FairLift: Label correction for fairer predictions in the presence of label bias

Anonymous Author(s)
Affiliation
Address
email

Abstract

When labels in the training data are the result of biased or discriminatory decisions, such biases are encoded in the machine learning models trained on the data. In this work, we revisit the problem of binary classification in the presence of label bias, where the individuals being qualified belong to either a privileged group or a historically unprivileged group. Previous works propose methods for learning classifiers that, in the absence of ground truth labels, maximize a proxy notion of fairness known as disparate impact or demographic parity, while suffering minimal utility loss. We identify two overlooked issues with these approaches. Firstly, depending on the type of label bias, their solutions can be sub-optimal in terms of the accuracy they achieve with respect to the ground truth, unbiased, labels. This is the case when the negative bias towards the unprivileged group is much larger than the positive bias exhibited towards the privileged group. Secondly, these methods can exhibit a leveling-down effect, where the accuracy for *each* group is lower than the accuracy of a *baseline* classifier trained with no fairness constraints.

We then propose FairLift, an algorithm that pre-processes the training data by flipping some of the labels of the unprivileged group, so that the label distribution for the unprivileged group mimics that of the privileged group. We show, both theoretically and empirically, that FairLift achieves superior accuracy to previous methods for the aforementioned scenario of disproportionate bias. In addition, the accuracy of FairLift for the privileged group is never lower than the accuracy of a baseline classifier, and thus FairLift does not exhibit leveling down.

1 Introduction

With machine learning models playing a key decision-making role in areas such as job applications [24], loan applications, and criminal justice decisions, it has become essential to study and mitigate sources of bias in the machine learning pipeline. A fundamental assumption in training such ML models is that the data is generated from a "ground truth" distribution that reflects the true features of individuals. However, societal bias and historical discrimination can manifest both in adverse decisions against a typically underprivileged group and favorable decisions for a privileged group. The biases in decisions, whether conscious or unconscious, lead to data with biased labels. If steps are not taken to mitigate such bias, machine learning models trained on such data and used in decision-making will continue to propagate biased decisions.

In this work, we revisit the problem of fair classification in the presence of label bias. We focus on binary classification as the simplest meaningful instance to explore this problem. We are given a training dataset $D = \{(x_i, s_i, y_i)\}_{i=1}^n$ of $n \in \mathbb{N}$ samples, with features x_i from a domain $\mathcal{X}$, sensitive features $s_i \in \{P, U\}$, and binary labels $y_i \in \{0, 1\}$, for $i \in [n]$. We assume that 1 represents a favorable outcome, e.g., low credit risk or low risk of recidivism. The features are sampled i.i.d from

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

a distribution $\mathcal{D}$ on the data domain $\mathcal{X} \times \{P, U\}$. For each $x \in \mathcal{X}$, there is an associated binary random variable $y_{\text{true}}(x)$ that gives the distribution of its *unbiased* label. The observed, potentially *biased*, labels are sampled from associated binary random variables $y_P(x)$ and $y_U(x)$. Specifically, given sample $(x_i, s_i) \sim \mathcal{X} \times \{P, U\}$ we observe label $y_i \sim y_P(x)$ if $s_i = P$ and label $y_i \sim y_U(x_i)$ if $s_i = U$. The goal it to learn a classifier $f((x, s))$ that predicts labels for unseen examples (x, s). Since only the biased labels can be observed, a classifier trained without any fairness constraints will learn to emulate the biased labels.

In the absence of ground truth labels, optimizing the disparate impact ratio is used as a proxy for fairer decisions. The disparate impact ratio (DIR), also referred to as demographic parity, is the ratio of the proportion of individuals with a favorable outcome of 1 in the privileged group versus the same proportion for the unprivileged group [2, 18, 12]. A classifier with DIR close to 1 has nearly identical distributions of outcomes between the two groups. Compared to other group-fairness definitions, the DIR has the benefit that it is not measured with respect to dataset labels, which, in the presence of label bias, can be unreliable. It is only measured via the classifier outputs and group membership. A measure like equalized odds [14] on the other hand considers false positive rates for each group, where the rates are measured with respect to the potentially biased dataset labels.

A classifier trained with fairness constraints such as disparate impact will aim to maximize DIR while suffering low accuracy loss compared to a baseline classifier trained with no fairness constraints. Usually, accuracy and some fairness measure are the only two measures reported empirically. However, there are infinitely many solutions that achieve the same DIR and accuracy (with respect to the biased labels), but have different utility when accuracy is measured with respect to the unbiased labels. To illustrate this, let $c = \mathbb{E}_x[y_{true}(x)]$ be the ratio of favorable outcomes when labels are unbiased (this ratio we assume is the same for both groups). On the other hand let $a = \mathbb{E}_x[y_P(x)] - \mathbb{E}_x[y_{true}(x)]$ and $b = \mathbb{E}_x[y_{true}(x)] - \mathbb{E}_x[y_U(x)]$. Suppose that $a, b \geq 0$, i.e., there is a surplus of favorable outcomes for the privileged groups when considering the biased labels, and there is a deficit for the unprivileged group. It is impossible to learn c given the observed labels, and a classifier that achieves a disparate impact ratio of 1 will learn a ratio $\hat{c}$ of favorable outcomes for both groups so that $\hat{c} \in [c - b, c + a]$. This classifier will have error of $a + b$ (with respect to the biased labels) for any choice of $\hat{c}$ in the interval. Therefore, infinitely many solution points exists with similar accuracy/DIR tradeoffs. However, these solutions attain different accuracies with respect to the unbiased labels $y_{\text{true}}(x)$.

We show that two previous works, [18] and [17], which study optimizing DIR in the presence of label bias, estimate $\hat{c}$ that lies approximately in the middle of the interval $[c - b, c + a]$, so that $\hat{c} \approx c + \frac{a-b}{2}$. We show such a result theoretically for [18] and empirically for [17], using synthetic data. Then, we observe that when accuracy is measured with respect to the unbiased labels $y_{\text{true}}(x)$ these methods can have unnecessarily high error in cases when $b >> a$ or $a >> b$.

We then introduce FairLift, a method for training classifiers with low disparate impact and show that it achieves superior accuracy to previous work [18, 17], where accuracy is measured with respect to unbiased labels $y_{\text{true}}(x)$, in cases when the bias against the unprivileged group is significantly higher than the bias favoring the privileged group. FairLift is a pre-processing method that corrects labels in the training dataset only for the unprivileged group. It does so by first learning a classifier f_{priv} trained on data from the privileged group only and then uses f_{priv} to predict labels for the unprivileged group. When f_{priv} predicts a label of 1, while the dataset label is 0, the label for the unprivileged individual is flipped from 0 to 1. We introduce a simple bias assumption that, in words, says that bias for the privileged group can only be favorable, while bias for the unprivileged group can only be negative. With this bias assumption, we show that the dataset returned by FairLift has low disparate impact, and even further the label distribution for the unprivileged group is close to the label distribution for the privileged group. Both the disparate impact and the distance between the two distributions is bounded by the error of the classifier f_{priv} in learning the labels of the unprivileged group.

FairLift provides a further advantage: it does not exhibit a leveling-down effect. The leveling-down effect refers to the case when a classifier learned with a fairness constraint lowers the accuracy for both the privileged and unprivileged group as compared to a baseline classifier learned with no fairness constraints [27]. Such an outcome is undesirable since it harms all groups conjointly, especially in situations where accuracy is essential, such as the medical field, and thus fairness approaches that exhibit leveling down would not gain support from any of the populations they are

serving. This recently studied phenomenon [34, 26], has been shown to be pervasive amongst fairness approaches in machine learning. We show both theoretically and empirically that FairLift never lowers the accuracy for the unprivileged group compared to the baseline classifier, and as such it does not achieve a leveling-down effect (here, we consider accuracy with respect to the biased labels).

We compare FairLift empirically to three other methods: the pre-processing technique of [18] and [17] which as already describe aim to achieve high DIR in the presence of label bias, and the post-processing technique of [14] due to its popularity in fairness literature. See Section 1.1 for a discussion of pre-processing vs post-processing methods. Comparing on four datasets, we find that FairLift and the two other pre-processing methods achieve similar DIR/accuracy tradeoffs, while the post-processing method has the lowest performance. For one of the datasets, we can clearly examine that the methods of [18] and [17] achieve a leveling-down effect, whereas FairLift maintains the same accuracy for the privileged group. We also compare these methods on synthetic datasets, which provide the advantage that we can control the type of bias and can measure accuracy with respect to unbiased labels. We can show that for bias regimes where the bias against the unprivileged group is significantly higher than the bias favoring the privileged group, FairLift outperforms the other methods in terms of the accuracy with respect to the unbiased labels.

1.1 Prior work

Most works in the fairness literature implicitly assume that the dataset labels are unbiased and mitigate other sources of bias in the learning process. Fewer works address the issue of label bias explicitly, which we hypothesize is due to the difficulty of measuring the success of bias mitigation measures while lacking access to ground truth labels.

In addition to [18] and [17] which have been discussed earlier, [25] also proposed a pre-processing approach for modifying the input dataset to mitigate label bias. They use the nearest neighbors algorithm to determine whether a label should be flipped. For each individual, its neighbors within a certain distance are examined, and if there is a discrepancy in outcomes amongst these "similar" individuals, then labels are flipped. Their algorithm takes as input the level of "unfairness" allowed (e.g., the level of DIR allowed), rather than optimizing for the best achievable DIR. Thus, we cannot compare to this method directly.

The study of algorithmic fairness necessitates a working notion of what it means for an algorithm to be "fair". Many such notions have been proposed that can be roughly categorized into group fairness measures, individual fairness, or counter-factual fairness [23] (see [6] for a survey). Group fairness measures include disparate impact ratio [2, 18], equalized odds [14], and calibration [28] and its extension, multi-calibration [15]. Such definitions posit that classifier outcomes (such as rate of positive outcomes, false positive rates, or mis-calibration rates) should be balanced between different groups in the population. While all such measures can be desirable, they cannot be simultaneously satisfied (for non-degenerate cases) [5, 21]. With the exception of disparate impact, all other group-fairness notions are measured with respect to the labels in the dataset. In the presence of label bias, such measures can be thus unreliable. Disparate impact is only measured with respect to the labels produced by the classifier, and thus does not rely on the potentially incorrect labels.

While group-fairness measures yield equitable outcomes across groups, they can result in harmful outcomes for specific individuals. Individuals fairness [11] and counter-factual fairness [23] were introduced to bypass the shortcomings of group-fairness measures. However, both frameworks require significant domain and expert knowledge to implement, namely, establishing a similarity metric between individuals or a casual model for the data generation. As such, they present an obstacle to implementation compared to group-fairness measures, which have been more readily adopted.

A diverse set of algorithmic approaches have been proposed to achieve good utility-fairness trade-offs with respect to group-fairness measures. They can be classified into *pre-processing* approaches, where the input dataset is modified prior to the learning process [18, 25, 33, 12, 4]; *in-processing*, where the learning process is modified to incorporate fairness constraints [19, 13, 32, 31, 1, 20, 10, 22, 9, 7, 8]; and *post-processing*, where only the outputs of a classifier are modified to maximize a group fairness objective. Post-processing approaches are the least intrusive to the learning process, however it has been shown that they do not achieve optimal utility-fairness tradeoffs [31]. We also observe this sub-optimality in our empirical evaluation, where we compare to the post-processing approach of [14]. In-processing methods can achieve optimal utility-fairness tradeoffs, however since they modify

the learning algorithm, issues of convergence and convexity arise. Pre-processing methods, such as those proposed in FairLift, [18], and [17], remain a desirable option due to the ease of implementation and the ability to achieve optimal tradeoffs.

2 The FairLift algorithm

In this section, we describe our proposed method FairLift. Then in Section 2.1 and Section 2.2 we show its theoretical guarantees. Namely, we show that after FairLift is applied, the label distribution of the unprivileged group is close to the label distribution of the privileged group. Via this result, we bound the disparate impact difference of the dataset returned by FairLift.

We are given a training dataset $D = \{(x_i, s_i, y_i)\}_{i=1}^n$ of $n \in \mathbb{N}$ samples, with features x_i from a domain $\mathcal{X}$, sensitive features $s_i \in \{P, U\}$, and binary labels $y_i \in \{0, 1\}$, for $i \in [n]$. We assume that 1 represents a favorable outcome, e.g., low credit risk or low risk of recidivism. The features are sampled i.i.d from a distribution $\mathcal{D}$ on the data domain $\mathcal{X} \times \{P, U\}$. For each $x \in \mathcal{X}$, there is an associated binary random variable $y_{\text{true}}(x)$ that gives the distribution of its *unbiased* label. The observed, potentially *biased*, labels are sampled from associated binary random variables $y_P(x)$ and $y_U(x)$. Specifically, given sample $(x_i, s_i) \sim \mathcal{X} \times \{P, U\}$ we observe label $y_i \sim y_P(x)$ if $s_i = P$ and label $y_i \sim y_U(x_i)$ if $s_i = U$.

FairLift, described in Algorithm 1, flips some of the 0 labels in the unprivileged group. It does so by first training a predictor f_{priv} on the privileged group only. The label of an unprivileged individual in the dataset is flipped from 0 to 1 if f_{priv} predicts a 1 for that individual.

Algorithm 1 FairLift

Input: $D = \{(x_i, s_i, y_i)\}_{i=1}^n$

1: Let $D_P = \{(x_i, s_i, y_i) \mid s_i = P, i \in [n]\}$
2: Train f_{priv} on D_P to predict binary labels
3: Let $D_U = \emptyset$
4: **for** $i \in [n]$ such that $s_i = U$: ▷ iterate through the unprivileged group
5: **if** $y_i = 0$ and $f_{\text{priv}}((x_i, P)) = 1$:
6: Let $\tilde{y}_i = 1$ and add $(x_i, U, \tilde{y}_i)$ to D_U
7: **else** add (x_i, U, y_i) to D_U
8: **Return** $D_P \cup D_U$.

In practice, we find that FairLift achieves a better performance if we incorporate a calibration step, where we learn a threshold for flipping labels in the unprivileged group. For completeness we state this algorithm (Algorithm 2) in Appendix A. It differs from Algorithm 1 only in Step 6, where the label of an unprivileged individual is flipped from 0 to 1 if f_{priv} predicts a 1 with probability above the learned threshold. The threshold is learned via a validation set, with the objective of minimizing the disparate impact difference in the dataset. For a threshold of 0.5, then Algorithm 2 is exactly Algorithm 1. Theoretically, the properties of Algorithm 1 are simpler to analyze.

2.1 FairLift equalizes label distributions for the two groups

In this section, we prove Theorem 2.3 which says that after FairLift is applied, the label distribution of the unprivileged group is close to the label distribution of the unprivileged group. The closeness parameter is upper-bounded by the error of the classifier f_{priv} in learning the labels of the privileged group. This result is obtained under the following assumption on label bias, which can be thought of as a counterfactual statement on the label generation process. It says that if a labeler were to provide a label 0 for an individual in the privileged group, they would provide the same label for an individual with the exact same features belonging to the unprivileged group. In Ex. C.1 we provide an example of random variables $y_P(x)$ and $y_U(x)$ that satisfy Assumption 2.1.

Assumption 2.1. *The random variables $y_P(x)$ and $y_U(x)$ that generate biased labels satisfy*

$$\Pr_{x \sim \mathcal{X}}[y_U(x) = 0 \mid y_P(x) = 0] = 1.$$

Dataset	#records	# features	Target variable	Protected Attribute
Compas	6,150	10	two-year recidivism	race
German Credit Risk	1,000	20	credit risk	age
UCI Adult	48,842	14	salary>$50k/yr	gender
Communities and Crime	1,994	22	per-capita capital crimes	race

Table 1: Summary of datasets.

In Theorem 2.3 we use total variation (TV) distance as the notion of distribution distances. We remark that Theorem 2.3 holds for other valid notions of distribution distances. However, TV distance is most amenable to obtaining our result in Section 2.2 on disparate impact difference.

Definition 2.2 (Total variation distance)**.** *Given two distributions P and Q over a domain $\mathcal{D}$, the total variation distance between P and Q is denoted by $\mathsf{TV}(P, Q)$ and is defined as*

$$\mathsf{TV}(P, Q) = \sup_{S \in \mathcal{D}} |P(S) - Q(S)|.$$

Theorem 2.3. *Let f_{priv} be the function learned on the privileged group in Step 2 of Algorithm 1. For $x \in \mathrm{X}$, let $\hat{y}_{\mathrm{U}}(x)$ be a binary random variable that gives the label distribution for individuals in the unprivileged group after FairLift (Algorithm 1) is applied. Let $\epsilon \geq 0$ be the error of f_{priv} in learning the labels of the privileged group, i.e.,*

$$\varepsilon = \mathbb{E}_{x \sim \mathcal{X}}[\mathsf{TV}(f_{\text{priv}}((x, P), y_P(x))]$$

If Assumption 2.1 holds then

$$\mathbb{E}_{x \sim \mathcal{X}}[\mathsf{TV}(\hat{y}_{\mathrm{U}}(x), y_P(x))] \leq \varepsilon.$$

Proof: See Appendix B.1.

See Appendix C for an example of biased label random variables $y_P(x)$ and $y_U(x)$ that satisfy Assumption 2.1.

2.2 FairLift achieves low disparate impact difference

In this section, we use Theorem 2.3 to prove Theorem 2.4. Theorem 2.4 bounds the disparate impact difference of the dataset returned by FairLift with the error of the classifier f_{priv} learned on the privileged group. When this error is low, the returned dataset has low disparate impact difference (in expectation).

Theorem 2.4 (Disparate impact difference after FairLift)**.** *Consider the setup of Theorem 2.3. The expected disparate impact difference of the dataset returned by FairLift (Algorithm 1) has disparate impact difference bounded by ε.*

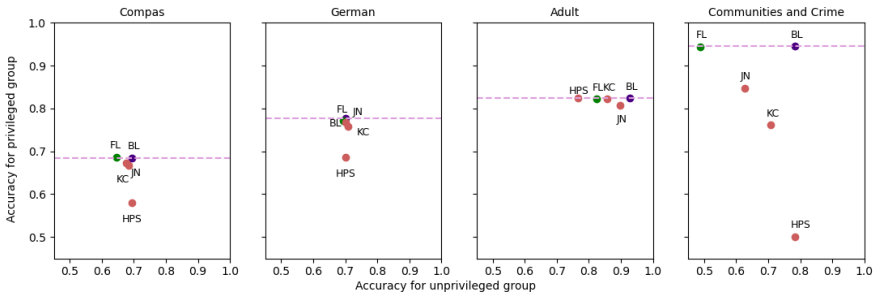
Proof: See Appendix B.2

2.3 FairLift does not level-down

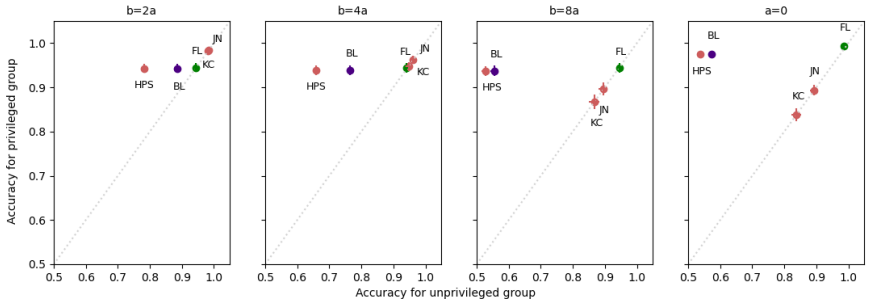
An interesting question not raised in the previous literature is the tradeoffs in the accuracies of privileged and underprivileged classes. We explore this topic both theoretically and empirically (using synthetic data). We derive the accuracies for the Baseline, Massaging method [18], and FairLift algorithms in Appendix E. A key insight of this analysis is that FairLift does not lower the accuracy of the privileged group to increase the accuracy for the underprivileged group.

Corollary 2.5 (FairLift does not level-down)**.** *Suppose we have access to classifiers that learn label distributions perfectly. Under mild assumptions, FairLift and the Massaging method of [18] have the same overall expect error of $a + b$ with respect to the biased labels, where a is the favorable bias for the privileged group and b is the unfavorable bias for the underprivileged group. However FairLift does not level-down, whereas the massaging method lowers the accuracy for both groups compared to the baseline method.*

Proof: See Appendix E.



(a) Accuracy for the privileged group vs unprivileged group of classifiers trained with each method (BL: baseline, FL: FairLift, KC: [18], JN: [17], HPS: [14]). The dashed pink line shows the accuracy of the Baseline method for the privileged group.



(b) Accuracy for the privileged group vs unprivileged group with respect to the unbiased labels in synthetic datasets (BL: baseline, FL: FairLift, KC: [18], JN: [17], HPS: [14]). The values a and b are the fraction of individuals with biased labels in the privileged and unprivileged groups, respectively.

Figure 1: Accuracy plots for privileged vs unprivileged groups on real and synthetic datasets.

3 Empirical Evaluation

In this section, we compare FairLift with three previous works: the pre-processing techniques of [18] and [17], and the post-processing technique of [14]. We also compare to a "Baseline" method where a classifier is trained without any fairness constraints. We compare on four existing datasets (Table 1) and on synthetic datasets. Synthetic datasets have the advantage that we can also measure accuracy of the classifiers with respect to the unbiased labels y_{true}.

For the pre-processing methods, a classifier is trained on the modified training dataset returned by each method. For the post-processing method, a classifier is trained on the training set and the threshold for binarizing the outputs of the classifier is chosen using the same set. The accuracy and disparate impact ratio of the obtained classifiers is then measured on a standalone test set. For all methods and experiments we train a logistic regression classifier with default hyperparameters from the package `scikit-learn`. Additional details can be found in Appendix F.2.

Datasets: We evaluate our method on four datasets commonly used in Fairness research: Compas, German Credit Risk, UCI Adult, and Communities and Crime. The details of the datasets are summarized in Table 1. Since the datasets are well-known in the fairness community, we defer the description of the datasets to Appendix F.1. For all datasets except the Adult dataset, we use random 80% of the data for training and the remaining 20% for testing. For the Adult dataset, we use the standard train-test split.

Metrics: We report Accuracy, Disparate Impact Ratio, and the harmonic mean of the two. We use the harmonic mean as the overall score.

System: We run all experiments on MacBook Pro laptops without GPUs. The experiments take a few minutes to complete.

(a)

	Overall Accuracy				Disparate Impact Ratio			
	Compas	German	Adult	Crime	Compas	German	Adult	Crime
Baseline	**0.6902**	0.7490*	**0.8582**	**0.8613**	0.7465	0.8344	0.8243	0.5059
FairLift	0.6611	0.7424	0.8227	0.7086	0.9689	0.9314*	0.9802*	**0.9966**
Kamiran and Calders [18]	0.6771	0.7398*	0.8333	0.7330	0.9721	0.9407*	0.9951*	0.9521
Jiang and Nachum [17]	0.6742	0.7428*	0.8377	0.7332	**0.9757**	0.9465*	0.9974*	0.9709
Hardt et al [14]	0.6484	0.6912	0.8049	0.6470	0.9749	0.9521	0.9951	0.9629

(b)

	Harmonic Mean			
	Compas	German	Adult	Crime
Baseline	0.7170	0.7883	0.8409	0.6366
FairLift	0.7858	0.8253*	0.8945	0.8282
Kamiran and Calders [18]	**0.7980**	0.8277	0.9070	0.8278*
Jiang and Nachum [17]	0.6742	0.8318*	**0.9106**	**0.8354**
Hardt et al [14]	0.7786	0.8002	0.8900	0.7735

Table 2: The disparate impact ratio and accuracy of classifiers (Table a) and their harmonic mean (Table b) obtained from four bias-correction methods. Comparison is performed on four datasets. All numbers are statistically significantly different at $p = 0.05$ compared to FairLift except those indicated by *. Number of trials is 50.

3.1 Results on real-world datasets

In Table 2 we show the harmonic mean of accuracy and disparate impact ratio of the five methods. We see that FairLift achieves higher or comparable DIR to the pre-processing methods of [17] and [18]. The DIR of the baseline method is as expected low for all 4 datasets. The post-processing method of [14] under-performs all other fairness approaches for 3 out of 4 of the datasets. In terms of accuracy, FairLift has comparable accuracy to the pre-processing methods. The post-processing method of [14] has the lowest accuracy across all datasets.

Then in Fig. 1a we examine the accuracy of each method for the privileged and unprivileged group respectively. For all datasets, FairLift never lowers the accuracy of the privileged group compared to the baseline classifier, which supports the claim that FairLift does not exhibit a leveling-down effect. For the "Communities and Crime" dataset, we especially see that the other two pre-processing methods lower the accuracy for both groups. On the other hand, we see that the post-processing method of [14] sacrifices accuracy on the privileged group while maintaining accuracy comparable to the baseline method for the unprivileged group.

3.2 Synthetic data generation

For each experiment, we generate a synthetic dataset of $n = 20000$ datapoints and 101 features, where one of the features is the protected attribute. The true labels y_{true} are generated following a linear regression model and are independent of the protected feature. Given feature vector $x \in \mathbb{R}^{10}$, coefficients $c \in \mathbb{R}^{10}$, and a noise random variable $\epsilon \in \mathbb{R}$, we let

$$y_{\text{true}}(x) = f(c^\intercal x + \varepsilon), \quad \text{where} \quad f(x) = \frac{1}{1+e^{-x}}.$$

The features $x[j]$ are sampled iid from $\mathcal{N}(0, 5)$, the coefficients $b[j]$ are sampled uniformly from the interval $[-4, 4]$, and the noise follows $\epsilon \sim \mathcal{N}(0, 1)$. We sample n datapoints $(x_i, y_{\text{true}}(x_i))$ with $i \in [n]$ in this way. To assign protected attributes s_i, a fraction $p_{\text{priv}} = 0.6$ of the dataset are randomly selected to belong to the privileged group (i.e., they have $s_i = 1$). The rest belong to the unprivileged group.

To obtain the biased labels y_i, let $a \in [0, 1]$ and $b \in [0, 1]$ denote the extent of bias for the privileged and unprivileged group respectively (see (10)). A random fraction a of the privileged individuals have their true label of 0 flipped to 1. A random fraction b of the unprivileged individuals have their true label of 1 flipped to 0. The final dataset is $\{(x_i, s_i, y_i)\}_{i=1}^n$. We consider different values of a and b in our experiments.

(a)

	Accuracy on unbiased labels				Disparate Impact Ratio			
	$b = 2a$	$b = 4a$	$b = 8a$	$a = 0$	$b = 2a$	$b = 4a$	$b = 8a$	$a = 0$
Baseline	0.9202	0.8695	0.7844	0.8138	0.8319	0.7048	0.4941	0.5747
FairLift	0.9444	0.9428	**0.9444**	**0.9895**	0.9783	0.9800	0.9812	0.9836
Kamiran and Calders [18]	0.9836	0.9472	0.8676	0.8380	0.9858	0.9850	0.9847*	0.9879
Jiang and Nachum [17]	**0.9814**	**0.9619**	0.8956	0.8935	0.9855	0.9858	0.9843*	0.9867
Hardt et al [14]	0.8274	0.7695	0.7226	0.7872	**0.9863**	**0.9879**	**0.9880**	0.9874

(b)

	Harmonic Mean			
	Compas	German	Adult	Crime
	$b = 2a$	$b = 4a$	$b = 8a$	$a = 0$
Baseline	0.8737	0.7784	0.6062	0.6736
FairLift	0.9610	0.9610	**0.9624**	**0.9865**
Kamiran and Calders [18]	**0.9847**	**0.9657**	0.9224	0.9067
Jiang and Nachum [17]	0.9834	0.9737	0.9378	0.9377
Hardt et al [14]	0.8999	0.8651	0.8346	0.8760

Table 3: The disparate impact ratio and accuracy of classifiers (Table a) and their harmonic mean (Table b) obtained from four bias-correction methods. Comparison is performed on synthetic datasets with varying levels of labels bias a and b for the privileged and unprivileged group respectively. All numbers are statistically significantly different at $p = 0.05$ compared to FairLift except those indicated by *. Number of trials is 100.

3.3 Results on synthetic datasets

For the synthetic data experiments, we vary $(a, b) \in \{(0.05, 0.1), (0.05, 0.2), (0.05, 0.4), (0, 0.4)\}$. That is, we experiment with b/a ratios of $2, 4, 6$ and ∞ (when $a = 0$). In Fig. 1b, we plot the accuracy of each method for the privileged vs unprivileged group with respect to the *unbiased* labels $y_{\text{true}}(x)$ (these labels are not used in any part of the training process). The accuracy is averaged over 50 trials and 95% confidence intervals are shown in error bars. Since the data generation process is quite simple and the logistic regression classifier fits the data fairly well, we are able to replicate the theoretical insights from Appendix E.

We first observe that, surprisingly, the post-processing method of [14] under-performs even the baseline method. All other methods improve on the accuracy of the baseline method. Note that while the baseline method has higher accuracy when using the biased labels, as is the case with the real-world datasets in Table 2, it performs worse than the pre-processing methods when accuracy is measured with respect to the unbiased labels, since the pre-processing methods account for the label bias. As the ratio b/a increases (i.e., the negative bias for the unprivileged group becomes proportionately higher than the positive bias for the privileged group), we observe that FairLift starts to outperform the two other pre-processing methods. The gap is most significant when $a = 0$, in which case FairLift achieves nearly perfect accuracy for both groups. On the other hand, when $b < 3a$, we see that FairLift is outperformed by the two other pre-processing methods, as predicted by Theorem E.4 in the Appendix.

In Table 3 we show the harmonic mean of accuracy and DIR of each method. Again, accuracy is measured with respect to the unbiased labels. All fairness methods achieve similar DIR, and the DIR is above 0.9. Accuracy-wise, FairLift outperforms the other pre-processing methods for higher ratios of b/a. Notice that while [14] has amongst the highest DIR, it also has the lowest accuracy compared to all other methods. This demonstrates the need for a more refined understanding of how various method achieve fairness, beyond a simple DIR/accuracy trade-off, especially when the accuracy is measured with respect to potentially biased labels. Synthetic data provide a good starting point for examining the advantages of each method for different sources of bias.

4 Limitations

FairLift is evaluated on standard benchmarks. The benchmarks contain only numerical and categorical data. We have not evaluated the approach on other modalities like text.

5 Conclusion

We have proposed and validated FairLift algorithm which identifies and corrects biased labels. We have also benchmarked the performance of FairLift, both theoretically and experimentally, with relevant approaches. FairLift is intuitively simple and the performance is competitive with related approaches. We have also explored bias regimes by defining bias for the privileged class and bias against the unprivileged class and how the two are related. Through the use of synthetically generated datasets, we have shown that different algorithms perform well in different bias regimes and that FairLift outperforms other approaches when the bias against the unprivileged class is high. These observations have the potential to lead to further studies to quantify the extent of bias and algorithms to address them. This will in turn have positive effects on the society.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In *Proceedings, International Conference on Machine Learning (ICML)*, volume 80, pages 60–69, 2018.

[2] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104:671, 2016.

[3] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[4] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, abs/1707.00075, 2017.

[5] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

[6] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018.

[7] Andrew Cotter, Maya R. Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Lutong Wang, Blake E. Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *Proceedings, International Conference on Machine Learning (ICML)*, volume 97, pages 1397–1405, 2019.

[8] Andrew Cotter, Heinrich Jiang, Maya R. Gupta, Serena Lutong Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *J. Mach. Learn. Res.*, 20:172:1–172:59, 2019.

[9] Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. In *Proceedings, International Conference on Algorithmic Learning Theory (ALT)*, volume 98, pages 300–332, 2019.

[10] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2796–2806, 2018.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Proceedings, Innovations in Theoretical Computer Science (ITCS)*, pages 214–226. ACM, 2012.

[12] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[13] Gabriel Goh, Andrew Cotter, Maya R. Gupta, and Michael P. Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2415–2423, 2016.

[14] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3315–3323, 2016.

[15] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings, International Conference on Machine Learning (ICML)*, volume 80, pages 1944–1953, 2018.

[16] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[17] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *Proceedings, International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 702–712, 2020.

[18] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1):1–33, 2011.

[19] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *ECML/PKDD (2)*, volume 7524, pages 35–50, 2012.

[20] Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings, International Conference on Machine Learning (ICML)*, volume 80, pages 2569–2577, 2018.

[21] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[22] Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. Nonconvex optimization for regression with fairness constraints. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2742–2751. PMLR, 2018.

[23] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4066–4076, 2017.

[24] Stella Lowry and Gordon Macpherson. A blot on the profession. *British Medical Journal (Clinical research ed.)*, 296:657 – 658, 1988.

[25] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 502–510, 2011.

[26] Gaurav Maheshwari, Aurélien Bellet, Pascal Denis, and Mikaela Keller. Fair without leveling down: A new intersectional fairness definition. 2023.

[27] Brent Daniel Mittelstadt, Sandra Wachter, and Chris Russell. The unfairness of fair machine learning: Levelling down and strict egalitarianism by default. *ArXiv*, abs/2302.02404, 2023.

[28] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *NIPS*, pages 5680–5689, 2017.

[29] ProPublica. Compas recidivism risk score data and analysis, 2016.

[30] Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C53W3X.

[31] Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. volume 65, pages 1920–1953, 2017.

[32] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings, International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 962–970, 2017.

[33] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings, International Conference on Machine Learning (ICML)*, volume 28, pages 325–333, 2013.

[34] Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Scholkopf, and Chris Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10400–10411, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the claims contained in both the abstract and the introduction are substantiated in the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 discusses the limitations of the approach.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The proofs are provided in the Appendix.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix F.2 provides additional details on experimental settings in addition to those listed in the main paper.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We are getting approval for open-sourcing the code and will release the code after organizational approval. [No] depicts the worse-case scenario of not getting the approval.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: **[TODO]**

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experimental results indicate statistical significance.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: All the experiments are performed on easily available resources. We have used MacBook Pro (M2) without GPUs for the experiments.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [NA]

   Justification: We have not used human subjects or confidential data sources.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We believe this works improves algorithmic fairness.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper uses small-scale publicly available datasets.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have used publicly available datasets which have been used in earlier studies.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code, if released, is well-documented.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We have not used crowdsourcing or human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We have not used human subjects.

A FairLift with calibration

In this section, we formally state our algorithm FairLift with the calibration step. The label of an unprivileged individual is flipped from 0 to 1 if f_{priv} predicts a 1 with probability above a learned threshold. The threshold is learned via a validation set, with the objective of minimizing the disparate impact difference in the dataset. For a threshold of 0.5, then Algorithm 2 is exactly Algorithm 1.

Algorithm 2 FairLift with Calibration

Input: $D = \{(x_i, s_i, y_i)\}_{i=1}^n$, flipping threshold

Let $D_P = \{(x_i, s_i, y_i) \mid s_i = P, i \in [n]\}$
Train f_{priv} on D_P to predict $\Pr[y_i = 1 \mid x_i, s_i = P]$
Let $D_U = \emptyset$
for $i \in [n]$ such that $s_i = U$: ▷ iterate through the unprivileged group
    if $y_i = 0$ and $f_{\text{priv}}((x_i, P)) >$ threshold :
        Let $\tilde{y}_i = 1$ and add $(x_i, U, \tilde{y}_i)$ to D_U
    else add (x_i, U, y_i) to D_U
Return $D_P \cup D_U$.

B Proofs

B.1 Proof of Theorem 2.3

Proof. We first observe that by Step 6 of Algorithm 1,

$$\hat{y}_{\text{U}}(x) = f_{\text{priv}}((x, P)) \vee y_U(x), \tag{1}$$

where $\vee$ denotes the logical OR function. By the law of total expectation:

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(\hat{y}_{\text{U}}(x), y_P(x))] \\
&= \mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(\hat{y}_{\text{U}}(x), y_P(x)) \mid y_P(x) = 1] \Pr[y_P(x) = 1] + \mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(\hat{y}_{\text{U}}(x), y_P(x)) \mid y_P(x) = 0] \Pr[y_P(x) = 0]
\end{aligned} \tag{2}$$

By Assumption 2.1, conditioned on $y_P(x) = 0$, then $y_U(x) = 0$. From the observation in (1), this implies that $\hat{y}_{\text{P}}(x) = f_{\text{priv}}(x)$, conditioned on $y_P(x) = 0$. We obtain

$$\mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(f_{\text{priv}}(x), y_P(x)) \mid y_P(x) = 0] = \mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(f_{\text{priv}}(x), y_P(x)) \mid y_P(x) = 0]. \tag{3}$$

We apply the law of total expectation one more time:

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(\hat{y}_{\text{U}}(x), y_P(x)) \mid y_P(x) = 1] \\
&\le \mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(\hat{y}_{\text{U}}(x), y_P(x)) \mid y_P(x) = 1, y_U(x) = 0] + \mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(\hat{y}_{\text{U}}(x), y_P(x)) \mid y_P(x) = 1, y_U(x) = 1] \\
&= \mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(f_{\text{priv}}(x), y_P(x)) \mid y_P(x) = 1, y_U(x) = 0] + \mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(1, 1)] && (4) \\
&= \mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(f_{\text{priv}}(x), y_P(x)) \mid y_P(x) = 1] + 0 && (5) \\
&= \mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(f_{\text{priv}}(x), y_P(x)) \mid y_P(x) = 1] && (6)
\end{aligned}$$

Eqn. (4) follows from the observation in (1), namely $\hat{y}_{\text{U}}(x) = f_{\text{priv}}(x)$ conditioned on $y_U(x) = 0$ and $\hat{y}_{\text{U}}(x) = 1$ conditioned on $y_U(x) = 1$. Eqn. (5) follows from the fact that $f_{\text{priv}}(x)$ and $y_P(x)$ are independent of $y_U(x)$ once $y_P(x)$ is fixed, and the fact that the TV distance of two identical distributions is 0.

Replacing (6) and (3) into (2) and applying the total law of expectation one more time we obtain:

$$\mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(\hat{y}_{\text{U}}(x), y_P(x))] \le \mathbb{E}_{x\sim\mathcal{X}}[\mathsf{TV}(f_{\text{priv}}(x), y_P(x))] \le \varepsilon.$$

■

B.2 Proof of Theorem 2.4

Proof. Recall that $D = \{(x_i, s_i, y_i)_{i=1}^n\}$ is the original training dataset. Let $\hat{y}_i$ for $i \in [n]$ be the labels of the dataset returned by FairLift. Note that $\hat{y}_i = y_i$ if $s_i = P$ and thus $y_i \sim y_P(x_i)$. If $s_i = U$, then $\hat{y}_i \sim \hat{y}_{\text{U}}(x_i)$. From an application of Theorem 2.3 we obtain:

$$\begin{aligned}
&\mathbb{E}_D\Big[\frac{\#\{i \mid s_i = P, \hat{y}_i = 1\}}{\#\{i \mid s_i = P\}} - \frac{\#\{i \mid s_i = U, \hat{y}_i = 1\}}{\#\{i \mid s_i = U\}}\Big] \\
&= \mathbb{E}_{(x,s,y)}\Big[\Pr[\hat{y} = 1 \mid x, s = P] - \Pr[\hat{y} = 1 \mid x, s = U]\Big] \\
&= \mathbb{E}_x[\Pr[y_P(x) = 1] - \Pr[\hat{y}_{\text{U}}(x) = 1]] \\
&\le \mathbb{E}_x[|\Pr[y_P(x) = 1] - \Pr[\hat{y}_{\text{U}}(x) = 1]|] \\
&= \mathbb{E}_x[\mathsf{TV}(y_P(x), \hat{y}_{\text{U}}(x))] \le \varepsilon. && \text{(by Theorem 2.3)}
\end{aligned}$$

■

C Example

This section provides an example of biased label random variables $y_P(x)$ and $y_U(x)$ that satisfy Assumption 2.1.

Example C.1. *Consider the following model of biased labels where for an individual in the privileged group their true label of* 0 *is flipped to* 1 *with some probability* $p_1 > 0$ *and for an individual in the unprivileged group their true label of* 1 *is flipped to* 0 *with some probability* p_2*. Formally, let* $Z_1 \sim \text{Ber}(p_1)$, $Z_2 \sim \text{Ber}(p_2)$*. Then for all* $x \in \mathcal{X}$*:*

$$y_P(x) = y_{\text{true}}(x) + Z_1 \cdot (1 - y_{\text{true}}(x)),$$
$$y_U(x) = (1 - Z_2) \cdot y_{\text{true}}(x).$$

Note that $y_P(x) = 1$ *conditioned on* $y_{\text{true}}(x) = 1$ *and* $y_P(x) = Z_1$ *if* $y_{\text{true}}(x) = 0$*. On the other hand,* $y_U(x) = 0$ *conditioned on* $y_{\text{true}}(x) = 0$*, and* $y_U(x) = 1 - Z_2$ *conditioned on* $y_{\text{true}}(x) = 1$*. Note that conditioned on* $y_P(x) = 0$*, then* $y_{\text{true}}(x) = 0$*. This fact, combined with the fact that* $y_U(x)$ *is independent of* $y_P(x)$ *when conditioning on* $y_{\text{true}}(x)$*, gives that we only need to consider* $\Pr[y_U(x) = 0 \mid y_{\text{true}}(x) = 0]$*. The latter equals* 1*, as desired. More formally:*

$$\begin{aligned}
&\Pr[y_U(x) = 0 \mid y_P(x) = 0] \\
&= \Pr[y_U(x) = 0 \cap y_{\text{true}}(x) = 0 \mid y_P(x) = 0] + \Pr[y_U(x) = 0 \cap y_{\text{true}}(x) = 1 \mid y_P(x) = 0] \\
&= \Pr[y_U(x) = 0 \mid y_P(x) = 0, y_{\text{true}}(x) = 0] \Pr[y_{\text{true}}(x) = 0 \mid y_P(x) = 0] + \frac{\Pr[y_U(x) = 0 \cap y_{\text{true}}(x) = 1 \cap y_P(x) = 0]}{\Pr[y_P(x) = 0]} && (7) \\
&= \Pr[y_U(x) = 0 \mid y_{\text{true}}(x) = 0] \Pr[y_{\text{true}}(x) = 0 \mid y_P(x) = 0] + 0 && (8) \\
&= 1 \cdot 1 = 1. && (9)
\end{aligned}$$

In (8) we apply the fact that $y_U(x) \perp y_P(x) \mid y_{\text{true}}(x)$ *and the fact that* $\Pr[y_P(x) = 0 \cap y_{\text{true}}(x) = 1] = 0$*. In (9) we use* $\Pr[y_U(x) = 0 \mid y_{\text{true}}(x) = 0] = 1$ *and* $\Pr[y_{\text{true}}(x) = 0 \mid y_P(x) = 0] = 1$*, which are true by the definition of* $y_U(x)$ *and* $y_P(x)$*.*

D Massaging method of Kamiran and Calders

In this section, we formally state the massaging method of [18]. Recall that the massaging method first learns a function $f_{\textbf{rank}}$ that predicts the probability that an individual receives a favorable outcome 1. Then m_P individuals from the privileged group with the lowest values of $f_{\textbf{rank}}$ have their labels in the dataset flipped from 1 to 0. Similarly, m_U individuals in the unprivileged group with the highest rank have their labels of 0 flipped to 1. [18] choose m_P and m_U so that after the label flipping the disparate impact difference in the training set is 1. There are many such choices of m_P and m_U, but they fix $m_P = m_U$. The method we state in Algorithm 3 differs slightly in that we select m_P and m_U so that an *equal ratio* of individuals are flipped within each group. This version of the algorithm is cleaner to analyze for our results in Theorem E.4 and Theorem E.6. Empirically. we find that this different choice of m_P and m_U has not effect on performance.

E Accuracy guarantees with respect to biased and unbiased labels

In this section, we analytically compare the expected accuracy of our method, FairLift, with the Massaging technique of [18], and the baseline method, where the classifier is trained without fairness constraints. To obtain our results, we make a simple assumption on the label bias present in the data:

Assumption E.1 (Bias assumption)**.** *The bias variables* $y_P(x)$ *and* $y_U(x)$ *satisfy*

$$\Pr_x[y_P(x) = 1 \mid y_{\text{true}}(x) = 1] = 1 \quad \textit{and} \quad \Pr_x[y_U(x) = 0 \mid y_{\text{true}}(x) = 0] = 1.$$

In addition, for all $x \in \mathcal{X}$ *we have* $y_P(x) \perp y_U(x) \mid y_{\text{true}}(x)$

In other words, Assumption E.1 says that labels for the privileged group can only be flipped in the positive direction (from 0 to 1) when compared to the true labels, whereas labels in the unprivileged group can only be flipped in the negative direction (from 1 to 0) when compared to the true labels. The assumption on conditional independence gives that Assumption E.1 implies Assumption 2.1.

Algorithm 3 Massaging method of [18]

Input: Training dataset $D = \{(x_i, s_i, y_i)\}_{i=1}^n$ with $s_i \in \{P, U\}$ and $y_i \in \{0, 1\}$ for all $i \in [n]$

1: Train a ranker $f_{\mathbf{rank}}$ on D to predict $\Pr[y_i = 1 \mid (x_i, s_i)]$
2: Let $n_P = \#\{i \in [n] \mid s_i = P\}$ and $n_U = \#\{i \in [n] \mid s_i = U\}$
3: Let $c = \frac{1}{n_P} \cdot \#\{i \mid s_i = P, y_i = 1\} - \frac{1}{n_U} \cdot \#\{i \mid s_i = U, y_i = 1\}$. ▷ disparate impact difference of D
4: Let $D_P \subseteq D$ be the subset of $m_P = c \cdot n_P$ privileged individuals with label 1 with the lowest values of f_{rank}
5: Let $D_U \subseteq D$ be the subset of $m_U = c \cdot n_U$ unprivileged individuals with label 0 with the highest values of f_{rank}
6: Let $D_{\text{unflipped}} = (D \setminus D_U) \setminus D_P$
7: Let $\tilde{D}_P = \{(x_i, s_i, 0) \mid (x_i, s_i, 1) \in D_P\}$ and $\tilde{D}_U = \{(x_i, s_i, 1) \mid (x_i, s_i, 0) \in D_U\}$ ▷ flip labels
8: Return the dataset with the flipped labels: $D_{\text{unflipped}} \cup \tilde{D}_U \cup \tilde{D}_P$

With this bias assumption, we provide guarantees on the accuracy of the three methods with respect to both the unbiased labels ($y_{\text{true}}(x)$) in Section E.1 and the biased labels ($y_P(x)$ and $y_U(x)$) in Section E.2. The biased labels $y_P(x)$ and $y_U(x)$ are the only labels that can be observed, and thus data analysts can only report accuracy with respect to these labels. On the other hand, while the labels $y_{\text{true}}(x)$ cannot be observed, we are still able to provide guarantees in terms of these labels, which allows us to compare the performance of the three different methods.

To obtain our results, we assume access to classifiers that learn label distributions perfectly, given access to data. While the assumption is unrealistic, recall that even with access to such classifiers, learning the unbiased labels in the presence of label bias is impossible. More importantly, this assumption allows us to understand how each method shifts the label distribution, and when one is preferable to the other depending on the extent of bias. Finally, as we show in Section 3 with our synthetic data experiments, when we have access to classifiers that learn with low error, the empirical results nearly match our theoretical results.

As it turns out, all our accuracy guarantees can be expressed in terms of the two quantities:

$$\begin{aligned} a &:= \mathbb{E}_x[y_P(x)] - \mathbb{E}_x[y_{\text{true}}(x)] \\ b &:= \mathbb{E}_x[y_{\text{true}}(x)] - \mathbb{E}_x[y_U(x)]. \end{aligned} \tag{10}$$

The quantities a and b represent the extent of bias in the privileged and unprivileged populations, respectively. Due to our bias assumption (Assumption E.1), we have that $a \in [0, 1]$ and $b \in [0, 1]$.

We end this section by proving two claims that will be useful in the proofs of Theorems E.4 and E.6.

Claim E.2. *If Assumption E.1 holds then*

$$a = \Pr_{x \sim \mathcal{X}}[y_P(x) = 1 \textit{ and } y_{\text{true}}(x) = 0] \quad \textit{and} \quad b = \Pr_{x \sim \mathcal{X}}[y_U(x) = 0 \textit{ and } y_{\text{true}}(x) = 1]. \tag{11}$$

Proof. We prove the statement for a. A similar proof holds for b.

$$\begin{aligned} a &= \mathbb{E}_x[y_P(x)] - \mathbb{E}_x[y_{\text{true}}(x)] = \Pr[y_P(x) = 1, y_{\text{true}}(x) = 1] + \Pr[y_P(x) = 1, y_{\text{true}} = 0] - \Pr[y_{\text{true}}(x) = 1] \\ &= \Pr[y_P(x) = 1 \mid y_{\text{true}}(x) = 1] \cdot \Pr[y_{\text{true}}(x) = 1] + \Pr[y_P(x) = 1, y_{\text{true}} = 0] - \Pr[y_{\text{true}}(x) = 1] \\ &= 1 \cdot \Pr[y_{\text{true}}(x) = 1] + \Pr[y_P(x) = 1, y_{\text{true}} = 0] - \Pr[y_{\text{true}}(x) = 1] \\ &= \Pr[y_P(x) = 1, y_{\text{true}}(x) = 0]. \end{aligned}$$

In the third equality we used the fact that $\Pr[y_P(x) = 1 \mid y_{\text{true}}(x) = 1] = 1$ from Assumption E.1. ■

Claim E.3. *The expected disparate impact difference of the dataset $D = \{(x_i, s_i, y_i)_{i=1}^n\}$ equals $a + b$.*

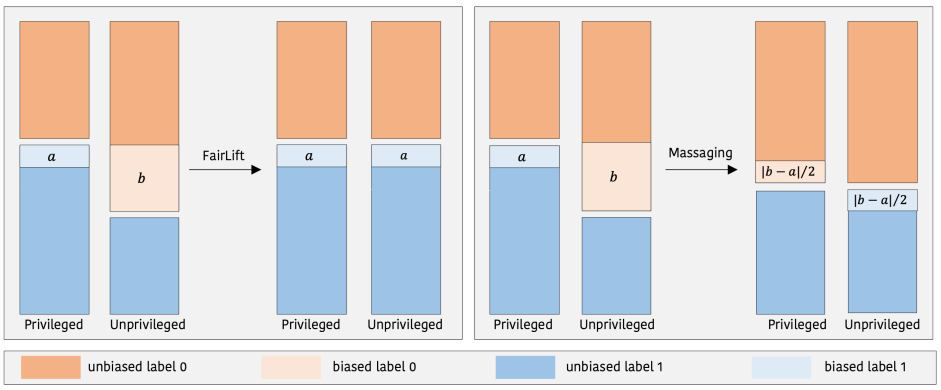


Figure 2: Illustration of the effect of FairLift and the Massaging method of [18] on the dataset labels for the privileged and unprivileged groups. The label distribution in the dataset is shown before and after each method is applied. Here, a is the fraction of privileged individuals with a biased label 1 (but a true label 0). The value b is the fraction of unprivileged individuals with a biased label 0 (but a true label 1).

Proof. Note that

$$\mathbb{E}_D\Big[\frac{\#\{i \mid s_i = P, y_i = 1\}}{\#\{i \mid s_i = P\}} - \frac{\#\{i \mid s_i = U, y_i = 1\}}{\#\{i \mid s_i = U\}}\Big] = \mathbb{E}_x[y_P(x)] - \mathbb{E}_x[y_U(x)]$$
$$= \mathbb{E}_x[y_P(x)] - \mathbb{E}_x[y_{\text{true}}(x)] + \mathbb{E}_x[y_{\text{true}}(x)] - \mathbb{E}_x[y_U(x)] = a + b.$$

■

E.1 Accuracy guarantees with respect to the unbiased labels

In this section, we prove Theorem E.4 on the accuracy guarantees of three methods (FairLift, [18], and baseline) with respect to the unbiased labels y_{true}. The Massaging method of [18] is described in detail in Section D. In this method, the dataset is pre-processed by flipping labels from 0 to 1 in the unprivileged group and from 1 to 0 in the privileged group. For each group, the fraction of flipped labels equals half the disparate impact difference of the dataset. See Fig. 2 for an illustration of how FairLift and the Massaging method change the label distribution in the training dataset.

Theorem E.4 (Per-group error with respect to unbiased labels)**.** *Let f_{BL}, f_{FL}, and f_{KC} be the classifiers learned by the baseline, FairLift, and the Massaging methods respectively. Suppose we have access to classifiers that learn label distributions perfectly and Assumption E.1 holds. Then the error of each classifier with respect to the true labels y_{true} is:*

	Expected error for the privileged group (w.r.t. y_{true})	Expected error for the unprivileged group (w.r.t. y_{true})
f_{BL}	a	b
f_{FL}	a	a
f_{KC}	$\geq \lvert b-a\rvert/2$	$\geq \lvert b-a\rvert/2$

From Theorem E.4, we can immediately obtain the following corollary on when FairLift outperforms the Massaging method in terms of overall accuracy. Consider in addition a counterpart of FairLift, call it FairPriv, which is the same as FairLift, but it flips the roles of the two groups and flips labels from 1 to 0 for the privileged group only. The expected error of this classifier is b for both groups. If we have knowledge of whether $a > b$, we can employ FairLift when $a \leq b$ and FairPriv otherwise. Let $\tilde{f}_{\text{FL}}$ be classifier learned after pre-processing the data in this way. Then $\tilde{f}_{\text{FL}}$ has expected error $\min(a, b)$ for both groups.

Corollary E.5 (When FairLift overperforms the massaging method)**.** *Consider the setup of Theorem E.4. The classifier f_{FL} learned by the FairLift method has higher accuracy than the classifier*

f_{KC} *learned by the massaging method of [18] when* $b > 3a$*. If in addition we have knowledge of whether* $a > b$*, we can learn a classifier* $\tilde{f}_{\mathrm{FL}}$ *that has higher accuracy than* f_{KC} *when* $b > 3a$ *and* $a > 3b$*.*

Proof of Theorem E.4. **(Baseline)** Since we have access to a perfect classifier, then $f_{\mathrm{BL}}((x,P)) = y_P(x)$ and $f_{\mathrm{BL}}((x,U)) = y_U(x)$. The error of f_{BL} for the privileged group is $\Pr_{x\sim\mathcal{X}}[f_{\mathrm{BL}}((x,P)) \neq y_{\mathrm{true}}(x)]$. We have:

$$\begin{aligned}\Pr_{x\sim\mathcal{X}}[f_{\mathrm{BL}}((x,P)) \neq y_{\mathrm{true}}(x)] &= \Pr_{x\sim\mathcal{X}}[f_{\mathrm{BL}}((x,P)) = 0, y_{\mathrm{true}}(x) = 1] + \Pr_{x\sim\mathcal{X}}[f_{\mathrm{BL}}((x,P)) = 1, y_{\mathrm{true}}(x) = 0] \\ &= \Pr_{x\sim\mathcal{X}}[y_P(x) = 0, y_{\mathrm{true}}(x) = 1] + \Pr_{x\sim\mathcal{X}}[y_P(x) = 1, y_{\mathrm{true}}(x) = 0] \\ &= 0 + a = a,\end{aligned}$$

where in the last equality we use Assumption E.1 and Claim E.2. A similar argument gives that the expected error of f_{BL} for the unprivileged group is b.

(FairLift) Note that f_{FL} returns the same labels as the baseline method for the privileged group, therefore the error of f_{FL} for the privileged group is a. Consider now the unprivileged group. Since y_P can be learned perfectly, then $f_{\mathrm{priv}}((x,P)) = y_P(x)$. In addition, notice that Assumption E.1 implies Assumption 2.1. From Theorem 2.3 we have $f_{\mathrm{FL}}((x,U)) = y_P(x)$ and thus the error of f_{FL} for the unprivileged group equals its error for the privileged group.

(Massaging method) Let $\hat{y}_{\mathrm{P}}(x)$ and $\hat{y}_{\mathrm{U}}(x)$ denote the label distribution induced by applying the Massaging technique in Algorithm 3. Since we consider perfect classifiers, then $f_{\mathrm{KC}}((x,P)) = \hat{y}_{\mathrm{P}}(x)$ and $f_{\mathrm{KC}}((x,U)) = \hat{y}_{\mathrm{U}}(x)$.

We lower bound the error of f_{KC} for the privileged group. A similar proof yields the same error lower bound for the unprivileged group. From the description of Algorithm 3,

$$\Pr[f_{\mathrm{KC}}((x,P)) = 0 \text{ and } y_P(x) = 1] = \Pr[\hat{y}_{\mathrm{P}}(x) = 0 \text{ and } y_P(x) = 1] = \frac{a+b}{2}. \tag{12}$$

From Assumption E.1 on the label bias, it is not too hard to see that

$$\Pr[f_{\mathrm{KC}}((x,P)) \neq y_{\mathrm{true}}(x)] = \Pr[y_{\mathrm{true}}(x) = 0, y_P(x) = 1, \hat{y}_{\mathrm{P}}(x) = 1] + \Pr[y_{\mathrm{true}}(x) = 1, y_P(x) = 1, \hat{y}_{\mathrm{P}}(x) = 0]. \tag{13}$$

First, consider $a \geq b$. We lower bound the first term of the sum in (13). We use the fact that for 3 events A, B, C we have $\Pr[A \cap B \cap C] \geq \Pr[A \cap B] - \Pr[B \cap \overline{C}]$. Then the first term of the sum is lower bounded by

$$\Pr[y_{\mathrm{true}}(x) = 0, y_P(x) = 1] - \Pr[y_P(x) = 1, \hat{y}_{\mathrm{P}}(x) = 0] = a - \frac{a+b}{2} = \frac{a-b}{2}, \tag{14}$$

after plugging in (12) and Claim E.2.

Next, consider $b < a$. Via a similar argument, the second term of the sum in (13) is lower bounded by

$$\Pr[y_P(x) = 1, \hat{y}_{\mathrm{P}}(x) = 0] - \Pr[y_P(x) = 1, y_{\mathrm{true}}(x) = 0] = \frac{a+b}{2} - a = \frac{b-a}{2}. \tag{15}$$

Combining (14) and (15), we conclude that

$$\Pr[f_{\mathrm{KC}}((x,P)) \neq y_{\mathrm{true}}(x)] \geq \frac{|a-b|}{2}. \tag{16}$$

■

E.2 Accuracy guarantees with respect to the observed (biased) labels

In this section we prove Theorem E.6 on the accuracy guarantees of the three methods (baseline, massaging, and FairLift) with respect to the observed labels. The observed labels result from potentially biased processes $y_P(x)$ and $y_U(x)$ that do not match $y_{\mathrm{true}}(x)$.

Theorem E.6 (Per-group error with respect to biased labels)**.** *Let* f_{BL}*,* f_{FL}*, and* f_{KC} *be the classifiers learned by the baseline, FairLift, and the Massaging method respectively. Suppose we have access to classifiers that learn label distributions perfectly and Assumption E.1 holds. Then the error of each classifier with respect to the biased labels* $y_P(x)$ *and* $y_U(x)$ *is:*

	Expected error for the privileged group (w.r.t. y_P)	Expected error for the unprivileged group (w.r.t. y_U)
f_{BL}	0	*0*
f_{FL}	0	$\leq a+b$
f_{KC}	$(a+b)/2$	$(a+b)/2$

From Theorem E.6 we immediately obtain the following corollary on the leveling-down effect of each classifier with respect to the baseline method.

Corollary E.7 (FairLift does not level-down)**.** *Consider the setup of Theorem E.6. FairLift and the Massaging method of [18] have the same overall expect error of $a+b$ with respect to the biased labels, however FairLift does not level-down, whereas the massaging method lowers the accuracy for both groups compared to the baseline method.*

Proof of Theorem E.6. **(Baseline)** Since we have access to classifiers that lean perfectly, then $f_{\mathrm{BL}}((x,P)) = y_P(x)$ and $f_{\mathrm{BL}}((x,U)) = y_U(x)$. The error of f_{BL} for the privileged group equals $\Pr_{x\sim\mathcal{X}}[f_{\mathrm{BL}}((x,P)) \neq y_P(x)] = 0$. A similar statement holds for the unprivileged group.

(FairLift) Note that FairLift does not modify labels for the privileged group, therefore $f_{\mathrm{FL}}((x,P)) = f_{\mathrm{BL}}((x,P)) = y_P(x)$ and the expected error of f_{FL} for the privileged group is 0. By Theorem 2.3, $f_{\mathrm{FL}}((x,U)) = y_P(x)$. Thus, the expected error of FairLift for the unprivileged group is $\Pr_{x\sim\mathcal{X}}[y_P(x) \neq y_U(x)]$. From Assumption 2.1, it is not too hard to see that $\Pr[y_P(x)=0, y_U(x)=1] = 0$ and thus

$$\begin{aligned}\Pr[y_P(x) \neq y_U(x)] &= \Pr[y_P(x)=1, y_U(x)=0] \\ &= \Pr[y_P(x)=1, y_U(x)=0, y_{\mathrm{true}}(x)=0] + \Pr[y_P(x)=1, y_U(x)=0, y_{\mathrm{true}}(x)=1] \\ &\leq a+b. \end{aligned} \tag{17}$$

The upper bound in (17) follows from Claim E.2.

(Massaging) From equation (12) in the proof of Theorem E.4 we have $\Pr[f_{\mathrm{KC}}((x,P)) \neq y_P(x)] = \frac{a+b}{2}$. A similar statement holds for the unprivileged group. ■

F Additional Experimental Details

F.1 Datasets

Adult [3]: The task is to predict whether an individual's income will exceed \$50K/yr based on Census data. We use sex as the protected attribute, with "male" being the privileged group. The dataset is prepared according to [32].

German Credit Risk [16]: The task is to classify individuals as good or bad credit risks based on their financial and demographic attributes. The protected attribute is age with individuals aged above 30 forming the privileged group. The choice of protected attribute follows [17].

Compas [29]: The task is to predict whether an individual will re-offend within two years based on criminal history, jail and prison time, and demographics. We only retain datapoints where the race is "Caucasian" (privileged group) or "African-American" (unprivileged group). This amounts to $\sim$1k dropped datapoints from a total of $\sim$7k.

Communities and Crime [30]: Each datapoint corresponds to a community and the task is to predict the number of violent crime per capita given socio-economic and crime data. The target variable is binarized so that 1 indicates a low rate of violent crimes per capita (below the 70% percentile) and 0 the opposite. The sensitive attribute is race, where the privileged group contains half of the communities with the highest percentage of white population and the unprivileged group contains the other half. The dataset is prepared according to [32].

F.2 Classifier Training

In this section, we provide further details on how we train classifiers for each method.

	Overall Accuracy				Disparate Impact Ratio				Harmonic Mean			
	Compas	German	Adult	Crime	Compas	German	Adult	Crime	Compas	German	Adult	Crime
Baseline	**0.6902**	0.7490*	**0.8582**	**0.8613**	0.7465	0.8344	0.8243	0.5059	0.7170	0.7883	0.8409	0.6366
FairLift	0.6611	0.7424	0.8227	0.7086	0.9689	0.9314*	0.9802*	**0.9966**	0.7858	0.8253*	0.8945	0.8282
Kamiran and Calders [18]	0.6771	0.7398*	0.8333	0.7330	0.9721	0.9407*	0.9951*	0.9521	**0.7980**	0.8277	0.9070	0.8278*
Jiang and Nachum [17]	0.6742	0.7428*	0.8377	0.7332	**0.9757**	0.9465*	0.9974*	0.9709	0.7972	0.8318*	**0.9106**	**0.8354**
Hardt et al [14]	0.6484	0.6912	0.8049	0.6470	0.9749	0.9521	0.9951	0.9629	0.7786	0.8002	0.8900	0.7735

Table 4: The disparate impact ratio and accuracy of classifiers obtained from four bias-correction methods. Comparison is performed on four datasets. All numbers are statistically significantly different at $p = 0.05$ compared to FairLift except those indicated by *. Number of trials is 50.

	Accuracy on unbiased labels				Disparate Impact Ratio				Harmonic Mean			
	$b = 2a$	$b = 4a$	$b = 8a$	$a = 0$	$b = 2a$	$b = 4a$	$b = 8a$	$a = 0$	$b = 2a$	$b = 4a$	$b = 8a$	$a = 0$
Baseline	0.9202	0.8695	0.7844	0.8138	0.8319	0.7048	0.4941	0.5747	0.8737	0.7784	0.6062	0.6736
FairLift	0.9444	0.9428	**0.9444**	**0.9895**	0.9783	0.9800	0.9812	0.9836	0.9610	0.9610	**0.9624**	**0.9865**
Kamiran and Calders [18]	0.9836	0.9472	0.8676	0.8380	0.9858	0.9850	0.9847*	**0.9879**	**0.9847**	**0.9657**	0.9224	0.9067
Jiang and Nachum [17]	**0.9814**	**0.9619**	0.8956	0.8935	0.9855	0.9858	0.9843*	0.9867	0.9834	0.9737	0.9378	0.9377
Hardt et al [14]	0.8274	0.7695	0.7226	0.7872	**0.9863**	**0.9879**	**0.9880**	0.9874	0.8999	0.8651	0.8346	0.8760

Table 5: The disparate impact ratio and accuracy of classifiers obtained from four bias-correction methods. Comparison is performed on synthetic datasets with varying levels of labels bias a and b for the privileged and unprivileged group respectively. All numbers are statistically significantly different at $p = 0.05$ compared to FairLift except those indicated by *. Number of trials is 100.

FairLift: The threshold hyper-parameter is chosen from the range $[2^{-4}, 2^{-1}]$ via grid search. Using a standalone validation set, we choose the threshold which achieves the highest disparate impact ratio. The validation set is randomly chosen from the training set to contain a 0.2 portion of the points.

Jiang and Nachum: The only hyper-parameter is the number of iterations, which we set to 100, similarly to the experiments in [17]. The algorithm of Jiang and Nachum requires as input a fairness objective that can be expressed as a linear constraint - we set the objective to minimizing disparate impact difference.

Kamiran and Calders: The function f_{rank} used to rank the dataset is obtained by training a Logistic Regression model with default hyperparameters from the sci-kit learn package, similarly to the classifiers trained on the training dataset.

Hardt et al: The threshold optimizer in the post-processing algorithm requires setting a fairness objective and an accuracy objective. We choose disparate impact ratio as the fairness objective. For the accuracy objective, one can choose between optimizing overall accuracy or balancing accuracy between groups. We find that both perform similarly and show results for the latter in this paper.