

# Inherent Tradeoffs in Learning Fair Representations

**Han Zhao\***

*University of Illinois at Urbana-Champaign*

HANZHAO@ILLINOIS.EDU

**Geoffrey J. Gordon**

*Carnegie Mellon University*

GGORDON@CS.CMU.EDU

**Editor:** Maya Gupta

## Abstract

Real-world applications of machine learning tools in high-stakes domains are often regulated to be fair, in the sense that the predicted target should satisfy some quantitative notion of parity with respect to a protected attribute. However, the exact tradeoff between fairness and accuracy is not entirely clear, even for the basic paradigm of classification problems. In this paper, we characterize an inherent tradeoff between statistical parity and accuracy in the classification setting by providing a lower bound on the sum of group-wise errors of any fair classifiers. Our impossibility theorem could be interpreted as a certain uncertainty principle in fairness: if the base rates differ among groups, then any fair classifier satisfying statistical parity has to incur a large error on at least one of the groups. We further extend this result to give a lower bound on the joint error of any (approximately) fair classifiers, from the perspective of learning fair representations. To show that our lower bound is tight, assuming oracle access to Bayes (potentially unfair) classifiers, we also construct an algorithm that returns a randomized classifier which is both optimal (in terms of accuracy) and fair. Interestingly, when the protected attribute can take more than two values, an extension of this lower bound does not admit an analytic solution. Nevertheless, in this case, we show that the lower bound can be efficiently computed by solving a linear program, which we term as the TV-Barycenter problem, a barycenter problem under the TV-distance.

On the upside, we prove that if the group-wise Bayes optimal classifiers are close, then learning fair representations leads to an alternative notion of fairness, known as the accuracy parity, which states that the error rates are close between groups. Finally, we also conduct experiments on real-world datasets to confirm our theoretical findings.

**Keywords:** Algorithmic fairness, representation learning, information theory

## 1. Introduction

With the prevalence of machine learning applications in high-stakes domains, e.g., criminal judgment, medical testing, online advertising, etc., it is crucial to ensure that the automated decision making systems do not propagate existing bias or discrimination that might exist in historical data (Barocas and Selbst, 2016; Berk et al., 2018). Among many recent proposals for achieving different notions of algorithmic fairness (Zemel et al., 2013; Dwork et al., 2012; Zafar et al., 2015; Hardt et al., 2016; Zafar et al., 2017), learning fair representations has received increasing attention due to recent advances in learning rich representations with deep neural networks (Edwards and

---

\*. This work is an extended version of an earlier paper with the same title appearing in NeurIPS 2019 by the same authors.

Storkey, 2015; Louizos et al., 2015; Madras et al., 2018; Zhang et al., 2018; Beutel et al., 2017; Song et al., 2019; Zhao et al., 2019c; Chi et al., 2021). In fact, a line of work has proposed to learn group-invariant representations with adversarial learning techniques in order to achieve statistical parity, also known as the demographic parity in the literature. This line of work dates at least back to Zemel et al. (2013) where the authors proposed to learn predictive models that are independent of the group membership attribute. At a high level, the underlying idea is that if representations of instances from different groups are similar to each other, then any predictive model on top of them will certainly make decisions independent of group membership.

On the other hand, it has long been observed that there is an underlying tradeoff between accuracy and statistical parity. In particular, it is easy to see that in an extreme case where the group membership coincides with the target variable to predict, a call for exact statistical parity will inevitably remove the perfect predictor (Hardt et al., 2016). Empirically, it has also been observed that a tradeoff exists between accuracy and fairness in binary classification (Zliobaite, 2015). Clearly, methods based on learning fair representations are also bound by such inherent tradeoff between accuracy and fairness. But before attempting to develop an algorithm to achieve a particular goal on fairness, it is natural to ask:

- Q1** How does the fairness constraint trade for accuracy? Without further assumptions on the data generating distributions, what is the exact price any fair classifiers have to pay for fairness?
- Q2** Furthermore, given the underlying distribution, can we construct an algorithm to return the optimal (in terms of accuracy) fair classifier?
- Q3** Will learning fair representations help to achieve other notions of fairness besides the statistical parity? If yes, what is the fundamental limit of accuracy that we can hope to achieve under such constraint?

To answer the above questions, through the lens of information theory, in this paper we provide the first result that quantitatively characterizes the tradeoff between demographic parity and the sum of group-wise accuracy across different population groups. Specifically, when the base rates differ between groups, we provide a tight information-theoretic lower bound on the joint error across these groups. Our lower bound is algorithm-independent so it holds for all methods that satisfy statistical parity. We also extend this result to prove a lower bound on the joint accuracy for any fair classifiers, and generalize it to the case where the protected attribute can take any finite number of values. Interestingly, when the number of groups defined by the protected attribute is more than two, we can no longer obtain an analytic lower bound. Nevertheless, we show that the lower bound can be efficiently computed by solving a linear program, which we term as the TV-Barycenter problem, a barycenter problem under the TV-distance. To show that our lower bound is tight, assuming oracle access to Bayes (potentially unfair) classifiers, we derive an algorithm that returns a randomized classifier which is both optimal (in terms of accuracy) and fair.

When only approximate statistical parity is achieved, we present a family of lower bounds to quantify the tradeoff of accuracy introduced by such approximate constraint. As a side contribution, our proof technique is simple but general, and we expect it to have broader applications in other learning problems using adversarial techniques, e.g., unsupervised domain adaptation (Ganin et al., 2016; Zhao et al., 2019b), privacy-preservation under attribute inference attacks (Hamm, 2017; Zhao et al., 2019a) and multilingual machine translation (Johnson et al., 2017; Zhao et al., 2020).

To complement our negative results, we show that if the (potentially unfair) Bayes optimal classifiers across different groups are close, then learning fair representations helps to achieve an alternative notion of fairness, i.e., the accuracy parity (Buolamwini and Gebru, 2018), which states that the error rates are close between different groups. Empirically, we conduct experiments on a real-world dataset to corroborate both our positive and negative results. We believe our theoretical insights contribute to better understanding of the tradeoff between accuracy and different notions of fairness, and they are also helpful in guiding the future design of representation learning algorithms to achieve algorithmic fairness.

## 2. Preliminaries

We first introduce the notation used throughout the paper and formally describe the problem setup. We then briefly discuss some information-theoretic concepts that will be used in our analysis.

### 2.1 Notation

We consider a general classification setting where there is a joint distribution  $\mu$  over the triplet  $T = (X, A, Y)$ , where  $X \in \mathcal{X} \subseteq \mathbb{R}^d$  is the input vector,  $A \in \{0, 1\}^1$  is the protected attribute, e.g., race, gender, etc., and  $Y \in \mathcal{Y} = \{0, 1\}$  is the target output. Lower case letters  $x$ ,  $a$  and  $y$  are used to denote the instantiation of  $X$ ,  $A$  and  $Y$ , respectively. Let  $\mathcal{H}$  be a hypothesis class of predictors from input to output space. Throughout the paper, we focus on the setting where the classifier *cannot* directly use the sensitive attribute  $A$  to form its prediction. However, note that even if the classifier does not explicitly take the protected attribute  $A$  as input, this *fairness through blindness* mechanism can still be biased due to the redundant encoding issue (Barocas et al., 2017). To keep the notation uncluttered, for  $a \in \{0, 1\}$ , we use  $\mu_a$  to mean the conditional distribution of  $\mu$  given  $A = a$ . We use  $\mu(Y)$  to denote the marginal distribution of  $Y$  from a joint distribution  $\mu$  over  $Y$  and some other random variables. With slight abuse of notation, occasionally we also use  $Y_{\#}\mu$  to denote the marginal distribution of  $Y$  from the joint distribution  $\mu$ , i.e., projection of  $\mu$  onto the  $Y$  coordinate.

For an event  $E$ ,  $\mu(E)$  denotes the probability of  $E$  under  $\mu$ . In particular, in the literature of fair machine learning, we call  $\mu(Y = 1)$  the *base rate* of distribution  $\mu$  and we use  $\Delta_{\text{BR}}(\mu, \mu') := |\mu(Y = 1) - \mu'(Y = 1)|$  to denote the difference of the base rates between two distributions  $\mu$  and  $\mu'$  over the same sample space.

Given a feature transformation function  $g : \mathcal{X} \rightarrow \mathcal{Z}$  that maps instances from the input space  $\mathcal{X}$  to feature space  $\mathcal{Z}$ , we define  $g_{\#}\mu := \mu \circ g^{-1}$  to be the induced (pushforward) distribution of  $\mu$  under  $g$ , i.e., for any event  $E' \subseteq \mathcal{Z}$ ,  $g_{\#}\mu(E') := \mu(g^{-1}(E')) = \mu(\{x \in \mathcal{X} \mid g(x) \in E'\})$ . The zero-one entropy of  $A$  (Grünwald et al., 2004, Section 3.5.3) is denoted as  $H_{0-1}(A) := 1 - \max_{a \in \{0,1\}} \Pr(A = a)$ . Furthermore, we use  $F_{\mu}$  to represent the cumulative distribution function of  $\mu$ , i.e., for  $z \in \mathbb{R}$ ,  $F_{\mu}(z) := \Pr_{\mu}((-\infty, z])$ .

### 2.2 Group Fairness

Given a joint distribution  $\mu$ , the error of a predictor  $h$  under  $\mu$  is defined as  $\text{Err}_{\mu}(h) := \Pr_{\mu}(Y \neq h(X))$ . To make the notation more compact, we may drop the subscript  $\mu$  when it is clear from the context. In this work we focus on group fairness where the group membership is given by the

---

1. Our main results could be extended to the case where  $A$  can take finitely many values. We show this extension in Section 3.2

sensitive attribute  $A$ . Even in this context there are many possible definitions of *fairness* (Narayanan, 2018), and in what follows we provide a brief review of the ones that are mostly relevant to this work.

**Definition 1 (Demographic Parity)** *Given a joint distribution  $\mu$ , a classifier  $\hat{Y}$  satisfies demographic parity if  $\hat{Y}$  is independent of  $A$ .*

Demographic parity reduces to the requirement that  $\mu_0(\hat{Y} = 1) = \mu_1(\hat{Y} = 1)$ , i.e., positive outcome is given to the two groups at the same rate. When exact equality does not hold, we use the absolute difference between them as an approximate measure:

**Definition 2 (DP Gap)** *Given a joint distribution  $\mu$ , the demographic parity gap of a classifier  $\hat{Y}$  is  $\Delta_{\text{DP}}(\hat{Y}) := |\mu_0(\hat{Y} = 1) - \mu_1(\hat{Y} = 1)|$ .*

Demographic parity is also known as *statistical parity*, and it has been adopted as definition of fairness in a series of seminal works (Calders et al., 2009; Edwards and Storkey, 2015; Johndrow et al., 2019; Kamiran and Calders, 2009; Kamishima et al., 2011; Louizos et al., 2015; Zemel et al., 2013; Madras et al., 2018). However, as we shall quantify precisely in Section 3, demographic parity may reduce the accuracy that we hope to achieve, especially in the common scenario where the *base rates* differ between two groups, e.g.,  $\mu_0(Y = 1) \neq \mu_1(Y = 1)$ . In light of this, an alternative definition is *accuracy parity*:

**Definition 3 (Accuracy Parity)** *Given a joint distribution  $\mu$ , a classifier  $h$  satisfies accuracy parity if  $\text{Err}_{\mu_0}(h) = \text{Err}_{\mu_1}(h)$ .*

In the literature, a break of accuracy parity is also known as disparate mistreatment (Zafar et al., 2017). Again, when  $h$  is a binary classifier, accuracy parity reduces to  $\mu_0(h(X) = Y) = \mu_1(h(X) = Y)$ . Different from demographic parity, the definition of accuracy parity does not eliminate the perfect predictor when  $Y = A$  when the base rates differ between two groups. When costs of different error types matter, more refined definitions exist:

**Definition 4 (Equalized Odds (Hardt et al., 2016))** *Given a joint distribution  $\mu$ , a classifier  $h$  satisfies equalized odds if  $\mu_0(h(X) = 1 \mid Y = y) = \mu_1(h(X) = 1 \mid Y = y)$ ,  $\forall y \in \{0, 1\}$ .*

Equalized odds essentially requires equal true positive and false positive rates between different groups. Furthermore, Hardt et al. (2016) also defined *true positive parity*, or *equal opportunity*, to be  $\mu_0(h(X) = 1 \mid Y = 1) = \mu_1(h(X) = 1 \mid Y = 1)$  when the positive outcome is more desirable in certain applications. For example, in school admission, the cost of denying a competent candidate is considerably higher than the other way around. Last but not least, *predictive rate parity*, also known as *test fairness* (Chouldechova, 2017), asks for equal chance of positive outcomes across groups given predictions:

**Definition 5 (Predictive Rate Parity)** *Given a joint distribution  $\mu$ , a probabilistic classifier  $h$  satisfies predictive rate parity if  $\mu_0(Y = 1 \mid h(X) = c) = \mu_1(Y = 1 \mid h(X) = c)$ ,  $\forall c \in [0, 1]$ .*

A closely related notion of predictive rate parity is known as *statistical calibration*. Formally, a classifier  $h$  is said to be *calibrated* if  $\mu(Y = 1 \mid h(X) = c) = c$ ,  $\forall c \in [0, 1]$ , i.e., if we look at the set of data that receive a predicted probability of  $c$  by  $h$ , we would like  $c$ -fraction of them to be positive instances according to  $Y$  (Pleiss et al., 2017). Hence, it is clear to see that if a classifier  $h$  is calibrated across different subgroups, then it also satisfies predictive rate parity.

In the special case when  $h$  is a deterministic binary classifier that only takes value in  $\{0, 1\}$ , Chouldechova (2017) showed an intrinsic incompatibility between predictive rate parity and equalized odds:

**Theorem 6 (Chouldechova (2017))** *Assume  $\mu_0(Y = 1) \neq \mu_1(Y = 1)$ , then for any deterministic classifier  $h : \mathcal{X} \rightarrow \{0, 1\}$  that is not perfect, i.e.,  $h(X) \neq Y$ , equalized odds and predictive rate parity cannot hold simultaneously.*

Similar incompatibility result for probabilistic classifier has also been proved by Kleinberg et al. (2016), where the authors showed that for any non-perfect predictors, statistical calibration and equalized odds cannot be achieved simultaneously if the base rates are different across groups.

### 2.3 $f$ -divergence

Introduced by Ali and Silvey (1966) and Csiszár (1964, 1967),  $f$ -divergence, also known as the Ali-Silvey distance, is a general class of statistical divergences to measure the difference between two probability distributions  $\mathcal{P}$  and  $\mathcal{Q}$  over the same probability space.

**Definition 7 ( $f$ -divergence)** *Let  $\mathcal{P}$  and  $\mathcal{Q}$  be two probability distributions over the same space and assume  $\mathcal{P}$  is absolutely continuous w.r.t.  $\mathcal{Q}$  ( $\mathcal{P} \ll \mathcal{Q}$ ). Then for any convex function  $f : (0, \infty) \rightarrow \mathbb{R}$  that is strictly convex at 1 and  $f(1) = 0$ , the  $f$ -divergence of  $\mathcal{Q}$  from  $\mathcal{P}$  is defined as*

$$D_f(\mathcal{P} \parallel \mathcal{Q}) := \mathbb{E}_{\mathcal{Q}} \left[ f \left( \frac{d\mathcal{P}}{d\mathcal{Q}} \right) \right]. \quad (1)$$

*The function  $f$  is called the generator function of  $D_f(\cdot \parallel \cdot)$ .*

Different choices of the generator function  $f$  recover popular statistical divergence as special cases, e.g., the KL-divergence. From Jensen's inequality it is easy to verify that  $D_f(\mathcal{P} \parallel \mathcal{Q}) \geq 0$  and  $D_f(\mathcal{P} \parallel \mathcal{Q}) = 0$  iff  $\mathcal{P} = \mathcal{Q}$  almost surely. Note that  $f$ -divergence does not necessarily leads to a distance metric, and it is not symmetric in general, i.e.,  $D_f(\mathcal{P} \parallel \mathcal{Q}) \neq D_f(\mathcal{Q} \parallel \mathcal{P})$  provided that  $\mathcal{P} \ll \mathcal{Q}$  and  $\mathcal{Q} \ll \mathcal{P}$ . We list some common choices of the generator function  $f$  and their corresponding properties in Table 1. Notably, Khosravifard et al. (2007) proved that among all the  $f$ -divergences, total variation is the only  $f$ -divergence that serves as a metric, i.e., satisfying the triangle inequality.

## 3. Tradeoff between Fairness and Accuracy

In this section we take a slight detour from learning fair representations to first provide general results on the tradeoff between fairness and accuracy that applies to any fair classifiers. As we briefly discussed in Section 2.2, it is impossible to have imperfect predictor that is both statistically calibrated and verifies equalized odds when the base rates differ between two groups. On the other hand, while it has long been observed that demographic parity may eliminate perfect predictor (Hardt et al., 2016), and previous work has empirically verified that tradeoff exists between accuracy and demographic parity (Calders et al., 2009; Kamiran and Calders, 2009; Zliobaite, 2015) on various datasets, so far a quantitative characterization on the exact tradeoff between accuracy and various notions of group fairness is still missing in the classification setting. In this section, we seek to answer the following intriguing and important question:

Table 1: List of different  $f$ -divergences and their corresponding properties.  $D_{\text{KL}}(\mathcal{P} \parallel \mathcal{Q})$  denotes the KL-divergence of  $\mathcal{Q}$  from  $\mathcal{P}$  and  $\mathcal{M} := (\mathcal{P} + \mathcal{Q})/2$  is the average distribution of  $\mathcal{P}$  and  $\mathcal{Q}$ . Symm. stands for symmetric and Tri. stands for triangle inequality.

Name	$D_f(\mathcal{P} \parallel \mathcal{Q})$	Generator $f(t)$	Symm.	Tri.
Kullback-Leibler	$D_{\text{KL}}(\mathcal{P} \parallel \mathcal{Q})$	$t \log t$	✗	✗
Reverse-KL	$D_{\text{KL}}(\mathcal{Q} \parallel \mathcal{P})$	$-\log t$	✗	✗
Jensen-Shannon	$D_{\text{JS}}(\mathcal{P}, \mathcal{Q}) := \frac{1}{2}(D_{\text{KL}}(\mathcal{P} \parallel \mathcal{M}) + D_{\text{KL}}(\mathcal{Q} \parallel \mathcal{M}))$	$t \log t - (t+1) \log(\frac{t+1}{2})$	✓	✗
Squared Hellinger	$H^2(\mathcal{P}, \mathcal{Q}) := \frac{1}{2} \int (\sqrt{d\mathcal{P}} - \sqrt{d\mathcal{Q}})^2$	$(1 - \sqrt{t})^2/2$	✓	✗
Total Variation	$d_{\text{TV}}(\mathcal{P}, \mathcal{Q}) := \sup_E  \mathcal{P}(E) - \mathcal{Q}(E) $	$ t - 1 /2$	✓	✓

*In the setting of classification, what is the minimum error that any fair algorithm has to incur, and how does this error depend on the coupling between the target and the protected attribute?*

In what follows we shall first provide a simple example to illustrate this general tradeoff. This example will give readers a flavor the kind of impossibility result we are interested in obtaining. We then proceed to formally present a family of information-theoretic lower bounds on the accuracy that hold for *all* algorithms, even if only approximate statistical parity is satisfied. We conclude this section by some discussions on the implications of our results.

**A Simple Example** As a warm-up, let us consider an example to showcase the potential tradeoff between statistical parity and accuracy in binary classification. But before our construction, it should be noted that the error  $\text{Err}_\mu(\hat{Y})$  bears an intrinsic lower bound for any classifier  $\hat{Y} = h(X)$ , i.e., the noise in the underlying data distribution  $\mu$ , e.g., the Bayes error rate. Hence to simplify our discussion, in this example we shall construct distributions such that the Bayes error rates are 0, i.e., for  $a \in \{0, 1\}$ , there exists a ground-truth labeling function  $h_a^*$  such that  $Y = h_a^*(X)$  on  $\mu_a$ . Realize that such simplification will only make it harder for us to prove lower bound on  $\text{Err}_{\mu_a}$  since there exists classifiers that are perfect.

**Example 1 (A bijection between the target and the protected attribute)** For  $a \in \{0, 1\}$ , let the marginal distribution  $X_{\#}\mu_a$  be a uniform distribution over  $\{0, 1\}$ . Let  $Y = a$  be a constant. Hence by construction, on the joint distribution, we have  $Y = A$  hold. Now for any fair predictor  $\hat{Y} = h(X)$ , the statistical parity asks  $\hat{Y}$  to be independent of  $A$ . However, no matter what value  $h(x)$  takes, we always have  $|h(x)| + |h(x) - 1| \geq 1$ . Hence for any predictor  $h : \mathcal{X} \rightarrow \{0, 1\}$ :

$$\begin{aligned} \text{Err}_{\mu_0}(h) + \text{Err}_{\mu_1}(h) &= \frac{1}{2}|h(0) - 0| + \frac{1}{2}|h(1) - 0| + \frac{1}{2}|h(0) - 1| + \frac{1}{2}|h(1) - 1| \\ &\geq \frac{1}{2} + \frac{1}{2} \\ &= 1. \end{aligned}$$

This shows that for any fair predictor  $h$ , the sum of the errors of  $h$  on both groups has to be at least 1. On the other hand, there exists a trivial unfair algorithm that makes no error on both groups by also taking the protected attribute into consideration:  $\forall x \in \{0, 1\}, h^*(x) = 0$  if  $A = 0$  else  $h^*(x) = 1$ .

### 3.1 An Accuracy Lower Bound for Fair Classifiers

In this subsection we generalize the above simple example to general cases without making explicit assumptions on the underlying data generating distributions. Essentially, every prediction function induces the following Markov chain:

$$X \xrightarrow{g} Z \xrightarrow{h} \hat{Y},$$

where  $g$  is the feature transformation,  $h$  is the classifier on feature space,  $Z$  is the feature and  $\hat{Y}$  is the predicted target variable by  $h \circ g$ . Note that simple models, e.g., linear classifiers, are also included by specifying  $g$  to be the identity map. With this notation, we first state the following theorem that quantifies an inherent tradeoff between fairness and accuracy.

**Theorem 8** *Let  $\hat{Y} = h(g(X))$  be a predictor. If  $\hat{Y}$  satisfies demographic parity, then  $\text{Err}_{\mu_0}(h \circ g) + \text{Err}_{\mu_1}(h \circ g) \geq \Delta_{\text{BR}}(\mu_0, \mu_1)$ .*

**Remark** It is worth pointing out that Theorem 8 holds for any representation function  $g$  and classifier  $h$ , as long as the final predictor  $\hat{Y}$  satisfies demographic parity. In particular, by restricting  $g$  to be the identity function, we see that the lower bound also holds for any classifier that directly acts on the original input data. We choose the current presentation using a composition function  $h \circ g$  only for consistency with the rest of the paper.

Next,  $\Delta_{\text{BR}}(\mu_0, \mu_1)$  is the difference of base rates across groups, and it achieves its maximum value of 1 iff there exists a bijection between  $Y$  and  $A$ , e.g.,  $Y = A$ . On the other hand, if  $Y$  is independent of  $A$ , then  $\Delta_{\text{BR}}(\mu_0, \mu_1) = 0$  so the lower bound does not make any constraint on the joint error. Hence, the lower bound could also be understood as an uncertainty principle in fairness, stating in general that when the difference of base rates is large, then any fair algorithm has to incur a large error on at least one of the subgroups. Lastly, although Theorem 8 asks for exact demographic parity, in Section 5 we shall extend the above theorem when only approximate demographic parity is met, via learning fair representations.

Note that from Example 1, we can see this lower bound is tight, in the sense that there exist problem instances where the equality is verified. Second, Theorem 8 applies to all possible feature transformation  $g$  and predictor  $h$ . In particular, if we choose  $g$  to be the identity map, then Theorem 8 says that when the base rates differ, *no algorithm* can achieve a small joint error on both groups, and it also recovers the previous observation that demographic parity can eliminate the perfect predictor (Hardt et al., 2016). Third, the lower bound in Theorem 8 is insensitive to the marginal distribution of  $A$ , i.e., it treats the errors from both groups equally. As a comparison, let  $\alpha := \mu(A = 1)$ , then  $\text{Err}_{\mu}(h \circ g) = (1 - \alpha)\text{Err}_{\mu_0}(h \circ g) + \alpha\text{Err}_{\mu_1}(h \circ g)$ . In this case  $\text{Err}_{\mu}(h \circ g)$  could still be small even if the minority group suffers a large error. More formally, for the joint error  $\text{Err}_{\mu}(h \circ g)$ , we have the following corollary hold:

**Corollary 9** *Let  $\hat{Y} = h(g(X))$  be a predictor. If  $\hat{Y}$  satisfies demographic parity, then the joint error has the following lower bound:  $\text{Err}_{\mu}(\hat{Y}) \geq H_{0-1}(A) \cdot \Delta_{\text{BR}}(\mu_0, \mu_1)$ .*

Compared with the lower bound in Theorem 8, the lower bound of the joint error in Corollary 9 additionally depends on the zero-one entropy of  $A$ . In particular, if the marginal distribution of  $A$  is skewed, then  $H_{0-1}(A)$  will be small, which means that fairness will not reduce the joint accuracy too much. This corollary further implies that when the demographic subgroups are imbalanced in the overall population, the joint accuracy is not an ideal metric to look at, since it may hide the

potentially large drop in accuracy of the minority group. In particular, by the pigeonhole principle, the following corollary holds:

**Corollary 10** *If the predictor  $\hat{Y} = h(g(X))$  satisfies demographic parity, then  $\max\{\text{Err}_{\mu_0}(h \circ g), \text{Err}_{\mu_1}(h \circ g)\} \geq \Delta_{\text{BR}}(\mu_0, \mu_1)/2$ .*

In words, this means that for fair predictors in the demographic parity sense, at least one of the subgroups has to incur an error of at least  $\Delta_{\text{BR}}(\mu_0, \mu_1)/2$ , which further emphasizes the fundamental role of the difference of base rates,  $\Delta_{\text{BR}}(\mu_0, \mu_1)$ , in the tradeoff between fairness and accuracy.

**Proofs of Theorem 8, Corollary 9** Before we present the proof, we first present a useful lemma that lower bounds the prediction error by the total variation distance.

**Lemma 11** *Let  $\hat{Y} = h(X)$  be a predictor, then for  $a \in \{0, 1\}$ ,  $d_{\text{TV}}(\mu_a(Y), \mu_a(\hat{Y})) \leq \text{Err}_{\mu_a}(h)$ .*

**Proof** For  $a \in \{0, 1\}$ , because both  $\mu_a(Y)$  and  $\mu_a(\hat{Y})$  are Bernoulli distributions, we have:

$$\begin{aligned} d_{\text{TV}}(\mu_a(Y), \mu_a(\hat{Y})) &= |\mu_a(Y = 1) - \mu_a(h(X) = 1)| \\ &= |\mathbb{E}_{\mu_a}[Y] - \mathbb{E}_{\mu_a}[h(X)]| \\ &\leq \mathbb{E}_{\mu_a}[|Y - h(X)|] \\ &= \text{Err}_{\mu_a}(h), \end{aligned}$$

where the last equality holds because  $\Pr_{\mu_a}(Y \neq \hat{Y}) = \mathbb{E}_{\mu_a}[|Y - h(X)|]$  when  $Y, \hat{Y} \in \{0, 1\}$ .  $\blacksquare$

Now we are ready to prove Theorem 8:

**Proof** [Proof of Theorem 8] First of all, we show that if  $\hat{Y} = h(g(X))$  satisfies demographic parity, then:

$$\begin{aligned} d_{\text{TV}}(\mu_0(\hat{Y}), \mu_1(\hat{Y})) &= \max\{|\mu_0(\hat{Y} = 0) - \mu_1(\hat{Y} = 0)|, |\mu_0(\hat{Y} = 1) - \mu_1(\hat{Y} = 1)|\} \\ &= |\mu_0(\hat{Y} = 1) - \mu_1(\hat{Y} = 1)| \\ &= |\mu(\hat{Y} = 1 | A = 0) - \mu(\hat{Y} = 1 | A = 1)| = 0, \end{aligned}$$

where the last equality follows from the definition of demographic parity. Now from Table 1,  $d_{\text{TV}}(\cdot, \cdot)$  is symmetric and satisfies the triangle inequality, we have:

$$\begin{aligned} d_{\text{TV}}(\mu_0(Y), \mu_1(Y)) &\leq d_{\text{TV}}(\mu_0(Y), \mu_0(\hat{Y})) + d_{\text{TV}}(\mu_0(\hat{Y}), \mu_1(\hat{Y})) + d_{\text{TV}}(\mu_1(\hat{Y}), \mu_1(Y)) \\ &= d_{\text{TV}}(\mu_0(Y), \mu_0(\hat{Y})) + d_{\text{TV}}(\mu_1(\hat{Y}), \mu_1(Y)). \end{aligned} \quad (2)$$

The last step is to bound  $d_{\text{TV}}(\mu_a(Y), \mu_a(\hat{Y}))$  in terms of  $\text{Err}_{\mu_a}(h \circ g)$  for  $a \in \{0, 1\}$  using Lemma 11:

$$d_{\text{TV}}(\mu_0(Y), \mu_0(\hat{Y})) \leq \text{Err}_{\mu_0}(h \circ g), \quad d_{\text{TV}}(\mu_1(Y), \mu_1(\hat{Y})) \leq \text{Err}_{\mu_1}(h \circ g).$$

Combining the above two inequalities and (2) completes the proof.  $\blacksquare$

We now provide the proof of Corollary 9 on the lower bound of the joint error.

**Proof** [Proof of Corollary 9] To simplify the notation used in the proof, define  $\varepsilon := \text{Err}_\mu(\widehat{Y})$ ,  $\varepsilon_0 := \text{Err}_{\mu_0}(\widehat{Y})$  and  $\varepsilon_1 := \text{Err}_{\mu_1}(\widehat{Y})$ . Let  $\alpha := \Pr_\mu(A = 0)$ . By Theorem 8, we know that  $\varepsilon_0 + \varepsilon_1 \geq \Delta_{\text{BR}}(\mu_0, \mu_1)$ . By definition of the joint error:

$$\begin{aligned} \varepsilon &= \alpha\varepsilon_0 + (1 - \alpha)\varepsilon_1 \\ &\geq \alpha\varepsilon_0 + (1 - \alpha)(\Delta_{\text{BR}}(\mu_0, \mu_1) - \varepsilon_0) \\ &= (1 - \alpha)\Delta_{\text{BR}}(\mu_0, \mu_1) + (2\alpha - 1)\varepsilon_0. \end{aligned}$$

Similarly, we can also lower bound the joint error by:

$$\varepsilon \geq \alpha\Delta_{\text{BR}}(\mu_0, \mu_1) + (1 - 2\alpha)\varepsilon_1.$$

Now we discuss in two cases. If  $\alpha \leq 1/2$ , considering the second inequality yields:

$$\varepsilon \geq \alpha\Delta_{\text{BR}}(\mu_0, \mu_1) + (1 - 2\alpha)\varepsilon_1 \geq \alpha\Delta_{\text{BR}}(\mu_0, \mu_1).$$

If  $\alpha > 1/2$ , using the first inequality we have:

$$\varepsilon \geq (1 - \alpha)\Delta_{\text{BR}}(\mu_0, \mu_1) + (2\alpha - 1)\varepsilon_0 \geq (1 - \alpha)\Delta_{\text{BR}}(\mu_0, \mu_1).$$

Combining the above two cases leads to:

$$\varepsilon \geq \min\{\alpha, 1 - \alpha\} \cdot \Delta_{\text{BR}}(\mu_0, \mu_1) = H_{0-1}(A) \cdot \Delta_{\text{BR}}(\mu_0, \mu_1),$$

completing the proof. ■

It is not hard to show that our lower bound in Theorem 8 is tight. To see this, consider the case  $A = Y$ , where the lower bound achieves its maximum value of 1. Now consider a constant predictor  $\widehat{Y} \equiv 1$  or  $\widehat{Y} \equiv 0$ , which clearly satisfies demographic parity by definition. But in this case either  $\text{Err}_{\mu_0}(h \circ g) = 1, \text{Err}_{\mu_1}(h \circ g) = 0$  or  $\text{Err}_{\mu_0}(h \circ g) = 0, \text{Err}_{\mu_1}(h \circ g) = 1$ , hence  $\text{Err}_{\mu_0}(h \circ g) + \text{Err}_{\mu_1}(h \circ g) \equiv 1$ , achieving the lower bound.

To conclude this section, we point out that the choice of total variation in the lower bound is not unique. As we will see shortly in Section 5, similar lower bounds could be attained using specific choices of the general  $f$ -divergence with some desired properties.

### 3.2 Extension to Multiple Subgroups under Multi-class Classification

The analytical tradeoff lower bound in Theorem 8 mainly works for the setting of binary classification ( $|\mathcal{Y}| = 2$ ) with binary protected attribute ( $|\mathcal{A}| = 2$ ). Hence it is natural to ask whether such lower bounds also exist in the general setting where the target variable  $Y$  is a discrete random variable that takes  $m \geq 2$  values and the protected attribute  $A$  is also a categorical random variable that takes  $n \geq 2$  different values. In what follows we shall provide an extension of Theorem 8 to this general setting, although in this case we can no longer obtain an analytical characterization of the lower bound. However, as we shall see shortly, the exact lower bound could still be efficiently computed by solving a linear program.

We first introduce some additional notation that will be used in this section. For a positive integer  $K \in \mathbb{N}^*$ , we use  $[K]$  to denote the set  $\{1, \dots, K\}$ . We use  $\Delta_K$  to denote the  $K$ -dimensional

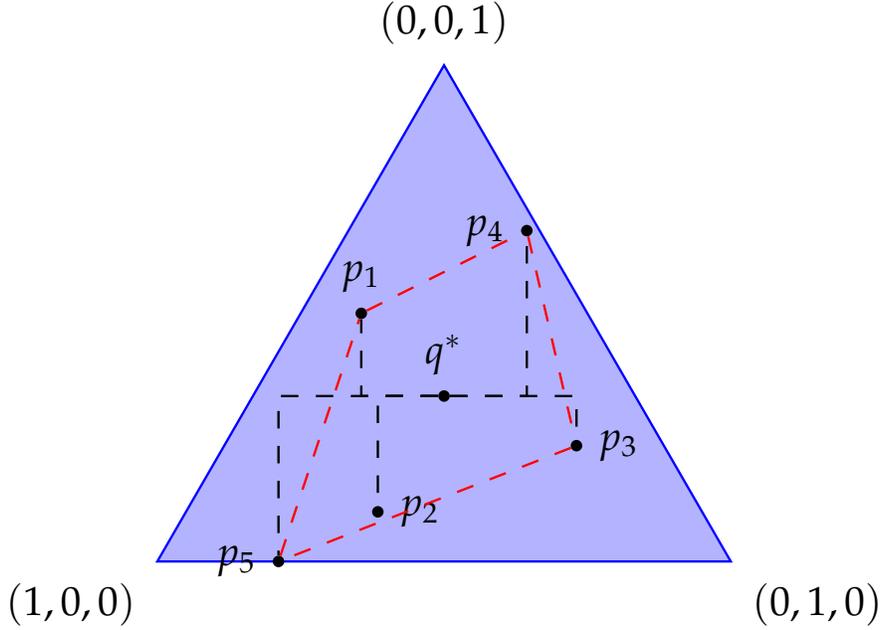


Figure 1: An example of the TV-Barycenter problem in (3) where  $n = 5$  and  $m = 3$ . The optimal solution vector  $q^*$  corresponds to the barycenter in the convex hull of  $\{p_1, \dots, p_5\}$  that minimizes the sum of  $\ell_1$  distances between  $p^*$  and  $p_i, i \in [5]$ .

probability simplex, i.e.,  $p \in \Delta^K$  if  $p \in \mathbb{R}_+^K$  and  $\sum_{i \in [K]} p_i = 1$ . For each subgroup  $A = i, i \in [n]$ , the corresponding marginal distribution of the target label  $Y$  could then be described by a  $m$ -dimensional vector  $p_i \in \Delta_m$ , i.e.,  $\mu_i(Y) = p_i$ . With these notation, we could then formally establish the following optimization problem:

$$\begin{aligned}
 \text{(TV-Barycenter):} \quad & \min_q \quad \frac{1}{2} \sum_{i=1}^n \|q - p_i\|_1 = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m |(p_i)_j - q_j| \\
 & \text{subject to} \quad q \in \Delta_m : q \geq 0, \sum_{j=1}^m q_j = 1
 \end{aligned} \tag{3}$$

We use  $\mathbf{OPT}(\{p_i\}_{i \in [n]})$  to denote the optimal value of the optimization problem in (3). As the name suggests, the above optimization problem computes the barycenter (under the  $\ell_1$  distance)  $q^*$  of the set of vectors  $\{p_i\}_{i \in [n]}$ , where each  $p_i \in \Delta_m$  corresponds to the marginal label distribution of  $Y$  within the group  $A = i$ . Clearly, the TV-Barycenter problem is a linear program, and hence its optimal solution could be efficiently computed in polynomial time.

We now state the extension of Theorem 8 using the optimal solution to the TV-Barycenter problem:

**Theorem 12** *Define  $p_i$  to be the probability mass vector of  $\mu_i(Y)$ :  $\forall i \in [n], j \in [m], (p_i)_j = \Pr_{\mu_i}(Y = j)$ . Let  $\hat{Y} = h(X)$  be a predictor. If  $\hat{Y}$  satisfies demographic parity, then  $\sum_{i=1}^n \text{Err}_{\mu_i}(h) \geq \mathbf{OPT}(\{p_i\}_{i \in [n]})$ .*

**Remark** To see that Theorem 12 is indeed a generalization of Theorem 8, note that when there are only two groups, i.e.,  $n = 2$ , we can readily read off the optimal solution  $q^*$  as:  $q^* = (p_1 + p_2)/2$  by realizing that the objective function is fully decomposable, and  $\mathbf{OPT}(\{p_i\}_{i \in [n]})$  is

$$\frac{1}{2} \left\| p_1 - \frac{p_1 + p_2}{2} \right\|_1 + \frac{1}{2} \left\| p_2 - \frac{p_1 + p_2}{2} \right\|_1 = \frac{1}{2} \|p_1 - p_2\|_1 = d_{\text{TV}}(\mu_1(Y), \mu_2(Y)).$$

Furthermore, when  $m = 2$ ,  $d_{\text{TV}}(\mu_1(Y), \mu_2(Y)) = \Delta_{\text{BR}}(\mu_1, \mu_2)$ . In general when  $n > 2$ , we cannot expect to have an analytic solution of  $\mathbf{OPT}(\{p_i\}_{i \in [n]})$ , but nevertheless it can be computed efficiently by solving a linear program. From this perspective, Theorem 12 builds an equivalent connection between the tradeoff problem in fairness and the barycenter problem in TV-distance.

**Proof** [Proof of Theorem 12] For  $\hat{Y} = h(X)$ , let  $C^{(i)} \in \mathbb{R}_+^{m \times m}$  be the confusion matrix between  $Y$  and  $\hat{Y}$  under  $\mu_i$ ,  $\forall i \in [n]$ . By definition of the confusion matrix, we have  $C_{jj'}^{(i)} = \Pr_{\mu_i}(Y = j, \hat{Y} = j')$ . Hence,

$$\Pr_{\mu_i}(Y \neq \hat{Y}) = 1 - \sum_{j=1}^m \Pr_{\mu_i}(Y = \hat{Y} = j) = 1 - \text{tr}(C^{(i)}).$$

On the other hand, consider  $d_{\text{TV}}(\mu_i(Y), \mu_i(\hat{Y}))$ , we have

$$\begin{aligned} d_{\text{TV}}(\mu_i(Y), \mu_i(\hat{Y})) &= \frac{1}{2} \sum_{j=1}^m \left| \Pr_{\mu_i}(Y = j) - \Pr_{\mu_i}(\hat{Y} = j) \right| \\ &= \frac{1}{2} \sum_{j=1}^m \left| \sum_{j' \in [m]} C_{jj'}^{(i)} - \sum_{j' \in [m]} C_{j'j}^{(i)} \right| = \frac{1}{2} \sum_{j=1}^m \left| \sum_{j' \neq j} (C_{jj'}^{(i)} - C_{j'j}^{(i)}) \right| \\ &\leq \frac{1}{2} \sum_{j=1}^m \sum_{j' \neq j} C_{jj'}^{(i)} + \frac{1}{2} \sum_{j=1}^m \sum_{j' \neq j} C_{j'j}^{(i)} = \sum_{j \neq j'} C_{jj'}^{(i)} \\ &= 1 - \text{tr}(C^{(i)}) = \Pr_{\mu_i}(Y \neq \hat{Y}). \end{aligned}$$

Next, since  $\hat{Y}$  satisfies demographic parity, so  $\hat{Y} \perp A$ , which means  $\nu := \mu_1(\hat{Y}) = \dots = \mu_n(\hat{Y})$ . Combining the above two arguments together yields

$$\begin{aligned} \inf_{\hat{Y}=h(X)} \sum_{i \in [n]} \text{Err}_{\mu_i}(\hat{Y}) &= \inf_{\hat{Y}=h(X)} \sum_{i \in [n]} \Pr_{\mu_i}(Y \neq \hat{Y}) \\ &\geq \inf_{\hat{Y}=h(X)} \sum_{i \in [n]} d_{\text{TV}}(\mu_i(Y), \mu_i(\hat{Y})) \\ &= \inf_{\hat{Y}=h(X)} \sum_{i \in [n]} d_{\text{TV}}(\mu_i(Y), \nu) \\ &\geq \inf_{\nu \in \Delta_m} \sum_{i \in [n]} d_{\text{TV}}(\mu_i(Y), \nu) \\ &= \min_{q \in \Delta_m} \frac{1}{2} \sum_{i=1}^n \|q - p_i\|_1 \\ &= \mathbf{OPT}(\{p_i\}_{i \in [n]}), \end{aligned}$$

completing the proof. ■

---

**Algorithm 1** Optimal fair classifier

---

**Input:** Oracle access to  $h_0^*$  and  $h_1^*$ , the Bayes optimal classifiers over  $\mu_0$  and  $\mu_1$

**Output:** A randomized optimal fair classifier  $h_{\text{Fair}}^* : \mathcal{X} \times \mathcal{A} \rightarrow \mathcal{Y}$

- 1: Compute  $\alpha := \Pr_{\mu_0}(Y = 1)$  and  $\beta := \Pr_{\mu_1}(Y = 1)$ . Without loss of generality assume  $\alpha \geq \beta$
- 2: For  $(x, a)$ , randomly sample  $s \sim U(0, 1)$ , the uniform distribution between  $(0, 1)$
- 3: Construct  $h_{\text{Fair}}^*(x, a)$  as

$$h_{\text{Fair}}^*(x, a) := \begin{cases} a = 0 : & \begin{cases} 0 & \text{If } h_0^*(x) = 0 \text{ or } h_0^*(x) = 1 \text{ and } s > \frac{\alpha + \beta}{2\alpha} \\ 1 & \text{If } h_0^*(x) = 1 \text{ and } s \leq \frac{\alpha + \beta}{2\alpha} \end{cases} \\ a = 1 : & \begin{cases} 0 & \text{If } h_1^*(x) = 0 \text{ and } s > \frac{\alpha - \beta}{2(1 - \beta)} \\ 1 & \text{If } h_1^*(x) = 1 \text{ or } h_1^*(x) = 0 \text{ and } s \leq \frac{\alpha - \beta}{2(1 - \beta)} \end{cases} \end{cases} \quad (4)$$

**return**  $h_{\text{Fair}}^*$

---

#### 4. An Optimal Fair Classifier

Theorem 8 provides an information-theoretic lower bound on the sum of group-wise errors for any fair classifiers. From the proof of Theorem 8, it is clear that the same lower bound also holds for fair classifiers that can have explicit access to the protected attribute  $A$ . To see this, consider a special case where the input  $X$  contains a redundant attribute that is a synonym of the protected attribute  $A$ . Since Theorem 8 holds for any distribution  $\mu$  over the triplet  $(X, A, Y)$ , this simple observation implies that our lower bound also holds for fair classifiers that take the protected attribute  $A$  as an input explicitly.

Although in the last section we briefly mention the tightness of Theorem 8 by constructing problem instances and fair classifiers where the equality verifies, it is still unclear whether it is possible to construct an algorithm such that:

1. For any distribution  $\mu$  over  $(X, A, Y)$ , the algorithm returns a (possibly randomized) fair classifier  $h_{\text{Fair}}^*$ .
2. The returned fair classifier  $h_{\text{Fair}}^*$  is optimal, in the sense that it verifies the lower bound in Theorem 8:  $\text{Err}_{\mu_0}(h_{\text{Fair}}^*) + \text{Err}_{\mu_1}(h_{\text{Fair}}^*) = \Delta_{\text{BR}}(\mu_0, \mu_1)$ .

It should be noted that in general it is relatively easy to construct a trivial classifier that is fair in the demographic parity sense. For example, any constant classifier that always outputs 0 or 1 is always fair for any distribution  $\mu$  over  $(X, A, Y)$ , but this classifier is not optimal in the sense of achieving the best possible accuracy. On the other hand, the problem of learning an optimal fair classifier is at least as hard as learning a Bayes optimal classifier, which we formally define as follows.

**Definition 13 (Bayes Optimal Classifier)** *Given random variables  $(X, A)$  and a target variable  $Y$ , the Bayes optimal classifier is  $h^* := \arg \min_{h(x, a)} \Pr(Y \neq h(x, a))$  with  $h^*(x, a) = 1$  iff  $\mathbb{E}[Y \mid X = x, A = a] \geq 1/2$  otherwise 0. We also use  $h_a^*(\cdot)$  to denote the restriction of  $h^*$  on  $A = a$ , respectively, i.e.,  $h_a^*(\cdot) := h^*(\cdot, a)$ .*

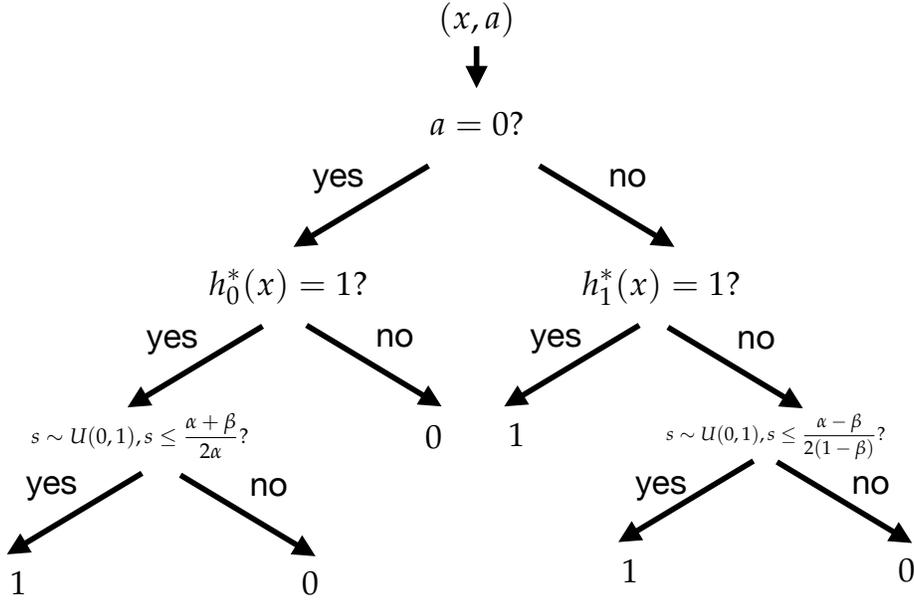


Figure 2: The decision tree diagram of the (randomized) optimal fair classifier  $h_{\text{Fair}}^*$ . The optimal fair classifier  $h_{\text{Fair}}^*$  has oracle access to  $h_0^*$  and  $h_1^*$ , the (possibly unfair) Bayes optimal classifier on  $\mu_0$  and  $\mu_1$ , respectively.  $\alpha, \beta$  are the base rates over  $\mu_0$  and  $\mu_1$ , i.e.  $\alpha := \Pr_{\mu_0}(Y = 1)$  and  $\beta := \Pr_{\mu_1}(Y = 1)$ .

Namely,  $h_a^*(\cdot)$  is the corresponding Bayes optimal classifier on the group  $A = a$  for  $a \in \{0, 1\}$ . With the definition of Bayes optimal classifier, we can formally argue that learning an optimal fair classifier is at least as hard as learning a Bayes optimal classifier by the following reduction.

**A Reduction** Given a distribution  $\mu'$  over  $\mathcal{X} \times \mathcal{Y}$ , let  $h'(\cdot)$  be the Bayes optimal classifier over  $\mu'$ . We can create a problem instance of learning the optimal fair classifier by constructing a distribution  $\mu$  over  $(X, A, Y)$  with  $\mu_{A=0} = \mu_{A=1} = \mu'$ . Now because  $\mu_{A=0} = \mu_{A=1} = \mu'$ , it is clear to see that the classifier  $h^*(\cdot, \cdot)$  where  $h^*(\cdot, 0) = h^*(\cdot, 1) = h'(\cdot)$  satisfies demographic parity. Furthermore, the optimality of  $h^*$  over  $\mu'$  implies the optimality of  $h^*$  over  $\mu$  as well. This shows that an oracle call to the problem of learning the optimal fair classifier over  $\mu$  can be used to solve the problem of learning a Bayes optimal classifier over  $\mu'$ , by a restriction of the returned  $h^*$  to either  $h_0^*$  or  $h_1^*$ .

Surprisingly, the other direction is also true. More specifically, in what follows we shall present an algorithm to construct a randomized classifier that is both fair and optimal, given oracle access to  $h_0^*$  and  $h_1^*$ , i.e., the group-wise Bayes optimal classifiers. We list the algorithm in Algorithm 1 and the corresponding decision diagram of  $h_{\text{Fair}}^*$  in Figure 2.

**Theorem 14** For any distribution  $\mu$  over  $(X, A, Y)$  such that  $Y_{A=0} = h_0^*(X)$  and  $Y_{A=1} = h_1^*(X)$ , the classifier  $h_{\text{Fair}}^*$  constructed by Algorithm 1 satisfies demographic parity and is optimal, i.e.,  $\text{Err}_{\mu_0}(h_{\text{Fair}}^*) + \text{Err}_{\mu_1}(h_{\text{Fair}}^*) = \Delta_{\text{BR}}(\mu_0, \mu_1)$ .

Before we present the proof of Theorem 14, we first briefly discuss one implication of its assumption that  $Y_{A=0} = h_0^*(X)$  and  $Y_{A=1} = h_1^*(X)$ . Essentially, this assumption says that there exists a perfect but potentially unfair classifier over  $\mu$ . Hence in this case  $\text{Err}_{\mu_0} + \text{Err}_{\mu_1}$  exactly corresponds to the price paid by enforcing demographic parity. In what follows we provide the proof for this theorem.

**Proof** [Proof of Theorem 14] We first show that  $h_{\text{Fair}}^*$  is fair in the demographic parity sense. Let  $\hat{Y} = h_{\text{Fair}}^*(X, A)$ ,  $\alpha := \Pr(Y = 1 \mid A = 0)$  and  $\beta := \Pr(Y = 1 \mid A = 1)$ . Without loss of generality, we assume  $\alpha \geq \beta$ .

For  $A = 0$ , consider the probability  $\Pr_{\mu_0}(\hat{Y} = 1)$ . Note that by construction in Algorithm 1,  $h_{\text{Fair}}^*(X, 0) = 0$  whenever  $h_0^*(X) = 0$ , so

$$\begin{aligned} \Pr_{\mu_0}(\hat{Y} = 1) &= \Pr_{\mu_0} \left( h_0^*(X) = 1, S \leq \frac{\alpha + \beta}{2\alpha} \right) = \Pr_{\mu_0} (h_0^*(X) = 1) \cdot \Pr \left( S \leq \frac{\alpha + \beta}{2\alpha} \right) \\ &= \Pr_{\mu_0}(Y = 1) \cdot \frac{\alpha + \beta}{2\alpha} = \alpha \cdot \frac{\alpha + \beta}{2\alpha} = \frac{\alpha + \beta}{2}. \end{aligned}$$

Similarly, for  $A = 1$ , recall that by construction,  $h_{\text{Fair}}^*(X, 1) = 1$  whenever  $h_1^*(X) = 1$ , so

$$\begin{aligned} \Pr_{\mu_1}(\hat{Y} = 1) &= \Pr_{\mu_1} \left( h_1^*(X) = 1 \vee \left( h_1^*(X) = 0 \wedge S \leq \frac{\alpha - \beta}{2(1 - \beta)} \right) \right) \\ &= \Pr_{\mu_1}(h_1^*(X) = 1) + \Pr(h_1^*(X) = 0) \cdot \Pr \left( S \leq \frac{\alpha - \beta}{2(1 - \beta)} \right) \\ &= \Pr_{\mu_1}(Y = 1) + \Pr_{\mu_1}(Y = 0) \cdot \frac{\alpha - \beta}{2(1 - \beta)} \\ &= \beta + (1 - \beta) \cdot \frac{\alpha - \beta}{2(1 - \beta)} = \frac{\alpha + \beta}{2}, \end{aligned}$$

where in the proof above we use the fact that  $S \sim U(0, 1)$  is drawn independently of  $X$ . This shows that  $\mu_0(\hat{Y}) = \mu_1(\hat{Y})$  so  $\hat{Y} = h_{\text{Fair}}^*(X, A)$  is fair.

Next, we prove that  $h_{\text{Fair}}^*$  is optimal. For  $A = 0$ ,

$$\text{Err}_{\mu_0}(\hat{Y}) = \Pr(h_{\text{Fair}}^*(X, 0) \neq Y \mid A = 0) = \Pr(h_{\text{Fair}}^*(X, 0) \neq h_0^*(X) \mid A = 0).$$

However, due to the construction of  $h_{\text{Fair}}^*$  in Algorithm 1,  $h_{\text{Fair}}^*(X, 0) \neq h_0^*(X)$  could only happen if  $h_0^*(X) = 1$  while  $h_{\text{Fair}}^*(X, 0) = 0$ , so

$$\begin{aligned} \text{Err}_{\mu_0}(\hat{Y}) &= \Pr(h_{\text{Fair}}^*(X, 0) \neq h_0^*(X) \mid A = 0) \\ &= \sum_x \Pr(h_{\text{Fair}}^*(x, 0) \neq h_0^*(x) \mid A = 0) \\ &= \sum_x \Pr(h_{\text{Fair}}^*(x, 0) = 0, h_0^*(x) = 1 \mid A = 0) \\ &= \sum_x \Pr(h_{\text{Fair}}^*(x, 0) = 0 \mid h_0^*(x) = 1, A = 0) \cdot \Pr(h_0^*(x) = 1 \mid A = 0) \\ &= \sum_x \Pr \left( S > \frac{\alpha + \beta}{2\alpha} \right) \cdot \Pr(h_0^*(x) = 1 \mid A = 0) \\ &= \left( 1 - \frac{\alpha + \beta}{2\alpha} \right) \cdot \sum_x \Pr(h_0^*(x) = 1 \mid A = 0) \\ &= \left( 1 - \frac{\alpha + \beta}{2\alpha} \right) \cdot \alpha \\ &= \frac{\alpha - \beta}{2}. \end{aligned}$$

Similarly, for  $A = 1$ ,

$$\text{Err}_{\mu_1}(\hat{Y}) = \Pr(h_{\text{Fair}}^*(X, 1) \neq Y \mid A = 1) = \Pr(h_{\text{Fair}}^*(X, 1) \neq h_1^*(X) \mid A = 1).$$

Again, due to the construction of  $h_{\text{Fair}}^*$  in Algorithm 1,  $h_{\text{Fair}}^*(X, 1) \neq h_1^*(X)$  could only happen if  $h_1^*(X) = 0$  while  $h_{\text{Fair}}^*(X, 0) = 1$ , so

$$\begin{aligned} \text{Err}_{\mu_0}(\hat{Y}) &= \Pr(h_{\text{Fair}}^*(X, 1) \neq h_0^*(X) \mid A = 1) \\ &= \sum_x \Pr(h_{\text{Fair}}^*(x, 1) \neq h_1^*(x) \mid A = 1) \\ &= \sum_x \Pr(h_{\text{Fair}}^*(x, 1) = 1, h_1^*(x) = 0 \mid A = 1) \\ &= \sum_x \Pr(h_{\text{Fair}}^*(x, 1) = 1 \mid h_1^*(x) = 0, A = 1) \cdot \Pr(h_1^*(x) = 0 \mid A = 1) \\ &= \sum_x \Pr\left(S \leq \frac{\alpha - \beta}{2(1 - \beta)}\right) \cdot \Pr(h_1^*(x) = 0 \mid A = 1) \\ &= \frac{\alpha - \beta}{2(1 - \beta)} \cdot \sum_x \Pr(h_1^*(x) = 0 \mid A = 1) \\ &= \frac{\alpha - \beta}{2(1 - \beta)} \cdot (1 - \beta) \\ &= \frac{\alpha - \beta}{2}. \end{aligned}$$

Combining the two cases above shows that  $\text{Err}_{\mu_0}(h_{\text{Fair}}^*) + \text{Err}_{\mu_1}(h_{\text{Fair}}^*) = \alpha - \beta = \Delta_{\text{BR}}(\mu_0, \mu_1)$ , which completes the proof.  $\blacksquare$

As a simple corollary of Theorem 14, we can now strengthen Theorem 8 as follows:

**Corollary 15** *Let  $\hat{Y} = h(X, A)$  be a predictor that satisfies demographic parity, then  $\inf_{\hat{Y}} \text{Err}_{\mu_0}(h) + \text{Err}_{\mu_1}(h) = \Delta_{\text{BR}}(\mu_0, \mu_1)$ .*

Note that one difference between Corollary 15 and Theorem 8 is that the fair predictor in Corollary 15 is allowed to have explicit access to the protected attribute  $A$  during decision making. This is necessary for our construction of the optimal fair classifier in Algorithm 1. It remains an open question whether it is possible to construct an optimal fair classifier to achieve the lower bound in Theorem 8 without having explicit access to the protected attribute, as in many practical applications of high stakes the automated decision making process is regulated to not directly use the protected attribute  $A$  (GDPR, Article 22 Paragraph 4).

## 5. Approximate Fairness via Learning Fair Representations

In the last section we show that there is an inherent tradeoff between fairness and accuracy when a predictor *exactly* satisfies demographic parity. In practice we may not be able to achieve demographic parity precisely. Instead, a line of recent algorithms (Edwards and Storkey, 2015; Louizos et al., 2015; Beutel et al., 2017; Adel et al., 2019; Zhao et al., 2019c) build an adversarial discriminator that takes as input the feature vector  $Z = g(X)$ , and the goal is to learn fair representations such that it is hard for the adversarial discriminator to infer the group membership from  $Z$ , typically by solving

a minimax objective between the feature encoder and the adversarial discriminator (Edwards and Storkey, 2015; Beutel et al., 2017; Zhang et al., 2018; Zhao et al., 2019c; Chi et al., 2021).

In these applications, due to the limit on the capacity of the adversarial discriminator, only approximate demographic parity can be achieved in practice. Hence it is natural to ask what is the tradeoff between fair representations and accuracy in this scenario? In this section we shall answer this question by generalizing our previous analysis with  $f$ -divergence to prove a family of lower bounds on the joint target prediction error. Our results also show how approximate DP helps to reconcile but not remove the tradeoff between fairness and accuracy. Before we state and prove the main results in this section, we first introduce the following lemma by Liese and Vajda (2006) as a generalization of the data processing inequality for  $f$ -divergence:

**Lemma 16 (Liese and Vajda (2006))** *Let  $\Delta(\mathcal{Z})$  be the space of all probability distributions over  $\mathcal{Z}$ . Then for any  $f$ -divergence  $D_f(\cdot \parallel \cdot)$ , any stochastic kernel  $\kappa : \mathcal{X} \rightarrow \Delta(\mathcal{Z})$ , and any distributions  $\mathcal{P}$  and  $\mathcal{Q}$  over  $\mathcal{X}$ ,  $D_f(\kappa\mathcal{P} \parallel \kappa\mathcal{Q}) \leq D_f(\mathcal{P} \parallel \mathcal{Q})$ .*

Roughly speaking, Lemma 16 says that data processing cannot increase discriminating information. Define  $d_{JS}(\mathcal{P}, \mathcal{Q}) := \sqrt{D_{JS}(\mathcal{P}, \mathcal{Q})}$  and  $H(\mathcal{P}, \mathcal{Q}) := \sqrt{H^2(\mathcal{P}, \mathcal{Q})}$ . It is well-known in information theory that both  $d_{JS}(\cdot, \cdot)$  and  $H(\cdot, \cdot)$  form a bounded distance metric over the space of probability distributions (Wu, 2017, Chapter 4). Realize that  $d_{TV}(\cdot, \cdot)$ ,  $H^2(\cdot, \cdot)$  and  $D_{JS}(\cdot, \cdot)$  are all  $f$ -divergence. The following corollary holds:

**Corollary 17** *Let  $h : \mathcal{Z} \rightarrow \mathcal{Y}$  be a classifier, and  $g_{\#}\mu_a$  be the pushforward distribution of  $\mu_a$  by  $g$ ,  $\forall a \in \{0, 1\}$ . Let  $\hat{Y} = h(g(X))$  be the predictor, then all the following inequalities hold:*

1.  $d_{TV}(\mu_0(\hat{Y}), \mu_1(\hat{Y})) \leq d_{TV}(g_{\#}\mu_0, g_{\#}\mu_1)$
2.  $H(\mu_0(\hat{Y}), \mu_1(\hat{Y})) \leq H(g_{\#}\mu_0, g_{\#}\mu_1)$
3.  $d_{JS}(\mu_0(\hat{Y}), \mu_1(\hat{Y})) \leq d_{JS}(g_{\#}\mu_0, g_{\#}\mu_1)$

Now we are ready to present the following main theorem of this section:

**Theorem 18** *Let  $\hat{Y} = h(g(X))$  be the predictor where  $h : \mathcal{Z} \rightarrow \{0, 1\}$  is any classifier acting on the feature space. Assume  $d_{JS}(g_{\#}\mu_0, g_{\#}\mu_1) \leq d_{JS}(\mu_0(Y), \mu_1(Y))$  and  $H(g_{\#}\mu_0, g_{\#}\mu_1) \leq H(\mu_0(Y), \mu_1(Y))$ , then the following three inequalities hold:*

1. *Total variation lower bound:*

$$\text{Err}_{\mu_0}(h \circ g) + \text{Err}_{\mu_1}(h \circ g) \geq d_{TV}(\mu_0(Y), \mu_1(Y)) - d_{TV}(g_{\#}\mu_0, g_{\#}\mu_1).$$

2. *Jensen-Shannon lower bound:*

$$\text{Err}_{\mu_0}(h \circ g) + \text{Err}_{\mu_1}(h \circ g) \geq (d_{JS}(\mu_0(Y), \mu_1(Y)) - d_{JS}(g_{\#}\mu_0, g_{\#}\mu_1))^2 / 2.$$

3. *Hellinger lower bound:*

$$\text{Err}_{\mu_0}(h \circ g) + \text{Err}_{\mu_1}(h \circ g) \geq (H(\mu_0(Y), \mu_1(Y)) - H(g_{\#}\mu_0, g_{\#}\mu_1))^2 / 2.$$

**Remark** All the three lower bounds in Theorem 18 imply a tradeoff between the joint error across demographic subgroups and learning group-invariant feature representations. When  $g_{\#}\mu_0 = g_{\#}\mu_1$ , due to the data-processing principle, we also have  $\mu_0(\hat{Y}) = \mu_1(\hat{Y})$ , and all three lower bounds get larger. In this case, we have

$$\begin{aligned} \max \left\{ d_{\text{TV}}(\mu_0(Y), \mu_1(Y)), \frac{1}{2}d_{\text{JS}}^2(\mu_0(Y), \mu_1(Y)), \frac{1}{2}H^2(\mu_0(Y), \mu_1(Y)) \right\} &= d_{\text{TV}}(\mu_0(Y), \mu_1(Y)) \\ &= \Delta_{\text{BR}}(\mu_0, \mu_1), \end{aligned}$$

and this reduces to Theorem 8. We now present the proof for Theorem 18.

**Proof** [Proof of Theorem 18] We prove the three inequalities respectively. The total variation lower bound follows the same idea as the proof of Theorem 8 and the inequality  $d_{\text{TV}}(\mu_0(\hat{Y}), \mu_1(\hat{Y})) \leq d_{\text{TV}}(g_{\#}\mu_0, g_{\#}\mu_1)$  from Corollary 17. To prove the Jensen-Shannon lower bound, realize that  $d_{\text{JS}}(\cdot, \cdot)$  is a distance metric over probability distributions. Combining with the inequality  $d_{\text{JS}}(\mu_0(\hat{Y}), \mu_1(\hat{Y})) \leq d_{\text{JS}}(g_{\#}\mu_0, g_{\#}\mu_1)$  from Corollary 17, we have:

$$d_{\text{JS}}(\mu_0(Y), \mu_1(Y)) \leq d_{\text{JS}}(\mu_0(Y), \mu_0(\hat{Y})) + d_{\text{JS}}(g_{\#}\mu_0, g_{\#}\mu_1) + d_{\text{JS}}(\mu_1(\hat{Y}), \mu_1(Y)).$$

Now by Lin's lemma (Lin, 1991, Theorem 3), for any two distributions  $\mathcal{P}$  and  $\mathcal{Q}$ , we have  $d_{\text{JS}}^2(\mathcal{P}, \mathcal{Q}) \leq d_{\text{TV}}(\mathcal{P}, \mathcal{Q})$ . Combine Lin's lemma with Lemma 11, we get the following lower bound:

$$\sqrt{\text{Err}_{\mu_0}(h \circ g)} + \sqrt{\text{Err}_{\mu_1}(h \circ g)} \geq d_{\text{JS}}(\mu_0(Y), \mu_1(Y)) - d_{\text{JS}}(g_{\#}\mu_0, g_{\#}\mu_1).$$

Apply the AM-GM inequality, we can further bound the L.H.S. by

$$\sqrt{2(\text{Err}_{\mu_0}(h \circ g) + \text{Err}_{\mu_1}(h \circ g))} \geq \sqrt{\text{Err}_{\mu_0}(h \circ g)} + \sqrt{\text{Err}_{\mu_1}(h \circ g)}.$$

Under the assumption that  $d_{\text{JS}}(g_{\#}\mu_0, g_{\#}\mu_1) \leq d_{\text{JS}}(\mu_0(Y), \mu_1(Y))$ , taking a square at both sides then completes the proof for the second inequality. The proof for Hellinger's lower bound follows exactly as the one for Jensen-Shannon's lower bound, except that instead of Lin's lemma, we need to use the fact that  $H^2(\mathcal{P}, \mathcal{Q}) \leq d_{\text{TV}}(\mathcal{P}, \mathcal{Q}) \leq \sqrt{2}H(\mathcal{P}, \mathcal{Q})$ ,  $\forall \mathcal{P}, \mathcal{Q}$ .  $\blacksquare$

As a simple corollary of Theorem 18, the following result shows how approximate DP (in terms of the DP gap) helps to reconcile the tradeoff between fairness and accuracy, by controlling the divergence between different groups of representations:

**Corollary 19** *Let  $g : \mathcal{X} \rightarrow \mathcal{Z}$  be a feature transformation. If  $d_{\text{TV}}(g_{\#}\mu_0, g_{\#}\mu_1) \leq \epsilon$ , then for any classifier  $h : \mathcal{Z} \rightarrow \mathcal{Y}$ , the DP gap  $\Delta_{\text{DP}}(h \circ g) \leq \epsilon$ , and  $\text{Err}_{\mu_0}(h \circ g) + \text{Err}_{\mu_1}(h \circ g) \geq \Delta_{\text{BR}}(\mu_0, \mu_1) - \epsilon$ .*

In a sense Corollary 19 means that in order to lower the joint error, the DP gap of the predictor cannot be too small. Of course, since the above inequality is a lower bound, it only serves as a necessary condition for small joint error. Hence an interesting question would be to ask whether it is possible to have a sufficient condition that guarantees a small joint error such that the DP gap of the predictor is no larger than that of the perfect predictor, i.e.,  $\Delta_{\text{BR}}(\mu_0, \mu_1)$ .

## 6. Fair Representations Lead to Accuracy Parity

In the previous sections we prove a family of information-theoretic lower bounds that demonstrate an inherent tradeoff between fair representations and joint error across groups. A natural question to ask then, is, what kind of parity can fair representations bring us? To complement our negative results, in this section we show that learning group-invariant representations help to reduce discrepancy of errors across groups.

First of all, since we work under the stochastic setting where  $\mu_a$  is a joint distribution over  $X$  and  $Y$  conditioned on  $A = a$ , then any function mapping  $h : \mathcal{X} \rightarrow \mathcal{Y}$  will inevitably incur an error due to the noise existed in the distribution  $\mu_a$ . In the case of binary classification, this error is also known as the Bayes error. Formally, for  $a \in \{0, 1\}$ , let  $h_a^* : \mathcal{X} \rightarrow \mathcal{Y}$  be the Bayes optimal classifier on  $\mu_a$ . Now define the noise of distribution  $\mu_a$  (the Bayes error on  $\mu_a$ ) to be  $n_{\mu_a} := \Pr_{\mu_a}(Y \neq h_a^*(X)) = \mathbb{E}_{\mu_a}[|Y - h_a^*(X)|]$ . We are now ready to present the following theorem:

**Theorem 20 (Error Decomposition Theorem)** *For any hypothesis  $\mathcal{H} \ni h : \mathcal{X} \rightarrow \mathcal{Y}$ , the following inequality holds:*

$$\begin{aligned} |\text{Err}_{\mu_0}(h) - \text{Err}_{\mu_1}(h)| &\leq (n_{\mu_0} + n_{\mu_1}) + d_{TV}(\mu_0(X), \mu_1(X)) \\ &\quad + \min \{ \mathbb{E}_{\mu_0}[|h_0^* - h_1^*|], \mathbb{E}_{\mu_1}[|h_0^* - h_1^*|] \}. \end{aligned}$$

**Remark** Theorem 20 upper bounds the discrepancy of accuracy across groups by three terms: the sum of group-wise noise, the distance of representations across groups and the discrepancy of the Bayes optimal classifiers. In an ideal setting where both distributions are noiseless, i.e., same individuals in the same group are always treated similarly, the upper bound simplifies to the latter two terms:

$$|\text{Err}_{\mu_0}(h) - \text{Err}_{\mu_1}(h)| \leq d_{TV}(\mu_0(X), \mu_1(X)) + \min \{ \mathbb{E}_{\mu_0}[|h_0^* - h_1^*|], \mathbb{E}_{\mu_1}[|h_0^* - h_1^*|] \}.$$

If we further require that the optimal decision functions  $h_0^*$  and  $h_1^*$  are close to each other, i.e., optimal decisions are insensitive to the group membership, then Theorem 20 implies that a sufficient condition to guarantee accuracy parity is to find group-invariant representation that minimizes  $d_{TV}(\mu_0(X), \mu_1(X))$ . Note, however, we should also pay attention to ensure that the first term as well as the third term in the upper bound do not increase drastically when learning the representations, as a change of the data representation will also change the noise term as well as the distance between the optimal decision functions based on the representations. We now present the proof for Theorem 20: **Proof** [Proof of Theorem 20] First, we show that for  $a \in \{0, 1\}$ ,  $\text{Err}_{\mu_a}(h)$  cannot be too large if  $h$  is close to  $h_a^*$ . Note that since we are focusing on binary classification problems, for any two classifiers  $h, h'$ ,  $\Pr(h(X) \neq h'(X)) = \mathbb{E}[|h(X) - h'(X)|]$ .

$$\begin{aligned} |\text{Err}_{\mu_a}(h) - \mathbb{E}_{\mu_a}[|h(X) - h_a^*(X)|]| &= |\mathbb{E}_{\mu_a}[|h(X) - Y|] - \mathbb{E}_{\mu_a}[|h(X) - h_a^*(X)|]| \\ &\leq \mathbb{E}_{\mu_a}[||h(X) - Y| - |h(X) - h_a^*(X)||] \\ &\leq \mathbb{E}_{\mu_a}[|Y - h_a^*(X)|] = n_{\mu_a}, \end{aligned}$$

where both inequalities are due to the triangle inequality. Next, we bound  $|\text{Err}_{\mu_0}(h) - \text{Err}_{\mu_1}(h)|$  by:

$$|\text{Err}_{\mu_0}(h) - \text{Err}_{\mu_1}(h)| \leq (n_{\mu_0} + n_{\mu_1}) + |\mathbb{E}_{\mu_0}[|h(X) - h_0^*(X)|] - \mathbb{E}_{\mu_1}[|h(X) - h_1^*(X)|]|.$$

In order to show this, define  $\varepsilon_a(h, h') := \mathbb{E}_{\mu_a}[|h(X) - h'(X)|]$  so that

$$|\mathbb{E}_{\mu_0}[|h(X) - h_0^*(X)|] - \mathbb{E}_{\mu_1}[|h(X) - h_1^*(X)|]| = |\varepsilon_0(h, h_0^*) - \varepsilon_1(h, h_1^*)|.$$

To bound  $|\varepsilon_0(h, h_0^*) - \varepsilon_1(h, h_1^*)|$ , realize that  $|h(X) - h_a^*(X)| \in \{0, 1\}$ . On one hand, we have:

$$\begin{aligned} |\varepsilon_0(h, h_0^*) - \varepsilon_1(h, h_1^*)| &= |\varepsilon_0(h, h_0^*) - \varepsilon_0(h, h_1^*) + \varepsilon_0(h, h_1^*) - \varepsilon_1(h, h_1^*)| \\ &\leq |\varepsilon_0(h, h_0^*) - \varepsilon_0(h, h_1^*)| + |\varepsilon_0(h, h_1^*) - \varepsilon_1(h, h_1^*)| \\ &\leq \varepsilon_0(h_0^*, h_1^*) + d_{\text{TV}}(\mu_0(X), \mu_1(X)), \end{aligned}$$

where the last inequality is due to  $|\varepsilon_0(h, h_1^*) - \varepsilon_1(h, h_1^*)| = |\mu_0(|h - h_1^*| = 1) - \mu_1(|h - h_1^*| = 1)| \leq \sup_E |\mu_0(E) - \mu_1(E)| = d_{\text{TV}}(\mu_0, \mu_1)$ . Similarly, by subtracting and adding back  $\varepsilon_1(h, h_0^*)$  instead, we can also show that  $|\varepsilon_0(h, h_0^*) - \varepsilon_1(h, h_1^*)| \leq \varepsilon_1(h_0^*, h_1^*) + d_{\text{TV}}(\mu_0(X), \mu_1(X))$ .

Combine the above two inequalities yielding:

$$|\varepsilon_0(h, h_0^*) - \varepsilon_1(h, h_1^*)| \leq \min\{\varepsilon_0(h_0^*, h_1^*), \varepsilon_1(h_0^*, h_1^*)\} + d_{\text{TV}}(\mu_0(X), \mu_1(X)).$$

Incorporating the sum of group-wise noise back to the above inequality by using one more triangle inequality then completes the proof.  $\blacksquare$

## 7. Empirical Results

Our theoretical results on the lower bound imply that over-training the feature transformation function to achieve group-invariant representations will inevitably lead to large joint errors. On the other hand, our upper bound also implies that group-invariant representations help to achieve accuracy parity. To verify these theoretical implications, in this section we conduct experiments on a real-world benchmark dataset, the UCI Adult dataset, to present empirical results with various metrics.

**Dataset** The Adult dataset contains 30,162/15,060 training/test instances for income prediction. Each instance in the dataset describes an adult from the 1994 US Census. Attributes include gender, education level, age, etc. In this experiment we use gender (binary) as the sensitive attribute, and we preprocess the dataset to convert categorical variables into one-hot representations. The processed data contains 114 attributes. The target variable (income) is also binary: 1 if  $\geq 50\text{K}/\text{year}$  otherwise 0. For the sensitive attribute  $A$ ,  $A = 0$  means Male otherwise Female. In this dataset, the base rates across groups are different:  $\Pr(Y = 1 | A = 0) = 0.310$  while  $\Pr(Y = 1 | A = 1) = 0.113$ . Also, the group ratios are different:  $\Pr(A = 0) = 0.673$ .

**Experimental Protocol** To validate the effect of learning group-invariant representations with adversarial debiasing techniques (Zhang et al., 2018; Madras et al., 2018; Beutel et al., 2017), we perform a controlled experiment by fixing the baseline network architecture to be a three hidden-layer feed-forward network with ReLU activations. The number of units in each hidden layer are 500, 200, and 100, respectively. The output layer corresponds to a logistic regression model. This baseline without debiasing is denoted as NoDebias. For debiasing with adversarial learning techniques, the adversarial discriminator network takes the feature from the last hidden layer as input, and connects it to a hidden-layer with 50 units, followed by a binary classifier whose goal is to predict the sensitive

attribute  $A$ . This model is denoted as AdvDebias. Compared with NoDebias, the only difference of AdvDebias in terms of objective function is that besides the cross-entropy loss for target prediction, the AdvDebias also contains a classification loss from the adversarial discriminator to predict the sensitive attribute  $A$ . In the experiment, all the other factors are fixed to be the same between these two methods, including learning rate, optimization algorithm, training epoch, and also batch size. To see how the adversarial loss affects the joint error, the demographic parity as well as the accuracy parity, we vary the coefficient  $\rho$  for the adversarial loss between 0.1, 1.0, 5.0 and 50.0.

**Results and Analysis** The experimental results are listed in Table 2. Note that in the table  $|\text{Err}_{\mu_0} - \text{Err}_{\mu_1}|$  could be understood as measuring an approximate version of accuracy parity, and similarly  $\Delta_{\text{DP}}(\hat{Y})$  measures the closeness of the classifier to satisfy demographic parity. From the table, it is then clear that with increasing  $\rho$ , both the overall error  $\text{Err}_{\mu}$  (sensitive to the marginal distribution of  $A$ ) and the joint error  $\text{Err}_{\mu_0} + \text{Err}_{\mu_1}$  (insensitive to the imbalance of  $A$ ) are increasing. As expected,  $\Delta_{\text{DP}}(\hat{Y})$  is drastically decreasing with the increasing of  $\rho$ . Furthermore,  $|\text{Err}_{\mu_0} - \text{Err}_{\mu_1}|$  is also gradually decreasing, but much slowly than  $\Delta_{\text{DP}}(\hat{Y})$ . This is due to the existing noise in the data as well as the shift between the optimal decision functions across groups, as indicated by our upper bound. To conclude, all the empirical results are consistent with our theoretical findings.

Table 2: Adversarial debiasing on demographic parity, joint error across groups, and accuracy parity.

	$\text{Err}_{\mu}$	$\text{Err}_{\mu_0} + \text{Err}_{\mu_1}$	$ \text{Err}_{\mu_0} - \text{Err}_{\mu_1} $	$\Delta_{\text{DP}}(\hat{Y})$
NoDebias	0.157	0.275	0.115	0.189
AdvDebias, $\rho = 0.1$	0.159	0.278	0.116	0.190
AdvDebias, $\rho = 1.0$	0.162	0.286	0.106	0.113
AdvDebias, $\rho = 5.0$	0.166	0.295	0.106	0.032
AdvDebias, $\rho = 50.0$	0.201	0.360	0.112	0.028

## 8. Related Work

**Tradeoff between Fairness and Accuracy** Although it has long been empirically observed that there is an inherent tradeoff between accuracy and statistical parity in both classification and regression problems (Calders et al., 2009; Zafar et al., 2015; Zliobaite, 2015; Corbett-Davies et al., 2017; Zhao et al., 2019c; Zhao, 2021), precise characterizations on such tradeoffs are less explored. Menon and Williamson (2018, Proposition 8) explored such tradeoff in terms of the fairness frontier function under the context of cost-sensitive binary classification. In this work the fair machine learning problem is reduced to learning a classifier which optimizes a difference between cost-sensitive risks, one with respect to the target variable and one with respect to the sensitive variable. Zhao and Gordon (2019) proved a lower bound of accuracy on both the sum of group-wise errors as well as the joint error that has to be incurred by any fair algorithm satisfying statistical parity. In this paper, assuming oracle access to Bayes optimal classifiers, we also give an algorithm to construct an optimal fair classifier that can verify the lower bound. Furthermore, we also extend the preliminary result in Zhao and Gordon (2019) for binary classification and binary protected attribute to the general multi-class classification setting where the protected attribute can take more than two values, i.e., there are more than two groups defined by the protected attribute. Recently, Chzhen et al.

(2020) and Le Gouic et al. (2020) concurrently derived an analytic bound to characterize the price of statistical parity in regression when the learner can take the sensitive attribute explicitly as an input for  $\ell_2$  loss. In this case, the lower bound is given by the optimal transportation distance from two group distributions to a common one, characterized by the  $W_2$  barycenter. Our result complements this line of works for the classification setting, where we show that the price paid by a fair classifier for general multi-class classification problems with a categorical protected attribute is given by the so-called TV-Barycenter problem.

On the upside, under certain data generative assumptions of the sampling bias, there is a line of recent works showing that fairness constraints could instead improve the accuracy of the predictor (Dutta et al., 2020; Blum and Stangl, 2020). In particular, Blum and Stangl (2020) prove that if the observable data are subject to labeling bias, then the Equality of Opportunity constraint could help recover the Bayes optimal classifier. Note that this does not contradict with our results, since in this work we do not make any assumptions on the underlying training distributions, and we mainly focus on statistical parity and accuracy parity, rather than equalized odds.

**Regularization Techniques** The line of work on fairness-aware learning through regularization dates at least back to Kamishima et al. (2012), where the authors argue that simple deletion of sensitive features in data is insufficient for eliminating biases in automated decision making, due to the possible correlations among attributes and sensitive information (Lum and Johndrow, 2016). In light of this, the authors proposed a *prejudice remover* regularizer that essentially penalizes the mutual information between the predicted goal and the sensitive information. In a more recent approach, Zafar et al. (2015) leveraged a measure of decision boundary fairness and incorporated it via constraints into the objective function of logistic regression as well as support vector machines. As discussed in Section 2, both approaches essentially reduce to achieving demographic parity through regularization.

**Fair Representations** In a pioneer work, Zemel et al. (2013) proposed to preserve both group and individual fairness through the lens of representation learning, where the main idea is to find a good representation of the data with two competing goals: to encode the data for accuracy maximization while at the same time to obfuscate any information about membership in the protected group. Due to the power of learning rich representations offered by deep neural nets, recent advances in building fair automated decision making systems focus on using adversarial techniques to learn fair representation that also preserves enough information for the prediction vendor to achieve his accuracy (Edwards and Storkey, 2015; Louizos et al., 2015; Beutel et al., 2017; Zhang et al., 2018; Adel et al., 2019; Song et al., 2019; Zhao et al., 2019c). Madras et al. (2018) further extended this approach by incorporating reconstruction loss given by an autoencoder into the objective function to preserve demographic parity, equalized odds, and equal opportunity.

## 9. Discussion and Conclusion

In this paper we theoretically and empirically study the important problem of quantifying the tradeoff between accuracy and statistical parity in algorithmic fairness. Specifically, we prove a novel lower bound to characterize the tradeoff between statistical parity and the joint accuracy across different population groups when the base rates differ between groups. In particular, our results imply that, in general, any method aiming to satisfy statistical parity admits an information-theoretic lower bound on the joint error. This holds even only approximate statistical parity is met. In light of this

impossibility result, under the statistical parity constraint, we can only hope to design algorithms that achieve the accuracy lower bound. To this end, assuming oracle access to the potentially unfair Bayes classifiers, we construct an algorithm that returns a randomized classifier, and we prove that this randomized classifier is both optimal (in terms of accuracy) and fair.

When the number of groups defined by the protected attribute is more than two, we can no longer obtain an analytic form of the lower bound. Nevertheless, we show that it can be efficiently computed by solving a linear program in polynomial time, which we term as the TV-Barycenter problem. This finding also builds a connection between the tradeoff problem in algorithmic fairness and the barycenter problem (under the TV-distance) in optimal transport. Complementary to our negative results, we also show that learning fair representations leads to accuracy parity if the Bayes optimal classifiers across different groups are close. Our theoretical findings are also confirmed empirically on a real-world dataset. We believe our results take an important step towards better understanding the tradeoff between accuracy and different notions of fairness.

## **Acknowledgments**

HZ and GG would like to acknowledge support from the DARPA XAI project, contract #FA87501720152 and a Nvidia GPU grant. HZ would also like to thank support from a Facebook research award. The authors are very grateful to the anonymous reviewers for the suggestions on improving the presentation of this work.

## References

- Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-network adversarial fairness. In *33rd AAAI Conference on Artificial Intelligence*, 2019.
- Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.
- Solon Barocas and Andrew D Selbst. Big data’s disparate impact. *Calif. L. Rev.*, 104:671, 2016.
- Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.
- Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.
- Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.
- Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? In *Symposium on Foundations of Responsible Computing (FORC)*, volume 1, 2020.
- Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.
- Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.
- Jianfeng Chi, Yuan Tian, Geoffrey J Gordon, and Han Zhao. Understanding and mitigating accuracy disparity in regression. In *International Conference on Machine Learning*, 2021.
- Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.
- Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. *arXiv preprint arXiv:2006.07286*, 2020.
- Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.
- Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108, 1964.
- Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

- Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*, pages 2803–2813. PMLR, 2020.
- Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.
- Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.
- GDPR. General data protection regulation. URL <https://gdpr-info.eu/art-22-gdpr/>. [Online; accessed 13-May-2021].
- Peter D Grünwald, A Philip Dawid, et al. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of statistics*, 32(4):1367–1433, 2004.
- Jihun Hamm. Minimax filter: Learning to preserve privacy from inference attacks. *The Journal of Machine Learning Research*, 18(1):4704–4734, 2017.
- Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.
- James E Johndrow, Kristian Lum, et al. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 2019.
- Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.
- Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.
- Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.
- Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.
- Mohammadali Khosravifard, Dariush Fooladivanda, and T Aaron Gulliver. Confliction of the convexity and metric properties in f-divergences. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 90(9):1848–1853, 2007.

- Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.
- Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to fairness in statistical learning. *arXiv e-prints*, pages arXiv–2005, 2020.
- Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.
- Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.
- Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.
- Kristian Lum and James Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.
- David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3381–3390, 2018.
- Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118, 2018.
- Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, 2018.
- Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.
- Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *Artificial Intelligence and Statistics*, pages 2164–2173, 2019.
- Yihong Wu. Lecture notes on information-theoretic methods for high-dimensional statistics. *Lecture Notes for ECE598YW*, 16, 2017.
- Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.
- Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.
- Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.
- Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.

- Han Zhao. Costs and benefits of wasserstein fair regression. *arXiv preprint arXiv:2106.08812*, 2021.
- Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 32:15675–15685, 2019.
- Han Zhao, Jianfeng Chi, Yuan Tian, and Geoffrey J. Gordon. Adversarial privacy preservation under attribute inference attack. *arXiv preprint arXiv:1906.07902*, 2019a.
- Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. On learning invariant representation for domain adaptation. In *International Conference on Machine Learning*, 2019b.
- Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional learning of fair representations. *arXiv preprint arXiv:1910.07162*, 2019c.
- Han Zhao, Junjie Hu, and Andrej Risteski. On learning language-invariant representations for universal machine translation. In *International Conference on Machine Learning*, pages 11352–11364. PMLR, 2020.
- Indre Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.