Beyond Adult and COMPAS: Fairness in Multi-Class Prediction

Wael Alghamdi [*1] **Hsiang Hsu** [*1] **Haewon Jeong** [*1] **Hao Wang** [1] **Peter Winston Michalak** [1] **Shahab Asoodeh** [2] **Flavio P. Calmon** [1]

Abstract

We produce fair probabilistic classifiers for multi-class prediction via "projecting" a pre-trained classifier onto the set of models that satisfy target group-fairness requirements. The new, projected model is given by post-processing the outputs of the pre-trained classifier by a multiplicative factor. We provide a parallelizable iterative algorithm for computing the projected classifier, and derive both sample complexity and convergence guarantees. Comprehensive numerical comparisons with state-of-the-art benchmarks demonstrate that our approach maintains competitive performance in terms of accuracy-fairness trade-off curves, while achieving favorable runtime on large datasets.

1. Introduction

Group-fairness interventions aim to ensure that a machine learning (ML) model does not discriminate based on, for example, race and/or sex. Extensive comparisons between discrimination control methods can be found in (Bellamy et al., 2018; Friedler et al., 2019; Wei et al., 2021). As these studies demonstrate, there is still no "best" fairness intervention for ML, with the majority of existing approaches tailored to binary classification tasks, binary population groups, or both. Moreover, discrimination control methods are often tested on overused datasets of modest size collected in either the US or Europe (e.g., UCI Adult (Lichman, 2013) and COMPAS (Angwin et al., 2016)). While binary classification covers a range of ML tasks of societal importance, there are many cases where the predicted variable is not binary, e.g., in education (grading scales) and healthcare (disease severity). Even the original COMPAS algorithm assigned a score between 1 to 10 to pre-trial defendants.

[*]Equal contribution [1]Harvard University, Cambridge, MA, USA [2]Department of Computing and Software, McMaster University, Hamilton, Ontario, Canada. Correspondence to: Wael Alghamdi <alghamdi@g.harvard.edu>, Flavio P. Calmon <flavio@seas.harvard.edu>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022.

We introduce a theoretically-grounded discrimination control method that ensures group fairness in multi-class prediction for several population groups. When restricted to two predicted classes, our method performs competitively against state-of-the-art fairness interventions tailored to binary classification. Our approach is based on the information theoretic formulation of *information projection*: Given a probability distribution P and a convex set of distributions $\mathcal{P}$, what is the "closest" distribution to P in $\mathcal{P}$? The study of information projection can be traced back to (Csiszár, 1975), which used KL-divergence to measure "closeness"; extensions followed for f-divergences (Csiszár, 1995), Rényi divergences (Kumar & Sason, 2016; Kumar & Sundaresan, 2015), and conditional distributions (Alghamdi et al., 2020).

Prior work on information projection relies on a critical—and limiting—assumption: the underlying distributions are *known exactly*. This is infeasible in practical ML applications, where only a set of training samples is available. We fill this gap by introducing an efficient procedure for computing the projected classifier given finite samples, called `FairProjection` (`FP`). We also establish convergence and sample complexity guarantees. Notably, our procedure is parallelizable (e.g., on a GPU). Thus, `FairProjection` scales to datasets with sizes comparable to the population of many US states ($> 10^6$ samples).

Related work. The differentiating factors from prior work are in Table 1. The fairness interventions that are the most similar to ours are the FairScoreTransformer (Wei et al., 2020; 2021, FST) and the pre-processing method in (Jiang & Nachum, 2020). The FST and (Jiang & Nachum, 2020) can be viewed as an instantiation of `FairProjection` restricted to binary classification and cross-entropy (for FST) or KL-divergence (for (Jiang & Nachum, 2020)) as the f-divergence of choice. Unlike our method, (Jiang & Nachum, 2020) requires retraining a classifier multiple times.

Notation. The entries of a vector $\boldsymbol{z}$ are denoted by z_j. We set $[N] := \{1, \cdots, N\}$ and $\mathbb{R}_+ \triangleq [0, \infty)$. The probability simplex over $[N]$ is denoted by $\boldsymbol{\Delta}_N$. If P is a Borel probability measure over $\mathbb{R}^N$, $Z \sim P$ is a random variable, and $f : \mathbb{R}^N \to \mathbb{R}^K$ is Borel, then the expectation of $f(Z)$ is denoted by $\mathbb{E}[f(Z)] = \mathbb{E}_P[f] = \mathbb{E}_{Z\sim P}[f(Z)]$.

Method	Feature						
	Multiclass	Multigroup	Scores	Curve	Parallel	Rate	Metric
Reductions (Agarwal et al., 2018)	✗	✓	✓	✓	✗	✓	SP, (M)EO
Reject-option (Kamiran et al., 2012)	✗	✓	✗	✓	✗	✗	SP, (M)EO
EqOdds (Hardt et al., 2016)	✗	✓	✗	✗	✗	✓	EO
LevEqOpp (Chzhen et al., 2019)	✗	✗	✗	✗	✗	✗	FNR
CalEqOdds (Pleiss et al., 2017)	✗	✗	✓	✗	✗	✓	MEO
FACT (Kim et al., 2020)	✗	✗	✗	✓	✗	✗	SP, (M)EO
Identifying (Jiang & Nachum, 2020)	✗	✓	✓	✓	✗	✗	SP, (M)EO
FST (Wei et al., 2020; 2021)	✗	✓	✓	✓	✗	✓	SP, (M)EO
Overlapping (Yang et al., 2020)	✓	✓	✓	✓	✗	✗	SP, (M)EO
Adversarial (Zhang et al., 2018)	✓	✓	N/A[1]	✓	✓	✗	SP, (M)EO
`FairProjection` (ours)	✓	✓	✓	✓	✓	✓	SP, (M)EO

Table 1. Comparison between benchmark methods. **Multiclass/multigroup**: implementation takes datasets with multiclass/multigroup labels; **Scores**: processes raw outputs of probabilistic classifiers; **Curve**: outputs fairness-accuracy tradeoff curves (instead of a single point); **Parallel**: parallel implementation (e.g., on GPU) is available; **Rate**: convergence rate or sample complexity guarantee is proved. **Metric**: applicable fairness metric, with SP↔Statistical Parity, EO↔Equalized Odds, MEO↔Mean EO, FNR↔False-Negative Rate.

2. Preliminaries and problem formulation

Fair ML. We fix two random variables X and Y, taking values in sets $\mathcal{X} \triangleq \mathbb{R}^d$ and $\mathcal{Y} \triangleq [C]$. A probabilistic classifier is a function $\boldsymbol{h} : \mathcal{X} \to \boldsymbol{\Delta}_C$, where $h_c(x)$ represents the probability of sample $x \in \mathcal{X}$ falling in class $c \in \mathcal{Y}$. Let S be a group attribute (e.g., race and/or sex), taking values in $\mathcal{S} \triangleq [A]$. We consider multi-class generalization of three commonly used group fairness criteria in Table 2. As observed by existing works (see, e.g., Agarwal et al., 2018; Menon & Williamson, 2018; Celis et al., 2019; Wei et al., 2020; Alghamdi et al., 2020), each of these fairness constraints can be written in the vector-inequality form $\mathbb{E}_{P_X}[\boldsymbol{Gh}] \leq \mathbf{0}$ for a closed-form matrix-valued function $\boldsymbol{G} : \mathcal{X} \to \mathbb{R}^{K \times C}$. For instance, for statistical parity, the $\boldsymbol{G}$ matrix evaluated at a fixed individual $x \in \mathcal{X}$ has $K = 2AC$ rows indexed by $(\delta, a, c') \in \{0,1\} \times [A] \times [C]$, where the (δ, a, c')-th row is equal to

$$\left(\frac{\sum_{c \in [C]} P_{S|X=x,Y=c}(a) h_c^{\text{base}}(x)}{(-1)^\delta P_S(a)} - (\alpha + (-1)^\delta) \right) \boldsymbol{e}_{c'}$$

with $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_C$ denoting the standard basis for $\mathbb{R}^C$.

Fairness via information-projection. For a search space $\mathcal{H} \subset \boldsymbol{\Delta}_C^{\mathcal{X}} \triangleq \{\boldsymbol{h} : \mathcal{X} \to \boldsymbol{\Delta}_C\}$, loss $\mathsf{err} : \boldsymbol{\Delta}_C^{\mathcal{X}} \times \boldsymbol{\Delta}_C^{\mathcal{X}} \to \mathbb{R}$, and base classifier $\boldsymbol{h}^{\text{base}} \in \boldsymbol{\Delta}_C^{\mathcal{X}}$, one seeks to solve:

$$\underset{\boldsymbol{h} \in \mathcal{H}}{\text{minimize}}\ \mathsf{err}\left(\boldsymbol{h}, \boldsymbol{h}^{\text{base}}\right) \text{ subject to } \mathbb{E}_{P_X}[\boldsymbol{Gh}] \leq \mathbf{0}. \quad (1)$$

[1](Zhang et al., 2018) is an in-processing method unlike other benchmarks in the table. It does not take a pre-trained classifier as an input.

Fairness Criterion	Expression
Statistical parity	$\left\lvert \frac{P_{\widehat{Y}\mid S=a}(c')}{P_{\widehat{Y}}(c')} - 1 \right\rvert \leq \alpha$
Equalized odds	$\left\lvert \frac{P_{\widehat{Y}\mid Y=c,S=a}(c')}{P_{\widehat{Y}\mid Y=c}(c')} - 1 \right\rvert \leq \alpha$
Overall accuracy equality	$\left\lvert \frac{P(\widehat{Y}=Y \mid S=a)}{P(\widehat{Y}=Y)} - 1 \right\rvert \leq \alpha$

Table 2. Standard group fairness criteria; one fixes $\alpha > 0$ and iterates over all $(a, c, c') \in [A] \times [C]^2$.

The function err quantifies "closeness" between the scores of $\boldsymbol{h}$ and $\boldsymbol{h}^{\text{base}}$, and one choice is f-divergence:

$$\mathsf{err}\left(\boldsymbol{h}, \boldsymbol{h}^{\text{base}}\right) = D_f(\boldsymbol{h} \| \boldsymbol{h}^{\text{base}} \mid P_X) \quad (2)$$
$$\triangleq \mathbb{E}_{P_X}\left[\sum_{c \in [C]} h_c^{\text{base}}(X) f\left(\frac{h_c(X)}{h_c^{\text{base}}(X)} \right) \right],$$

where f is a convex function over $(0, \infty)$ satisfying $f(1) = 0$. By varying different choices of f, we can obtain, e.g., cross-entropy (CE, $f(t) = -\log t$) and KL-divergence ($f(t) = t \log t$). For a chosen f-divergence, the optimization problem (1) becomes a generalization of *information projection* (Csiszár, 1975).

Recall the definition of the convex conjugate.

Definition 2.1. For $\boldsymbol{p} \in \boldsymbol{\Delta}_C$, let $D_f^{\text{conj}}(\cdot, \boldsymbol{p})$ denote the convex conjugate of $D_f(\cdot \| \boldsymbol{p})$:

$$D_f^{\text{conj}}(\boldsymbol{v}, \boldsymbol{p}) \triangleq \sup_{\boldsymbol{q} \in \boldsymbol{\Delta}_C} \boldsymbol{v}^T \boldsymbol{q} - D_f(\boldsymbol{q} \| \boldsymbol{p}). \quad (3)$$

In (Alghamdi et al., 2020), it is shown that (1) under an f-divergence loss (2) reduces to solving[2]

$$D^* \triangleq \min_{\boldsymbol{\lambda}\in\mathbb{R}^K_+} \mathbb{E}\left[D_f^{\text{conj}}\left(-\boldsymbol{G}(X)^T\boldsymbol{\lambda}, \boldsymbol{h}^{\text{base}}(X)\right)\right]. \quad (4)$$

Using gradient-based methods to optimize (4) would be hindered due to the intractability of ∇D_f^{conj} in most cases. We address this problem next.

Problem formulation. In practice, only data points $\mathbb{X} \triangleq \{X_i\}_{i\in[N]} \subset \mathcal{X}$ drawn from P_X are available. Thus, we propose the following fairness intervention problem. We search for a (multi-class) classifier $\boldsymbol{h} : \mathbb{X} \to \boldsymbol{\Delta}_C$ that solves:

$$\begin{aligned} \underset{\substack{\boldsymbol{h}:\mathbb{X}\to\boldsymbol{\Delta}_C \\ \boldsymbol{a}:\mathbb{X}\to\mathbb{R}^C, \boldsymbol{b}\in\mathbb{R}^K}}{\text{minimize}} \quad & D_f\left(\boldsymbol{h} \,\|\, \boldsymbol{h}^{\text{base}} \mid \widehat{P}_X\right) + \tau_1 \cdot (\|\boldsymbol{a}\|_2^2 + \|\boldsymbol{b}\|_2^2) \\ \text{subject to} \quad & \mathbb{E}_{\widehat{P}_X}\left[\boldsymbol{G}\cdot(\boldsymbol{h} + \tau_2\boldsymbol{a})\right] \le \tau_2\boldsymbol{b}, \end{aligned} \quad (5)$$

with $\widehat{P}_X$ the empirical measure, $\tau_1, \tau_2 > 0$ are prescribed constants, and $\|\boldsymbol{a}\|_2^2 \triangleq \mathbb{E}_{X\sim\widehat{P}_X}\left[\|\boldsymbol{a}(X)\|_2^2\right]$.[3]

Strong duality. We show that the unique solution for our fairness optimization problem (5) is a tilt of $\boldsymbol{h}^{\text{base}}$, under the following assumption.

Assumption 2.2. $f \in \mathcal{C}^2(\mathbb{R})$, $f(1) = 0$, $f'(0^+) = -\infty$, $f''(t) > 0$ for $t > 0$, and $h_c^{\text{base}}(x) > 0$ for $(x, c) \in \mathbb{X}\times[C]$.

Theorem 2.3. *Suppose Assumption 2.2 holds, and set* $\zeta \triangleq \tau_2^2/(2\tau_1)$. *There exists a unique solution* $\boldsymbol{h}^{\text{opt},N}$ *to* (5), *and it is given by the formula*

$$h_c^{\text{opt},N}(x) = h_c^{\text{base}}(x)\phi\left(\boldsymbol{v}_c(x;\boldsymbol{\lambda}^*_{\zeta,N}) + \gamma(x;\boldsymbol{\lambda}^*_{\zeta,N})\right) \quad (6)$$

$(x, c) \in \mathbb{X}\times[C]$, *where:* **(i)** *the function* $\boldsymbol{v} : \mathbb{X}\times\mathbb{R}^K \to \mathbb{R}^C$ *is defined by* $\boldsymbol{v}(x;\boldsymbol{\lambda}) \triangleq -\boldsymbol{G}(x)^T\boldsymbol{\lambda}$; **(ii)** *the function* ϕ *denotes the inverse of* f'; **(iii)** *the function* $\gamma : \mathbb{X}\times\mathbb{R}^K \to \mathbb{R}$ *is characterized by* $\mathbb{E}_{c\sim\boldsymbol{h}^{\text{base}}(x)}\left[\phi\left(\gamma(x;\boldsymbol{\lambda}) + v_c(x;\boldsymbol{\lambda})\right)\right] = 1$ *and* **(iv)** $\boldsymbol{\lambda}^\star_{\zeta,N} \in \mathbb{R}^K$ *is the unique solution to the strongly convex problem*

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^K_+} \mathbb{E}_{\widehat{P}_X}\left[D_f^{\text{conj}}\left(\boldsymbol{v}(X;\boldsymbol{\lambda}), \boldsymbol{h}^{\text{base}}(X)\right)\right] + \frac{\zeta}{2}\left\|\boldsymbol{\mathcal{G}}_N^T\boldsymbol{\lambda}\right\|_2^2 \quad (7)$$

where $\boldsymbol{\mathcal{G}}_N \triangleq \left(\frac{\boldsymbol{G}(X_1)}{\sqrt{N}}, \cdots, \frac{\boldsymbol{G}(X_N)}{\sqrt{N}}, \boldsymbol{I}_K\right) \in \mathbb{R}^{K\times(NC+K)}$.

3. Fair projection and theoretical guarantees

Our algorithm. Theorem 2.3 yields a *practical* procedure for solving the functional optimization in equation (5): (i) compute the dual variables by solving the strongly convex optimization in (7); (ii) tilt the base classifier by using the dual variables according to (6). The `FairProjection` algorithm uses ADMM (Boyd et al., 2011) to solve the convex program (7). Algorithm 1 presents the steps of `FairProjection`. A salient feature of `FairProjection` is its *parallelizability*. Each step that is done for i varying over $[N]$ can be executed for each i in parallel. In particular, this applies to the most computationally intensive step, the $\boldsymbol{v}_i$-update step. For the KL-divergence case, minimizing $\boldsymbol{v} \mapsto D_f^{\text{conj}}(\boldsymbol{v}, \boldsymbol{p}_i) + \xi\|\boldsymbol{v}\|_2^2 + \boldsymbol{a}_i^T\boldsymbol{v}$ reduces to a fixed-point equation for the softmax function, whose Lipschitzness yields exponentially fast convergence for fixed-point methods. For a general f-divergence, we reduce the $\boldsymbol{v}_i$-update step to a tractable 1-dimensional root-finding problem that can also be solved efficiently.

Algorithm 1 : `FairProjection` for solving (7).

Input: divergence f, predictions $\{\boldsymbol{p}_i \triangleq \boldsymbol{h}^{\text{base}}(X_i)\}_{i\in[N]}$, constraints $\{\boldsymbol{G}_i \triangleq \boldsymbol{G}(X_i)\}_{i\in[N]}$, regularizer ζ, ADMM penalty ρ, and initializers $\boldsymbol{\lambda}$ and $(\boldsymbol{w}_i)_{i\in[N]}$.
Output: $h_c^{\text{opt},N}(x) \triangleq h_c^{\text{base}}(x)\cdot\phi(\gamma(x;\boldsymbol{\lambda}) + v_c(x;\boldsymbol{\lambda}))$.

$\boldsymbol{Q} \leftarrow \frac{\zeta}{2}\boldsymbol{I} + \frac{\rho}{2N}\sum_{i\in[N]}\boldsymbol{G}_i\boldsymbol{G}_i^T$
for $t = 1, 2, \cdots, t'$ **do**
  $\boldsymbol{a}_i \leftarrow \boldsymbol{w}_i + \rho\boldsymbol{G}_i^T\boldsymbol{\lambda}$ $\quad i \in [N]$
  $\boldsymbol{v}_i \leftarrow \underset{\boldsymbol{v}\in\mathbb{R}^C}{\text{argmin}}\, D_f^{\text{conj}}(\boldsymbol{v}, \boldsymbol{p}_i) + \frac{\rho+\zeta}{2}\|\boldsymbol{v}\|_2^2 + \boldsymbol{a}_i^T\boldsymbol{v}$ $\quad i \in [N]$
  $\boldsymbol{q} \leftarrow \frac{1}{N}\sum_{i\in[N]}\boldsymbol{G}_i\cdot(\boldsymbol{w}_i + \boldsymbol{v}_i)$
  $\boldsymbol{\lambda} \leftarrow \underset{\boldsymbol{\ell}\in\mathbb{R}^K_+}{\text{argmin}}\, \boldsymbol{\ell}^T\boldsymbol{Q}\boldsymbol{\ell} + \boldsymbol{q}^T\boldsymbol{\ell}$
  $\boldsymbol{w}_i \leftarrow \boldsymbol{w}_i + \rho\cdot(\boldsymbol{v}_i + \boldsymbol{G}_i^T\boldsymbol{\lambda})$ $\quad i \in [N]$
end for

Convergence guarantees. We show that our algorithm, `FairProjection`, converges exponentially fast, and that its output $\boldsymbol{\lambda}^{(\log N)}_{N^{-1/2},N}$ performs well for the *population* problem (4).

Theorem 3.1. *Under Assumption 2.2,* `FairProjection` *for KL-divergence converges in*[4] $t' = O(\log N)$ *steps, and runs in time* $O(N\log N)$, *to the unique solution* $\boldsymbol{\lambda}^*_{\zeta,N}$ *of* (7). *If* $\boldsymbol{\lambda}^{(t)}_{\zeta,N}$ *and* $\boldsymbol{h}^{(t)}$ *are the* t*-th iteration outputs of* `FairProjection`, *then* $\|\boldsymbol{\lambda}^{(t)}_{\zeta,N} - \boldsymbol{\lambda}^*_{\zeta,N}\|_2 = O(e^{-t})$ *and* $\boldsymbol{h}^{(t)}(x) = \boldsymbol{h}^{\text{opt},N}(x)\cdot(1 + O(e^{-t}))$ *uniformly as* $t \to \infty$.

Theorem 3.2. *Suppose Assumption 2.2 holds, and consider the KL-divergence case. Then, choosing* $\zeta = \Theta\left(N^{-1/2}\right)$ *and* $t = \Omega(\log N)$ *we obtain for any* $\delta \in (0, 1)$ *and* $N = \Omega(\log\frac{1}{\delta})$ *that, with probability* $1 - \delta$, *(see* (4)*)*

$$\mathbb{E}_X\left[D_{\mathsf{KL}}^{\text{conj}}\left(\boldsymbol{v}\left(X;\boldsymbol{\lambda}^{(t)}_{\zeta,N}\right), \boldsymbol{h}^{\text{base}}(X)\right)\right] \le D^\star + O\left(\frac{1}{\sqrt{N}}\right).$$

[2] See Theorems 1–2 in (Alghamdi et al., 2020) for the details.
[3] The terms $\boldsymbol{a}$ and $\boldsymbol{b}$ are added to circumvent infeasibility issues and aid convergence of our numerical procedure.
[4] We use the standard asymptotic notation O, Ω, and Θ.

Method	Reduction (Agarwal et al., 2018)	Rejection (Kamiran et al., 2012)	EqOdds (Hardt et al., 2016)	LevEqOpp (Chzhen et al., 2019)	CalEqOdds (Pleiss et al., 2017)	FP (ours)
Runtime	223.6	16.9	5.9	7.9	5.3	11.2

Table 3. Execution time of `FP-CE` on the ENEM (with 1.4M samples) compared with five baseline methods (time shown in minutes). Methods in **bold** are capable of producing the full fairness-accuracy trade-off curves.

Benefit of parallelization. The parallelizability of `FairProjection` provides significant speedup. We perform an ablation study comparing the speedup due to parallelization. For the ENEM dataset (discussed next section), parallelization yields a 15-fold reduction in runtime.

4. Empirical study

We show that `FP` (the constrained optimization in (5)) has competitive performance in terms of runtime and fairness-accuracy trade-off curves compared to benchmarks in Table 1 (if reproducible codes are available). We use cross-entropy (`FP-CE`) as the loss function, and mean equalized odds (MEO) as the fairness constraint (cf. Table 2).

We adopt two datasets: HSLS (Ingels et al., 2011) and a novel dataset ENEM (INEP, 2020). The HSLS dataset is collected from high school students in the USA, whose features include student and parent information, and the binary label Y is students' 9$^{\text{th}}$-grade math test scores. The ENEM dataset contains Brazilian college entrance exam scores with students' demographic and socio-economic information. It contains $\sim$1.4 million samples with 139 features, and the multiclass label Y is the Humanities exam score. For both datasets, race is used as the group attribute S, and is included as a feature for training (Agarwal et al., 2018).

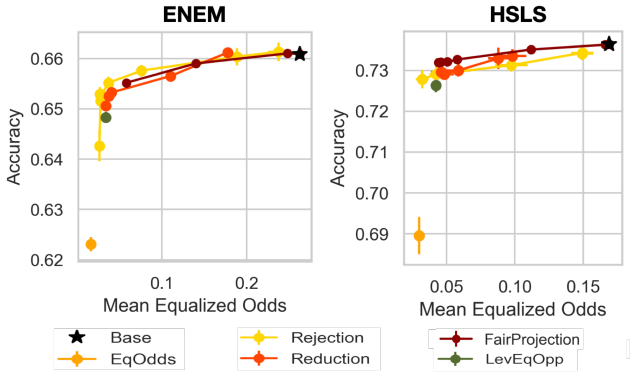

Figure 1. Fairness-accuracy trade-off comparisons between `FP-CE` and five baselines on ENEM and HSLS for binary class prediction. For all methods, random forest is the base classifier.

Binary-classes/groups. `FP-CE` is compared with benchmarks in terms of the MEO-accuracy trade-off (by varying α) on ENEM[5] and HSLS in Fig. 1. `FP-CE` and `Reduction` have the overall best and most consistent performances; although `EqOdds` achieves the best fairness, that fairness comes at the cost of 4% accuracy drop. `FP-CE` has the smallest accuracy drop whilst improving MEO from 0.17 to 0.04 on HSLS. `LevEqOpp` achieves comparable MEO with a slight accuracy drop. Note that in the high fairness regime, the accuracy of `Rejection` deteriorates. We exclude `CalEqOdds` since it yields inconsistent performance when the strict calibration requirement is not met.

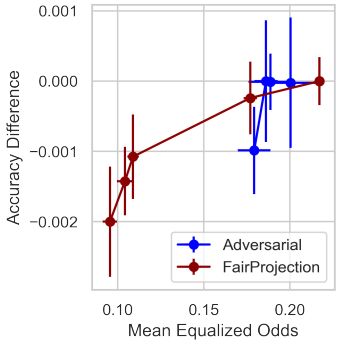

Figure 2. MEO-accuracy trade-off for multi-class prediction on ENEM. `FairProjection-CE` has a logistic regression base classifier. Base accuracy for `FP-CE` $= 0.336$, `Adversarial` $= 0.307$, and random guessing accuracy $= 0.2$.

Multi-classes. Fig. 2 shows MEO-accuracy trade-off of `FP-CE` and `Adversarial` (Zhang et al., 2018) on multi-class prediction with ENEM (Y quantized into 5 classes). As their base classifiers are different (`Adversarial` is based on GANs), we plot accuracy difference compared to the base classifier instead of the absolute value of accuracy. Compared to `Adversarial`, `FP-CE` improves MEO significantly with very small loss in accuracy.

Runtime. Table 3 report the runtime on ENEM with $\sim$ 1.4M samples. `EqOdds`, `LevEqOpp`, and `CalEqOdds` are faster than `FP-CE` as they only produce one trade-off point. The benchmarks that produce full fairness-accuracy trade-off curves have slower runtime than `FP`.

[5] We downsample ENEM with 50k samples, and binarize the labels and groups for consistent comparison.

References

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

Alghamdi, W., Asoodeh, S., Wang, H., Calmon, F. P., Wei, D., and Ramamurthy, K. N. Model projection: Theory and applications to fair machine learning. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2711–2716, 2020. doi: 10.1109/ISIT44484.2020.9173988.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica*, 2016.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, jan 2011. ISSN 1935-8237. doi: 10.1561/2200000016. URL https://doi.org/10.1561/2200000016.

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

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32, 2019.

Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pp. 146–158, 1975.

Csiszár, I. Generalized projections for non-negative functions. In *Proceedings of 1995 IEEE International Symposium on Information Theory*, pp. 6. IEEE, 1995.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 329–338, 2019.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

INEP. Instituto nacional de estudos e pesquisas educaionais anísio teixeira, microdados do ENEM. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem, 2020. Accessed: 2022-05-23.

Ingels, S. J., Pratt, D. J., Herget, D. R., Burns, L. J., Dever, J. A., Ottem, R., Rogers, J. E., Jin, Y., and Leinwand, S. High school longitudinal study of 2009 (hsls: 09): Base-year data file documentation. nces 2011-328. *National Center for Education Statistics*, 2011.

Jiang, H. and Nachum, O. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.

Kamiran, F., Karim, A., and Zhang, X. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929, Dec 2012. doi: 10.1109/ICDM.2012.45.

Kim, J. S., Chen, J., and Talwalkar, A. Fact: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*, pp. 5264–5274. PMLR, 2020.

Kumar, M. A. and Sason, I. Projection theorems for the rényi divergence on α-convex sets. *IEEE Transactions on Information Theory*, 62(9):4924–4935, 2016.

Kumar, M. A. and Sundaresan, R. Minimization problems based on relative α-entropy i: Forward projection. *IEEE Transactions on Information Theory*, 61(9):5063–5080, 2015.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

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

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017.

Wei, D., Ramamurthy, K. N., and Calmon, F. P. Optimized score transformation for fair classification. In *23rd International Conference on Artificial Intelligence and Statistics*, 2020.

Wei, D., Ramamurthy, K. N., and Calmon, F. P. Optimized score transformation for consistent fair classification. *Journal of Machine Learning Research*, 22(258): 1–78, 2021.

Yang, F., Cisse, M., and Koyejo, O. O. Fairness with overlapping groups; a probabilistic perspective. *Advances in Neural Information Processing Systems*, 33, 2020.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.