}_t$ is the imitated target behaviors. A meta-IRL method considers the available demonstrations $\mathcal{D}$ as a prior for $\mathcal{D}_t$ that can then be used

to efficiently learn the structure of new demonstrations with new intentions. As the reward function is able to succinctly capture the structure of demonstrations, Meta-IRL methods aim at fast inferring the new reward functions governing the new demonstrations $\mathcal{D}_t$. The goal, therefore, is not to acquire good reward functions that explain the expert's demonstration set $\mathcal{D}$, but rather to learn reward functions that can be quickly and efficiently adapted to the new target demonstration set $\mathcal{D}_t$.

In the LfD literature, the Meta-learning methods are commonly developed by extending the single-intention and multi-intention approaches into the meta-learning frameworks. Table 1 shows an overview of various Meta-learning methods and demonstrates their relations with their single-intention and multi-intention counterparts. In contrast to the multi-intention IRL approaches, the meta-IRL methods normally assume that the demonstration sets are labeled (Gleave & Habryka, 2018; Seyed Ghasemipour et al., 2019; Xu et al., 2019; Wang et al., 2021). Nevertheless, in the case of unlabeled demonstrations, the ideas from multi-intention IRL are employed as done by Yu et al. (2019). The authors have proposed a meta-IRL method by extending the Adversarial IRL (Fu et al., 2018) model to a meta-learning framework with unlabeled demonstrations. Similar to InfoGAIL (Li et al., 2017; Yu et al., 2019) infer the intentions by optimizing the mutual information between the latent intention variable and the induced demonstrations. They further employ intention-specific regularization terms in the reward functions for more efficient reward adaptation in the case of new intentions. The authors compare their proposed meta-IRL method with a meta-learning variant of InfoGAIL (Li et al., 2017), called Meta-InfoGAIL, and demonstrate superior performance in various meta-learning tasks (Yu et al., 2019).

As one can see, this line of research is parallel to the research direction of multi-intention IRL, and any direct comparison between the methods of these two research directions requires modification on the problem formulation level. In this paper, we consider the setting of model-free IRL with multi-intention and unlabeled demonstrations, model-free IRL-MUD, and consequently, we only compare our method with the methods of this line of research. In the following sections, we define our problems in more detail and propose our novel approach.

3 Problem definition

In this section, we provide the exact definitions of the IRL problems discussed in Sect. 1, namely (1) model-free IRL-MUD and (2) model-free IRL-MUD-O. To accomplish this, we first define the task of RL with Multiple Reward Functions (RL-MRF), i.e., multiple intentions where each intention corresponds to one reward function.

A Markov Decision Process (MDP) in RL-MRF is a tuple $(S, A, T, \kappa, d_0, \mathcal{I}, R)$ where S is the state space, A is the action space, $T : S \times A \times S \to [0, 1]$ is the dynamics model, κ is the discount factor, $d_0 : S \to [0, 1]$ is the initial state distribution, $\mathcal{I}$ is the set of intentions with prior probability $p(i)$, and R is the set of intention specific reward functions $R = \{r_i | \forall i \in \mathcal{I}\}$ where $r_i : S \times A \to \mathbb{R}$. We further define a policy as $\pi : S \times A \to [0, 1]$. In model-free RL-MRF, the dynamics model and the reward functions are not available in explicit forms. Therefore, model-free RL approaches are employed to approximate the optimal policy:

Definition 1 The task of model-free RL-MRF is defined as approximating the optimal policy $\pi^*(a|s) = \mathbb{E}_{i \sim p(i)} \pi^*(a|s, i)$ that maximizes the expected discounted reward over the

Table 1 The connection between single-intention LfD methods with their multi-intention and meta-learning variants

Single-intention	Multi-intention		Meta-learning
		Unlabeled Demonstrations	
Behavior Cloning Michie et al. (1990)	VAE-BC Hsiao et al. (2019)	✓	DAML Yu et al. (2018)
	L-MLP Morton and Kochenderfer (2017)	✓	
	ILPO Edwards et al. (2019)	✓	
MaxLikelihood IRL Babes et al. (2011)	EM-MLIRL Babes et al. (2011)	✓	-
MaxEnt IRL Ziebart et al. (2008)	EM-MaxEnt Babes et al. (2011)	✓	MCEIRL Gleave and Habryka (2018)
	SEM-MIIRL Bighashdel et al. (2021)	✓	MandRIL Xu et al. (2019)
	NP-BCIRL Rajasekaran et al. (2017)	✓	
Bayesian IRL Ramachandran and Amir (2007)	DPM-BIRL Choi and Kim (2012)	✓	MURAL Li et al. (2021)
	BNIRL Michini et al. (2015)	✓	
	NPBRS Ranchod et al. (2015)	✓	
GAIL Ho and Ermon (2016)	InfoGAIL Li et al. (2017)	✓	Meta-InfoGAIL Yu et al. (2019)
	IntentionGAIL Hausman et al. (2017)	✓	
	ACGAIL Lin and Zhang (2018)	-	
	GoalGAIL Ding et al. (2019)	-	
	EM-GAIL (ours)	✓	
Adversarial IRL	-		Meta-AIRL Wang et al. (2021)
			PEMIRL Yu et al. (2019)
			SMILe Seyed Ghasemipour et al. (2019)

induced state-action pairs: $\mathbb{E}_{i\sim p(i)}\mathbb{E}_{d_0\pi^i T}[\sum_t \kappa^t r_i(s_t, a_t)]$, where $d_0\pi^i T$ stands for ($s_0 \sim d_0$, $a_t \sim \pi(.|s_t, i)$, $s_{t+1} \sim T(.|s_t, a_t)$).

For the optimal policy $\pi^*(a|s)$, its occupancy measure can be defined as Ho and Ermon (2016):

$$p(s,a) \propto \mathbb{E}_{i\sim p(i)}\left[\pi^*(a|s,i)\sum_t \kappa^t T(s|s_t,a_t)\right]. \tag{1}$$

The occupancy measure is interpreted as the distribution of state-action pairs that an agent with optimal policy (referred to as an expert) encounters in navigating the environment. Therefore, we are allowed to write (Ho & Ermon, 2016):

$$\mathbb{E}_{i\sim p(i)}\mathbb{E}_{d_0\pi^{i^*}T}\left[\sum_t \kappa^t\, r_i(s_t,a_t)\right] = \mathbb{E}_{i\sim p(i),s,a\sim p(s,a|i)}\left[r_i(s,a)\right]. \tag{2}$$

In IRL-MUD, the reward functions $R = \{r_i|\forall i \in \mathcal{I}\}$ are unknown, and instead a set of experts' unlabeled demonstrations $\mathcal{D}$, i.e., without the intention label, are given. An expert's demonstration is defined as a sequence of state-action pairs, $(s_0, a_0), (s_1, a_1), ...$, induced by an optimal policy which is assumed to be the solution of RL-MRF (Definition 1). Given the definition of the occupancy measure, the experts' demonstrations can be perceived as a set of state-action pairs $\mathcal{D} = \{s, a \sim p(s, a)\}$ where $p(s, a)$ corresponds to the optimal policy $\pi^*(a|s)$ through Eq. (1).

Problem 1 The problem of *model-free IRL-MUD* is defined as finding the pseudo-reward function $\mathcal{R}(s,a) = \mathbb{E}_{i\sim p(i)}\left[\mathcal{R}_i(s,a;\theta)\right]$ from a set of experts' unlabeled demonstrations $\mathcal{D}$ such that the background distribution $q(s,a) = \mathbb{E}_{i\sim p(i)}q(s,a|i)$, induced by the approximated optimal policy under the the pseudo-reward function $\pi^*_{\mathcal{R}}(a|s) = \mathbb{E}_{i\sim p(i)}\pi^*_{\mathcal{R}}(a|s,i)$, matches $p(s, a)$.

Similar to Eq. (1), the background distribution relates to the policy $\pi^*_{\mathcal{R}}(a|s)$ as:

$$q(s,a) \propto \mathbb{E}_{i\sim p(i)}\left[\pi^*_{\mathcal{R}}(a|s,i)\sum_t \kappa^t T(s|s_t,a_t)\right]. \tag{3}$$

As discussed in 2.2, a number of studies have proposed solutions for the problem of model-free IRL-MUD via non-clustering approaches. These studies lack any evidence to support the validity of their proposed solutions in the environments where the demonstrations overlap. The overlapping demonstrations can be characterized by the existence of shared pairs:

Definition 2 Given a set of demonstrated state-action pairs $\mathcal{D} = \{\mathcal{D}_i|\forall i \in \mathcal{I}\}$ where $\mathcal{D}_i = \{s, a \sim p(s, a|i)\}$, an imitated state-action pair $\hat{s}, \hat{a} \sim q(\hat{s}, \hat{a}|k)$ where $k \in \hat{\mathcal{I}} \subseteq \mathcal{I}$, and $|\hat{\mathcal{I}}| > 1$, is called a *shared pair* when:

$$\forall i \in \hat{\mathcal{I}},\quad \exists s, a \in \mathcal{D}_i \mid \quad s, a = \hat{s}, \hat{a}. \tag{4}$$

In other words, an imitated state-action pair is a shared pair when it is shared between more than one intention. Using the definition above, we can define the problem that is the focus of this work:

Problem 2 The problem of *model-free IRL-MUD-O* is defined as a model-free IRL-MUD problem with the presence of shared pairs.

To solve this problem of model-free IRL-MUD-O, we propose and research a clustering-based approach that is detailed in the next section.

4 Approach

In this section, we first introduce our clustering-based solution for Problem 1: model-free IRL-MUD, see Sect. 4.1. Then in Sect. 4.2, we verify the validity of our solution for Problem 2: model-free IRL-MUD with overlapping demonstrations. This done by providing the theoretical analysis as shortly shown in Theorem 2 and Corollary 2.

4.1 Model-free IRL-MUD

To provide a clustering-based solution for the problem of model-free IRL-MUD, we first use parametric estimation of the occupancy measure $p(s, a)$. Then, we indicate the intractability of the solution by addressing the aforementioned challenges (see Sect. 1) and propose a new tractable approach using a mixture of logistic regressions.

Due to the multi-intention nature of the demonstrations, we model the occupancy measure $p(s, a)$ as a mixture of Boltzmann distributions, $p(s, a;\psi)$, parameterized by ψ, where the energy is given by the parameterized, intention-specific reward functions $r_i(s, a;\psi)$:

$$p(s, a;\psi) = \mathbb{E}_{i\sim p(i)}\left[\frac{\exp(r_i(s, a;\psi))}{Z_i(\psi)}\right], \tag{5}$$

where $Z_i(\psi)$ are the intention-specific partition functions.

Definition 3 The Parametric Density Estimation (PDE) approach for the problem of model-free IRL-MUD is defined as minimizing the following loss function:

$$\mathcal{L}_{PDE}(\psi) = -\mathbb{E}_{s,a\sim p(s,a)}[\log \mathbb{E}_{i\sim p(i)}\left[\frac{\exp(r_i(s, a;\psi))}{Z_i(\psi)}\right]. \tag{6}$$

The partition functions can not be obtained analytically for high dimensional domains without knowing the system dynamics, i.e., model-free IRL-MUD. Therefore, a sampling-based approach is employed where the partition functions are estimated from a background distribution $q(s, a) = \mathbb{E}_{i\sim p(i)}[q(s, a|i)]$. The background distribution is adaptively refined using the current reward functions in a policy optimization procedure. However, especially in the early stages, where the reward estimates experience high errors, the background distribution may fail to generate high-reward state-action pairs, which can lead the non-convergent behavior. As proposed by Finn et al. (2016a, b this problem is addressed by mixing the background distribution with an estimated distribution of the occupancy measure that has naturally high rewards:

Definition 4 The mixed sampling estimation of the partition function is defined as

$$Z_i(\psi) = \mathbb{E}_{s,a\sim\rho(s,a|i)}\left[\frac{\exp(r_i(s,a;\psi))}{\rho(s,a|i)}\right], \tag{7}$$

where $\rho(s,a|i) = \frac{1}{2}\tilde{p}(s,a|i) + \frac{1}{2}q(s,a|i)$ is the mixed sampler, and $\tilde{p}(s,a|i)$ is an estimate of the occupancy measure $p(s,a|i)$.

The mixed sampling estimation can help us to reach a solution via the EM algorithm:

Proposition 1 *Given the mixed sampling estimation, the PDE solution for the problem of model-free IRL-MUD constitutes the following iterative steps*:

- E-step:

$$\gamma(i|s,a) = \frac{p(s,a|i;\psi)p(i)}{p(s,a;\psi)}. \tag{8}$$

- M-step:

$$\begin{aligned}\partial_\psi \mathcal{L}_{PDE}(\psi) = &- \mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}[\partial_\psi r_i(s,a;\psi)] \\ &+ \mathbb{E}_{s,a\sim\rho(s,a),i\sim\rho(i|s,a)}\left[\frac{\frac{\exp(r_i(s,a;\psi))}{Z_i(\psi)}\partial_\psi r_i(s,a;\psi)}{\rho(s,a|i)}\right].\end{aligned} \tag{9}$$

Proof See Appendix A.1. □

The PDE solution is merely conceptual, and in order to come to practical implementation, we first need to address the following two key challenges: (1) estimation of the partition function (Sect. 4.1.1), and (2) estimation of the posterior distribution (Sect. 4.1.2).

4.1.1 Estimation of the partition function

The mixed sampling estimation for approximating the partition function requires an estimation of the true distribution of the experts' state-action pairs. Due to the multi-modality of the experts' multi-intention behavior, this estimation can be quite challenging. Inspired by Gutmann and Hyvärinen (2010), we take a different approach and consider the intention-specific partition functions as a set of additional parameters of the model $\omega = \{\omega_i|\forall i \in \mathcal{I}\}$ to avoid the explicit estimation:

$$p(s,a;\theta) = \mathbb{E}_{i\sim p(i)}\left[\frac{\exp(r(s,a,i;\psi))}{\omega_i}\right], \tag{10}$$

where $\theta = \{\psi, \omega\}$.

By defining the partition functions as the model parameters, we can learn the partition functions rather than explicit estimation, e.g., via mixed sampling (Definition 4). However, this approach is no longer consistent with the loss function of the PED approach (Eq. 6), as the maximum likelihood can lead to arbitrarily large numbers by

making the partition parameters, ω_i, reach zero. To address this, we first define a substitute loss function for the problem of model-free IRL-MUD (Problem 1) using a Mixture of Logistic Regressions (MLR) (Definition 5) and obtain the solution via EM (Proposition 2). Then, we prove that the MLR approach results in the same solution as the PDE approach (Theorem 2) and, therefore, is a valid substitute for the PDE approach.

Definition 5 The Mixture of Logistic Regressions (MLR) approach for the problem of model-free IRL-MUD is defined as minimizing the following loss function:

$$\mathcal{L}_{MLR}(\theta) = -\mathbb{E}_{s,a\sim\rho(s,a)}\left[\mathbb{E}_{c\sim\tilde{D}(c|s,a)}\left[\log D(c|s,a;\theta)\right]\right], \tag{11}$$

where $c \in \{real, fake\}$ is the class label, $\tilde{D}(c|s,a)$ is the true class label distribution, and $D(c|s,a;\theta) = \mathbb{E}_{i\sim p(i)}[D(c|s,a,i;\theta)]$ is the parameterized class label distribution, with

$$D(c = real|s,a,i;\theta) = \frac{1}{1+\exp(-f(s,a,i;\theta))} \tag{12}$$

$$f(s,a,i;\theta) = r_i(s,a;\psi) - \log\omega_i - \log q(s,a|i). \tag{13}$$

The idea of the MLR approach is to estimate the parameters θ by learning to discriminate between the real state-action pairs, which are sampled from intention-specific occupancy measures $p(s,a|i;\theta)$, and the fakes ones, sampled from the intention-specific background distributions $q(s,a|i)$. Once again, we can obtain a solution by employing the EM algorithm:

Proposition 2 *The MLR solution for the problem of model-free IRL-MUD is the E-step, defined in Eq.* (8)*, and the following M-step*:

$$\begin{aligned}\partial_\theta\mathcal{L}_{MLR}(\theta) = &-\mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}\left[\partial_\theta\log D(s,a,i;\theta)\right]\\ &-\mathbb{E}_{s,a\sim q(s,a),i\sim q(i|s,a)}\left[\partial_\theta\log(1-D(s,a,i;\theta))\right],\end{aligned} \tag{14}$$

where

$$D(s,a,i;\theta) \equiv D(c = real|s,a,i;\theta). \tag{15}$$

Proof See Appendix A.2. □

To show the resemblances between the PDE (Proposition 1) and MLR solutions (Proposition 2) for the problem of model-free IRL-MUD (Problem 1), we first address the relationship between the partition functions and the optimal partition parameters $\omega_i^* = \text{argmax}_{\omega_i}\,\mathcal{L}_{MLR}(\theta)$:

Lemma 1 $\omega_i^* = \mathbb{E}_{s,a\sim\rho(s,a|i)}\left[\frac{\exp(r(s,a,i;\psi))}{\rho(s,a|i)}\right]$.

Proof See Appendix A.3. □

In other words, the parameter set w reaches the mixed sampling estimation of the partition functions. Now, we are ready to conclude that:

Theorem 1 *The PDE solution is equal to the MLR solution for the problem of model-free IRL-MUD, when* $\omega \to \omega^*$.

Proof See Appendix A.4. □

Applying Theorem 1, we can avoid the explicit, mixed sampling estimation of the partition functions by employing the MLR solution and assuming $\omega \approx \omega^*$ at each iteration.

4.1.2 Estimation of the posterior distribution

The goal of the E-step is to estimate the posterior probabilities of the *a priori* unknown intention labels, given the model parameters. This estimation is typically done via the Bayes' rule as shown in Eq. (8). However, according to the Bayes' rule, the intention-specific likelihood functions $p(s,a|i;\psi)$ are required, which further depend on the availability of the intention-specif partition functions $Z_i(\psi)$. Even if the partition functions can be estimated via the partition parameters in the MLR approach, the problem's dimensionality makes it impractical to compute the posterior probabilities for all demonstrations in all intentions. To address this, we propose a posterior network P, parameterized by ϕ, to obtain an estimate of the posterior probabilities. The posterior network can be trained in a supervised fashion by maximizing the likelihood of the state-action pairs sampled from the background distribution. Since intention-specific policies generate them, they have known intention labels.

Definition 6 The Posterior Probability Estimation (PPE) is defined as minimizing the following loss function:

$$\mathcal{L}_{PPE}(\phi) = -\mathbb{E}_{s,a\sim q(s,a),i\sim q(i|s,a)}\big[\log P(i|s,a;\phi)\big]. \tag{16}$$

In each E-step, we first update the parameters ϕ and then we set $\gamma(i|s,a) \approx P(i|s,a;\phi)$.

4.1.3 A practical clustering-based solution

Although the defined MLR approach for the problem of model-free IRL-MUD eliminates the need for the mixed sampling estimation of the partition functions, it still requires an specific parameterization of the nonlinear functions $f(s,a,i;\theta)$, as the inputs to the logistic functions $D(s,a,i;\theta)$ (Eqs. 11 and 12). In order to avoid this, the logistic functions $D(s,a,i;\theta)$ can be directly parameterized in a more general manner, e.g. a neural network with a logistic function as the final layer, to output the classification probabilities. Given the E-step, our proposed solution, referred to as EM-GAIL, can be seen as a GAN where in the M-step, the discriminator $D(s,a;\theta) = \mathbb{E}_{i\sim p(i)}[D(s,a,i;\theta)]$ is trained to minimize Eq. (11), and in the subsequent B-step (Background-step), the background distribution $q(s,a) = \mathbb{E}_{i\sim p(i)}\big[q(s,a|i)\big]$ is trained to maximize Eq. (11). Given the definition of the background distribution in Eq. (3), maximizing Eq. (11) is equal to training a policy under the pseudo-reward function $\mathcal{R}(s,a)$:

$$\begin{aligned}\mathcal{R}(s,a) &= -\mathbb{E}_{i\sim p(i)}\big[\log(1 - D(s,a,i;\theta))\big] \\ &= \mathbb{E}_{i\sim p(i)}\big[\mathcal{R}_i(s,a;\theta)\big],\end{aligned} \tag{17}$$

where $\mathcal{R}_i(s,a;\theta)$ are the intention-specific pseudo-reward functions.

By direct parameterization of the logistic functions $D(s,a,i;\theta)$ in Eq. (11), a practical solution to the problem of model-free IRL-MUD is obtained:

Solution 1 The EM-GAIL solution for problem of model-free IRL-MUD (Problem 1) is the following iterative steps:

- E-step:

$$\begin{aligned} &\partial_\phi \mathcal{L}_{PPE}(\phi) = -\mathbb{E}_{s,a\sim q(s,a),i\sim q(i|s,a)}\big[\partial_\phi \log P(i|s,a;\phi)\big], \\ &\gamma(i|s,a) = P(i|s,a;\phi). \end{aligned} \tag{18}$$

- M-step:

$$\begin{aligned} \partial_\theta \mathcal{L}_{MLR}(\theta) = &- \mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}\big[\partial_\theta \log D(s,a,i;\theta)\big] \\ &- \mathbb{E}_{s,a\sim q(s,a),i\sim q(i|s,a)}\big[\partial_\theta \log(1-D(s,a,i;\theta))\big]. \end{aligned} \tag{19}$$

- *B-step with the pseudo-reward function*:

$$\mathcal{R}(s,a) = -\mathbb{E}_{i\sim p(i)}\big[\log(1-D(s,a,i;\theta))\big]. \tag{20}$$

The following corollary shows the resemblance of our clustering-based solution to the GAIL solution (Ho & Ermon, 2016), which is proposed for model-free IRL with single-intention demonstrations:

Corollary 1 *The EM-GAIL solution reduces to GAIL solution when* $|\mathcal{I}| = 1$, *i.e., the number of intentions is one*.

In the following, first, we define the IntentionGAIL and InfoGAIL solutions as the non-clustering alternative for the problem of IRL-MUD in Sect. 4.1.4. Then, we theoretically analyze our clustering-based solution, along with the non-clustering alternatives, in the presence of overlapping demonstrations (see Sect. 4.2).

4.1.4 Non-clustering alternatives

InfoGAIL (Li et al., 2017) and IntentionGAIL (Hausman et al., 2017) are non-clustering approaches for the problem of model-free IRL-MUD, proposed as the extensions of GAIL Ho and Ermon (2016). Inspired by Chen et al. (2016), both methods infer the latent intentions by optimizing the mutual information between the intentions and the demonstrations in the GAN training. Despite the alternative derivations, the objective functions are identical (see Section A.5).

Solution 2 The Info/IntentionGAIL solution for the problem of model-free IRL-MUD (Problem 1) is the following iterative steps (Li et al., 2017; Hausman et al., 2017):

- Posterior-step:

$$\partial_{\phi}\mathcal{L}_{PDE}(\phi) = -\mathbb{E}_{s,a\sim q(s,a),i\sim q(i|s,a)}\big[\partial_{\phi}\log P(i|s,a;\phi)\big]. \tag{21}$$

- Discriminator-step:

$$\begin{aligned}\partial_{\theta}\mathcal{L}_{D}(\theta) = &-\mathbb{E}_{s,a\sim p(s,a)}\big[\partial_{\phi}\log D(s,a;\theta)\big]\\ &-\mathbb{E}_{s,a\sim q(s,a)}\big[\partial_{\theta}\log(1-D(s,a;\theta))\big].\end{aligned} \tag{22}$$

- Generator-step with the pseudo-reward function:

$$\mathcal{R}(s,a) = -\log(1-D(s,a;\theta)) + \lambda\mathbb{E}_{i\sim q(i|s,a)}\big[\log P(i|s,a;\phi)\big]. \tag{23}$$

4.2 Model-free IRL-MUD-O

In this section, we apply our clustering-based solution (Solution 1) to the problem of model-free IRL-MUD-O (Problem 2). Then, we provide theoretical comparisons with the Info/IntentionGAIL solution (Solution 2).

Given Definition 2, a shared pair is an imitated state-action pair that is shared between more than one intention. Since a shared pair mimics a demonstrated state-action pair, it should receive a high reward, regardless of the number of intentions in which it is shared. In other words, the reward of an imitated state-action pair should not be decreased by increasing the number of intentions in which the imitated state-action pair is shared. To research this behavior, we define *multi-intention reward error* to analyze the reward assignment sensitivity in various solutions for the problem of model-free IRL-MUD-O.

Definition 7 The multi-intention reward error for state-action pair s, a is defined as:

$$\mathcal{ER}(s,a) = \mathcal{R}_{|\mathcal{I}|=1}(s,a) - \mathcal{R}(s,a), \tag{24}$$

where $\mathcal{R}_{|\mathcal{I}|=1}$ is the pseudo-reward function when $|\mathcal{I}| = 1$.

The multi-intention reward error can be used as an indication of solution validity for the problem of model-free IRL-MUD-O. When the multi-intention reward error of a state-action pair is zero, it means that the reward assignment for the state-action pair is not sensitive to the number of intentions, whether the state-action pair in question is a shared pair or not. In other words, the reward assignment only depends on the imitation quality of the state-action pairs and not on the number of intentions in which they are shared. Therefore, a solution is considered to be valid for the problem of model-free IRL-MUD-O if the multi-intention reward errors of the state-action pairs are zero.

Theorem 2 *In global optimality condition, for each shared pair* $\hat{s},\hat{a}$ *we have* $\mathcal{ER}_{\text{EM-GAIL}}(\hat{s},\hat{a}) = 0$ *and* $\mathcal{ER}_{\text{Info/IntentionGAIL}}(\hat{s},\hat{a}) \propto log|\hat{\mathcal{I}}|$.

To prove Theorem 2, we first need to identify the condition for achieving a global optimality solution. The global optimality condition is defined when the discriminator cannot differentiate between a demonstrated and a generated state-action pair, and the intention predictions match the true intentions. Therefore:

Lemma 2 (Theorem 1 in Goodfellow et al. (2014)) *The global optimality of EM-GAIL is achieved if and only if the $D(s,a,i;\theta)$ and $P(i|s,a;\phi)$ are optimal and $p(s,a|i) = q(s,a|i)$, $\forall i \in \mathcal{I}$. The global optimality of Info/IntentionGAIL is achieved if and only if $D(s,a;\theta)$ and $P(i|s,a;\phi)$ are optimal and $p(s,a) = q(s,a)$.*

Proof See Appendix A.6. □

Now we are ready to give a direct proof of Theorem 2. A shared pair $\hat{s}, \hat{a}$ is similar to one state-action pair $\forall i \in \hat{\mathcal{I}}$. Therefore, the optimal posterior function outputs similar probabilities for all intentions in $\hat{\mathcal{I}}$ including intention k, i.e $P(k|\hat{s},\hat{a}) = 1/|\hat{\mathcal{I}}|$. In global optimality condition, we have $D(\hat{s},\hat{a}) = D(\hat{s},\hat{a},k) = 1/2$ (see Lemma 2). Therefore, the multi-intention reward error in EM-GAIL for the shared pair $\hat{s}, \hat{a}$ is:

$$\mathcal{ER}_{\text{EM-GAIL}}(\hat{s},\hat{a}) = -\log\frac{1}{2} + \log\frac{1}{2} = 0. \tag{25}$$

In other words, the reward assignment for a state-action pair in EM-GAIL is not sensitive to the number of intentions. As long as an intention-specific state-action pair is similar to a demonstrated one with the same intention, it receives a high reward in EM-GAIL. Therefore:

Corollary 2 *EM-GAIL is a solution for the problem of model-free IRL-MUD-O* (*Problem* 2).

In the case of Info/IntentionGAIL, the multi-intention reward error of the shared pair $\hat{s}, \hat{a}$ is:

$$\mathcal{ER}_{\text{Info/IntentionGAIL}}(\hat{s},\hat{a}) = -\log\frac{1}{2} + \log\frac{1}{2} - \lambda\log\frac{1}{|\hat{\mathcal{I}}|} = \lambda\log|\hat{\mathcal{I}}|. \tag{26}$$

Since $|\hat{\mathcal{I}}| \geq 1$, the error is always greater than zero, with equality when $|\hat{\mathcal{I}}| = 1$. This means that the pseudo-reward of a perfectly imitated state-action pair for a specific intention is reduced by increasing the number of intentions where the pair is shared.

The intuition behind these behaviors can be seen in the role of the discriminators. The goal of the policy optimizer in Info/IntentionGAIL is to fool the discriminator by making $q(s,a) \to p(s,a)$, and consequently, $\pi^*_{\mathcal{R}}(a|s) \to \pi^*(a|s)$, which is the direct result of the policy and occupancy measure uniqueness property (Syed et al., 2008). This is while no constraints are put upon the intention-specific background distributions $q(s,a|i)$ and the corresponding intention-specific policies $\pi^*_{\mathcal{R}}(a|s,i)$. The lack of sufficient constraints on the policy may lead to possible errors in the policy optimization procedure in multi-intention tasks. In EM-GAIL, on the other hand, the more stricter goal is to trick the intention-specific discriminators by making $q(s,a|i) \to p(s,a|i)$, i.e., $\pi^*_{\mathcal{R}}(a|s,i) \to \pi^*(a|s,i)$. In other words, the discriminator is viewing each individual intention-specific policy instead of the combined policy as in Info/IntentionGAIL. This way, not only the Info/IntentionGAIL goal is satisfied, but also more constraints are set for the policy, leading to more stable behaviors in multi-intention environments.

5 Experimental analyses

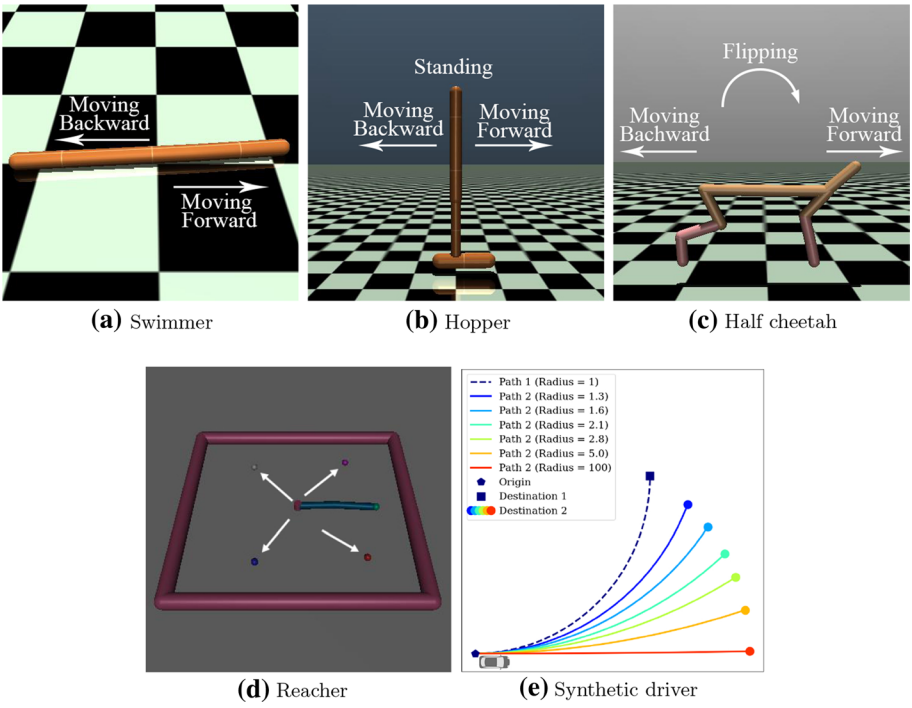
In this section, we compare the experimental results of our proposed EM-GAIL with two well-known baselines: (1) GAIL, proposed by Ho and Ermon (2016), which was originally developed for single-intention imitation learning, and (2) Info/IntentionGAIL, proposed by

Li et al. (2017) and Hausman et al. (2017) for multi-intention imitation learning. The main goal of this section is twofold: (1) to emphasize the limitation of the algorithms with the single-intention assumption, and (2) to experimentally validate our theoretical outcomes. The former is accomplished by evaluating the GAIL algorithm in various environments with a different number of intentions. For the second part, we introduce the metric of separability, which indicates the level of overlap in the demonstrations (the lower the separability, the higher the number of shared pairs), and we compare the algorithms in several environments with various levels of separability.

5.1 Environments

The experiments are conducted in four robotics environments, implemented in OpenAI Gym (Brockman et al., 2016) with the MuJoCo (Todorov et al., 2012) physics engine. For the purpose of this work, they are extended to cover multiple intentions by defining additional movements (see Figs. 3a–d). Note that, as in their standard single-intention versions, the horizontal location of the robot is excluded from the state space in all multi-intention environments. As a result, it is not possible for the algorithms to separate the demonstrations by simply observing the moving direction of the robot.

Swimmer is a 2-intention environment with an 8D state space and 2D action space. We have defined the intentions as moving forward and backward. For each intention, a



(a) Swimmer **(b)** Hopper **(c)** Half cheetah

(d) Reacher **(e)** Synthetic driver

Fig. 3 Environments. **a** Swimmer, which is a 2-intention environment. **b** Hopper, which is a 3-intention environment. **c** Half Cheetah, which is a 3-intention environment. **d** Reacher, which is a 4-intention environment. **e** Synthetic driver, which is a 2-intention environment (Color figure online).

linear reward is assigned for moving progress. We set the maximum time-step of each episode to 500.

Hopper is a 3-intention environment with an 11D state space and 3D action space. We have defined the intentions as moving forward, backward, and standing. A positive linear reward is assigned for moving progress for intended moving, and a negative moving reward for intended standing. The maximum time-step of each episode is set to 500.

Half cheetah is a 3-intention environment, with a 17D state space, 6D action space. We have defined the intentions as moving forward, backward, and flipping. A positive linear reward is assigned for moving progress for intended moving, and a negative moving reward for intended flipping, along with a positive linear reward for increasing the body angle. The maximum time-step is 500.

Reacher is a 4-intention environment, with a 26D state space and 2D action space. We have defined the intentions as reaching to separate colored balls. The balls are randomly located at separate quarters and the reward is defined as the distance to each intended ball. The maximum time-steps is 50.

We additionally introduce a synthetic driver task (see Fig. 3e) in which the separability between multiple intentions can be controlled directly.

Synthetic driver is a 2-intention environment in which the agent can move freely from the origin at a constant speed in a 2D environment by controlling the steering angle α_t at discrete time t. For the agent, the state at time t is the 2D positions from $t-4$ to t. As shown in Fig. 3e, the intentions are defined as two destinations where each destination has a curvature path with a specific radius. While the first destination is always fixed, the second destination and its corresponding path radius vary in different experiments. As will be shown later (Fig. 4b), each radius of the second destination results in an specific level of overlap in the demonstrations.

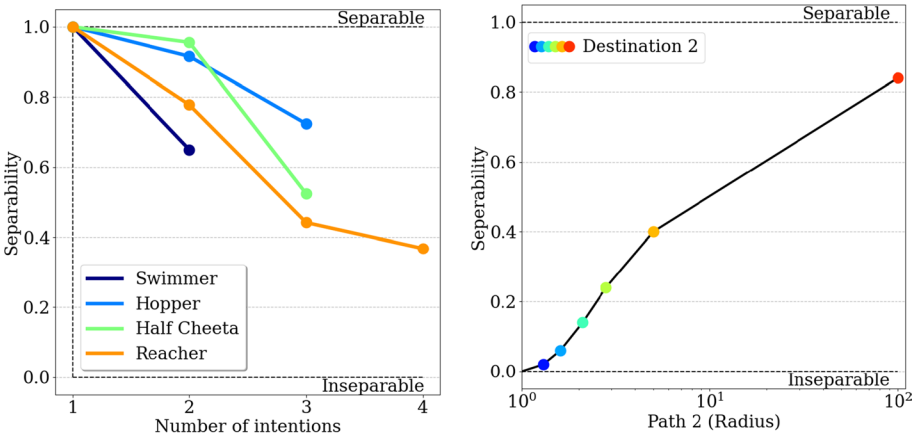


Fig. 4 Separability in the expert's state-action space for (a) robotics environments with respect to the number of intentions, and (b) Synthetic driver with respect to the path radius of the second destination (Color figure online)

5.2 Metrics

The performances of IRL algorithms are evaluated by the following metrics:

Expected Reward Difference (ERD), which is inspired by Choi and Kim (2012), Bighashdel et al. (2021), is a measure of how accurate the optimal policy, trained on the learned pseudo-reward function, performs under the ground truth reward function. We train an agent on the learned pseudo-reward function and evaluate its behavior on the ground truth reward function by computing the cumulative ground truth reward. The expected difference between the agent's cumulative ground truth reward and the expert's cumulative ground truth reward is referred to as ERD. The ERD avoids having to define task-specific metrics for each unique robotics environment (e.g., a swimming metric for the Swimmer environment or a hopping metric for the Hopper environment), thereby allowing a uniform approach to compare methods concerning different robotics tasks. The ERD is defined as:

$$\text{ERD} = |\mathbb{E}_{i\sim p(i),s,a\sim p(s,a|i)}\left[r_i(s,a)\right] - \mathbb{E}_{i\sim p(i),s,a\sim q(s,a|i)}\left[r_i(s,a)\right]|, \tag{27}$$

where r_i is the ground truth intention-specific reward function. Given Eq. (2), ERD can be further written as:

$$\text{ERD} = \mathbb{E}_{i\sim p(i)}\left|\mathbb{E}_{d_0\pi^{i^*}T}\left[\sum_t \kappa^t r_i(s_t,a_t)\right] - \mathbb{E}_{d_0\pi^{i^*}_{\mathcal{R}}T}\left[\sum_t \kappa^t r_i(s_t,a_t)\right]\right|, \tag{28}$$

where $d_0\pi^{i^*}T$ stands for ($s_0 \sim d_0$, $a_t \sim \pi^*(.|s_t,i)$, $s_{t+1} \sim T(.|s_t,a_t)$), and $d_0\pi^{i^*}_{\mathcal{R}}T$ stands for ($s_0 \sim d_0$, $a_t \sim \pi^*_{\mathcal{R}}(.|s_t,i)$, $s_{t+1} \sim T(.|s_t,a_t)$). The only unknowns in Eq. (28) are the intention-specific policies $\pi^*_{\mathcal{R}}(a|s,i)$. When the IRL training is completed and the intention-specific, pseudo-reward functions $\mathcal{R}_i(s,a;\theta)$ are learned, the intention-specific policies are obtained by training an agent on the learned pseudo-reward functions using RL. We normalize the ERD values between 0 and 1, where 1 corresponds to random policy and 0 represents the experts' policy π^*.

Unsupervised clustering accuracy is evaluated in experts' state-action space and is defined as Yang et al. (2010):

$$\text{Accuracy} = \max_{g\in\mathcal{G}} \frac{\sum_{m=1}^{M} \mathbf{1}\{\bar{k}^m = g(k^m)\}}{M}, \tag{29}$$

where $\bar{k}^m$ and k^m are the ground truth and predicted intention label of the expert's state-action pair m, respectively, M is the total number of experts' state-action pairs, and $\mathcal{G}$ is the set of all possible one-to-one mappings between ground truth and predicted intention labels. Unsupervised clustering accuracy can evaluate the performance of the posterior networks.

We further introduce a metric to measure the level of overlap in the demonstrations of each environment:

Separability is obtained by first fitting a Gaussian Mixture Model (GMM) on the expert's state-action space and then predicting the unsupervised clustering accuracy. Then, the separability is computed by normalizing the accuracy between 1 and $1/|\mathcal{I}|$. The separability is a measure of how well the intention-specific state-action pairs are represented by separate distributions. Low separability indicates a high level of overlap in the demonstrations and consequently increases the probability of generating the shared pairs.

The separability of the experts' demonstrations in simulated robotics and the synthetic driver environments are depicted in Fig. 4a, b, respectively. As can be seen in Fig. 4a, the separability differs in various robotics environments with the same number of intentions, and in all cases, the separability is reduced by increasing the number of intentions. Figure 4b also shows that the separability in the synthetic driver environment gets lower by reducing the distance between the two destinations, which is the result of decreasing the radius of path 2.

5.3 Implementation details

Algorithm 1 EM-GAIL

Input: $\mathcal{D}, \pi_{\mathcal{R}_0}, \theta_0, \phi_0, p(i), d_0, T$
Output: $\mathcal{R}(s,a)$
for $i \leftarrow 1, n$ **do**
Sample a batch of expert's state-action pairs $\chi_n \sim \mathcal{D}$
Sample a batch of state-action-intention triples $\hat{\chi}_n \sim \{i \sim p(i), d_0 \pi^i_{\mathcal{R}_n} T\}$
E-step:
Update ϕ_n to ϕ_{n+1} with the gradient:
$$\mathbb{E}_{s,a,i\sim\chi_n}\left[\partial_{\phi_n} \log P(i|s,a;\ \phi_n)\right]$$
Set $\gamma(i|s,a) = P(i|s,a;\ \phi_n) \qquad \forall s,a \in \chi_n$
M-step:
Update θ_n to θ_{n+1} with the gradient:
$$\mathbb{E}_{s,a\sim\chi_n,i\sim\gamma(i|s,a)}\left[\partial_{\theta_n} \log D(s,a,i;\ \theta_n)\right] + \mathbb{E}_{s,a,i\sim\hat{\chi}_n}\left[\partial_{\theta_n} \log(1 - D(s,a,i;\ \theta_n))\right]$$
B-step:
Update $\pi_{\mathcal{R}_n}$ to $\pi_{\mathcal{R}_{n+1}}$ using TRPO rule with following objective:
$$\mathbb{E}_{s,a,i\sim\hat{\chi}_n}\left[\mathcal{R}_i(s,a;\ \theta)\right]$$
end for

We used one implementation for InfoGAIL (Li et al., 2017) and IntentionGAIL (Hausman et al., 2017), referred to as Info/IntentionGAIL, as both approaches follow exactly the same procedure (See Sect. 4.1.4 and A.5), although they were presented in parallel studies. We used policies and discriminators with the same neural network architecture in all IRL algorithms (GAIL by Ho and Ermon (2016), Info/IntentionGAIL by Li et al. (2017), Hausman et al. (2017) and our EM-GAIL) for all experiments. We employed two hidden layers of dimension 64 for policies, discriminators, and posterior networks, with Tanh nonlinear functions in between. The policies of both Info/IntentionGAIL and EM-GAIL, as well as the discriminator of EM-GAIL, accept an additional, one-hot, intention assignment vector.

The number of learnable parameters of GAIL, Info/IntentionGAIL, and EM-GAIL, as well as their averaged training time for one iteration, is depicted in Table 2 for the evaluated robotics environments. As can be seen, both Info/IntentionGAIL and EM-GAIL have approximately the same number of learnable parameters, leading to similar

Table 2 Number of Learnable Parameters (LP) and an average training times in various robotic environments

Env.	GAIL		Info/Intention-GAIL		EM-GAIL	
	#LP	Time (s)	#LP	Time (s)	#LP	Time (s)
Swimmer	14.6K	8.5	19.9K	10.2	20.0K	10.3
Hopper	15.3K	12.4	21.1K	15.3	21.2K	15.6
Half cheetah	16.9K	13.3	23.2K	16.1	23.4K	16.5
Reacher	18.1K	5.5	24.9K	8.4	25.1K	8.7

training times for both methods. Due to additional posterior networks, both Info/IntentionGAIL and EM-GAIL have higher numbers of learnable parameters and training times compared to GAIL.

For each experiment, we first obtain the expert's policy by running Trust Region Policy Optimization (TRPO) algorithm (Schulman et al., 2015) on the true reward functions to generate the expert's demonstrations, i.e., solving the task of model-free RL-MRF (Definition 1). The number of expert's state-action pairs per intention is fixed to 2500 for swimmer, hopper, and half cheetah environments, 1500 for reacher, and 500 for the synthetic driver. Then, the IRL algorithms are trained for 400 iterations on the expert's demonstrations by running Adam optimizer (Kingma & Ba, 2015) with a fixed learning rate of 0.001 (after testing for the range of 0.1 to 0.0001) for the discriminators and posterior networks, and TRPO with the reported hyperparameters in Schulman et al. (2015) for the policies. The optimization outline is depicted in Algorithm 1. All the experiments are run on a GPU-enabled computer (NVIDIA GeForce RTX 2080 Ti).

5.4 Results

Each experiment is repeated four times, and the results are shown in the form of means (lines) and standard deviations (shadings). The results are normalized between 0, corresponding to the expert's behavior, and 1, corresponding to the random behavior.

5.4.1 Synthetic driver task

Figure 5a shows the ERD for various levels of separability in the synthetic driver task. The results in this Figure show that the GAIL algorithm, which is developed with the single-intention assumption, is not able to imitate the expert's behaviors in any of the experiments. What can be further seen in Fig. 5a is the insensitivity of our purposed EM-GAIL to the separability of the demonstrations. The Info/IntentionGAIL shows insensitive behavior as long as the separability, i.e., the distance between the two destinations, is more than a certain limit (≈ 0.2). When the separability of the demonstrations gets lower than 0.2, which according to Fig. 4b corresponds to the second destination path radius of lower than ≈ 2, the number of shared pairs grows exponentially, leading to higher multi-intention reward errors in Info/IntentionGAIL. This will consequently result in a performance drop in Info/IntentionGAIL for the separability of lower than 0.2.

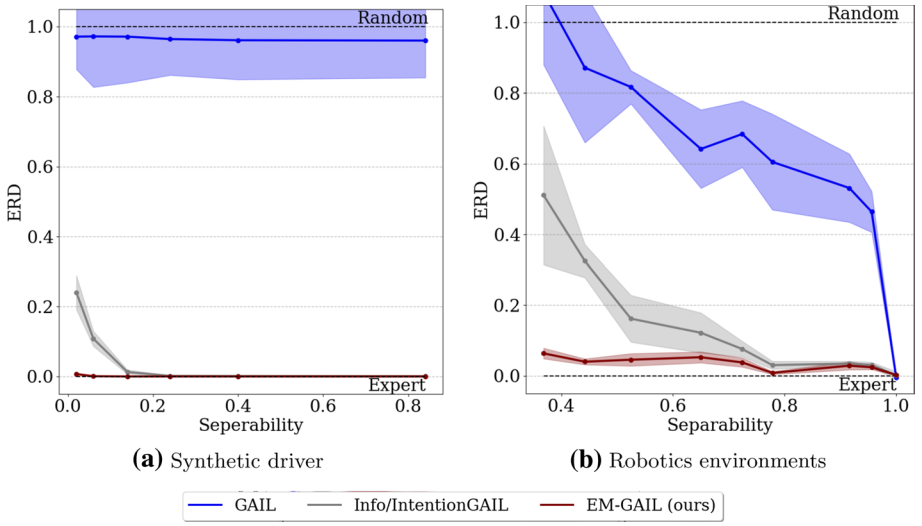


Fig. 5 ERD with respect to the separability of the demonstrations, where a lower ERD is better and an ERD of 1 is equal to random behavior. **a** Synthetic driver, where each separability point corresponds to a specific path 2 radii (see Fig. 4a). **b** Robotics environments, where each separability point corresponds to a specific robotic environment with a specific number of intentions (see Fig. 4a) (Color figure online)

5.4.2 Simulated robotics tasks

Figure 6 shows the sensitivity of ERD to the number of intentions in four robotics environments. The results once again show the incapability of GAIL in environments with more than one intention. What stands out in Fig. 6 is that when the number of intentions gets higher, which according to Fig. 4a leads to a lower separability and consequently a growth in the number of shared pairs, the performance of Info/IntentionGAIL drops significantly. This is while EM-GAIL has shown more stable behavior regardless of the number of intentions. To make the conclusion more clear, we further demonstrate the performances of the algorithms directly with respect to the separability in Fig. 5b. As shown, Info/IntentionGAIL shows insensitive behavior as long as the separability is more than a certain limit. These results are also consistent with the sub-optimal behavior of Info/IntentionGAIL in the synthetic driver task in Fig. 5a. This is while it is apparent from Fig. 5b that the performance of EM-GAIL is less sensitive to the separability of the demonstrations.

Further experiments are conducted to evaluate the unsupervised clustering performance of the multi-intention IRL solutions. Since GAIL is not built for multi-intention IRL, it has been excluded from these experiments. To evaluate the clustering performance of Info/IntentionGAIL, the intention labels of the expert's state-action pairs are obtained using the posterior network. Table 3 indicates the average unsupervised clustering accuracy of the methods for the varying number of intentions. What is interesting in Table 3 is that even though both Info/IntentionGAIL and EM-GAIL are trained with the same loss function (Eq. 16), EM-GAIL outperforms Info/IntentionGAIL in unsupervised clustering, especially in environments with a higher number of intentions. The reason lies in the fact that the prediction performance of the posterior network highly depends on the quality of the generated samples, which are considered the training dataset for the posterior function. According to Theorem 2, the higher number of intentions results in higher multi-intention

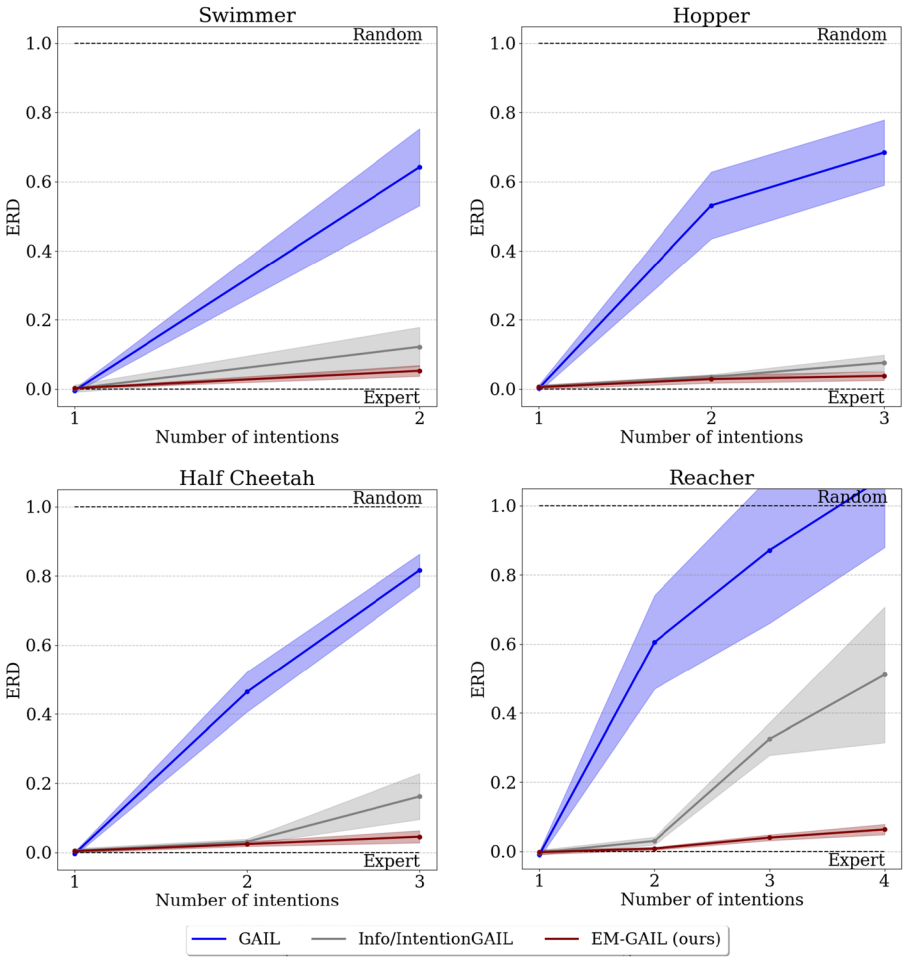


Fig. 6 ERD with respect to the number of intentions. A lower ERD is better and an ERD of 1 is equal to random behavior (Color figure online)

reward errors in Info/IntentionGAIL. As a direct consequence, the policy of Info/IntentionGAIL is less able to generate perfectly imitated state-action pairs. Once again, these results emphasize the strength of EM-GAIL in imitation from overlapping demonstrations.

6 Conclusions

We proposed a model-free IRL framework to imitate the multi-intention behaviors of the experts, from unlabeled (without intention label) and partially overlapping (shared between multiple intentions) demonstrations. This is realized by a novel clustering-based approach, EM-GAIL, using a mixture of logistic regressions and a posterior network. The mixture of logistic regression avoids the direct estimation of the partition function, and the posterior network approximates the posterior distributions that are impractical to be expressed and

Table 3 Accuracy in unsupervised clustering of the expert's state-action pairs

Models	Hopper		Half Cheetah	
	2 int. (%)	3 int. (%)	2 int. (%)	3 int. (%)
GMM	95.8	81.6	97.8	68.3
Info/IntentionGAIL	**99.3**	96.3	**99.9**	79.5
EM-GAIL (ours)	99.2	**98.1**	99.8	**96.5**
Models	Swimmer	Reacher		
	2 int. (%)	2 int. (%)	3 int. (%)	4 int. (%)
GMM	82.5	88.9	62.8	52.6
Info/IntentionGAIL	89.1	98.3	72.0	52.4
EM-GAIL (ours)	**97.3**	**99.2**	**97.1**	**96.8**

The highest accuracy in each environment is indicated in bold

estimated analytically. We addressed the problem of overlapping demonstrations by defining the concept of *shared pair* and analytically proved that, under the global optimality condition, EM-GAIL is a solution. We further compared EM-GAIL with well-known baselines on a set of simulated tasks. We conclude that first, the algorithms with single-intention assumption are incapable of imitating the experts' multi-intention demonstrations, and second, the performance of the non-clustering approaches, which focus on the direct inferring of the latent intention, significantly depends on the separability of the intentions. This is while our proposed clustering-based approach shows stable behavior, regardless of the separability and the number of intentions.

Having shown the benefits of our approach with textita priori known number of experts' intentions, we aim to extend the same approach to also be able to infer the intentions when the number of intentions is a priori not known.

Appendix A

This section provides the proofs for all the statements in the main paper.

A.1 Proof of Proposition 1

Minimizing the negative log likelihood leads to the following loss function:

$$\mathcal{L}_{PDE}(\psi) = -\mathbb{E}_{s,a\sim p(s,a)}[\log \mathbb{E}_{i\sim p(i)}[p(s,a|i;\psi)]. \tag{A1}$$

Taking the derivatives results in:

$$\begin{aligned} \partial_\psi \mathcal{L}_{PDE}(\psi) &= -\mathbb{E}_{s,a\sim p(s,a)}\left[\frac{\partial_\psi \mathbb{E}_{i\sim p(i)}[p(s,a|i;\psi)]}{\mathbb{E}_{i\sim p(i)}[p(s,a|i;\psi)]}\right] \\ &= -\mathbb{E}_{s,a\sim p(s,a)}\left[\mathbb{E}_{i\sim p(i)}\left[\frac{\partial_\psi p(s,a|i;\psi)}{p(s,a;\psi)}\right]\right]. \end{aligned} \tag{A2}$$

We can use the definition of the intention posterior probability which is computed in the E-step:

$$\gamma(i|s,a) = p(i|s,a;\psi) = \frac{p(s,a|i;\psi)p(i)}{p(s,a;\psi)}. \tag{A3}$$

Multiplying the nominator and the denominator in the derivatives by $p(s,a|i;\psi)p(i)$ yields:

$$\begin{aligned}\partial_\psi \mathcal{L}_{PDE}(\psi) &= -\mathbb{E}_{s,a\sim p(s,a)}\left[\mathbb{E}_{i\sim p(i)}\left[\frac{p(s,a|i;\psi)p(i)}{p(s,a;\psi)}\frac{\partial_\psi p(s,a|i;\psi)}{p(s,a|i;\psi)p(i)}\right]\right] \\ &= -\mathbb{E}_{s,a\sim p(s,a)}\mathbb{E}_{i\sim p(i)}\left[\frac{\gamma(i|s,a)}{p(i)}\partial_\psi \log p_\psi(s,a|i;\psi)\right].\end{aligned} \tag{A4}$$

Using the definition of Boltzmann distribution results in:

$$\begin{aligned}\partial_\psi \mathcal{L}_{PDE}(\psi) &= -\mathbb{E}_{i\sim p(i)}\mathbb{E}_{s,a\sim p(s,a)}\left[\frac{\gamma(i|s,a)}{p(i)}\partial_\psi \log \frac{1}{Z_i(\psi)}\exp(r_i(s,a;\psi))\right] \\ &= -\mathbb{E}_{i\sim p(i)}\mathbb{E}_{s,a\sim p(s,a)}\left[\frac{\gamma(i|s,a)}{p(i)}\partial_\psi r_i(s,a;\psi)\right. \\ &\quad \left. -\frac{\gamma(i|s,a)}{p(i)}\partial_\psi \log Z_i(\psi)\right] \\ &= -\mathbb{E}_{i\sim p(i)}\left[\mathbb{E}_{s,a\sim p(s,a)}\left[\frac{\gamma(i|s,a)}{p(i)}\partial_\psi r_i(s,a;\psi)\right]\right. \\ &\quad \left. -\mathbb{E}_{s,a\sim p(s,a)}\left[\frac{\gamma(i|s,a)}{p(i)}\partial_\psi \log Z_i(\psi)\right]\right] \\ &= -\mathbb{E}_{i\sim p(i)}\left[\mathbb{E}_{s,a\sim p(s,a)}\left[\frac{\gamma(i|s,a)}{p(i)}\partial_\psi r_i(s,a;\psi)\right]\right. \\ &\quad \left. -\partial_\psi \log Z_i(\psi)\mathbb{E}_{s,a\sim p(s,a)}\left[\frac{\gamma(i|s,a)}{p(i)}\right]\right].\end{aligned} \tag{A5}$$

In the E-step of the EM algorithm, we set $\gamma(i|s,a) \approx p(i|s,a)$ where ψ is the current set of parameters. Assuming a constant $p(i)$ and given that $p(s,a) = \frac{p(s,a|i)p(i)}{p(i|s,a)}$ and $\mathbb{E}_{s,a\sim p(s,a|i)}[1] = 1$, we have:

$$\partial_\psi \mathcal{L}_{PDE}(\psi) = -\mathbb{E}_{i\sim p(i)}\left[\mathbb{E}_{s,a\sim p(s,a)}\left[\frac{\gamma(i|s,a)}{p(i)}\partial_\psi r_i(s,a;\psi)\right] - \partial_\psi \log Z_i(\psi)\right]. \tag{A6}$$

Using the definition of mixed sampling estimation, $Z_i(\psi) = \mathbb{E}_{s,a\sim\rho(s,a|i)}[\frac{\exp(r_i(s,a;\psi))}{\rho(s,a|i)}]$, we have:

$$
\begin{aligned}
\partial_\psi \mathcal{L}_{PDE}(\psi) &= -\mathbb{E}_{i\sim p(i)}\left[\mathbb{E}_{s,a\sim p(s,a)}\left[\frac{\gamma(i|s,a)}{p(i)}\partial_\psi r_i(s,a;\psi)\right]\right. \\
&\quad \left. -\partial_\psi \log(\mathbb{E}_{s,a\sim\rho(s,a|i)}\left[\frac{\exp(r_i(s,a;\psi))}{\rho(s,a|i)}\right])\right] \\
&= -\mathbb{E}_{i\sim p(i)}[\mathbb{E}_{s,a\sim p(s,a)}[\frac{\gamma(i|s,a)}{p(i)}\partial_\psi r_i(s,a;\psi)] \\
&\quad -\frac{\mathbb{E}_{s,a\sim\rho(s,a|i)}[\frac{\exp(r_i(s,a;\psi))\partial_\psi r_i(s,a;\psi)}{\rho(s,a|i)}]}{\mathbb{E}_{s,a\sim\rho(s,a|i)}[\frac{\exp(r_i(s,a;\psi))}{\rho(s,a|i)}]}] \\
&= -\mathbb{E}_{i\sim p(i)}[\mathbb{E}_{s,a\sim p(s,a)}[\frac{\gamma(i|s,a)}{p(i)}\partial_\psi r_i(s,a;\psi)] \\
&\quad -\mathbb{E}_{s,a\sim\rho(s,a|i)}[\frac{\frac{1}{Z_i(\psi)}\exp(r_i(s,a;\psi))\partial_\psi r_i(s,a;\psi)}{\rho(s,a|i)}]].
\end{aligned}
\tag{A7}
$$

We can use the definition of the expectation as follows:

$$
\begin{aligned}
\mathbb{E}_{i\sim p(i)}[\mathbb{E}_{s,a\sim\rho(s,a|i)}[g(s,a,i)]] &= \int p(i)\int \rho(s,a|i)g(s,a,i)d_i d_{s,a} \\
&= \int\int p(i)\rho(s,a|i)g(s,a,i)d_i d_{s,a} \\
&= \int\int \rho(s,a)\rho(i|s,a)g(s,a,i)d_i d_{s,a} \\
&= \int \rho(s,a)\int \rho(i|s,a)g(s,a,i)d_i d_{s,a} \\
&= \mathbb{E}_{s,a\sim\rho(s,a)}[\mathbb{E}_{i\sim\rho(i|s,a)}[g(s,a,i)]].
\end{aligned}
\tag{A8}
$$

where g is a function. This property of the expectation has been frequently used throughout the paper. Now, we can reach the final equation for the M-step:

$$
\begin{aligned}
\partial_\psi \mathcal{L}_{PDE}(\psi) = &-\mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}[\partial_\psi r_i(s,a;\psi)] \\
&+\mathbb{E}_{s,a\sim\rho(s,a),i\sim\rho(i|s,a)}\left[\frac{\frac{1}{Z_i(\psi)}\exp(r_i(s,a;\psi))\partial_\psi r_i(s,a;\psi)}{\rho(s,a|i)}\right].
\end{aligned}
\tag{A9}
$$

A.2 Proof of Proposition 2

The mixture of logistic regressions is defined as:

$$
\mathcal{L}_{MLR}(\theta) = -\mathbb{E}_{s,a\sim\rho(s,a)}\left[\mathbb{E}_{c\sim\tilde{D}(c|s,a)}\left[\log D(c|s,a;\theta)\right]\right],
\tag{A10}
$$

where $\tilde{D}(c|s,a)$ is the true class label of the state-action pair s, a, and

$$
D(c|s,a;\theta) = \mathbb{E}_{i\sim D(i)}[D(c|s,a,i;\theta)],
\tag{A11}
$$

where we have defined a constant prior $D(i) = p(i)$. Taking the derivatives with respect to θ yields:

$$\begin{aligned}\partial_\theta \mathcal{L}_{MLR}(\theta) &= -\mathbb{E}_{s,a\sim\rho(s,a)}\left[\mathbb{E}_{c\sim\tilde{D}(c|s,a)}\left[\partial_\theta \log D(c|s,a;\theta)\right]\right] \\ &= -\mathbb{E}_{s,a\sim\rho(s,a)}\left[\mathbb{E}_{c\sim\tilde{D}(c|s,a)}\left[\partial_\theta \log \mathbb{E}_{i\sim D(i)}[D(c|s,a,i;\theta)]\right]\right] \\ &= -\mathbb{E}_{s,a\sim\rho(s,a)}\left[\mathbb{E}_{c\sim\tilde{D}(c|s,a)}\left[\frac{\partial_\theta \mathbb{E}_{i\sim D(i)}[D(c|s,a,i;\theta)]}{\mathbb{E}_{i\sim D(i)}[D(c|s,a,i;\theta)]}\right]\right] \\ &= -\mathbb{E}_{s,a\sim\rho(s,a)}\left[\mathbb{E}_{c\sim\tilde{D}(c|s,a)}\left[\mathbb{E}_{i\sim D(i)}\left[\frac{\partial_\theta D(c|s,a,i;\theta)}{D(c|s,a;\theta)}\right]\right]\right].\end{aligned} \tag{A12}$$

We can use the definition of the posterior probability which is computed in the E-step:

$$D(i|c,s,a;\theta) = \frac{D(c|s,a,i;\theta)D(i)}{D(c|s,a;\theta)}. \tag{A13}$$

Multiplying the nominator and the denominator in the derivatives by $D(c|s,a,i;\theta)D(i)$ yields:

$$\begin{aligned}\partial_\theta \mathcal{L}_{MLR}(\theta) &= -\mathbb{E}_{s,a\sim\rho(s,a)}\left[\mathbb{E}_{c\sim\tilde{D}(c|s,a)}\left[\mathbb{E}_{i\sim D(i)}\left[\frac{D(c|s,a,i;\theta)D(i)}{D(c|s,a;\theta)}\right.\right.\right. \\ &\quad \left.\left.\left.\times\frac{\partial_\theta D(c|s,a,i;\theta)}{D(c|s,a,i;\theta)D(i)}\right]\right]\right] \\ &= -\mathbb{E}_{s,a\sim\rho(s,a)}\left[\mathbb{E}_{c\sim\tilde{D}(c|s,a)}\left[\mathbb{E}_{i\sim D(i)}\left[\frac{D(i|c,s,a;\theta)}{D(i)}\right.\right.\right. \\ &\quad \left.\left.\left.\times\partial_\theta \log D(c|s,a,i;\theta)\right]\right]\right] \\ &= -\mathbb{E}_{s,a\sim\rho(s,a)}\left[\mathbb{E}_{c\sim\tilde{D}(c|s,a)}\left[\mathbb{E}_{i\sim D(i|c,s,a;\theta)}\left[\partial_\theta \log D(c|s,a,i;\theta)\right].\right]\right]\end{aligned} \tag{A14}$$

Separating the real state-action pairs from the fake ones yields:

$$\begin{aligned}\partial_\theta \mathcal{L}_{MLR}(\theta) = &-\mathbb{E}_{s,a\sim p(s,a)}\left[\mathbb{E}_{i\sim D(i|c=real,s,a;\theta)}\left[\partial_\theta \log D(c=real|s,a,i;\theta)\right]\right] \\ &-\mathbb{E}_{s,a\sim q(s,a)}\left[\mathbb{E}_{i\sim D(i|c=fake,s,a;\theta)}\left[\partial_\theta \log D(c=fake|s,a,i;\theta)\right]\right].\end{aligned} \tag{A15}$$

Given that:

$$D(i|c=real,s,a;\theta) = p(i|s,a;\theta) = \gamma(i|s,a) \tag{A16}$$

$$D(i|c=fake,s,a;\theta) = q(i|s,a), \tag{A17}$$

and setting:

$$D(c=real|i,s,a;\theta) \equiv D(s,a,i;\theta) \tag{A18}$$

$$D(c=fake|i,s,a;\theta) \equiv 1 - D(s,a,i;\theta), \tag{A19}$$

We reach the final equation for the M-step:

$$\begin{aligned}\partial_\theta \mathcal{L}_{MLR}(\theta) = &-\mathbb{E}_{s,a\sim p(s,a)}\left[\mathbb{E}_{i\sim\gamma(i|s,a)}\left[\partial_\theta \log D(s,a,i;\theta)\right]\right] \\ &-\mathbb{E}_{s,a\sim q(s,a)}\left[\mathbb{E}_{i\sim q(i|s,a)}\left[\partial_\theta \log(1 - D(s,a,i;\theta))\right]\right].\end{aligned} \tag{A20}$$

A.3 Proof of Lemma 1

The mixture of logistic regressions for the multi-intention IRL problem is defined as:

$$\begin{aligned}\mathcal{L}_{MLR}(\theta) = &- \mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}\big[\log D(s,a,i;\theta)\big] \\ &- \mathbb{E}_{s,a\sim q(s,a),i\sim q(i|s,a)}\big[\log(1-D(s,a,i;\theta))\big].\end{aligned} \tag{A21}$$

The nonlinear logistic function $D(s,a,i;\theta)$ is further defined as:

$$\begin{aligned}D(s,a,i;\theta) &= \frac{1}{1+\exp(-f(s,a,i;\theta))} \\ &= \frac{1}{1+\exp(-r_i(s,a;\psi)+\log\omega_i+\log q(s,a|i))} \\ &= \frac{1}{1+\frac{q(s,a|i)}{\frac{1}{\omega_i}\exp(r_i(s,a;\psi))}} \\ &= \frac{\frac{1}{\omega_i}\exp(r_i(s,a;\psi))}{\frac{1}{\omega_i}\exp(r_i(s,a;\psi))+q(s,a|i)}.\end{aligned} \tag{A22}$$

Given that $\rho(s,a|i)=\frac{1}{2}\tilde{p}(s,a|i)+\frac{1}{2}q(s,a|i)$ and by setting $p(s,a|i;\psi)$ as an estimate of $p(s,a|i)$ i.e. $\tilde{p}(s,a|i)=p(s,a|i;\psi)$, we have:

$$D(s,a,i;\theta) = \frac{\frac{1}{\omega_i}\exp(r_i(s,a;\psi))}{2\rho(s,a|i)}. \tag{A23}$$

Replacing Eq. (A23) in Eq. (A21) and separating the terms independent of ω_i yields:

$$\begin{aligned}\mathcal{L}_{MLR}(\theta) &= -\mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}\left[\log\frac{\frac{1}{\omega_i}\exp(r_i(s,a;\psi))}{2\rho(s,a|i)}\right] \\ &\quad - \mathbb{E}_{s,a\sim q(s,a),i\sim q(i|s,a)}\left[\log\frac{q(s,a|i)}{2\rho(s,a|i)}\right] \\ &= -\mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}\big[r_i(s,a;\psi)-\log\omega_i-\log 2\rho(s,a|i)\big] \\ &\quad - \mathbb{E}_{s,a\sim q(s,a),i\sim q(i|s,a)}\big[\log q(s,a|i)-\log 2\rho(s,a|i)\big] \\ &= \mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}[\log\omega_i]+2\mathbb{E}_{s,a\sim\rho(s,a),i\sim\rho(i|s,a)}[\log 2\rho(s,a|i)] \\ &\quad + g(\psi) \\ &= \log\omega_i + 2\mathbb{E}_{s,a\sim\rho(s,a|i)}[\log 2\rho(s,a|i)] + f(\psi,\omega_{-i}),\end{aligned} \tag{A24}$$

where ω_{-i} is the set ω excluding ω_i, and f and g are some functions. Taking the derivatives with respect to ω_i leads to:

$$\partial_{\omega_i}\mathcal{L}_{MLR}(\theta) = -\frac{1}{\omega_i} + 2\mathbb{E}_{s,a\sim\rho(s,a|i)}\left[\frac{\frac{\exp(r_i(s,a;\psi))}{\omega_i^2}}{2\rho(s,a|i)}\right]. \tag{A25}$$

The optimal value results by setting the derivative to zero:

$$\omega_i^* = \mathbb{E}_{s,a\sim\rho(s,a|i)}\left[\frac{\exp(r_i(s,a;\psi))}{\rho(s,a|i)}\right]. \tag{A26}$$

A.4 Proof of Theorem 1

Replacing Eq. (A23) in Eq. (A21) and separating the terms independent of ψ yields:

$$\begin{aligned}
\mathcal{L}_{MLR}(\theta) &= -\mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}\left[\log\frac{\frac{1}{\omega_i}\exp(r_i(s,a;\psi))}{2\rho(s,a|i)}\right] \\
&\quad - \mathbb{E}_{s,a\sim q(s,a),i\sim q(i|s,a)}\left[\log\frac{q(s,a|i)}{2\rho(s,a|i)}\right] \\
&= -\mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}\left[r_i(s,a;\psi) - \log\omega_i - \log 2\rho(s,a|i)\right] \\
&\quad - \mathbb{E}_{s,a\sim q(s,a),i\sim q(i|s,a)}\left[\log q(s,a|i) - \log 2\rho(s,a|i)\right] \\
&= -\mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}[r_i(s,a;\psi)] \\
&\quad + 2\mathbb{E}_{s,a\sim\rho(s,a),i\sim\rho(i|s,a)}[\log 2\rho(s,a|i)] + g(\omega).
\end{aligned} \tag{A27}$$

Taking the derivative with respect to ψ leads to:

$$\begin{aligned}
\partial_\psi\mathcal{L}_{MLR}(\theta) &= -\mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}[\partial_\psi r_i(s,a;\psi)] \\
&\quad + \mathbb{E}_{s,a\sim\rho(s,a),i\sim\rho(i|s,a)}\left[\frac{\frac{\exp(r_i(s,a;\psi))}{\omega_i}\partial_\psi r(s,a;\psi)}{\rho(s,a|i)}\right].
\end{aligned} \tag{A28}$$

Now by setting $\omega = \omega^*$, i.d $\omega_i = Z_i(\psi)$, we have:

$$\begin{aligned}
\partial_\psi\mathcal{L}_{MLR}(\psi,\omega^*) &= -\mathbb{E}_{s,a\sim p(s,a),i\sim\gamma(i|s,a)}[\partial_\psi r_i(s,a;\psi)] \\
&\quad + \mathbb{E}_{s,a\sim\rho(s,a),i\sim\rho(i|s,a)}\left[\frac{\frac{\exp(r_i(s,a;\psi))}{Z_i(\psi)}\partial_\psi r(s,a;\psi)}{\rho(s,a|i)}\right] \\
&= \partial_\psi\mathcal{L}_{PDE}(\psi).
\end{aligned} \tag{A29}$$

A.5 Equivalency of InfoGAIL and IntentionGAIL objectives

The main objective function of the InfoGAIL is (equation 3 in Li et al. (2017) without the constant coefficients and the entropy term):

$$\max_{\pi}\min_{Q,D}\mathbb{E}_{\pi}[\log D(s,a)] + \mathbb{E}_{\pi_E}[\log(1-D(s,a)] + L_1(\pi,Q) \tag{A30}$$

where:

$$L_1(\pi,Q) = \mathbb{E}_{c\sim p(c),a\sim\pi(.|s,c)}\left[\log Q(c|\tau)\right] + H(c) \tag{A31}$$

Please note that on page 4, paragraph 4 of the InfoGAIL paper (Li et al., 2017), the authors of InfoGAIL use a simplified posterior approximation $Q(c|\tau) \approx Q(c|s,a)$, to avoid working

with entire trajectories. Furthermore, in the InfoGAIL paper (Li et al., 2017), the authors have used the letters "c" and "Q" to address the intention and posterior function, respectively. In order to have more consistent notations, the letters "c" and "Q" are replaced with the letters "i" and "p", respectively. Given these, the final objective function will be:

$$\max_{\pi} \min_{D} \underbrace{\mathbb{E}_{\pi}[\log D(s,a)]}_{(1)} + \underbrace{\mathbb{E}_{\pi_E}[\log(1-D(s,a)]}_{(2)} + \underbrace{\mathbb{E}_{i\sim p(i),a\sim\pi(.|s,i)}[\log p(i|s,a)]}_{(3)} + \underbrace{H(i)}_{(4)} \tag{A32}$$

On the other hand, the main objective function of IntentionGAIL is (equation 8 in Hausman et al. (2017) without the constant coefficients and the entropy term):

$$\max_{\theta} \min_{\omega} \underbrace{\mathbb{E}_{\pi_\theta}[\log D_\omega(s,a)]}_{(1)} + \underbrace{\mathbb{E}_{\pi_E}[\log(1-D_\omega(s,a)]}_{(2)} + \underbrace{\mathbb{E}_{i\sim p(i),(s,a)\sim\pi_\theta}[\log p(i|s,a)]}_{(3)} + \underbrace{H(i)}_{(4)} \tag{A33}$$

where the identical terms with respect to the Eq. (A32) are labeled. As can be seen, the objective functions of both InfoGAIL and IntentionGAIL are identical.

A.6 Proof of Lemma 2

The Info/IntentionGAIL (Li et al., 2017; Hausman et al., 2017) corresponds to the following max-min game:

$$\begin{aligned}\max_{q,P} \min_{D} V(D,q,P) = &-\mathbb{E}_{s,a\sim p(s,a)}\big[\log D(s,a)\big] - \mathbb{E}_{s,a\sim q(s,a)}[\log(1-D(s,a)) \\ &+ \lambda\mathbb{E}_{i\sim q(i|s,a)}[\log P(i|s,a)]].\end{aligned} \tag{A34}$$

For a fixed $q(s,\ a)$ and $P(i|s,a)$, the optimal discriminator $D^*(s,a)$ is Goodfellow et al. (2014):

$$D^*(s,a) = \frac{p(s,a)}{p(s,a)+q(s,a)}. \tag{A35}$$

The max–min game can now be reformulated as:

$$\begin{aligned}\max_{q,P} V(D^*,q,P) = &-\mathbb{E}_{s,a\sim p(s,a)}\left[\log\frac{p(s,a)}{p(s,a)+q(s,a)}\right] \\ &- \mathbb{E}_{s,a\sim q(s,a)}\left[\log\frac{q(s,a)}{p(s,a)+q(s,a)} + \lambda\mathbb{E}_{i\sim q(i|s,a)}[\log P(i|s,a)]\right].\end{aligned} \tag{A36}$$

For a fixed $P(i|s,a)$, the maximum with respect to $q(s,\ a)$ happens at $p(s,a) = q^*(s,a)$ (Goodfellow et al., 2014). At this point, $V(D^*,q^*,P)$ achieves:

$$\max_{P} V(D^*, q^*, P) = \log 4 + \mathbb{E}_{s,a\sim q(s,a)}\left[\lambda \mathbb{E}_{i\sim q(i|s,a)}[\log P(i|s,a)]\right]. \tag{A37}$$

Finally, the max-min game is reached to the global optimal, when the log likelihood of posterior function is maximized, i.e. optimal posterior $P^*(i|s,a)$:

$$V(D^*, q^*, P^*) = \log 4 + \mathbb{E}_{s,a\sim q(s,a)}\left[\lambda \mathbb{E}_{i\sim q(i|s,a)}[\log P^*(i|s,a)]\right]. \tag{A38}$$

For EM-MIRL with an optimal posterior $P^*(i|s,a)$, i.e. $p(i|s,a) = P^*(i|s,a)$ we have the following max-min game:

$$\begin{aligned}\max_{q}\min_{D} V(D, q, P^*) = &- \mathbb{E}_{s,a\sim p(s,a),i\sim p(i|s,a)}\left[\log D(s,a,i)\right] \\ &- \mathbb{E}_{s,a\sim q(s,a),i\sim q(i|s,a)}\left[\log(1 - D(s,a,i))\right].\end{aligned} \tag{A39}$$

For a fixed $q(s,a|i)$, the optimal discriminator $D^*(s,a,i)$ is Goodfellow et al. (2014):

$$D^*(s,a,i) = \frac{p(s,a|i)}{p(s,a|i) + q(s,a|i)}. \tag{A40}$$

The max-min game can now be reformulated as:

$$\begin{aligned}\max_{q} V(D^*, q, P^*) = &- \mathbb{E}_{s,a\sim p(s,a),i\sim p(i|s,a)}\left[\log \frac{p(s,a|i)}{p(s,a|i) + q(s,a|i)}\right] \\ &- \mathbb{E}_{s,a\sim q(s,a),i\sim q(i|s,a)}\left[\log \frac{q(s,a|i)}{p(s,a|i) + q(s,a|i)}\right].\end{aligned} \tag{A41}$$

The maximum of $V(D^*, q^*, P^*)$ is achieved for $p(s,a|i) = q^*(s,a|i)$ (Goodfellow et al., 2014) with the value of $\log 4$:

$$V(D^*, q^*, P^*) = \log 2 + \log 2 = \log 4. \tag{A42}$$

Author contributions The authors' contributions are: Methodology, formal analysis and investigation: AB; Writing—original draft preparation: AB; Writing—review and editing: PJ and GD, Supervision: PJ and GD; Resources: GD.

Funding This research has received funding from ECSEL JU project PRYSTINE in collaboration with the European Union's 2020 Framework Programme and National Authorities, under grant agreement no. 783190.

Data availibility Necessary data and materials are available with the code.

Code availability For easing the reproducibility of our work, the code of our method is shared with the community https://github.com/tue-mps/EM-GAIL.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Consent to participate Not applicable.

Consent for publication Not applicable.

Ethical approval Not applicable.

References

Almingol, J., Montesano, L., & Lopes, M. (2013). Learning multiple behaviors from unlabeled demonstrations in a latent controller space. In *International conference on machine learning* (pp. 136–144).

Babes, M., Marivate, V., Subramanian, K., & Littman, M. L. (2011). Apprenticeship learning about multiple intentions. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 897–904).

Belogolovsky, S., Korsunsky, P., Mannor, S., Tessler, C., & Zahavy, T. (2021). Inverse reinforcement learning in contextual MDPs. *Machine Learning*, 1–40.

Bighashdel, A., Meletis, P., Jancura, P., & Dubbelman, G. (2021). Deep adaptive multi-intention inverse reinforcement learning. In *Proceeding of joint European conference on machine learning and knowledge discovery in databases* (pp. 206-221).

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). Openai gym. arXiv preprint arXiv:1606.01540

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*, 2172–2180.

Chen, L., Paleja, R., Ghuy, M., & Gombolay, M. (2020). Joint goal and strategy inference across heterogeneous demonstrators via reward network distillation. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction* (pp. 659–668).

Chen, S.-A., Tangkaratt, V., Lin, H.-T., & Sugiyama, M. (2020). Active deep Q-learning with demonstration. *Machine Learning, 109*(9), 1699–1725.

Choi, J., & Kim, K. -E. (2012). Nonparametric Bayesian inverse reinforcement learning for multiple reward functions. *Advances in neural information processing systems* (pp. 305–313).

Ding, Y., Florensa, C., Abbeel, P., & Phielipp, M. (2019). Goal-conditioned imitation learning. Advances in Neural Information Processing Systems *32*.

Edwards, A., Sahni, H., Schroecker, Y., & Isbell, C. (2019). Imitating latent policies from observation. *International Conference on Machine Learning*, 1755–1763.

Fang, B., Jia, S., Guo, D., Xu, M., Wen, S., & Sun, F. (2019). Survey of imitation learning for robotic manipulation. *International Journal of Intelligent Robotics and Applications, 3*(4), 362–369.

Finn, C., Christiano, P., Abbeel, P., & Levine, S. (2016). A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. arXiv preprint arXiv:1611.03852

Finn, C., Levine, S., & Abbeel, P. (2016). Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning* (pp. 49–58).

Fu, J., Luo, K., & Levine, S. (2018). Learning robust rewards with adversarial inverse reinforcement learning. *International Conference on Learning Representations*.

Gleave, A., & Habryka, O. (2018). Multi-task maximum entropy inverse reinforcement learning. arXiv preprint arXiv:1805.08882

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2672–2680.

Gutmann, M., & Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 297–304). JMLR Workshop and Conference Proceedings.

Hausman, K., Chebotar, Y., Schaal, S., Sukhatme, G., & Lim, J. J. (2017). Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets. *Advances in neural information processing systems*, 235–1245.

Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. *Advances in Neural Information Processing Systems*, 4565–4573.

Hsiao, F. -I., Kuo, J. -H., & Sun, M. (2019). Learning a multi-modal policy via imitating demonstrations with mixed behaviors. arXiv preprint arXiv:1903.10304

Hussein, A., Gaber, M. M., Elyan, E., & Jayne, C. (2017). Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR), 50*(2), 1–35.

Kangasrääsiö, A., & Kaski, S. (2018). Inverse reinforcement learning from summary data. *Machine Learning, 107*(8), 1517–1535.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations*.

Li, K., Gupta, A., Reddy, A., Pong, V. H., Zhou, A., Yu, J., & Levine, S. (2021). MURAL: Meta-learning uncertainty-aware rewards for outcome-driven reinforcement learning. In *International conference on machine learning* (pp. 6346–6356). PMLR.

Li, Y., Song, J., & Ermon, S. (2017). Infogail: Interpretable imitation learning from visual demonstrations. *Advances in Neural Information Processing Systems*, 3812–3822.

Likmeta, A., Metelli, A. M., Ramponi, G., Tirinzoni, A., Giuliani, M., & Restelli, M. (2021). Dealing with multiple experts and non-stationarity in inverse reinforcement learning: an application to real-life problems. *Machine Learning*, 1–36.

Lin, J., & Zhang, Z. (2018). Acgail: Imitation learning about multiple intentions with auxiliary classifier gans. In *Pacific rim international conference on artificial intelligence* (pp. 321–334). Springer.

Michie, D., Bain, M., & Hayes-Miches, J. (1990). Cognitive models from subcognitive skills. *IEE Control Engineering Series, 44,* 71–99.

Michini, B., & How, J.P. (2012). Bayesian nonparametric inverse reinforcement learning. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 148–163). Springer.

Michini, B., Walsh, T. J., Agha-Mohammadi, A.-A., & How, J. P. (2015). Bayesian nonparametric reward learning from demonstration. *IEEE Transactions on Robotics, 31*(2), 369–386.

Morton, J., & Kochenderfer, M. J. (2017). Simultaneous policy learning and latent state inference for imitating driver behavior. In *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)* (pp. 1–6). IEEE.

Neu, G., & Szepesvári, C. (2009). Training parsers by inverse reinforcement learning. *Machine Learning, 77*(2–3), 303.

Ng, A.Y., Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *International conference on machine learning* (Vol. 1, p. 2).

Nikolaidis, S., Ramakrishnan, R., Gu, K., & Shah, J. (2015) Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *2015 10th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 189–196). IEEE.

Rajasekaran, S., Zhang, J., & Fu, J. (2017). Inverse reinforce learning with nonparametric behavior clustering. arXiv preprint arXiv:1712.05514

Ramachandran, D., & Amir, E. (2007). Bayesian inverse reinforcement learning. *IJCAI, 7,* 2586–2591.

Ramponi, G., Likmeta, A., Metelli, A. M., Tirinzoni, A., & Restelli, M. (2020). Truly batch model-free inverse reinforcement learning about multiple intentions. In *International conference on artificial intelligence and statistics* (pp. 2359–2369). PMLR.

Ranchod, P., Rosman, B., & Konidaris, G. (2015). Nonparametric bayesian reward segmentation for skill discovery using inverse reinforcement learning. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 471–477). IEEE.

Ratliff, N. D., Bagnell, J. A., & Zinkevich, M. A. (2006). Maximum margin planning. In *Proceedings of the 23rd international conference on machine learning* (pp. 729–736). ACM.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning* (pp. 1889–1897).

Seyed Ghasemipour, S. K., Gu, S. S., & Zemel, R. (2019). Smile: Scalable meta inverse reinforcement learning through context-conditional policies. *Advances in Neural Information Processing Systems*, *32*.

Syed, U., Bowling, M., & Schapire, R .E. (2008). Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on machine learning* (pp. 1032–1039). ACM.

Tangkaratt, V., Charoenphakdee, N., & Sugiyama, M. (2021). Robust imitation learning from noisy demonstrations. In *AISTATS*.

Tangkaratt, V., Han, B., Khan, M. E., & Sugiyama, M. (2020). Variational imitation learning with diverse-quality demonstrations. In *International Conference on Machine Learning* (pp. 9407–9417). PMLR.

Todorov, E., Erez, T., & Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems* (pp. 5026–5033). IEEE.
Wang, P., Li, H., & Chan, C. -Y. (2021). Meta-adversarial inverse reinforcement learning for decision-making tasks. In *2021 IEEE international conference on robotics and automation (ICRA)* (pp. 12632–12638). IEEE.
Wang, Z., Merel, J. S., Reed, S. E., de Freitas, N., Wayne, G., & Heess, N. (2017). Robust imitation of diverse behaviors. *Advances in Neural Information Processing Systems*, 5320–5329.
Wu, Y. -H., Charoenphakdee, N., Bao, H., Tangkaratt, V., & Sugiyama, M. (2019). Imitation learning from imperfect demonstration. In *International Conference on Machine Learning* (pp. 6818–6827). PMLR.
Wulfmeier, M., Ondruska, P., & Posner, I. (2015). Maximum entropy deep inverse reinforcement learning. arXiv preprint arXiv:1507.04888
Xu, K., Ratner, E., Dragan, A., Levine, S., & Finn, C. (2019). Learning a prior over intent via meta-inverse reinforcement learning. In *International conference on machine learning* (pp. 6952–6962). PMLR.
Yang, Y., Xu, D., Nie, F., Yan, S., & Zhuang, Y. (2010). Image clustering using local discriminant models and global integration. *IEEE Transactions on Image Processing, 19*(10), 2761–2773.
Yu, T., Finn, C., Xie, A., Dasari, S., Zhang, T., Abbeel, P., & Levine, S. (2018). One-shot imitation from observing humans via domain-adaptive meta-learning. arXiv preprint arXiv:1802.01557
Yu, X., Lyu, Y., & Tsang, I. (2020). Intrinsic reward driven imitation learning via generative model. In *International conference on machine learning* (pp. 10925–10935). PMLR.
Yu, L., Yu, T., Finn, C., & Ermon, S. (2019). Meta-inverse reinforcement learning with probabilistic context variables. *Advances in Neural Information Processing Systems32*.
Zheng, B., Verma, S., Zhou, J., Tsang, I., & Chen, F. (2021). Imitation learning: progress, taxonomies and opportunities. arXiv preprint arXiv:2106.12177
Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations*.

Authors and Affiliations

Ariyan Bighashdel[1] · Pavol Jancura[1] · Gijs Dubbelman[1]

✉ Ariyan Bighashdel
a.bighashdel@tue.nl

Pavol Jancura
p.jancura@tue.nl

Gijs Dubbelman
g.dubbelman@tue.nl

[1] Electrical Engineering, Eindhoven University of Technology, 5612AZ Eindhoven, The Netherlands