Using Property Elicitation to Understand the Impacts of Fairness Regularizers

Jessie Finocchiaro
finocch@bc.edu
Harvard University, CRCS
Allston, Massachusetts, USA

ABSTRACT

Predictive algorithms are often trained by optimizing some loss function, to which regularization functions are added to impose a penalty for violating constraints. As expected, the addition of such regularization functions can change the minimizer of the objective. It is not well-understood which regularizers change the minimizer of the loss, and, when the minimizer does change, *how* it changes. We use *property elicitation* to take first steps towards understanding the joint relationship between the loss and regularization functions and the optimal decision for a given problem instance. In particular, we give a necessary and sufficient condition on loss and regularizer pairs for when a property changes with the addition of the regularizer, and examine some regularizers satisfying this condition standard in the fair machine learning literature. We empirically demonstrate how algorithmic decision-making changes as a function of both data distribution changes and hardness of the constraints.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning algorithms*; **Classification and regression trees**; Machine learning approaches.

KEYWORDS

Loss function design, property elicitation, fairness regularizers

ACM Reference Format:
Jessie Finocchiaro. 2024. Using Property Elicitation to Understand the Impacts of Fairness Regularizers. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3630106.3658540

1 INTRODUCTION

Machine learning is increasingly being used for prediction and resource allocation tasks pertaining to human livelihood; algorithms often make predictions based on patterns in historical data to make or supplement decisions about future events. For example, algorithms are commonly used to determine a whether or not a loan applicant should receive a loan [2, 34, 35], estimate a patient's risk of heart disease [15, 28, 32], and estimate need for public assistance [24], among other settings. Typically, an algorithm tries to predict something like the probability of an applicant repaying the loan if granted one, and then uses this prediction to assign a treatment to the applicant, such as granting or not granting a loan. Implicit in this model is the use of an underlying distribution to assign a treatment by computing some underlying summary statistic, or *property*, of the distribution over outcomes. Property elicitation studies the relationship between the choice of objective function, treatment assignments, and various statistics. For example, minimizing squared loss corresponds to predicting the *expected value* of the outcome (the probability of repayment) and deciding whether or not to give a loan based on the expected value being above a given threshold. This contrasts with minimizing the 0-1 loss, which corresponds to learning the *mode*, of whether the person is more likely than not to repay a loan, and the assigned treatment is simply the decision to grant a loan.

In most practical optimization and allocation tasks, however, one faces constraints on the treatment space, especially when the treatments impact human livelihood and when resources are scarce. In particular, fairness constraints are often employed to enforce the (approximately) equal algorithmic treatment of different predefined groups. Instead of minimizing the original loss function, these algorithms often instead minimize `loss + weight * regularizer`, where the regularization term adds a penalty for violating certain desiderata about community-level outcomes.

However, to date, there is little understanding of how adding regularization functions into the optimization problem changes the property of the data distribution learned. We give a necessary and sufficient condition for regularizers to preserve an elicited property: the property elicited by the fairness regularizer must be equivalent to the property elicited by the original loss. However, this condition is rather strong: equivalence holds *regardless of the underlying data distribution.* Therefore, we further characterize for which data distributions the optimal treatments differ or are the same. We demonstrate our results on group fairness regularizers, though other regularization functions can be used as well (e.g., [29]).

To this end, we introduce the notion of *regularized property elicitation*, and what it means for two properties to be equivalent. In Theorem 1 we show that, under mild conditions on the regularizer, a regularized property is equivalent to the original property if and only if the property elicited by the regularizer is equivalent to the original property. We apply Theorem 1 to a handful of popular fairness regularizers– the absolute difference of demographic parity, expected equality of opportunity, and equalized false positive rates– and demonstrate they are not equivalent to cost-sensitive classifications. However, it is not necessarily the case that

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658540

a regularizer changes the elicited property: many additive regularizers yield regularized properties equivalent to the original, namely (multi)calibration and bounded group loss.[1] In these cases, while the property does not change, using the regularizer is still effective because of practical limitations on the experessivity of the hypothesis class $\mathcal{H}$, among other optimization challenges. It does suggest in some sense that these equivalent regularizers value "accuracy as fairness," in line with sentiment from the original works.

In § 3, we present Theorem 1, which gives the necessary and sufficient condition for the equivalence of properties, and in § 4 demonstrate these conditions on common fairness regularizers for binary classification. For those regularizers that do change an elicited property, we additionally provide examples and geometric intuition about *for which data distributions* the regularizers change (or do not change) the optimal decision, enforcing the imposed constraints. Finally, in § 5, we demonstrate our results with empirical evaluation on synthetic data, a heart attack risk analysis dataset [32], and the German lending dataset [21].

1.1 Literature review

In machine learning, a variety of pre-, in-, and post-processing techniques have emerged in recent years to make algorithmic decision-making more fair or equitable. We focus on one algorithmic aspect of in-processing wherein one modifies the learning algorithm itself by adding a soft constraint to the objective function, which is some weighted metric of the fairness violation. The addition of fairness regularizers is one common approach to try to improve algorithmic decision-making in practice, though their effects are generally not well-understood (cf. [3, 4, 7, 8, 14, 18, 19, 22, 38]). In particular, while we study exact formulations of different fairness metrics, a handful of works in the literature have studied and used convex relaxations of the metrics of interest [37, 39, 40]. We posit that our work could also serve as a tool to examine how the first-order behavior of a parity metric and its convex relaxation align. While many proposed fairness metrics are situated in binary classification settings, extensions beyond the binary setting have been studied more recently [7, 9, 23, 38, 40]. Our framework is general enough to handle a variety of prediction tasks and regularizers beyond the fair machine learning literature.

We study the impact of regularization functions on the "right" decision an algorithm should make as a function of the underlying data distribution through the lens of property elicitation. Property elicitation is well understood on an individual basis for a variety of discrete prediction tasks [10, 25–27] and continuous estimation problems [6, 11, 12, 33, 36] on an individual level. Recently, Jung et al. [20] and Noarov and Roth [30] relate property elicitation to the notion of multicalibration. These works extend the canonical understanding of multicalibration to estimate values beyond the mean, and provide (multicalibrated) algorithms to estimate higher moments, showing a strong equivalence between calibration and property elicitation. These results align with some of the intuition provided in Theorem 1, but our result goes beyond the scope of calibration as a fairness concept. Regularizers considering community-level outcomes and group membership requires we extend traditional notions of property elicitation.

2 BACKGROUND

We are primarily concerned with evaluating the optimal treatment for various prediction tasks. Consider an agent $i \in \{1, 2, \ldots, m\} = [m]$ who will achieve some outcome $y^{(i)} \in \mathcal{Y}$ with probability $p^{(i)} \in \Delta_{\mathcal{Y}}$, where $\Delta_{\mathcal{Y}}$ is the simplex over a finite set of outcomes $\mathcal{Y}$. A central decision-maker (often a principal or algorithm) assigns a treatment $t^{(i)} \in \mathcal{T}$ to the agent, and their error is scored according to a loss function $L : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}_{+}$. As shorthand, denote $L(t^{(i)}; p^{(i)}) := \mathbb{E}_{Y \sim p^{(i)}} L(t^{(i)}, Y)$ as the expected loss over $p^{(i)}$. Moreover, we assume each agent i is a member of a group $s^{(i)} \in \mathcal{S}$, and want to ensure agents of different groups are treated fairly by the centralized decision-maker. Let $n_g := |\{i \in [m] : s^{(i)} = g\}|$ be the number of agents belonging to group g, which we assume is positive for each $g \in \mathcal{S}$. Often, we are concerned with possibly set-valued functions, $\Gamma : \Delta_{\mathcal{Y}} \to 2^{\mathcal{T}} \setminus \{\emptyset\}$; for shorthand, we denote this $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{T}$.

In supervised machine learning, predictions are made by learning a hypothesis function $h : \mathcal{X} \to \mathcal{T}$ mapping features $x \in \mathcal{X}$ to treatments $t \in \mathcal{T}$. We assume $\mathcal{T}$ is a finite set unless otherwise stated. If the class of hypotheses $\mathcal{H}$ is sufficiently expressive, then t encapsulates how the optimal hypothesis *should assign treatment*, given an input x. Equivalently, we are concerned with optimal decisions under $p^{(i)} = \Pr[Y \mid X = x^{(i)}]$. For simplicity, we abstract away $\mathcal{X}$ and proceed with $p^{(i)} \in \Delta_{\mathcal{Y}}$ and $t^{(i)} \in \mathcal{T}$ in the sequel.

2.1 Regularization functions

Often, "fair" algorithms constrain optimization to ensure certain desiderata are satisfied. However, some standard optimization algorithms such as stochastic gradient descent often soften these constraints, adding an additional penalty to the loss function for violating the constraints. We study how the addition of regularization functions $\mathcal{R} : \mathcal{T}^m \times \mathcal{S}^m \times \Delta^m_{\mathcal{Y}} \to \mathbb{R}_{+}$ (henceforth: regularizers) change the optimal treatment assigned by minimizing the expected loss.

For example, imposing group fairness constraints, one might aim to ensure treatments are independent of the sensitive statistic (as in demographic parity) or treatments are calibrated to line up with the true probabilities of positive classification (as in multicalibration). In this setting, given a collection of individuals $\{(s^{(i)}, p^{(i)})\}$, we aim to optimize

$$\min_{\mathbf{t} \in \mathcal{T}^m} L^{\mathcal{R},\lambda}(\mathbf{t}; \mathbf{s}; \mathbf{p}) := (1 - \lambda) \underbrace{\left[\frac{1}{m} \sum_{i=1}^{m} L(t^{(i)}; p^{(i)}) \right]}_{\text{expected loss over } m \text{ agents}} + \lambda \mathcal{R}(\mathbf{t}; \mathbf{s}; \mathbf{p}) \; . \tag{1}$$

Because the regularizer might not be additive in $\mathbf{t}$, the treatment of an individual is not necessarily independent of the treatment of others. This necessitates the optimization of $\mathbf{t} \in \mathcal{T}^m$ rather than considering each data point individually, as is standard in unregularized property elicitation.

[1] We are not assigning a value judgment to whether or not a regularizer changes a property.

2.2 Property elicitation

When making predictions, a decision-maker often aims to learn a *property* $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{T}$, which is simply a function mapping probability distributions to treatments. Examples of commonly sought properties include the expected value $EV(p) = \{\mathbb{E}_{Y\sim p}[Y]\}$, the mode $\text{mode}(p) = \arg\max_y p_y$, α-quantiles, and rankings.

Definition 1 (Property, elicits). *A* property *is a function* $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{T}$ *mapping probability distributions to reports. If* $|\mathcal{T}|$ *is finite, we call* Γ *a* finite property. *Moreover, a minimizable*[2] *loss* $L : \mathcal{T} \times \mathcal{Y} \to \mathbb{R}_+$ elicits *a property* Γ *if, for all* $p \in \Delta_{\mathcal{Y}}$,

$$\Gamma(p) = \arg\min_{t\in\mathcal{T}} L(t;p) \ .$$

Conversely, we denote the level set *of a property* $\Gamma_t = \{p \in \Delta_{\mathcal{Y}} \mid t \in \Gamma(p)\}$ *as the set of distributions yielding the same optimal treatment.*

Throughout, we assume that properties are *nonredundant*, meaning that the level set Γ_t is full-dimensional[3] for all $\mathbf{t} \in \mathcal{T}$ and for each $p \in \text{relint}(\hat{\Gamma}_\mathbf{t})$, we have $|\hat{\Gamma}(\mathbf{p})| = 1$. This precludes the consideration of treatments that are rarely optimal, or only optimal if and only if another treatment is optimal as well.

Every minimizable loss elicits some property; we denote $\text{prop}[L]$ as the (unique) property elicited by the loss L. For example, the squared loss elicits the expected value [6, 33], and the level set $\Gamma_0 = \{p \in \Delta_{\mathcal{Y}} : \mathbb{E}_p[Y] = \{0\}\}$ of the expected value is the set of distributions with zero mean. We will later study the geometry of the level sets of various properties to characterize the how the minimizers of unregularized losses differ from those of their regularized counterparts. In order to do so, we consider the property Γ evaluated on a population. Given $\mathbf{p} \in \Delta^m_{\mathcal{Y}}$, we consider the extension $\hat{\Gamma}(\mathbf{p}) := [\Gamma(p^{(i)})]_i$ with level sets $\hat{\Gamma}_\mathbf{t} := \bigcap_i \{\mathbf{p} \in \Delta^m_{\mathcal{Y}} \mid t^{(i)} \in \Gamma(p^{(i)})\}$.

We now extend Definition 1 to include population-level reports for loss functions to encapsulate the case where the regularizer is not additive in $\mathbf{t}$ and/or is dependent on $\mathbf{s}$.

Definition 2 (Regularized property elicitation). *A* regularized property *is a function* $\Theta^{\mathcal{R},\lambda} : \mathcal{S}^m \times \Delta^m_{\mathcal{Y}} \rightrightarrows \mathcal{T}^m$ *mapping beliefs over outcomes to population-level treatments. Similarly, an objective function* L *regularized by* $\mathcal{R}$ *(weighted by* λ*), denoted* $L^{\mathcal{R},\lambda}$*, elicits a regularized property if, for all* $\mathbf{s} \in \mathcal{S}^m$ *and* $\mathbf{p} \in \Delta^m_{\mathcal{Y}}$,

$$\Theta^{\mathcal{R},\lambda}(\mathbf{s};\mathbf{p}) = \arg\min_{\mathbf{t}\in\mathcal{T}^m} L^{\mathcal{R},\lambda}(\mathbf{t};\mathbf{s};\mathbf{p}).$$

We let $\text{prop}[L^{\mathcal{R},\lambda}]$ *denote the regularized property elicited by* $L^{\mathcal{R},\lambda}$.

Denoting the level set of a regularized property requires some nuance because we are concerned with the change in optimal treatments as a function outcome distributions $\mathbf{p}$, but the regularized property is a function of $\mathbf{s}$ as well as $\mathbf{p}$. Therefore, we denote the level set $\Theta^{\mathcal{R},\lambda}_{\mathbf{t};\mathbf{s}} = \{\mathbf{p} \in \Delta^m_{\mathcal{Y}} \mid \mathbf{t} \in \Theta^{\mathcal{R},\lambda}(\mathbf{s};\mathbf{p})\}$ denote the level set of the regularized property $\Theta^{\mathcal{R},\lambda}$. If $\mathbf{s}$ is clear from context, we sometimes omit it and write $\Theta^{\mathcal{R},\lambda}_\mathbf{t}$. We now define a trivial, constant regularizer $\mathcal{R}$ as non-enforcing, since it never enforces any constraints.

Definition 3. *A regularizer* $\mathcal{R}$ *is* nonenforcing *if* $\text{prop}[\mathcal{R}]_\mathbf{t} = \Delta^m_{\mathcal{Y}}$ *for all* $\mathbf{t} \in \mathcal{T}$*, and* enforcing *otherwise.*

3 EQUIVALENCE OF (REGULARIZED) PROPERTIES

With an understanding of regularized property elicitation, we are now equipped to ask when a property "changes" with the addition of a regularizer to a loss; this requires us to consider what it means for properties to be unchanged, or equivalent.

Definition 4 (Equivalence of properties). *A property* $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{T}$ is equivalent to *a regularized property* $\Theta : \mathcal{S}^m \times \Delta^m_{\mathcal{Y}} \rightrightarrows \mathcal{T}^m$ *on* $\mathbf{s}$ *(denoted* $\Gamma \equiv_\mathbf{s} \Theta$ *or* $\hat{\Gamma} \equiv_\mathbf{s} \Theta$*) if, for all* $\mathbf{p} \in \Delta^m_{\mathcal{Y}}$*, we have* $\mathbf{t} \in \Gamma(\mathbf{p}) \iff \mathbf{t} \in \Theta(\mathbf{s};\mathbf{p})$.

In general, but particularly for large sets of agents, equivalence of a regularized property to its unregularized counterpart is a rather strong condition: when there is a "universally fair" report, equivalence holds if (and only if) the regularizer elicits essentially the same property as the original loss.

The proof relies on the relationship between subgradients of the Bayes risk and property values [13, Theorem 4.5]: if L elicits Γ, then there is a choice of subgradients D of the Bayes risk of L, $\underline{L}(\mathbf{p}) := \inf_{\mathbf{t}\in\mathcal{T}^m} L(\mathbf{t};\mathbf{p})$ such that there is a bijection from property values to D. Therefore, points of nondifferentiability of $\underline{L}$ form the intersection of level sets.

Theorem 1. *Fix* $\lambda \in (0,1)$ *and* $\mathbf{s} \in \mathcal{S}^m$*. Let loss* L *elicit* Γ*,* $L^{\mathcal{R},\lambda}$ *elicit* Θ*, and* $\mathcal{R}$ *elicit* H*. Then (1)* $\hat{\Gamma} \equiv_\mathbf{s} H \implies \hat{\Gamma} \equiv_\mathbf{s} \Theta$*. (2) If* H *is nonredundant, then additionally assume* $H_t \cap \hat{\Gamma}_t \cap \Theta_t \neq \emptyset$ *for all* $t \in \mathcal{T}$*. If* $\hat{\Gamma} \equiv_\mathbf{s} \Theta$*, then* $\mathcal{R}$ *is nonenforcing or* $\hat{\Gamma} \equiv_\mathbf{s} H$.

Proof. (1) The first statement is immediate as $H \equiv_\mathbf{s} \hat{\Gamma}$ implies

$$\begin{aligned} \mathbf{t} \in \arg\min_{\mathbf{t}'} \mathcal{R}(\mathbf{t}';\mathbf{s};\mathbf{p}) &\iff \mathbf{t} \in \arg\min_{\mathbf{t}'} L(\mathbf{t}';\mathbf{p}) \\ \iff \mathbf{t} \in \arg\min_{\mathbf{t}'} \lambda\mathcal{R}(\mathbf{t}';\mathbf{s};\mathbf{p}) &\iff \mathbf{t} \in \arg\min_{\mathbf{t}'} (1-\lambda)L(\mathbf{t}';\mathbf{p}) \\ &\implies \mathbf{t} \in \arg\min_{\mathbf{t}'} \lambda\mathcal{R}(\mathbf{t}';\mathbf{s};\mathbf{p}) + (1-\lambda)L(\mathbf{t}';\mathbf{p}) \end{aligned}$$

Now $\mathbf{t} \in \hat{\Gamma}(\mathbf{p}) \implies \mathbf{t} \in \Theta(\mathbf{p})$. If $\mathbf{t} \in \Theta(\mathbf{p})$, then consider two cases: if $\mathbf{t} \in H(\mathbf{p})$, we are done by assumption. If $\mathbf{t} \notin H(\mathbf{p})$, then $\mathbf{t} \notin \hat{\Gamma}(\mathbf{p})$. However, the two are equivalent, so there is some $\mathbf{t}' \in H(\mathbf{p}) \cap \hat{\Gamma}(\mathbf{p})$, so we contradict $\mathbf{t} \in \Theta(\mathbf{p})$.

(2) Observe that since $\mathcal{T}$ is finite, so is $\mathcal{T}^m$, and the function $\underline{L} : \mathbf{p} \mapsto \inf_{\mathbf{t}\in\mathcal{T}^m} L(\mathbf{t};\mathbf{p})$ is piecewise linear and concave, as it is the pointwise infimum of a finite set of affine functions (since expectation is linear). Moreover, the function $\eta^L_t : \mathbf{p} \mapsto L(\mathbf{t};\mathbf{p})$ is affine and supports $\underline{L}$ on $\hat{\Gamma}_t$ for every $\mathbf{t} \in \mathcal{T}^m$. Observe that if $\Theta \equiv_\mathbf{s} \hat{\Gamma}$, then η^L_t and $\eta^{L^{\mathcal{R},\lambda}}_t$ support $\underline{L}$ and $\underline{L^{\mathcal{R},\lambda}}$ respectively on the same sets for all $\mathbf{t} \in \mathcal{T}^m$.

Consider $\mathbf{nondiff}(f : \Delta^m_{\mathcal{Y}} \to \mathbb{R}_+) := \{\mathbf{p} \in \Delta^m_{\mathcal{Y}} \mid f \text{ is not differentiable at } \mathbf{p}\}$. Since $\lambda \in (0,1)$, then $\mathbf{nondiff}(\underline{L^{\mathcal{R}}}) = \mathbf{nondiff}(\underline{L}) \cup \mathbf{nondiff}(\underline{\mathcal{R}})$ [4]. The assumption $\Gamma \equiv_\mathbf{s} \Theta$ implies that $\mathbf{nondiff}(\underline{L}) = \mathbf{nondiff}(\underline{L^{\mathcal{R}}})$, which in turn implies $\mathbf{nondiff}(\underline{\mathcal{R}}) \subseteq \mathbf{nondiff}(\underline{L})$. If $\mathcal{R}$ is enforcing and $\mathbf{nondiff}(\underline{\mathcal{R}}) \subsetneq \mathbf{nondiff}(\underline{L})$, there must be some $\mathbf{t}' \in \mathcal{T}^m$ such that $H_{\mathbf{t}'} = \emptyset$, and therefore, H is redundant.

[2] One that attains the infimum in its first argument for all $y \in \mathcal{Y}$

[3] The affine dimension of the set equals the affine dimension of the simplex

[4] This is true regardless of $\lambda \in (0,1)$; see [10, Lemma 5]

Consider $\mathbf{p}' \in \mathbf{nondiff}(\underline{L}) \setminus \mathbf{nondiff}(\underline{\mathcal{R}})$. Observe that $\mathbf{p}' \in \hat{\Gamma}_{\mathbf{t}} \cap \hat{\Gamma}_{\mathbf{t}'}$ for some $\mathbf{t} \neq \mathbf{t}'$. There exists a $\mathbf{p} \in B(\mathbf{p}', \epsilon)$ for small $\|\epsilon\| > 0$ such that $\eta_t^{L^{\mathcal{R},\lambda}}$ supports $\underline{L^{\mathcal{R},\lambda}}$ on $\mathbf{conv}(\{p, p'\})$.

$$\begin{aligned}
&\mathbf{p} \in \hat{\Gamma}_{\mathbf{t}'} \setminus \hat{\Gamma}_{\mathbf{t}} \iff \mathbf{p} \in \Theta_{\mathbf{t}'} \setminus \Theta_{\mathbf{t}} \\
&\implies (1-\lambda)L(\mathbf{t}'; \mathbf{p}) + \lambda\mathcal{R}(\mathbf{t}'; \mathbf{p}) < (1-\lambda)L(\mathbf{t}; \mathbf{p}) + \lambda\mathcal{R}(\mathbf{t}; \mathbf{p}) \\
&\iff (1-\lambda)(L(\mathbf{t}'; \mathbf{p}') + c^T\epsilon) + \lambda\mathcal{R}(\mathbf{t}'; \mathbf{p}) < (1-\lambda)(L(\mathbf{t}; \mathbf{p}') + d^T\epsilon) + \lambda\mathcal{R}(\mathbf{t}; \mathbf{p}) \qquad \underline{L} \text{ affine on } \mathbf{conv}(\{p, p'\}) \\
&\implies \lambda\mathcal{R}(\mathbf{t}'; \mathbf{p}) \leq \lambda\mathcal{R}(\mathbf{t}; \mathbf{p}) \qquad \epsilon \to \mathbf{0}\,,
\end{aligned}$$

which implies $\mathbf{p} \in H_{t'}$, and therefore, $H_{t'} \neq \emptyset$, yielding a contradiction.

Therefore, we must either have $\mathcal{R}$ nonenforcing or $\mathbf{nondiff}(\underline{\mathcal{R}}) = \mathbf{nondiff}(\underline{L})$, the latter of which implies that H is nonredundant. We avoid permutations of level sets by the assumption that $H_t \cap \Gamma_t \cap \Theta_t$ is nonempty, and must have equivalence of the properties.

□

Intuitively, Theorem 1 says that the property elicited by a regularized loss function is the same as the unregularized loss if and only if the regularizer elicits the same property as the loss itself. Since loss functions are measurements of accuracy, then equivalence of properties implies an algorithm values accuracy as fairness.

4 (NON)EQUIVALENCE OF COMMON FAIRNESS METRICS FOR BINARY CLASSIFICATION

We now evaluate a handful of common fairness regularizers, and apply Theorem 1 to show nonequivalence between binary classification tasks and their regularized counterparts. For each regularizer, we give restrictions on $\Delta_{\mathcal{Y}}^m$ such that the regularized property is equivalent to the original under these restrictions.

To build intuition, we examine simple cases of how regularizers change elicited properties with populations of $m = 2$ agents belonging to different groups $\mathbf{s} = (a, b)$.

Figure 1 provides some additional intuition for the proof of Theorem 1. Each subfigure gives the level sets of the property elicited by the mode regularized by the demographic parity violation (DP), where each point in $[0, 1]^2$ represents $\mathbf{p} \in \Delta_{\mathcal{Y}}^2$ by $(\Pr_{p^{(1)}}[Y = 1], \Pr_{p^{(2)}}[Y = 1])$. Each colored cell depicts a different level set of a regularized property $\Theta^{DP,\lambda}$. This regularized property is overlaid on the (unregularized) mode, so that, upon visual inspection, one observes the regions where the two properties differ. As $\lambda \to 0$, the regularized property becomes increasingly similar to the unregularized, and as $\lambda \to 1$, the regularized property increasingly resembles the property elicited by $\mathcal{R}$.

4.1 Demographic parity

In the context of binary classification, one might be interested in regularizing their loss with the demographic parity violation, measured by the absolute difference of the rates at which agents are assigned the positive treatment from each of two groups. Any treatment that assigns the positive treatment at the same rate optimizes the demographic parity regularizer, which is not equivalent to the mode. That is, $H(\mathbf{s}; \mathbf{p}) \supseteq \{\mathbf{0}, \mathbb{1}\}$ for all $\mathbf{p} \in \Delta_{\mathcal{Y}}^m$ and $\mathbf{s} \in \mathcal{S}^m$. Thus, if $\mathcal{S} = \{a, b\}$[5], we can apply Theorem 1 to conclude the DP-regularized mode is not equivalent to the unregularized mode.

$$L^{DP,\lambda}(\mathbf{t}; \mathbf{s}; \mathbf{p}) = \frac{1-\lambda}{m}\sum_{i=1}^{m} L(t^{(i)}; p^{(i)}) + \lambda\left|\frac{1}{n_a}\sum_{i:s^{(i)}=a} t^{(i)} - \frac{1}{n_b}\sum_{i:s^{(i)}=b} t^{(i)}\right| \tag{DP}$$

Now, with $\mathcal{T} = \{0, 1\}$, if L is the 0-1 loss[6], we can evaluate $L^{DP,\lambda}$ for each treatment in $\mathcal{T}^2 = \{(1, 1), (0, 1), (1, 0), (0, 0)\}$.

$$\begin{aligned}
L^{DP,\lambda}((1,1); (p^{(1)}, p^{(2)})) &= \frac{1-\lambda}{2}\left[(1-p^{(1)}) + (1-p^{(2)})\right] \\
L^{DP,\lambda}((0,1); (p^{(1)}, p^{(2)})) &= \frac{1-\lambda}{2}\left[p^{(1)} + (1-p^{(2)})\right] + \lambda \\
L^{DP,\lambda}((1,0); (p^{(1)}, p^{(2)})) &= \frac{1-\lambda}{2}\left[(1-p^{(1)}) + p^{(2)}\right] + \lambda \\
L^{DP,\lambda}((0,0); (p^{(1)}, p^{(2)})) &= \frac{1-\lambda}{2}\left[p^{(1)} + p^{(2)}\right].
\end{aligned}$$

These expected losses now enable us to study the level sets $\Theta_{\mathbf{t};\mathbf{s}}^{DP,\lambda} = \{\mathbf{p} \in \Delta_{\mathcal{Y}}^m \mid \mathbf{t} \in \Theta^{DP,\lambda}(\mathbf{s}; \mathbf{p})\}$.

We have have $(0, 0) \in \arg\min_{\mathbf{t} \in \mathcal{T}^2}$ if

$$\begin{aligned}
&\frac{1-\lambda}{2}\left[(1-p^{(1)}) + (1-p^{(2)})\right] \leq \frac{1-\lambda}{2}\left[p^{(1)} + (1-p^{(2)})\right] + \lambda \\
&\qquad \iff \frac{1-3\lambda}{2(1-\lambda)} \leq p^{(1)} \\
&\frac{1-\lambda}{2}\left[(1-p^{(1)}) + (1-p^{(2)})\right] \leq \frac{1-\lambda}{2}\left[p^{(2)} + (1-p^{(1)})\right] + \lambda \\
&\qquad \iff \frac{1-3\lambda}{2(1-\lambda)} \leq p^{(2)} \\
&\frac{1-\lambda}{2}\left[(1-p^{(1)}) + (1-p^{(2)})\right] \leq \frac{1-\lambda}{2}\left[p^{(1)} + p^{(2)}\right] \\
&\qquad \iff p^{(1)} + p^{(2)} \leq 1\,.
\end{aligned}$$

Therefore, the level set $\Theta_{(0,0)}^{DP,\lambda}$ can be described by the polyhedron

$$\Theta_{(0,0)}^{DP,\lambda} = \left\{ p \in [0,1]^2 \mid \begin{bmatrix} 0 & -1 & \frac{1-3\lambda}{2(1-\lambda)} \\ -1 & 0 & \frac{1-3\lambda}{2(1-\lambda)} \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} p^{(1)} \\ p^{(2)} \\ 1 \end{bmatrix} \geq \mathbf{0} \right\}.$$

Observe that the final constraint is actually one on the marginal $P[Y]$: the expected outcome over the whole population should be less likely to be 1 than 0. We can evaluate the rest of the level sets in a similar manner.

Now let us gain some geometric intuition for how these level sets change by referencing Figure 1. For two agents belonging to different groups, each point in the figure represents a pair $\mathbf{p} := (p^{(1)}, p^{(2)})$ of true probabilities for the two agents. The pair $\mathbf{p} \in [0, 1]^2$, and the region $[0, 1]^2$ can be divided into up to $|\mathcal{T}^m|$ regions for which each $\mathbf{t} \in \mathcal{T}^m$ is contained in $\Theta^{DP,\lambda}(\mathbf{p})$. The sequence of figures in Figure 1 denotes the level sets of $\Theta^{DP,\lambda}$ as one varies $\lambda \in [0, 1]$. For intuition, one can observe that the regions where the players receive the same treatment (blue and red) grows as λ

[5] This is simply for ease of exposition, and can be relaxed.

[6] These derivations also hold if L is squared loss, hinge loss, and many other losses for binary classification.

increases, starting with $1/2$ of the $[0,1]^2$ space, and increasing to all of $[0,1]^2$ as $\lambda \to 1$.

We now turn our attention towards the regions of $\Delta_{\mathcal{Y}}^m$ where the regularized and unregularized properties are equivalent with a demographic parity regularizer. First, we observe that if uniform treatment of a population is optimal on the unregularized property, it is also optimal on the regularized property.

Proposition 1. *Fix $\lambda \in (0,1)$. Let L elicit Γ, $L^{\mathcal{R},\lambda}$ elicit Θ, and $\mathcal{R}$ elicit H. For all $\mathbf{t} \in \mathcal{T}^m$ and $\mathbf{s} \in \mathcal{S}^m$, $\hat{\Gamma}_{\mathbf{t}} \cap H_{\mathbf{t};\mathbf{s}} \subseteq \Theta_{\mathbf{t}}$.*

Proof. $\mathbf{p} \in \hat{\Gamma}_{\mathbf{t}} \cap H_{\mathbf{t};\mathbf{s}} \implies L(\mathbf{t};\mathbf{p}) \leq L(\mathbf{t}';\mathbf{p})$ and $\mathcal{R}(\mathbf{t};\mathbf{s};\mathbf{p}) \leq \mathcal{R}(\mathbf{t}';\mathbf{s};\mathbf{p})$ for all $\mathbf{t}' \in \mathcal{T}^m$, which in turn implies $L(\mathbf{t};\mathbf{p}) + \mathcal{R}(\mathbf{t};\mathbf{s};\mathbf{p}) \leq L(\mathbf{t}';\mathbf{p}) + \mathcal{R}(\mathbf{t}';\mathbf{s};\mathbf{p}) \implies (1-\lambda)L(\mathbf{t};\mathbf{p}) + \lambda\mathcal{R}(\mathbf{t};\mathbf{s};\mathbf{p}) \leq (1-\lambda)L(\mathbf{t}';\mathbf{p}) + \lambda\mathcal{R}(\mathbf{t}';\mathbf{s};\mathbf{p})$ for all $t' \in \mathcal{T}^m$. □

We apply this result to the "universally fair" reports via demographic parity $\mathbf{0}$ and $\mathbb{1}$.

Corollary 1. *Fix $\mathbf{s} \in \mathcal{S}^m$ and $\lambda \in [0,1]$. Let L elicit Γ and $\mathbf{L}^{DP,\lambda}$ elicit Θ. $\hat{\Gamma}_{\mathbf{0}} \subseteq \Theta_{\mathbf{0};\mathbf{s}}$. Moreover, $\hat{\Gamma}_{\mathbb{1}} \subseteq \Theta_{\mathbf{s};\mathbb{1}}$.*

Proof. Let $H := \text{prop}[L^{DP,\lambda}]$. For all $\mathbf{p} \in \Delta_{\mathcal{Y}}^m$, we have $\{\mathbf{0},\mathbf{1}\} \subseteq H(\mathbf{p})$. Therefore, $\hat{\Gamma}_{\mathbf{0}} \cap H_{\mathbf{0}} = \hat{\Gamma}_{\mathbf{0}}$ (and similarly with $\hat{\Gamma}_{\mathbb{1}} \cap H_{\mathbb{1}}$). Therefore, $\hat{\Gamma}_{\mathbf{0}} = \hat{\Gamma}_{\mathbf{0}} \cap H_{\mathbf{0}} \subseteq \Theta_{\mathbf{0}}$ and $\hat{\Gamma}_{\mathbb{1}} = \hat{\Gamma}_{\mathbb{1}} \cap H_{\mathbb{1}} \subseteq \Theta_{\mathbb{1}}$. □

We now turn our attention to the opposite case: if, while regularized, treating different groups differently (and uniformly within the groups) is optimal, then it is also optimal in the unregularized setting. In particular, this holds for treatments maximizing $\mathcal{R}$.

Proposition 2. *Fix $\mathbf{s} \in \{a,b\}^m$ and $\lambda \in [0,1]$. Fix $\mathbf{t} = \mathbb{1}_a$ (or $\mathbb{1}_b$ without loss of generality). Let L elicit Γ over outcomes $\mathcal{Y} = \{0,1\}$. $\Theta_{\mathbf{t};\mathbf{s}}^{DP,\lambda} \subseteq \hat{\Gamma}_{\mathbf{t}}$.*

Proof. With $\mathbf{s}$ fixed, $t \in \arg\max_{\mathbf{t}'} DP(\mathbf{t}';\mathbf{p})$ for all $\mathbf{p} \in \Delta_{\mathcal{Y}}^m$. Therefore,

$$\begin{aligned}(1-\lambda)L(\mathbf{t};\mathbf{p}) + \lambda DP(\mathbf{t};\mathbf{p}) &\leq (1-\lambda)L(\mathbf{t}';\mathbf{p}) + \lambda DP(\mathbf{t}';\mathbf{p}) \forall \mathbf{t}' \\ \implies (1-\lambda)L(\mathbf{t};\mathbf{p}) &\leq (1-\lambda)L(\mathbf{t}';\mathbf{p}) \qquad \forall \mathbf{t}' ,\end{aligned}$$

which implies the result. □

With that, we partially characterize the relationship between the unregularized and DP-regularized level sets for standard binary classification. In the simple case with $m = 2$ agents, this characterization is complete: if the optimal treatment is uniform, it stays uniform. Moreover, if the most "unfair" treatment wherein all the members of one group receive the treatment, and none of the second group is optimal in the regularized setting, it is also optimal in the unregularized setting. In any other setting, the optimal treatment changes with the addition of a DP regularizer.

4.2 Equalized FPR

Following a similar process to § 4.1, we now consider the regularizer that measures the absolute difference of false positive rates across groups, where the false positive rate is given by $FPR_g(\mathbf{t};\mathbf{s};\mathbf{p}) = \Pr[Y^{(i)} = 0 \mid t^{(i)} = 1, s^{(i)} = g] = \frac{1}{|\{i:t^{(i)}=1,s^{(i)}=g\}|} \sum_{i:s^{(i)}=g,t^{(i)}=1}(1-p^{(i)})$. The optimization problem then becomes

$$L^{FPR,\lambda}(\mathbf{t};\mathbf{s};\mathbf{p}) = \frac{1-\lambda}{m}\sum_i L(t^{(i)};p^{(i)}) + \lambda\,|FPR_a(\mathbf{t};\mathbf{s};\mathbf{p}) - FPR_b(\mathbf{t};\mathbf{s};\mathbf{p})| \tag{FPR}$$

The FPR regularizer computes the difference of false positive rates between groups, so one can observe that the false positive rate of a group is is reduced by assigning more negative treatments $t^{(i)} = 0$. We can see in Figure 2 that the FPR regularizer then makes it worse for an algorithm to assign the positive treatment to an agent i even if $p^{(i)}$ is slightly greater than $1/2$, as marked by the ★ in Figure 2(R).

As in § 4.1, we can apply Proposition 1 to show that if assigning everyone the negative treatment is optimal in the unregularized setting, it is also the optimal treatment with the FPR regularizer.

Corollary 2. *Fix $\mathbf{s} \in \mathcal{S}^m$. Let L elicit Γ and $L^{FPR,\lambda}$ elicit $\Theta^{FPR,\lambda}$. $\hat{\Gamma}_{\mathbf{0}} \subseteq \Theta_{\mathbf{0};\mathbf{s}}^{FPR,\lambda}$.*

Proof. For all $\mathbf{p} \in \Delta_{\mathcal{Y}}^m$, we have $\mathbf{0} \in H(p)$. Therefore $\hat{\Gamma}_{\mathbf{0}} = \hat{\Gamma}_{\mathbf{0}} \cap H_{\mathbf{0};\mathbf{s}} \subseteq \Theta_{\mathbf{0}}$ by Proposition 1. □

4.3 Expected equality of opportunity

While standard equality of opportunity (cf. [16]) requires access to observed labels, we are interested in equality of opportunity in expectation, and consider a variant that does not require access to labels proposed by Blandin and Kash [5]. Consider the treatment space $\mathcal{T} = \{0,1\}^m$ and regularizer $\mathcal{R}(\mathbf{t};\mathbf{s};\mathbf{p}) = |EEO_a(\mathbf{t};\mathbf{s};\mathbf{p}) - EEO_b(\mathbf{t};\mathbf{s};\mathbf{p})|$, where

$$\begin{aligned}EEO_g(\mathbf{t};\mathbf{s};\mathbf{p};g) &= \Pr_{i\sim[m]}[t^{(i)} = 1 \mid y^{(i)} = 1, s^{(i)} = g] \qquad \text{(EEO)} \\ &= \frac{\Pr[Y^{(i)} = 1 \mid t^{(i)} = 1, s^{(i)} = g]\Pr[t^{(i)} = 1]}{\Pr[Y^{(i)} = 1]} \\ &= \frac{\left(\frac{1}{|\{i:s^{(i)}=g,t^{(i)}=1\}|}\sum_{i:t^{(i)}=1,s^{(i)}=g}(p^{(i)})\right)\left(\sum_i t^{(i)}\right)}{\sum_i p^{(i)}} .\end{aligned}$$

We can apply Proposition 1 to show that uniform treatment being optimal in the unregularized case implies it is also optimal with the EEO regularizer as well.

Corollary 3. *Fix $\mathbf{s} \in \mathcal{S}^m$, and let L elicit Γ over outcomes $\mathcal{Y} = \{0,1\}$ and $L^{EEO,\lambda}$ elicit $\Theta^{EEO,\lambda}$. $\hat{\Gamma}_{\mathbf{0}} \subseteq \Theta_{\mathbf{0}}^{EEO,\lambda}$.*

4.4 Equivalent regularizers

In the previous section, we use Theorem 1 to show the nonequivalence of reguliarized properties, and examine a few common regularizers to show some restrictions that recover equivalence under certain distributional assumptions on the outcomes. We examine two regularizers that elicit the mode, and thus the regularized property is equivalent to the unregularized on all of $\Delta_{\mathcal{Y}}^m$: (multi)calibration [17, 20, 30, 31] and bounded group loss [1]. In some sense, this suggests that these regularizers value accuracy as fairness. If models are as accurate as they could possibly be, the most "fair" treatments to assign are also the most accurate. In

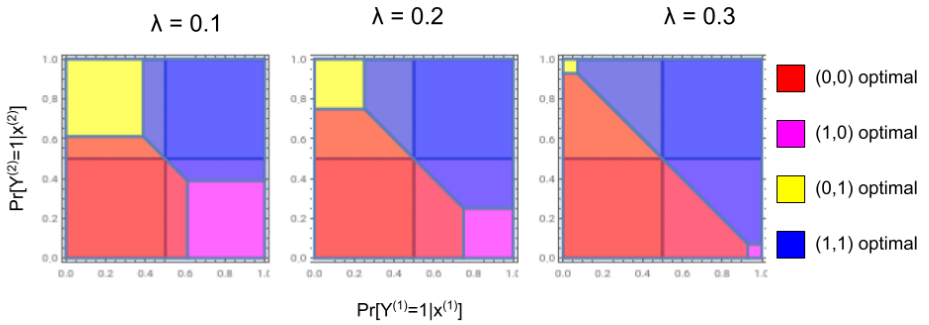

Figure 1: Visualizing the level sets of the DP-regularized property $\Theta^{DP,\lambda}$ for different values of $\lambda \in [0,1]$, where $m = 2$ and $\mathrm{s} = (a,b)$. Each point $(p^{(1)}, p^{(2)})$ in a square represents $(\Pr_{p^{(1)}}[Y = 1], \Pr_{p^{(2)}}[Y = 1])$, and each colored cell represents sets of $(p^{(1)}, p^{(2)})$ pairs such that the optimal treatment is the same for all points in the cell. For example, the magenta cell (lower right) is the set of distributions where the decision-maker prefers to attribute the positive treatment ($t^{(i)} = 1$) to the first agent, and the negative treatment ($t^{(i)} = 0$) to the second agent.

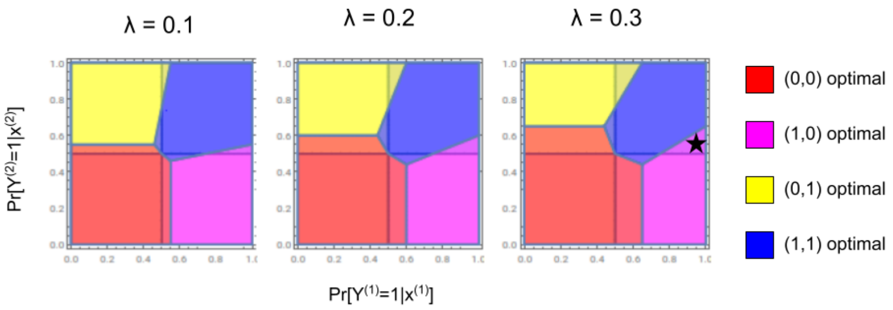

Figure 2: Visualizing the level sets of the FPR-regularized property $\Theta^{FPR,\lambda}$ for different values of $\lambda \in [0,1]$, where $m = 2$ and $s = (a,b)$. Each point $(p^{(1)}, p^{(2)})$ in a square represents $(\Pr_{p^{(1)}}[Y = 1], \Pr_{p^{(2)}}[Y = 1])$, and each colored cell represents sets of $(p^{(1)}, p^{(2)})$ pairs such that the optimal treatment is the same for all points in the cell. For example, the magenta cell (lower right) is the set of distributions where the decision-maker prefers to attribute the positive treatment ($t^{(1)} = 1$) to the first, and the negative treatment ($t^{(2)} = 0$) to the second agent.

practice, the regularizers mitigate unfairness arising from limited expressivity of the model: if the model was perfectly expressive and could predict the mode perfectly, it would assign the same treatments even with heavy penalties for "unfairness."

4.4.1 *Calibration.*

Calibration constraints ensure that the predicted value $t^{(i)}$ most closely lines up with the true probability $p^{(i)}$, regularizing the loss by the sums of the absolute differences $|t^{(i)} - p^{(i)}|$. The absolute difference elicits the $1/2$-quantile, which is also the mode on $\mathcal{Y} = \{0,1\}$, so the regularizer $\mathcal{R}(\mathbf{t};\mathbf{s};\mathbf{p}) = \sum_g \frac{1}{n_g} \sum_{i:s^{(i)}=g} |t^{(i)} - p^{(i)}|$ elicits the mode in binary classification problems.

Formally, consider the objective

$$L^{Cal,\lambda}(\mathbf{t};\mathbf{s};\mathbf{p}) = \frac{1-\lambda}{m} \sum_i L(t^{(i)};p^{(i)}) + \lambda \sum_g \frac{1}{n_g} \sum_{i:s^{(i)}=g} |t^{(i)} - p^{(i)}| \tag{Cal}$$

This constraint does not include any comparisons across group averages, so the optimal report is obtained by giving individual predictions. In binary classification, the $1/2$-quantile is the same as the mode, so the property is given $\Theta^{Cal,\lambda}(\mathbf{s};\mathbf{p}) = \mathrm{mode}(\mathbf{p})$.

This observation holds even with different weightings for specific subgroups, as in multicalibration a lá Hebert-Johnson et al. [17].

4.4.2 *Bounded group loss.*

We now consider the constraint on bounded group loss: $\mathbb{E}_{Y|S=s} L(r,Y) < \epsilon$ for all $s \in \mathcal{S}$, introduced by Agarwal et al. [1]. To model bounded group loss as a soft constraint, we simply weigh the expected loss conditioned on the group size as a regularizer, so accuracy is more incentivized on small groups.

$$L^{BGL,\lambda}(\mathbf{t};\mathbf{s};\mathbf{p}) = \frac{1-\lambda}{m} \sum_i L(t^{(i)};p^{(i)}) + \sum_g \frac{\lambda}{n_g} \sum_{i:s^{(i)}=g} L(t^{(i)};p^{(i)})$$

Adding this constraint as a fairness regularizer does not change the property elicited (e.g., $\Theta^{BGL,\lambda}(\mathbf{s};\mathbf{p}) = \hat{\Gamma}(\mathbf{p})$ for all $p \in \Delta^m_{\mathcal{Y}}$). In part this is because it still encourages the model to learn what is best for each individual in the population, where other constraints add a regularizer that compares the deviation between two groups.

Corollary 4. *Let L elicit Γ. $\hat{\Gamma} \equiv_{\mathbf{s}} \Theta^{BGL,\lambda}$ for all $\mathbf{s} \in \mathcal{S}^m$ and $\lambda \in [0,1]$.*

Proof. The regularizer $\mathcal{R}(\mathbf{t};\mathbf{s};\mathbf{p}) := \sum_g \frac{1}{n_g} \sum_{i:s^{(i)}=g} L(t^{(i)};p^{(i)})$ is additive in $\mathbf{t}$, and elicits the same property as L since it is simply a reweighing of L. □

5 EXPERIMENTS

While property elicitation allows us to reason about what a treatment an algorithm *should* assign, we examine whether or not these decisions are consistent with the treatments assigned by algorithms in practice with simple models. We first generate a set of synthetic datasets to understand how a classifier's decisions change as one navigates the space of data distributions. Moving through this space demonstrates the relationship between loss and regularizer in the synthetic setting as the data distribution over changes in $\Delta^m_{\mathcal{Y}}$. We then evaluate the effect of the regularizer weight λ on treatment assignment in cardiovascular disease risk prediction [32] and lending [21] datasets, where the data distribution is fixed. In both settings, we train a linear classifier over 30 trials with binary cross entropy loss with (a) no regularizer, (b) demographic parity difference (c) false positive rate difference (d) false negative rate difference, (e) equality of opportunity difference, and compute the fairness violations of the classifier trained on each of the four losses, where elicited property values are shown in Figure 7.

5.1 Effect of the data distribution

Recall that we applied Theorem 1 and its intuition in Figures 1 and 2 to conclude the mode is not equivalent to $\Theta^{\mathcal{R},\lambda}$ for various regularizers including (DP), (FPR), False Negative Rates (in § A), and Expected Equality of Opportunity (EEO). However, the equivalence of regularized properties and their unregularized counterparts is a rather strong condition, as pointwise equivalence must hold for *every* set of data distributions. In practice, the true data distribution may be somewhere in the space of distributions where the property value does not change for the chosen value of λ. With the knowledge that equivalent distributions have no endogeneous differences in hand, we generate a set of synthetic distributions to understand tradeoffs to regularizers as we move though the space of data distributions.

We generate generate synthetic datasets for binary classification as follows: there are two groups, $\mathcal{S} = \{a, b\}$ with $\Pr[a] = \Pr[b] = 1/2$, a member of each group has $\Pr[Y = 1 \mid S = g] = p_g \in [0,1]$. Each set of agents is represented by $x = \{p_a, p_b, r_1, \ldots, r_k\}$, where $r_1, \ldots, r_k$ are uniformly random values in $[-1,1]$. We then train a logistic regressor via stochastic gradient descent (30 trials with learning rate = 0.001, 1500 epochs, 10000 (p_a, p_b) pairs, $k = 3$), that minimizes the binary cross entropy loss regularized by either demographic parity, false positive rate, false negative rate, or difference in equality of opportunity with $\lambda = 0.15$. The simplicity of features is intentional: the "perfect" decision should be fully realizable in the unregularized setting, so the benchmark accuracy should be relatively high. Fixing the probability for a positive outcome $p_a = 0.3$ for a member of group a, we vary the probability of a positive outcome p_b for a member of group b to observe how fairness violations change as the underlying data distribution changes. For intuition, by design of the datasets, we reason about the "average member" of the population and reference the level sets drawn in Figure 4. Fixing p_a and varying p_b can be thought of as understanding what happens in decision making as one moves vertically up the line $\{(0.3, p_b) \mid p_b \in [0,1]\}$, denoted by the black dashed lines in Figure 4. In Figure 3 (L), we observe a a significant difference in DP violation rate only when $p_b \geq 1/2$. Similarly, the false positive rate violation gap "opens up" for $p_b \in [0.5, 0.65]$ in Figure 3 (ML), in line with Figure 4, and no significant different in FNR violations is observed as decision-making on this axis does not change in Figure 4. Finally, for EEO, this gap opens for $p_b \geq 1/2$, then closes again later.

5.2 The effect of choice of λ

Conversely to the interpretation of the experiments in § 5.1, to gain intuition for why decisions might change as a function of λ, we now consider each dataset representing a (p_a, p_b) point in one of Figures 1–2, and consider how the level set it belongs to changes as one changes λ. We examine two datasets, German lending [21] and heart disease risk prediction [32]. For both datasets, we train 30 linear models with 15000 epochs, learning rate of 0.001.

German lending. In the German lending dataset, we treat age as the sensitive attribute, using an indicator thresholded at 25 years old. On the entire dataset, we have $\Pr[Y = 1 \mid S \geq 25] = 0.728$ and $\Pr[Y = 1 \mid S < 25] = 0.578$, and an unbalanced group representation with $\Pr[S < 25] = 0.191$.

Perhaps surprisingly, we observe little impact of the choice of λ; moreover, in Figure 5, we observe no significant difference in the performance across fairness metrics from regularized and unregularized losses. Upon closer inspection, this can be explained partly by the observation that the "average" group members $(p_a, p_b) = (0.728, 0.578)$: a distribution that warrants treating the average member of each subpopulation the same, which aligns with most fairness regularizers. This is demonstrated in Figure 7, where the (p_a, p_b) coordinate is denoted by a g, for German. For every subfigure in Figure 7, the g coordinate is in the blue cell, implying that the "average member" of each group receives the same treatment with a fairness-regularized loss as with an unregularized loss, suggesting that for the probability distribution underlying this dataset, data subjects are already treated approximately fairly by the unregularized loss.

Heart disease risk. In the heart disease risk prediction dataset, we treat sex as the sensitive attribute, and observe $\Pr[Y = 1 \mid S = 0] = 0.75$ and $\Pr[Y = 1 \mid S = 1] = 0.449$ yields a p_a, p_b pair warranting different treatments for the "average" member of each group, and $\Pr[S = 1] = 0.63$ for a more sensitive-attribute-balanced dataset. The relationship between the optimal treatment of the "average" member of both groups as λ changes can be seen in Figure 7.

Figure 8 shows the tradeoffs incurred by large weights on fairness violations, as accuracy of regularized losses tends to drop for $\lambda > 0.3$, which aligns with some of the improvements in fairness violations– namely for demographic parity and false positive rates. There is no significant difference in the FNR violation, regardless of λ, and an increase in the EEO violation; we conjecture this is due to numerical stability as the baseline EEO violation is very small. In

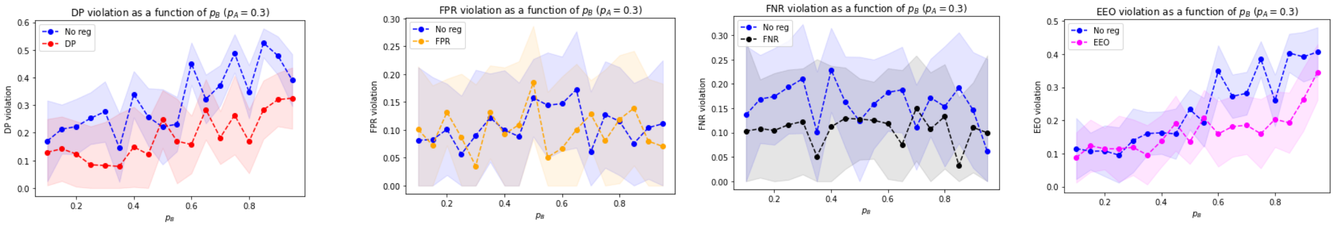

Figure 3: Regularizer values with synthetic data generated via $\Pr[Y = 1 \mid g = a] = 0.3$ and $\Pr[Y = 1 \mid g = b]$ on the horizontal axis. 95% confidence intervals over 30 randomizations included.

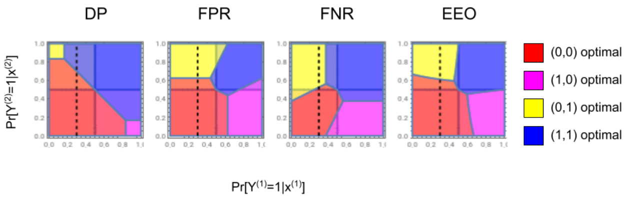

Figure 4: Fixing $p_a = 0.3$, examining how the property value changes as a function of p_b for different regularizers. Demographic parity results in different decisions only if $p_b \in [1/2, 3/4]$, FPR if $p_b \in [1/2, 2/3]$, FNR has essentially the same property values on the line $p_a = 0.3$, and EEO leads to a small region where optimal decisions change for $p_b \in [1/2, 2/3]$.

Figure 9 (R), this is supported by a higher range of EEO violations in the regularized models.

6 DISCUSSION AND CONCLUSION

In this work, we extend the notion of property elicitation to consider regularized loss functions, and give a necessary and sufficient condition on a regularizer to be equivalent to the original property. We apply this condition to demonstrate the (non-)equivalence of properties with a handful of regularizers common in the fair machine learning literature. Finally, we show how the choice and weight of regularization function can change decision-making on synthetic data as well as the German lending and heart disease risk datasets.

Limitations and considerations. The main intent of this work is to provide conceptual insight about how fairness regularizers change algorithmic decision-making and predictions. The insights provided rely on the hypothesis class being sufficiently expressive, and should not be solely used to justify the use of a regularizer. The addition of a regularizer and insights given are agnostic to the data itself and therefore agnostic to pre-processing and post-processing of data. Additional pre- or post-processing of the data may change the elicited property, though we leave this to future work.

Future work. There are many directions for future work. This work serves as a proof of concept for the extension of property elicitation to accommodate regularization functions, demonstrated on a handful of regularizers, but applying the necessary and sufficient condition on the equivalence of properties under different regularizers and more general prediction tasks remains an open direction of work. Moreover, it is important to understand how model complexity as well as pre- and post-processing of data can affect results.

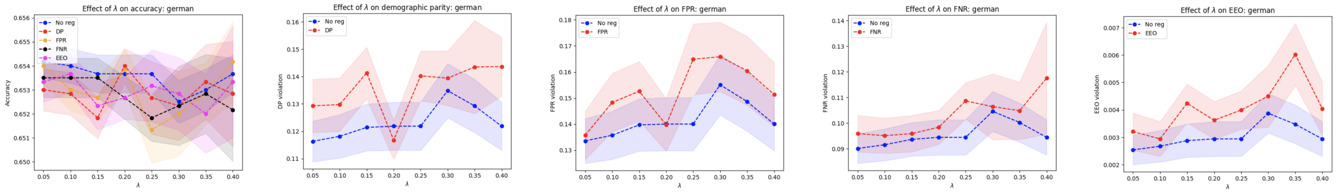

Figure 5: Effect of λ on regularizer values on the German lending dataset [21]. Because the (p_a, p_b) point summarizing group differences in the dataset are at a point where regularized decisions are the same as unregluarized decisions, it is unsurprising that regularizers do not significantly reduce unfairness, regardless of λ.

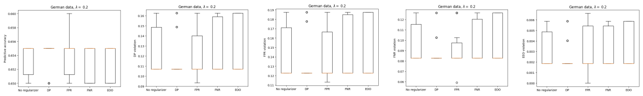

Figure 6: Distributions of accuracy and fairness violations in lending data. In general, it seems the models are tending to make similar predictions, which often nearly equal medians.

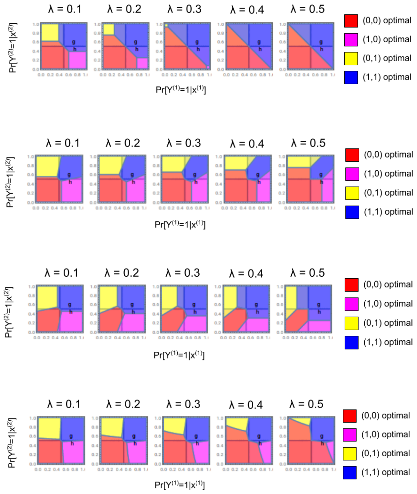

Figure 7: The level sets of different regularized properties as λ changes. (Top to bottom: DP, FPR, FNR, EEO). The g represents the "average" members of each group in the German lending dataset, and h the heart disease risk dataset.

7 ETHICAL CONSIDERATIONS, POSITIONALITY, AND ADVERSE IMPACT

Ethical considerations. While this paper is theoretical in nature, we hope it provides some helpful first steps in evaluating the different decisions recommended by algorithms subject to different fairness criterion. Being largely theoretical, the choice of datasets in the experimental section were based largely on three criterion: (1) public access, for reproducibility, (2) relevant domains where the FAccT community has implemented fairness-constrained algorithms, and (3) the underlying data distributions do eventually lead to some change in decision for some regularizers. To this third point, there is some merit to verifying that decisions *do not* change when they are not supposed to, but we view the main point of the experimental section as giving a proof of concept that fairness increases roughly in line with the regularized properties, even in imperfect circumstances.

Moreover, we view this work as an initial step towards understanding *how* fairness considerations in algorithm development change decision-making, and are interested to further see if lessons from this work can be shared with human-in-the-loop decision-makers to help them make the most informed decisions possible. In its current form, the theoretical nature of this work is limited in understanding how human decision-makers will use this information.

Positionality statement. The authors are white women based at universities in North America. As such, we occupy identities that are often seen as both the "advantaged" and "disadvantaged" groups in group fairness codifications. As algorithms might "fairness gerrymander," we are some of the most likely beneficiaries of more favorable decision-making, depending on the choice of sensitive attributes. Academically, the authors' backgrounds lie historically in property elicitation and in the algorithmic consequences of objective function choice.

Adverse impacts. While we hope this is not the use case, this work provides a framework for understanding when different fairness regularizers change decision-making, disagree with each other, and conversely, agree with each other. One adverse impact of this might mislead practitioners to conclude that when regularizers agree often, the choice of regularizer is inconsequential.

Additionally, since this work focuses on one in-processing technique, practitioners might be inclined or encouraged to solely use this in-processing technique instead of additionally using pre- and post-processing to make algorithms more fair. We highly encourage this framework to be used in conjuction with pre- and post-processing techniques.

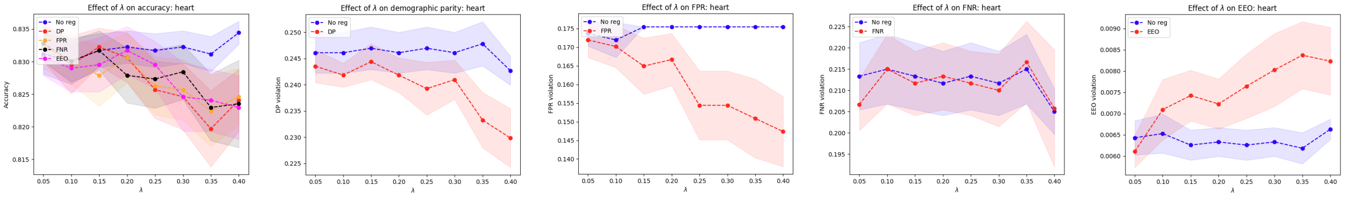

Figure 8: Effect of λ on regularizer values on the heart disease risk dataset [32].

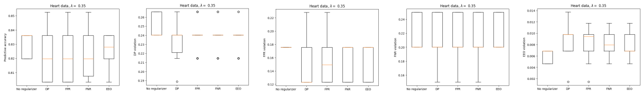

Figure 9: Distributions of accuracy and fairness violations in heart disease data. In general, it seems the models are tending to make similar predictions, which often nearly equal medians.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Award No. 2202898. Thanks to Yiling Chen, Francisco Marmolejo Cossio, Esther Rolf, and Arpita Biswas for feedback and comments, as well as participants in the EC Gender Inclusion Workshop.

REFERENCES

[1] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*. PMLR, 120–129.

[2] G Arutjothi and C Senthamarai. 2017. Prediction of loan status in commercial bank using machine learning classifier. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 416–419.

[3] Yahav Bechavod and Katrina Ligett. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044* (2017).

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[5] Jack Blandin and Ian Kash. 2022. Fairness Over Utilities Via Multi-Objective Rewards. (2022).

[6] Glenn W Brier et al. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78, 1 (1950), 1–3.

[7] Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. 2021. Fairness guarantee in multi-class classification. *arXiv preprint arXiv:2109.13642* (2021).

[8] Hyungrok Do, Preston Putzel, Axel S Martin, Padhraic Smyth, and Judy Zhong. 2022. Fair Generalized Linear Models with a Convex Penalty. In *International Conference on Machine Learning*. PMLR, 5286–5308.

[9] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems* 31 (2018).

[10] Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. 2019. An Embedding Framework for Consistent Polyhedral Surrogates. https://doi.org/10.48550/ARXIV.1907.07330

[11] Tobias Fissler. 2017. *On higher order elicitability and some limit theorems on the Poisson and Wiener space.* Ph. D. Dissertation.

[12] Rafael Frongillo and Ian Kash. 2014. General truthfulness characterizations via convex analysis. In *Web and Internet Economics*. Springer, 354–370.

[13] Rafael M. Frongillo and Ian A. Kash. 2019. General Truthfulness Characterizations Via Convex Analysis. arXiv:1211.3043 [cs.GT]

[14] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[15] Benjamin A Goldstein, Ann Marie Navar, and Rickey E Carter. 2017. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal* 38, 23 (2017), 1805–1814.

[16] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[17] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 1939–1948. https://proceedings.mlr.press/v80/hebert-johnson18a.html

[18] Lingxiao Huang and Nisheeth Vishnoi. 2019. Stable and Fair Classification. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2879–2890. https://proceedings.mlr.press/v97/huang19e.html

[19] Christopher Jung, Sampath Kannan, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. 2020. Fair prediction with endogenous behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation*. 677–678.

[20] Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. 2021. Moment Multicalibration for Uncertainty Estimation. In *Proceedings of Thirty Fourth Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 134)*, Mikhail Belkin and Samory Kpotufe (Eds.). PMLR, 2634–2678. https://proceedings.mlr.press/v134/jung21a.html

[21] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*. IEEE, 1–6.

[22] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 35–50.

[23] Nikola Konstantinov and Christoph H. Lampert. 2021. Fairness Through Regularization for Learning to Rank. *CoRR* abs/2102.05996 (2021). arXiv:2102.05996 https://arxiv.org/abs/2102.05996

[24] Amanda R Kube, Sanmay Das, and Patrick J Fowler. 2023. Community- and data-driven homelessness prevention and service delivery: Optimizing for equity. *Journal of the American Medical Informatics Association* 30, 6 (04 2023), 1032–1041. https://doi.org/10.1093/jamia/ocad052

[25] Nicolas S. Lambert. 2018. Elicitation and Evaluation of Statistical Forecasts. (2018). https://web.stanford.edu/~nlambert/papers/elicitability.pdf

[26] Nicolas S. Lambert, David M. Pennock, and Yoav Shoham. 2008. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*. 129–138.

[27] Nicolas S. Lambert and Yoav Shoham. 2009. Eliciting truthful answers to multiple-choice questions. In *Proceedings of the 10th ACM conference on Electronic commerce*. 109–118.

[28] Donald M Lloyd-Jones. 2010. Cardiovascular risk prediction: basic concepts, current status, and future directions. *Circulation* 121, 15 (2010), 1768–1777.

[29] Fatemehsadat Mireshghallah, Huseyin A. Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert Sim. 2021. Privacy Regularization: Joint Privacy-Utility Optimization in Language Models. arXiv:2103.07567 [cs.LG]

[30] Georgy Noarov and Aaron Roth. 2023. The Statistical Scope of Multicalibration. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 26283–26310. https://proceedings.mlr.press/v202/noarov23a.html

[31] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30 (2017).

[32] Rashik Rahman. 2021. Heart Attack Analysis and Prediction Dataset. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

[33] Leonard J Savage. 1971. Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* 66, 336 (1971), 783–801.

[34] Mohammad Ahmad Sheikh, Amit Kumar Goel, and Tapas Kumar. 2020. An approach for prediction of loan approval using machine learning algorithm. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 490–494.

[35] Vishal Singh, Ayushman Yadav, Rajat Awasthi, and Guide N Partheeban. 2021. Prediction of modernized loan approval system based on machine learning approach. In *2021 International Conference on Intelligent Technologies (CONIT)*. IEEE, 1–4.

[36] Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. 2014. Elicitation and Identification of Properties. In *Proceedings of The 27th Conference on Learning Theory*. 482–526.

[37] Zeyu Tang and Kun Zhang. 2022. Attainability and optimality: The equalized odds fairness revisited. In *Conference on Causal Learning and Reasoning*. PMLR, 754–786.

[38] Robert Williamson and Aditya Menon. 2019. Fairness risk measures. In *International Conference on Machine Learning*. PMLR, 6786–6797.

[39] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference*. 3356–3362.

[40] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*. PMLR, 962–970.

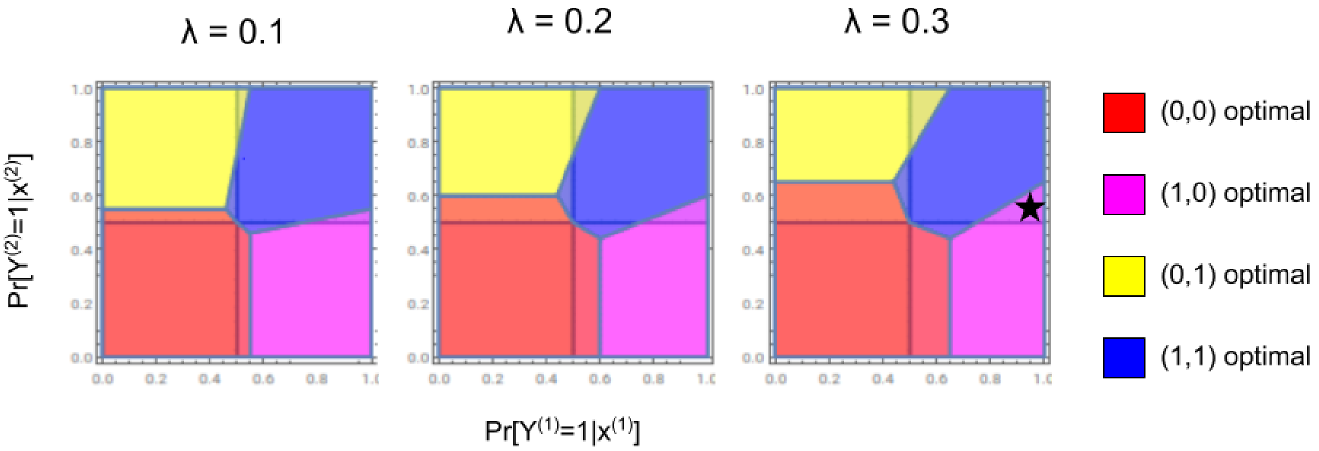

Figure 10: Visualizing the level sets of the FNR-regularized property $\Theta^{FNR,\lambda}$ for different values of $\lambda \in [0,1]$, where $m = 2$ and $s = (a,b)$. Each point $(p^{(1)}, p^{(2)})$ in a square represents $(\Pr_{p^{(1)}}[Y = 1], \Pr_{p^{(2)}}[Y = 1])$, and each colored cell represents sets of $(p^{(1)}, p^{(2)})$ pairs such that the optimal treatment is the same for all points in the cell. For example, the magenta cell is the set of distributions where the decision-maker prefers to attribute the positive treatment ($t^{(i)} = 1$) to the agent in group a, and the negative treatment ($t^{(i)} = 0$) to the agent in group b.

A ADDITIONAL EXAMPLE OF NON-EQUIVALENT PROPERTIES

A.1 Equalized FNR

Similarly, we consider false negative rates. Our objective is

$$L^{FNR,\lambda}(\mathbf{t};\mathbf{s};\mathbf{p}) = \frac{1}{m}\sum_i L(t^{(i)}, p^{(i)}) + \lambda \left| \frac{1}{n_a} \sum_{i:s^{(i)}=a, t^{(i)}=0} p^{(i)} - \frac{1}{n_b} \sum_{i:s^{(i)}=b, t^{(i)}=0} p^{(i)} \right|$$

Like the FPR regularizer, since the FNR regularizer computes the difference of false negative rates between groups, one can observe that a way to reduce the false negative rate of a group is to assign more positive treatments $t^{(i)} = 1$. Again, we see in figure 10 that the FNR regularizer then makes it worse for an algorithm to assign the negative treatment to an agent i even if $p^{(i)}$ slightly less than 1/2.

Corollary 5. *Let $L : [0,1] \times \{0,1\} \to [0,1]$ and $\psi : r \mapsto \mathbf{1}\{r \geq 1/2\}$ indirectly elicit the mode over $\mathcal{Y} = \{0,1\}$ such that $L(y,y) = 0$, and let $\Theta^{FPR,\lambda} := \psi \circ \text{prop}[L^{FPR,\lambda}]$. Then $\Theta^{FPR,\lambda}$ is not equivalent to the mode for $\lambda > 0$.*

Received January 2024