Addressing Fairness in Classification with a Model-Agnostic Multi-Objective Algorithm (Supplementary material)

Kirtan Padh[*1] **Diego Antognini**[2] **Emma Lejal-Glaude**[3] **Boi Faltings**[2] **Claudiu Musat**[3]

[1]Helmholtz AI, Germany
[2]Ecole Polytechnique Fédérale de Lausanne, Switzerland
[3]Swisscom, Switzerland

1 TOY DATASET DESCRIPTION

Figure 1 provides a visualization of the toy dataset used in comparing the relaxations in Figure 1 in the main paper.

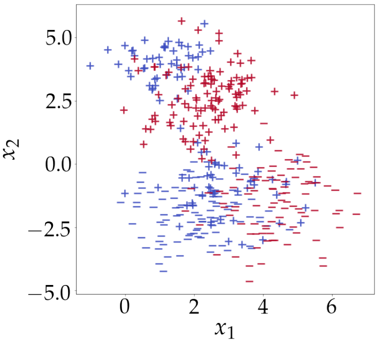


Figure 1: A visualization of the toy dataset used in Figure 1 in the main paper. The class labels are $(+)$ and $(-)$. The color represents group membership for a binary sensitive attribute, so the two groups are *red* and *blue*. So the goal is to separate the class labels, and remain fair with respect to the colors. The dataset contains 600 points, but only 400 are shown for clarity.

Dataset construction: The dataset is taken directly from (Lohaus et al., 2020). The points are drawn from various Gaussian distributions.

- *Protected sensitive attribute.* We draw 150 points with a negative label from a Gaussian with mean $\mu_1 = [2, -1]$ and covariance $\Sigma_1 = [[1, 0], [0, 1]]$. For the positive label we draw 150 points from a mixture of two Gaussians, with $\mu_2 = [3, -1]$ and $\Sigma_2 = [[1, 0], [0, 1]]$ and $\mu_3 = [1, 4]$ and $\Sigma_3 = [[0.5, 0], [0, 0.5]]$.
- *Unprotected sensitive attribute:* For the unprotected sensitive attribute, we draw 150 points with a positive label from a Gaussian with mean $\mu_4 = [2.5, 2.5]$ and covariance $\Sigma_4 = [[1, 0], [0, 1]]$. For the negative label we draw 150 points from a Gaussian with $\mu_5 = [4.5, -1.5]$ and $\Sigma_5 = [[1, 0], [0, 1]]$.

2 MAMO-FAIR ALGORITHM

Here we provide some further details on the multi-objective algorithm described in Section 5 of the main paper.

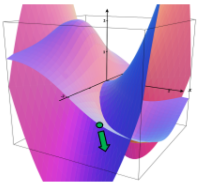


Figure 2: The figure gives an intuitive visualization of the common descent vector for two objectives. The two surfaces can be interpreted as loss functions for two objectives. The arrow points to the direction that minimizes both loss functions simultaneously.

Figure 2 gives an intuition for a key ingredient of the multi-objective algorithm, the common descent vector. Algorithm 1 provides the pseudocode for the algorithm.

3 SUPPORTED METRICS

Since the method is based on relaxing the indicator function, it supports all error-rate based metrics. We formally define some of them here. Table 1 in (Celis et al., 2019) provides an even more complete list. The Figure 3 defines metrics based on mis-classification rates of the prediction. We formally define some of the supported metrics next to give a general picture.

Definition 1 (False Positive Rate). Parity of false positive rate

$$\mathbb{P}[\hat{y} = 1 \mid a = -1,\, y = -1] = \mathbb{P}[\hat{y} = 1 \mid a = 1,\, y = -1]$$

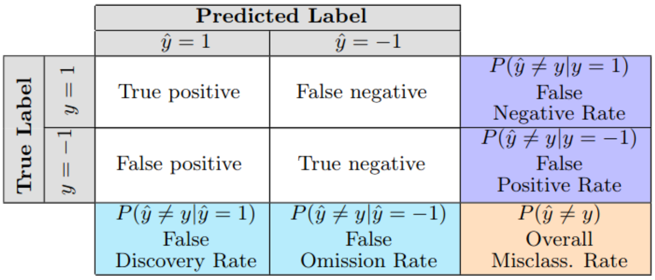
[*] Work done while at EPFL and Swisscom.

Table 1: **Results Table**: MF1 is the MAMO-fair algorithm optimizing separately for DEO and DDP, and MF2 is the algorithm optimizing simultaneously for DDP and DEO. SFa is the SearchFair algorith, Zaf is Zafar, Cot is Cotter, Unc is the unconstrained model and Con is the constant model

	Adult				Compas			
	Demographic parity		**Equality of opportunity**		**Demographic parity**		**Equality of opportunity**	
	|DDP|	Error	|DEO|	Error	|DDP|	Error	|DEO|	Error
MF1	0.09 ± 0.03	**0.18 ± 0.01**	0.05 ± 0.03	**0.18 ± 0.01**	0.04 ± 0.01	**0.32 ± 0.01**	0.08 ± 0.04	**0.33 ± 0.01**
MF2	0.08 ± 0.03	0.19 ± 0.02	0.04 ± 0.02	0.19 ± 0.02	0.11 ± 0.06	0.33 ± 0.01	0.11 ± 0.07	0.33 ± 0.01
SFa	0.00 ± 0.00	0.24 ± 0.00	0.05 ± 0.03	0.20 ± 0.01	**0.03 ± 0.01**	0.45 ± 0.02	0.01 ± 0.01	0.45 ± 0.01
Zaf	0.20 ± 0.01	0.18 ± 0.00	0.09 ± 0.06	0.20 ± 0.02	**0.03 ± 0.01**	0.42 ± 0.01	0.21 ± 0.06	0.33 ± 0.02
Cot	0.00 ± 0.00	0.24 ± 0.00	0.06 ± 0.04	0.20 ± 0.01	0.04 ± 0.02	0.40 ± 0.01	0.01 ± 0.01	0.45 ± 0.02
Unc	0.19 ± 0.01	0.17 ± 0.00	0.18 ± 0.03	0.17 ± 0.00	0.20 ± 0.02	0.32 ± 0.01	0.23 ± 0.05	0.32 ± 0.03
Con	0.00 ± 0.00	0.24 ± 0.00	0.00 ± 0.00	0.24 ± 0.00	0.00 ± 0.00	0.46 ± 0.01	0.00 ± 0.00	0.46 ± 0.01

Table 2: **Results Table**: MF1 is the MAMO-fair algorithm optimizing separately for DEO and DDP, and MF2 is the algorithm optimizing simultaneously for DDP and DEO. SFa is the SearchFair algorith, Zaf is Zafar, Cot is Cotter, Unc is the unconstrained model and Con is the constant model

	Dutch				CelebA			
	Demographic parity		**Equality of opportunity**		**Demographic parity**		**Equality of opportunity**	
	|DDP|	Error	|DEO|	Error	|DDP|	Error	|DEO|	Error
MF1	0.08 ± 0.03	**0.19 ± 0.01**	0.03 ± 0.01	**0.18 ± 0.00**	0.06 ± 0.02	0.16 ± 0.01	0.06 ± 0.02	**0.15 ± 0.03**
MF2	0.14 ± 0.01	0.19 ± 0.00	0.08 ± 0.02	0.19 ± 0.00	0.04 ± 0.01	0.16 ± 0.00	0.01 ± 0.01	0.16 ± 0.00
SFa	**0.02 ± 0.01**	0.23 ± 0.00	**0.01 ± 0.00**	0.18 ± 0.00	0.01 ± 0.01	0.17 ± 0.00	**0.01 ± 0.01**	0.17 ± 0.00
Zaf	0.03 ± 0.01	0.23 ± 0.00	0.01 ± 0.01	0.18 ± 0.00	0.17 ± 0.01	0.15 ± 0.00	0.16 ± 0.01	0.15 ± 0.00
Cot	0.01 ± 0.01	0.25 ± 0.01	**0.00 ± 0.00**	0.19 ± 0.00	**0.01 ± 0.01**	0.18 ± 0.00	0.03 ± 0.01	0.16 ± 0.00
Unc	0.16 ± 0.01	0.18 ± 0.01	0.08 ± 0.01	0.18 ± 0.01	0.20 ± 0.01	0.15 ± 0.00	0.16 ± 0.01	0.15 ± 0.00
Con	0.00 ± 0.00	0.48 ± 0.00	0.00 ± 0.00	0.48 ± 0.00	0.00 ± 0.00	0.48 ± 0.00	0.00 ± 0.00	0.48 ± 0.00



		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y \mid y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y \mid y = -1)$ False Positive Rate
		$P(\hat{y} \neq y \mid \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y \mid \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

Figure 3: Table from Zafar et al. (2017) on disparate mistreatment based measures. The table defines the rates, the measure of fairness corresponding to each rate is the parity of that rate across groups

Definition 2 (**False Negative Rate**). Parity of false negative rate across groups

$$\mathbb{P}[\hat{y} = -1 \,|\, a = -1,\, y = 1] = \mathbb{P}[\hat{y} = -1 \,|\, a = 1,\, y = 1]$$

Definition 3 (**True Positive Rate**). Parity of true positive rates across groups

$$\mathbb{P}[\hat{y} = 1 \,|\, a = -1,\, y = 1] = \mathbb{P}[\hat{y} = 1 \,|\, a = 1,\, y = 1]$$

Definition 4 (**True Negative Rate**). Parity of true positive rate across groups

$$\mathbb{P}[\hat{y} = -1 \,|\, a = -1,\, y = -1] = \mathbb{P}[\hat{y} = -1 \,|\, a = 1,\, y = -1]$$

The relaxation procedure follows the same principle as described in the main content, where each fairness notion is written as a difference of expectation, further relaxed to an empirical estimate of the expectation. As a last step $\mathbb{1}_{x>0}$ is relaxed to $\tanh(c * \max(0, x))$ and $\mathbb{1}_{x<0}$ is relaxed to $\tanh(c * \min(0, x))$.

4 RESULT TABLES

Table 1 and Table 2 provide full tables for the results described in Figure 2 in the main paper. We note that in a few cases both the error and fairness value are identical for more

Algorithm 1 Final algorithm with gradient normalization

```
for $i \in 1, ..., k$ do
    $EL_i = \ell_i(w)$
end for
for $epoch \in 1, ..., M$ do
    for $batch \in 1, ..., B$ do
        $forward\_pass()$
        $evaluate\_model()$
        for $i \in 1, ..., n$ do
            $loss = \ell_i(w)$
            $loss\_gradient = \nabla \ell_i(w)$
            $\nabla \overline{\ell_i(w)} = \frac{\nabla_w \ell_i(w)}{EL_i}$
        end for
        $\alpha_1, ..., \alpha_k = \text{QCOPSolver}\left(\nabla_w \overline{\ell_1(w)}, ..., \nabla_w \overline{\ell_k(w)}\right)$
        $\nabla_w L(w) = \sum_{i=1}^{k} \alpha_i \nabla_w \overline{\ell_i(w)}$
        $w = w - \eta \nabla_w L(w)$
    end for
end for
```

than one baseline method. In this case we slightly perturb one of the values to ensure that all points are visible in the figure in the main paper. The tables in this appendix provide the values without this perturbation.

5 PROOF OF THEOREM 1

Here we provide the proof of Theorem 1 from the main paper. First we give a reminder of the definition of the sign function

$$sign(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \end{cases} \tag{1}$$

Observation 1. *The hyperbolic tangent is an odd function, which is to say that*

$$\tanh(-x) = -\tanh(x)$$

Observation 2 (The quotient law of convergent series)**.** *Let (a_n) and (b_n) be convergent series such that $\lim_{n\to\infty} a_n = A$ and $\lim_{n\to\infty} b_n = B$. Then we have*

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{\lim_{n\to\infty} a_n}{\lim_{n\to\infty} b_n} = \frac{A}{B}$$

provided that $B \neq 0$.

Observation 2 is a commonly used result in real analysis. See Theorem C in (Freiwald, 2014) for a proof.

Theorem 1. *The hyperbolic tangent of $n * x$ converges to the sign of x for every fixed $x \in \mathbb{R}$ as n goes to infinity. Formally,*

$$\lim_{n\to\infty} \tanh(nx) = \text{sign}(x) \, \forall x \in \mathbb{R} \tag{2}$$

Proof. We know from the definition of the hyperbolic tangent function that

$$\tanh(nx) = \frac{1 - e^{-2nx}}{1 + e^{-2nx}} \tag{3}$$

The theorem requires pointwise convergence, meaning that the convergence in n should hold for each value of x. Therefore x can be though of as a constant for the purpose of the proof. Assuming x to be a constant let $a_n = 1 - e^{-2nx}$ and $b_n = 1 + e^{-2nx}$. Then we have

$$\tanh(nx) = \frac{a_n}{b_n} \tag{4}$$

We divide into cases by the value of x.

Case 1: $\mathbf{x > 0}$. In this case we have $\lim_{n\to\infty} e^{-2nx} = 0$. Therefore it follows that $\lim_{n\to\infty} a_n = 1$ and $\lim_{n\to\infty} b_n = 1$. From Equation 4 we know that $\tanh(nx)$ is a ratio of a_n and b_n. Therefore it follows from Observation 2 that

$$\lim_{n\to\infty} \tanh(nx) = \frac{\lim_{n\to\infty} a_n}{\lim_{n\to\infty} b_n} = 1$$

Case 2: $\mathbf{x < 0}$. Since $x < 0$, we have $-x > 0$. Therefore from case 1 we know $\lim_{n\to\infty} \tanh(n(-x)) = 1$. We have from Observation 1 that $\tanh(-nx) = -\tanh(nx)$. Therefore,

$$\lim_{n\to\infty} \tanh(nx) = -\lim_{n\to\infty} \tanh(n(-x)) = -1$$

Case 3: $\tanh(nx) = 0$ for $x = 0$. Therefore

$$\lim_{n\to\infty} \tanh(nx) = 0$$

Putting the three cases together we have

$$tanh(nx) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \end{cases}$$

This is identical to the definition of the sign function (Equation 1. Therefore,

$$\lim_{n\to\infty} \tanh(nx) = \text{sign}(x) \, \forall x \in \mathbb{R}$$

□

6 PROOF OF LEMMA 1

Lemma 1. $\tanh(n * \max(0, x))$ *converges to the indicator function of* $x > 0$ *as* n *goes to infinity. Formally,*

$$\lim_{n\to\infty} \tanh(n * \max(0, x)) = \mathbb{1}_{x>0} \, \forall x \in \mathbb{R} \tag{5}$$

Proof. We know from Theorem 1 that

$$\lim_{n\to\infty} \tanh(n * \max(0, x)) = \text{sign}(\max(0, x)) \tag{6}$$

Case 1: $x > 0$. When $x > 0$, $\max(0, x) = x$. Therefore we have $\text{sign}(\max(0, x)) = \text{sign}(x) = 1$. So $\text{sign}(\max(0, x)) = 1$ when $x > 0$.

Case 2: $x \leq 0$. When $x \leq 0$, $\max(0, x) = 0$ and therefore $\text{sign}(\max(0, x)) = 0$.

So we have that $\text{sign}(\max(0, x)) = 0$ for $x \leq 0$ and $\text{sign}(\max(0, x)) = 1$ for $x > 0$. But this is by definition the indicator function of $x > 0$, $\mathbb{1}_{x>0}$. Hence, $\text{sign}(\max(0, x)) = \mathbb{1}_{x>0}$ and we can conclude that $\lim_{n\to\infty} \tanh(n * \max(0, x)) = \mathbb{1}_{x>0}$. □

References

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*, pages 319–328, 2019.

Ron Freiwald. Lecture notes in analysis, 2014. URL `https://www.math.wustl.edu/~freiwald/310sequences2.pdf`.

Michael Lohaus, Michaël Perrot, and Ulrike von Luxburg. Too relaxed to be fair. In *International Conference on Machine Learning*, 2020.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.