Learning Fair Classifiers in Online Stochastic Settings

Yi Sun [1] Ivan Ramirez [1] Alfredo Cuesta-Infante [2] Kalyan Veeramachaneni [1]
[1]Massachusetts Institute of Technology [2]Universidad Rey Juan Carlos

Introduction and Motivation

What makes policy-driven machine learning different?

- Data might not be available upfront $\longrightarrow$ Online Algorithm
- Human decision makers in the loop $\longrightarrow$ Meta Algorithm
- Fair and Accurate

Online Binary Classification With Fairness

Given: A set of experts $f \in \mathcal{F}$, where $f : (\mathcal{X}, \mathcal{Z}) \longrightarrow \{0, 1\}$
At each round:

- An individual arrives with sensitive attributes z, and non-sensitive attributes x
- Sample an expert and use it's prediction
- Observe true label and update weights on experts

Goal:

1. Regret

$$\sum_{t=1}^{T} \ell(f^t(x^t, z^t), y^t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x^t, z^t), y^t)$$

2. Equalized Odds [2]

$$|\mathbb{E}[\hat{Y} = 1 | Y = 1, Z = A] - \mathbb{E}[\hat{Y} = 1 | Y = 1, Z = B]| \leq \epsilon$$

Methodology

Key Ideas

- Running separate instances of Multiplicative Weights algorithm for each group and label combination
- Randomize between instances help with fairness
- Obtain optimal selection probability between instances by optimizing regret and fairness bound

Algorithm

Initialize $w^1_{f,z,k} = 1 \ \forall f, z, k,$ and $q^1_{z,k} = \frac{1}{2} \ \forall z, k$
for $t \leftarrow 1, \ldots, T$ **do**
Each classifier obtains $\hat{y}^t_f$
Obtain the optimal q^*

$$\pi^t(f|z^t) = \begin{cases} \dfrac{w^t_{f,z^t,+}}{\sum_{f \in \mathcal{F}} w^t_{f,z^t,+}} & \text{with probability } q^*_{z^t,+} \\ \dfrac{w^t_{f,z^t,-}}{\sum_{f \in \mathcal{F}} w^t_{f,z^t,-}} & \text{with probability } q^*_{z^t,-} \end{cases}$$

Select classifier f according to probability π^t and update the regret
Obtain loss $\ell^t_f = \ell(\hat{y}^t_f, y^t)$ for each classifier f
Update weights $w^{t+1}_{f,z,k} = w^t_{f,z,k}(1-\eta)^{\ell^t_f 1\{z^t = z\} 1\{y^t = k\}} \quad \forall f, z, k$
end

Algorithm 1: Fairness-Aware MW algorithm

Figure 1: Fairness-aware RMW algorithm

Theorem 0.1: Upper Bound On Regret

Let $\alpha^t_{z,-} = \sum_{f \in \mathcal{F}} \frac{w^t_{f,z,-}}{\sum_f w^t_{f,z,-}} \cdot \ell^t_{f,z,+} - \sum_{f \in \mathcal{F}} \frac{w^t_{f,z,+}}{\sum_f w^t_{f,z,+}} \cdot \ell^t_{f,z,+}$ and $\alpha^t_{z,-}$ defined similarly. Thus the expected total loss of the algorithm is:

$$\mathbb{E}[L] \leq (1+\eta) L_f + 4\frac{\ln d}{\eta} + \alpha \tag{1}$$

where $\alpha = \sum_{z \in \{A,B\}, y \in \{+,-\}} q_{z,y} \sum_t \alpha^t_{z,y}$

Theorem 0.2: Fairness Bound

In the stochastic setting, there exists $q_{A,-}$ and $q_{B,-}$ such that the absolute difference in FPR can be bounded as:

$$\mathbb{E}_{x,y,z}\left[\frac{\mathbb{E}[L_{A,-}]}{C_{A,-}} - \frac{\mathbb{E}[L_{B,-}]}{C_{B,-}}\right] \leq (1+\eta-\gamma(\eta))\,\mathbb{E}_{x,y,z}\left[\frac{L_{f^*(B,-),B,-}}{C_{B,-}}\right] + \epsilon(1+\eta) + \left(\frac{q_{A,-} \cdot \sum_t \alpha^t_{A,-}}{p \cdot (1-\mu_{A,+}) \cdot T} - \frac{q_{B,-} \cdot \sum_t \alpha^t_{B,-}}{(1-p) \cdot (1-\mu_{B,+}) \cdot T}\right) \tag{2}$$

Optimal Balance Between Regret and Fairness

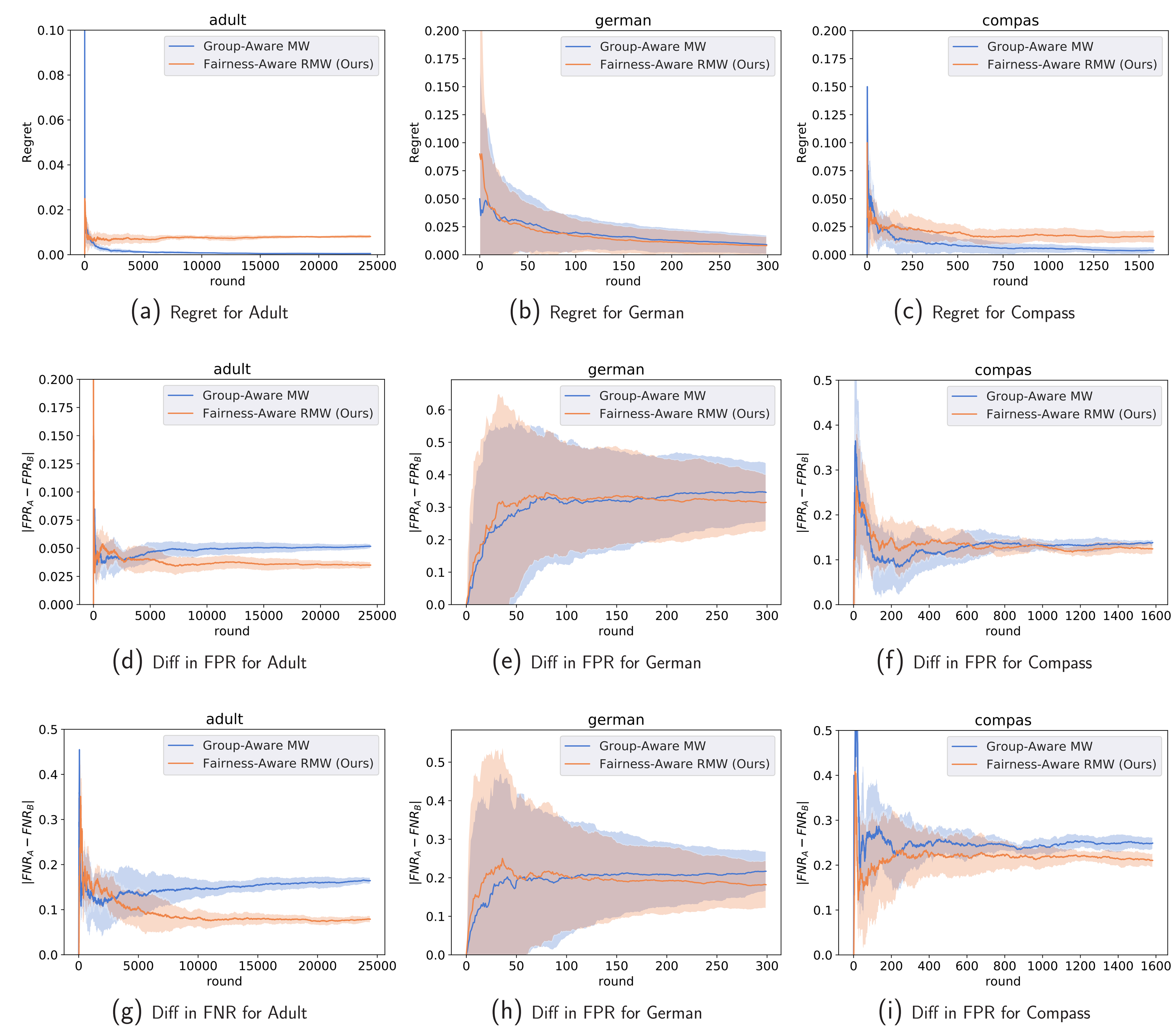
At each round, we solve the following optimization problem to minimize the fairness and regret upper bound:

$$\mathbf{q}^* = \arg\min_{\mathbf{q}} ||\boldsymbol{\lambda}(\mathbf{Aq} - \mathbf{b})||^2 \tag{3}$$

where $\boldsymbol{\lambda}$ is a vector of balancing the importance of equalized FPR, equalized FNR and regret that can be provided on a case-by-case basis based on different potential applications.

Experiments

The set of classifiers $\mathcal{F}$ in our hypothesis sets are as follows: Logistic Regression (LR), Linear SVM (L SVM), RBF SVM, Decision Tree (DT), Multi-Layer Perceptron (MLP). We pre-trained each classifier for each trial by splitting the data set, with 70% for training and 30% for testing. During the simulations, the examples in the testing set arrived one by one. We compare with [1], which achieves equalized error rates by running seperate instance of MW algorithm for each sensitive group.



(a) Regret for Adult (b) Regret for German (c) Regret for Compas

(d) Diff in FPR for Adult (e) Diff in FPR for German (f) Diff in FPR for Compas

(g) Diff in FNR for Adult (h) Diff in FPR for German (i) Diff in FPR for Compas

Conclusion

- Improvement in fairness both in terms of equalized FPR and FNR, along with a small increase in regret
- Randomization help overcome biases of experts

References

[1] A. Blum, S. Gunasekar, T. Lykouris, and N. Srebro.
On preserving non-discrimination when combining expert advice.
In *NeurIPS*, 2018.

[2] P. E. S. N. e. a. Hardt, Moritz.
Equality of opportunity in supervised learning.
In *In Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.