Active Reward Learning from Multiple Teachers

Peter Barnett[1,*], Rachel Freedman[1], Justin Svegliato[1] and Stuart Russell[1]

[1]*Center for Human-Compatible AI, University of California, Berkeley, CA 94720, USA*

Abstract

Reward learning algorithms utilize human feedback to infer a reward function, which is then used to train an AI system. This human feedback is often a preference comparison, in which the human teacher compares several samples of AI behavior and chooses which they believe best accomplishes the objective. While reward learning typically assumes that all feedback comes from a single teacher, in practice these systems often query multiple teachers to gather sufficient training data. In this paper, we investigate this disparity, and find that algorithmic evaluation of these different sources of feedback facilitates more accurate and efficient reward learning. We formally analyze the value of information (VOI) when reward learning from teachers with varying levels of rationality, and define and evaluate an algorithm that utilizes this VOI to actively select teachers to query for feedback. Surprisingly, we find that it is often more informative to query comparatively irrational teachers. By formalizing this problem and deriving an analytical solution, we hope to facilitate improvement in reward learning approaches to aligning AI behavior with human values.

Keywords

Reward Learning, Active Learning, Preference Learning, Value of Information

1. Introduction

Standard AI and machine learning algorithms require the designer to specify a cost or reward function. This objective incentivizes desired behavior and penalizes mistakes, teaching the system how to perform the task. While such objectives are easy to manually specify for problems with clear win conditions, such as games [1, 2, 3] and tasks with clear goals, such as image classification [4, 5], they can be challenging to formalize for more nuanced tasks [6]. For example, Lee et al. [7] find that humans struggle to define an objective that incentivizes bipedal locomotion, despite being experts in both machine learning and walking. By incentivizing incorrect behavior, misspecified objectives can lead to useless or even dangerous outcomes [8]. Ensuring that AI systems optimize objectives that align with our own is a crucial part of building safe and beneficial AI.

Reward learning techniques enable AI systems to learn their objectives by observing and interacting with humans instead of requiring their designers to specify these objectives manually [9]. Humans can train reward learning systems using a variety of feedback modalities, including demonstrations [10, 11, 12], pairwise comparisons [7, 13, 14], natural language [15], numeric values [16], corrections [17], and proxy rewards [18, 19]. Reward learning from pairwise comparisons in particular has proven remarkably effective across a variety of tasks, including complex physical maneuvers for continuous control systems [7, 14] and text summarization

for language language models [20, 21]. In the future, it may even be possible to use reward learning to train AI systems to assist humans in researching safe AI [8, 22].

However, to infer reward functions from human feedback, reward learning systems must model human decision-making, and incorrect human decision-making models often leads to poor inference [23, 24, 25]. Moreover, reward learning systems typically assume that all feedback comes from a single distribution or teacher, despite querying multiple teachers to generate sufficient feedback. However, humans often vary in their expertise, focus, and intelligence, affecting the noisiness of their feedback. The practice of conflating all feedback implicitly disregards the differences between different teachers, increasing the likelihood of human model misspecification and the limitations of reward learning [26].

In this work, we extend reward learning to take advantage of differences between teachers. We develop a Bayesian reward learning algorithm that actively selects which teacher to query based on the noisiness of their feedback and the learner's current belief. We find that querying a *less* rational teacher can often be more informative than querying a *more* rational teacher, since teacher mistakes inform the agent of the relative values of alternatives. For example, imagine that two teachers are comparing two alternatives, A and B. A is worth more than B, but only slightly. If the first teacher is perfectly rational, they will always select A over B. The learner can infer from this that A is preferable to B, but has no way to learn how significant the distinction is. However, assume that the second teacher is somewhat less rational, and occasionally mixes up alternatives of similar value. Then they will typically choose A, but sometimes choose B, and this allows the learner to infer that the gap between A and B is small. Section 3 formalizes this

---

rationality model and inference procedure.

The rest of the paper is as follows. In Section 2, we discuss prior work on reward learning, active learning, and human modeling. In Section 3, we describe the mechanics of reward learning, including the model of human rationality and the metrics that will be used to measure the value of information (VOI) of teacher feedback. In Section 4, we propose a teacher selection algorithm that selects which teacher to query for feedback at each time step based on the modeled rationality of each teacher and the learner's belief distribution over the reward function. In Sections 5 and 6, we present theoretical and empirical results, showing that the learner's belief will eventually converge to the true reward function under the teacher selection algorithm, that querying less rational teachers can often be more informative, and that our teacher selection method outperforms simple heuristics like always querying the most rational teacher. By formalizing the problem of learning from multiple teachers and deriving an analytical solution, we hope to facilitate improvement in reward learning approaches to value alignment.

2. Related Work

Reward Learning   Reward learning techniques allow AI systems to learn reward functions by observing or interacting with humans. For example, *inverse reinforcement learning* agents observe human behavior or policies, and then infer an underlying reward function that the behavior optimizes [10, 11, 12]. Recent advances in reward learning have focused on learning from preference comparisons. Here, human teachers observe paired samples of system behavior, then choose which sample they prefer out of each pair. The system learns a reward model that maximizes the likelihood of these preferences, then uses that model to generate a reward signal to guide its behavior. This technique has been successfully applied to many domains, from continuous control [7, 14] to language generation tasks [20, 21]. Reward learning can also use a variety of other feedback modalities, including preference comparisons [7, 13, 14], natural language [15], numeric values [16], corrections [17], and proxy rewards [18, 19], but we focus on preference comparisons in this paper due to its recent success.

Active Reward Learning   Human feedback is expensive and time-consuming to generate, so reward learning algorithms must learn efficiently from limited data. They do this in part by actively selecting the queries that are sent to human teachers in order to maximize the expected VOI of human feedback. Sadigh et al. [13] assume that the system is a Bayesian learner, actively synthesizing queries that maximize the expected volume removed from the learner's posterior. Bıyık and Sadigh [27] de-
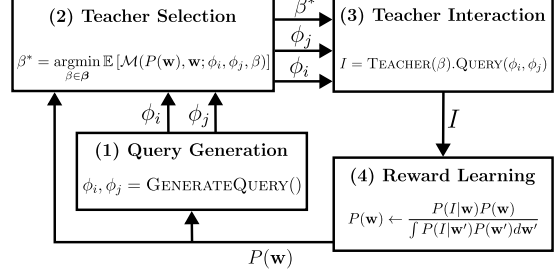

Figure 1: Our active reward learning approach.

velop efficient approximations to this method and show how to integrate active query selection and reward learning in practice. Lee et al. [7] take a different approach, empirically evaluating various heuristic strategies for query selection and finding that uncertainty-based sampling methods tend to perform the best. However, all of this previous work focuses on choosing which *queries* to send to the teachers. In this paper, we instead consider which *teachers* to send these queries to.

Human Modeling   To infer reward functions, AI systems must model the behavior of humans. Early work on reward learning assumed that human behavior was perfectly rational and that human teachers always chose the alternative that maximized their reward [10]. Later work models human behavior as pedagogic [24], systematically biased [28], and noisily or Boltzmann-rational [9, 12]. We will follow recent work on learning from human preferences [7, 9, 12, 14] and model human teachers as Boltzmann-rational, making choices according to a well-known probability model specified later in the paper.

3. Active Reward Learning

In this section, we formalize the problem of selecting the most informative teacher to query in order to gradually learn the correct reward model. In particular, we are interested in greedily selecting the teacher to query at each time step such that the reward model of the agent efficiently converges to the correct reward model.

At a high level, the teacher selection problem begins with a set of items or trajectories to compare, along with a set of human teachers to evaluate those comparisons. The human teachers each have a different level of rationality that is known *a priori*, meaning that the probability of a given human teacher making a mistake by preferring a less valuable item over a more valuable item is known in advance. During each time step of our approach depicted in Figure 1, two items are sampled from the set of items (*Step 1*) and then a human teacher is selected to be queried based on these items and the current belief about the

reward model (*Step 2*). The human teacher is asked which of the two items they prefer (*Step 3*), and their preference is used to update the reward model (*Step 4*). This process of selecting a query and a teacher is repeated until the reward model converges to the correct reward model.

Query selection is the problem of choosing which items to present to the teacher [7]. Some approaches to query selection include choosing the pair of items for which the preference predictors are most uncertain [7, 14]. Other approaches to query selection include selecting the pair of items that ensure that the space of queries is well covered. Finally, there are more active methods that actively synthesize queries in order learn more efficiently [13, 29]. Since our focus is on teacher selection rather than query selection, for the purposes of our analysis we will assume that queries are sampled uniformly at random. However, existing methods for query selection can be easily combined with our teacher selection algorithm to further improve reward learning.

To formalize the problem of teacher selection, this section proceeds as follows. We (1) provide a representation of items and rewards, (2) apply a well-known model of human rationality to our problem, (3) offer a method for updating belief distributions that uses preference comparisons from a human teacher, and (4) propose two metrics that measure the correctness of a belief distribution.

Representing Items and Rewards   Intuitively, each item can be represented as a set of features. For example, a book could be described by the number of pages and the number of positive reviews or a maneuver made by a self-driving car could be described by its position and distance from other vehicles at each time step. Hence, each item i can formally be represented by a feature vector $\phi_i \in \mathbb{R}^d$ where d is the number of features that describe the ith item.

Given this representation of an item, the reward $R(i)$ for an item i can be expressed as a dot product between the feature vector ϕ_i and the weight vector $\mathbf{w} \in \mathbb{R}^d$ for the reward model that is being learned:

$$R(i) = \mathbf{w}^\top \phi_i. \qquad (1)$$

If the items cannot be expressed by a feature vector, this approach can still be used by treating the feature vector ϕ_i as a one-hot vector: given the ith item, the ith entry of the feature vector ϕ_i would be 1 and every other entry would be 0 while the ith entry of the weight vector $\mathbf{w}$ would be the reward $R(i)$ for the ith item.

During reward learning, the human teacher is presented with two items and the probability of the human choosing one item over another item depends on the difference in reward between the two items at hand. We therefore express the difference in the reward between two items i and j as the equation

$$R(i) - R(j) = \mathbf{w}^\top (\phi_i - \phi_j) = \mathbf{w}^\top \varphi_{ij}, \qquad (2)$$

where $\varphi_{ij} = \phi_i - \phi_j$ is the difference in the feature vectors of the two items.

Modeling Human Rationality   Human teachers can be represented as Boltzmann-rational agents following a large body of existing work on reward learning [7, 9, 12, 14, 30, 31, 32, 33, 34]. Moreover, we assume that each teacher has a different known rationality parameter β rather than assuming $\beta = 1$ for all teachers. Boltzmann-rational teachers are more likely to choose the higher reward item if they are "more rational" (i.e., a higher β), or if the difference in reward between the two items is greater. The probability that the teacher chooses an item i over and an item j is given by

$$P(i \succ j; \beta) = \frac{\exp(\beta R(i))}{\exp(\beta R(i)) + \exp(\beta R(j))}. \qquad (3)$$

We thus model the human choice probabilistically:

$$P(I|\mathbf{w}; \varphi_{ij}, \beta) = \frac{1}{1 + \exp(-I\beta\mathbf{w}^\top \varphi_{ij})}, \qquad (4)$$

where $I = +1$ if the human prefers item i over item j and $I = -1$ if the human prefers item j over item i. This reflects the difference in value of the two items but not their absolute value. Equation 4 is a logistic model of the probability of the human preference I, where β determines the slope. As the difference in reward between the two items increases, the probability that the teacher chooses the higher reward item approaches 1.

Updating Belief Distributions   The goal of reward learning is to learn the weight vector $\mathbf{w}$ of the reward model. Given the preference of a teacher I, the difference in feature vectors φ_{ij}, and the teacher's rationality parameter β, the learner updates its belief over the weights of the reward model. That is, the belief over the weights of the reward model is updated such that the reward model now predicts that the item selected by the teacher is more valuable than it was prior to the belief update. Formally, we begin with the current belief distribution $P(\mathbf{w})$, which we treat as the prior distribution, and update it according to Bayes' theorem in the following way:

$$P(\mathbf{w}|I; \varphi_{ij}, \beta) = \frac{P(I|\mathbf{w}; \varphi_{ij}, \beta)P(\mathbf{w})}{\int P(I|\mathbf{w}'; \varphi_{ij}, \beta)P(\mathbf{w}')d\mathbf{w}'}, \qquad (5)$$

where $P(I|\mathbf{w}; \varphi_{ij}, \beta)$ is given by Equation 4.

Measuring Belief Distribution Error   After querying a teacher and updating the belief over the weights of the reward model $\mathbf{w}$, the belief distribution can be evaluated on a metric that measures the "correctness" or the distance of this belief distribution to the true belief distribution. Here, we consider two such metrics: the

Table 1
The general form of an expected metric $\mathcal{M}$ along with the expected metrics for mean squared error (MSE) and log loss (LL).

Expected Metric	Equation
$\mathbb{E}_{\substack{\mathbf{w}\sim P_{\mathbf{w}} \\ I\sim P_{I\mid\mathbf{w}}}}\left[\mathcal{M}(P_{\mathbf{w}\mid I},\mathbf{w};\varphi_{ij},\beta)\right]$	$\int P_{\mathbf{w}}\sum_I P_{I\mid\mathbf{w}}\mathcal{M}(P_{\mathbf{w}\mid I},\mathbf{w})d\mathbf{w}$
$\mathbb{E}_{\substack{\mathbf{w}\sim P_{\mathbf{w}} \\ I\sim P_{I\mid\mathbf{w}}}}\left[\mathrm{MSE}(P_{\mathbf{w}\mid I},\mathbf{w};\varphi_{ij},\beta)\right]$	$2\sum_I \frac{2}{\int f_I(\mathbf{w})d\mathbf{w}}\times\left[\int f_I(\mathbf{w})d\mathbf{w}\int f_I(\mathbf{w})\left\|\mathbf{w}\right\|^2 d\mathbf{w}-\left\|\int f_I(\mathbf{w})\mathbf{w}d\mathbf{w}\right\|^2\right]$
$\mathbb{E}_{\substack{\mathbf{w}\sim P_{\mathbf{w}} \\ I\sim P_{I\mid\mathbf{w}}}}\left[\mathrm{LL}(P_{\mathbf{w}\mid I},\mathbf{w};\varphi_{ij},\beta)\right]$	$-\sum_I \int f_I(\mathbf{w})\log\left(\frac{f_I(\mathbf{w})}{\int f_I(\mathbf{w}')d\mathbf{w}'}\right)d\mathbf{w}$

mean squared error (MSE) and the log loss (LL). The MSE measure represents how "far away" the belief distribution is from the true value while the LL measure represents the height of the belief distribution at the true value. In both cases, a lower score indicates a more accurate distribution. Using $Q(\mathbf{w})$ as the belief distribution over the weight vector $\mathbf{w}$ and $\mathbf{w}_{true}$ as the true weight vector, the MSE and LL measures are given as follows.

$$\mathrm{MSE}(Q(\mathbf{w}),\mathbf{w}_{true})=\int Q(\mathbf{w})\|\mathbf{w}-\mathbf{w}_{true}\|^2 d\mathbf{w} \quad (6)$$

$$\mathrm{LL}(Q(\mathbf{w}),\mathbf{w}_{true})=-\log(Q(\mathbf{w}_{true})) \quad (7)$$

Note that we will describe a greedy approach that selects the teacher that in expectation leads to our belief distribution scoring the best on one of these metrics after a single update in the next section.

Work on active learning from human preferences uses volume removal (i.e., removing as much of the integral of the unnormalized distribution as possible) as a metric [13, 27, 33]. However, this may not be an appropriate metric for teacher selection. This is because a larger Boltzmann rationality parameter β results in a larger volume of the belief distribution being removed but may not necessarily lead to a more accurate belief distribution.

4. Teacher Selection

We propose a method for selecting and querying the teacher that produces the best immediate improvement in the expectation of a given metric, which approximates the expected VOI of the teacher feedback. The metrics evaluate how similar the posterior belief is to the ground truth reward, so lower scores indicate improvements in the learned reward model. The algorithm considers uncertainty over two variables: the ground-truth parameterization of the reward model and the item from the query that the teacher prefers. In particular, the expectation of the metric must be taken over the current belief distribution $P(\mathbf{w})$ and the probability $P(I\mid\mathbf{w};\varphi_{ij},\beta)$ of the teacher preferring each item. Formally, we express the expectation of a given metric $\mathcal{M}$ in Table 1. Note that we

use the notation $P_{\mathbf{w}}=P(\mathbf{w})$, $P_{I\mid\mathbf{w}}=P(I\mid\mathbf{w};\varphi_{ij},\beta)$, and $P_{\mathbf{w}\mid I}=P(\mathbf{w}\mid I,\varphi_{ij},\beta)$ throughout this section.

Importantly, the expected value of a given metric only depends on the known variables φ_{ij} and β along with the current belief distribution $P_{\mathbf{w}}$ given a straightforward substitution of Equations 4 and 5. This enables our method to calculate the expected value of the metric for a given teacher with the rationality parameter β. This will be used to find the teacher to query at each time step: the teacher with the lowest metric in expectation should be selected as that would result in a weight vector that is closest to the true weight vector in expectation.

Finally, given the general form of an expected metric, Table 1 defines the expectations of the MSE and LL metrics using the function $f_I(\mathbf{w})=P_{\mathbf{w}}/(1+\exp(-I\beta\mathbf{w}^\top\varphi_{ij}))$.

Selecting a Teacher   To select the teacher to query, we first calculate the expected metric for each teacher β given the current belief distribution $P(\mathbf{w})$ and then select the teacher that would result in the lowest expected metric score. Formally, the rationality parameter β^* that leads to the largest reduction in the expectation of the metric is defined as follows:

$$\beta^*=\underset{\beta\in\boldsymbol{\beta}}{\mathrm{argmin}}\left[\mathbb{E}_{\substack{\mathbf{w}\sim P_{\mathbf{w}} \\ I\sim P_{I\mid\mathbf{w}}}}\left[\mathcal{M}(P_{\mathbf{w}\mid I},\mathbf{w};\varphi_{ij},\beta)\right]\right], \quad (8)$$

where $\boldsymbol{\beta}$ is a vector of the β values of the teachers.

Learning a Reward Model   To learn the reward model, the learner begins with an initial belief distribution $P_{\mathbf{w}}$ over the reward function parameterization and then updates it according to Algorithm 1. First, the algorithm generates queries of paired items and calculates β^*, which is the rationality parameter that leads to the largest improvement in the expectation over the correctness metric. The algorithm queries the teacher with this rationality parameter, and the teacher responds with a *preference* indicating which of the two items in the query they prefer. This preference is used to update the

---
Algorithm 1: LEARNREWARDMODEL($\cdot$)

Input: An initial belief distribution $P(\mathbf{w})$, a list of the teachers' Boltzmann rationality parameters $\boldsymbol{\beta}$, an expected metric function $\mathbb{E}[\mathcal{M}]$, and an entropy convergence threshold ϵ

Output: A posterior belief distribution $P(\mathbf{w})$
---
1 *converged* $\leftarrow$ *False*
2 **while not** *converged* **do**
3 $\quad$ ϕ_i, ϕ_j $\leftarrow$ GENERATEQUERY()
4 $\quad$ φ_{ij} $\leftarrow$ $\phi_i - \phi_j$
5 $\quad$ $\beta^* \leftarrow \operatorname{argmin}_{\beta \in \boldsymbol{\beta}} \mathbb{E}\left[\mathcal{M}(P(\mathbf{w}), \mathbf{w}; \varphi_{ij}, \beta)\right]$
6 $\quad$ $I \leftarrow$ TEACHER(β^*).QUERY(ϕ_i, ϕ_j)
7 $\quad$ $P(\mathbf{w}) \leftarrow$ NORMALIZE($P(\mathbf{w}) \cdot P(I|\mathbf{w}, \varphi_{ij}, \beta^*)$)
8 $\quad$ *entropy* $\leftarrow -\int P(\mathbf{w}) \log P(\mathbf{w}) d\mathbf{w}$
9 $\quad$ *converged* $\leftarrow$ *entropy* $< \epsilon$
10 **return** $P(\mathbf{w})$
---

belief distribution $P_\mathbf{w}$. The algorithm iterates until convergence, which is when the entropy of the distribution $P_\mathbf{w}$ becomes lower than a specified threshold ϵ.

5. Theoretical Analysis

In this section, we first prove that the belief distribution will converge to the true distribution and then show that, under certain conditions, querying a less rational teacher can result in more informative feedback.

Convergence Algorithm 1 queries multiple teachers with different β values until the reward estimate convergences. Here, we show that this process will make the belief distribution over $\mathbf{w}$ converge to the true value.

Theorem 1. *In the limit of $N \to \infty$ random queries to Boltzmann-rational teachers with positive, finite β values, the posterior distribution over $\mathbf{w}$ converges to the true value.*

Proof. The likelihood of a sequence of human choices $\underline{I} \in [\pm 1]^N$ from humans with rationality parameters $\underline{\beta}$ is $P(\underline{I}|\mathbf{w}; \underline{\beta}) = \prod_{i=1}^N P(\underline{I}_i|\mathbf{w}; \underline{\beta}_i)$. The posterior distribution over $\mathbf{w}$ after a sequence of queries is

$$P(\mathbf{w}|\underline{I}; \underline{\beta}) \propto \prod_i^N P(\underline{I}_i|\mathbf{w}; \underline{\beta}_i)P(\mathbf{w}).$$

We will show that $P(\mathbf{w}|\underline{I}; \underline{\beta}) \to 0$ as $N \to \infty$ for all $\mathbf{w} \neq \mathbf{w}_{true}$. The Bayes factor between $\mathbf{w}$ and $\mathbf{w}_{true}$ is

$$\text{BF} = \frac{P(\mathbf{w}|\underline{I}; \underline{\beta})}{P(\mathbf{w}_{true}|\underline{I}; \underline{\beta})} = \frac{\prod_i^N P(\underline{I}_i|\mathbf{w}; \underline{\beta}_i)P(\mathbf{w})}{\prod_i^N P(\underline{I}_i|\mathbf{w}_{true}; \underline{\beta}_i)P(\mathbf{w}_{true})},$$

where $P(\mathbf{w}_{true}|\underline{I}; \underline{\beta})$ is the posterior distribution at $\mathbf{w}_{true}$. We can show that $\text{BF} \to 0$ as $N \to \infty$ except when

$\mathbf{w} = \mathbf{w}_{true}$. This implies $P(\mathbf{w}|\underline{I}; \underline{\beta}) \to 0$ except when $\mathbf{w} = \mathbf{w}_{true}$. We require $P(\mathbf{w}_{true}) \neq 0$ as BF is undefined otherwise. Trivially, $\text{BF} = 1$ when $\mathbf{w} = \mathbf{w}_{true}$.

We now consider $\mathbf{w} \neq \mathbf{w}_{true}$. We can define the negative logarithm of BF, which approaches ∞ as $\text{BF} \to 0$:

$$-\log(\text{BF})$$
$$= -\log\left(\frac{\prod_i^N P(\underline{I}_i|\mathbf{w}; \underline{\beta}_i)P(\mathbf{w})}{\prod_i^N P(\underline{I}_i|\mathbf{w}_{true}; \underline{\beta}_i)P(\mathbf{w}_{true})}\right)$$
$$= -\sum_i^N \log\left(\frac{P(\underline{I}_i|\mathbf{w}; \underline{\beta}_i)}{P(\underline{I}_i|\mathbf{w}_{true}; \underline{\beta}_i)}\right) - \log\left(\frac{P(\mathbf{w})}{P(\mathbf{w}_{true})}\right).$$

The first term is the sum of many terms. If this term approaches ∞ as $N \to \infty$ then $\text{BF} \to 0$. We now examine each term in the sum and show that in expectation they are each positive. All of these terms are independent as they are only depend on the likelihood and not on the current distribution. Hence, they will not decay with additional steps, and so the sum will diverge if the individual terms are positive in expectation. The expected value for each term in the sum is

$$\mathbb{E}\left[-\log\left(\frac{P(\underline{I}_i|\mathbf{w}; \underline{\beta}_i)}{P(\underline{I}_i|\mathbf{w}_{true}; \underline{\beta}_i)}\right)\right]$$
$$= -\sum_{\underline{I}_i \in +1, -1} P(\underline{I}_i|\mathbf{w}_{true}; \underline{\beta}_i) \log\left(\frac{P(\underline{I}_i|\mathbf{w}; \underline{\beta}_i)}{P(\underline{I}_i|\mathbf{w}_{true}; \underline{\beta}_i)}\right).$$

This is the KL divergence between $P(\underline{I}_i|\mathbf{w}_{true}; \underline{\beta}_i)$ and $P(\underline{I}_i|\mathbf{w}; \underline{\beta}_i)$. This is strictly non-negative and only equal to zero when $P(\underline{I}_i|\mathbf{w}; \underline{\beta}_i) = P(\underline{I}_i|\mathbf{w}_{true}; \underline{\beta}_i)$. When $\beta = 0$, each of these terms equals 0. As $\beta \to \infty$, $P(\underline{I}_i|\mathbf{w}; \underline{\beta}_i) \to H(I\mathbf{w}^\top \varphi)$, where $H(\cdot)$ is the Heaviside step function. In this case, it holds that $P(\underline{I}_i|\mathbf{w}; \underline{\beta}_i) = P(\underline{I}_i|\mathbf{w}_{true}; \underline{\beta}_i)$ whenever the values $\mathbf{w}^\top \varphi$ and $\mathbf{w}_{true}^\top \varphi$ have the same sign.

Therefore, for positive, finite β each of the terms in the sum is positive, so the sum diverges, and so the $P(\mathbf{w}|\underline{I}; \underline{\beta}) \to 0$ for all $\mathbf{w} \neq \mathbf{w}_{true}$. $\qquad \square$

Bigger β isn't always more informative Querying a more rational teacher (with a larger β value) does not always lead to faster convergence to the true value, as measured by lower MSE or LL, because the magnitude of $\mathbf{w}^\top \varphi_{ij}$ can be learned from the teacher making mistakes.

We empirically observe this in Figure 2, where we demonstrate that if our current belief distribution $P(\mathbf{w})$ is a normal distribution characterized by μ and σ, a lower β value is more informative for certain values of μ and σ. Specifically, when the distribution is symmetric ($\mu = 0$) then a larger value of β is better, and as the distribution gets broader (larger σ) larger β is also better. If the distribution is very wide then a large β allows us to quickly
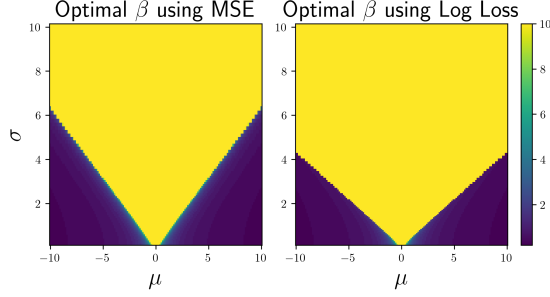
Figure 2: For some prior beliefs over **w**, querying a teacher with a lower β parameter is more informative. The plots show the most informative β value (according to the mean squared error and log loss metrics, respectively) for a range of beliefs. Each belief is a Gaussian, parameterized by μ (horizontal axis) and σ (vertical axis). The purple regions of the plots indicate beliefs where it is most informative to query a teacher with a β of approximately 1.

remove a lot of probability mass, while if the distribution is narrow (and asymmetric) then we learn about the value of $\mathbf{w}^\top \varphi_{ij}$ from the humans making mistakes, which requires the human to be less than perfectly rational. For example, if $\mathbf{w}^\top \varphi_{ij} > 0$ then a perfectly rational human would always choose item i over item j, and we would not learn about the actual *value* of $\mathbf{w}^\top \varphi_{ij}$.

6. Restaurant Recommendation

We now discuss how our method for reward learning using feedback from multiple teachers can be applied to a simplified restaurant recommendation domain. In this domain, the goal is to learn a reward function that can be used to recommend restaurants to a user. This reward model must be learned from feedback from multiple teachers, in this case by asking which of two restaurants a human prefers. It is important to highlight that our approach is compatible with a variety of popular recommendation tasks, including entertainment [35, 36], news [37], and shopping [38] recommendations.

More formally, the problem of restaurant recommendation has a set of restaurants $\rho = \{\rho_1, \rho_2, \ldots, \rho_n\}$ that can be recommended to a user. Moreover, there is a set of users $U = \{U_1, U_2, \ldots, U_m\}$ who can be queried about their restaurant preferences. Each restaurant is expressed as a set of features $F = \{\text{Cleanliness}, \text{Vegan}, \text{Spiciness}\}$ where $\text{Cleanliness} \in [1, 10]$ describes the cleanliness of the restaurant, $\text{Vegan} \in \{0, 1\}$ describes whether the restaurant is vegan-friendly, and $\text{Spiciness} \in [1, 10]$ describes the spiciness of the food. The preference rating for each restaurant is denoted by $\mathbf{w}^\top \rho_i$, where $\mathbf{w} \in \mathbb{R}^3$ is a weight vector that parameterizes the reward model. The

aim is to learn the weights **w** using feedback from multiple users to provide useful restaurant recommendations.

We can represent the restaurant recommendation domain using our approach. The set of items $\phi_1, \phi_2, \ldots, \phi_n$ is the set of restaurants ρ. The set of human users U is the set of human teachers. The users are modelled as Boltzmann-rational, and have known rationality parameters $\beta_1, \beta_2, \ldots, \beta_m$. Beginning with an initial distribution $P(\mathbf{w})$, we will use Algorithm 1 to converge to the weight values for the reward function that represents the user preferences. First, we select a pair of restaurants for a user to compare (in this case randomly selected) and apply Equation 8 describing which user should be queried in order to achieve the lowest metric score in expectation after a single update. Next, this user is selected and then asked which of the two restaurants they prefer. Finally, using the selected user's preference, the reward model weights are updated according to Equation 5 to generate a new belief distribution. The process is repeated until the belief distribution converges.

7. Experiments

We now show that our approach method for selecting β outperforms several baseline methods, using the simple restaurant recommendation domain. In Figure 3, we compare: (1) selecting the largest β value to see if the result that larger β is not always better is true in practice; (2) selecting β randomly to ensure that the advantage over selecting the largest β is not just due to the randomness of the selection; and (3) always selecting $\beta = 1$ because this is often what is assumed to be the rationality parameter in other work.

In this experiment, the size of the weight vector is $d = 3$ and the domain of the weights is $W = [-10, 10]^3$, which is discretized. The prior distribution of the weights is a uniform distribution over this domain $P(\mathbf{w}) = \mathcal{U}(W)$ and the true weight $\mathbf{w}_{true} \in W$ is sampled from this prior. There are 21 teachers, with β values uniformly spaced between 0 and 4. For 100 steps, two restaurant feature vectors $\phi = \{\text{Cleanliness}, \text{Vegan}, \text{Spiciness}\}$ are generated randomly, where $\text{Cleanliness}, \text{Spiciness} \sim \mathcal{U}(1, 10)$, and Vegan are uniformly drawn from $\{0, 1\}$. While we generate our samples randomly in order to isolate the the effect of teacher selection, any of the active query selection methods from previous work could be used here. The teacher is selected and then queried using one of the various methods and the belief distribution is updated based on the preference of that teacher. The same ϕ vectors are used for each method, so that the only difference between the methods is the selection of β. This procedure is repeated 100 times, each time sampling a new true weight vector $\mathbf{w}_{true}$.

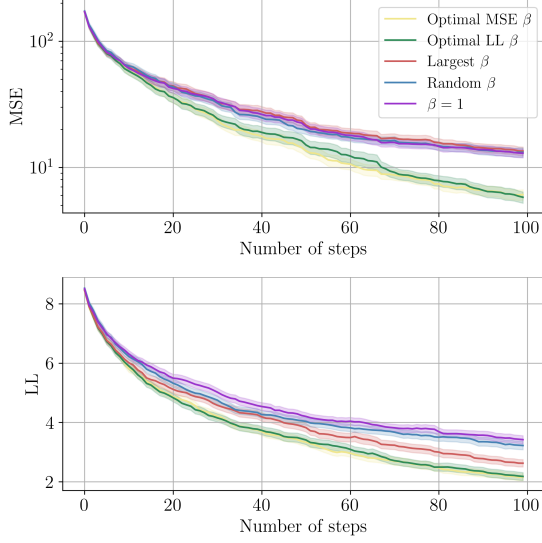Overall, we observe that the active teacher selection

Figure 3: Active teacher selection improves reward inference. These plots show the expected mean squared error and expected log loss over the course of 100 iterations of reward inference using various teacher selection methods. The solid line is the mean, and the shading is the standard deviation. Selecting teacher β w.r.t. mean square error most effectively minimizes mean square error, while selecting β w.r.t. log loss most effectively minimizes log loss. In both cases, selecting teachers according to Equation 8 clearly outperforms the heuristic of always selecting the most rational teacher (largest β) and the baselines (random β and $\beta = 1$).

Figure 4: This plot shows the most informative values of β during training, averaged across 100 runs (given the expected mean squared error and expected log loss respectively). The solid line is the mean and the shaded area is the standard deviation. β decreases over the course of training, as the learner's belief distribution over $\mathbf{w}$ becomes more confident.

methods (MSE and LL) outperform the baseline methods. Moreover, we examine how the most informative value of β changes with additional queries in Figure 4. As expected, the optimal β value decreases with additional queries, as the distribution gets less broad. At beginning of training, our approach queries the teachers with large β values because this enables it to determine the sign of $\mathbf{w}^\top \varphi_{ij}$, and then our approach queries the teachers with smaller β values to determine the magnitude of $\mathbf{w}^\top \varphi_{ij}$ as it gets more information.

8. Limitations and Future Work

For the sake of conceptual clarity and mathematical formalism, we have used relatively simple human decision-making and reward models. Future work should extend these results by increasing model complexity.

For example, this analysis assumes that humans are Boltzmann-rational decision-makers with constant, known β values. While more nuanced than optimal models, Boltzmann-rational models fail to account for systematic biases in human judgement [28, 39, 40]. This work could be improved by using more complex, realistic mod-
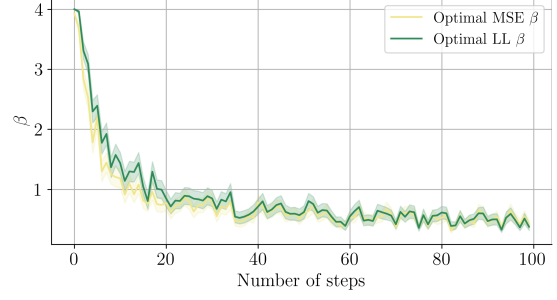
els of human decision-making, for example by allowing each human's β parameter to vary across the state space to capture teacher specialization or by measuring and explicitly modeling systematic cognitive biases. Moreover, this analysis assumes that the teacher β parameters are given, whereas in reality the agent may not have access to this information. Future work should also examine ways of modeling this part of human decision-making alongside learning the reward function.

Finally, future work could extend these results to non-linear reward models, such as ensembles of neural networks. Moreover, it could explore convergence properties and optimal querying strategies for learning from teachers with different reward functions. For example, variations in individual taste might lead teachers to disagree on which restaurants are best. Future work should explore the ramifications of such inter-teacher variance on teacher selection and reward learning.

9. Conclusion

In this work, we motivated, specified, and evaluated an algorithm for selecting which teacher to query during active reward learning with multiple teachers. Our algorithm models the teachers as Boltzmann-rational with known β parameters. At each time step, it queries the teacher that will be most informative in expectation. Interestingly, we find that the most informative teacher is not always the most rational one. We prove and demonstrate that the reward learner's belief will eventually collapse to the true reward function under our algorithm. Our hope is that this method and analysis will improve reward learning in domains where feedback is gathered from multiple teachers with varying levels of rationality.

Acknowledgments

References

[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of Go with deep neural networks and tree search, Nature 529 (2016) 484–489.

[2] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of Go without human knowledge, Nature 550 (2017) 354–359.

[3] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, et al., Dota 2 with large scale deep reinforcement learning, arXiv preprint arXiv:1912.06680 (2019).

[4] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, Communications of the ACM 60 (2017) 84–90.

[5] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.

[6] V. Krakovna, Specification gaming examples in AI, 2018.

[7] K. Lee, L. M. Smith, P. Abbeel, PEBBLE: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training, in: 38th International Conference on Machine Learning, PMLR, 2021, pp. 6152–6163.

[8] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, S. Legg, Scalable agent alignment via reward modeling: A research direction, arXiv preprint arXiv:1811.07871 (2018).

[9] H. J. Jeon, S. Milli, A. D. Dragan, Reward-rational (implicit) choice: A unifying formalism for reward learning, arXiv preprint arXiv:2002.04833 (2020).

[10] A. Y. Ng, S. J. Russell, Algorithms for inverse reinforcement learning, in: International Conference on Machine Learning, 2000, pp. 663–670.

[11] P. Abbeel, A. Y. Ng, Apprenticeship learning via inverse reinforcement learning, in: 21st International Conference on Machine Learning, 2004, p. 1.

[12] B. D. Ziebart, Modeling purposeful adaptive behavior with the principle of maximum causal entropy, Ph.D. thesis, Carnegie Mellon University, 2010.

[13] D. Sadigh, A. Dragan, S. Sastry, S. Seshia, Active preference-based learning of reward functions, in: Robotics: Science and Systems XIII, 2017, pp. 53–63.

[14] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, Neural Information Processing Systems (2017) 4300–4308.

[15] P. Goyal, S. Niekum, R. J. Mooney, Using natural language for reward shaping in reinforcement learning, arXiv preprint arXiv:1903.02020 (2019).

[16] D. Arumugam, J. K. Lee, S. Saskin, M. L. Littman, Deep reinforcement learning from policy-dependent human feedback, arXiv preprint arXiv:1902.04257 (2019).

[17] A. Bajcsy, D. P. Losey, M. K. O'Malley, A. D. Dragan, Learning robot objectives from physical human interaction, Machine Learning Research 78 (2017) 217–226.

[18] D. Hadfield-Menell, S. Milli, P. Abbeel, S. J. Russell, A. Dragan, Inverse reward design, in: Neural Information Processing Systems, 2017, pp. 6765–6774.

[19] S. Mindermann, R. Shah, A. Gleave, D. Hadfield-Menell, Active inverse reward design, arXiv preprint arXiv:1809.03060 (2018).

[20] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, P. F. Christiano, Learning to summarize with human feedback, Neural Information Processing Systems 33 (2020) 3008–3021.

[21] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, G. Irving, Fine-tuning language models from human preferences, arXiv preprint arXiv:1909.08593 (2019).

[22] J. Leike, J. Schulman, J. Wu, Our approach to alignment research, 2022. URL: https://openai.com/blog/our-approach-to-alignment-research/.

[23] J. Skalse, A. Abate, Misspecification in inverse reinforcement learning, arXiv preprint arXiv:2212.03201 (2022).

[24] S. Milli, A. D. Dragan, Literal or pedagogic human? Analyzing human model misspecification in objective learning, in: Uncertainty in Artificial Intelligence, 2020, pp. 925–934.

[25] R. Freedman, R. Shah, A. Dragan, Choice set misspecification in reward inference, arXiv preprint arXiv:2101.07691 (2021).

[26] O. Daniels-Koch, R. Freedman, The expertise problem: Learning from specialized feedback, arXiv preprint arXiv:2211.06519 (2022).

[27] E. Bıyık, D. Sadigh, Batch active preference-based learning of reward functions, arXiv preprint arXiv:1810.04303 (2018).

[28] O. Evans, A. Stuhlmüller, N. D. Goodman, Learning

the preferences of ignorant, inconsistent agents, in: 30th AAAI Conference on Artificial Intelligence, 2016, pp. 323–329.

[29] E. Bıyık, M. Palan, N. C. Landolfi, D. P. Losey, D. Sadigh, Asking easy questions: A user-friendly approach to active reward learning, in: Conference on Robot Learning, 2020, pp. 1177–1190.

[30] R. A. Bradley, M. E. Terry, Rank analysis of incomplete block designs: I. The method of paired comparisons, Biometrika 39 (1952) 324–345.

[31] X. Liang, K. Shu, K. Lee, P. Abbeel, Reward uncertainty for exploration in preference-based reinforcement learning, arXiv preprint arXiv:2205.12401 (2022).

[32] D. Ramachandran, E. Amir, Bayesian Inverse Reinforcement Learning., in: International Joint Conference on Artificial Intelligence, volume 7, 2007, pp. 2586–2591.

[33] M. Palan, G. Shevchuk, N. Charles Landolfi, D. Sadigh, Learning reward functions by integrating human demonstrations and preferences, in: Robotics: Science and Systems XV, 2019, pp. 23–33.

[34] R. Freedman, J. S. Borg, W. Sinnott-Armstrong, J. P. Dickerson, V. Conitzer, Adapting a kidney exchange algorithm to align with human values, Artificial Intelligence 283 (2020) 103261.

[35] C. A. Gomez-Uribe, N. Hunt, The netflix recommender system: Algorithms, business value, and innovation, ACM Transactions on Management Information Systems (TMIS) 6 (2015) 1–19.

[36] M. Perano, G. L. Casali, Y. Liu, T. Abbate, Professional reviews as service: A mix method approach to assess the value of recommender systems in the entertainment industry, Technological Forecasting and Social Change 169 (2021) 120800.

[37] S. Raza, C. Ding, News recommender system: A review of recent progress, challenges, and opportunities, Artificial Intelligence Review (2021) 1–52.

[38] P. M. Alamdari, N. J. Navimipour, M. Hosseinzadeh, A. A. Safaei, A. Darwesh, A systematic study on the recommender systems in the E-commerce, IEEE Access 8 (2020) 115694–115716.

[39] R. Shah, N. Gundotra, P. Abbeel, A. Dragan, On the feasibility of learning, rather than assuming, human biases for reward inference, in: 36th International Conference on Machine Learning, PMLR, 2019, pp. 5670–5679.

[40] L. Chan, A. Critch, A. Dragan, Human irrationality: Both bad and good for reward inference, arXiv preprint arXiv:2111.06956 (2021).