A Reductions Approach to Fair Classification

Alekh Agarwal
Microsoft Research
alekha@microsoft.com

Alina Beygelzimer
Yahoo! Research
beygel@gmail.com

Miroslav Dudík
Microsoft Research
mdudik@microsoft.com

John Langford
Microsoft Research
jcl@microsoft.com

ABSTRACT

We present a systematic method for training-time enforcement of definitions of fairness in binary classification with protected attributes. Taking two popular definitions, *demographic parity* and *equalized odds*, we show how to reduce fairness-constrained classification to cost-sensitive classification. The reduction is agnostic to the representation and form of the cost-sensitive classifier, allowing a range of existing algorithms. Empirical evaluation shows that the systematic incorporation of fairness constraints allows us to outperform prior baselines for both definitions.

ACM Reference format:
Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, and John Langford. 2017. A Reductions Approach to Fair Classification. In *Proceedings of FAT ML, Halifax, Nova Scotia, Canada, 2017 (FATML'17),* 5 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Despite the relatively recent origin, there is already a thriving research program on incorporating fairness into machine learning algorithms, with the largest body of work focusing on binary classification. The typical setup involves a pre-determined *protected attribute* and seeks to achieve fairness across all values of this attribute, using one of many definitions of fairness [5, 7, 10]. To avoid the design of a new algorithm for each definition of fairness, it is desirable to have a general scheme for mapping fairness definitions into algorithms. The scheme should work for a large class of fairness definitions. Also, to allow modeling flexibility, it should be agnostic to classifier representations and algorithms, for instance, including neural networks trained with backprop, ridge regression solved in closed form, or SVM trained with stochastic gradient descent. Finally, we want the resulting algorithms to be effective, that is, they should obtain near-optimal classification performance subject to fairness constraints.

We describe an approach to achieve these desiderata through the framework of learning reductions [2]. Specifically, we assume access to a cost-sensitive classification algorithm, which does not incorporate any notion of fairness. Starting with the fairness notion of *demographic parity*, we show how to turn binary classification with demographic-parity constraints into cost-sensitive classification.[1] The key property of demographic parity that allows our transformation is that it can be fully described via (conditional) moments involving the protected attribute, the true label, and our predictions. This means that an analogous approach can be applied to the *equalized odds* notion of fairness [5]. The overall scheme provides a systematic route from a fairness definition to an algorithm, unlike several prior works which use heuristic arguments to justify intuitive mechanisms for operationalizing specific definitions.

[1] Cost-sensitive classification can be further reduced to binary classification [1] or regression [11].

FATML'17, 2017, Halifax, Nova Scotia, Canada
© 2017 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

We empirically evaluate our method and compare with prior approaches. We find that our method generally achieves better fairness-accuracy trade-offs with substantial improvements in several problems.

Related work. Unlike previous pre-processing and post-processing approaches [5, 8], our method directly optimizes classification error under fairness constraints. Prior works that incorporate fairness constraints into training, for instance, via regularization, rely on specific representations such as linear or generalized linear models [6, 9, 13, 14]. These methods often relax the definition of fairness, only enforcing weaker constraints, such as lack of correlation, and obtain fairness guarantees only under strong distributional assumptions.

2 SETTING

We consider a binary classification setting where the training examples consist of triples (X, A, Y), where X is a feature vector, $A \in \{1, \ldots, K\}$ is a protected attribute and $Y \in \{0, 1\}$ is the label. We aim to learn a classifier $h : X \mapsto \{0, 1\}$ from a set $\mathcal{H}$ of candidate classifiers.[2]

We assume access to a cost-sensitive classification algorithm for the set $\mathcal{H}$. The input to such an algorithm is a set of points $\{(X_i, C_i^0, C_i^1)\}$, where C_i^0 and C_i^1 denote the losses (called *costs* in this setting) for predicting the labels 0 or 1 on X_i, respectively. A *cost-sensitive classification algorithm* takes such a dataset and outputs

$$\hat{h} = \operatorname*{arg\,min}_{h \in \mathcal{H}} \sum_{i=1}^{n} h(X_i) C_i^1 + (1 - h(X_i)) C_i^0. \tag{1}$$

This abstraction allows specifying different costs on each example, which is essential in incorporating fairness constraints. At the same time, practical algorithms for cost-sensitive classification are readily available for several common representations [1, 4, 11].

We consider two definitions of fairness. The first one, called *demographic parity* [7], is based on prior legal literature, including the so-called 80/20 rule.

*Definition 2.1 (Demographic parity—*DP*).* A classifier h satisfies demographic parity under a distribution over (X, A, Y) if $\exists c > 0$ such that $\mathbb{E}[h(X) \mid A = a] = c$ for all a. That is, the prediction $h(X)$ and the attribute A are independent.

More recently, Hardt et al. [5] proposed an alternative called *equalized odds*, which can be a better definition if the label proportions for $Y = 0$ or 1 are drastically different across the different values of A.

*Definition 2.2 (Equalized odds—*EO*).* A classifier h satisfies equalized odds under a distribution over (X, A, Y) if $\exists c_y > 0$ such that $\mathbb{E}[h(X) \mid A = a, Y = y] = c_y$ for all $a \in \{1, 2, \ldots, K\}$ and $y \in \{0, 1\}$. That is, $h(X)$ and A are conditionally independent given Y.

3 REDUCTIONS APPROACH

We next show how the problem of binary classification under demographic parity can be turned into a cost-sensitive classification problem. The steps in our derivation are general and can easily be used to obtain a reduction for other notions of fairness based on (conditional) moment constraints, such as EO.

The accuracy of a classifier h is measured by 0-1 error, $\mathbf{1}\{h(X) \neq Y\}$, where $\mathbf{1}\{\cdot\}$ is the binary indicator, and fairness corresponds to DP (Defn. 2.1). If we knew the distribution over triples (X, A, Y) exactly, we would find the best fair classifier by solving

$$\begin{aligned} \min_{h \in \mathcal{H}} \quad & \mathbb{P}[h(X) \neq Y] \\ \text{s.t. } \forall a : \quad & \underbrace{\mathbb{E}[h(X) \mid A{=}a] - \frac{1}{K} \sum_{a'=1}^{K} \mathbb{E}[h(X) \mid A{=}a']}_{=: \gamma_a(h)} = 0. \end{aligned} \tag{2}$$

In words, we seek to minimize the misclassification error under the constraint that the probability of predicting $Y = 1$ conditioned on $A = a$ is the same for each value a of the protected attribute, and equals the average probability of predicting $Y = 1$ over all a. While this criterion appears more specific than our definition of DP, it is easily seen to be equivalent.

With finite samples, we replace the expectations with sample averages and define $\widehat{\gamma}_a(h)$ to be the empirical constraint violation. We relax the fairness constraint to $|\widehat{\gamma}_a(h)| \leq \varepsilon$ to account for sampling errors in estimating the fairness violation. In summary, our goal is to solve the following optimization problem

$$\min_{h \in \mathcal{H}} \widehat{\mathbb{P}}[h(X) \neq Y] \qquad \text{s.t.} \quad \forall a : |\widehat{\gamma}_a(h)| \leq \varepsilon. \tag{3}$$

To solve this problem, we introduce Lagrange multipliers $\lambda_a^+ \geq 0$ and $\lambda_a^- \geq 0$ corresponding to upper and lower bounds on the expression inside the absolute

[2] The classifier does not explicitly depend on the protected attribute A, but X can contain A as one of the features, or have other features arbitrarily indicative of A.

value, and form the Lagrangian $L(h, \boldsymbol{\lambda})$ as

$$\widehat{\mathbb{P}}[h(X) \neq Y] + \sum_a \Big[\lambda_a^+\big(\widehat{\gamma}_a(h) - \varepsilon\big) + \lambda_a^-\big(-\widehat{\gamma}_a(h) - \varepsilon\big)\Big],$$

where we denote the vector of all Lagrange multipliers as $\boldsymbol{\lambda}$. Under standard non-degeneracy assumptions, such as the set $\mathcal{H}$ being closed under convex combinations [12], the optimization problem (3) is equivalent to the saddle-point problem

$$\max_{\lambda_a^+ \geq 0, \lambda_a^- \geq 0} \; \min_{h \in \mathcal{H}} L(h, \boldsymbol{\lambda}) \quad . \tag{4}$$

Our scheme proceeds by iteratively optimizing over Lagrange multipliers $\boldsymbol{\lambda}$ in the outer maximization loop (for instance by grid search or subgradient descent), while invoking a cost-sensitive learner for the class $\mathcal{H}$ in the inner loop. We know that at the solution, at most one of λ_a^+ and λ_a^- is non-zero, so we assume that the outer loop iterates only over λ of that form. We next argue that for any such vector of Lagrange multipliers, the inner minimization can be written as an instance of cost-sensitive learning.

Reduction to cost-sensitive learning. We first express all of the dependence on h via unconditional expectations. We replace $\widehat{\mathbb{P}}[h(X) \neq Y]$ by $\widehat{\mathbb{E}}[\mathbf{1}\{h(X) \neq Y\}]$, and $\widehat{\mathbb{E}}[h(X)|A{=}a]$, appearing in $\widehat{\gamma}_a(h)$, by $\widehat{\mathbb{E}}[h(X)\mathbf{1}\{A{=}a\}]/p_a$ where $p_a = \widehat{\mathbb{P}}[A = a]$. We also introduce the shorthands

$$\lambda_a = \lambda_a^+ - \lambda_a^- \quad , \qquad \mu = \frac{1}{K}\sum_a \lambda_a \quad ,$$

and further simplify the expression for $L(h, \boldsymbol{\lambda})$, using the fact that $\lambda_a^+ + \lambda_a^- = |\lambda_a|$, because at most one of λ_a^+ and λ_a^- is non-zero,

$$\begin{aligned} L(h, \boldsymbol{\lambda}) &= \widehat{\mathbb{E}}\Big[\mathbf{1}\{h(X) \neq Y\}\Big] - \varepsilon \sum_a |\lambda_a| \\ &\quad + \widehat{\mathbb{E}}\left[\sum_a \lambda_a \frac{h(X)\mathbf{1}\{A{=}a\}}{p_a} - \sum_{a'} \mu \frac{h(X)\mathbf{1}\{A{=}a'\}}{p_{a'}}\right] \\ &= -\varepsilon \sum_a |\lambda_a| + \widehat{\mathbb{E}}\left[\mathbf{1}\{h(X) \neq Y\} + (\lambda_A - \mu)\frac{h(X)}{p_A}\right], \end{aligned} \tag{5}$$

where in the last step we used the fact that the indicators $\mathbf{1}\{A{=}a\}$ and $\mathbf{1}\{A{=}a'\}$ select the terms of the sum with $a = A$ and $a' = A$, respectively.

Assume that we are given a data set $\{(X_i, A_i, Y_i)\}$. Eq. (5) means that the minimization of $L(h, \boldsymbol{\lambda})$ over h is the same as the minimization of the objective:

$$\sum_i \mathbf{1}\{h(X_i) \neq Y_i\} + (\lambda_{A_i} - \mu)\frac{h(X_i)}{p_{A_i}} \quad ,$$

which corresponds to cost-sensitive classification on the data $\{(X_i, C_i^0, C_i^1)\}$ with costs

$$C_i^0 = \mathbf{1}\{Y_i \neq 0\}, \quad C_i^1 = \mathbf{1}\{Y_i \neq 1\} + \frac{\lambda_{A_i} - \mu}{p_{A_i}}.$$

Thus, we can use a cost-sensitive classifier (1) to minimize over h for a fixed $\boldsymbol{\lambda}$, and find $\boldsymbol{\lambda}$ using grid search if K is small or subgradient descent if K is large.

Equalized odds. If the fairness is represented as equalized odds, the optimization problem takes form

$$\begin{aligned} &\min_{h \in \mathcal{H}} \; \mathbb{P}[h(X) \neq Y] \\ &\text{s.t. } \forall a, y : \\ &\left| \widehat{\mathbb{E}}[h(X) \mid A{=}a, Y{=}y] - \sum_{a'=1}^{K} \frac{\widehat{\mathbb{E}}[h(X) \mid A{=}a', Y{=}y]}{K} \right| \leq \varepsilon. \end{aligned}$$

In this case we end up with more Lagrange multipliers, specifically, we have $\lambda_{ay}^+ \geq 0$ and $\lambda_{ay}^- \geq 0$ across $a \in \{1, \ldots, K\}$ and $y \in \{0, 1\}$. These, again, give rise to $\lambda_{ay} = \lambda_{ay}^+ - \lambda_{ay}^-$, and $\mu_y = \frac{1}{K}\sum_a \lambda_{ay}$. The costs derived from the example (X_i, A_i, Y_i) are then

$$C_i^0 = \mathbf{1}\{Y_i \neq 0\}, \quad C_i^1 = \mathbf{1}\{Y_i \neq 1\} + \frac{\lambda_{A_i Y_i} - \mu_{Y_i}}{p_{A_i Y_i}}, \tag{6}$$

where $p_{ay} = \widehat{\mathbb{P}}[A = a, Y = y]$.

4 EXPERIMENTAL RESULTS

We examine how different methods perform for two notions of fairness, DP and EO. In both cases, we evaluate the best accuracy without fairness constraints. For DP, we additionally evaluate the *reweighting* and *relabeling* techniques of Kamiran and Calders [8]. For EO, we include the approach of Hardt et al. [5]. Our method and the relabeling approach have a tuning parameter giving a trade-off between fairness and accuracy, so we can compute a full Pareto frontier. For other methods, we get a single solution with a specific trade-off. For fairness violation, we report the absolute value of disparity on the test set, according to DP or EO. The experiments were performed on the following datasets:

Adult data (UCI) is extracted from the 1994 census bureau database. The task is to predict whether a person makes more than $50k per year, with gender as the protected attribute.

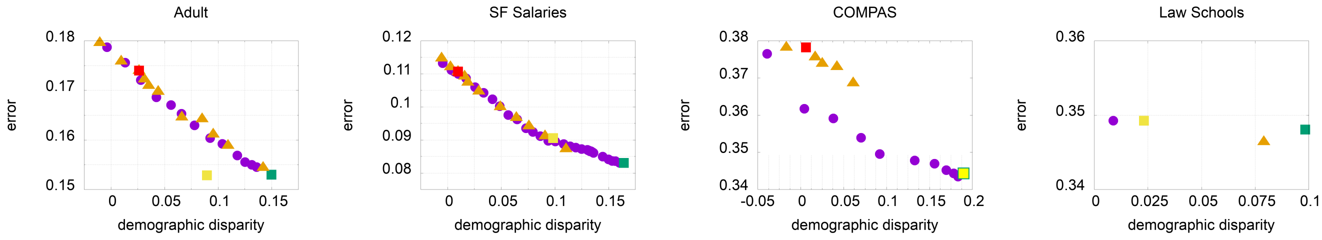

Figure 1: Demographic parity vs error rate. *Legend:* ●—DP reduction, 🟥—relabeling [8], ▲—relabeling with varying relabeling amounts, 🟨—reweighting [8], 🟩—unconstrained.

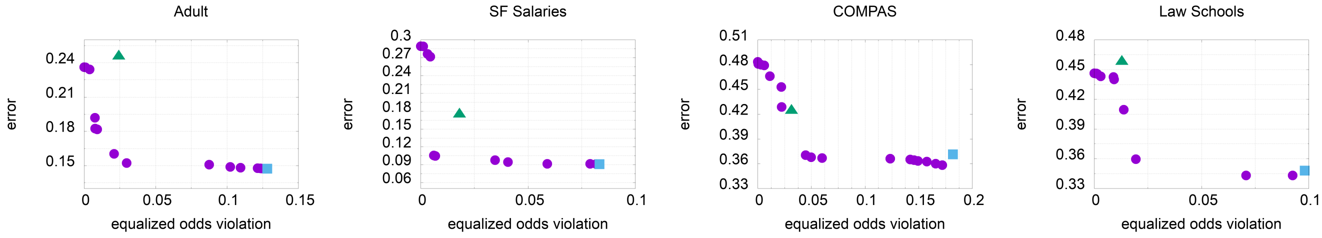

Figure 2: Equalized odds violation vs error rate. *Legend:* ●—EO reduction, ▲—Hardt et al. [5], 🟦—unconstrained.

San Francisco salary data (Kaggle) contains the names, job titles, and compensation for San Francisco city employees on an annual basis from 2011 to 2014. The goal is to predict if an employee makes at least $100k per year, with gender as the protected attribute.

COMPAS Recidivism Risk Score Data (Florida Department of Corrections) is based on ProPublica's study.[3] The goal is to predict recidivism based on criminal history, jail and prison time, demographics and COMPAS risk scores for defendants from Broward County, with race as the protected attribute.

In data from *Law School Admissions Council's National Longitudinal Bar Passage Study* the task is to predict whether applicants end up going to the law school of their choice, with gender as the protected attribute.

We used the Vowpal Wabbit (VW) [4] open source machine learning system as our base learning algorithm. The cost-sensitive classification algorithm used is a reduction to regression described in Algorithms 4 and 5 of [3], with a single pass of online learning used to minimize the resulting regression objective approximately. The Pareto frontier was computed using grid search.

Overall, we find that our method is always competitive, and often substantially outperform baselines for both fairness measures. For DP, the reduction approach uniformly achieves the lowest demographic disparities. With one exception (reweighting on the Adult dataset) the reduction approach performs similar to or better than the Pareto frontier of all other approaches for DP.[5] In general, the fixed threshold in reweighting baseline does not offer the flexibility of picking a performance-fairness tradeoff of our approach. Relabeling allows this, but appears to perform poorly on the COMPAS dataset, possibly due to the noisiness of the underlying statistical problem. Note that the relabeling heuristic is not guaranteed to find the most accurate and fair classifier in general, so this is not surprising.

For EO, results are uniformly positive, partially reflecting the weakness of post-processing based techniques like Hardt et al. [5]. The classification performance generally degrades more under the EO than the DP constraint, reflecting that this is a more stringent notion. In general, the results demonstrate the utility of having an easy recipe to tailor algorithms to fairness definitions.

[3] https://github.com/propublica/compas-analysis
[4] https://github.com/JohnLangford/vowpal_wabbit

[5] We did not get full Pareto frontiers for *Law Schools* data as the accuracy varied in a very non-smooth manner, causing one point to dominate the entire curve for both our reduction and reweighting.

REFERENCES

[1] Alina Beygelzimer, Varsha Dani, Thomas P. Hayes, John Langford, and Bianca Zadrozny. 2005. Error limiting reductions between classification tasks. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*. 49–56. https://doi.org/10.1145/1102351.1102358

[2] Alina Beygelzimer, Hal Daumé, John Langford, and Paul Mineiro. 2016. Learning reductions that really work. *Proc. IEEE* 104, 1 (2016), 136–147.

[3] Alina Beygelzimer and John Langford. 2009. The Offset Tree for Learning with Partial Labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*.

[4] Wei Fan, Salvatore J. Stolfo, Junxin Zhang, and Philip K. Chan. 1999. AdaCost: Misclassification Cost-Sensitive Boosting. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, June 27 - 30, 1999*. 97–105.

[5] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Neural Information Processing Systems (NIPS)*.

[6] Kory D Johnson, Dean P Foster, and Robert A Stine. 2016. Impartial Predictive Modeling: Ensuring Fairness in Arbitrary Models. *arXiv preprint arXiv:1608.00528* (2016).

[7] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*. IEEE, 1–6.

[8] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[9] T. Kamishima, S. Akaho, and J. Sakuma. 2011. Fairness-aware Learning through Regularization Approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. 643–650.

[10] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[11] John Langford and Alina Beygelzimer. 2005. Sensitive Error Correcting Output Codes. In *Learning Theory, 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*. 158–172. https://doi.org/10.1007/11503415_11

[12] Maurice Sion. 1958. On general minimax theorems. *Pacific J. Math.* 8, 1 (1958), 171–176. http://projecteuclid.org/euclid.pjm/1103040253

[13] Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. (2017).

[14] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2015. Fairness constraints: A mechanism for fair classification. *stat* 1050 (2015), 19.