Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification

Emmanouil Krasanakis
CERTH-ITI, Thessaloniki, Greece
maniospas@iti.gr

Eleftherios Spyromitros-Xioufis
CERTH-ITI, Thessaloniki, Greece
espyromi@iti.gr

Symeon Papadopoulos
CERTH-ITI, Thessaloniki, Greece
papadop@iti.gr

Yiannis Kompatsiaris
CERTH-ITI, Thessaloniki, Greece
ikom@iti.gr

ABSTRACT

Machine learning bias and fairness have recently emerged as key issues due to the pervasive deployment of data-driven decision making in a variety of sectors and services. It has often been argued that unfair classifications can be attributed to bias in training data, but previous attempts to "repair" training data have led to limited success. To circumvent shortcomings prevalent in data repairing approaches, such as those that weight training samples of the sensitive group (e.g. gender, race, financial status) based on their misclassification error, we present a process that iteratively adapts training sample weights with a theoretically grounded model. This model addresses different kinds of bias to better achieve fairness objectives, such as trade-offs between accuracy and disparate impact elimination or disparate mistreatment elimination. We show that, compared to previous fairness-aware approaches, our methodology achieves better or similar trades-offs between accuracy and unfairness mitigation on real-world and synthetic datasets.

CCS CONCEPTS

• **Computing methodologies** → **Cost-sensitive learning**; • **Theory of computation** → *Boosting*; • **Applied computing** → Law;

ACM Reference Format:
Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3178876.3186133

1 INTRODUCTION

As machine learning systems are currently deployed in an ever-growing number of services that affect people's lives, fairness concerns have become increasingly important. Such concerns are well justified, since automated decision-making systems can be biased against sensitive groups, if not properly constrained. For example, training a logistic regression classifier on the ProPublica COMPAS dataset of crime recidivism [33] yields differences between black and white defendants that amount to 17% for false positive and 25% for false negative rates. Hence, it is understandable why legal measures are in place to explicitly protect the right of minorities to not be subjected to different policies [1].

Fairness concern formulations compare aspects of a classifier between sensitive and non-sensitive groups (see Subsection 2.1). The measure of choice often depends on the respective legal setting, as well as on whether the ground truth is biased. For example, if the ground truth is historically unbiased, it is preferable to mitigate misclassification differences between sensitive and non-sensitive groups [44].

Researchers have previously recognized that classification bias is often caused by data rather than classifiers [13, 31]. For example, Kamishima et al. [29] categorize sources of unfair labeling as prejudice stemming from correlations between features and a sensitive attribute, underestimation due to inadequate convergence of the training algorithm and negative legacy of (historical) human biases in labeling training data. Hence, it has often been argued that we should look for ways to remove bias from training data (instead of constraining the training process), either through massaging the training labels [24] or reweighting training samples according to an estimated probability that they belong to a sensitive group [25].

Methods based on removing bias from training data usually fail to perform on par with the state-of-the-art (e.g. covariance-based models by Zafar et al. [41, 44]). However, we argue that this happens not due to an inherent inability to treat datasets, but rather due to methodological deficiencies in previous approaches. In Subsection 2.4 we discuss some of the most common deficiencies, such as preprocessing limitations, heuristic statistical models and inability to justify all types of fairness-aware edits. If those deficiencies are appropriately handled, we expect dataset editing methods to perform on the same level or even better than state-of-the-art, since they directly work on the source of bias instead of its outcome.

In this paper we propose an adaptive sensitive reweighting mechanism and a weight estimation model that do not suffer from these shortcomings. Our approach assumes that there exists an (unobservable) underlying set of class labels corresponding to training samples that, if predicted, would yield unbiased classification with respect to a fairness objective. It then searches for sample weights that make weighted training on the original dataset also train towards those labels, without explicitly observing them. To obtain those weights our approach employs a non-linear probability inference model, which we call *CULEP*, standing for Convex Underlying Label Error Perturbation. This model can be trained to convert

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04.
https://doi.org/10.1145/3178876.3186133

classification error to a probability that the estimated labels approach the desired underlying labels. We then use it to infer training weights based on classifier outputs and iteratively retrain the classifier on these new weights.

This bias mitigation mechanism encapsulates both fairness- and classifier-related information and thus allows a more precise stochastic analysis. Furthermore, it avoids concerns that arise from label editing approaches, since training is still conducted on the original labels. Finally, in Subsection 4.4 we explain that different CULEP parameters can help achieve different fairness goals, such as obtaining designated trade-offs between accuracy and various fairness metrics, such as those outlined in Subsection 2.2.

The novelty of our approach lies in the ability of CULEP to help an iterative reweighting process recognize sources of bias and diminish their impact without affecting features or labels. This way, the classification model is trained on *original* (possibly biased) dataset labels while still achieving designated fairness goals.[1] The CULEP model improves previous reweighting mechanisms with regards to estimating compliance between estimated and underlying labels and can be used to mitigate various types of unfairness.

2 BACKGROUND

Throughout this work, we consider binary classifiers that produce label estimations $\hat{y}_i \in \{0, 1\}$ for samples i of features x_i and labels $y_i \in \{0, 1\}$. A certain group of samples $\mathcal{S}$ is recognized as *sensitive* compared to its non-sensitive complement $\mathcal{S}'$ based on a sensitive real-world attribute, such as gender, race or financial status. Bias arises when a statistical property for the distribution of $\{\hat{y}_i, i \in \mathcal{S}\}$ is different for the distribution of $\{\hat{y}_i, i \in \mathcal{S}'\}$. Fairness-aware classification methods attempt to mitigate such differences.

2.1 Types of Unfairness

As outlined by Zafar et al. [44], classification unfairness is often expressed through the notions of disparate treatment, disparate impact and disparate mistreatment. Fairness objectives aim to eliminate these types of unfairness.

Disparate treatment elimination reflects the ability of a trained classifier to yield the same outputs $\hat{y}_i$ for features x_i regardless of whether the sample belongs to the sensitive group $\mathcal{S}$ or not:

$$P(\hat{y}_i|x_i, i \in \mathcal{S}) = P(\hat{y}_i|x_i) \tag{1}$$

Effectively, this fairness objective requires samples with similar features to be similarly classified. For example, if gender is a sensitive attribute for a classifier, males and females with otherwise similar features should be assigned to the same class under the principle of disparate treatment elimination.

A simple way to avoid disparate treatment is refraining from using information about the sensitive group for classification. This avoids discrimination or reverse discrimination [38], but the accuracy cost can sometimes be too high [28].

Disparate impact elimination reflects the ability of a classifier to achieve statistical parity [25, 27, 28], i.e. assign the same portion of users to a class for sensitive and non-sensitive groups:

$$P(\hat{y}_i = 1|i \in \mathcal{S}) = P(\hat{y}_i = 1|i \notin \mathcal{S}) \tag{2}$$

For example, if financial status is a sensitive attribute for term deposit predictions [37], disparate impact elimination would ensure that the portion of positive predictions is the same between low-income and high-income clients.

Disparate mistreatment elimination reflects the ability of a classifier to achieve equal misclassification rates across *sound ground truth labels* (i.e. not suffering from dataset construction problems, such as historical biases) [41, 44]. For example if race is a sensitive attribute for prediction of criminal behavior [33], disparate mistreatment elimination would ensure the same error rate between white and non-white defendants.

Recent works [7, 31] have shown that it is impossible to simultaneously satisfy all notions of disparate mistreatment elimination, unless the classifier is 100% accurate. More commonly adopted are the disparate mistreatment elimination constraints of equal false positive rates (FPR) and equal false negative rates (FNR):

$$P(\hat{y}_i \neq y_i|y_i = 1, i \in \mathcal{S}) = P(\hat{y}_i \neq y_i|y_i = 1, i \notin \mathcal{S}) \tag{3a}$$

$$P(\hat{y}_i \neq y_i|y_i = 0, i \in \mathcal{S}) = P(\hat{y}_i \neq y_i|y_i = 0, i \notin \mathcal{S}) \tag{3b}$$

2.2 Metrics

Following earlier fairness-aware approaches, in this work we measure classifier performance using accuracy, i.e. the proportion of correctly classified samples, disparate impact using the p% rule and disparate mistreatment using the difference between sensitive and non-sensitive FPR and FNR.

The p% rule [2] is an empirical rule which does not allow sensitive group identification to be lower than a set percentage of non-sensitive group identification:

$$pRule = \min\left\{\frac{P(\hat{y}_i = 1|i \in \mathcal{S})}{P(\hat{y}_i = 1|i \notin \mathcal{S})}, \frac{P(\hat{y}_i = 1|i \notin \mathcal{S})}{P(\hat{y}_i = 1|i \in \mathcal{S})}\right\} \tag{4}$$

This metric is correlated to the Calders-Verwer measure [5] which calculates the disparity between those two percentages. Since these two measures share the same optimal point, we prefer reporting the *pRule*, for which there exists a set legal context. More specifically, the Uniform Guidelines on Employee Selection Procedures require at least 80% rule adherence [2].

To measure disparate mistreatment, it is common to measure how deviation from set goals differs between the sensitive and non-sensitive group. In alignment with the common disparate mistreatment elimination conditions outlined in Eq. 3, we employ the following measures of disparate mistreatment:

$$D_{FPR} = P(\hat{y}_i \neq y_i|y_i = 1, i \in \mathcal{S}) - P(\hat{y}_i \neq y_i|y_i = 1, i \notin \mathcal{S}) \tag{5a}$$

$$D_{FNR} = P(\hat{y}_i \neq y_i|y_i = 0, i \in \mathcal{S}) - P(\hat{y}_i \neq y_i|y_i = 0, i \notin \mathcal{S}) \tag{5b}$$

To report the overall disparate mistreatment, we combine those two metrics into the quantity:

$$|D_{FPR}| + |D_{FNR}| \tag{5c}$$

2.3 Previous Work

Works aiming to reduce classification unfairness can be categorized in the following approaches: a) preprocessing training data, b) training under fairness constraints, c) attempts to 'fix' posteriors.

[1] Learning training weights is hardly a new concept in machine learning, but usually weights are learned to help boost weak learner accuracy [40] and only seldom to satisfy other training objectives [16, 21, 34].

Approaches based on training data preprocessing aim to remove disparate impact from training data, under the assumption that the disparate impact of the trained classifier follows the disparate impact of training data. Such approaches include massaging the dataset [3, 13, 16, 24–26] by changing class labels that are identified as mislabeled due to bias and reweighting (usually heuristically) training samples so that more importance is placed on sensitive ones [3, 11, 25]. Concerning massaging techniques, it must be noted that altering labels for training, even under bias concerns, can result in legal implications [1].

Approaches training under fairness constraints select a disparate impact or mistreatment metric and attempt to properly adjust the training rules, either via editing the rules themselves [4, 44] (e.g. by inserting an appropriate regularization term towards fairness) or by introducing appropriate linear program constraints [6, 11, 18, 39, 42–44] that reflect the desired optimization goals.

Finally, certain approaches attempt to edit posteriors in a way that satisfies fairness constraints [9, 10, 12, 20]. Such strategies are typically centered around some form of group-based thresholding. It must be noted that such systems require information about the sensitive group to make an appropriate decision. Although Hardt et al. [20] argue that privacy concerns can be alleviated by remotely obtaining the different decision-making rules between sensitive and non-sensitive groups and locally applying the appropriate rule, such practices may still be inapplicable under certain legal settings, since they introduce disparate treatment.

2.4 Discussion on Dataset Editing Deficiencies

In this subsection we discuss three common shortcomings across previous dataset editing fairness-aware mechanisms.

Limitations of preprocessing. Dataset editing approaches are commonly formulated by defining types of bias in the training data and then trying to statistically eliminate them. This process is indeed suitable for mitigating simple dataset-related biases, but fails to take into account more intricate sources of unfairness. For example, there may exist weaker feature correlations (e.g. through a chain of correlation of unobserved features, which may require external explanatory attributes to identify [4]) that cause bias against only a subset of the sensitive group. Furthermore, certain data could cause biases to only certain types of classifiers. For example, linear classifiers may have difficulty eliminating non-linear types of bias. Since biases emerge through systems of high complexity, which often tend to exhibit non-linear behavior, it is difficult to identify them through simple stochastic analysis and develop specific elimination strategies using only training data. Instead, it could be more informative to directly observe the effect of biases on the classifier and suitably perform adjustments while training. Following this line of thought, in this work, we propose an adaptive scheme, which iteratively adapts training data until stable behavior is achieved with respect to the classifier trained on this data.

Heuristic statistical models. Another shortcoming of dataset editing approaches is the introduction of ad hoc assumptions on the nature of unfairness, the most prominent one being that classifier bias closely follows the bias of training data. Although there often exists a high degree of correlation between the two, other structural difficulties may cause inadequate bias elimination or introduction of inverse bias, even if statistically optimal methods are employed to remove dataset bias. As a result, statistical models often arrive at a minimum condition that guarantees correct but not necessarily full treatment of training bias. For example, it is common practice to assume that higher prediction errors of robust classifiers indicate mislabeled data [16], but we show (see Subsection 4.2) that this -otherwise intuitive- assumption discards cases where classification error relates differently to mislabeling for sensitive and non-sensitive groups. In Subsection 4.3 we propose a statistical model that takes such differences into account and can be trained to satisfy various fairness objectives.

Inability to justify disparate mistreatment elimination. Disparate mistreatment is an emerging fairness concern that is attributed to difficulties in reaching similar misclassification rates between groups rather than direct dataset biases. Since we cannot attribute such concerns to biased data, it is difficult to justify a disparate mistreatment elimination method which attempts to treat the dataset. Furthermore, since disparate mistreatment is not necessarily caused by disparate impact, constructing datasets unbiased with respect to disparate impact does not treat disparate mistreatment. In other words, the relation between dataset bias and disparate mistreatment remains unclear to date. This detriment is more apparent when trying to develop massaging approaches that edit dataset labels with a target other than disparate impact elimination; to the authors' knowledge, there exists no clear (ethical or legal) justification to develop disparate mistreatment elimination procedures in a label editing process, as it cannot be attributed to any previously proposed source of dataset bias. For example, even in the well-formulated partial dataset repair mechanism of Feldman et al. [13], it becomes impossible to legally justify further label editing to remove disparate mistreatment, since dataset-related bias has already been treated. In this work, we try to bypass such limitations through a reweighting scheme, thus discovering sample weights that train towards desired objectives instead of editing training labels towards the same objectives.

3 ADAPTIVE SENSITIVE REWEIGHTING

3.1 Training Objective

Our analysis is conducted on a binary probabilistic classifier, which produces probability estimates $\hat{P}(Y = y_i) = 1 - \hat{P}(Y \neq y_i)$ for samples i (with features x_i) and each class label $Y \in \{0, 1\}$. Such a classifier estimates class labels as:

$$\hat{y_i} = \underset{Y \in \{0,1\}}{\operatorname{argmax}} \hat{P}(Y = y_i) = \underset{Y \in \{0,1\}}{\operatorname{argmin}} \hat{P}(Y \neq y_i) \tag{6}$$

For ease of understanding, we prefer referring to the estimated label error $\hat{P}(Y \neq y_i)$, since for a well-calibrated classifier, $\hat{P}(\hat{y_i} \neq y_i)$ approaches the misclassification error $P(\hat{y_i} \neq y_i)$, which is usually the desired minimization target of the learning process.

As previously described, this work assumes that unfairness is often caused by skewed group and label distributions in the dataset. However, the available ground truth may not always suffer from biases but yield disparate mistreatment due to other reasons, such as correlations between the sensitive group and certain attributes. To avoid confusion, we propose a common formulation for different fairness goals on a classifier. For training samples i, features x_i

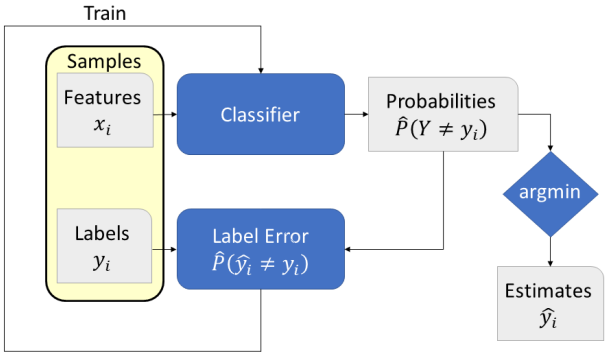

Figure 1: Probabilistic classifier training.

and class labels y_i, there exist underlying (i.e. unobservable) class labels $\tilde{y}_i$ that yield estimated labels $\hat{y}_i$ which conform to designated fairness and accuracy trade-offs.

In this setting, training goals are twofold: a) make the classifier yield accurate predictions, i.e. minimize $\hat{P}(\hat{y}_i \neq y_i)$ and b) make classifier predictions approach the underlying labels, i.e. minimize $\hat{P}(\hat{y}_i \neq \tilde{y}_i)$.[2] Obviously, there is difficulty in simultaneously training towards both of these objectives when data labels and underlying labels do not coincide. Training towards data labels could be achieved through the scheme demonstrated in Fig. 1 and training towards underlying labels could be achieved through the scheme demonstrated in Fig. 2.

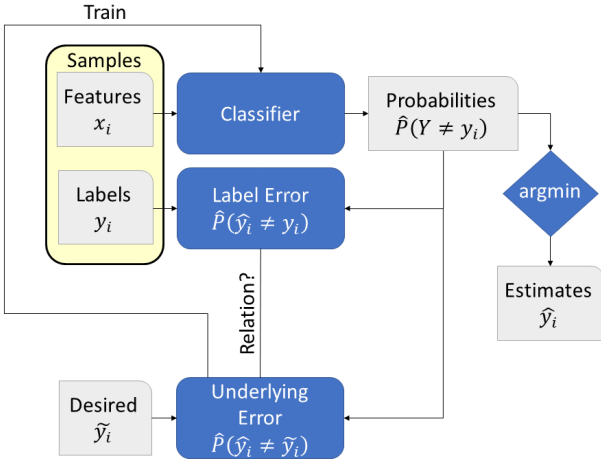

Figure 2: Directly training on observable desired labels. This can be ethically or legally questionable.

Furthermore, estimating the underlying labels and directly using them for training can be argued to be an act of data falsification under certain legal settings. Therefore, not only should training be conducted on the original data labels, it is also desirable to fully abstain from any observation of underlying labels. In this respect, the scheme demonstrated in Fig. 2 is inadequate.

[2]Probability maximization is equivalent to loss minimization: for the loss $\mathcal{L}_i$ calculated for sample i we can formulate $\mathcal{L}_i \propto \hat{P}(\hat{y}_i \neq y_i)$. We prefer a probabilistic formulation, since it allows us to infer conditional relations in a theoretically sound manner.

To solve these contradictions, we propose selecting weights w_i for training samples i that make weighted training on data labels equivalent to unweighted training on underlying labels. This way, we can focus on estimating weights w_i that help achieve designated fairness objectives without observing underlying labels. In other words, we try to minimize both weighted error on observed labels as well as the distance between weighted observed labels and unweighted underlying labels:

$$\min \sum_i w_i \hat{P}(\hat{y}_i \neq y_i)$$

$$\min \sum_i \left(w_i \hat{P}(\hat{y}_i \neq y_i) - \hat{P}(\hat{y}_i \neq \tilde{y}_i) \right)^2$$

For simplification purposes, in this paper we set the second minimization goal to 0 and attempt to analytically derive the weights w_i rather than tuning towards them with a gradient-based method, as per Eq. 7. In future work, training towards minimizing differences between underlying and weighted estimation could be conducted in place of analytical calculation, so as to make convergence more robust against noise.

$$\min \sum_i w_i \hat{P}(\hat{y}_i \neq y_i) \tag{7a}$$

$$w_i \hat{P}(\hat{y}_i \neq y_i) = \hat{P}(\hat{y}_i \neq \tilde{y}_i) \, \forall i \tag{7b}$$

Contrary to previous works on treating dataset bias, we assume that $\hat{P}(\hat{y}_i \neq \tilde{y}_i)$ cannot be estimated through simple sample- or group-specific dependencies. Instead, in Section 4 we propose a model which, in addition to knowledge of whether samples i belong to the sensitive group, employs conditional probabilities to make more informed estimations based on $\hat{P}(\hat{y}_i \neq y_i)$. Such a model allows the classifier to satisfy all previously outlined goals, namely estimating the underlying labels while training on an appropriately weighted original training label objective. This is more clearly demonstrated in Fig. 3.

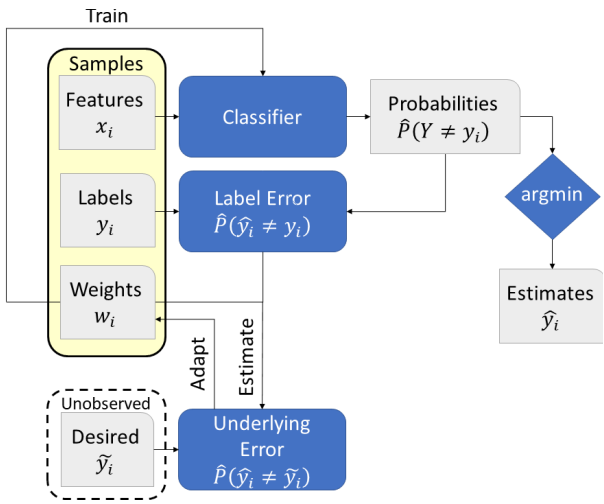

Figure 3: Training on unobservable desired labels.

Such a process shifts the focus of the training scheme to discovering a probability estimation model $\hat{P}(\hat{y}_i \neq \tilde{y}_i)$ that can train

towards the desired goals rather than searching for the underlying labels themselves. This is a substantial improvement compared to heuristically defining label editing procedures.

3.2 Why Unobserved Underlying Labels?

Essentially, the training objectives in Eq. 7 are equivalent to training the classifier on underlying labels. However, if we adequately estimate $\hat{P}(\hat{y}_i \neq \tilde{y}_i)$ using only $\hat{P}(\hat{y}_i \neq y_i)$, we can train for those labels using *only* the original training labels y_i. Previous works (e.g. the original massaging approach of Calders et al. [3]) try to infer and directly utilize underlying labels. However, we refrain from doing so, as not directly observing those labels yields three significant advantages.

Firstly, the classifier cannot be accused of being trained on falsified data. This practice can be ethically or legally questionable, however well-intended the 'falsification' as a form of label editing is. Instead, the classifier is trained over the labels y_i for weights that help it achieve its target objective, which is a widely accepted process in machine learning.

Secondly, we can select models for the estimation $\hat{P}(\hat{y}_i \neq \tilde{y}_i)$ that allow training towards objectives that could not be formulated as deficiencies in training data. In fact, training objectives include unbiased underlying label discovery instead of training data label discovery. Since discovered underlying labels are not directly utilized in training, we can then select probability estimation models that train towards objectives other than simple disparate impact, such as disparate mistreatment or fairness and accuracy trade-offs.

Thirdly, there is no need to introduce massaging-like heuristics about how relabeling should be distributed across classes and/or groups. This way, the underlying label *distribution* becomes more important than identifying which sample labels are biased. This property is important, as the relation between data and certain notions of unfairness is not yet clear, but there exist clear definitions on whether a label distribution adheres to a notion of fairness.

3.3 Training Algorithm

To simultaneously adjust training weights alongside classifier training using Eq. 7, we adopt a classifier-agnostic iterative approach, in which we first fully train a classifier based on uniform weights and then appropriately readjust those weights, repeating these steps until convergence. This process is specified in Algorithm 1.

Algorithm 1 Adaptive Sensitive Reweighting

function REWEIGHT(classifier C, data $\mathcal{D}$, sensitive group $\mathcal{S}$)
  $w_i \leftarrow 1 \; \forall i \in \mathcal{D}$
  $w_{i,prev} \leftarrow 1 + \sqrt{\epsilon} \; \forall i \in \mathcal{D}$
  while $\sum_{i \in \mathcal{D}} (w_i - w_{i,prev})^2 \geq \epsilon$ **do**
    train C on samples $i = (x_i, y_i) \in \mathcal{D}$ and weights $\frac{w_i}{\sum_{j \in \mathcal{D}} w_j}$
    use C to obtain $\hat{P}(\hat{y}_i \neq y_i)$.
    estimate $\hat{P}(\hat{y}_i \neq \tilde{y}_i)$ using $\hat{P}(\hat{y}_i \neq y_i) \; \forall i \in \mathcal{D}$
    $w_{i,prev} \leftarrow w_i \; \forall i \in \mathcal{D}$
    $w_i \leftarrow P(\hat{y}_i \neq \tilde{y}_i)/P(\hat{y}_i \neq y_i) \; \forall i \in \mathcal{D}$ (see Section 4)
  return trained classifier C, $\{w_i\}$

As per our previous formulation, we directly set new weight estimations rather than partially editing existing ones. We do so under the assumption that the adaptation model fails to converge only if the underlying probability estimation model does not adequately model the desired underlying labels. Otherwise, as long as the estimator model is convex, locality is preserved and thus the updating process should eventually converge to points or tracks of optimal weights. We experimentally assert this behavior in Subsection 6.1.

Recent works [17, 44] have occasionally proposed similar iterative methods as baselines to compare themselves to. However, our work differs in that it employs an inferred rather than a heuristic model to produce bias-related probabilities (see Section 4).

4 UNDERLYING LABEL ERROR ESTIMATION

4.1 Motivation

In the previous section's methodology, it is important to accurately model the error $\hat{P}(\hat{y}_i \neq \tilde{y}_i)$ of classified labels $\hat{y}_i$ deviating from underlying labels $\tilde{y}_i$ as a function of the classifier error $\hat{P}(\hat{y}_i \neq y_i)$ for original training labels y_i.

To do so, we propose a model that performs convex perturbations of classifier error to parameterize the deviation between original and underlying labels for sensitive and non-sensitive group samples. This model can then be used to estimate weights that help achieve various fairness objectives. In this section we explain why this process is superior to simpler error-based weighting (e.g. boosting) and why it can be fine-tuned towards the more common fairness objectives.

4.2 Weighting by Error is Inadequate

Previous attempts on sensitive reweighting (e.g. baselines employed by [17, 44]) propose that weights in Eq. 7a should be proportional to classifier error. However, by solving Eq. 7b this leads to:

$$w_i \approx \hat{P}(\hat{y}_i \neq y_i) \Leftrightarrow \hat{P}(\hat{y}_i \neq \tilde{y}_i) \approx \hat{P}^2(\hat{y}_i \neq y_i)$$

Furthermore, this assumption ignores differences in classifier error stemming from matching vs. non-matching dataset and underlying labels. The Bayes rule yields:

$$\begin{aligned} &\hat{P}(\hat{y}_i \neq \tilde{y}_i) \\ &\quad = \hat{P}(\hat{y}_i \neq y_i | y_i = \tilde{y}_i)\hat{P}(y_i = \tilde{y}_i) \\ &\quad + \hat{P}(\hat{y}_i \neq y_i | y_i \neq \tilde{y}_i)\hat{P}(y_i \neq \tilde{y}_i) \end{aligned} \tag{8}$$

Therefore, since $\hat{P}(y_i = \tilde{y}_i) + \hat{P}(y_i \neq \tilde{y}_i) = 1$, the proposed condition can always hold true only if:

$$\hat{P}(\hat{y}_i \neq y_i | y_i = \tilde{y}_i) \approx \hat{P}(\hat{y}_i \neq y_i | y_i \neq \tilde{y}_i) \approx \hat{P}(\hat{y}_i \neq y_i)$$

However, the above conditions are impossible to uphold for every dataset, since any sequence of datasets on which misclassification progressively becomes independent of desired underlying labels (e.g. progressively becomes unbiased) converges to the contradictory $\hat{P}(\hat{y}_i \neq y_i | y_i = \tilde{y}_i) \approx \hat{P}(\hat{y}_i \neq y_i | y_i \neq \tilde{y}_i) \approx \hat{P}(\hat{y}_i \neq y_i) \neq \hat{P}^2(\hat{y}_i \neq y_i)$. In other words, the previously proposed heuristic cannot always be met with success in removing bias.

4.3 Convex Underlying Label Error Perturbation (CULEP) Model

In this work, we recognize that conditional classifier error can be different when underlying labels coincide with original labels compared to when they do not; if classifier error would be overestimated compared to its estimation under the condition that original and underlying labels coincide, then it would be underestimated under the condition that they do not coincide and conversely. In other words:

$$\big(\hat{P}(\hat{y}_i \neq y_i | y_i = \tilde{y}_i) - \hat{P}(\hat{y}_i = y_i)\big)\big(\hat{P}(\hat{y}_i \neq y_i | y_i \neq \tilde{y}_i) - \hat{P}(\hat{y}_i \neq y_i)\big) < 0$$

To satisfy this property, we propose estimating those conditional probabilities by perturbing classifier error of training samples i. To do so, we multiply it with values obtained through a non-decreasing convex function $L_{\beta_i}(p_i) \geq 0$, $L_{\beta_i}(0) = 1$ of perturbation parameters $p_i \in [-1, 1]$, whose Lipschitz constant is proportional to β_i.[3]

Without loss of generality, we model perturbation parameters as $|p_i| = \hat{P}(\hat{y}_i \neq y_i)$, where their signs depend on whether conditional probabilities are overestimated or underestimated. Since probability spaces are continuous, whether to overestimate or underestimate original and underlying label coincidence during perturbations should be maintained throughout training samples. Adopting the $\pm, \mp$ notation[4] the above can be written as:

$$\hat{P}(\hat{y}_i \neq y_i | y_i = \tilde{y}_i) = L_{\beta_i}\big(\pm \hat{P}(\hat{y}_i \neq y_i)\big)\hat{P}(\hat{y}_i \neq y_i)$$
$$\hat{P}(\hat{y}_i \neq y_i | y_i \neq \tilde{y}_i) = L_{\beta_i}\big(\mp \hat{P}(\hat{y}_i \neq y_i)\big)\hat{P}(\hat{y}_i \neq y_i)$$

The sensitive group $\mathcal{S}$ and the non-sensitive group $\mathcal{S}'$ can adhere to different misclassification biases, which skew error in different ways (e.g. the coefficient of variation for classifier error for females mislabeled as males may be different for males mislabeled as females). Hence, we select different Lipschitz constants between those groups to produce different perturbations:

$$\beta_i = \left\{ \begin{array}{ll} \beta_{\mathcal{S}} & \text{if } i \in \mathcal{S} \\ \beta_{\mathcal{S}'} & \text{if } i \notin \mathcal{S} \end{array} \right\}$$

Finally, conditional probability estimations require the probability of biased or inadequate labeling. These probabilities may differ between the sensitive and non-sensitive groups (e.g. dataset construction may have been impartial between males and only biased against females) and can be affected by many unknown social- and dataset-related parameters. However, as long as these parameters remain approximately constant during dataset creation (e.g. because all data were gathered from the same regions during the same time period), their cumulative effect also remains approximately constant. Hence, data mislabeling would occur with a fixed probability, depending on whether samples belong to the sensitive group and can be modeled as two Bernoulli processes, one for sensitive group samples with mean value $q_{\mathcal{S}}$ and another for non-sensitive group samples with mean value $q_{\mathcal{S}'}$:

$$\hat{P}(y_i \neq \tilde{y}_i) = q_i = \left\{ \begin{array}{ll} q_{\mathcal{S}} & \text{if } i \in \mathcal{S} \\ q_{\mathcal{S}'} & \text{if } i \notin \mathcal{S} \end{array} \right\}$$

[3] If the derivative of a function exists, its Lipschitz constant coincides with the derivative's supremum. Convex functions, such as $exp(\beta_i p)$, are Lipschitz-continuous in bounded sets [22].

[4] $\pm$ represents either the positive or the negative sign and $\mp$ its opposite sign.

Substituting the above in Eq. 7b we obtain:

$$\begin{aligned} w_i \hat{P}(\hat{y}_i \neq y_i) &= \hat{P}(\hat{y}_i \neq \tilde{y}_i) \\ \Leftrightarrow w_i \hat{P}(\hat{y}_i \neq y_i) &= L_{\beta_i}\big(\pm \hat{P}(\hat{y}_i \neq y_i)\big)\hat{P}(\hat{y}_i \neq y_i) q_i \\ &\quad + L_{\beta_i}\big(\mp \hat{P}(\hat{y}_i \neq y_i)\big)\hat{P}(\hat{y}_i \neq y_i)(1 - q_i) \end{aligned}$$

This Convex Underlying Label Error Perturbation (CULEP) model obtained through the previous propositions can be rewritten as:

$$\boxed{\begin{array}{l} w_i = \alpha_i L_{\beta_i}\big(\hat{P}(\hat{y}_i \neq y_i)\big) + (1 - \alpha_i) L_{\beta_i}\big(-\hat{P}(\hat{y}_i \neq y_i)\big) \\ \beta_i = \left\{ \begin{array}{ll} \beta_{\mathcal{S}} & \text{if } i \in \mathcal{S} \\ \beta_{\mathcal{S}'} & \text{if } i \notin \mathcal{S} \end{array} \right\} \geq 0 \quad \alpha_i = \left\{ \begin{array}{ll} \alpha_{\mathcal{S}} & \text{if } i \in \mathcal{S} \\ \alpha_{\mathcal{S}'} & \text{if } i \notin \mathcal{S} \end{array} \right\} \in [0, 1] \end{array}} \quad (10)$$

For each selection of $(q_i, \pm)$, parameters α_i can be calculated as $\alpha_i = q_i$ or $\alpha_i = 1 - q_i$ depending on the sign of $\pm$. Therefore, when tuning Eq. 10, it suffices to search only for values of α_i instead of both the values of q_i and the sign of $\pm$.

4.4 Achieving Fairness with the CULEP Model

In this subsection, we discuss how the CULEP model allows us to select parameters in Eq. 10 such that we can train towards accuracy, disparate impact elimination and disparate mistreatment elimination objectives. As a result, it is possible to tune those parameters (see Subsection 5.4) to satisfy various such objectives or trade-offs between them.

Accuracy objectives. Training towards maximal accuracy of the classification model is achieved when all training weights are equal, i.e. $w_i = 1 \, \forall i \Leftrightarrow \beta_{\mathcal{S}} = \beta_{\mathcal{S}'} = 0$.

Disparate mistreatment objectives. As $\alpha_i \to 1$ we obtain $w_i \to L_{\beta_i}\big(\hat{P}(\hat{y}_i \neq y_i)\big)$ and hence place higher importance on misclassified samples. Whereas as $\alpha_i \to 0$ we obtain $w_i \to L_{\beta_i}\big(-\hat{P}(\hat{y}_i \neq y_i)\big)$ and hence place higher importance on correctly classified samples. Therefore, $\alpha_i \in [0, 1]$ interpolate between the importance of correct vs. incorrect classification for each sample. As $\beta_i \to \infty$, these trade-offs dominate classifier training pertaining to respective samples.

Based on these observations, we recognize two cases of disparate mistreatment with respect to the signs of D_{FPR} and D_{FNR}:

a) $D_{FPR} D_{FNR} > 0$, i.e. false positives and false negatives are either both overestimated or both underestimated for the sensitive group. In this case, as $(\alpha_{\mathcal{S}}, \alpha_{\mathcal{S}'}) \to (0, 1)$ more importance is placed on sensitive compared to non-sensitive group sample misclassification. The opposite happens as $(\alpha_{\mathcal{S}}, \alpha_{\mathcal{S}'}) \to (1, 0)$. This means that $|D_{FPR}|$ and $|D_{FNR}|$ are reduced as values of α_i move towards one of those two antipodal points. Large enough $\beta_{\mathcal{S}}$ and/or $\beta_{\mathcal{S}'}$ can magnify this effect in a way that minimizes either of those metrics or trade-offs between them.

b) $D_{FPR} D_{FNR} < 0$, i.e. false positives and false negatives are not overestimated or underestimated simultaneously for the sensitive group. In this case, we obtain opposite increments to D_{FPR} and D_{FNR} as either $(\alpha_{\mathcal{S}}, \alpha_{\mathcal{S}'}) \to (0, 0)$ or $(\alpha_{\mathcal{S}}, \alpha_{\mathcal{S}'}) \to (1, 1)$. Similarly to before, for large enough $\beta_{\mathcal{S}}$ and/or $\beta_{\mathcal{S}'}$, $|D_{FPR}|$ and $|D_{FNR}|$ or a trade-off between them can be minimized as a_i move towards one of those two antipodal points.

Disparate impact objectives. Positive discovery is more sensitive either towards higher or lower misclassification weights for each

group. Hence, there exist parameters $\alpha_{\mathcal{S}}, \alpha_{\mathcal{S}'}$ that either increase or decrease positive discoveries. Therefore, there also exist large enough $\beta_{\mathcal{S}}, \beta_{\mathcal{S}'}$ that maximize the *pRule*.

Summarizing the above, we can see that the CULEP model introduces four degrees of freedom (one for each of its parameters), with regards to positive or negative importance of misclassification rates and the degree of this importance for sensitive and non-sensitive groups. Therefore, those parameters are able to place different importances on accuracy and mitigation of sensitive and non-sensitive group differences on quantities correlated with misclassification (e.g. disparate mistreatment metrics) or discovery (e.g. disparate impact metrics).

5 EXPERIMENTS SETUP

5.1 Datasets

To assert the validity of our approach, we experiment with two synthetic datasets suffering from disparate mistreatment previously proposed by Zafar et al. [41], as well as with three well-known real world datasets: the *Adult* income dataset [32], the *Bank* marketing dataset [37] from the UCI repository [35] and the ProPublica *COMPAS* dataset [33] of criminal recidivism.

The two synthetic datasets suffering from disparate mistreatment comprise 10,000 samples with 2 features, a binary sensitive label and a binary classification label. Their features are obtained through bivariate normal distributions, chosen so that their sensitive labels yield $D_{FPR}D_{FNR} < 0$ and opposite-sign $D_{FPR}D_{FNR} > 0$, respectively, for a logistic regression classifier. This way, we can explore the ability of our approach to handle the two different cases of disparate mistreatment recognized in Subsection 4.4. The synthetic dataset with opposite signs of disparate mistreatment between FPR and FNR, which we call *SynthOpp*, is constructed by sampling the following distributions 2,500 times each:

$$x_{i,y_i=1,i\notin\mathcal{S}} \sim N([2,0],[5,1;1,5])$$
$$x_{i,y_i=1,i\in\mathcal{S}} \sim N([2,3],[5,1;1,5])$$
$$x_{i,y_i=0,i\notin\mathcal{S}} \sim N([-1,-3],[5,1;1,5])$$
$$x_{i,y_i=0,i\in\mathcal{S}} \sim N([-1,0],[5,1;1,5])$$

The synthetic dataset with same signs of disparate mistreatment between FPR and FNR, which we call *SynthSame*, is constructed by sampling the following distributions, 2,500 times each:

$$x_{i,y_i=1,i\notin\mathcal{S}} \sim N([1,2],[5,2;2,5])$$
$$x_{i,y_i=1,i\in\mathcal{S}} \sim N([2,0],[10,1;1,4])$$
$$x_{i,y_i=0,i\notin\mathcal{S}} \sim N([0,-1],[7,1;1,7])$$
$$x_{i,y_i=0,i\in\mathcal{S}} \sim N([-5,0],[5,1;1,5])$$

The Adult dataset comprises 48,842 test samples with 14 features and a binary label indicating whether income is above 50K. For this dataset, we consider gender as the sensitive feature.

The Bank dataset comprises 41,188 samples with 20 features and a binary label, indicating whether a client has subscribed to a term deposit. For this dataset, ages less than 25 and more than 60 years are considered sensitive.

We select a subset of the COMPAS dataset previously used for fairness experiments [44], which comprises 6,150 samples with five features (age category, gender, race, priors count and charge degree) and a binary label indicating whether the defendant reoffended within two years. The race is considered as the sensitive attribute and, to make it binary, we follow earlier approaches in selecting only Black and White individuals. It must be noted that the selected features aim to facilitate fairness experiments comparable to previous approaches rather than high predictive accuracy.

5.2 Fairness Objectives

Fairness-aware classifiers are usually able to train towards mitigating various fairness metrics. At the same time, they need to preserve the accuracy (*acc*) of the base classification model as much as possible. Otherwise, it could be possible for their outputs to be misleading.

When a classifier targets multiple objectives [19], it can employ either linear scalarization, where a linear trade-off is set between the objectives, or ϵ-constraints, which bound individual objectives. Since there usually exist legal bounds for disparate impact (e.g. the 80% rule) but not for mistreatment, it is easier to formulate disparate impact as an ϵ-constraint and disparate mistreatment as linear scalarization. However, Miglierina et al. [36] theoretically show the duality between those two types of objectives. Furthermore, it is easier to tune the parameters of Eq. 10 in a linear than in a constrained space. Therefore, we opt for relaxing training bounds by setting linear scalarization goals for all fairness objectives.

In particular, the Adult and Bank datasets are commonly considered to suffer from disparate impact and thus we train the CULEP model towards eliminating disparate treatment while preserving accuracy:

$$\max(acc + pRule)$$

On the other hand, the COMPAS and synthetic datasets are considered to contain sound ground truth and thus we place more emphasis on overall disparate mistreatment elimination, as previously discussed in Subsection 2.2. For our experiments, we consider accuracy to be equally important to each fairness constraint:

$$\max(2acc - |D_{FPR}| - |D_{FNR}|)$$

5.3 Validation

For Adult and Bank dataset experiments we perform a $70:30$ random split and for the COMPAS and synthetic dataset experiments we perform a $50:50$ random data split to obtain training and test data. These splits are used by previous works exploring those datasets and thus allow our results to be comparable across approaches. In both cases, we use the training set to tune the CULEP model on Algorithm 1 and then train the base classifier on the training set. We use the evaluation set only to calculate the accuracy and disparate impact and mistreatment elimination of the resulting classifier. For robustness, we again follow the validation methodology of previous approaches, which repeat this process 5 times and report the average measures across experiments.

We employ *logistic regression* without regularization as our base classifier of choice. To speed up training time we normalize numeric attributes in real-world datasets by dividing with their mean value. We encode nominal attributes using a one-hot scheme to convert them to binary arrays. Finally, CULEP conditional probabilities are modeled as Gaussian processes, which is a popular generic model

in stochastic analysis, as it often arises in all sorts of physical and theoretical systems:

$$L_{\beta_i}(p) = \exp(\beta_i p)$$

5.4 Training the CULEP Model

The CULEP model outlined in Eq. 10 is non-linear parametric model and thus tuning needs to be accurate. Furthermore, Algorithm 1, can exhibit non-smooth behavior, since a different number of adjustments can arise from different parameter selections. Hence, to optimize CULEP parameters, we employ the DIvided RECTangles (DIRECT) method [14, 15, 23], which is guaranteed to yield globally optimal solutions in Lipschitz-continuous objective spaces. Since $a_{\mathcal{S}}, a_{\mathcal{S}'}$ lie in $[0, 1]$ as either probabilities or complements of probabilities, $b_{\mathcal{S}}, b_{\mathcal{S}'}$ are non-negative constants and $\exp(\beta p)$ quickly converges to higher variation coefficients for larger β, we search for optimal parameters in the space $(a_{\mathcal{S}}, a_{\mathcal{S}'}, b_{\mathcal{S}}, b_{\mathcal{S}'}) \in [0, 1]^2 \times [0, 3]^2$. Each combination of parameters is evaluated with a full run of Algorithm 1 on the training set.

5.5 Compared Methods

Zafar et al. previously tested various fairness-aware approaches for disparate impact elimination on the Adult dataset [44] and disparate mistreatment elimination on the COMPAS dataset [41]. We compare our method with those they report to yield superior results to the rest. These 'best' methods also happen to use logistic regression. The methods compared in our experiments are:

- **ASR+CULEP**. Adaptive Sensitive Reweighting using the CULEP model as described throughout this paper, which can be used to mitigate disparate impact and mistreatment. Our implementation is available online.[5]
- **Covariance**. Models proposed by Zafar et al. [41, 44] employing covariance to approximate linear program constraints that mitigate disparate impact [41] and mistreatment [44].
- **Group Thresholding**. Model proposed by Hardt et al. [20] for disparate mistreatment elimination.
- **Regularizer**. Approach proposed by Kamishima et al. [28] to remove prejudice-related disparate impact. It suffers from disparate treatment, since it takes into account whether samples are sensitive during classification.

6 RESULTS

6.1 Exploring Convergence

In this subsection, we explore the convergence of Algorithm 1 towards optimal weights and the impact on the objective function. This way, we can get a general idea about convergence speed, as well as the effect of multiple iterations in our scheme.

To explore convergence after selecting CULEP model parameters, we measure the objective functions formulated in Subsection 5.2 for each dataset on training data. We also measure the root mean square weight edits on each iteration of Algorithm 1:

$$\sqrt{\frac{1}{N}\sum_i (w_i - w_{i,prev})^2}$$

where N is the number of training samples.

[5] https://github.com/MKLab-ITI/adaptive-fairness

As root mean square edits approach zero, Algorithm 1 approaches (locally) optimal weights. On the other hand, if root mean square edits approach a fixed constant, the adaptive scheme alternates between similar weights in a locally unstable way which is close to a global optimum. Hence, we can consider that weights converge to a stable state as long as weight edits approach a fixed value.

In Fig. 4 we can see that weight edits converge in very few iterations for the studied datasets. However, they do not stabilize immediately but need a small number of repetitions to converge to a fixed value. Furthermore, we can see that, after weights converge, the objective functions yield substantial improvements compared to the first iteration. These findings align with our hypothesis that single-step methodologies are insufficient to fully discover appropriate weights and that more iterations should be performed.

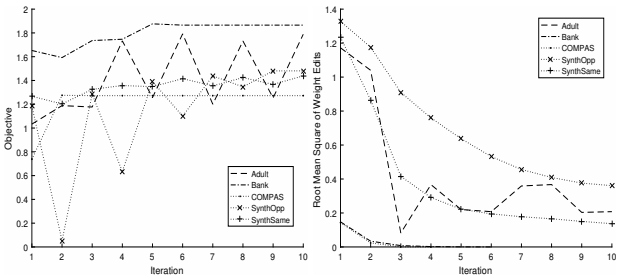

Figure 4: Objective and weight editing across datasets for each iteration in Algorithm 1 using the trained CULEP model to re-estimate underlying label error (disparate treatment was avoided for the synthetic datasets).

ASR retrains the base classifier only a few times (~ 5) before converging. Furthermore, DIRECT training achieves precise estimation up to the fifth decimal point (which empirically suffices) for the four CULEP parameters within at most $\lceil \log_2(3/0.000005) \rceil 2^4 = 320$ evaluations of ASR. Therefore, ASR+CULEP trains the base classifier at most $320 \cdot 5 = 1,600$ times. Although this computational cost could be prohibitive for more complex base classifiers, it scales linearly without further approximations and hence is suited to simpler classifiers, such as logistic regression.

6.2 Results for Disparate Mistreatment

Our experiments for disparate mistreatment attempt to explore mistreatment elimination both when disparate treatment is avoided and when it is not. In the first case, we do not include information about the sensitive group in the training and validation datasets, whereas in the second case we do. It must be noted that not all classifier and datasets can account for disparate treatment. For example, in the COMPAS dataset, removing the sensitive group feature, i.e. race, yields inadequate levels of prediction for the explored dataset, whereas group thresholding approaches inherently require information about the sensitive group. Although Hardt et al. [20] argue that thresholding can be performed without information disclosure (e.g. locally), avoiding disparate treatment may still be important in certain legal settings. Our ASR+CULEP model is compared to previous ones based on its ability to eliminate overall disparate mistreatment (i.e. minimize both $|D_{FPR}|$ and $|D_{FNR}|$).

	Disparate Treatment									Avoiding Disparate Treatment					
	COMPAS			SynthOpp			SynthSame			SynthOpp			SynthSame		
Fairness Approach	acc	D_{FPR}	D_{FNR}	acc	D_{FPR}	D_{FNR}	acc	D_{FPR}	D_{FNR}	acc	D_{FPR}	D_{FNR}	acc	D_{FPR}	D_{FNR}
None	66%	17%	-25%	78%	-16%	19%	80%	25%	14%	78%	-16%	19%	80%	25%	14%
ASR+CULEP $2acc - \|D_{FPR}\| - \|D_{FNR}\|$	65%	-1%	-1%	81%	0%	0%	77%	0%	-16%	77%	0%	-1%	75%	0%	-13%
Covariance [41]	66%	3%	-11%	80%	1%	2%	77%	14%	6%	75%	-1%	1%	69%	-1%	6%
Group Thresholding [20]	65%	-1%	-1%	79%	0%	-1%	67%	2%	0%	-	-	-	-	-	-

Table 1: Disparate mistreatment elimination for logistic regression on both $|D_{FPR}|$ and $|D_{FNR}|$ constraints for disparate treatment (i.e. the sensitive group is a feature) and avoiding disparate treatment.

In Table 1 we can see that, when information about the sensitive group is available, ASR+CULEP outperforms covariance-based constraints in eliminating disparate mistreatment and yields equally favorable results to group thresholding. On the COMPAS dataset it reduces overall mistreatment by 12% more in exchange for 1% accuracy compared to covariance-based methods and yields identical results to group thresholding. It also performs slightly better in all respects on the SynthOpp dataset and manages to maintain high accuracy on the SynthSame dataset while reducing overall mistreatment 4% more compared to covariance-based constraints.

Table 1 also shows that, when avoiding disparate treatment, ASR+CULEP produces better accuracy vs. overall mistreatment elimination trade-offs compared to covariance-based linear constraints. In particular, it yields slightly better (∼1%) overall disparate mistreatment elimination while better preserving accuracy on the SynthOpp dataset and trades 6% overall disparate mistreatment to gain 6% accuracy on the SynthSame dataset.

Considering all comparisons, ASR+CULEP is only inferior to other methods in eliminating mistreatment on the SynthSame dataset. However, this can be attributed to the choice of optimization goals while training the CULEP model. Indeed, in every instance where residual mistreatment is worse compared to other methods, significantly higher accuracy is retained to compensate.

6.3 Results for Disparate Impact

Table 2 demonstrates the ability of our methodology to eliminate disparate impact compared to state-of-the-art approaches. ASR+CULEP is able to achieve higher pRule for smaller accuracy trade-off than the best two previous approaches on the Adult dataset and yields better pRule than Covariance but worse than the Regularizer approach on the Bank dataset. Hence, it has merit, especially if disparate treatment is important (Regularizer employs disparate treatment to make results more fair). In this case, compared with Covariance, it attains 6% pRule gain for the same accuracy on the Adult dataset and 16% pRule gain for 2% accuracy loss on the Bank dataset.

It must be noted that smaller ASR+CULEP accuracies result from the fairness objective $acc+pRule$, which incentivizes small accuracy losses in favor of significant fairness improvements. However, this method is superior in that it allows higher margins in optimizing towards pRule and yield better trade-offs for this objective.

	Adult		Bank	
Fairness Approach	pRule	acc	pRule	acc
None	27%	85%	31%	91%
ASR+CULEP $acc + pRule$	100%	82%	99%	89%
Covariance [44]	94%	82%	83%	91%
Regularizer [28]	85%	83%	100%	91%

Table 2: Adult dataset disparate impact elimination for logistic regression. We compare our method to the highest pRule obtained by other methods.

7 CONCLUSIONS AND FUTURE WORK

In this work we presented an Adaptive Sensitive Reweighting (ASR) scheme that uses a convex model (CULEP) to estimate distributions of underlying labels with which to adapt weights. Our method can be applied on multiple types of fairness objectives and can also avoid disparate treatment. Experiments on logistic regression classifiers show that it performs similarly to covariance-based methods in trading-off accuracy and bias if disparate treatment is avoided. If we do not avoid disparate treatment and provide information about the sensitive group in evaluation data though, our approach performs better than these methods in trading-off disparate impact and mistreatment elimination for small accuracy losses and is comparable (even superior in some aspects) to methods specifically designed to take advantage of disparate treatment.

Our results indicate that there is merit in further developing non-heuristic dataset editing mechanisms as competent alternatives to existing fairness-aware approaches - i.e. such approaches are equally valid to existing ones.

For future work, we propose exploring ways to faster train or estimate the CULEP model parameters or adjust them during the training process, as well as developing methods with theoretically guaranteed convergence by training towards optimal sample weights rather than analytically deriving them. Furthermore, our methodology can be tested on more datasets, including multiclass and regression tasks. Finally, since recent works [8, 30] argue that stochastic methods can be inefficient in removing differences pertaining to certain sub-groups, further work should be conducted to examine whether the proposed CULEP model also suffers from the same limitations and, if so, extend it to multimodal distribution formulations, which can adequately model subgroups.

ACKNOWLEDGMENTS

This work was supported by the InVID and hackAIR projects under contract nr. 687786 and 688363 respectively, funded by the European Commission. The authors thank Nikolaos Nikolaou from the University of Manchester for his valuable feedback on this paper.

REFERENCES

[1] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. (2016).
[2] Dan Biddle. 2006. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd.
[3] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*. IEEE, 13–18.
[4] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. 2013. Controlling attribute effect in linear regression. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 71–80.
[5] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.
[6] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with Fairness Constraints. *arXiv preprint arXiv:1704.06840* (2017).
[7] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056* (2017).
[8] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. *arXiv preprint arXiv:1701.08230* (2017).
[9] Georges Dionne and Casey Rothschild. 2014. Economic effects of risk classification bans. *The Geneva Risk and Insurance Review* 39, 2 (2014), 184–221.
[10] Neil A Doherty, Anastasia V Kartasheva, and Richard D Phillips. 2012. Information effect of entry into credit ratings market: The case of insurers' ratings. *Journal of Financial Economics* 106, 2 (2012), 308–330.
[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226.
[12] Michael Feldman. 2015. Computational Fairness: Preventing Machine-Learned Discrimination. (2015).
[13] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.
[14] Daniel E Finkel. 2003. DIRECT optimization algorithm user guide. *Center for Research in Scientific Computation, North Carolina State University* 2 (2003).
[15] Daniel E Finkel and CT Kelley. 2004. Convergence analysis of the DIRECT algorithm. *Optimization Online* 14, 2 (2004), 1–10.
[16] Benjamin Fish, Jeremy Kun, and Adám D Lelkes. 2015. Fair boosting: a case study. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*.
[17] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 144–152.
[18] Kazuto Fukuchi and Jun Sakuma. 2015. Fairness-Aware Learning with Restriction of Universal Dependency using f-Divergences. *arXiv preprint arXiv:1506.07721* (2015).
[19] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. 2016. Satisfying Real-world Goals with Dataset Constraints. In *Advances in Neural Information Processing Systems*. 2415–2423.
[20] Moritz Hardt, Eric Price, Nati Srebro, and others. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. 3315–3323.
[21] Qinghua Hu, Pengfei Zhu, Yongbin Yang, and Daren Yu. 2011. Large-margin nearest neighbor classifiers via sample weight learning. *Neurocomputing* 74, 4 (2011), 656–660.
[22] Anatoli Iouditski and Yuri Nesterov. 2014. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv preprint arXiv:1401.1792* (2014).
[23] Donald R Jones, Cary D Perttunen, and Bruce E Stuckman. 1993. Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications* 79, 1 (1993), 157–181.
[24] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*. IEEE, 1–6.
[25] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.
[26] Faisal Kamiran, Toon Calders, and others. 2011. Handling conditional discrimination. In *Proc. of the 11th IEEE Int'l Conf. on Data Mining*.
[27] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*. IEEE, 869–874.
[28] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.
[29] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 643–650.
[30] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2017. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *arXiv preprint arXiv:1711.05144* (2017).
[31] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
[32] Ron Kohavi. 1996. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid.. In *KDD*, Vol. 96. Citeseer, 202–207.
[33] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. 2017. COMPAS dataset. (2017). https://github.com/propublica/compas-analysis
[34] Yuan Li, Chang Huang, and Ram Nevatia. 2009. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2953–2960.
[35] M. Lichman. 2013. UCI Machine Learning Repository. (2013). http://archive.ics.uci.edu/ml
[36] Enrico Miglierina and Elena Molho. 2002. Scalarization and stability in vector optimization. *Journal of Optimization Theory and Applications* 114, 3 (2002), 657–670.
[37] Sergio Moro, Raul Laureano, and Paulo Cortez. 2011. Using data mining for bank direct marketing: An application of the crisp-dm methodology. In *Proceedings of European Simulation and Modelling Conference-ESM'2011*. Eurosis, 117–121.
[38] Shelly L Peffer. 2009. Title VII and disparate-treatment discrimination versus disparate-impact discrimination: The Supreme Court's decision in Ricci v. DeStefano. *Review of Public Personnel Administration* 29, 4 (2009), 402–410.
[39] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 05 (2014), 582–638.
[40] Robert E Schapire. 2003. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*. Springer, 149–171.
[41] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1171–1180. DOI: http://dx.doi.org/10.1145/3038912.3052660
[42] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2015. Fairness constraints: A mechanism for fair classification. *stat* 1050 (2015), 19.
[43] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2015. Learning Fair Classifiers. *stat* 1050 (2015), 29.
[44] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Artificial Intelligence and Statistics*. 962–970.