Model-Agnostic Characterization of Fairness Trade-offs

Joon Sik Kim[*1], Jiahao Chen[†2], and Ameet Talwalkar[‡1,3]

[1]Machine Learning Department, Carnegie Mellon University
[2]J.P.Morgan AI Research
[3]Determined AI

Abstract

There exist several inherent trade-offs in designing a fair model, such as those between the models predictive performance and fairness, or even among different notions of fairness. In practice, exploring these trade-offs requires significant human and computational resources. We propose a diagnostic that enables practitioners to explore these trade-offs without training a single model. Our work hinges on the observation that many widely-used fairness definitions can be expressed via the fairness-confusion tensor, an object obtained by splitting the traditional confusion matrix according to protected data attributes. Optimizing accuracy and fairness objectives directly over the elements in this tensor yields a data-dependent yet model-agnostic way of understanding several types of trade-offs. We further leverage this tensor-based perspective to generalize existing theoretical impossibility results to a wider range of fairness definitions. Finally, we demonstrate the usefulness of the proposed diagnostic on synthetic and real datasets.

1 Introduction

Machine learning continues to be more widely used for applications with important societal consequences, such as credit decisioning, predictive policing, healthcare, and employment applicant screening. In these applications, model developers face regulatory, ethical, and legal challenges to prove whether or not their models are fair [1]. To provide quantitative tests of model fairness, practitioners further need to choose between multiple definitions of fairness that exist in the machine learning literature [2, 3, 4]. Furthermore, these various notions of fairness have been shown to conflict with one another, and in some situations, enforcement of fairness come with a necessary cost in loss of accuracy [5, 6]. Such considerations complicate the practical development and assessment of fair machine learning models, as applications do not always clearly point to a specific definition of fairness as appropriate for the use case, and the conditions under which performance loss must necessarily occur can be too abstract to test in practice.

As an example, suppose an engineer is responsible for training a loan prediction model from a large user dataset, subject to mandatory fairness requirements as set by internal bank policy, which itself is shaped by regulatory and business concerns. She has many choices for how to train a fair model, with fairness enforced before model training [7, 8, 9, 10, 11, 12], during model training [13, 14], or after model training [15, 16, 17]. However, she currently must resort to trial and error to determine which of these myriad approaches, if any, will produce a compliant model with sufficient performance[1] to satisfy business needs. It may even turn out that despite her best efforts, the fairness constraints are impossible to satisfy, due to limitations intrinsic to the prediction task and data at hand.

Motivated by such practical considerations, we propose the FACT (**FA**irness-**C**onfusion **T**ensor) diagnostic for exploring these fairness–performance trade-offs systematically *before* building any model. The

---

[*]joonsikk@cs.cmu.edu; part of the work done while at J.P.Morgan

[†]jiahao.chen@jpmchase.com

[‡]talwalkar@cmu.edu

[1]In this paper, *performance* refers to classical metrics derived from confusion matrix analysis, like accuracy, precision, and recall.

fairness-confusion tensor, the underlying object considered by our proposed diagnostic, is simply the ordinary confusion matrix augmented with an additional axis for each protected attribute. Our diagnostic builds upon the observation that many metrics for model performance and many fairness definitions are simple rational functions of the elements of this tensor. The diagnostic solves an optimization problem defined over this tensor (and not over model parameters), which we call the *performance–fairness optimality problem* (PFOP), thus enabling a data-dependent, yet model-agnostic analysis of the trade-offs, simply by considering the geometry of valid fairness–confusion tensors for all models that satisfy a specified set of performance and/or fairness conditions. In particular, by noting that many settings involve only linear notions of fairness, we focus on the *least-squares accuracy–fairness optimality problem* (LAFOP) in this work, a specific convex instantiation of PFOP.

Our contributions are:

1. to identify many definitions of fairness that can be expressed as functions of the fairness–confusion tensor,

2. to formulate the FACT diagnostic as a PFOP/LAFOP over the fairness–confusion tensor that explores model-agnostic fairness trade-offs,

3. to derive general fairness incompatibility results, and

4. to demonstrate the practical use of the FACT diagnostic on synthetic and real datasets.

2   Related Work

Quantitative definitions of fairness exist in many different variations [4, 5, 6, 15, 17, 18, 19]; however, less work exists to categorize these notions. [20] categorized the existing fairness definitions based on entries and rates derived from the confusion matrix, but did not fully explore trade-offs and incompatibilities among different notions. Our work extends this perspective and provides a simple geometric formalism to study these trade-offs. Yet some other definitions of fairness defy quantification using confusion matrix notions, such as individual-level fairness [15] or representation-level fairness [9, 10, 11, 12], and many models can be trained to satisfy them using different algorithms. Nevertheless, the FACT diagnostic being model-agnostic allows it to be applied even to models trained to satisfy these definitions of fairness, as long as the final criteria for evaluation involves fairness definitions that rely on the fairness-confusion tensor.

Fairness–confusion tensor is indeed not a completely new or ground-breaking notion — several work has mentioned it in passing, most of the times disregarded it as a simple computational tool that eases the computation of several definitions of fairness on the implementation level [21, 22]. It is also a natural object considered in several post-processing methods in fairness [17, 23], a group of algorithms that fine-tune a trained model to mitigate the unfairness while keeping the performance change minimal. Unlike the previous approaches, we take a closer look at the fairness–confusion tensor itself and study how this object brings together several notions of fairness, simplifying and generalizing the analysis of inherent trade-offs in fairness.

Fairness–performance trade-offs have been studied in many specific cases [2, 3, 7, 16, 24, 25, 26], for some definition of fairness, some definition of performance, and for some model. To our knowledge, these trade-offs have not been studied in the general, model-agnostic way we present below. [13, 14] presented an optimization-based analysis of the trade-offs, albeit over the parameter space of a particular model.

Fairness–fairness trade-offs are less well-studied; most existing results describe the impossibility of satisfying multiple notions of fairness simultaneously. The result presented in [5, 19, 23] is perhaps the most well-known, which states that no classifier can satisfy class balance and calibration simultaneously, except under some strong assumptions about the data and the model. We prove below a more general form of this theorem, which yields many related generalizations. [6] also showed a similar type of incompatibility between predictive parity and class balance; we re-derive this result using the fairness–confusion tensor perspective with a more detailed set of conditions for compatibility. Nevertheless, to the best of our knowledge, our work is the first to provide a general approach to diagnose both fairness–fairness and fairness–performance trade-offs together under the same formalism.

3   The fairness–confusion tensor

Our key insight is that the elements of the fairness–confusion tensor encode all the information needed to study many notions of performance and fairness. The fairness–confusion tensor is simply the stack of confusion matrices for each protected attribute a. We focus on the simplest case, with one binary protected class $a \in \{0, 1\}$, and a binary classifier $\hat{y} \in \{0, 1\}$ for a binary label $y \in \{0, 1\}$, which is shown in Table 1[2].

$a = 1$	$y = 1$	$y = 0$		$a = 0$	$y = 1$	$y = 0$
$\hat{y} = 1$	TP_1	FP_1		$\hat{y} = 1$	TP_0	FP_0
$\hat{y} = 0$	FN_1	TN_1		$\hat{y} = 0$	FN_0	TN_0

Table 1: The fairness–confusion tensor, showing the two planes corresponding to the confusion matrix for each of the favored ($a = 1$) and disfavored groups ($a = 0$).

By convention, we call $a = 1$ the favored group and $a = 0$ the disfavored group. We augment the conventional abbreviations TP, FP, FN, and TN with a subscript indicating the group variable a.

Let N be the number of data points, $N_a = TP_a + FN_a + FP_a + TN_a$ be the number of data points in each group $a \in \{0, 1\}$, $M_a = TP_a + FN_a$ be the number of postive-class instances ($y = 1$) for each group. Assume N, N_a and M_a are known constants. Unraveling the fairness–confusion tensor into an 8-dimensional vector, we denote it as

$$\mathbf{z} = (TP_1, FN_1, FP_1, TN_1, TP_0, FN_0, FP_0, TN_0)^T / N,$$

normalized and constrained to lie on a simplicial complex $\mathcal{K} = \{\mathbf{z} \geq 0 : \mathbf{A}_{\mathsf{const}}\mathbf{z} = \mathbf{b}_{\mathsf{const}}, \|\mathbf{z}\|_1 = 1\}$, where $\mathbf{A}_{\mathsf{const}}$ and $\mathbf{b}_{\mathsf{const}}$ encode marginal sum constraints of the dataset (e.g., $TP_a + FN_a = M_a$) in matrix notation:

$$\mathbf{A}_{\mathsf{const}} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix},$$

$$\mathbf{b}_{\mathsf{const}} = (N_1, M_1, N_0, M_0)^T / N.$$

The following examples show how some typical fairness definitions can be reformulated as functions of $\mathbf{z}$.

Demographic parity (DP) states that each protected group should receive positive prediction at an equal rate:

$$\Pr(\hat{y} = 1 | \mathbf{a} = 1) = \Pr(\hat{y} = 1 | \mathbf{a} = 0). \tag{1}$$

DP is equivalent to

$$(TP_1 + FP_1)/N_1 = (TP_0 + FP_0)/N_0, \tag{2}$$

or also the linear system $\mathbf{A}_{\mathrm{DP}}\mathbf{z} = 0$, where

$$\mathbf{A}_{\mathrm{DP}} = \begin{pmatrix} N_0 & 0 & N_0 & 0 & -N_1 & 0 & -N_1 & 0 \end{pmatrix} / N. \tag{3}$$

The choice of normalization, $1/N$, ensures that the matrix coefficients are in $[0, 1]$. This choice does not affect the solution $\mathbf{z}$, but will be important for having a uniform choice of "units" for ϵ and δ.

We now introduce a graphical notation to help visualize which components of the fairness–confusion tensor participate in the fairness definition. Depict the fairness–confusion tensor as ⊞ ⊞ , with the left matrix for the favored class ($a = 1$) and the right matrix for the disfavored class ($a = 0$). Since each component of $\mathbf{z}$ corresponds to some element of the fairness–confusion tensor, we shade each component that appears in the equation. Blue shading denotes the favored class, while red shading denotes the disfavored class. We further distinguish two kinds of dependencies. Components that have a nonzero coefficient in the matrix are shaded fully. However, the values of these coefficients themselves can depend on other components, albeit implicitly. In this example, $N_0 = TP_0 + FP_0 + FN_0 + TN_0$ implicitly touches the entire confusion matrix for

Name of fairness	Definition and linear system	Terms in fairness–confusion tensor
Demographic parity (DP)[18]	$\Pr(\hat{y}=1\|\mathbf{a}=1) = \Pr(\hat{y}=1\|\mathbf{a}=0)$ $\mathbf{A}_{\mathrm{DP}} = \frac{1}{N}\left(N_0\ 0\ N_0\ 0\ -N_1\ 0\ -N_1\ 0 \right)$	
Equalized opportunity (EOp)[15, 17]	$\Pr(\hat{y}=1\|y=1,\mathbf{a}=1) = \Pr(\hat{y}=1\|y=1,\mathbf{a}=0)$ $\mathbf{A}_{\mathrm{EOP}} = \frac{1}{N}\left(M_0\ 0\ 0\ 0\ -M_1\ 0\ 0\ 0 \right)$	
Predictive equality (PE)[6]	$\Pr(\hat{y}=1\|y=0,\mathbf{a}=1) = \Pr(\hat{y}=1\|y=0,\mathbf{a}=0)$ $\mathbf{A}_{\mathrm{PE}} = \frac{1}{N}\left(0\ 0\ N_0-M_0\ 0\ 0\ 0\ -N_1+M_1\ 0 \right)$	
Equalized odds (EOd)[17]	EOp $\wedge$ PE	
Calibration within groups (CG)[5]	$\Pr(y=1\|P_\theta(\mathbf{x})=s,\mathbf{a}=1) = \Pr(y=1\|P_\theta(\mathbf{x})=s,\mathbf{a}=0)=s$ $\mathbf{A}_{\mathrm{CG}} = \begin{pmatrix} 1-v_1 & 0 & -v_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1-v_0 & 0 & -v_0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1-v_1 & 0 & -v_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1-v_0 & 0 & -v_0 \end{pmatrix}$	
Positive class balance (PCB)[5]	$\mathbb{E}(P_\theta\|y=1,\mathbf{a}=1) = \mathbb{E}(P_\theta\|y=1,\mathbf{a}=0)$ $\mathbf{A}_{\mathrm{PCB}} = \min_a(M_a)\left(\frac{v_1}{M_1}\ \frac{v_0}{M_1}\ 0\ 0\ -\frac{v_1}{M_0}\ -\frac{v_0}{M_0}\ 0\ 0 \right)$	
Negative class balance (NCB)[5]	$\mathbb{E}(P_\theta\|y=0,\mathbf{a}=1) = \mathbb{E}(P_\theta\|y=0,\mathbf{a}=0)$ $\mathbf{A}_{\mathrm{NCB}} = \min_a(N_a-M_a)\left(0\ 0\ \frac{v_1}{N_1-M_1}\ \frac{v_0}{N_1-M_1}\ 0\ 0\ -\frac{v_1}{N_0-M_0}\ -\frac{v_0}{N_0-M_0} \right)$	
Predictive parity (PP)[6]	$\Pr(y=1\|\hat{y}=1,\mathbf{a}=1) = \Pr(y=1\|\hat{y}=1,\mathbf{a}=0)$ $\frac{1}{2}\mathbf{z}^T\mathbf{B}_{\mathrm{PP}}\mathbf{z} = (TP_1 FP_0 - TP_0 FP_1)/N^2$	
Equal false omission rate (EFOR) [1]	$\Pr(y=1\|\hat{y}=0,\mathbf{a}=1) = \Pr(y=1\|\hat{y}=0,\mathbf{a}=0)$ $\frac{1}{2}\mathbf{z}^T\mathbf{B}_{\mathrm{EFOR}}\mathbf{z} = (TN_1 FN_0 - TN_0 FN_1)/N^2$	
Conditional accuracy equality (CA)[19]	PP $\wedge$ EFOR	
Equal false positive rate (EFPR) [2]	$\Pr(\hat{y}=1\|y=0,a=1) = \Pr(\hat{y}=1\|y=0,a=0)$ $\frac{1}{2}\mathbf{z}^T\mathbf{B}_{\mathrm{EFPR}}\mathbf{z} = (FP_0 TN_1 - FP_1 TN_0)/N^2$	
Equal false negative rate (EFNR) [2]	$\Pr(\hat{y}=0\|y=1,a=1) = \Pr(\hat{y}=0\|y=1,a=0)$ $\frac{1}{2}\mathbf{z}^T\mathbf{B}_{\mathrm{EFNR}}\mathbf{z} = (FN_0 TP_1 - FN_1 TP_0)/N^2$	

[1] To our knowledge, EFOR has not been described in literature in isolation, but is used in the definition of conditional accuracy equality (CA)[19].
[2] Defined implicitly in [6].

Table 2: Some common fairness definitions in terms of linear functions $\phi(\mathbf{z}) = \mathbf{A}\mathbf{z}$ or quadratic functions $\phi(\mathbf{z}) = \frac{1}{2}\mathbf{z}^T\mathbf{B}\mathbf{z}$ that appear in the performance–fairness optimality problem (8). There are two groups separated by the horizontal line: those that are specified by linear functions (above), or quadratic functions (below). The graphical notation is described in Section 3. P_θ is the probability produced by a model (parameterized by θ) of $\hat{y}=1$. The fairness functions ϕ are uniquely defined only up to a normalization factor and overall sign.

$a=0$, so we shade these implicit components in a lighter shade. Putting this all together, we can represent (3) graphically as .

Predictive parity (PP) [6] states that the likelihood of being in the positive class given the positive prediction is the same for each protected class:

$$\Pr(y=1\|\hat{y}=1,\mathbf{a}=1) = \Pr(y=1\|\hat{y}=1,\mathbf{a}=0), \tag{4}$$

which is equivalent to

$$\frac{TP_1}{TP_1 + FP_1} = \frac{TP_0}{TP_0 + FP_0} \iff \frac{TP_1}{TP_0} = \frac{FP_1}{FP_0}. \tag{5}$$

Unlike for DP, the marginal sum constraints do not relate TP_a and FP_a, so this notion of fairness is *not* linear in the fairness–confusion tensor. Nevertheless, PP can be expressed using the quadratic form

$$\frac{1}{2}\mathbf{z}^T\mathbf{B}_{\mathrm{PP}}\mathbf{z} = 0, \quad \mathbf{B}_{\mathrm{PP}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{6}$$

We can write down the corresponding graphical notation as $(\ \ )^{(2)}$, with the superscript denoting the quadratic order of the term.

Calibration within groups (CG) [5], when specialized to binary classifiers and binary protected classes, can be written as the system of equations

$$\begin{aligned} FN_a &= v_0(FN_a + TN_a), \\ TP_a &= v_1(TP_a + FP_a), \end{aligned} \tag{7}$$

---

[2] The arguments naturally generalize to multiple and non-binary protected attributes with high-dimensional tensors.

where the v_is are scores satisfying $0 \leq v_0 < v_1 \leq 1$ and have no implicit dependence on any entries of the fairness–confusion tensor. We can rewrite this linear system explicitly as the matrix equation $\mathbf{A}_{\mathrm{CG}}\mathbf{z} = 0$ with

$$\mathbf{A}_{\mathrm{CG}} = \begin{pmatrix} 1-v_1 & 0 & -v_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1-v_0 & 0 & -v_0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1-v_1 & 0 & -v_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1-v_0 & 0 & -v_0 \end{pmatrix},$$

which can be written in our graphical notation as $\blacksquare \boxplus \wedge \blacksquare \boxplus \wedge \boxplus \blacksquare \wedge \boxplus \blacksquare$. This notation makes clear that calibration is a notion that touches all eight entries of the fairness–confusion tensor, but only two at a time.

Equalized odds (EOd) [17] can be expressed as two simultaneous linear equations, or $\blacksquare\blacksquare \wedge \blacksquare\blacksquare$ in our notation.

The graphical notation gives us a simple, intuitive, yet powerful way to organize many different fairness definitions in the literature, as summarized in Table 2.

4 Optimization problems over the fairness-confusion tensor

As presented in the previous section, the fairness-confusion tensor $\mathbf{z}$ allows for a succinct linear and quadratic characterization of multiple fairness definitions in the literature. We can then naturally consider the following family of optimization problems over $\mathbf{z} \in \mathcal{K}$, where the objective function is constructed in a way that the solution space reflects potential trade-offs between fairness and performance.

Definition 1 (Performance-fairness optimality problem (PFOP)). *Let $f^{(i)}(\mathbf{z}) : \mathcal{K} \to [0,1]$ be performance metrics (indexed by i) with best performance 0 and worst performance 1, $\phi^{(j)}(\mathbf{z}) = 0$ be fairness constraints (indexed by j) with $\phi^{(j)}(\mathbf{z}) : \mathcal{K} \to [0,1]$, $L_f^{(i)}(x) \geq 0$ and $L_\phi^{(j)}(x) \geq 0$ be two series of loss functions specific to each performance metrics and fairness constraints respectively, and μ_i, λ_j be real constants with $\mu_0 = 1$. Then, the* performance–fairness optimality problem (PFOP) *is a class of optimization problem of form:*

$$\arg\min_{\mathbf{z}\in\mathcal{K}} \sum_{i\geq 0} \mu_i L_f^{(i)}(f^{(i)}(\mathbf{z})) + \sum_{j\geq 0} \lambda_j L_\phi^{(j)}(\phi^{(j)}(\mathbf{z})) \tag{8}$$

PFOP is a general optimization problem that contains two groups of terms; the first quantifies performance loss; the second quantifies unfairness. The restriction $\mathbf{z} \in \mathcal{K}$ is necessary to ensure that $\mathbf{z}$ is a valid fairness–confusion tensor that obeys the requisite marginal sums.

In our discussion below, it will be convenient to consider solutions with explicit bounds on their optimality.

Definition 2. *Let $\epsilon \geq 0$ and $\delta \geq 0$. Then, a (ϵ, δ)-solution to the PFOP is a $\mathbf{z}$ that satisfies (8) such that $\sum_i L_\phi^{(i)}(\phi^{(i)}(\mathbf{z})) \leq \epsilon$ and $\sum_j L_f^{(j)}(f^{(j)}(\mathbf{z})) \leq \delta$.*

The parameters ϵ and δ therefore represent the sum total of deviation from perfect fairness and perfect predictive performance respectively.

Unless otherwise stated, the rest of this paper is dedicated to analyzing one of the simplest instantiations of PFOP, defined below.

Definition 3. *The* least-squares accuracy–fairness optimality problem (LAFOP) *is a PFOP with accuracy as the performance function, and $K \geq 1$ fairness constraints with*

$$L_\phi^{(j)}(x) = L_f^{(i)}(x) = \|x\|_2^2,$$
$$\phi(\mathbf{z}) = \left(\phi^{(0)}(\mathbf{z}), \ldots, \phi^{(K-1)}(\mathbf{z})\right)^T = \mathbf{A}\mathbf{z},$$
$$f^{(0)}(\mathbf{z}) = \mathbf{c} \cdot \mathbf{z}, \tag{9}$$
$$\mathbf{c} = (0,1,1,0,0,1,1,0)^T, \text{ and}$$
$$\lambda = \lambda_0 = \ldots = \lambda_{K-1}.$$

In other words, LAFOP is the problem

$$\arg\min_{\mathbf{z}\in\mathcal{K}} \left(\mathbf{c} \cdot \mathbf{z}\right)^2 + \lambda\|\mathbf{A}\mathbf{z}\|_2^2, \tag{10}$$

Sets of fairness definitions	Necessary conditions
{CG, PP, DP, and any of EOp, PE, PCB, NCB, EFOR}	$M_0 = M_1$ and $N_0 = N_1$
{CG, DP, and any of EOp, PE, PCB, NCB, EFOR}	EBR only
{CG,EOp}, {CG,PCB}, {CG,EOp,PCB},{CG,EFOR,EOp}, {CG,EFOR,PCB},{CG,EFOR,EOp,PCB}	$v_0 = 0$ or EBR
{CG,PE}, {CG,NCB}, {CG,EOp,NCB}, {CG,EFOR,PE}, {CG,EFOR,NCB}, {CG,EFOR,EOp,NCB}	$v_1 = 1$ or EBR
{CG,EOd}, {CG, PCB, NCB} [5],{CG,EOd,PCB,NCB}, {CG,EFOR,EOd}, {CG,EFOR,PCB,NCB},{CG,EFOR,EOd,PCB,NCB}	($v_0 = 0$ and $v_1 = 1$) or EBR

Table 3: Some example sets of fairness definitions containing CG, which are rank-0 compatible (i.e. incompatible) in the sense of Definition 4 (left-column), together with their necessary conditions to be compatible (right column). EBR is the equal base rate condition, $M_0/N_0 = M_1/N_1$. These are all special cases of Theorem 1, while not exhaustive.

where $\mathbf{c} \cdot \mathbf{z}$ encodes the usual notion of classification error. We consider only a single hyperparameter, λ, which specifies the relative importance of satisfying the fairness constraints while optimizing classification performance. This parameter can be interpreted as a regularization strength, with $\lambda = 0$ considering only performance and disabling all fairness constraints, and $\lambda = \infty$ imposing fairness constraints without regard to accuracy. Also, we can stack K linear fairness functions together into one linear system (encoded in the matrix $\mathbf{A}$) as the regularizer for considering multiple linear notions of fairness.

LAFOP is a convex optimization problem which is straightforward to analyze and solve. Despite its simplicity, LAFOP still encompasses many situations involving linear notions of fairness, allowing us to reason about multiple fairness constraints as well as fairness–accuracy trade-offs, presented in more detail in Section 6.

5   Incompatible fairness definitions

In this section, we show how PFOPs (usually LAFOPs) can be used to derive existing and new fairness incompatibility results. Proofs are given in the Appendix.

5.1   The incompatibility of {CG, PCB, NCB}, and generalizations thereof

The well-known impossibility theorem of [5] can be proved simply by asking if and when solutions exist to its corresponding LAFOP. The proof uses only elementary linear algebra.

Definition 4. *Let* $\Phi = \{\phi^{(i)}(\mathbf{z})\}_{i=0}^{K-1}$ *be a set of linear fairness functions, and ρ be the dimensionality of the vector space* $\{\mathbf{z} \in \mathcal{K} : \phi^{(i)}(\mathbf{z}) = 0 \forall i\}$*. Then, Φ is a set of rank-ρ compatible fairnesses. If $\rho > 0$, then Φ is a set of compatible fairnesses. Otherwise, Φ is termed a set of incompatible notions of fairness. An incompatible set Φ is termed* rank-0 compatible *if it is not empty.*

The notion of rank-0 compatibility captures situations when the set of fairness notions is compatible but only under exceptional, data-dependent constraints (i.e. incompatible).

For linear fairness functions $\phi^{(i)}(\mathbf{z}) = \mathbf{A}^{(i)}\mathbf{z}$, compatibility is equivalent to having infinitely many solutions to

$$\begin{pmatrix} \mathbf{A}^{(0)} \\ \vdots \\ \mathbf{A}^{(K-1)} \\ \mathbf{A}_{\text{const}} \end{pmatrix} \mathbf{z} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{b}_{\text{const}} \end{pmatrix}, \mathbf{z} \geq 0. \tag{11}$$

Below, we introduce a general version of the impossibility theorem of [5] that leads to many other results.

Theorem 1. *Let $B = 2$ be the number of bins in the definition of calibration within groups fairness (CG) [5], and v_0, v_1 be the scores, with $0 \leq v_0 < v_1 \leq 1$, and $K > 1$ with $\phi^{(0)}(\mathbf{z}) = \mathbf{A}_{\text{CG}}\mathbf{z}$. Then, the corresponding*

PFOP (8) *has the only solution*

$$z_0 = \frac{1}{N(v_1 - v_0)} \begin{pmatrix} v_1(M_1 - N_1 v_0) \\ v_0(-M_1 + N_1 v_1) \\ (1 - v_1)(M_1 - N_1 v_0) \\ (1 - v_0)(-M_1 + N_1 v_1) \\ v_1(M_0 - N_0 v_0) \\ v_0(-M_0 + N_0 v_1) \\ (1 - v_1)(M_0 - N_0 v_0) \\ (1 - v_0)(-M_0 + N_0 v_1) \end{pmatrix}, \tag{12}$$

and only when

$$0 \leq v_0 \leq \min_a \left(\frac{M_a}{N_a} \right) \leq \max_a \left(\frac{M_a}{N_a} \right) \leq v_1 \leq 1. \tag{13}$$

Otherwise, no solution exists.

Corollary 1 ([5])**.** *Consider a classifier that satisfies CG, PCB and NCB fairness simultaneously. Then, at least one of the following statements is true:*

  1. the data have equal base rates for each class a, i.e. $M_0/N_0 = M_1/N_1$, or
  2. the classifier has perfect prediction, i.e. $v_0 = 0$ and $v_1 = 1$.
In other words, {CG, PCB, NCB} is rank-0 compatible.

Theorem 1 yields many similar results regarding the incompatibility of CG and other notions of fairness.

Corollary 2. *Consider a classifier that satisfies CG and DP fairness simultaneously. Then, the data have equal base rates for each group a. In other words, {CG, DP} is rank-0 compatible.*

The same approach can be extended easily to quadratic notions of fairness.

Corollary 3. *Consider a classifier that satisfies CG and PP fairness simultaneously. Then, at least one of the following is true:*

  1. $v_0 = (M_1 - M_0)/(N_1 - N_0)$.
  2. $v_1 = 1$.
In other words, {CG, PP} is rank-0 compatible.

Similar constructions yield many other sets of fairness definitions in Table 2 that have similar impossibility criteria, as summarized in Table 3. To our knowledge, all cases other than {CG, PCB, NCB} are new.

5.2   The incompatibility of {EFPR, EFNR, PP}

In this section, we re-derive an impossibility result that is described in [6] using LAFOP formulation just like the previous section, and provide more precise conditions for compatibility. Details of the proof are in the Appendix.

Theorem 2 ([6])**.** $\Phi = ${*EFPR, EFNR, PP*} *is a rank-0 compatible set of fairness notions. For a classifier that satisfies Φ, at least one of these statements must be true:*

  1. The classifier has no true positives.
  2. The classifier has no false positives.
  3. Each protected class has the same base rate.

Theorem 2 shows that equal false positive rates, equal false negative rates, and predictive parity are compatible only under exceptional, data-dependent circumstances.

6 Experiments

In this section, we show how the FACT diagnostic can be useful in practice by considering various scenarios. First we compare the FACT Pareto frontier of equalized odds to existing published results of fair models (Section 6.2). The FACT Pareto frontier characterizes a model's achievable accuracy for a given set of fairness constraints, and helps us to assess the suboptimality of existing methods and to contextualize the FACT diagnostic in terms of recent work. Next, we further explore the cases when multiple fairness constraints are imposed in LAFOP, both in exact (Section 6.3) and relaxed (Sections 6.4, 6.5) settings. By observing these cases we draw a clear picture of how several notions of fairness and accuracy interact with one another, and demonstrate the relative impact of individual notions of fairness when imposing multiple notions simultaneously.

6.1 Datasets

We study a **synthetic dataset** similar to that in [13], consisting of two-dimensional features along with a single binary protected attribute that is either sampled from an independent Bernoulli distribution ("unbiased" variant, denoted **S(U)**), or sampled dependent on the features ("biased" variant, denoted **S(B)**). More details on sampling distribution are in Appendix H.1. We also study the **UCI Adult dataset** [27], a census dataset with protected attributes often used to predict whether or not an individual has high income. We consider sex as the only protected attribute.
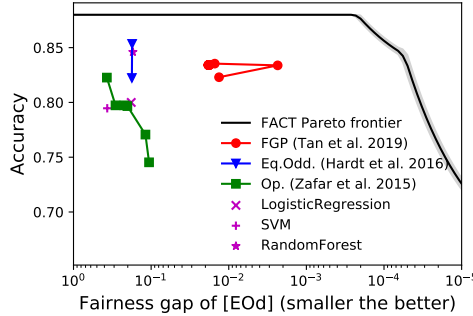
6.2 Interpreting the FACT Pareto frontier

Figure 1: Model-agnostic FACT Pareto frontier of accuracy and equalized odds fairness for the Adult dataset. Three fair models (FGP, Eq.Odd., Op.) are plotted by varying the the strength of the fairness criteria imposed, along with the baseline models (LR, SVM, RF). The FACT Pareto frontier here is plotted by taking the reasonable empirical estimate of the Bayes error into account.

With LAFOP, one can naturally consider a FACT Pareto frontier by plotting (ϵ, δ) values of (ϵ, δ)-solutions obtained under a certain fairness constraint. In this section, we want to highlight the importance of this frontier in the context of several published fair models in the literature as well as its implications.

We consider three fair models: **FGP** [12], **Op.** [13], and **Eq.Odd.** [17], individually representing three different approaches one can take in training fair models (imposing fairness before, during, or after training). We also train some baseline models (logistic regression, SVM, random forest) for reference.

The FACT Pareto frontier can be interpreted as characterizing the model's achievable accuracy *relative* to the Bayes error (i.e., the degree to which the added fairness constraints adversely impact the Bayes error). A wide range of ML models have been applied to the UCI Adult dataset [28], and thus we have a reasonable empirical estimate of the Bayes error (around 0.12). Combining this empirical estimate of the Bayes error with the output of FACT diagnostic, we obtain the FACT Pareto frontier shown in Figure 1 for the Adult dataset with respect to Equalized Odds. Indeed, we observe that the FACT Pareto frontier provides a reasonable upper bound for the three existing fair models considered, and moreover informs that only for quite small fairness gaps (e.g. 10^{-4} and smaller) that the accuracy will start to suffer. In the rest of the

following sections and figures, because reasonable estimates of the Bayes error are not available for most datasets, δ should be interpreted in reference to the Bayes error, i.e $\delta = 0$ means that the upper bound of the best-achievable accuracy is the accuracy of the Bayes classifier, not 1.

6.3   Multiple exact fairness constraints

We are now interested in a more complicated setting, when a *group* of fairness constraints are required to be *exactly* satisfied. In the context of LAFOP, we are interested in the δ values of $(\epsilon = 0, \delta)$-solutions. The optimal δ values are shown in Table 4 in terms of the best attainable accuracy $(1 - \delta)$ under several fairness constraints, which are imposed as strict equality constraints of the form $\mathbf{Az} = 0$. Again, as in Section 6.2, the reported values indicate the relative accuracy when we set 1.0 to be the accuracy of the Bayes classifier. For instance, in the unbiased synthetic dataset (first column), while {EOd, DP} are compatible, the best attainable accuracy under these fairness constraints drops by over 60 percent relative to the accuracy of the Bayes classifier. Another observation is that under EOd and DP, we can additionally have CB, PE, and PCB "for free", with no additional drop in the best attainable accuracy. Similar trends hold for the biased synthetic and the Adult dataset as well (middle, right columns), where {EOd, DP} determines δ for the group of fairness definitions in color blue. As expected, the bottom two rows of the table verify the impossibility results from Section 5 for all three datasets. Note that this table is generalizable to an arbitrary number of fairness constraints imposed on the optimization.

Fairnesses	S(U)	S(B)	Adult
PCB, CB	1.000	1.000	1.000
PE, NCB	1.000	1.000	1.000
PCB, DP	0.979	0.793	0.935
EOd, DP	0.392	0.392	0.763
EOd, PCB, DP	0.392	0.392	0.763
EOd, CB, PE, DP	0.392	0.392	0.763
EOd, CB, PE, EOp, DP	0.392	0.392	0.763
PCB, NCB, CG	-	-	-
CG, CB, EOp, DP	-	-	-

Table 4:   The best accuracy achievable (1.0 being the accuracy of the Bayes classifier) for any classifier that satisfies multiple fairnesses exactly on the synthetic datasets (unbiased, denoted as S(U), and biased, denoted as S(B)) and the Adult dataset (Section 6.1). The numbers are the largest $1 - \delta$ (and hence the smallest δ) attainable in an $(\epsilon = 0, \delta)$-solution to the performance–fairness optimality problem, as in Definition 2. A dash indicates that the optimization could not find feasible solutions, i.e. the set of fairness definitions are incompatible. This table shows which fairness comes "for free", in the sense that additional fairness can be satisfied without extra drop in accuracy (increase in δ). The groups of fairnesses are colored according to their maximum attainable δ values (from top to bottom: magenta, green, blue, black, and red).

6.4   Multiple approximate fairness constraints

In practice, we may have situations where fairness constraints need not be satisfied exactly, either because the application does not require exact fairness, or because exact fairness is impossible to attain. To account for approximate solutions with $\epsilon > 0$, we examine how (ϵ, δ)-solutions for different groups of fairness constraints change with varying amount of relaxation. Figure 2 shows this in two different ways:

1. (ϵ, δ)-solutions obtained when we impose fairness conditions as inequality constraints instead of as regularizers, i.e. solving

$$\underset{\mathbf{z} \in \mathcal{K}}{\arg\min} \, (\mathbf{c} \cdot \mathbf{z})^2 \quad \text{s.t. } \|\mathbf{Az}\|_2^2 \leq \epsilon \tag{14}$$

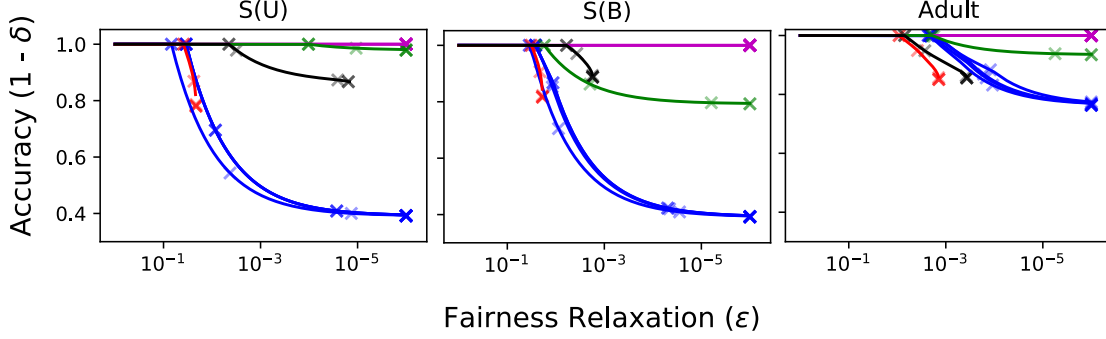while varying ϵ (drawn as solid lines), and

Figure 2: (ϵ,δ) curves for the unbiased synthetic dataset (Section 6.1, left), biased synthetic dataset (Section 6.1, middle), and Adult dataset (Section 6.1, right). Accuracy axes are equal on all plots. Note that the ϵ-axis is inverted (ϵ decreases from left to right), and the δ values should be considered in reference to the Bayes error (i.e. $\delta = 0$ represents accuracy of the Bayes classifier rather than a perfect accuracy). The colors are consistent with the groups of fairness definitions in Table 4 – groups that demonstrate similar behavior are plotted under the same color. The solid line is a trajectory of (ϵ,δ)-solutions when directly varying the ϵ and optimizing (14), and crosses are (ϵ,δ)-solutions obtained from (10) with varying λs. All other groups of definitions except for those in red and black (which are incompatible in the sense of Definition 4) converge to the value reported in Table 4 as $\epsilon \to 0$. Incompatible groups (red, black) have halted trajectories before hitting smaller values of ϵ, meaning the optimization is not feasible.

2. (ϵ,δ)-solutions obtained from the LAFOP (10) while varying λs (drawn as crosses).

It confirms that the regularization approach works just like hard constraints on the optimization.

In Figure 2, we show (ϵ,δ) curves for the groups of fairness definitions used in Table 4 with corresponding colors in all three datasets. Looking side by side, the curves in Figure 2 all converge to the values reported in Table 4 for each group, and our findings from Table 4 regarding the dominance of EOd and DP in the blue group is verified as the blue curves overlap with each other showing similar behaviors.

Comparing the blue and green groups, we can see how imposing EOd fairness instead of PCB fairness jointly with DP fairness can introduce a comparably larger drop in accuracy (EOd indeed is a stronger definition of fairness than PCB). Groups that contain incompatible fairness definitions, like {PCB, NCB, CG} and {CG, CB, EOp, DP}, have halted trajectories before hitting much smaller ϵ values for all three datasets. The trajectories confirm the theoretical impossibility results ($\epsilon = 0$ is unattainable), while also showing how much relaxation is needed for these definitions to be approximately compatible. Because the unbiased synthetic dataset has approximately equal base rates compared to the Adult or the biased synthetic dataset, the incompatibility is less strict (black line). All such analyses on trade-offs can be performed before training any models on these datasets— through the lens of LAFOP, we are able to quantify and visualize the trade-offs using changes in (ϵ, δ) values for different groups of fairness definitions, which is otherwise not captured in such model-agnostic manner.

6.5 Multi-way fairness–accuracy trade-offs

Until now, we have only considered situations where zero or one parameter is sufficient to simultaneously specify the fairness strength for every fairness function, i.e. $\lambda = \lambda_0 = \cdots = \lambda_{K-1}$. In this section, we generalize this and allow each regularization parameter to vary freely. It is then natural to consider the multi-linear least-squares accuracy–fairness optimality problem (MLAFOP):

$$\arg\min_{\mathbf{z} \in \mathcal{K}} (\mathbf{c} \cdot \mathbf{z})^2 + \sum_{i=0}^{K-1} \lambda_i \|\mathbf{A}^{(i)} \mathbf{z}\|_2^2, \tag{15}$$

where the regularization parameters λ_i now take different values across each of the K fairness constraints. This allows for a general inspection of the individual effect of fairness constraints in a group.
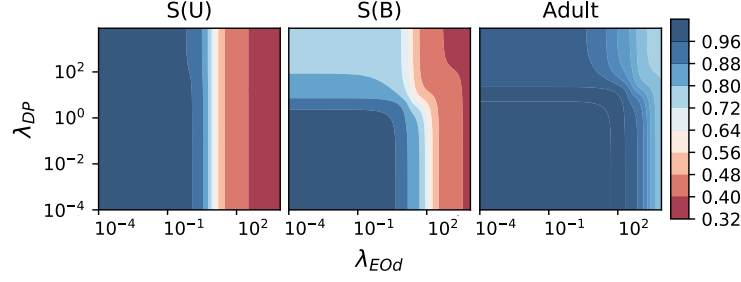
10

Figure 3: Fairness–fairness–accuracy trade-off analysis using contour plot of accuracy with varying regularization strengths of Demographic Parity (DP) and Equalized Odds (EOd) for the unbiased synthetic dataset (left), biased synthetic dataset (middle), and Adult dataset (right). The contours show how the regularization strength of each fairness individually influence the accuracy $(1 - \delta)$ given the other (accuracy of 1.0 being the accuracy of the Bayes classifier). For the unbiased synthetic data, the accuracy change along the vertical axis (DP) is practically nonexistent given EOd, while along the horizontal axis (EOd) the change is drastic. Other datasets demonstrate more complex relationships.

As an example, a three-way trade-off between EOd, DP, and accuracy can be visualized as a contour plot, similar to the ones shown in Figure 3. Such contours allow us to interpolate multiple points in the plane and observe the changes in δ more effectively for different settings for fairness constraints. For the unbiased synthetic dataset (left), EOd dominates DP in terms of the sensitivity of δ to the changes in regularization strength, as the variation of δ is the strongest along the horizontal axis. On the other hand, the relationship is relatively less one-sided for the biased synthetic and the Adult dataset.

For general $(K + 1)$-way trade-offs involving K fairness constraints and accuracy, it may be convenient to visualize specifically at two-dimensional slices along the general four-dimensional surface. We plot K different plots using the $1 - \delta$ values obtained from optimizing (15), by varying one λ for one fairness constraint while keeping the other λ values fixed. By observing how δ changes in these slices, it is possible to implicitly rank the K fairness measures in terms of their influence on δ by considering the sensitivity of δ to changes in the corresponding λ_i.
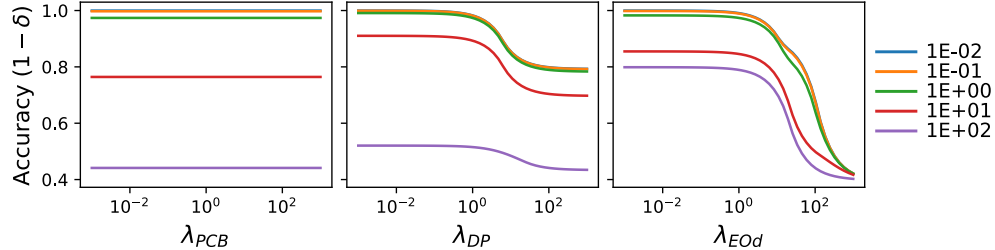

Figure 4: The four-way trade-off between accuracy, PCB, EOd, and DP in the biased synthetic dataset (Section 6.1). Shown here is the $(1-\delta)$ value as a function of some regularization strength λ_ϕ for some fairness function ϕ, while holding all other $\lambda_{\phi'}$s constant (accuracy of 1.0 being the accuracy of the Bayes classifier). The value next to each colored line in the legend represents constant values for the fixed $\lambda_{\phi'}$s. Sweeping through PCB while keeping DP and EOd fixed (left) does not change the accuracy, whereas the other plots show multiple levels of variations. For EOd (right), the accuracy levels converge quickly to the limiting value of 0.392 as reported in Table 4, suggesting that the accuracy is more sensitive to changes in EOd constraint strength compared to the others.

For example, consider a four-way trade-off between a group of three fairness definitions (DP, EOd, PCB) and accuracy. Table 4 already showed that imposing PCB given (DP, EOd) does not affect δ, which may imply that PCB is the weakest in terms of its influence on δ. To get more information, for the biased synthetic dataset, we show in Figure 4 three cases of sweeping through the regularization strength of each fairness

constraints while keeping that of the others fixed. Sweeping through PCB condition (left) does not affect $1 - \delta$ at fixed EOd and DP levels, confirming the observation from Table 4. Sweeping through DP conditions while keeping PCB and EOd strengths fixed (middle) results in a slight drop, but not big enough to make all levels to converge to $1 - \delta$ value reported on Table 4 (0.392). Sweeping through EOd while keeping PCB and DP strengths fixed (right) on the other hand results in significant changes for all levels and convergence to the value 0.392, suggesting EOd is stronger than DP in terms of its influence on changing δ. This notion of relative influence of fairness deserves further investigation, to see if these preliminary results are robust across other slices and datasets. Nonetheless, such analysis allows the practitioners to better understand the limitations of any models they train or evaluate, while also demonstrating a clear picture of how different notions of fairness interact with one another when they are to be imposed together.

7 Conclusions

The FACT diagnostic facilitates model-agnostic reasoning about different kinds of trade-offs involving arbitrarily many notions of performance and fairness which can be expressed as functions of the fairness–confusion tensor. In our formalism, many fairness definitions in the literature are in fact linear or quadratic, thus are easy to be imposed as constraints to the PFOP. The FACT formalism further allows us to use elementary linear algebra and convex optimization theory to derive many existing and new theoretical results regarding fairness–fairness trade-offs and fairness–performance trade-offs. We have also empirically validated the practical use of the FACT diagnostic in several scenarios. Many of the presented results require only linear fairness functions, as in the LAFOP setting. Nevertheless, it is easy to extend this to quadratic fairness functions with more varied performance metrics depending on different use cases. Furthermore, analyzing multi-way trade-offs suggest additional data-dependent properties to investigate, such as the relative ordering of fairness notions by sensitivity.

Acknowledgements

References

[1] Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. Ai now 2019 report, 2019.

[2] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, 2009.

[3] Indrė Žliobaitė. On the relation between accuracy and fairness in binary classification. In *Proceedings of the Second Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2015.

[4] Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the Conference on Fairness, Accountability and Transparency, New York, USA*, 2018.

[5] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017.

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

[7] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, 2010.

[8] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[9] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[10] Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair PCA: One extra dimension. In *Advances in Neural Information Processing Systems*, 2018.

[11] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *Proceedings of Machine Learning Research*, 2019.

[12] Zilong Tan, Samuel Yeom, Matt Fredrikson, and Ameet Talwalkar. Learning fair representations for kernel models. In *Proceedings of The 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

[13] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the Second Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2015.

[14] 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*, 2017.

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

[16] 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*, KDD '15, 2015.

[17] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016.

[18] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2), September 2010.

[19] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

[20] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, 2018.

[21] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[22] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[23] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, 2017.

[24] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018.

[25] Lydia T. Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[26] Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems*, 2019.

[27] Dheeru Dua and Casey Graff. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2017.

[28] Navoneel Chakrabarty and Sanket Biswas. A statistical approach to adult census income level prediction. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, 2018.

[29] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001.

[30] George B. Dantzig. Linear programming and extensions. Technical Report R-366-PR, RAND Corporation, Santa Monica, California, 1963.

[31] Dieter Kraft. A software package for sequential quadratic programming. Technical Report DFVLR-FB 88-28, Institut fr Dynamik der Flugsysteme, Deutsche Forschungs- und Versuchsanstalt fr Luft- und Raumfahrt (DFVLR), 1988.

[32] Dieter Kraft. Algorithm 733: TOMPFortran modules for optimal control calculations. *ACM Transactions on Mathematical Software*, 20(3), 1994.

A   Proof of Theorem 1

A useful strategy is to solve (11) for a set of solutions, then ask if any of these solutions satisfies an additional fairness constraint $\phi^{(K)}(\mathbf{z}) = 0$. This proof, as well as many of the ones below, illustrate this strategy in practice.

Proof. First, set $K = 1$ and $\mathbf{A}^{(0)} = \mathbf{A}_{\mathrm{CG}}$ in (11). Since $v_0 \neq v_1$, the matrix $\mathbf{A}$ is full rank and therefore admits the solution (12). Considering $\mathbf{z}_0 \geq 0$ yields immediately the condition (13).

Next, set $K > 1$. Then either $\mathbf{z}_0$ is a solution (which is the case when all other fairnesses are linear and linearly dependent on $\begin{pmatrix} \mathbf{A}_{\mathrm{CG}} \\ \mathbf{A}_{\mathrm{const}} \end{pmatrix}$), or otherwise no solution exists to both (11) and $\phi^{(1)}(\mathbf{z}) = \cdots = \phi^{(K-1)}(\mathbf{z}) = 0$ simultaneously. □

This theorem states that $\Phi = \{\mathrm{CG}\}$ is incompatible when $v_0 \neq v_1$, since it is a singleton set of rank-0 compatible fairness.

The condition $v_0 \neq v_1$ is necessary in Theorem 1, which is reasonable to assume as we would expect the positive class to have a higher score than the negative class in the definition of CG. We can prove the necessity of this condition by contradiction. In the degenerate case $v_0 = v_1 = v$, $\Phi = \{\mathrm{CG}\}$ is a set of rank-2 compatible fairnesses. It turns out that (11) with $K = 1$ is only on rank 6. Denoting $\textcircled{\scriptsize i}$ as the ith row of the matrix, we have two linear dependencies, $\textcircled{\scriptsize 5} + \textcircled{\scriptsize 6} + v\textcircled{\scriptsize 1} = \textcircled{\scriptsize 2}$ and $\textcircled{\scriptsize 7} + \textcircled{\scriptsize 8} + v\textcircled{\scriptsize 3} = \textcircled{\scriptsize 4}$. There is no longer a unique solution to the LAFOP; instead, we have a two-parameter family of solutions,

$$\mathbf{z}(\alpha, \beta) = \frac{1}{N(1-v)} \begin{pmatrix} v(N_1(1-v) - \alpha) \\ v\alpha \\ (1-v)(N_1(1-v) - \alpha) \\ (1-v)\alpha \\ v(N_0(1-v) - \beta) \\ v\beta \\ (1-v)(N_0(1-v) - \beta) \\ (1-v)\beta \end{pmatrix}, \tag{16}$$

$$0 \leq \alpha \leq (1-v)N_1, \quad 0 \leq \beta \leq (1-v)N_0.$$

Furthermore, this family of solutions satisfies $\mathbf{A}_{\mathrm{const}}\mathbf{z}_0 = \mathbf{b}_{\mathrm{const}}$ if and only if $v = M_0/N_0 = M_1/N_1$, i.e. the base rates are equal and furthermore the score for both bins is equal to the base rate.

B   Proof of Corollary 1

Proof. Consider the product

$$\begin{pmatrix} \mathbf{A}_{\mathrm{PCB}} \\ \mathbf{A}_{\mathrm{NCB}} \end{pmatrix} z_0 = \frac{M_1 N_0 - M_0 N_1}{N} \begin{pmatrix} \frac{v_0 v_1}{M_0 M_1} \\ \frac{(1-v_0)(1-v_1)}{(M_0 - N_0)(M_1 - N_1)} \end{pmatrix}. \tag{17}$$

This product equals the zero vector (and hence satisfies both PCB and NCB) if and only if either of the conditions of the Corollary hold. (The last solution, $v_0 = 1$ and $v_1 = 0$, is inadmissible since $v_0 < v_1$ by assumption.) □

C   Proof of Corollary 2

Proof. The result follows from solving

$$\mathbf{A}_{\mathrm{DP}}\mathbf{z}_0 = \frac{M_1 N_0 - M_0 N_1}{N^2(v_1 - v_0)} = 0. \tag{18}$$

□

D   Proof of Corollary 3

Proof. The result follows from solving

$$\phi_{\text{PP}}(\mathbf{z}_0) = v_1(1 - v_1)\left((M_1 - N_1 v_0)^2 - (M_0 - N_0 v_0)^2\right) = 0 \tag{19}$$

which is true if and only if either condition in the Corollary is true. (The last case, $v_1 = 0$, is inadmissible by assumption.) $\qquad\square$

E   Proof of Corollary 4

In addition, here is a situation of fairness "for free", in the sense that one notion of fairness automatically implies another.

Corollary 4. *Consider a classifier that satisfies CG fairness. Then, the classifier also satisfies EFOR fairness. In other words, {CG, EFOR} is rank-0 compatible.*

Proof. $\phi_{\text{EFOR}}(\mathbf{z}_0) = 0$ vanishes identically. $\qquad\square$

F   Proof of Theorem 2

Proof. Finding the solution to $\phi_{\text{PP}}(\mathbf{z}) = \phi_{\text{EFPR}}(\mathbf{z}) = \phi_{\text{EFNR}}(\mathbf{z}) = 0$ and also the linear system $\mathbf{A}_{\text{const}}\mathbf{z} = \mathbf{b}_{\text{const}}$ yields the three conditions of the Theorem. $\qquad\square$

G   CG–accuracy trade-offs

In the paper, we have only considered the case when $\lambda = \infty$ in the LAFOP: we only consider when the fairness criteria are satisfied exactly yielding several fairness–fairness trade-off results without heed to the accuracy of the classifiers. Nonetheless, recall that LAFOP allows us to express both fairness–accuracy and fairness–fairness trade-offs by introducing an accuracy objective along with a fairness regularizer. In this section, we show how the LAFOP can be used to theoretically analyze a simple fairness–accuracy trade-off. We now describe a small result that is relevant to the CG–accuracy trade-off considered in [25].

Theorem 3. *Let $\alpha = (M_0 + M_1)/N$ be the base rate. Consider a classifier that satisfies CG with $0 \leq v_0 < v_1 \leq 1$. Then, perfect accuracy is attained if and only if*

$$\frac{v_0(1 - 2v_1)}{1 - v_1 + v_0} = \alpha \leq \frac{1}{8}, \quad \left|v_0 - \frac{1}{4}\right| \leq \frac{\sqrt{1 - 8\alpha}}{4}. \tag{20}$$

Proof. The case of necessity ($\Rightarrow$) follows immediately from solving $\mathbf{c} \cdot \mathbf{z}_0 = 0$, where $\mathbf{z}_0$ is defined in Theorem 1. The inequality conditions follow immediately from the constraint $0 \leq v_0 < v_1 \leq 1$. The case of sufficiency ($\Leftarrow$) follows immediately from Theorem 1 and substituting the equality condition. $\qquad\square$

The condition of this theorem relates the scores v_0 and v_1 to the base rate of the data, thus providing simple, explicit data dependencies that are necessary and sufficient.

H   Experiment Details

H.1   Dataset

The synthetic dataset consists of two-dimensional data $\mathbf{x} = (x_0, x_1)$ that follow the Gaussian distributions

$$\mathbf{x}|y = 1 \sim \mathcal{N}\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 5 & 1 \\ 1 & 5 \end{pmatrix}\right)$$
$$\mathbf{x}|y = 0 \sim \mathcal{N}\left(\begin{pmatrix} -2 \\ -2 \end{pmatrix}, \begin{pmatrix} 10 & 1 \\ 1 & 3 \end{pmatrix}\right). \tag{21}$$

We further introduce two cases for the protected attribute.

Unbiased synthetic dataset The protected attribute value is independent of $\mathbf{x}$ and y, and is instead distributed according to the Bernoulli distribution $a \sim \mathcal{B}\left(\frac{1}{2}\right)$. This notion of fairness was described in [2].

Biased synthetic dataset The protected attribute value is assigned as $a|\mathbf{x} = \text{sgn}(x_0)$. This dataset models a situation when some features (but not all) encode a protected attribute.

H.2   Optimization

For solving the optimization problems, we used solvers as implemented in the `scipy` package for Python [29]. For linear fairness constraints, we used the simplex algorithm [30], and for other constrained optimization forms, we used sequential least-squares programming (SLSQP) solver [31, 32].