arXiv:1910.07870v1 [stat.ML] 17 Oct 2019

An Information-Theoretic Perspective on the Relationship Between Fairness and Accuracy

Sanghamitra Dutta,[1,2] Dennis Wei,[1] Hazar Yueksel,[1]
Pin-Yu Chen,[1] Sijia Liu,[1] and Kush R. Varshney[1]
[1]IBM Research, [2]Carnegie Mellon University

Abstract

Our goal is to understand the so-called trade-off between fairness and accuracy. In this work, using a tool from information theory called *Chernoff information*, we derive fundamental limits on this relationship that explain why the accuracy on a given dataset often decreases as fairness increases. Novel to this work, we examine the problem of fair classification through the lens of a *mismatched hypothesis testing* problem, i.e., where we are trying to find a classifier that distinguishes between two "ideal" distributions but instead we are given two mismatched distributions that are biased. Based on this perspective, we contend that measuring accuracy with respect to the given (possibly biased) dataset is a problematic measure of performance. Instead one should also consider accuracy with respect to an ideal dataset that is unbiased. We formulate an optimization to find such ideal distributions and show that the optimization is feasible. Lastly, when the Chernoff information for one group is strictly less than another in the given dataset, we derive the information-theoretic criterion under which collection of more features can actually improve the Chernoff information and achieve fairness without compromising accuracy on the available data.

1 Introduction

With machine learning being applied in highly consequential domains such as hiring, criminal justice, and lending, it is becoming increasingly important to ensure that the decisions are fair and trustworthy [1]. To address this issue, several fairness measures and bias mitigation algorithms have been proposed that either pre-process training data to remove bias, train with a fairness regularization term in the loss function, or perform post-processing of an algorithmic decision for fairness [2–26]. In this work, we focus on group fairness measures[1] that quantify bias in allocation decisions against entire groups of people defined by their *protected attributes* such as race and gender.

It has often been observed that improving group fairness leads to a reduction in accuracy on the given dataset [4, 5, 27, 28]. This motivates a fundamental question as follows:

Is there a trade-off between fairness and accuracy?

Author Contacts: S. Dutta (sanghamd@andrew.cmu.edu), D. Wei (dwei@us.ibm.com), H. Yueksel (hazar.yueksel@ibm.com), P.-Y. Chen (pin-yu.Chen@ibm.com), S. Liu (sijia.liu@ibm.com) and K. R. Varshney (krvarshn@us.ibm.com).

This work was done when S. Dutta was a research intern at IBM Research.

[1] Other measures of fairness, e.g., individual fairness [3] are outside the scope of this work.

Towards answering this question, we first demonstrate that the main reason behind the trade-off is a discrepancy across groups in the amount of information available in the given dataset to *separate* the positive and negative class labels. Using a tool from information theory called *Chernoff information* [29], we theoretically quantify the "separability information" for each group (see Theorem 1).

Our information-theoretic quantification of separability helps demonstrate that if the dataset inherently has less Chernoff information for an unprivileged group compared to a privileged group, then an algorithm attempting to achieve fairness, e.g., by equalizing the true positive rate across groups (equality of opportunity [5]), can often choose a sub-optimal classifier for one or both the groups, thereby reducing the overall accuracy on the given dataset (see Theorem 2). As [30] write, "an algorithm is only as good as the data it works with."

Typically, accuracy with respect to the given dataset is the performance metric of a trained model. However, in the fairness context, this dataset itself is an inaccurate representation due to historical prejudices or sample biases [30]. This motivates us to an alternate interpretation of the problem of fair classification through the lens of *mismatched hypothesis testing* [31]. The goal is to find a classifier that distinguishes between two "ideal" hypotheses, but due to practical limitations, we only have access to the biased or mismatched data. The accuracy of a classifier with respect to the given dataset may be different from its accuracy with respect to an ideal dataset that is free from biases. We therefore contend that in the fairness context, accuracy with respect to the given dataset is a problematic measure of performance, and instead one should also consider accuracy with respect to an ideal data distribution.

We propose an optimization problem to find these ideal distributions such that: (i) they are a useful representation of the given data distributions; and (ii) the Bayes optimal classifier [29] with respect to the ideal distributions is fair on the available data.

We show that this optimization is feasible (see Theorem 3), demonstrating that a classifier that is sub-optimal with respect to the given dataset can be optimal with respect to an ideal dataset that is free from biases, and therefore deemed *fair*. We also clarify how most given methods of fairness explicitly (e.g., data pre-processing) or implicitly (e.g., regularized training) adopt this approach of finding a fair classifier, thereby explaining the accuracy-fairness trade-off on the given dataset while improving accuracy on the ideal dataset.

Lastly, in the typical case that the separability (Chernoff information) of one group is lower than the other in the given dataset, we derive the information-theoretic criterion (see Theorem 4) under which collection of more features can actually improve separability, and achieve fairness without compromising accuracy on the given data. This analysis further explains the benefits of the active fairness paradigm recently proposed by [32].

To summarize, our main contributions are as follows:

- **Quantifying Separability (Theorem 1):** We quantify the separability of each group of people in a given dataset using Chernoff information, an information-theoretic bound on the best exponent of the probability of error in binary classification.
- **Explaining the Accuracy-Fairness Trade-Off (Theorem 2):** Next, we demonstrate that if the Chernoff information of one group is lower than that of the other in the given dataset, then modifying the classifier using a group fairness criterion compromises the error exponent (accuracy) of one or both the groups with respect to the given dataset.
- **Alternate Perspective on Accuracy based on Mismatched Detection (Theorem 3):** Because the given data distribution is inherently biased, we propose an optimization problem

to find an ideal data distribution that is free from biases while still being a reasonably good representation of the given data, and show that this optimization is feasible.

- **Information-Theoretic Criterion to Improve Separability (Theorem 4):** If the Chernoff information for one group is lower than the other, we derive the criterion under which collecting more features improves separability, achieving fairness without compromising accuracy on the available data.

Related Work: We note that some existing works such as [6,27,28] have addressed the problem of an accuracy-fairness trade-off or pointed out that collecting more data is the key to improving fairness. Our novelty lies in adopting a mismatched hypothesis testing viewpoint, which also helps us demonstrate how existing methods without active data collection choose a sub-optimal detector with respect to the given dataset, explaining the so-called accuracy-fairness trade-off on the given dataset, while improving the classification accuracy with respect to an ideal dataset. We also provide the information-theoretic criterion that describes when active data collection methods, as mentioned in [28, 32] actually improve accuracy.

Compared to existing methods of pre-processing data to generate a fair dataset [20, 33, 34], in this work our goal is to quantify the best possible accuracy of any classifier on the given dataset, while satisfying a fairness criterion. Our method of generating an ideal dataset is based on equal opportunity rather than a statistical parity based criterion in [33].

Our tools share similarities with [35] (which demonstrates how explainability can improve Chernoff information), as well as the theory of hypothesis testing in general [29, 31]. Our contribution lies in using these tools in the context of fairness and trustworthy machine learning, that has not received much attention.

Remark 1 (Population Setting)**.** *In this work, we operate in the population setting, 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 [29]) 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 [29]). The population setting abstracts away practical issues of finite data and choice of classification algorithm. This setting has been used widely in machine learning by, e.g., [36–38].*

2 Background and Problem Formulation

2.1 Notation and Preliminaries

Let Z denote the protected attribute, X denote the feature vector, Y denote the true label (i.e., takes value 0 or 1) and $\hat{Y}$ denote the predicted label. Without loss of generality, let $Z = 0$ be the unprivileged group and $Z = 1$ be the privileged group. Let the features in the given dataset 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 $X|_{Y=1,Z=1}{\sim}Q_1(x)$.

We use the notation H_0 and H_1 to denote the hypothesis that the true label is 0 or 1. For each group $Z = z$, we will be considering classifiers of the form $T_z(x) \overset{H_1}{\geq} \tau_z$, i.e., the prediction label is $\mathbb{1}(T_z(x) \geq \tau_z)$ where $\mathbb{1}(\cdot)$ is the indicator function. Thus, the predicted label for $(X, Z) = (x, z)$ is given by:

$$\hat{Y}(x,z) = \begin{cases} \mathbb{1}(T_0(x) \geq \tau_0), & z=0 \\ \mathbb{1}(T_1(x) \geq \tau_1), & z=1. \end{cases} \tag{1}$$

Remark 2 (Decoupled Classifiers)**.** *While such models may qualify as disparate treatment (due to the explicit use of Z), the intent and effect is to better mitigate disparate impact resulting from the model using the protected attribute Z explicitly in the decision making. In this respect the classifier in* (1) *shares the same spirit with fair affirmative action [3, 39]. Furthermore, a classifier that does not use Z explicitly becomes a special case of* (1) *if we choose the same classifier for both the groups.*

Next, we state two important assumptions.

Assumption 1 (Absolute Continuity)**.** *Let $P_0(x)$, $P_1(x)$, $Q_0(x)$ and $Q_1(x)$ be greater than 0 everywhere in the range of x.*

This assumption ensures that likelihood ratio detectors of the form $T_0(x) = \log \frac{P_1(x)}{P_0(x)}$ as well as the Kullback-Leibler (KL) divergences between any two of these distributions are well-defined.

Assumption 2 (Distinct Hypotheses)**.** *Let $\mathrm{D}(P_0||P_1)$, $\mathrm{D}(P_1||P_0)$, $\mathrm{D}(Q_0||Q_1)$ and $\mathrm{D}(Q_1||Q_0)$ be strictly greater than 0, where $\mathrm{D}(\cdot||\cdot)$ is the KL divergence between two distributions.*

We use the notations $P_{\mathrm{FP}}^{(T_z)}(\tau_z)$ and $P_{\mathrm{FN}}^{(T_z)}(\tau_z)$ to denote the probability of false positive and the probability of false negative for a particular classifier of the form $T_z(x) \overset{H_1}{\geq} \tau_z$ over the members of the group $Z=z$. For any group, $P_{\mathrm{FP}}^{(T_z)}(\tau_z)$ is the probability of wrongful acceptance of negative class labels, given by:

$$P_{\mathrm{FP}}^{(T_z)}(\tau_z) = \Pr\left(T_z(X) \geq \tau_z | Y=0, Z=z\right). \tag{2}$$

Similarly, $P_{\mathrm{FN}}^{(T_z)}(\tau_z)$ is the probability of wrongful rejection of positive class labels, given by:

$$P_{\mathrm{FN}}^{(T_z)}(\tau_z) = \Pr\left(T_z(X) < \tau_z | Y=1, Z=z\right). \tag{3}$$

The overall probability of error of a group is given by:

$$P_e^{(T_z)}(\tau_z) = \pi_0 P_{\mathrm{FP}}^{(T_z)}(\tau_z) + \pi_1 P_{\mathrm{FN}}^{(T_z)}(\tau_z), \tag{4}$$

where π_0 and π_1 are the prior probabilities of hypotheses H_0 and H_1 given $Z=z$.

One well-known definition of fairness by [5] is *equalized odds*, which states that an algorithm is fair if it has equal probabilities of false positive and false negative for the two groups, i.e., $Z=0$ and 1. A relaxed variant of this definition, widely used in the fairness literature, is *equal opportunity*, which enforces only equal false negative rate (or equivalently, equal true positive rate) for the two groups.

We consider binary classification here for the sake of simplicity, but our techniques extend to more than two classes. Similarly, while we consider only two groups defined by the protected attribute, the techniques also apply when there are more groups.

Next, we provide a brief background on the error exponents of a binary classifier.

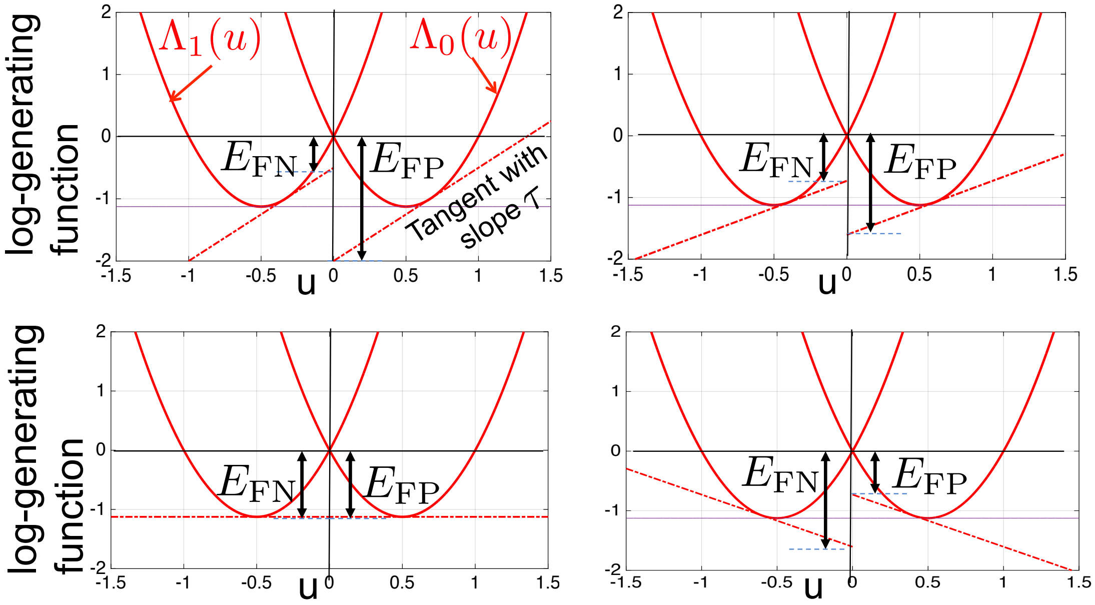

Figure 1: Let $P_0(x)\sim\mathcal{N}(1,1)$ and $P_1(x)\sim\mathcal{N}(4,1)$. For a likelihood ratio detector $T(x)=\log\frac{P_1(x)}{P_0(x)}\overset{H_1}{\geq}\tau$, we plot $\Lambda_0(u)$ and $\Lambda_1(u)$. Note that, $\Lambda_0(u)$ is strictly convex with zeros at $u=0$ and $u=1$, and $\Lambda_1(u)=\Lambda_0(u+1)$. We obtain $E^{(T)}_{\mathrm{FP}}(\tau)$ and $E^{(T)}_{\mathrm{FN}}(\tau)$ as the negative of the y-intercepts for tangents to $\Lambda_0(u)$ and $\Lambda_1(u)$ respectively with slope τ. As we vary τ, there is a trade-off between $E^{(T)}_{\mathrm{FP}}(\tau)$ and $E^{(T)}_{\mathrm{FN}}(\tau)$ until they both become equal at $\tau=0$. The value of the exponent at $\tau=0$ is defined as the Chernoff Information $(\mathrm{C}(P_0,P_1):=E^{(T)}_{\mathrm{FP}}(0)=E^{(T)}_{\mathrm{FN}}(0))$.

2.2 Error Exponents of a Binary Classifier

The error exponents of the probability of false positive and false negative are given by $-\log P_{\text{FP}}^{(T_z)}(\tau_z)$ and $-\log P_{\text{FN}}^{(T_z)}(\tau_z)$ respectively. In many applications, we may not be able to obtain a simple closed-form expression for the exact error probabilities or their exponents [40], but the exponents are approximated using a well-known lower bound called the *Chernoff bound*, that is known to be pretty tight (see Remark 3 and also [40–45]). Compared to exact error exponents, the Chernoff exponents yield more insight because of their connection to Fenchel-Legendre (FL) transforms (see Property 5 in Appendix A.2).

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

$$E_{\text{FP}}^{(T_z)}(\tau_z) = \sup_{u>0}(u\tau_z - \Lambda_0(u))$$
$$E_{\text{FN}}^{(T_z)}(\tau_z) = \sup_{u<0}(u\tau_z - \Lambda_1(u)),$$

where $\Lambda_0(u)$ and $\Lambda_1(u)$ are log-generating functions:

$$\Lambda_0(u) = \log \mathbb{E}[e^{uT_z(X)}|Y=0, Z=z]$$
$$\Lambda_1(u) = \log \mathbb{E}[e^{uT_z(X)}|Y=1, Z=z].$$

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

$$P_{\text{FP}}^{(T_z)}(\tau_z) \leq e^{-E_{\text{FP}}^{(T_z)}(\tau_z)} \tag{5}$$
$$P_{\text{FN}}^{(T_z)}(\tau_z) \leq e^{-E_{\text{FN}}^{(T_z)}(\tau_z)}. \tag{6}$$

The proof is provided in Appendix A.1.

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 [41]. The Bhattacharya bound (a special case of the Chernoff bound) both upper and lower bounds the Bayes error probability [43–45].*

Geometric Interpretation: 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., $\mathbb{E}[T_z(X)|H_0, Z{=}z]{<}0$ and $\mathbb{E}[T_z(X)|H_1, Z{=}z]{>}0$, 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 $\tau_z \in (\mathbb{E}[T_z(X)|H_0, Z{=}z], \mathbb{E}[T_z(X)|H_1, Z{=}z])$).

For the ease of understanding, consider the following numerical example:

Example 1. *Let the data distributions 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 $T_0(x) = \log \frac{P_1(x)}{P_0(x)} \overset{H_1}{\geq} \tau_0$ be a likelihood ratio detector for the group $Z = 0$. The likelihood ratio detector is well-behaved if $\mathrm{D}(P_0||P_1) > 0$. The log-generating functions for this detector can be computed as follows: $\Lambda_0(u) = \frac{9}{2}u(u-1)$ and $\Lambda_1(u) = \frac{9}{2}u(u+1)$ (see Appendix A.3 for more details). We refer to Fig. 1 for an illustration of the Chernoff exponents for this detector.

[2] For a detector $T_z(x) \overset{H_1}{\geq} \tau_z$, we would expect $T_z(X)$ to be high when H_1 is true, and low when H_0 is true.

Next, we quantify the error exponent of the overall probability of error ($P_e^{(T_z)}(\tau_z)$). For the sake of simplicity, first consider the case where $\pi_0 = \pi_1 = \frac{1}{2}$, and $P_e^{(T_z)}(\tau_z) = \frac{1}{2}P_{\text{FP}}^{(T_z)}(\tau_z) + \frac{1}{2}P_{\text{FN}}^{(T_z)}(\tau_z)$. 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 their Chernoff bounds.

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

$$E_e^{(T_z)}(\tau_z) = \min\{E_{\text{FP}}^{(T_z)}(\tau_z), E_{\text{FN}}^{(T_z)}(\tau_z)\}. \tag{7}$$

Relation with Accuracy: 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 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.

Remark 4 (Unequal Priors)**.** *When the prior probabilities are unequal, we can write $P_e^{(T_z)}(\tau_z)$ as $P_e^{(T_z)}(\tau_z)$=$\frac{1}{2}(2\pi_0 P_{\text{FP}}^{(T_z)}(\tau_z))$+$\frac{1}{2}(2\pi_1 P_{\text{FN}}^{(T_z)}(\tau_z))$, and define the Chernoff exponent of $P_e^{(T_z)}(\tau_z)$, i.e., $E_e^{(T_z)}(\tau_z)$ more generally as follows (see Appendix E for details):*

$$\min\{E_{\text{FP}}^{(T_z)}(\tau_z) - \log 2\pi_0, E_{\text{FN}}^{(T_z)}(\tau_z) - \log 2\pi_1\}.$$

Lemma 2 (Relation with Accuracy)**.** *Let Assumptions 1 and 2 hold, and $T_z(x)$ be the likelihood ratio detector for the group $Z = z$. Then, the value of τ_z that maximizes $E_e^{(T_z)}(\tau_z)$, i.e.,*

$$\max_{\tau_z} \min\{E_{\text{FP}}^{(T_z)}(\tau_z) - \log 2\pi_0, E_{\text{FN}}^{(T_z)}(\tau_z) - \log 2\pi_1\},$$

is given by $\tau_z^* = \log \frac{\pi_0}{\pi_1}$, which is the same as the value of τ_z that minimizes $P_e^{(T_z)}(\tau_z)$, i.e.,

$$\min_{\tau_z} \pi_0 P_{\text{FP}}^{(T_z)}(\tau_z) + \pi_1 P_{\text{FN}}^{(T_z)}(\tau_z).$$

This likelihood ratio detector $T_z(x) \overset{H_1}{\geq} \log \frac{\pi_0}{\pi_1}$ is the Bayes optimal detector for the group. The proof of Lemma 2 is provided in Appendix E.1.

For the rest of the paper, we will assume that $\pi_0 = \pi_1 = \frac{1}{2}$ for simplicity even though our techniques generalize to unequal priors as well (as already demonstrated in Lemma 2). Equal priors also correspond to the balanced accuracy measure [46] which is often favored over ordinary accuracy.

2.3 Problem Setup

We aim to derive fundamental limits (bounds) that explain the accuracy-fairness trade-off of a classifier. In particular, our metrics of interest are:

1. $E_e^{(T_0)}(\tau_0)$ and $E_e^{(T_1)}(\tau_1)$: A higher value of the Chernoff exponent of $P_e^{(T_z)}(\tau_z)$ implies a higher accuracy for group z.
2. $|E_{\text{FN}}^{(T_0)}(\tau_0) - E_{\text{FN}}^{(T_1)}(\tau_1)|$: Inspired by equal opportunity, we consider the absolute difference of the Chernoff exponents of the probability of false negative as our measure of fairness.

Our *goal* is to understand the trade-off between $E_e^{(T_0)}(\tau_0)$, $E_e^{(T_1)}(\tau_1)$ and $|E_{\text{FN}}^{(T_0)}(\tau_0) - E_{\text{FN}}^{(T_1)}(\tau_1)|$.

As stated before in Remark 1 (Section 1), we assume that we are operating in the population setting, i.e., we have access to the probability distributions of the data. This assumption enables us to represent $T_0(x)$ and $T_1(x)$ as *likelihood ratio detectors* for each group.

This modeling can be further justified using the NP Lemma [29, Theorem 11.7.1] which states that given any classifier $T'(x) \overset{H_1}{\geq} \tau'$ with a certain $(P_{\text{FP}}^{(T')}(\tau'), P_{\text{FN}}^{(T')}(\tau'))$, there exists a likelihood ratio detector $T(x) \overset{H_1}{\geq} \tau$ with $P_{\text{FP}}^{(T)}(\tau) = P_{\text{FP}}^{(T')}(\tau')$ and $P_{\text{FN}}^{(T)}(\tau) \leq P_{\text{FN}}^{(T')}(\tau')$. Given enough labeled samples from each of the distributions, the probabilities of false positive and false negative for a classifier learned from training data converges to that of a likelihood ratio detector [47].

3 Fundamental Limits on the Trade-Off Between Accuracy and Fairness

3.1 Quantification of Separability

We quantify the separability of the positive and negative labels of a group in the given dataset using a tool from information theory called *Chernoff information*. We first state a well-known result from [29] here.

Lemma 3. *Given two hypotheses, $H_0 : P_0(x)$ and $H_1 : P_1(x)$, the Chernoff exponent of the Bayes optimal classifier is given by the Chernoff information:*

$$\mathrm{C}(P_0, P_1) = -\min_{u \in (0,1)} \log \left(\sum_x P_0(x)^{1-u} P_1(x)^u \right). \tag{8}$$

For completeness, a proof is provided in Appendix B.1.

For a proof sketch, the reader may refer to Fig. 1. The Chernoff information essentially quantifies $E_e^{T_z}(\tau_z)$ for the best likelihood ratio detector. It is therefore well-suited as an information-theoretic quantification of the separability of the two hypotheses.

Using Lemma 3, we can quantify the separability of the group with $Z = 0$ in the given dataset as $\mathrm{C}(P_0, P_1)$, and that of the group with $Z = 1$ as $\mathrm{C}(Q_0, Q_1)$. Since $Z = 0$ is assumed to be the unprivileged group without loss of generality, $\mathrm{C}(P_0, P_1) \leq \mathrm{C}(Q_0, Q_1)$.

The following theorem summarizes the two cases in which $\mathrm{C}(P_0, P_1)$ is either equal to or less than $\mathrm{C}(Q_0, Q_1)$.

Theorem 1 (Quantification of Separability)**.** *Under Assumptions 1 and 2, one of the following is true:*

- $\mathrm{C}(P_0, P_1) = \mathrm{C}(Q_0, Q_1)$*, and there exist likelihood ratio detectors for the two groups such that the Chernoff exponents of the probability of error, false positive and false negative are all equal to* $\mathrm{C}(Q_0, Q_1)$*.*
- $\mathrm{C}(P_0, P_1) < \mathrm{C}(Q_0, Q_1)$*, and no likelihood ratio detector can improve the Chernoff exponent of the probability of error for the unprivileged group beyond* $\mathrm{C}(P_0, P_1)$ *to match* $\mathrm{C}(Q_0, Q_1)$*.*

The proof is provided in Appendix B.2. Under the first scenario, the classifier with the best accuracy is also a classifier that meets the fairness criterion of equal opportunity. Thus, it is possible to achieve fairness without compromising the accuracy on the given dataset.

The second scenario, which occurs more commonly in practice, is where discrimination is caused due to an inherent limitation of the dataset, i.e., it does not have enough separability information about one group when compared to the other. For the rest of the paper, we will focus on this scenario, i.e., $\mathrm{C}(P_0, P_1) < \mathrm{C}(Q_0, Q_1)$. Under this scenario, no classifier with any group fairness criteria whatsoever can actually improve the Chernoff exponent of the unprivileged group beyond $\mathrm{C}(P_0, P_1)$. On the other hand, the Chernoff exponent of the best detector for the privileged group is higher. An attempt to use any alternate likelihood ratio detector for any of the groups, e.g., detectors that meet fairness criteria, will only reduce accuracy for that group. We formalize this intuition in the following subsection.

3.2 Explaining the so-called accuracy-fairness trade-off

Theorem 2. *Let Assumptions 1 and 2 hold and* $\mathrm{C}(P_0, P_1) < \mathrm{C}(Q_0, Q_1)$. *Suppose that there are two likelihood ratio detectors* $T_0(x) \overset{H_1}{\geq} \tau_0$ *and* $T_1(x) \overset{H_1}{\geq} \tau_1$, *one for each group, such that*

$$|E_{\mathrm{FN}}^{(T_0)}(\tau_0) - E_{\mathrm{FN}}^{(T_1)}(\tau_1)| = 0.$$

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

(i) $E_e^{(T_0)}(\tau_0) < \mathrm{C}(P_0, P_1)$.

(ii) $E_e^{(T_1)}(\tau_1) < \mathrm{C}(Q_0, Q_1)$.

The proof is provided in Appendix B.2. As a proof sketch, we refer to Fig. 2. When both the groups use their Bayes optimal detectors, i.e., $T_0(x) \overset{H_1}{\geq} 0$ and $T_1(x) \overset{H_1}{\geq} 0$, then,

$$E_{\mathrm{FN}}^{(T_0)}(0) = E_{\mathrm{FP}}^{(T_0)}(0) = \mathrm{C}(P_0, P_1),$$

and

$$E_{\mathrm{FN}}^{(T_1)}(0) = E_{\mathrm{FP}}^{(T_1)}(0) = \mathrm{C}(Q_0, Q_1) > \mathrm{C}(P_0, P_1).$$

Thus, $E_{\mathrm{FN}}^{(T_0)}(0)$ cannot be equal to $E_{\mathrm{FN}}^{(T_1)}(0)$. To make them equal, at least one of τ_0 or τ_1 should be non-zero, leading to a sub-optimal detector for that group.

This result explains the accuracy-fairness trade-off in the view of Chernoff information, i.e., why accuracy on the given dataset decreases if fairness is increased.

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.,* $T_0(x) \overset{H_1}{\geq} 0$ *and* $T_1(x) \overset{H_1}{\geq} 0$ *do not satisfy a fairness criterion (e.g. demographic parity), 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 next two results provide intuition on each of the two cases in Theorem 2.

Towards understanding existing methods of finding a fair classifier [9, 48], first consider the following optimization problem, where the goal is to find classifiers of the form $T_0(x) \overset{H_1}{\geq} \tau_0$ and $T_1(x) \overset{H_1}{\geq} \tau_1$ for the two groups that maximize the Chernoff exponent of the probability of error under the constraint that they satisfy *equal opportunity* on the given dataset.

$$\begin{aligned} &\max_{T_0, \tau_0, T_1, \tau_1} \min\{E_{\mathrm{FP}}^{T_0}(\tau_0), E_{\mathrm{FN}}^{T_0}(\tau_0), E_{\mathrm{FP}}^{T_1}(\tau_1), E_{\mathrm{FN}}^{T_1}(\tau_1)\} \\ &\text{such that } E_{\mathrm{FN}}^{T_0}(\tau_0) = E_{\mathrm{FN}}^{T_1}(\tau_1). \end{aligned} \tag{9}$$

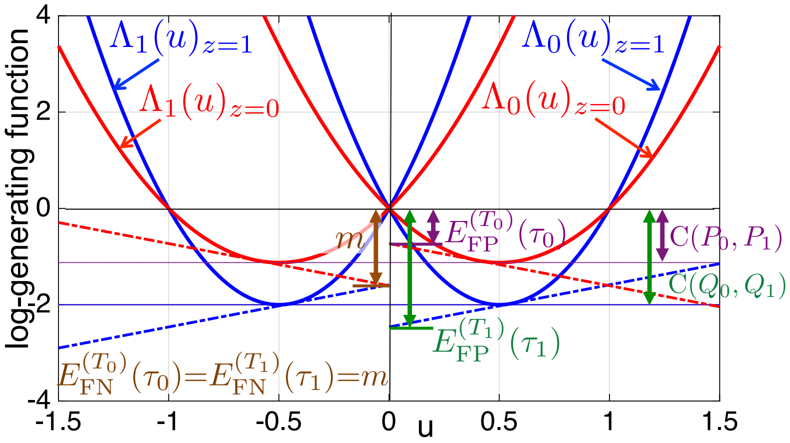

Figure 2: The plot shows the log-generating functions for the two groups corresponding to Example 1. Note that, $\mathrm{C}(P_0, P_1)<\mathrm{C}(Q_0, Q_1)$. The detectors satisfy equal opportunity, i.e., $E_{\mathrm{FN}}^{T_0}(\tau_0) = E_{\mathrm{FN}}^{T_1}(\tau_1)$.

Remark 6 (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 $Z = 1$.*

Recall that the NP Lemma states that given any classifier, there exists a likelihood ratio detector which is at least as good in terms of accuracy. Based on this lemma, 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 there exists a unique solution (τ_0^*, τ_1^*) to (9).

Proposition 1. *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 $T_0(x) \overset{H_1}{\geq} \tau_0^*$ and $T_1(x) \overset{H_1}{\geq} \tau_1^*$ that satisfy the optimization (9) 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. Let $\tau_0^* = 0$, which ensures $E_{\mathrm{FN}}^{(T_0)}(0) = E_{\mathrm{FP}}^{(T_0)}(0) = \mathrm{C}(P_0, P_1)$. Now, the only value of 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

$$\min\{E_{\mathrm{FP}}^{T_0}(0), E_{\mathrm{FN}}^{T_0}(0), E_{\mathrm{FP}}^{T_1}(\tau_1^*), E_{\mathrm{FN}}^{T_1}(\tau_1^*)\}{<}\mathrm{C}(P_0, P_1).$$

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 criteria, 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 instead we try to choose a more fair detector only for the unprivileged group. We propose the following optimization:

$$\begin{aligned} &\max_{T_0, \tau_0} \min\{E_{\mathrm{FP}}^{T_0}(\tau_0), E_{\mathrm{FN}}^{T_0}(\tau_0)\} \\ &\text{such that } E_{\mathrm{FN}}^{T_0}(\tau_0) = \mathrm{C}(Q_0, Q_1). \end{aligned} \tag{10}$$

Again, if we restrict $T_0(x)$ to be the likelihood ratio detector of the form $\log \frac{P_1(x)}{P_0(x)}$, then there exists a unique solution $\tau_0 = \tau_0^*$ to the optimization (10).

Proposition 2. *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* $T_0(x) \overset{H_1}{\geq} \tau_0^*$ *that satisfies the optimization* (10) *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 again refer to Fig. 2. If we choose $\tau_0^* \neq 0$, we get a sub-optimal detector for the unpriviledged group with $E_e^{T_0}(\tau_0^*) < \mathrm{C}(P_0, P_1)$. The proofs for Propositions 1 and 2 are provided in Appendix B.3.

4 A Mismatched Hypothesis Testing Perspective

We view the classification problem on a biased dataset through the lens of mismatched hypothesis testing, i.e., where we wish to find a classifier that distinguishes between two ideal hypotheses, but due to practical limitations, we only have access to data from two mismatched distributions. Existing methods of incorporating fairness, e.g., data pre-processing and regularized training (the latter for example as in Proposition 1), explicitly or implicitly choose a classifier that may be sub-optimal for one or both the groups on the given dataset, but improves the accuracy with respect to an ideal dataset distribution. In particular, we consider the variant where we choose the Bayes optimal detector for the privileged group with respect to the given dataset, while for the unprivileged group, we choose a sub-optimal detector that is fair on the given dataset (optimization (10)). We show that there exist two ideal distributions $\widetilde{P}_0(x)$ and $\widetilde{P}_1(x)$ such that, the Bayes optimal detector with respect to the ideal distributions is fair on the given dataset.

Measure of Representation: In general, given only $P_0(x)$ and $P_1(x)$, the ideal distributions $\widetilde{P}_0(x)$ and $\widetilde{P}_1(x)$ cannot be uniquely specified unless further assumptions are made about their desirable properties. One desirable property of such an ideal dataset is that it should also be a useful representative of the given dataset. This leads to 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 dataset distributions are small.

Based on this perspective, we formulate the following optimization for specifying two ideal distributions $\widetilde{P}_0$ and $\widetilde{P}_1$ for the unprivileged group:

$$\min_{\widetilde{P}_0, \widetilde{P}_1} \pi_0 \mathrm{D}(\widetilde{P}_0 || P_0) + \pi_1 \mathrm{D}(\widetilde{P}_1 || P_1)$$
$$\text{such that, } E_{\mathrm{FN}}^{\widetilde{T}}(0) = \mathrm{C}(Q_0, Q_1), \tag{11}$$

where $\widetilde{T}(x) = \log \frac{\widetilde{P}_1(x)}{\widetilde{P}_0(x)} \overset{H_1}{\geq} 0$ is the Bayes optimal detector with respect to the ideal distributions and $E_{\mathrm{FN}}^{\widetilde{T}}(0)$ is the Chernoff exponent of the probability of false negative for this detector when evaluated on the given distributions $P_0(x)$ and $P_1(x)$. We show that the aforementioned optimization is feasible.

Theorem 3 (Construction of Ideal Distributions)**.** *Under Assumptions 1 and 2, 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:*

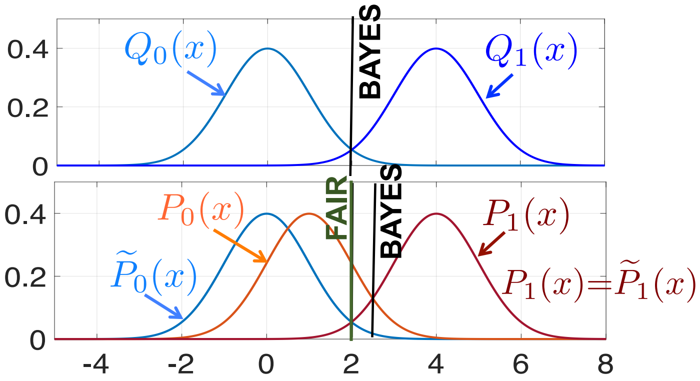

Figure 3: We consider optimization (10) for the distributions in Example 1. For $Z = 1$, we fix the Bayes optimal detector ($\log \frac{Q_1(x)}{Q_0(x)} \overset{H_1}{\geq} 0$). Now, for $Z = 0$, the optimal detector $\log \frac{P_1(x)}{P_0(x)} \overset{H_1}{\geq} 0$ does not satisfy equal opportunity on the given dataset but a sub-optimal detector does (notice the equal area corresponding to false negative for two groups). The fair detector is optimal w.r.t. ideal distributions $\widetilde{P}_0 = Q_0$ and $\widetilde{P}_1 = P_1 = Q_1$.

- *The Bayes optimal detector for the ideal distributions, i.e.,* $\widetilde{T}(x) = \log \frac{\widetilde{P}_1(x)}{\widetilde{P}_0(x)} \overset{H_1}{\geq} 0$ *satisfies equal opportunity on the given dataset, i.e.,* $E_{\text{FN}}^{\widetilde{T}}(0) = \mathrm{C}(Q_0, Q_1)$.
- $\mathrm{C}(P_0, P_1) < \mathrm{C}(\widetilde{P}_0, \widetilde{P}_1) = \mathrm{C}(Q_0, Q_1)$.

The proof is provided in Appendix C. We also refer to Fig. 3 for a pictorial illustration. Theorem 3 demonstrates that both accuracy and fairness of a detector can improve simultaneously when the accuracy is measured with respect to an ideal dataset. Theorem 3 also provides an explicit method of constructing such an ideal dataset when the measure of fairness is equal opportunity. The formulation can be extended to optimization (9) as well as to other measures of fairness altogether, e.g., demographic parity.

Remark 7 (Explicit Use of an Ideal Dataset)**.** *There are several existing methods [33, 34, 49] that address the problem of fairness by first generating an alternate dataset from the given dataset via data pre-processing such that the new dataset satisfies certain fairness properties and utility (representation) metrics along with other constraints if desired, e.g., individual distortion. Next, a detector is trained on the alternate dataset. The trained detector may be sub-optimal with respect to the given dataset but is deemed to be fair. In this work, our contribution lies in explaining this so-called accuracy-fairness trade-off on the given dataset and demonstrating that both accuracy and fairness can improve simultaneously when the accuracy is measured with respect to an ideal dataset.*

Remark 8 (Implicit Use of an Ideal Dataset)**.** *Existing methods that fall under 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 criteria on the given dataset, with minimal compromise of accuracy on the given dataset (recall optimizations* (9) *and* (10)*). In this work, 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.*

5 Improving the Trade-Off by Using More Features

The inherent limitation of disparate separability between groups in the given dataset, discussed in Section 3, can in fact be overcome but with an associated cost. In this section, we demonstrate how 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, and is the idea behind active fairness [32, 50]. Our analysis below serves as a technical explanation for the success of active fairness.

Let X' denote the additional features so that now (X, X') is used for classification of the unprivileged group ($Z = 0$). Let the features (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 4 (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. Here, we first provide an intuition on when $\mathrm{C}(W_0, W_1)$ is strictly greater than $\mathrm{C}(P_0, P_1)$. First, 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 or error probability than 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 H_0 and H_1 better, that essentially leads to X' not being independent of Y given X and $Z = 0$.

In active fairness and in our framing of this section, X' contains additional features to be measured. However, X' could also easily be other forms of additional information including extra explanations to go along with the data or decision, similar to [35].

6 Conclusion

We provide a novel perspective on the so-called accuracy-fairness trade-off on a given dataset using information-theoretic tools such as Chernoff information and mismatched detection that have not been used in this context before. Our results provide insights on existing methods of fair classification and their fundamental limits. Under the perspective of mismatched hypothesis testing, there exist ideal datasets such that a fair yet sub-optimal classifier on a biased dataset is optimal on the ideal dataset. Lastly, we also derive the information-theoretic criterion under which active feature collection can actually improve the accuracy-fairness trade-off.

Acknowledgements

We would like to thank Pulkit Grover for his feedback on this work.

References

[1] Kush R. Varshney. Trustworthy machine learning and artificial intelligence. *ACM XRDS Magazine*, 25(3):26–29, 2019.

[2] Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[3] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

[4] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the International Conference on Machine Learning*, pages 60–69, 2018.

[5] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[6] Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, pages 107–118, 2018.

[7] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012.

[8] Junpei Komiyama and Hajime Shimao. Two-stage algorithm for fairness-aware machine learning. arXiv:1710.04924, 2017.

[9] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

[10] AmirEmad Ghassami, Sajad Khodadadian, and Negar Kiyavash. Fairness in supervised learning: An information theoretic approach. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 176–180, 2018.

[11] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*, pages 1265–1276, 2018.

[12] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[13] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[14] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

[15] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7801–7808, 2019.

[16] Anupam Datta, Matthew Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Use privacy in data-driven systems: Theory and experiments with machine learnt programs. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 1193–1210, 2017.

[17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

[18] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

[19] Jiachun Liao, Chong Huang, Peter Kairouz, and Lalitha Sankar. Learning generative adversarial representations (gap) under fairness and censoring constraints. *arXiv preprint arXiv:1910.00411*, 2019.

[20] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[21] Samuel Yeom, Anupam Datta, and Matt Fredrikson. Hunting for discriminatory proxies in linear regression models. In *Advances in Neural Information Processing Systems*, pages 4568–4578, 2018.

[22] Robert C Williamson and Aditya Krishna Menon. Fairness risk measures. *arXiv preprint arXiv:1901.08665*, 2019.

[23] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio du Pin Calmon. Optimized score transformation for fair classification. *arXiv preprint arXiv:1906.00066*, 2019.

[24] Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2125–2126, 2016.

[25] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568, 2008.

[26] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2239–2248, 2018.

[27] Sumegha Garg, Michael P. Kim, and Omer Reingold. Tracking and improving information in the service of fairness. In *Proceedings of the ACM Conference on Economics and Computation*, pages 809–824, 2019.

[28] Irene Y. Chen, Fredrik D. Johansson, and David Sontag. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pages 3539–3550, 2018.

[29] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[30] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104:671–732, 2016.

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

[32] Alejandro Noriega-Campero, Michiel A. Bakker, Bernardo Garcia-Bulle, and Alex 'Sandy' Pentland. Active fairness in algorithmic decision making. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, pages 77–83, 2019.

[33] Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1106–1119, 2018.

[34] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[35] Kush R. Varshney, Prashant Khanduri, Pranay Sharma, Shan Zhang, and Pramod K. Varshney. 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*, pages 15–20, 2018.

[36] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2007.

[37] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

[38] Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Proceedings of the Conference On Learning Theory*, pages 489–511, 2013.

[39] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 119–133, New York, NY, USA, 23–24 Feb 2018. PMLR.

[40] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Cambridge university press, 1995.

[41] François D Côté, Ioannis N Psaromiligkos, and Warren J Gross. A chernoff-type lower bound for the gaussian q-function. *arXiv preprint arXiv:1202.6483*, 2012.

[42] Daniel Berend and Aryeh Kontorovich. A finite sample analysis of the naive bayes classifier. *Journal of Machine Learning Research*, 16:1519–1545, 2015.

[43] Visar Berisha, Alan Wisler, Alfred O Hero, and Andreas Spanias. Empirically estimable classification bounds based on a nonparametric divergence measure. *IEEE Transactions on Signal Processing*, 64(3):580–591, 2015.

[44] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946.

[45] Thomas Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE transactions on communication technology*, 15(1):52–60, 1967.

[46] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124. IEEE, 2010.

[47] Clayton Scott and Robert Nowak. A neyman-pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, 2005.

[48] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[49] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[50] Michiel A. Bakker, Alejandro Noriega-Campero, Duy Patrick Tu, Prasanna Sattigeri, Kush R. Varshney, and Alex 'Sandy' Pentland. On fairness in budget-constrained decision making. In *Proceedings of the KDD Workshop on Explainable Artificial Intelligence*, 2019.

[51] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[52] RG Gallager. Detection, decisions, and hypothesis testing. `http://web.mit.edu/gallager/www/papers/chap3.pdf`, 2012.

A Background on Chernoff Bound

In this section, we provide a brief background on Chernoff bounds and Chernoff information, leading to the derivation of the results under equal priors, i.e., $\pi_0 = \pi_1 = \frac{1}{2}$. We discuss the case of unequal priors in Appendix E.

Consider a detector of the form $T(x) \overset{H_1}{\geq} \tau$ for classification between two hypothesis $H_0 : X \sim P_0(x)$ and $H_1 : X \sim P_1(x)$. Recall that the log-generating functions for this detector are defined as follows:

$$\Lambda_0(u) = \log \mathbb{E}[e^{uT(X)}|H_0] \tag{12}$$

$$\Lambda_1(u) = \log \mathbb{E}[e^{uT(X)}|H_1]. \tag{13}$$

A.1 Proof of Equation (6)

We first state the Chernoff bound [51, Chapter 2.2] here, which is a well-known tight bound for approximating error probabilities. For a random variable T,

$$\Pr(T \geq \tau) = \Pr(e^{uT} \geq e^{u\tau}) \leq \frac{\mathbb{E}[e^{uT}]}{e^{u\tau}} \quad \forall u > 0. \tag{14}$$

Proof of Equation (6). Using the Chernoff bound, we can bound $P^{(T)}_{\text{FP}}(\tau)$ as follows:

$$\begin{aligned} P^{(T)}_{\text{FP}}(\tau) &= \Pr(T(X) \geq \tau | H_0) \\ &\leq \frac{\mathbb{E}[e^{uT(X)}|H_0]}{e^{u\tau}} \quad \forall u > 0 \\ &= \frac{e^{\Lambda_0(u)}}{e^{u\tau}} \quad \forall u > 0. \end{aligned} \tag{15}$$

This leads to,

$$-\log P^{(T)}_{\text{FP}}(\tau) \geq \sup_{u>0} (u\tau - \Lambda_0(u)) = E^{(T)}_{\text{FP}}(\tau). \tag{16}$$

Similarly,

$$\begin{aligned} P^{(T)}_{\text{FN}}(\tau) &= \Pr(T(X) < \tau | H_1) \\ &\leq \frac{\mathbb{E}[e^{uT(X)}|H_1]}{e^{u\tau}} \quad \forall u < 0 \\ &= \frac{e^{\Lambda_1(u)}}{e^{u\tau}} \quad \forall u < 0. \end{aligned} \tag{17}$$

This leads to,

$$-\log P^{(T)}_{\text{FN}}(\tau) \geq \sup_{u<0} (u\tau - \Lambda_1(u)) = E^{(T)}_{\text{FN}}(\tau). \tag{18}$$

□

A.2 Properties of log-generating functions

Here, we state some useful properties of the log-generating functions that are used later in the other proofs/explanations.

Property 1 (Convexity)**.** *The log-generating functions $\Lambda_0(u)$ and $\Lambda_1(u)$ are convex in u.*

Proof of Property 1. The proof follows directly using Hölder's inequality. For any u and v, and $\alpha \in [0, 1]$,

$$\begin{aligned}
&\mathbb{E}[e^{(\alpha u+(1-\alpha)v)T(X)}|H_0] \\
&= \mathbb{E}[e^{\alpha uT(X)}e^{(1-\alpha)vT(X)}|H_0] \\
&\leq \left(\mathbb{E}[|e^{\alpha uT(X)}|^{\frac{1}{\alpha}}|H_0]\right)^{\alpha} \left(\mathbb{E}[|e^{(1-\alpha)vT(X)}|^{\frac{1}{1-\alpha}}|H_0]\right)^{1-\alpha}. \qquad (19)
\end{aligned}$$

This leads to,

$$\begin{aligned}
&\Lambda_0(\alpha u + (1-\alpha)v) \\
&= \log \mathbb{E}[e^{(\alpha u+(1-\alpha)v)T(X)}|H_0] \\
&\leq \alpha \log \mathbb{E}[e^{uT(X)}|H_0] + (1-\alpha)\log \mathbb{E}[e^{vT(X)}|H_0] \\
&= \alpha\Lambda_0(u) + (1-\alpha)\Lambda_1(u). \qquad (20)
\end{aligned}$$

The proof is similar for $\Lambda_1(u)$. □

Property 2 (Zero at origin)**.** *The log-generating functions $\Lambda_0(u)$ and $\Lambda_1(u)$ are both 0 at $u = 0$.*

Proof of Property 2. The proof follows by substituting $u = 0$ in the expressions of $\Lambda_0(u)$ and $\Lambda_1(u)$. □

Next, we prove some properties for the log-generating functions when the detector is *well-behaved.* In general, when using a detector of the form $T(x) \overset{H_1}{\geq} \tau$, we would expect $T(X)$ to be high when H_1 is true, and low when H_0 is true. We call a detector *well-behaved* if $\mathbb{E}[T(X)|H_0] < 0$ and $\mathbb{E}[T(X)|H_1] > 0$. The next property provides more intuition on what the log-generating functions look like for *well-behaved* detectors.

Property 3 (Log-generating functions of well-behaved detectors)**.** *Suppose that $\mathbb{E}[T(X)|H_0] < 0$ and $\mathbb{E}[T(X)|H_1] > 0$, and $P_0(x)$ and $P_1(x)$ are non-zero for all x. Then, the following holds:*

- *$\Lambda_0(u)$ and $\Lambda_1(u)$ are strictly convex.*
- *$\Lambda_0(u) > 0$ if $u < 0$. $\Lambda_1(u) > 0$ if $u > 0$.*

Proof of Property 3. The convexity of $\Lambda_0(u)$ is proved in Property 1. Now $\Lambda_0(u)$ is strictly convex if, for all distinct reals u and v,

$$\Lambda_0(\alpha u + (1-\alpha)v) < \alpha\Lambda_0(u) + (1-\alpha)\Lambda_1(u).$$

For the sake of contradiction, let us assume that there exists u and v with $v > u$ such that,

$$\Lambda_0(\alpha u + (1-\alpha)v) = \alpha\Lambda_0(u) + (1-\alpha)\Lambda_1(u).$$

This indicates that Hölder's inequality holds with exact equality in (19), which could happen if and only if $ae^{uT(x)} = be^{vT(x)}$ almost everywhere with respect to the probability measure $P_0(x)$ for constants a and b, i.e., $(v-u)T(x) = \log a/b$. Thus,

$$\begin{aligned}&\mathbb{E}[T(X)|H_0] \\ &= \frac{1}{(v-u)} \log a/b \\ &= \mathbb{E}[T(X)|H_1],\end{aligned} \tag{21}$$

where the last line holds because $P_1(x)$ and $P_0(x)$ are both non-zero everywhere (absolutely continuous with respect to each other).

But, this is a contradiction since $\mathbb{E}[T(X)|H_0] < 0 < \mathbb{E}[T(X)|H_1]$. Thus, $\Lambda_0(u)$ is strictly convex. A similar proof can be done for $\Lambda_1(u)$.

For proving the next claim, consider the derivative of $\Lambda_0(u)$.

$$\frac{d\Lambda_0(u)}{du} = \frac{\mathbb{E}[e^{uT(X)}T(X)|H_0]}{e^{\Lambda_0(u)}}. \tag{22}$$

The derivative of $\Lambda_0(u)$ at $u = 0$ is given by $\mathbb{E}[T(X)|H_0]$ which is strictly less than 0. Because $\Lambda_0(u)$ is strictly convex in u and $\Lambda_0(0) = 0$, if $\frac{d\Lambda_0(u)}{du}|_{u=0} < 0$, then $\Lambda_0(u) > 0$ for all $u < 0$.

A similar proof holds for the last claim as well, since the derivative of $\Lambda_1(u)$ at $u = 0$ is given by $\mathbb{E}[T(X)|H_1]$ which is strictly greater than 0, and $\Lambda_1(0) = 0$.

□

Next, we examine the properties of the log-generating functions for likelihood ratio detectors. Consider the likelihood ratio detector $T_0(x) = \log \frac{P_1(x)}{P_0(x)}$. The two conditions $\mathbb{E}[T(X)|H_0] < 0$ and $\mathbb{E}[T(X)|H_1] > 0$ become equivalent to $\mathrm{D}(P_0||P_1) > 0$ and $\mathrm{D}(P_1||P_0) > 0$ where $\mathrm{D}(\cdot||\cdot)$ denotes the Kullback-Leibler (KL) divergence between the two distributions $P_0(x)$ and $P_1(x)$. Thus, a likelihood ratio detector always satisfies these conditions as long as the KL divergences are well-defined and non-zero.

Property 4. *(Log-generating functions of likelihood ratio detectors) Let* $T_0(x) = \log \frac{P_1(x)}{P_0(x)}$*, and* $P_0(x)$ *and* $P_1(x)$ *be non-zero for all* x *with* $\mathrm{D}(P_0||P_1)$ *and* $\mathrm{D}(P_1||P_0)$ *strictly greater than* 0*. Then, the following properties hold:*

- $\Lambda_0(u)$ *is* 0 *at* $u = 0$ *and* 1*.*
- $\Lambda_1(u)$ *is* 0 *at* $u = 0$ *and* -1*.*
- $\Lambda_1(u) = \Lambda_0(u+1)$*.*
- $\mathrm{C}(P_0, P_1) > 0$*.*
- $\Lambda_0(u)$ *and* $\Lambda_1(u)$ *are continuous, differentiable and strictly convex.*
- *The derivatives of* $\Lambda_0(u)$ *and* $\Lambda_1(u)$ *are continuous, monotonically increasing and take all values between* $-\infty$ *and* ∞*.*
- $\Lambda_0(u)$ *attains its global minima for* u *in* $(0,1)$*.*
- $\Lambda_1(u)$ *attains its global minima for* u *in* $(-1,0)$*.*

We first introduce the arithmetic mean-geometric mean (AM-GM) inequality.

Lemma 4 (AM-GM inequality)**.** *The following inequality is satisfied for $u \in (0,1)$ and $a, b \geq 0$:*

$$a^{1-u}b^u \leq (1-u)a + ub, \tag{23}$$

where the equality holds if and only if $a = b$.

Proof of Property 4. The first two claims can be verified by direct substitution.

To show that $\Lambda_1(u) = \Lambda_0(u+1)$, observe that,

$$\begin{aligned}\Lambda_1(u) &= -\log \sum_x P_1(x)^{1+u} P_0(x)^u \\ &= -\log \sum_x P_1(x)^{1+u} P_0(x)^{1-(1+u)} = \Lambda_0(u+1).\end{aligned}$$

Next, we will show that $\mathrm{C}(P_0, P_1) > 0$. Observe that, $\mathrm{C}(P_0, P_1) = -\log \sum_x P_0(x)^{1-u^*} P_1(x)^{u^*}$ for some $u^* \in (0,1)$. From the AM-GM inequality (Lemma 4), there exists at least one x' with $P_0(x) > 0$ and $P_1(x) > 0$ such that,

$$P_0(x')^{1-u^*} P_1(x')^{u^*} < (1-u^*)P_0(x') + u^* P_1(x'),$$

where the inequality is strict because $P_0(x') \neq P_1(x')$ for at least one x' since $\mathrm{D}(P_0||P_1) > 0$ and $\mathrm{D}(P_1||P_0) > 0$.

For all other $x \neq x'$,

$$P_0(x)^{1-u^*} P_1(x)^{u^*} \leq (1-u^*)P_0(x) + u^* P_1(x).$$

Thus,

$$\begin{aligned}&\sum_x P_0(x)^{1-u^*} P_1(x)^{u^*} \\ &< \sum_x (1-u^*)P_0(x) + u^* P_1(x) = 1 \\ &\implies -\log \sum_x P_0(x)^{1-u^*} P_1(x)^{u^*} > 0.\end{aligned} \tag{24}$$

Thus, $\mathrm{C}(P_0, P_1) > 0$. A similar proof extends for continuous distributions as well where the strict inequality holds at least over a set of x's that is not measure 0.

We move on to the next claim. Since both $P_0(x)$ and $P_1(x)$ are strictly greater than 0 for all x, we have $P_0(x)^{1-u} P_1(x)^u$ to be well-defined and continuous for all values of u, including $u = 0$ and $u = 1$. Thus, $\Lambda_0(u)$ is continuous over the range $(-\infty, \infty)$.

The derivative of $\Lambda_0(u)$ is given by:

$$\frac{d\Lambda_0(u)}{du} = \frac{\sum_x P_0(x)^{1-u} P_1(x)^u \log \frac{P_1(x)}{P_0(x)}}{e^{\Lambda_0(u)}}, \tag{25}$$

which is well-defined for all values of u.

The strict convexity of $\Lambda_0(u)$ can be proved using Property 3, because the two conditions $\mathbb{E}[T(X)|H_0] < 0$ and $\mathbb{E}[T(X)|H_1] > 0$ become equivalent to $\mathrm{D}(P_0||P_1) > 0$ and $\mathrm{D}(P_1||P_0) > 0$.

A similar proof extends to $\Lambda_1(u)$.

Now, we move on to the next claim. Observe from (25) that, the derivative is also continuous for all values of u since both $P_0(x)$ and $P_1(x)$ are strictly greater than 0 for all x. It is monotonically increasing because $\Lambda_0(u)$ is strictly convex. Also note that, as $u \to -\infty$, its derivative tends to $-\infty$. Similarly, as $u \to \infty$, its derivative tends to ∞.

A similar proof extends to $\Lambda_1(u)$.

Lastly, because $\Lambda_0(u)$ is 0 at $u = 0$ and $u = 1$, and is a continuous and strictly convex function, it attains its minima for u in $(0, 1)$.

A similar proof extends to $\Lambda_1(u)$, validating the last claim as well. □

Property 5 (Connection to FL transforms)**.** *For well-behaved detectors, the following properties hold:*

- *If* $\tau < \mathbb{E}[T(X)|H_1]$*, then* $\sup_{u<0} (u\tau - \Lambda_1(u)) = \sup_{u\in\mathbb{R}} (u\tau - \Lambda_1(u))$.
- *If* $\tau > \mathbb{E}[T(X)|H_0]$*, then* $\sup_{u>0} (u\tau - \Lambda_0(u)) = \sup_{u\in\mathbb{R}} (u\tau - \Lambda_0(u))$.

Before the proof, we introduce a lemma that will be used in the proof.

Lemma 5 (Supporting line of a strictly convex function)**.** *For a strictly convex and differentiable function* $f(u) : \mathcal{R} \to \mathcal{R}$,

$$u_a \frac{df(u)}{du}|_{u=u_a} - f(u_a) = \sup_{u\in\mathcal{R}} \left(u\frac{df(u)}{du}|_{u=u_a} - f(u) \right).$$

The proof holds from the definition of strict convexity.

Proof of Property 5.

$$\begin{aligned}
&\sup_{u\in\mathcal{R}} (u\tau - \Lambda_1(u)) \\
&\overset{(a)}{=} \sup_{u\in\mathcal{R}} \left(u\frac{d\Lambda_1(u)}{du}|_{u=u_a} - \Lambda_1(u) \right) \\
&\overset{(b)}{=} u_a \frac{d\Lambda_1(u)}{du}|_{u=u_a} - \Lambda_1(u_a) \\
&\overset{(c)}{=} \sup_{u<0} \left(u\frac{d\Lambda_1(u)}{du}|_{u=u_a} - \Lambda_1(u) \right) \\
&\overset{(d)}{=} \sup_{u<0} (u\tau - \Lambda_1(u)) .
\end{aligned} \tag{26}$$

Here (a) holds because the derivative of $\Lambda_1(u)$ is continuous, monotonically increasing and takes all values from $(-\infty, \infty)$ (see Property 4). Thus, for any τ, there exists a single u_a such that $\frac{d\Lambda_1(u)}{du}|_{u=u_a} = \tau$. Next, (b) holds from Lemma 5, whereas (c) holds because $\frac{d\Lambda_1(u)}{du}|_{u=u_a} = \tau < \mathbb{E}[T(X)|H_1] = \frac{d\Lambda_1(u)}{du}|_{u=0}$ and the derivative is monotonically increasing (see Property 4) implying $u_a < 0$. Lastly (d) holds by again substituting $\tau = \frac{d\Lambda_1(u)}{du}|_{u=u_a}$.

Similarly,

$$\begin{aligned}
&\sup_{u\in\mathcal{R}} (u\tau - \Lambda_0(u))\\
&\overset{(a)}{=} \sup_{u\in\mathcal{R}} \left(u\frac{d\Lambda_0(u)}{du}|_{u=u_a} - \Lambda_0(u)\right)\\
&\overset{(b)}{=} u_a \frac{d\Lambda_0(u)}{du}|_{u=u_a} - \Lambda_0(u_a)\\
&\overset{(c)}{=} \sup_{u>0} \left(u\frac{d\Lambda_0(u)}{du}|_{u=u_a} - \Lambda_0(u)\right)\\
&\overset{(d)}{=} \sup_{u>0} (u\tau - \Lambda_0(u)). \qquad (27)
\end{aligned}$$

Here (a) holds because the derivative of $\Lambda_0(u)$ is continuous, monotonically increasing and takes all values from $(-\infty,\infty)$ (see Property 4). Thus, for any τ, there exists a single u_a such that $\frac{d\Lambda_0(u)}{du}|_{u=u_a} = \tau$. Next, (b) holds from Lemma 5, whereas (c) holds because $\frac{d\Lambda_0(u)}{du}|_{u=u_a} = \tau > \mathbb{E}[T(X)|H_0] = \frac{d\Lambda_0(u)}{du}|_{u=0}$ and the derivative is monotonically increasing (see Property 4) implying $u_a > 0$. Lastly (d) holds by again substituting $\tau = \frac{d\Lambda_0(u)}{du}|_{u=u_a}$.

□

A.3 Log Generating Functions for Gaussians

Let $P_0(x) \sim \mathcal{N}(\mu_0, \sigma^2)$ and $P_1(x) \sim \mathcal{N}(\mu_1, \sigma^2)$. We derive the log-generating functions for likelihood ratio detectors corresponding to these two distributions.

$$\begin{aligned}
\Lambda_0(u) &= \int \log \frac{P_1(x)}{P_0(x)} P_0(x) dx\\
&= \frac{1}{2\sigma^2}(\mu_0 - \mu_1)^2 u(u-1). \qquad (28)
\end{aligned}$$

B Appendix to Section 3

B.1 Proof of Lemma 3

Proof of Lemma 3. The detector that minimizes the Bayesian probability of error, i.e., $P_e^{(T)}(\tau) = \pi_0 P_{\text{FP}}^{(T)}(\tau) + \pi_1 P_{\text{FN}}^{(T)}(\tau)$ is the likelihood ratio detector given by $T(x) = \log \frac{P_1(x)}{P_0(x)} \overset{H_1}{\geq} 0$ (for $\pi_0 = \pi_1 = \frac{1}{2}$). The proof is available in [52, Theorem 3.1].

Here, we will show that the Chernoff exponent of the probability of error for this detector, i.e., $E_e^{(T)}(0)$ is equal to $\mathrm{C}(P_0, P_1) = -\min_{u\in(0,1)} \log \sum_x P_0(x)^{(1-u)} P_1(x)^u$.

Note that,

$$\begin{aligned}
E_{\text{FP}}^{(T)}(0) &= \sup_{u>0} -\Lambda_0(u)\\
&= -\min_{u\in(0,1)} \log \sum_x P_0(x)^{(1-u)} P_1(x)^u,
\end{aligned}$$

where the last line follows because $\Lambda_0(u)$ attains its minima in the range $u \in (0,1)$ (see Property 4).

$$\begin{aligned}E^{(T)}_{\mathrm{FN}}(0) &= \sup_{u<0} -\Lambda_1(u)\\ &\overset{(a)}{=} -\min_{u\in(-1,0)} \log \sum_x P_0(x)^{(-u)} P_1(x)^{(1+u)}\\ &= -\min_{u'=u+1\in(0,1)} \log \sum_x P_0(x)^{(1-u')} P_1(x)^{(u')},\end{aligned}$$

where (a) also holds because $\Lambda_1(u)$ attains its minima in the range $u \in (-1,0)$ (see Property 4).

Lastly,

$$E^{(T)}_e(0) = \min\{E^{(T)}_{\mathrm{FP}}(0), E^{(T)}_{\mathrm{FN}}(0)\} = \mathrm{C}(P_0,P_1). \tag{29}$$

□

B.2 Proofs of Theorem 1 and Theorem 2

Before the proofs, we introduce a lemma that will be used in the proofs.

Lemma 6. *Let $P_0(x)$ and $P_1(x)$ be non-zero for all x and $\mathrm{D}(P_0||P_1)$ and $\mathrm{D}(P_1||P_0)$ be strictly greater than 0. For likelihood ratio detectors of the form $T_0(x) = \log \frac{P_1(x)}{P_0(x)} \overset{H_1}{\geq} \tau_0$, if $\tau_0 \neq 0$, then one of the following statements is true:*

$E^{(T_0)}_{\mathrm{FN}}(\tau_0) < \mathrm{C}(P_0,P_1) < E^{(T_0)}_{\mathrm{FP}}(\tau_0)$, or

$E^{(T_0)}_{\mathrm{FP}}(\tau_0) < \mathrm{C}(P_0,P_1) < E^{(T_0)}_{\mathrm{FN}}(\tau_0)$.

Proof of Lemma 6. Let us analyze the scenario where $\tau_0 > 0$. Observe that,

$$\begin{aligned}E^{(T_0)}_{\mathrm{FP}}(\tau_0) &= \sup_{u>0}(u\tau_0 - \Lambda_0(u))\\ &\geq u_0^*\tau_0 - \Lambda_0(u_0^*) && \text{[for any } u_0^*>0]\\ &> -\Lambda_0(u_0^*) && \text{[since } u_0^*\tau_0>0]\\ &\overset{(a)}{=} \mathrm{C}(P_0,P_1),\end{aligned} \tag{30}$$

where (a) follows if we choose $u_0^* = \arg\min \Lambda_0(u)$ (from Property 4, $\Lambda_0(u)$ attains its minima for some $u \in (0,1)$) and $\Lambda_0(u_0^*) = -\mathrm{C}(P_0,P_1)$ (by definition).

Now, we will show that $E^{(T_0)}_{\mathrm{FN}}(\tau_0) < \mathrm{C}(P_0,P_1)$ when $\tau_0 > 0$.

Case 1: $\tau_0 \geq \frac{d\Lambda_1(u)}{du}|_{u=0} = \mathrm{D}(P_1||P_0)$

$$\begin{aligned}E^{(T_0)}_{\mathrm{FN}}(\tau_0) &= \sup_{u<0}(u\tau_0 - \Lambda_1(u))\\ &\leq \sup_{u<0}(u\mathrm{D}(P_1||P_0) - \Lambda_1(u)) \text{ [since } \tau_0 \geq \mathrm{D}(P_1||P_0)]\\ &\leq \sup_{u\in\mathcal{R}}(u\mathrm{D}(P_1||P_0) - \Lambda_1(u))\\ &\overset{(a)}{=} (0\cdot \mathrm{D}(P_1||P_0) - \Lambda_1(0))\\ &\overset{(b)}{=} 0\\ &\overset{(c)}{<} \mathrm{C}(P_0,P_1),\end{aligned} \tag{31}$$

where (a) holds from Lemma 5 because $\frac{d\Lambda_1(u)}{du}|_{u=0} = \mathrm{D}(P_1||P_0)$, and (b) and (c) hold from Property 4 since $\Lambda_1(0) = 0$ and $\mathrm{C}(P_0, P_1) > 0$.

Case 2: $0 < \tau_0 < \frac{d\Lambda_1(u)}{du}|_{u=0} = \mathrm{D}(P_1||P_0)$

$$\begin{aligned}
&E_{\mathrm{FN}}^{(T_0)}(\tau_0) = \sup_{u<0}(u\tau_0 - \Lambda_1(u)) \\
&\leq \sup_{u\in\mathcal{R}}(u\tau_0 - \Lambda_1(u)) \\
&\overset{(a)}{=} \sup_{u\in\mathcal{R}}(u\tau_0 - \Lambda_1(u)) \quad [\text{where } \frac{d\Lambda_1(u)}{du}|_{u=u_a} = \tau_0] \\
&\overset{(b)}{=} u_a\tau_0 - \Lambda_1(u_a) \\
&\overset{(c)}{<} -\Lambda_1(u_a) \qquad [\text{since } u_a\tau_0 < 0] \\
&\leq -\min_u \Lambda_1(u) \\
&\overset{(d)}{=} -\min_{u\in(-1,0)} \Lambda_1(u) = \mathrm{C}(P_0, P_1) \qquad (32)
\end{aligned}$$

Here, (a) holds because the derivative of $\Lambda_1(u)$ is continuous, monotonically increasing and takes all values from $-\infty$ to ∞ (see Property 4). Thus, for any τ_0, there exists a single u_a such that $\frac{d\Lambda_1(u)}{du}|_{u=u_a} = \tau_0$. Next, (b) holds from Lemma 5, (c) holds because $\frac{d\Lambda_1(u)}{du}|_{u=u_a} = \tau_0 < \frac{d\Lambda_1(u)}{du}|_{u=0}$, and the derivative is monotonically increasing, implying $u_a < 0$. Lastly (d) holds because $\Lambda_1(u)$ attains its minima in the range $u \in (-1, 0)$ (see Property 4).

Thus, for $\tau_0 > 0$, we get
$E_{\mathrm{FN}}^{(T_0)}(\tau_0) < \mathrm{C}(P_0, P_1) < E_{\mathrm{FP}}^{(T_0)}(\tau_0)$.

The proof is similar for the scenario where $\tau_0 < 0$, and leads to $E_{\mathrm{FP}}^{(T_0)}(\tau_0) < \mathrm{C}(P_0, P_1) < E_{\mathrm{FN}}^{(T_0)}(\tau_0)$. □

Proof of Theorem 1. The first claim follows directly from Lemma 3 by choosing the likelihood ratio detectors for the two groups with thresholds $\tau_0 = \tau_1 = 0$, i.e., the Bayes optimal detector under equal priors.

To prove the next claim, assume (for the sake of contradiction) that there is a likelihood ratio detector such that $E_e^{(T_0)}(\tau_0) > \mathrm{C}(P_0, P_1)$.

Now, if $\tau_0 = 0$, then we have $E_e^{(T_0)}(\tau_0) = \mathrm{C}(P_0, P_1)$ (from Lemma 3). Alternately, if $\tau_0 \neq 0$, then we either have $E_{\mathrm{FN}}^{(T_0)}(\tau_0) < \mathrm{C}(P_0, P_1) < E_{\mathrm{FP}}^{(T_0)}(\tau_0)$ or $E_{\mathrm{FP}}^{(T_0)}(\tau_0) < \mathrm{C}(P_0, P_1) < E_{\mathrm{FN}}^{(T_0)}(\tau_0)$ (from Lemma 6). Thus,

$$E_e^{(T_0)}(\tau_0) = \min\{E_{\mathrm{FP}}^{(T_0)}(\tau_0), E_{\mathrm{FN}}^{(T_0)}(\tau_0)\} < \mathrm{C}(P_0, P_1). \qquad (33)$$

For both cases, we have a contradiction, implying that $E_e^{(T_0)}(\tau_0) \leq \mathrm{C}(P_0, P_1) < \mathrm{C}(Q_0, Q_1)$ for all likelihood ratio detectors $T_0 \overset{H_1}{\geq} \tau_0$. □

Proof of Theorem 2. In Lemma 6, we show that for likelihood ratio detectors of the form $T_0(x) = \log \frac{P_1(x)}{P_0(x)} \overset{H_1}{\geq} \tau_0$, if $\tau_0 \neq 0$, then we either have $E_{\mathrm{FN}}^{(T_0)}(\tau_0) < \mathrm{C}(P_0, P_1) < E_{\mathrm{FP}}^{(T_0)}(\tau_0)$ or $E_{\mathrm{FP}}^{(T_0)}(\tau_0) < \mathrm{C}(P_0, P_1) < E_{\mathrm{FN}}^{(T_0)}(\tau_0)$.

Thus, $\max_{\tau_0} E_e^{(T_0)}(\tau_0) = \max_{\tau_0} \min\{E_{\mathrm{FP}}^{(T_0)}(\tau_0), E_{\mathrm{FN}}^{(T_0)}(\tau_0)\}$ is attained when both $E_{\mathrm{FN}}^{(T_0)}(\tau_0)$ and $E_{\mathrm{FP}}^{(T_0)}(\tau_0)$ are equal to $\mathrm{C}(P_0, P_1)$, which holds for $\tau_0 = 0$ (proved in Lemma 3).

Similarly, $\max_{\tau_1} \min\{E_{\text{FP}}^{(T_1)}(\tau_1), E_{\text{FN}}^{(T_1)}(\tau_1)\}$ is attained when both $E_{\text{FN}}^{(T_1)}(\tau_1)$ and $E_{\text{FP}}^{(T_1)}(\tau_1)$ are equal to $\mathrm{C}(Q_0, Q_1)$, which holds for $\tau_1 = 0$.

Now, suppose there exists two detectors for the two classes such that, $E_{\text{FN}}^{(T_0)}(\tau_0) = E_{\text{FN}}^{(T_1)}(\tau_1)$. Since $\mathrm{C}(P_0, P_1) < \mathrm{C}(Q_0, Q_1)$, *at most* one of the two exponents $E_{\text{FN}}^{(T_0)}(\tau_0)$ and $E_{\text{FN}}^{(T_1)}(\tau_1)$ can be equal to their corresponding Chernoff information $\mathrm{C}(P_0, P_1)$ or $\mathrm{C}(Q_0, Q_1)$. Without loss of generality, we may assume that $E_{\text{FN}}^{(T_0)}(\tau_0) \neq \mathrm{C}(P_0, P_1)$. Therefore, $\tau_0 \neq 0$, which implies (from Lemma 6) that we either have $E_{\text{FN}}^{(T_0)}(\tau_0) < \mathrm{C}(P_0, P_1) < E_{\text{FP}}^{(T_0)}(\tau_0)$ or $E_{\text{FP}}^{(T_0)}(\tau_0) < \mathrm{C}(P_0, P_1) < E_{\text{FN}}^{(T_0)}(\tau_0)$. Thus,

$$E_e^{(T_0)}(\tau_0) = \min\{E_{\text{FP}}^{(T_0)}(\tau_0), E_{\text{FN}}^{(T_0)}(\tau_0)\} < \mathrm{C}(P_0, P_1). \tag{34}$$

□

B.3 Proofs of Proposition 1 and Proposition 2

Proof of Proposition 1. Let $\tau_0^* = 0$. Using Lemma 3, this ensures,

$$E_{\text{FN}}^{(T_0)}(0) = E_{\text{FP}}^{(T_0)}(0) = \mathrm{C}(P_0, P_1).$$

Now, the only value of τ_1^* that will satisfy $E_{\text{FN}}^{(T_1)}(\tau_1^*) = E_{\text{FN}}^{(T_0)}(0)$ is a $\tau_1^* > 0$ such that $E_{\text{FN}}^{(T_1)}(\tau_1^*) = \mathrm{C}(P_0, P_1)$. To prove that such a τ_1^* exists, consider the function:

$$g(u) = u\frac{d\Lambda_1(u)}{d(u)} - \Lambda_1(u),$$

where $\Lambda_1(u)$ is the log-generating transform for $z = 1$. The function $g(u)$ is continuous. At $u = 0$, $g(u) = 0$ and at $u = u_1^*$ (where $u_1^* = \arg\min \Lambda_1(u)$ and lies in $(-1, 0)$ from Property 4) we have $g(u) = \mathrm{C}(Q_0, Q_1)$. Because $g(u)$ is continuous, there exists a $u_a \in (u_1^*, 0)$ such that $g(u_a) = \mathrm{C}(P_0, P_1)$ which lies between 0 and $\mathrm{C}(Q_0, Q_1)$. If we set $\tau_1^* = \frac{d\Lambda_1(u)}{d(u)}|_{u=u_a}$, we have (using Lemma 5)

$$\mathrm{C}(P_0, P_1) = g(u_a) = \sup_{u<0}(u\tau_1^* - \Lambda_1(u)).$$

Also note that $\tau_1^* > 0$ because the derivative of $\Lambda_1(u)$ is monotonically increasing and $u_a > u_1^*$, leading to $\tau_1^* = \frac{d\Lambda_1(u)}{d(u)}|_{u=u_a} > \frac{d\Lambda_1(u)}{d(u)}|_{u=u_1^*} = 0$.

Now that we have a τ_1^* such that $E_{\text{FN}}^{(T_1)}(\tau_1^*) = \mathrm{C}(P_0, P_1)$ which is strictly less that $\mathrm{C}(Q_0, Q_1)$, we must have $E_{\text{FP}}^{(T_1)}(\tau_1^*) > \mathrm{C}(Q_0, Q_1)$ (from Lemma 6).

This leads to,

$$\min\{E_{\text{FP}}^{T_0}(0), E_{\text{FN}}^{T_0}(0), E_{\text{FP}}^{T_1}(\tau_1^*), E_{\text{FN}}^{T_1}(\tau_1^*)\} = \mathrm{C}(P_0, P_1).$$

For any other choice of $\tau_0^* \neq 0$, we either have $E_{\text{FP}}^{(T_0)}(\tau_0^*) < \mathrm{C}(P_0, P_1) < E_{\text{FN}}^{(T_0)}(\tau_0^*)$, or $E_{\text{FN}}^{(T_0)}(\tau_0^*) < \mathrm{C}(P_0, P_1) < E_{\text{FP}}^{(T_0)}(\tau_0^*)$, implying

$$\min\{E_{\text{FP}}^{T_0}(\tau_0^*), E_{\text{FN}}^{T_0}(\tau_0^*), E_{\text{FP}}^{T_1}(\tau_1^*), E_{\text{FN}}^{T_1}(\tau_1^*)\} < \mathrm{C}(P_0, P_1).$$

□

Proof of Proposition 2. We are given that,

$$E_{\text{FN}}^{(T_1)}(\tau_1) = E_{\text{FP}}^{(T_1)}(\tau_1) = \mathrm{C}(Q_0, Q_1).$$

Now, the only value of τ_0^* that will satisfy $E_{\mathrm{FN}}^{(T_0)}(\tau_0^*) = \mathrm{C}(Q_0, Q_1)$ is a $\tau_0^* < 0$. To prove that such a τ_0^* exists, consider the function

$$g(u) = u\frac{d\Lambda_1(u)}{d(u)} - \Lambda_1(u),$$

where $\Lambda_1(u)$ is the log-generating transform for $z = 0$. The function $g(u)$ is continuous. At $u = u_1^*$ (where $u_1^* = \arg\min \Lambda_1(u)$ and lies in $(-1, 0)$ from Property 4), we have $g(u_1^*) = \mathrm{C}(P_0, P_1)$ and as $u \to -\infty$, we have $g(u) \to \infty$. Because $g(u)$ is continuous, there exists a $u_a \in (-\infty, u_1^*)$ such that $g(u_a) = \mathrm{C}(Q_0, Q_1)$ which lies between $\mathrm{C}(P_0, P_1)$ and ∞. If we set $\tau_0^* = \frac{d\Lambda_1(u)}{d(u)}|_{u=u_a}$, we have (using Lemma 5)

$$\mathrm{C}(Q_0, Q_1) = g(u_a) = \sup_{u<0}(u\tau_0^* - \Lambda_1(u)).$$

This τ_0^* is less than 0 because the derivative of $\Lambda_1(u)$ is monotonically increasing and $u_a < u_1^*$, leading to $\tau_0^* = \frac{\Lambda_1(u)}{d(u)}|_{u=u_a} < \frac{\Lambda_1(u)}{d(u)}|_{u=u_1^*} = 0$.

Now that we have a τ_0^* such that $E_{\mathrm{FN}}^{(T_0)}(\tau_0^*) = \mathrm{C}(Q_0, Q_1)$ which is strictly greater that $\mathrm{C}(P_0, P_1)$, we must have $E_{\mathrm{FP}}^{(T_0)}(\tau_0^*) < \mathrm{C}(P_0, P_1)$ (from Lemma 6).

This leads to,

$$\min\{E_{\mathrm{FP}}^{T_0}(\tau_0^*), E_{\mathrm{FN}}^{T_0}(\tau_0^*)\} < \mathrm{C}(P_0, P_1).$$

□

C Appendix to Section 4

Proof of Theorem 3. From Proposition 2, there exists a likelihood ratio detector of the form $T_0(x) = \log \frac{P_1(x)}{P_0(x)} \overset{H_1}{\geq} \tau_0^*$ such that

$$E_{\mathrm{FN}}^{T_0}(\tau_0^*) = \mathrm{C}(Q_0, Q_1). \tag{35}$$

In the proof of Proposition 2, we showed that this $\tau_0^* < 0$.

Now, we will show that there exists $\widetilde{P}_0(x)$ and $\widetilde{P}_1(x)$ such that their optimal detector $\widetilde{T}(x) = \log \frac{\widetilde{P}_1(x)}{\widetilde{P}_0(x)} \overset{H_1}{\geq} 0$ is equivalent to the detector $T_0(x) \overset{H_1}{\geq} \tau_0^*$.

Let $\widetilde{P}_0(x) = \frac{P_0(x)^{(1-w)}P_1(x)^w}{\sum_x P_0(x)^{(1-w)}P_1(x)^w}$ 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 some $w, v \in \mathcal{R}$ with $w \neq v$. Observe that,

$$\begin{aligned}
\widetilde{T}(x) &= \log \frac{\widetilde{P}_1(x)}{\widetilde{P}_0(x)} \\
&= (v - w)\log \frac{P_1(x)}{P_0(x)} + \log \frac{\sum_x P_0(x)^{(1-w)}P_1(x)^w}{\sum_x P_0(x)^{(1-v)}P_1(x)^v} \\
&= (v - w)\log \frac{P_1(x)}{P_0(x)} + \Lambda_0(w) - \Lambda_0(v) \\
&= (v - w)\left(\log \frac{P_1(x)}{P_0(x)} - \frac{\Lambda_0(v) - \Lambda_0(w)}{v - w}\right).
\end{aligned} \tag{36}$$

Because $\Lambda_0(u)$ is strictly convex with its derivative taking all values from $-\infty$ to ∞, one can always find a tangent to $\Lambda_0(u)$ that has a slope τ_0^* at (say) $u = u_a$. Thus, one can always find pairs of points (w, v) on either sides of $u = u_a$ such that $\tau_0^* = \frac{\Lambda_0(v) - \Lambda_0(w)}{v - w}$, which are essentially pairs of points (w, v) at which a straight line with slope τ_0^* cuts $\Lambda_0(u)$. In particular, we can fix $v = 1$ and always find a $w < 0$ such that

$$\tau_0^* = \frac{\Lambda_0(v) - \Lambda_0(w)}{v - w} = \frac{-\Lambda_0(w)}{1 - w}, \tag{37}$$

because $\Lambda_0(u)$ is continuous taking values 0 at $u = 0$ and $u = 1$, and takes all values from $(0, \infty)$ in the range $(-\infty, 0)$.

Thus, the first claim is proved.

Now, we calculate $\mathrm{C}(\widetilde{P}_0, \widetilde{P}_1)$.

$$\begin{aligned}
&\mathrm{C}(\widetilde{P}_0, \widetilde{P}_1)\\
&= \max_{u \in (0,1)} - \log \sum_x \widetilde{P}_0(x)^{1-u} \widetilde{P}_1(x)^u\\
&\overset{(a)}{=} \max_{u \in \mathcal{R}} - \log \sum_x \widetilde{P}_0(x)^{1-u} \widetilde{P}_1(x)^u\\
&\overset{(b)}{=} \max_{u \in \mathcal{R}} - \log \sum_x P_0(x)^{(1-w)(1-u)} P_1(x)^{w(1-u)+u}\\
&\qquad\qquad\qquad + (1-u)\Lambda_0(w)\\
&\overset{(c)}{=} \max_{u \in \mathcal{R}} - \log \sum_x P_0(x)^{(1-w)(1-u)} P_1(x)^{w(1-u)+u}\\
&\qquad\qquad\qquad + (1-u)(w-1)\tau_0^*\\
&\overset{(d)}{=} \max_{u \in \mathcal{R}} (1-u)(w-1)\tau_0^* - \Lambda_1((1-u)(w-1))\\
&\overset{(e)}{=} \sup_{u' \in \mathcal{R}} (u'\tau_0^* - \Lambda_1(u')) \quad [u' = (1-u)(w-1)]\\
&\overset{(f)}{=} \sup_{u' < 0} (u'\tau_0^* - \Lambda_1(u')) \quad [u' = (1-u)(w-1)]\\
&\overset{(g)}{=} \mathrm{C}(Q_0, Q_1).
\end{aligned} \tag{38}$$

Here (a) holds because the log-generating function $-\log \sum_x \widetilde{P}_0(x)^{1-u} \widetilde{P}_1(x)^u$ of a likelihood ratio detector attains its global minima at $(0, 1)$ (see Property 4) and (b) holds by substituting $\widetilde{P}_0(x) = \frac{P_0(x)^{(1-w)} P_1(x)^w}{\sum_x P_0(x)^{(1-w)} P_1(x)^w}$ and $\widetilde{P}_1(x) = \frac{P_0(x)^{(1-v)} P_1(x)^v}{\sum_x P_0(x)^{(1-v)} P_1(x)^v}$ with $v = 1$. Next, (c) holds by using $\tau_0^* = \frac{\Lambda_0(v) - \Lambda_0(w)}{v-w} = \frac{-\Lambda_0(w)}{1-w}$ (see (37)), (d) holds from the definition of $\Lambda_1((1-u)(w-1))$, (e) holds by a change of variable $u' = (1-u)(w-1)$, (f) holds because $\tau_0^* < 0 \leq \mathrm{D}(\widetilde{P}_1 || \widetilde{P}_0) = \mathbb{E}[\widetilde{T}(X) | \widetilde{H}_1]$ and the detector is well-behaved (see Property 5), and lastly (g) holds because $E_{\mathrm{FN}}^{T_0}(\tau_0^*) = \mathrm{C}(Q_0, Q_1)$ (see (35)). □

D Appendix to Section 5

D.1 Proof of Theorem 4

Proof of Theorem 4. We remind the readers that,

$$\frac{W_0(x,x')}{P_0(x)} = \Pr\left(X' = x' | X = x, Z = 0, Y = 0\right), \tag{39}$$

$$\frac{W_1(x,x')}{P_1(x)} = \Pr\left(X' = x' | X = x, Z = 0, Y = 1\right). \tag{40}$$

We would like to prove:
$I(X';Y|X,Z=0) > 0 \implies \mathrm{C}(W_0,W_1) > \mathrm{C}(P_0,P_1)$.

Suppose that X' is not independent of Y given X and $Z=0$, i.e., $I(X';Y|X,Z=0) > 0$. This implies that there exists at least one $X = x_a$ such that the distributions of $X'|_{X=x_a,Z=0,Y=0}$ and $X'|_{X=x_a,Z=0,Y=1}$ are different. Therefore, there exists at least one pair $(x',x) = (x'_a, x_a)$ for which the following AM-GM inequality (Lemma 4) holds with strict inequality for all $u \in (0,1)$, i.e,

$$\begin{aligned}&\left(\frac{W_0(x_a,x'_a)}{P_0(x_a)}\right)^{1-u}\left(\frac{W_1(x_a,x'_a)}{P_1(x_a)}\right)^{u}\\ &< (1-u)\frac{W_0(x_a,x'_a)}{P_0(x_a)} + u\frac{W_1(x_a,x'_a)}{P_1(x_a)}.\end{aligned} \tag{41}$$

For all other $(x',x) \neq (x'_a,x_a)$, we have (from the AM-GM inequality in Lemma 4):

$$\begin{aligned}&\left(\frac{W_0(x,x')}{P_0(x)}\right)^{1-u}\left(\frac{W_1(x,x')}{P_1(x)}\right)^{u}\\ &\leq (1-u)\frac{W_0(x,x')}{P_0(x)} + u\frac{W_1(x,x')}{P_1(x)}.\end{aligned} \tag{42}$$

Using (41) and (42),

$$\begin{aligned}&\sum_{x'}\left(\frac{W_0(x_a,x')}{P_0(x_a)}\right)^{1-u}\left(\frac{W_1(x_a,x')}{P_1(x_a)}\right)^{u}\\ &< \sum_{x'}\left((1-u)\frac{W_0(x_a,x')}{P_0(x_a)} + u\frac{W_1(x_a,x')}{P_1(x_a)}\right) = 1.\end{aligned} \tag{43}$$

This leads to,

$$\sum_{x'} W_0(x_a,x')^{1-u} W_1(x_a,x')^{u} < P_0(x_a)^{1-u} P_1(x_a)^{u}. \tag{44}$$

For all other $x \neq x_a$, we have (using (42) alone),

$$\begin{aligned}&\sum_{x'}\left(\frac{W_0(x,x')}{P_0(x)}\right)^{1-u}\left(\frac{W_1(x,x')}{P_1(x)}\right)^{u}\\ &\leq \sum_{x'}\left((1-u)\frac{W_0(x,x')}{P_0(x)} + u\frac{W_1(x,x')}{P_1(x)}\right) = 1,\end{aligned} \tag{45}$$

leading to

$$\sum_{x'} W_0(x,x')^{1-u} W_1(x,x')^u \leq P_0(x)^{1-u} P_1(x)^u. \tag{46}$$

Lastly, using (44) and (46),

$$\sum_x \sum_{x'} W_0(x,x')^{1-u} W_1(x,x')^u < \sum_x P_0(x)^{1-u} P_1(x)^u, \tag{47}$$

leading to the claim:

$$\begin{aligned}
&\mathrm{C}(W_0, W_1)\\
&= -\min_{u\in(0,1)} \log \sum_x \sum_{x'} W_0(x,x')^{1-u} W_1(x,x')^u\\
&> -\min_{u\in(0,1)} \log \sum_x P_0(x)^{1-u} P_1(x)^u\\
&= \mathrm{C}(P_0, P_1).
\end{aligned} \tag{48}$$

We would now like to prove:
$\mathrm{C}(W_0,W_1) > \mathrm{C}(P_0,P_1) \implies I(X';Y|X,Z=0) > 0$, or, $I(X';Y|X,Z=0)$=$0 \Longrightarrow \mathrm{C}(W_0,W_1)\not>\mathrm{C}(P_0,P_1)$.

First note that, from the previous proof, $\mathrm{C}(W_0,W_1) \geq \mathrm{C}(P_0,P_1)$ always holds using the AM-GM inequality. Thus, $\mathrm{C}(W_0,W_1)\not>\mathrm{C}(P_0,P_1)$ is same as $\mathrm{C}(W_0,W_1)$=$\mathrm{C}(P_0,P_1)$.

Suppose that X' is independent of Y given X and $Z=0$.

$$\begin{aligned}
&I(X';Y|X,Z=0) = 0\\
&\Rightarrow \Pr(X'=x'|X,Z=0,Y=0)\\
&\qquad\qquad = \Pr(X'=x'|X,Z=0,Y=1)\ \forall x'\\
&\Rightarrow \frac{W_0(x,x')}{P_0(x)} = \frac{W_1(x,x')}{P_1(x)} \qquad \forall x', x\\
&\Rightarrow \sum_{x'} (\frac{W_0(x,x')}{P_0(x)})^{1-u} (\frac{W_1(x,x')}{P_1(x)})^u = 1\ \forall x\\
&\Rightarrow \sum_x \sum_{x'} W_0(x,x')^{1-u} W_1(x,x')^u\\
&\qquad\qquad = \sum_x P_0(x)^{1-u} P_1(x)^u.
\end{aligned} \tag{49}$$

This leads to

$$\begin{aligned}
&\mathrm{C}(W_0, W_1)\\
&= -\min_{u\in(0,1)} \log \sum_x \sum_{x'} W_0(x,x')^{1-u} W_1(x,x')^u\\
&= -\min_{u\in(0,1)} \log \sum_x P_0(x)^{1-u} P_1(x)^u\\
&= \mathrm{C}(P_0, P_1).
\end{aligned} \tag{50}$$

□

E Unequal Priors

E.1 Proof of Lemma 2

First observe that,

$$\begin{aligned}
&u\tau_0 - \Lambda_0(u) - \log 2\pi_0 \\
&= u(\tau_0 - \log\frac{\pi_0}{\pi_1}) + u\log\frac{\pi_0}{\pi_1} - \Lambda_0(u) - \log 2\pi_0 && (51)\\
&= u\tau' - \widetilde{\Lambda}_0(u) - \log 2, && (52)
\end{aligned}$$

where $\tau' = \tau_0 - \log\frac{\pi_0}{\pi_1}$, and $\widetilde{\Lambda}_0(u) = \Lambda_0(u) - u\log\frac{\pi_0}{\pi_1} + \log\pi_0$. Similarly,

$$\begin{aligned}
&u\tau_0 - \Lambda_1(u) - \log 2\pi_1 \\
&= u(\tau_0 - \log\frac{\pi_0}{\pi_1}) + u\log\frac{\pi_0}{\pi_1} - \Lambda_1(u) - \log 2\pi_1 && (53)\\
&= u\tau' - \widetilde{\Lambda}_1(u) - \log 2, && (54)
\end{aligned}$$

where $\tau' = \tau_0 - \log\frac{\pi_0}{\pi_1}$, and $\widetilde{\Lambda}_1(u) = \Lambda_1(u) - u\log\frac{\pi_0}{\pi_1} + \log\pi_1$.

We first derive some properties of $\widetilde{\Lambda}_0(u)$ and $\widetilde{\Lambda}_1(u)$.

Lemma 7. *Let $P_0(x)$ and $P_1(x)$ be strictly greater than 0 everywhere and $\mathrm{D}(P_0||P_1)$ and $\mathrm{D}(P_1||P_0)$ be strictly greater than 0 and π_0 and π_1 lie in $(0,1)$. Then, the following properties hold:*

- *$\widetilde{\Lambda}_0(u)$ and $\widetilde{\Lambda}_1(u)$ are continuous, differentiable and strictly convex.*
- *The derivatives of $\widetilde{\Lambda}_0(u)$ and $\widetilde{\Lambda}_1(u)$ are continuous, monotonically increasing, and take all values from $-\infty$ to ∞.*
- *$\widetilde{\Lambda}_1(u) = \widetilde{\Lambda}_0(u+1)$.*

Proof of Lemma 7. Note that, $\widetilde{\Lambda}_0(u)$ is the sum of $\Lambda_0(u)$ and an affine function $-u\log\frac{\pi_0}{\pi_1} + \log\pi_0$. Because $\Lambda_0(u)$ is continuous, differentiable and strictly convex (from Property 4), $\widetilde{\Lambda}_0(u)$ also satisfies those properties. The second claim also holds for the same reason because the derivative of $\Lambda_0(u)$ satisfies all these properties (from Property 4).

Lastly,

$$\begin{aligned}
&\widetilde{\Lambda}_0(u+1) = \Lambda_0(u+1) - (u+1)\log\frac{\pi_0}{\pi_1} + \log\pi_0 \\
&= \Lambda_0(u+1) - u\log\frac{\pi_0}{\pi_1} + \log\pi_1 \\
&\overset{(a)}{=} \Lambda_1(u) - u\log\frac{\pi_0}{\pi_1} + \log\pi_1 = \widetilde{\Lambda}_1(u), && (55)
\end{aligned}$$

where (a) holds because $\Lambda_1(u) = \Lambda_0(u+1)$ from Property 4.

□

Proof of Lemma 2. We specifically consider the case where $\pi_0 \neq \pi_1$ in this proof because the case of equal priors $\pi_0 = \pi_1$ can be proved using Lemma 3 and Lemma 6.

Without loss of generality, we assume $\pi_0 > \pi_1$. Thus, $\log\frac{\pi_0}{\pi_1} > 0$.

Case 1: $\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} = \mathrm{D}(P_1||P_0) - \log\frac{\pi_0}{\pi_1} > 0$.

Observe that, $\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=-1} = -\mathrm{D}(P_0||P_1) - \log\frac{\pi_0}{\pi_1} < 0$ and $\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} = \mathrm{D}(P_1||P_0) - \log\frac{\pi_0}{\pi_1} > 0$. Thus, the strictly convex function $\widetilde{\Lambda}_1(u)$ attains its minima in $(-1,0)$ (using Lemma 7). Next, using $\widetilde{\Lambda}_0(u+1) = \widetilde{\Lambda}_1(u)$ (also from Lemma 7), we have $\widetilde{\Lambda}_0(u)$ attaining its minima in $(0,1)$.

For $\tau' = 0$ (equivalently $\tau_0 = \log\frac{\pi_0}{\pi_1}$), we have

$$\begin{aligned}
&E_{\mathrm{FP}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_0 = \\
&\overset{(a)}{=} \sup_{u>0}(u\cdot 0 - \widetilde{\Lambda}_0(u) - \log 2) \\
&\overset{(b)}{=} -\min_u \widetilde{\Lambda}_0(u) - \log 2 \\
&\overset{(c)}{=} -\min_u \widetilde{\Lambda}_1(u) - \log 2 \\
&\overset{(d)}{=} \sup_{u<0}(u\cdot 0 - \widetilde{\Lambda}_1(u) - \log 2) \\
&\overset{(e)}{=} E_{\mathrm{FN}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_1.
\end{aligned} \tag{56}$$

Here, (a) holds from (52), (b) holds because $\widetilde{\Lambda}_0(u)$ attains its minima in $(0,1)$, (c) holds from $\widetilde{\Lambda}_0(u+1) = \widetilde{\Lambda}_1(u)$ (see Lemma 7), (d) holds because $\widetilde{\Lambda}_1(u)$ attains its minima in $(-1,0)$, and (e) holds from (54).

Next, we will show that, for any other value of $\tau' \neq 0$ ($\tau_0 \neq \log\frac{\pi_0}{\pi_1}$), we either have

$$\begin{aligned}
&E_{\mathrm{FP}}^{(T_0)}(\tau_0) - \log 2\pi_0 < E_{\mathrm{FP}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_0 \\
&< E_{\mathrm{FN}}^{(T_0)}(\tau_0) - \log 2\pi_1,
\end{aligned}$$

or,

$$\begin{aligned}
&E_{\mathrm{FN}}^{(T_0)}(\tau_0) - \log 2\pi_1 < E_{\mathrm{FP}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_0 \\
&< E_{\mathrm{FP}}^{(T_0)}(\tau_0) - \log 2\pi_0.
\end{aligned}$$

Let $\tau' > 0$. Then,

$$\begin{aligned}
&E_{\mathrm{FP}}^{(T_0)}(\tau_0) - \log 2\pi_0 \\
&\overset{(a)}{=} \sup_{u>0}(u\tau' - \widetilde{\Lambda}_0(u) - \log 2) \\
&\overset{(b)}{\geq} (u_0^*\tau' - \widetilde{\Lambda}_0(u_0^*) - \log 2) \\
&\overset{(c)}{>} -\widetilde{\Lambda}_0(u_0^*) - \log 2 \\
&\overset{(d)}{=} E_{\mathrm{FP}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_0.
\end{aligned} \tag{57}$$

Here (a) holds from (52), (b) holds for any $u_0^* > 0$, (c) holds because $u_0\tau' > 0$, and (d) holds if we set $u_0^* = \arg\min \widetilde{\Lambda}_0(u)$ since $\widetilde{\Lambda}_0(u)$ attains its minima in $(0,1)$.

Sub-case 1a: $\tau' \geq \frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} = \mathrm{D}(P_1||P_0) - \log\frac{\pi_0}{\pi_1}$

$$\begin{aligned}
&E_{\mathrm{FN}}^{(T_0)}(\tau_0) - \log 2\pi_1 = \sup_{u<0}(u\tau' - \widetilde{\Lambda}_1(u) - \log 2) \\
&\overset{(a)}{\leq} \sup_{u<0}(u\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} - \widetilde{\Lambda}_1(u) - \log 2) \\
&\leq \sup_{u\in\mathcal{R}}(u\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} - \widetilde{\Lambda}_1(u) - \log 2) \\
&\overset{(b)}{=} (0\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} - \widetilde{\Lambda}_1(0) - \log 2) \\
&= (-\widetilde{\Lambda}_1(0) - \log 2) \\
&\overset{(c)}{<} -\min_u \widetilde{\Lambda}_1(u) - \log 2 \\
&\overset{(d)}{=} E_{\mathrm{FP}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_0,
\end{aligned} \tag{58}$$

where (a) holds because $\tau' \geq \frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0}$, (b) holds from Lemma 5, (c) holds from the strict convexity of $\widetilde{\Lambda}_1(u)$ because it attains its minima in $(-1, 0)$, and (d) holds from (56).

Sub-case 1b: $0 < \tau' < \frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0}$

$$\begin{aligned}
&E_{\mathrm{FN}}^{(T_0)}(\tau_0) - \log 2\pi_0 = \sup_{u<0}(u\tau' - \widetilde{\Lambda}_1(u) - \log 2) \\
&\leq \sup_{u\in\mathcal{R}}(u\tau' - \widetilde{\Lambda}_1(u) - \log 2) \\
&\overset{(a)}{=} u_a\tau' - \widetilde{\Lambda}_1(u_a) - \log 2 \\
&\overset{(b)}{<} -\widetilde{\Lambda}_1(u_a) - \log 2 \qquad [\text{since } u_a\tau' < 0] \\
&\leq -\min_u \Lambda_1(u) - \log 2 \\
&\overset{(c)}{=} E_{\mathrm{FP}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_0
\end{aligned} \tag{59}$$

Here, (a) holds from Lemma 5 because $\widetilde{\Lambda}_1(u)$ is a strictly convex and differentiable function, and its derivative is also continuous, monotonically increasing and takes all values from $-\infty$ to ∞ (see Lemma 7). Thus, there exists a single u_a such that $\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=u_a} = \tau'$. Next, (b) holds because $\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=u_a} = \tau' < \frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0}$, and the derivative is monotonically increasing, implying $u_a < 0$. Lastly (c) holds from (56).

Thus,

$$\begin{aligned}
&E_{\mathrm{FN}}^{(T_0)}(\tau_0) - \log 2\pi_1 < E_{\mathrm{FP}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_0 \\
&< E_{\mathrm{FP}}^{(T_0)}(\tau_0) - \log 2\pi_0.
\end{aligned} \tag{60}$$

For $\tau' < 0$, a similar proof holds, leading to

$$\begin{aligned}
&E_{\mathrm{FP}}^{(T_0)}(\tau_0) - \log 2\pi_0 < E_{\mathrm{FP}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_0 \\
&< E_{\mathrm{FN}}^{(T_0)}(\tau_0) - \log 2\pi_1,
\end{aligned} \tag{61}$$

Then, the value of τ_0 that maximizes the Chernoff exponent $E_e^{(T_0)}(\tau_0)$, i.e.,

$$\max_{\tau_0}\ \min\{E_{\text{FP}}^{(T_0)}(\tau_0) - \log 2\pi_0, E_{\text{FN}}^{(T_0)}(\tau_0) - \log 2\pi_1\},$$

is given by $\tau_0^* = \log\frac{\pi_0}{\pi_1}$ ($\tau' = 0$).

This matches with the detector that minimizes the Bayesian probability of error under unequal priors (see [52, Theorem 3.1]).

Case 2: $\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} = \text{D}(P_1||P_0) - \log\frac{\pi_0}{\pi_1} \leq 0.$

For this case, note that, both $\widetilde{\Lambda}_1(u)$ and $\widetilde{\Lambda}_0(u)$ attain their minima in $u \in [0, \infty)$.

For $\tau' = 0$ (equivalently $\tau_0 = \log\frac{\pi_0}{\pi_1}$), we have

$$\begin{aligned}
&E_{\text{FN}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_1 \\
&= \sup_{u<0}(u\cdot 0 - \widetilde{\Lambda}_1(u) - \log 2) \qquad\qquad = -\widetilde{\Lambda}_1(0) - \log 2.
\end{aligned} \tag{62}$$

And,

$$\begin{aligned}
&E_{\text{FP}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_0 = \\
&= \sup_{u>0}(u\cdot 0 - \widetilde{\Lambda}_0(u) - \log 2) \\
&= -\min_u \widetilde{\Lambda}_0(u) - \log 2 \\
&= -\min_u \widetilde{\Lambda}_1(u) - \log 2 \\
&\geq -\widetilde{\Lambda}_1(0) - \log 2.
\end{aligned} \tag{63}$$

Thus,

$$\begin{aligned}
&\min\{E_{\text{FP}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_0, E_{\text{FN}}^{(T_0)}(\log\frac{\pi_0}{\pi_1}) - \log 2\pi_1\} \\
&= -\widetilde{\Lambda}_1(0) - \log 2.
\end{aligned} \tag{64}$$

Now, we will show that any other value of $\tau' \neq 0$ (equivalently $\tau_0 \neq \log\frac{\pi_0}{\pi_1}$) cannot increase the Chernoff exponent of the probability of error beyond $-\widetilde{\Lambda}_1(0) - \log 2$.

Sub-case 2a: $\tau' \geq \frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} = \text{D}(P_1||P_0) - \log\frac{\pi_0}{\pi_1}$

$$\begin{aligned}
E_{\text{FN}}^{(T_0)}(\tau_0) - \log 2\pi_1 &= \sup_{u<0}(u\tau' - \widetilde{\Lambda}_1(u) - \log 2) \\
&\overset{(a)}{\leq} \sup_{u<0}(u\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} - \widetilde{\Lambda}_1(u) - \log 2) \\
&\leq \sup_{u\in\mathcal{R}}(u\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} - \widetilde{\Lambda}_1(u) - \log 2) \\
&\overset{(b)}{=} (0\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} - \widetilde{\Lambda}_1(0) - \log 2) \\
&= (-\widetilde{\Lambda}_1(0) - \log 2),
\end{aligned} \tag{65}$$

where (a) holds because $\tau' \geq \frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0}$ and (b) holds from Lemma 5.

Thus,

$$\begin{aligned}
&\min\{E^{(T_0)}_{\text{FP}}(\tau_0) - \log 2\pi_0, E^{(T_0)}_{\text{FN}}(\tau_0) - \log 2\pi_1\} \\
&\leq -\widetilde{\Lambda}_1(0) - \log 2.
\end{aligned} \tag{66}$$

Sub-case 2b: $\tau' < \frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} = \text{D}(P_1||P_0) - \log\frac{\pi_0}{\pi_1}$

$$\begin{aligned}
&E^{(T_0)}_{\text{FP}}(\tau_0) - \log 2\pi_0 \\
&= \sup_{u>0}(u\tau' - \widetilde{\Lambda}_0(u) - \log 2) \\
&\overset{(a)}{\leq} \sup_{u>0}(u\frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} - \widetilde{\Lambda}_0(u) - \log 2) \\
&\overset{(b)}{\leq} \sup_{u>0}(u\frac{d\widetilde{\Lambda}_0(u)}{du}|_{u=1} - \widetilde{\Lambda}_0(u) - \log 2) \\
&\overset{(c)}{\leq} \sup_{u\in\mathcal{R}}(u\frac{d\widetilde{\Lambda}_0(u)}{du}|_{u=1} - \widetilde{\Lambda}_0(u) - \log 2) \\
&\overset{(d)}{=} \frac{d\widetilde{\Lambda}_0(u)}{du}|_{u=1} - \widetilde{\Lambda}_0(1) - \log 2 \\
&\overset{(e)}{\leq} -\widetilde{\Lambda}_0(1) - \log 2 \\
&\overset{(f)}{=} -\widetilde{\Lambda}_1(0) - \log 2.
\end{aligned} \tag{67}$$

Here (a) holds because $\tau' < \frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0}$, (b) holds from Lemma 7 since $\widetilde{\Lambda}_1(u) = \widetilde{\Lambda}_0(u+1)$, (c) holds because the supremum is taken over a larger superset, (d) holds from Lemma 5, (e) holds because $\frac{d\widetilde{\Lambda}_0(u)}{du}|_{u=1} = \frac{d\widetilde{\Lambda}_1(u)}{du}|_{u=0} = \text{D}(P_1||P_0) - \log\frac{\pi_0}{\pi_1} \leq 0$, and (f) holds again from from Lemma 7 since $\widetilde{\Lambda}_1(u) = \widetilde{\Lambda}_0(u+1)$.

Thus,

$$\begin{aligned}
&\max_{\tau_0}\min\{E^{(T_0)}_{\text{FP}}(\tau_0) - \log 2\pi_0, E^{(T_0)}_{\text{FN}}(\tau_0) - \log 2\pi_1\} \\
&= -\widetilde{\Lambda}_1(0) - \log 2,
\end{aligned} \tag{68}$$

which is attained at $\tau_0 = \log\frac{\pi_0}{\pi_1}$.

□

E.2 Unequal priors on Z

Here we discuss a modification of optimization (9) proposed in Section 3 to account for the case of unequal priors on both Z and Y.

Let $\Pr(Z=0) = \lambda_0$ and $\Pr(Z=1) = \lambda_1$. Also let, $\Pr(Y=0|Z=0) = \pi_{00}$, $\Pr(Y=1|Z=0) = \pi_{10}$, $\Pr(Y=0|Z=1) = \pi_{01}$ and $\Pr(Y=1|Z=1) = \pi_{11}$.

Then, the overall probability of error considering both groups together is given by:

$$\begin{aligned}
&\lambda_0 P_e^{T_0}(\tau_0) + \lambda_1 P_e^{T_1}(\tau_1) \\
&= \frac{1}{2}(2\lambda_0) P_e^{T_0}(\tau_0) + \frac{1}{2}(2\lambda_1) P_e^{T_1}(\tau_1) \\
&= \frac{1}{4}(4\lambda_0\pi_{00}) P_{\mathrm{FP}}^{T_0}(\tau_0) + \frac{1}{4}(4\lambda_0\pi_{10}) P_{\mathrm{FN}}^{T_0}(\tau_0) + \\
&\qquad \frac{1}{4}(4\lambda_1\pi_{01}) P_{\mathrm{FP}}^{T_1}(\tau_1) + \frac{1}{4}(4\lambda_1\pi_{11}) P_{\mathrm{FN}}^{T_1}(\tau_1).
\end{aligned} \tag{69}$$

Then, the error exponent of the overall probability of error considering both groups is defined as:

$$\begin{aligned}
&\min\{E_{\mathrm{FP}}^{T_0}(\tau_0) - 4\pi_{00}\lambda_0, E_{\mathrm{FN}}^{T_0}(\tau_0) - 4\pi_{10}\lambda_0, \\
&\qquad E_{\mathrm{FP}}^{T_1}(\tau_1) - 4\pi_{01}\lambda_1, E_{\mathrm{FN}}^{T_1}(\tau_1) - 4\pi_{11}\lambda_1\}.
\end{aligned} \tag{70}$$

These log-generating functions can be plotted, and the intercepts made by their tangents can be examined again to obtain the error exponents, leading to the optimal detector.