Achieving Fairness at No Utility Cost via Data Reweighing with Influence

Peizhao Li [1] **Hongfu Liu** [1]

Abstract

With the fast development of algorithmic governance, fairness has become a compulsory property for machine learning models to suppress unintentional discrimination. In this paper, we focus on the pre-processing aspect for achieving fairness, and propose a data reweighing approach that only adjusts the weight for samples in the training phase. Different from most previous reweighing methods which usually assign a uniform weight for each (sub)group, we granularly model the influence of each training sample with regard to fairness-related quantity and predictive utility, and compute individual weights based on influence under the constraints from both fairness and utility. Experimental results reveal that previous methods achieve fairness at a non-negligible cost of utility, while as a significant advantage, our approach can empirically release the tradeoff and obtain cost-free fairness for equal opportunity. We demonstrate the cost-free fairness through vanilla classifiers and standard training processes, compared to baseline methods on multiple real-world tabular datasets. Code available at https://github.com/brandeis-machine-learning/influence-fairness.

1. Introduction

For artificial intelligence technology deployed in high-stakes applications like welfare distribution or school admission, it is essential to regulate algorithms and prevent unaware discrimination and unfairness in decision making (Barocas & Selbst, 2016; Goodman & Flaxman, 2017; Ferguson, 2017; Civil Rights Act, 1964). Even though general data-driven algorithms are not designed to be unfair, the outcomes can still violate the AI principle of equality unintentionally (Chouldechova, 2017). Typically learning from historically biased data, the learner can retain or even amplify the inherent bias if there is no proper constraint on data or algorithms. As a consequence, the decisions from these algorithms may disadvantage users in certain sensitive groups (*e.g.* female and African Americans), therefore raising societal concerns.

To mitigate unfairness algorithmically, solutions can be divided into three main categories: pre-processing, in-processing, and post-processing. Pre-processing approaches adjust the input data or training sample weights, and expect a vanilla learner can deliver fair results from the fair data transformation (Kamiran & Calders, 2012; Krasanakis et al., 2018; Calmon et al., 2017; Jiang & Nachum, 2020; Feldman et al., 2015; Yan et al., 2020; Zemel et al., 2013; Rastegarpanah et al., 2019; Chhabra et al., 2021). In-processing approaches insert fair constraints or penalties into the training pipeline, so the fair performance can be generalized to inference as achieved during training (Zhang et al., 2018; Agarwal et al., 2018; Zhao et al., 2019; Zafar et al., 2017; Jiang et al., 2020; Kearns et al., 2018; Goh et al., 2016; Li et al., 2020; 2021; Song et al., 2021). As a drawback, adding non-convex constraints or balancing the additional fair penalties with the primary training objective could rise optimization complexity, and sometimes incur instability (Cotter et al., 2019; Roth et al., 2017). Post-processing approaches calibrate the outcomes independently from a model (Hardt et al., 2016; Pleiss et al., 2017), but would result in sub-optimal solutions (Woodworth et al., 2017) and request sensitive attributes during the inference stage.

In this work, we advocate the pre-processing category since it directly diagnoses and corrects the source of bias, and can be easily adapted to existing data analytic pipelines. Pre-processing approaches can be further divided into two subcategories. The first one modifies original data or injects extra antidote data (Calmon et al., 2017; Feldman et al., 2015), but will expose to the risk of learning from falsification which could be legally questionable (Barocas & Selbst, 2016; Krasanakis et al., 2018). Conservatively, we consider sample reweighing, the second subcategory that purely adjusts the weights for samples in the training objective, and intend to achieve fairness through vanilla classifiers.

To prevent the model from disadvantaging a certain sensitive group and violating group-level fairness (Dwork et al., 2012; Hardt et al., 2016), most previous reweighing methods statis-

[1]Brandeis University. Correspondence to: Peizhao Li <peizhaoli@brandeis.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

tically identify the unprivileged groups from a heuristic or a learning process. Then, they assign an identical weight to all samples in a sensitive (sub)group, where the group membership is conditional on both sensitive attributes and the target label (Kamiran & Calders, 2012; Jiang & Nachum, 2020). The philosophy is to amplify the error from an underrepresented group in training, so optimization can equally update a model for different groups. Differently, we hypothesize the inherent bias in data can be traced to a biased and unconfident labeling function (Chen et al., 2018; Jiang & Nachum, 2020). The labeling might be unconsciously affected by sensitive attributes and therefore lead some training samples to be assigned with an improper label. By this means, some unqualified samples are assigned with positive labels, camouflaging themselves to be privileged, and correspondingly, some qualified samples are labeled as negative, inducing their groups to be unprivileged. Instead of equally treating every sample, we believe that an ideal pre-processing method is to downweight the mislabeled samples in training while keeping other good samples unchanged to preserve the predictive utility.

To this end, we propose a one-pass data reweighing method that granularly computes a weight for every sample in the training set. We use influence function (Cook & Weisberg, 1980; Hampel et al., 2011; Koh & Liang, 2017; Giordano et al., 2019) from robust statistics to estimate the effect of changing the weight of one sample without explicitly retraining the model. Specifically, we measure the sample influence in terms of both fairness and predictive utility, and theoretically prove that in a general case on influential approximation, a soft downweighting over some training samples can always enhance fairness while avoiding a cost in utility. We realize our findings through a proposed reweighing algorithm that estimates the individual weight through linear programming under both utility and fairness constraints. On multiple tabular datasets and in most cases, our empirical results achieve good group fairness at no utility cost compared to vanilla classifiers with original training data. We deem fairness at no utility cost as a significant advantage since it could help to popularize fair algorithms for extensive utility-driven products.

2. Characterizing Sample Influence

In this section, we introduce preliminaries on influence function (Cook & Weisberg, 1980; Hampel et al., 2011; Koh & Liang, 2017; Giordano et al., 2019). Influence function from robust statistics is to quantitatively measure the impact of an infinitesimal fraction of samples to an estimator. Consider a classifier with parameters $\theta \in \mathbb{R}^D$ mapping instances from input space $x \in \mathcal{X}$ to output space $y \in \mathcal{Y}$. The model is trained on a training set $\mathcal{T} = \{z_i = (x_i, y_i)\}_{i=1}^{N_\mathcal{T}}$ with some loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}$ through empirical risk minimization on the training set:

$$\hat{\theta}(\mathbf{1}) = \arg\min_{\theta} \sum_i \ell(z_i; \theta). \tag{1}$$

The all-one vector $\mathbf{1}$ indicates an equal assignment of a unit weight to every training sample. A reweighing of samples followed by a retraining can be expressed as

$$\hat{\theta}(\mathbf{1} - \mathbf{w}) = \arg\min_{\theta} \sum_i (1 - w_i)\ell(z_i; \theta), \tag{2}$$

where w_i denotes the deviation from a unit weight for z_i, and $\mathbf{w} \in \mathbb{R}^{N_\mathcal{T}}$. Note that here $(1 - w_i)$ is the weight of z_i, rather than w_i. A larger w_i indicates its less importance of z_i to the model training, and $w_i = 1$ is equivalents to an entire removal of z_i from the training set.

It is of high interest and value to know in a counterfactual how the model will change with regard to some typical measurements, *e.g.* fairness or utility, if there is a reweighing. The actual influence derived from a reweighing can be expressed as follows:

$$\mathcal{I}_f^*(\mathbf{w}) = f(\hat{\theta}(\mathbf{1} - \mathbf{w})) - f(\hat{\theta}(\mathbf{1})), \tag{3}$$

where the function $f : \mathbb{R}^D \to \mathbb{R}$ evaluates the quantity of interest. In the next section, we shall realize f with functions describing utility and fairness, and study the correlations between these two influences. To compute $\mathcal{I}^*$, one can always separately train the model two times with and without the reweighing. However, to reach an ideal value in f, retraining is a brute force approach and can be prohibitively expensive to find the optimal $\mathbf{w}$. To this end, a first-order approximation called influence function helps to estimate the actual influence $\mathcal{I}^*$ while getting rid of retraining (Cook & Weisberg, 1980; Hampel et al., 2011).

Influence function measures the effect of changing an infinitesimal weight from samples in $\mathbf{w}$, then linearly extrapolates to complete all of $\mathbf{w}$. It assumes ℓ to be twice-differentiable and strictly convex in θ, and f to be differentiable as well. These assumptions are mild and feasible for many pipelines involving classifiers like logistic regression to process tabular data. Having $\theta(\mathbf{1})$ and $\theta(\mathbf{1} - \mathbf{w})$ satisfied their first-order optimality conditions, and by taking a Taylor approximation, the actual influence $\mathcal{I}_f^*(\mathbf{w})$ can be approximated by an estimation

$$\begin{aligned} \mathcal{I}_f(\mathbf{w}) &= \nabla_\theta f(\hat{\theta}(\mathbf{1}))^\top \left[\frac{d}{dt} \theta(\mathbf{1} - t\mathbf{w})|_{t=0} \right] \\ &= \nabla_\theta f(\hat{\theta}(\mathbf{1}))^\top \mathbf{H}_{\hat{\theta}(\mathbf{1})}^{-1} \left[\sum_i w_i \nabla_\theta \ell(z_i; \hat{\theta}(\mathbf{1})) \right], \end{aligned} \tag{4}$$

where $\mathbf{H}_{\hat{\theta}(\mathbf{1})} = \sum_{i=1}^{N_\mathcal{T}} \nabla_\theta^2 \ell(z_i; \hat{\theta}(\mathbf{1}))$ is the Hessian matrix of ℓ, and the convexity ensures its invertibility. Influence

function has been empirically proofed to be valid for representing the actual influence obtained by model retraining when the above assumptions are satisfied or violated by a small degree (Koh & Liang, 2017). The additivity of $\mathcal{I}_f(\mathbf{w})$ regarding a set of $\mathbf{w}$ should follow the additivety of f. A negative value of $\mathcal{I}_f(\mathbf{w})$ instructs a reduction in f if the model $\hat{\theta}(\mathbf{1})$ is retrained with the weight $\mathbf{w}$ by Equation (2).

3. Fairness at No Utility Cost

We consider group fairness in this work. Group fairness articulates the equality of some statistics like predictive rate or true positive rate between certain groups. And here a group is specifically constructed based on some sensitive attributes like gender or race. Consider a binary classification problem $\mathcal{Y} = \{0, 1\}$, where predictions are affected by a binary sensitive attribute $a \in \{0, 1\}$. The sensitive attribute divides samples into a privileged group and an unprivileged group. An initial notion of group fairness is called Demographic Parity (Dwork et al., 2012), requesting equality on the rate of positive predictions:

$$\Pr(\hat{y} \mid a = 0) = \Pr(\hat{y} \mid a = 1). \tag{5}$$

Demographic Parity enforces a group fairness on the outcomes regardless of the gap in base rate, *i.e.* $\Pr(y \mid a = 0) \neq \Pr(y \mid a = 1)$. If sensitive groups have a gap in the base rate, there will always be a tradeoff between DP and utility for any classifier. To this end, another notion called Equal Opportunity (Hardt et al., 2016) has been raised to measure the equality on true positive rate:

$$\Pr(\hat{y} \mid a = 0, y = 1) = \Pr(\hat{y} \mid a = 1, y = 1). \tag{6}$$

Other notions (Mehrabi et al., 2021) also help to characterize the fairness problem including Equalized Odds (equality on both true positive rate and true negative rate), Accuracy Parity (equality on the predictive error rate), Predictive Equality (equality on the false positive rate), etc. We focus on Equal Opportunity and Demographic Parity while the rest notions could be incremental extensions to our framework.

To formulate the fairness issue as an optimization problem, one can intuitively quantify the inequality and turn the subjection of fairness into an objective function. For instance, with $\ell_{0/1}$ denotes zero-one loss, the gap in Equal Opportunity is:

$$\Big| \mathbb{E}\left[\ell_{0/1}(z;\theta) | a = 1, y = 1\right] - \mathbb{E}\left[\ell_{0/1}(z;\theta) | a = 0, y = 1\right] \Big|. \tag{7}$$

To make the function differentiable, a surrogate function is necessary to replace $\ell_{0/1}$ and here we substitute it with the training loss ℓ for Equal Opportunity, following previous works (Zafar et al., 2017; Donini et al., 2018). Demographic Parity can be derived similarly. The fair loss over a sample set $\mathcal{S}$ for these two notions are:

$$\begin{aligned} f^{\mathcal{S}}_{\text{eop}}(\theta) &= \Big| \mathbb{E}_{\mathcal{S}}\left[\ell(z;\theta) \mid a = 1, y = 1\right] \\ &\quad - \mathbb{E}_{\mathcal{S}}\left[\ell(z;\theta) \mid a = 0, y = 1\right] \Big|, \\ f^{\mathcal{S}}_{\text{dp}}(\theta) &= \Big| \mathbb{E}_{\mathcal{S}}\left[\hat{y} \mid a = 1\right] - \mathbb{E}_{\mathcal{S}}\left[\hat{y} \mid a = 0\right] \Big|. \end{aligned} \tag{8}$$

We now begin to elaborate on our framework. Instead of evaluating the fairness and utility on the training set $\mathcal{T}$, we are interested in the change of this two-side performance for the classifier on a validation set $\mathcal{V} = \{z_j = (x_j, y_j)\}_{j=1}^{N_{\mathbf{v}}}$ before and after a reweighing over $\mathcal{T}$, where each sample z_j in the validation set is associated with a binary sensitive attribute $a_j \in \{0, 1\}$. Measuring the influence of training samples on $\mathcal{V}$ towards fairness and utility can help us identify which sample has a positive impact on training, and to be precise, in which direction and to what extent.

To clarify the source of a sample, we use i to denote the index belonging to the training set, and j for the validation set. Let $\mathbf{e}_i$ stand for an all-zero vector except entry i equal to 1. Deploying $\mathbf{e}_i$ as a weight for retraining means a hard removal of z_i while preserving the rest samples unchanged. The influence function on fairness can be derived by realizing the function f in Equation (4) with Equation (8):

$$\mathcal{I}_{\text{eop/dp}}(\mathbf{e}_i) = \nabla_\theta f^{\mathcal{V}}_{\text{eop/dp}}(\hat{\theta}(\mathbf{1}))^\top \mathbf{H}^{-1}_{\hat{\theta}(\mathbf{1})} \nabla_\theta \ell(z_i; \hat{\theta}(\mathbf{1})), \tag{9}$$

and similarly, the influence on utility is expressed as

$$\mathcal{I}_{\text{util}}(\mathbf{e}_i) = \sum_j \nabla_\theta \ell(z_j; \hat{\theta}(\mathbf{1}))^\top \mathbf{H}^{-1}_{\hat{\theta}(\mathbf{1})} \nabla_\theta \ell(z_i; \hat{\theta}(\mathbf{1})). \tag{10}$$

Here we use a truncated subscript 'eop/dp' to express the feasibility for either functions, and in what follows we may use $f^{\mathcal{V}}_{\text{fair}}$ to unify $f^{\mathcal{V}}_{\text{eop/dp}}$. The influence function satisfies additivity followed by the additivity of $\nabla_\theta f^{\mathcal{V}}_{\text{fair}}$ and $\sum_j \nabla_\theta \ell(z_j; \hat{\theta}(\mathbf{1}))$ w.r.t. $\mathbf{e}_i$, *i.e.* $\mathbf{w} = \mathbf{e}_1 + \mathbf{e}_2$ implies $\mathcal{I}(\mathbf{w}) = \mathcal{I}(\mathbf{e}_1) + \mathcal{I}(\mathbf{e}_2)$.

We study the correlation between $\mathcal{I}_{\text{eop/dp}}$ and $\mathcal{I}_{\text{util}}$ over the reweighing of $\mathcal{T}$. We first state our assumption for the remaining gradient of z_i at $\hat{\theta}(\mathbf{1})$.

Assumption 3.1. The gradient matrix of training samples in $\mathcal{T}$ at $\hat{\theta}(\mathbf{1})$: $\begin{bmatrix} \nabla \ell_\theta(z_1; \hat{\theta}(\mathbf{1})) \\ \vdots \\ \nabla \ell_\theta(z_{N_\mathcal{T}}; \hat{\theta}(\mathbf{1})) \end{bmatrix} \in \mathbb{R}^{N_\mathcal{T} \times D}$ has rank D.

Assumption 3.1 is easy to satisfy if $N_\mathcal{T} \gg D$ in a general training case, *i.e.* one set a model with a proper dimension to fit an adequate number of training samples. The assumption is mild for large-scale data with sufficient variety.

Theorem 3.2. *If $f^{\mathcal{V}}_{eop}(\hat{\theta}(\mathbf{1}))$ is not in local optimum, $\nabla_\theta f^{\mathcal{V}}_{eop}(\hat{\theta}(\mathbf{1}))$ and $\sum_j \nabla_\theta \ell(z_j; \hat{\theta}(\mathbf{1}))$ are linearly independent, then under assumptions of influence function and Assumption 3.1, there are conical combinations $\mathbf{w}$ of $\{\mathbf{e}_i\}_{i=1}^{N_\mathcal{T}}$ to construct a reweighing for training samples such that $\mathcal{I}_{eop}(\mathbf{w}) < 0$ and $\mathcal{I}_{util}(\mathbf{w}) \leq 0$, with $\|\mathbf{w}\|_\infty \leq 1$.*

Remark 3.3. Theorem 3.2 states that, if the loss of Equal Opportunity on $\mathcal{V}$ remains a space to improve, then there exists reweighing on $\mathcal{T}$ to improve $f^{\mathcal{V}}_{\text{eop}}$ but will not increase the utility loss under some conditions. $\mathbf{w}$ is an element-wise non-negative vector and have an upper bound on entries, which is a nice property we will use in our algorithmic design in the next section. Proofs can be found in Appendix A.

Here we see a proper downweighting can be expected to enhance fairness while keeping the utility not going worse at the influence function level. This finding raises an opportunity to *achieve fairness at no utility cost*. Similarly, we present a corollary that states the same conclusion for Demographic Parity.

Corollary 3.4. *If $\nabla_\theta f^{\mathcal{V}}_{dp}(\hat{\theta}(\mathbf{1})) \neq 0$, $\nabla_\theta f^{\mathcal{V}}_{dp}(\hat{\theta}(\mathbf{1}))$ and $\sum_j \nabla_\theta \ell(z_j; \hat{\theta}(\mathbf{1}))$ are linearly independent, then under assumptions of influence function and Assumption 3.1 satisfied, there are conical combinations $\mathbf{w}$ of $\{\mathbf{e}_i\}_{i=1}^{N_\mathcal{T}}$ to construct a reweighing for training samples such that $\mathcal{I}_{dp}(\mathbf{w}) < 0$ and $\mathcal{I}_{util}(\mathbf{w}) \leq 0$, with $\|\mathbf{w}\|_\infty \leq 1$.*

Note that a classifier that satisfies Equal Opportunity with an optimal predictive utility performance, might not be optimal for Demographic Parity. A perfect classifier will still encounter a tradeoff between Demographic Parity and predictive utility induced by the difference in base rate between sensitive groups. Corollary 3.4 is not against this impossibility since it only declares a cost-free improvement on $f^{\mathcal{V}}_{\text{dp}}$ when it is not locally optimal.

Discussion on Fairness-Utility Tradeoff There is literature theoretically discussing the intrinsic tradeoff between fairness and utility, and such a tradeoff has been empirically revealed in many experiments. Zhao & Gordon (2019) point out the tradeoff between Demographic Parity and the joint error rate across sensitive groups, and Menon & Williamson (2018) also characterize the accuracy-fairness tradeoff holding for an arbitrary classifier. However, they discuss under the context of a given fixed distribution that does not correspond with us. Our findings, from a reweighing perspective, match with the conclusion in Dutta et al. (2020) that characterizing with Chernoff information from information theory, there exists an ideal distribution where fairness and utility are in accord. However, they do not consider a more practical model learning and inference setting, but primarily restrict to likelihood ratio detectors with synthetic experiments. Our data reweighing via influence function lets us move beyond a bias distribution and release the tradeoff.

4. Fairness via Data Reweighing

The foregoing theoretical analysis raises an opportunity to mitigate the unfairness towards different notions while preserving the predictive utility. In this section, we concretize this idea and convert it into an algorithm to find the optimal weights $\mathbf{w}^*$ through linear programming.

Having $\mathcal{I}_{\text{fair}}(\mathbf{e}_i)$ and $\mathcal{I}_{\text{util}}(\mathbf{e}_i)$ in hands, the ideal reweighing is to completely close the gap towards a fairness notion, *e.g.* Equal Opportunity, or reach a user-defined threshold, while preserving the utility not decrease. The searching for $\mathbf{w}^*$ can be cast into a linear program as follows:

$$\begin{aligned} \text{minimize} \quad & \sum_i w_i \\ \text{subject to} \quad & \sum_i w_i \mathcal{I}_{\text{fair}}(\mathbf{e}_i) = -f^{\mathcal{V}}_{\text{fair}}, \\ & \sum_i w_i \mathcal{I}_{\text{util}}(\mathbf{e}_i) \leq 0, \\ & w_i \in [0, 1]. \end{aligned} \tag{11}$$

Equation (11) tends to minimize the total amount of perturbation from an initially uniform weight. This prevents downweighting too many samples, thus keeping the generalization capacity from the validation set to the test set. The first subjection is to find the weights that perfectly close the fairness gap in Equation (8), where the minimum of an absolute value is zero or it can also be set to a user-defined threshold. The second subjection is to keep the utility loss non-increasing. In the last subjection we set a range for w_i from 0 to 1, inheriting the property of $\mathbf{w}$ in Theorem 3.2. We set the lower bound as 0 to ensure all the reweighing is downweighting instead of upweighting. Here an upweighting is definitely practicable which emphasizes some particular samples and could achieve the same effect as downweighting on other samples. For the simultaneous upweighting and downweighting, it leads the objective function to be the sum of a set of absolute values between w_i and 1. The linear programming problem with absolute values needs to be transformed into a standard one by introducing additional variables (Dantzig, 2016). Also, we do not observe a conspicuous benefit from introducing upweighting in practice. For simplicity and efficiency, we only consider the downweighting for data reweighing. The upper bound of weights prevents the linear programming from turning a sample into a negative sample which affords negative loss. This is for training stability concerns and avoids the reweighing overly concentrating on one sample.

However, Equation (11) may not offer a feasible solution since there might exist a case that $\min_{\mathbf{w}} \sum_i w_i \mathcal{I}_{\text{fair}}(\mathbf{e}_i) > -f^{\mathcal{V}}_{\text{fair}}$ with the rest constraints held. In this case, we substitute Equation (11) with Equation (12) stated as follow:

$$\begin{aligned} \text{minimize} \quad & \sum_i w_i \mathcal{I}_{\text{fair}}(\mathbf{e}_i) \\ \text{subject to} \quad & \sum_i w_i \mathcal{I}_{\text{util}}(\mathbf{e}_i) \leq 0, \\ & \sum_i w_i \leq \alpha N_{\mathcal{T}}, \\ & w_i \in [0,1], \end{aligned} \tag{12}$$

with $\alpha \in (0,1]$. Equation (12) rotates the first constraint to its objective function and replenishes a constraint on the quantity of perturbation w.r.t the total number of training samples. α here is a hyperparameter indicating the proportion of weights to be changed. Note that we set a higher priority of Equation (11) over Equation (12), since if $\min_{\mathbf{w}} \sum_i w_i \mathcal{I}_{\text{fair}}(\mathbf{e}_i) < -f^{\mathcal{V}}_{\text{fair}}$, the objective values goes to negative and cause unfairness again, *i.e.,* it flips the original unprivileged group into a privileged group. If one wants to support Equation (12) as the priority, a lower bound of $-f^{\mathcal{V}}_{\text{fair}}$ should be added to the objective function but that can cause dual degeneracy for a minimization problem in LP.

In practice, although the solution seems to be optimal in terms of influential approximation, $\mathbf{w}^*$ cannot imply absolute fairness at no utility cost. The error comes from two sides. (1) Even though influence function is almost precise for the removal of an individual sample, transferring the influence function to the actual influence still suffers some additional errors related to the quantity of perturbation. Koh et al. (2019) empirically reveal that when removing a group of samples from the training set, influence function tends to underestimate or overestimate the actual influence, but still keeps a high correlation. Colloquially, the larger the group removed from the training set, the more imprecision can be expected. (2) There is a gap between the surrogate function (for differentiable purpose) and the real value of Equal Opportunity gap or Demographic Parity gap, as well as the utility. As a remedy, we complement some extra hyperparameters in the constraints of Equation (11) to alleviate such errors, and we restate it as follows:

$$\begin{aligned} \text{minimize} \quad & \sum_i w_i \\ \text{subject to} \quad & \sum_i w_i \mathcal{I}_{\text{fair}}(\mathbf{e}_i) \leq -(1-\beta)\ell^{\mathcal{V}}_{\text{fair}}, \\ & \sum_i w_i \mathcal{I}_{\text{util}}(\mathbf{e}_i) \leq \gamma(\min_{\mathbf{v}} \sum_i v_i \mathcal{I}_{\text{util}}(\mathbf{e}_i)), \\ & w_i \in [0,1]. \end{aligned} \tag{13}$$

We linearly tighten or relax the constraint with hyperparameters β and γ since influence function under group effect could enlarge the scale of deflection but still under a high correlation. If there is a deflection, we can observe it on the validation set and regulate the objectives to compensate for the group effect. We conduct a grid search for β and γ on the validation set, and demonstrate the performance on the test set. We summarize the algorithmic pipeline in Algorithm 1 and show its property of feasibility.

Algorithm 1 No Utility-Cost Fairness via Data Reweighing

Input: Training set $\mathcal{T} = \{z_i\}_{i=1}^{N_{\mathcal{T}}}$, validation set $\mathcal{V} = \{z_j, a_j\}_{j=1}^{N_{\mathcal{V}}}$.
Train $\hat{\theta}(\mathbf{1}) = \arg\min_\theta \sum_i \ell(z_i;\theta)$ by Equation (1);
Compute $f^{\mathcal{V}}_{\text{fair}}$ by Equation (8);
Compute $\sum_j \nabla_\theta \ell(z_j;\hat{\theta}(\mathbf{1}))$;
Compute Hessian vector product $\mathbf{H}^{-1}_{\hat{\theta}(\mathbf{1})} \nabla_\theta \ell(z_i;\hat{\theta}(\mathbf{1}))$ for every training sample $i \in \mathcal{T}$;
Compute $\mathcal{I}_{\text{fair}}(\mathbf{e}_i)$ and $\mathcal{I}_{\text{util}}(\mathbf{e}_i)$ for every training sample $i \in \mathcal{T}$ by Equations (9) and (10);
Solve the linear programming problem in Equation (13);
if Equation (13) is infeasible **then**
  Solve the linear programming problem in Equation (12);
end if
Retrain the model with $\mathbf{w}^*$ by Equation (2) and obtain $\hat{\theta}(\mathbf{1} - \mathbf{w}^*)$;
Evaluate the test set with $\hat{\theta}(\mathbf{1} - \mathbf{w}^*)$.

Corollary 4.1. *If assumptions in Theorem 3.2 and Corollary 3.4 are satisfied, then Algorithm 1 could reach a feasible and non-trivial solution.*

Remark 4.2. A trivial solution means the solved $\mathbf{w} = 0$ that does not reweigh any training samples.

5. Related Work

We chronologically review the pre-processing methods for algorithmic fairness, and introduce related works on influence function. We include many of these pre-processing methods in experimental comparison.

Fair Algorithms with Pre-processing Previous works develop pre-processing techniques to ensure a nice property that the fairness-oriented improvement is independent of existing learning pipelines. Kamiran & Calders (2012) propose several reweighing and label flipping approaches based on data statistics to expand the underrepresented group or benefit the unprivileged group. Some of these approaches further consider the original predictive confidence. Zemel et al. (2013) learn a fair intermediate representation for data towards both group and individual fairness. Feldman et al. (2015) satisfy Demographic Parity by transforming the input data into fair features while preserving their original rank. Calmon et al. (2017) follow the data fair transformation and optimize the transformation with individual least distance constraints. Krasanakis et al. (2018) and Jiang & Nachum (2020) introduce reweighing approaches and update weights

iteratively through a continuous learning process. The iterative update asks for retraining the model within each learning iteration. Wang et al. (2019) use a descent algorithm to learn a counterfactual distribution from data to close the fairness gap for a black-box predictor, and build the pre-processing via optimal transport. Yan et al. (2020) and Lahoti et al. (2020) focus on a case where the sensitive attribute is missing. Yan et al. (2020) use clustering to find out underrepresented groups and complements them with nearest neighbor searching. Lahoti et al. (2020) considers Rawlsian Max-Min Fairness with a special interest in the worst-case accuracy. Many works consider fairness subjecting to predictive utility constraints in design, and the intention to preserve the utility helps to popularize fair algorithms. Our algorithmic design considers both fairness and predictive utility via a one-pass reweighing approach that is friendly for subsequent learning pipelines which might be time-consuming. We granularly characterize each training sample using their fairness and utility influence, and compute individual weights by solving linear programs.

Influence Function Influence function originates from diagnostic in statistics (Cook, 1977; Cook & Weisberg, 1980). It approximates the actual effect brought by the removal of training points from models or other perturbations on data. Influence function in machine learning has been used to study model robustness (Christmann & Steinwart, 2004; Hampel et al., 2011; Liu et al., 2014). Recently, Koh & Liang (2017) extend influence function to various large-scale machine learning models, and introduce its applications in adversarial attack, data interpretation, and label fixing. More works follow up to develop this statistical tool. Giordano et al. (2019) provides finite-sample error bounds on the leave-k-out case of the asymptotic results. Koh et al. (2019) characterize the error when removing a group of data and provide interesting empirical findings. They reveal that when influence function is measuring the effect of removing a random or a certain group of samples, the absolute and relative error will be large, but still correlates well with the actual influence. In this work, we consider an innovative application of influence function in considering both algorithmic fairness and utility.

6. Experiment

6.1. Dataset

We use the following real-world tabular datasets for experiments (Dua & Graff, 2017). We provide statistics in Appendix B. **Adult**. The Adult dataset (Kohavi & Becker) contains 45,222 census personal records. It includes 14 attributes such as age, education, race, etc. The goal is to predict if the personal annual income exceeds 50k. We set gender as the sensitive attribute. **Compas**. The Compas dataset (Julia Angwin & Kirchner, 2016) records information like criminal history, jail and prison time, demographic, etc. The dataset is to predict a recidivism risk score for defendants. We set race as the sensitive attribute. **Communities and Crime**. The Communities and Crime dataset (Redmond & Baveja, 2002) describes communities with the percent of the population considered urban, the median family income, etc. The goal is to predict violent crimes and we set the percentage of the black population as the sensitive attribute. **German Credit**. The German Credit dataset (Hofmann) classifies people as good or bad credit risks using their profile and history. We set age as the sensitive attribute with a threshold at 30.

6.2. Implementation and Protocol

For fair classification, we consider pre-processing baselines: Massaging, Reweigh, Preferential from Kamiran & Calders (2012), and Disparate Impact Remover (Dis. Remover) (Feldman et al., 2015), Label Bias (Jiang & Nachum, 2020), and Adversarial Reweighted Learning (ARL) (Lahoti et al., 2020). We also involve two in-processing adversarial training methods: Adversarial Debiasing (Adv.) (Zhang et al., 2018) and Conditional Adversarial Debiasing (Cond. Adv.) (Zhao et al., 2019). For comprehensiveness, we consider two kinds of base models: Logistic Regression and two-layer non-linear Neural Networks, where we use Logistic Regression for all four datasets and use Neural Networks for the Adult and Compas datasets due to their large sample sizes. Most pre-processing approaches work independently of the base model and can be wrapped on both Logistic Regression and Neural Networks, while adversarial training methods only work for Neural Networks. We equip our methods to Neural Networks by only computing the influence to the last layer of Neural Networks and retraining this part of weights, thus the convexity is guaranteed. The specific parameters of base models and input data are used exactly the same across all baselines and our methods. Input data are standardized by removing the mean and scaling to unit variance. For methods with a hyperparameter to directly control the fairness-utility tradeoff, we tune the hyperparameter and show the Pareto curve in figures. Linear programs in Algorithm 1 are solved using Gurobi (Gurobi Optimization, LLC, 2021) under an academic license. We divide all the datasets into training set (60%), validation set (20%), and test set (20%), except for the Adult dataset that has a pre-defined split on training/validation/test set. More details are reported in Appendix C.

We explicitly compute the Hessian matrix and its inverse to get the influence function. A typical complexity for computing Hessian is $\mathcal{O}(nd^2)$, where n and d are the numbers of samples and model dimension, and its inversion takes $\mathcal{O}(d^3)$. To address high-dimensional models, we can apply conjugate gradients or stochastic estimation of Hessian-vector products, resulting in $\mathcal{O}(nd)$ (Koh & Liang, 2017).

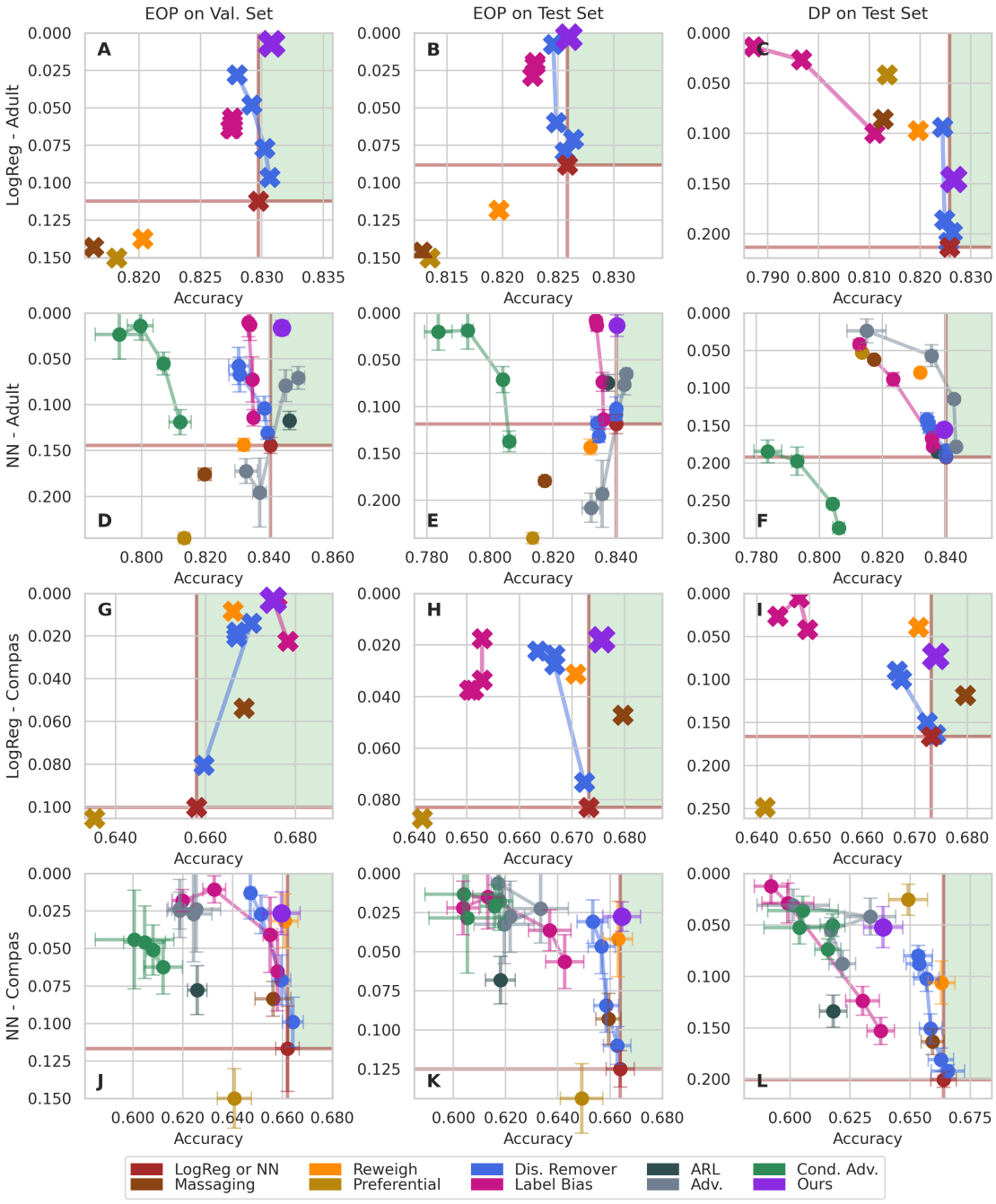

Figure 1. Experimental results of five pre-processing and two in-processing fair classification methods on the Adult and Compas datasets. For each figure, we indicate the corresponding base model, experimental dataset, fairness metric, and evaluation set in the left and top header (EOP→ Equal Opportunity, DP→ Demographic Parity). Y-axis for fairness is inverted and shows the absolute value of the gap in fairness between the privileged and unprivileged groups (the higher position means the smaller fairness metric value and the better performance towards fairness), while X-axis shows the predictive accuracy. According to the utility and fairness performance of the base model LogReg (Logistic Regression) or NN (Neural Networks), we plot a horizontal and a vertical line in each figure and divide the space by fairness and utility results into four regions, where the space in green means a fairer and more accurate model compared to the base model. A point closer to the top right indicates better performance in both utility and fairness. The values of some points by the Preferential method are out of the current scale of these figures; for better visualization, we put these points at the boundary of a figure. We do not observe randomness for the Logistic Regression model, while we plot the standard deviation for Neural Networks with results obtained by five random seeds. For methods with a hyperparameter to control the tradeoff between fairness and utility, the line connecting the same method indicates the monotonously increased change on the hyperparameter. For details please refer to Appendix C.

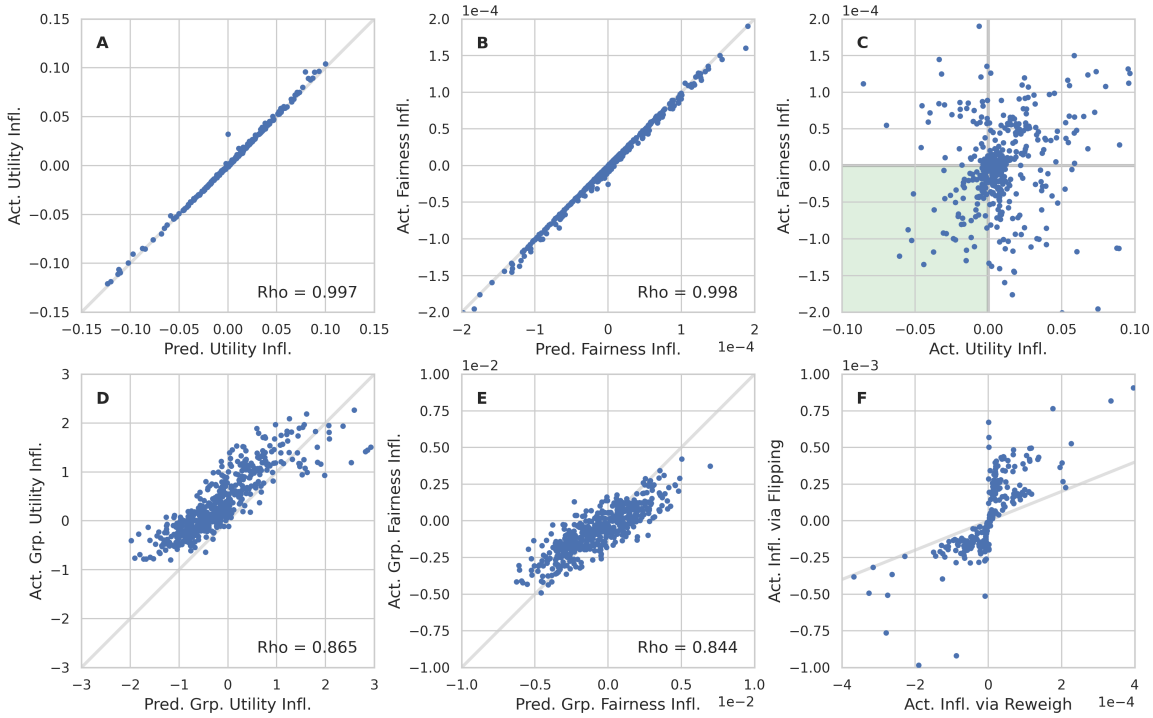

Figure 2. Studies on influence function and its actual effect. All results are obtained on the Adult dataset. **A** and **B**: Predictive influence v.s. actual influence by leave-one-out retraining in terms of utility and Equal Opportunity loss; **C**: Actual influence of utility loss v.s. Equal Opportunity loss by leave-one-out retraining. **A**, **B** and **C** are conducted on 500 randomly selected training samples; **D** and **E**: Influence function v.s. actual effect by leave-group-out retraining in terms of utility and Equal Opportunity loss with a group size equal to 250. **F**: Actual influence of Equal Opportunity loss by target label flip and retraining v.s. reweighing and retraining from our algorithm. 'Rho' indicates Pearson correlation coefficient in figures.

6.3. Experimental Comparison

We present our experimental results on the Adult and Compas datasets in Figure 1, and defer other results to Appendix D. Through experiments we have several findings: (1) In general, except in Figure **G** where the majority of these fair classifiers deliver simultaneous and non-trivial improvements in both fairness and utility, the experimental results of baseline methods are mainly located at the top left region, indicating that their improved fairness is achieved at a non-negligible cost of utility. (2) Specifically, heuristic pre-processing methods including Massaging, Reweigh, and Preferential fail to guarantee a stable improvement over fairness, and sometimes even decrease Equal Opportunity. The learning approaches Disparate Impact Remover and Label Bias induce a tradeoff between fairness and utility, and unfortunately lead to either unsatisfied fairness or an unacceptable utility cost in many cases. The adversarial approaches Adversarial Debiasing and Conditional Adversarial Debiasing also bring an obligatory tradeoff so that they need careful tuning. Moreover, they induce a large variance in Neural Networks model. (3) In most cases (See **A**-**I**), our method is able to improve fairness while keeping the utility unchanged or even slightly improved. This nice property and its pre-processing fashion can play a significant advantage in popularizing fair algorithms. These empirical results also verify our theoretical findings. It is worthy to note that our method cannot escape the upper bound of Demographic Parity at no utility cost since this scenario resists the impossibility from the difference in the base rate. In a few cases (See **J**-**K**), our method sacrifices a little utility to greatly improve the fairness. Due to the nature of influence approximation, when removing a group of samples from the training set, there exists a gap between actual and predicted influence. There is still some deviation even though we add hyperparameters to fulfill the gap in Equation (13).

6.4. Sample Influence and Actual Effect

We investigate the influence function and its actual influence in Figure 2. In **A** and **B**, we visualize influence function and corresponding actual change after leave-one-out retraining, and show that influence function can predict the actual influence with high precision for both utility and fairness. In **C**, we show the distribution of actual influence for individual samples. Some parts of points are located at the green regions, meaning that removing or downweighting these points can simultaneously improve fairness and utility. This is well in accord with our Theorem 3.2. In **D** and **E**, we show how the influence function differs from removing an individual sample to removing a group of samples. The influence function still remains a high correlation but re-

sults in a larger error. The imprecision justifies our design in Equation (13) where we use hyperparameters to mitigate this effect. In **F**, we compare the influence of removing samples to target label flipping, complementing the label bias hypothesis in the introduction. Flipping the label can incur a larger change in loss value compared to simply reweighing them.

7. Conclusion

We proposed a data reweighing approach for improving algorithmic fairness and utility. We granularly measured the influence of every training sample towards fairness and utility on a validation set. We demonstrated that under some mild assumptions, there exists some reweighing on training samples that can improve fairness at no utility cost. We solved the reweighing strategy through linear programming with constraints of both fairness and utility, and guaranteed its feasibility. We empirically verified our algorithms on multiple tabular datasets and showed that the traditional fairness-utility tradeoff could be released in most cases.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *International Conference on Machine Learning*, 2018.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *California Law Review*, 2016.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, 2017.

Chen, I. Y., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, 2018.

Chhabra, A., Singla, A., and Mohapatra, P. Fair clustering using antidote data. *arXiv preprint arXiv:2106.00600*, 2021.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 2017.

Christmann, A. and Steinwart, I. On robustness properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, 2004.

Civil Rights Act. Civil rights act of 1964, title vii, equal employment opportunities. 1964.

Cook, R. D. Detection of influential observation in linear regression. *Technometrics*, 1977.

Cook, R. D. and Weisberg, S. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 1980.

Cotter, A., Jiang, H., Gupta, M. R., Wang, S., Narayan, T., You, S., and Sridharan, K. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 2019.

Dantzig, G. *Linear programming and extensions*. Princeton University Press, 2016.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, 2018.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Dutta, S., Wei, D., Yueksel, H., Chen, P.-Y., Liu, S., and Varshney, K. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*, 2020.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, 2012.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

Ferguson, A. G. *The rise of big data policing*. New York University Press, 2017.

Giordano, R., Stephenson, W., Liu, R., Jordan, M., and Broderick, T. A swiss army infinitesimal jackknife. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Goh, G., Cotter, A., Gupta, M., and Friedlander, M. P. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, 2016.

Goodman, B. and Flaxman, S. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 2017.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2021. URL https://www.gurobi.com.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 2016.

Hofmann, H. Statlog (german credit data) data set. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

Jiang, H. and Nachum, O. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, 2020.

Julia Angwin, Jeff Larson, S. M. and Kirchner, L. Machine bias, 2016.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 2012.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, 2018.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 2017.

Koh, P. W., Ang, K.-S., Teo, H. H., and Liang, P. On the accuracy of influence functions for measuring group effects. In *Advances in Neural Information Processing Systems*, 2019.

Kohavi, R. and Becker, B. Adult data set. https://archive.ics.uci.edu/ml/datasets/adult.

Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., and Kompatsiaris, Y. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *The World Wide Web Conference*, 2018.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Li, P., Zhao, H., and Liu, H. Deep fair clustering for visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

Li, P., Wang, Y., Zhao, H., Hong, P., and Liu, H. On dyadic fairness: Exploring and mitigating bias in graph connections. In *International Conference on Learning Representations*, 2021.

Liu, Y., Jiang, S., and Liao, S. Efficient approximation of cross-validation for kernel methods using bouligand influence function. In *International Conference on Machine Learning*, 2014.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 2021.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, 2018.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, 2017.

Rastegarpanah, B., Gummadi, K. P., and Crovella, M. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *ACM International Conference on Web Search and Data Mining*, 2019.

Redmond, M. and Baveja, A. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 2002.

Roth, K., Lucchi, A., Nowozin, S., and Hofmann, T. Stabilizing training of generative adversarial networks through regularization. *Advances in Neural Information Processing Systems*, 2017.

Song, H., Li, P., and Liu, H. Deep clustering based fair outlier detection. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.

Wang, H., Ustun, B., and Calmon, F. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, 2019.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In *Conference on Learning Theory*, 2017.

Yan, S., Kao, H.-t., and Ferrara, E. Fair class balancing: enhancing model fairness without observing sensitive attributes. In *ACM International Conference on Information & Knowledge Management*, 2020.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *The 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

Zhao, H. and Gordon, G. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems*, 2019.

Zhao, H., Coston, A., Adel, T., and Gordon, G. J. Conditional learning of fair representations. In *International Conference on Learning Representations*, 2019.

A. Proof and Discussion

Assumption 3.1. The gradient matrix of training samples in $\mathcal{T}$ at $\hat{\theta}(\mathbf{1})$: $\begin{bmatrix} \nabla \ell_\theta(z_1; \hat{\theta}(\mathbf{1})) \\ \vdots \\ \nabla \ell_\theta(z_{N_\mathcal{T}}; \hat{\theta}(\mathbf{1})) \end{bmatrix} \in \mathbb{R}^{N_\mathcal{T} \times D}$ has rank D.

We discuss the feasibility of the assumption here. The assumption requires that we have training samples such that the gradient vector matrix has full rank. This is a mild assumption for a large training dataset with sufficient sample diversity and a proper model dimension. The assumption could be violated when (1). the model has a very large and improper dimension such that some parameters are not even activated by the training data, and (2). there are very limited training points. For the second scenario, consider an extreme case if we only have one training sample for the model, therefore we cannot reweight this sample to change the model's parameter, but luckily it is not likely to happen in general cases. In Figure **C** from Figure 2 we show some hard removals of samples are bringing positive impact to both utility and fairness, which confirm our theorem and the feasibility of this assumption.

Theorem 3.2. *If $f^{\mathcal{V}}_{eop}(\hat{\theta}(\mathbf{1}))$ is not in local optimum, $\nabla_\theta f^{\mathcal{V}}_{eop}(\hat{\theta}(\mathbf{1}))$ and $\sum_j \nabla_\theta \ell(z_j; \hat{\theta}(\mathbf{1}))$ are linearly independent, then under assumptions of influence function and Assumption 3.1, there are conical combinations $\mathbf{w}$ of $\{\mathbf{e}_i\}_{i=1}^{N_\mathcal{T}}$ to construct a reweighing for training samples such that $\mathcal{I}_{eop}(\mathbf{w}) < 0$ and $\mathcal{I}_{util}(\mathbf{w}) \leq 0$, with $\|\mathbf{w}\|_\infty \leq 1$.*

Proof. For completeness, we restate the Equal Opportunity loss over the validation set $\mathcal{V}$ in Equation (8) as follows:

$$\begin{aligned} f^{\mathcal{V}}_{\text{eop}}(\hat{\theta}(\mathbf{1})) &= |\mathbb{E}_\mathcal{V}[\ell(z; \hat{\theta}(\mathbf{1})) \mid a = 1, y = 1] - \mathbb{E}_\mathcal{V}[\ell(z; \hat{\theta}(\mathbf{1})) \mid a = 0, y = 1]| \\ &= |\frac{1}{|\mathcal{V}_{a=1,y=1}|} \sum_{j: a_j = 1, y_j = 1} \ell(z_j; \hat{\theta}(\mathbf{1})) - \frac{1}{|\mathcal{V}_{a=0,y=1}|} \sum_{j: a_j = 0, y_j = 1} \ell(z_j; \hat{\theta}(\mathbf{1}))| > 0, \end{aligned} \tag{14}$$

where $\mathcal{V}_{a=1}$ denotes the set conditional on $a = 1$, and $|\mathcal{V}|$ is the cardinality of set $\mathcal{V}$. We have $\nabla_\theta f^{\mathcal{V}}_{\text{eop}}(\hat{\theta}(\mathbf{1})) \neq 0$ since $\theta f^{\mathcal{V}}_{\text{eop}}(\hat{\theta}(\mathbf{1}))$ is not in local optimum, and would guarantees $\nabla_\theta f^{\mathcal{V}}_{\text{eop}}(\hat{\theta}(\mathbf{1}))^\top \mathbf{H}^{-1}_{\hat{\theta}(\mathbf{1})} \neq 0$ in $\mathcal{I}_{\text{eop}}$.

Now let $\mathbf{G} = \begin{bmatrix} \nabla \ell_\theta(z_1; \hat{\theta}(\mathbf{1})) \\ \vdots \\ \nabla \ell_\theta(z_{N_\mathcal{T}}; \hat{\theta}(\mathbf{1})) \end{bmatrix} \in \mathbb{R}^{N_\mathcal{T} \times D}$ with full rank D (Assumption 3.1), we can write the influence function towards Equal Opportunity and utility in Equation (9) and Equation (10) as follows:

$$\begin{aligned} \mathcal{I}_{\text{eop}}(\mathbf{w}) &= (\nabla_\theta f^{\mathcal{V}}_{\text{eop}}(\hat{\theta}(\mathbf{1}))^\top \mathbf{H}^{-1}_{\hat{\theta}(\mathbf{1})})(\mathbf{G}^\top \mathbf{w}), \\ \mathcal{I}_{\text{util}}(\mathbf{w}) &= (\sum_j \nabla_\theta \ell(z_j; \hat{\theta}(\mathbf{1}))^\top \mathbf{H}^{-1}_{\hat{\theta}(\mathbf{1})})(\mathbf{G}^\top \mathbf{w}). \end{aligned} \tag{15}$$

Since $\nabla_\theta f^{\mathcal{V}}_{\text{eop}}(\hat{\theta}(\mathbf{1}))$ and $\sum_j \nabla_\theta \ell(z_j; \hat{\theta}(\mathbf{1}))$ are linearly independent, we have $\nabla_\theta f^{\mathcal{V}}_{\text{eop}}(\hat{\theta}(\mathbf{1}))^\top \mathbf{H}^{-1}_{\hat{\theta}(\mathbf{1})}$ and $\sum_j \nabla_\theta \ell(z_j; \hat{\theta}(\mathbf{1}))^\top \mathbf{H}^{-1}_{\hat{\theta}(\mathbf{1})}$ to be linearly independent, since $\mathbf{H}^{-1}_{\hat{\theta}(\mathbf{1})}$ is non-singular. Because $\mathbf{G}$ has rank D, so the gradient vectors in $\mathbf{G}$ span $\mathbb{R}^D$, thus we can always find $\mathbf{w}$ on $N_\mathcal{T}$ training samples such that $\mathcal{I}_{\text{eop}}(\mathbf{w}) < 0$ and $\mathcal{I}_{\text{util}}(\mathbf{w}) \leq 0$.

Recall the first-order optimality $\sum_i \nabla \ell(z_i; \hat{\theta}(\mathbf{1})) = 0 \rightarrow \mathbf{1}^\top \mathbf{G} = 0$ for the model. Once we have $\mathbf{w}$ satisfy $\mathcal{I}_{\text{eop}}(\mathbf{w}) < 0$ and $\mathcal{I}_{\text{util}}(\mathbf{w}) \leq 0$ on $N_\mathcal{T}$, we should have

$$\begin{aligned} \mathcal{I}_{\text{eop}}(\mathbf{w}) &= (\nabla_\theta f^{\mathcal{V}}_{\text{eop}}(\hat{\theta}(\mathbf{1}))^\top \mathbf{H}^{-1}_{\hat{\theta}(\mathbf{1})})(\mathbf{G}^\top (\mathbf{w} + c\mathbf{1})) < 0, \\ \mathcal{I}_{\text{util}}(\mathbf{w}) &= (\sum_j \nabla_\theta \ell(z_j; \hat{\theta}(\mathbf{1}))^\top \mathbf{H}^{-1}_{\hat{\theta}(\mathbf{1})})(\mathbf{G}^\top (\mathbf{w} + c\mathbf{1})) \leq 0, \end{aligned} \tag{16}$$

where c is an arbitrary real number. If $\mathbf{w}$ has negative entries, let $c = -\min_i w_i$ and $\mathbf{w} \leftarrow (\mathbf{w} + c\mathbf{1})$ helps to convert all entries into non-negative. A normalization $\mathbf{w} \leftarrow \mathbf{w}/\|\mathbf{w}\|$ helps to bound $\|\mathbf{w}\|_\infty \leq 1$, hence completing the proof. □

Corollary 3.4. *If $\nabla_\theta f_{dp}^{\mathcal{V}}(\hat{\theta}(\mathbf{1})) \neq 0$, $\nabla_\theta f_{dp}^{\mathcal{V}}(\hat{\theta}(\mathbf{1}))$ and $\sum_j \nabla_\theta \ell(z_j; \hat{\theta}(\mathbf{1}))$ are linearly independent, then under assumptions of influence function and Assumption 3.1 satisfied, there are conical combinations $\mathbf{w}$ of $\{\mathbf{e}_i\}_{i=1}^{N_\mathcal{T}}$ to construct a reweighing for training samples such that $\mathcal{I}_{dp}(\mathbf{w}) < 0$ and $\mathcal{I}_{util}(\mathbf{w}) \leq 0$, with $\|\mathbf{w}\|_\infty \leq 1$.*

Proof. The proof follows the proof for Theorem 3.2 by replacing $\nabla_\theta f_{\text{eop}}^{\mathcal{V}}(\hat{\theta}(\mathbf{1}))$ with $\nabla_\theta f_{\text{dp}}^{\mathcal{V}}(\hat{\theta}(\mathbf{1}))$. □

Corollary 4.1. *If assumptions in Theorem 3.2 and Corollary 3.4 are satisfied, then Algorithm 1 could reach a feasible and non-trivial solution.*

Proof. For any $\alpha > 0$, following the results from Theorem 3.2 and Corollary 3.4, we shall always find a non-negative $\mathbf{w}^*$ satisfying constraints in Equation (12) and have $\mathcal{I}_{\text{fair}}(\mathbf{w}^*) < 0$. This induces the objective function value in Equation (12) with $\mathbf{w}^*$ smaller than 0. Note that a trivial solution with an all-zero $\mathbf{w}$ makes the objective function zero. □

B. Dataset Statistics

We summarize some key statistics of four datasets we use in experiments in Table 1. For two numerical sensitive variables' %Black Popluation' and 'Age,' we use them and set a threshold to divide the privileged and unprivileged groups. 'Group Pos. Rate' calculates the proportion of samples which has a positive label in the privileged and unprivileged group, respectively. 'ℓ_2 reg.' is L2 regularization strength for the Logistic Regression model, obtained by a grid search over the validation set. The former number is L2 regularization normalized by the number of training samples, while the latter one is the value we set for models. We set 1e-3 as the L2 regularization for Neural Networks as default.

Table 1. Dataset Statistics

Dataset	#Sample (Train / Val. / Test)	#Dim.	Sensitive Attribute	Group Pos. Rate	ℓ_2 reg. for LogReg
Adult	22,622 / 7,540 / 15,060	102	Gender - Male / Female	0.312 / 0.113	1.00e-4 → 2.26
Compas	3,700 / 1,234 / 1,233	433	Race - White / Non-white	0.609 / 0.518	1.00e-2 → 37.00
Comm. and Crime	1,196 / 399 / 399	122	%Black Popluation - 0.06	0.887 / 0.537	2.15e-2 → 25.79
German Credit	600 / 200 / 200	56	Age - 30	0.742 / 0.643	9.75e-3→ 5.85

C. Hyperparameter Selection

We use grid search on the validation to set hyperparameters in Equation (12) and Equation (13). The procedure for Equation (13) is: first, we set $\beta = 0$ and $\gamma = 0$, and observe the performance of fairness. After reaching fairness to a desirable level, we tune γ to control the utility and try to keep its original utility on the validation set. A finer tuning on β is conducted after we fix γ. The minimum interval of β and γ between values is 0.1, with range $[0, 0.9]$ and $[0, 0.4]$, respectively. For α in Equation (12), increasing it from 0 with interval 0.01 with a maximum value of 0.15 can reach our results. The final selections are listed as follows.

Adult: LogReg - EOP: $\beta = 0.5$, $\gamma = 0.2$; NN - EOP: $\beta = 0.5$, $\gamma = 0.2$; LogReg - DP: $\beta = 0.8$, $\gamma = 0.3$; NN - DP: $\alpha = 0.02$.

Compas: LogReg - EOP: $\beta = 0.2, \gamma = 0.1$; NN - EOP: $\beta = 0.2, \gamma = 0.1$; LogReg - DP: $\beta = 0.3, \gamma = 0.1$; NN - DP: $\beta = 0.3, \gamma = 0.1$.

Comm.: LogReg - EOP: $\alpha = 0.1$; LogReg - DP: $\alpha = 0.1$.

German.: LogReg - EOP: $\beta = 0.0, \gamma = 0.0$; LogReg - DP: $\beta = 0.5, \gamma = 0.0$.

For baseline methods with a controllable tradeoff, we use the hyperparamters listed as follows: Dis. remover: repair level $= \{0.25, 0.5, 0.75, 1.0\}$; Label bias: learning rate $= \{0.05, 0.1, 0.5, 1.0\}$; Adv. and Cond. Adv.: $\alpha = \{0.1, 1.0, 5.0, 10.0\}$.

We show a hyperparameters tuning process on the Adult dataset with logistic regression in Figure 3.

D. Additional Results

We present experimental results for the Community and Crime dataset and the German Credit datasets in Figure 4. Our method works consistently across all datasets, and in most cases, we obtain an improvement in fairness at no utility cost.

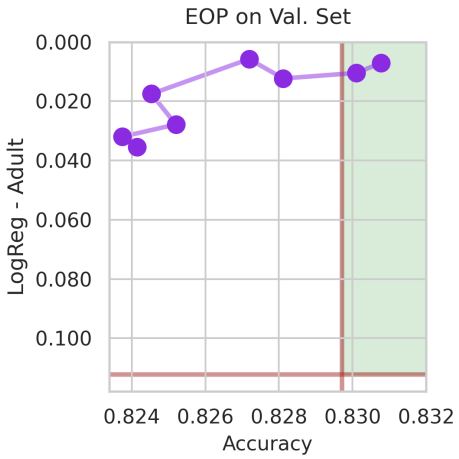

Figure 3. Hyperparameters tuning for Algorithm 1. The lines between nodes show the process of tuning hyperparameters. Initially, the model suffers from group effect of influence function, and the hyperparameters help to mitigate such effect and help the model reach the green region.

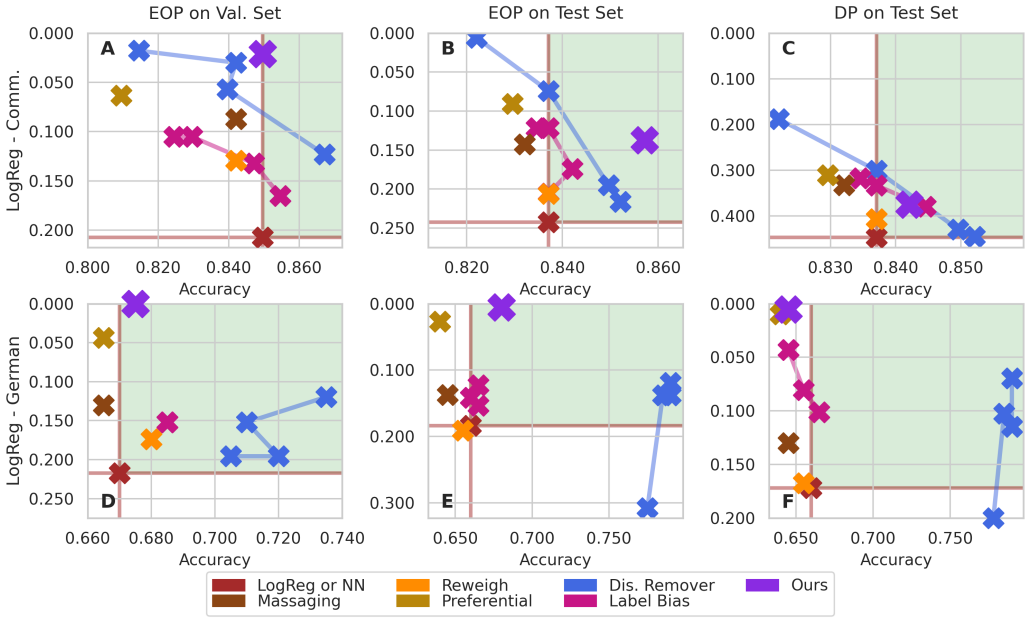

Figure 4. Experimental results on the Communities and Crime and German Credit datasets with LogReg (Logistic Regression) as the base model. For each figure, we indicate the corresponding base model, experimental dataset, fairness metric, and evaluation set in the left and top header (EOP→ Equal Opportunity, DP→ Demographic Parity). Y-axis for fairness is inverted and shows the absolute value of the gap in fairness between the privileged and unprivileged groups (the higher position means the smaller fairness metric value and the better performance towards fairness), while X-axis shows the predictive accuracy. According to the utility and fairness performance of the base model, we plot a horizontal and a vertical line in each figure and divide the space by fairness and utility results into four regions, where the space in green means a fairer and more accurate model compared to the base model. A point closer to the top right indicates better performance in both utility and fairness. For methods with a hyperparameter to control the tradeoff between fairness and utility, the line connecting the same method indicates the monotonously increased change on the hyperparameter.