FAIRRET: A FRAMEWORK FOR DIFFERENTIABLE FAIRNESS REGULARIZATION TERMS

Maarten Buyl
Ghent University
maarten.buyl@ugent.be

MaryBeth Defrance
Ghent University
marybeth.defrance@ugent.be

Tijl De Bie
Ghent University
tijl.debie@ugent.be

ABSTRACT

Current fairness toolkits in machine learning only admit a limited range of fairness definitions and have seen little integration with automatic differentiation libraries, despite the central role these libraries play in modern machine learning pipelines. We introduce a framework of fairness regularization terms (FAIRRETs) which quantify bias as modular, flexible objectives that are easily integrated in automatic differentiation pipelines. By employing a general definition of fairness in terms of linear-fractional statistics, a wide class of FAIRRETs can be computed efficiently. Experiments show the behavior of their gradients and their utility in enforcing fairness with minimal loss of predictive power compared to baselines. Our contribution includes a PyTorch implementation of the FAIRRET framework.

1 INTRODUCTION

Many machine learning *fairness* methods aim to enforce mathematical formalizations of non-discrimination principles (Mehrabi et al., 2021), often by requiring statistics to be equal between groups (Agarwal et al., 2018). For example, we may require that men and women receive positive decisions at equal rates in binary classification (Dwork et al., 2012). The main interest in fairness tools is to meet such constraints without destroying the accuracy of the ML model.

A large class of these fairness tools utilizes *regularization terms*, i.e. quantifications of unfairness that can be added to the existing error term of an unfair ML model (Kamishima et al., 2012; Berk et al., 2017; Zafar et al., 2019; Padala & Gujar, 2021; Padh et al., 2021; Buyl & De Bie, 2022). The modularity of such loss terms appears to align well with the paradigm of automatic differentiation libraries like PyTorch (Paszke et al., 2019), which have become the bedrock of modern machine learning pipelines. However, the practical use of this modularity has seen little interest thus far.

Contributions Hence, we formalize a framework of fairness regularization terms (**FAIRRETs**) that consolidates research in differentiable fairness methods. A FAIRRET quantifies a model's unfairness as a single value that is minimized like any other objective through automatic differentiation.

We implement two types of FAIRRETs: FAIRRETs that directly penalize the *violation* of fairness constraints and FAIRRETs that minimize the distance between a model and its *projection* onto the set of fair models. These FAIRRETs support any fairness notion defined through linear-fractional statistics (Celis et al., 2019), which is a far wider range than the exclusively linear statistics typically considered in literature (Zafar et al., 2019; Agarwal et al., 2018). Moreover, our framework generalizes to the simultaneous handling of multiple sensitive traits and (a weaker form of) fairness with respect to continuous sensitive variables. By design, FAIRRETs are both modular and extensible such that future work that can benefit from their wide applicability. The Appendix contains a code example.

We visualize the FAIRRETs' gradients and evaluate their empirical performance in enforcing fairness notions compared to baselines. We infer this is far more difficult for fairness notions with linear-fractional statistics, which were rarely studied in prior work, than those with linear statistics.

The framework is available as a package at `https://github.com/aida-ugent/fairret`.

Related Work Fairness tools are classified as *preprocessing*, *inprocessing* or *postprocessing* (Mehrabi et al., 2021). FAIRRETs perform inprocessing, as they are minimized during training.

A popular approach to fairness regularization is to penalize the violation of fairness constraints (Zemel et al., 2013; Padala & Gujar, 2021; Wick et al., 2019), which we formalize as a FAIRRET. We also take inspiration from postprocessing methods that project classifiers onto a fair set (Alghamdi et al., 2020; Wei et al., 2020) and penalize the cost of this projection (Buyl & De Bie, 2021) as a FAIRRET. Fair representation learning (McNamara et al., 2019; Oneto et al., 2020; Franco et al., 2022) finds intermediate representations that minimally contain sensitive information. An example is the adversarial approach of Adel et al. (2019), which is a baseline in our experiments.

Celis et al. (2019) observed that many fairness definitions express a parity between linear-fractional statistics. They propose a meta-algorithm to find optimal classifiers that satisfy this constraint. Instead, we employ a simpler (yet sufficiently expressive) linear-fractional form and propose an algorithm to use them in the construction of linear constraints that does not require a meta-algorithm.

Popular fairness toolkits such as Fairlearn (Bird et al., 2020) and AIF360 (Bellamy et al., 2018) expect the underlying model in the form of *scikit-learn Estimators*[1] that can be retrained at-will in fairness meta-algorithms. Instead, our proposed FAIRRETs act as a loss term that can simply be added *within* a training step. The aforementioned toolkits have some integration with automatic differentiation libraries in adversarial fairness approaches (Zhang et al., 2018), yet these still require full control over the training process and lack generality in the fairness notions they can enforce.

Two PyTorch-specific projects with similar goals as our paper are FairTorch (Masashi, 2020) and the Fair Fairness Benchmark (FFB) (Han et al., 2023). However, neither present a formal framework and both only support a limited range of fairness definitions.

2 FAIRNESS IN BINARY CLASSIFICATION

In fair binary classification, we are provided with random variables $(\mathbf{X}, \mathbf{S}, Y)$ with $\mathbf{X} \in \mathbb{R}^{d_x}$ the feature vector of an individual, $\mathbf{S} \in \mathbb{R}^{d_s}$ their *sensitive* feature vector and $Y \in \{0, 1\}$ the binary output label. In what remains, all expectations are taken over the joint distribution of $(\mathbf{X}, \mathbf{S}, Y)$.

The goal is to learn a classifier f such that its predictions $f(\mathbf{X})$ match Y while avoiding discrimination with respect to $\mathbf{S}$. In this section, we will assume f directly provides binary decisions, i.e. $f : \mathbb{R}^{d_x} \to \{0, 1\}$, as this is expected in traditional formalizations of fairness. However, since such 'hard' classifiers are not differentiable, we will instead be learning probabilistic classifiers in Sec. 3.

Further note that our definition of sensitive features $\mathbf{S}$ as real-valued and d_s-dimensional vectors is a generalization of typical fairness definitions which assume a categorical (or binary) domain for sensitive features (Verma & Rubin, 2018). We will one-hot encode such categorical traits, e.g. by encoding 'white' or 'non-white' as the vectors $\mathbf{S} = (1, 0)^\top$ and $\mathbf{S} = (0, 1)^\top$ respectively. Our generalization allows us to take multiple non-exclusive sensitive traits into account by mapping them to different values S_k in the same vector $\mathbf{S}$ for $k \in [d_s] = \{0, ..., d_s - 1\}$. Additionally, by letting $S_k \in \mathbb{R}$, we allow soft specifications of identity rather than requiring hard discretization.

2.1 PARTITION FAIRNESS

Though we will allow any feature vector $\mathbf{S} \in \mathbb{R}^{d_s}$ in our framework, popular fairness definitions require every person to belong to exactly one demographic group. We call this *partition fairness*.

Definition 1 *In **partition fairness**, $\mathbf{S}$ is a one-hot encoding, i.e. $S_k \in \{0, 1\}$ and $\sum_{k \in [d_s]} S_k = 1$.*

Example 1 *A straightforward, popular definition in partition fairness is Demographic Parity (DP), also known as statistical parity (Dwork et al., 2012; Verma & Rubin, 2018). It enforces*

$$\forall k \in [d_s] : P(f(\mathbf{X}) = 1 \mid S_k = 1) = P(f(\mathbf{X}) = 1) \tag{1}$$

which states that all groups ought to get positive predictions at the same rate (i.e. the overall rate).

Let $\gamma(k; f) \triangleq \frac{\mathbb{E}[S_k f(\mathbf{X})]}{\mathbb{E}[S_k]}$. It is easily shown that $\gamma(k; f) = P(f(\mathbf{X}) = 1 \mid S_k = 1)$. Thus also

$$P(f(\mathbf{X}) = 1 \mid S_k = 1) = P(f(\mathbf{X}) = 1) \iff \gamma(k; f) = \mathbb{E}[f(\mathbf{X})]. \tag{2}$$

[1] `https://scikit-learn.org/1.3/developers/develop.html` describes these *Estimators*

Table 1: Fairness definitions and their α and β functions. Conditional Demographic Parity encompasses many notions with an arbitrary function ζ conditioned on the input $\mathbf{X}$.

Fairness Definition	α_0	β_0	α_1	β_1
Demographic Parity (Dwork et al., 2012)	0	1	1	0
Conditional Demographic Parity (Wachter et al., 2020)	0	$\zeta(\mathbf{X})$	$\zeta(\mathbf{X})$	0
Equal Opportunity (Hardt et al., 2016)	0	Y	Y	0
False Positive Parity (Hardt et al., 2016)	0	1 - Y	1 - Y	0
Predictive Parity (Chouldechova, 2017)	0	Y	0	1
False Omission Parity	Y	-Y	1	-1
Accuracy Equality (Berk et al., 2021)	1 - Y	2Y - 1	1	0
Treatment Equality (Berk et al., 2021)	Y	-Y	0	1 - Y

In Example 1, fairness is formalized by requiring a statistic γ to be equal across groups. This principle can be generalized to a wide class of parity-based fairness notions. In particular, we consider those expressed through *linear-fractional* statistics (Celis et al., 2019).

Definition 2 *A **linear-fractional** statistic γ computes values $\gamma(k; f) \in \mathbb{R}$ for sensitive variable S_k and classifier $f : \mathbb{R}^{d_x} \to \{0, 1\}$. We assume γ is differentiable with respect to f. It takes the form*

$$\gamma(k; f) = \frac{\mathbb{E}[S_k(\alpha_0(\mathbf{X}, Y) + f(\mathbf{X})\beta_0(\mathbf{X}, Y))]}{\mathbb{E}[S_k(\alpha_1(\mathbf{X}, Y) + f(\mathbf{X})\beta_1(\mathbf{X}, Y))]} \tag{3}$$

with α_0, α_1, β_0, and β_1 all functions that do not depend on $\mathbf{S}$ or f. Let Γ denote all such statistics. Also, let $\overline{\gamma}(f) \triangleq \frac{\mathbb{E}[\alpha_0(\mathbf{X},Y)+f(\mathbf{X})\beta_0(\mathbf{X},Y)]}{\mathbb{E}[\alpha_1(\mathbf{X},Y)+f(\mathbf{X})\beta_1(\mathbf{X},Y)]}$ denote the overall statistic value without conditioning on $\mathbf{S}$.

Definition 3 *A **fairness notion** is expressed through a statistic $\gamma \in \Gamma$. The set $\mathcal{F}_\gamma$ of classifiers that adhere to the fairness notion is defined as*

$$\mathcal{F}_\gamma \triangleq \left\{ f : \mathbb{R}^{d_x} \to \{0, 1\} \mid \forall k \in [d_s] : \gamma(k; f) = \overline{\gamma}(f) \right\} \tag{4}$$

i.e. the statistic $\gamma(k; f)$ for each S_k equals the overall statistic $\overline{\gamma}(f)$.

Indeed, the DP fairness notion in Example 1 is expressed as a fairness notion as defined in Def. 3 with linear-fractional statistics as defined in Def. 2. The same holds for the following notions.

Example 2 *Equalized Opportunity (EO) (Hardt et al., 2016) only computes DP for actual positives $Y = 1$. Its statistic γ is thus the recall $P(f(\mathbf{X}) = 1 \mid Y = 1, S_k = 1)$, i.e. $\gamma(k; f) = \frac{\mathbb{E}[S_k f(\mathbf{X})Y]}{\mathbb{E}[S_k Y]}$.*

Example 3 *Predictive Parity (PP) (Chouldechova, 2017), which compares the precision statistic $P(Y = 1 \mid f(\mathbf{X}) = 1, S_k = 1)$, i.e. $\gamma(k; f) = \frac{\mathbb{E}[S_k f(\mathbf{X})Y]}{\mathbb{E}[S_k f(\mathbf{X})]}$.*

Example 4 *Treatment Equality (TE) (Berk et al., 2021) balances the ratios of false negatives over false positives, i.e. $\gamma(k; f) = \frac{\mathbb{E}[S_k(1-f(\mathbf{X}))Y]}{\mathbb{E}[S_k f(\mathbf{X})(1-Y)]}$. Unlike the other notions, its γ is not a probability.*

Table 1 summarizes the α and β functions of several fairness notions (Verma & Rubin, 2018) with linear-fractional statistics. Their derivations are found in the Appendix.

Definition 4 *A linear-fractional statistic $\gamma \in \Gamma$ is **linear** when $\beta_1(\mathbf{X}, Y) \equiv 0$.*

Let $\Gamma_{\mathrm{L}} \subset \Gamma$ denote the set of all linear statistics.

Fairness notions with linear statistics $\gamma \in \Gamma_{\mathrm{L}}$ are thus identified in Table 1 by checking the column for β_1. Such notions are especially useful because the fairness constraint in Def. 3 is easily written as a linear constraint over classifier f. In turn, this makes the set of fair classifiers $\mathcal{F}_\gamma$ a convex set, which leads to convex optimization problems (Boyd & Vandenberghe, 2004). Thus, the constrained optimization of f can be efficiently performed if f is itself linear (Zafar et al., 2019).

However, fairness notions with linear-fractional statistics $\gamma \in \Gamma \setminus \Gamma_{\mathrm{L}}$ do not directly lead to linear constraints in Def. 3. To facilitate optimization, we therefore propose to narrow the set of fair classifiers $\mathcal{F}_\gamma$ to the subset where the statistics are all equal in a particular value c.

Definition 5 *Fix a $c \in \mathbb{R}$. A c-**fixed fairness notion** is expressed through a linear-fractional statistic $\gamma \in \Gamma$ such that the set $\mathcal{F}_\gamma(c)$ of classifiers f that adhere to the fairness notion is defined as*

$$\mathcal{F}_\gamma(c) \triangleq \left\{ f : \mathbb{R}^{d_x} \to \{0,1\} \mid \forall k \in [d_s] : \gamma(k; f) = c \right\}. \tag{5}$$

Proposition 1 *With $\gamma \in \Gamma$, the c-fixed fairness notion $\mathcal{F}_\gamma(c)$ enforces **linear** constraints:*

$$\gamma(k; f) = c \iff \mathbb{E}[S_k(\alpha(\mathbf{X}, Y, c) + f(\mathbf{X})\beta(\mathbf{X}, Y, c))] = 0 \tag{6}$$

where $\alpha(\mathbf{X}, Y, c) = \alpha_0(\mathbf{X}, Y) - c\alpha_1(\mathbf{X}, Y)$ and $\beta(\mathbf{X}, Y, c) = \beta_0(\mathbf{X}, Y) - c\beta_1(\mathbf{X}, Y)$.

Using Prop. 1, we can still obtain linear constraints for fairness notions $\mathcal{F}_\gamma$ with linear-fractional statistics $\gamma \in \Gamma \setminus \Gamma_{\mathrm{L}}$ by considering their c-fixed variant $\mathcal{F}_\gamma(c)$ instead. This sacrifices a degree of freedom because statistics $\gamma(k; f)$ are no longer allowed to be equal for any overall statistic $\overline{\gamma}(f)$, they must now do so for the specific case where $\overline{\gamma}(f) = c$. However, there are c values that still lead to interesting sets $\mathcal{F}_\gamma(c)$. In the FAIRRETs we propose, we take an unfair classifier h and fix $c = \overline{\gamma}(h)$ to construct the set of all *fair* classifiers $\mathcal{F}_\gamma(\overline{\gamma}(h))$ that would result from a fair redistribution of scores in h over the sensitive groups.

Though Prop. 1 is inspired by Celis et al. (2019), our use of this result vastly differs. Instead of fixing the statistics to a single value c, they set many pairs of upper and lower bounds for each group's statistics, giving rise to as many optimization programs. They then propose a meta-algorithm that searches the best classifier over each of these programs. A meta-algorithm is not necessary in our framework, as we will allow c to evolve during training. While we have no formal convergence guarantees for this approach, empirical results show it works well in practice.

2.2 Beyond Partition Fairness

Having firmly rooted our definitions in partition fairness (Def. 1), we now abandon its assumptions. First, we allow $S_k \in \mathbb{R}$. Second, we extend to multiple sensitive features with $\sum_k S_k \in \mathbb{R}$.

2.2.1 Continuous Sensitive Values

Admitting continuous values for someone's sensitive trait, i.e. $S_k \in \mathbb{R}$ allows us to take naturally continuous features, such as age, into account. Also, it provides an opportunity for an imprecise specification of demographic group membership.

For instance, instead of exactly knowing the gender of an individual, we may only have a probability available, e.g. because it is noisily predicted by a third-party classifier, or to protect the individual's privacy. By allowing $S_k \in (0, 1)$, the attribute S_k could then express 'woman-ness' instead of a binary 'woman' or 'not woman'. Thus, we also allow individuals to themselves quantify how strongly they identify with a group, rather than requiring a binary membership.

Our notation already generalizes to non-binary S_k values; they can simply be filled in for linear-fractional statistics $\gamma \in \Gamma$ as defined in Def. 2. Fairness as formalized in Def. 3 can then still be enforced through $\gamma(k; f) = \overline{\gamma}(f)$.

Remark 1 *Partition fairness constraints stem from the principle of treating distinct groups equally. This does not directly apply for a non-binary S_k. For example, if there is only one, continuous sensitive variable $(S_0) = \mathbf{S}$ such as the age of an individual, then we cannot compare $\gamma(0; f)$ to another group's statistics. Instead, $\gamma(0; f)$ must be compared to a value independent of $\mathbf{S}$.*

Enforcing $\gamma(k; f) = \overline{\gamma}(f)$ is then a sensible choice, as it satisfies key properties one can expect from a fairness measure. First, the constraint is met when $S_k \equiv s$, i.e. when S_k is a deterministic constant. Second, it holds if S_k has no linear influence on the numerator and denominator of γ, i.e.

$\operatorname{cov}(S_k, \alpha_0(\mathbf{X}, Y) + f(\mathbf{X})\beta_0(\mathbf{X}, Y)) = \operatorname{cov}(S_k, \alpha_1(\mathbf{X}, Y) + f(\mathbf{X})\beta_1(\mathbf{X}, Y)) = 0 \implies \gamma(k; f) = \overline{\gamma}(f)$.

For a full derivation of this result, we refer to the Appendix.

2.2.2 Multiple Axes of Discrimination

By allowing $\sum_k S_k \in \mathbb{R}$, we support that $\mathbf{S}$ contains information about people from several sensitive traits, e.g. gender, ethnicity, and religion. Because these each form a possible axis of discrimination, we can 'sum' these sources of discrimination by combining the constraints.

For example, if pairs of sensitive features (S_0, S_1) and (S_2, S_3) each partition the dataset, then fairness requires both $\gamma(0; f) = \gamma(1; f) = \overline{\gamma}(f)$ and $\gamma(2; f) = \gamma(3; f) = \overline{\gamma}(f)$. Combined, these constraints make up the fairness definition in Def. 3. The use of one-hot notations for sensitive values thus already allows us to combine axes of discrimination for categorical sensitive traits.

Remark 2 *An important limitation is that we only view fairness separately per axis of discrimination. Outside the partition fairness setting, this means that some intersections of sensitive groups, e.g. 'black woman', will not be represented in the constraints that enforce fairness with respect to 'black' and 'woman' separately (Kearns et al., 2018). A toy example is given in the Appendix.*

3 Fairness Regularization Terms

The popular approach to modern machine learning is to construct pipelines consisting of modular, parameterized components that are differentiable from the objective to the input. We therefore use *probabilistic* classifier models $h : \mathbb{R}^{d_x} \to (0, 1)$ from now on, where decisions are sampled from a Bernoulli distribution with parameter $h(\mathbf{X})$. Let $\mathcal{H}$ denote the hypothesis class of these models.

Remark 3 *Fairness statistics $\gamma(k; h)$ over the output of a* probabilistic *classifier h only approximately verify their respective fairness notions, as these were only defined for* hard *classifiers with a binary output (Lohaus et al., 2020). In the Appendix, we discuss the impact of this approximation and how its fidelity can be traded-off with the quality of the gradient of $\gamma(k; h)$ with respect to h.*

In binary classification, we minimize a loss $\mathcal{L}_Y(h)$ over the probabilistic classifier h given output labels Y, e.g. the cross-entropy. In *fair* binary classification we additionally pursue $h \in \mathcal{F}_\gamma$:

$$\min_{h \in \mathcal{F}_\gamma} \mathcal{L}_Y(h). \tag{7}$$

For linear-fractional statistics, the constraint is linear when considering the c-fixed variant of $\mathcal{F}_\gamma$ (using Prop. 1). However, for non-convex models h, the constrained optimization of h will remain non-convex as well. In the general case, we thus relax $h \in \mathcal{F}_\gamma$ and instead incur a cost to $h \notin \mathcal{F}_\gamma$.

Definition 6 *A **fairness regularization term** (**FAIRRET**) $R_\gamma(h) : \mathcal{H} \to \mathbb{R}_{\geq 0}$ quantifies the unfairness of the model $h \in \mathcal{H}$ with respect to the fairness notion defined through statistic γ.*

A FAIRRET *is **strict** if it holds that $h \in \mathcal{F}_\gamma \iff R_\gamma(h) = 0$.*

The objective in Eq. (7) is then relaxed as

$$\min_h \mathcal{L}_Y(h) + \lambda R_\gamma(h) \tag{8}$$

with λ a hyperparameter. The objective in Eq. (8) is equivalent to Eq. (7) for $\lambda \to \infty$ if R_γ is strict.

Remark 4 *We call R_γ a regularization term, yet its purpose is* not *to reduce model complexity or improve generalization performance, in contrast to traditional regularization in machine learning (Kukačka et al., 2017). Instead, we aim to limit the hypothesis class of h to the set of fair classifiers.*

In what follows, we introduce two archetypes of FAIRRETs: *violation* and *projection*. We visualize $\nabla_h R_\gamma$ for each FAIRRET in Fig. 1 with γ the positive rate statistic (thereby enforcing DP).

3.1 Violation FAIRRETs

To quantify $h \notin \mathcal{F}_\gamma$, we can start from the *violation* $\mathbf{v}(h)$ of the constraint that defines $\mathcal{F}_\gamma$:

$$\mathbf{v}_k(h) = \left| \frac{\gamma(k; h)}{\overline{\gamma}(h)} - 1 \right| \tag{9}$$

with $\mathbf{v} : \mathcal{H} \to \mathbb{R}^{d_s}$ a vector-valued function with components $\mathbf{v}_k$. Clearly, $\mathbf{v}(h) = \mathbf{0} \iff h \in \mathcal{F}_\gamma$.

Note that $\mathbf{v}(h)$ is normalized[2] by $\overline{\gamma}(h)$ such that a classifier cannot minimize $\mathbf{v}(h)$ by uniformly downscaling its statistics γ without reducing relative differences between groups (Celis et al., 2019).

[2]In cases where $\overline{\gamma}(h) = 0$, we can simply use $\mathbf{v}_k(h) = |\gamma(k; h)|$ instead. We assume $h(\mathbf{X}) \in (0, 1)$, so this only occurs in degenerate cases for the notions in Table 1 (like when all $Y = 0$ for Equal Opportunity).

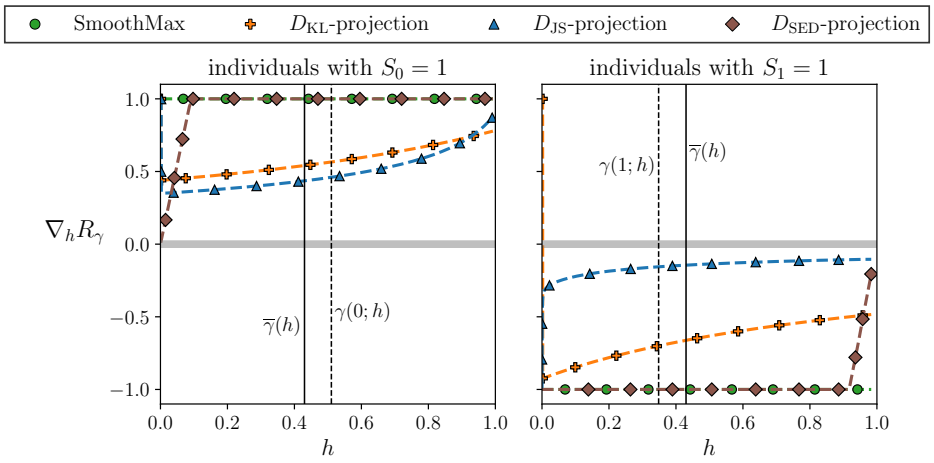


Figure 1: The model h was trained on the *ACSIncome* dataset without FAIRRET (i.e. $\lambda = 0$) and ends up with disparate positive rates $\gamma(0; h) > \overline{\gamma}(h) > \gamma(1; h)$ for the one-hot encoded sensitive variables (S_0, S_1). These should be brought closer to the overall positive rate $\overline{\gamma}(f)$. We show probability scores h and the gradients[3] of several FAIRRETs R_γ with respect to h. The gradients are normalized by dividing them by their maximum absolute value per FAIRRET and per group. They are positive for samples with $S_0 = 1$, implying their scores should decrease, and vice versa for $S_1 = 1$.

Definition 7 *We define the **Norm*** FAIRRET *as* $R_\gamma(h) \triangleq \|\mathbf{v}(h)\|$, *with* $\|\cdot\|$ *a norm over* $\mathbb{R}^{d_s}$.

Many variants of the Norm FAIRRET have been proposed, e.g. by Zemel et al. (2013), Padala & Gujar (2021), Wick et al. (2019) and Chuang & Mroueh (2020). However, fairness evaluation metrics often only consider the maximal violation. Hence, we propose the SmoothMax variant.

Definition 8 *We define the **SmoothMax*** FAIRRET *as* $R_\gamma(h) \triangleq \log \sum_{k \in [d_s]} \exp(\mathbf{v}_k(h)) - \log d_s$

Because the SmoothMax performs the log-sum-exp operation over the violation, it can be considered a smooth approximation of the maximum. We subtract $\log d_s$ to ensure the FAIRRET is strict.

Generally, violation FAIRRETs can be characterized as functions of the violation $\mathbf{v}(h)$. This lends them interpretability, but it also means that the gradient[3] $\nabla_h R_\gamma$ decomposes as

$$\nabla_h R_\gamma = \left(\frac{\partial \mathbf{v}}{\partial h}\right)^\top \nabla_{\mathbf{v}} R_\gamma \tag{10}$$

with $\frac{\partial \mathbf{v}}{\partial h}$ the Jacobian[3] of $\mathbf{v}(h)$. The gradients of violation FAIRRETs R_γ thus only differ in the $\nabla_{\mathbf{v}} R_\gamma$ gradient. Hence, the Norm FAIRRET is excluded from Fig. 1 because its gradients equal those of SmoothMax after normalization. Figure 1 also suggests that violation FAIRRETs convey little information on how each individual $h(\mathbf{X})$ score should be modified. Instead, they merely direct scores to uniformly increase or decrease within each group.

3.2 PROJECTION FAIRRETS

Recent postprocessing approaches to fairness redistribute all individual probability scores of a model $h(\mathbf{X})$ to a fair scores vector with a minimal loss in predictive power. For example, Alghamdi et al. (2020) project the scores onto the fair set $\mathcal{F}_\gamma$ as a postprocessing step. Yet, the cost of this projection can be seen as a quantification of unfairness that may be minimized as a FAIRRET during training.

Given a statistical divergence or distance D, we can generally define such a *projection* FAIRRET as

$$R_\gamma(h) \triangleq \min_{f \in \mathcal{F}_\gamma(\overline{\gamma}(h))} \mathbb{E}[D(f(\mathbf{X}) \parallel h(\mathbf{X}))]. \tag{11}$$

[3]There is some abuse of notation here. When taking the gradient or Jacobian with respect to h, we take it with respect to the vector of n outputs of h for a set of n input features sampled from the distribution over $\mathbf{X}$.

Importantly, we do not project h onto the general fair set $\mathcal{F}_\gamma$, but on the c-fixed subset $\mathcal{F}_\gamma(c)$ with $c = \overline{\gamma}(h)$. The c-fixing is done such that the projection only requires linear constraints for linear-fractional statistics (see Prop. 1). A projection FAIRRET (Eq. 11) is then a convex optimization problem if we limit ourselves to a D that is convex with respect to f, which is the case for all projections discussed here. In particular, we c-fix to the overall statistic $\overline{\gamma}(h)$ of h because this ensures h can always be projected onto itself if it is already fair, as then $h \in \mathcal{F}_\gamma(\overline{\gamma}(h))$.

Definition 9 *The D_{KL}-**projection** uses the binary Kullback-Leibler divergence*

$$D_{\mathrm{KL}}(f(\mathbf{X}) \parallel h(\mathbf{X})) \triangleq f(\mathbf{X}) \log \frac{f(\mathbf{X})}{h(\mathbf{X})} + (1 - f(\mathbf{X})) \log \frac{1 - f(\mathbf{X})}{1 - h(\mathbf{X})}. \tag{12}$$

The D_{KL}-divergence is both a Csiszar divergence and a Bregman divergence (Amari, 2009). Also, the cross-entropy error minimized in $\mathcal{L}_Y(h)$ equals $D_{\mathrm{KL}}(Y \parallel h(\mathbf{X}))$ up to a constant. The minimization of Eq. (8) thus comes down to simultaneously minimizing D_{KL} between $h(\mathbf{X})$ and the data Y, and between $h(\mathbf{X})$ and the closest $f \in \mathcal{F}_\gamma(\overline{\gamma}(h))$ (Buyl & De Bie, 2021).

Definition 10 *The D_{JS}-**projection** uses the binary Jensen-Shannon divergence.*

$$D_{\mathrm{JS}}(f(\mathbf{X}) \parallel h(\mathbf{X})) \triangleq \frac{1}{2} D_{\mathrm{KL}}(f(\mathbf{X}) \parallel m(\mathbf{X})) + \frac{1}{2} D_{\mathrm{KL}}(h(\mathbf{X}) \parallel m(\mathbf{X})) \tag{13}$$

with $m(\mathbf{X}) = \frac{1}{2} f(\mathbf{X}) + \frac{1}{2} h(\mathbf{X})$.

Just like D_{KL}, the D_{JS}-divergence is a Csiszar divergence. However, the D_{JS}-divergence is symmetric with respect to its arguments f and h, which is not the case for the D_{KL}-divergence.

Definition 11 *The D_{SED}-**projection** uses the squared Euclidean distance between the two points $(1 - f(\mathbf{X}), f(\mathbf{X}))$ and $(1 - h(\mathbf{X}), h(\mathbf{X}))$:*

$$D_{\mathrm{SED}}(f(\mathbf{X}) \parallel h(\mathbf{X})) \triangleq 2(f(\mathbf{X}) - h(\mathbf{X}))^2. \tag{14}$$

D_{SED} is a Bregman divergence between the Bernoulli distributions with parameters $f(\mathbf{X})$ and $h(\mathbf{X})$.

In practice, we evaluate projection FAIRRETs $R_\gamma(h)$ in two steps.

$$\text{(i)} \quad f^* = \operatorname*{arg\,min}_{f \in \mathcal{F}_\gamma(\overline{\gamma}(h))} \mathbb{E}[D(f(\mathbf{X}) \parallel h(\mathbf{X}))] \tag{15}$$

$$\text{(ii)} \quad R_\gamma(h) = \mathbb{E}[D(f^*(\mathbf{X}) \parallel h(\mathbf{X}))] \tag{16}$$

While keeping h fixed, step (i) computes the overall statistic $\overline{\gamma}(h)$ and then finds the projection f^* through constrained optimization. Subsequently, step (ii) keeps f^* fixed and computes $\mathbb{E}[D(f^*(\mathbf{X}) \parallel h(\mathbf{X}))]$ as a function of h, which we use to compute the gradient with respect to h. This gradient differs from the actual gradient of the optimization as a function of h in a projection FAIRRET (Eq. 11), because the latter would require us to treat f^* as a function of h. However, by treating f^* as fixed instead (without backpropagating through it), we significantly simplify the FAIRRET's implementation. The optimization in step (i) can then be solved generically using specialized libraries such as `cvxpy` (Agrawal et al., 2018; Diamond & Boyd, 2016). In our experiments, we find that only 10 optimization steps is enough to get a reasonable approximation of the solution. We refer to the Appendix for a discussion of this approximation and for a visualization of each projection f^*.

Figure 1 shows that the gradients of the projection FAIRRETs increase with higher values of h. We hypothesize this occurs when $\gamma(k; h) > \overline{\gamma}(h)$ because $\gamma(k; h)$ is more easily decreased by reducing higher h values than lower ones. Conversely, when $\gamma(k; h) < \overline{\gamma}(h)$, there is more to gain from increasing lower h values than higher ones. The sharp bend of the gradients of the D_{SED}-projection is explained in the Appendix through an analysis of the projected distributions.

3.3 Analysis

Proposition 2 *All FAIRRETs presented in this paper (i.e. Def. 7, 8, 9, 10 and 11) are strict.*

Hence, all proposed FAIRRETs can indeed be properly regarded as quantifications of unfairness.

They are differentiable with respect to h. Violation FAIRRETs owe this to the differentiability of γ and projection FAIRRETs to the differentiability of D. Hence, FAIRRETs are easily implemented with an automatic differentiation library like PyTorch. The computational overhead is unaffected by the complexity of the parameters $\boldsymbol{\theta}$ of h, as the gradients $\nabla_{\boldsymbol{\theta}} \mathcal{L}_Y = \left(\frac{\partial h}{\partial \boldsymbol{\theta}}\right)^\top \nabla_h \mathcal{L}_Y$ and $\nabla_{\boldsymbol{\theta}} R_\gamma = \left(\frac{\partial h}{\partial \boldsymbol{\theta}}\right)^\top \nabla_h R_\gamma$ of both loss functions in the joint objective (Eq. 8) share the computation of the Jacobian $\frac{\partial h}{\partial \boldsymbol{\theta}}$.

It is common to minimize $\mathcal{L}_Y$ using mini-batches; the same batches can be used to minimize R_γ. Indeed, this is done in our experiments. Though this makes FAIRRETs scalable, insufficient batch sizes will lead to poor approximations of the statistics γ. Clearly, the mean violation $\mathbf{v}(h)$ in a violation FAIRRET (Eq. 9) computed over mini-batches is not an unbiased estimate of the actual violation over all data. We report the mean SmoothMax loss for increasing batch sizes in the Appendix.

4 Experiments

4.1 Setup

Experiments were conducted on the *Bank* (Moro et al., 2014), *CreditCard* (Yeh & hui Lien, 2009), *LawSchool*[4], and *ACSIncome* (Ding et al., 2021) datasets. Each has multiple sensitive features, including some continuous. The classifier h was a fully connected neural net with hidden layers of sizes [256, 128, 32] followed by a sigmoid and did not take sensitive features $\mathbf{S}$ as input. We trained with all FAIRRETs discussed in Sec. 3 but only report results of Norm, D_{JS}-projection and D_{SED}-projection in the Appendix to avoid clutter here. The remaining FAIRRETs, SmoothMax and D_{KL}-projection, were representative for their archetype. These are compared against three baselines implemented in the Fair Fairness Benchmark (FFB) by Han et al. (2023), as their implementation provides these baselines as loss terms in idiomatic PyTorch. They are *PRemover* (Kamishima et al., 2012), *HSIC* (Pérez-Suay et al., 2017), and *AdvDebias* (Adel et al., 2019) (where the reverse of the adversary's loss is the fairness loss term). In contrast to the FAIRRET implementations, they only accept a single, categorical sensitive attribute. Each FAIRRET and FFB fairness loss was added to the cross-entropy loss in the objective (Eq. 8) in a separate training run for a range of strengths $\lambda > 0$.

We measured fairness over the four statistics γ in Table 1 that relate to Demographic Parity (DP), Equal Opportunity (EO), Predictive Parity (PP), and Treatment Equality (TE) respectively. Violation of each fairness notion is computed as $\max_k \mathbf{v}_k(h)$ (see Eq. (9)). Each FAIRRET was minimized with respect to each γ in a separate training run (and only the optimized violation is reported). The three FFB baselines only consider one fairness notion, which is to maximize independence between the model's output and the sensitive attributes. Their violation is reported for each statistic γ.

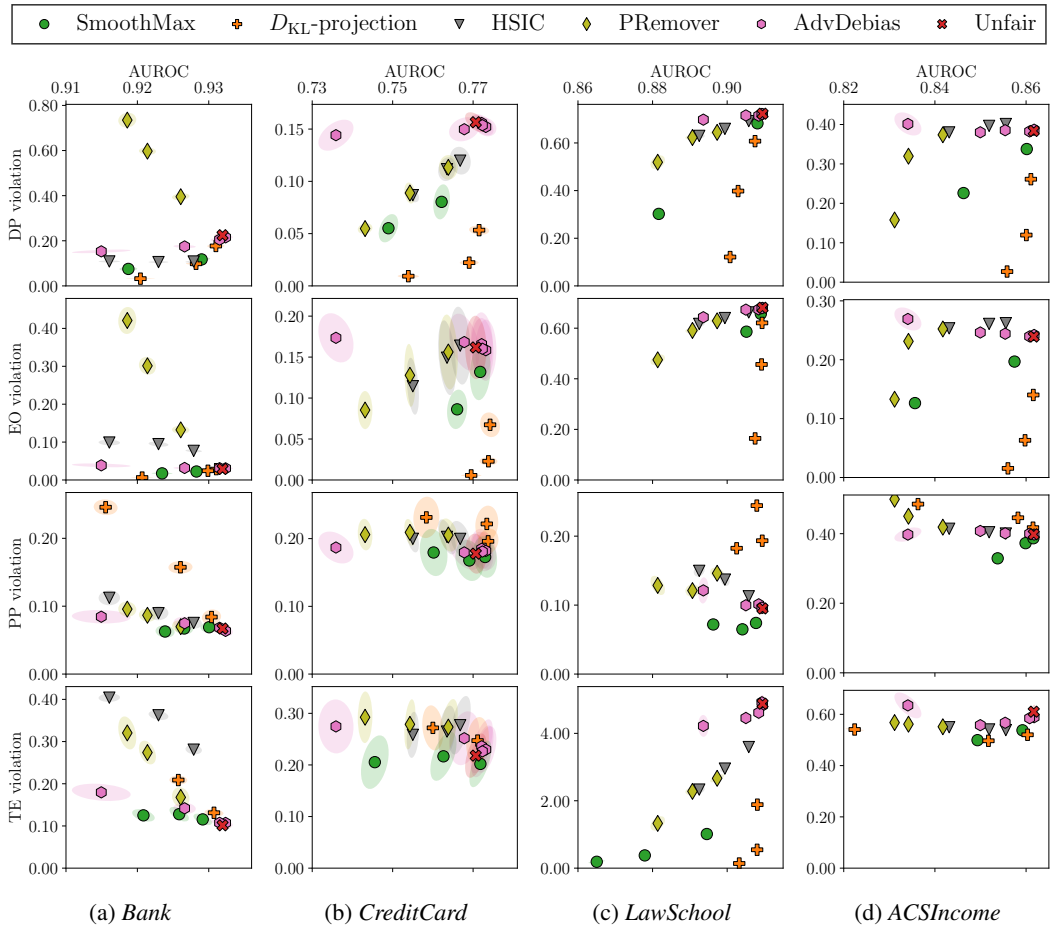
In summary, there was an experiment run for each dataset, fairness method, fairness strength λ, and statistic γ (except for the FFB baselines). Finally, we also use the *Unfair* baseline with $\lambda = 0$. Each of these combinations was repeated across 10 random seeds with each different train/test splits.

4.2 Results

Test set results are visualized in Fig. 2; train set results are found in the Appendix (and display the same trends). We separately discuss the notions with linear and with linear-fractional statistics.

For DP and EO, which have *linear* statistics, both the SmoothMax and D_{KL}-projection FAIRRETs are effectively used to minimize the fairness violation with respect to multiple sensitive attributes while minimally suffering a loss in AUROC scores, though the projection FAIRRET clearly performs better than the violation-based SmoothMax FAIRRET. As expected, the FFB baselines perform worse than the methods implemented in our FAIRRET framework, since they cannot be configured to optimize the same general range of fairness definitions. Also, their implementation only minimizes bias with respect to a single sensitive attribute, and so they are oblivious to some of the components in $\mathbf{S}$ that the violation in Fig. 2 measures. We report their violations on this single attribute in the Appendix, though the FAIRRETs still outperform them there as well.

[4]Curated and published by the SEAPHE project



(a) *Bank* (b) *CreditCard* (c) *LawSchool* (d) *ACSIncome*

Figure 2: Mean test set results with confidence ellipse for the standard error. Each marker is a separate combination of dataset, FAIRRET, FAIRRET strength, and statistic. Results in the lower right are optimal. Failed runs (with an AUROC far worse than the rest) are omitted.

For PP and TE, which have *linear-fractional statistics*, all methods appear to struggle far more. SmoothMax is most consistent and never makes the fairness violation worse, yet the D_{KL}-projection in most cases makes both the fairness violation and the AUROC worse. The same occurs for the FFB baselines. To some extent, this can be attributed to overfitting, as SmoothMax leads to a significantly more consistent reduction of the train set fairness violation than the test set (see Appendix). Still, non-linear fairness notions are clearly harder to optimize, which aligns with the results of Celis et al. (2019). Though Barocas et al. (2019) conclude that sufficiency (a notion related to PP) 'often comes for free', further work is needed to better understand how such notions can be consistently achieved.

5 Conclusion

The FAIRRET framework allows for a wide range of fairness definitions by comparing linear-fractional statistics for each sensitive feature. We implement several FAIRRETs and show how they are easily integrated in existing machine learning pipelines utilizing automatic differentiation.

Empirically, violation FAIRRETs like SmoothMax consistently lead to trade-offs between fairness and AUROC, though the more involved projection FAIRRETs like the D_{KL}-projection clearly outperform them on fairness definitions with linear statistics. However, all methods struggle with fairness notions that have linear-fractional statistics like PP and TE, which have mostly been ignored in prior work. This signals a lucrative direction for future research.

ETHICS STATEMENT

The FAIRRET framework was made as a technical tool to help unveil and address a mathematical formalization of fairness in machine learning systems. However, such tools should never be considered a sufficient solution to truly achieve fairness in real-world decision processes because the social, human component of fairness is completely outside the control of this framework (Selbst et al., 2019). There is a significant risk that technologies such as ours may anyway be abused to suggest discriminatory bias has been 'removed' from a decision process without actually addressing underlying injustices (Hoffmann, 2019).

REPRODUCIBILITY

All proofs, i.e. for Table 1, Prop. 1 and Prop. 2, are found in the Appendix A. Appendix B contains additional context for Remarks 1, 2 and 3. Appendix C provides experiments referred to in Sec. 3.2: a visualization of the projections of the projection FAIRRETs and an empirical assessment of their approximation for fewer optimization iterations. It also evaluates the mean SmoothMax loss for smaller batch sizes mentioned in Sec. 3.3. Furthermore, Appendix C extends the main experiment results of Sec. 4.2 by providing the metrics of the other FAIRRETs, the train set results and fairness violations computed for only a single sensitive attribute. Finally, Appendix D further explains the experiment setup already summarized in Sec. 4.1, i.e. the datasets, hyperparameters, the baselines implementation, the computation of the confidence ellipses and runtimes.

The code for our full experiment pipeline is found in the rest of the supplementary material.

The streamlined package is available at `https://github.com/aida-ugent/fairret`.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the Special Research Fund (BOF) of Ghent University (BOF20/DOC/144 and BOF20/IBF/117), from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme, and from the FWO (project no. G0F9816N, 3G042220, G073924N).

REFERENCES

Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-Network Adversarial Fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2412–2420, July 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33012412.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 60–69. PMLR, July 2018.

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

Wael Alghamdi, Shahab Asoodeh, Hao Wang, Flavio P. Calmon, Dennis Wei, and Karthikeyan Natesan Ramamurthy. Model Projection: Theory and Applications to Fair Machine Learning. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2711–2716. IEEE, June 2020. doi: 10.1109/ISIT44484.2020.9173988.

Shun-Ichi Amari. α -Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, November 2009. ISSN 1557-9654. doi: 10.1109/TIT.2009.2030485.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar,

Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL https://arxiv.org/abs/1810.01943.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A Convex Framework for Fair Regression, June 2017.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50 (1):3–44, 2021. doi: 10.1177/0049124118782533. URL https://doi.org/10.1177/0049124118782533.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020. URL https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK ; New York, 2004. ISBN 978-0-521-83378-3.

Maarten Buyl and Tijl De Bie. The KL-Divergence Between a Graph Model and its Fair I-Projection as a Fairness Regularizer. In *Machine Learning and Knowledge Discovery in Databases*, pp. 351–366. Springer International Publishing, 2021.

Maarten Buyl and Tijl De Bie. Optimal Transport of Classifiers to Fairness. *Advances in Neural Information Processing Systems*, 35:33728–33740, December 2022.

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*, pp. 319–328, Atlanta GA USA, January 2019. ACM. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287586.

Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, June 2017. ISSN 2167-6461. doi: 10.1089/big.2016.0047.

Ching-Yao Chuang and Youssef Mroueh. FAIR MIXUP: FAIRNESS VIA INTERPOLATION. *International Conference on Learning Representations*, 2020.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

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*, ITCS '12, pp. 214–226, New York, NY, USA, January 2012. Association for Computing Machinery. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255.

Danilo Franco, Nicolò Navarin, Michele Donini, Davide Anguita, and Luca Oneto. Deep fair models for complex data: Graphs labeling and explainable face recognition. *Neurocomputing*, 470:318–334, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.05.109. URL https://www.sciencedirect.com/science/article/pii/S0925231221011140.

Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. FFB: A Fair Fairness Benchmark for In-Processing Group Fairness Methods, June 2023.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Anna Lauren Hoffmann. Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7):900–915, June 2019. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2019.1573912.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-Aware Classifier with Prejudice Remover Regularizer. In Peter A. Flach, Tijl De Bie, and Nello Cristianini (eds.), *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pp. 35–50, Berlin, Heidelberg, 2012. Springer. ISBN 978-3-642-33486-3. doi: 10.1007/978-3-642-33486-3_3.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2564–2572. PMLR, July 2018.

Jan Kukačka, Vladimir Golkov, and Daniel Cremers. Regularization for Deep Learning: A Taxonomy, October 2017.

Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. Too Relaxed to Be Fair. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6360–6369. PMLR, November 2020.

Sode Masashi. Fairtorch. `https://github.com/wbawakate/fairtorch`, Dec 2020. Version 0.1.2.

Daniel McNamara, Cheng Soon Ong, and Robert C. Williamson. Costs and Benefits of Fair Representation Learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 263–270, Honolulu HI USA, January 2019. ACM. ISBN 978-1-4503-6324-2. doi: 10.1145/3306618.3317964.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):115:1–115:35, July 2021. ISSN 0360-0300. doi: 10.1145/3457607.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2014.03.001. URL `https://www.sciencedirect.com/science/article/pii/S016792361400061X`.

Luca Oneto, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, and Massimiliano Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15360–15370. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/af9c0e0c1dee63e5acad8b7ed1a5be96-Paper.pdf`.

Manisha Padala and Sujit Gujar. FNNC: Achieving fairness through neural networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, pp. 2277–2283, Yokohama, Yokohama, Japan, January 2021. ISBN 978-0-9992411-6-5.

Kirtan Padh, Diego Antognini, Emma Lejal-Glaude, Boi Faltings, and Claudiu Musat. Addressing fairness in classification with a model-agnostic multi-objective algorithm. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 600–609. PMLR, December 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair Kernel Learning. In Michelangelo Ceci, Jaakko Hollmén, Ljupčo Todorovski, Celine Vens, and Sašo Džeroski (eds.), *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pp. 339–355, Cham, 2017. Springer International Publishing. ISBN 978-3-319-71249-9. doi: 10.1007/978-3-319-71249-9_21.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68, Atlanta GA USA, January 2019. ACM. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287598.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pp. 1–7, Gothenburg Sweden, May 2018. ACM. ISBN 978-1-4503-5746-3. doi: 10.1145/3194770.3194776.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law and Security Review 41*, (3547922), Mar 2020. doi: 10.2139/ssrn.3547922. URL https://papers.ssrn.com/abstract=3547922.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio Calmon. Optimized Score Transformation for Fair Classification. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 1673–1683. PMLR, June 2020.

Michael Wick, swetasudha panda, and Jean-Baptiste Tristan. Unlocking Fairness: A Trade-off Revisited. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

I-Cheng Yeh and Che hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2, Part 1):2473–2480, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2007.12.020. URL https://www.sciencedirect.com/science/article/pii/S0957417407006719.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019. ISSN 1533-7928.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 325–333. PMLR, May 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 335–340, New York, NY, USA, December 2018. Association for Computing Machinery. ISBN 978-1-4503-6012-8. doi: 10.1145/3278721.3278779.