The Cost of Fairness in Binary Classification

Aditya Krishna Menon ADITYA.MENON@ANU.EDU.AU
The Australian National University, Canberra, Australia

Robert C. Williamson BOB.WILLIAMSON@ANU.EDU.AU
The Australian National University and DATA61, Canberra, Australia

Editors: Sorelle Friedler and Christo Wilson

Abstract

Binary classifiers are often required to possess *fairness* in the sense of not overly discriminating with respect to a feature deemed sensitive, e.g. race. We study the inherent tradeoffs in learning classifiers with a fairness constraint in the form of two questions: what is the best accuracy we can expect for a given level of fairness?, and what is the nature of these optimal fairness-aware classifiers? To answer these questions, we provide three main contributions. First, we relate two existing fairness measures to cost-sensitive risks. Second, we show that for such cost-sensitive fairness measures, the optimal classifier is an instance-dependent thresholding of the class-probability function. Third, we relate the tradeoff between accuracy and fairness to the alignment between the target and sensitive features' class-probabilities. A practical implication of our analysis is a simple approach to the fairness-aware problem which involves suitably thresholding class-probability estimates.

1. Introduction

Suppose we wish to learn a classifier to determine if an applicant will repay a loan. That is, given various input features about the applicant – such as their employment status, income, and credit history – we wish to predict the target feature, namely likelihood of repaying the loan. Suppose however that one of the input features is deemed sensitive, e.g. their race. Then, we might be required to constrain the classifier to not be overly discriminative with respect to this sensitive feature; subject to this constraint, we would of course like our classifier to be as accurate at predicting the target feature as possible. This *fairness-aware learning* problem has received considerable attention in the machine learning community of late (Pedreshi et al., 2008; Kamiran and Calders, 2009; Calders and Verwer, 2010; Dwork et al., 2012; Kamishima et al., 2012; Fukuchi et al., 2013; Zafar et al., 2016; Hardt et al., 2016; Zafar et al., 2017).

Existing work on fairness-aware learning has largely focussed on two central questions: (a) how does one formally measure the fairness of a classifier?, and (b) given such a measure, how does one learn a classifier that achieves fairness? For the former, the challenge is that seemingly sensible definitions of fairness can have subtle, undesirable consequences (Žliobaitė et al., 2011; Dwork et al., 2012); to address this, a range of progressively refined measures have been designed (Calders and Verwer, 2010; Dwork et al., 2012; Hardt et al., 2016; Zafar et al., 2017). For the latter, the challenge is that merely ignoring the sensitive feature is inadmissible, owing to it potentially being predictable by other features (Pedreshi et al., 2008); to address this, approaches based on post-hoc correction (Calders and Verwer, 2010; Hardt et al., 2016), regularisation (Kamishima et al., 2012), and surrogate loss minimisation have been proposed (Zafar et al., 2016, 2017).

1.1. The limits of the possible in fairness-aware learning

Despite the impressive advances detailed above, some basic theoretical aspects of the fairness-aware problem have received less attention. For example, before attempting to design an algorithm targetting a particular measure of fairness, it is natural to ask:

Q1: **What is the best we can do?** There is typically an unavoidable tradeoff between *how accurate* our classifier is with respect to the target feature, and *how fair* it is with respect to the sensitive feature. One may seek to *quantify* this tradeoff in terms of properties of the data source, giving inherent limits of what is possible for *any* possible algorithm.

Q2: **How do we achieve the best?** Having determined the inherent accuracy-fairness tradeoff, one may seek to find *what classifiers achieve this limit*. This is not purely of theoretical import, as we may then seek to design methods to approximate these optimal classifiers.

In machine learning terminology, **Q2** concerns the *Bayes-optimal classifiers* for the fairness-aware learning problem. The Bayes-optimal classifier attains the lowest possible average error for a given problem: thus, *no algorithm, no matter how clever or sophisticated, can attain lower average error than this classifier*. Such classifiers are foundational in the study of standard binary classification (Devroye et al., 1996), and provide a "limit of the possible" in a manner similar to what Shannon's information theory did for practical problems of telecommunication (Gleick, 2011), or what the science of thermodynamics did for heat engines in the 19th century (Bryant, 1973).

1.2. A Mathematics of Morality?

> *We should strive for a kind of moral geometry with all the rigor which this name connotes.* (Rawls, 1971, pg. 121)

In this paper, we present three contributions in the study of **Q1** and **Q2**. First, we show that two popular fairness measures can be seen as instances of a more general scheme. We then show that for this general scheme of fairness measures, the Bayes-optimal classifier can be explicitly computed. Intuitively, this classifier deems an instance to be positive if the probability of the target feature being active is sufficiently higher than the probability of the sensitive feature being active. Finally, we use the explicit form of this optimal classifier to provide an analytical expression for the fundamental tradeoff curve of accuracy versus fairness. This curve is shown to depend on measure of similarity between the target and sensitive features.

Our approach in answering **Q1** and **Q2** is mathematical in nature, but does not purport to provide a complete answer to the probems of fairness! It is motivated by statements such as that due to John Rawls quoted above. Such an approach provides the ability to make precise statements at a considerable level of generality, and is comparable to the formal analysis of problems (including fairness) in welfare economics (Harsanyi, 1955; Sen, 2009). In the context of this literature, our work follows the precept of Sen (2009, Chapter 18) that mere identification of "fully just social arrangements is neither necessary nor sufficient." We embrace Sen's pragmatism by focussing on the *quantifiable tradeoffs* one might make to approach (certain notions of) fairness. However, we do *not* claim to derive the "right" tradeoffs, neither in the choice of loss functions to be used, nor even their relative weights in contrast to Rawls (1971, pg 37ff) who not only ackowledges there will be tradeoffs between overall social utility and fairness but tries to argue what the "right" tradeoff is. See Appendix H for further relations to the philosophical and economic literature on fairness.

Formally, our main contributions **C1**—**C3** are:

C1: we reduce two popular fairness measures (disparate impact and mean difference) to cost-sensitive risks (Lemmas 1, 2).

C2: we show that for cost-sensitive fairness measures, the optimal fairness-aware classifier is an *instance-dependent thresholding* of the class-probability function (Propositions 4, 6).

C3: we quantify the intrinsic, method-independent impact of the fairness requirement on accuracy via a notion of *alignment* between the target and sensitive feature (Proposition 8).

Our results deal with the theoretical limits of what is possible for *any* fairness-aware learning method (for the class of fairness measures we consider), given access to the theoretical population distribution. This leaves some important questions unanswered, such as how one can construct a classifier that is optimal for a *given* dataset (rather than the theoretical population). While we do not provide a complete answer to this matter, we *do* provide a practical means of *approximating* the Bayes-optimal classifier. Specifically, the form of the optimal fairness-aware classifiers (**C2**) lets us derive a practically usable algorithm, wherein we separately estimate class-probabilities for the target and sensitive features, e.g. by logistic regression, and combine them suitably (§5.2).

2. Background and notation

We fix notation and review background. Table 1 summarises some symbols that we frequently use.

2.1. Standard learning from binary labels

Let $\mathcal{X}$ be a measurable instance space, e.g. characteristics of an applicant for a loan. In standard learning from binary labels, we have samples from a distribution D over $\mathcal{X} \times \{0, 1\}$, with $(\mathsf{X}, \mathsf{Y}) \sim D$. Here,

Y is some *target feature* we would like to predict, e.g. whether to grant a loan. Our goal is to output a measurable *randomised classifier*[1] parametrised by $f\colon \mathcal{X} \to [0,1]$ that distinguishes between positive ($\mathsf{Y}=1$) and negative ($\mathsf{Y}=0$) instances. A randomised classifier predicts any $x \in \mathcal{X}$ to be positive with probability $f(x)$; the quality of any such classifier is assessed by a *statistical risk*[2] $R(\cdot; D)\colon [0,1]^{\mathcal{X}} \to \mathbb{R}_+$. Often, this is some function of the *false negative* and *false positive rates*

$$\begin{aligned} \mathrm{FNR}(f; D) &\doteq \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=1}[1 - f(\mathsf{X})] \\ \mathrm{FPR}(f; D) &\doteq \mathop{\mathbb{E}}_{\mathsf{X}|\mathsf{Y}=0}[f(\mathsf{X})], \end{aligned} \tag{1}$$

viz. the class-conditional error probabilities when classifying x as positive with probability $f(x)$. In the sequel, we will drop dependence of quantities on the underlying distribution when it is clear from context; e.g. we will write $\mathrm{FNR}(f)$ in place of $\mathrm{FNR}(f; D)$.

Remark 2.1: A randomised classifier parameterised by $f\colon \mathcal{X} \to [0,1]$ should not be confused with two related objects: a deterministic classifier $g\colon \mathcal{X} \to \{0,1\}$, and a class-probability estimator $h\colon \mathcal{X} \to [0,1]$. The former produces deterministic (non-random) classifications for each input. Consequently, e.g., in contrast to Equation 1,

$$\mathrm{FNR}(g; D) = \mathop{\mathbb{P}}_{\mathsf{X}|\mathsf{Y}=1}[g(\mathsf{X}) = 0]\,.$$

The latter emits the confidence that an instance has positive label, as per $\mathbb{P}(\mathsf{Y}=1 \mid \mathsf{X}=x)$. However, one typically uses this to make deterministic classifications, e.g. by constructing the classifier $g(x) = [\![h(x) > 0.5]\!]$. Thus, while h has the same type as f, the resulting predictions and their evaluation are different; e.g., in contrast to Equation 1,

$$\mathrm{FNR}(h; D) = \mathop{\mathbb{P}}_{\mathsf{X}|\mathsf{Y}=1}[h(\mathsf{X}) < 0.5]\,.$$

A canonical risk is the *cost-sensitive risk*, which for cost parameter $c \in (0,1)$ and $\pi \doteq \mathbb{P}(\mathsf{Y}=1)$ is

$$\mathrm{CS}(f; c) \doteq \pi\cdot(1-c)\cdot\mathrm{FNR}(f) + (1-\pi)\cdot c\cdot\mathrm{FPR}(f). \tag{2}$$

When $c = \pi$, this is a scaled version of the *balanced error*, $\mathrm{BER}(f) = (\mathrm{FNR}(f) + \mathrm{FPR}(f))/2$.

A *Bayes-optimal randomised classifier* for a risk is any minimiser $f^* \in \operatorname{Argmin} R(f; D)$. For the cost-sensitive risk with parameter c, the Bayes-optimal classifier is $f^*(x) = [\![\eta(x) > c]\!] + \alpha \cdot [\![\eta(x) = c]\!]$ (Elkan, 2001), where $\eta(x) \doteq \mathbb{P}(\mathsf{Y}=1 \mid \mathsf{X}=x)$ is the class-probability function, $[\![\mathrm{E}]\!] = 1$ if E is true, and zero otherwise, and $\alpha \in [0,1]$ is arbitrary. Consequently, for the 0-1 loss corresponding to $c = 1/2$, we classify as deterministically positive those instances whose label is on average more likely to be positive.

2.2. Fairness-aware learning

In fairness-aware learning, one modifies the standard binary label learning problem in two ways. The statistical setup is modified by assuming that in addition to the target feature Y, there is some binary *sensitive feature* $\bar{\mathsf{Y}}$ we would like to treat in a special way, e.g. the race of an applicant. The classifier evaluation is modified by assuming that we reward classifiers that are "fair" in the treatment of $\bar{\mathsf{Y}}$.[3] To make this goal concrete, the literature has studied notions of *perfect* and *approximate fairness*.

Perfect fairness. We consider two notions of perfect fairness, stated in terms of $\mathsf{Y}, \bar{\mathsf{Y}}$, and classifier prediction $\hat{\mathsf{Y}} \mid \mathsf{X} \sim \mathrm{Bern}(f(\mathsf{X}))$. The first is *demographic parity* (*DP*) (Calders and Verwer, 2010), which requires the predictions to be independent of the sensitive feature:

$$\mathbb{P}(\hat{\mathsf{Y}}=1 \mid \bar{\mathsf{Y}}=0) = \mathbb{P}(\hat{\mathsf{Y}}=1 \mid \bar{\mathsf{Y}}=1). \tag{3}$$

The second is *equality of opportunity* (*EO*) (Hardt et al., 2016), which requires the predictions to be independent of the sensitive feature, but only for the positive instances:

$$\mathbb{P}(\hat{\mathsf{Y}}=1 \mid \mathsf{Y}=1, \bar{\mathsf{Y}}=0) = \mathbb{P}(\hat{\mathsf{Y}}=1 \mid \mathsf{Y}=1, \bar{\mathsf{Y}}=1). \tag{4}$$

Other notions include equalised odds (Hardt et al., 2016), and lack of disparate mistreatment (Zafar et al., 2017). Demographic parity has received the most study, but is known to have deficiencies (Dwork et al., 2012; Hardt et al., 2016; Zafar et al., 2017).

Approximate fairness. We consider two notions of approximate fairness, via *fairness measures* that quantify the degree of fairness for a classifier. The

1. Randomised classifiers are commonly used in rectifying local non-concavities in ROC curves (Fawcett, 2006).
2. Strictly, this is an abuse of terminology, as risks conventionally refer to expected losses. In our context, this corresponds to using functionals that only linearly combine the false positive and negative rates, i.e. , cost-sensitive risks. Fortunately, these are precisely the class of $R(\cdot)$ that we consider in the sequel.
3. The sensitive feature may or may not be available during training (see §5.1.1); even if not available, other correlated features may induce discrimination (Pedreshi et al., 2008).

Symbol	Meaning	Symbol	Meaning	Symbol	Meaning	Symbol	Meaning
X	Instance	D	Distribution $\mathbb{P}(\mathsf{X}, \mathsf{Y})$	f	Randomised classifier	CS	Cost-sensitive risk
Y	Target feature	$\bar{D}$	One of $\{\bar{D}_{\mathrm{DP}}, \bar{D}_{\mathrm{EO}}\}$	$\eta(x)$	$\mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x)$	$\mathrm{CS}_{\mathrm{bal}}$	Balanced CS risk
$\bar{\mathsf{Y}}$	Sensitive feature	$\bar{D}_{\mathrm{DP}}$	Distribution $\mathbb{P}(\mathsf{X}, \bar{\mathsf{Y}})$	$\bar{\eta}_{\mathrm{DP}}(x)$	$\mathbb{P}(\bar{\mathsf{Y}} = 1 \mid \mathsf{X} = x)$	DI	Disparate impact
$\hat{\mathsf{Y}}$	Prediction	$\bar{D}_{\mathrm{EO}}$	Distribution $\mathbb{P}(\mathsf{X}, \bar{\mathsf{Y}} \mid \mathsf{Y} = 1)$	$\bar{\eta}_{\mathrm{EO}}(x, y)$	$\mathbb{P}(\bar{\mathsf{Y}} = 1 \mid \mathsf{X} = x, \mathsf{Y} = y)$	MD	Mean difference

Table 1: Glossary of commonly used symbols.

first is the *disparate impact* (*DI*) *factor* (Feldman et al., 2015), which is the ratio of the probabilities appearing in Equation 5,

$$\mathrm{DI}(f) \doteq \mathbb{P}(\hat{\mathsf{Y}} = 1 \mid \bar{\mathsf{Y}} = 0) / \mathbb{P}(\hat{\mathsf{Y}} = 1 \mid \bar{\mathsf{Y}} = 1). \quad (5)$$

The second is the *mean difference* (*MD*) *score* (Calders and Verwer, 2010),

$$\mathrm{MD}(f) \doteq \mathbb{P}(\hat{\mathsf{Y}} = 1 \mid \bar{\mathsf{Y}} = 0) - \mathbb{P}(\hat{\mathsf{Y}} = 1 \mid \bar{\mathsf{Y}} = 1). \quad (6)$$

We refer the reader to Žliobaitė (2017) for a survey of other fairness measures. While some scenarios may demand perfect fairness, having a strong degree of approximate fairness may be acceptable in others; for example, disparate impact has its roots in the 80% rule of the U.S. Equal Employment Opportunity Commission (EEOC, 1979).

Remark 2.2: The DI factor takes values in the range $[0, \infty]$, and the MD score in $[-1, +1]$. Demographic parity is achieved when $\mathrm{DI}(f) = 1$, or $\mathrm{MD}(f) = 0$; thus, for both measures, it is undesirable for the score to be too small *or* too large.

It is often implicitly assumed that f is such that $\mathrm{DI}(f) \leq 1$, or $\mathrm{MD}(f) \leq 0$, in which case a large score is desirable. Alternately, one can work with "symmetrised" versions of these measures, e.g. $\mathrm{DI}^{\diamond}(f) \doteq \min(\mathrm{DI}(f), \mathrm{DI}(1 - f))$. Maximising this ensures that one cannot predict the sensitive feature merely by flipping outputs. Formally, observe that e.g. $\mathrm{DI}^{\diamond}(f) \in [0, 1]$, and $\mathrm{DI}^{\diamond}(f) = 1 \iff \mathrm{DI}(f) = 1$, i.e., we have perfect fairness; see Appendix C.

3. Fairness-aware learning and risk difference minimisation

We now formalise the fairness-aware learning problem by viewing fairness measures as statistical risks. We assume the following statistical setup. Let random variables $(\mathsf{X}, \bar{\mathsf{Y}}, \mathsf{Y}) \sim D_{\mathrm{jnt}}$ for some joint distribution D_{jnt} over $\mathcal{X} \times \{0, 1\} \times \{0, 1\}$, where X represents the instance, Y the target feature, and $\bar{\mathsf{Y}}$ the sensitive feature. Suppose D refers to the distribution $\mathbb{P}(\mathsf{X}, \mathsf{Y})$, and $\bar{D}$ to a suitable distribution over $(\mathsf{X}, \bar{\mathsf{Y}})$; concretely, in the demographic parity setting $\bar{D} = \bar{D}_{\mathrm{DP}} \doteq \mathbb{P}(\mathsf{X}, \bar{\mathsf{Y}})$, while in the equality of opportunity setting $\bar{D} = \bar{D}_{\mathrm{EO}} \doteq \mathbb{P}(\mathsf{X}, \bar{\mathsf{Y}} \mid \mathsf{Y} = 1)$.

3.1. Existing fairness measures as risks

We begin with a simple observation: the DI and MD fairness measures are transformations of suitable false positive and negative rates, and thus are statistical risks. Specifically, for any randomised classifier[4] $f \colon \mathcal{X} \to [0, 1]$ with predictions $\hat{\mathsf{Y}} \mid \mathsf{X} \sim \mathrm{Bern}(f(\mathsf{X}))$, we have $\mathrm{FNR}(f; \bar{D}_{\mathrm{DP}}) = \mathbb{P}(\hat{\mathsf{Y}} = 0 \mid \bar{\mathsf{Y}} = 1)$ and $\mathrm{FPR}(f; \bar{D}_{\mathrm{DP}}) = \mathbb{P}(\hat{\mathsf{Y}} = 1 \mid \bar{\mathsf{Y}} = 0)$. Consequently, for $\bar{D} = \bar{D}_{\mathrm{DP}}$,

$$\begin{aligned} \mathrm{DI}(f; \bar{D}) &= \mathrm{FPR}(f; \bar{D}) / (1 - \mathrm{FNR}(f; \bar{D})) \\ \mathrm{MD}(f; \bar{D}) &= \mathrm{FPR}(f; \bar{D}) + \mathrm{FNR}(f; \bar{D}) - 1. \end{aligned} \quad (7)$$

Observe that one can equally choose $\bar{D} = \bar{D}_{\mathrm{EO}}$ so as to yield approximate fairness measures for the equality of opportunity setting; clearly, when e.g. $\mathrm{DI}(f; \bar{D}_{\mathrm{EO}}) = 1$, we recover Equation 4. The notion that fairness measures are risks on $\bar{D}$ is implicit in prior surveys, e.g. (Žliobaitė, 2017). By making this notion explicit, we may now succinctly state the fairness-aware learning problem.

3.2. Fairness-aware learning: general case

Informally, fairness-aware learning involves finding a randomised classifier $f \colon \mathcal{X} \to [0, 1]$ so that Y is well predicted, but $\bar{\mathsf{Y}}$ is not. Formally, suppose that we measure how well f can predict Y and $\bar{\mathsf{Y}}$ via two risks: a *performance measure* $R_{\mathrm{perf}}(\cdot; D) \colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$, and *fairness measure* $R_{\mathrm{fair}}(\cdot; \bar{D}) \colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$. For example, we might pick R_{perf} to be a cost-sensitive risk, and R_{fair} to be one of DI or MD. Then, we seek to minimise the *difference* of two statistical risks, as below.

Problem 3.1 (Fairness-aware learning): For trade-off $\lambda \in \mathbb{R}$, minimise the fairness-aware objective

$$R_{\mathrm{perf}}(f; D) - \lambda \cdot R_{\mathrm{fair}}(f; \bar{D}). \quad (8)$$

4. This is understood to mean a randomised classifier parametrised by f.

The tradeoff parameter λ determines how we balance the competing goals of accuracy and fairness. We do *not* constrain $\lambda > 0$ for a subtle, but important reason: as per Remark 2.2, for the DI and MD scores, it is undesirable for the fairness measure to be too small *or* too large. Equation 8 effectively considers a constrained version of the problem, where $R_{\text{fair}}(f) \in [\tau_0, \tau_1]$, and converts it to Lagrangian form; λ corresponds to the difference in Lagrange multipliers for the two bounds, which can be negative. In §4.2, we will see that another interpretation for the DI and MD scores is that we constrain their "symmetrised" versions (per Remark 2.2) to be large.

Remark 3.1: In practical settings, one only has access to finite samples from the joint distribution D_{jnt}. By focussing on Equation 8, our analysis is thus asymptotic. We emphasise that our motivating questions **Q1** and **Q2** are nonetheless non-trivial: even with access to the underlying distributions, it is not obvious what the accuracy-fairness tradeoff is, nor what the form of the optimal classifier is. Further, we will show how to approximate the latter given only finite samples (§5.2).

3.3. Fairness-aware learning: cost-sensitive case

The generality of Problem 3.2 has conceptual appeal, but presents challenges if we are to proceed with our intended analysis of **Q1** and **Q2**. To make progress, we assume both the performance and fairness measures are cost-sensitive risks (Equation 2).

Problem 3.2 (Cost-sensitive fairness-aware learning): For tradeoff $\lambda \in \mathbb{R}$, and cost parameters $c, \bar{c} \in (0, 1)$, minimise the fairness-aware cost-sensitive risk[a]

$$R_{\text{FA}}(f; D, \bar{D}) \doteq \text{CS}(f; D, c) - \lambda \cdot \text{CS}(f; \bar{D}, \bar{c}). \quad (9)$$

a. More precisely, this is $R_{\text{FA}}(f; D, \bar{D}, c, \bar{c}, \lambda)$; we suppress $(c, \bar{c}, \lambda)$ as their scope will be clear from context.

Using a cost-sensitive risk as performance measure is not strongly limiting, as it can encompass more complex performance measures with distribution-dependent costs (Parambath et al., 2014; Narasimhan et al., 2015). However, it would be disappointing if this choice makes us unable to accommodate the popular DI and MD fairness measures. We now allay this concern.

4. A cost-sensitive view of existing fairness measures

The previous section cast the DI and MD measures as statistical risks. We now show they may be further related to cost-sensitive risks, implying that it suffices to analyse the latter. In the following, recall that for brevity we write e.g. $\text{FPR}(f)$ in place of $\text{FPR}(f; \bar{D})$.

4.1. Relating DI and MD to cost-sensitive risks

Underpinning our results is the *balanced cost-sensitive risk*,

$$\text{CS}_{\text{bal}}(f; c) \doteq (1 - c) \cdot \text{FNR}(f) + c \cdot \text{FPR}(f). \quad (10)$$

When $c = 1/2$, we get the balanced error. For general c, this is simply a scaled and re-parameterised version of the standard cost-sensitive risk; it will however prove more convenient in our analysis.

Our first result relates the disparate impact factor (Equation 7) and balanced cost-sensitive risk.

Lemma 1 *Pick any randomised classifier f. Then, for any $\tau \in (0, \infty)$, if $\kappa \doteq \frac{\tau}{1+\tau} \in (0, 1)$,*

$$\text{DI}(f) \geq \tau \iff \text{CS}_{\text{bal}}(f; 1 - \kappa) \geq \kappa.$$

Lemma 1 does not imply that disparate impact *equals* a cost-sensitive risk; rather, it says that a disparate impact *constraint* is equivalent to a cost-sensitive *constraint*, i.e. their super-level sets are equivalent. We shall shortly see that this is sufficient to justify learning with a cost-sensitive risk.

We next show an analogous (in fact stronger) result for the mean difference score (Equation 7).

Lemma 2 *Pick any randomised classifier f. Then, for any $\tau \in (-1, 1)$, if $\kappa \doteq \frac{1+\tau}{2} \in (0, 1)$,*

$$\text{MD}(f) = 2 \cdot \text{CS}_{\text{bal}}(f; 1/2) - 1 \quad (11)$$

$$\text{MD}(f) \geq \tau \iff \text{CS}_{\text{bal}}(f; 1/2) \geq \kappa. \quad (12)$$

Note that Equation 11 implies an equivalence of risks, and not just super-level sets. Note also that for the MD score, the corresponding balanced cost-sensitive risk has a cost-parameter that does *not* depend on τ. This proves beneficial for the purposes of learning, as we shall see in §5.2.

4.2. Implications for learning with DI and MD

Lemmas 1 and 2 establish the versatility of cost-sensitive fairness measures: we can reduce learning with the DI and MD scores to the cost-sensitive Problem 3.3 for a suitable choice of cost $\bar{c}$.

Lemma 3 *Pick any distributions $D, \bar{D}$, and fairness measure $R_{\text{fair}} \in \{\text{DI}, \text{MD}\}$. Pick any $c, \tau \in (0,1)$. Then, $\exists \lambda \in \mathbb{R}, \bar{c} \in (0,1)$ with*

$$\min_f \text{CS}(f; D, c) \colon \min(R_{\text{fair}}(f; \bar{D}), R_{\text{fair}}(1-f; \bar{D})) \geq \tau \equiv \min_f \text{CS}(f; D, c) - \lambda \cdot \text{CS}_{\text{bal}}(f; \bar{D}, \bar{c}). \tag{13}$$

The first objective in Equation 13 constrains that the "symmetrised" fairness measure (in the sense of Remark 2.2) is large; recall that maximising this quantity ensures perfect fairness, and thus the objective is sensible. The equivalence to the second objective is a consequence of the Lagrangian principle (see Appendix E), combined with Lemmas 1 and 2.

We emphasise that our focus on the cost-sensitive version of the fairness-aware learning problem is owing to Lemma 3: it implies that by analysing this special case, we encompass the use of the MD and DI scores as fairness measures. Moving beyond cost-sensitive risks is only beneficial if one is interested in a more exotic fairness measure that is inexpressible as such a risk.

Remark 4.1: One subtlety with Lemma 3 is that for the DI, we have to tune both λ *and* $\bar{c}$: this is because the cost in Lemma 1 depends on τ. For the MD, however, we can set $\bar{c} = 1/2$.

Related work. Lemma 1 can be seen a special case of a broader relationship between fractional performance measures and "level-finder" functions (Parambath et al., 2014, Theorem 1), (Narasimhan et al., 2015, Lemma 7). We are not however aware of prior results relating the DI and MD scores to cost-sensitive risks. The closest analogue is Feldman et al. (2015), who related (for $\tau = 0.8$) the DI to the balanced error BER via $\text{DI}(f) \leq 0.8 \iff \text{BER}(f) \leq (0.4 + 0.1 \cdot \text{FNR}(f))$. This bound depends on the distribution and classifier, while our bound on CS_{bal} uses a constant κ. Our result is thus simpler, and allows for tractable analysis of optimal classifiers (see §5).

Remark 4.2: Although not our primary focus, Lemma 1 also lets us address the problem of certifying whether a dataset is free of disparate impact Feldman et al. (2015) (i.e. for fixed τ, every classifier f satisfies $\text{DI}(f) \geq \tau$). By Lemma 1, this is equivalent to asking whether for every f, $\text{CS}_{\text{bal}}(f; 1-\kappa) \geq \kappa$, where $\kappa = \tau/(1+\tau)$. See Appendix F and G.1.

5. Optimal fairness-aware classifiers

Having justified the broad applicability of the cost-sensitive fairness-aware learning problem (Problem 3.3), we are in a position to examine its inherent tradeoffs. We begin by studying the question: what are the theoretically (i.e. Bayes-) optimal randomised classifiers, and how do they differ in the fairness-*un*aware problem? Formally, fix some D corresponding to $\mathbb{P}(\mathsf{X}, \mathsf{Y})$, and pick $\bar{D} \in \{\bar{D}_{\text{DP}}, \bar{D}_{\text{EO}}\}$. Then, for costs $c, \bar{c} \in (0,1)$ and tradeoff $\lambda \in \mathbb{R}$, we seek

$$f^* \in \operatorname*{Argmin}_{f \in [0,1]^{\mathcal{X}}} R_{\text{FA}}(f; D, \bar{D})$$

for the fairness-aware risk of Equation 8. Such Bayes-optimal classifiers represent the gold-standard for the learning problem, and computing them is thus of theoretical import. Further, we will shortly see that their explicit form suggests a simple practical algorithm.

Our results will be expressed in terms of three sets of quantities: the *base rates* $\pi \doteq \mathbb{P}(\mathsf{Y} = 1)$, $\bar{\pi} \doteq \mathbb{P}(\bar{\mathsf{Y}} = 1)$, the *class-probability functions* for the target and sensitive feature under demographic parity and equality of opportunity,

$$\begin{aligned} \eta(x) &\doteq \mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x) \\ \bar{\eta}_{\text{DP}}(x) &\doteq \mathbb{P}(\bar{\mathsf{Y}} = 1 \mid \mathsf{X} = x) \\ \bar{\eta}_{\text{EO}}(x) &\doteq \mathbb{P}(\bar{\mathsf{Y}} = 1 \mid \mathsf{X} = x, \mathsf{Y} = 1), \end{aligned} \tag{14}$$

and the modified Heaviside (or step) function $H_\alpha(z) \doteq [\![z > 0]\!] + \alpha \cdot [\![z = 0]\!]$ for parameter $\alpha \in [0,1]$.

5.1. Computing the Bayes-optimal classifiers

Thus far, we have not seen any distinction between the demographic parity and equality of opportunity settings; however, in computing the Bayes-optimal classifiers, a separate analysis is necessary.

5.1.1. Demographic parity

We begin with the explicit form of the optimal solutions under demographic parity, $\bar{D} = \bar{D}_{\text{DP}}$.

Proposition 4 *Pick any costs* $c, \bar{c} \in (0,1)$*, and* $\lambda \in \mathbb{R}$*. Then,*

$$\operatorname*{Argmin}_{f \in [0,1]^{\mathcal{X}}} R_{\mathrm{FA}}(f; D, \bar{D}_{\mathrm{DP}}) = \{H_\alpha \circ s^*(x) \mid \alpha \in [0,1]\} \tag{15}$$

$$\text{for } s^*(x) \doteq \eta(x) - c - \lambda \cdot (\bar{\eta}_{\mathrm{DP}}(x) - \bar{c}).$$

Three comments are in order. First, as a sanity check, when $\lambda = 0$, the optimal f^* comprises the familiar Bayes-optimal classifiers for a cost-sensitive risk which thresholds the class-probability η around c via $f^*(x) = [\![\eta(x) > c]\!] + \alpha \cdot [\![\eta(x) = c]\!]$. Here, α is arbitrary, since for instances at the threshold boundary $\eta(x) = c$, the risk is a constant and so any prediction is optimal.

Second, for $\lambda \neq 0$, the optimal f^* modifies the $\lambda = 0$ solution with an *instance dependent threshold correction*, which depends on $\bar{\eta}(x)$. The correction increases the standard threshold of c whenever $\bar{\eta}(x) > \bar{c}$; intuitively, when we are confident in the sensitive feature being active for an instance, we are more conservative in classifying the instance as positive for the target feature.

Third, the optimal classifier above is in fact *deterministic*, except when $\eta(x) = c + \lambda \cdot (\bar{\eta}(x) - \bar{c})$; in general, for a given λ, we expect this to only hold for few or no $x \in \mathcal{X}$.

In the above, we made no explicit assumption as to whether or not the sensitive feature is provided as input to the classifier. If we assume this feature is in fact available, the form of the optimal classifier simplifies dramatically.

Corollary 5 *Pick any costs* $c, \bar{c} \in (0,1)$*, and* $\lambda \in \mathbb{R}$*. Suppose* D *is over* $(\bar{\mathsf{X}}, \mathsf{Y})$*, where* $\bar{\mathsf{X}} = (\mathsf{X}, \bar{\mathsf{Y}})$*, and* $\eta(x, \bar{y}) \doteq \mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x, \bar{\mathsf{Y}} = \bar{y})$*. Then,*

$$\operatorname*{Argmin}_{f \in [0,1]^{\mathcal{X} \times \{0,1\}}} R_{\mathrm{FA}}(f; D, \bar{D}_{\mathrm{DP}}) = \left\{H_\alpha \circ s^* \mid \alpha \in [0,1]\right\}$$

$$s^*(x, 0) \doteq \eta(x, 0) - c + \lambda \cdot \bar{c}$$

$$s^*(x, 1) \doteq \eta(x, 1) - c - \lambda \cdot (1 - \bar{c}).$$

Here, we simply apply a constant threshold to the class-probabilities for each value of the sensitive feature. This is a simple consequence of Proposition 4, as we can simply consider one of the features of X to be perfectly predictive of the sensitive feature, which makes $\bar{\eta}_{\mathrm{DP}}(x, \bar{y}) \in \{0, 1\}$.

5.1.2. Equality of opportunity

We next turn to the optimal classifiers for the equality of opportunity setting, $\bar{D} = \bar{D}_{\mathrm{EO}}$. The result here is similar to Proposition 4, but with a multiplicative threshold correction. It is surprising that a simple change in fairness setting results in such a non-trivial modification of optimal solutions.

Proposition 6 *Pick any costs* $c, \bar{c} \in (0,1)$*, and* $\lambda \in \mathbb{R}$*. Then,*

$$\operatorname*{Argmin}_{f \in [0,1]^{\mathcal{X}}} R_{\mathrm{FA}}(f; D, \bar{D}_{\mathrm{EO}}) = \{H_\alpha \circ s^* \mid \alpha \in [0,1]\}$$

$$s^*(x) \doteq \left(1 - \lambda \cdot \pi^{-1} \cdot (\bar{\eta}_{\mathrm{EO}}(x, 1) - \bar{c})\right) \cdot \eta(x) - c. \tag{16}$$

When the sensitive feature is available, an analogous result to the previous section holds.

Corollary 7 *Pick any costs* $c, \bar{c} \in [0,1]$*, and* $\lambda \in \mathbb{R}$*. Then,*

$$\operatorname*{Argmin}_{f \in [0,1]^{\mathcal{X} \times \{0,1\}}} R(f; D, \bar{D}_{\mathrm{EO}}) = \left\{H_\alpha \circ s^* \mid \alpha \in [0,1]\right\}$$

$$s^*(x, 0) \doteq \left(1 + \lambda \cdot \pi^{-1} \cdot \bar{c}\right) \cdot \eta(x, 0) - c$$

$$s^*(x, 1) \doteq \left(1 - \lambda \cdot \pi^{-1} \cdot (1 - \bar{c})\right) \cdot \eta(x, 1) - c,$$

where $\eta(x, \bar{y}) = \mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x, \bar{\mathsf{Y}} = \bar{y})$.

Related work. Computing the Bayes-optimal classifiers as above is not without precedent: Hardt et al. (2016); Corbett-Davies et al. (2017) considered the same question, but in the case of *exact* fairness measures. We are not aware of prior work on computing the optimal classifiers for *approximate* fairness measures. While our results have a similar flavour to the exact fairness case, explicating them is important to understand the full tradeoff between accuracy and fairness (§6.1), and also suggests a simple algorithm as we now see.

5.2. A plugin approach to the fairness problem

The form of the Bayes-optimal classifiers above is not only of theoretical interest: they enable the derivation of practical classifiers suitable which are suitable for learning from finite samples, and rely on nothing more than logistic regression. As a warm up, recall that for standard cost-sensitive learning (i.e. $\lambda = 0$), the optimal classifier $f^*(x) = [\![\eta(x) > c]\!] + \alpha \cdot [\![\eta(x) = c]\!]$ suggests the following plugin approach: estimate η, e.g. by logistic regression, and then threshold the resulting predictions around c. This approach is intuitive, and provably consistent Narasimhan et al. (2014).

For the cost-sensitive fairness problem, the Bayes-optimal classifiers similarly rely on thresholding a suitable combination of the class-probabilities η and $\bar{\eta} \in \{\bar{\eta}_{\text{DP}}, \bar{\eta}_{\text{EO}}\}$. Thus, we can analogously construct a plugin classifier by estimating $\eta, \bar{\eta}$ separately, e.g. by logistic regression, and then combining them per Equations 15, 16. When the sensitive feature is available, we only need a single model for $\eta(x, \bar{y})$, which is thresholded separately for each sensitive feature value (Equation 16).

Algorithm 1 summarises this procedure for the demographic parity setting. Appendix G.2 presents some illustrative experiments for this algorithm.

Algorithm 1 Plugin approach to fairness-aware learning, demographic parity setting.

Input: Samples $\{(x_i, y_i, \bar{y}_i)\}_{i=1}^N$ from distribution D_{jnt}; cost parameters $c, \bar{c}$; tradeoff parameter λ
Output: Fairness-aware randomised classifier $f\colon \mathcal{X} \to [0, 1]$

Estimate $\eta\colon \mathcal{X} \to [0, 1]$ via logistic regression on $\{(x_i, y_i)\}_{i=1}^N$
Estimate $\bar{\eta}_{\text{DP}}\colon \mathcal{X} \to [0, 1]$ via logistic regression on $\{(x_i, \bar{y}_i)\}_{i=1}^N$
Compute $s\colon x \mapsto \hat{\eta}(x) - c - \lambda \cdot (\hat{\bar{\eta}}_{\text{DP}}(x) - \bar{c})$ from above estimates

Return $f\colon x \mapsto H_\alpha(s(x))$ for any $\alpha \in [0, 1]$

5.3. Strengths and weaknesses of the plugin approach

The plugin approach has several salient features:

(a) it reduces the problem to two calls of a logistic regression (or other class-probability estimation) solver, and avoids the need for any bespoke algorithms;

(b) as a consequence of (a), it involves a convex optimisation, unlike some existing approaches to the problem (Kamishima et al., 2012);

(c) tuning of the tradeoff parameter λ does *not* require any retraining: we can simply learn $\eta, \bar{\eta}$ once, and appropriately change how they are thresholded. The same holds true for tuning the cost parameter $\bar{c}$ when learning with the DI score.

Despite these appealing properties, we caution that in practice, the result of Algorithm 1 may be sub-optimal. This is because our estimates of the class-probabilities may be imperfect, owing to at least two distinct possible sources of error:

(a) there are limits on how accurately we can estimate these probabilities from a finite sample;

(b) the true class-probabilities may not be expressible in our chosen class of models.

These are manifestations of the more general issues of estimation and approximation error (Devroye et al., 1996) that plague any machine learning algorithm, even for standard binary classification. There are principled means of at least partially mitigating both issues, for example by employing suitable regularisation to prevent overfitting to a finite sample, and employing nonparameteric estimators to allow modelling of arbitrarily complex functions. Quantifying the degradation resulting from using imperfect probabilities is nonetheless an important but non-trivial task; indeed, until the recent work of Woodworth et al. (2017), we are not aware of *any* prior fairness-aware algorithm with finite sample guarantees.

Related work. The idea of correcting outputs to ensure fairness goes back to at least Calders and Verwer (2010), who proposed to modify the output of naïve Bayes so as to minimise the MD score. However, their approach does not have any optimality guarantees. More recently, Hardt et al. (2016) proposed to post-process the outputs of a classifier trained on the original problem, and argued for its optimality. They however worked in the exact rather than approximate fairness setting, and did not consider separate training procedures for predicting the target and sensitive features. Woodworth et al. (2017) established limits on such post-processing; extending this to approximate fairness measures would be of interest. Zafar et al. (2016, 2017) proposed to approximately solve Equation 9 by picking convex *surrogate losses* $\ell, \bar{\ell}\colon \{0, 1\} \times \mathbb{R} \to \mathbb{R}_+$, and find

$$s^* \in \operatorname*{Argmin}_{s\colon \mathcal{X} \to \mathbb{R}} \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [C_{\mathsf{Y}} \cdot \ell(\mathsf{Y}, s(\mathsf{X}))] - \lambda \cdot \mathop{\mathbb{E}}_{(\mathsf{X},\bar{\mathsf{Y}}) \sim \bar{D}} \left[\bar{C}_{\bar{\mathsf{Y}}} \cdot \bar{\ell}(\bar{\mathsf{Y}}, s(\mathsf{X}))\right] \tag{17}$$

for $C_1 = 1 - c, C_0 = c$. For nonlinear $\bar{\ell}$, this objective will be non-convex in s. Similar problems plague related approaches based on regularisation (Kamishima et al., 2012; Fukuchi et al., 2013).

Remark 5.1: We emphasise a conceptual difference between our plugin approach, and the surrogate approach of Equation 17. The latter aims to directly design a differentiable approximation to Equation 9, but it is non-trivial task for the resulting objective to be convex. By contrast, the plugin approach side-steps direct minimisation of this Equation 9, and instead follows a procedure that asymptotically results in the same theoretical optimal solution.

6. The accuracy-fairness tradeoff

As our final contribution, we now study the tradeoff between performance on our base problem and fairness, and show it is quantifiable by a measure of *alignment* of the target and sensitive variables.

6.1. The fairness frontier

Our definition of the cost-sensitive fairness-aware learning problem (Problem 3.3) was in terms of a linear tradeoff between the performance and fairness measures. To quantify the tradeoff imposed by a fairness constraint, we will study the following explicitly constrained problem: for $\tau \in [0,1]$, let

$$f^*_\tau \in \operatorname*{Argmin}_{f\colon \mathcal{X}\to[0,1]} \mathrm{CS}(f; D, c)\colon \mathrm{CS}(f; \bar{D}, \bar{c}) \geq \tau \tag{18}$$

$$F(\tau) = \mathrm{CS}(f^*_\tau; D, c) - \mathrm{CS}(f^*_0; D, c). \tag{19}$$

Equations 9 and 18 are related by the Lagrangian principle (see Appendix E). The function $F\colon \mathbb{R} \to \mathbb{R}_+$ represents the *fairness frontier*: for a given lower bound on fairness, it measures the *best excess risk* over the solution *without* a fairness constraint. Evidently, $F(\cdot)$ is non-decreasing since the constraints are nested as τ increases; i.e., demanding more fairness can never improve performance.

Remark 6.1: The tradeoff measured here is one *inherent* to the problem, rather than one owing to the specific technique one uses. By computing $F(\cdot)$, we determine the fundamental limits of what accuracy is achievable by *any* classifier, no matter how sophisticated it may be. Intuitively, this tradeoff captures the level of "contention" between utility and fairness in the distribution.

6.2. The frontier and probability alignment

The behaviour of $F(\cdot)$ summarises the tradeoff between performance and fairness: the steeper its growth, the more we have to sacrifice per unit increase in fairness. Can we relate this behaviour to properties of the underlying distributions $D, \bar{D}$? Note that since cost-sensitive risks are linear in the classifier, $F(\cdot)$ can be computed *empirically* via a linear program (see Appendix D); however, we seek a more abstract understanding of its behaviour. By analogy, we seek something akin to the notion of *compatibility functions* in semi-supervised learning (Balcan and Blum, 2010), wherein one can guarantee that unlabelled data is useful when there is an alignment of the marginal data distribution with one's function class.

Intuitively, we expect that the tradeoff between fairness and performance depends on how "similar" the target and sensitive features are: in the extreme case where they are one and the same, there is an inescapable linear penalty, while if they are completely dissimilar, we expect the penalty to be milder. This idea can be formalised via a notion of *disalignment* between the features. (For concreteness, the following is for demographic parity.)

Proposition 8 *Pick any cost parameters $c, \bar{c} \in (0,1)$. For any $\tau \in [0,1]$, there is some $\lambda \in \mathbb{R}$ and Bayes-optimal randomised classifier $f^* \in \operatorname{Argmin}_{f\in[0,1]^{\mathcal{X}}} R_{\mathrm{full}}(f; D, \bar{D}_{\mathrm{DP}})$ so that the frontier is*

$$F(\tau) = \mathop{\mathbb{E}}_{\mathsf{X}} \left[(c - \eta(\mathsf{X})) \cdot (f^*(\mathsf{X}) - [\![\eta(\mathsf{X}) > c]\!]) \right].$$

If further this f^* is deterministic i.e. $\mathrm{Im}(f^*) \subseteq \{0,1\}$, *the frontier is*

$$F(\tau) = \mathop{\mathbb{E}}_{\mathsf{X}} \left[B_\lambda(\eta(\mathsf{X}), \bar{\eta}_{\mathrm{DP}}(\mathsf{X})) \right] \tag{20}$$

$$B_\lambda(\eta, \bar{\eta}_{\mathrm{DP}}) \doteq |\eta - c| \cdot [\![(\eta - c) \cdot (\eta - c - \lambda \cdot (\bar{\eta}_{\mathrm{DP}} - \bar{c})) < 0]\!]. \tag{21}$$

Unpacking the above, Equation 21 gives a concrete notion of disalignment between η and $\bar{\eta}_{\mathrm{DP}}$ – how much they disagree around the respective thresholds c and $\bar{c}$ – and Equation 20 shows that when this disalignment is high, the fairness constraint has less of an effect. The requirement that f^* be deterministic may be dropped, at the expense of an additional term in Equation 20 that depends on the alignment of the non-deterministic component and η. Appendix G.3 presents some illustrations of this frontier on synthetic datasets.

Our prior analysis of the cost-sensitive fairness problem is crucial to establishing Proposition 8: we

exploit the explicit form of Bayes-optimal classifiers to Problem 3.3, and existing results about cost-sensitive risks. In particular, we use the fact that the frontier (Equation 19) is simply the *cost-sensitive regret* (or *excess risk*) of the classifier f^*_τ. This quantity can be related to a Bregman divergence (Reid and Williamson, 2009), and combined with our explicit form for the Bayes-optimal classifier, we obtain our Equation 21.

Relation to existing work. That there is in general a tradeoff between accuracy and fairness has been long recognized (Kamiran et al., 2010). Žliobaitė (2015) has considered the subtleties of empirically determining trade-offs between fairness and accuracy, but did not provide a theoretical analysis as above. Some work has empirically studied the impact of varying τ on problems akin to Equation 18 (e.g. (Zafar et al., 2016)); however, we are not aware of a result analogous to Proposition 8 that analytically quantifies the performance-fairness tradeoff.

7. Conclusion and future work

We studied the tradeoffs inherent (i.e. not specific to any algorithm) in the problem of learning with a fairness constraint, showing that for cost-sensitive *approximate fairness* measures, the optimal classifier is an *instance-dependent thresholding* of the class-probability function, and quantifying the degradation in performance by a measure of *alignment* of the target and sensitive variable. We used our analysis to derive a simple plugin approach for the fairness problem.

7.1. An intuitive summary of our results

In order to grasp our main results at a more intuitive level, let us consider their implications for a concrete task of designing a classifier to determine whether an applicant should be given a loan (the target feature), while not discriminating based on gender (the sensitive feature).

C1 says that if we measure fairness using either the DI or MD score, then this is equivalent to measuring a particular cost-sensitive error. This simply means that we use our classifier to predict each applicant's gender based on the available features, and look at the error rates amongst both males and females; by summing a suitably weighted combination of these error rates, we get a quantity that is reflective of the underlying fairness measure.

C2 says that, given access to the entire population of males and females, the classifier which attains the best tradeoff between accuracy and fairness has a simple form: one computes for a given applicant the probability of them repaying the loan, and determines if this probability is larger than a threshold based on the probability of them being male or female. To understand this intuitively, assume that women are inherently more likely to repay a loan than men, and that the sensitive feature is available as an input. Then, for a given choice of accuracy-fairness tradeoff, the optimal classifier aims to make it easier for men and harder for women to be granted a loan, so as to make the proportions of men and women amongst the accepted and rejected pool more commensurate; and clearly, it does so at some expense in accuracy of predicting the probability of repayment.

C3 says that if gender perfectly coincided with loan repayment (e.g. women were guaranteed to repay loans, and men guaranteed to default), then we can either have maximum accuracy but no fairness, or maximum fairness but random-level accuracy. At the other extreme, if gender were perfectly independent of loan repayment (e.g. both women and men were equally likely to repay loans), then we can have maximum accuracy and fairness simultaneously. Finally, if gender partially correlates with loan repayment, then the tradeoff between accuracy and fairness is determined by the strength of this correlation, and falls in between the previous two extremes.

The ability to theoretically compute the tradeoffs between fairness and utility is perhaps the most interesting aspect of our technical results. We stress that the tradeoff is *intrinsic to the underlying data* (in fact it is intrinisc to the underlying distributions that generated that data). That is, any fairness or unfairness, is a *property of the data, not of any particular technique*. This raises interesting philosophical issues, which are briefly touched upon in Appendix H. The tradeoffs we can theoretically compute precisely quantify what price one has to pay (in utility) in order to achieve a desired degree of fairness: in other words, we have computed the cost of fairness.

7.2. Future work

There are several possible directions for future work: we believe it valuable to study Bayes-optimal scorers for ranking measures such as AUC; establish consistency and finite-sample guarantees of the plugin

estimators of §5; and extend our analysis to the case of multi-category sensitive features.

Acknowledgements

The authors thank the Asian Office of Aerospace Research and Development (AOARD), the Australian Research Council, and DATA61 for supporting this research, and the FAT*18 referees and shepherd for helping improve the paper. This work was performed while AKM was with DATA61.

References

Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM*, 57(3):19:1–19:46, March 2010.

Lynwood Bryant. The role of thermodynamics in the evolution of heat engines. *Technology and Culture*, 14(2):152–165, 1973.

Toon Calders and Sicco Verwer. Three Naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2): 277–292, 2010.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *CoRR*, abs/1701.08230, 2017. URL http://arxiv.org/abs/1701.08230.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference (ITCS)*, pages 214–226, 2012.

EEOC. Uniform guidelines on employee selection procedures. https://www.eeoc.gov/policy/docs/qanda_clarify_procedures.html, 1979.

Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on Artificial Intelligence (IJCAI)*, pages 973–978, 2001.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006. ISSN 0167-8655.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 259–268, 2015.

Kazuto Fukuchi, Jun Sakuma, and Toshihiro Kamishima. Prediction with model-based neutrality. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 499–514, 2013.

James Gleick. *The Information: A History, a Theory, a Flood*. Fourth Estate, 2011.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, December 2016.

John C. Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *The Journal of Political Economy*, 63(4):309–321, 1955.

Faisal Kamiran and Toon Calders. Classification without discrimination. In *IEEE International Conference on Computer, Control and Communication (IEEE-IC4)*, 2009.

Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 869–874. IEEE, 2010.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 35–50, 2012.

Harikrishna Narasimhan, Rohit Vaish, and Shivani Agarwal. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1493–1501, 2014.

Harikrishna Narasimhan, Purushottam Kar, and Prateek Jain. Optimizing non-decomposable performance measures: A tale of two classes. In *International Conference on Machine Learning (ICML)*, pages 199–208, 2015.

Shameem Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. Optimizing F-measures by cost-sensitive classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2123–2131, 2014.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 560–568, 2008.

John Rawls. *A Theory of Justice*. Harvard University Press, 1971.

Mark D. Reid and Robert C. Williamson. Surrogate regret bounds for proper losses. In *International Conference on Machine Learning (ICML)*, pages 897–904, 2009.

Amartya K. Sen. *The Idea of Justice*. Harvard University Press, 2009.

Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory (COLT)*, pages 1920–1953, 2017.

Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International World Wide Web Conference (WWW)*, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna Gummadi. Learning fair classifiers. *arXiv preprint arXiv:1507.05259*, 2016.

Indrė Žliobaitė. On the relation between accuracy and fairness in binary classification. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2015.

Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(1060–1089), 2017.

Indrė Žliobaitė, Faisal Kamiran, and Toon Calders. Handling conditional discrimination. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 992–1001. IEEE, 2011.