Beyond Adult and COMPAS: Fairness in Multi-Class Prediction

Wael Alghamdi*, Hsiang Hsu*, Haewon Jeong*, Hao Wang,
Peter Winston Michalak, Shahab Asoodeh, Flavio P. Calmon
Harvard University
alghamdi@g.harvard.edu, flavio@seas.harvard.edu; (* Equal contribution)

Main Idea

Given an unfair base classifier $\boldsymbol{h}^{\text{base}} : \mathcal{X} \to \boldsymbol{\Delta}_C$ from dataset $\mathcal{X} \subset \mathbb{R}^d$ to C possible classes, we produce the closest fair classifier $\boldsymbol{h}^{\text{opt}}$.

Several fairness criteria (e.g., Statistical Parity, Equalized Odds, Overall Accuracy Equality) can be written in linear form:

$$\mathbb{E}[\mathbf{G}(X) \cdot \boldsymbol{h}(X)] \leq \mathbf{0}.$$

We find $\boldsymbol{h}^{\text{opt}}$ from i.i.d. samples $\mathbb{X} = \{X_i\}_{i\in[N]} \subset \mathbb{R}^d$ by solving:

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

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

We transform the fairness problem into a finite-dimensional, strongly convex optimization problem, and propose a fast ADMM-based algorithm to solve it.

Strong Duality

The optimal classifier is given by a tilt:

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

where $\boldsymbol{v}(x,\boldsymbol{\lambda}) = -\mathbf{G}(x)^T\boldsymbol{\lambda}$, $\phi = (f')^{-1}$, and $\boldsymbol{\lambda}_{\zeta,N}^*$ is the unique solution to the strongly convex problem:

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

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

Proposed Parallel Algorithm

Algorithm 1 : `FairProjection`.

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**

$\quad \boldsymbol{a}_i \leftarrow \boldsymbol{w}_i + \rho\boldsymbol{G}_i^T\boldsymbol{\lambda} \qquad i \in [N]$

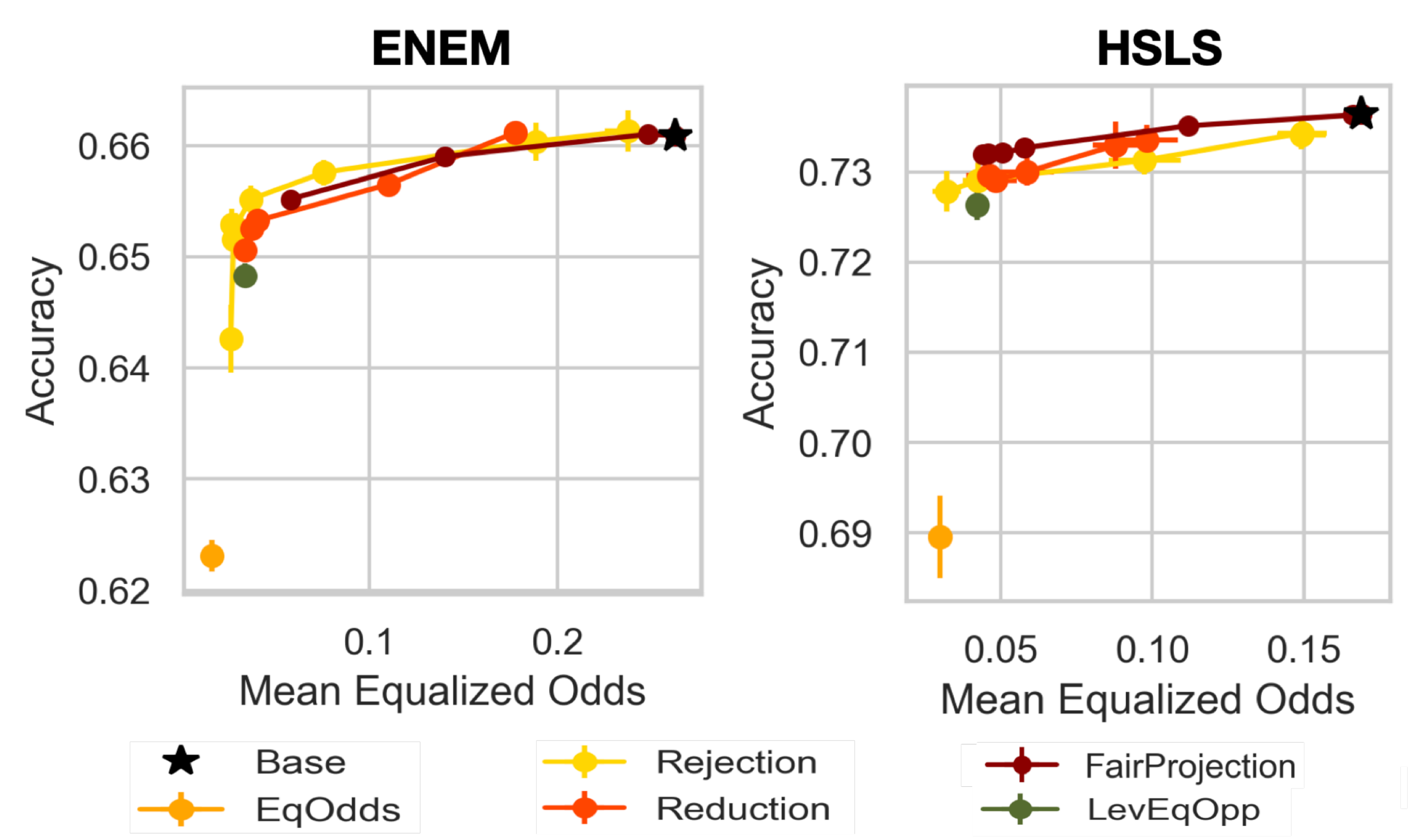
$\quad \boldsymbol{v}_i \leftarrow \underset{\boldsymbol{v}\in\mathbb{R}^C}{\text{argmin}}\, D_f^{\text{conj}}(\boldsymbol{v}, \boldsymbol{p}_i) + \frac{\rho+\zeta}{2}\|\boldsymbol{v}\|_2^2 + \boldsymbol{a}_i^T\boldsymbol{v} \qquad i \in [N]$

$\quad \boldsymbol{q} \leftarrow \frac{1}{N}\sum_{i\in[N]} \boldsymbol{G}_i \cdot (\boldsymbol{w}_i + \boldsymbol{v}_i)$

$\quad \boldsymbol{\lambda} \leftarrow \underset{\boldsymbol{\ell}\in\mathbb{R}_+^K}{\text{argmin}}\, \boldsymbol{\ell}^T\boldsymbol{Q}\boldsymbol{\ell} + \boldsymbol{q}^T\boldsymbol{\ell}$

$\quad \boldsymbol{w}_i \leftarrow \boldsymbol{w}_i + \rho \cdot (\boldsymbol{v}_i + \boldsymbol{G}_i^T\boldsymbol{\lambda}) \qquad i \in [N]$

end for



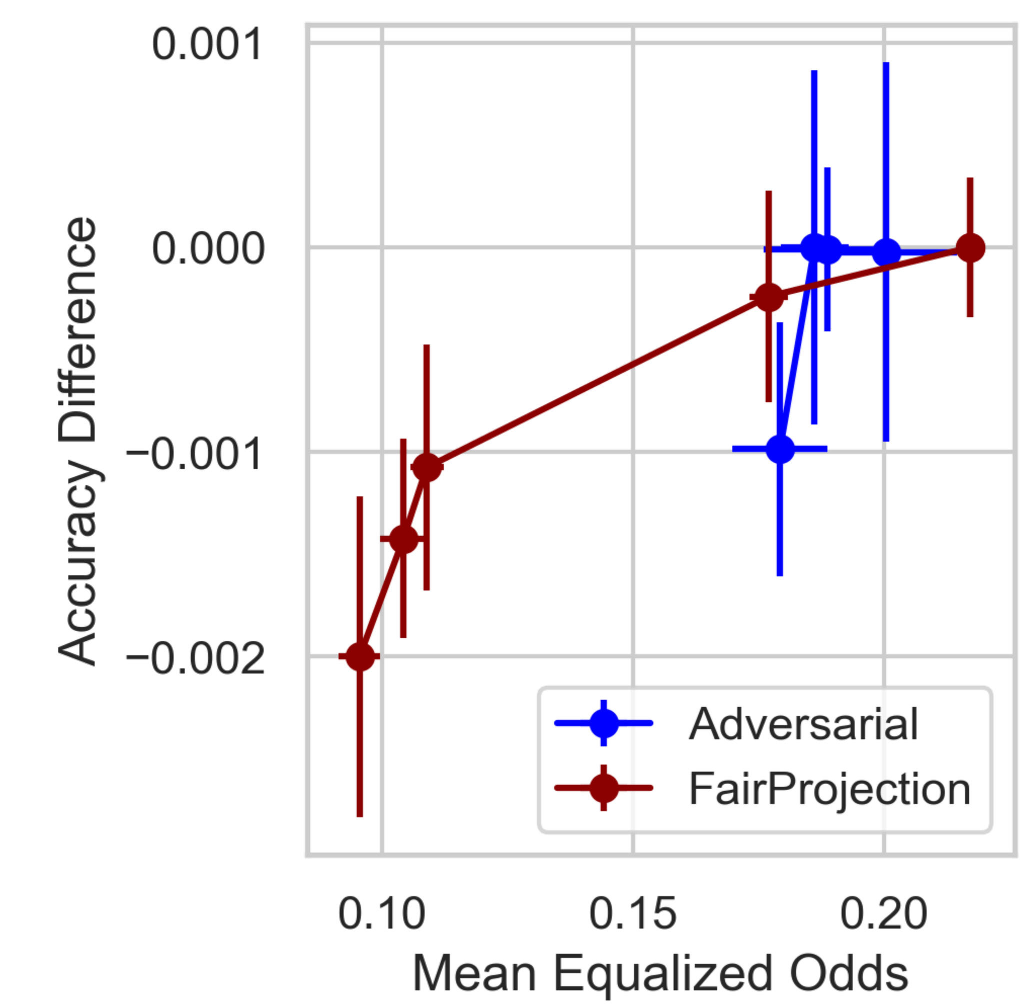
Fairness-accuracy trade-off comparisons between FairProjection-Cross Entropy and five baselines on the ENEM and HSLS datasets for binary class prediction. For all methods, random forest is the base classifier.

Theoretical Guarantees

For the KL-divergence:

1. `FairProjection` converges in $O(\log N)$ steps,
2. `FairProjection` runs in time $O(N\log N)$,
3. `FairProjection` converges to the unique solution $\boldsymbol{\lambda}_{\zeta,N}^*$,
4. the t-th iteration satisfies $\|\boldsymbol{\lambda}_{\zeta,N}^{(t)} - \boldsymbol{\lambda}_{\zeta,N}^*\|_2 = O(e^{-t})$,
5. the t-th iteration satisfies $\boldsymbol{h}^{(t)}(x) = \boldsymbol{h}^{\text{opt},N}(x) \cdot (1 + O(e^{-t}))$ uniformly,
6. for $\zeta = \Theta(N^{-1/2})$, `FairProjection` is within $O(N^{-1/2})$ from solving the population problem.

`FairProjection` is parallelizable: the inner routines of `FairProjection` can be executed in parallel for each sample X_i, $i \in [N]$.



MEO-accuracy trade-off for multi-class prediction on the ENEM dataset. FairProjection-Cross Entropy has a logistic regression base classifier. Base accuracy for FairProjection is = 0.336, Adversarial = 0.307, and random guessing accuracy = 0.2.