Learning Fair Classifiers in Online Stochastic Settings

Yi Sun [*†] MIT

Ivan Ramirez [*‡] MIT

Alfredo Cuesta-Infante [‡] Universidad Rey Juan Carlos

Kalyan Veeramachaneni [†] MIT

Abstract

In this paper, we try to accomplish approximate group fairness in an online decision-making process where examples are sampled *i.i.d* from an underlying distribution, which could be useful for implementing new public policy. Our work follows from the classical learning-from-experts scheme, extending the Randomized Multiplicative Weights algorithm by keeping separate weights for label classes as well as groups, where the probability of choosing each weights is optimized for both fairness and regret.

1 Introduction

Recently, there have been growing concerns about potential bias and discrimination in machine learning models. In many situations, machine learning may accentuate preexisting human biases, affecting policy-driven decision making in various areas including policing, college admissions, and loan approvals. Ideally, ensuring fairness *via* a mathematical framework would not only prevent prejudice within algorithms and models, but also help quantitatively overcome human biases. This possibility has motivated researchers in the machine learning community to develop numerous methods for making models *fair*. There have been many existing work on achieving fairness on a pre-existing dataset. Zafar *et al.* (2017) incorporates equalized odds as a constraint while solving optimization problems, while Hardt (2016) removes discrimination at post-processing steps.

One thing that differentiates policy-driven machine learning is that new public policies are often implemented in a trial-and-error fashion, as data might not be available upfront. Thus it is important to have a system that makes accurate and fair real-time decisions. Moreover, designing new public policy usually involves bringing together different parts with diverse and even conflicting goals. It is also generally accepted that there is often a trade-off between predictive accuracy and fairness Corbett-Davies *et al.* (2017).

At the group level, fairness can be defined as balancing some statistical metrics approximately across different demographic groups (such as gender groups, racial groups, etc.). Equalized odds Zafar *et al.* (2017), or "disparate mistreatment," requires that no error type is disproportionate for any one or more groups. This could be achieved by equalizing false positive rates, commonly referred as equal opportunity Hardt (2016), or equalizing classification errors. A predictor exhibits equalized odds if it achieves both an equalized false-positive rate (FPR) and an equalized false-negative rate (FNR).

In this paper, we consider a setting where individuals arrive in a sequential and stochastic manner from an underlying distribution, and the goal is to make real-time decisions *fair* by combining decisions from experts. This setting could be useful for implementing new policies in many areas, including improving the fairness of clinical trial participants recruitment or online loan applications. The algorithm we propose is a meta algorithm and can be used as a central coordinator to fairly combine different parts of the system. We demonstrate the performance of the algorithm on real data sets commonly used by the fairness community.

AI for Social Good workshop at NeurIPS (2019), Vancouver, Canada.

2 Preliminaries and Model

2.1 Online Classification with Sensitive Attribute

We consider a binary classification problem where $Y = \{0, 1\}$, and we assume that there is a finite set of classifiers $\mathcal{F}$ to choose from, where $\mathcal{F} = \{f_1, ..., f_d\}$. As is typical in an online learning setting, the algorithms run through rounds $t = 1, .., T$. At each round t, the classifier receives a joint example vector $(x^t, z^t) \in \mathbb{R}^n$, where x^t are the unprotected attributes and z^t are the protected or sensitive attributes. We consider the case where there are two sensitive groups $Z = \{A, B\}$, with a generic element denoted as z. We denote the base rates for outcomes as $\mu_{z,+} = \mathbb{P}(Y = 1|Z = z)$.

An important metric of online learning is regret, which compares the performance of the algorithm with the best expert in hindsight. At each time step t, the classifier first produces $\hat{y}^t = f^t(x^t, z^t)$, a predicted label for the input example. Then, at the end of the round and always after having produced its prediction, it observes the true label y^t and suffers a loss $\ell(\hat{y}^t, y^t)$. After T rounds, regret is formally expressed as $Regret(T) = \sum_{t=1}^{T} \ell(f^t(x^t, z^t), y^t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x^t, z^t), y^t)$ where the first term is the cumulative loss of the algorithm, and the second term is the cumulative loss of the best fixed classifier in hindsight. The typical goal of online learning is to design an algorithm that achieves sub-linear regret compared with the best hindsight over the T rounds; i.e. $\lim_{T \to \infty} \frac{Regret(T)}{T} = 0$. In this paper, we add in a fairness constraint, which requires the online learning algorithm to satisfy an approximate ϵ-fairness on average.

Definition 2.1 (ϵ-fairness) *A randomized algorithm satisfies ϵ-fairness if:*

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

2.2 Randomized Multiplicative Weights Algorithm

The *Multiplicative Weights* (MW) Arora *et al.* (2012) method is a frequently used meta-algorithm for achieving no-regret by following the experts. In the MW algorithm, a decision maker has a choice of d experts. We denote the probability that classifier i being selected at round t as π_i^t. Initially, each classifier i has an even chance of being selected. After each round of decisions, the decision maker maintains weights on the experts based on their performances so far. Higher weights indicate a higher chance of being selected in the next round.

Arora *et al.* (2012) bounded the total expected loss of the MW algorithm by the total loss of the best experts with the following theorem:

Theorem 2.1 *Assume that the loss ℓ_i^t for classifier i at round t is bounded in [0,1] and $\eta \leq \frac{1}{2}$. Then after T rounds, for any classifier i among the d classifiers we have:*

$$\sum_{t=1}^{T} \ell^t \pi^t \leq (1 + \eta) \sum_{t=1}^{T} \ell_i^t + \frac{\ln d}{\eta}$$

This powerful theorem shows that the expected cumulative loss achieved by the MW algorithm is upper bounded by the cumulative loss of the best fixed expert in hindsight asymptotically. In other words, the MW algorithm achieves sub-linear regret.

2.3 Randomized Multiplicative Weights Algorithm With Fairness

Blum *et al.* (2018) proposed a group-aware version of the MW algorithm to achieve equalized error rates in an adversarial setting. Their idea was to run separate instances of the original MW algorithm on each group, and they demonstrated that this is necessary to achieve equalized error rates across groups.One potential drawback of the group-aware algorithm is that it only bounds the performance of the overall algorithm errors for each group, without a guarantee of how the errors will distribute across the label classes. In order to satisfy equalized odds, we also need a bound on the number of false positives and false negatives made by the algorithm on each group.

Our proposal extends Blum's Group-aware MW algorithm to *Fairness-Aware* RMW by keeping a table of weights for each possible 3-tuple (f, z, y) with $f \in \mathcal{F}$, $z \in \{A, B\}$ and $y \in \{+, -\}$. At each

round, we maintain a probability distribution of selecting different copies of weights. We indicate the probability of selecting $w_{f,z,+}$ as $q_{z,+}$, and the probability of selecting $w_{f,z,-}$ as $q_{z,-}$. As shown in the fairness bound later, $q_{z,+}$ and $q_{z,-}$ can be explicitly set to balance regret and fairness loss of the algorithm.

Thus we propose the following variant of the MW algorithm:

Initialize $w^1_{f,z,k} = 1 \ \ \forall f, z, k, \ \ \text{and} \ \ q^1_{z,k} = \frac{1}{2} \ \ \forall z, k$
for $t \leftarrow 1, ..., T$ **do**
  Each classifier obtains $\hat{y}^t_f$
  Obtain the optimal q^* according to equation 4

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

  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

Lemma 2.2 (Upper Bound) *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:*

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

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

In the algorithm 1, since the learner only knows the group but not the label when it makes a decision, it selects by sampling from the estimated label estimation. Therefore, at each round, the losses the learner obtains can be decomposed to the losses of original MW algorithm and the losses from choosing the wrong copy of weights when selecting classifiers due to randomization (as the extra α term). We argue that randomization is the key to improving fairness on a biased distribution. We omit the proof due to space constraints.

Theorem 2.3 (Fairness Bound) *In the stochastic setting, there exists* $q_{A,-}$ *and* $q_{B,-}$ *such that the absolute difference in FPR can be bounded as:*

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

For the fairness bound, since $\alpha^t_{z,y}$ and $\mu_{z,+}$ can be tracked and estimated, the last two terms involving $q_{A,-}$ and $q_{B,-}$ can be cancelled out by specifically setting $q_{A,-}$ and $q_{B,-}$ at each round. we have:

$$\begin{bmatrix} -\sum_t \alpha^t_{A,+} & \sum_t \alpha^t_{B,+} \\ \hline p\mu_{A,+}T & (1-p)\mu_{B,+}T \end{bmatrix} \begin{bmatrix} q_{A,+} \\ q_{B,+} \end{bmatrix} = 0. \tag{3}$$

We can obtain something similar for the upper bound for regret.

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

$$\mathbf{q}^* = \arg\min_{\mathbf{q}} ||\boldsymbol{\lambda}(\mathbf{A}\mathbf{q} - \mathbf{b})||^2 \tag{4}$$

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.

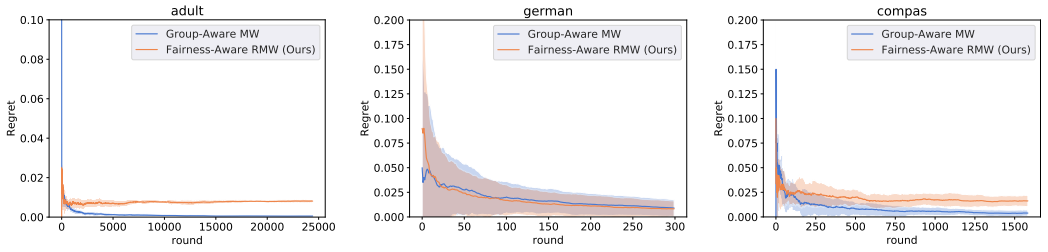
3 Experiments

We test our algorithms on the `Adult`, `German Credit` and datasets, all of which are commonly used by the fairness community. "Adult" consists of individuals' annual income measurements based on different factors. In the "German Credit" dataset, people applying for credit from a bank have been classified as "good" or "bad" credit risks based on their attributes. "COMPAS" (Correctional Offender Management Profiling for Alternative Sanctions) provides a likelihood of recidivism based on a criminal defendant's history and other aspects.

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. In practice, crafting policy might involve balancing the effects of multiple parts of systems with diverse goals. Thus, in the experiment, we did not explicitly require that each individual classifier satisfies ϵ-fairness, and each classifier could be biased.

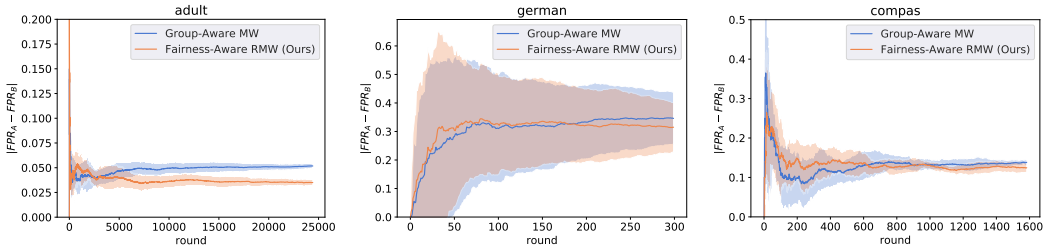
For each data set, the first plot shows the discrepancy of averaged regret between the algorithm and the best classifiers in hindsight, the second plot shows the absolute differences of FPR between the two groups. Results depicted in figures show a general improvement in fairness over Blum's, both in terms of equalized FPR and FNR, along with a small increase in regret. In all datasets, difference in FPR and FNR are lower than the group-aware MW algorithm. This justifies the use of different instances of the MW algorithm for each subset of group and label combination. Though the standard deviation (the shadowed part) is larger for smaller data set(german), there is still a trend towards convergence. As a result of our proposed method, the probability distribution $\mathbf{q}$ for sampling each instance of the weights (at label level) can be adjusted to obtain the optimal probabilities that satisfy the required constraints, which can be provided a on case to case basis.

Future research could take on the more realistic case in which feedback from the implemented policy is delayed for some number of rounds. For example, during the college admissions process, the performance of a student is generally evaluated at the end of each term, while colleges typically offer admission decisions in mid-year. Similarly, when an individual applies for a loan, the bank often needs to wait for some time to know whether the applicant will default or not.



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

Figure 1: Comparison of Average Regret



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

Figure 2: Comparison of Absolute Difference of FPR

References

Arora, Sanjeev, Hazan, Elad, & Kale, Satyen. 2012. The Multiplicative Weights Update Method: a Meta-Algorithm and Applications. *Theory of Computing*, **8**, 121–164.

Blum, Avrim, Gunasekar, Suriya, Lykouris, Thodoris, & Srebro, Nathan. 2018. On preserving non-discrimination when combining expert advice. *In: NeurIPS*.

Corbett-Davies, Sam, Pierson, Emma, Feller, Avi, Goel, Sharad, & Huq, Aziz. 2017. Algorithmic Decision Making and the Cost of Fairness. *In: KDD*.

Hardt, Moritz, Price Eric Srebro Nati et al. 2016. Equality of opportunity in supervised learning. *In: In Advances in Neural Information Processing Systems*. pp. 3315–3323.

Zafar, Muhammad Bilal, Valera, Isabel, Gomez-Rodriguez, Manuel, & Gummadi, Krishna P. 2017. Fairness Beyond Disparate Treatment and Disparate Impact: Learning Classification without Disparate Mistreatment. *In: WWW*.