Costs and Benefits of Fair Representation Learning

Daniel McNamara, Cheng Soon Ong and Robert C. Williamson
Australian National University and CSIRO Data61
Canberra, ACT, Australia

Abstract

Machine learning algorithms are increasingly used to make or support important decisions about people's lives. This has led to interest in the problem of fair classification, which involves learning to make decisions that are non-discriminatory with respect to a sensitive variable such as race or gender. Several methods have been proposed to solve this problem, including fair representation learning, which cleans the input data used by the algorithm to remove information about the sensitive variable. We show that using fair representation learning as an intermediate step in fair classification incurs a cost compared to directly solving the problem, which we refer to as the *cost of mistrust*. We show that fair representation learning in fact addresses a different problem, which is of interest when the data user is not trusted to access the sensitive variable. We quantify the benefits of fair representation learning, by showing that any subsequent use of the cleaned data will not be too unfair. The benefits we identify result from restricting the decisions of adversarial data users, while the costs are due to applying those same restrictions to other data users.

1 Introduction

Machine learning algorithms are used to make or support decisions in a wide variety of contexts including financial and judicial risk assessments, applicant screening for employment, and online ad selection. Concerns about the fairness of these algorithms have arisen as a result (O'Neil 2017; Barocas and Selbst 2016; Angwin et al. 2016; Datta, Tschantz, and Datta 2015). Decisions made by machine learning algorithms typically cannot be controlled or interpreted as straightforwardly as those made by rule-based systems. Furthermore, artefacts of previous discrimination in an algorithm's training data may affect its decisions. Researchers have responded by developing techniques to incorporate fairness into the design of machine learning algorithms (Barocas, Hardt, and Narayanan 2018; Zliobaite 2015; Romei and Ruggieri 2014). While these techniques often focus on achieving *group fairness* – i.e. not discriminating against particular groups – another important consideration is *individual fairness* – i.e. giving similar treatment to individuals who are similar (Dwork et al. 2012).

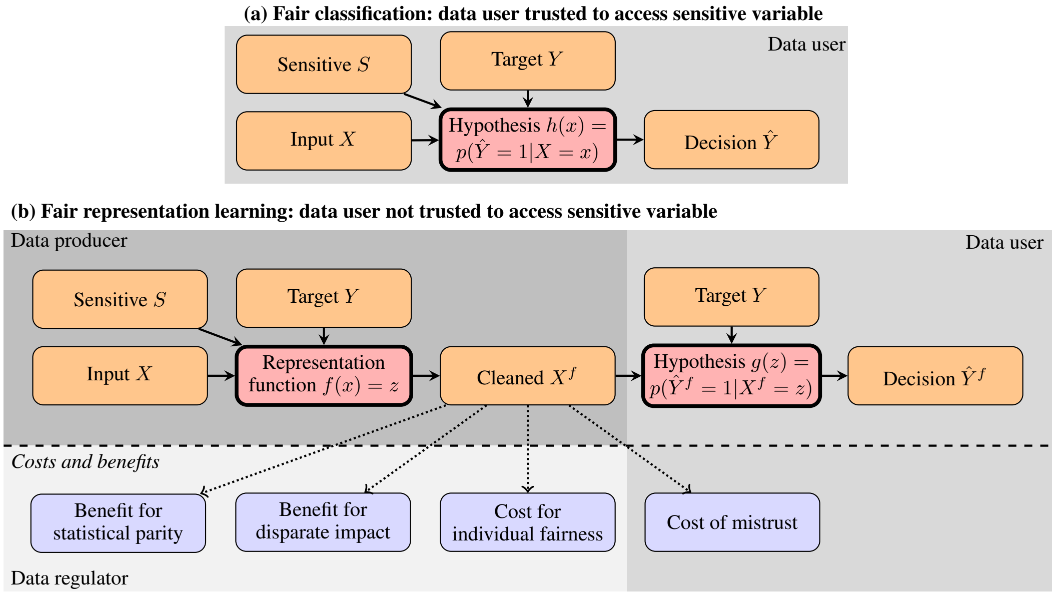
The problem of fair classification involves making a *decision* (e.g. whether to grant a loan) based on an *input* (e.g. individual financial and demographic information) which accurately predicts a *target* of interest (e.g. loan default), while at the same time avoiding discrimination on the basis of an individual's group membership (e.g. race, gender) encoded in a *sensitive* variable. The data user is trusted to access the sensitive variable in training and is responsible for making decisions that appropriately consider accuracy and fairness.

In contrast (see Figure 1), the problem of fair representation learning involves producing a *cleaned* version of the input which remains useful for predicting the target, but suppresses information which could be used to discriminate based on the sensitive variable. We now assume the data user is not trusted to access the sensitive variable in training, which may be appropriate if the data user could be either *adversarial*, i.e. interested in being unfair, or *indifferent*, i.e. interested only in target accuracy (Madras et al. 2018). This problem setting involves three parties: a *data producer* who cleans 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, when 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 within an organization, this separation of concerns has the advantage of providing checks and balances.

1.1 Contributions of This Paper

This paper offers contributions that are have both scientific and policy significance, and are technically novel.

Scientific significance: A plethora of methods use fair representation learning (Zemel et al. 2013; Feldman et al. 2015; Edwards and Storkey 2016; Louizos et al. 2016; Johndrow and Lum 2017; Beutel et al. 2017; Madras et al. 2018) as a *technique* for fair classification. Recent work (Menon and Williamson 2018) has solved in analytical form a canonical version of the fair classification problem. Is fair representation learning then to be relegated to a sub-optimal technique for a problem better solved through other means? Developing more fair representation learning techniques does not address this question. Instead, we show that fair representation learning in fact solves a different *problem* – i.e. how to guarantee that decisions made by an untrusted data user can be accurate but will not be unfair – and quantify the costs and benefits of such representations in terms of fairness and utility. This represents a progression in our scientific understanding, given that this problem had never previously been formally posed or analysed.


(a) Fair classification: data user trusted to access sensitive variable

(b) Fair representation learning: data user not trusted to access sensitive variable

Data user
Sensitive S
Target Y
Input X
Hypothesis $h(x) = p(\hat{Y} = 1|X = x)$
Decision $\hat{Y}$
Data producer
Data user
Sensitive S
Target Y
Input X
Representation function $f(x) = z$
Cleaned X^f
Target Y
Hypothesis $g(z) = p(\hat{Y}^f = 1|X^f = z)$
Decision $\hat{Y}^f$
Costs and benefits
Benefit for statistical parity
Benefit for disparate impact
Cost for individual fairness
Cost of mistrust
Data regulator

Figure 1: Summary of (a) fair classification and (b) fair representation learning, showing train time data processing for both, and costs and benefits of (b).

Policy significance: Our approach makes possible a governance model involving a separation of concerns between a data producer, data user and data regulator (previous work assumes a single trusted data user). The model enables a regulator to guarantee fairness even if the data user is adversarial. This is an advance in the regulation of algorithmic fairness, given that no alternatives currently exist in the realistic setting where a data user is not trusted to be fair.

Novel technical results: We formalize the problem of fair representation learning as distinct from fair classification (Section 3). By stating the data producer's optimization problem in (5) and showing that a proxy problem can be solved without access to the target variable (Theorem 1), we derive a principled way to select a fair representation learning objective function (this is heuristic in prior work).

We present a novel quantification of the costs of using a given representation (Section 4), a topic which had not previously been investigated. We identify costs both in terms of the accuracy-fairness trade-off (i.e. the *cost of mistrust* given in closed form in Theorem 2 and bounded without requiring access to the target variable in Theorem 3), and in terms of individual fairness (Theorem 4).

We present novel guarantees of the benefits of a given representation (Section 5). We do this for two common measures of fairness: statistical parity (Theorem 5) and disparate impact (Theorem 6), by computing the unfairness of an optimal adversary. Conditioning on the target variable, our analysis can be used to guarantee quantified versions of equality of opportunity and equalized odds as well.

We provide a proof idea in the main paper for each result. The Appendices contain complete proofs, along with a summary of the problems, costs and benefits we consider, and examples and experiments.

2 Background

We provide a brief summary of relevant work on parity-based definitions of fairness, fair classification, and fair representation learning.

Parity measures of quantitative fairness compare an algorithm's average decisions for different groups, for example taking their difference – known as *statistical parity* (Calders and Verwer 2010; Dwork et al. 2012) – or ratio – known as *disparate impact* (United States Equal Opportunity Employment Commission 1978; Feldman et al. 2015). We may wish to compute a parity measure only on a population subset. Constructing subsets using values of the target variable yields variants such as *equality of opportunity* and *equalized odds* (Hardt, Price, and Srebro 2016) – i.e. avoiding *disparate mistreatment* (Zafar et al. 2017a). However, when the training data labels are themselves affected by discrimination, conditioning on the target variable may not be suitable (Zafar et al. 2017a). If the population subset consists of individuals who are similar according to some metric, we have *individual fairness* – i.e. avoiding *disparate treatment* (Dwork et al. 2012; Mitchell and Shadlen 2018).

Methods for fair classification can be divided into *pre-processing* – i.e. fair representation learning – which modifies the data that the algorithm learns from (Zemel et al. 2013); *in-processing*, which modifies the algorithm's objective function to incorporate a fairness constraint or penalty (Menon and Williamson 2018; Donini et al. 2018; Zafar et al. 2017b; 2017a; Dwork et al. 2018; Bechavod and Ligett 2017); and *post-processing*, which modifies the predictions produced by the algorithm (Hardt, Price, and Srebro 2016). We show that fair representation learning in fact addresses a distinct problem, which is of interest when the data user is not trusted to access the sensitive variable.

A common approach to fair representation learning is to clean the data such that conditioning on different sensitive variable values yields similar distributions (Louizos et al. 2016; Feldman et al. 2015; Johndrow and Lum 2017). Adversarial approaches (Edwards and Storkey 2016; Beutel et al. 2017; Madras et al. 2018) use a neural network to learn a representation function such that an adversary network cannot accurately predict the sensitive variable from the cleaned data. A problem variant, where the target is also modified and the input is discrete, has been formulated as a convex optimization (Calmon et al. 2017). What existing approaches typically do not offer (Theorem 4.1 from Feldman et al. 2015 is an exception) is a guarantee that all uses of the cleaned data will be fair, or a quantification of the costs of the cleaning process. We seek to provide a stronger theoretical foundation for fair representation learning. This objective is similar in spirit to that of privacy aware learning, which is concerned with the mathematical trade-off between the privacy and utility of data (Wainwright, Jordan, and Duchi 2012).

3 Fair Classification vs Fair Representation Learning

We introduce and compare the problems of fair classification and fair representation learning. This formal comparison is itself novel and is necessary for our subsequent analysis of the costs and benefits of fair representation learning.

3.1 Fair Classification

In fair classification (Figure 1(a)), the data user trains on samples of input variable X, target variable Y and sensitive variable S. The samples are drawn from a distribution over $\mathcal{X}\times\mathcal{Y}\times\mathcal{S}$, where $\mathcal{X}$ is the set of possible inputs, $\mathcal{Y}$ is the set of possible labels and $\mathcal{S}$ is the set of possible sensitive variable values. We focus on the setting where $\mathcal{Y}\in\{0,1\}$, corresponding to binary classification, and $\mathcal{S}\in\{0,1\}$, corresponding to some common sensitive variable examples such as gender or race. Let $\pi_Y := p(Y=1)$ and $\pi_S := p(S=1)$ be prior probabilities, and $\eta_Y(x) := p(Y=1|X=x)$ and $\eta_S(x) := p(S=1|X=x)$ be conditional probabilities, for the positive classes of Y and S.

The data user learns stochastic hypothesis $h : \mathcal{X}\to[0,1]$ which is used to construct decision variable $\hat{Y}\in\{0,1\}$, where $h(x) := p(\hat{Y}=1|X=x)$. At test time, the data user makes a decision using a sample of X, which may contain information about S. The quality of a hypothesis h in predicting Y can be measured by a risk $R_Y : [0,1]^{\mathcal{X}}\to[0,1]$, where we prefer hypotheses with a small value of $R_Y(h)$. A common choice is the cost-sensitive risk.

Definition 1 (Cost-sensitive risk (Elkan 2001; Zhao et al. 2013; Menon and Williamson 2018))**.** *The cost-sensitive risk of hypothesis h with respect to Y is*

$$\begin{aligned}R_Y(h) &:= \pi_Y(1-c_Y)p(\hat{Y}=0|Y=1)\\ &\qquad + (1-\pi_Y)c_Y p(\hat{Y}=1|Y=0)\end{aligned}$$

where $c_Y\in[0,1]$, $p(\hat{Y}=0|Y=1)$ is known as the false negative rate *and $p(\hat{Y}=1|Y=0)$ as the* false positive rate*.*

We also wish to ensure that the hypothesis we learn is fair. Two common fairness measures are statistical parity and disparate impact, which compare outcomes for different sensitive variable groups using their difference and ratio respectively. In the analysis that follows we focus on the case where statistical parity and disparate impact are computed on the joint distribution over $\mathcal{X}\times\mathcal{Y}\times\mathcal{S}\times\{0,1\}$. However, computing these measures only on part of the distribution yields other variants of interest, such as conditioning on $Y=1$ for quantified versions of equality of opportunity.

Definition 2 (Statistical parity (Calders and Verwer 2010; Dwork et al. 2012))**.** *The statistical parity of a hypothesis h is*

$$SP(h) := p(\hat{Y}=1|S=1) - p(\hat{Y}=1|S=0).$$

Definition 3 (Disparate impact (United States Equal Opportunity Employment Commission 1978; Feldman et al. 2015))**.** *The disparate impact of a hypothesis h is*

$$DI(h) := \frac{p(\hat{Y}=1|S=0)}{p(\hat{Y}=1|S=1)}.$$

Notice that $SP(h)\in[-1,1]$, with equality of outcome corresponding to 0, while $DI(h)\in[0,\infty)$, with equality of outcome corresponding to 1. In both cases we want a value that is neither too low nor too high. It has been shown that this is equivalent to requiring that h and the ‘anti-classifier’ $1-h$ both have values that are not too low (Appendix C of Menon and Williamson 2018).

The fair classification problem then naturally takes the form, for some $R_{\text{fair}}\in\{SP, DI\}$:

$$\min_{h\in\mathcal{H}} R_Y(h) \text{ subject to } \min[R_{\text{fair}}(h), R_{\text{fair}}(1-h)] \geq \tau, \quad (1)$$

where $\mathcal{H} := [0,1]^{\mathcal{X}}$ and τ is a constant measuring the required level of fairness. For DI, $\tau\in[0,\infty)$, while for SP, $\tau\in[-1,0]$ since $SP(1-h) = -SP(h)$.

It has been shown that a constraint on SP or DI of the type in (1) is equivalent to a constraint on a cost sensitive risk with respect to S (see Lemmas 1 and 2 of Menon and Williamson 2018 for details). Using Definition 1, this cost sensitive risk is written as:

$$\begin{aligned}R_S(h) &:= \pi_S(1-c_S)p(\hat{Y}=0|S=1)\\ &\qquad + (1-\pi_S)c_S p(\hat{Y}=1|S=0), \quad (2)\end{aligned}$$

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

It is more convenient to work with an unconstrained variant of the fair classification problem:

$$\min_{h\in\mathcal{H}}[R_Y(h) - \lambda R_S(h)], \quad (3)$$

where λ is a constant (not necessarily non-negative) controlling the trade-off between accuracy with respect to Y and fairness with respect to S. It has been shown (Menon and Williamson 2018) that for some choice of λ, some solution to (3) is also a solution to (1).

Definition 4 (Optimal fair classification)**.** *Let the combined risk $R_{YS}(h) := R_Y(h) - \lambda R_S(h)$. Let $R_{YS}(h^*)$ be the value of* (3) *and h^* be a corresponding hypothesis.*

Subsequently we will compare optimal fair classification to the case where we instead use fair representation learning as an intermediate step in fair classification.

3.2 Fair Representation Learning

In fair representation learning (Figure 1(b)), the data producer trains on samples of X, S and Y (we also examine the case where the data producer does not access Y), and learns the representation function $f : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is the set of possible cleaned variable values. The data producer samples X and applies f to each sample to produce cleaned variable $X^f := f(X)$. The data producer learns f so that X^f is still useful for predicting Y but suppresses information about S. Let $\eta_Y^f(z) := p(Y = 1|X^f = z)$ and $\eta_S^f(z) := p(S = 1|X^f = z)$ be conditional probabilities of the positive classes of Y and S induced by f. The data user trains on samples of X^f and Y and learns a stochastic hypothesis $g : \mathcal{Z} \to [0, 1]$, which is used to construct modified decision variable $\hat{Y}^f \in \{0, 1\}$ where $g(z) := p(\hat{Y}^f = 1|X^f = z)$. At test time, the data producer samples X and passes it through f to produce a sample of X^f, from which the data user makes a decision.

When the data user is not trusted, we are interested in constraining how unfair an *adversarial* user can be with the cleaned data. As in the fair classification case, this is equivalent to a constraint on an adversary's cost-sensitive risk with respect to S. We are also interested in ensuring that the cleaned data is still useful for predicting the target. We are therefore interested in the following problem:

$$\min_{f \in \mathcal{F}} R_Y(g_Y^* \circ f) \text{ subject to } R_S(g_S^* \circ f) \geq \tau, \tag{4}$$

where τ is a constant measuring the required level of fairness, $\circ$ is function composition, $g_Y^* \in \arg\min_{g \in \mathcal{G}} R_Y(g \circ f)$ is an optimal indifferent user of the cleaned data, $g_S^* \in \arg\min_{g \in \mathcal{G}} R_S(g \circ f)$ is an optimal adversary using the cleaned data, $\mathcal{G} := [0, 1]^{\mathcal{Z}}$ and $\mathcal{F} := \mathcal{Z}^{\mathcal{X}}$.

It is more convenient to work with the following unconstrained problem variant:

$$\min_{f \in \mathcal{F}} [R_Y(g_Y^* \circ f) - \lambda R_S(g_S^* \circ f)]. \tag{5}$$

Using the form of the minimum cost-sensitive risk (Zhao et al. 2013), we may express the terms in (5) as follows:

$$\begin{aligned} &R_Y(g_Y^* \circ f) \\ &= \mathbb{E}_{X^f}[\min((1 - c_Y)\eta_Y^f(X^f), c_Y(1 - \eta_Y^f(X^f)))] \end{aligned} \tag{6}$$

$$\begin{aligned} &R_S(g_S^* \circ f) \\ &= \mathbb{E}_{X^f}[\min((1 - c_S)\eta_S^f(X^f), c_S(1 - \eta_S^f(X^f)))]. \end{aligned} \tag{7}$$

Adversarial neural networks have previously been used to estimate g_S^* and g_Y^* (Edwards and Storkey 2016; Beutel et al. 2017; Madras et al. 2018). We observe that (6) and (7) simplify the fair representation learning cost function (5) by removing the two inner minimizations. Of course, there remains the task of estimating the underlying distribution and computing the outer minimization.

We focus on the case where the data producer learns a representation without using the target variable. This allows a single fair representation to be learned that can be used for multiple target tasks. It also covers the situation where the data producer does not have access to the target variable. For example, Y contains commercially confidential information (e.g. defaults on a specific type of loan) known to the data user (e.g. a bank) but not the data producer (e.g. a credit bureau). Furthermore, we focus on the case $\mathcal{Z} = \mathcal{X}$ is a Euclidean space, which facilitates our analysis and covers many practical applications. In this case, we define *average reconstruction error* and show its use as a proxy for target task performance.

Definition 5 (Average reconstruction error)**.** *Suppose $\mathcal{Z} = \mathcal{X}$ is a Euclidean space. Let $\mathbb{E}_X \|X - f(X)\|_2$ be the average reconstruction error of f with respect to X, where $\|\cdot\|_2$ is the Euclidean vector norm.*

Assuming the data producer does not access the target variable, we propose the following variant of the fair representation learning problem:

$$\min_{f \in \mathcal{F}} [\mathbb{E}_X \|X - f(X)\|_2 - \lambda R_S(g_S^* \circ f)]. \tag{8}$$

We relate (8) and (5) as follows. This result allows us to select a principled objective function for the data producer.

Theorem 1 (Fair representation learning without accessing target variable)**.** *Suppose $\mathcal{Z} = \mathcal{X}$ and we have the Lipschitz condition that for some non-negative constant l_Y*

$$\forall x, x' \in \mathcal{X}, |\eta_Y(x) - \eta_Y(x')| \leq l_Y \|x - x'\|_2.$$

Then any $f \in \mathcal{F}$ minimizing

$$\mathbb{E}_X \|X - f(X)\|_2 - \lambda R_S(g_S^* \circ f)$$

also minimizes an upper bound on

$$R_Y(g_Y^* \circ f) - l_Y \lambda R_S(g_S^* \circ f).$$

Proof idea. We upper bound $R_Y(g_Y^* \circ f) - l_Y \lambda R_S(g_S^* \circ f)$ by re-expressing the risks using Lemma 9 from Menon and Williamson 2018, and making use of the Lipschitz condition. We then observe that the f minimizing this upper bound also minimizes $\mathbb{E}_X \|X - f(X)\|_2 - \lambda R_S(g_S^* \circ f)$. □

4 Costs of Fair Representation Learning

We now identify two costs of fair representation learning relative to the case of a single trusted data user. These costs are incurred by data users who optimally incorporate fairness into their decisions, as well as individuals about whom these users make decisions. The first cost is the difference in the optimal fairness-accuracy trade-off available with the cleaned data compared to the original data, known as the *cost of mistrust*. This cost is of interest to the data user – as well as potentially the data regulator. The second cost is for individual fairness, which is primarily of interest to the data regulator. We show that they can both be estimated by a data producer without accessing the target variable.

4.1 Cost of Mistrust

Suppose that after cleaning the data with the representation function f, we solve the following fair classification problem, which is equivalent to (3) but using the cleaned data.

$$\min_{g\in\mathcal{G}}[R_Y(g\circ f)-\lambda R_S(g\circ f)] \tag{9}$$

Definition 6 (Cost of mistrust)**.** *The cost of mistrust for a representation function f is $R_{YS}(g^*\circ f)-R_{YS}(h^*)$, where g^* and h^* are hypotheses minimizing* (9) *and* (3) *respectively and the value of λ is the same in both equations.*

The cost of mistrust is non-negative because f restricts the hypothesis class. If $\lambda=0$ in (9) and (3), f may incur a cost for the target accuracy of the indifferent user, which seems unsurprising. However, for general λ we see that f may also incur a cost for fair classification. Without access to S the data user has no way to estimate $R_S(g\circ f)$ in (9). However, even if they could somehow guess this quantity, f may create a suboptimal trade-off between fairness and accuracy compared to the trade-off available to a trusted data user using the original input. See Appendix C for examples where the cost of mistrust is either zero or positive.

We now show in Theorem 2 that we can express the cost of mistrust in analytical form. We build on an existing result (Proposition 4 of Menon and Williamson 2018) which yields the expressions $h^*(x)=\mathbf{1}(\eta_Y(x)-c_Y\geq\lambda(\eta_S(x)-c_S))$ and $g^*(z)=\mathbf{1}(\eta^f_Y(z)-c_Y\geq\lambda(\eta^f_S(z)-c_S))$.

Theorem 2 (Analytical form of cost of mistrust)**.** *The cost of mistrust may be expressed as*

$$\begin{aligned}&R_{YS}(g^*\circ f)-R_{YS}(h^*)\\&\quad=\mathbb{E}_X[\min(\eta^f_Y(f(X))-c_Y,\lambda(\eta^f_S(f(X))-c_S))\\&\qquad\quad-\min(\eta_Y(X)-c_Y,\lambda(\eta_S(X)-c_S))].\end{aligned} \tag{10}$$

The cost of mistrust may be decomposed into accuracy and fairness differences, where the accuracy difference is

$$\begin{aligned}&R_Y(g^*\circ f)-R_Y(h^*)\\&=\mathbb{E}_X[h^*(X)(\eta_Y(X)-c_Y)-g^*(f(X))(\eta^f_Y(f(X))-c_Y)],\end{aligned}$$

and the fairness difference is

$$\begin{aligned}&R_S(g^*\circ f)-R_S(h^*)\\&=\mathbb{E}_X[h^*(X)(\eta_S(X)-c_S)-g^*(f(X))(\eta^f_S(f(X))-c_S)],\end{aligned}$$

which are combined in the overall cost of mistrust

$$\begin{aligned}&R_{YS}(g^*\circ f)-R_{YS}(h^*)\\&=R_Y(g^*\circ f)-R_Y(h^*)-\lambda(R_S(g^*\circ f)-R_S(h^*)).\end{aligned}$$

Proof idea. We apply Lemma 9 of Menon and Williamson 2018 to express each of $R_Y(g^*\circ f)$, $R_Y(h^*)$, $R_S(g^*\circ f)$ and $R_S(h^*)$. Combining these yields a compact expression for $R_{YS}(g^*\circ f)-R_{YS}(h^*)$. □

The expression (10) for the cost of mistrust allows us to measure the quality of the fairness-accuracy trade-off available using f compared to using the original data. The decomposition reveals that the signs of the accuracy and fairness differences may vary. However, since the cost of mistrust is non-negative, for a fixed value of R_S we incur a value of R_Y that is at least as large using f as with the original data.

For intuition about the expression (10) for the cost of mistrust in Theorem 2, consider some point $z\in\mathcal{Z}$ and its preimage $\mathcal{X}_z:=\{x\in\mathcal{X}|f(x)=z\}$. If for all $x\in\mathcal{X}_z$, we have the same value of $\mathbf{1}(\eta_Y(x)-c_Y\geq\lambda(\eta_S(x)-c_S))$, then the expectation conditioned on $x\in\mathcal{X}_z$ will be zero, otherwise it will be positive. Hence the cost of mistrust will be small when points mapped to the same value of z tend to have the same value of $\mathbf{1}(\eta_Y(x)-c_Y\geq\lambda(\eta_S(x)-c_S))$.

We are interested in situations where the data producer can guarantee that the cost of mistrust is small without accessing Y. When $\mathcal{Z}=\mathcal{X}$ and the conditional distributions $\eta_Y(x)$ and $\eta_S(x)$ are smooth, the cost of mistrust can be upper bounded in terms of average reconstruction error. This result, shown in Theorem 3, allows the data producer to bound the cost of mistrust using only X and X^f.

Theorem 3 (Upper bound on cost of mistrust with smooth conditional distributions)**.** *Suppose $\mathcal{Z}=\mathcal{X}$ is a Euclidean space and we have the Lipschitz conditions that for some non-negative constants l_Y and l_S*

$$\forall x,x'\in\mathcal{X},|\eta_Y(x)-\eta_Y(x')|\leq l_Y\|x-x'\|_2$$

and

$$\forall x,x'\in\mathcal{X},|\eta_S(x)-\eta_S(x')|\leq l_S\|x-x'\|_2.$$

Then

$$R_{YS}(g^*\circ f)-R_{YS}(h^*)\leq(l_Y+\lambda l_S)\mathbb{E}_X\|X-f(X)\|_2.$$

Proof idea. We observe that $R_{YS}(h^*\circ f)$ is an upper bound on $R_{YS}(g^*\circ f)$. We use Lemma 9 of Menon and Williamson 2018 to re-express R_{YS}. We then use the Lipschitz conditions to upper bound $R_{YS}(h^*\circ f)-R_{YS}(h^*)$. □

4.2 Cost for Individual Fairness

We investigate the cost of using a given representation in terms of *individual fairness* (Dwork et al. 2012). This notion requires that similar decisions should be made for similar individuals, i.e. decisions are smooth.

Definition 7 (Individual fairness (Dwork et al. 2012))**.** *Let D and d be subadditive functions. Hypothesis h is $D,d-$individually fair if*

$$\forall x,x'\in\mathcal{X},D(h(x),h(x'))\leq d(x,x').$$

We also give a quantitative notion of individual *unfairness* by measuring the probability that a pair of randomly selected individuals will be treated unfairly according to Definition 7.

Definition 8 (Individual unfairness)**.** *Hypothesis h has $D,d-$individual unfairness with respect to X defined as*

$$IU_{D,d}(h):=p(D(h(x),h(x'))>d(x,x')),$$

where x and x' are independent random samples of X.

It is possible that a representation function maps points that are nearby in the input space to points that are distant from each other in the feature space. Therefore, smooth hypotheses may not be individually fair when applied to the cleaned data. We wish to quantify this cost for individual fairness by upper bounding the individual unfairness of an arbitrary smooth hypothesis applied to the cleaned data. We show that it is possible for a data user to provide this kind of certification to a data regulator by inspecting X^f and X. To do this we introduce the following definition.

Definition 9 (Large reconstruction error rate)**.** *Suppose $\mathcal{Z} = \mathcal{X}$. Let ϵ be a non-negative constant. Let $p(d(X, f(X)) > \epsilon)$ be the large reconstruction error rate of f.*

In Theorem 4 we show that if the large reconstruction error rate is small, then any hypothesis that is smooth (i.e. individually fair when applied to the original data) will not be too individually unfair when applied to the cleaned data. We observe that there is a tension between guaranteeing group fairness, which involves removing information to protect an adversary from inferring the sensitive variable, and individual fairness, which requires preserving information from the original data.

Theorem 4 (Upper bound on individual unfairness)**.** *Suppose $\mathcal{Z} = \mathcal{X}$. Let $d_\epsilon(x, x') := d(x, x') + 2\epsilon$ and let h be any individually fair hypothesis. Then the $D, d_\epsilon-$individual unfairness of $h \circ f$ is upper bounded as follows:*

$$IU_{D,d_\epsilon}(h \circ f) \leq 2p(d(X, f(X)) > \epsilon).$$

Proof idea. Let $\delta := p(d(X, f(X)) > \epsilon)$. For randomly drawn x and x', $d(x, f(x)) \leq \epsilon$ and $d(x', f(x')) \leq \epsilon$ with probability at least $1 - 2\delta$ by the union bound. If these statements hold, we may use the triangle inequality to conclude that $D(h(f(x), h(f(x')) \leq d(x, x') + 2\epsilon$. □

5 Benefits of Fair Representation Learning

We quantify the benefits of some f by measuring the discrimination achieved by an optimal adversary using X^f. We show that a data producer can do this for both statistical parity and disparate impact. We can compute these two quantities directly for a given f, so that unlike in the optimization problems we considered earlier there is no need to use a cost-sensitive risk. The quantities we obtain can be given to a data regulator to certify that any use of the cleaned data will not be too unfair. If the data producer has access to the target variable, these quantities can also be evaluated on subsets of the data with the same value of the target, to measure quantified versions of equality of opportunity (conditioning on $Y = 1$) and equalized odds (conditioning separately on $Y = 1$ and $Y = 0$) (Hardt, Price, and Srebro 2016).

5.1 Benefit for Statistical Parity

We certify that any decision using the cleaned data has statistical parity (Definition 2) that is neither too small nor too large. In Theorem 5, we show that the maximum and minimum statistical parity of an adversary using X^f can be expressed in closed form. The maximum and minimum will be closer if the induced conditional probability $\eta^f_S(z)$ does not deviate too much on average from the prior π_S. If $\eta^f_S(z) = \pi_S$ everywhere, we have statistical parity of zero, i.e. exact equality of outcome.

Theorem 5 (Statistical parity of optimal adversary)**.** *An adversarial user of X^f achieves maximum and minimum statistical parity*

$$\max_{g\in\mathcal{G}} SP(g \circ f) = 1 - \mathbb{E}_{X^f}[\min(\frac{\eta^f_S(X^f)}{\pi_S}, \frac{1 - \eta^f_S(X^f)}{1 - \pi_S})]$$

$$\min_{g\in\mathcal{G}} SP(g \circ f) = -1 + \mathbb{E}_{X^f}[\min(\frac{\eta^f_S(X^f)}{\pi_S}, \frac{1 - \eta^f_S(X^f)}{1 - \pi_S})].$$

Proof idea. Observe that statistical parity is a linear transformation of balanced error rate. Apply the minimum balanced error rate from Equation 32 of Zhao et al. 2013. □

5.2 Benefit for Disparate Impact

We certify that any decision using the cleaned data has disparate impact (Definition 3) that is neither too small nor too large. In Theorem 6, we show that the maximum and minimum disparate impact of an adversary using X^f can be expressed in closed form. The maximum and minimum will be closer if the induced conditional probability $\eta^f_S(z)$ never deviates too much from the prior π_S. If $\eta^f_S(z) = \pi_S$ everywhere, we have disparate impact of one, i.e. exact equality of outcome. Observe how disparate impact is more sensitive than statistical parity, since it requires $\eta^f_S(z)$ to be close to π_S *everywhere* rather than only in expectation.

Theorem 6 (Disparate impact of optimal adversary)**.** *Let $\overline{\eta}^f_S := \max_{z\in\mathcal{Z}} \eta^f_S(z)$ and $\underline{\eta}^f_S := \min_{z\in\mathcal{Z}} \eta^f_S(z)$. An adversarial user of X^f achieves maximum and minimum disparate impact*

$$\max_{g\in\mathcal{G}} DI(g \circ f) = \frac{\pi_S(1 - \underline{\eta}^f_S)}{\underline{\eta}^f_S(1 - \pi_S)}$$

$$\min_{g\in\mathcal{G}} DI(g \circ f) = \frac{\pi_S(1 - \overline{\eta}^f_S)}{\overline{\eta}^f_S(1 - \pi_S)}.$$

Proof idea. Re-express $DI(g \circ f)$ using the law of total probability, the fact that $\hat{Y}^f$ and S are conditionally independent given X^f, and Bayes' rule. Using this form we obtain the maximum and minimum values of $DI(g \circ f)$ and the corresponding choices of g. □

6 Conclusion

We have quantified the costs – an inferior fairness-accuracy trade-off and an increase in individual unfairness – incurred by a given representation. We have also quantified the benefits – reduced statistical parity and disparate impact achievable by an adversary – of such a representation. The benefits result from restricting the decisions of adversarial data users, while the costs are due to applying those same restrictions to other data users. We showed how a data producer can estimate these costs and benefits, even without access to the target variable, to support a novel three-party governance model entailing a separation of concerns between fairness and accuracy. Future directions of interest include extending our results to finite samples, stochastic representation functions, multiple sensitive groups and variables, more general representation spaces, and other fairness definitions.

Acknowledgments We would like to thank the reviewers for their useful feedback. We would also like to thank Aditya Menon for discussions during the development of this work. The research was supported by an Australian Government Research Training Program Scholarship and a CSIRO Data61 Top-Up Scholarship.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, S., and Selbst, A. D. 2016. Big Data's Disparate Impact. *California Law Review* 104:671.

Barocas, S.; Hardt, M.; and Narayanan, A. 2018. *Fairness and Machine Learning*. fairmlbook.org.

Bechavod, Y., and Ligett, K. 2017. Penalizing Unfairness in Binary Classification. *arXiv preprint arXiv:1707.00044*.

Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *FAT/ML Workshop*.

Calders, T., and Verwer, S. 2010. Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery* 21(2):277–292.

Calmon, F. P.; Wei, D.; Ramamurthy, K. N.; and Varshney, K. R. 2017. Optimized Data Pre-Processing for Discrimination Prevention. *Advances in Neural Information Processing Systems*.

Datta, A.; Tschantz, M. C.; and Datta, A. 2015. Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies* 2015(1):92–112.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.; and Pontil, M. 2018. Empirical Risk Minimization under Fairness Constraints. *arXiv preprint arXiv:1802.08626*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Dwork, C.; Immorlica, N.; Kalai, A. T.; and Leiserson, M. D. 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *Conference on Fairness, Accountability and Transparency*, 119–133.

Edwards, H., and Storkey, A. 2016. Censoring Representations with an Adversary. *International Conference on Learning Representations*.

Elkan, C. 2001. The Foundations of Cost-Sensitive Learning. In *International Joint Conference on Artificial Intelligence*, volume 17, 973–978.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, 3315–3323.

Johndrow, J., and Lum, K. 2017. An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction. *arXiv preprint arXiv:1703.04957*.

Louizos, C.; Swersky, K.; Li, Y.; Zemel, R.; and Welling, M. 2016. The Variational Fair Autoencoder. *International Conference on Learning Representations*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning*, 3384–3393.

Menon, A. K., and Williamson, R. C. 2018. The Cost of Fairness in Binary Classification. In *Conference on Fairness, Accountability and Transparency*, 107–118.

Mitchell, S., and Shadlen, J. 2018. Mirror Mirror: Reflections on Quantitative Fairness. https://speak-statistics-to-power.github.io/fairness.

O'Neil, C. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Romei, A., and Ruggieri, S. 2014. A Multidisciplinary Survey on Discrimination Analysis. *The Knowledge Engineering Review* 29(5):582–638.

United States Equal Opportunity Employment Commission. 1978. Uniform Guidelines on Employee Selection Procedures.

Wainwright, M. J.; Jordan, M. I.; and Duchi, J. C. 2012. Privacy Aware Learning. In *Advances in Neural Information Processing Systems*, 1430–1438.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017a. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017b. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *International Conference on Machine Learning*, 325–333.

Zhao, M.-J.; Edakunni, N.; Pocock, A.; and Brown, G. 2013. Beyond Fano's Inequality: Bounds on the Optimal F-Score, BER, and Cost-Sensitive Risk and their Implications. *Journal of Machine Learning Research* 14:1033–1090.

Zliobaite, I. 2015. A Survey on Measuring Indirect Discrimination in Machine Learning. *arXiv preprint arXiv:1511.00148*.

A Summary of Problems, Costs and Benefits Considered

We summarize the problems we consider in Table 1. In Tables 2 and 3, we summarize the costs and benefits respectively of a given representation function f. In each table we distinguish between cases where access to the target variable Y is required, and cases where it is not required. Observe that if access to the target variable is available, the benefits described in Table 3 can be computed for specific subsets of the data based on the target variable (e.g. conditioning on $Y = 1$ or $Y = 0$). Definitions of all terms can be found in the main text.

Table 1: Problems

Problem	**Reference**	**Optimization Problem**
Access to target variable required		
Fair classification	(3)	$\min_{h \in \mathcal{H}} [R_Y(h) - \lambda R_S(h)]$
Fair representation learning	(5)	$\min_{f \in \mathcal{F}} [R_Y(g_Y^* \circ f) - \lambda R_S(g_S^* \circ f)]$
Access to target variable not required		
Fair representation learning without accessing target variable	(8)	$\min_{f \in \mathcal{F}} [\mathbb{E}_X \lVert X - f(X) \rVert_2 - \lambda R_S(g_S^* \circ f)]$

Table 2: Costs of a representation function f

Cost	**Reference**	**Analytical Form**
Access to target variable required		
Cost of mistrust	Theorem 2	$\mathbb{E}_X[\min(\eta_Y^f(f(X)) - c_Y, \lambda(\eta_S^f(f(X)) - c_S)) - \min(\eta_Y(X) - c_Y, \lambda(\eta_S(X) - c_S))]$
Access to target variable not required		
Upper bound on cost of mistrust using average reconstruction error	Theorem 3	$(l_Y + \lambda l_S)\mathbb{E}_X \lVert X - f(X) \rVert_2$
Upper bound on individual unfairness using large reconstruction error rate	Theorem 4	$2p(d(X, f(X)) > \epsilon)$

Table 3: Benefits of a representation function f

Benefit	**Reference**	**Analytical Form**
Access to target variable not required		
Maximum and minimum statistical parity	Theorem 5	$\max_{g \in \mathcal{G}} SP(g \circ f) = 1 - \mathbb{E}_{X^f}[\min(\frac{\eta_S^f(X^f)}{\pi_S}, \frac{1-\eta_S^f(X^f)}{1-\pi_S})]$ $\min_{g \in \mathcal{G}} SP(g \circ f) = -1 + \mathbb{E}_{X^f}[\min(\frac{\eta_S^f(X^f)}{\pi_S}, \frac{1-\eta_S^f(X^f)}{1-\pi_S})]$
Maximum and minimum disparate impact	Theorem 6	$\max_{g \in \mathcal{G}} DI(g \circ f) = \frac{\pi_S(1-\underline{\eta}_S^f)}{\underline{\eta}_S^f(1-\pi_S)}$ $\min_{g \in \mathcal{G}} DI(g \circ f) = \frac{\pi_S(1-\overline{\eta}_S^f)}{\overline{\eta}_S^f(1-\pi_S)}$

B Theorem Proofs

We present complete proofs of our theoretical results.

B.1 Proof of Theorem 1 (Fair Representation Learning Without Accessing Target Variable)

Proof. We derive an upper bound on

$$R_Y(g_Y^* \circ f) - l_Y \lambda R_S(g_S^* \circ f).$$

Let $h_Y^* \in \underset{h \in \mathcal{H}}{\arg\min} R_Y(h)$, which takes the form $h_Y^*(x) = \mathbf{1}(\eta_Y(x) \geq c_Y)$ (Zhao et al. 2013).

$$\begin{aligned}
& R_Y(g_Y^* \circ f) - l_Y \lambda R_S(g_S^* \circ f) \\
& \leq R_Y(h_Y^* \circ f) - l_Y \lambda R_S(g_S^* \circ f) \\
& = R_Y(h_Y^* \circ f) - R_Y(h_Y^*) + R_Y(h_Y^*) - l_Y \lambda R_S(g_S^* \circ f) \\
& = \mathbb{E}_X[(c_Y - \eta_Y(X)) h_Y^*(f(X))] - \mathbb{E}_X[(c_Y - \eta_Y(X)) h_Y^*(X)] + R_Y(h_Y^*) - l_Y \lambda R_S(g_S^* \circ f) \\
& = \mathbb{E}_X[(c_Y - \eta_Y(X))(h_Y^*(f(X)) - h_Y^*(X))] + R_Y(h_Y^*) - l_Y \lambda R_S(g_S^* \circ f) \\
& \leq l_Y \mathbb{E}_X \|X - f(X)\|_2 + R_Y(h_Y^*) - l_Y \lambda R_S(g_S^* \circ f).
\end{aligned}$$

For the second equality we apply Lemma 9 from (Menon and Williamson 2018). For the third equality we apply linearity of expectation.

For the final inequality, for any x where $h_Y^*(x) \neq h_Y^*(f(x))$, there must exist some x' on the decision boundary such that $c_Y - \eta_Y(x') = 0$ and $\|x - x'\|_2 \leq \|x - f(x)\|_2$. Combining with the Lipchitz condition yields

$$c_Y - \eta_Y(x) \leq c_Y - \eta_Y(x') + l_Y \|x - x'\|_2 \leq l_Y \|x - f(x)\|_2.$$

Since this is true for every x it is also true in expectation.

We then observe

$$\begin{aligned}
& \underset{f \in \mathcal{F}}{\arg\min} [l_Y \mathbb{E}_X \|X - f(X)\|_2 + R_Y(h_Y^*) - l_Y \lambda R_S(g_S^* \circ f)] \\
= & \underset{f \in \mathcal{F}}{\arg\min} [\mathbb{E}_X \|X - f(X)\|_2 - \lambda R_S(g_S^* \circ f)].
\end{aligned}$$

□

B.2 Proof of Theorem 2 (Analytical Form of Cost of Mistrust)

Proof. First we show the analytical expression for the cost of mistrust. Applying Proposition 4 of Menon and Williamson 2018, we have that $h^*(x) = \mathbf{1}(\eta_Y(x) - c_Y \geq \lambda(\eta_S(x) - c_S))$ and $g^*(z) = \mathbf{1}(\eta_Y^f(z) - c_Y \geq \lambda(\eta_S^f(z) - c_S))$ are solutions to (3) and (9) respectively.

$$\begin{aligned}
& \mathbb{E}_X[\min(\eta_Y(X) - c_Y, \lambda(\eta_S(X) - c_S))] \\
& = \mathbb{E}_X[(1 - h^*(X))(\eta_Y(X) - c_Y)] + \lambda \mathbb{E}_X[h^*(X)(\eta_S(X) - c_S)] \\
& = \mathbb{E}_X[\eta_Y(X) - c_Y] - \mathbb{E}_X[h^*(X)(\eta_Y(X) - c_Y)] + \lambda \mathbb{E}_X[h^*(X)(\eta_S(X) - c_S)] \\
& = \pi_Y - c_Y - \mathbb{E}_X[h^*(X)(\eta_Y(X) - c_Y)] + \lambda \mathbb{E}_X[h^*(X)(\eta_S(X) - c_S)] \\
& = \pi_Y - c_Y + R_Y(h^*) - (1 - c_Y)\pi_Y + \lambda \mathbb{E}_X[h^*(X)(\eta_S(X) - c_S)] \\
& = \pi_Y - c_Y + R_Y(h^*) - (1 - c_Y)\pi_Y - \lambda R_S(h^*) + \lambda(1 - c_S)\pi_S \\
& = R_Y(h^*) - \lambda R_S(h^*) - c_Y(1 - \pi_Y) + \lambda(1 - c_S)\pi_S.
\end{aligned}$$

The second and third last lines both involve substitutions based on Lemma 9 from Menon and Williamson 2018.

Similarly,

$$\begin{aligned}
& \mathbb{E}_X[\min(\eta_Y^f(f(X)) - c_Y, \lambda(\eta_S^f(f(X)) - c_S))] \\
& = \mathbb{E}_{X^f}[\min(\eta_Y^f(X^f) - c_Y, \lambda(\eta_S^f(X^f) - c_S))] \\
& = R_Y(g^* \circ f) - \lambda R_S(g^* \circ f) - c_Y(1 - \pi_Y) + \lambda(1 - c_S)\pi_S.
\end{aligned}$$

The result follows by substituting the former expression from the latter and applying linearity of expectation.

The decomposed form straightforwardly follows from applying Lemma 9 of Menon and Williamson 2018 to each of $R_Y(g^* \circ f)$, $R_Y(h^*)$, $R_S(g^* \circ f)$ and $R_S(h^*)$, then applying linearity of expectation to express $R_Y(g^* \circ f) - R_Y(h^*)$ and $R_S(g^* \circ f) - R_S(h^*)$. □

B.3 Proof of Theorem 3 (Upper Bound on Cost of Mistrust with Smooth Conditional Distributions)

Proof. Let $h^*(x) := \mathbf{1}(\eta_Y(x) - c_Y \geq \lambda(\eta_S(x) - c_S))$, which is a solution to (3) (Proposition 4 of Menon and Williamson 2018).

$$\begin{aligned}
&R_{YS}(g^* \circ f) - R_{YS}(h^*) \\
&\leq R_{YS}(h^* \circ f) - R_{YS}(h^*) \\
&= R_Y(h^* \circ f) - R_Y(h^*) - \lambda(R_S(h^* \circ f) - R_S(h^*)) \\
&= \mathbb{E}_X[(c_Y - \eta_Y(X))(h^*(f(X)) - h^*(X))] - \lambda \mathbb{E}_X[(c_S - \eta_S(X))(h^*(f(X)) - h^*(X))] \\
&= \mathbb{E}_X[(c_Y - \eta_Y(X) - \lambda(c_S - \eta_S(X)))(h^*(f(X)) - h^*(X))] \\
&\leq (l_Y + \lambda l_S)\mathbb{E}_X\|X - f(X)\|_2.
\end{aligned}$$

The second equality is by Lemma 9 from Menon and Williamson 2018 and linearity of expectation. The third equality is by linearity of expectation.

For the final inequality, for any x where $h^*(x) \neq h^*(f(x))$, there must exist some x' on the decision boundary such that $c_Y - \eta_Y(x') - \lambda(c_S - \eta_S(x')) = 0$ and $\|x - x'\|_2 \leq \|x - f(x)\|_2$. Combining with the Lipchitz conditions yields

$$\begin{aligned}
&c_Y - \eta_Y(x) - \lambda(c_S - \eta_S(x)) \\
&\leq c_Y - \eta_Y(x') + l_Y\|x - x'\|_2 - \lambda(c_S - \eta_S(x') - l_S\|x - x'\|_2) \\
&\leq (l_Y + \lambda l_S)\|x - f(x)\|_2.
\end{aligned}$$

Since this is true for every x it is also true in expectation. □

B.4 Proof of Theorem 4 (Upper Bound on Individual Unfairness)

Proof. Let $\delta := p(d(X, f(X)) > \epsilon)$. Let h be a $D, d-$individually fair hypothesis (see Definition 7).

Consider points x and x' drawn independently at random using the input X.With probability $1-\delta$, $d(x, f(x)) \leq \epsilon$. Similarly, with probability $1-\delta$, $d(x', f(x')) \leq \epsilon$. By the union bound, both statements hold with probability at least $1 - 2\delta$. In that case, the following statements also hold:

$$\begin{aligned}
&D(h(f(x), h(f(x')) \\
&\leq D(h(f(x)), h(x)) + D(h(x), h(f(x'))) \\
&\leq \epsilon + D(h(x), h(f(x'))) \\
&\leq \epsilon + D(h(x), h(x')) + D(h(x'), h(f(x'))) \\
&\leq 2\epsilon + D(h(x), h(x')) \\
&\leq 2\epsilon + d(x, x').
\end{aligned}$$

The first and third inequalities apply the triangle inequality since D is subadditive. The second and fourth inequalities hold as assumed above. The final inequality applies Definition 7.

Therefore $IU_{D,d_\epsilon}(h \circ f) \leq 2\delta$. □

B.5 Proof of Theorem 5 (Statistical Parity of Optimal Adversary)

Proof. Let

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

be the balanced error rate of a hypothesis h. Observe that we have $SP(h) = 1 - 2BER(h)$ for all h.

Therefore

$$\begin{aligned}
&\max_{g \in \mathcal{G}} SP(g \circ f) \\
&= 1 - 2\min_{g \in \mathcal{G}} BER(g \circ f) \\
&= 1 - \mathbb{E}_{X^f}[\min(\frac{\eta_S^f(X^f)}{\pi_S}, \frac{1 - \eta_S^f(X^f)}{1 - \pi_S})]
\end{aligned}$$

where the final equality uses Equation 32 from Zhao et al. 2013.

Similarly,

$$\begin{aligned}
&\min_{g\in\mathcal{G}} SP(g\circ f)\\
&= 1 - 2\max_{g\in\mathcal{G}} BER(g\circ f)\\
&= 1 - 2\max_{g\in\mathcal{G}}[1 - BER(1 - g\circ f)]\\
&= -1 + \min_{g\in\mathcal{G}} BER(1 - g\circ f)\\
&= -1 + \min_{g\in\mathcal{G}} BER(g\circ f)\\
&= -1 + \mathbb{E}_{X^f}[\min(\frac{\eta_S^f(X^f)}{\pi_S}, \frac{1-\eta_S^f(X^f)}{1-\pi_S})]
\end{aligned}$$

where for the second equality we used the fact that $BER(h) = 1 - BER(1-h)$ for all h. □

B.6 Proof of Theorem 6 (Disparate Impact of Optimal Adversary)

Proof. Disparate impact can be expressed as follows:

$$\begin{aligned}
&DI(g\circ f)\\
&= \frac{p(\hat{Y}^f = 1|S=0)}{p(\hat{Y}^f = 1|S=1)}\\
&= \frac{\int_z p(X^f = z|S=0)p(\hat{Y}^f = 1|S=0, X^f = z)dz}{\int_z p(X^f = z|S=1)p(\hat{Y}^f = 1|S=1, X^f = z)dz}\\
&= \frac{\int_z p(X^f = z|S=0)g(z)dz}{\int_z p(X^f = z|S=1)g(z)dz}\\
&= \frac{\pi_S \int_z p(X^f = z)(1-\eta_S^f(z))g(z)dz}{(1-\pi_S)\int_z p(X^f = z)\eta_S^f(z)g(z)dz}\\
&= \frac{\pi_S \mathbb{E}_{X^f}[(1-\eta_S^f(X^f))g(X^f)]}{(1-\pi_S)\mathbb{E}_{X^f}[\eta_S^f(X^f)g(X^f)]}.
\end{aligned}$$

For the third equality we used the fact that $\hat{Y}^f$ and S are conditionally independent given X^f. For the fourth equality we used Bayes' rule.

Recall that $\overline{\eta}_S^f := \max_{z\in\mathcal{Z}} \eta_S^f(z)$ and $\underline{\eta}_S^f := \min_{z\in\mathcal{Z}} \eta_S^f(z)$. Let γ be an arbitrary constant in the range $(0,1]$. Using the form of $DI(g\circ f)$ above, we have:

$$\max_{g\in\mathcal{G}} DI(g\circ f) = \frac{\pi_S(1-\underline{\eta}_S^f)}{\underline{\eta}_S^f(1-\pi_S)}$$

where the maximum is obtained for

$$g(z) = \begin{cases} \gamma & \text{if } \eta_S^f(z) = \underline{\eta}_S^f \\ 0 & \text{otherwise.} \end{cases}$$

Similarly,

$$\min_{g\in\mathcal{G}} DI(g\circ f) = \frac{\pi_S(1-\overline{\eta}_S^f)}{\overline{\eta}_S^f(1-\pi_S)}$$

where the minimum is obtained for

$$g(z) = \begin{cases} \gamma & \text{if } \eta_S^f(z) = \overline{\eta}_S^f \\ 0 & \text{otherwise.} \end{cases}$$

□

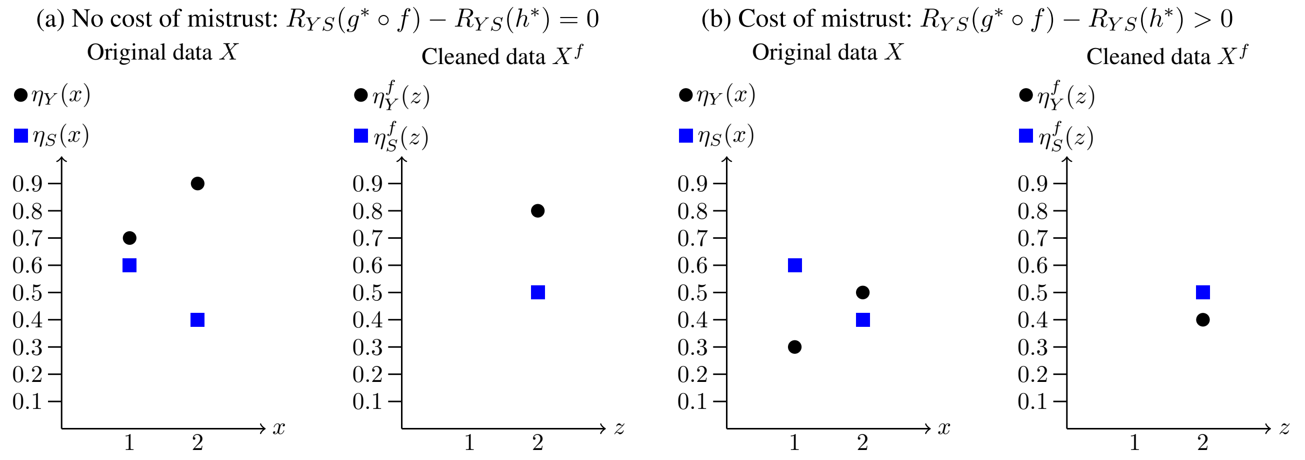


Figure 2: Two examples illustrating the cost of mistrust.

C Examples of the Cost of Mistrust

We use examples to demonstrate that the cost of mistrust may be either zero or positive, as depicted in Figure 2. Let $\mathcal{X} = \mathcal{Z} = \{1, 2\}$, $c_Y = c_S = 0.5$, $p(X = 1) = p(X = 2) = 0.5$, $\lambda = 1$ in (3), (5) and (9). In both examples, $\eta_S(1) = 0.6$ and $\eta_S(2) = 0.4$. In (a) $\eta_Y(1) = 0.7$ and $\eta_Y(2) = 0.9$, while in (b) $\eta_Y(1) = 0.3$ and $\eta_Y(2) = 0.5$. While setting f to map all points to a constant is a crude example, it suffices for our illustration. In (a) the cost of mistrust is 0, while in (b) it is 0.05. This is because in (a), h^* predicts the same value for the two points combined by f and is hence unaffected by f, while in (b), h^* predicts different values and is hence affected by f.

We compare the representation $f(x) = 2$ to the identity representation $f_I(x) = x$, which makes our analysis sufficiently general to cover all choices of representation. First we show that in both examples f is a solution to (5). In the first example, we may show $R_Y(g_Y^* \circ f) = R_Y(g_Y^* \circ f_I) = 0.1$ by applying (6). However, using (7) we may show $R_S(g_S^* \circ f) = 0.25$, while $R_S(g_S^* \circ f_I) = 0.2$. Similarly in the second example, $R_Y(g_Y^* \circ f) = R_Y(g_Y^* \circ f_I) = 0.2$, while $R_S(g_S^* \circ f) = 0.25$ and $R_S(g_S^* \circ f_I) = 0.2$.

Now we compute the cost of mistrust for both cases by applying Theorem 2. In the first example,

$$R_{YS}(g^* \circ f) - R_{YS}(h^*) = (1 \times 0) - (0.5 \times 0.1 + 0.5 \times -0.1) = 0 - 0 = 0.$$

In the second example,

$$R_{YS}(g^* \circ f) - R_{YS}(h^*) = (1 \times -0.1) - (0.5 \times -0.2 + 0.5 \times -0.1) = 0.05.$$

Thus we observe that it is straightforward to construct examples where the cost of mistrust is both zero as well as those where the cost of mistrust is positive.

The examples in Figure 2 can be used as intuition for interpreting the expression for the cost of mistrust in Theorem 2. For some point $z \in \mathcal{Z}$, define its preimage $\mathcal{X}_z := \{x \in \mathcal{X} | f(x) = z\}$. If for all $x \in \mathcal{X}_z$, we have the same value of $\mathbf{1}(\eta_Y(x) - c_Y \geq \lambda(\eta_S(x) - c_S))$ as in example (a), then the expectation conditioned on $x \in \mathcal{X}_z$ will be zero. Otherwise, as in example (b), the conditional expectation will be positive.

D Experiments

We conducted experiments with two objectives in mind. First, to show that the formalization of the fair representation learning problem we suggested in Section 3 can be used in practice. Second, to illustrate how the costs and benefits identified in Sections 4 and 5 can be estimated and interpreted without requiring access to the target variable.

D.1 Datasets

We used the UCI Adult and ProPublica recidivism datasets, which are both well-known in the fair machine learning literature (e.g. Calmon et al. 2017). These datasets are located at `https://archive.ics.uci.edu/ml/datasets/adult` and `https://github.com/propublica/compas-analysis` respectively. We selected S to be gender for Adult and whether the person is of African-American ethnicity for ProPublica. Our experiments do not depend on a particular choice of Y. We learn f using 70% of the data and report results on the remaining 30%.

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 S as gender, while a possible choice of Y is 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 ProPublica dataset contains information about 7214 criminal offences committed in Broward County, Florida and contains 79 input columns once categorical features are binarized. 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', then binarized the categorical variable. We selected S as whether the person is of African-American ethnicity, while a possible choice of Y is whether the person reoffended 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.

D.2 Method

We approximate (8) by estimating f with an *encoder* neural network, testing several values of λ. We use a finite sample to estimate the average reconstruction error component of the cost function. To estimate the $R_S(g^* \circ f)$ component of the cost function we use the form given by (7) with $c_S = 0.5$, train another *evaluator* neural network to estimate $\eta_S^f(z)$, and use a finite sample to approximate the expectation. The evaluator is comparable to the 'adversary' in (Edwards and Storkey 2016; Beutel et al. 2017; Madras et al. 2018) since we alternate updating its weights with those of the encoder. However, the evaluator is used to estimate (7) which is used to evaluate f, rather than its performance directly being used to evaluate f. This approach is motivated by the fact that (7) gives us the performance of the optimal adversary.

Our training set consists of N points, where the input and sensitive variable values for the nth point are given by x_n and s_n respectively. We estimate f using a fully-connected *encoder* neural network with one softplus ($\text{softplus}(x) := \ln(1 + e^x)$) hidden layer of 100 units and a linear output layer with the same number of units as the input layer. To approximate (8) we update f to minimize the following cost function.

$$\frac{1}{N}\sum_{n=1}^{N}\|x_n - f(x_n)\|_2 - \frac{\lambda}{2N}\sum_{n=1}^{N}\min[\widetilde{\eta}_S^f(f(x_n)), 1 - \widetilde{\eta}_S^f(f(x_n))]$$

We compute $\widetilde{\eta}_S^f$ to estimate $\eta_S^f(z)$ using a fully-connected *evaluator* neural network with one softplus hidden layer of 100 units and a single sigmoidal output unit. The output layer of the encoder, which corresponds to the variable X^f, is the input layer for the evaluator. For the evaluator network we update $\widetilde{\eta}_S^f$ to minimize the following cost function.

$$-\frac{1}{N}\sum_{n=1}^{N}[s_n \ln \widetilde{\eta}_S^f(f(x_n)) + (1 - s_n)\ln(1 - \widetilde{\eta}_S^f(f(x_n)))]$$

We alternate updates of the weights in the encoder and evaluator networks, as in adversarial methods (Edwards and Storkey 2016; Beutel et al. 2017; Madras et al. 2018). We use the Adam Optimizer with a learning rate of 0.0001, a batch size of 100 and set the training set epochs to 100. We implemented the model in Python using the TensorFlow library.

We set the large reconstruction error rate threshold $\epsilon := 0.1 \times \frac{1}{N}\sum_{n=1}^{N}\|x_n\|_2$ and set the individual fairness distance function $d(x, x') := \|x - x'\|_2$. We evaluate the costs and benefits of a representation f using the following empirical estimates computed over a test set of N' points:

$$\mathbb{E}_X\|X - f(X)\|_2 \approx \frac{1}{N'}\sum_{n=1}^{N'}\|x_n - f(x_n)\|_2$$

$$p(\|X - f(X)\|_2 > \epsilon) \approx \frac{1}{N'}\sum_{n=1}^{N'}\mathbf{1}(\|x_n - f(x_n)\|_2 > \epsilon)$$

$$\max_{g \in \mathcal{G}} SP(g \circ f) \approx 1 - \frac{1}{N'}\sum_{n=1}^{N'}\min[\frac{\widetilde{\eta}_S^f(f(x_n))}{\widetilde{\pi}_S}, \frac{1 - \widetilde{\eta}_S^f(f(x_n))}{1 - \widetilde{\pi}_S}]$$

$$\max_{g \in \mathcal{G}} DI(g \circ f) \approx \frac{\widetilde{\pi}_S(1 - \underline{\widetilde{\eta}}_S^f)}{\underline{\widetilde{\eta}}_S^f(1 - \widetilde{\pi}_S)}$$

where $\widetilde{\pi}_S := \frac{1}{N'}\sum_{n=1}^{N'} s_n$ and $\underline{\widetilde{\eta}}_S^f := \min_{n \le N'} \widetilde{\eta}_S^f(f(x_n))$.

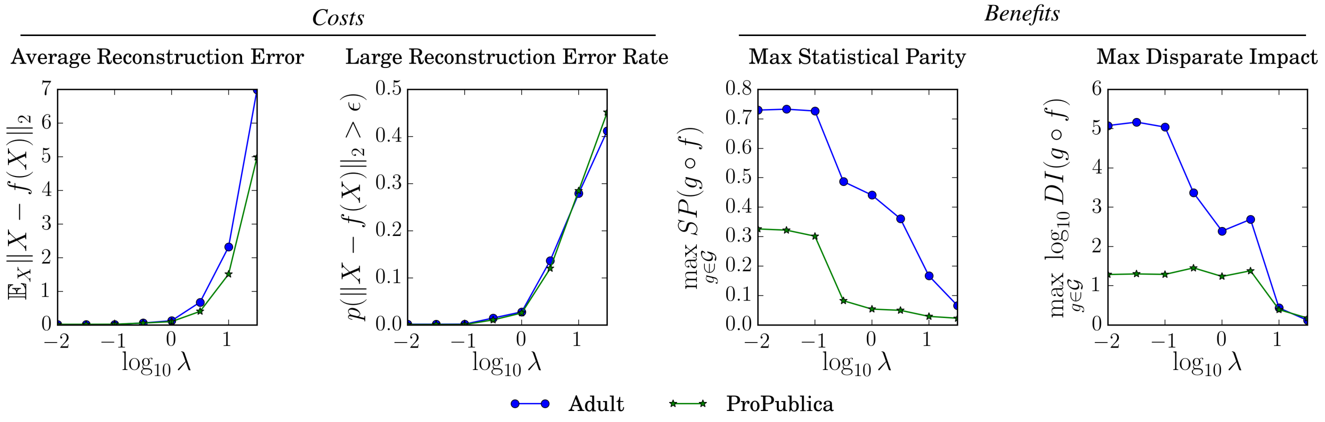


Figure 3: Estimates of costs and benefits of fair representation learning on Adult and ProPublica datasets. Lower is better on all plots. See text for discussion.

D.3 Results

We show our results in Figure 3. For several values of λ, we estimate the costs and benefits of the learned representation f. The trends for both datasets are similar. A subtlety is that we report proxies for the costs motivated by our theoretical results, which can be estimated by the data user without access to the target variable.

Recall that we may use average reconstruction error to upper bound the cost of mistrust (Theorem 3) and the large reconstruction error rate to upper bound the cost for individual fairness (Theorem 4). Estimates of both of these cost proxies increase with λ, as the cleaned data becomes more distorted by f.

Furthermore, recall that we have the closed form of an adversary's maximum statistical parity (Theorem 5) and disparate impact (Theorem 6). Estimates of both of these quantities decline as λ increases, indicating benefits from using f. For disparate impact we use a log scale on the y-axis for clarity, and observe that its empirical estimates appear noisier than those of statistical parity, since it requires us to estimate the minimum rather than the expectation of $\eta_S^f(z)$.

Our experiments, when combined with our theoretical results, reveal that the choice of λ in (8) starkly determines the relative costs and benefits of fair representation learning.