What Is It You Really Want of Me?
Generalized Reward Learning with Biased Beliefs about Domain Dynamics

Ze Gong and **Yu Zhang**
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University, Tempe, AZ 85281 USA
{zgong11, yzhan442}@asu.edu

Abstract

Reward learning as a method for inferring human intent and preferences has been studied extensively. Prior approaches make an implicit assumption that the human maintains a correct belief about the robot's domain dynamics. However, this may not always hold since the human's belief may be biased, which can ultimately lead to a misguided estimation of the human's intent and preferences, which is often derived from human feedback on the robot's behaviors. In this paper, we remove this restrictive assumption by considering that the human may have an inaccurate understanding of the robot. We propose a method called *Generalized Reward Learning with biased beliefs about domain dynamics (GeReL)* to infer both the reward function and human's belief about the robot in a Bayesian setting based on human ratings. Due to the complex forms of the posteriors, we formulate it as a variational inference problem to infer the posteriors of the parameters that govern the reward function and human's belief about the robot simultaneously. We evaluate our method in a simulated domain and with a user study where the user has a bias based on the robot's appearances. The results show that our method can recover the true human preferences while subject to such biased beliefs, in contrast to prior approaches that could have misinterpreted them completely.

Introduction

With the rapid advancement in AI and robotics, intelligent agents begin to play an important role in our lives in many different areas. Robots will soon be expected to not only achieve tasks alone, but also engage in tasks that require close collaboration with their human teammates. In such situations, the ability of the robot to understand the human's intent and preferences becomes a determinant for achieving effective human-robot teaming. The problem of inferring human's intent and preferences has been studied extensively before. Some researchers (Ng and Russell 2000) formulated this problem as an Inverse Reinforcement Learning (IRL) problem (Russell 1998). The reward function is recovered from optimal policies or behaviors demonstrated by human experts. Such expert demonstrations, however, are often difficult to obtain in real-world tasks. To address this problem, learning methods based on non-expert user ratings of the robot's behaviors (Daniel et al. 2014; Dorsa Sadigh, Sastry, and Seshia 2017; Cui and Niekum 2018) are developed. A common assumption made implicitly in all these prior works is that the human always maintains a correct understanding of the robot's domain dynamics. This, however, may not be the case in many scenarios especially with non-expert users. Having a biased belief about the robot could lead to biased (not just noisy) ratings for the robot's behaviors, resulting in an inaccurate estimation of the human's reward function.

Consider a robot vacuum cleaner that is tasked to clean the floors in a house. Suppose that the robot vacuum is designed to clean most floor types except for hardwood since it is too slippery for the robot to grasp onto (so it may be stuck in a room with a hardwood floor once entered). Consider a user who is asked to rate the robot's behaviors. Given a set of trajectories of the robot cleaner (with most of the areas covered except for the living room with a hardwood floor), the robot may get low ratings even though it should have received high ratings had the user known about the robot's capabilities (which are expressed in terms of domain dynamics). On the other hand, the robot may receive high ratings (even though it should not have) when it stays (stuck) in the living room but somehow manages to clean it (albeit much less efficiently), if the user had the belief that the robot was designed to clean only one room at a time.

In this paper, we remove the restrictive assumption that humans have a correct belief about the robot's domain dynamics. Our goal is to recover the true reward function under biased beliefs. We refer to this problem as Generalized Reward Learning (GRL) and propose a method called *Generalized Reward Learning with biased beliefs about domain dynamics (GeReL)* that infers the latent variables governing both the reward function and human's belief together in a Bayesian setting based on human ratings of the robot's behaviors. Due to the complex forms of the posteriors, the problem is formulated in a variational inference framework (Jordan et al. 1999; Bishop 2006). The variational posterior distribution of the latent variables for estimating the true posterior is optimized using a black-box optimization method (Ranganath, Gerrish, and Blei 2014). To reduce the variance of Monte Carlo estimates of the variational gradients, we factorize the updating rules according to the independence of the latent variables and apply control variate to make the optimization converge faster. By inferring the reward function and the human's belief about the robot simultaneously in this way, our learning method is able to recover the true human preferences while at the same time maintain an estimate of the human's biased belief. As such, our method addresses a key limitation of the existing methods and hence has broad impacts on improving the applicability and safety of robotic systems that work closely with humans.

To evaluate our method, we perform experiments in a simulated navigation domain and with a user study in the Coffee Robot domain (Boutilier, Dearden, and Goldszmidt 2000; Sigaud and Buffet 2013) where biases are introduced by varying the robot's appearances. We compare GeReL with a variant of Simultaneous Estimation of Rewards and Dynamics (SERD) (Herman et al. 2016), Maximum Entropy IRL (MaxEnt-IRL) (Ziebart et al. 2008), and another baseline approach that uses our inference method but maintains the same assumption as in MaxEnt-IRL (that the human's understanding of the robot is correct). In the latter two methods, the true domain dynamics is used and held fixed during learning. Results show that GeReL can better recover the true reward function under such biased beliefs when compared to these other methods. Furthermore, when biases are present, the learned preferences could be completely opposite to the ground truth, suggesting that such a method is indeed valuable for addressing biases in robotic applications.

Related Work

Researchers have formulated the problem of inferring the human's intent and preferences as an IRL problem (Russell 1998) where the goal is to recover the human's preferences as a reward function. IRL is often solved using various optimization techniques with expert demonstrations as the input (Ng and Russell 2000). IRL has also been applied to apprenticeship learning (Abbeel and Ng 2004) to directly approximate the expert's policy. In order to deal with noise in the demonstrations, (Ziebart et al. 2008; Boularias, Kober, and Peters 2011) proposed a probabilistic approach based on the principle of maximum entropy. Furthermore, Bayesian IRL (Ramachandran and Amir 2007) is introduced that incorporates prior knowledge.

However, expert demonstrations (with or without noise) are often difficult to obtain in real-world tasks. More recently, researchers start focusing on learning with non-expert feedback on the queries of the robot's behaviors, often in the forms of ratings (Daniel et al. 2014), comparisons (Dorsa Sadigh, Sastry, and Seshia 2017), or critiques (Cui and Niekum 2018; Zhang and Dragan 2019). All these prior works rely on an implicit assumption that the non-expert user maintains a correct understanding of the robot's domain dynamics. However, when the user is biased, which is likely under such a non-expert setting, it may lead to learning a wrong reward function.

There exists prior work that considers differences between the human and robot in the forms of differences in domain dynamics (Zhang et al. 2016; 2017; Chakraborti et al. 2017) as well as reward functions (Arnold, Kasenberg, and Scheutz 2017; Russell, Dewey, and Tegmark 2015), where the first direction is particularly relevant to our work. (Zhang et al. 2016; 2017; Zakershahrak et al. 2018) approximate the human's understanding of the robot's behaviors using a learning approach and integrate it into task planning to generate explicable plans to bridge the differences in domain dynamics. Researchers have also investigated directly learning the human's understanding of the domain dynamics using model learning. (Unhelkar and Shah 2019) leverage a factored model of behaviors and partial specifications to recover the agent's true generative model of behaviors. (Reddy, Dragan, and Levine 2018) learn the human's belief about the domain dynamics via inverse soft Q-learning given the human's reward function. In this paper, we consider the human's biased belief about the robot in reward learning. We use the human's ratings of the robot's behaviors to infer both the reward function and human's belief without assuming either is available. Although simultaneously learning the rewards and domain dynamics has been studied before (Herman et al. 2016), it is not applicable to our problem setting where non-expert human ratings are used. Compared to optimizing the likelihood of expert demonstrations by computing the derivatives directly, the posteriors for our method assume more complex forms. Furthermore, our method does not require the value functions to take certain forms (e.g., using soft Bellman equation) to perform well and hence it is more general and effective. Thereby, the smoothing effect of entropy regularized reward function caused by soft Bellman equation is avoided, as we will show.

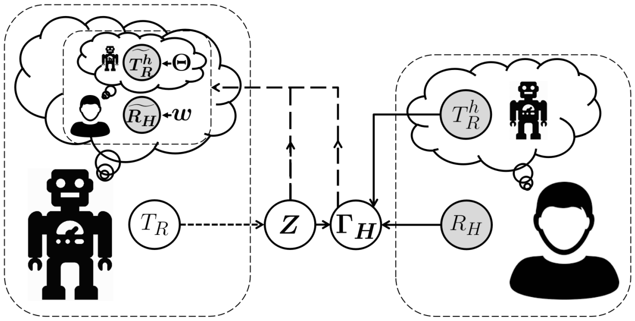

Figure 1: Workflow of GeReL. Using the robot's true transition model T_R, the robot randomly generates a set of demonstrations $\boldsymbol{Z}$ which are evaluated by the human. The human is assumed to provide his rating $\gamma_H \in \boldsymbol{\Gamma_H}$ for each instance $\zeta \in \boldsymbol{Z}$ according to the reward function R_H and his belief T_R^h about the robot's domain dynamics. The ratings will be used to update the estimated reward function $\widetilde{R_H}$ (governed by the parameters $\boldsymbol{w}$) and human's understanding of the robot $\widetilde{T_R^h}$ (governed by the parameters $\boldsymbol{\Theta}$). Gray circles denote the latent variables while the observed are in white.

GeReL

The workflow of GeReL is presented in Figure 1. The robot will first randomly generate a set of demonstrations for querying the human for ratings. Then, the ratings of the demonstrations will be used to infer both the human's reward function and his belief. The system terminates when it meets the convergence criterion. Similar to prior work on reward learning, we assume that the human is to always maximize the rewards (Ng and Russell 2000; Abbeel and Ng 2004), so that his ratings can be estimated given the reward function and his belief of the domain dynamics.

Problem Formulation

More specifically, given a robot's demonstration ζ, we assume that the human would rate it according to two factors, the reward function R_H and his belief about the robot's domain dynamics T_R^h. When the human's belief is different from the true robot's domain dynamics, the rating may be biased and could then lead to a wrong interpretation of the human's preference. This setting introduces the *Generalized Reward Learning (GRL)* problem as follows:

Given:

- Robot's demonstrations $\boldsymbol{Z}$;
- Human's ratings $\boldsymbol{\Gamma_H}$ for each instance in $\boldsymbol{Z}$.

To determine:

- Human's true reward function R_H;
- Human's belief T_R^h about robot's domain dynamics.

To solve this problem, we formulate the environment as a Markov Decision Processes (MDP). An MDP is defined by a tuple (S, A, R, T, λ) where S is a finite set of states, A is a finite set of actions, and $R : S \mapsto \mathbb{R}$ is the reward function that maps each state to a utility value. $T : S \times A \times S \mapsto [0, 1]$ is the transition function that specifies the probability of transitioning to the next state when you take an action in the current state. λ is the discount factor that determines how the agent favors current rewards over future rewards.

Similar to prior work on reward learning (Ng and Russell 2000; Abbeel and Ng 2004), we formulate the reward function R_H for a state s as follows:

$$R_H(s) = \boldsymbol{w} \cdot \Phi(s)$$

where $\Phi = [\phi_0, \phi_1, \ldots, \phi_k]^T$ denotes a set of predefined features for states and $\boldsymbol{w} = [w_0, w_1, \ldots, w_k]^T$ denotes a set of weights for the features. The robot's domain dynamics (i.e., the true domain dynamics) is captured by a transition function and assumed to be given. Likewise, the human's belief about the robot's domain dynamics is modeled also as a transition function T_R^h, which is hidden. T_R^h is assumed to follow a set of probability distributions $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{|S|\times|A|}]$ where $\boldsymbol{\theta}_i = p(s'|s, a)$ is a distribution for a fixed s and a. These distributions capture the human's prior belief about the robot.

To rate a robot's behavior ζ, we assume that the human will first generate his expectation of the robot's behavior as an optimal policy generated using his reward function R_H and belief about the robot T_R^h. Then the behavior of the robot is compared with the optimal policy to generate a rating γ_H. Hence, our learning task in this paper becomes to learn the weights $\boldsymbol{w}$ and transition probability distributions $\boldsymbol{\Theta}$.

Methodology

The inference problem above is often solved by optimizing the posterior probability with respect to the latent variables. However, due to the complex forms of the posteriors, we formulate the problem in a variational inference framework (Jordan et al. 1999; Bishop 2006). Our goal is to approximate the posterior distribution $p(\boldsymbol{w}, \boldsymbol{\Theta}|\boldsymbol{\Gamma_H}, \boldsymbol{Z})$, where $\boldsymbol{\Gamma_H}, \boldsymbol{Z}$ are the observations and $\boldsymbol{w}, \boldsymbol{\Theta}$ the latent variables.

We assume that $\boldsymbol{w}$ follows a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For simplicity, we assume that $\boldsymbol{\Sigma}$ is given a priori. For $\boldsymbol{\Theta}$, we need to select a prior for each $\boldsymbol{\theta}_i$ as a probability distribution. We assume that each $\boldsymbol{\theta}_i$ follows a Dirichlet distribution $\textsc{Dir}(\boldsymbol{\alpha}_i)$, which encodes a distribution over distributions. Let $\boldsymbol{\mathcal{A}} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{|S|\times|A|}]$, and thereby, $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{A}}$ are the parameters we need to learn. As a result, our variational posterior distribution becomes $q(\boldsymbol{w}, \boldsymbol{\Theta}|\boldsymbol{\mu}, \boldsymbol{\mathcal{A}})$, which is the posterior of the latent variables that correspond to the reward function and human's belief about the robot's domain dynamics governed by $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{A}}$. It thus transforms the problem of inferring R_H and T_R^h into a problem of finding $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{A}}$ to make $q(\boldsymbol{w}, \boldsymbol{\Theta}|\boldsymbol{\mu}, \boldsymbol{\mathcal{A}})$ to be close to $p(\boldsymbol{w}, \boldsymbol{\Theta}|\boldsymbol{\Gamma_H}, \boldsymbol{Z})$.

As a variational inference problem, we optimize the Evidence Lower BOund (ELBO):

$$\mathcal{L}(q) = \mathbb{E}_{q(\boldsymbol{w},\boldsymbol{\Theta})}\left[\log p(\boldsymbol{\Gamma_H}, \boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta}) - \log q(\boldsymbol{w}, \boldsymbol{\Theta})\right]$$

where $p(\boldsymbol{\Gamma_H}, \boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta})$ is the joint probability of the observations $\boldsymbol{\Gamma_H}, \boldsymbol{Z}$ and latent variables $\boldsymbol{w}, \boldsymbol{\Theta}$. In order to make it computable in our task, we apply black-box variational inference (Ranganath, Gerrish, and Blei 2014) to maximize ELBO via stochastic optimization:

$$\langle \boldsymbol{\mu}, \boldsymbol{\mathcal{A}} \rangle = \langle \boldsymbol{\mu}, \boldsymbol{\mathcal{A}} \rangle + \rho \cdot \nabla_{\langle \boldsymbol{\mu}, \boldsymbol{\mathcal{A}} \rangle} \mathcal{L}(q)$$

where the learning rate ρ follows the Robbins-Monro rules (Robbins and Monro 1951). We compute the gradient of ELBO with respect to the free parameters $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{A}}$ and $\nabla_{\langle \boldsymbol{\mu}, \boldsymbol{\mathcal{A}} \rangle} \mathcal{L}(q)$ is derived as follows:

$$\begin{aligned} \nabla_{\langle \boldsymbol{\mu}, \boldsymbol{\mathcal{A}} \rangle} \mathcal{L}(q) = \mathbb{E}_{q(\boldsymbol{w},\boldsymbol{\Theta})}[&\nabla_{\langle \boldsymbol{\mu}, \boldsymbol{\mathcal{A}} \rangle} \log q(\boldsymbol{w}, \boldsymbol{\Theta}|\boldsymbol{\mu}, \boldsymbol{\mathcal{A}}) \\ &\cdot (\log p(\boldsymbol{\Gamma_H}, \boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta}) - \log q(\boldsymbol{w}, \boldsymbol{\Theta}|\boldsymbol{\mu}, \boldsymbol{\mathcal{A}}))] \end{aligned} \quad (1)$$

From Equation (1), we can see that the gradient of ELBO is the expectation of the multiplication of the *score function* (Hinkley and Cox 1979) (i.e., $\nabla_{\langle \boldsymbol{\mu}, \boldsymbol{\mathcal{A}} \rangle} \log q(\boldsymbol{w}, \boldsymbol{\Theta}|\boldsymbol{\mu}, \boldsymbol{\mathcal{A}})$) and *instantaneous ELBO* (i.e., $\log p(\boldsymbol{\Gamma_H}, \boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta}) - \log q(\boldsymbol{w}, \boldsymbol{\Theta}|\boldsymbol{\mu}, \boldsymbol{\mathcal{A}})$) with respect to our variational posterior distribution. The detailed derivation of $\nabla_{\langle \boldsymbol{\mu}, \boldsymbol{\mathcal{A}} \rangle} \mathcal{L}(q)$ is presented in (Ranganath, Gerrish, and Blei 2014).

The form of $\nabla_{\langle \boldsymbol{\mu}, \boldsymbol{\mathcal{A}} \rangle} \mathcal{L}(q)$ is not directly computable. Given that $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{A}}$ are independent parameters in our setting, we compute $\nabla_{\boldsymbol{\mu}} \mathcal{L}(q)$ and $\nabla_{\boldsymbol{\mathcal{A}}} \mathcal{L}(q)$ respectively and update them separately:

$$\begin{aligned} \boldsymbol{\mu} &= \boldsymbol{\mu} + \rho_{\boldsymbol{\mu}} \cdot \nabla_{\boldsymbol{\mu}} \mathcal{L}(q) \\ \boldsymbol{\mathcal{A}} &= \boldsymbol{\mathcal{A}} + \rho_{\boldsymbol{\mathcal{A}}} \cdot \nabla_{\boldsymbol{\mathcal{A}}} \mathcal{L}(q) \end{aligned}$$

This also allows us to apply the mean-field assumption that gives the following factorization:

$$q(\boldsymbol{w}, \boldsymbol{\Theta}|\boldsymbol{\mu}, \boldsymbol{\mathcal{A}}) = q(\boldsymbol{w}|\boldsymbol{\mu}) \cdot q(\boldsymbol{\Theta}|\boldsymbol{\mathcal{A}})$$

Then we can rewrite the gradient of ELBO as follows:

$$\begin{aligned}&\nabla_{\langle \boldsymbol{\mu}, \boldsymbol{\mathcal{A}} \rangle}\mathcal{L}(q) = \\ &\quad \mathbb{E}_{q(\boldsymbol{w})}\mathbb{E}_{q(\boldsymbol{\Theta})}\left[\nabla_{\langle \boldsymbol{\mu}, \boldsymbol{\mathcal{A}} \rangle}(\log q(\boldsymbol{w}|\boldsymbol{\mu}) + \log q(\boldsymbol{\Theta}|\boldsymbol{\mathcal{A}}))\right. \\ &\quad \left.\cdot (\log p(\boldsymbol{\Gamma_H}, \boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta}) - \log q(\boldsymbol{w}|\boldsymbol{\mu}) - \log q(\boldsymbol{\Theta}|\boldsymbol{\mathcal{A}}))\right]\end{aligned}$$

Take $q(\boldsymbol{w}|\boldsymbol{\mu})$ as an example, following the derivations in (Ranganath, Gerrish, and Blei 2014), the gradient of ELBO with respect to $\boldsymbol{\mu}$ becomes:

$$\begin{aligned}&\nabla_{\boldsymbol{\mu}}\mathcal{L}(q) = \mathbb{E}_{q(\boldsymbol{w})}\left[\nabla_{\boldsymbol{\mu}}(\log q(\boldsymbol{w}|\boldsymbol{\mu}))\right. \\ &\left.\cdot \mathbb{E}_{q(\boldsymbol{\Theta})}\left[\log p(\boldsymbol{\Gamma_H}, \boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta}) - \log q(\boldsymbol{w}|\boldsymbol{\mu}) - \log q(\boldsymbol{\Theta}|\boldsymbol{\mathcal{A}})\right]\right]\end{aligned}$$

Note that the first term $\mathbb{E}_{q(\boldsymbol{w})}\left[\nabla_{\boldsymbol{\mu}}\left(\log q(\boldsymbol{w}|\boldsymbol{\mu})\right] = 0\right.$ (Ranganath, Gerrish, and Blei 2014). Hence the last term in the *instantaneous ELBO* can be considered as a constant with respect to $q(\boldsymbol{w})$ and canceled out. $\nabla_{\boldsymbol{\mu}}\mathcal{L}(q)$ then becomes:

$$\begin{aligned}&\nabla_{\boldsymbol{\mu}}\mathcal{L}(q) = \mathbb{E}_{q(\boldsymbol{w})}\left[\nabla_{\boldsymbol{\mu}}(\log q(\boldsymbol{w}|\boldsymbol{\mu}))\right. \\ &\qquad \left.\cdot (\mathbb{E}_{q(\boldsymbol{\Theta})}\left[\log p(\boldsymbol{\Gamma_H}, \boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta})\right] - \log q(\boldsymbol{w}|\boldsymbol{\mu}))\right] \quad (2)\end{aligned}$$

Different from (Ranganath, Gerrish, and Blei 2014), in our problem, the expectation of the log joint probability $\log p(\boldsymbol{\Gamma_H}, \boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta})$ cannot be canceled out since $\boldsymbol{w}$ and $\boldsymbol{\Theta}$ happen to be in the Markov blanket of each other. Based on the relationship among these variables as shown in Figure 1, the log probability can be factorized as follows:

$$\begin{aligned}&\log p(\boldsymbol{\Gamma_H}, \boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta}) \\ &= \log p(\boldsymbol{\Gamma_H}|\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta}) + \log p(\boldsymbol{w}) + \log p(\boldsymbol{\Theta}) + \log p(\boldsymbol{Z}) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3)\end{aligned}$$

Putting Equation (3) back into Equation (2), we obtain:

$$\begin{aligned}&\nabla_{\boldsymbol{\mu}}\mathcal{L}(q) = \mathbb{E}_{q(\boldsymbol{w})}\left[\nabla_{\boldsymbol{\mu}}(\log q(\boldsymbol{w}|\boldsymbol{\mu}))\right. \\ &\left.\cdot (\mathbb{E}_{q(\boldsymbol{\Theta})}\left[\log p(\boldsymbol{\Gamma_H}|\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta})\right] + \log p(\boldsymbol{w}) - \log q(\boldsymbol{w}|\boldsymbol{\mu}))\right] \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (4)\end{aligned}$$

where the expectation of the terms $\log p(\boldsymbol{\Theta})$ and $\log p(\boldsymbol{Z})$ with respect to $q(\boldsymbol{\Theta})$ are constants and can be canceled out since $\mathbb{E}_{q(\boldsymbol{w})}[\nabla_{\boldsymbol{\mu}}(\log q(\boldsymbol{w}|\boldsymbol{\mu})] = 0$. Now we have obtained the gradient of ELBO with respect to the latent variable $\boldsymbol{\mu}$ as presented in Equation (4). Similarly, the gradient of ELBO with respect to each $\boldsymbol{\alpha}_i \in \boldsymbol{\mathcal{A}}$ is as follows:

$$\begin{aligned}&\nabla_{\boldsymbol{\alpha}_i}\mathcal{L}(q) = \mathbb{E}_{q(\boldsymbol{\theta}_i)}\left[\nabla_{\boldsymbol{\alpha}_i}(\log q(\boldsymbol{\theta}_i|\boldsymbol{\alpha}_i))\right. \\ &\left.\cdot (\mathbb{E}_{q(\boldsymbol{w})}\left[\log p(\boldsymbol{\Gamma_H}|\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta})\right] + \log p(\boldsymbol{\theta}_i) - \log q(\boldsymbol{\theta}_i|\boldsymbol{\alpha}_i))\right]\end{aligned}$$

Both $p(\boldsymbol{w})$ and $p(\boldsymbol{\theta_i})$ are priors, which are assumed to follow a multivariate Gaussian distribution and a Dirichlet distribution respectively.

In the equations above, $\log p(\boldsymbol{\Gamma_H}|\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\Theta}) = \sum \log p(\gamma_H|\zeta, \boldsymbol{w}, \boldsymbol{\Theta})$ since the demonstrations and ratings are conditionally independent from each other. $p(\gamma_H|\zeta, \boldsymbol{w}, \boldsymbol{\Theta})$ indicates how likely the human would give a rating γ_H for the demonstration ζ given the parameters for the reward function and human's belief about the robot. We assume a Gaussian distribution $\mathcal{N}(\gamma_H|\widetilde{\gamma_H}, \Sigma_{\gamma_H})$ where $\widetilde{\gamma_H}$ is the estimated mean of the human's ratings given $\boldsymbol{w}$ and $\boldsymbol{\Theta}$, and Σ_{γ_H} is assumed to be given to simplify the discussion. As discussed earlier, the estimated mean rating of a demonstration is assumed to depend on two factors, the reward function and the human's belief about the robot's domain dynamics. They together determine the human's expectation of the robot's behavior, which corresponds to the optimal policy for the robot in the human's mind. In this paper, we assume that the rating is proportional to the geometric mean of the human's softmax policy applied to the demonstration. Moreover, we define $\gamma_{\max}$ to be a constant that represents the highest rating that may be given. Following our discussion, the estimated human's rating can be computed for a demonstration $\zeta = \{(s_1, a_1), (s_2, a_2) \ldots (s_n, a_n)\}$ as:

$$\widetilde{\gamma_H} = \gamma_{\max} \cdot \left(\prod_{i=1}^{n} \widetilde{\pi}(a_i|s_i)\right)^{\frac{1}{n}}$$

where n is the length of the demonstration, and $\widetilde{\pi}$ is the estimated human's softmax policy computed using $\boldsymbol{w}$ and $\boldsymbol{\Theta}$.

Variance Reduction: The computation of $\nabla_{\boldsymbol{\mu}}\mathcal{L}(q)$ and $\nabla_{\boldsymbol{\alpha}_i}\mathcal{L}(q)$ above cannot be performed directly due to the intractability of computing the expectations. Hence, we approximate the gradients using sampling methods (Hastings 1970). With Monte Carlo samples, the gradients are estimated as follows:

$$\begin{aligned}&\hat{\nabla}_{\boldsymbol{\mu}}\mathcal{L}(q) \triangleq \frac{1}{S}\sum_{s=1}^{S}\left[\nabla_{\boldsymbol{\mu}}\left(\log q(\boldsymbol{w}_s|\boldsymbol{\mu})\right)\right. \\ &\left.\cdot \left(\log p(\boldsymbol{\Gamma_H}|\boldsymbol{Z}, \boldsymbol{w}_s, \boldsymbol{\Theta}_s) + \log p(\boldsymbol{w}_s) - \log q(\boldsymbol{w}_s|\boldsymbol{\mu})\right)\right]\end{aligned}$$

where S is the number of samples, and $\boldsymbol{w}_s \sim q(\boldsymbol{w})$, $\boldsymbol{\Theta}_s \sim q(\boldsymbol{\Theta})$.

These estimated gradients, however, may have a large variance which could hinder the convergence of our approach. Therefore, it is necessary to reduce the variance. (Ross 2002) introduced control variate that represents a family of functions with equivalent expectations. With control variate, we can instead compute the expectation of an alternative function which has a smaller variance. Let f be the function to be approximated, function $\hat{f}$ is defined as:

$$\hat{f} = f - a \cdot (g - \mathrm{E}[g])$$

where g serves as an auxiliary function that has a finite first moment. $\hat{f}$ can be proven to have smaller variances with an equivalent expectation, where the factor a is computed to minimize the variance (Ranganath, Gerrish, and Blei 2014) as $a = \frac{\mathrm{cov}(f,g)}{\mathrm{var}(g)}$. In this paper, we select the expectation of the *score function* (i.e., $\mathbb{E}_{q(\boldsymbol{w})}\left[\nabla_{\boldsymbol{\mu}}\left(\log q(\boldsymbol{w}|\boldsymbol{\mu})\right)\right]$ and $\mathbb{E}_{q(\boldsymbol{\Theta})}\left[\nabla_{\boldsymbol{\alpha}_i}\left(\log q(\boldsymbol{\Theta}|\boldsymbol{\alpha}_i)\right)\right]$) to be g.

We present GeReL in Algorithm 1. Given the robot's demonstrations and the corresponding ratings, we leverage the human's ratings to update the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{A}}$ of our variational posteriors $q(\boldsymbol{w})$ and $q(\boldsymbol{\Theta})$ via stochastic optimization. The gradients of the ELBO with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{A}}$ are approximated using Monte Carlo sampling. Furthermore, we take advantage of control variate to reduce the

Algorithm 1 Generalized Reward Learning with Biased Belief about Domain Dynamics (GeReL)

```
Input: the robot's demonstrations $\boldsymbol{Z}$, human's ratings $\boldsymbol{\Gamma_H}$, variational posteriors $q(\boldsymbol{w})$ and $q(\boldsymbol{\Theta})$, MaxIter
Output: $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{A}}$
Initialize: free parameters $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{A}}$ for $q(\boldsymbol{w})$ and $q(\boldsymbol{\Theta})$
Let $t = 1$.
while $t <$ MaxIter or convergence not met do
    Draw $S$ samples from $q(\boldsymbol{w})$ and $q(\boldsymbol{\Theta})$
    for $s = 1$ to $n$ do
        $\widetilde{\pi} \leftarrow$ the human's expected policy for the robot
        $\widetilde{\Gamma_H} \leftarrow$ estimated human's ratings for $\boldsymbol{Z}$ given $\widetilde{\pi}$
        Compute $f_{\boldsymbol{\mu}}$, $g_{\boldsymbol{\mu}}$, $f_{\boldsymbol{\alpha}_i}$, and $g_{\boldsymbol{\alpha}_i}$
    end for
    Compute $a_{\boldsymbol{\mu}}$ and $a_{\boldsymbol{\alpha}_i}$
    Approximate $\hat{\nabla}_{\boldsymbol{\mu}} L \triangleq \frac{1}{S}\sum_{s=1}^{S}[f_{\boldsymbol{\mu}} - a_{\boldsymbol{\mu}} g_{\boldsymbol{\mu}}]$ and $\hat{\nabla}_{\boldsymbol{\alpha}_i} L \triangleq \frac{1}{S}\sum_{s=1}^{S}[f_{\boldsymbol{\alpha}_i} - a_{\boldsymbol{\alpha}_i} g_{\boldsymbol{\alpha}_i}]$
    Compute learning rates $\rho_{\boldsymbol{\mu}}$ and $\rho_{\boldsymbol{\alpha}_i}$ with $\hat{\nabla}_{\boldsymbol{\mu}} L$, $\hat{\nabla}_{\boldsymbol{\alpha}_i} L$
    Update $\boldsymbol{\mu} = \boldsymbol{\mu} + \rho_{\boldsymbol{\mu}} \hat{\nabla}_{\boldsymbol{\mu}} L$ and $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_i + \rho_{\boldsymbol{\alpha}_i} \hat{\nabla}_{\boldsymbol{\alpha}_i} L$
end while
return $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{A}}$
```

variance of the gradient estimates. Lastly, the parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\mathcal{A}}$, are updated in each iteration with an adapted learning rate based on AdaGrad (Duchi, Hazan, and Singer 2011). GeReL terminates when the convergence criterion is met.

Evaluation

To evaluate our approach, we conduct two sets of experiments in a simulated grid-world navigation domain and a Coffee Robot domain (Boutilier, Dearden, and Goldszmidt 2000; Sigaud and Buffet 2013) with a user study. The simulation will be focusing on validating our learning method under biased beliefs. The user study will serve two purposes, showing that 1) human users are easily biased in our problem setting; 2) our algorithm learns the correct human preferences under such biases, while prior methods that ignore such biases would fail.

Simulated Navigation Domain

In the first experiment, we test the performance of GeReL in a grid-world navigation domain which contains $7 \times 7 = 49$ states. We set one reward state (i.e., location) that has a large positive weight (i.e. 5) and one penalty location with a large negative weight (i.e., -5). They are randomly located at corners of the grid-world. The robot starts at a random state and its goal is to maximize the rewards. There are four actions, $\{1$ (Up), 2 (Down), 3 (Left), 4 (Right)$\}$, which can transfer the agent from the current state to another state.

To test our algorithm, we simulate two types of biased human beliefs about the robot's domain dynamics. 1) *Reversed Up & Down* : the human believes that action 1 would take the robot down and action 2 would move it up instead. 2) *Rotated Belief*: human believes that the action 1 would move the robot left, the action 2 would move it right, the action 3 would move it up and the action 4 would move it down. The human's reward function for each state is defined as a weighted summation of an inverse distance metric to the reward and penalty states (i.e., the closer it is to the state, the more influence that state has on its reward). The demonstrations are randomly generated via the robot's true dynamic model. The human's ratings are simulated using the true human reward function and biased belief about the robot's domain dynamics following a Gaussian distribution.

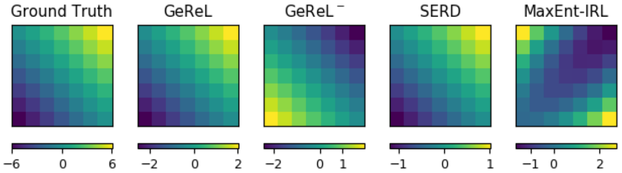

Figure 2: Rewards learned by different approaches.

We compare GeReL with two baseline methods that assume that the human maintains the correct belief about the robot's domain dynamics, namely MaxEnt-IRL (Ziebart et al. 2008) and GeReL$^-$, with the latter basically uses GeReL without updating the domain dynamics. In both baseline methods, the true robot's domain dynamics is used during learning. In addition, a variant of the Simultaneous Estimation of Rewards and Dynamics (SERD) algorithm (Herman et al. 2016) implemented that learns both the reward function and dynamics based on ratings (the original method does not apply to our problem setting) is used in the comparison, which relies on soft value iteration that requires the value functions to assume certain shapes to perform well. To obtain demonstrations for MaxEnt-IRL, we generate them based on the softmax policy of the human. All of the four methods are provided with the same amount of demonstrations. All the results are averaged over multiple runs.

Figure 3 shows the result for the *Reversed Up & Down* setting. The result shows that GeReL can successfully recover the human's reward function and belief about the robot's domain dynamics while GeReL$^-$ and MaxEnt-IRL converge in the completely opposite direction since they do not consider that the human's belief could be biased. On the other hand, SERD converges in the right direction, but the learned values are farther from the ground truth than GeReL in all cases. This is due to the smoothing effect of soft value iteration. In addition, we compute the KL divergence of the softmax policy generated by the estimated reward function and human's belief with that of the ground truth to examine how well we can estimate the human's expectation of the robot's behaviors. Similar trends are observed among all the methods. The comparison of the rewards learned by these four methods with the ground truth is presented in Figure 2. Both GeReL and SERD converge to the correct pattern of rewards in terms of their relative magnitudes. SERD shows less sensitivity to the magnitudes since soft Bellman equation would lead to an entropy augmented reward function (Haarnoja et al. 2017). The adverse effect of learning from biased ratings is clear from the figure for GeReL$^-$ and MaxEnt-IRL, which both fail to recover the true preferences. The results for both settings are presented in Table 1, which

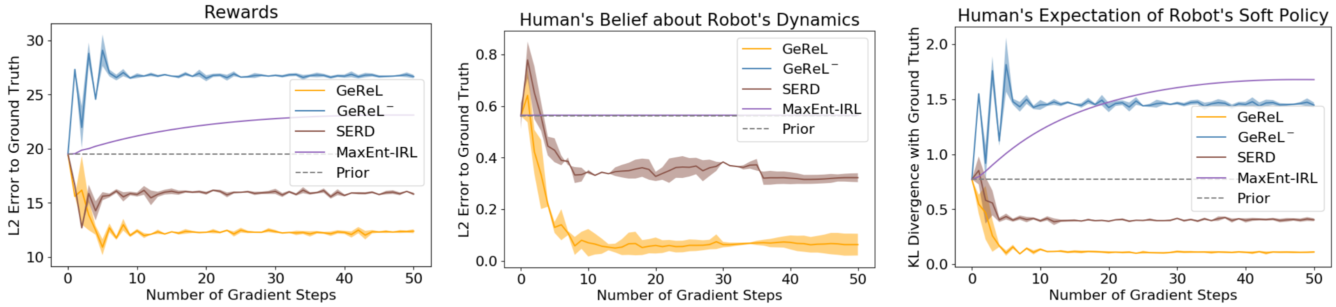

Figure 3: Comparison of the performance among GeReL, SERD, GeReL$^-$, MaxEnt-IRL with the prior also shown. **Left**: The L_2 distance between the learned rewards (which is computed using $\boldsymbol{w}$) and the ground truth. **Middle**: The L_2 distance between the estimated human's belief (i.e., $\boldsymbol{\Theta}$) and the ground truth. **Right**: The KL divergence between the estimated human's expectation of robot's (softmax) policy (which is computed using $\boldsymbol{w}$ and $\boldsymbol{\Theta}$) and that under the ground truth.

	$d(\boldsymbol{R})$	$d(\boldsymbol{\Theta})$	$d(\pi)$	$d(\boldsymbol{R})$	$d(\boldsymbol{\Theta})$	$d(\pi)$
	Reversed Up & Down			*Rotated Belief*		
GeReL	**12.17**	**0.06**	**0.11**	**12.61**	**0.08**	**0.23**
SERD	16.43	0.38	0.47	17.62	0.57	0.63
GeReL$^-$	26.72	0.56	1.46	23.96	0.91	1.43
MaxEnt-IRL	23.32	0.56	1.68	28.02	0.91	1.55

Table 1: Comparison of GeReL, SERD, GeReL$^-$, and MaxEnt-IRL for the two settings in our simulation with respect to the L_2 distance between the estimated values and the ground truth (i.e., $d(\boldsymbol{R})$, $d(\boldsymbol{\Theta})$). The third column (i.e., $d(\pi)$) shows the KL divergence between the estimated human's expectation of the robot's softmax policy and that under the ground truth.

show similar performances in between the two settings. It confirms that GeReL can effectively estimate the human's reward function under biased beliefs.

User Study with the Coffee Robot Domain

Besides the experiments in a simulated domain, we also conduct a user study. Through the study, we hope to demonstrate that humans can be easily biased in our problem setting, which may lead to biased ratings that could have led to a wrong interpretation of the human preference. In such cases, we will show that GeReL can accurately identify the situation. We apply the Coffee Robot domain (Boutilier, Dearden, and Goldszmidt 2000; Sigaud and Buffet 2013) in this user study, which is illustrated in Figure 4. This is a typical factored MDP domain described by 6 binary features which represent whether it is raining, whether the robot has a coffee, etc. The task of the robot is to buy a cup of coffee from a cafe and deliver it to a person in his office. When it is raining, the robot could choose to either operate in the rain or use an umbrella to stay dry. However, using an umbrella while holding the coffee cup may cause the coffee to spill.

To create a situation where biases may be present, we design two experimental settings with two different types of robots: a mobile robot and a humanoid, as seen in Figure 4. We anticipate that the appearance would introduce human biases (Haring et al. 2018) in terms of their capabilities of handling the task. To reduce the effects that the human subject would improve their understanding over time, we generate only 7 demonstrations for each robot that include various scenarios that may occur, such as for a sunny or rainy day, for whether or not the robot takes the umbrella, and for whether or not the robot spills the coffee. *The ground truth for the domain dynamics is set such that the humanoid is less likely to spill the coffee while using the umbrella than the mobile robot.*

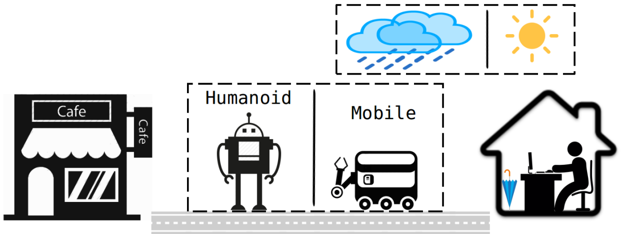

Figure 4: The Coffee Robot domain. The weather could be rainy or sunny, and the robot may choose to use an umbrella or operate in the rain. We use two types of robots (humanoid vs. mobile) to perform the same set of demonstrations that cover the various situations that may occur.

We publish the experiments on Amazon Mechanical Turk (MTurk). To remove invalid responses, we insert a sanity check demonstration with random actions, which should have received the lowest rating. We recruited 20 participants for each setting. After removing those that failed the sanity check or with very short response time (< 3 min), we obtained 12 valid responses for each setting with ages ranging from 23 to 61 (the ratio of males to females is $2 : 1$). Each participant is provided with instructions about the domain at the beginning. To avoid the influence from viewing the demonstrations, immediately after the instructions, we ask the participants two questions as follows:

- **Q1:** *How much more likely do you feel that the robot may spill the coffee while using an umbrella?*
- **Q2:** *How much do you care about the robot being wet?*

The first question is designed to illicit the participant's belief

Question	p-value	Mobile	Humanoid
(Q1) Domain Dynamics	0.047	2.92	3.58
(Q2) Weight Preference	0.027	2.83	3.67

Table 2: Averaged participants' responses to the two questions for each setting before viewing the demonstrations. They are asked in a 5-point Likert scale where 1 is the lowest and 5 the highest.

about the robot's domain dynamics while the second question is for the participant's preference. Their feedback for each setting is presented in Table 2. The participants of the mobile robot setting believed that the robot would be less likely to spill the coffee while holding the umbrella than the participants of the humanoid setting. *Notice that this is in contrast to the ground truth.* Meanwhile, the participants expressed more concern about the robot getting wet in the humanoid setting than the mobile robot setting.

After the questions, we asked the participants to rate the demonstrations. Accordingly, we find that the ratings for the demonstrations where the robot operates in the rain without an umbrella, or takes an umbrella in a sunny day to be rated low in the humanoid setting. In contrast, in the mobile setting, fewer demonstrations received low ratings. These results supported our assumption that the human are easily biased when working with robots.

Next, we run our method under each setting to see whether our method can recover from such biased beliefs. For comparison, we also run GeReL$^-$, which performed similarly to MaxEnt-IRL in our simulation task. We run each method for each participant in both settings. The ratings are normalized to remove inconsistencies across different participants. The results are presented in Figure 5. We observed that the learned probability of spilling coffee while holding an umbrella by GeReL for the humanoid robot setting is generally larger than the mobile robot setting. This represents the estimated human understanding of the domain dynamics, which is consistent with the participant's feedback shown in Table 2. Furthermore, GeReL learned that the participants cared more about the robot getting wet in the humanoid setting than the mobile robot setting, which is also consistent with the participant's true preference. In contrast, GeReL$^-$ discovered just the opposite!

Discussions

Once the biased domain dynamics is obtained, the next question is how to use it. The simplest method of course is to inform the human about his biases and hope that it would work. An alternative method that is often considered in the area of human-aware planning is that the robot could, instead of always pursue optimal behaviors, behave to match with the human's expectation whenever feasible, so as to behave in an explicable manner. In contrast to the multi-objective MDP problem (Roijers et al. 2013; Chatterjee, Majumdar, and Henzinger 2006) which has more than one reward function to optimize, in this problem, the robot maintains two transition functions, one for its own dynamics and the other for the human's belief of it. There already exists work that looks at this problem (Zhang et al. 2017; Chakraborti et al. 2017).

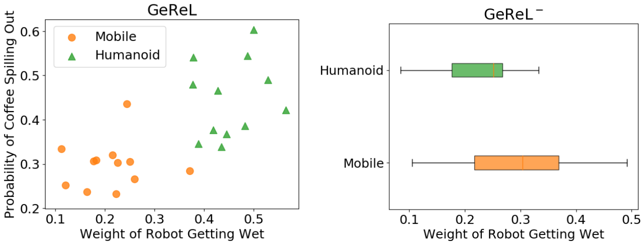

Figure 5: Feature weights learned by GeReL and GeReL$^-$.

Like all reward learning problems, the solution is not unique. This is commonly known as the non-identifiability issue. In general reward learning (GRL), an additional complexity is the learning of the transition function, which unfortunately only aggravates the issue. So far, we are not aware of any solutions to this problem except for the ones that introduce inductive biases on the priors or the error functions, such as Bayesian IRL (Ramachandran and Amir 2007) and apprenticeship learning methods (Abbeel and Ng 2004). In this regard, our work also introduces an inductive bias by assuming a form of the posterior as a multivariate Gaussian distribution.

In terms of simultaneously learning different factors, there exist prior results (Armstrong and Mindermann 2018) that argue against it and prove that it is impossible to determine one without assuming some form of the other. However, we note that the negative results apply only to the situation where one of the factors is the computational process. Consider the function $C(R, M) = \Gamma$. When C is given, the choices of $r \in R$ and $m \in M$ are connected to the corresponding value of $\gamma \in \Gamma$. However, if only m is given, we may choose any r and then simply remap (choosing a $c \in C$) (r, m)'s to their corresponding γ's. This flexibility of the computational process is the core reason of the negative results. However, the non-identifiability issue is still there.

Conclusion

In this paper, we looked the Generalized Reward Learning (GRL) problem and proposed a method called GeReL to address it. GeReL removes the assumption that the human always maintains the true belief about the robot's domain dynamics. To develop the method, we formulated the GRL problem in a variational inference framework to infer the parameters governing the reward function and the human's belief about the robot simultaneously. To reduce the effort for obtaining training samples, we used the human's ratings of robot demonstrations. We evaluated our approach experimentally using a simulated domain and with a user study. The results showed that GeReL outperformed prior approaches that could have misinterpreted the human preferences when such biases are not considered. We showed that GeReL could recover the true human preferences effectively even under such a challenging setting.

Acknowledgments

This research is supported in part by the NSF grant IIS-1844524, the NASA grant NNX17AD06G, and the AFOSR grant FA9550-18-1-0067.

References

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

Armstrong, S., and Mindermann, S. 2018. Occam's razor is insufficient to infer the preferences of irrational agents. In *Advances in Neural Information Processing Systems*, 5598–5609.

Arnold, T.; Kasenberg, D.; and Scheutz, M. 2017. Value alignment or misalignment–what will keep systems accountable? In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

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

Boularias, A.; Kober, J.; and Peters, J. 2011. Relative entropy inverse reinforcement learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 182–189.

Boutilier, C.; Dearden, R.; and Goldszmidt, M. 2000. Stochastic dynamic programming with factored representations. *Artificial intelligence* 121(1-2):49–107.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan explanations as model reconciliation: moving beyond explanation as soliloquy. In *IJCAI*, 156–163. AAAI Press.

Chatterjee, K.; Majumdar, R.; and Henzinger, T. A. 2006. Markov decision processes with multiple objectives. In *Annual Symposium on Theoretical Aspects of Computer Science*, 325–336. Springer.

Cui, Y., and Niekum, S. 2018. Active reward learning from critiques. In *ICRA*, 6907–6914. IEEE.

Daniel, C.; Viering, M.; Metz, J.; Kroemer, O.; and Peters, J. 2014. Active reward learning. In *RSS*.

Dorsa Sadigh, A. D. D.; Sastry, S.; and Seshia, S. A. 2017. Active preference-based learning of reward functions. In *RSS*.

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR* 12(Jul):2121–2159.

Haarnoja, T.; Tang, H.; Abbeel, P.; and Levine, S. 2017. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, 1352–1361. JMLR.org.

Haring, K. S.; Watanabe, K.; Velonaki, M.; Tossell, C. C.; and Finomore, V. 2018. Ffabthe form function attribution bias in human–robot interaction. *IEEE Transactions on Cognitive and Developmental Systems* 10(4):843–851.

Hastings, W. K. 1970. Monte carlo sampling methods using markov chains and their applications.

Herman, M.; Gindele, T.; Wagner, J.; Schmitt, F.; and Burgard, W. 2016. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. In *Artificial Intelligence and Statistics*, 102–110.

Hinkley, D. V., and Cox, D. 1979. *Theoretical statistics*. Chapman and Hall/CRC.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K. 1999. An introduction to variational methods for graphical models. *Machine learning* 37(2):183–233.

Ng, A. Y., and Russell, S. 2000. Algorithms for inverse reinforcement learning. In *in Proc. 17th International Conf. on Machine Learning*.

Ramachandran, D., and Amir, E. 2007. Bayesian inverse reinforcement learning. In *IJCAI*, 2586–2591.

Ranganath, R.; Gerrish, S.; and Blei, D. 2014. Black box variational inference. In *Artificial Intelligence and Statistics*, 814–822.

Reddy, S.; Dragan, A.; and Levine, S. 2018. Where do you think you're going?: Inferring beliefs about dynamics from behavior. In *Advances in Neural Information Processing Systems*, 1461–1472.

Robbins, H., and Monro, S. 1951. A stochastic approximation method. *The annals of mathematical statistics* 400–407.

Roijers, D. M.; Vamplew, P.; Whiteson, S.; and Dazeley, R. 2013. A survey of multi-objective sequential decision-making. *JAIR* 48:67–113.

Ross, S. M. 2002. *Simulation*. Elsevier.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. *AI Magazine* 36(4):105–114.

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

Sigaud, O., and Buffet, O. 2013. *Markov decision processes in artificial intelligence*. John Wiley & Sons.

Unhelkar, V. V., and Shah, J. A. 2019. Learning models of sequential decision-making with partial specification of agent behavior. In *AAAI*.

Zakershahrak, M.; Sonawane, A.; Gong, Z.; and Zhang, Y. 2018. Interactive plan explicability in human-robot teaming. In *RO-MAN*, 1012–1017. IEEE.

Zhang, J., and Dragan, A. 2019. Learning from extrapolated corrections. In *ICRA*, 7034–7040. IEEE.

Zhang, Y.; Sreedharan, S.; Kulkarni, A.; Chakraborti, T.; Zhuo, H. H.; and Kambhampati, S. 2016. Plan explicability for robot task planning. In *Proceedings of the RSS Workshop on Planning for Human-Robot Interaction: Shared Autonomy and Collaborative Robotics*.

Zhang, Y.; Sreedharan, S.; Kulkarni, A.; Chakraborti, T.; Zhuo, H. H.; and Kambhampati, S. 2017. Plan explicability and predictability for robot task planning. In *ICRA*, 1313–1320. IEEE.

Ziebart, B. D.; Maas, A.; Bagnell, J. A.; and Dey, A. K. 2008. Maximum entropy inverse reinforcement learning. In *AAAI*.