Cooperative Inverse Decision Theory for Uncertain Preferences

Zachary Robertson
Stanford

Hantao Zhang
University of Illinois Urbana Champaign

Oluwasanmi Koyejo
Stanford

Abstract

Inverse decision theory (IDT) aims to learn a performance metric for classification by eliciting expert classifications on examples. However, elicitation in practical settings may require many classifications of potentially ambiguous examples. To improve the efficiency of elicitation, we propose the cooperative inverse decision theory (CIDT) framework as a formalization of the performance metric elicitation problem. In cooperative inverse decision theory, the expert and a machine play a game where both are rewarded according to the expert's performance metric, but the machine does not initially know what this function is. We show that optimal policies in this framework produce active learning that leads to an exponential improvement in sample complexity over previous work. One of our key findings is that a broad class of sub-optimal experts can be represented as having uncertain preferences. We use this finding to show such experts naturally fit into our proposed framework extending inverse decision theory to efficiently deal with decision data that is sub-optimal due to noise, conflicting experts, or systematic error.

1 INTRODUCTION

Computer-assisted detection (CAD) systems are increasingly used as a source for efficient screening, especially in the medical domain [Lindsey et al., 2018, Jin et al., 2020, Calisto et al., 2021]. However, it is still not clear how this influences decision-making or if there is a best approach to maximize the benefits of Human-AI collaboration [Calisto et al., 2021]. Indeed, while AI approaches have been able to make use of large amounts of data to learn classification models they frequently lack mechanisms to adapt to particular human preferences. Inverse decision theory (IDT) and performance metric elicitation (PME) are promising approaches to this problem [Davies, 2005, Hiranandani et al., 2019b]. By incorporating human preferences through a small amount of feedback, these approaches aim to calibrate a classification model to better reflect human values and judgment, reducing the problem to a fine-tuning or few-shot problem. IDT in particular has proven useful in real-world settings since it addresses the description-experience gap, the observation that it is relatively difficult for humans to reason abstractly about preferences versus directly decide classifications for examples [Hertwig and Erev, 2009, Swartz et al., 2006].

In practice, these methods can require a large number of samples and have limited ability to deal with conflicting preferences from multiple experts or non-realizable preference models [Paolacci et al., 2010]. In this work, we propose a novel framework that can reduce human sample complexity in these settings. The central contributions are as follows.

1. **Cooperative Inverse Decision Theory.** We propose the cooperative inverse decision theory (CIDT) framework as a formalization of the performance metric elicitation problem. In CIDT the expert and a machine play a game where both are rewarded according to a surrogate of the expert's performance metric, but the machine does not initially know what this function is. In our analysis, we show optimal joint policies in the CIDT framework have exponential improvement in sample complexity over IDT. See Figure 1 for an overview of our framework.
2. **Uncertain Preferences.** We extend our framework to handle experts which may be represented as having uncertain preferences or, equivalently, a random decision threshold. Our study suggests that this may create a situation where it is optimal to teach differently than classify, creating a "description-experience gap."
3. **Practical Implementation.** We address uncertain preference elicitation with a practical implementation that is robust to demonstrations that are sub-optimal due to noise, conflicting experts, or the presence of missing/additional information. In our theoretical analysis and experiments, we demonstrate advantages over passive approaches.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

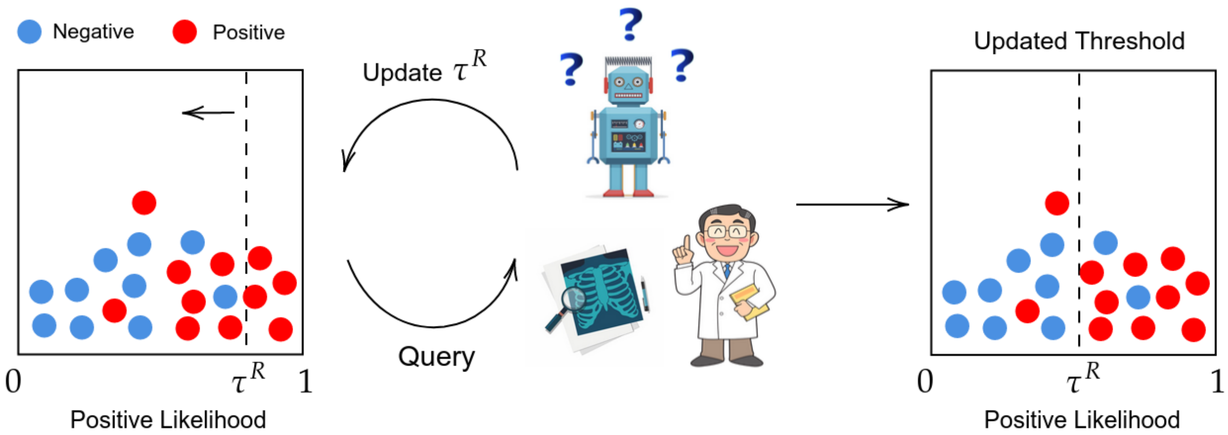


Figure 1: An illustration of the cooperative inverse decision theory framework. A doctor (H) and computer (R) cooperate to elicit the human's decision rule. The computer can not directly observe the decision rule used by the doctor. Left, there is a dataset of examples ordered by positive likelihood. The computer has a prior τ^R for the doctor's decision rule. Middle, the computer selects examples to show to the doctor and observes their decisions. The computer updates their belief about the doctor's decision rule. Right, the computer has an updated estimate for the decision rule that better reflects the doctor's preferences for false-positive and false-negative errors.

2 RELATED WORK

Metric selection is often critical to the practicality of machine learning as optimizing the wrong metric can lead to inappropriate models and critical errors [Dmitriev and Wu, 2016, Lindsey et al., 2018]. In the binary setting, one could select a suitable operating point on the ROC curve [Hajian-Tilaki, 2013]. However, humans are poor at providing such absolute preferences – which motivates the use of alternative elicitation strategies to compare metrics and classifiers [Qian et al., 2013]. Inverse decision theory uses decisions on a few classification examples to produce estimates of the underlying performance metric [Davies, 2005, Laidlaw and Russell, 2021]. However, this method doesn't make use of the interactivity available in the computer-assisted detection setting and fails to account for noise, diverse opinions, or missing/additional context.

Our work extends inverse decision theory to be useful in more cooperative settings, such as computer-assisted detection, by introducing the cooperative inverse decision theory framework (CIDT). Our framework can be considered as a special case of cooperative inverse reinforcement learning (CIRL) [Hadfield-Menell et al., 2016]. In CIRL, the demonstrator interacts cooperatively in a two-player game of partial information with the imitator to communicate the reward function. Such work is closely related to optimal teaching and active learning, demonstrations that optimally train an imitator [Balbach and Zeugmann, 2009, Settles, 2009]. However, our main purpose is not to present improvements to CIRL algorithms. Instead, we present an analysis of the difference between interactive and passive settings for the special case of classifier preference learning.

Another approach to the metric selection is performance metric elicitation, which uses pairwise comparisons, as represented by their confusion matrices, to estimate the underlying performance metric generating the comparisons [Hiranandani et al., 2019a]. There have been various extensions of the basic framework along with application in real-world settings [Hiranandani et al., 2019b, Robertson, 2022, Ali et al., 2022]. However, this method suffers from the description-experience gap, that is, it may be relatively difficult for humans to compare confusion matrices versus decide classifications for examples [Hertwig and Erev, 2009]. In general, inferring preferences from proxy tasks may not generalize well to tasks of interest [Buçinca et al., 2020].

In our analysis, we make extensive use of the fact that the experts and preferences are naturally ordered along a one-dimensional axis of decision thresholds. In particular, this suggests we are looking for a median preference [Black, 1948, Rowley, 1984]. While previous work has approached the problem of online median estimation using techniques such as stochastic approximation [Hanson and Russo, 1981, Pap, 2010, Meister and Nietert, 2021], our work extends this line of research by using the probabilistic bisection algorithm (PBA) to implement the CIDT algorithm. This approach offers advantages such as linear convergence in low-noise settings and ease of setup in parallel environments [Horstein, 1963]. The situation is more complicated for stochastic approximation [Pallone et al., 2014, Kushner and Yin, 1987b, Kushner and Yin, 1987a].

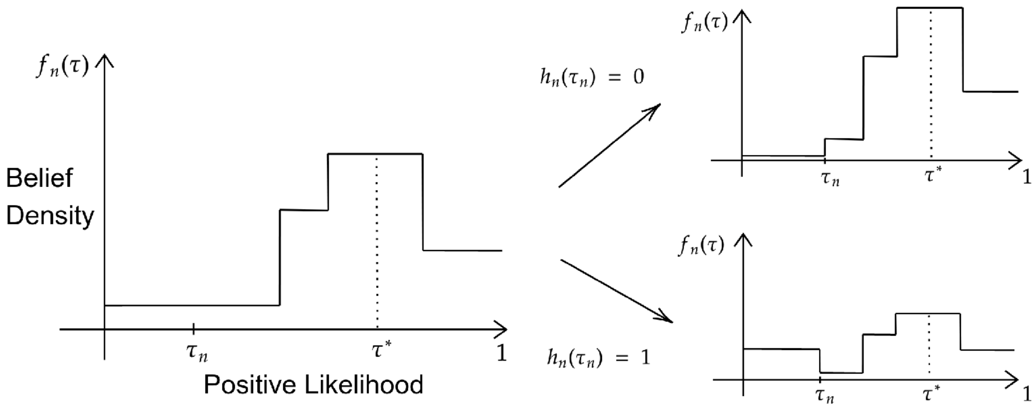


Figure 2: How we update the belief distribution in our cooperative inverse decision theory algorithm implementation. The belief density f_n for the optimal threshold is a piece-wise constant function. On the right, we observe a classification $h_n(\tau_n)$ from the decision maker of an example with positive likelihood τ_n. We learn about the relative location of the optimal decision threshold τ^* and the belief distribution for τ^* is updated.

3 LEARNING PREFERENCES FROM HUMAN DECISIONS

In this section, we'll introduce the inverse decision theory framework. Following this, we introduce our extension to the cooperative setting. In order to simplify our analysis, we consider the case of binary decisions in this work. However, our results are also applicable to decision problems with a larger number of choices. This is because, under the assumption of irrelevance from independent alternatives (i.e., the independence axiom [Luce, 1977]), a decision among many choices can be reduced to a series of binary choices between pairs of alternatives [Laidlaw and Russell, 2021]. To introduce the problem we'll give some notation for the basic variables of interest in classification. In a binary decision problem, an agent receives an observation $x \in \mathcal{X}$ and returns a decision $y \in \mathcal{Y} = \{0, 1\}$. Agents choose decision rules $h : \mathcal{X} \to \mathcal{Y}$ from a hypothesis class $\mathcal{H}$ such that,

$$h \in \text{argmin}_{h \in \mathcal{H}} \mathbb{E}_{(X,Y) \sim \mathcal{D}}[l(h(X), Y)] \tag{1}$$

with respect to some distribution $\mathcal{D}$ of examples paired with labels and loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. We can tabulate errors over a population of examples by using a confusion matrix. A confusion matrix C is a two-by-two array where the on-diagonal elements give the correct classification rates and the off-diagonal elements are the error rates. We can define $C_{\mathcal{D}} : \mathcal{H} \to [0, 1]^{2 \times 2}$ as the confusion matrix generated from using a particular classifier $h \in \mathcal{H}$ on the distribution $\mathcal{D}$. A performance metric is a function $\Psi : \mathcal{C} \to \mathbb{R}$ from the space of confusion matrices to an evaluation score. A simple example are the linear performance metrics. These are the linear functionals $\Psi(C) = \langle \phi, C \rangle$ defined on $\mathcal{C}$. For example, accuracy could be measured by $\phi = \text{Id}$. It is known that the optimal classifier according to a quasi-concave performance metric produces a confusion matrix equal to that of an optimal classifier under a linear performance metric [Hiranandani et al., 2019a] – a class sufficiently broad to include most of the metrics in common use.

Concretely, we can consider threshold decision rules $h \in \mathcal{H}$ of the form $h_\tau(x) = \mathbb{I}(p(y = 1|x) \geq \tau)$ where $p(y = 1|x)$ is the positive class-conditional probability for the decision problem and $\mathbb{I}(\cdot)$ is an indicator function. We'll commonly abbreviate the positive class-conditional probability as positive likelihood $p(x)$ and the image over the example space as $p(\mathcal{X})$. This choice well-justified, as in the following result.

Lemma 3.1. *[Hiranandani et al., 2019a] An optimal decision rule h over $\mathcal{D}$ for some quasi-convex performance metric $\phi : \mathcal{C} \to \mathbb{R}$ is given by a threshold function of the positive class-conditional probability for the observation.*

Thus, the elicitation problem reduces to determining a decision threshold $\tau^* \in (0, 1)$. Intuitively, the threshold τ^* indicates a trade-off between the cost of false-positives and false-negatives which can then be converted to a loss function [Hiranandani et al., 2019a, Laidlaw and Russell, 2021]. We can write a performance metric for a distribution $\mathcal{D}$ as a function $\Phi_{\mathcal{D}} : [0, 1] \to \mathbb{R}$ from the space of thresholds to an evaluation score. In particular, $\Phi(\tau) = \Psi \circ \mathcal{C}_{\mathcal{D}}(h_\tau)$. Because optimal classifiers are characterized by their decision threshold, we can also represent the performance metric as a function from a decision threshold to an evaluation score.

3.1 Inverse Decision Theory

In the inverse decision theory framework we assume we know everything about the sample distribution $(X, Y) \sim \mathcal{D}$. This means we have access to class conditional statistics and classifications from a decision maker optimal with respect to the decision problem. The goal of IDT is to search for a threshold probability τ^* that best explains a decision

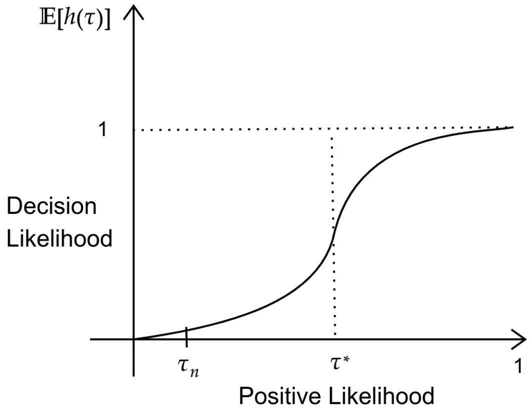


Figure 3: Illustration of a decision rule that our framework can estimate from observations. Shown is the likelihood $\mathbb{E}[h(\tau)]$ that the decision rule h makes a detection on a random example with positive likelihood τ_n. We consider cumulative decision rules in this work, decision rules that are monotone in expectation as a function of the example's true likelihood.

maker's classifications. Formally, inverse decision theory takes individual decisions $\hat{y} \in \mathcal{Y}$ on samples $\hat{x} \in \mathcal{X}$ as the base object of study rather than the population-based confusion matrix.

Definition 3.2. A decision problem is a pair $(\mathcal{D}, \tau^*)$ where $\mathcal{D}$ is a joint distribution over observations $x \in \mathcal{X}$ paired with ground truth decisions $\mathcal{Y} \in \{0, 1\}$ and $\tau^* \in (0, 1)$ is a decision threshold.

As discussed in the previous section, the unknown τ^* is equivalent to having an unknown quasi-convex loss function. Overall, the goal of inverse decision theory is to identify a decision map $h : p(\mathcal{X}) \to \mathcal{Y}$ in a hypothesis space $\mathcal{H}$ that matches the decisions of the human. When the decision maker is optimal then they act as a threshold classifier for some $\tau^* \in (0, 1)$ that maximizes an implicit performance metric $\Phi(\tau)$. A central result is that we can identify the decision problem given $\mathcal{D}$ and observations of classifications $(X, \hat{Y}) \sim \hat{\mathcal{D}}$ from the decision maker,

Theorem 3.3. *[Laidlaw and Russell, 2021] Assume that there exists $p_{\tau^*} > 0$ such that $\mathbb{P}(p(\mathcal{X}) \in (\tau^*, \tau^* + \epsilon]) \geq p_{\tau^*}\epsilon$ and $\mathbb{P}(p(\mathcal{X}) \in [\tau^* - \epsilon, \tau^*)) \geq p_{\tau^*}\epsilon$. Let $\epsilon > 0$ and $\delta > 0$. Let $\hat{\tau}$ be chosen to be consistent with m observed decisions of an optimal decision maker following a distribution $\hat{\mathcal{D}}$ corresponding to threshold τ^*. Then $|\hat{\tau} - \tau^*| \leq \epsilon$ with probability at least $1 - \delta$ as long as the number of samples satisfies $m \geq \frac{\log(2/\delta)}{p_c \epsilon}$.*

This bound is tight to constant factors. However, this analysis can be extended to address experts with suboptimal decision rules in a restricted choice set for a convergence rate of $O(m^{-1/2})$ which is also tight [Laidlaw and Russell, 2021].

Algorithm 1: Probabilistic Bisection Algorithm

```
Input  : Classification expert's decision rule $h_n$,
         prior density $f_0$, number of iterations $T$,
         update confidence level $\alpha$
Output : Density, Estimator for optimum
for $n \in [0, \dots, T-1]$ do
    $F_n = \mathrm{CDF}(f_n)$ ;
    $\tau_n = F_n^{-1}(1/2)$ ;
    if $h_n(\tau_n) = 1$ then
        $f_{n+1}(\tau) = \begin{cases} 2(1-\alpha) f_n(\tau) \text{ if } \tau \leq \tau_n, \\ 2\alpha f_n(\tau) \text{ if } \tau \geq \tau_n \end{cases}$ ;
    else
        $f_{n+1}(\tau) = \begin{cases} 2\alpha f_n(\tau) \text{ if } \tau \leq \tau_n, \\ 2(1-\alpha) f_n(\tau) \text{ if } \tau \geq \tau_n \end{cases}$ ;
    end
end
return $f_n, F_n^{-1}(1/2)$ ;
```

3.2 How Classifications Reveal Preferences

It is natural to investigate what information about an underlying threshold can be obtained from observing a classification. While inverse decision theory is able to learn preferences from observing classifications this process is framed in terms of matching a decision rule. Here we'll present a more Bayesian perspective that puts individual classification decisions as the fundamental data.

From a single example we obtain directional information about the optimal threshold. The positive likelihood $\tau \in (0, 1)$ of an example determines the agent's classification. Accordingly, we will define the random variable $h(\tau)$ as the outcome of $h(x)$ conditional on $p(y = 1|x) = \tau$. For an optimal decision rule, $h(\tau)$ is a parameterization of h in terms of positive likelihood that indicates the direction of the optimal threshold τ^* relative to τ. The expected value $\mathbb{E}[h(\tau)|\tau]$ is illustrated in Figure 3. If $h(x) = 0$ we should conclude that $p(y = 1|x) \leq \tau^*$ more than likely. Otherwise, $h(x) = 1$ and we should conclude that $p(y = 1|x) \geq \tau^*$ more than likely.

We will frequently make observations of $h(\tau)$ without knowing the underlying decision rule and it is entirely plausible that such observations will be corrupted or noisy in one manner or another. Our ultimate goal is to use the expected behavior of $h(\tau)$ to learn about the underlying decision rule. The Probabilistic Bisection Algorithm (PBA) is a useful way to proceed in such situations [Horstein, 1963]. In this approach, we take $h(\tau)$ as a black-box oracle and perform Bayesian updating for the position of an underlying decision threshold parameter by attaching a confidence score $\alpha \in (1/2, 1]$ to obtained classifications at a given threshold. We start with uniform prior density f_0 over the unit interval and query at the current median of the belief distribution for the optimal decision threshold. Let F_0 denote the corresponding cumulative distribution function.

This updating scheme, shown in Algorithm 1, gives the Bayesian posterior distribution of the optimum assuming f_0 has support on the unit interval. We form our estimate as the current measurement point.

4 COOPERATIVE INVERSE DECISION THEORY

Here we'll introduce cooperative inverse decision theory (CIDT) as a way to understand the benefits of cooperation for preference learning. See Figure 1 for a illustration of the procedure. We also propose Algorithm 2 as a baseline implementation of CIDT which comes with theoretical guarantees, but there are many other possibilities that could be explored, such as using reinforcement learning. Finally, we demonstrate conditions under which cooperation generates expert behavior that differs from IDT and has better human sample-complexity properties.

To facilitate our analysis, we focus on binary decisions in this work. However, our findings can be applied to decision problems involving a larger number of options as well. This is because we can decompose a decision among many choices into a series of binary choices between pairs of alternatives under the assumption of independence among the options [Luce, 1977, Laidlaw and Russell, 2021].

4.1 The Framework

To define the cooperative inverse decision theory game we assume the imitator R can select a random example from the set $p^{-1}(\tau)$ for any $\tau \in (0,1)$ and that the demonstrator H has an underlying decision rule h^* and threshold τ^* they consider optimal. The goal of R is to present examples to H and learn a threshold decision rule that closely matches H's preferences. The goal of H is pick a classification rule h, not necessarily equal to h^*, that will induce R to learn τ^*. Specifically, we assume that both H and R both aim to maximize the following performance metric

$$\Phi_{\tau^*}(\tau^R) = -|\tau^R - \tau^*|, \tag{2}$$

where τ^R is the learned threshold and τ^* is the optimal threshold according to H. Using a surrogate allows for an evaluation independent of the underlying decision problem, making it more robust to distribution shift [Laidlaw and Russell, 2021].

In each stage, R queries H with an example, and H responds with a classification based on h. R updates its belief distribution for τ^* based on this response. The state S_n at the n^{th} stage is the set of examples and responses generated so far. A policy π for R is a mapping from any given state to a query for H.

Algorithm 2: Implementation of CIDT

```
Input  : Sampled decision rules $h_n$, prior density
         $f_0$, iterations $T$, desired confidence level
         $\alpha$ for obtained signals, weight $\delta$
Output : Density, Estimator for optimum
$n = 0$ ;
$\tau_0 = \text{CDF}(f_0)^{-1}(1/2)$;
while $\text{sum}(N_n) \leq T$ do
    /* Empirical majority vote    */
    $\text{Result}_n, N_n = \text{Vote}(h_n, \alpha, \tau_n)$ ;
    $f_{n+1}, \tau_{n+1} = \text{PBA}_\alpha(\text{Result}_n, f_n, 1)$ ;
    $F_{n+1} = \text{CDF}(f_{n+1})$;
    $n = n + 1$ ;
end
return $\sum_{i=0}^{n-1} N_i^\delta \tau_i / \sum_{i=0}^{n-1} N_i^\delta$ ;
```

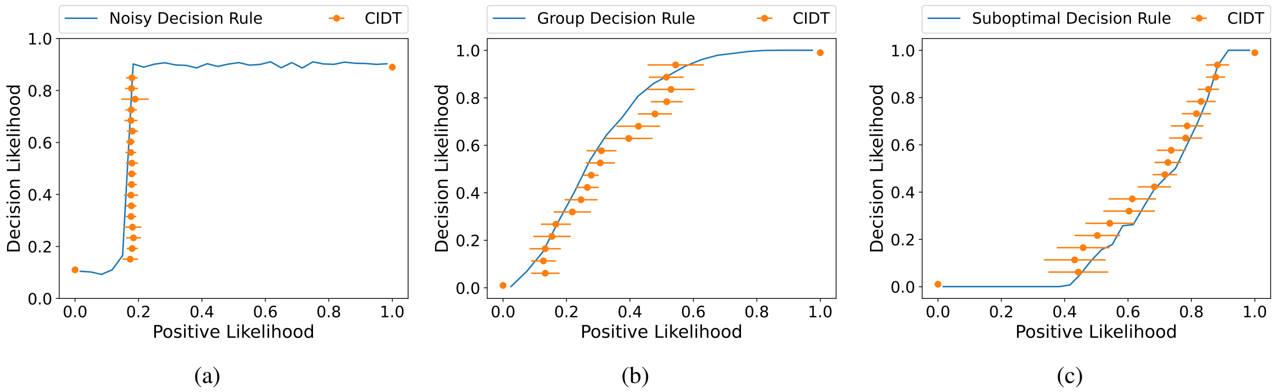
The policy π induced by R will generate an allocation of probability measures $f_0, \ldots, f_T$ for the optimal threshold where f_{n+1} is $\mathcal{G}_n$-measurable and $\mathcal{G}_n$ is the σ-algebra generated by the examples and classifications. Under certain conditions, a joint policy consisting of a specific decision rule for the demonstrator and a particular algorithm for the imitator is optimal in expected performance.

Theorem 4.1. *Consider a cooperative inverse decision theory game between R and H to determine the threshold τ^*. Suppose H plays the decision rule $h = \mathbb{I}(\tau \geq \tau^*)$ and the imitator uses the probabilistic bisection algorithm (PBA) with $\alpha = 1$ to select examples and update their belief distribution f_T for each step $T \in \mathbb{N}$ of the game. If the initial prior f_0 for the performance parameter τ^* is uniform on the unit interval, then this joint policy is optimal in the sense that it maximizes the expected performance metric $\mathbb{E}_{\tau^* \sim f_T}[\Phi_{\tau^*}(\tau^R)]$ for any $T \in \mathbb{N}$.*

This theorem shows that, in order to optimize performance, the demonstrator should classify examples based on their position relative to their optimal threshold τ^*, while the imitator should select examples that it believes are likely to be near the optimal threshold. It is not difficult to see that this policy has a geometric convergence rate, which is significantly faster than the $\Omega(1/n)$ lower bound on the sample efficiency of elicitation obtained in previous work on inverse decision theory [Laidlaw and Russell, 2021].

4.2 Sub-optimal Teaching

The optimal policy introduced in the previous section can be quite complex for the demonstrator to use due to the requirement for the human demonstrator to make Bayes optimal classifications. Here we partially relax this assumption by considering a more general class of decision rules that allows for sub-optimal teaching from H. When H is a rational, noisy, or group of decision makers this will be sufficient for a practical implementation. In the most general setting,



(a)

(b)

(c)

Figure 4: From left to right we show reconstructions predicted by cooperative inverse decision theory for noisy, group, and suboptimal decision rules. Error bars are the standard deviation of the estimated quantiles over multiple experiments. The results show that CIDT can recover the decision rule at various possible thresholds. In other words, CIDT can efficiently estimate the full preference distribution pointwise, in noisy settings. Further details in text.

we need to make an assumption that H has the ability to demonstrate a decision rule such that R can identify τ^*.

4.2.1 Uncertain Preferences

Our first assumption is that H has and demonstrates a decision rule that increases in expectation with the underlying likelihood.

Definition 4.2. Let h be a decision rule. We define the random variable $h(\tau) = h(X)$ where X is a random example such that $p(y = 1|X) = \tau$. We say that h is cumulative if $\mathbb{E}[h(\tau)|\tau]$ is monotone as a function of τ.

Indeed, it is possible to interpret any cumulative decision rule as implying that the decision maker has uncertain preferences following some distribution.

Lemma 4.3. *Every expert with a cumulative decision rule h has a unique representation as an expert that classifies examples using a random decision threshold variable T.*

The main implication of this result is that we can model experts with cumulative decision rules as having uncertain preferences. As a reminder, H views some decision rule h^* as optimal, demonstrates a rule h, and then R will elicit a decision threshold from observations of h acting on selected examples such that $\mathbb{E}[h(\tau_q)] = q$. From Lemma 4.3 we see this is some quantile of the decision threshold distribution used by H on examples.

Cumulative decision rules are fairly general. There are several classes of sub-optimal experts that can be represented as an expert with uncertain preferences.

Noisy Experts: An expert whose optimal decisions are corrupted with some constant probability corresponds to an expert whose preferred decision threshold is uncertain between 0, τ^*, and 1.

Multiple Experts: A mixture of experts with a distribution of thresholds can be represented as a single expert with uncertain thresholds following the same distribution.

Sub-optimal Experts: Sometimes the expert has an altered context with either missing or additional information. If they are approximately optimal with respect to the given context we can expect their decision likelihood to increase with the positive likelihood of given examples.

4.2.2 Description-Experience Gap

The expert (denoted H) may alter its behavior by producing sub-optimal classifications in order to more effectively communicate its preferences to the algorithm (denoted R). For the general setting, we make an assumption that H can steer R towards the correct underlying preference. However, we can remove this assumption when H is a rational, noisy, or group of decision makers which we discuss in Section 4.3.

Specifically, we will assume that $\tau_q = \tau^*$. However, we do not assume that $\tau_q^* = \tau^*$. As a reminder, τ_q^* is the threshold such that $\mathbb{E}[h^*(\tau_q)] = q$. In fact, H may misreport on a disagreement set defined by examples such that $p(X) \in [\min(\tau_q^*, \tau^*), \max(\tau_q^*, \tau^*)]$.

The description-experience gap, the difference between classifying and demonstrating can be measured using the size of this set. If the measure is larger, then H will classify more examples sub-optimally to steer R to the correct underlying preference. If the measure is zero, then H is well-aligned with R and does not alter its behavior.

Uncertainty in preferences also corresponds with the size of this disagreement set. For example, we could quantify uncertainty with the narrowness of the support of decision thresholds used by H. In the limit, this approximates a Bayes optimal decision rule and the size of the disagreement set goes to zero. Accordingly, the size of this disagreement set can be seen as a measure of the description-experience gap, and highlights the role that the implicit bias of the algorithm plays in determining its size. In practice, we may want the description-experience gap to be small so that the human does not have the added burden of accounting for the dynamics of the machine it is interacting with.

4.3 Practical Implementation

In this section, we describe our approach (Algorithm 2) for implementing a practical solution for CIDT that accounts for sub-optimal demonstration from H. Our proposed solution is to have the demonstrator(s) classify multiple examples with similar positive likelihood and then process the responses using a voting mechanism. Since we use a finite number of observations we also estimate the confidence of the result. We then have R perform a Bayesian update for the optimal threshold using Algorithm 1. We repeat until we reach desired accuracy.

The voting mechanism we will consider is to take the q^{th} quantile of the classification responses. This is as a truthful mechanism for noisy and group expert decision rules which means that if we apply it as a processing step the demonstrator(s) gain nothing from deviating away from their true underlying decision rule [Black, 1948, Rowley, 1984]. In this case we achieve demonstrator expressiveness, discussed in Section 4.2, via a voting mechanism.

We can interpret voting as a test of whether or not the expected decision of an example with positive likelihood τ satisfies $\mathbb{E}[h(\tau)] \geq q$. For a proof see the appendix. However, we need to estimate the confidence $\alpha(\cdot)$ which is the chance our election draws the wrong conclusion given some decisions on examples with a given positive likelihood. Formally, we obtain a sequence of responses $\{h_i(\tau)\}$. If we can confidently determine the sign of the drift θ formed by the random walk $S_m(\tau) = \sum_{i=0}^{m} 2h_i(\tau) - 1$ then we'll be able to construct a new signal with less noise. Sequential tests of power-one are useful for this task and allow us to determine the sign of the drift θ [Lai, 1977].

We follow [Frazier et al., 2019] in discussing our statistical test. A test of power one can be defined through a positive sequence k_m and a stopping time $N(\tau) = \inf\{m \in \mathbb{N} : |S_m(\tau)| \geq k_m\}$. If $S_{N(\tau)} \geq k_{N(\tau)}$ then the test decides $\theta > 0$ and if $S_{N(\tau)} \leq -k_{N(\tau)}$ then the test decides $\theta < 0$. We may take for $\gamma \in (0, 1)$,

$$k_m = (2m(\log(m+1) - \log(\gamma)))^{1/2} \tag{3}$$

At the $n+1$ iteration we'll observe a random walk with m^{th} term $S_{n,m} = \sum_{i=1}^{m} 2 \cdot (h_i(\tau_n) - q)$ which will almost surely have a finite stopping time whenever the confidence we have in the signal satisfies $\alpha > 1/2$. Thus, we can define a new signal,

Elicitation Error	Sample size = 15		Sample size = 30	
Feature Subset	IDT	CIDT	IDT	CIDT
15	0.176	**0.063**	0.174	**0.046**
25	0.171	**0.061**	0.157	**0.061**
30	0.194	**0.048**	0.195	**0.047**

Table 1: We show the mean elicitation error (cooperative) inverse decision theory on the Breast Cancer dataset. This is measured as the absolute difference between the elicited and optimal threshold. We model decision makers as reasoning over a random subset of the available features.

$$\text{Vote}_n(\tau_n) = \begin{cases} 1, & \text{if } S_{n,N_n} > 0, \\ 0, & \text{if } S_{n,N_n} < 0 \end{cases} \tag{4}$$

From [Frazier et al., 2019] we know that conditioned on the event the drift is above/below q we have,

$$P(\text{Vote}_n = 1 | h_q(\tau_n) = 0) \leq \gamma/2 \tag{5}$$

$$P(\text{Vote}_n = 0 | h_q(\tau_n) = 1) \leq \gamma/2 \tag{6}$$

Thus, we have $\alpha(\tau) \geq 1 - \gamma/2$. As a final remark, the closer an example's positive likelihood is to the median threshold preference the larger the error rate. This will lead to a gap between the number of iterations and the number of expert queries used.

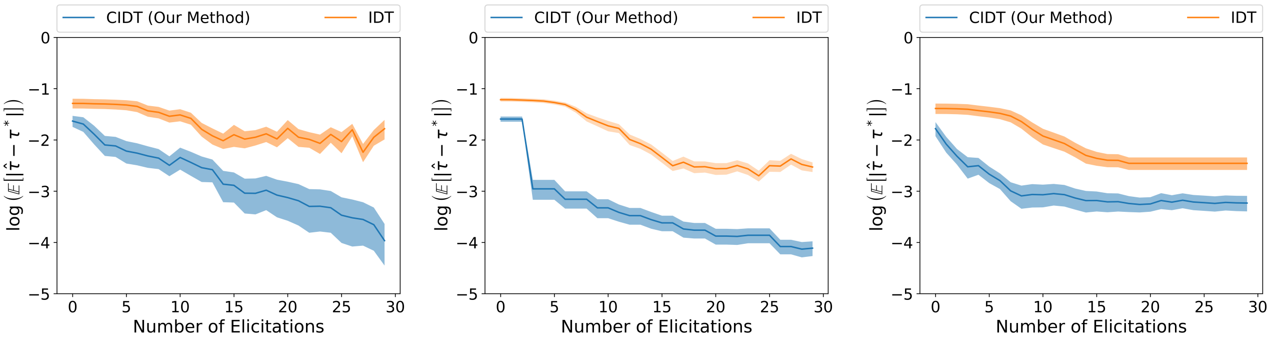
4.4 Performance Guarantees

In this section, we analyze the efficiency of the PBA implementation for CIDT under the assumptions of sub-optimal teaching discussed in Section 4.2. We provide theoretical guarantees for the convergence of the algorithm when responses are (noisely) optimal, demonstrating that it converges to the optimal solution at a geometric rate similar to the optimal policy of Theorem 4.1. Furthermore, we also provide a sample complexity analysis for the general case of a cumulative decision rule and show it is arbitrarily close to the minimax rate [Massart and Nédélec, 2006].

When H is (noisely) optimal, the PBA implementation for CIDT converges to the optimal solution at a geometric rate.

Theorem 4.4. *[Frazier et al., 2019] Suppose H makes accurate classifications with confidence $\alpha \in (1/2, 1]$. Consider having R run the PBA starting with uniform prior f_0 to produce a sequence of outputs $\tau_n = F_n^{-1}(1/2)$. Then there exists an $r > 0$ such that $e^{rn}(\tau_n - \tau^*) \to 0$ as $n \to \infty$ almost surely.*

The implications of this theorem are that CIDT can converge to the optimal solution at rate similar to the optimal joint policy which is much faster than other existing methods, such as IDT, which only have a lower bound of $\Omega(1/n)$ for sample efficiency of elicitation. However, when H demonstrates



(a)

(b)

(c)

Figure 5: We show the elicitation error for (cooperative) inverse decision theory as a function of the number of seen examples. From left to right, we display the performance of (cooperative) inverse decision theory for noisy, group, and suboptimal decision rules. We display uncertainty as standard deviation over experiment trials. Choice of the underlying preference threshold, group decision rule, and suboptimality are randomized across trials.

with a cumulative decision rule, we also need to estimate our confidence in the responses. To do this, we repeat queries on decisions some N_i times to satisfy a statistical test and then run the PBA. This is described in Algorithm 2. This implementation for CIDT converges to the optimal, but at a slower rate than Theorem 4.4.

Theorem 4.5. *Suppose H demonstrates a cumulative decision rule to R which runs the PBA with repeated queries. Let N_i be the number of comparisons used for the power-one test at each stage and let $T_i = \sum_{i=0}^{n-1} N_i$ be the total number of comparisons used up to time n. Suppose that,*

$$\alpha(\tau_n) - 1/2 \geq c|\tau_n - \tau^*|^\gamma \qquad (7)$$

for some $c > 0$ and $\gamma > 1$. Also, for $\delta \in (0, 1/2\gamma)$ define the estimator,

$$\hat{\tau}_n(\delta) = \frac{\sum_{i=0}^{n-1} N_i^\delta \tau_i}{\sum_{i=0}^{n-1} N_i^\delta} \qquad (8)$$

where τ_i is the median of the i^{th} posterior distribution. This estimator satisfies,

$$\mathbb{E}\left[(T_{n-1})^\delta \cdot |\hat{\tau}_n(\delta) - \tau^*|\right] \rightarrow_{a.s.} 0 \qquad (9)$$

This theorem applies directly to the setting where our expert has a cumulative decision rule so it can be used for elicitation of noisy, group, and suboptimal decision rules. For a proof see the appendix. However, it is slower than the geometric rate of convergence of Theorem 4.4 because the noise-condition degrades as the algorithm approaches the optimum. However, this noise-condition can be related to Massart noise, which can be used to show that the rate can be made arbitrarily close to optimal [Massart and Nédélec, 2006]. Despite this slower rate of convergence, we can still use Theorem 4.4 to argue that the algorithm can elicit the optimal solution within any fixed error $\epsilon > 0$ at a geometric rate.

5 EXPERIMENTS

In this section, we have two aims. First, we empirically validate the advantage of CIDT over IDT by presenting a comparison. IDT serves as an appropriate comparison in this setting as it is equivalent to learning from a static data-set. Second, we demonstrate that the expected behavior of cumulative decision rules are identifiable [1]. More experimental details are provided in the Appendix.

We define synthetic decision problems for our first two experiments and then consider real decision problems for our third experiment. We assume a joint probability for $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{-1, 1\}$ given by a uniform distribution $f_{\mathcal{X}} = \mathbb{U}[-\sqrt{10}, \sqrt{10}]^{10}$ and logistic noise model $\eta(x) = p(y = 1|x) = \frac{1}{1+e^{\langle a, x\rangle}}$ where a is randomly selected in our experiments. To avoid giving an unfair advantage to our approach we fix a training set of data before running algorithms. When we query for an example such that $p(y = 1|x) = \tau$ we look through the data set for the closest match and then query with this example.

We consider noisy, group, and suboptimal decision rules. The noisy decision rule randomly flips the classification from an underlying optimal decision rule with probability $1/4$. The group decision rule uses a random decision threshold for each query. These distributions are randomly sampled beta distributions. The suboptimal decision rule estimates $\eta(x)$ with $\hat{a}$ which is a random perturbation of a. This is meant to simulate having a systematically suboptimal representation.

We aim to recover a best approximating threshold decision rule and a reconstruction of the underlying decision rule as a function of positive likelihood. In the first set of experiments, we use CIDT to elicit quantiles of the decision rules and then compare them with ground truth. We use thirty

[1] https://github.com/zrobertson466920/Cooperative-IDT

elicitations to elicit each quantile. We construct ground truth by classifying a large number of examples and then constructing a histogram based on class-conditional probability and likelihood of detection. We repeat the experiment 30 times and average the results. In the second set of experiments, we compare the convergence rate of CIDT with IDT for eliciting optimal decision thresholds. We use thirty elicitations to produce an estimate. We repeat these experiments 30 times and average the results.

For our final experiment we work with the Breast Cancer Wisconsin Diagnostic dataset containing 569 instances. This is a real-world dataset that is particularly challenging due to its relatively small size. For this experiment we work with sub-optimal decision makers that reason over a random subset of the available 30 features. We take ground-truth to be IDT elicitation using the entire training set. We then measure elicitation error as the absolute difference between the elicited and the ground-truth optimal threshold. We run our experiment 30 times and then report the mean elicitation error.

From Figues 4 and 5 we see that we are able to recover approximations of decision rules with uncertain underlying preferences. We also see that approximations of the underlying class-conditional probabilities suffices for the corresponding decision rule to be cumulative. Finally we observe that our CIDT algorithm converges at a faster rate than IDT for noisy, group, and suboptimal decision rules. We also report results from applying CIDT to our real-world dataset in Table 1. We see improved elicitation using our proposed cooperative inverse decision theory approach, suggesting that CIDT may be an effective strategy in this few-shot setting. We also see that our algorithm is effective even when the decision maker is using a sparse subset of features which provides some evidence that our monotone decision rule may hold in practical settings.

6 LIMITATIONS

While our investigation makes advances in applying preference elicitation in settings with noise, group, or suboptimal considerations there are some limitations that may be addressed by future work. Foremost, we do not perform experiments with real human subjects. While our approach is robust to a fairly general class of decision rules, it is not clear to what degree human error or more complicated settings would complicate elicitation. Related to this, our study of preference elicitation assumes that preference distributions are well-supported. However, our experiments seem to indicate our approach may not work for low-support regions of a distribution. Finally, while our theoretical results are fairly general, we do not provide optimal algorithms for the general setting, implement our solution using RL, or provide full analysis for the averaging parameter (δ in Algorithm 2) used for the crowdsourced experiments. Addressing these limitations may be promising for further work.

7 CONCLUSIONS

In this paper, we proposed cooperative inverse decision theory that is able to leverage cooperation between the human and agent to improve the sample complexity of elicitation. This approach is flexible enough to consistently reconstruct a broad collection of decision rules from observations of classifications. This bypasses the description-experience gap present in other approaches. While our work could lead to reproducing human bias we think that understanding human values is central in the long-term development of human-centered AI. We hope that our work will contribute towards AI systems that better take into account human values in their decision making.

Acknowledgements

This work is partially supported by NSF III 2046795, IIS 1909577, CCF 1934986, NIH 1R01MH116226-01A, NIFA award 2020-67021-32799, the Alfred P. Sloan Foundation, and Google Inc.

References

[Ali et al., 2022] Ali, S., Upadhyay, S., Hiranandani, G., Glassman, E. L., and Koyejo, O. (2022). Metric elicitation; moving from theory to practice. *arXiv preprint arXiv:2212.03495*.

[Balbach and Zeugmann, 2009] Balbach, F. J. and Zeugmann, T. (2009). Recent developments in algorithmic teaching. In *International Conference on Language and Automata Theory and Applications*, pages 1–18. Springer.

[Black, 1948] Black, D. (1948). On the rationale of group decision-making. *Journal of political economy*, 56(1):23–34.

[Buçinca et al., 2020] Buçinca, Z., Lin, P., Gajos, K. Z., and Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 454–464.

[Calisto et al., 2021] Calisto, F. M., Santiago, C., Nunes, N., and Nascimento, J. C. (2021). Introduction of human-centric ai assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies*, 150:102607.

[Davies, 2005] Davies, K. R. (2005). *Inverse decision theory with medical applications*. Rice University.

[Dmitriev and Wu, 2016] Dmitriev, P. and Wu, X. (2016). Measuring metrics. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 429–437.

[Frazier et al., 2019] Frazier, P. I., Henderson, S. G., and Waeber, R. (2019). Probabilistic bisection converges almost as quickly as stochastic approximation. *Mathematics of Operations Research*, 44(2):651–667.

[Gut and Gut, 2005] Gut, A. and Gut, A. (2005). *Probability: a graduate course*, volume 200. Springer.

[Hadfield-Menell et al., 2016] Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29.

[Hajian-Tilaki, 2013] Hajian-Tilaki, K. (2013). Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2):627.

[Hanson and Russo, 1981] Hanson, D. and Russo, R. P. (1981). A new stochastic approximation procedure using quantile curves. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 56(2):145–162.

[Hertwig and Erev, 2009] Hertwig, R. and Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, 13(12):517–523.

[Hiranandani et al., 2019a] Hiranandani, G., Boodaghians, S., Mehta, R., and Koyejo, O. (2019a). Performance metric elicitation from pairwise classifier comparisons. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 371–379. PMLR.

[Hiranandani et al., 2019b] Hiranandani, G., Boodaghians, S., Mehta, R., and Koyejo, O. O. (2019b). Multiclass performance metric elicitation. *Advances in Neural Information Processing Systems*, 32:9356–9365.

[Horstein, 1963] Horstein, M. (1963). Sequential transmission using noiseless feedback. *IEEE Transactions on Information Theory*, 9(3):136–143.

[Jin et al., 2020] Jin, S., Wang, B., Xu, H., Luo, C., Wei, L., Zhao, W., Hou, X., Ma, W., Xu, Z., Zheng, Z., et al. (2020). Ai-assisted ct imaging analysis for covid-19 screening: Building and deploying a medical ai system in four weeks. *MedRxiv*.

[Kushner and Yin, 1987a] Kushner, H. and Yin, G. (1987a). Stochastic approximation algorithms for parallel and distributed processing. *Stochastics: An International Journal of Probability and Stochastic Processes*, 22(3-4):219–250.

[Kushner and Yin, 1987b] Kushner, H. J. and Yin, G. (1987b). Asymptotic properties of distributed and communicating stochastic approximation algorithms. *SIAM Journal on Control and Optimization*, 25(5):1266–1290.

[Lai, 1977] Lai, T. L. (1977). Power-one tests based on sample sums. *The Annals of Statistics*, 5(5):866–880.

[Laidlaw and Russell, 2021] Laidlaw, C. and Russell, S. (2021). Learning the preferences of uncertain humans with inverse decision theory. *arXiv preprint arXiv:2106.10394*.

[Lindsey et al., 2018] Lindsey, R., Daluiski, A., Chopra, S., Lachapelle, A., Mozer, M., Sicular, S., Hanel, D., Gardner, M., Gupta, A., Hotchkiss, R., et al. (2018). Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences*, 115(45):11591–11596.

[Luce, 1977] Luce, R. D. (1977). The choice axiom after twenty years. *Journal of mathematical psychology*, 15(3):215–233.

[Massart and Nédélec, 2006] Massart, P. and Nédélec, É. (2006). Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366.

[Meister and Nietert, 2021] Meister, M. and Nietert, S. (2021). Learning with comparison feedback: Online estimation of sample statistics. In *Algorithmic Learning Theory*, pages 983–1001. PMLR.

[Pallone et al., 2014] Pallone, S., Frazier, P. I., and Henderson, S. G. (2014). Multisection: parallelized bisection. In *Proceedings of the Winter Simulation Conference 2014*, pages 3773–3784. IEEE.

[Paolacci et al., 2010] Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.

[Pap, 2010] Pap, Z. (2010). Estimation of a median point by stochastic approximation. In *IEEE 8th International Symposium on Intelligent Systems and Informatics*, pages 197–201. IEEE.

[Qian et al., 2013] Qian, B., Wang, X., Wang, F., Li, H., Ye, J., and Davidson, I. (2013). Active learning from relative queries. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

[Robertson, 2022] Robertson, Z. (2022). *Probabilistic performance metric elicitation*. PhD thesis.

[Rowley, 1984] Rowley, C. K. (1984). The relevance of the median voter theorem. *Zeitschrift für die gesamte Staatswissenschaft/Journal of Institutional and Theoretical Economics*, (H. 1):104–126.

[Settles, 2009] Settles, B. (2009). Active learning literature survey.

[Swartz et al., 2006] Swartz, R. J., Cox, D. D., Cantor, S. B., Davies, K., and Follen, M. (2006). Inverse decision theory: characterizing losses for a decision rule with applications in cervical cancer screening. *Journal of the American Statistical Association*, 101(473):1–8.

A FURTHER EXPERIMENTAL DETAILS

To evaluate, we define a classification task. We assume a joint probability for $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{-1, 1\}$ given by a uniform distribution $f_\mathcal{X} = \mathbb{U}[-5, 5]^2$ and logistic noise model $\eta(x) = p(y = 1|x) = \frac{1}{1+e^{\langle a,x\rangle}}$ where a is selected as $a = [1, 1]$. To avoid giving an unfair advantage to our approach we fix a training set of data before running algorithms. We use a dataset with 10k sampled examples, but only have the human-in-the-loop label 30. When we query for an example such that $p(y = 1|x) = \tau$ we look through the data set for a close match and then query with a randomly selected example. In our experiments we look for the nearest twenty examples and select from this subset randomly.

We consider noisy, group, and suboptimal decision rules. The noisy decision rule flips the classification from an underlying optimal decision rule with probability $1/4$. The group decision rule uses a random decision threshold for each query. These distributions are randomly sampled beta distributions. Specifically, the shape parameters are sampled uniformly from $[20, 40]$ and $[40, 100]$. This produces concentrated unimodal distributions with mean roughly uniformly distributed across the unit interval. The suboptimal decision rule estimates $\eta(x)$ with $\hat{a}$ which is a random perturbation of $[1, 1]$. Specifically, we select the elements for the vector uniformly at random from the interval $[0.8, 1.2]$. This is meant to simulate having a systematically sub-optimal representation.

We aim to recover a best approximating threshold decision rule and a reconstruction of the underlying decision rule. In the first set of experiments, we use CIDT to elicit quantiles of the decision rules and then compare them with ground truth. We construct ground truth by classifying a (100k) large number of examples and then constructing a histogram based on class-conditional probability and likelihood of detection. We run the experiment 30 times. In the second set of experiments, we compare the convergence rate of CIDT with IDT for eliciting optimal decision thresholds. We run these experiments 30 times.

B REDUCTION TO CIRL

Definition B.1. A CIRL game M is a two-player Markov game with identical payoffs between a demonstrator H and an imitator R. The game is described by a tuple, $M = \langle S, \{A^H, A^R\}, T(\cdot|\cdot, \cdot, \cdot), \{\Theta, R(\cdot, \cdot, \cdot, \cdot)\}, P_0(\cdot, \cdot), \gamma\rangle$ with the following definitions. S is a set of world states: $s \in S$. A^R is a set of actions for H: $a^H \in A^H$. A^R is a set of actions for R: $a^R \in A^R$. $T(s'|s, a^H, a^R)$ is a conditional distribution on the next world state s', given previous state s and actions (a^H, a^R) for both agents. Θ is a set of possible static reward parameters, only observed by H where $\theta \in \Theta$. $R(s, a^H, a^R, \theta)$ is a parameterized reward function that maps world states, actions, and reward parameters to real numbers. $P_0(s_0, \theta)$ is a distribution over the initial state. $\gamma \in [0, 1]$ is a discount factor.

We will assume that the demonstrator H has uncertain preferences indicated by some distribution of decision thresholds ν. At each stage, the imitator R can present any example $x \in \mathcal{X}$ to the demonstrator. Accordingly, the state S_n at the n-th stage simply consists of the examples and responses generated so far. The demonstrator will receive this example and classify it according to its positive likelihood $p(x) = \mathbb{P}(y = 1|x)$ and ν.

Corollary 1 of [Hadfield-Menell et al., 2016] allows us to represent the optimal policy solely in terms of the current state and R's belief distribution. Thus, we can describe the imitator as taking the result of classification and updating its belief distribution f_n for the demonstrator's preferred decision threshold. For the reward, we use the elicitation error.

$$R(\tau^H, \tau^R(f_n), \nu) = -\mathbb{E}_{\tau\sim\nu}[|\tau^R - \tau^*|]$$

$$\tau^* = \text{argmin}_{\tau^H} - \mathbb{E}_{\tau\sim\nu}[|\tau^H - \tau|]$$

The goal of R is to present examples to H such that after a number of rounds its reward is maximized.

C PROOFS

C.1 Proof of Theorem 4.1

Theorem C.1. *Consider a cooperative inverse decision theory game between R and H to determine the threshold τ^*. Suppose H plays the decision rule $h = \mathbb{I}(\tau \geq \tau^*)$ and the imitator uses the probabilistic bisection algorithm (PBA) with*

$\alpha = 1$ *to select examples and update their belief distribution* f_T *for each step* $T \in \mathbb{N}$ *of the game. If the initial prior* f_0 *for the performance parameter* τ^* *is uniform on the unit interval, then this joint policy is optimal in the sense that it maximizes the expected performance metric* $\mathbb{E}_{\tau^* \sim f_T}[\Phi_{\tau^*}(\tau^R)]$ *for any* $T \in \mathbb{N}$.

Proof. It's not hard to see noisier responses result in sub optimal performance. To be concrete consider an instance with $\tau^* \in \{0, 1\}$ and uniform prior f_0. The performance $\Omega(-2^{-n})$ where n is the number of stages is only obtained for $\alpha = 1$. What remains is to show that bisection with $\alpha = 1$ is in fact optimal as opposed to some other approach. This is relatively straight-forward because query responses are constrained.

Given a belief distribution f_0 for τ^* we can show the best estimator under the performance metric is given by the median of f_0. This follows from a sub-gradient calculation given by,

$$\partial_{\tau^*}\mathbb{E}_{\tau^* \sim f_0}[|\tau^* - \tau|] = \mathbb{E}_{\tau^* \sim f_0}[\text{sign}(\tau^* - \tau)] = \text{CDF}_\nu(\tau^*) - (1 - \text{CDF}_\nu(\tau^*)) = 2\text{CDF}_\nu(\tau^*) - 1 = 0$$

However, this only equals zero for $\tau^* = \text{CDF}_{f_0}^{-1}(1/2)$. In fact, this is equivalent to selecting the next example point to have true positive likelihood,

$$\text{argmax}_\tau \mathbb{E}_{\tau^* \sim f_0}[\Phi_{\tau^*}(\tau)] = \text{argmin}_\tau \mathbb{E}_{\tau^* \sim f_0}[|\tau^* - \tau|] = \text{CDF}_{f_0}^{-1}(1/2)$$

This establishes that the estimator we return at each stage is optimal with respect to the belief distribution.

What we are interested to show is that we allocate examples optimally so that the belief distribution maximizes performance. Note that when we show an example to the demonstrator they aim to return a response that maximizes the final performance metric. Consider the case where they report accurately then all the belief distributions will be uniform and positive support on a sub-intervals of $[0, 1]$. Thus, the expected error reduces by a factor of two during each stage. Up to scaling, these sub-intervals will be equivalent to the unit-interval and so we can consider the performance of selecting a true positive rate τ for the examples. Up to scaling this equals,

$$\tau \cdot \text{CDF}_\nu(\tau) + (1 - \tau) \cdot (1 - \text{CDF}_\nu(\tau)) = \tau^2 + (1 - \tau)^2$$

It's clear that the mid-point $\tau = 1/2$ is optimal. So we can conclude that accurate responses starting with uniform prior allow for geometric reduction in the elicitation error across different stages. Specifically, it is optimal to use mid-point bisection at each stage.

□

C.2 Proof of Lemma 4.3

Lemma C.2. *Every expert with a cumulative decision rule* h *has a unique representation as an expert that classifies examples using a random decision threshold variable* T.

Proof. Recall that $Z(\tau)$ is equal to the expected outcome of $h(x)$ such that the sampled $x \in \mathcal{X}$ satisfies $p(y = 1|x) = \tau$. Without loss of generality we assume that this function is right-continuous. If not, we can make it so by modifying the function on at most a countable subset of points.

Since $Z(0) = 0$ and $Z(1) = 1$ we claim there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $T : \Omega \to (0, 1)$ defined on this space such that $Z(\tau)$ is the cumulative distribution function of T. To see this take $\Omega = (0, 1)$, $\mathcal{F}$ as the restriction of the Borel sets to $(0, 1)$, and $\mathbb{P}$ as the restriction of Lebesgue measure to $(0, 1)$, which is a probability measure. For $\omega \in \Omega$ we define $\hat{\tau}(\omega) := \inf\{\tau \in \mathbb{R} : Z(\tau) \geq \omega\}$. Now, if $\hat{\tau}(\omega) \leq \tau$ then there is a sequence $\tau_n \downarrow \tau$ such that $\hat{\tau}(\tau_n) \geq \omega$ for each n. By the right-continuity of $Z(\tau)$ this implies that $Z(\tau) \geq \omega$. Right-continuous monotone increasing functions are measurable so we have that $\hat{\tau}$ is measurable and that $\mathbb{P}(\hat{\tau} \leq \tau) = \mathbb{P}((0, Z(\tau)]) = Z(\tau)$.

Note that this probability $\mathbb{P}(\hat{\tau} \leq \tau)$ is a measure over the interval $[0, 1]$. In particular, the intersection of these half-space sets is closed under intersection. Without loss of generality assume $\tau < \tau'$ then $\{\omega : T(\omega) \leq \tau\}$ and $\{\omega : T(\omega) \leq \tau'\}$ intersect to $\{\omega : T(\omega) \leq \tau'\}$. Such sets are known as π-systems and the half-spaces also generate the Borel σ-algebra on $[0, 1]$. By Dynkin's π-λ theorem knowledge of the measure on these subsets implies knowledge on all Borel subsets of

$[0, 1]$ [Gut and Gut, 2005]. Therefore, $Z(\tau)$ can be represented uniquely as the cumulative distribution of an underlying preference measure ν. So we conclude that $h(\tau)$ can be represented as an expert that classifies examples using a random decision threshold variable T.

□

C.3 Proof of Theorem 4.5

Theorem C.3. *Let N_i be the number of comparisons used for the power-one test at each stage and let $T_i = \sum_{i=0}^{n-1} N_i$ be the total number of comparisons used up to time n. Suppose that,*

$$\alpha(\tau_n) - 1/2 \geq c|\tau_n - \tau^*|^\gamma = \Delta_n \tag{10}$$

for some $c > 0$ and $\gamma > 1$. Also, for $\delta \in (0, 1/2\gamma)$ define the estimator,

$$\hat{\tau}_n(\delta) = \frac{\sum_{i=0}^{n-1} N_i^\delta \tau_i}{\sum_{i=0}^{n-1} N_i^\delta} \tag{11}$$

where τ_i is the median of the i^{th} posterior distribution. This estimator satisfies,

$$\mathbb{E}\left[(T_{n-1})^\delta \cdot |\hat{\tau}_n(\delta) - \tau^*|\right] \to_{a.s.} 0 \tag{12}$$

This requires Lemma 10 from [Frazier et al., 2019] which we state here for completeness.

Lemma C.4. *[Frazier et al., 2019] Let $W = (W_n : n \geq 0)$ be a sequence non-negative random variables, and let $(\mathcal{F}_n : n \geq 0)$ be a filtration. If $\sum_{n=0}^\infty \mathbb{E}(W_n|\mathcal{F}_n)$ almost surely, then $\sum_{n=0}^\infty W_n < \infty$ almost surely.*

Proof. We'll abbreviate $c|\tau_n - \tau^*|^\gamma = \Delta_n$. First, we'll make some manipulations to use our lemma.

$$T_{n-1}^\delta |\hat{\tau}_n(\epsilon) - \tau^*| = \frac{T_{n-1}^\delta}{\sum_{i=0}^{n-1} N_i^\delta} \left| \sum_{i=0}^{n-1} N_i^\delta (\tau_i - \tau^*) \right| \tag{13}$$

$$\leq \frac{T_{n-1}^\delta}{\sum_{i=0}^{n-1} N_i^\delta} \sum_{i=0}^{n-1} N_i^\delta |\tau_i - \tau^*| \leq \sum_{i=0}^{n-1} N_i^\delta |\tau_i - \tau^*| \tag{14}$$

If we show that this is a finite-valued random variable then we have a good bound. Recall that the N_i indicate the number of samples used in a statistical test which depends on the drift $\theta_i = 2\alpha(\tau_i) - 1$. We have that [Frazier et al., 2019],

$$\text{lim sup}_{\theta \to 0} \mathbb{E}[N(\theta)\theta^2(\log \theta)^{-1}] < \infty \tag{15}$$

Moreover, a sample-path argument establishes that $\mathbb{E}[N(\theta)]$ is decreasing in $\theta > 0$ so it follows that for any $\eta > 0$ there is a $\theta_0(\gamma) > 0$ such that for $\theta \neq 0$ we have,

$$\mathbb{E}[N(\theta)] \leq c_1 |\theta|^{-(2+\eta)} I(|\theta| \leq \theta_0) + c_2 I(|\theta| \geq \theta_0) \tag{16}$$

Now we take a sequence $\{\theta_i\}$ from the PTA and then it follows from assumption that,

$$\mathbb{E}[N_i] \leq c_3 \Delta_i^{-(2+\eta)} I(\Delta_i \leq c_4) + c_5 I(\Delta_i > c_4) \tag{17}$$

$$\Rightarrow \mathbb{E}[N_i^\delta] \leq c_3 \Delta_i^{-\delta(2+\eta)} I(\Delta_i \leq c_4) + c_6 I(\Delta_i > c_4) \tag{18}$$

We know that there is an $r > 0$ such that $e^{r\gamma i} \cdot \Delta_i \to 0$. So we could define index sets,

$$J(1) = \{\Delta_i > c_4 : i \geq 0\}, \tag{19}$$

$$J(2) = \{\Delta_i > e^{-r\gamma i} : i \geq 0 \wedge i \notin J(1)\}, \tag{20}$$

$$J(3) = \{i \notin J(1) \vee J(2)\} \tag{21}$$

This gives us,

$$\mathbb{E}[N_i^{\delta}|\tau_i - \tau^*||\mathcal{G}_i] \tag{22}$$

$$\leq I(i \in J(1))c_6|\tau_i - \tau^*| \tag{23}$$

$$+I(i \in J(2))c_3|\tau_i - \tau^*| \cdot \Delta_i^{-\delta(2+\eta)} \tag{24}$$

$$+I(i \in J(3))c_3|\tau_i - \tau^*| \cdot e^{\delta(2+\eta)r\gamma i} \tag{25}$$

The first and second sets have finite cardinality almost surely. The third set will have infinite cardinality so we must rely on the last term being geometric. We know that, for some $r' > 0$ we have $e^{ri} \cdot |\tau_i - \tau^*| \to 0$. Reassign $r = \min(r', r)$. Thus, if we have $1 - \delta(2+\eta)\gamma > 0$ the last term will be summable. This implies that $\delta < \dfrac{1}{(2+\eta)\gamma} \leq \dfrac{1}{2\gamma}$. Summing yields,

$$\mathbb{E}[N_i^{\delta}|\tau_i - \tau^*||\mathcal{G}_i] \leq c_6|J(1)| + c_3|J(2)| + c_7 \tag{26}$$

and our result follows from our previous lemma. □

C.4 Further Proofs

Theorem C.5. *Suppose a decision maker has uncertain preferences distributed according to ν and we observe decisions on examples with likelihood τ according to the random variable $h(\tau)$. The following threshold decision rule always indicates the threshold closer to $\tau_q = \text{CDF}_\nu^{-1}(q)$,*

$$h_q(\tau) = \mathbb{I}\left(\mathbb{E}[h(\tau)] \geq q\right) \tag{27}$$

Proof. Since $Z(\tau) = \mathbb{E}[h(\tau)]$ is monotone we see that $h_q(\tau)$ is determined by whether or not $\tau \leq \text{CDF}^{-1}(\tau)$. When $\tau \leq \text{CDF}^{-1}(q)$ then $h_q(\tau) = 0$ and otherwise $h_q(\tau) = 1$. Therefore, $h_q(\tau)$ always indicates the direction of $\text{CDF}^{-1}(q)$ relative to τ. □

Lemma C.6. *Whenever the population distribution of thresholds ν has full support on $[0, 1]$ the noise generated from the preference uncertainty respects the conditions of Theorem 4.5.*

Proof. Allow $\tau^* = \text{CDF}^{-1}(q)$. To see that this is in fact the case note that for sufficiently small $\delta\tau > 0$,

$$\text{CDF}_\nu(\tau^* + \delta\tau) \geq 1/2 + c\delta\tau \tag{28}$$

$$\Rightarrow \alpha(\tau_n) - 1/2 \geq c\delta\tau \tag{29}$$

for some $c > 0$. Similarly, whenever $\delta\tau < 0$ then we have,

$$1 - \text{CDF}_\nu(\tau^* - \delta\tau) \geq 1/2 + c\delta\tau \tag{30}$$

$$\Rightarrow \alpha(\tau_n) - 1/2 \geq c\delta\tau \tag{31}$$

for the same $c > 0$. □