Machine learning systems are increasingly used to make decisions about people’s lives, such as whether to give someone a loan or whether to interview someone for a job. This has led to considerable interest in making such machine learning systems fair. One approach is to transform the input data used by the algorithm. This can be achieved by passing each input data point through a representation function prior to its use in training or testing. Techniques for learning such representation functions from data have been successful empirically, but typically lack theoretical guarantees. We show that it is possible to prove that a representation function is fair according to common measures of both group and individual fairness, as well as useful with respect to a target task. These provable properties can be used in a governance model involving a data producer, a data user and a data regulator, where there is a separation of concerns between fairness and target task utility to ensure transparency and prevent perverse incentives. We formally define the ‘cost of mistrust’ of using this model compared to the setting where there is a single trusted party, and provide bounds on this cost in particular cases. We present a practical approach to learning fair representation functions and apply it to financial and criminal justice datasets. We evaluate the fairness and utility of these representation functions using measures motivated by our theoretical results.

1 INTRODUCTION

Machine learning algorithms play an ever-increasing role in society. Such algorithms are widely used to make decisions about individuals, such as whether to grant someone a loan or what advertisements to serve someone online. Other relevant examples are automated applicant screening used by employers and educational institutions, automated assignment grading, and algorithmic advice on criminal sentencing. A risk of this trend is that such algorithms may be unfair in some way, for example by discriminating against particular groups. Even if not intended by the algorithm designer, discrimination is possible because the reasoning behind the algorithm’s decisions is often difficult for humans to interpret. Furthermore, artefacts of previous discrimination present in the algorithm’s training data may increase this tendency in the algorithm’s decisions.

To remove or minimize discrimination effects caused by the use of machine learning systems, a fairness objective may be incorporated into algorithm design. This approach benefits users of such algorithms, particularly those in social groups that are potentially the subject of discrimination on grounds such as race or gender.

We consider the task of producing a decision variable (e.g., whether to grant a loan) which predicts a target variable of interest (e.g., loan default), while at the same time avoiding discrimination on the basis of an individual’s group membership (e.g., race, gender) encoded in a sensitive variable. We adopt the approach of passing input data about individuals through a representation function, so that subsequent use of the cleaned data will not be able to discriminate based on the sensitive variable. By incorporating fairness as a data pre-processing step, we achieve a separation of concerns between fairness and target task utility which offers governance and regulatory advantages.

Our work provides theoretical guarantees for several performance measures of interest:

1. **Group fairness**: Using common measures of the similarity of decisions for one group compared to another, we show that it is possible for a representation function to improve group fairness.
2. **Individual fairness**: We quantify the effect the representation function has on the assignment of similar decisions to individuals who are similar.
3. **Target utility**: We quantify the cost a representation function incurs in terms of target variable utility.
4. **Cost of mistrust**: We quantify the cost of this separation of concerns, relative to a setting where the decision-making party is given full access to the sensitive variable.

Previous work on fairness in machine learning has mostly considered the case where there is a single trusted data user [8, 15, 17, 18]. This baseline scenario is shown in Figure 1(a). The data user learns decision variable \(\hat{Y}\) by training on samples of input variable \(X\), target variable \(Y\) and sensitive variable \(S\). We assume that \(S\), \(Y\) and \(\hat{Y}\) are binary variables, a restriction which still allows us to work on many problems of interest, as we show. \(X\) may be discrete or continuous. At test time, on input \(X = x\), the data user makes decision \(P(\hat{Y} = 1|X = x)\). We assume that \(S\) is not available at test time. However, this is not overly restrictive since \(X\) may be constructed to perfectly predict \(S\) (e.g., \(S\) is one of the attributes included in \(X\)).

In our scenario of interest, shown in Figure 1(b), the data user is treated as untrusted. This may be appropriate in a setting where the data user’s incentives are aligned to target task utility, possibly at the expense of fairness. Our scenario of interest involves three

---

\[1\] Some other related notions of fairness, such as compliance by search engines with antitrust legislation, may also be readily incorporated within our framework.

\[2\] For example, the European Union has proposed 'right to explanation' laws scheduled to commence in 2018, allowing a user to ask for an explanation of an algorithmic decision made about them. [10]
parties: a data producer who prepares the input data, a data user who makes decisions from the data, and a data regulator who oversees fair use of the data. For example, in the context of deciding whether to give an individual a loan, the data producer might be a credit bureau, the data user a bank and the data regulator a government authority. Even if the data producer and data user (and potentially even the data regulator) are the same organization, this conceptual framework provides improved transparency.

We consider the case where the data user only receives access to cleaned data prepared by the data producer, in order to ensure the fairness of the decisions made. The data producer passes the input variable $X$ through a representation function $f$ to produce cleaned variable $X_f$, which is then made available to the data user who learns decision variable $\hat{Y}_f$ using $X_f$ and $Y$. We assume $\hat{Y}_f$ is also binary. We assume $X_f$ has the same domain as $X$, which means that the cleaned data is interpretable using existing domain knowledge. The representation function $f$ is learned by the data producer from samples of the input variable $X$ and sensitive variable $S$. At test time, on input $X = x$, the data producer computes $X_f = f(x)$, after which the data user makes decision $P(\hat{Y}_f = 1|X_f = f(x))$.

The data producer computes performance guarantees (as described above) about group fairness and individual fairness to the data regulator and about target task utility and the cost of mistrust to the data user. In practice, the statistics required for the guarantees must be estimated from a finite data sample.

It is not necessary for the data producer to have access to target variable labels $Y$ which may be used by the data user for supervised learning. This is a realistic situation in the case where the labels are constructed using proprietary transactions (e.g. previous loans) conducted by the data user. Furthermore, assuming that $Y$ is not modified ensures that the data user still has an incentive to accurately predict these labels. The properties of $f$ we guarantee hold for any target task, despite the fact that $Y$ may not be accessible when $f$ is learned.

The remainder of the paper is structured as follows. In Section 2 we introduce related work. In Section 3 we present our theoretical results. In Section 4 we propose a practical algorithm for learning fair representation functions, apply it to real-world financial and criminal justice datasets, and analyze the key performance measures suggested by our theoretical results. We conclude and present ideas for future work in Section 5. We defer proofs of our theorems to Section 6.

## 2 BACKGROUND

Fairness in machine learning has emerged as a significant field of research in recent years [6, 9, 11, 20]. A seminal study providing theoretical foundations for fairness in machine learning [7] proposed two notions of fairness. Group fairness can be defined, roughly, as similar (or at least more similar than a baseline\(^3\)) decisions for one group compared to another. This is the type of fairness we focus on improving. The individual fairness of a machine learning algorithm can be defined, roughly, as the assignment of similar decisions to individuals who are similar by some metric. The two notions of fairness are potentially in tension with each other: group fairness achieves equal outcomes for each group regardless of the characteristics of the individuals that make up the groups, while individual fairness provides individuals who are similar with equal treatment regardless of their group membership. We are interested in the problem of improving group fairness, without too great a

---

\(^3\)For example, an affirmative action system may introduce quotas which do not aim for ‘equality of outcome’ but rather ‘reduced inequality of outcome’.
cost to individual fairness or to performance on the target task of interest.

A common approach to measuring group fairness is to compare the decision variable probabilities conditioned on different values of the sensitive variable. The difference in these probabilities was proposed as statistical parity \[7\] and is commonly in use \[8, 15\]. The ratio of these probabilities, known as disparate impact, is favoured in certain applications such as the 80% rule advocated by the US Equal Employment Opportunity Commission \[4\] and has also been studied in the literature \[9, 15\]. It may be appropriate to require group fairness only on some subset of the input data \[3\], such as requiring equal employment outcomes for men and women with suitable qualifications rather than for all men and women. Disparate mistreatment, which compares the proportion for all men and utility. Subsequently the ‘adversarial’ approach \[2, 8\] to learning terms for each of individual fairness, group fairness and target task cluster centers. A target task linear classifier using the cleaned data therefore focus on modifying the input data only.

to accurately predict the target label if it has been modified. We inputs and the target labels has also been proposed \[3\]. However, a subsequent decision variable, relying only on the fact that some particular learner failed to be unfair using the cleaned data. In fact, a data user focused only on target task utility may have a perverse incentive to use a weak learner which makes the representation function appear fair. This problem highlights the need for provably fair representation functions.

3 THEORETICAL GUARANTEES FOR FAIR AND USEFUL REPRESENTATIONS

We propose theoretical guarantees for learning representation functions which improve group fairness, while preserving individual fairness and target task utility.

3.1 Guaranteeing group fairness

We would like to guarantee that a representation function \( f \) achieves group fairness if used to transform \( X \) to \( X_f \). To formalize group fairness, we first consider statistical parity \[7\], which measures the difference in outcomes for different sensitive variable groups.

**Definition 3.1 (Statistical parity).** Decision variable \( \hat{Y} \) and sensitive variable \( S \) have statistical parity

\[
SP(\hat{Y}, S) := p(\hat{Y} = 1|S = 1) - p(\hat{Y} = 1|S = 0).
\]

We also consider the related notion of normalized disparate impact, which measures the ratio of outcomes for different sensitive variable groups.\(^4\)

**Definition 3.2 (Normalized disparate impact).** Decision variable \( \hat{Y} \) and sensitive variable \( S \) have normalized disparate impact

\[
DI(\hat{Y}, S) := 1 - \frac{p(\hat{Y} = 1|S = 1)}{p(\hat{Y} = 1|S = 1)}.
\]

We assume \( p(\hat{Y} = 1|S = 1) \geq p(\hat{Y} = 1|S = 0) \) without loss of generality. Since the assignment of sensitive variable groups to the values 0 and 1 is arbitrary. Therefore \( SP(\hat{Y}, S), DI(\hat{Y}, S) \in [0, 1] \) where 0 is perfectly fair and 1 is maximally unfair.

3.1.1 Statistical parity guarantee. The statistical parity of a decision variable is closely related to the balanced error rate with respect to the sensitive variable, as observed in \[15\].

\[^{4}\text{Disparate impact defined as } \frac{p(Y = 1|S = 1)}{p(Y = 1|S = 0)} \text{ is commonly used, although we prefer normalized disparate impact so that like statistical parity, smaller values are fairer.}\]
Definition 3.3 (Balanced error rate). Decision variable $\hat{Y}$ has balanced error rate with respect to sensitive variable $S$ defined as

$$BER(\hat{Y}, S) := \frac{1}{2}p(\hat{Y} = 1|S = 0) + \frac{1}{2}p(\hat{Y} = 0|S = 1).$$

While in our problem setting the role of the decision variable is to predict the target variable $Y$ rather than $S$, an adversary may instead make decisions using predictions of $S$. We may tightly upper bound the statistical parity of any decision made using $X_f$ using the minimum achievable balanced error rate, as shown in Theorem 3.4. The relationship between $BER$ and $SP$ for a particular classifier was previously observed by [15], although our result provides an $SP$ guarantee for any classifier using $X_f$.

Theorem 3.4 (Tight bound on statistical parity). For all decisions $\hat{Y}_f$ made using $X_f$ as input,

$$SP(\hat{Y}_f, S) \leq 1 - 2\text{BER}(\hat{Y}_f^{BER}, S).$$

The bound is tight when $\hat{Y}_f = \hat{Y}_f^{BER}$, where

$$\hat{Y}_f^{BER}(x) := I(p(S = 1|X_f = x) \geq p(S = 1)).$$

3.1.2 Alternative statistical parity guarantee using conditional entropy. Applying Theorem 3.4 involves computing the decision variable which minimizes balanced error rate when predicting $S$ from $X_f$. This involves making decisions by thresholding the distribution $p(S = 1|X_f = x)$. Because of the discontinuity introduced by this thresholding function, the method may be sensitive to errors in the estimation of $p(S = 1|X_f = x)$ and $p(S = 1)$ from a finite sample. We present an alternative upper bound on statistical parity, which avoids both the thresholding function and the need to estimate $P(S = 1)$ by instead estimating the conditional entropy of $S$ given $X_f$.

We introduce two definitions from information theory which we will require for our statistical parity guarantee.

Definition 3.5 (Binary entropy). For a binary variable with success probability $p$, the binary entropy is

$$H_b(p) := -p \log_2 p - (1 - p) \log_2(1 - p).$$

Definition 3.6 (Conditional entropy). For a binary variable $S$ and input variable $X$, the conditional entropy

$$H(S|X) := \mathbb{E}_x[H_b(p(S = 1|X = x))].$$

In Theorem 3.7 we upper bound statistical parity using $H(S|X_f)$.

Hence, making $H(S|X_f)$ large ensures that any decision-making procedure using $X_f$ will have small statistical parity.

Theorem 3.7 (Bound on statistical parity using conditional entropy). Let $H_b^{-1}$ be the inverse of $H_b(p)$ where $p \in [0, \frac{1}{2}]$. For all decisions $\hat{Y}_f$ made using $X_f$ as input,

$$SP(\hat{Y}_f, S) \leq 1 - \frac{H_b^{-1}(H(S|X_f))}{\max(p(S = 1), p(S = 0))}.$$

3.1.3 Normalized disparate impact guarantee. We may tightly upper bound the normalized disparate impact of any decision made using $X_f$, as shown in Theorem 3.8.

Theorem 3.8 (Tight bound on normalized disparate impact). Let $\eta_f := \max_x p(S = 1|X_f = x)$. Let

$$\hat{Y}_f^{DI}(x) := \begin{cases} y & \text{if } p(S = 1|X_f = x) = \eta_f, \\ 0 & \text{otherwise.} \end{cases}$$

where $y$ is an arbitrary positive constant. For all decisions $\hat{Y}_f$ made using $X_f$ as input,

$$DI(\hat{Y}_f, S) \leq 1 - \frac{p(S = 1)(1 - \eta_f)}{p(S = 0)\eta_f}.$$

The bound is tight when $\hat{Y}_f = \hat{Y}_f^{DI}$.

Previous work has provided upper bounds on $DI$ for a particular decision using balanced error rate and the related notion of cost sensitive risk [9, 15]. However, our result applies to any decision made using $X_f$ as input. Furthermore, $p(\hat{Y}_f^{DI} = 0|S = 1) \approx 1$, $p(\hat{Y}_f^{DI} = 1|S = 0) \approx 0$ and so $BER(\hat{Y}_f^{DI}, S) \approx \frac{1}{2}$ regardless of $X_f$, while if $\eta_f \approx 1$ then $DI(\hat{Y}_f^{DI}, S) \approx 1$. Hence ensuring $BER(\hat{Y}_f, S)$ is large for all $\hat{Y}_f$ is not sufficient to guarantee that $DI(\hat{Y}_f, S)$ is small for all $\hat{Y}_f$. However, we can guarantee $DI(\hat{Y}_f)$ will be small for all $\hat{Y}_f$ if $\eta_f$ is not too large. Furthermore we observe that because the proof of Theorem 3.7 relies on the particular relationship of $SP$ and $BER$, there is no immediate analog to this theorem in the case of $DI$.

3.2 Guaranteeing individual fairness

We introduce the notion of individual fairness from [7], which roughly states that if two individuals are similar then a similar decision should be made about them. Another term used for this type of (un)fairness is disparate treatment [17]. This is familiar mathematically as a smoothness property.

Definition 3.9 (Individual fairness [7]). Decision variable $\hat{Y}$ is $D, d$ individually fair with respect to input variable $X$ iff $\forall x, x'$, $\text{DI}(\hat{Y}(\hat{Y} = 1|X = x), p(\hat{Y} = 1|X = x')) \leq d(x, x')$, where $D$ and $d$ are subadditive functions.

We also give a quantitative notion of individual unfairness by measuring the probability that a pair of randomly selected individuals will be treated fairly.

Definition 3.10 (Individual unfairness). Decision variable $\hat{Y}$ and input variable $X$ have $D, d$ individual unfairness $\text{IU}_{D, d}((\hat{Y}, X)) := p(D(\hat{Y}(\hat{Y} = 1|X = x), p(\hat{Y} = 1|X = x')) > d(x, x'))$ where $x$ and $x'$ are independent random samples of $X$.

We would like to understand the effect on individual fairness of moving from $X$ to $X_f$. We assume the motivation for using $X_f$ is group rather than individual fairness. However, we would like to ensure that group fairness is not achieved at too great a cost to individual fairness. To do this we introduce the following definition.

Definition 3.11 (Large reconstruction error rate). Let $\epsilon$ be a constant and $x$ be a random sample of $X$. Let $p(d(x, f(x)) > \epsilon)$ be the large reconstruction error rate.
In Theorem 3.12 we provide a result guaranteeing that if $X_f$ approximately reconstructs $X$ for most individuals, as measured by the large reconstruction error rate $p(d(x, f(x)) > \epsilon)$, then a decision variable that is individually fair using $X$ can be used to construct a decision variable that is not too individually unfair using $X_f$.

**Theorem 3.12 (Guarantee for individual fairness).** Suppose decision variable $\hat{Y}$ is individually fair with respect to $X$ and the large reconstruction error rate $p(d(x, f(x)) > \epsilon) \leq \delta$. Let $\hat{Y}_f$ be the random variable formed by setting $p(\hat{Y}_f = 1|X_f = x) = p(\hat{Y} = 1|X = x)$. Let $d_e(x, x') := d(x, x') + 2\epsilon$. Then $\mathbb{I}U_{D_d} (\hat{Y}_f, X) \leq 2\delta$.

### 3.3 Guaranteeing target task utility

We would also like to guarantee that using $X_f$ does not have too great an impact on target task utility compared to using $X$. We first define a measure of target task utility, known as risk.

**Definition 3.13 (Target variable risk).** Decision variable $\hat{Y}$ has risk with respect to target variable $Y$ defined as

$$R_Y(\hat{Y}) := \mathbb{E}_X[D(p(\hat{Y} = 1|X = x), p(Y = 1|X = x))]$$

where $D$ is a subadditive function.

We would like to quantify the effect on risk of using a decision variable $\hat{Y}_f$ using $X_f$ as input rather than decision variable $\hat{Y}$ using $X$ as input. To do this we introduce the following definition.

**Definition 3.14 (Average reconstruction error).** Let the average reconstruction error incurred by $f$ for input variable $X$ be

$$\mathbb{E}_X[d(x, f(x))].$$

We would like to guarantee that if $X_f$ approximately reconstructs $X$, measured using the average reconstruction error $\mathbb{E}_X[d(x, f(x))]$, then $R_Y(\hat{Y}_f)$ is not too much larger than $R_Y(\hat{Y})$. We show this in Theorem 3.15. The theorem requires a smoothness property of $\hat{Y}$, which is equivalent to our previous definition of individual fairness. Our result applies to any such $\hat{Y}$, so that we do not need to know $\hat{Y}$ in order to provide the guarantee.

**Theorem 3.15 (Guarantee for target task utility).** Suppose decision variable $\hat{Y}$ is individually fair with respect to $X$ and has risk $R_Y(\hat{Y})$ with respect to target variable $Y$. Let $\hat{Y}_f$ be the decision variable formed by setting $p(\hat{Y}_f = 1|X_f = x) = p(\hat{Y} = 1|X = x)$. Then

$$R_Y(\hat{Y}_f) \leq R_Y(\hat{Y}) + \mathbb{E}_X[d(x, f(x))].$$

### 3.4 Guaranteeing the cost of mistrust

The problem of making accurate predictions of the target variable with a group fairness penalty has previously been posed as minimizing a risk $R_{YS}$ defined as the weighted difference of cost-sensitive risks for the target and sensitive variables [15].

**Definition 3.16 (Combined risk $R_{YS}$ for group fairness and target utility [15]).** For a decision variable $\hat{Y}$, a target variable $Y$ and a sensitive variable $S$, let target variable risk have the form $R_Y(\hat{Y}) := c_Y p(Y = 0)p(\hat{Y} = 1|Y = 0) + (1 - c_Y) p(Y = 1)p(\hat{Y} = 0|Y = 1)$ and sensitive variable risk $R_S(\hat{Y}) := c_S p(S = 0)p(\hat{Y} = 1|S = 0) + (1 - c_S) p(S = 1)p(\hat{Y} = 0|S = 1)$

where $c_Y, c_S \in [0, 1]$. Let

$$R_{YS}(\hat{Y}) := R_Y(\hat{Y}) - \lambda R_S(\hat{Y})$$

where $\lambda$ is a non-negative constant.

We observe that $R_{YS}$ may be negative. Both $R_Y$ and $R_S$ are cost-sensitive risks. This form of $R_Y$, which differs from Definition 3.13, will be used for the remainder of the paper. Our choices of the parameters $c_Y$ and $c_S$ affect the evaluation of the decision variable. We saw in Theorem 3.4 that balanced error rate, which is a special case of $R_S$, is closely related to statistical parity. Furthermore, [15] showed that $R_S$ is also related to disparate impact.
Table 1: Characteristics of the Adult and ProPublica datasets

| Dataset   | Y = 1 definition | S = 1 definition | p(Y = 1) | p(S = 1) | p(Y = 1|S = 1) | p(Y = 1|S = 0) | SP(Y,S) | DH(Y,S) |
|-----------|-----------------|-----------------|---------|---------|---------------|---------------|--------|---------|
| Adult     | Income > $50,000 | Gender male     | 0.243   | 0.671   | 0.308         | 0.111         | 0.197  | 0.639   |
| ProPublica| Reoffended within two years | Ethnicity | African-American | 0.639 | 0.587 | 0.439 | 0.148 | 0.252 |

The decision variable minimizing $R_{YS}$ was found in [15].

**Definition 3.17 (Minimizer of $R_{YS}$ [15]).** Let $\hat{Y}^*$ be the minimizer of $R_{YS}$ for all decisions made using $X$ as input and let $\hat{Y}^*_f$ be the minimizer of $R_{YS}$ for all decisions made using $X_f$ as input.

**Lemma 3.18 (Analytic form of minimizer of $R_{YS}$ [15]).**

\[
\hat{Y}^*(x) = 1(p(Y = 1|X = x) - c_Y > \lambda(p(S = 1|X = x) - c_S)).
\]

Furthermore $R_{YS}(\hat{Y}^*) = \mathbb{E}_X[(c_Y - p(Y = 1|X = x) - \lambda(c_S - p(S = 1|X = x)))\hat{Y}^*(x)] + \max_{R_{YS}}$ where

\[
R_{YS} = (1 - c_y)p(Y = 1) - \lambda(1 - c_S)p(S = 1).
\]

We showed that organizational governance reasons motivate a separation of concerns between a data producer who ensures fairness by producing cleaned data, and a data user learns the target task from the cleaned data. Compared to the minimizer of $R_{YS}$ computed in [15] which may be applied in the case of a single trusted party, we incur some excess risk which we refer to as the cost of mistrust.

**Definition 3.19 (Cost of mistrust).** Let the cost of mistrust be defined as $R_{YS}(Y^*_f) - R_{YS}(\hat{Y}^*)$.

We would like to quantify what this cost of mistrust is. The form of $R_{YS}$, along with the fact that the minimizer of $R_{YS}$ is a thresholding function and is therefore not individually fair as per Definition 3.9, means that we cannot apply Theorem 3.15. We provide a different guarantee in Theorem 3.20 using smoothness properties of the target and sensitive variable distributions conditioned on the input, and the form of the minimizer of $R_{YS}$.

**Theorem 3.20 (Guarantee for cost of mistrust).** Suppose for some subadditive function $d$ we have the Lipschitz conditions $\forall x, x'$,

\[
|p(Y = 1|X = x) - p(Y = 1|X = x')| \leq l_Y(d(x, x'))
\]

and

\[
|p(S = 1|X = x) - p(S = 1|X = x')| \leq l_S(d(x, x')).
\]

Then

\[
R_{YS}(Y^*_f) - R_{YS}(\hat{Y}^*) \leq (l_Y + \lambda l_S)\mathbb{E}_X[d(x, f(x))].
\]

Considering the general case without Lipschitz assumptions, the cost of mistrust may take on values in the range $[0, \max_{R_{YS}} - R_{YS}(\hat{Y}^*)]$ as shown in Figure 2. In example (a), $R_{YS}(\hat{Y}^*) - R_{YS}(\hat{Y}^*) = 0$, so the minimum cost of mistrust (zero) is achieved. This is achieved if and only if $\hat{Y}^*$ agrees for all points $x$ mapped to the same value $f(x)$, where invertible maps $f$ are a special case. In example (b), $R_{YS}(Y^*_f) - R_{YS}(\hat{Y}^*) = \max_{R_{YS}} - R_{YS}(\hat{Y}^*)$ so the maximum cost of mistrust is achieved for a given value of $R_{YS}(\hat{Y}^*)$.

4 Practical Approach

Our theoretical results have shown us the desirable properties of a representation function $f$ in terms of both fairness and utility. In practice, we would like to use real-world data to construct such a representation function and estimate the performance measures identified in Section 3.

4.1 Learning a representation function

We encode the desirable properties of a representation function $f$ in a cost function $J(f)$, shown in Definition 4.1, which we would like to minimize.

**Definition 4.1 (Cost of representation function $f$).** Let $f$ be function transforming input variable $X$ to $X_f$. Let $d$ be a function measuring reconstruction error. Let $\hat{S}_f$ be an estimate of sensitive variable $S$ using $X_f$ and let $J'(\hat{S}_f, S)$ be a sensitive variable cost function. The cost of $f$ is

\[
J(f) := \mathbb{E}_x[d(x, f(x))] - \lambda J'(\hat{S}_f, S)
\]

where $\lambda$ is a non-negative constant.

The reconstruction error term promotes target task utility and individual fairness,\(^6\) while the $J'$ term promotes group fairness since we would like to make it difficult to estimate $S$ from $X_f$. The parameter $\lambda$ controls the trade-off between the two terms.

We propose the following specific form of the cost function, which facilitates the optimization of $f$. We use Euclidean distance for the reconstruction error

\[
d(x, f(x)) := \|x - f(x)\|_2
\]

and cross-entropy for the sensitive variable cost function

\[
J'(\hat{S}_f, S) := -\mathbb{E}_x[p(S = 1|X = x)\log_2 p(\hat{S}_f = 1|X = x)] + p(S = 0|X = x)\log_2 p(\hat{S}_f = 0|X = x).
\]

We estimate and attempt to minimize $J(f)$ over a finite training sample. We estimate $f$ using a fully-connected neural network with one softplus\(^6\) hidden layer of 100 units and a linear output layer with the same number of units as the input layer. Hence $f$ is an autoencoder learned using a training objective given by $J$. We estimate $\hat{S}_f$ using a fully-connected neural network with one softplus hidden layer of 100 units and a single sigmoidal output unit. The output layer of $f$, which corresponds to the variable

\[^6\]Another variant would include a $p(d(x, f(x)) > e)$ term to enforce individual fairness. We omit this term for simplicity, since the reconstruction error term should already promote individual fairness.

\[^6\]The softplus activation function is $\text{softplus}(x) = \log_e(1 + e^x)$. 
$X_f$, is the input layer for the network estimating $\hat{S}_f$. We alternate updates of the weights in the $f$ and $\hat{S}$ networks. The scheme is comparable to [8], although we do not use the target variable $Y$ when learning $f$ since in general this may not be accessible to the data producer and is not required to learn a representation function with the properties we desire. We optimize the networks using 100 iterations of the Adam Optimizer with a learning rate of 0.0001 and a batch size of 100. We implemented the model in Python using the TensorFlow library.

4.2 Learning a decision variable

The results presented in Section 3 show that it is possible to provide performance guarantees for a representation function for any target task. In practice, we would like to understand how a decision variable constructed using cleaned input $X_f$ performs on the target task of interest. We are also interested to compare this to a decision variable constructed using the original input $X$.

Using Lemma 3.18, we know the form of the minimizer of the combined risk $R_{YS}$ for group fairness and target utility. If we are only interested in the target variable risk $R_T$, this equals $R_{YS}$ in the case $\lambda = 0$. We produce decision variable $\hat{Y}_f$ by applying the analytic form of $\hat{Y}_f^*$ and separately training networks to estimate $p(Y|X_f)$ and $p(S|X_f)$. We perform this estimation by minimizing the cost functions $f^*(\cdot, S)$ and $f^*(\cdot, Y)$ respectively over the cleaned training set, in both cases using a fully-connected neural network with one softplus hidden layer of 100 units and a single sigmoidal output unit, and the same training settings previously described. In the same way we produce decision variable $\hat{Y}$ by applying the analytic form of $\hat{Y}^*$ and estimating $p(Y|X)$ and $p(S|X)$ using the original training set.

4.3 Experiments

We conducted experiments to test the representation learning approach described in Section 4.1 and to illustrate how our theoretical results suggest useful performance measures for the learned representation function $f$. We used two datasets that are well-known in the fair machine learning literature (for example in [3]), the UCI Adult and ProPublica recidivism datasets. In both datasets we used a training set of 70% of the data and test on the remaining 30%. These datasets are summarized in Table 1, including statistics of interest computed for the training set.

The Adult dataset contains financial and demographic information compiled from a census about 32561 people and contains 110 input columns once categorical features are binarized. We selected the sensitive variable as gender and the target variable as whether the person’s income is at least $50,000. This setting is similar to a situation where a financial institution makes an algorithmic decision about whether to grant an individual a loan based on a prediction of their income. The training set estimates of $SP(Y, S)$ and $DI(Y, S)$ indicate that there is a clear relationship between the sensitive and target variables.

The ProPublica dataset contains information about 7214 criminal offences committed in Broward County, Florida and contains 79 input columns once categorical features are binarized. We selected the sensitive variable as whether the person is of African-American ethnicity and the target variable as whether the person reoffended.

---

7 Available at https://archive.ics.uci.edu/ml/datasets/adult and https://github.com/propublica/compas-analysis respectively.
8 We processed the free text crime description column by converting it to a categorical variable where descriptions occurring at least 20 times have their own category (covering 82.9% of all offences) and all other descriptions are marked as ‘other’. The categorical variable is then binarized.
within two years. This setting is similar to a situation where sentencing decisions are made based on an algorithmic assessment of the individual’s likelihood of reoffending, as in the case of the COMPAS tool [12]. The training set estimates of $SP(Y,S)$ and $DI(Y,S)$ indicate that there is a relationship between the sensitive and target variables, but weaker than in the case of the Adult dataset.

We present the results of our experiments in Figure 3 measured over the test set. We plot a series of performance measures of interest for both datasets, varying the parameter $\lambda$ which controls the extent to which group fairness is considered in the training objective for learning $f$. For all measures, lower is better.

The top row contains measures of group and individual fairness. We compute the statistical parity $SP$ and normalized disparate impact $DI$ using estimates of $\hat{Y}^{BER}_f$ and $\hat{Y}^{DI}$ respectively (see Theorems 3.4 and 3.8). For both datasets, these two measures decrease as $\lambda$ increases, which provides evidence that group fairness is improving. We measure the large reconstruction error rate $p(\|x - f(x)\|_2 > \epsilon)$, which is of interest due to its relationship with individual fairness established in Theorem 3.12. A smaller value of this quantity means that greater individual fairness is possible using $X_f$. We select $\epsilon$ to be $\frac{1}{10}$ of the average training set value of $\|x\|_2$. The results show that there is a trade-off between group and individual fairness, with $p(\|x - f(x)\|_2 > \epsilon)$ increasing and hence individual fairness becoming worse as $\lambda$ increases.

The bottom row contains measures of target task utility. We show the average reconstruction error, which is related to a particular form of target task risk as shown in Theorem 3.15, which increases with $\lambda$ since the impact of protecting the sensitive variable has a greater influence on $f$. We estimate the target cost-sensitive risk $R_f$ (see Definition 3.16) with $\hat{Y}_f$ using the decision variable $\hat{Y}_f$ described in Section 4.2. Interestingly, this risk only slightly increases with $\lambda$ for both datasets, and for the Adult dataset the effect is particularly small. This indicates that removing knowledge of $S$ may only slightly affect target variable prediction performance if other factors in the input data that are not correlated with $S$ are predictive of $Y$. We also estimate the cost of mistrust with parameters $c_S = c_Y = \frac{1}{2}$ (see Definitions 3.16 and 3.19) using the decision variables $\hat{Y}_f$ and $\hat{Y}$ described in Section 4.2, which increases with $\lambda$ as $X_f$ differs more from $X$.

5 CONCLUSION AND FUTURE WORK

We proposed learning a representation function to prepare a cleaned version of the input data made available to a decision-making data user. We showed that this approach improves group fairness using common measures such as statistical parity and disparate impact. Furthermore, we showed that it is possible to quantify the cost of such a representation function on individual fairness and target task utility. We also evaluated the cost of mistrust, by comparing performance using our approach with a baseline scenario of a trusted data user given access to the original input data and sensitive variable at training time.

We used our results to motivate a practical algorithm that learns representation functions to achieve both fairness and target task utility. Using real-world datasets in financial and criminal justice domains, we found that increasing the fairness parameter $\lambda$ controls the trade-off between group fairness on the one hand, and individual fairness and target task utility on the other. The results highlight the value of the conceptual distinction between group and individual fairness. They also indicate that when the target variable can be predicted using other factors not correlated with the sensitive variable, group fairness can be achieved without too great a cost to target task utility.

Promising future directions include exploring other individual similarity measures besides Euclidean distance and analyzing the ability to extend theoretical guarantees to a finite sample regime. From a regulatory perspective it is interesting to more closely consider cases where the governance model and separation of concerns we propose would be suitable in practice. Protecting the anonymity of user records is closely related to the task of protecting a sensitive variable, indicating the potential to use analysis developed in the context of fairness to privacy applications.

6 PROOFS OF THEORETICAL GUARANTEES

We present proofs of the theoretical results stated in Section 3.

Proof of Theorem 3.4.

\[
SP(\hat{Y}_f, S) = 1 - 2BER(\hat{Y}^{BER}_f, S).
\]

Therefore it is sufficient to show

\[
2BER(\hat{Y}^{BER}_f, S) \geq \frac{H_b^{-1}(H(S|X_f))}{\max(p(S=1), p(S=0))}.
\]

It is helpful to observe that $BER(\hat{Y}^{BER}_f, S) \leq \frac{1}{2}$. We consider the cases (a) $p(S=1) = \frac{1}{2}$ and (b) $p(S=1) > \frac{1}{2}$.

(a) First consider the case $p(S=1) = \frac{1}{2}$. If in addition we have

\[
2BER(\hat{Y}^{BER}_f, S) \geq \frac{1}{2}
\]

we have

\[
2BER(\hat{Y}^{BER}_f, S) \geq \frac{1}{2p(S=0)}\left(1 - \frac{1}{2\max(p(S=1), p(S=0))}\right)\]

\[
\geq \frac{H_b^{-1}(H(S|X_f))}{\max(p(S=1), p(S=0))}.
\]

Otherwise, using Corollary 8 from [19]

\[
H(S|\hat{Y}_f) \leq H_b(p(S=0)2BER(\hat{Y}^{BER}_f, S)).
\]

Therefore

\[
2BER(\hat{Y}^{BER}_f, S) \geq \frac{H_b^{-1}(H(S|X_f))}{p(S=0)}\frac{H_b^{-1}(H(S|\hat{Y}_f))}{\max(p(S=1), p(S=0))}.
\]
where for the first inequality we used the fact that $H^{-1}_b$ is a non-decreasing function.

(b) The case $p(S = 1) > \frac{1}{2}$ is similar. If in addition

$$p(S = 1) \geq \frac{1}{2},$$

we have

$$2 \text{BER}(\hat{Y}_f, S) \geq \frac{1}{2p(S = 1)}$$

Using the first result established above

$$\leq \frac{H^{-1}_b(1 - \eta_f) + H^{-1}_b(1 - \eta_f)}{\max(p(S = 1), p(S = 0))}.$$ 

Again since $\eta_f \geq p(S = 1|x = x)$, with equality when $\hat{Y}_f = \hat{Y}_f^{DI}$

$$= 1 - \frac{p(S = 1) - p(S = 0)\eta_f}{p(S = 0)\eta_f}.$$ 

Simplifying

\[\square\]

**Proof of Theorem 3.12.** Consider points $x$ and $x'$ drawn independently at random using the input variable $X$.

$$p(\hat{Y}_f = 1|x = x)$$

$$= p(\hat{Y}_f = 1, X_f = f(x)|X = x) + p(\hat{Y}_f = 1, X_f \neq f(x)|X = x)$$

law of total probability

$$= p(\hat{Y}_f = 1, X_f = f(x)|X = x)$$

$f$ is deterministic

$$= p(X_f = f(x)|X = x)p(\hat{Y}_f = 1|X_f = f(x), X = x)$$

definition of conditional probability

$$= p(\hat{Y}_f = 1|X_f = f(x), X = x)$$

from definition of $X_f$

$$= p(\hat{Y}_f = 1|X_f = f(x))$$

$\hat{Y}_f$ is conditionally independent of $X$ given $X_f$

$$= p(\hat{Y}_f = 1|X = x'), p(\hat{Y}_f = 1|X = f(x'))$$

With probability at least $1 - \delta, d(x, f(x)) \leq \epsilon$. Similarly, with probability at least $1 - \delta, d(x', f(x')) \leq \epsilon$. By the union bound, both statements hold with probability at least $1 - \delta$.

Having established these facts, we prove the result:

$$D(p(\hat{Y}_f = 1|x = x), p(\hat{Y}_f = 1|x = x'))$$

$$= D(p(\hat{Y}_f = 1|X = f(x)), p(\hat{Y}_f = 1|X = f(x')))$$

using the first result established above

$$\leq D(p(\hat{Y}_f = 1|X = f(x)), p(\hat{Y}_f = 1|X = x) + D(p(\hat{Y}_f = 1|X = x), p(\hat{Y}_f = 1|X = f(x')))$$

triangle inequality

$$\leq \epsilon + D(p(\hat{Y}_f = 1|X = x), p(\hat{Y}_f = 1|X = f(x')))$$

Using the triangle inequality

$$\leq 2\epsilon + d(x, x').$$

Once again using the high probability bound above plus the fact that $\hat{Y}$ is individually fair with respect to $X$.

We have thus established $\Delta_{D, \epsilon}(\hat{Y}_f, X) \leq 2\delta$ as required.
Proof of Theorem 3.20.

\[ R_{YS}(\hat{Y}^*) = R_{YS}(\hat{Y}^*) \]

\[ \leq R_{YS}(\hat{Y}) - R_{YS}(\hat{Y}^*) \]

using the definition \( \hat{Y}_f(x) := \hat{Y}_f(x) \)

\[ = R_Y(\hat{Y}) - R_Y(\hat{Y}^*) - \lambda(R_S(\hat{Y}_f) - R_S(\hat{Y}^*)) \]

Definition of \( R_{YS} \)

\[ = \mathbb{E}_x[(c_Y - p(S = 1|X = x))((\hat{Y}_f(x) - \hat{Y}^*(x)))] \]

We may write

\[ R_Y(\cdot) = (1-c)p(Y = 1) + \mathbb{E}_x[(c_Y - p(Y = 1|X = x))p(\cdot = 1|X = x)] \]

and

\[ R_S(\cdot) = (1-c_S)p(S = 1) + \mathbb{E}_x[(c_S - p(S = 1|X = x))p(\cdot = 1|X = x)] \]

(Lemma 10 from [15]). Then we apply linearity of expectation.

\[ = \mathbb{E}_x[(c_Y - p(Y = 1|X = x)) - \lambda(c_S - p(S = 1|X = x))((\hat{Y}_f(x) - \hat{Y}^*(x))] \]

Factorizing, linearity of expectation

\[ \leq \mathbb{E}_x[(l_Y + \lambda l_S)d(x, f(x))] \]

See below for discussion

\[ = (l_Y + \lambda l_S)\mathbb{E}_x[d(x, f(x))] \]

factorizing Discussion of second last line: consider the minimizer of \( R_{YS} \) from Lemma 3.18. For any \( x \) such that \( \hat{Y}_f(x) \neq \hat{Y}^*(x) \), there must be some point \( x' \) such that

\[ c_Y - p(Y = 1|X = x') - \lambda(c_S - p(S = 1|X = x')) = 0 \]

and \( d(x, x') \leq d(x, f(x)) \). Hence, applying the Lipschitz conditions we have:

\[ c_Y - p(Y = 1|X = x) - \lambda(c_S - p(S = 1|X = x)) \leq c_Y - p(Y = 1|X = x') + l_Y d(x, x') - \lambda(c_S - p(S = 1|X = x') - l_S d(x, x')) \]

\[ = (l_Y + \lambda l_S)d(x, x') \]

\[ \leq (l_Y + \lambda l_S)d(x, f(x)). \]

\[ \square \]

REFERENCES

[1] Churu C Aggarwal and S Yu Philip. 2008. A General Survey of Privacy-Preserving Data Mining Models and Algorithms. In Privacy-Preserving Data Mining. 11–32.

[2] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. Proceedings of 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning (2017).

[3] Flavio P Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Data Pre-Processing for Discrimination Prevention. arXiv preprint arXiv:1704.03554 (2017).

[4] United States Equal Opportunity Employment Commission. 1978. Uniform Guidelines on Employee Selection Procedures. (1978).

[5] Thomas M Cover and Joy A Thomas. 2012. Elements of Information Theory.

[6] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings. Proceedings on Privacy Enhancing Technologies, 1(2015), 92–112.

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. 214–226.

[8] Harrison Edwards and Amos Storkey. 2016. Censoring Representations with an Adversary. International Conference on Learning Representations (2016).

[9] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 259–268.

[10] Bryce Goodman and Seth Flaxman. 2016. EU Regulations on Algorithmic Decision-Making and a ‘Right to Explanation’. In ICMIL Workshop on Human Interpretability in Machine Learning.

[11] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In Advances in Neural Information Processing Systems 29. 3315–3323.

[12] Northpoint Inc. 2012. Practitioners Guide to COMIXS. (2012). http://www.northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf

[13] James Johndrow and Krishan Lum. 2017. An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction. arXiv preprint arXiv:1703.04957 (2017).

[14] Christos Louizos, Kevin Swersky, Yuja Li, Richard Zemel, and Max Welling. 2016. The Variational Fair Autoencoder. International Conference on Learning Representations (2016).

[15] Aditya Menon and Robert Williamson. 2017. The cost of fairness in classification. arXiv preprint arXiv:1705.09055 (2017).

[16] John Morrison. 2014. The Social License: How to Keep Your Organization Legitimate.

[17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In Proceedings of the 26th International Conference on World Wide Web. 1171–1180.

[18] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In Proceedings of the 30th International Conference on Machine Learning. 325–333.

[19] Ming-Jie Zhao, Narayanan Edakunni, Adam Pocock, and Gavin Brown. 2013. Beyond Fano’s Inequality: Bounds on the Optimal F-score, BER, and Cost-Sensitive Risk and their Implications. The Journal of Machine Learning Research 14, 1 (2013), 1033–1090.

[20] Indre Zliobaite. 2015. A Survey on Measuring Indirect Discrimination in Machine Learning. arXiv preprint arXiv:1511.00148 (2015).