Fairness Constraints: Mechanisms for Fair Classification

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi
Max Planck Institute for Software Systems (MPI-SWS), Germany

Abstract

Algorithmic decision making systems are ubiquitous across a wide variety of online as well as offline services. These systems rely on complex learning methods and vast amounts of data to optimize the service functionality, satisfaction of the end user and profitability. However, there is a growing concern that these automated decisions can lead, even in the absence of intent, to a lack of fairness, *i.e.*, their outcomes can disproportionately hurt (or, benefit) particular groups of people sharing one or more sensitive attributes (*e.g.*, race, sex). In this paper, we introduce a flexible mechanism to design fair classifiers by leveraging a novel intuitive measure of decision boundary (un)fairness. We instantiate this mechanism with two well-known classifiers, logistic regression and support vector machines, and show on real-world data that our mechanism allows for a fine-grained control on the degree of fairness, often at a small cost in terms of accuracy.
A Python implementation of our mechanism is available at **fate-computing.mpi-sws.org**

1 INTRODUCTION

Algorithmic decision making processes are increasingly becoming automated and data-driven in both online (*e.g.*, spam filtering, product personalization) as well as offline (*e.g.*, pretrial risk assessment, mortgage approvals) settings. However, as automated data analysis replaces human supervision in decision making, and the scale of the analyzed data becomes "big", there are growing concerns from civil organizations [Bhandari, 2016], governments [Podesta et al., 2014, Muñoz et al., 2016], and researchers [Sweeney, 2013] about potential loss of transparency, accountability and fairness.

Anti-discrimination laws in many countries prohibit unfair treatment of people based on certain attributes, also called sensitive attributes, such as gender or race [Civil Rights Act, 1964]. These laws typically evaluate the fairness of a decision making process by means of two distinct notions [Barocas and Selbst, 2016]: ***disparate treatment*** and ***disparate impact***. A decision making process suffers from disparate treatment if its decisions are (partly) based on the subject's sensitive attribute information, and it has disparate impact if its outcomes disproportionately hurt (or, benefit) people with certain sensitive attribute values (*e.g.*, females, blacks).

While it is desirable to design decision making systems free of disparate treatment as well as disparate impact, controlling for both forms of unfairness *simultaneously* is challenging. One could avoid disparate treatment by ensuring that the decision making process does not have access to sensitive attribute information (and hence cannot make use of it). However, ignoring the sensitive attribute information may still lead to disparate impact in outcomes: since automated decision-making systems are often trained on historical data, if a group with a certain sensitive attribute value was unfairly treated in the past,[1] this unfairness may persist in future predictions through *indirect discrimination* [Pedreschi et al., 2008], leading to disparate impact. Similarly, avoiding disparate impact in outcomes by using sensitive attribute information while making decisions would constitute disparate treatment, and may also lead to *reverse discrimination* [Ricci vs. DeStefano, 2009].

In this work, our goal is to design classifiers—specifically, convex margin-based classifiers like logistic regression and support vector machines (SVMs)—that avoid *both* disparate treatment and disparate impact, and can additionally accommodate the "business necessity" clause of disparate impact doctrine. Accord-

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

[1] Like earlier studies on fairness aware-learning, we assume that while historical class labels may be biased against group(s) with certain sensitive attribute value(s), they still contain *some* degree of information on the *true* (unbiased) labels. Assuming the class labels to be completely biased and having no information on the true labels would render a learning task on such a dataset pointless.

ing to the business necessity clause, an employer can justify a certain degree of disparate impact in order to meet certain performance-related constraints [Barocas and Selbst, 2016]. However, the employer needs to ensure that the current decision making incurs the *least possible* disparate impact under the given constraints. To the best of our knowledge, this clause has not been addressed by any prior study on fairness-aware learning.

While there is no specific numerical formula laid out by anti-discrimination laws to quantify disparate impact, here we leverage a specific instantiation supported by the U.S. Equal Employment Opportunity Commission: the *"80%-rule"* (or more generally, the p%-rule) [Biddle, 2005]. The p%-rule states that the ratio between the percentage of subjects having a certain sensitive attribute value assigned the positive decision outcome and the percentage of subjects not having that value also assigned the positive outcome should be no less than p:100. Since it is very challenging to directly incorporate this rule in the formulation of convex margin-based classifiers, we introduce a novel intuitive measure of decision boundary (un)fairness as a tractable proxy to the rule: the covariance between the sensitive attributes and the (signed) distance between the subjects' feature vectors and the decision boundary of the classifier.

Our measure of fairness allows us to derive two complementary formulations for training fair classifiers: one that maximizes accuracy subject to fairness constraints, and enables compliance with disparate impact doctrine in its basic form (*i.e.*, the p%-rule); and another that maximizes fairness subject to accuracy constraints, and ensures fulfilling the business necessity clause of disparate impact. Remarkably, both formulations also avoid disparate treatment, since they do not make use of sensitive attribute information while making decisions. Our measure additionally satisfies several desirable properties: (i) for a wide variety of convex margin-based (linear and non-linear) classifiers, it is convex and can be readily incorporated in their formulation without increasing their complexity; (ii) it allows for clear mechanism to trade-off fairness and accuracy; and, (iii) it can be used to ensure fairness with respect to several sensitive attributes. Experiments with two well-known classifiers, logistic regression and support vector machines, using both synthetic and real-world data show that our fairness measure allows for a fine-grained control of the level of fairness, often at a small cost in terms of accuracy, and provides more flexibility than the state-of-the-art (see Table 1).

Related Work. A number of prior studies have focused on controlling disparate impact and/or disparate treatment-based discrimination in the context of binary classification [Romei and Ruggieri, 2014]. These studies have typically adopted one of the two following strategies:

The first strategy consists of pre-processing the training data [Dwork et al., 2012, Feldman et al., 2015, Kamiran and Calders, 2009, 2010]. This typically involves (i) changing the value of the sensitive attributes or class labels of individual items in the training data, or (ii) mapping the training data to a transformed space where the dependencies between sensitive attributes and class labels disappear. However, these approaches treat the learning algorithm as a *black box* and, as a consequence, the pre-processing can lead to unpredictable losses in accuracy.

The second strategy consists of modifying existing classifiers to limit discrimination [Calders and Verwer, 2010, Kamishima et al., 2011, Goh et al., 2016]. Among them, the work by Kamishima et al. [Kamishima et al., 2011] is the most closely related to ours: it introduces a regularization term to penalize discrimination in the formulation of the logistic regression classifier.

Recently, Zemel et al. [2013], building on Dwork et al. [2012], combined both strategies by jointly learning a fair representation of the data and the classifier parameters. This approach has two main limitations: i) it leads to a non-convex optimization problem and does not guarantee optimality, and ii) the accuracy of the classifier depends on the dimension of the fair representation, which needs to be chosen rather arbitrarily.

Many of the prior studies suffer from one or more of the following limitations: (i) they are restricted to a narrow range of classifiers, (ii) they only accommodate a single, binary sensitive attribute, and (iii) they cannot eliminate disparate treatment and disparate impact *simultaneously*. Table 1 compares the capabilities of different methods while achieving fairness.

Finally, as discussed earlier, disparate impact is particularly well suited as a fairness criterion when historical decisions used during the training phase are biased against certain social groups. In such contexts, proportionality in outcomes (*e.g.*, p%-rule) may help mitigate these historical biases. However, in cases where the (unbiased) ground-truth is available for the training phase, *i.e.*, one can tell whether a historical decision was *right* or *wrong*, disproportionality in outcomes can be explained by the means of the ground-truth. In those cases, disparate impact may be a rather misleading notion of fairness, and other recently proposed criteria like "disparate mistreatment" by Zafar et al. [2017], may be better suited notions of fairness. For more discussion into this alternative notion, we point the reader to our companion paper [Zafar et al., 2017].

Method	Disp. Treat.	Disp. Imp.	Business. Necessity	Polyvalent sens. attrs.	Multiple sens. attrs.	Range of classifiers
Our method	✓	✓	✓	✓	✓	Any convex margin-based
Kamiran and Calders [2010]	✓	✓	✗	✗	✗	Any score-based
Calders and Verwer [2010]	✓	✓	✗	✗	✗	Naive Bayes
Luong et al. [2011]	✓	✗	✗	✗	✗	Any score-based
Kamishima et al. [2011]	✗	✓	✗	✗	✗	Logistic Regression
Zemel et al. [2013]	✓	✓	✗	✗	✗	Log loss
Feldman et al. [2015]	✓	✓	✗	✓	✓	Any (only numerical non-sens. attrs.)
Goh et al. [2016]	✓	✓	✗	✓	✓	Ramp loss

Table 1: Capabilities of different methods in eliminating disparate impact and/or disparate treatment. None of the prior methods addresses disparate impact's business necessity clause. Many of the methods do not generalize to multiple (*e.g.*, gender and race) or polyvalent sensitive attributes (*e.g.*, race, that has more than two values).

2 FAIRNESS IN CLASSIFICATION

For simplicity, we consider binary classification tasks in this work. However, our ideas can be easily extended to m-ary classification.

In a binary classification task, one needs to find a mapping function $f(\mathbf{x})$ between user feature vectors $\mathbf{x} \in \mathbb{R}^d$ and class labels $y \in \{-1, 1\}$. This task is achieved by utilizing a training set, $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, to construct a mapping that works *well* on an *unseen* test set. For margin-based classifiers, finding this mapping usually reduces to building a decision boundary in feature space that separates users in the training set according to their class labels. One typically looks for a decision boundary, defined by a set of parameters $\boldsymbol{\theta}^*$, that achieves the greatest classification accuracy in a test set, by minimizing a loss function over a training set $L(\boldsymbol{\theta})$, *i.e.*, $\boldsymbol{\theta}^* = \text{argmin}_{\boldsymbol{\theta}}\, L(\boldsymbol{\theta})$. Then, given an *unseen* feature vector $\mathbf{x}_i$ from the test set, the classifier predicts $f_{\boldsymbol{\theta}}(\mathbf{x}_i) = 1$ if $d_{\boldsymbol{\theta}^*}(\mathbf{x}_i) \geq 0$ and $f_{\boldsymbol{\theta}}(\mathbf{x}_i) = -1$ otherwise, where $d_{\boldsymbol{\theta}^*}(\mathbf{x})$ denotes the signed distance from the feature vector $\mathbf{x}$ to the decision boundary.

If class labels in the training set are correlated with one or more sensitive attributes $\{\mathbf{z}_i\}_{i=1}^N$ (*e.g.*, gender, race), the percentage of users with a certain sensitive attribute having $d_{\boldsymbol{\theta}^*}(\mathbf{x}_i) \geq 0$ may differ dramatically from the percentage of users without this sensitive attribute value having $d_{\boldsymbol{\theta}^*}(\mathbf{x}_i) \geq 0$ (*i.e.*, the classifier may suffer from disparate impact). Note that this may happen even if sensitive attributes are not used to construct the decision boundary but are correlated with one or more of user features, through indirect discrimination [Pedreschi et al., 2008].

2.1 Fairness Definition

First, to comply with **disparate treatment** criterion we specify that sensitive attributes are not used in decision making, *i.e.*, $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{z}_i\}_{i=1}^N$ consist of disjoint feature sets.

Next, as discussed in Section 1, our definition of **disparate impact** leverages the "80%-rule" [Biddle, 2005]. A decision boundary satisfies the "80%-rule" (or more generally the "p%-rule"), if the ratio between the percentage of users with a particular sensitive attribute value having $d_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0$ and the percentage of users without that value having $d_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0$ is no less than 80:100 (p:100). For a given binary sensitive attribute $z \in \{0, 1\}$, one can write the p%-rule as:

$$\min\left(\frac{P(d_{\boldsymbol{\theta}}(\mathbf{x})\geq 0|z=1)}{P(d_{\boldsymbol{\theta}}(\mathbf{x})\geq 0|z=0)}, \frac{P(d_{\boldsymbol{\theta}}(\mathbf{x})\geq 0|z=0)}{P(d_{\boldsymbol{\theta}}(\mathbf{x})\geq 0|z=1)}\right) \geq \frac{p}{100}. \tag{1}$$

Unfortunately, it is very challenging to directly incorporate the p%-rule in the formulation of convex margin-based classifiers, since it is a non-convex function of the classifier parameters $\boldsymbol{\theta}$ and, therefore, it would lead to non-convex formulations, which are difficult to solve efficiently. Secondly, as long as the user feature vectors lie on the same side of the decision boundary, the p%-rule is invariant to changes in the decision boundary. In other words, the p%-rule is a function having saddle points. The presence of saddle points furthers complicate the procedure for solving non-convex optimization problems [Dauphin et al., 2014]. To overcome these challenges, we next introduce a novel measure of decision boundary (un)fairness which can be used as a proxy to efficiently design classifiers satisfying a given p%-rule.

3 OUR APPROACH

In this section, we first introduce our measure of decision boundary (un)fairness, the decision boundary covariance. We then derive two complementary formulations. The first formulation ensures compliance with disparate impact doctrine in its basic form (ensure a given p%-rule) by maximizing accuracy subject to fairness constraints. The second formulation guarantees fulfilling disparate impact's "business necessity" clause by maximizing fairness subject to accuracy constraints.

For conciseness, we append a constant 1 to all feature vectors ($\mathbf{x_i}$) so that the linear classifier decision boundary equation $\boldsymbol{\theta}^T\mathbf{x} + b = 0$ reduces to $\boldsymbol{\theta}^T\mathbf{x} = 0$.

3.1 Decision Boundary Covariance

Our measure of decision boundary (un)fairness is defined as the covariance between the users' sensitive attributes, $\{\mathbf{z}_i\}_{i=1}^N$, and the signed distance from the users' feature vectors to the decision boundary, $\{d_{\boldsymbol{\theta}}(\mathbf{x}_i)\}_{i=1}^N$, *i.e.*:

$$\begin{aligned} \text{Cov}(\mathbf{z}, d_{\boldsymbol{\theta}}(\mathbf{x})) &= \mathbb{E}[(\mathbf{z}-\bar{\mathbf{z}})d_{\boldsymbol{\theta}}(\mathbf{x})] - \mathbb{E}[(\mathbf{z}-\bar{\mathbf{z}})]\bar{d}_{\boldsymbol{\theta}}(\mathbf{x}) \\ &\approx \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i-\bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i), \end{aligned} \tag{2}$$

where $\mathbb{E}[(\mathbf{z}-\bar{\mathbf{z}})]\bar{d}_{\boldsymbol{\theta}}(\mathbf{x})$ cancels out since $\mathbb{E}[(\mathbf{z}-\bar{\mathbf{z}})] = 0$. Since in linear models for classification, such as logistic regression or linear SVMs, the decision boundary is simply the hyperplane defined by $\boldsymbol{\theta}^T\mathbf{x} = 0$, Eq. (2) reduces to $\frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i-\bar{\mathbf{z}})\,\boldsymbol{\theta}^T\mathbf{x}_i$.

In contrast with the p%-rule (Eq. 1), the decision boundary covariance (Eq. 2) is a convex function with respect to the decision boundary parameters $\boldsymbol{\theta}$, since $d_{\boldsymbol{\theta}}(\mathbf{x}_i)$ is convex with respect to $\boldsymbol{\theta}$ for all linear, convex margin-based classifiers.[2] Hence, it can be easily included in the formulation of these classifiers without increasing the complexity of their training.

Moreover, note that, if a decision boundary satisfies the "100%-rule", *i.e.*,

$$P(d_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0 | z = 0) = P(d_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0 | z = 1), \tag{3}$$

then the (empirical) covariance will be approximately zero for a sufficiently large training set.

3.2 Maximizing Accuracy Under Fairness Constraints

In this section, we design classifiers that maximize accuracy subject to fairness constraints (*e.g.*, a specific p%-rule), and thus may be used to ensure compliance with the disparate impact doctrine in its basic form.

To this end, we find the decision boundary parameters $\boldsymbol{\theta}$ by minimizing the corresponding loss function over the training set under fairness constraints, *i.e.*:

$$\begin{aligned} \text{minimize} \quad & L(\boldsymbol{\theta}) \\ \text{subject to} \quad & \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\mathbf{z}_i-\bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i) \leq \mathbf{c}, \\ & \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\mathbf{z}_i-\bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq -\mathbf{c}, \end{aligned} \tag{4}$$

where $\mathbf{c}$ is the covariance threshold, which specifies an upper bound on the covariance between each sensitive attribute and the signed distance from the feature vectors to the decision boundary. In this formulation, $\mathbf{c}$ trades off fairness and accuracy, such that as we decrease $\mathbf{c}$ towards zero, the resulting classifier will satisfy a larger p%-rule but will potentially suffer from a larger loss in accuracy. Note that since the above optimization problem is convex, our scheme ensures that the trade-off between the classifier loss function and decision boundary covariance is Pareto optimal.

Remarks. It is important to note that the distance to the margin, $d_{\boldsymbol{\theta}}(\mathbf{x})$, only depends on the non-sensitive features $\mathbf{x}$ and, therefore, the sensitive features $\mathbf{z}$ are not needed while making decisions. In other words, we account for *disparate treatment*, by removing the sensitive features from the decision making process and, for *disparate impact*, by adding fairness constraints during the training process of the classifier. Additionally, the constrained optimization problem (4), can also be written as a regularized optimization problem by making use of its dual form, in which the fairness constraints are moved to the objective and the corresponding Lagrange multipliers act as regularizers.

Next, we specialize problem (4) for logistic regression classifiers.

Logistic Regression. In logistic regression classifiers, one maps the feature vectors $\mathbf{x}_i$ to the class labels y_i by means of a probability distribution:

$$p(y_i = 1 | \mathbf{x}_i, \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T\mathbf{x}_i}}, \tag{5}$$

where $\boldsymbol{\theta}$ is obtained by solving a maximum likelihood problem over the training set, *i.e.*, $\boldsymbol{\theta}^* = \text{argmin}_{\boldsymbol{\theta}} -\sum_{i=1}^{N}\log p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$. Thus, the corresponding loss function is given by $-\sum_{i=1}^{N}\log p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$, and problem (4) adopts the form:

$$\begin{aligned} \text{minimize} \quad & -\textstyle\sum_{i=1}^{N}\log p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \\ \text{subject to} \quad & \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\mathbf{z}_i-\bar{\mathbf{z}})\,\boldsymbol{\theta}^T\mathbf{x}_i \leq \mathbf{c}, \\ & \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\mathbf{z}_i-\bar{\mathbf{z}})\,\boldsymbol{\theta}^T\mathbf{x}_i \geq -\mathbf{c}, \end{aligned} \tag{6}$$

Appendix A presents the specialization of our formulation for both linear and non-linear SVM classifiers.

3.3 Maximizing Fairness Under Accuracy Constraints

In the previous section, we designed classifiers that maximize accuracy subject to fairness constraints. However, if the underlying correlation between the class labels and the sensitive attributes in the training set is very high, enforcing fairness constraints may results in underwhelming performance (accuracy) and thus be unacceptable in terms of business objectives. Disparate impact's "business necessity" clause accounts for such scenarios by allowing *some* degree of disparate impact in order to meet performance constraints. However, the employer needs to ensure that the decision making causes *least possible* disparate impact under the given performance (accuracy) constraints [Barocas and Selbst, 2016]. To accommodate

[2] For non-linear convex margin-based classifiers like nonlinear SVM, equivalent of $d_{\boldsymbol{\theta}}(\mathbf{x}_i)$ is still convex in the transformed kernel space. See Appendix A for details.

such scenarios, we now propose an alternative formulation that maximizes fairness (minimizes disparate impact) subject to accuracy constraints.

To this aim, we find the decision boundary parameters $\boldsymbol{\theta}$ by minimizing the corresponding (absolute) decision boundary covariance over the training set under constraints on the classifier loss function, *i.e.*:

$$\begin{array}{ll} \text{minimize} & |\frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i)| \\ \text{subject to} & L(\boldsymbol{\theta}) \leq (1+\gamma)L(\boldsymbol{\theta}^*), \end{array} \tag{7}$$

where $L(\boldsymbol{\theta}^*)$ denotes the optimal loss over the training set provided by the unconstrained classifier and $\gamma \geq 0$ specifies the maximum additional loss with respect to the loss provided by the unconstrained classifier. Here, we can ensure maximum fairness with no loss in accuracy by setting $\gamma = 0$. As in Section 3.2, it is possible to specialize problem (7) for the same classifiers and show that the formulation remains convex.

Fine-Grained Accuracy Constraints. In many classifiers, including logistic regression and SVMs, the loss function (or the dual of the loss function) is additive over the points in the training set, *i.e.*, $L(\boldsymbol{\theta}) = \sum_{i=1}^{N} L_i(\boldsymbol{\theta})$, where $L_i(\boldsymbol{\theta})$ is the individual loss associated with the i-th point in the training set. Moreover, the individual loss $L_i(\boldsymbol{\theta})$ typically tells us how *close* the predicted label $f(\mathbf{x}_i)$ is to the true label y_i, by means of the signed distance to the decision boundary. Therefore, one may think of incorporating loss constraints for a certain set of users, and consequently, prevent individual users originally classified as positive (by the unconstrained classifier) from being classified as negative by the constrained classifier. To do so, we find the decision boundary parameters $\boldsymbol{\theta}$ as:

$$\begin{array}{ll} \text{minimize} & |\frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\, \boldsymbol{\theta}^T\mathbf{x}_i| \\ \text{subject to} & L_i(\boldsymbol{\theta}) \leq (1+\gamma_i)L_i(\boldsymbol{\theta}^*) \quad \forall i \in \{1,\ldots,N\}, \end{array} \tag{8}$$

where $L_i(\boldsymbol{\theta}^*)$ is the individual loss associated to the i-th user in the training set provided by the unconstrained classifier and $\gamma_i \geq 0$ is her allowed additional loss. For example, in the case of logistic regression classifier, $\boldsymbol{\theta}^* = \text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{N} -\log p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$ and the losses for individual points are $L_i(\boldsymbol{\theta}) = -\log p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$. Now, if we set $\gamma_i = 0$, we are enforcing that the probability of the i-th user to be mapped in the positive class to be equal or higher than in the original (unconstrained) classifier.

4 EVALUATION

We evaluate our framework on several synthetic and real-world datasets. We first experiment with our first formulation and show that it allows for fine-grained fairness control, often at a minimal loss in accuracy. Then, we validate our second formulation, which allows maximizing fairness under accuracy constraints, and also provides guarantees on avoiding negative classification of certain individual users or group of users.

Here, we adopt the p%-rule [Biddle, 2005] as our true measure of fairness. However, as shown in Appendix B.2, we obtain similar results if we consider another measure of fairness used by some of the previous studies in this area.

4.1 Experiments on Synthetic Data

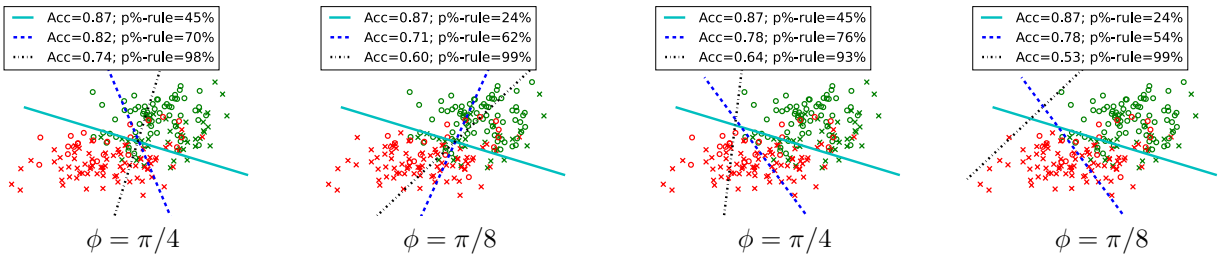
Fairness constraints vs accuracy constraints. To simulate different degrees of disparate impact in classification outcomes, we generate two synthetic datasets with different levels of correlation between a single, binary sensitive attribute and class labels. We then train two types of logistic regression classifiers: one type maximizes accuracy subject to fairness constraints (Section 3.2), and the other maximizes fairness under fine-grained accuracy constraints (Section 3.3).

Specifically, we generate 4,000 binary class labels uniformly at random and assign a 2-dimensional user feature vector per label by drawing samples from two different Gaussian distributions: $p(x|y{=}1) = N([2;2],[5,\ 1;1,\ 5])$ and $p(x|y{=}{-}1) = N([-2;-2],[10,\ 1;1,\ 3])$. Then, we draw each user's sensitive attribute z from a Bernoulli distribution: $p(z=1) = p(\mathbf{x}'|y=1)/(p(\mathbf{x}'|y=1) + p(\mathbf{x}'|y=-1))$, where $\mathbf{x}' = [\cos(\phi),\ -\sin(\phi); \sin(\phi),\ \cos(\phi)]\mathbf{x}$ is simply a rotated version of the feature vector, $\mathbf{x}$.

We generate datasets with two values for the parameter ϕ, which controls the correlation between the sensitive attribute and the class labels (and hence, the resulting disparate impact). Here, the closer ϕ is to zero, the higher the correlation. Finally, we trained both types of constrained classifiers on each dataset.

Fig. 1a shows the decision boundaries provided by the classifiers that maximize accuracy under fairness constraints for two different correlation values and two (successively decreasing) covariance thresholds, $\mathbf{c}$. We compare these boundaries against the unconstrained decision boundary (solid line). As expected, given the data generation process, fairness constraints map into a rotation of the decision boundary (dashed lines), which is greater as we decrease threshold value $\mathbf{c}$ or increase the correlation in the original data (from $\phi = \pi/4$ to $\phi = \pi/8$). This movement of the decision boundaries shows that our fairness constraints are successfully undoing (albeit in a highly controlled setting) the rotations we used to induce disparate impact in the dataset. Moreover, a smaller covariance threshold (a larger rotation) leads to a more fair solution, although, it comes at a larger cost in accuracy.

Fig. 1b shows the decision boundaries provided by the classifiers that maximize fairness under fine-grained accuracy constraints. Here, the fine-grained accuracy



(a) Maximizing accuracy under fairness constraints (b) Maximizing fairness under accuracy constraints

Figure 1: The solid light blue lines show the decision boundaries for logistic regressors without fairness constraints. The dashed lines show the decision boundaries for fair logistic regressors trained (a) to maximize accuracy under fairness constraints and (b) to maximize fairness under fine-grained accuracy constraints, which prevents users with $z = 1$ (circles) labeled as positive by the unconstrained classifier from being moved to the negative class. Each column corresponds to a dataset, with different correlation value between sensitive attribute values (crosses vs circles) and class labels (red vs green).

constraints ensure that the users with $z = 1$ classified as positive by the unconstrained classifier (circles above the solid line) are not labeled as negative by the fair classifier. The decision boundaries provided by this formulation, in contrast to the previous one, are rotated *and shifted* versions of the unconstrained boundary. Such shifts enable the constrained classifiers to avoid negatively classifying users specified in the constraints.

We also illustrate how the decision boundary of a non-linear classifier, a SVM with RBF kernel, changes under fairness constraints in Appendix B.1.

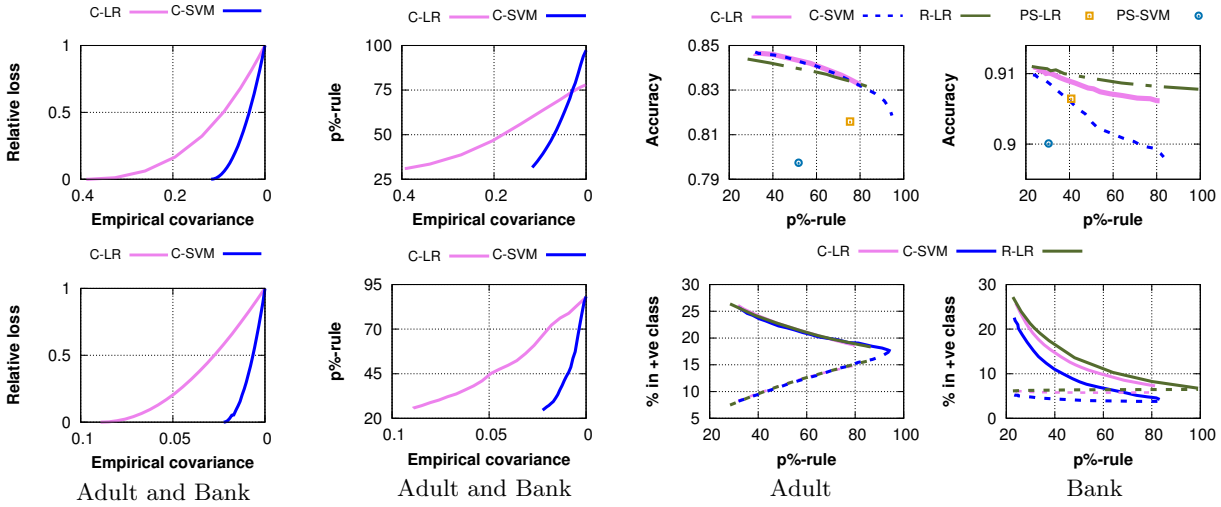
4.2 Experiments on Real Data

Experimental Setup. We experiment with two real-world datasets: The Adult income dataset [Adult data, 1996] and the Bank marketing dataset [Bank data, 2014]. The Adult dataset contains a total of 45,222 subjects, each with 14 features (e.g., age, educational level) and a binary label, which indicates whether a subject's incomes is above (positive class) or below (negative class) 50K USD. For this dataset, we consider gender and race, respectively, as binary and non-binary (polyvalent) sensitive attributes. The Bank dataset contains a total of 41,188 subjects, each with 20 attributes (e.g., marital status) and a binary label, which indicates whether the client has subscribed (positive class) or not (negative class) to a term deposit. In this case, we consider age as (binary) sensitive attribute, which is discretized to indicate whether the client's age is between 25 and 60 years. For detailed statistics about the distribution of different sensitive attributes in positive class in these datasets, we refer the reader to Appendix B.2.

For the sake of conciseness, while presenting the results for binary sensitive attributes, we refer to females and males, respectively, as protected and non-protected groups in Adult data. Similarly, in Bank data, we refer to users between age 25 and 60 as protected and rest of the users as non-protected group. In our experiments, to obtain more reliable estimates of accuracy and fairness, we repeatedly split each dataset into a train (70%) and test (30%) set 5 times and report the average statistics for accuracy and fairness.

Maximizing accuracy under fairness constraints. First, we experiment with a **single binary** sensitive attribute, gender and age, for respectively, the Adult and Bank data. For each dataset, we train several logistic regression and SVM classifiers (denoted by 'C-LR' and 'C-SVM', respectively), each subject to fairness constraints with different values of covariance threshold $\mathbf{c}$ (Section 3.2), and then empirically investigate the trade-off between accuracy and fairness. Fig. 2a shows the (empirical) decision boundary covariance against the relative loss incurred by the classifier. The 'relative loss' is normalized between the loss incurred by an unconstrained classifier and by the classifier with a covariance threshold of 0. Here, each pair of (covariance, loss) values is guaranteed to be Pareto optimal, since our problem formulation is convex. Additionally, Fig. 2b investigates the correspondence between decision boundary covariance and p%-rule computed on the training set, showing that, as desired: i) the lower the covariance, the higher the p%-rule the classifiers satisfy and (ii) a 100%-rule maps to zero covariance.

Then, we compare our approach to a well-known competing method from each of the two categories discussed in Section 1: preferential sampling approach [Kamiran and Calders, 2010], applied to logistic regression ('PS-LR') and SVM ('PS-SVM'), as an example of data pre-processing, and the regularized logistic regression ('R-LR') [Kamishima et al., 2011], as an example of modifying a classifier to limit unfairness. Fig. 2c summarizes the results: the top panel shows av-



(a) Loss vs. Cov. (b) Cov. vs. p%-rule (c) Single binary sensitive attribute

Figure 2: [Maximizing accuracy under fairness constraints: single, binary sensitive attribute] Panels in (a) show the trade-off between the empirical covariance in Eq. 2 and the relative loss (with respect to the unconstrained classifier), for the Adult (top) and Bank (bottom) datasets. Here each pair of (covariance, loss) values is guaranteed to be Pareto optimal by construction. Panels in (b) show the correspondence between the empirical covariance and the p%-rule for classifiers trained under fairness constraints. Panels in (c) show the accuracy against p%-rule value (top) and the percentage of protected (dashed) and non-protected (solid) users in the positive class against the p%-rule value (bottom).

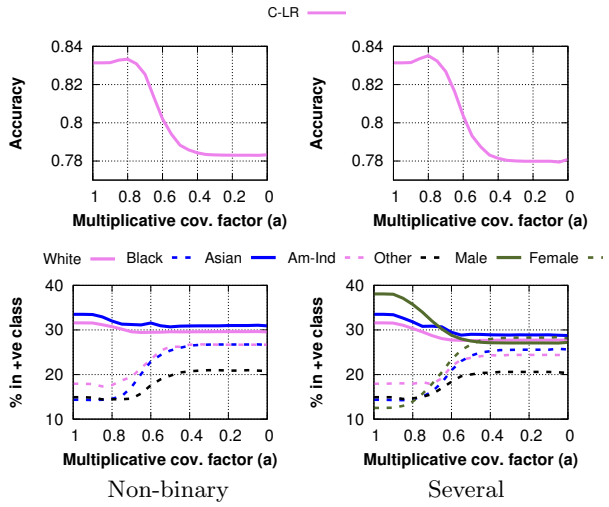
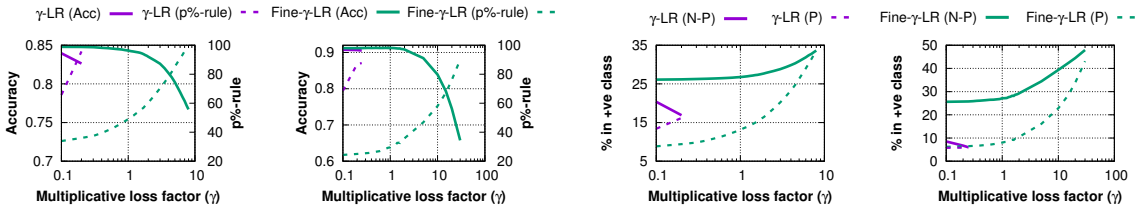


Figure 3: [Maximizing accuracy under fairness constraints: non-binary and several sensitive attributes] The figure shows accuracy (top) and percentage of users in positive class (bottom) against a multiplicative factor $a \in [0, 1]$ such that $\mathbf{c} = a\mathbf{c}^*$, where $\mathbf{c}^*$ denotes the unconstrained classifier covariance.

erage accuracy and the bottom panel the percentage of protected (dashed lines) and non-protected (solid lines) users in positive class against the average p%-rule, as computed on test sets. We observe that: i) the performance of our classifiers (C-LR, C-SVM) and regularized logistic regression (R-LR) is comparable, ours are slightly better for Adult data (left column) while slightly worse for Bank data (right column). However, R-LR uses sensitive attribute values from the test set to make predictions, failing the disparate treatment test and potentially allowing for reverse discrimination [Ricci vs. DeStefano, 2009]; ii) the preferential sampling presents the worst performance and always achieves p%-rules under 80%; and, (iii) in the Adult data, all classifiers move non-protected users (males) to the negative class and protected users (females) to the positive class to achieve fairness, in contrast, in the Bank data, they only move non-protected (young and old) users originally labeled as positive to the negative class since it provides a smaller accuracy loss. However, the latter can be problematic: from a business perspective, a bank may be interested in finding potential subscribers rather than losing existing customers. This last observation motivates our second formulation (Section 3.3), which we experiment with later in this section.

Finally, we experiment with **non-binary** (race) and **several** (gender and race) sensitive attributes in Adult dataset. We do not compare with competing methods since they cannot handle non-binary or several sensitive attributes. Fig. 3 summarizes the results by showing the accuracy and the percentage of subjects sharing each sensitive attribute value classified as positive against a multiplicative covariance factor $a \in [0, 1]$ such that $\mathbf{c} = a\mathbf{c}^*$, where $\mathbf{c}^*$ is the unconstrained classifier covariance[3] (note that p%-rule is only defined for

[3]For several sensitive features, we compute the initial covariance $\mathbf{c}_k^*$ for each of the sensitive feature k, and then



(a) Acc and p%-rule (b) % in +ve class

Figure 4: [Maximizing fairness under accuracy constraints] Panels in (a) show accuracy (solid) and p%-rule value (dashed) against γ. Panels in (b) show the percentage of protected (P, dashed) and non-protected (N-P, solid) users in the positive class against γ.

a binary sensitive feature). As expected, as the value of $\mathbf{c}$ decreases, the percentage of subjects in the positive class from sensitive attribute value groups become nearly equal[4] while the loss in accuracy is modest.

Maximizing fairness under accuracy constraints. Next, we demonstrate that our second formulation (Section 3.3) can maximize fairness while precisely controlling loss in accuracy. To this end, we first train several logistic regression classifiers (denoted by 'γ-LR'), which minimize the decision boundary covariance subject to accuracy constraints over the entire dataset by solving problem (7) with increasing values of γ. Then, we train logistic regression classifiers (denoted by 'Fine-γ-LR') that minimize the decision boundary covariance subject to *fine-grained* accuracy constraints by solving problem (8). Here, we prevent the non-protected users that were classified as positive by the unconstrained logistic regression classifier from being classified as negative by constraining that their distance from decision boundary stays positive while learning the fair boundary. We then increase $\gamma_i = \gamma$ for the remaining users. In both cases, we increased the value of γ until we reach a 100%-rule during training. Fig. 4 summarizes the results for both datasets, by showing (a) the average accuracy (solid curves) and p%-rule (dashed curves) against γ, and (b) the percentage of non-protected (N-P, solid curves) and protected (P, dashed curves) users in the positive class against γ. We observe that, as we increase γ, the classifiers that constrain the overall training loss (γ-LR) remove non-protected users from the positive class and add protected users to the positive class, in contrast, the classifiers that prevent the non-protected users that were classified as positive in the unconstrained classifier from being classified as negative (Fine-γ-LR) add both protected and non-protected users to the positive class. As a consequence, the latter achieves lower accuracy for the same p%-rule.

5 DISCUSSION & FUTURE WORK

In this paper, we introduced a novel measure of decision boundary fairness, which enables us to ensure fairness with respect to one or more sensitive attributes, in terms of both disparate treatment and disparate impact. We leverage this measure to derive two complementary formulations: one that maximizes accuracy subject to fairness constraints, and helps ensure compliance with a non-discrimination policy or law (*e.g.*, a given p%-rule); and another one that maximizes fairness subject to accuracy constraints, and ensures fulfilling certain business needs (*e.g.*, disparate impact's business necessity clause).

Our framework opens many avenues for future work. For example, one could include fairness constraints in other supervised (*e.g.*, regression, recommendation) as well as unsupervised (*e.g.*, set selection, ranking) learning tasks. Further, while we note that a decreasing covariance threshold corresponds to an increasing (more fair) p%-rule, the relation between the two is only empirically observed. A precise mapping between covariance and p%-rule is quite challenging to derive analytically since it depends on the specific classifier and the dataset being used. Such a theoretical analysis would be an interesting future direction. Finally, in this paper we consider disparate impact as our notion of fairness with the assumption that the historical training data may contain biases against certain group(s). Since the actual proportions in positive class from different groups (*e.g.*, males, females) in the ground-truth dataset are not available (we only have access to a biased dataset), ensuring equal proportions from each group in the positive class (removing disparate impact) serves as an attractive notion of fairness. However, in cases where the historical ground truth decisions are available, disparate impact can be explained by the means of the ground truth, and alternative notions of fairness, *e.g.*, disparate mistreatment [Zafar et al., 2017], might be more suitable.

compute the covariance threshold separately for each sensitive feature as $a\mathbf{c}_k^*$.

[4]The scarce representation of the race value 'Other' (only 0.8% of the data) hinders an accurate estimation of the decision boundary covariance and, as a result, the classifier does not reach perfect fairness with respect to this sensitive attribute value.

References

Adult data. http://tinyurl.com/UCI-Adult, 1996.

Bank data. http://tinyurl.com/UCI-Bank, 2014.

S. Barocas and A. D. Selbst. Big Data's Disparate Impact. *California Law Review*, 2016.

E. Bhandari. Big Data Can Be Used To Violate Civil Rights Laws, and the FTC Agrees. https://www.aclu.org/blog/free-future/big-data-can-be-used-violate-civil-rights-laws-and-ftc-agrees, 2016.

D. Biddle. *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing*. Gower, 2005.

T. Calders and S. Verwer. Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery*, 2010.

Civil Rights Act. Civil Rights Act of 1964, Title VII, Equal Employment Opportunities, 1964.

Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and Attacking the Saddle Point Problem in High-dimensional Non-convex Optimization. In *NIPS*, 2014.

C. Dwork, M. Hardt, T. Pitassi, and O. Reingold. Fairness Through Awareness. In *ITCSC*, 2012.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, 2015.

G. Goh, A. Cotter, M. Gupta, and M. Friedlander. Satisfying Real-world Goals with Dataset Constraints. In *NIPS*, 2016.

F. Kamiran and T. Calders. Classifying without Discriminating. In *IC4*, 2009.

F. Kamiran and T. Calders. Classification with No Discrimination by Preferential Sampling. In *BENELEARN*, 2010.

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware Classifier with Prejudice Remover Regularizer. In *PADM*, 2011.

B. T. Luong, S. Ruggieri, and F. Turini. kNN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *KDD*, 2011.

C. Muñoz, M. Smith, and D. Patil. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. *Executive Office of the President. The White House.*, 2016.

D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware Data Mining. In *KDD*, 2008.

J. Podesta, P. Pritzker, E. Moniz, J. Holdren, and J. Zients. Big Data: Seizing Opportunities, Preserving Values. *Executive Office of the President. The White House.*, 2014.

Ricci vs. DeStefano. U.S. Supreme Court, 2009.

A. Romei and S. Ruggieri. A Multidisciplinary Survey on Discrimination Analysis. *KER*, 2014.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

L. Sweeney. Discrimination in Online Ad Delivery. *ACM Queue*, 2013.

M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*, 2017.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning Fair Representations. In *ICML*, 2013.