Bachelor Thesis in Computer Science

Fairness Behind a Veil of Ignorance: A Welfare Analysis for Supervised Learning

Learning & Adaptive Systems Group - ***ETH***

Author:
Claudio Ferrari
ferraric@ethz.ch

Supervisors:

Dr. Hoda Heidari
Prof. Dr. Andreas Krause

June 26, 2018

Contents

1. Overview

This thesis represents an extension of the paper "Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making" by Heidari et al. [20]. As a result of that, most of it is included in unchanged form, with consent from Dr. Hoda Heidari.

Traditionally, data-driven decision making systems have been designed with the sole purpose of maximizing some system-wide measure of performance, such as accuracy or revenue. Today, these systems are increasingly employed to make consequential decisions for human subjects—examples include employment [26], credit lending [28], policing [32], and criminal justice [5]. Decisions made in this fashion have long-lasting impact on people's lives and—absent a careful ethical analysis—may affect certain individuals or social groups negatively [36, 3, 24]. This realization has recently spawned an active area of research into quantifying and guaranteeing fairness for machine learning [14, 22, 17].

Virtually all existing formulations of algorithmic fairness focus on guaranteeing *equality* of some notion of *benefit* across different individuals or socially salient groups. For instance, demographic parity [14, 22, 8] seeks to equalize the percentage of people receiving the positive outcome across different groups. Equality of opportunity [17] requires the equality of false positive/false negative rates. Individual fairness [14] demands that people who are equal with respect to the task at hand receive equal outcomes. In essence, the debate so far has mostly revolved around identifying the right notion of benefit and a tractable mathematical formulation for equalizing it.

The view of fairness as some form of equality is indeed an important consideration in evaluating algorithmic decision making systems from a moral standpoint—people would compare their decision outcomes with other similarly situated individuals, and these *interpersonal comparisons* must be key in shaping our ethical judgement of the system. However, in this work, we argue that equality is not the only factor at play that is worth an in-depth investigation. Here, we draw attention to two important, yet largely overlooked aspects when evaluating fairness of automated decision making systems—namely *risk* and *welfare* considerations, illustrated in Section 2.1 Our proposed family of measures corresponds to the long-established formulations of cardinal social welfare in economics. We come to this proposal by taking the perspective of a rational, risk-averse individual who is going to be subject to algorithmic decision making and is faced with the task of choosing between several algorithmic alternatives *behind a Rawlsian veil of ignorance* (see Section 2.1). The convex formulation of our measures allows us to integrate them as a constraint into any convex loss minimization pipeline. We formulate an optimization procedure for both individual-level fairness (Section 2) and group-level fairness (Section 3). Our empirical analysis reveals interesting trade-offs between our proposal and (a) prediction accuracy, (b) group discrimination, and (c) Dwork et al.'s notion of individual fairness. Furthermore and perhaps most importantly, our work provides both theoretical and empirical evidence suggesting that a lower-bound on our measures often leads to bounded inequality in algorithmic outcomes; hence presenting the first computationally feasible mechanism for bounding individual-level (un)fairness.

In Section 4 we propose a heuristic to make Dwork et al.'s notion of individual fairness tractable as another possible way for ensuring individual fairness for reference. Empirical results show that while unfairness indeed decreases, that approach cannot be used to increase fairness beyond a certain threshold.

1.1. Related Work

Much of the existing work on algorithmic fairness has been devoted to the study of *discrimination* (also called *statistical*- or *group*-level fairness by the machine learning community). Group fairness notions require that given a classifier, a certain fairness metric is equal across all protected groups. Different choices for the metric have led to different naming of the corresponding fairness notions (see e.g., demographic parity [22, 14, 8], disparate impact [38, 15], equality of opportunity [17], and calibration [22]). Statistical notions of fairness fail to guarantee fairness at the individual level.

Dwork et al. [14] first formalized the notion of individual fairness for classification learning tasks, requiring that two individuals who are similar with respect to the task at hand receive similar classification outcomes. The definition relies on the existence of a suitable similarity metric between individuals and is computationally prohibitive to solve exactly—it requires adding $O(n^2)$ constraints to the loss minimization program, where n is the number of individuals in the training set. Furthermore, as pointed out by Speicher et al. [35], the formulation does not take into account the variation in *social desirability* of various outcomes and people's merit for different decisions. More recently, Speicher et al. [35] proposed a new measure for quantifying individual unfairness, utilizing income *inequality indices* from economics (in particular, they propose the use of generalized entropy) and applying them to algorithmic benefit distributions. Both of these formulations focus solely on the *inter-personal* comparison of algorithmic outcomes across individuals and do not account for *risk* and *welfare* considerations. Moreover, there do not exist efficient, exact mechanisms for bounding either of these formulations.

Zafar et al. [39] recently proposed two preference-based notions of fairness at the group-level, called *preferred treatment* and *preferred impact*. A group-conditional classifier satisfies preferred treatment if no group collectively prefers another group's classifier to their own (in terms of average misclassification rate). This definition is based on the notion of *envy-freeness* [37] in economics and applies to group-conditional classifiers only. A classifier satisfies preferred impact if it Pareto-dominates an existing impact parity classifier (i.e. every group is better off using the former classifier compared to the latter). Pareto-dominance (to be defined precisely in Section 2.2) leads to a *partial* ordering among alternatives and usually in practice, does not have much bite. Similar to [39], our work can be thought of as a preference-based notions of fairness, but unlike their proposal our measures lead to a *total* ordering among all available algorithmic alternatives, and is applicable to quantify both individual and group-level (un)fairness.

Also related to our work is [8], where authors propose maximizing an objective called "immediate utility" while satisfying existing fairness constraints. Immediate utility is meant to capture the impact of a decision rule on the society (e.g. on public safety when the task is to predict recidivism), and is composed of two terms: the first term is the expected number of true positives under the decision rule (e.g. number of crimes prevented), and the second term is the expected cost of positive labels (e.g. cost of detention). Note that our proposal is conceptually different from immediate utility in that we are concerned with the utility an individual derives as the result of being *subject* to algorithmic decision making, whereas immediate utility captures the impact of these decisions on the society/system as a whole. For example, while it might be beneficial from the perspective of a high-risk defendant to be released, the societal cost of releasing him/her into the community is regarded as high. Furthermore and from a normative perspective, immediate utility is proposed as a replacement for prediction accuracy, whereas our measures are meant to capture desirability of algorithmic outcomes from the perspective of individuals subject to it.

Several papers in economics have studied the relationship between inequality aversion and risk aversion. At a high level, the larger the relative risk aversion is, the more an individual choosing between different societies behind a "veil of ignorance" will be willing to trade-off expected benefit in order to achieve a more equal distribution. The following papers attempt to further clarify the link between evaluating risk ex ante and evaluating inequality ex post: Cowell and Schokkaert [10] and Carlsson et al. [7] empirically measure individuals' perceptions and preferences for risk and inequality through human-subject experiments. Amiel and Cowell [2] establish a general relationship between the standard form of the social-welfare function and the "reduced-form" version that is expressed in terms of inequality and mean income.

2. Social Welfare as a Measure of Individual Fairness

2.1. Motivation

We try to highlight the role of risk and welfare considerations in the context of fairness in automated decision making. The importance of these issues is perhaps best illustrated via a simple example: Suppose we have four decision making models A, B, C, D each resulting in a different benefit distribution across 5 groups/individuals i_1, i_2, i_3, i_4, i_5 (we will precisely define in Section 2.2 how benefits are computed, but for the time being, suppose benefits are equivalent to salary predictions made through different regression models). Figure 1 illustrates the setting. Suppose one is tasked with determining which one of these alternatives is *ethically more desirable*. From an inequality minimizing perspective, A is clearly more desirable than B (note that both A, B result in the same total benefit of 4, and A distributes it equally across i_1, ..., i_5). With a similar reasoning, C is preferred to D. Notice, however, that by focusing on equality alone, one would also deem A more desirable than D, but there is a problem with this judgement: almost everyone (expect for i_1 who sees a negligible drop of less than 2% in their benefit) is significantly better off under D compared to A. In other words, even though D results in unequal benefits and it does *not* Pareto-dominate A, collectively it results in higher *welfare* and *lower risk*, and therefore, both intuitively and from a *rational* point of view it should

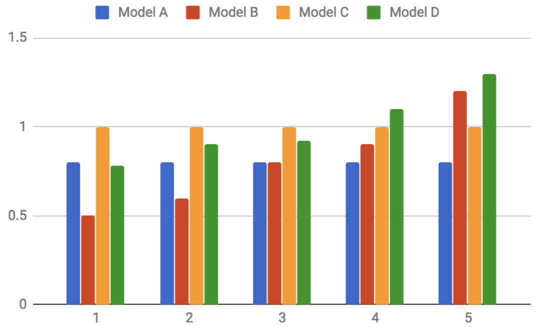


Figure 1: Model A assigns the same benefit of 0.8 to everyone; model C assigns the same benefit of 1 to everyone; model B results in benefits $(0.5, 0.6, 0.8, 0.9, 1.2)$, and model D, $(0.78, 0.9, 0.92, 1.1, 1.3)$. Our proposed measures prefer A to B, C to D, and D to A.

be considered more desirable. With a similar reasoning, one would conclude C is more desirable than A, even though both provide benefits equally to all individuals. In light of this example and inspired by the long line of research on distributive justice in economics, in this work we propose a natural family of measures for evaluating algorithmic fairness corresponding to the well-studied notions of *cardinal social welfare* in economics [18, 19]. Our proposed measures indeed prefer A to B, C to D, and D to A.

The interpretation of social welfare as a measure of fairness is justified by the Rawlsian theory of justice [30] and through the concept of *veil of ignorance*. Consider the following thought experiment—originally proposed by John Rawls and recast here to reflect our setting of interest: imagine an individual who knows nothing about the particular position they will be born in within the society, and is asked to select among a set of algorithmic alternatives (e.g. to choose whether salary predictions are made using neural networks or decision trees). In this hypothetical original/ex-ante position, if the individual is *rational*, they would aim to minimize risk and insure against unlucky events in which they turn out to assume the position of a low-benefit individual. In the example above, if one is to choose between models A, D without knowing which one of the 5 individuals they will be, then the risk associated with alternative D is much less than that of A—under A the individual is going to receive a (relatively low) benefit of 0.8 with certainty, whereas under D with high probability (i.e. 4/5) they gain a (relatively large) benefit of 0.9 or more, and with low probability (1/5) they receive a benefit of 0.78, roughly the same as the level of benefit they would attain under A. Such considerations of risk is precisely what our proposal seeks to quantify. Our core idea consists of computing the expected *utility* of a randomly chosen, *risk-averse* individual as the result of being subject to algorithmic decision making.

We remark that in comparing two benefit distributions of the *same mean* (e.g. A, B or C, D in our earlier example), our measures always prefer the more equal one (A is preferred to B and C is preferred to D). See Proposition 2 for the formal statement. Thus, our measures are inherently equality preferring. However, the key advantage of our measures of social welfare over those focusing on inequality manifests when, as we saw in the above example, comparing two benefit distributions of different means. In such conditions, inequality based measures are insufficient and may result in misleading conclusions, while measures of social welfare are better suited to identify the fairest (most desirable) alternative. When comparing two benefit distributions of the same mean, social welfare and inequality would always yield identical conclusions.

2.2. Our Proposed Family of Measures

We consider the standard supervised learning setting: A learning algorithm receives the training data set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ consisting of n instances, where $\mathbf{x}_i \in \mathcal{X}$ specifies the feature vector for individual i and $y_i \in \mathcal{Y}$, the **ground truth** label for him/her. The training data is sampled i.i.d. from a distribution P on $\mathcal{X} \times \mathcal{Y}$. Unless specified otherwise, we assume $\mathcal{X} \subseteq \mathbb{R}^k$, where k denotes the number of features. To avoid introducing extra notation for an intercept, we assume feature vectors are in homogeneous form, i.e. the k-th feature value is 1 for every instance. The goal of a learning algorithm is to use the training data to fit a *model* (or hypothesis) $h : \mathcal{X} \rightarrow \mathcal{Y}$ that accurately predicts the label for new instances. Let $\mathcal{H}$ be the hypothesis class consisting of all the models the learning algorithm can choose from. A learning algorithm receives D as the input; then utilizes the data to select a model $h \in \mathcal{H}$ that minimizes some notion of loss. For instance, in classification the (0-1) loss of a model h on the training data D is defined as $\sum_{i=1}^{n} \mathbf{1}[y_i \neq \hat{y}_i]$, where $\hat{y}_i = h(\mathbf{x}_i)$. The learning algorithm outputs $h^* \in \mathcal{H}$ that minimizes the empirical loss; i.e., $h^* = \arg\min_h L_D(h)$.
We assume there exists a benefit function $b : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ that quantifies the benefit an individual with ground truth label y receives, if the trained model assigns label $\hat{y}$ to them.[1] The benefit function is meant to capture the **signed discrepancy** between an individual's assigned outcome and their true/deserved outcome. Throughout, for simplicity we assume higher values of $\hat{y}$ correspond to more desirable outcomes (e.g. loan or salary amount). With this assumption in place, a benefit function must assign a high value to an individual if their assigned label is greater (better) than their deserved label, and a low value if an individual receives a label less (worse) than their deserved label. The following are a few examples of benefit functions that satisfy this: $b(y, \hat{y}) = \hat{y} - y$; $b(y, \hat{y}) = log\left(1 + e^{\hat{y}-y}\right)$; $b(y, \hat{y}) = \hat{y}/y$. In order to maintain the convexity of our fairness constraints, throughout this work, we will focus on benefit functions that are positive and affine in $\hat{y}$. For binary classification, this restriction is without loss of any generality:

Proposition 1 *For $y, \hat{y} \in \{0, 1\}$, let $\bar{b}_{y,\hat{y}} \in \mathbb{R}$ be arbitrary constants specifying the benefit an individual with ground truth label y receives when they are assigned label $\hat{y}$. Then there exists a linear benefit function of form $c_y\hat{y} + d_y$ such that for all $y, \hat{y} \in \{0, 1\}$, $b(y, \hat{y}) = \bar{b}_{y,\hat{y}}$.*

[1] Note that the benefit function can potentially depend on $\mathbf{x}$ and other available information about the individual—as long the formulation is linear in the predicted label, our approach remains effective. For simplicity and ease of interpretation, however, we will focus on benefit functions that depend on y and $\hat{y}$, only.

Proof Solving the following system of equations,

$$\forall y, \hat{y} \in \{0,1\} : c_y \hat{y} + d_y = \bar{b}_{y,\hat{y}}$$

we obtain: $c_0 = \bar{b}_{0,1} - \bar{b}_{0,0}$, $c_1 = \bar{b}_{1,1} - \bar{b}_{1,0}$, $d_0 = \bar{b}_{0,0}$, and $d_1 = \bar{b}_{1,0}$. ∎

Of course, for $\bar{b}$'s in the above proposition to reflect the signed discrepancy between y and $\hat{y}$, it must hold that $\bar{b}_{1,0} < \bar{b}_{0,0} \leq \bar{b}_{1,1} < \bar{b}_{0,1}$. Given a decision making algorithm we can compute its corresponding benefit profile $\mathbf{b} = (b_1, \cdots, b_n)$ where b_i denotes individual i's benefit. A benefit profile $\mathbf{b}$ *Pareto-dominates* $\mathbf{b}'$ (or in short $\mathbf{b} \succeq \mathbf{b}'$), if for all $i = 1, \cdots, n$, $b_i \geq b'_i$.

Following the economic models of risk attitude, we assume the existence of a utility function $u : \mathbb{R} \rightarrow \mathbb{R}$, where $u(b)$ represent the utility derived from algorithmic benefit b. We will focus on *Constant Relative Risk Aversion (CRRA)* utility functions. In particular, we take $u(b) = (sign(\alpha) \cdot b)^{\alpha}$ where $\alpha = 1$ corresponds to risk-neutral, $\alpha > 1$ corresponds to risk-seeking, and $\alpha < 1$ corresponds to risk-averse preferences. Our main focus in this work is on values of $\alpha < 1$: the larger one's initial algorithmic benefit is, the smaller the added utility he/she derives from an increase in his/her benefit. While in principle our model can allow for different risk parameters for different individuals (α_i for individual i), for simplicity throughout we assume all individuals have the same risk parameter. Our measures assess the fairness of a decision making model via the expected *utility* a randomly chosen, *risk-averse* individual receives as the result of being subject to decision making through that model. Formally, our measure is defined as follows: $\mathcal{U}_P(h) = \mathbb{E}_{(\mathbf{x}_i,y_i)\sim P}\left[u\left(b(y_i, h(\mathbf{x}_i)\right)\right]$. We estimate this expectation by $\mathcal{U}_D(h) = \frac{1}{n}\sum_{i=1}^{n} u(b(y_i, h(\mathbf{x}_i)))$. To guarantee fairness, we propose minimizing loss ($L_D(h)$) subject to $\mathcal{U}_D(h) \geq \tau$, where the parameter τ specifies a lower bound on our measure. For instance, when the learning task is linear regression[2], $b(y, \hat{y}) = \hat{y} - y + 1$, and the degree of risk aversion in α, this optimization amounts to:

$$\begin{aligned} \min_{\boldsymbol{\theta} \in \mathcal{H}} \quad & \sum_{i=1}^{n} (\boldsymbol{\theta}.\mathbf{x}_i - y_i)^2 \\ \text{s.t.} \quad & \sum_{i=1}^{n} (sign(\alpha) \cdot (\boldsymbol{\theta}.\mathbf{x}_i - y_i + 1))^{\alpha} \geq \tau n \end{aligned} \tag{1}$$

Note that both the objective function and the constraint in (1) are convex functions of $\boldsymbol{\theta}$, therefore, the optimization can be solved efficiently and exactly.

Connection to Cardinal Welfare Our proposed family of measures corresponds to a particular subset of cardinal social welfare functions. At a high level, a cardinal social welfare function is meant to rank different distributions of welfare across individuals, as more or less desirable in terms of distributive justice [27]. More precisely, let $\mathcal{W}$ be a

[2] The problem formulation for classification can be found in Appendix A

welfare function defined over benefit vectors, such that given any two benefit vectors $\mathbf{b}$ and $\mathbf{b}'$, $\mathbf{b}$ is considered more desirable than $\mathbf{b}'$ if and only if $\mathcal{W}(\mathbf{b}) \geq \mathcal{W}(\mathbf{b}')$. The rich body of work on welfare economics offers several axioms to characterize the set of all welfare functions that pertain to collective rationality or fairness. Any such function, $\mathcal{W}$, must satisfy the following axioms [34, 31]:

1. **Monotonicity:** If $\mathbf{b}' \succ \mathbf{b}$, then $\mathcal{W}(\mathbf{b}') > \mathcal{W}(\mathbf{b})$. If one individual's benefit increases while everyone else's benefits remain at the same level, $\mathcal{W}$ should strictly prefer the latter distribution. Monotonicity is closely related to Pareto-efficiency.
2. **Symmetry:** $\mathcal{W}(b_1, \ldots, b_n) = \mathcal{W}\left(b_{(1)}, \cdots, b_{(n)}\right)$. In other words, $\mathcal{W}$ does not depend on the identity of the individuals, only their benefit levels.
3. **Independence of unconcerned agents:** $\mathcal{W}$ should be independent of individuals whose benefits remain at the same level. Formally, let $(\mathbf{b}|^i a)$ be a benefit vector that is identical to $\mathbf{b}$, expect for the ith component which has been replaced by a. The property requires that for all $\mathbf{b}, \mathbf{b}', a, c$, $(\mathbf{b}|^i a) \succeq (\mathbf{b}'|^i a) \Leftrightarrow (\mathbf{b}|^i c) \succeq (\mathbf{b}'|^i c)$.

It has been shown that every *continuous*[3] social welfare function $\mathcal{W}$ with properties 1–3 is additive and can be represented as $\sum_{i=1}^{n} w(b_i)$. According to the Debreu-Gorman theorem [13, 16], if in addition to 1–3, $\mathcal{W}$ satisfies:

4. **Independence of common scale:** For any $c > 0$, $\mathcal{W}(\mathbf{b}) \geq \mathcal{W}(\mathbf{b}') \Leftrightarrow \mathcal{W}(c\mathbf{b}) \geq \mathcal{W}(c\mathbf{b}')$. The simultaneous rescaling of every individual benefit in $\mathbf{b}, \mathbf{b}'$, should not affect their relative order.

then it belongs to the following one-parameter family: $\mathcal{W}_\alpha(b_1, \ldots, b_n) = \sum_{i=1}^{n} w_\alpha(b_i)$, where (a) for $\alpha > 0$, $w_\alpha(b) = b^\alpha$ (b) for $\alpha = 0$, $w_\alpha(b) = \ln(b)$; and (c) for $\alpha < 0$, $w_\alpha(b) = -b^\alpha$—note that the limiting case of $\alpha \rightarrow -\infty$ is equivalent the leximin ordering (or Rawlsian max-min welfare).

Our focus in this work is on $\alpha < 1$, because for this choice of parameters, our measures exhibit aversion to pure inequality. More precisely, they satisfy the following important property:

5. **Pigou-Dalton transfer principle [29, 12]:** Transferring benefit from a high-benefit to a low-benefit individual must increase social welfare, that is, for any $1 \leq i < j \leq n$ and $0 < \delta < \frac{b_{(j)} - b_{(i)}}{2}$, $\mathcal{W}(b_{(1)}, \cdots, b_{(i)} + \delta, \cdots, b_{(j)} - \delta, \cdots, b_{(n)}) > \mathcal{W}(\mathbf{b})$.

Connection to Inequality Measures Speicher et al. [35] recently proposed quantifying individual-level unfairness utilizing income *inequality indices.* Their proposed index—generalized entropy—satisfies four important axioms: symmetry, population invariance, 0-normalization[4], and the Pigou-Dalton transfer principle. Our measures satisfy all

[3] That is, for every vector $\mathbf{b}$, the set of vectors weakly better than $\mathbf{b}$ (i.e. $\{\mathbf{b}' : \mathbf{b}' \succeq \mathbf{b}\}$) and the set of vectors weakly worse than $\mathbf{b}$ (i.e. $\{\mathbf{b}' : \mathbf{b}' \preceq \mathbf{b}\}$) are closed sets.

[4] 0-normalization requires the inequality index to be 0 if and only if the distribution is perfectly equal/uniform.

the aforementioned axioms, except for 0-normalization. Additionally and in contrast with measures of inequality—where the goal is to capture interpersonal comparison of benefits—our measure is monotone and independent of unconcerned agents. The latter two are the fundamental properties that set our proposal apart from measures of inequality.

Next, we observe that under certain conditions, our proposed measure of fairness results in the same total ordering as the Atkinson's index [4]. Atkinson's inequality index is defined as:

$$A_\beta(b_1, \dots, b_n) = \begin{cases} 1 - \frac{1}{\mu}\left(\frac{1}{n}\sum_{i=1}^n b_i^{1-\beta}\right)^{1/(1-\beta)} & \text{for } 0 \leq \beta \neq 1 \\ 1 - \frac{1}{\mu}\left(\prod_{i=1}^n b_i\right)^{1/n} & \text{for } \beta = 1, \end{cases}$$

where $\mu = \frac{1}{n}\sum_{i=1}^n b_i$ is the mean benefit. Atkinson's inequality index is a *welfare*-based measure of inequality: The measure compares the actual average benefit individuals receive under benefit distribution $\mathbf{b}$ (i.e. μ) with its Equally Distributed Equivalent (EDE)—the level of benefit that if obtained by every individual, would result in the same level of welfare as that of $\mathbf{b}$ (i.e. $\frac{1}{n}\sum_{i=1}^n b_i^{1-\beta}$). It is easy to verify that for $0 < \alpha < 1$, the generalized entropy and Atkinson index result in the same total ordering among benefit distributions (see Proposition 3). Furthermore, for a fixed mean benefit μ, our proposed measure of fairness results in the same indifference curves and total ordering as the Atkinson index with $\beta = 1 - \alpha$.

Proposition 2 *Consider two benefit vectors* $\mathbf{b}, \mathbf{b}' \succ \mathbf{0}$ *with equal means* ($\mu = \mu'$). *For* $0 < \alpha < 1$, $A_{1-\alpha}(\mathbf{b}) \geq A_{1-\alpha}(\mathbf{b}')$ *if and only if* $\mathcal{W}_\alpha(\mathbf{b}) \leq \mathcal{W}_\alpha(\mathbf{b}')$.

Proof We have that:

$$\begin{aligned} A_{1-\alpha}(\mathbf{b}) \geq A_{1-\alpha}(\mathbf{b}') \quad &\Rightarrow \quad 1 - \frac{1}{\mu}\left(\frac{1}{n}\sum_{i=1}^n b_i^\alpha\right)^{1/\alpha} \geq 1 - \frac{1}{\mu'}\left(\frac{1}{n}\sum_{i=1}^n b_i'^\alpha\right)^{1/\alpha} \\ &\Leftrightarrow \quad \frac{1}{\mu}\left(\frac{1}{n}\sum_{i=1}^n b_i^\alpha\right)^{1/\alpha} \leq \frac{1}{\mu'}\left(\frac{1}{n}\sum_{i=1}^n b_i'^\alpha\right)^{1/\alpha} \\ &\Leftrightarrow \quad \left(\frac{1}{n}\sum_{i=1}^n b_i^\alpha\right)^{1/\alpha} \leq \left(\frac{1}{n}\sum_{i=1}^n b_i'^\alpha\right)^{1/\alpha} \\ &\Leftrightarrow \quad \sum_{i=1}^n b_i^\alpha \leq \sum_{i=1}^n b_i'^\alpha \\ &\Leftrightarrow \quad \mathcal{W}_\alpha(\mathbf{b}) \leq \mathcal{W}_\alpha(\mathbf{b}') \end{aligned}$$

■

Generalized entropy vs. Atkinson index Let $\mathcal{G}_\alpha(\mathbf{b})$ specify the generalized entropy, where

$$\mathcal{G}_\alpha(\mathbf{b}) = \frac{1}{n\alpha(\alpha-1)}\sum_{i=1}^n \left[\left(\frac{b_i}{\mu}\right)^\alpha - 1\right]$$

Proposition 3 *Suppose $0 < \alpha < 1$. For any two benefit distributions $\mathbf{b}, \mathbf{b}'$, $\mathcal{A}_{1-\alpha}(\mathbf{b}) \geq \mathcal{A}_{1-\alpha}(\mathbf{b}')$ if and only if $\mathcal{G}_\alpha(\mathbf{b}) \geq \mathcal{G}_\alpha(\mathbf{b}')$.*

Proof First note that for any distribution $\mathbf{b}$, $\mathcal{A}_{1-\alpha}(\mathbf{b}) = 1 - (\alpha(\alpha-1)\mathcal{G}_\alpha(\mathbf{b}) + 1)^{1/\alpha}$. We have that

$$\begin{aligned}
\mathcal{A}_{1-\alpha}(\mathbf{b}) \geq \mathcal{A}_{1-\alpha}(\mathbf{b}') \quad &\Leftrightarrow \quad 1 - \mathcal{A}_{1-\alpha}(\mathbf{b}) \leq 1 - \mathcal{A}_{1-\alpha}(\mathbf{b}') \\
&\Leftrightarrow \quad \alpha \ln\left(1 - \mathcal{A}_{1-\alpha}(\mathbf{b})\right) \leq \alpha \ln\left(1 - \mathcal{A}_{1-\alpha}(\mathbf{b}')\right) \\
&\Leftrightarrow \quad \ln\left(\alpha(\alpha-1)\mathcal{G}_\alpha(\mathbf{b}) + 1\right) \leq \ln\left(\alpha(\alpha-1)\mathcal{G}_\alpha(\mathbf{b}') + 1\right) \\
&\Leftrightarrow \quad \alpha(\alpha-1)\mathcal{G}_\alpha(\mathbf{b}) + 1 \leq \alpha(\alpha-1)\mathcal{G}_\alpha(\mathbf{b}') + 1 \\
&\Leftrightarrow \quad \mathcal{G}_\alpha(\mathbf{b}) \geq \mathcal{G}_\alpha(\mathbf{b}')
\end{aligned}$$

$\blacksquare$

Tradeoffs among Various Measures of Fairness Finally, we demonstrate the existence of multilateral tensions among accuracy, social welfare, individual, and statistical measures of (un)fairness by looking at the predictive models that would optimize each. See Table 1. In the realizable case, we assume the existence of a hypothesis $h^* \in \mathcal{H}$ such that $y = h^*(\mathbf{x})$ (perfect prediction accuracy); for the unrealizable case, we assume the existence of a hypothesis $h^* \in \mathcal{H}$, such that $h^*(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$). Also we define $y_{\max} = \max_{h \in \mathcal{H}, \mathbf{x} \in \mathcal{X}} h(\mathbf{x})$ and $y_{\min} = \min_{h \in \mathcal{H}, \mathbf{x} \in \mathcal{X}} h(\mathbf{x})$.

Table 1: Optimal predictions with respect to different notions of fairness

	Classification		Regression	
	Realizable	Unrealizable	Realizable	Unrealizable
Social welfare	$\hat{y} \equiv 1$	$\hat{y} \equiv 1$	$\hat{y} \equiv y_{\max}$	$\hat{y} \equiv y_{\max}$
Atkinson index	$\hat{y} = h^*(\mathbf{x})$	$\hat{y} \equiv 1$	$\hat{y} = h^*(\mathbf{x})$	$\hat{y} \equiv y_{\max}$
Dwork et al.	$\hat{y} \equiv 0$ or 1	$\hat{y} \equiv 1$ or 0	$\hat{y} \equiv c$	$\hat{y} \equiv c$
Mean diff.	$\hat{y} \equiv 0$ or 1	$\hat{y} \equiv 1$ or 0	$\hat{y} \equiv c$	$\hat{y} \equiv c$
Pos. res. diff.	$\hat{y} \equiv 0$ or $\hat{y} = h^*(\mathbf{x})$	$\hat{y} \equiv 0$	$\hat{y} \equiv y_{\min}$ or $\hat{y} = h^*(\mathbf{x})$	$\hat{y} \equiv y_{\min}$
Neg. res. diff.	$\hat{y} \equiv 1$ or $\hat{y} = h^*(\mathbf{x})$	$\hat{y} \equiv 1$	$\hat{y} \equiv y_{\max}$ or $\hat{y} = h^*(\mathbf{x})$	$\hat{y} \equiv y_{\max}$

To see the tradeoffs among various notions of fairness, and between accuracy and fairness, take the unrealizable classification case as an example. Optimizing for accuracy obviously requires the predictions to follow the Bayes optimal classifier. On the one hand, a lower bound on social welfare pulls the model in the direction of allocating the desirable outcome (i.e. 1) to a larger fraction of the population. On the other hand, to guarantee low positive residual difference, individuals should ideally all receive a label less than their ground truth (e.g. 0).

2.3. Experiments

In this section, we illustrate the trade-offs between our family of measures and accuracy, as well as existing definitions of group discrimination and individual fairness. Details on how the different measures were computed can be found in Appendix A. All experiments in this work were done on Propublica's *COMPAS classification dataset* [23].[5] We ran logistic regression and the benefit function was defined as follows: $b(y, \hat{y}) = c_y\hat{y} + d_y$ where $y \in \{-1, 1\}$, $c_1 = 0.5, d_1 = 0.5$, and $c_{-1} = 0.25, d_{-1} = 1.25$. This results in benefit levels 0 (false negative), 1 (true positive and true negative), and 1.5 (false positives).

Welfare as a Measure of Fairness We start by illustrating that our proposed measures can be applied to compare and rank different algorithmic alternatives. We trained the following models: a multi-layered perceptron, fully connected with one hidden layer with 100 units (NN), the AdaBoost classifier (Ada), Logistic Regression (LR), a decision tree classifier (Tree), a nearest neighbor classifier (KNN). Figure 2 illustrates how these learning models compare with one another according accuracy, Atkinson index, and social welfare. All values were computed using 20-fold cross validation. The confidence intervals are formed assuming samples come from Student's t distribution. As shown in Figure 2, the rankings obtained from Atkinson index and social welfare are identical. Note that this is consistent with Proposition 2. Given the fact that all models result in similar mean benefits, we expect the rankings to be consistent.

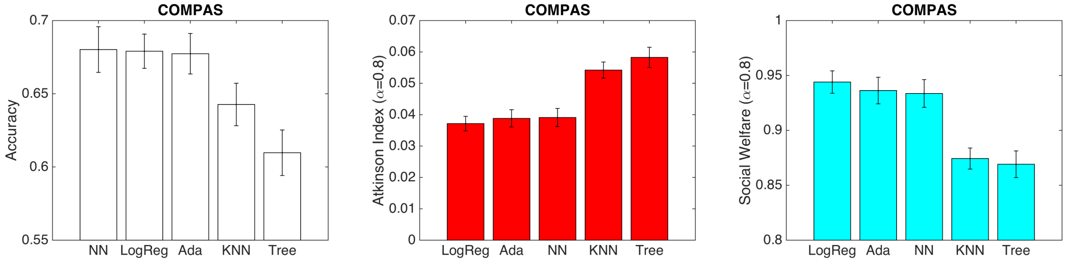
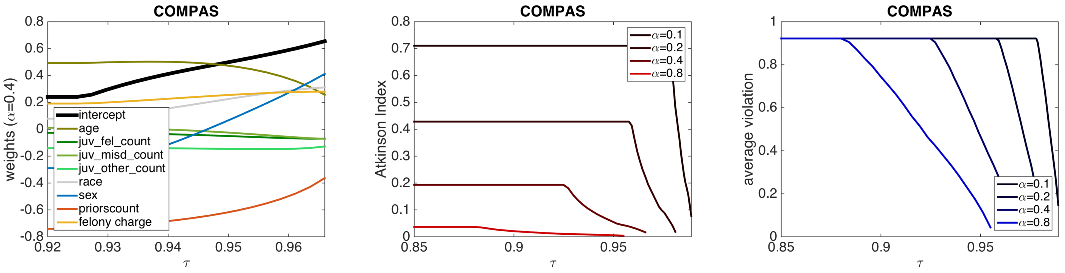


Figure 2: Comparison of different learning models according to accuracy, social welfare and Atkinson index. The mean benefits are 0.97 for LogReg, 0.96 for NN, 0.96 for AdaBoost, 0.89 for KNN, and 0.89 for Tree. Note that for the Atkinson measure, smaller values correspond to fairer outcomes, where as for social welfare larger values reflect greater fairness.

Impact on Model Parameters Next, we study the impact of changing τ on the trained model parameters (see Figure 3a). We observe that as τ increases, the intercept continually rises to guarantee high levels of benefit and social welfare. An interesting trend for the binary feature sex (0 is female, 1 is male) can be observed; initially being male has a negative weight and thus a negative impact on the classification outcome, but as τ is increased, the sign changes to positive to ensure men also get high benefits.

[5] A more detailed description of the data set and our preprocessing steps can be found in Appendix A.



(a) **(b)** **(c)**

Figure 3: (a) Changes in weights—$\boldsymbol{\theta}$ in logistic regression—as the function of τ. Note the continuous rise of the intercept with τ. (b) Atkinson index as a function of the threshold τ. Note the consistent decline in inequality as τ increases. (c) Dwork measure as a function of τ.

Trade-offs with Accuracy Here we illustrate the trade-offs between our proposed measure and prediction accuracy. As expected and evident in Figure 4, imposing more restrictive fairness constraints (larger τ and smaller α), results in higher loss of accuracy.

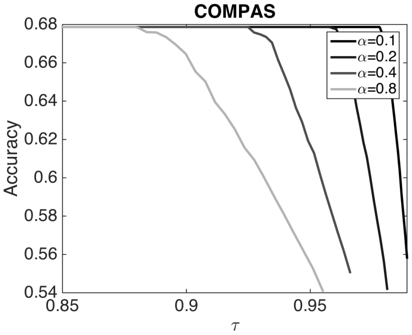
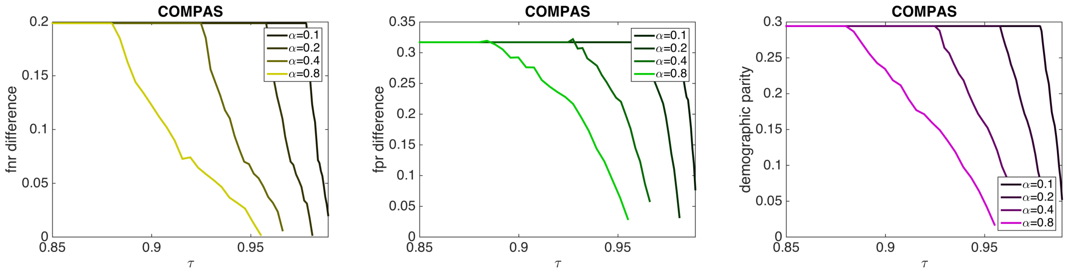


Figure 4: Accuracy as the function of τ for different values of α

Trade-offs with Individual Notions Figures 3b, 3c illustrate the impact of bounding our measure on existing individual measures of fairness. As expected, we observe that higher values of τ (i.e. social welfare) consistently result in lower inequality. Note that τ cannot be arbitrarily large (due to the infeasibility of achieving arbitrarily large social welfare levels). Also as expected, smaller α values (i.e. higher degrees of risk aversion) lead to a faster drop in inequality. The average violation of Dwork's pairwise constraints go down as τ increases until the measure reaches 0—which is what we expect the measure to amount to once almost every individual receives the positive label.

Trade-offs with Statistical Notions Next, we illustrate the impact of bounding our measure on statistical measures of fairness, such as demographic parity and difference between false positive/negative rates across groups. The groups were determined by the sensitive feature "race".

Figure 5a shows the impact of τ and α on false negative rate difference. As expected, the quantity decreases with τ until it reaches 0—once everyone receives a label at least as large as their ground truth. The trends are similar for false positive rate difference (Figure 5b). Figure 5c shows the impact of τ and α on demographic parity. Not surprisingly, also this measure goes to 0 as τ increases.



(a) **(b)** **(c)**

Figure 5: Group discrimination as a function of τ for different values of α. (a) Difference in false negative rates is decreasing with τ and approaches 0. (b) Difference in false positive rates also goes to 0. (c) The same happens for demographic parity.

3. Social welfare as a Measure of Group Fairness

In this section we explore how the notion of social welfare can be used as a means to ensure group fairness. We previously advocated for social welfare as an individual fairness measure, but the same concept could be applied to groups as well. Consider the thought experiment by John Rawls outlined in Section 2.1: an individual has to choose between different algorithms that could be used for decision making in a world in which the individual would be born into. One does not know a priori to which group one will belong. The only information given is the group-benefits and the sizes of the respective groups. In this position a rational, risk-averse person would try to maximize the expected group utility.

Therefore we can solve an analogous optimization problem as in Section 2, where we replace the individual utility by the group utility. To compute utilities for groups we first need to define benefit on a group level. One of the most natural ways to define group benefit is to take the average benefit of all individuals in a group. Let G be a group, we then define the group benefit $b_g(G)$ as follows:

$$b_g(G) = \frac{1}{|G|} \sum_{i \in G} b(y_i, \hat{y}_i)$$

We will work with the same family of utility functions $u(b) = (sign(\alpha) \cdot b)^\alpha$ as proposed in Section 2.2.

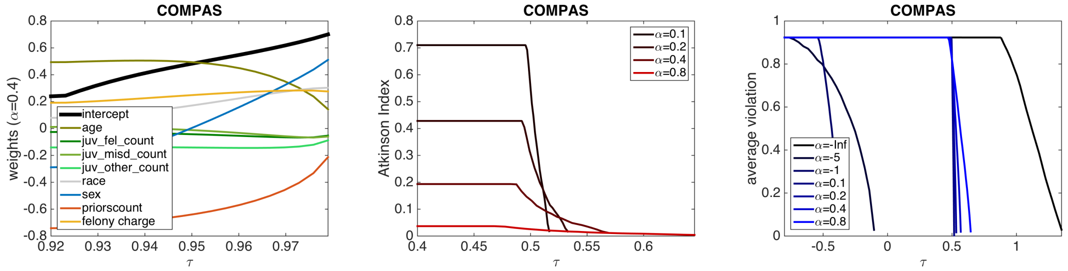
Now suppose we have groups $G_1, G_2, ..., G_m$ and a loss function ℓ. Then we want to solve the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\theta} \in \mathcal{H}} \quad & \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \mathbf{x}_i, y_i) \\ \text{s.t.} \quad & \frac{1}{m} \sum_{j=1}^{m} \frac{|G_j|}{n} \; u(b_g(G_j)) \geq \tau \end{aligned}$$

3.1. Experiments

Here we illustrate how different measures of individual fairness and group discrimination behave when we solve the optimization problem above for different values of τ. The experiments were also done using logistic regression and the same benefit function for individuals as in Section 2.3. Also analogous to the previous section, we used $\hat{y} = \boldsymbol{\theta}.\mathbf{x}$ instead of $\hat{y} = sign(\boldsymbol{\theta}.\mathbf{x})$ to compute benefits in order to have convex constraints.

Evaluating Individual Level Unfairness Figure 6a shows the effect of raising the lower bound τ on the weights θ of our logistic regression problem. One can observe the same patterns as in Figure 3a. Most notably, the intercept continually rises when increasing τ. In Figures 6b and 6c we can see that enforcing a lower bound on social welfare on a group level also leads to lower individual unfairness.



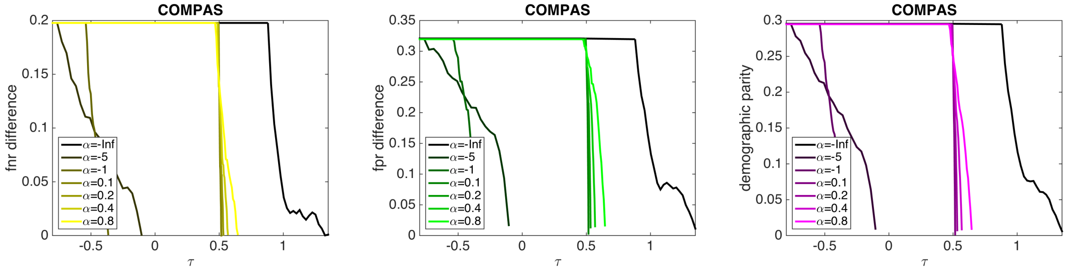
(a) **(b)** **(c)**

Figure 6: (a) Changes in weights—$\boldsymbol{\theta}$ in logistic regression—as the function of τ. Note the continuous rise of the intercept with τ. (b) Atkinson index as a function of the threshold τ. Note that it is only defined for $\alpha \geq 0$. (c) Dwork measure as a function of τ. Note the consistent decline in inequality as τ increases.

Evaluating Group Level Unfairness Next, we look at how measures of group discrimination behave (Figure 7). As expected, bounding our measure leads to decreasing group discrimination. In most cases this happens in a rather smooth manner, except for $\alpha = -\infty$. This could be attributed to the fact that we compute benefits w.r.t. the distance from the hyperplane $\boldsymbol{\theta}.\mathbf{x}$, as previously mentioned.

This can lead to a scenario where the ranking of benefits of two groups is actually inverted depending on if the benefits are (i) computed with $\boldsymbol{\theta}.\mathbf{x}$ or (ii) with $sign(\boldsymbol{\theta}.\mathbf{x})$. Let's construct a small example to show this: consider groups A and B, each containing 3 data points with the following values of $(\boldsymbol{\theta}.\mathbf{x}, y)$: $\{(0.1, 1), (0.1, 1), (-1, 1)\}$ for group A and $\{(1, 1), (-0.1, 1), (-0.1, 1)\}$ for group B. For simplicity we take $b(y, \hat{y}) = \hat{y} - y$ although such an example could also easily be constructed for the benefit function used in the experiments. The benefits computed with (i) are $\{-0.9, -0.9, -2\}$ and $\{0, -1.1, -1.1\}$ for groups A and B respectively. This is what the solver sees during the optimization process, so when τ is increased, the benefits of group A would be increased. But if we look at the benefits computed with (ii) $\{0, 0, -2\}$ and $\{0, -2, -2\}$ we see that actually group A had a higher average benefit to begin with. This can lead to locally increasing measures of group discrimination. For $\alpha = -\infty$ the constraints will *always* act on the group with lower benefit (Rawlsian max-min welfare), therefore this behavior is most clearly visible in that setting.

It is important to note however, that this only leads to some local fluctuations, the general trend is still showing decreasing unfairness.



(a) **(b)** **(c)**

Figure 7: Group discrimination as a function of τ for different values of α. (a) The difference in false negative rate is decreasing with τ and approaches 0. (b) The same happens for the difference in false positive rate. (c) The difference in positive prediction percentage also goes to 0.

4. A Computationally Feasible Heuristic for Dwork et al.'s Individual Fairness

An early proposal on individual fairness was made by Dwork et al. [14]. As an alternative approach to our social welfare constrained classifier one could think of trying to directly solve the optimzation problem they propose. In this section we will present a possible heuristic approach for doing that. Informally their constraints say that for any two individuals the distance between their outcomes should be at most the difference between the two individuals, so similar individuals should receive similar outcomes.

More formally, let ℓ be a loss function, $d_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a distance function between individuals and $d_y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a distance function between outcomes. The optimization problem then looks as follows, assuming a linear predictor:

$$\begin{aligned} \min_{\boldsymbol{\theta} \in \mathcal{H}} \quad & \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \mathbf{x}_i, y_i) \\ \text{s.t.} \quad & \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X} \quad d_y(\boldsymbol{\theta}.\mathbf{x}_i, \boldsymbol{\theta}.\mathbf{x}_j) \leq d_x(\mathbf{x}_i, \mathbf{x}_j) \end{aligned}$$

Contained in that proposal is the assumption that we can compute distances both between individuals and between outcomes. We therefore need to instantiate both d_x and d_y. The distance between individuals we took to be the euclidean distance between their input vectors, normalized by the largest euclidean distance between any two input vectors.

$$d_x(\mathbf{x}_i, \mathbf{x}_j) = \frac{||\mathbf{x}_i - \mathbf{x}_j||}{\max\limits_{\mathbf{x}_k, \mathbf{x}_l \in \mathcal{X}} ||\mathbf{x}_k - \mathbf{x}_l||}$$

Different choices of measures are imaginable but we thought this would be a reasonable default choice.

For the distance between outcomes Dwork et al. [14] propose the statistical distance, which in case of a dichotomy is just the difference in probabilities of being assigned to the positive class. In the case of logistic regression, for individuals $\mathbf{x}_i$ and $\mathbf{x}_j$, that is:

$$d_y(\boldsymbol{\theta}.\mathbf{x}_i, \boldsymbol{\theta}.\mathbf{x}_j) = \left| \frac{1}{(1 + exp(-\boldsymbol{\theta}.\mathbf{x}_i))} - \frac{1}{(1 + exp(-\boldsymbol{\theta}.\mathbf{x}_j))} \right|$$

However, this constraint is not convex in our weight vector $\boldsymbol{\theta}$. Another natural idea would be to take the absolute difference between predicted class labels as a distance measure, namely:

$$d_y(\boldsymbol{\theta}.\mathbf{x}_i, \boldsymbol{\theta}.\mathbf{x}_j) = |sign(\boldsymbol{\theta}.\mathbf{x}_i) - sign(\boldsymbol{\theta}.\mathbf{x}_j)|$$

Unfortunately also this is non-convex. Therefore we used the following as a convex surrogate

$$d_y(\boldsymbol{\theta}.\mathbf{x}_i, \boldsymbol{\theta}.\mathbf{x}_j) := \left| \frac{\boldsymbol{\theta}.\mathbf{x}_i}{c} - \frac{\boldsymbol{\theta}.\mathbf{x}_j}{c} \right|$$

With c being a constant that ensures that the range of the values is comparable to the normalized distance. For logistic regression this gives us the following *convex* optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\theta} \in \mathcal{H}} \quad & \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{\theta}.\mathbf{x}_i)) \\ \text{s.t.} \quad & \forall \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X} \quad d_y(\boldsymbol{\theta}.\mathbf{x}_i, \boldsymbol{\theta}.\mathbf{x}_j) \leq d_x(\mathbf{x}_i, \mathbf{x}_i) \end{aligned}$$

As noted in Section 1.1 however, the number of required constraints is prohibitive. We therefore propose a heuristic approach that works in the following manner:

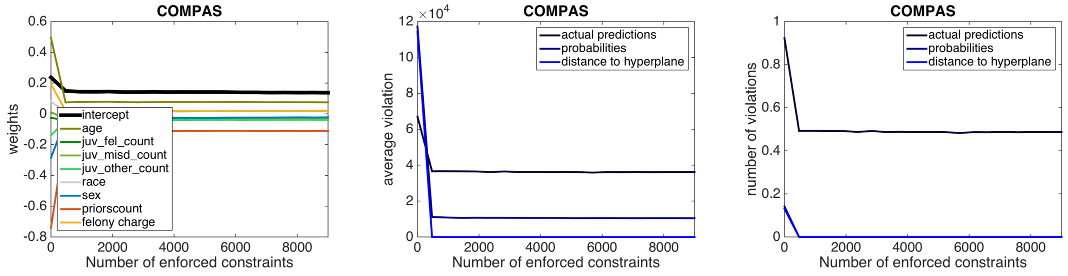
1. Train an unconstrained classifier
2. List all constraints that are violated by more than a margin of δ
3. Shuffle the list randomly, take the first k constraints, where k is chosen such that the optimization is still feasible, add those to the optimization problem
4. Train a classifier with the added constraints

So we enforce the constraint that are violated badly and hope that by doing this, other constraints will be satisfied as well. Afterwards we would like to evaluate how well our constrained classifier satisfies all the initial pairwise constraints. For that we have to ask ourselves how to measure violations of the constraints. Note that for evaluating them we don't need convexity so all the three options above are on the table. Taking the probabilities would be closest to the original idea of the paper. We argue taking the actual outcomes may capture fairness most intuitively since in real life people might not care about probabilities of a decision rather than the actual decision. And lastly one could argue that if we optimize with the distance to the hyperplane as constraints, it would make sense to evaluate the violations in this manner as well. Because an argument can be made for each one of the options, we included all three in the experiments.

4.1. Experiments

We implemented the heuristic outlined above. Then we chose $\delta = 0.55$, k was initially chosen to be 0 and then increased in 20 steps of equal size up to 9000. The list of violated constraints stayed the same, and in each iteration we enforced the first k of them. All evaluations were done on the validation sets that were obtained by doing 10-fold cross-validation.

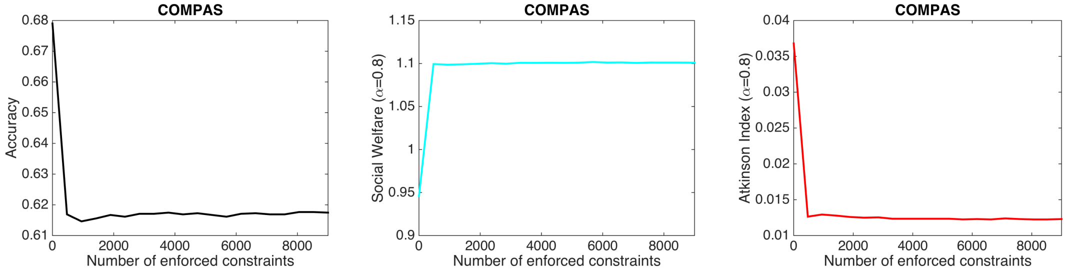
In Figure 8 one can observe how the weights and also both the average and number of violations on the validation data behave as more constraints are enforced on the training data. It is apparent that the weights change as the first few constraints are enforced but afterwards remain almost the constant. A possible explanation for this is that if we enforce the first few constraints of our list of constraints that we possibly could to enforce, most of the constraints that we add in later iterations are already satisfied and thus the classifier does not change much. Both violation measures sharply decline initially, and also remain unchanged after that. We see that all three violation measures exhibit the same trends. However the violation w.r.t. the actual predictions decrease much slower than the other two. At this point it is noteworthy that for a classification task with labels $\{-1, 1\}$ and this choice of a normalized distance function this measure only vanishes if all data points are assigned to the same class. That is because if two individuals are assigned to different classes, one gets the label 1 and the other one gets the label -1. The distance between those is 2 and it should be smaller than the normalized distance between the two individuals (which can't be larger than 1).



(a) **(b)** **(c)**

Figure 8: (a) Weights $\boldsymbol{\theta}$ as a function of the number of enforced constraints during training. Interestingly the intercept does not rise as it did when enforcing social welfare constraints. (b) The number of violated constraints on the validation set as a function of the number of enforced constraints. (c) The average violation of constraints on the validation set as a function of the number of enforced constraints.

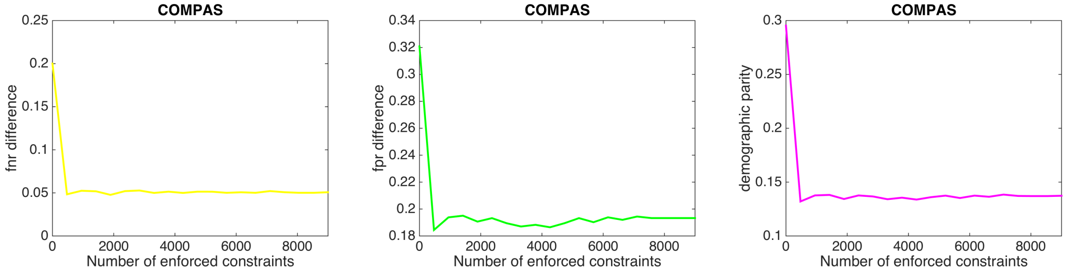
Figures 9 and 10 tell the same story. Initially the classifier becomes more fair w.r.t. to all tested measures of both individual and group fairness but after the initial decline the classifier and therefore the measures do not change much.



(a) **(b)** **(c)**

Figure 9: Changes in accuracy and measures of individual fairness with increasing numbers of enforced constraints. Accuracy and Atkinson's Index sharply decline at first but stay at a constant non-zero rate afterwards. For Social Welfare the trend is exactly reversed, which is what we expect because larger is better in this case as opposed to (a) and (c).

On one hand this is good news from a computational point of view, because we apparently only have to include very few constraints in our optimization procedure. On the other hand we have to conclude that this approach will not allow us to decrease unfairness arbitrarily as was for example possible by constraining social welfare.



(a) **(b)** **(c)**

Figure 10: Group discrimination as a function of the number of enforced constraints. Also here, all measures of group discrimination sharply decline at first but remain at roughly the same level afterwards.

5. Conclusions and Future Directions

Our work makes an important connection between the growing literature on fairness for machine learning, and the long-established notion of cardinal social welfare in economics. We formulated an optimization procedure based on our model for both individual and group level fairness. Thanks to their convexity, our measures can be bounded as part of any convex loss minimization program. We provided both theoretical justification and empirical evidence suggesting that constraining our measures often leads to bounded inequality in algorithmic outcomes, hence presenting the first computationally feasible mechanism for bounding individual-level (un)fairness. Additionally we compared our approach to a heuristic for directly optimizing an existing individual-level fairness definition and found that the heuristic could not decrease unfairness beyond a certain threshold.

Our focus in this work was on a normative theory of how rational individuals should react to different algorithmic alternatives. Descriptive behavioural theories, such as the prospect theory [21], may be instrumental in understanding the *perception of fairness* through our framework. Another interesting avenue for future work would be exploring the possibility of defining fairness via ordinal social welfare.

Acknowledgements

First and foremost I would like to thank Dr. Hoda Heidari for supporting me throughout the whole project, answering all my questions, discussing the main findings and giving me a great insight into the world of research. Furthermore I am grateful for the opportunity provided by Prof. Dr. Andreas Krause to write this thesis in the Learning and Adaptive Systems group.

References

[1] Dennis J Aigner and A J Heins. A social welfare view of the measurement of income equality. *Review of Income and Wealth*, 13(1):12–25, 1967.

[2] Yoram Amiel and Frank Cowell. Inequality, welfare and monotonicity. 1997.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. *Propublica*, 2016.

[4] Anthony B Atkinson. On the measurement of inequality. *Journal of economic theory*, 2(3):244–263, 1970.

[5] Anna Barry-Jester, Ben Casselman, and Dana Goldstein. The New Science of Sentencing. *The Marshall Project*, aug 2015. URL `https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing`.

[6] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 71–80. IEEE, 2013.

[7] Fredrik Carlsson, Dinky Daruvala, and Olof Johansson-Stenman. Are People Inequality-Averse, or Just Risk-Averse? *Economica*, 72(287):375–396, 2005.

[8] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *arXiv preprint arXiv:1701.08230*, 2017.

[9] Frank A Cowell. Measurement of inequality. *Handbook of income distribution*, 1: 87–166, 2000.

[10] Frank A Cowell and Erik Schokkaert. Risk perceptions and distributional judgments. *European Economic Review*, 45(4-6):941–952, 2001.

[11] Camilo Dagum. On the relationship between income inequality measures and social welfare functions. *Journal of Econometrics*, 43(1-2):91–102, 1990.

[12] Hugh Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30(119):348–361, 1920.

[13] Gerard Debreu and Others. Topological methods in cardinal utility theory. Technical report, Cowles Foundation for Research in Economics, Yale University, 1959.

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[15] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[16] William M Gorman. The structure of utility functions. *The Review of Economic Studies*, 35(4):367–390, 1968.

[17] Moritz Hardt, Eric Price, Nati Srebro, and Others. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[18] John C Harsanyi. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, 61(5):434–435, 1953.

[19] John C Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, 63(4):309–321, 1955.

[20] Hoda Heidari, Claudio Ferrari, Krishna P. Gummadi, and Andreas Krause. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. pages 1–15, 2018. URL `http://arxiv.org/abs/1806.04959`.

[21] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[22] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[23] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. Data and analysis for '{How} we analyzed the {COMPAS} recidivism algorithm'. \url{https://github.com/propublica/compas-analysis}, 2016.

[24] Sam Levin. A beauty contest was judged by {AI} and the robots didn't like dark skin. *The Guardian*, 2016.

[25] M Lichman. {UCI} Machine Learning Repository: Communities and Crime Data Set. \url{http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime}, 2013.

[26] Clair Miller. Can an Algorithm Hire Better than a Human? *The New York Times*, jun 2015. URL `http://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html/`.

[27] Hervé Moulin. *Fair division and collective welfare*. MIT press, 2004.

[28] Kevin Petrasic, Benjamin Saul, James Greig, and Matthew Bornfreund. Algorithms and bias: What lenders need to know. *White & Case*, 2017.

[29] Arthur Cecil Pigou. *Wealth and welfare.* Macmillan and Company, limited, 1912.

[30] John Rawls. *A theory of justice.* Harvard university press, 2009.

[31] Kevin W S Roberts. Interpersonal comparability and social choice theory. *The Review of Economic Studies*, pages 421–439, 1980.

[32] Cynthia Rudin. Predictive Policing Using Machine Learning to Detect Patterns of Crime. *Wired Magazine*, aug 2013. URL `http://www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-patterns-of-crime/`.

[33] Joseph Schwartz and Christopher Winship. The welfare approach to measuring inequality. *Sociological methodology*, 11:1–36, 1980.

[34] Amartya Sen. On weights and measures: informational constraints in social welfare analysis. *Econometrica: Journal of the Econometric Society*, pages 1539–1572, 1977.

[35] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual and Group Unfairness via Inequality Indices. 2018.

[36] Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.

[37] Hal R Varian. Equity, envy, and efficiency. *Journal of economic theory*, 9(1):63–91, 1974.

[38] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *arXiv preprint arXiv:1507.05259*, 2017.

[39] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From Parity to Preference-based Notions of Fairness in Classification. In *Advances in Neural Information Processing Systems*, pages 228–238, 2017.

A. Omitted Experimental Details

Data set We used the *COMPAS dataset* originally compiled by Propublica [23]. The data consists of 5278 observations each made up of the following features: intercept, severity of charge (felony or misdemeanour), number of priors, juvenile felony count, juvenile misdemeanor count, other juvenile offense count, race (African-American or white), age, gender, COMPAS scores (not included in our analysis). The target variable indicates the actual recidivism within 2 years. The data was filtered following the original study: If the COMPAS score was not issued within 30 days from the time of arrest, because of data quality reasons the instance was omitted. The recidivism flag is assumed to be -1 if no COMPAS case could be found at all. Ordinary traffic offences were removed. We standardized the non binary features to have mean 0 and variance 1. Also, we negated the labels to make sure higher y-values correspond to more desirable outcomes.

Optimization program for classification Ideally we would like to find the optimum of the following constrained optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}\in\mathcal{H}} \quad & \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i\boldsymbol{\theta}.\mathbf{x}_i)) \\ \text{s.t.} \quad & \frac{1}{n}\sum_{i=1}^{n}u(b(y_i, sign(\boldsymbol{\theta}.\mathbf{x}_i))) \geq \tau \end{aligned}$$

However, the sign function makes the constraint non-convex, therefore we instead solve the following:

$$\begin{aligned} \min_{\boldsymbol{\theta}\in\mathcal{H}} \quad & \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i\boldsymbol{\theta}.\mathbf{x}_i)) \\ \text{s.t.} \quad & \frac{1}{n}\sum_{i=1}^{n}u(b(y_i, \frac{\boldsymbol{\theta}.\mathbf{x}_i}{c})) \geq \tau, \\ & \|\boldsymbol{\theta}\|^2 = 1 \end{aligned}$$

The constant c ensures that the argument ($\frac{\boldsymbol{\theta}.\mathbf{x}_i}{c}$) of the benefit function is in $[-1, 1]$ which keeps our benefit non negative. For Section 2 we chose $c = 5$, for Section 3 we chose $c = 1$. We constrain $\boldsymbol{\theta}$ to be unit-length since otherwise one could increase the benefit without changing the classification outcome by just increasing the length of $\boldsymbol{\theta}$.

How we computed the different fairness measures Suppose we have two groups G_1, G_2 and our labels for classification are in $\{-1, 1\}$. Also let

$$G^+ := \sum_{i\in G}\mathbf{1}[\hat{y}_i > y_i]$$

and similarly

$$G^- := \sum_{i\in G}\mathbf{1}[\hat{y}_i < y_i]$$

- **Dwork et al. measure**

  $$\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\max\{0,|\hat{y}_i-\hat{y}_j|-d(i,j)\}$$

  At a high level, the measure is equal to the average of the amount by which each pairwise constraint is violated. We took $d(i,j)$ to be the Euclidean distance between $\mathbf{x}_i,\mathbf{x}_j$ divided by the maximum Euclidean distance between any two points in the dataset. The normalization step is performed to make sure the range of $|\hat{y}_i-\hat{y}_j|$ and $d(i,j)$ are similar. For Sections 2 and 3 we took $\hat{y}_i = sign(\theta.\mathbf{x}_i)$.

- **Atkinson's Index** is computed as described in Section 2.2

- **Demographic parity** is computed by taking the absolute difference between percentage of positive predictions across groups:

  $$\left|\frac{1}{|G_1|}\sum_{i\in G_1}\mathbf{1}[\hat{y}_i=1]-\frac{1}{|G_2|}\sum_{i\in G_2}\mathbf{1}[\hat{y}_i=1]\right|$$

- **Difference in false positive rate** is computed by taking the absolute difference of the false positive rates across groups:

  $$|f_{fpr}(G_1)-f_{fpr}(G_2)|$$

  where:

  $$f_{fpr}(G) := \sum_{i\in G}\frac{\mathbf{1}[\hat{y}_i=1\wedge y_i=-1]}{\mathbf{1}[y_i=-1]}$$

- **Difference in false negative rate** is computed by taking the absolute difference of the false negative rates across groups:

  $$|f_{fnr}(G_1)-f_{fnr}(G_2)|$$

  where:

  $$f_{fnr}(G) := \sum_{i\in G}\frac{\mathbf{1}[\hat{y}_i=-1\wedge y_i=1]}{\mathbf{1}[y_i=1]}$$