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 al- most 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 ex- amine fair classiﬁcation through the lens of *mis- matched hypothesis testing*: trying to ﬁnd a clas- siﬁer that distinguishes between two ideal dis- tributions when given two mismatched distribu- tions that are biased. Using Chernoff informa- tion, a tool in information theory, we theoreti- cally demonstrate that, contrary to popular belief, there always exist ideal distributions such that op- timal fairness and accuracy (with respect to the ideal distributions) are achieved simultaneously: there is no trade-off. Moreover, the same clas- siﬁer 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 fundamen- tal 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 problem- atic to measure accuracy with respect to data that reﬂects bias, and instead, we should be consider- ing accuracy with respect to ideal, unbiased data.

# Introduction

This work addresses a fundamental question in the ﬁeld of algorithmic fairness ([Calmon et al.](#_bookmark42), [2017](#_bookmark42); [Dwork et al.](#_bookmark32),

1IBM Research 2Department of Electrical and Computer Engi- neering, Carnegie Mellon University. Correspondence to: Sang- hamitra Dutta *<*[sanghamd@andrew.cmu.edu](mailto:sanghamd@andrew.cmu.edu)*>*.

*Proceedings of the 97 th International Conference on Machine Learning*, Online, PMLR 119, 2020. Copyright 2020 by the au- thor(s).

[2012](#_bookmark32); [Agarwal et al.](#_bookmark29), [2018](#_bookmark29); [Hardt et al.](#_bookmark46), [2016](#_bookmark46); [Ghassami](#_bookmark43) [et al.](#_bookmark43), [2018](#_bookmark43); [Kusner et al.](#_bookmark55), [2017](#_bookmark55); [Kilbertus et al.](#_bookmark53), [2017](#_bookmark53); [Zemel et al.](#_bookmark56), [2013](#_bookmark56)):

*Is there a trade-off between fairness and accuracy?*

The existence of this trade-off has been pointed out in sev- eral existing works ([Menon & Williamson](#_bookmark59), [2018](#_bookmark59); [Chen et al.](#_bookmark27), [2018](#_bookmark27); [Zhao & Gordon](#_bookmark57), [2019](#_bookmark57)) that also propose different theoretical approaches to characterize it. Yet, it is not preor- dained as to why such a trade-off should exist between fair- ness and accuracy. For instance, [Friedler et al.](#_bookmark39) ([2016](#_bookmark39)) and [Yeom & Tschantz](#_bookmark51) ([2018](#_bookmark51)) 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 be- tween accuracy and fairness (in particular, equal opportu- nity ([Hardt et al.](#_bookmark46), [2016](#_bookmark46))) 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” distribu- tions, 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 ob- served dataset, we quantify the “separability” into positive and negative class labels using Chernoff information, an information-theoretic approximation to the best exponent of

the probability of error in binary classiﬁcation. We demon- strate (in Theorem [1](#_bookmark14)) 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 exist- ing works ([Menon & Williamson](#_bookmark59), [2018](#_bookmark59); [Chen et al.](#_bookmark27), [2018](#_bookmark27)) 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](#_bookmark26) in Section [4](#_bookmark22)). 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 ac- cord:* Novel to this work, we examine the problem of fair classiﬁcation through the lens of mismatched hypothesis testing. We show (in Theorem [2](#_bookmark20)) 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 in- terpretation 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](#_bookmark17)).

*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, specif- ically through additional data collection. We derive an information-theoretic criterion (in Theorem [3](#_bookmark24)) 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.](#_bookmark61), [2019](#_bookmark61); [Bakker](#_bookmark33) [et al.](#_bookmark33), [2019](#_bookmark33); [Chen et al.](#_bookmark27), [2018](#_bookmark27)) 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.](#_bookmark41) ([2019](#_bookmark41)), [Menon & Williamson](#_bookmark59) ([2018](#_bookmark59)), [Chen](#_bookmark27) [et al.](#_bookmark27) ([2018](#_bookmark27)), and [Zhao & Gordon](#_bookmark57) ([2019](#_bookmark57)), have also used in-

formation 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](#_bookmark16)). This enables us to numerically compute the accuracy-fairness trade-off (Fig. [3](#_bookmark26)), 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 ex- actly compute it, motivating us to introduce the additional tool of Chernoff information to do so approximately. Fur- thermore, 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.](#_bookmark49) ([2019](#_bookmark49)) and [Sharma et al.](#_bookmark64) ([2020](#_bookmark64)) further elucidate the signiﬁcance of Theorem [2](#_bookmark20) 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.](#_bookmark49) ([2019](#_bookmark49)) and [Sharma et al.](#_bookmark64) ([2020](#_bookmark64)), clari- fying 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.](#_bookmark44), [2018](#_bookmark44); [Feld-](#_bookmark36) [man et al.](#_bookmark36), [2015](#_bookmark36); [Zemel et al.](#_bookmark56), [2013](#_bookmark56)). Here, our goal is not to propose another competing strategy of fairness through pre-processing. Instead, our focus is to theoretically demon- strate that there exists an ideal dataset such that a fair classi- ﬁer 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.](#_bookmark44) ([2018](#_bookmark44))).

Our tools share similarities with [Varshney et al.](#_bookmark65) ([2018](#_bookmark65)) (that demonstrates how explainability can improve Chernoff in- formation), as well as the theory of hypothesis testing in general ([Lee & Sung](#_bookmark58), [2012](#_bookmark58); [Cover & Thomas](#_bookmark30), [2012](#_bookmark30)). Our contribution lies in using these tools in fair machine learn- ing, where they have not been used to the best of our knowl- edge (e.g., in the previous analyses of [Menon & Williamson](#_bookmark59) ([2018](#_bookmark59)); [Zhao & Gordon](#_bookmark57) ([2019](#_bookmark57)); [Chen et al.](#_bookmark27) ([2018](#_bookmark27))).

**Remark 1** (Population Setting)**.** *In this work, we operate in the population setting (motivated from* [*Gretton et al.*](#_bookmark45) *(*[*2007*](#_bookmark45)*);* [*Ravikumar et al.*](#_bookmark62) *(*[*2009*](#_bookmark62)*);* [*Scott et al.*](#_bookmark63) *(*[*2013*](#_bookmark63)*)), 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*](#_bookmark30) *(*[*2012*](#_bookmark30)*)).*

# 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](#_bookmark51) ([2018](#_bookmark51)) and [Friedler et al.](#_bookmark39) ([2016](#_bookmark39)), 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](#_bookmark51) ([2018](#_bookmark51)).[1](#_bookmark2) The observed features are

derived from features in the construct space as follows: X = f*Y,Z*(X*a*) where f*Y,Z*( ) 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*~*P0(x) and

X*|Y* =1*,Z*=0*~*P1(x). Similarly, X*|Y* =0*,Z*=1*~*Q0(x) and

X*|Y* =1*,Z*=1*~*Q1(x). For each group Z = z, we will be denoting classiﬁers as T*z*(x) *>* τ*z*, i.e., the prediction label is 1 when T*z*(x) *>* τ*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.*](#_bookmark32)*,* [*2012*](#_bookmark32)*;* [*2018*](#_bookmark34)*)). Fur-*

*thermore, 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 Con-

tinuity: P (x), P (x), Q (x) and Q (x) are greater than

We let PFP*,Ts* (τ*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., PFP*,Ts* (τ*z*) = Pr (T*z*(X) τ*z* Y = 0, Z = z). Sim- ilarly, PFN*,Ts* (τ*z*) is the probability of false negative (wrongful rejection of positive class labels; also called

*> |*

false negative rate (FNR)), given by: PFN*,Ts* (τ*z*) = Pr (T*z*(X) < τ*z* Y = 1, Z = z). The overall probability of error of a group is given by: P*e,Ts* (τ*z*) = π0PFP*,Ts* (τ*z*) + π1PFN*,Ts* (τ*z*), where π0 and π1 are the prior probabilities of Y = 0 and Y = 1 given Z = z. For the sake of simplic- ity, we consider the case where π0 = π1 = 1 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

*|*

2

unequal priors in Appendix E. Equal priors also correspond to the balanced accuracy measure ([Brodersen et al.](#_bookmark40), [2010](#_bookmark40)) which is often favored over ordinary accuracy.

A well-known deﬁnition of fairness is *equalized odds* ([Hardt](#_bookmark46) [et al.](#_bookmark46), [2016](#_bookmark46)), 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., sta- tistical parity ([Agarwal et al.](#_bookmark29), [2018](#_bookmark29)).

We assume that in the construct space, there is no trade-off between accuracy and equal opportunity, i.e., the Bayes optimal ([Cover & Thomas](#_bookmark30), [2012](#_bookmark30)) 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 PFP*,Ts* (τ*z*) and *-* log PFN*,Ts* (τ*z*). Often, we may

0 1 0 1

0 everywhere in range of x. This ensures that likelihood ratio detectors such as log *P*1(*x*) *>* τ0 and Kullback-Leibler

*P*0(*x*)

(KL) divergences between any two of these distributions are well-deﬁned. (**A2**) Distinct Hypotheses: D(P0*||*P1),

D(P1*||*P0), D(Q0*||*Q1) and D(Q1*||*Q0) are strictly greater

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](#_bookmark4); proof in Appendix A.1), that is known to be pretty tight (see Remark [3](#_bookmark5) and also [Motwani](#_bookmark60) [& Raghavan](#_bookmark60) ([1995](#_bookmark60)); [Berend & Kontorovich](#_bookmark35) ([2015](#_bookmark35))).

than 0, where D(*.||.*) is the KL divergence.

1This is consistent with the “What You See Is What You Get” worldview in [Yeom & Tschantz](#_bookmark51) ([2018](#_bookmark51)) where label bias can be

**Deﬁnition 1.** *The Chernoff exponents of* P PFN*,Ts* (τ*z*) *are deﬁned as:*

FP*,Ts*

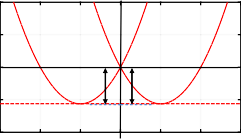
(τ*z*) *and*

ignored and our chosen measure of fairness, i.e., equal opportunity is justiﬁed as a measure of fairness.

EFP*,Ts* (τ*z*) = sup(uτ*z* Λ0(u)), *and*

*u>*0

*-*

2 2 2 2



五FP

EFN

）

〇(u

⇤

)

l(u

⇤



**log-generating function**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  | 历 | FN | 历 | FP |  |
|  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  | E | FN | E | FP |  |
|  |  |  |  |  |  |

1 1

0 0

-1 -1

-2 u -2

1 1

0 0

EFN E

-1

FP

-1

-2 -2

-1.5 -1 -0.5 0 0.5 1 1.5

-1.5 -1 -0.5

u 0 0.5 1 1.5

-1.5 -1 -0.5

u 0 0.5 1 1.5

-1.5 -1 -0.5 u 0 0.5 1 1.5

*Figure 1.* Let *P*0(*x*)*～N* (1*,* 1) and *P*1(*x*)*～N* (4*,* 1). For a likelihood ratio detector *T* (*x*)= log *P*1(*x*) *> τ* , we can compute the log-

*P*0(*x*)

generating functions as follows: Λ0(*u*) = 9 *u*(*u \_* 1) and Λ1(*u*) = 9 *u*(*u* + 1) (derived in Appendix A.3). Note that, Λ0(*u*) is strictly

2

2

convex with zeros at *u*=0 and *u*=1, and Λ1(*u*)=Λ0(*u* + 1). We obtain *E*FP*,T* (*τ* ) and *E*FN*,T* (*τ* ) as the negative of the y-intercepts for tangents to Λ0(*u*) and Λ1(*u*) respectively with slope *τ* . As we vary the slope of the tangent (*τ* ), there is a trade-off between *E*FP*,T* (*τ* ) and *E*FN*,T* (*τ* ) until they both become equal at *τ* = 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: C(*P*0*, P*1):=*E*FP*,T* (0)=*E*FN*,T* (0)*,* which is equal to 9*/*8 for this particular example.

EFN*,Ts* (τ*z*) = sup(uτ*z* Λ1(u)).

*-*

*u<*0

*Here,* Λ0(u) *and* Λ1(u) *are called log-generating func- tions, given by* Λ0(u) = log E[e*uTs* (*X*)*|*Y = 0, Z = z] *and* Λ1(u) = log E[e*uTs* (*X*)*|*Y = 1, Z = z].

*bility of error* P*e,Ts* (τ*z*) *is deﬁned as:*

E*e,Ts* (τ*z*) = min*{*EFP*,Ts* (τ*z*), EFN*,Ts* (τ*z*)*}*.

Recall that, under equal priors, we have P*e,Ts* (τ*z*) =

1 PFP*,T* (τ*z*) + 1 PFN*,T* (τ*z*). The exponent of P*e,T* (τ*z*)

**Lemma 1** (Chernoff Bound)**.** *The exponents satisfy:* 2 *s* 2 *s s*

PFP*,T* (τ*z* )*<*e*-E*FP*,Ts* (*τs* ) *and* PFN*,T* (τ*z* )*<*e*-E*FN*,Ts* (*τs* ).

is dominated by the minimum of the error exponents of

P

*s s* (τ ) and P (τ ), which in turn is bounded by

**Remark 3** (Tightness of the Chernoff Bound)**.** *For Gaus-*

FP*,Ts z*

FN*,Ts z*

*sian 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 (*[*Coˆte´ et al.*](#_bookmark28)*,* [*2012*](#_bookmark28)*). The Bhattacharya bound (a special case of Chernoff bound) both upper and lower bounds the Bayes error exponent (*[*Berisha*](#_bookmark37)[*et al.*](#_bookmark37)*,* [*2015*](#_bookmark37)*;* [*Bhattacharyya*](#_bookmark38)*,* [*1946*](#_bookmark38)*;* [*Kailath*](#_bookmark50)*,* [*1967*](#_bookmark50)*).*

**Geometric Interpretation of Chernoff Exponents:** Cher- noff 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](#_bookmark3) where we il- lustrate 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](#_bookmark7), i.e., E[T*z*(X) Y =1, Z=z]>0 and E[T*z*(X) Y =0, Z=z]<0, then Λ0(u) and Λ1(u) are strictly convex and attain their minima on either sides

*|*

*|*

the minimum of the Chernoff exponents of FPR and FNR

(Deﬁnition [1](#_bookmark1)). A higher E*e,Ts* (τ*z*) indicates higher accuracy, i.e., lower P*e,Ts* (τ*z*). To understand this, ﬁrst consider likeli- hood ratio detectors of the form T0(x) = log *P*1(*x*) for Z =

*P*0(*x*)

0. As we vary τ0, there is a trade-off between PFP*,T*0 (τ0) and PFN*,T*0 (τ0), i.e., as one increases, the other decreases. A similar trade-off is also observed in their Chernoff exponents

(see Fig. [1](#_bookmark3)). P*e,T*0 (τ0) is minimized when τ0 = 0 (for equal priors) and PFP*,T*0 (0)=PFN*,T*0 (0). For this optimal value of τ0 = 0, the Chernoff exponents of FPR and FNR also become equal, i.e., EFP*,T*0 (0)=EFN*,T*0 (0), and the max- imum value of E*e,T*0 (τ0)= min EFP*,T*0 (τ0), EFN*,T*0 (τ0) is attained. This exponent is called the Chernoff informa-

*{ }*

tion ([Cover & Thomas](#_bookmark30), [2012](#_bookmark30)). For completeness, we include a well-known result on Chernoff information from [Cover &](#_bookmark30) [Thomas](#_bookmark30) ([2012](#_bookmark30)) with the proof in Appendix A.4.

**Lemma 2.** *For two hypotheses* P0(x) *under* Y = 0 *and* P1(x) *under* Y = 1*, the Chernoff exponent of the probabil- ity of error of the Bayes optimal classiﬁer is given by the*

of the origin. The Chernoff exponents E

FP*,Ts*

(τ*z*) and

*Chernoff information*[3](#_bookmark8)*:*

EFN*,Ts* (τ*z*) can be obtained as the negative of the y- intercepts for tangents to Λ0(u) and Λ1(u) with slope τ*z* (for τ *e* (E[T (X)*|*Y =0, Z=z], E[T (X)*|*Y =1, Z=z])).

*z*

*z*

*z*

C(P0, P1) = min log P0(x)

*- u* l

*∈*(0*,*1) *x*

1*-u*

P1(x)

*u*、. (1)

**Deﬁnition 2.** *The Chernoff exponent of the overall proba-*

2For a detector *Tz* (*x*) *τz,* we would expect *Tz* (*X*) to be high when *Y* =1, and low when *Y* =0 justifying the crite- ria E[*Tz* (*X*) *Y* =1*, Z*=*z*]*>*0 and E[*Tz* (*X*) *Y* =0*, Z*=*z*]*<*0

*>*

*| |*

for being well-behaved. A likelihood ratio detector

**Goals:** Our metrics of interest for *accuracy* are E*e,T*0 (τ0) and E*e,T*1 (τ1) because a higher value of the Chernoff expo- nent of P*e,Ts* (τ*z*) implies a higher accuracy for the respec- tive groups Z=0 and Z=1. Our metric of interest for *fair-*

*ness* is the difference of the Chernoff exponents of FNR, i.e.,

*T*0(*x*)= log *P*1(*x*) *>τ*0 is well-behaved under assumption A2 in

*P*0(*x*)

Section [2](#_bookmark0) because we have E[*Tz* (*X*)*|Y* =1*, Z*=*z*]=*D*(*P*1*||P*0)

and E[*Tz* (*X*)*|Y* =0*, Z*=*z*]= *\_ D*(*P*0*||P*1).

3When *P*0(*x*) and *P*1(*x*) are continuous distributions, the sum- mation is replaced by an integral over *x* (see Appendix A.3).

EFN*,T*0 (τ0) EFN*,T*1 (τ1) (inspired from equal opportu- nity). A model is *fair* when EFN*,T*0 (τ0) EFN*,T*1 (τ1) = 0, and progressively becomes more and more unfair as this

*| - |*

*| - |*

quantity *|*EFN*,T*0 (τ0) *-* EFN*,T*1 (τ1)*|* increases.

Our ﬁrst goal is to quantify fundamental limits on the best accuracy-fairness trade-off in terms of our metrics of inter- est on an existing real-world dataset, i.e., given observed

distributions P0(x), P1(x), Q0(x), and Q1(x). Next, our goal is to ﬁnd ideal distributions where fairness and accu-

racy are in accord when accuracy is measured with respect to the ideal distributions.

# Main Results

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

Given the setup in Section [2](#_bookmark0), 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

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 oc- curs more commonly in practice, is where discrimination is caused due to an inherent limitation of the dataset: the map- pings 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 C(P0, P1) < C(Q0, Q1). Under this scenario, the Chernoff exponents of FNR of the Bayes optimal detectors for the two groups are C(P0, P1) and C(Q0, Q1) 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](#_bookmark10) (used in proof of Theorem [1](#_bookmark14); see Appendix B).

**Lemma 3.** *Let* C(P0, P1)<C(Q0, Q1)*. Suppose that there are two likelihood ratio detectors* T0(x) τ0 *and* T1(x) τ1*,*

*> >*

*one for each group, such that* EFN*,T* (τ0)=EFN*,T* (τ1).

0 1

us ﬁrst formally deﬁne our intuitive notion of separability.

**Deﬁnition 3.** *For a group of people with distributions* P0(x) *and* P1(x) *under hypotheses* Y =0 *and* Y =1*, we deﬁne the separability as their Chernoff information* C(P0, P1)*.*

Deﬁnition [3](#_bookmark11) is motivated from Lemma [2](#_bookmark6) because Chernoff information essentially provides an information-theoretic approximation to the best classiﬁcation accuracy (in an ex- ponent sense) for a group of people in a given dataset. Next, we deﬁne unbiased mappings from a separability standpoint.

*Then, at least one of the following statements is true:*

*(i)* E*e,T*0 (τ0) < C(P0, P1)*, or (ii)* E*e,T*1 (τ1) < C(Q0, Q1)*.*

The next two results show how current and reasonable ap- proaches to fair classiﬁcation can give rise to each of the two cases in Lemma [3](#_bookmark10). Consider the following optimization problem, where the goal is to ﬁnd classiﬁers of the form

T0(x) τ0 and T1(x) τ1 for the two groups that maxi- mize the Chernoff exponent of the probability of error under

*> >*

the constraint that they are *fair* on the given dataset.

**Deﬁnition 4.** *Consider the setup in Section* [*2*](#_bookmark0)*. The mapping*

X = f*Y,Z*(X*a*) *from the construct space to the observed* *space is said to be unbiased if* C(P0, P1) = C(Q0, Q1).

max

*T*0*,τ*0*,T*1*,τ*1

min *{*EFP*,T*0 (τ0), EFN*,T*0 (τ0),

EFP*,T*1 (τ1), EFN*,T*1 (τ1)*}*

Our next result demonstrates that the trade-off between fair- ness and accuracy arises due to a bias in the mappings from a separability standpoint, i.e., C(P0, P1) C(Q0, Q1). Be-

cause we assumed that Z = 0 is the unprivileged group, we

let C(P0, P1) be either equal to, or less than C(Q0, Q1).

**Theorem 1** (Explaining the Trade-Off)**.** *For the setup in Section* [*2*](#_bookmark0)*, one of the following is true:*

1. *Unbiased Mappings, i.e.,* C(P0, P1)=C(Q0, Q1)*: The*

*Bayes optimal detectors* T (x) *>* τ *and* T (x) *>* τ *for*

such that EFN*,T*0 (τ0) = EFN*,T*1 (τ1). (2)

This optimization is in the spirit of existing works ([Zafar](#_bookmark54) [et al.](#_bookmark54), [2017](#_bookmark54); [Agarwal et al.](#_bookmark29), [2018](#_bookmark29); [Donini et al.](#_bookmark31), [2018](#_bookmark31); [Celis](#_bookmark47) [et al.](#_bookmark47), [2019](#_bookmark47)) that maximize accuracy under fairness con- straints. 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 T0(x) and T1(x) to be likelihood ratio detectors of the form log *P*1(*x*)

*P*0(*x*)

and log *Q*1(*x*) , then ([2](#_bookmark13)) has a unique solution (τ *\**, τ *\**).

*Q* (*x*) 0 1

0

0 0 1 1

**Lemma 4.** *Let* C(P , P )<C(Q , Q ) *and* T (x) *and*

*the two groups with Chernoff exponents of the probability of error* C(Q0, Q1)(= C(P0, P1)) *also attain fairness,*

*i.e., |*EFN*,T* (τ0) *-* EFN*,T* (τ1)*|* = 0*.*

0 1 0 1 0

T1(x) *be restricted to be likelihood ratio detectors. Then the detectors* T0(x) *>* τ0*\* and* T1(x) *>* τ1*\* that solve the*

0 1 *optimization* ([2](#_bookmark13)) *are the Bayes optimal detector for the un-*

1. *Biased Mappings, i.e.,* C(P0, P1) < C(Q0, Q1)*: The Bayes optimal detectors* T0(x) *>* τ0 *and* T1(x) *>* τ1

*privileged group (*τ0*\** = 0*) and a sub-optimal detector for*

*the privileged group (*τ1*\** > 0*) with* E*e,T* (τ1*\**) < C(Q0, Q1)*.*

1

*for the two groups are not fair, i.e., |*E

FN*,T*0

(τ0) *-*

EFN*,T*1 (τ1)*|* = 0*. Furthermore, no likelihood ratio de-*

*tector can improve the Chernoff exponent of the probabil- ity of error for the unprivileged group beyond* C(P0, P1)*.*

As a proof sketch, we refer to Fig. [2](#_bookmark16) (Left). Let τ0*\** = 0, which ensures EFN*,T*0 (0) = EFP*,T*0 (0) = C(P0, P1).

4 4



⇤=(u)*Z*丰u

⇤0(u〉*之*二0

-

𝐸!",$! 𝜏% =

𝐸!",$"(𝜏&)0

𝐸!',$"(𝜏

1

&)

1之*2*～．

ī(u）*之*=ó

⇤．已

⇤



⇤0(u〉*)*=1

⇤0(0）*2*=0

𝐸!",$! 𝜏% =

-1

𝐸!",$"(𝜏&)

0

1

𝐸!',$!(𝜏%)

u〉*之*=1

(Z)*2*=0

⇤1(

⇤1

2 2

0 0

-2 -2

-4 -4

1

*Figure 2.* Let the distributions for the unprivileged group (*Z* = 0) be *P*0(*x*) (1*,* 1) and *P*1(*x*) (4*,* 1). Also, let the distributions of the privileged group be *Q*0(*x*) (0*,* 1) and *Q*1(*x*) (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

*～N ～N*

*～N ～N*

have Λ0(*u*)*z*=1 = 8*u*(*u \_* 1) and Λ1(*u*)*z*=1 = 8*u*(*u* + 1). Also, Λ0(*u*)*z*=0 = 9 *u*(*u \_* 1)*,* and Λ1(*u*)*z*=0 = 9 *u*(*u* + 1). Note that,

2

2

C(*P*0*, P*1)*<*C(*Q*0*, Q*1). (**Left**) This plot corresponds to the scenario of Lemma [4](#_bookmark15). The detector for the group *Z* = 0 is the Bayes optimal detector with *τ*0*\** = 0 and *E*FN*,T*0 (*τ*0*\**) = *E*FP*,T*0 (*τ*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*FN*,T*1 (*τ*1*\**) = *E*FN*,T*0 (*τ*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](#_bookmark19). The detector for the group *Z* = 1 is the Bayes optimal detector with *τ*1*\** = 0 and *E*FN*,T*1 (*τ*1*\**) = *E*FP*,T*1 (*τ*1*\**) = *C*(*Q*0*, Q*1)*.* In order to satisfy equal opportunity, we have to choose *τ*0*\** such that *E*FN*,T*0 (*τ*0*\**) = *E*FN*,T*1 (*τ*1*\**) = *C*(*Q*0*, Q*1) which is strictly greater that *C*(*P*0*, P*1). However, this threshold *τ*0*\** makes *E*FP*,T*0 (*τ*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 τ1*\** that will sat-

group with E*e,T*0 (τ0*\**) < C(P0, P1). The full proofs for

isfy EFN*,T*1 (τ1*\**)=EFN*,T*0 (0) is a τ1*\**>0 such that EFN*,T*1 (τ1*\**)=C(P0, P1)<C(Q0, Q1), and

hence EFP*,T*1 (τ1*\**)>C(Q0, Q1). This leads to, min*{*EFP*,T*0 (0), EFN*,T*0 (0), EFP*,T*1 (τ1*\**), EFN*,T*1 (τ1*\**)*}* = C(P0, P1).

For τ0*\**=0, either EFP*,T*0 (τ0*\**)<C(P0, P1)<EFN*,T*0 (τ0*\**),

or EFN*,T*0 (τ0*\**)<C(P0, P1)<EFP*,T*0 (τ0*\**), implying that, min*{*EFP*,T*0 (τ0*\**), EFN*,T*0 (τ0*\**), EFP*,T*1 (τ1*\**), EFN*,T*1 (τ1*\**)*}* < C(P0, P1).

This situation of reducing the accuracy of the privileged group is often interpreted as causing *active harm* to the priv- ileged 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 priv- ileged group (i.e., EFN*,T*1 (τ1) = EFP*,T*1 (τ1) = C(Q0, Q1) for the privileged group), but only vary the detector for the

unprivileged group to achieve fairness. We propose the following optimization:

max min EFP*,T* (τ0), EFN*,T* (τ0)

*{ }*0 0

*T*0*,τ*0

such that EFN*,T*0 (τ0) = C(Q0, Q1). (3)

Again, if we restrict T0(x) to be a likelihood ratio detector, then there exists a unique solution τ0*\** to optimization ([3](#_bookmark18)).

**Lemma 5.** *Let* T0(x) = log *P*1(*x*) *and we have* C(P0, P1) < C(Q0, Q1)*. The detector* T0(x) τ0*\* that solves optimization* ([3](#_bookmark18)) *is a sub-optimal detector for the unprivileged group with* E*e,T*0 (τ0*\**) < C(P0, P1)*.*

*>*

*P*0(*x*)

As a proof sketch, we refer to Fig. [2](#_bookmark16) (Right). If we choose

τ0*\* /*= 0, we get a sub-optimal detector for the unprivileged

Lemmas [4](#_bookmark15) and [5](#_bookmark19) 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.,* T0(x) 0 *and*

*>*

*>*

T1(x) 0 *do not satisfy statistical parity (*[*Agarwal et al.*](#_bookmark29)*,*

[*2018*](#_bookmark29)*), 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.*

### 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 de- tector that achieved fairness with sub-optimal accuracy in Lemma [5](#_bookmark19) 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*](#_bookmark0)*, let* C(P0, P1) < C(Q0, Q1)*. Let us choose*

*the Bayes optimal detector* T1(x) = log *Q*1(*x*) *>*0 *for the*

*Q*0(*x*)

are small. Building on this perspective, we formulate the following optimization for specifying two ideal distributions

*group* Z = 1*. Then, for group* Z = 0*, there exist* P←0(x) ← ←

*and*

*and* P←1(x) *of the form* P←0(x) =

1

*z P*0(*x*)(1*\_o*)*P*1(*x*)*o*

*P*0(*x*)(1*\_w*)*P*1(*x*)*w*

*z P*0(*x*)(1*\_w*)*P*1(*x*)*w*

←P (x) =

*P*←0*,P*←1

P0 and P1 for the unprivileged group:

*P*0(*x*)(1*\_o*)*P*1(*x*)*o*

← ← *>*0

←

~FN*,T*

*for* w, v *e R such that:*

min π0D(P←0*||*P0) + π1D(P←1*||*P1)

* *(Fairness on given data) The Bayes optimal detector for*

*·*

*P*0(*x*)

0

*the ideal distributions, i.e.,* T (x)= log *P*1(*x*) 0 *is equiv-*

*P*0(*x*)

*alent to the detector* T0(x) = log *P*1(*x*) *>*τ *\* of Lemma* [*5*](#_bookmark19)

such that, E (0) = C(Q0, Q1), (4)

0

← ← *>*0

~0

tor with respect to the ideal distributions and E

(0) is

where T (x) = log *P*1(*x*) 0 is the Bayes optimal detec-

*P*←0(*x*)

*that satisﬁes equal opportunity on the given dataset, i.e.,*

← ←

← *z z*

*P*0(*x*)(1*\_w*)*P*1(*x*)*w*

*P*0(*x*)(1*\_o*)*P*1(*x*)*o*

←

EFN*,T*0 (τ0) = EFN*,T*1 (0) = C(Q0, Q1)*.*

*(Accuracy and Fairness on ideal data) The Chernoff ex-*

*ponent of the probability of error of the Bayes optimal*

*detector on the ideal distributions, i.e.,* C(P0, P1) =

C(Q0, Q1)*, and is hence greater than* C(P0, P1)*.*

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 distri-

butions (see Lemma [5](#_bookmark19)) is in fact the Bayes optimal de-

tector with respect to the ideal distributions. Note that

there exist multiple pairs of (v, w) such that P0(x) =

*P*0(*x*)(1*\_w*)*P*1(*x*)*w* and P1(x) = *P*0(*x*)(1*\_o*)*P*1(*x*)*o* sat-

isfy 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(P0, P1) = C(Q0, Q1)). The Bayes optimal detector for the unprivileged group with re-

spect to the ideal distributions, i.e., T←0(x) = log *P*←1(*x*) *>*0

is thus not only *fair*

*P*←0(*x*)

on the given dataset but also satisﬁes

FN*,T*

the Chernoff exponent of the probability of false negative

for this detector when evaluated on the given distributions

P0(x) and P1(x). Theorem [2](#_bookmark20) already shows that the afore-

mentioned optimization is feasible.

The results of this subsection can be extended to optimiza-

tion ([2](#_bookmark13)), 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 interpreta-

tions: (i) They could be viewed as plausible distributions in

the observed space if the mappings were unbiased from a

separability standpoint (recall Deﬁnition [4](#_bookmark12)). (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 di- mensionality) except through the observed data. It is not unfathomable to assume they would have a separability of

at least C(Q , Q ), which is the separability exhibited by

equal opportunity on the ideal data because its Chernoff ex-

← ←*|| ||*

ponent of FNR is also equal to that of the privileged group,

i.e., C(Q0, Q1). 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*](#_bookmark20) *provides a proof of ex-*

*istence of ideal distributions along with an explicit construc-*

*tion. In general, there may exist other pairs of distributions,*

*which are not of the particular form mentioned in Theo-*

*rem* [*2*](#_bookmark20)*, but might satisfy the two conditions of the theorem.*

*Therefore, given only* P0(x) *and* P1(x)*, the ideal distribu-*

*tions are not necessarily unique unless further assumptions*

*are made about their desirable properties.*

In order to go about ﬁnding such ideal distributions in prac-

tice, 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 π0D(P0 P0) + π1D(P1 P1) be as small

as possible, i.e., the KL divergences of the ideal distribu-

tions from their respective given real-world distributions

the privileged group in the observed space. Theorem [2](#_bookmark20) thus

also demonstrates that the construct space is non-empty.

0

1

**Remark 7** (Explicit Use of an Ideal Dataset)**.** *Several ex-*

*isting methods (*[*Calmon et al.*](#_bookmark44)*,* [*2018*](#_bookmark44)*;* [*Feldman et al.*](#_bookmark36)*,* [*2015*](#_bookmark36)*;*

[*Kamiran & Calders*](#_bookmark52)*,* [*2012*](#_bookmark52)*) propose pre-processing the given*

*dataset to generate an alternate dataset that satisﬁes cer-*

*tain fairness and utility (representation) properties, in the*

*same spirit as optimization* ([4](#_bookmark21))*, 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](#_bookmark21)) *is also remi-*

*niscent of the formulation of* [*Jiang & Nachum*](#_bookmark48) *(*[*2019*](#_bookmark48)*), 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 eX as opposed to conditional feature*

*distributions* p*XeY . Furthermore,* [*Jiang & Nachum*](#_bookmark48) *(*[*2019*](#_bookmark48)*)*

*do not analytically characterize trade-offs.*

**Remark 8** (Implicit Use of an Ideal Dataset)**.** *Existing meth- ods 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 gen- erating an ideal dataset, these methods aim to ﬁnd a clas- siﬁer that satisﬁes a fairness criterion on the given dataset, with minimal compromise of accuracy on the given dataset (recall optimizations* ([2](#_bookmark13)) *and* ([3](#_bookmark18))*). Here, we show that there exist ideal distributions corresponding to these fair detec- tors such that a sub-optimal detector on the given dataset can be optimal with respect to the ideal dataset.*

### 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](#_bookmark9), 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 informa- tion 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 ad- ditional separability information that was not present in the initial dataset. In fact, this is the idea behind active fairness

([Noriega-Campero et al.](#_bookmark61), [2019](#_bookmark61); [Bakker et al.](#_bookmark33), [2019](#_bookmark33); [Chen](#_bookmark27)

C(W0, W1)=C(P0, P1). 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 C(W0, W1)>C(P0, P1), 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](#_bookmark26) in

Section [4](#_bookmark22)). New ideal distributions can also be found using the techniques of Section [3.2](#_bookmark17) that are more plausible as ei- ther 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.

# Numerical Example

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

**Example 1.** *Let the exam score for* Z = 0 *be*

P0(x)*~N* (1, 1) *and* P1(x)*~N* (4, 1)*, and that for* Z = 1

*be* Q0(x)*~N* (0, 1) *and* Q1(x)*~N* (4, 1)*.*

Let us restrict ourselves to likelihood ratio detectors of the form T0(x) = log *P*0(*x*) *>* τ0 and T1(x) = log *Q*0(*x*) *>* τ1

*P*1(*x*)

*Q*1(*x*)

[et al.](#_bookmark27), [2018](#_bookmark27)). 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

for the two groups. The log generating functions for Z = 1

can be computed analytically as: Λ0(u)*z*=1 = 8u(u 1) and Λ1(u)*z*=1 = 8u(u+1) (derivation in Appendix A.3) and the Chernoff information can be computed as C(Q0, Q1) = 2.

*-*

Now, for the unprivileged group Z = 0, the log generat-

ing functions can be computed as Λ (u) = 9 u(u *-* 1)

including extra explanations to go along with the data or

and 9

0 *z*=0 2

decision, similar to [Varshney et al.](#_bookmark65) ([2018](#_bookmark65)). Let (X, X*/*) have

the following distributions: (X, X*/*)*|Y* =0*,Z*=0 *~* W0(x, x*/*)

and (X, X*/*)*|Y* =1*,Z*=0 *~* W1 (x, x*/*), where Y is the true

label. Note that, P0(x) =

*x/* W0(x, x*/*) and P1(x) =

Λ1(u)*z*=0 = 2 u(u + 1) (again see Appendix A.3 for

derivation). The Chernoff information is C(P0, P1) = 9/8.

**Accuracy-Fairness Trade-off in Real World:** We restrict

the detector for the privileged group to be the Bayes optimal

*x/* W1(x, x*/*). Our goal is to derive the conditions un-

detector T1(x)= log *Q*1(*x*) *>* 0 (equivalent to x *>* 2). For

der which the separability improves with addition of more

features, i.e., C(W , W ) > C(P , P ).

this detector,

*Q*0(*x*)

EFP*,T*1 (0)=EFN*,T*1 (0) = C(Q0, Q1) = 2.

0 1 0 1

**Theorem 3** (Improving Separability)**.** *The Chernoff infor- mation* C(W0, W1) *is strictly greater than* C(P0, P1) *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

C(W , W ) *>* C(P , P ) because separability can only

0

1

0

1

Now, for Z=0, the Bayes optimal detector T0(x)= log *P*1(*x*) *>*0 (or, x*>*1.5) will be unfair since EFN*,T*0 (0)=C(P0, P1)<EFN*,T*1 (0). Using the geometric interpretation of Chernoff information (recall Fig. [2](#_bookmark16)),

we can compute the Chernoff exponents of FPR and FNR, i.e., EFP*,T*0 (τ0) and EFN*,T*0 (τ0) as the negative

*P*0(*x*)

of the y-intercept of the tangents to Λ0(u)*z*=0 and

Λ1(u)*z*=0 for detectors T0(x)= log *P*1(*x*) *>*τ0. This enables

*P*0(*x*)

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 W0(x, x*/*)=P0(x) if x*/*=f(x), and 0 otherwise. Similarly, W1(x, x*/*)=P1(x) if x*/*=f(x), and 0 otherwise, leading to

us to numerically plot the trade-off between accuracy

(E*e,T*0 (τ0)= min EFP*,T*0 (τ0), EFN*,T*0 (τ0) ) and fairness ( EFN*,T*0 (τ0) EFN*,T*1 (τ0) ) by varying τ0 as shown by the blue curve in Fig. [3](#_bookmark26).

*| - |*

*{ }*

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

1.4

𝜏% )

1.2



1

Accuracy (𝐸!,#!

0.8

)PwC, w1(

C(,0～ ,1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  | **BA**  **but** |
|  |  |  |  | **exi** |
| **FAIR** | **but SUBOP** | **TIMAL on** | **existing** |  |
| **data a** | **fter data** | **collection** |  | **BAYES O** |
| **FAIR but**  **on existin** | **SUBOPTIM**  **g data** | **AL** |  | **UNFAIR**  **after dat** |

**YES OPTIMAL UNFAIR on**

**sting data**

0.4

0.2



0

0.4

0.2

z)

P0

(z)

P (z)

Q0(

**FAIR**

**BAYES**

Q1

P1

(z)

P (z)=P (z)

(z)

**BAYES**

0.6

0.4

**PTIMAL but**

0

0

-4 -2 0 2

1 1

4 6 8

0.2

0

**on existing data a collection**

*Figure 4.* (Top) For the distributions in Example [1](#_bookmark23), we denote the

Bayes optimal detector log *Q*1(*x*) *>*0 (equivalent to *x >* 2) for the

0 0.2 0.4 0.6 0.8 1



𝜏(

privileged group *Z* = 1

*Q*0(*x*)

*Z* = 0, the optimal detec-

## |𝐸

− 𝐸

| Decrease in

*P*1(*x*)

*>*0

. (Bottom) For

&',#!



𝜏%

&',#"

Fairness

tor log *P* (*x*) 0 does not satisfy equal opportunity on the given dataset but a sub-optimal detector does (notice the equal area cor- responding to false negative rate for two groups). However, there

*Figure 3.* Computation of the trade-off between fairness and ac- curacy using a numerical example: For the unprivileged group, we let *P*0(*x*) (1*,* 1) and *P*1(*x*) (4*,* 1)*.* We restrict the de- tector 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 fea- ture *X/* for the unprivileged group such that (*X, X/*) *Y* =0*,Z*=0

*～N ～N*

*| ～*

～ ～

((1*,* 1)*,* **1**) and (*X, X/*) *Y* =1*,Z*=0 ((4*,* 2)*,* **1**), where **1** is

*N | ～ N*

the 2 2 identity matrix. The green curve shows how active data collection alleviates the trade-off between fairness and accuracy.

*x*

portunity) on the given distributions can also be com-

puted analytically as log *P*1(*x*) *>*τ *\** where τ *\**=*-*3/2 (equiv-

exist ideal distributions given by *P*0 = *Q*0 and *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 ((1, 1), **I**) and (X, X*/*) *Y* =1*,Z*=0

*| ~ N | ~*

((4, 2), **I**), where **I** is the 2 2 identity matrix. The log generating functions can be derived as: Λ0(u) = 5u(u 1) and Λ1(u) = 5u(u + 1). Note that, the Chernoff informa- tion (separability) C(W0, W1) = 5/4 which is greater than C(P0, P1) = 9/8. Thus, the collection of the new feature has improved the separability of the unprivileged group.

*-*

*N ×*

Now, we examine the effect of active data collection on the

alent to

*P*0(*x*) 0 0

accuracy-fairness trade-off in the real world. We again

x 2). This leads to equal exponent of FNR, i.e., EFN*,T*0 ( 3/2)=2=EFN*,T*1 (0) but for this detec-

*-*

*>*

tor EFP*,T* (τ *\**)=1/2 leading to reduced Chernoff ex-

refer to Fig. [3](#_bookmark26) (green curve). Consider the likelihood

ratio detector for Z = 0 based on the total set of fea-

0

0 0 tures, i.e., T (x, x*/*) = log *W*0(*x,x/*)

ponent of overall error probability (represents accu-

0

*W*1(*x,x/*)

*>* τ . To satisfy

racy), i.e., E*e,T*0 (τ0*\**)= min*{*EFP*,T*0 (τ0*\**), EFN*,T*0 (τ0*\**)*}* =

min*{*1/2, 2*}* = 1/2 which is less than C(P0, P1) = 9/8.

our fairness constraint, we need to choose a τ0*\** such that

EFN*,T*0 (τ0*\**)=EFN*,T*1 (0) = C(Q*′*0, Q1) = 2. Upon solv-

ing, we obtain that τ0*\** = 5 *-* 40 *s -′*1.32. For this

**Ideal Distributions:** We refer to Fig. [4](#_bookmark25). It turns out that one pair of ideal distributions prescribed by Theorem [2](#_bookmark20) is P0=Q0 and P1=P1=Q1. The Bayes optimal detector with

← ←

*- s*

respect to the ideal distributions for Z = 0 is given by

log *P*←1(*x*) *>*0 (equivalent to x*>*2). Note that, this is equiva-

←*P* (*x*) 0

lent to the detector log *P*1(*x*) *>* τ *\** where τ *\**= *-* 3/2 which satisﬁed equal opportunity on the given dataset. This detec-

*P*0(*x*)

0

0

tor is now Bayes optimal with respect to the ideal distribu- tions P0 and P1, and has a Chernoff exponent of the overall

← ←

← ←

probability of error equal to C(P0, P1) = 2 when measured with respect to the ideal distributions. Thus, we demon-

strate 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](#_bookmark21)) 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

value of τ0*\**, we obtain EFP*,T*0 (τ0*\**)=7 40) 0.68. The Chernoff exponent of the probability of error for this *fair* detector is given by min EFN*,T*0 (τ0*\**), EFP*,T*0 (τ0*\**) =

min 2, 0.68 = 0.68 which is greater than 0.5 (the Cher-

*{ }*

*{ }*

noff exponent of the probability of error for the fair detector

before collection of the additional feature X*/*).

# 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.](#_bookmark64), [2020](#_bookmark64); [Wick et al.](#_bookmark49), [2019](#_bookmark49)). 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., Dud´ık, M., Langford, J., and Wallach, H. A reductions approach to fair classiﬁca- tion. 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. Em- pirically estimable classiﬁcation bounds based on a non- parametric divergence measure. *IEEE Transactions on Signal Processing*, 64(3):580–591, 2015.

Bhattacharyya, A. On a measure of divergence between two multinomial populations. *Sankhya¯ : 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 Infor- mation Processing Systems*, pp. 3992–4001, 2017.

Calmon, F. P., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Data pre-processing for discrim- ination 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 Trans- parency*, pp. 319–328, 2019.

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

Coˆte´, 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 Process- ing Systems*, pp. 2791–2801, 2018.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel,

1. 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,

1. 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/](http://web.mit.edu/gallager/www/papers/chap3.pdf) [papers/chap3.pdf](http://web.mit.edu/gallager/www/papers/chap3.pdf), 2012.

Garg, S., Kim, M. P., and Reingold, O. Tracking and improv- ing 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. Fair- ness in supervised learning: An information theoretic approach. In *Proceedings of the IEEE International Sym- posium on Information Theory*, pp. 176–180, 2018.

Gretton, A., Borgwardt, K., Rasch, M., Scho¨ lkopf, 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 Informa- tion Processing Systems*, pp. 3315–3323, 2016.

Jiang, H. and Nachum, O. Identifying and correct- ing 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 Com- munication 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 Scho¨lkopf, B. Avoiding discrimination through causal reasoning. In *Advances in Neural Infor- mation Processing Systems*, pp. 656–666, 2017.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Coun- terfactual 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 dis- criminatory: A comparison of fairness deﬁnitions under different worldviews. *arXiv preprint arXiv:1808.08619*, 2018.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gum- madi, K. P. Fairness beyond disparate treatment & dis- parate impact: Learning classiﬁcation without disparate mistreatment. In *Proceedings of the International Con- ference 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.