Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing

Sanghamitra Dutta [1 2] **Dennis Wei** [1] **Hazar Yueksel** [1] **Pin-Yu Chen** [1] **Sijia Liu** [1] **Kush R. Varshney** [1]

Abstract

A trade-off between accuracy and fairness is almost taken as a given in the existing literature on fairness in machine learning. Yet, it is not preordained that accuracy should decrease with increased fairness. Novel to this work, we examine fair classification through the lens of *mismatched hypothesis testing*: trying to find a classifier that distinguishes between two ideal distributions when given two mismatched distributions that are biased. Using Chernoff information, a tool in information theory, we theoretically demonstrate that, contrary to popular belief, there always exist ideal distributions such that optimal fairness and accuracy (with respect to the ideal distributions) are achieved simultaneously: there is no trade-off. Moreover, the same classifier yields the lack of a trade-off with respect to ideal distributions while yielding a trade-off when accuracy is measured with respect to the given (possibly biased) dataset. To complement our main result, we formulate an optimization to find ideal distributions and derive fundamental limits to explain why a trade-off exists on the given biased dataset. We also derive conditions under which active data collection can alleviate the fairness-accuracy trade-off in the real world. Our results lead us to contend that it is problematic to measure accuracy with respect to data that reflects bias, and instead, we should be considering accuracy with respect to ideal, unbiased data.

1. Introduction

This work addresses a fundamental question in the field of algorithmic fairness (Calmon et al., 2017; Dwork et al., 2012; Agarwal et al., 2018; Hardt et al., 2016; Ghassami et al., 2018; Kusner et al., 2017; Kilbertus et al., 2017; Zemel et al., 2013):

Is there a trade-off between fairness and accuracy?

The existence of this trade-off has been pointed out in several existing works (Menon & Williamson, 2018; Chen et al., 2018; Zhao & Gordon, 2019) that also propose different theoretical approaches to characterize it. Yet, it is not preordained as to why such a trade-off should exist between fairness and accuracy. For instance, Friedler et al. (2016) and Yeom & Tschantz (2018) suggest that the observed features in a machine learning model (e.g., test scores) are a possibly noisy mapping from features in an abstract construct space (e.g., true ability) where there is no such trade-off. Then, why does correcting for biases worsen predictive accuracy in the real world? We believe there is value in stepping back and reposing the fundamental question.

In this work, our main assertion is that the trade-off between accuracy and fairness (in particular, equal opportunity (Hardt et al., 2016)) in the real world is due to noisier (and hence biased) mappings for the unprivileged group due to historic differences in opportunity, representation, etc., making their positive and negative labels "less separable." To concretize this idea, we adopt a novel viewpoint on fair classification: the perspective of mismatched hypothesis testing. In mismatched hypothesis testing, the goal is to find a classifier that distinguishes between two "ideal" distributions, but instead, one only has access to two mismatched distributions that are biased. Our most important result is to theoretically show that for a fair classifier with sub-optimal accuracy on the given biased data distributions, there always exist ideal distributions such that fairness and accuracy are in accord when accuracy is measured with respect to the ideal distributions. Through this perspective, there is no trade-off between fairness and accuracy.

Our contributions in this work are as follows:

Concept of separability to quantify accuracy-fairness trade-off in the real world: For a group of people in an observed dataset, we quantify the "separability" into positive and negative class labels using Chernoff information, an information-theoretic approximation to the best exponent of

[1]IBM Research [2]Department of Electrical and Computer Engineering, Carnegie Mellon University. Correspondence to: Sanghamitra Dutta <sanghamd@andrew.cmu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

the probability of error in binary classification. We demonstrate (in Theorem 1) that if the Chernoff information of one group is lower than that of the other in the observed dataset, then modifying the best classifier using a group fairness criterion compromises the error exponent (representative of accuracy) of one or both the groups, explaining the accuracy-fairness trade-off. Not only do these tools demonstrate the existence of a trade-off (as also demonstrated in some existing works (Menon & Williamson, 2018; Chen et al., 2018) using alternative formulations), but they also enable us to approximately quantify the trade-off, e.g., how close can we bring the probabilities of false negative for two groups in an attempt to attain equal opportunity for a certain compromise on accuracy (see Fig. 3 in Section 4). The existence of this trade-off prompts us to contend that accuracy of a classifier with respect to the existing (possibly biased) dataset is a problematic measure of performance. Instead, one should consider accuracy with respect to an ideal dataset that is an unbiased representation of the population.

Ideal distributions where fairness and accuracy are in accord: Novel to this work, we examine the problem of fair classification through the lens of mismatched hypothesis testing. We show (in Theorem 2) that there exist ideal distributions such that both fairness (in the sense of equal opportunity on both the existing and the ideal distributions) and accuracy (with respect to the ideal distributions) are in accord. We also formulate an optimization to show how to go about finding such ideal distributions in practice. The ideal distributions provide a target to shift the given biased distributions toward and to evaluate accuracy on. Their interpretation can be two-fold: (i) plausible distributions in the observed space resulting from an "unbiased" mapping from the construct space; or (ii) candidate distributions in the construct space itself (discussed further in Section 3.2).

Criterion to alleviate the accuracy-fairness trade-off in the real world: Next, we also address another important question, i.e., when can we alleviate the accuracy-fairness trade-off in the real world that we must work in, specifically through additional data collection. We derive an information-theoretic criterion (in Theorem 3) under which collecting more features improves separability, and hence, accuracy in the real world, alleviating the trade-off. This can also inform our choice of the ideal distributions. Our analysis serves as a technical explanation for the success of active fairness (Noriega-Campero et al., 2019; Bakker et al., 2019; Chen et al., 2018) that uses additional features to improve fairness.

Numerical example: We demonstrate how the analysis works through an example (with analytical closed-forms).

Related Work: We note that several existing works, such as Garg et al. (2019), Menon & Williamson (2018), Chen et al. (2018), and Zhao & Gordon (2019), have also used information theory or Bayes risk to characterize the accuracy-fairness trade-off. However, computing Bayes risk is not straightforward. Indeed, even for Gaussians, one resorts to Chernoff bounds to approximate the Q-function. Chernoff information is an approximation for Bayes risk that has a tractable geometric interpretation (see Fig. 2). This enables us to numerically compute the accuracy-fairness trade-off (Fig. 3), and also understand "how much" accuracy can be improved by data collection, going beyond the assertion that there is some improvement. To the best of our knowledge, existing works have pointed out the existence of a trade-off based on Bayes risk but have not provided a method to exactly compute it, motivating us to introduce the additional tool of Chernoff information to do so approximately. Furthermore, this work goes beyond characterizing the trade-off imposed by the given dataset. Our novelty lies in adopting the perspective of mismatched detection and demonstrating that there exist ideal distributions such that both fairness and accuracy are in accord when accuracy is measured with respect to the ideal distributions.

The recent works of Wick et al. (2019) and Sharma et al. (2020) further elucidate the significance of Theorem 2 and how it presents an insight that contradicts "the prevailing wisdom," i.e., there exists an ideal dataset for which fairness and accuracy are in accord. In a sense, our work provides a theoretical foundation that complements the empirical results of Wick et al. (2019) and Sharma et al. (2020), clarifying when a trade-off exists and when it does not.

There are also several existing methods of pre-processing data to generate a fair dataset (Calmon et al., 2018; Feldman et al., 2015; Zemel et al., 2013). Here, our goal is not to propose another competing strategy of fairness through pre-processing. Instead, our focus is to theoretically demonstrate that there exists an ideal dataset such that a fair classifier is also optimal in terms of accuracy, which has not been formally shown before. We also focus on equal opportunity rather than statistical parity (as in Calmon et al. (2018)).

Our tools share similarities with Varshney et al. (2018) (that demonstrates how explainability can improve Chernoff information), as well as the theory of hypothesis testing in general (Lee & Sung, 2012; Cover & Thomas, 2012). Our contribution lies in using these tools in fair machine learning, where they have not been used to the best of our knowledge (e.g., in the previous analyses of Menon & Williamson (2018); Zhao & Gordon (2019); Chen et al. (2018)).

Remark 1 (Population Setting)**.** *In this work, we operate in the population setting (motivated from Gretton et al. (2007); Ravikumar et al. (2009); Scott et al. (2013)), i.e., the limit as the number of samples goes to infinity, allowing use of the probability distributions of the data. This allows us to represent binary classifiers as likelihood ratio detectors (also called Neyman-Pearson (NP) detectors) and quantify*

the fundamental limits on the accuracy-fairness trade-off. Indeed, given any classifier, there always exists a likelihood ratio detector which is at least as good (see NP Lemma in Cover & Thomas (2012)).

2. Preliminaries

Setup: In this work, we focus on binary classification, which arises commonly in practice in the fairness literature, e.g., in deciding whether a candidate should be accepted or rejected in applications such as hiring, lending, etc. We let Z denote the protected attribute, e.g., gender, race, etc. Without loss of generality, let $Z = 0$ be the unprivileged group and $Z = 1$ be the privileged group.

Inspired by Yeom & Tschantz (2018) and Friedler et al. (2016), we assume that there is an abstract construct space where X_a is the feature (e.g., true ability) and Y_a is the true label (i.e., takes value 0 or 1). The construct space is not directly accessible to us. In the real world, we instead have access to an observed space where X denotes the feature vector and Y denotes the true label (i.e., takes value 0 or 1). For the sake of simplicity, we assume $Y_a = Y$ based on Yeom & Tschantz (2018).[1] The observed features are derived from features in the construct space as follows: $X = f_{Y,Z}(X_a)$ where $f_{Y,Z}(\cdot)$ is a possibly noisy mapping that can depend on Y and Z.

Let the features in the given dataset in the observed space have the following distributions: $X|_{Y=0,Z=0} \sim P_0(x)$ and $X|_{Y=1,Z=0} \sim P_1(x)$. Similarly, $X|_{Y=0,Z=1} \sim Q_0(x)$ and $X|_{Y=1,Z=1} \sim Q_1(x)$. For each group $Z = z$, we will be denoting classifiers as $T_z(x) \geq \tau_z$, i.e., the prediction label is 1 when $T_z(x) \geq \tau_z$ and 0 otherwise.

Remark 2 (Decoupled Classifiers)**.** *While such models may exhibit disparate treatment (explicit use of Z), the intent is to better mitigate disparate impact using the protected attribute explicitly in the decision making (along the spirit of fair affirmative action (Dwork et al., 2012; 2018)). Furthermore, a classifier that does not use Z becomes a special case of our classifier if T_z and τ_z are same for both groups.*

Next, we state two basic assumptions: (**A1**) Absolute Continuity: $P_0(x)$, $P_1(x)$, $Q_0(x)$ and $Q_1(x)$ are greater than 0 everywhere in range of x. This ensures that likelihood ratio detectors such as $\log \frac{P_1(x)}{P_0(x)} \geq \tau_0$ and Kullback-Leibler (KL) divergences between any two of these distributions are well-defined. (**A2**) Distinct Hypotheses: $\mathrm{D}(P_0||P_1)$, $\mathrm{D}(P_1||P_0)$, $\mathrm{D}(Q_0||Q_1)$ and $\mathrm{D}(Q_1||Q_0)$ are strictly greater than 0, where $\mathrm{D}(\cdot||\cdot)$ is the KL divergence.

[1] This is consistent with the "What You See Is What You Get" worldview in Yeom & Tschantz (2018) where label bias can be ignored and our chosen measure of fairness, i.e., equal opportunity is justified as a measure of fairness.

We let $P_{\mathrm{FP},T_z}(\tau_z)$ be the probability of false positive (wrongful acceptance of negative class labels; also called false positive rate (FPR)) over the group $Z = z$, i.e., $P_{\mathrm{FP},T_z}(\tau_z) = \Pr\left(T_z(X) \geq \tau_z | Y = 0, Z = z\right)$. Similarly, $P_{\mathrm{FN},T_z}(\tau_z)$ is the probability of false negative (wrongful rejection of positive class labels; also called false negative rate (FNR)), given by: $P_{\mathrm{FN},T_z}(\tau_z) = \Pr\left(T_z(X) < \tau_z | Y = 1, Z = z\right)$. The overall probability of error of a group is given by: $P_{e,T_z}(\tau_z) = \pi_0 P_{\mathrm{FP},T_z}(\tau_z) + \pi_1 P_{\mathrm{FN},T_z}(\tau_z)$, where π_0 and π_1 are the prior probabilities of $Y = 0$ and $Y = 1$ given $Z = z$. For the sake of simplicity, we consider the case where $\pi_0 = \pi_1 = \frac{1}{2}$ given $Z = z$, and also equal priors on all groups $Z = z$. We include a discussion on how to extend our results for the case of unequal priors in Appendix E. Equal priors also correspond to the balanced accuracy measure (Brodersen et al., 2010) which is often favored over ordinary accuracy.

A well-known definition of fairness is *equalized odds* (Hardt et al., 2016), which states that an algorithm is fair if it has equal probabilities of false positive (wrongful acceptance of true negative class labels) and false negative (wrongful acceptance of true positive class labels) for the two groups, i.e., $Z = 0$ and 1. A relaxed variant of this measure, widely used in the literature, is *equal opportunity*, which enforces only equal false negative rate (or equivalently, equal true positive rate) for the two groups. In this work, we focus primarly on equal opportunity, although the arguments can be extended to other measures of fairness as well, e.g., statistical parity (Agarwal et al., 2018).

We assume that in the construct space, there is no trade-off between accuracy and equal opportunity, i.e., the Bayes optimal (Cover & Thomas, 2012) classifiers for the groups $Z = 0$ and $Z = 1$ also satisfy equal opportunity (equal probabilities of false negative). In this work, our objective is to explain the accuracy-fairness trade-off in the observed space and attempt to find ideal distributions with respect to which there is no trade-off. We now provide a brief background on error exponents of a classifier to help follow the rest of the paper.

Background on Error Exponents of a Classifier: The error exponents of the FPR and FNR are given by $-\log P_{\mathrm{FP},T_z}(\tau_z)$ and $-\log P_{\mathrm{FN},T_z}(\tau_z)$. Often, we may not be able to obtain a closed-form expression for the exact error probabilities or their exponents, but the exponents are approximated using a well-known lower bound called the *Chernoff bound* (see Lemma 1; proof in Appendix A.1), that is known to be pretty tight (see Remark 3 and also Motwani & Raghavan (1995); Berend & Kontorovich (2015)).

Definition 1. *The Chernoff exponents of $P_{\mathrm{FP},T_z}(\tau_z)$ and $P_{\mathrm{FN},T_z}(\tau_z)$ are defined as:*

$$E_{\mathrm{FP},T_z}(\tau_z) = \sup_{u>0}(u\tau_z - \Lambda_0(u)), \text{ and}$$

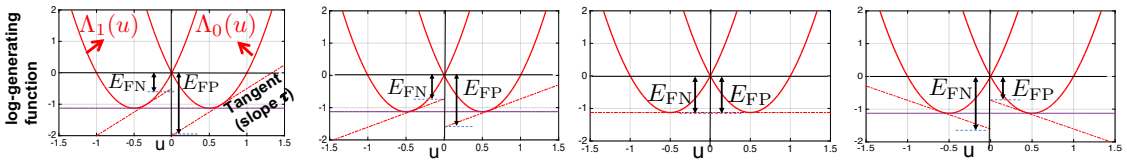

Figure 1. Let $P_0(x)\sim\mathcal{N}(1,1)$ and $P_1(x)\sim\mathcal{N}(4,1)$. For a likelihood ratio detector $T(x)=\log\frac{P_1(x)}{P_0(x)}\geq\tau$, we can compute the log-generating functions as follows: $\Lambda_0(u)=\frac{9}{2}u(u-1)$ and $\Lambda_1(u)=\frac{9}{2}u(u+1)$ (derived in Appendix A.3). Note that, $\Lambda_0(u)$ is strictly convex with zeros at u=0 and u=1, and $\Lambda_1(u)$=$\Lambda_0(u+1)$. We obtain $E_{\text{FP},T}(\tau)$ and $E_{\text{FN},T}(\tau)$ as the negative of the y-intercepts for tangents to $\Lambda_0(u)$ and $\Lambda_1(u)$ respectively with slope τ. As we vary the slope of the tangent (τ), there is a trade-off between $E_{\text{FP},T}(\tau)$ and $E_{\text{FN},T}(\tau)$ until they both become equal at $\tau=0$ (third figure from left). The value of the exponent at τ=0 (negative of the y-intercepts for tangents with 0-slope) is defined as the Chernoff Information, given by: $\text{C}(P_0,P_1)$:=$E_{\text{FP},T}(0)$=$E_{\text{FN},T}(0)$, which is equal to $9/8$ for this particular example.

$$E_{\text{FN},T_z}(\tau_z)=\sup_{u<0}(u\tau_z-\Lambda_1(u)).$$

Here, $\Lambda_0(u)$ and $\Lambda_1(u)$ are called log-generating functions, given by $\Lambda_0(u)=\log\mathbb{E}[e^{uT_z(X)}|Y=0,Z=z]$ and $\Lambda_1(u)=\log\mathbb{E}[e^{uT_z(X)}|Y=1,Z=z]$.

Lemma 1 (Chernoff Bound)**.** *The exponents satisfy: $P_{\text{FP},T_z}(\tau_z)\leq e^{-E_{\text{FP},T_z}(\tau_z)}$ and $P_{\text{FN},T_z}(\tau_z)\leq e^{-E_{\text{FN},T_z}(\tau_z)}$.*

Remark 3 (Tightness of the Chernoff Bound)**.** *For Gaussian distributions, the tail probabilities are characterized by the Q-function which has both upper and lower bounds in terms of Chernoff exponents with constant factors that do not affect the exponent significantly (Côté et al., 2012). The Bhattacharya bound (a special case of Chernoff bound) both upper and lower bounds the Bayes error exponent (Berisha et al., 2015; Bhattacharyya, 1946; Kailath, 1967).*

Geometric Interpretation of Chernoff Exponents: Chernoff exponents yield more insight than exact error exponents because of their geometric interpretation, as we discuss here (more details in Appendix A.2).

For ease of understanding, we refer to Fig. 1 where we illustrate the idea of Chernoff exponents with a numerical example. In general, the log-generating functions are convex and become 0 at $u=0$ (see Appendix A.2). Furthermore, if a detector is well-behaved[2], i.e., $\mathbb{E}[T_z(X)|Y{=}1,Z{=}z]{>}0$ and $\mathbb{E}[T_z(X)|Y{=}0,Z{=}z]{<}0$, then $\Lambda_0(u)$ and $\Lambda_1(u)$ are strictly convex and attain their minima on either sides of the origin. The Chernoff exponents $E_{\text{FP},T_z}(\tau_z)$ and $E_{\text{FN},T_z}(\tau_z)$ can be obtained as the negative of the y-intercepts for tangents to $\Lambda_0(u)$ and $\Lambda_1(u)$ with slope τ_z (for $\tau_z\in(\mathbb{E}[T_z(X)|Y{=}0,Z{=}z],\mathbb{E}[T_z(X)|Y{=}1,Z{=}z])$).

Definition 2. *The Chernoff exponent of the overall probability of error $P_{e,T_z}(\tau_z)$ is defined as:*

$$E_{e,T_z}(\tau_z)=\min\{E_{\text{FP},T_z}(\tau_z),E_{\text{FN},T_z}(\tau_z)\}.$$

Recall that, under equal priors, we have $P_{e,T_z}(\tau_z)=\frac{1}{2}P_{\text{FP},T_z}(\tau_z)+\frac{1}{2}P_{\text{FN},T_z}(\tau_z)$. The exponent of $P_{e,T_z}(\tau_z)$ is dominated by the minimum of the error exponents of $P_{\text{FP},T_z}(\tau_z)$ and $P_{\text{FN},T_z}(\tau_z)$, which in turn is bounded by the minimum of the Chernoff exponents of FPR and FNR (Definition 1). A higher $E_{e,T_z}(\tau_z)$ indicates higher accuracy, i.e., lower $P_{e,T_z}(\tau_z)$. To understand this, first consider likelihood ratio detectors of the form $T_0(x)=\log\frac{P_1(x)}{P_0(x)}$ for $Z=0$. As we vary τ_0, there is a trade-off between $P_{\text{FP},T_0}(\tau_0)$ and $P_{\text{FN},T_0}(\tau_0)$, i.e., as one increases, the other decreases. A similar trade-off is also observed in their Chernoff exponents (see Fig. 1). $P_{e,T_0}(\tau_0)$ is minimized when $\tau_0=0$ (for equal priors) and $P_{\text{FP},T_0}(0)$=$P_{\text{FN},T_0}(0)$. For this optimal value of $\tau_0=0$, the Chernoff exponents of FPR and FNR also become equal, i.e., $E_{\text{FP},T_0}(0)$=$E_{\text{FN},T_0}(0)$, and the maximum value of $E_{e,T_0}(\tau_0)$=$\min\{E_{\text{FP},T_0}(\tau_0),E_{\text{FN},T_0}(\tau_0)\}$ is attained. This exponent is called the Chernoff information (Cover & Thomas, 2012). For completeness, we include a well-known result on Chernoff information from Cover & Thomas (2012) with the proof in Appendix A.4.

Lemma 2. *For two hypotheses $P_0(x)$ under $Y=0$ and $P_1(x)$ under $Y=1$, the Chernoff exponent of the probability of error of the Bayes optimal classifier is given by the Chernoff information[3]:*

$$\text{C}(P_0,P_1)=-\min_{u\in(0,1)}\log\left(\sum_x P_0(x)^{1-u}P_1(x)^u\right). \quad (1)$$

Goals: Our metrics of interest for *accuracy* are $E_{e,T_0}(\tau_0)$ and $E_{e,T_1}(\tau_1)$ because a higher value of the Chernoff exponent of $P_{e,T_z}(\tau_z)$ implies a higher accuracy for the respective groups Z=0 and Z=1. Our metric of interest for *fairness* is the difference of the Chernoff exponents of FNR, i.e.,

[2]For a detector $T_z(x)\geq\tau_z$, we would expect $T_z(X)$ to be high when Y=1, and low when Y=0 justifying the criteria $\mathbb{E}[T_z(X)|Y{=}1,Z{=}z]{>}0$ and $\mathbb{E}[T_z(X)|Y{=}0,Z{=}z]{<}0$ for being well-behaved. A likelihood ratio detector $T_0(x)=\log\frac{P_1(x)}{P_0(x)}\geq\tau_0$ is well-behaved under assumption A2 in Section 2 because we have $\mathbb{E}[T_z(X)|Y{=}1,Z{=}z]$=$D(P_1||P_0)$ and $\mathbb{E}[T_z(X)|Y{=}0,Z{=}z]$=$-D(P_0||P_1)$.

[3]When $P_0(x)$ and $P_1(x)$ are continuous distributions, the summation is replaced by an integral over x (see Appendix A.3).

$|E_{\text{FN},T_0}(\tau_0) - E_{\text{FN},T_1}(\tau_1)|$ (inspired from equal opportunity). A model is *fair* when $|E_{\text{FN},T_0}(\tau_0) - E_{\text{FN},T_1}(\tau_1)| = 0$, and progressively becomes more and more unfair as this quantity $|E_{\text{FN},T_0}(\tau_0) - E_{\text{FN},T_1}(\tau_1)|$ increases.

Our first goal is to quantify fundamental limits on the best accuracy-fairness trade-off in terms of our metrics of interest on an existing real-world dataset, i.e., given observed distributions $P_0(x)$, $P_1(x)$, $Q_0(x)$, and $Q_1(x)$. Next, our goal is to find ideal distributions where fairness and accuracy are in accord when accuracy is measured with respect to the ideal distributions.

3. Main Results

3.1. Concept of Separability: Fundamental Limits on Accuracy-Fairness Trade-Off in the Real World

Given the setup in Section 2, we show that the trade-off between accuracy and equal opportunity in the observed space is due to noisier mappings for the unprivileged group making their positive and negative labels less separable. Let us first formally define our intuitive notion of separability.

Definition 3. *For a group of people with distributions $P_0(x)$ and $P_1(x)$ under hypotheses $Y{=}0$ and $Y{=}1$, we define the separability as their Chernoff information $\mathrm{C}(P_0, P_1)$.*

Definition 3 is motivated from Lemma 2 because Chernoff information essentially provides an information-theoretic approximation to the best classification accuracy (in an exponent sense) for a group of people in a given dataset. Next, we define unbiased mappings from a separability standpoint.

Definition 4. *Consider the setup in Section 2. The mapping $X = f_{Y,Z}(X_a)$ from the construct space to the observed space is said to be unbiased if $\mathrm{C}(P_0, P_1) = \mathrm{C}(Q_0, Q_1)$.*

Our next result demonstrates that the trade-off between fairness and accuracy arises due to a bias in the mappings from a separability standpoint, i.e., $\mathrm{C}(P_0, P_1) \neq \mathrm{C}(Q_0, Q_1)$. Because we assumed that $Z = 0$ is the unprivileged group, we let $\mathrm{C}(P_0, P_1)$ be either equal to, or less than $\mathrm{C}(Q_0, Q_1)$.

Theorem 1 (Explaining the Trade-Off)**.** *For the setup in Section 2, one of the following is true:*

1. *Unbiased Mappings, i.e., $\mathrm{C}(P_0, P_1){=}\mathrm{C}(Q_0, Q_1)$: The Bayes optimal detectors $T_0(x) \geq \tau_0$ and $T_1(x) \geq \tau_1$ for the two groups with Chernoff exponents of the probability of error $\mathrm{C}(Q_0, Q_1)(= \mathrm{C}(P_0, P_1))$ also attain fairness, i.e., $|E_{\text{FN},T_0}(\tau_0) - E_{\text{FN},T_1}(\tau_1)| = 0$.*
2. *Biased Mappings, i.e., $\mathrm{C}(P_0, P_1) < \mathrm{C}(Q_0, Q_1)$: The Bayes optimal detectors $T_0(x) \geq \tau_0$ and $T_1(x) \geq \tau_1$ for the two groups are not fair, i.e., $|E_{\text{FN},T_0}(\tau_0) - E_{\text{FN},T_1}(\tau_1)| \neq 0$. Furthermore, no likelihood ratio detector can improve the Chernoff exponent of the probability of error for the unprivileged group beyond $\mathrm{C}(P_0, P_1)$.*

The first scenario is where the mappings are unbiased from a separability standpoint, and there is no trade-off between accuracy and fairness. The second scenario, which occurs more commonly in practice, is where discrimination is caused due to an inherent limitation of the dataset: the mappings from the construct space are biased and do not have enough separability information about one group compared to the other. For the rest of the paper, we will focus on the case of $\mathrm{C}(P_0, P_1) < \mathrm{C}(Q_0, Q_1)$. Under this scenario, the Chernoff exponents of FNR of the Bayes optimal detectors for the two groups are $\mathrm{C}(P_0, P_1)$ and $\mathrm{C}(Q_0, Q_1)$ which are unequal, and hence *unfair*. An attempt to ensure fairness by using any alternate likelihood ratio detector for any of the groups will therefore only reduce accuracy (Chernoff exponent of the probability of error) for that group below the Bayes optimal (best) classifier for that group, explaining the accuracy-fairness trade-off. We formalize this intuition in Lemma 3 (used in proof of Theorem 1; see Appendix B).

Lemma 3. *Let $\mathrm{C}(P_0, P_1){<}\mathrm{C}(Q_0, Q_1)$. Suppose that there are two likelihood ratio detectors $T_0(x){\geq}\tau_0$ and $T_1(x){\geq}\tau_1$, one for each group, such that $E_{\text{FN},T_0}(\tau_0){=}E_{\text{FN},T_1}(\tau_1)$. Then, at least one of the following statements is true: (i) $E_{e,T_0}(\tau_0) < \mathrm{C}(P_0, P_1)$, or (ii) $E_{e,T_1}(\tau_1) < \mathrm{C}(Q_0, Q_1)$.*

The next two results show how current and reasonable approaches to fair classification can give rise to each of the two cases in Lemma 3. Consider the following optimization problem, where the goal is to find classifiers of the form $T_0(x) \geq \tau_0$ and $T_1(x) \geq \tau_1$ for the two groups that maximize the Chernoff exponent of the probability of error under the constraint that they are *fair* on the given dataset.

$$\max_{T_0,\tau_0,T_1,\tau_1} \min\{E_{\text{FP},T_0}(\tau_0), E_{\text{FN},T_0}(\tau_0), E_{\text{FP},T_1}(\tau_1), E_{\text{FN},T_1}(\tau_1)\} \quad \text{such that } E_{\text{FN},T_0}(\tau_0) = E_{\text{FN},T_1}(\tau_1). \tag{2}$$

This optimization is in the spirit of existing works (Zafar et al., 2017; Agarwal et al., 2018; Donini et al., 2018; Celis et al., 2019) that maximize accuracy under fairness constraints. From the NP Lemma, we know that given any classifier, there exists a likelihood ratio detector which is at least as good in terms of accuracy. If we restrict $T_0(x)$ and $T_1(x)$ to be likelihood ratio detectors of the form $\log \frac{P_1(x)}{P_0(x)}$ and $\log \frac{Q_1(x)}{Q_0(x)}$, then (2) has a unique solution (τ_0^*, τ_1^*).

Lemma 4. *Let $\mathrm{C}(P_0, P_1){<}\mathrm{C}(Q_0, Q_1)$ and $T_0(x)$ and $T_1(x)$ be restricted to be likelihood ratio detectors. Then the detectors $T_0(x) \geq \tau_0^*$ and $T_1(x) \geq \tau_1^*$ that solve the optimization (2) are the Bayes optimal detector for the unprivileged group ($\tau_0^* = 0$) and a sub-optimal detector for the privileged group ($\tau_1^* > 0$) with $E_{e,T_1}(\tau_1^*) < \mathrm{C}(Q_0, Q_1)$.*

As a proof sketch, we refer to Fig. 2 (Left). Let $\tau_0^* = 0$, which ensures $E_{\text{FN},T_0}(0) = E_{\text{FP},T_0}(0) = \mathrm{C}(P_0, P_1)$.

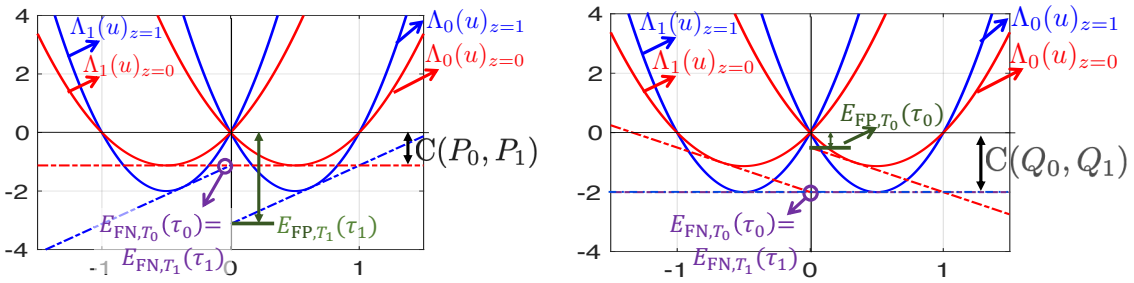

Figure 2. Let the distributions for the unprivileged group ($Z=0$) be $P_0(x)\sim\mathcal{N}(1,1)$ and $P_1(x)\sim\mathcal{N}(4,1)$. Also, let the distributions of the privileged group be $Q_0(x)\sim\mathcal{N}(0,1)$ and $Q_1(x)\sim\mathcal{N}(4,1)$. In both the figures, the red and blue curves denote the log-generating functions for the likelihood ratio detectors for the groups $Z=0$ and $Z=1$ respectively (see Appendix A.3 for derivation). We have $\Lambda_0(u)_{z=1}=8u(u-1)$ and $\Lambda_1(u)_{z=1}=8u(u+1)$. Also, $\Lambda_0(u)_{z=0}=\frac{9}{2}u(u-1)$, and $\Lambda_1(u)_{z=0}=\frac{9}{2}u(u+1)$. Note that, $\mathrm{C}(P_0,P_1)<\mathrm{C}(Q_0,Q_1)$. (**Left**) This plot corresponds to the scenario of Lemma 4. The detector for the group $Z=0$ is the Bayes optimal detector with $\tau_0^*=0$ and $E_{\mathrm{FN},T_0}(\tau_0^*)=E_{\mathrm{FP},T_0}(\tau_0^*)=C(P_0,P_1)$. The detector for the group $Z=1$ is a sub-optimal detector because in order to satisfy equal opportunity, we have to choose τ_1^* such that $E_{\mathrm{FN},T_1}(\tau_1^*)=E_{\mathrm{FN},T_0}(\tau_0^*)=C(P_0,P_1)$ and this is strictly less than $C(Q_0,Q_1)$. (**Right**) This plot corresponds to the scenario of Lemma 5. The detector for the group $Z=1$ is the Bayes optimal detector with $\tau_1^*=0$ and $E_{\mathrm{FN},T_1}(\tau_1^*)=E_{\mathrm{FP},T_1}(\tau_1^*)=C(Q_0,Q_1)$. In order to satisfy equal opportunity, we have to choose τ_0^* such that $E_{\mathrm{FN},T_0}(\tau_0^*)=E_{\mathrm{FN},T_1}(\tau_1^*)=C(Q_0,Q_1)$ which is strictly greater that $C(P_0,P_1)$. However, this threshold τ_0^* makes $E_{\mathrm{FP},T_0}(\tau_0^*)$ lower that $C(P_0,P_1)$, leading to a sub-optimal detector for the group $Z=0$.

Now, the only value of slope τ_1^* that will satisfy $E_{\mathrm{FN},T_1}(\tau_1^*)=E_{\mathrm{FN},T_0}(0)$ is a $\tau_1^*>0$ such that $E_{\mathrm{FN},T_1}(\tau_1^*)=\mathrm{C}(P_0,P_1)<\mathrm{C}(Q_0,Q_1)$, and hence $E_{\mathrm{FP},T_1}(\tau_1^*)>\mathrm{C}(Q_0,Q_1)$. This leads to, $\min\{E_{\mathrm{FP},T_0}(0), E_{\mathrm{FN},T_0}(0), E_{\mathrm{FP},T_1}(\tau_1^*), E_{\mathrm{FN},T_1}(\tau_1^*)\} = \mathrm{C}(P_0,P_1)$.

For $\tau_0^*\neq 0$, either $E_{\mathrm{FP},T_0}(\tau_0^*)<\mathrm{C}(P_0,P_1)<E_{\mathrm{FN},T_0}(\tau_0^*)$, or $E_{\mathrm{FN},T_0}(\tau_0^*)<\mathrm{C}(P_0,P_1)<E_{\mathrm{FP},T_0}(\tau_0^*)$, implying that, $\min\{E_{\mathrm{FP},T_0}(\tau_0^*), E_{\mathrm{FN},T_0}(\tau_0^*), E_{\mathrm{FP},T_1}(\tau_1^*), E_{\mathrm{FN},T_1}(\tau_1^*)\} < \mathrm{C}(P_0,P_1)$.

This situation of reducing the accuracy of the privileged group is often interpreted as causing *active harm* to the privileged group. To avoid causing active harm while satisfying a fairness criterion, we may also consider a variant where we do not alter the optimal detector (or accuracy) of the privileged group (i.e., $E_{\mathrm{FN},T_1}(\tau_1)=E_{\mathrm{FP},T_1}(\tau_1)=\mathrm{C}(Q_0,Q_1)$ for the privileged group), but only vary the detector for the unprivileged group to achieve fairness. We propose the following optimization:

$$\max_{T_0,\tau_0} \min\{E_{\mathrm{FP},T_0}(\tau_0), E_{\mathrm{FN},T_0}(\tau_0)\}$$
$$\text{such that } E_{\mathrm{FN},T_0}(\tau_0)=\mathrm{C}(Q_0,Q_1). \tag{3}$$

Again, if we restrict $T_0(x)$ to be a likelihood ratio detector, then there exists a unique solution τ_0^* to optimization (3).

Lemma 5. *Let* $T_0(x)=\log\frac{P_1(x)}{P_0(x)}$ *and we have* $\mathrm{C}(P_0,P_1)<\mathrm{C}(Q_0,Q_1)$. *The detector* $T_0(x)\geq\tau_0^*$ *that solves optimization* (3) *is a sub-optimal detector for the unprivileged group with* $E_{e,T_0}(\tau_0^*)<\mathrm{C}(P_0,P_1)$.

As a proof sketch, we refer to Fig. 2 (Right). If we choose $\tau_0^*\neq 0$, we get a sub-optimal detector for the unprivileged group with $E_{e,T_0}(\tau_0^*)<\mathrm{C}(P_0,P_1)$. The full proofs for Lemmas 4 and 5 are provided in Appendix B.3.

Remark 4 (Equal priors on Z)**.** *Along the lines of balanced accuracy measures, the optimization assumes equal priors on* $Z=0$ *and* $Z=1$ *as well. We refer to Appendix E.2 for modification of the optimization to account for unequal priors on* $Z=0$ *and* $Z=1$.

Remark 5 (Generalization to other fairness measures)**.** *While we focus on equal opportunity here, the idea extends to other fairness measures as well. For example, if the best likelihood detectors for each group, i.e.,* $T_0(x)\geq 0$ *and* $T_1(x)\geq 0$ *do not satisfy statistical parity (Agarwal et al., 2018), while there are other pairs of detectors for the two groups that do satisfy the criterion, then for at least one of the two groups, a sub-optimal detector is being used.*

3.2. The Mismatched Hypothesis Testing Perspective: Ideal Distributions with no Accuracy-Fairness Trade-Off

Here, we will show that there exist ideal distributions such that fairness and accuracy are in accord. Since the trade-off arises due to insufficient separability of the unprivileged group in the observed space, we are specifically interested in finding ideal distributions for the unprivileged group that match the separability of the privileged, and the same detector that achieved fairness with sub-optimal accuracy in Lemma 5 now achieves optimal accuracy with respect to the ideal distributions. We show the existence of such ideal distributions and also provide an explicit construction.

Theorem 2 (Existence of Ideal Distributions)**.** *For the setup in Section 2, let* $\mathrm{C}(P_0,P_1)<\mathrm{C}(Q_0,Q_1)$. *Let us choose*

the Bayes optimal detector $T_1(x) = \log \frac{Q_1(x)}{Q_0(x)} \geq 0$ *for the group* $Z = 1$. *Then, for group* $Z = 0$, *there exist* $\widetilde{P}_0(x)$ *and* $\widetilde{P}_1(x)$ *of the form* $\widetilde{P}_0(x) = \frac{P_0(x)^{(1-w)}P_1(x)^w}{\sum_x P_0(x)^{(1-w)}P_1(x)^w}$ *and* $\widetilde{P}_1(x) = \frac{P_0(x)^{(1-v)}P_1(x)^v}{\sum_x P_0(x)^{(1-v)}P_1(x)^v}$ *for* $w, v \in \mathcal{R}$ *such that:*

- *(Fairness on given data) The Bayes optimal detector for the ideal distributions, i.e.,* $\widetilde{T_0}(x) = \log \frac{\widetilde{P}_1(x)}{\widetilde{P}_0(x)} \geq 0$ *is equivalent to the detector* $T_0(x) = \log \frac{P_1(x)}{P_0(x)} \geq \tau_0^*$ *of Lemma 5 that satisfies equal opportunity on the given dataset, i.e.,* $E_{\mathrm{FN},T_0}(\tau_0) = E_{\mathrm{FN},T_1}(0) = \mathrm{C}(Q_0, Q_1)$.
- *(Accuracy and Fairness on ideal data) The Chernoff exponent of the probability of error of the Bayes optimal detector on the ideal distributions, i.e.,* $\mathrm{C}(\widetilde{P}_0, \widetilde{P}_1) = \mathrm{C}(Q_0, Q_1)$, *and is hence greater than* $\mathrm{C}(P_0, P_1)$.

The proof is provided in Appendix C. The first criterion demonstrates that one can always find ideal distributions such that the *fair* detector with respect to the given distributions (see Lemma 5) is in fact the Bayes optimal detector with respect to the ideal distributions. Note that there exist multiple pairs of (v, w) such that $\widetilde{P}_0(x) = \frac{P_0(x)^{(1-w)}P_1(x)^w}{\sum_x P_0(x)^{(1-w)}P_1(x)^w}$ and $\widetilde{P}_1(x) = \frac{P_0(x)^{(1-v)}P_1(x)^v}{\sum_x P_0(x)^{(1-v)}P_1(x)^v}$ satisfy the first criterion of the theorem.

The second criterion goes a step further and demonstrates that among such pairs of ideal distributions, one can always find at least one pair such that they are just as separable as the privileged group (i.e.,$C(\widetilde{P}_0, \widetilde{P}_1) = C(Q_0, Q_1)$). The Bayes optimal detector for the unprivileged group with respect to the ideal distributions, i.e., $\widetilde{T_0}(x) = \log \frac{\widetilde{P}_1(x)}{\widetilde{P}_0(x)} \geq 0$ is thus not only *fair* on the given dataset but also satisfies equal opportunity on the ideal data because its Chernoff exponent of FNR is also equal to that of the privileged group, i.e., $C(Q_0, Q_1)$. Note that, in order to satisfy the second criterion, we restrict ourselves to choosing $v = 1$ which leads to an appropriate value of w.

Remark 6 (Uniqueness)**.** *Theorem 2 provides a proof of existence of ideal distributions along with an explicit construction. In general, there may exist other pairs of distributions, which are not of the particular form mentioned in Theorem 2, but might satisfy the two conditions of the theorem. Therefore, given only* $P_0(x)$ *and* $P_1(x)$, *the ideal distributions are not necessarily unique unless further assumptions are made about their desirable properties.*

In order to go about finding such ideal distributions in practice, we therefore propose an additional desirable property of such an ideal dataset. We require the ideal dataset to be a useful representative of the given dataset. This motivates a constraint that $\pi_0 \mathrm{D}(\widetilde{P}_0 || P_0) + \pi_1 \mathrm{D}(\widetilde{P}_1 || P_1)$ be as small as possible, i.e., the KL divergences of the ideal distributions from their respective given real-world distributions are small. Building on this perspective, we formulate the following optimization for specifying two ideal distributions $\widetilde{P}_0$ and $\widetilde{P}_1$ for the unprivileged group:

$$\min_{\widetilde{P}_0, \widetilde{P}_1} \pi_0 \mathrm{D}(\widetilde{P}_0 || P_0) + \pi_1 \mathrm{D}(\widetilde{P}_1 || P_1)$$
$$\text{such that, } E_{\mathrm{FN},\widetilde{T_0}}(0) = \mathrm{C}(Q_0, Q_1), \tag{4}$$

where $\widetilde{T_0}(x) = \log \frac{\widetilde{P}_1(x)}{\widetilde{P}_0(x)} \geq 0$ is the Bayes optimal detector with respect to the ideal distributions and $E_{\mathrm{FN},\widetilde{T_0}}(0)$ is the Chernoff exponent of the probability of false negative for this detector when evaluated on the given distributions $P_0(x)$ and $P_1(x)$. Theorem 2 already shows that the aforementioned optimization is feasible.

The results of this subsection can be extended to optimization (2), or to other measures of fairness altogether, e.g., statistical parity, or to other kinds of constraints such as minimal individual distortion.

Relation to the construct space: The ideal distributions for the unprivileged group, in conjunction with the given distributions of the privileged group, have two interpretations: (i) They could be viewed as plausible distributions in the observed space if the mappings were unbiased from a separability standpoint (recall Definition 4). (ii) Given our limited knowledge of the construct space, they could also be viewed as candidate distributions in the construct space itself if the mappings for the group $Z = 1$ were identity mappings. This can be justified because we do not have much knowledge about the construct space (or even its dimensionality) except through the observed data. It is not unfathomable to assume they would have a separability of at least $\mathrm{C}(Q_0, Q_1)$, which is the separability exhibited by the privileged group in the observed space. Theorem 2 thus also demonstrates that the construct space is non-empty.

Remark 7 (Explicit Use of an Ideal Dataset)**.** *Several existing methods (Calmon et al., 2018; Feldman et al., 2015; Kamiran & Calders, 2012) propose pre-processing the given dataset to generate an alternate dataset that satisfies certain fairness and utility (representation) properties, in the same spirit as optimization (4), and train models on them. The trained detector may be sub-optimal with respect to the given dataset but is deemed to be fair. The results in this subsection help to explain why these approaches result in an accuracy-fairness trade-off on the given dataset, and also demonstrate that both accuracy and fairness can improve simultaneously when the accuracy is measured with respect to the alternate/ideal dataset. Optimization (4) is also reminiscent of the formulation of Jiang & Nachum (2019), who posit that a given biased label function is closest to an ideal unbiased label function in terms of KL divergence. In that work however, the KL divergence is applied to conditional label distributions* $p_{Y|X}$ *as opposed to conditional feature distributions* $p_{X|Y}$. *Furthermore, Jiang & Nachum (2019)*

do not analytically characterize trade-offs.

Remark 8 (Implicit Use of an Ideal Dataset)**.** *Existing methods that fall in this category include training with fairness regularization in the loss function or post-processing the output to meet a fairness criterion. Instead of explicitly generating an ideal dataset, these methods aim to find a classifier that satisfies a fairness criterion on the given dataset, with minimal compromise of accuracy on the given dataset (recall optimizations (2) and (3)). Here, we show that there exist ideal distributions corresponding to these fair detectors such that a sub-optimal detector on the given dataset can be optimal with respect to the ideal dataset.*

3.3. Active Data Collection: Alleviating Real-World Trade-Offs with Improved Knowledge

The inherent limitation of disparate separability between groups in the given dataset, discussed in Section 3.1, can in fact be overcome but with an associated cost: active data collection. In this section, we demonstrate when gathering more features can help in improving the Chernoff information of the unprivileged group without affecting that of the privileged group. Gathering more features helps us classify members of the unprivileged group more carefully with additional separability information that was not present in the initial dataset. In fact, this is the idea behind active fairness (Noriega-Campero et al., 2019; Bakker et al., 2019; Chen et al., 2018). Our analysis below also serves as a technical explanation for the success of active fairness.

Let X' denote the additional features so that (X, X') is now used for classification of the group $Z{=}0$. Note that X' could also easily be other forms of additional information including extra explanations to go along with the data or decision, similar to Varshney et al. (2018). Let (X, X') have the following distributions: $(X, X')|_{Y=0,Z=0} \sim W_0(x, x')$ and $(X, X')|_{Y=1,Z=0} \sim W_1(x, x')$, where Y is the true label. Note that, $P_0(x) = \sum_{x'} W_0(x, x')$ and $P_1(x) = \sum_{x'} W_1(x, x')$. Our goal is to derive the conditions under which the separability improves with addition of more features, i.e., $\mathrm{C}(W_0, W_1) > \mathrm{C}(P_0, P_1)$.

Theorem 3 (Improving Separability)**.** *The Chernoff information $\mathrm{C}(W_0, W_1)$ is strictly greater than $\mathrm{C}(P_0, P_1)$ if and only if X' and Y are not independent of each other given X and $Z = 0$, i.e., the conditional mutual information $I(X'; Y|X, Z = 0) > 0$.*

The proof is provided in Appendix D. Note that, in general $\mathrm{C}(W_0, W_1) \geq \mathrm{C}(P_0, P_1)$ because separability can only improve or remain the same (see Appendix D). We identify the scenario where the inequality is strict.

Let x' be a deterministic function of x, i.e., $f(x)$. Then $W_0(x, x'){=}P_0(x)$ if $x'{=}f(x)$, and 0 otherwise. Similarly, $W_1(x, x'){=}P_1(x)$ if $x'{=}f(x)$, and 0 otherwise, leading to $\mathrm{C}(W_0, W_1){=}\mathrm{C}(P_0, P_1)$. This agrees with the intuition that if X' is fully determined by X, then it does not improve the separability beyond what one could achieve using X alone. Therefore, for $\mathrm{C}(W_0, W_1){>}\mathrm{C}(P_0, P_1)$, we require X' to contribute some information that helps in separating hypotheses $Y = 0$ and $Y = 1$ better, that essentially leads to X' not being independent of Y given X and $Z = 0$.

If new data improves the separability of the group $Z = 0$, its accuracy-fairness trade-off is alleviated (see Fig. 3 in Section 4). New ideal distributions can also be found using the techniques of Section 3.2 that are more plausible as either candidate observed-space distributions under unbiased mappings or construct-space distributions. The new ideal distributions will also have better separability if the new data improves the separability of both groups.

4. Numerical Example

We use a simple numerical example to show how our theoretical concepts and results can be computed in practice.

Example 1. *Let the exam score for $Z = 0$ be $P_0(x){\sim}\mathcal{N}(1, 1)$ and $P_1(x){\sim}\mathcal{N}(4, 1)$, and that for $Z = 1$ be $Q_0(x){\sim}\mathcal{N}(0, 1)$ and $Q_1(x){\sim}\mathcal{N}(4, 1)$.*

Let us restrict ourselves to likelihood ratio detectors of the form $T_0(x) = \log \frac{P_0(x)}{P_1(x)} \geq \tau_0$ and $T_1(x) = \log \frac{Q_0(x)}{Q_1(x)} \geq \tau_1$ for the two groups. The log generating functions for $Z = 1$ can be computed analytically as: $\Lambda_0(u)_{z=1} = 8u(u{-}1)$ and $\Lambda_1(u)_{z=1} = 8u(u{+}1)$ (derivation in Appendix A.3) and the Chernoff information can be computed as $\mathrm{C}(Q_0, Q_1) = 2$.

Now, for the unprivileged group $Z = 0$, the log generating functions can be computed as $\Lambda_0(u)_{z=0} = \frac{9}{2}u(u - 1)$ and $\Lambda_1(u)_{z=0} = \frac{9}{2}u(u + 1)$ (again see Appendix A.3 for derivation). The Chernoff information is $\mathrm{C}(P_0, P_1) = 9/8$.

Accuracy-Fairness Trade-off in Real World: We restrict the detector for the privileged group to be the Bayes optimal detector $T_1(x){=}\log \frac{Q_1(x)}{Q_0(x)} \geq 0$ (equivalent to $x \geq 2$). For this detector, $E_{\mathrm{FP},T_1}(0){=}E_{\mathrm{FN},T_1}(0) = \mathrm{C}(Q_0, Q_1) = 2$.

Now, for $Z{=}0$, the Bayes optimal detector $T_0(x){=}\log \frac{P_1(x)}{P_0(x)}{\geq}0$ (or, $x{\geq}1.5$) will be unfair since $E_{\mathrm{FN},T_0}(0){=}\mathrm{C}(P_0, P_1){<}E_{\mathrm{FN},T_1}(0)$. Using the geometric interpretation of Chernoff information (recall Fig. 2), we can compute the Chernoff exponents of FPR and FNR, i.e., $E_{\mathrm{FP},T_0}(\tau_0)$ and $E_{\mathrm{FN},T_0}(\tau_0)$ as the negative of the y-intercept of the tangents to $\Lambda_0(u)_{z=0}$ and $\Lambda_1(u)_{z=0}$ for detectors $T_0(x){=}\log \frac{P_1(x)}{P_0(x)}{\geq}\tau_0$. This enables us to numerically plot the trade-off between accuracy ($E_{e,T_0}(\tau_0){=}\min\{E_{\mathrm{FP},T_0}(\tau_0), E_{\mathrm{FN},T_0}(\tau_0)\}$) and fairness ($|E_{\mathrm{FN},T_0}(\tau_0){-}E_{\mathrm{FN},T_1}(\tau_0)|$) by varying τ_0 as shown by the blue curve in Fig. 3.

Note that, the detector that satisfies fairness (equal op-

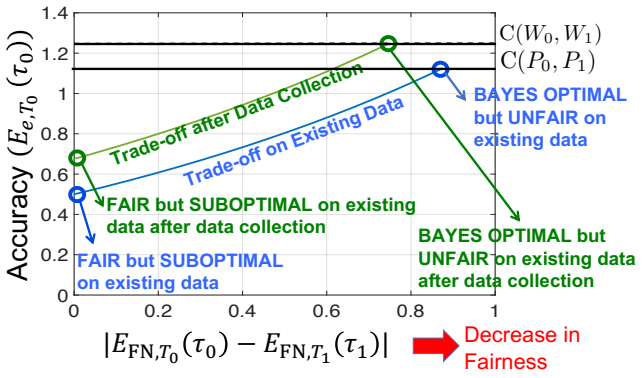

Figure 3. Computation of the trade-off between fairness and accuracy using a numerical example: For the unprivileged group, we let $P_0(x) \sim \mathcal{N}(1,1)$ and $P_1(x) \sim \mathcal{N}(4,1)$. We restrict the detector of the privileged group to its Bayes optimal detector with $C(Q_0, Q_1) = 2$. The blue curve denotes the trade-off between accuracy and fairness in the existing dataset for the unprivileged group. Now suppose we are able to collect an additional feature X' for the unprivileged group such that $(X, X')|_{Y=0,Z=0} \sim \mathcal{N}((1,1), \mathbf{I})$ and $(X, X')|_{Y=1,Z=0} \sim \mathcal{N}((4,2), \mathbf{I})$, where $\mathbf{I}$ is the 2×2 identity matrix. The green curve shows how active data collection alleviates the trade-off between fairness and accuracy.

portunity) on the given distributions can also be computed analytically as $\log \frac{P_1(x)}{P_0(x)} \geq \tau_0^*$ where $\tau_0^* = -3/2$ (equivalent to $x \geq 2$). This leads to equal exponent of FNR, i.e., $E_{\text{FN},T_0}(-3/2) = 2 = E_{\text{FN},T_1}(0)$ but for this detector $E_{\text{FP},T_0}(\tau_0^*) = 1/2$ leading to reduced Chernoff exponent of overall error probability (represents accuracy), i.e., $E_{e,T_0}(\tau_0^*) = \min\{E_{\text{FP},T_0}(\tau_0^*), E_{\text{FN},T_0}(\tau_0^*)\} = \min\{1/2, 2\} = 1/2$ which is less than $\mathrm{C}(P_0, P_1) = 9/8$.

Ideal Distributions: We refer to Fig. 4. It turns out that one pair of ideal distributions prescribed by Theorem 2 is $\widetilde{P}_0 = Q_0$ and $\widetilde{P}_1 = P_1 = Q_1$. The Bayes optimal detector with respect to the ideal distributions for $Z = 0$ is given by $\log \frac{\widetilde{P}_1(x)}{\widetilde{P}_0(x)} \geq 0$ (equivalent to $x \geq 2$). Note that, this is equivalent to the detector $\log \frac{P_1(x)}{P_0(x)} \geq \tau_0^*$ where $\tau_0^* = -3/2$ which satisfied equal opportunity on the given dataset. This detector is now Bayes optimal with respect to the ideal distributions $\widetilde{P}_0$ and $\widetilde{P}_1$, and has a Chernoff exponent of the overall probability of error equal to $C(\widetilde{P}_0, \widetilde{P}_1) = 2$ when measured with respect to the ideal distributions. Thus, we demonstrate that both fairness (in the sense of equal opportunity on existing dataset as well as ideal dataset) and accuracy (with respect to the ideal distributions) are in accord. Note that, one may also find alternate pairs of ideal distributions using optimization (4) or any variant of the optimization, e.g., using statistical parity.

Active Data Collection: Now suppose we are able to collect an additional feature X' for $Z = 0$ such that

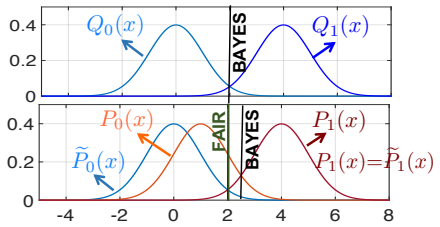

Figure 4. (Top) For the distributions in Example 1, we denote the Bayes optimal detector $\log \frac{Q_1(x)}{Q_0(x)} \geq 0$ (equivalent to $x \geq 2$) for the privileged group $Z = 1$. (Bottom) For $Z = 0$, the optimal detector $\log \frac{P_1(x)}{P_0(x)} \geq 0$ does not satisfy equal opportunity on the given dataset but a sub-optimal detector does (notice the equal area corresponding to false negative rate for two groups). However, there exist ideal distributions given by $\widetilde{P}_0 = Q_0$ and $\widetilde{P}_1 = P_1 = Q_1$ such that this detector is optimal w.r.t. the ideal distributions, and also achieves fairness w.r.t. both existing and ideal distributions.

$(X, X')|_{Y=0,Z=0} \sim \mathcal{N}((1,1), \mathbf{I})$ and $(X, X')|_{Y=1,Z=0} \sim \mathcal{N}((4,2), \mathbf{I})$, where $\mathbf{I}$ is the 2×2 identity matrix. The log generating functions can be derived as: $\Lambda_0(u) = 5u(u-1)$ and $\Lambda_1(u) = 5u(u+1)$. Note that, the Chernoff information (separability) $C(W_0, W_1) = 5/4$ which is greater than $C(P_0, P_1) = 9/8$. Thus, the collection of the new feature has improved the separability of the unprivileged group.

Now, we examine the effect of active data collection on the accuracy-fairness trade-off in the real world. We again refer to Fig. 3 (green curve). Consider the likelihood ratio detector for $Z = 0$ based on the total set of features, i.e., $T_0(x, x') = \log \frac{W_0(x,x')}{W_1(x,x')} \geq \tau_0$. To satisfy our fairness constraint, we need to choose a τ_0^* such that $E_{\text{FN},T_0}(\tau_0^*) = E_{\text{FN},T_1}(0) = \mathrm{C}(Q_0, Q_1) = 2$. Upon solving, we obtain that $\tau_0^* = 5 - \sqrt{40} \approx -1.32$. For this value of τ_0^*, we obtain $E_{\text{FP},T_0}(\tau_0^*) = 7 - \sqrt{40}) \approx 0.68$. The Chernoff exponent of the probability of error for this *fair* detector is given by $\min\{E_{\text{FN},T_0}(\tau_0^*), E_{\text{FP},T_0}(\tau_0^*)\} = \min\{2, 0.68\} = 0.68$ which is greater than 0.5 (the Chernoff exponent of the probability of error for the fair detector before collection of the additional feature X').

5. Conclusion

Our results provide novel analytical insights that explain and characterize accuracy-fairness trade-offs on real datasets. Our Chernoff information based analysis can help quantify the separability of a dataset, even before any classification algorithm is applied. We believe that our demonstration that fairness and accuracy are in accord with respect to ideal datasets will motivate the use of accuracy with respect to an ideal dataset as a performance metric in algorithmic fairness research (Sharma et al., 2020; Wick et al., 2019). Lastly, our results also inform how and when active data collection can alleviate the trade-off in the real world.

Acknowledgements

The authors would like to thank Pulkit Grover, Shubham Sharma, and the anonymous reviewers for their valuable suggestions.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *Proceedings of the International Conference on Machine Learning*, pp. 60–69, 2018.

Bakker, M. A., Noriega-Campero, A., Tu, D. P., Sattigeri, P., Varshney, K. R., and Pentland, A. S. On fairness in budget-constrained decision making. In *Proceedings of the KDD Workshop on Explainable Artificial Intelligence*, 2019.

Berend, D. and Kontorovich, A. A finite sample analysis of the naive Bayes classifier. *Journal of Machine Learning Research*, 16:1519–1545, 2015.

Berisha, V., Wisler, A., Hero, A. O., and Spanias, A. Empirically estimable classification bounds based on a nonparametric divergence measure. *IEEE Transactions on Signal Processing*, 64(3):580–591, 2015.

Bhattacharyya, A. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, 7(4):401–406, 1946.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. The balanced accuracy and its posterior distribution. In *Proceedings of the International Conference on Pattern Recognition*, pp. 3121–3124, 2010.

Calmon, F. P., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017.

Calmon, F. P., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1106–1119, 2018.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 319–328, 2019.

Chen, I. Y., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pp. 3539–3550, 2018.

Côté, F. D., Psaromiligkos, I. N., and Gross, W. J. A Chernoff-type lower bound for the Gaussian Q-function. *arXiv preprint arXiv:1202.6483*, 2012.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, 2012.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, pp. 119–133, 2018.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

Gallager, R. Detection, decisions, and hypothesis testing. http://web.mit.edu/gallager/www/papers/chap3.pdf, 2012.

Garg, S., Kim, M. P., and Reingold, O. Tracking and improving information in the service of fairness. In *Proceedings of the ACM Conference on Economics and Computation*, pp. 809–824, 2019.

Ghassami, A., Khodadadian, S., and Kiyavash, N. Fairness in supervised learning: An information theoretic approach. In *Proceedings of the IEEE International Symposium on Information Theory*, pp. 176–180, 2018.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pp. 513–520, 2007.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Jiang, H. and Nachum, O. Identifying and correcting label bias in machine learning. *arXiv preprint arXiv:1901.04966*, 2019.

Kailath, T. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, 15(1):52–60, 1967.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pp. 656–666, 2017.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Lee, Y. and Sung, Y. Generalized Chernoff information for mismatched Bayesian detection and its application to energy detection. *IEEE Signal Processing Letters*, 19 (11):753–756, 2012.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, pp. 107–118, 2018.

Motwani, R. and Raghavan, P. *Randomized Algorithms*. Cambridge University Press, 1995.

Noriega-Campero, A., Bakker, M. A., Garcia-Bulle, B., and Pentland, A. S. Active fairness in algorithmic decision making. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, pp. 77–83, 2019.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

Scott, C., Blanchard, G., and Handy, G. Classification with asymmetric label noise: Consistency and maximal denoising. In *Proceedings of the Conference On Learning Theory*, pp. 489–511, 2013.

Sharma, S., Zhang, Y., Aliaga, J. M. R., Bouneffouf, D., Muthusamy, V., and Varshney, K. R. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, pp. 358–364, 2020.

Varshney, K. R., Khanduri, P., Sharma, P., Zhang, S., and Varshney, P. K. Why interpretability in machine learning? An answer using distributed detection and data fusion theory. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*, pp. 15–20, 2018.

Wick, M., Panda, S., and Tristan, J.-B. Unlocking fairness: a trade-off revisited. In *Advances in Neural Information Processing Systems*, pp. 8780–8789, 2019.

Yeom, S. and Tschantz, M. C. Discriminative but not discriminatory: A comparison of fairness definitions under different worldviews. *arXiv preprint arXiv:1808.08619*, 2018.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the International Conference on World Wide Web*, pp. 1171–1180, 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of the International Conference on Machine Learning*, pp. 325–333, 2013.

Zhao, H. and Gordon, G. J. Inherent tradeoffs in learning fair representation. *arXiv preprint arXiv:1906.08386*, 2019.