Costs and Benefits of Fair Representation Learning

Daniel McNamara
Australian National University and CSIRO Data61
Canberra, ACT, Australia
dpmcna@gmail.com

Cheng Soon Ong
Australian National University and CSIRO Data61
Canberra, ACT, Australia
chengsoon.ong@anu.edu.au

Robert C. Williamson
Australian National University and CSIRO Data61
Canberra, ACT, Australia
bob.williamson@anu.edu.au

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.

CCS CONCEPTS

• **Computing methodologies** → **Learning latent representations**; *Machine learning*; • **Social and professional topics** → *Socio-technical systems*; • **Theory of computation** → *Machine learning theory*.

KEYWORDS

fairness; representation learning; machine learning

ACM Reference Format:
Daniel McNamara, Cheng Soon Ong, and Robert C. Williamson. 2019. Costs and Benefits of Fair Representation Learning. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3306618.3317964

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 [1, 3, 8, 21]. 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 [2, 22, 29]. 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 [10].

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 [18]. 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 [5, 12, 14, 16–18, 27] as a *technique* for fair classification. Recent work [19] 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

AIES '19, January 27–28, 2019, Honolulu, HI, USA

ACM ISBN 978-1-4503-6324-2/19/01…$15.00
https://doi.org/10.1145/3306618.3317964

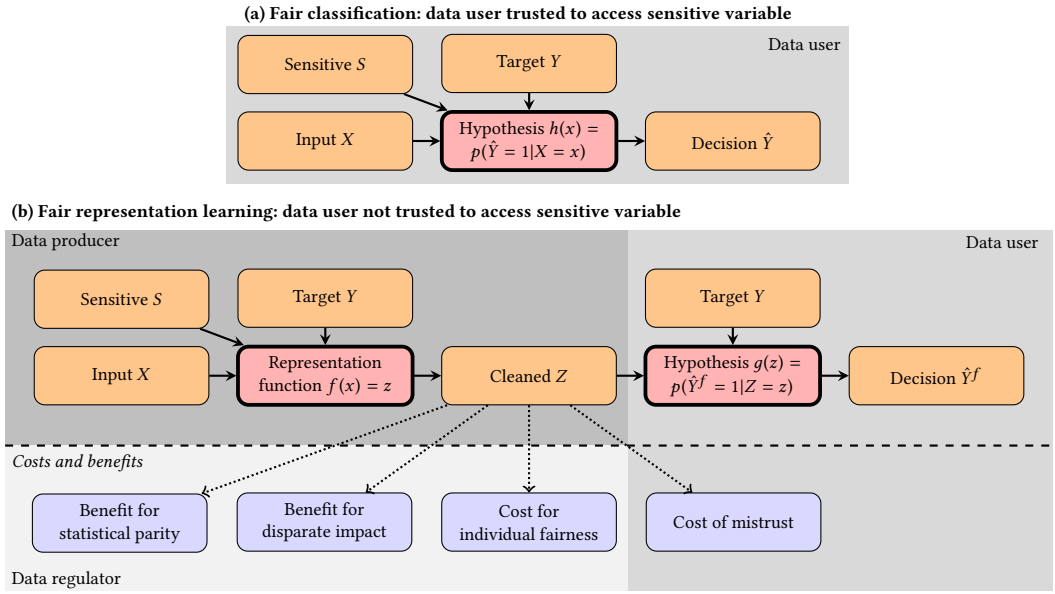

Figure 1: Summary of (a) fair classification and (b) fair representation learning, showing train time data processing for both, and costs and benefits of (b).

utility. This represents a progression in our scientific understanding, given that this problem had never previously been formally posed or analyzed.

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. 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 also be used to guarantee quantified versions of two other well-known fairness definitions: *equality of opportunity* and *equalized odds*. We provide a proof idea in the main paper for each result, while Appendix A provides complete proofs.

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, we may take the difference between groups – known as *statistical parity* [6, 10] – or the ratio between groups – known as *disparate impact* [14, 23]. We may wish to compute a parity measure only on a population subset. Constructing subsets by conditioning on particular values of the target variable yields variants [15] such as *equality of opportunity* (conditioning only on the positive class) and *equalized odds* (conditioning separately on the positive and negative classes).[1] However, when the training data labels are themselves affected by discrimination, conditioning on the target variable may not be suitable [25]. If the population subset consists of individuals who are similar according to some metric, we have *individual fairness*, also known as avoiding *disparate treatment* [10, 20].

Methods for fair classification can be divided into *pre-processing* – i.e. fair representation learning – which modifies the data that the algorithm learns from [27]; *in-processing*, which modifies the algorithm's objective function to incorporate a fairness constraint or penalty [4, 9, 11, 19, 25, 26]; and *post-processing*, which modifies the predictions produced by the algorithm [15].

Several fair representation learning techniques have been proposed. One such approach is to design the cleaned variable Z such that the distributions of Z conditioned on different values of the sensitive variable S are similar [14, 16]. In addition to this requirement, the pre-processing procedure may optimize the independence of Z

[1] Satisfying equalized odds has also previously been referred to as avoiding *disparate mistreatment* [25].

and S [17]. Adversarial approaches [5, 12, 18] 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 problem [7].

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 [24]. We also show that fair representation learning in fact addresses a problem that is distinct from fair classification, which is of interest when the data user is not trusted to access the sensitive variable.

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 respectively.

The data user learns a stochastic hypothesis $h : \mathcal{X} \to [0, 1]$ which is used to construct decision variable $\hat{Y} \in \{0, 1\}$, where $h(x) := p(\hat{Y} = 1|X = x)$. Let $\mu_{XYS\hat{Y}}$ be the joint distribution of the input, target, sensitive and decision variables.

At test time, the data user makes a decision using a sample of X, which may contain information about S. The quality of an 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 [13, 19, 28]). *The cost-sensitive risk of hypothesis h with respect to Y is*

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

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 $\mu_{XYS\hat{Y}}$. 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, or conditioning separately on $Y = 1$ and $Y = 0$ for quantified versions of equalized odds.

Definition 2 (Statistical parity [6, 10]). *The statistical parity of an hypothesis h is*

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

Definition 3 (Disparate impact [14, 23]). *The disparate impact of an 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 (see Appendix C of [19]).

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

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

where $H := [0, 1]^{\mathcal{X}}$ 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 [19]). Using Definition 1, this cost sensitive risk is written as:

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

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

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

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

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 [19] 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 $Z := f(X)$. The data producer learns f so that Z is still useful for predicting Y but suppresses information about S.

Let $\eta_Y^f(z) := p(Y = 1|Z = z)$ and $\eta_S^f(z) := p(S = 1|Z = z)$ be conditional probabilities of the positive classes of Y and S induced by f. The data user trains on samples of Z 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 $g(z) := p(\hat{Y}^f = 1|Z = z)$. At test time, the data producer samples X and passes it through f to produce a sample of Z, 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:

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

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

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

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

Using the form of the minimum cost-sensitive risk from [28], we may express the terms in (5) as follows:

$$R_Y(g_Y^* \circ f) = \mathbb{E}_Z[\min((1 - c_Y)\eta_Y^f(Z), c_Y(1 - \eta_Y^f(Z)))] \tag{6}$$

$$R_S(g_S^* \circ f) = \mathbb{E}_Z[\min((1 - c_S)\eta_S^f(Z), c_S(1 - \eta_S^f(Z)))]. \tag{7}$$

Adversarial neural networks have previously been used to estimate g_Y^* and g_S^* [5, 12, 18]. 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 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 task performance.

Definition 5 (Average reconstruction error). *Suppose that $\mathcal{Z} = \mathcal{X}$ 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:

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

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 that $\mathcal{Z} = \mathcal{X}$ and we have the Lipschitz condition that for some non-negative constant l_Y*

$$\forall x, x' \in \mathcal{X}, |\eta_Y(x) - \eta_Y(x')| \leq l_Y\|x - x'\|_2. \tag{9}$$

Then any $f \in F$ 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 [19], and making use of the Lipschitz condition. We then observe that the f minimizing this upper bound also minimizes $\mathbb{E}_X\|X - f(X)\|_2 - \lambda R_S(g_S^* \circ f)$. See Appendix A.1 for complete proof. □

4 COSTS OF FAIR REPRESENTATION LEARNING

We identify and quantify two costs of using fair representation learning rather than entrusting a single trusted data user to make decisions. These costs are incurred by decision-makers, as well as individuals about whom decisions are made. The first cost, which we refer to as the *cost of mistrust*, is the difference in the optimal fairness-accuracy trade-off available with the cleaned data produced by a representation function f compared to the original input. This cost is of interest to the data user – as well as potentially the data regulator. The second cost quantifies the extent to which individual fairness is violated by using a representation function f, which is primarily of interest to the data regulator. We show that both of these costs can 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.

$$\min_{g \in G} [R_Y(g \circ f) - \lambda R_S(g \circ f)] \tag{10}$$

Definition 6 (Cost of mistrust). *Let g^* and h^* be hypotheses minimizing* (10) *and* (3) *respectively, where the value of λ is the same in both equations. The cost of mistrust for a representation function f is defined as*

$$R_{YS}(g^* \circ f) - R_{YS}(h^*).$$

The cost of mistrust is non-negative because f restricts the hypothesis class to a subset of H. If $\lambda = 0$ in (10) 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 the sensitive variable S the data user has no way to estimate $R_S(g \circ f)$ in (10). 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.

We now show in Theorem 2 that we can express the cost of mistrust in analytical form. In our result, we use the expressions

$$h^*(x) = \mathbf{1}(\eta_Y(x) - c_Y \geq \lambda(\eta_S(x) - c_S)) \tag{11}$$

and

$$g^*(z) = \mathbf{1}(\eta_Y^f(z) - c_Y \geq \lambda(\eta_S^f(z) - c_S)), \tag{12}$$

obtained from Proposition 4 of [19].

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

$$\begin{aligned} R_{YS}(g^* \circ f) - R_{YS}(h^*) = \\ \mathbb{E}_X[\min(\eta_Y^f(f(X)) - c_Y, \lambda(\eta_S^f(f(X)) - c_S)) \\ - \min(\eta_Y(X) - c_Y, \lambda(\eta_S(X) - c_S))]. \end{aligned} \tag{13}$$

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

$$\begin{aligned} R_Y(g^* \circ f) - R_Y(h^*) = \\ \mathbb{E}_X[h^*(X)(\eta_Y(X) - c_Y) - g^*(f(X))(\eta_Y^f(f(X)) - c_Y)], \end{aligned} \tag{14}$$

and the fairness difference is

$$\begin{aligned} R_S(g^* \circ f) - R_S(h^*) = \\ \mathbb{E}_X[h^*(X)(\eta_S(X) - c_S) - g^*(f(X))(\eta_S^f(f(X)) - c_S)], \end{aligned} \tag{15}$$

which are combined in the overall cost of mistrust

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

Proof idea. Apply Lemma 9 of [19] 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 $R_{YS}(g^* \circ f) - R_{YS}(h^*)$. See Appendix A.2 for complete proof. □

The expression (13) 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 input. 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 input.

For intuition about the expression (13) for the cost of mistrust, consider a point $z \in \mathcal{Z}$ and its preimage $\mathcal{X}_z := \{x \in \mathcal{X} | f(x) = z\}$. If all $x \in \mathcal{X}_z$ 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 Z.

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*

$$\forall x, x' \in \mathcal{X}, |\eta_Y(x) - \eta_Y(x')| \leq l_Y \|x - x'\|_2 \tag{16}$$

and

$$\forall x, x' \in \mathcal{X}, |\eta_S(x) - \eta_S(x')| \leq l_S \|x - x'\|_2. \tag{17}$$

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 [19] to re-express R_{YS}. We then use the Lipschitz conditions to upper bound $R_{YS}(h^* \circ f) - R_{YS}(h^*)$. See Appendix A.3 for complete proof. □

4.2 Cost for Individual Fairness

We investigate the cost of using a given representation in terms of *individual fairness* [10]. This notion requires that similar decisions should be made for similar individuals, i.e. decisions are smooth. 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 representation 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 Z and X.

First, we restate a previous definition of individual fairness.

Definition 7 (Individual fairness [10]). *Let D and d be sub-additive 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 novel 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.

In order to bound the level of individual unfairness induced by a representation, 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 input) 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 input.

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, by the triangle inequality $D(h(f(x), h(f(x')) \leq d(x, x') + 2\epsilon$. See Appendix A.4 for complete proof. □

5 BENEFITS OF FAIR REPRESENTATION LEARNING

We quantify the benefits of some representation function f by measuring the discrimination achieved by an optimal adversary using Z, the representation variable induced by 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 $Y = 0$) [15].

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 Z can be expressed in closed form. The maximum and minimum will be closer if the induced conditional probability $\eta_S^f(z)$ does not deviate too much on average from the prior π_S. If $\eta_S^f(z) = \pi_S$ everywhere, we have statistical parity of zero, i.e. exact equality of outcome.

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

$$\max_{g \in G} SP(g \circ f) = 1 - \mathbb{E}_Z[\min(\frac{\eta_S^f(Z)}{\pi_S}, \frac{1 - \eta_S^f(Z)}{1 - \pi_S})]$$

$$\min_{g \in G} SP(g \circ f) = -1 + \mathbb{E}_Z[\min(\frac{\eta_S^f(Z)}{\pi_S}, \frac{1 - \eta_S^f(Z)}{1 - \pi_S})].$$

Proof idea. Observe that statistical parity is a linear transformation of balanced error rate. Apply the minimum balanced error rate from Equation 32 of [28]. See Appendix A.5 for complete proof. □

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 Z can be expressed in closed form. The maximum and minimum will be closer if the induced conditional probability $\eta_S^f(z)$ never deviates too much from the prior π_S. If $\eta_S^f(z) = \pi_S$ 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 $\eta_S^f(z)$ to be close to π_S *everywhere* rather than only in expectation.

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

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

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

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 Z, and Bayes' rule. Using this form we obtain the maximum and minimum values of $DI(g \circ f)$ and the corresponding choices of g. See Appendix A.6 for complete proof. □

6 CONCLUSION

We have quantified the costs – an inferior fairness-accuracy tradeoff and an increase in individual unfairness – incurred by a given representation. We have also quantified the benefits – narrower bands of 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

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. *Fairness and Machine Learning*. fairmlbook.org.

[3] Solon Barocas and Andrew D Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671.

[4] Yahav Bechavod and Katrina Ligett. 2017. Penalizing Unfairness in Binary Classification. arXiv 1707.00044.

[5] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. In *FAT/ML Workshop*.

[6] Toon Calders and Sicco Verwer. 2010. Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[7] Flavio P Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Data Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*.

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

[9] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical Risk Minimization Under Fairness Constraints. In *Advances in Neural Information Processing Systems*.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Innovations in Theoretical Computer Science Conference*.

[11] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Mark DM Leiserson. 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *Conference on Fairness, Accountability and Transparency*.

[12] Harrison Edwards and Amos Storkey. 2016. Censoring Representations with an Adversary. In *International Conference on Learning Representations*.

[13] Charles Elkan. 2001. The Foundations of Cost-Sensitive Learning. In *International Joint Conference on Artificial Intelligence*.

[14] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *International Conference on Knowledge Discovery and Data Mining*.

[15] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*.

[16] James Johndrow and Kristian Lum. 2017. An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction. arXiv 1703.04957.

[17] Christos Louizos, Kevin Swersky, Yujia Li, Richard Zemel, and Max Welling. 2016. The Variational Fair Autoencoder. In *International Conference on Learning Representations*.

[18] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *International Conference on Machine Learning*.

[19] Aditya Krishna Menon and Robert C. Williamson. 2018. The Cost of Fairness in Binary Classification. In *Conference on Fairness, Accountability and Transparency*.

[20] Shira Mitchell and Jackie Shadlen. 2018. Mirror Mirror: Reflections on Quantitative Fairness. https://speak-statistics-to-power.github.io/fairness.

[21] Cathy O'Neil. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

[22] Andrea Romei and Salvatore Ruggieri. 2014. A Multidisciplinary Survey on Discrimination Analysis. *The Knowledge Engineering Review* 29, 5 (2014), 582–638.

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

[24] Martin J Wainwright, Michael I Jordan, and John C Duchi. 2012. Privacy Aware Learning. In *Advances in Neural Information Processing Systems*.

[25] 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 *International Conference on World Wide Web*.

[26] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *International Conference on Artificial Intelligence and Statistics*.

[27] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *International Conference on Machine Learning*.

[28] 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. *Journal of Machine Learning Research* 14 (2013), 1033–1090.

[29] Indre Zliobaite. 2015. A Survey on Measuring Indirect Discrimination in Machine Learning. arXiv 1511.00148.

A THEOREM PROOFS

We present complete proofs of our theoretical results.

A.1 Proof of Theorem 1

Proof. Let $h^*_Y \in \arg\min_{h \in H} R_Y(h)$, which is given by the expression $h^*_Y(x) = \mathbf{1}(\eta_Y(x) \geq c_Y)$ [28].

$$
\begin{aligned}
& R_Y(g^*_Y \circ f) - l_Y \lambda R_S(g^*_S \circ f) \\
&\leq R_Y(h^*_Y \circ f) - l_Y \lambda R_S(g^*_S \circ f) \\
&= R_Y(h^*_Y \circ f) - R_Y(h^*_Y) + R_Y(h^*_Y) - l_Y \lambda R_S(g^*_S \circ f) \\
&= \mathbb{E}_X[(c_Y - \eta_Y(X))h^*_Y(f(X))] - \mathbb{E}_X[(c_Y - \eta_Y(X))h^*_Y(X)] \\
&\quad + R_Y(h^*_Y) - l_Y \lambda R_S(g^*_S \circ f) && (18) \\
&= \mathbb{E}_X[(c_Y - \eta_Y(X))(h^*_Y(f(X)) - h^*_Y(X))] \\
&\quad + R_Y(h^*_Y) - l_Y \lambda R_S(g^*_S \circ f) && (19) \\
&\leq l_Y \mathbb{E}_X \|X - f(X)\|_2 + R_Y(h^*_Y) - l_Y \lambda R_S(g^*_S \circ f). && (20)
\end{aligned}
$$

For (18) we apply Lemma 9 from [19]. For (19) we apply linearity of expectation. For (20), for any x where $h^*_Y(x) \neq h^*_Y(f(x))$, there must exist some x' on the decision boundary of h^* such that

$$c_Y - \eta_Y(x') = 0 \tag{21}$$

and

$$\|x - x'\|_2 \leq \|x - f(x)\|_2. \tag{22}$$

Combining (21) and (22) with the Lipschitz condition (9) 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

$$
\begin{aligned}
& \arg\min_{f \in F} [l_Y \mathbb{E}_X \|X - f(X)\|_2 + R_Y(h^*_Y) - l_Y \lambda R_S(g^*_S \circ f)] \\
&= \arg\min_{f \in F} [\mathbb{E}_X \|X - f(X)\|_2 - \lambda R_S(g^*_S \circ f)].
\end{aligned}
$$

□

A.2 Proof of Theorem 2

Proof. First we show the analytical expression for the cost of mistrust (13). Applying Proposition 4 of [19], we have that (11) and (12) are hypotheses h^* and g^* corresponding to solutions to (3) and (10) respectively.

$$
\begin{aligned}
& \mathbb{E}_X[\min(\eta_Y(X) - c_Y, \lambda(\eta_S(X) - c_S))] && (23) \\
&= \mathbb{E}_X[(1 - h^*(X))(\eta_Y(X) - c_Y)] + \lambda \mathbb{E}_X[h^*(X)(\eta_S(X) - c_S)] && (24) \\
&= \pi_Y - c_Y - \mathbb{E}_X[h^*(X)(\eta_Y(X) - c_Y)] + \lambda \mathbb{E}_X[h^*(X)(\eta_S(X) - c_S)] \\
&= \pi_Y - c_Y + R_Y(h^*) - (1 - c_Y)\pi_Y + \lambda \mathbb{E}_X[h^*(X)(\eta_S(X) - c_S)] && (25) \\
&= \pi_Y - c_Y + R_Y(h^*) - (1 - c_Y)\pi_Y - \lambda R_S(h^*) + \lambda(1 - c_S)\pi_S && (26) \\
&= R_Y(h^*) - \lambda R_S(h^*) - c_Y(1 - \pi_Y) + \lambda(1 - c_S)\pi_S.
\end{aligned}
$$

(24) follows from the form of h^* given in (11). (25) and (26) both involve substitutions using Lemma 9 from [19].

Similarly, using the form of g^* from (12) we conclude that

$$
\begin{aligned}
& \mathbb{E}_X[\min(\eta^f_Y(f(X)) - c_Y, \lambda(\eta^f_S(f(X)) - c_S))] && (27) \\
&= R_Y(g^* \circ f) - \lambda R_S(g^* \circ f) - c_Y(1 - \pi_Y) + \lambda(1 - c_S)\pi_S. && (28)
\end{aligned}
$$

The result (13) follows by subtracting (23) from (27) and applying linearity of expectation.

The decomposed form follows from applying Lemma 9 of [19] 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 $R_Y(g^* \circ f) - R_Y(h^*)$ as in (14) and $R_S(g^* \circ f) - R_S(h^*)$ as in (15). □

A.3 Proof of Theorem 3

Proof.

$$
\begin{aligned}
& R_{YS}(g^* \circ f) - R_{YS}(h^*) \\
& \leq R_{YS}(h^* \circ f) - R_{YS}(h^*) \\
& = R_Y(h^* \circ f) - R_Y(h^*) - \lambda(R_S(h^* \circ f) - R_S(h^*)) \\
& = \mathbb{E}_X[(c_Y - \eta_Y(X))(h^*(f(X)) - h^*(X))] \\
& \quad - \lambda \mathbb{E}_X[(c_S - \eta_S(X))(h^*(f(X)) - h^*(X))] && (29) \\
& = \mathbb{E}_X[(c_Y - \eta_Y(X) - \lambda(c_S - \eta_S(X)))(h^*(f(X)) - h^*(X))] && (30) \\
& \leq (l_Y + \lambda l_S)\mathbb{E}_X\|X - f(X)\|_2. && (31)
\end{aligned}
$$

(29) is by Lemma 9 from [19] and linearity of expectation. (30) is by linearity of expectation. For (31), using the form of h^* from (11), for any x where $h^*(x) \neq h^*(f(x))$, there must exist some x' on the decision boundary of h^* such that

$$c_Y - \eta_Y(x') - \lambda(c_S - \eta_S(x')) = 0 \tag{32}$$

and

$$\|x - x'\|_2 \leq \|x - f(x)\|_2. \tag{33}$$

Combining (32) and (33) with the Lipschitz conditions (16) and (17),

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

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

A.4 Proof of Theorem 4

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 $1 - \delta$,

$$d(x, f(x)) \leq \epsilon. \tag{34}$$

Similarly, with probability $1 - \delta$,

$$d(x', f(x')) \leq \epsilon. \tag{35}$$

By the union bound, both statements hold with probability at least $1 - 2\delta$. In that case, the following statements also hold:

$$
\begin{aligned}
& D(h(f(x), h(f(x')) \\
& \leq D(h(f(x)), h(x)) + D(h(x), h(f(x'))) && (36) \\
& \leq \epsilon + D(h(x), h(f(x'))) && (37) \\
& \leq \epsilon + D(h(x), h(x')) + D(h(x'), h(f(x'))) && (38) \\
& \leq 2\epsilon + D(h(x), h(x')) && (39) \\
& \leq 2\epsilon + d(x, x'). && (40)
\end{aligned}
$$

(36) and (38) use the triangle inequality since D is subadditive. (37) and (39) hold due to (34) and (35) respectively, along with Definition 7. (40) uses Definition 7. Therefore $IU_{D,d_\epsilon}(h \circ f) \leq 2\delta$. □

A.5 Proof of Theorem 5

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 an hypothesis h. Observe the fact that $SP(h) = 1 - 2BER(h)$ for all h. Therefore

$$
\begin{aligned}
& \max_{g \in G} SP(g \circ f) \\
& = 1 - 2\min_{g \in G} BER(g \circ f) \\
& = 1 - \mathbb{E}_Z[\min(\frac{\eta_S^f(Z)}{\pi_S}, \frac{1 - \eta_S^f(Z)}{1 - \pi_S})], && (41)
\end{aligned}
$$

where (41) uses Equation 32 from [28].

Similarly, we may show that

$$\min_{g \in G} SP(g \circ f) = -1 + \mathbb{E}_Z[\min(\frac{\eta_S^f(Z)}{\pi_S}, \frac{1 - \eta_S^f(Z)}{1 - \pi_S})],$$

using the fact that $BER(h) = 1 - BER(1 - h)$ for all h. □

A.6 Proof of Theorem 6

Proof. Disparate impact can be expressed as follows:

$$
\begin{aligned}
& DI(g \circ f) \\
& = \frac{\int_z p(Z = z|S = 0)p(\hat{Y}^f = 1|S = 0, Z = z)dz}{\int_z p(Z = z|S = 1)p(\hat{Y}^f = 1|S = 1, Z = z)dz} \\
& = \frac{\int_z p(Z = z|S = 0)g(z)dz}{\int_z p(Z = z|S = 1)g(z)dz} && (42) \\
& = \frac{\pi_S \int_z p(Z = z)(1 - \eta_S^f(z))g(z)dz}{(1 - \pi_S)\int_z p(Z = z)\eta_S^f(z)g(z)dz} && (43) \\
& = \frac{\pi_S \mathbb{E}_Z[(1 - \eta_S^f(Z))g(Z)]}{(1 - \pi_S)\mathbb{E}_Z[\eta_S^f(Z)g(Z)]}. && (44)
\end{aligned}
$$

For (42) we used the fact that $\hat{Y}^f$ and S are conditionally independent given Z. For (43) 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 (44), we have:

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

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,

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

where the minimum is obtained for

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

□