A Convex Framework for Fair Regression*

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph
Michael Kearns, Jamie Morgenstern, Seth Neel, Aaron Roth
University of Pennsylvania

Abstract

We introduce a flexible family of fairness regularizers for (linear and logistic) regression problems. These regularizers all enjoy convexity, permitting fast optimization, and span the range from group fairness to strong individual fairness. We study the accuracy-fairness trade-off on any given dataset, and we measure the severity of this trade-off via a numerical quantity we call the Price of Fairness (PoF). The centerpiece of our results is an extensive comparative study of the PoF across six different datasets in which fairness is a primary consideration.

1 Introduction

The widespread use of machine learning to make consequential decisions about individual citizens (including in domains such as credit, employment, education and criminal sentencing [3, 4, 25, 27]) has been accompanied by increased reports of instances in which the algorithms and models employed can be unfair or discriminatory in a variety of ways [2, 28]. As a result, research on fairness in machine learning and statistics has seen rapid growth in recent years [1, 5–7, 9–11, 13, 14, 17–20, 24, 26], and several mathematical formulations have been proposed as metrics of (un)fairness for a number of different learning frameworks. While much of the attention to date has focused on (binary) classification, where standard fairness notions include equal false positive or negative rates across different populations, less attention has been paid to fairness in (linear and logistic) regression, where the target and/or predicted values are continuous, and the same value may not occur even twice in the training.

In this work, we introduce a rich family of fairness metrics for regression models that take the form of a fairness regularizer and apply them to the standard loss functions for linear and logistic regression. Since these loss functions and our fairness regularizer are convex, the combined objective functions obtained from our framework are also convex, and thus permit efficient optimization. Furthermore, our family of fairness metrics covers the spectrum from the type of *group* fairness that is common in classification formulations (where e.g. false arrests in one racial group can be "compensated" for by false arrests in another racial group) to much stronger notions of *individual* fairness (where such cancellations are forbidden, and every injustice is charged to the model). Our framework also permits one to either forbid the use of a "protected" variable (such as race), by demanding that a single model be learned across all groups, or to build different group-dependent models.

Most importantly, by varying the weight on the fairness regularizer, our framework permits us to study the trade-off between predictive accuracy and fairness. This is important to examine and understand in a domain-specific manner as demanding fairness of models always come at a cost of reduced accuracy [8, 11, 15, 32], it behooves practitioners working with fairness-sensitive data sets to understand just how mild or severe this trade-off is in their particular arena, permitting them to make informed modeling and policy decisions.

Our central results take the form of an extensive comparative empirical case study across six distinct datasets in which fairness is a primary concern. We introduce an intuitive quantity called the *Price of Fairness (PoF)*, which numerically quantifies the extent to which increased fairness degrades accuracy. We compare the PoF across all of our 6 datasets, fairness notions, and treatments of protected variables.

Our primary contributions are: (1) The introduction of a flexible but convex family of fairness regularizers of varying strength that spans the spectrum from group to individual fairness. (2) The introduction of a quantitative, data-dependent measure of the severity of the accuracy-fairness tradeoff. (3) An extensive empirical comparative study across six fairness-sensitive data sets.

While our empirical study does reveal some reasonably consistent findings across datasets, perhaps the most important message is a cautionary one: the detailed trade-off between accuracy and fairness, and the comparison of different fairness notions, appears to be quite domain-dependent and lacking prescriptive "universals". This is perhaps consistent with the emerging theoretical literature demonstrating the lack of a single

*The full technical version of this paper is available at https://arxiv.org/abs/1706.02409.

"right" definition of fairness [7, 12, 21].

2 The Regression Setting

Consider the standard (linear and logit) regression setting: denote the *explanatory* variables (or *instances*) by $\mathbf{x} \in \mathcal{X} = \mathbb{R}^d$ and the *target* variables (or *labels*) by $y \in \mathcal{Y} = [-1, 1]$. Note that for both linear and logit models, the target values are continuous. Let $\mathcal{P}$ denote the joint distribution over $\mathcal{X} \times \mathcal{Y}$. Suppose every instance $\mathbf{x}$ belongs to exactly one of 2 groups, denoted by 1 and 2.[1] This partition of $\mathcal{X}$ into groups (e.g. into different races or genders) is encoded in a "sensitive" feature $\mathcal{X}_{d+1}$. Let $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n\}$ be a training set of n samples drawn i.i.d. from $\mathcal{P}$, separated by groups into S_1 and S_2. Let $n_1 = |S_1|$ and $n_2 = |S_2|$. ($n = n_1 + n_2$.)

We study the trade-off between fairness and accuracy for the class of linear and logit regression models. Given a pair of explanatory and target variables $(\mathbf{x}, y)$, we treat y as the ground truth description of $\mathbf{x}$'s merit for the regression task at hand: two pairs $(\mathbf{x}, y), (\mathbf{x}', y')$ with $y \approx y'$ have similar observed outcomes. We aim to design models which treat two such instances with similar observed outcomes similarly, a notion we refer to as *fairness* with respect to the ground truth. For a given accuracy loss ℓ and fairness loss (or *penalty*) f, we define the λ-weighted fairness loss of a regressor $\mathbf{w}$ on $\mathcal{P}$ to be $\ell_{\mathcal{P}}(\mathbf{w}) + \lambda f_{\mathcal{P}}(\mathbf{w})$. For a sample S, we analogously define the λ-weighted fairness training loss of $\mathbf{w}$ as $\ell(\mathbf{w}, S) + \lambda f(\mathbf{w}, S)$. For linear regression, we let ℓ be mean-squared error; for logistic regression, we let ℓ be the log loss. Finally, we use ℓ_2 regularization, so the overall loss is $\ell_{\mathcal{P}}(\mathbf{w}) + \lambda f_{\mathcal{P}}(\mathbf{w}) + \gamma||\mathbf{w}||_2$.

2.1 A Convex Family of Fairness Penalties

Our definitions of fairness all measure how similarly a model treats two similarly labeled instances, one from each group. In particular, all of our definitions have a term for each "cross-group" *pair* of instances/labels, weighted as a function of $|y_i - y_j|$ and $|\mathbf{w} \cdot \mathbf{x}_i - \mathbf{w} \cdot \mathbf{x}_j|$. For shorthand, we refer to pairs of instances/labels (one from each group) as *cross pairs*, and cross pairs with similar labels as *similar cross pairs*. Our definitions differ in precisely which cross pair disparities can counteract one another. In one extreme (individual fairness), every cross pair disparity increases the fairness penalty of a model. In the other (group fairness), the disparities in different similar cross pairs can counterbalance each other. Hence, our fairness notions for regression align closely to individual and group fairness for classification, both common threads in the fairness literature.

[1] The generalization to more than 2 groups is straightforward.

Remark 1. *We assume the sensitive feature $\mathcal{X}_{d+1}$ is available to the learning procedure in one of two ways. In the "single model" setting, the algorithm should build a single linear model $\mathbf{w}$ for all of $\mathcal{X}$ (over all but the sensitive features), but can measure the empirical fairness loss of $\mathbf{w}$ using the sensitive feature. In the "separate models" setting, the algorithm can build two distinct linear models $\mathbf{w}_1, \mathbf{w}_2$ for the two groups, thus directly observing the sensitive feature when building these models.*

We define our fairness penalties for single model but the extension to separate models is straightforward.

Individual Fairness The first fairness penalty we propose is the following:

$$f_1(\mathbf{w}, S) = \frac{1}{n_1 n_2} \sum_{\substack{(\mathbf{x}_i, y_i) \in S_1 \\ (\mathbf{x}_j, y_j) \in S_2}} d(y_i, y_j)\big(\mathbf{w} \cdot \mathbf{x}_i - \mathbf{w} \cdot \mathbf{x}_j\big)^2,$$

for some fixed non-negative function d (assumed to be decreasing in $|y_i - y_j|$). Since $d(y_i, y_j)$ does not depend upon the decision variables ($\mathbf{w}$), we treat these values as constants in an optimization procedure for selecting $\mathbf{w}$.

f_1 corresponds to *individual fairness*; for every cross pair $(\mathbf{x}, y) \in S_1, (\mathbf{x}', y') \in S_2$, a model $\mathbf{w}$ is penalized for how differently it treats $\mathbf{x}$ and $\mathbf{x}'$ (weighted by $d(y, y')$). No cancellation occurs: the penalty for overestimating several of one group's labels cannot be mitigated by overestimating several of the other group's labels.

Group Fairness The second fairness penalty we propose is the following:

$$f_2(\mathbf{w}, S) = \Big(\frac{1}{n_1 n_2} \sum_{\substack{(\mathbf{x}_i, y_i) \in S_1 \\ (\mathbf{x}_j, y_j) \in S_2}} d(y_i, y_j)\big(\mathbf{w} \cdot \mathbf{x}_i - \mathbf{w} \cdot \mathbf{x}_j\big)\Big)^2.$$

f_2 corresponds to *group fairness*: on average, the two groups' instances should have similar labels (weighted by function d). Unlike f_1, f_2 allows for *compensation*: informally, if $\mathbf{w}$ over-values some instances of one group 1 relative to group 2 in similar cross pairs, it can compensate on other similar cross-pairs by over-valuing those instances from group 2 relative to group 1.

In both of the above formulations, for any cross pair $(\mathbf{x}_i, y_i) \in S_1$ and $(\mathbf{x}_j, y_j) \in S_2$, any regressor $\mathbf{w}$ will have penalty that increases as $|\mathbf{w} \cdot \mathbf{x}_i - \mathbf{w} \cdot \mathbf{x}_j|$ increases, weighted by $d(y_i, y_j)$. If the cross pair is similar (y_i is close to y_j and $d(y_i, y_j)$ is large), a regressor which makes very different predictions for $\mathbf{x}_i$ and $\mathbf{x}_j$ will incur large loss. If the cross pair is less similar (y_i is far from y_j and $d(y_i, y_j)$ is smaller), there is less penalty for having a regressor for which $|\mathbf{w} \cdot \mathbf{x}_i - \mathbf{w} \cdot \mathbf{x}_j|$ is large.

Group and individual fairness correspond to two extremes and one could define different notions of fairness by grouping the cross pairs in different manners.

2.2 Discussion of Our Notions of Fairness

Difference between fairness and accuracy: All of our fairness penalties are small for any perfect regressor: for a similar cross pair, $y_i \approx y_j$ and also $\mathbf{w}\cdot\mathbf{x}_i \approx \mathbf{w}\cdot\mathbf{x}_j$ for a perfect regressor $\mathbf{w}$. Our fairness regularizers might then be interpreted as an unusual proxy for standard accuracy rather than as fairness notions. However, perfect (linear or otherwise) regressors almost never exist in practice; and between two models with similar accuracy, these definitions bias a learning procedure towards those which have similar treatment of similar cross pairs.

Fairness minimizers: Any constant regressor exactly minimizes all of our fairness regularizers. As we seen empirically, this implies that as λ increases, we transition from an unfair model with minimum accuracy loss to a constant (and perfectly fair but trivial) model, whose accuracy is the best any constant model can achieve.

3 Related Work

In classification, one line of work aims to achieve the group fairness notion known as *statistical parity*, i.e. to avoid disparate impact (see e.g. [1, 5, 10, 11, 13, 17–20, 24, 26, 31]). Statistical parity can be at odds with accuracy especially when the two groups are inherently different. Hardt et al. [14] introduced a new notion of group fairness called *equality of odds*, partially to alleviate this friction. Optimizing for accuracy subject to equality of odds was shown to be NP-hard [29]; work following this result presented efficient heuristics/relaxations [29, 32].

Calders et al. [6] were the first to study the statistical parity's analog in regression (called *equal means* and *balanced residuals*). Recently Johnson et al. [16] formalized several notions of *impartial estimates* for regression. Both papers consider group fairness. Our group fairness notion however incorporates the similarity of pairs (through the function d) though the specific choice of d as a constant would recover the equal means [6].

To achieve any of these fairness notions, one needs to decide whether or not to allow for *disparate treatment* (allowing for different treatment or models for different groups)[2], and where in the learning process to enforce fairness: preprocessing (e.g. [17]); inprocessing (e.g. [11, 20, 31]); or postprocessing (e.g. [14]). Our approach falls into the inprocessing by encoding fairness as a regularizer (an approach previously studied in e.g. [20, 31, 32]). Our work differs from previous work in several aspects mainly by focusing on regression rather classification. Moreover, our fairness measures draw inspiration from the idea that *similar* instances should be treated *similarly* [9, 33] though we define similarity of two instances based on their ground truth label.

[2]Any classifier that uses sensitive attributes is implicitly fitting separate models to the 2 groups. While this might seem unfair, it has been argued that it is actually necessary for fairness [9].

4 A Comparative Case Study

In this section we describe an empirical case study in which we apply our regularization framework to six different datasets in which fairness is a central concern. These datasets include cases in which the observed labels are real-valued, and cases in which they are binary-valued. We applied linear and logistic regression with our various fairness regularizers for the real-valued and binary datasets, respectively. For datasets with real-valued targets we normalized the inputs and outputs to be zero mean and unit variance, and we set the cross-group fairness weights as $d(y_i, y_j) = e^{-(y_i - y_j)^2}$; for datasets with binary targets we set $d(y_i, y_j) = \mathbb{1}[y_i = y_j]$.

For each dataset S, we solved optimization problems of the form $\min_{\mathbf{w}} \ell(\mathbf{w}, S) + \lambda f(\mathbf{w}, S) + \gamma ||\mathbf{w}||_2$ for variable values of λ, where $\ell(\mathbf{w}, S)$ is either MSE (linear regression) or the logistic regression loss. For each λ we picked γ as a function of λ by cross validation. Optimization problems are solved using the CVX solver.

The datasets themselves are summarized in Table 1, where we specify the size and dimensionality of each, along with the "protected" feature (race or gender) that thus defines the subgroups across which we apply our fairness criteria. The datasets vary considerably in the number of observations, their dimensionality, and the relative size of the minority subgroup.

The *Adult* dataset [22, 23] contains 1994 Census data, and the goal is to predict whether the income of an individual is more than 50K per year or not.[3] The *Communities and Crime* dataset [23] includes features relevant to per capita violent crime rates, and the goal is to predict this crime rate. The *COMPAS* dataset[4] contains data from Broward County, Florida, originally compiled by ProPublica [2], in which the goal is to predict whether a convicted individual would commit a violent crime in the following two years or not. The *Default* dataset [23, 30] contains data from Taiwanese credit card users, and the goal is to predict whether an individual will default on payments. The *Law School* dataset (`http://www2.law.ucla.edu/sander/Systemic/Data.htm`) consists of the records of law students who went on to take the bar exam and the goal is to predict whether a student will pass the exam. The *Sentencing* dataset contains information from a state department of corrections regarding inmates in 2010. The goal is to predict the sentence length given by the judge.

[3]We only used the data in Adult.data in our experiments.

[4]filtered similar to that of Corbett-Davies et al. [8].

Dataset	Type	n	d	Min n	Protected
Adult	logit	32561	14	10771	gender
Communities and Crime	linear	1994	128	227	race
COMPAS	logit	3373	19	1455	race
Default	logit	30000	24	11888	gender
Law School	logit	27478	36	12079	gender
Sentencing	linear	5969	17	385	gender

Table 1: Summary of datasets. Type indicates whether regression is logistic or linear; n is the number of data points; d is dimensionality; Min n is the number of data points in the smaller population; Protected indicates the protected feature.

4.1 Price of Fairness

To study the accuracy vs. fairness trade-off, we vary the weight λ on the fairness regularizer, and for each value of λ find the model which minimizes the associated regularized loss. For the logistic regression, we extract probabilities from the learned model $\mathbf{w}$ as $\Pr[y_i = 1] = 1/(1 + \exp(-\mathbf{w} \cdot x_i))$ and evaluate these probabilities as predictions for the binary labels using MSE.[5] In all of the datasets, as λ increases, the models converge to the best constant predictor, with minimum fairness penalty.

We now propose a measure for cross-dataset comparison we call *Price of Fairness* which has the effect of normalizing the fairness loss to the same scale. This is because otherwise comparing the trade-offs in a quantitative manner would be difficult since the scale of the fairness loss differs substantially from dataset to dataset.

For a given data set and regression type (linear or logistic), let $\mathbf{w}^*$ be the optimal model absent any fairness penalty (i.e. the empirical risk minimizer when $\lambda = 0$). This model will suffer some fairness penalty. For each dataset, we fix a normalization such that this fairness penalty is rescaled to be 1, and ask for the cost (in terms of the relative increase in mean squared error) of constraining our predictor to have fairness penalty $\alpha \leq 1$. Formally, let $\mathbf{w}^* = \arg\min_{\mathbf{w}} \ell_{\mathcal{P}}(\mathbf{w})$. For any value of $\alpha \in [0, 1]$ we define the *price of fairness* (PoF) as:

$$\mathrm{PoF}(\alpha) = \frac{\min_{\mathbf{w}} \ell_{\mathcal{P}}(\mathbf{w}) \text{ subject to } f_{\mathcal{P}}(\mathbf{w}) \leq \alpha f_{\mathcal{P}}(\mathbf{w}^*)}{\ell_{\mathcal{P}}(\mathbf{w}^*)}.$$

In theory, group fairness is strictly less costly than individual fairness for any particular model (by Jensen's inequality), and using separate models (one for each group) should strictly improve the fairness/accuracy trade-off for any of our fairness notions. However, PoF asks for the cost of *relative* improvements over the unconstrained optimum. So it can be that the PoF for one fairness penalty case is larger than for another, even if the *absolute* fairness loss for both the numerator and the denominator is smaller in the second case. With this observation in mind, we move to the empirical findings.

Figure 1 displays the PoF on each of the 6 datasets, for each fairness regularizer, and the single and separate models. Note that we see diversity of trade-offs across datasets. For some (e.g. COMPAS and Sentencing), increasing the fairness constraint by decreasing α has only a mild cost on accuracy. For others (e.g. Communities and Crime, and Law School), the cost increases steadily.

Next, we observe that with this normalization, although the difference between separate and single models remains small on most datasets, on two datasets, differences emerge. In the Law School dataset, restricting to a single model leads to a significantly higher PoF when considering the group fairness metric, compared to allowing separate models. In contrast, on the Adult dataset, restricting to a single model substantially *reduces* the PoF when considering individual fairness.

Finally, this normalization allows us to observe variation across fairness penalties in the *rate of change* in the PoF as α is decreased. In some datasets (e.g. Communities and Crime, and Sentencing), the PoF changes in lock-step across all measures of unfairness. However, for others (e.g. Default), the PoF increases substantially with α when we consider group or hybrid fairness measures, but is much more stable for individual fairness.

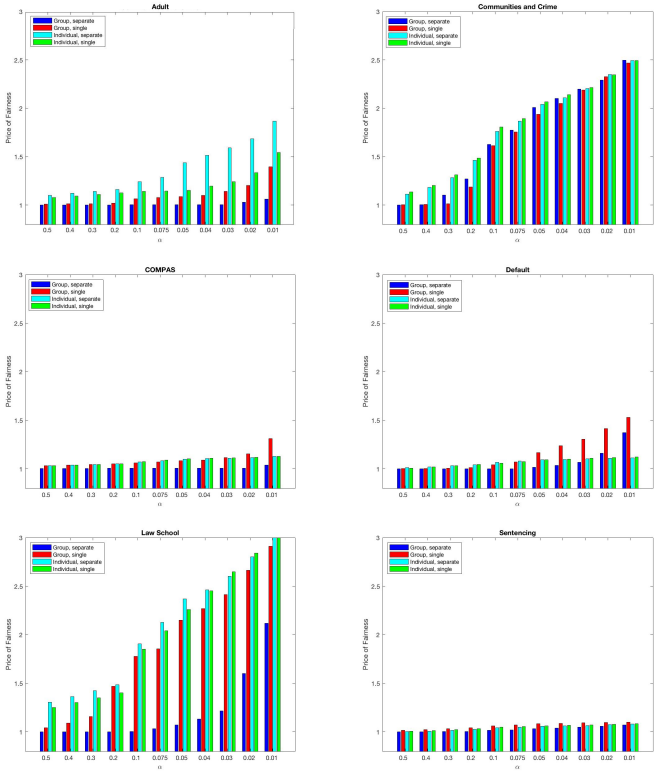

Figure 1: The PoF across data sets, for each type of fairness and single/separate models.

[5]Note that assessing the MSE of these probabilities, interpreted as predictions, is a sensible choice. Since squared error is a proper scoring rule, if the labels are indeed generated according to a logistic regression model, minimizing the squared error of a predictor using mean squared error will elicit the true model as its minimizer.

References

[1] P. Adler, C. Falk, S. Friedler, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian. Auditing black-box models for indirect influence. In *ICDM*, pages 1–10, 2016.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *Propublica*, 2016.

[3] A. Barry-Jester, B. Casselman, and D. Goldstein. The new science of sentencing. *The Marshall Project*, 2015.

[4] N. Byrnes. Artificial intolerance. *MIT Technology Review*, 2016.

[5] T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[6] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling attribute effect in linear regression. In *ICDM*, pages 71–80, 2013.

[7] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *coRR*, abs/1703.00056, 2017.

[8] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. *CoRR*, abs/1701.08230, 2017.

[9] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *ITCS*, pages 214–226, 2012.

[10] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268, 2015.

[11] B. Fish, J. Kun, and Á. D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *SDM*, pages 144–152, 2016.

[12] S. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im)possibility of fairness. *CoRR*, abs/1609.07236, 2016.

[13] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1445–1459, 2013.

[14] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016.

[15] Z. Jelveh and M. Luca. Towards diagnosing accuracy loss in discrimination-aware classification: An application to predictive policing. In *FATML*, 2015.

[16] K. Johnson, D. Foster, and R. Stine. Impartial predictive modeling: Ensuring fairness in arbitrary models. *CoRR*, abs/1608.00528, 2016.

[17] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2011.

[18] F. Kamiran, T. Calders, and M. Pechenizkiy. Discrimination aware decision tree learning. In *ICDM*, pages 869–874, 2010.

[19] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *ICDM*, pages 924–929, 2012.

[20] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *ECML/PKDD*, pages 35–50, 2012.

[21] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, 2017.

[22] R. Kohavi. Scaling up the accuracy of NB classifiers: A decision-tree hybrid. In *ICDM*, pages 202–207, 1996.

[23] M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

[24] B. T. Luong, S. Ruggieri, and F. Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *KDD*, pages 502–510, 2011.

[25] C. Miller. Can an algorithm hire better than a human? *The New York Times*, 2015.

[26] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *KDD*, pages 560–568. ACM, 2008.

[27] C. Rudin. Predictive policing using machine learning to detect patterns of crime. *Wired Magazine*, 2013.

[28] L. Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013.

[29] B. Woodworth, S. Gunasekar, M. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. *CoRR*, abs/1702.06081, 2017.

[30] I.-C. Yeh and C.-h. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

[31] M. Zafar, I. Valera, M. Gomez-Rodriguez, and K. Gummadi. Fairness constraints: A mechanism for fair classification. *CoRR*, abs/1507.05259, 2015.

[32] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*, pages 1171–1180, 2017.

[33] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML*, pages 325–333, 2013.