2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

A Generalized Acquisition Function for Preference-based Reward Learning

Evan Ellis[1], Gaurav R. Ghosal[1,2], Stuart J. Russell[1], Anca Dragan[1], Erdem Bıyık[1,3]

***Abstract*— Preference-based reward learning is a popular technique for teaching robots and autonomous systems how a human user wants them to perform a task. Previous works have shown that actively synthesizing preference queries to maximize information gain about the reward function parameters improves data efficiency. The information gain criterion focuses on precisely identifying all parameters of the reward function. This can potentially be wasteful as many parameters may result in the same reward, and many rewards may result in the same behavior in the downstream tasks. Instead, we show that it is possible to optimize for learning the reward function up to a behavioral equivalence class, such as inducing the same ranking over behaviors, distribution over choices, or other related definitions of what makes two rewards similar. We introduce a tractable framework that can capture such definitions of similarity. Our experiments in a synthetic environment, an assistive robotics environment with domain transfer, and a natural language processing problem with real datasets demonstrate the superior performance of our querying method over the state-of-the-art information gain method.**

I. Introduction

Reliably inducing desirable behavior in robots is an important prerequisite for their real-world deployment. As desirability is fundamentally human-dependent and subjective, prior work has extensively studied the problem of learning reward functions from human feedback. Although this approach is intuitive, it can be costly to collect sufficient human data to identify a faithful reward function. Moreover, errors in the identified reward function can often lead to highly undesirable consequences [1].

A popular approach for improving data efficiency in reward learning is active learning, where a robot generates the queries that are the most informative for identifying the reward function. This approach focuses human labeling efforts on resolving presently under-specified aspects of the reward function, making it appealing for avoiding hard-to-detect misalignment in reward functions. Prior works have proposed various objectives for quantifying how informative a query is, such as the volume removed from the robot's belief over the reward function [2] or the information gain about the reward function [3].

Existing objectives such as volume removal and information gain optimize for reducing the uncertainty over the reward function parameters. However, what we care about is not the exact weights over features, but instead the correctness of the reward function for the downstream task, as measured by the behavior it induces – be this the distribution over trajectories, how possible candidate trajectories rank or compare, or what the optimal trajectory/policy is. These characteristics are rarely unique in parameter space, and there are often families of reward parameters that are indistinguishable in terms of these properties [4], [5]. Previous methods ignore this fact and often ask queries that yield little benefit in the downstream application.

Our key insight in this work is that *the active learning algorithm should encourage learning the true reward function only up to an equivalence class of statistics over the induced behavior.* Doing so enables the algorithm to focus on learning what matters, such as how the true reward ranks trajectories, and skip questions that are irrelevant to this objective.

To this end, we introduce a novel framework that allows active learning policies to focus on learning the true reward function for an *alignment* metric that captures the functional characteristics we care about when comparing rewards. This alignment metric can leverage prior works in reward distance metrics [6] and can encode distributional information about the domain where we wish to deploy our reward function. Despite the flexibility of our framework, we provide a tractable approximation that holds under mild assumptions.

To validate our approach, we run experiments on three different tasks: a synthetic environment, a simulated assistive robotics environment, and a natural language setting. We show results using three different measures of reward alignment, which induce different equivalent classes over the induced behavior: log-likelihood (does the reward choose the same query answer with the same probability?), EPIC distance [7] (a state-of-the art metric for alignment that goes beyond induced optimal policies), and trajectory rankings (does the reward rank trajectories in the same order with the same magnitudes?). We outperform state-of-the-art performance by up to 85% in learning rewards that transfer well to new domains, using both linear and nonlinear rewards.

II. Related Work

Reward Learning from Human Feedback. Hand-designing a reward function that induces desired behavior is generally challenging and prior works have consequently proposed methods for *learning* reward functions from various forms of human input, such as demonstrations [8], comparisons or rankings [2], [9], [10], physical corrections [11], [12], and

This work was supported by Open Philantrophy and ONR Young Investigator Program (YIP).

Email addresses: {evan.ellis, gauravrghosal, russell, anca}@berkeley.edu, biyik@usc.edu

[1]Department of Electrical Engineering and Computer Sciences, UC Berkeley

[2]Machine Learning Department, Carnegie Mellon University

[3]Thomas Lord Department of Computer Science, University of Southern California

emergency stops [13]. In this work, we primarily consider learning from pairwise trajectory comparisons, although our work generalizes to other feedback types when they can be modeled appropriately.

Active Reward Learning. Data efficiency is crucial in reward learning, as human feedback is typically costly to obtain. *Active learning* techniques seek to alleviate this burden by identifying the queries with the most informative feedback. [14]. Prior works have considered multiple metrics for query informativeness including the volume removed from the hypothesis space [2] and mutual information [3]. Recently, Wilde *et al.* [15] demonstrated existing reward learning objectives that largely depend on parameter-space uncertainty can be suboptimal and proposed a regret-based technique. However, this technique relies on the existence of an efficient method that provides the optimal policy for any given reward function. Similarly, [16] proposes to maximize information gain about the differences among a set of plausibly optimal policies. In this work, we extend their observations and provide a tractable and general method for addressing them.

Reward Metrics. While learned reward functions can be evaluated by computing the ground-truth returns of policies optimized on the learned reward [9], [17], this can introduce confounding factors from policy learning failures. An alternative approach is measuring the distance between the learned and ground-truth reward functions. Simple parameter-space metrics are typically insufficient as they correlate poorly with functional differences between reward functions [6]. For example, transformations of reward functions often yield the same optimal policy and ranking over trajectories [4], [5]. To address this, Gleave *et al.* [7] introduced the EPIC reward psuedometric which provably respects the equivalence class of reward functions inducing the same optimal policy. Wulfe *et al.* [18] refined EPIC to better take into account the likelihood of transitions under an approximate dynamics model. Finally, Balakrishnan *et al.* [19] introduced a reward function projection under which nearby points induce a similar likelihood on expert demonstrations. In this work, we demonstrate how reward metrics can be harnessed for efficient active reward learning.

III. Problem Formulation

We consider a Markov decision process represented with the tuple: $\langle \mathcal{S}, \mathcal{A}, \mu, \Delta, r, T \rangle$ where $\mathcal{S}$ and $\mathcal{A}$ are the state and the action spaces, respectively. The initial state of the system is distributed according to μ, i.e., $s_0 \sim \mu(\cdot)$. The distribution Δ is the transition dynamics of the system such that when action $a_t \in \mathcal{A}$ is taken at state $s_t \in \mathcal{S}$ at timestep t, the next state is $s_{t+1} \sim \Delta(\cdot \mid s_t, a_t)$. T represents the finite horizon. We denote the set of dynamically feasible trajectories as Ξ.

The reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is only known by a human user. We use R to denote the cumulative reward over trajectories: $R(\xi) = \sum_{(s,a)\in\xi} r(s,a)$. The human prefers trajectories with higher cumulative rewards, i.e.,

$$\xi \succ \xi' \iff R(\xi) > R(\xi') \tag{1}$$

for any $\xi, \xi' \in \Xi$. Our goal is to efficiently learn the cumulative reward function R from human feedback.

Human Feedback. We adopt pairwise comparison feedback where the human is presented with a pair of trajectories and is asked to select which one they prefer. We use $Q \in \Xi^2$ to denote the query and $q \in Q$ for the human's response to query Q. We assume access to a human response model conditioned on reward function, i.e., $P(q \mid Q, R)$. While we use this type of feedback in our experiments, our discussion of the existing solutions and the method we propose extend to any type of human feedback for which we have a human response model [20], such as rankings [10], [21], ordinal feedback [22], [23], and corrections [11], [12].

Learning from Human Feedback. Given a dataset of query responses $\mathcal{D}_K = \{(Q_k, q_k)\}_{k=1}^K$, we learn a reward function $R_w : \Xi \to \mathbb{R}$, parameterized by w, in a Bayesian way:

$$p(w \mid \mathcal{D}_K) \propto p(w) p(\mathcal{D}_K \mid w) = p(w) \prod_{k=1}^{K} p(q_k \mid Q_k, R_w) \tag{2}$$

by assuming the human's responses are conditionally independent from each other given the reward function. While we use an uninformative prior $p(w)$ in our experiments, it is possible to inject domain knowledge or other forms of human data (e.g., demonstrations [24]) into learning via the prior. Samples from the posterior distribution give estimates of the true reward function R.

Adaptive Querying. We are interested in an adaptive, human-in-the-loop setting, where we decide the query we will make to the human based on the previous queries and the human's answers to them:

$$Q_k = \pi(\mathcal{D}_{k-1}) \,, \tag{3}$$

where π is the adaptive querying policy.

Objective. Our goal is to actively learn a reward function $R_w : \Xi \to \mathbb{R}$ with a small number of queries such that the learned reward function aligns with the true reward function R under some metric f. Mathematically, we want to find an adaptive querying policy by solving

$$\begin{aligned} &\text{maximize}_\pi \; \mathbb{E}_{w \sim P(w \mid \mathcal{D}_K)} f(R_w, R) \\ &\text{s.t. } Q_k = \pi(\mathcal{D}_{k-1}) \text{ for } k = 1, 2, \ldots, K \end{aligned} \tag{4}$$

We start by reviewing the state-of-the-art active learning method in which an information-theoretic objective is greedily maximized. We then show its drawbacks, and propose our solution in Section VI.

IV. Existing Solutions

The state-of-the-art method for actively choosing queries is mutual information maximization [3]. Given any dataset $\mathcal{D}_{k-1}$ of $k-1$ query-response pairs, the next query is selected to greedily maximize the mutual information between its expected response and the reward function parameters:

$$Q_k^{\text{MI}} = \pi^{\text{MI}}(\mathcal{D}_{k-1}) = \operatorname*{argmax}_{Q \in \Xi^2} I(q; w \mid Q, \mathcal{D}_{k-1}) \,. \tag{5}$$

The rationale behind this method is that the mutual information is the difference between prior and posterior entropies

over w, and maximizing it is equivalent to minimizing the posterior entropy. Therefore, a query that has the highest mutual information is expected to decrease the uncertainty about the reward function parameters w the most.

In fact, this approach outperformed the previous methods (e.g., volume removal [2]) in various metrics, such as the cosine similarity between w and w^* [2], or the loglikelihood over a validation set of queries [25], [26]. However, Wilde *et al.* [15] showed entropy minimization may not be the correct objective for the querying policy if the goal is not minimizing the entropy $H(w \mid \mathcal{D}_K)$. We describe their method and why π^{MI} fails in the next section.

V. Failure Cases of Mutual Information Based Querying Policy

The interesting observation made by Wilde *et al.* [15] is that when the goal is to find the optimal trajectory $\xi^* = \text{argmax}_{\xi \in \Xi} R(\xi)$, the following maximum-regret based greedy optimization performs better than π^{MI}:

$$\begin{aligned} Q_k^{\text{MR}} &= \pi^{\text{MR}}(\mathcal{D}_{k-1}) = (\xi^A, \xi^B) \text{ such that} \\ \xi^A &= \operatorname*{argmax}_{\xi \in \Xi} R_{w^A}(\xi)\,, \xi^B = \operatorname*{argmax}_{\xi \in \Xi} R_{w^B}(\xi)\,, \quad (6) \\ w^A, w^B &= \operatorname*{argmax}_{w^a, w^b} P(w^a \mid \mathcal{D}_{k-1}) P(w^b \mid \mathcal{D}_{k-1}) \\ &\quad (R_{w^a}(\xi^A) - R_{w^a}(\xi^B) + R_{w^b}(\xi^B) - R_{w^b}(\xi^A))\,. \end{aligned}$$

Here, the regret is defined between two reward functions: if the true reward is parameterized by w^A but the system learns w^B, then the regret is $R_{w^A}(\xi^A) - R_{w^A}(\xi^B)$ where ξ^A and ξ^B are the optimal trajectories under R_{w^A} and R_{w^B}, respectively. This method implicitly makes the user choose between a pair of rewards, R_{w^A} and R_{w^B}, which maximizes some regret metric by querying them with (ξ^A, ξ^B).

The reason why maximum regret based policy π^{MR} finds optimal trajectories better/faster than the mutual information based policy π^{MI} is that the mutual information based method tries to greedily minimize the entropy over the reward function parameters, i.e., $H(w \mid \mathcal{D}_k)$. However, many different parameters may lead to the same optimal trajectory (reward ambiguity problem [27]). Therefore, the mutual information based policy π^{MI} may make queries that are wasteful if the true goal is not to reduce entropy over parameters w. On the other hand, even if the goal is to find the optimal trajectory, the maximum regret based policy π^{MR} has practical problems as the first two constraints in Equation (6) require the ability to optimize trajectories for any given reward function. These indicate a need for an efficient and general method that can handle various learning objectives, not just entropy minimization or trajectory optimization.

As another example of a common failure case of π^{MI}, suppose our goal in reward learning is to be able to compare any two trajectories in terms of their rewards, which is indeed true to the motivation of why reward functions exist. In such a case, if some parameters of the reward function are relevant only for a small subset of trajectories, then those parameters are less important than the others, as they will not affect most of the trajectory comparisons. However, π^{MI} will give the same importance to all parameters, because they contribute to the entropy in the same way.

The mutual information based policy π^{MI} is also not suitable for problems that involve domain transfer, i.e., the reward is learned in one domain and then transferred to the other. One may want to learn reward functions in simulation and then deploy the learned reward on a real robot due to safety concerns. Or it may be cheaper to collect human feedback in one domain than the other, e.g., if we think of each text as a trajectory a natural language processing (NLP) system takes, it is easier for most people to compare the writing quality of two paragraphs rather than two scientific articles. Due to the potential distribution shift between the domains, π^{MI} will produce suboptimal queries as it is agnostic to the trajectory distribution of the systems.

These motivate us to develop our novel querying approach that lets designers plug their true alignment objective f.

VI. Our Approach

As we stated in Equation (4), our objective is to find an adaptive querying policy that maximizes

$$\mathbb{E}_{w \sim P(w \mid \mathcal{D}_K)} \left[f(R_w, R) \right]$$

for some f that captures the alignment between the true reward and the learned reward. However, in the most general case, we cannot compute this because we simply do not know the true reward function R.

Approximation. Our approach is based on the observation that $P(w \mid \mathcal{D}_k)$ will give a high probability for w^* that best aligns with the true reward, i.e.,

$$w^* = \operatorname*{argmax}_w f(R_w, R)\,. \quad (7)$$

This observation holds under the mild assumption that $P(q \mid Q, R_{w^*}) \approx P(q \mid Q, R)$ for any query-response pair (Q, q). Practically, this assumption only enforces that f is really an alignment metric to be maximized, instead of adversarial metrics, e.g., one that is maximized when R and R_w make opposite predictions about user responses.

As a result of this observation, our insight is to solve

$$\begin{aligned} &\text{maximize}_\pi \; \mathbb{E}_{w' \sim P(w \mid \mathcal{D}_K)} \left[\mathbb{E}_{w \sim P(w \mid \mathcal{D}_K)} \left[f(R_w, R_{w'}) \right] \right] \quad (8) \\ &\text{s.t. } Q_k = \pi(\mathcal{D}_{k-1}) \text{ for } k = 1, 2, \ldots, K \end{aligned}$$

as a proxy to the original problem, because the outer expectation in the objective is an expectation over cases where w' is the target parameters w^*.

Intuitively, this gives birth to the notion of identifying the reward function up to a certain equivalence class. Say f cared about the induced ranking over trajectories – the original objective incentivizes that we find a reward which ranks trajectories similarly to the true reward; this proxy objective, which is computable based on the information we know, incentivizes that we identify the reward up to rewards that produce the same ranking; at that point, $f(R_w, R_{w'})$ becomes 0 and further queries are no longer helpful.

Greedy soluton. Solving for the optimal adaptive policy π^* is intractable, as it requires planning over K queries each of which is answered stochastically by the user. We follow the prior work by taking a greedy approach to the problem. Merging the expectations in Equation (8), we let

$$\pi^f(\mathcal{D}_{k-1}) = \operatorname*{argmax}_Q \mathbb{E}_{q\sim P(q|Q,\mathcal{D}_{k-1})}\big[\mathbb{E}_{w,w'\sim P(w|\mathcal{D}_{k-1},Q,q)}\left[f(R_w,R_{w'})\right]\big], \quad (9)$$

where we greedily optimize for the expected alignment of the posterior $P(w \mid \mathcal{D}_{k-1}, Q, q)$ under f. The inner expectation is not trivial to compute, as it requires sampling parameter pairs from the posterior for the given query-response pair (Q, q). Given the optimization is over Q, we would need to repeat sampling many times. However, as we show in the appendix, this greedy solution is simplified as:

$$\operatorname*{argmax}_Q \sum_{q\in Q} \frac{\mathbb{E}_{w,w'}\left[P(q \mid Q,R_w)P(q \mid Q,R_{w'})f(R_w,R_{w'})\right]}{\mathbb{E}_{w''}P(q \mid Q,R_{w''})} \quad (10)$$

where all expectations are over the prior, i.e., $P(w \mid \mathcal{D}_{k-1})$, and we only need to evaluate $f(R_w, R_{w'})$ and $P(q \mid Q, R_w)$, which are already given. Therefore, our approach requires computing expectations over the prior and we compute them by sampling from that prior only once for each query Q_k.

Example fs. We present below three useful examples of alignment metrics f that we use in our experiments. Roughly, they correspond to inducing the same answers to trajectory comparisons, mapping to a (canonically) shaped version of the same reward function (EPIC distance), as well as inducing the same ranking over trajectories.

First, we consider an f based on loglikelihood, a popularly used metric in preference-based reward learning [25], [26], [28]. Under this metric, two reward functions align with each other if human response predictions under one of them get high probabilities under the other:

$$f^{\text{LL}}(R_w,R_{w'}) = g^{\text{LL}}(R_w,R_{w'}) + g^{\text{LL}}(R_{w'},R_w) \quad (11)$$
$$g^{\text{LL}}(R_w,R_{w'}) = \sum_{Q\in\mathcal{Q}} \log P(q = \operatorname*{argmax}_{\xi\in Q} R_{w'}(\xi) \,|\, Q, R_w)$$

for some set of queries $\mathcal{Q}$. Ideally, this set should be representative of the environment the learned reward will be deployed to. For example, if the human will give their preferences on a simulator but the learned reward will be deployed on a real robot, the search space of the querying optimization is the simulator trajectories, but $\mathcal{Q}$ should consist of real robot trajectories. We call this variant of our querying policy π^{LL}.

Secondly, distance functions developed to measure the misalignment between reward functions are a natural fit for f. We use EPIC distance [7] to measure alignment in π^{EPIC}:

$$f^{\text{EPIC}}(R_w, R_{w'}) = -\text{EPIC}(R_w, R_{w'}). \quad (12)$$

We refer to [7] for the details on computing EPIC distance.

Similarly, Balakrishnan *et al.* [19] developed a technique called ρ-projection to project reward functions to a space in which L2-distance can be used as a misalignment measure between those functions:

$$f^{\rho}(R_w, R_{w'}) = -\|\rho(R_w), \rho(R_{w'})\|_2 \quad (13)$$
$$\rho(R_\omega) = \frac{[\exp R_w(\xi_1), \exp R_w(\xi_2), \ldots, \exp R_w(\xi_N)]}{\sum_{i=1}^N \exp R_w(\xi_i)}$$

for some trajectories $\xi_1, \xi_2, \ldots, \xi_N$. Intuitively, ρ-projection metric compares how two reward functions rank the given set of trajectories under the Boltzmann rational model. Similar to the query set $\mathcal{Q}$ in π^{LL}, this trajectory set should be representative of the deployment environment. We denote the querying policy that uses this alignment metric with π^{ρ}.

Having presented our active querying policy and example alignment metrics, we will analyze some of its useful properties in the next section.

VII. Analysis

We will make two remarks in our analysis. First, our method can be seen as a generalization of the mutual information based method. Second, for choices of f that makes $\mathbb{E}_{w,w'\sim P(w|\mathcal{D}_k)}\left[f(R_w,R_{w'})\right]$ adaptive monotone and adaptive submodular over the sets $\mathcal{D}_k$, our method is a near-optimal solution to the problem stated in (8).

Remark 1: Slightly abusing the notation to let f depend on $\mathcal{D}_k$, the mutual information based policy π^{MI} can be seen as a special case of our approach, since it is the solution to (10) when $f(R_w, R) = \log P(w \mid \mathcal{D}_k)$.

Proof: Plugging $\log P(w \mid \mathcal{D}_k)$ in (10), we get

$$\operatorname*{argmax}_Q \sum_{q\in Q} \frac{\mathbb{E}_{w,w'}[P(q|Q,R_w)P(q|Q,R_{w'})\log P(w|\mathcal{D}_k)]}{\mathbb{E}_{w''}P(q|Q,R_{w''})}$$

We separate the w and w' terms in the numerator, and note w' term is equivalent to the expectation in the denominator:

$$\operatorname*{argmax}_Q \sum_{q\in Q} \mathbb{E}_{w\sim P(w|\mathcal{D}_{k-1})}\left[P(q \mid Q,R_w)\log P(w \mid \mathcal{D}_k)\right]$$
$$= \operatorname*{argmax}_Q \mathbb{E}_{w,q\sim P(w,q|\mathcal{D}_{k-1},Q)}\left[\log P(w \mid \mathcal{D}_k)\right]$$
$$= \operatorname*{argmax}_Q \mathbb{E}_{q\sim P(q|\mathcal{D}_{k-1},Q)}\left[\mathbb{E}_{w\sim P(w|\mathcal{D}_k)}\left[\log P(w \mid \mathcal{D}_k)\right]\right]$$

where we used the fact that $\mathcal{D}_k = (\mathcal{D}_{k-1}, Q, q)$. Noting the inner expectation is the posterior entropy and the prior entropy $H(w \mid \mathcal{D}_{k-1})$ does not depend on the optimization variable Q, we equivalently write:

$$\operatorname*{argmax}_Q \mathbb{E}_{q\sim P(q|\mathcal{D}_{k-1},Q)}\left[H(w|\mathcal{D}_{k-1}) - H(w|\mathcal{D}_{k-1},Q,q)\right]$$
$$= \operatorname*{argmax}_Q I(q; w \mid Q, \mathcal{D}_{k-1}),$$

which is equal to the optimization in Equation (5). ∎

Generalizing the mutual information based solution introduces some desirable theoretical properties. Bıyık *et al.* [24] noted that there is no known theoretical guarantee for π^{MI}. However, for adaptive monotone and adaptive submodular objectives, our method enjoys the following guarantee.

Remark 2: If $\mathbb{E}_{w,w'\sim P(w|\mathcal{D}_k)}\left[f(R_w,R_{w'})\right]$ is adaptive monotone and adaptive submodular over the sets $\mathcal{D}_k$, the

greedy solution we presented in (10) will, in expectation, achieve at least $(1 - \frac{1}{\epsilon})\text{OPT}_K$ improvement over the objective after K queries, where OPT_K is the theoretical upper bound (due to the optimal but intractable adaptive policy).

Proof: Directly follows from Golovin and Krause [29]. ■

While common acquisition functions in active reward learning do not satisfy adaptive submodularity, Golovin and Krause [29] discuss some possibilities, including works that use adaptive submodular objectives in active learning, e.g., [30], [31]. In the next section, we empirically demonstrate how our method, which is tailored to the objective f of the application, achieves the best results against the baselines including the mutual information optimization.

VIII. Experiments

We conduct experiments in three different domains: a synthetic environment, Assistive Gym that simulates an assistive robot [32], and a natural language processing (NLP) task with datasets curated from Reddit [33]. Following prior work, we use a probabilistic human model to simulate human responses to the preference queries [3], [10], [15].

A. Human Response Model

We simulate the human response q to query Q using a probabilistic model conditioned on the reward function. For this, we use the standard Boltzmann rational model, parameterized by a rationality coefficient β:

$$P(q = \xi \mid Q, R) = \frac{\exp \beta \cdot R(\xi)}{\sum_{\xi' \in Q} \exp \beta \cdot R(\xi')} \tag{14}$$

for any trajectory $\xi \in Q$. For our experiments, we tune β such that around 95% of the simulated responses align with the reward functions as in (1).

B. Metrics

We claim that when the alignment function is f, one should use our active querying method with π^f. Therefore, we use three metrics each of which corresponds to one of the methods: loglikelihood, EPIC distance, and ρ-projection distance. We expect each variant of our algorithm to be the most successful under the corresponding metric.

IX. Results

A. Synthetic Environment

We first evaluate our methods in comparison with π^{MI} on a synthetic environment. This experiment demonstrates the data-efficiency of our methods in learning a reward function that can be transferred to new domains.

For this, we simulate trajectories from a source and a target domain. Trajectories in the source domain have 15 features sampled i.i.d. from $\mathcal{N}(0, 1)$. To simulate the distribution shift between different domains, we let 10 of the features of the target domain have a mixture distribution $\frac{1}{2}\mathcal{N}(-1, 10^{-4}) + \frac{1}{2}\mathcal{N}(1, 10^{-4})$, and let the remaining 5 features have the same distribution as in the source domain. The reward function is a linear combination of these features. The parameters to learn $w \in \mathbb{R}^{15}$ correspond to the weights of the features in the reward function.

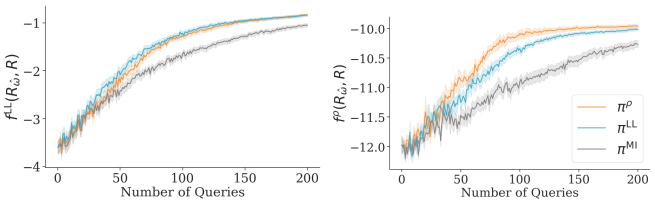

Fig. 1: Results of the synthetic environment experiment over 50 seeds (mean±se).

We randomly generate 50 different true reward parameters w^* to evaluate the methods. This enables us to compute metrics against the true reward function R_{w^*}. This procedure, as well as the linear reward structure, is common in preference-based reward learning literature [2], [3], [15]. We compare π^{MI}, π^{LL}, π^{ρ}, and exclude π^{EPIC} as it depends on more granular information in the trajectories, e.g., state-action-next state tuples, which do not exist in the synthetic data.

Figure 1 shows the results of this experiment. Both π^{ρ} and π^{LL} significantly outperform π^{MI} in both loglikelihood and ρ-projection based alignment metrics. These results strongly support the hypothesis that π^{MI} is suboptimal when the learned reward is deployed in a different environment than the training (source) environment. It also supports the argument that we should use the variant of our algorithm that corresponds to the metric we want to optimize: π^{ρ} outperforms all other methods on the ρ-projection based metric, and π^{LL} outperforms all others on the loglikelihood.

B. Assistive Gym

Next, we evaluate our methods in the Assistive Gym [32] simulated robotics environment. This experiment demonstrates the ability of our methods to learn a nonlinear reward that can be transferred between realistic robotics domains. We consider a robotic arm feeding a patient. The robotic arm must learn where to place the spoon, which is attached to the end effector, by asking preference queries. The experimenters have access to a Sawyer robotic arm (Rethink Robotics) but wish to learn a reward that applies to a Jaco arm (Kinova). Therefore, in this setting, the source domain involves a Sawyer arm and the target domain involves a Jaco arm.

We randomly sample 20 different goal positions for the spoon and let the reward function be the negative distance between the end-effector and the goal positions. However, the goal is not known by the robot and must be learned via preference queries. Hence, the learnable parameters of the reward function w correspond to the goal position.

Figure 2 shows the results of this experiment. π^{LL}, π^{EPIC}, and π^{ρ} all outperform π^{MI} in loglikelihood, rho-projection alignment, and EPIC-distance alignment score, showing that our algorithm succeeds in learning nonlinear rewards for the target domain. The inefficiency of π^{MI} at learning this simple nonlinear reward suggests that taking the deployment environment into account is essential to solving more complex active learning problems.

C. NLP Task

Finally, we evaluate our methods in an NLP task. This is inspired by InstructGPT [34], which fine-tunes a large language model with preferences, a popular practice in NLP.

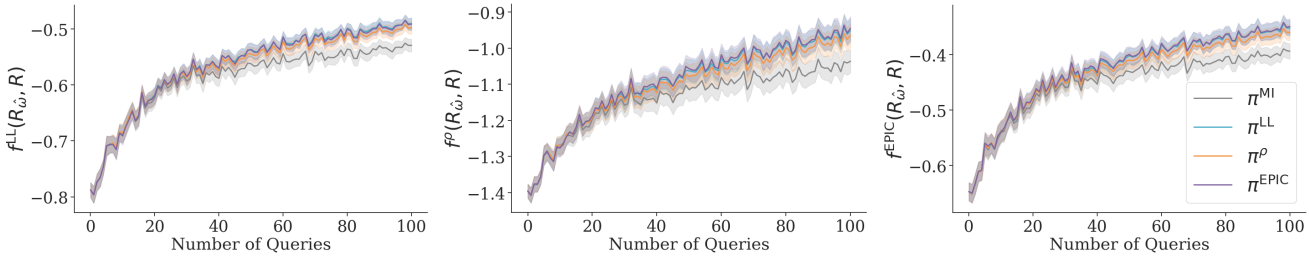

Fig. 2: Results of the Assistive Gym experiment over 100 seeds (mean±se).

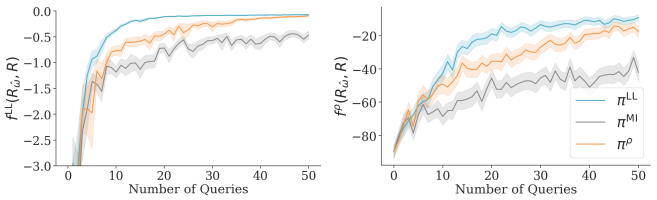

Fig. 3: NLP experiments results over 50 seeds (mean±se).

We consider a setting where a human user may easily compare the quality of texts in one domain, but it is costly in some other domains. For example, texts in the target domain may be much longer than those in the source domain, or the user may be less knowledgeable about the target domain. Both of these cases have been pointed out as limitations of reinforcement learning from human feedback [35], but as we will show, our methods alleviate these problems by enabling data-efficient reward learning on a simpler source domain.

For this, we employ Stanford Human Preferences Dataset [33], which has curated data from Reddit. Specifically, we take r/askvet and r/askphilosophy subreddits, which contain discussions on completely different topics. Our goal is to learn a reward function for writing quality by using preferences on r/askvet and then check if the learned reward aligns with the true preferences in r/askphilosophy. To this end, we model each comment in the subreddits as a trajectory ξ, and we label each of them with: sentiment analysis over "emotion", "hate", "irony", "offensive", "sentiment" [36], Flesch-Kincaid grade level [37], Flesch-Kincaid reading ease [37], Dale-Chall readability [38], Coleman–Liau index [39], automated readability index [40], and the relevance between the comment and its main post according to the model by Liu *et al.* [41]. The reward of each comment is then a linear combination of these features after normalization.

We randomly generate 50 parameter vectors w^*, representing different views on writing quality. We restrict the query space such that we can query the user only with comments that belong to the same main post.

Results from this experiment are shown in Figure 3. Both π^{ρ} and π^{LL} outperform π^{MI} in the log-likelihood and ρ-projection based alignment metrics (π^{EPIC} was excluded as there is no granular information about the trajectories). Surprisingly, π^{LL} performs better than π^{ρ} even when the alignment metric is f^{ρ}. Noting that this is not the case early in the training with a smaller number of queries, we posit this may be because of the greedy approximation to the original optimization problem (see Equation (9)). It is also possible that certain alignment metrics are better suited for some domains than others.

X. Conclusion

We introduced a new method for active preference-based learning of a reward that behaves similarly to the true reward in terms of a user-defined alignment metric. We have shown results comparing our method using three different alignment metrics with the state-of-the-art baseline on various environments. The results demonstrated the advantages of our method in learning both linear and nonlinear rewards.

Future works may investigate different alignment metrics, and their implications on the learned rewards. They may also explore how our methodology can be extended to gradient-based learning methods (as opposed to Bayesian) so that it can be applied to settings where rewards are modeled with a large number of parameters, e.g., deep neural networks.

Acknowledgments

This work was supported by Cocosys SRC center, and an ONR YIP. The authors also acknowledge a gift from Open Philanthropy to support the work of the Center for Human-Compatible AI at UC Berkeley,

Appendix

A. *Derivation of Equation* (10)

We start from Equation (9) and expand the expectations, where integrals are over the entire parameter space for w:

$$\underset{Q}{\operatorname{argmax}} \sum_{q\in Q} \int\int P(q \mid Q, \mathcal{D}_{k-1}) P(w \mid \mathcal{D}_{k-1}, Q, q) \\ P(w' \mid \mathcal{D}_{k-1}, Q, q) f(R_w, R_{w'}) dw dw' .$$

Using Bayes rule, $w \perp Q \mid \mathcal{D}_{k-1}$, and $q \perp \mathcal{D}_{k-1} \mid w, Q$ to replace $P(w \mid \mathcal{D}_{k-1}, Q, q)$ with $\frac{P(w|\mathcal{D}_{k-1})P(q|w,Q)}{P(q|\mathcal{D}_{k-1},Q)}$, we get

$$\underset{Q}{\operatorname{argmax}} \sum_{q\in Q} \frac{1}{P(q \mid \mathcal{D}_{k-1}, Q)} \int\int P(w|\mathcal{D}_{k-1}) P(w'|\mathcal{D}_{k-1}) \\ P(q \mid w, Q) P(q \mid w', Q) f(R_w, R_{w'}) dw dw' .$$

Rewriting the integrals as expectations gives

$$\underset{Q}{\operatorname{argmax}} \sum_{q\in Q} \frac{\mathbb{E}_{w,w'|\mathcal{D}_{k-1}}[P(q|w,Q)P(q|w',Q)f(R_w,R_{w'})]}{P(q \mid \mathcal{D}_{k-1}, Q)}$$

Finally, we note $P(q \mid \mathcal{D}_{k-1}, Q) = \int P(w, q|\mathcal{D}_{k-1}, Q) dw = \int P(w \mid \mathcal{D}_{k-1}) P(q \mid Q, R_w) dw$. Plugging this final expression as an expectation into the objective, we reach the final objective we presented in (10):

$$\underset{Q}{\operatorname{argmax}} \sum_{q\in Q} \frac{\mathbb{E}_{w,w'|\mathcal{D}_{k-1}}[P(q|w,Q)P(q|w',Q)f(R_w,R_{w'})]}{\mathbb{E}_{w''|\mathcal{D}_{k-1}} P(q \mid Q, R_{w''})}$$

References

[1] J. Skalse, N. Howe, D. Krasheninnikov, and D. Krueger, "Defining and characterizing reward gaming," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9460–9471, 2022.

[2] D. Sadigh, A. D. Dragan, S. S. Sastry, and S. A. Seshia, "Active preference-based learning of reward functions," in *Proceedings of Robotics: Science and Systems (RSS)*, Jul. 2017. DOI: 10.15607/RSS.2017.XIII.053.

[3] E. Biyik, M. Palan, N. C. Landolfi, D. P. Losey, and D. Sadigh, "Asking easy questions: A user-friendly approach to active reward learning," in *Proceedings of the 3rd Conference on Robot Learning (CoRL)*, 2019.

[4] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Icml*, Citeseer, vol. 99, 1999, pp. 278–287.

[5] J. Skalse, M. Farrugia-Roberts, S. Russell, A. Abate, and A. Gleave, *Invariance in policy optimisation and partial identifiability in reward learning*, 2023. arXiv: 2203.07475 [cs.LG].

[6] E. Jenner, J. M. V. Skalse, and A. Gleave, "A general framework for reward function distances," in *NeurIPS ML Safety Workshop*, 2022. [Online]. Available: https://openreview.net/forum?id=Hn21kZHiCK.

[7] A. Gleave, M. D. Dennis, S. Legg, S. Russell, and J. Leike, "Quantifying differences in reward functions," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=LwEQnp6CYev.

[8] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 1.

[9] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[10] V. Myers, E. Biyik, N. Anari, and D. Sadigh, "Learning multimodal rewards from rankings," in *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

[11] A. Bajcsy, D. P. Losey, M. K. O'malley, and A. D. Dragan, "Learning robot objectives from physical human interaction," in *Conference on Robot Learning*, PMLR, 2017, pp. 217–226.

[12] A. Bajcsy, D. P. Losey, M. K. O'Malley, and A. D. Dragan, "Learning from physical human corrections, one feature at a time," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 141–149.

[13] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "The off-switch game," in *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[14] D. J. MacKay, "Information-based objective functions for active data selection," *Neural computation*, vol. 4, no. 4, pp. 590–604, 1992.

[15] N. Wilde, D. Kulić, and S. L. Smith, "Active preference learning using maximum regret," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2020, pp. 10 952–10 959.

[16] D. Lindner, M. Turchetta, S. Tschiatschek, K. Ciosek, and A. Krause, *Information directed reward learning for reinforcement learning*, 2022. arXiv: 2102.12466 [cs.LG].

[17] Y. Wang, Z. Sun, J. Zhang, Z. Xian, E. Biyik, D. Held, and Z. Erickson, *Rl-vlm-f: Reinforcement learning from vision language foundation model feedback*, 2024. arXiv: 2402.03681 [cs.RO].

[18] B. Wulfe, A. Balakrishna, L. Ellis, J. Mercat, R. McAllister, and A. Gaidon, *Dynamics-aware comparison of learned reward functions*, 2022. arXiv: 2201.10081 [cs.LG].

[19] S. Balakrishnan, Q. P. Nguyen, B. K. H. Low, and H. Soh, *Efficient exploration of reward functions in inverse reinforcement learning via bayesian optimization*, 2020. arXiv: 2011.08541 [cs.LG].

[20] H. J. Jeon, S. Milli, and A. Dragan, "Reward-rational (implicit) choice: A unifying formalism for reward learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4415–4426, 2020.

[21] D. Brown, W. Goo, P. Nagarajan, and S. Niekum, "Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations," in *International conference on machine learning*, PMLR, 2019, pp. 783–792.

[22] W. Chu and Z. Ghahramani, "Gaussian processes for ordinal regression," *Journal of Machine Learning Research*, vol. 6, pp. 1–48, 2005.

[23] K. Li, M. Tucker, E. Bıyık, E. Novoseller, J. W. Burdick, Y. Sui, D. Sadigh, Y. Yue, and A. D. Ames, "Roial: Region of interest active learning for characterizing exoskeleton gait preference landscapes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 3212–3218.

[24] E. Bıyık, D. P. Losey, M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh, "Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences," *The International Journal of Robotics Research*, vol. 41, no. 1, pp. 45–67, 2022.

[25] E. Biyik, N. Huynh, M. J. Kochenderfer, and D. Sadigh, "Active preference-based gaussian process regression for reward learning," in *Proceedings of Robotics: Science and Systems (RSS)*, Jul. 2020. DOI: 10.15607/rss.2020.xvi.041.

[26] E. Bıyık, N. Huynh, M. J. Kochenderfer, and D. Sadigh, "Active preference-based gaussian process regression for reward learning and optimization," *The International Journal of Robotics Research*, p. 02 783 649 231 208 729, 2023.

[27] A. Y. Ng, S. Russell, *et al.*, "Algorithms for inverse reinforcement learning.," in *Icml*, vol. 1, 2000, p. 2.

[28] N. Wilde, E. Biyik, D. Sadigh, and S. L. Smith, "Learning reward functions from scale feedback," in *Proceedings of the 5th Conference on Robot Learning (CoRL)*, 2021.

[29] D. Golovin and A. Krause, "Adaptive submodularity: Theory and applications in active learning and stochastic optimization," *Journal of Artificial Intelligence Research*, vol. 42, pp. 427–486, 2011.

[30] D. Golovin, A. Krause, and D. Ray, "Near-optimal bayesian active learning with noisy observations," *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[31] C. S. Gowtham Bellala, "Modified group generalized binary search with near-optimal performance guarantees," 2010.

[32] Z. Erickson, V. Gangaram, A. Kapusta, C. K. Liu, and C. C. Kemp, *Assistive gym: A physics simulation framework for assistive robotics*, 2019. arXiv: 1910.04700 [cs.RO].

[33] K. Ethayarajh, Y. Choi, and S. Swayamdipta, "Understanding dataset difficulty with $\mathcal{V}$-usable information," in *International Conference on Machine Learning*, PMLR, 2022, pp. 5988–6008.

[34] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[35] S. Casper, X. Davies, C. Shi, *et al.*, "Open problems and fundamental limitations of reinforcement learning from human feedback," *Transactions on Machine Learning Research (TMLR)*, 2023.

[36] F. Barbieri, J. Camacho-Collados, L. E. Anke, and L. Neves, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1644–1650.

[37] R. F. Flesch and A. J. Gould, *The art of readable writing*. 1949.

[38] E. Dale and J. S. Chall, "A formula for predicting readability: Instructions," *Educational research bulletin*, pp. 37–54, 1948.

[39] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring.," *Journal of Applied Psychology*, vol. 60, no. 2, p. 283, 1975.

[40] R. Senter and E. A. Smith, "Automated readability index," Technical report, DTIC document, Tech. Rep., 1967.

[41] B. Liu, H. Zamani, X. Lu, and J. S. Culpepper, "Generalizing discriminative retrieval models using generative tasks," in *Proceedings of the Web Conference 2021*, 2021, pp. 3745–3756.