A Probabilistic Approach to Fairness under Label Bias

Saurav Anchlia[1] **YooJung Choi**[1]

[1]School of Computing and Augmented Intelligence, Arizona State University

Abstract

Machine learning systems are widely used in our daily lives, making fairness an important concern when designing and deploying these systems. Moreover, the data that we use to train or audit them often contain biased labels. In this paper, we show that not only does label bias in training data affect model performance, but it also misrepresents fairness of classifiers at test time. To tackle this problem, we propose a framework to audit and learn fair classifiers by using a probabilistic model to *infer the hidden fair labels* and estimating the *expected fairness* under this distribution. In particular, we provide (i) a "data clean-up" method which replaces biased labels with the fair ones—which can be used as pre-processing at train time or for better auditing at test time—and (ii) a reweighting method that directly estimates statistical fairness notions with respect to the inferred fair labels. Experimental results demonstrate the effectiveness of our proposed approach on synthetic data, with controlled ground truth labels and their biased versions, as well as on real-world benchmark datasets.

1 INTRODUCTION

As machine learning systems are being used to assist or make decisions in areas that affect our lives—advertising, credit scoring, hiring, education, and even criminal risk assessments [2, 6–8]—measuring and ensuring algorithmic fairness have received much attention in the recent years. This includes various definitions to quantify fairness and algorithms to mitigate bias [6, 10, 11, 14–16, 23, 26]. In particular, the presence of noisy or biased labels in data may exacerbate discrimination and make bias mitigation techniques more challenging [13, 14, 18]. For instance, a biased classifier may incorrectly be deemed fair, and vice versa. In many domains especially relevant for algorithmic fairness, it may be impossible or at least highly infeasible to observe the true target variable, and only a biased proxy may be available. For example, risk assessment tools whose aim is to predict re-offense would be trained with data containing re-arrest information instead. In this paper, we show that explicit consideration of such label bias is necessary for fairness and propose probabilistic approaches to reliably audit and learn fair classifiers from biased labels.

The key component of our approach is to infer the probability of hidden fair labels given the observations about the features and biased labels. This requires efficient inference of conditional probabilities given different evidence. We leverage a recent fair probabilistic modeling approach [5], which learns the bias mechanism in a way that best explains the observed data as well as allowing tractable inference on the learned distribution.

Our framework uses this distribution to address label bias both at *test (audit)* time and *train* time, through *data cleaning* and *reweighting*. For data cleaning, we replace the potentially biased labels with inferred fair labels, obtained either by sampling from aforementioned probability distribution or by thresholding. The clean data can then be used to train classifiers that are accurate and fair with respect to these generated labels, as well as to evaluate given classifiers at test time. In addition, we also propose an importance reweighting approach which does not alter the dataset, but rather incorporates the fair label probabilities as weights when estimating performance metrics and fairness violations.

Related Work Several recent works also studied quantifying and mitigating the impact of biased or noisy labels, assuming group-dependent [3, 12, 24, 27] or instance-dependent noise rates [25]. In particular, Jiang and Nachum [17] propose to reweigh the dataset as pre-processing, with the assumption that the biased labeling function is the one that is closest to the fair one. On the other hand, we learn the weights through a probabilistic modeling approach (Section 3.1). Moreover, Wang et al. [24] and Wu et al. [25]

Accepted for the 6th Workshop on Tractable Probabilistic Modeling at UAI (TPM 2023).

derive weighted classification loss and fairness metrics to incorporate group- and instance-dependent noise rates, respectively. Instead, our reweighting utilizes the inferred probabilities of fair labels for each observed instance, which also allows us to derive a pre-processing data cleaning. This in fact subsumes the group-dependent noise setting, while relaxing an implicit independence assumption (Appendix A.1).

	Y		$\widetilde{Y}$	
S, X	#pos	#neg	#pos	#neg
$1,1$	30	40	35	35
$1,0$	10	20	20	10
$0,1$	30	10	20	20
$0,0$	10	50	20	40

(a) Hidden (Y) and observed label ($\widetilde{Y}$)

	$P(Y=1 \mid S, X, \widetilde{Y})$	
S, X	$\widetilde{Y}=1$	$\widetilde{Y}=0$
$1,1$	0.86	0
$1,0$	0.5	0
$0,1$	1	0.5
$0,0$	0.5	0

(b) Fair label probabilities

Figure 1: Example dataset with label bias

2 FAIRNESS UNDER LABEL BIAS

We use uppercase letters (X) for random variables and lowercase letters (x) for their assignments; bold letters denote sets of random variables ($\mathbf{X}$) and assignments ($\mathbf{x}$), respectively. Let S denote a sensitive attribute defining the demographic group assignment, $\mathbf{X}$ a set of non-sensitive features, and $Y \in \{0,1\}$ a binary label. The set of possible values for $\mathbf{X}$ and S are denoted by $\mathcal{X}$ and $\mathcal{S}$, respectively. For simplicity, we assume that $\mathcal{S} = \{0,1\}$, but our method can easily be applied to multi-valued sensitive attributes. Moreover, $\widetilde{Y}$ denotes the noisy or biased version of Y that is actually observed, and $P(\mathbf{X}, S, \widetilde{Y}, Y)$ the joint distribution over all random variables. The observed data $\mathcal{D} = \{(\mathbf{x}_i, s_i, \tilde{y}_i)\}_{i=1}^n$ consists of n i.i.d. samples drawn from $P(\mathbf{X}, S, \widetilde{Y})$.

Our goal is to train a classifier $f : \mathcal{X} \to \{0,1\}$ to minimize a loss function $l(.)$ with some fairness constraint:[1]

$$\min_f \mathbb{E}_P[l(f(\mathbf{X}), Y)] \text{ s.t. } f \text{ is fair w.r.t. } P(\mathbf{X}, S, Y) \quad (1)$$

Among many statistical notions of fairness [6, 11, 15, 19], we focus on the effect of label bias on *equal opportunity (EOp)* and *equalized odds (EO)* [15]. A binary classifier f satisfies equalized odds if the *true positive rate (*$\mathrm{TPR}_{Y,s}$*)* and *false positive rate (*$\mathrm{FPR}_{Y,s}$*)* are equal across the demographic groups; i.e., for each $y \in \{0,1\}$:

$$P(f(\mathbf{X}) = 1 \mid S = 1, Y = y) = P(f(\mathbf{X}) = 1 \mid S = 0, Y = y).$$

Equal opportunity only requires the true positive rates to be equalized ($y = 1$ in above equation). These notions are loosely based on an intuition that a perfect classifier is fair, which no longer holds in the presence of label bias: *a classifier that perfectly predicts biased labels is clearly not fair*.

Example 2.1. Consider an example dataset shown in Figure 1a, over a single feature X, a sensitive attribute S, fair label Y, and biased observed label $\widetilde{Y}$. The number of positive and negative labels are shown for each feature assignment, and the highlighted entries indicate the observed data $\{(x_i, s_i, \tilde{y}_i)\}_{i=1}^{200}$. Suppose we have a classifier $f(X) = \mathbb{1}[X = 1]$. It satisfies EO w.r.t. the fair label Y: $\mathrm{TPR}_{Y,1} = \mathrm{TPR}_{Y,0} = {}^{30}/_{40} = 0.75$. However, an audit w.r.t. observed data would conclude that it violates EOp: $\mathrm{TPR}_{\widetilde{Y},1} = {}^{35}/_{55} = 0.64, \mathrm{TPR}_{\widetilde{Y},0} = {}^{20}/_{40} = 0.50$. Thus, data containing label bias may lead to incorrect fairness assessment of classifiers.

While we could compute the TPR and FPR of f with respect to the true labels Y with access to the underlying distribution $P(\mathbf{X}, S, \widetilde{Y}, Y)$, it is generally unavailable in practice. More importantly, even if such distribution is given, exactly computing the TPR and FPR corresponds to the *expected prediction task* [20, 21] $\mathbb{E}_{P(\mathbf{X}|s,y)}[f(\mathbf{X})]$ which is known to be NP-hard even in restricted cases such as when f is a logistic regression classifier and P a naive Bayes model. Instead, we use the fact that if we can reliably infer $P(Y \mid \mathbf{x}, s, \tilde{y})$, using the data drawn from $P(\mathbf{X}, S, \widetilde{Y})$ we can estimate the fairness violation and empirical loss as if we were sampling from the joint distribution $P(\mathbf{X}, S, \widetilde{Y}, Y)$. Based on this intuition, we propose a data *pre-processing* method and an *importance reweighting* approach to reliably estimate the expected fairness violation of existing classifiers and to enforce fairness constraints w.r.t. the hidden fair labels. For the instance-specific probabilities $P(Y \mid \mathbf{x}, s, \tilde{y})$, we show that the fair probabilistic modeling proposed by Choi et al. [5] can infer the fair labels both accurately and efficiently. We now describe our proposed approaches in detail.

3 INFERRING FAIR LABELS

This section describes our approach to "clean up" a given dataset by replacing the biased labels with a probabilistically inferred hidden fair label for each data sample. The clean dataset can then be used by downstream (fair) classification algorithms or evaluation of learned classifiers.

Suppose we had access to the conditional distribution $P(Y \mid \mathbf{X}, S, \widetilde{Y})$. Then we can augment our data to obtain $\{(\mathbf{x}_i, s_i, \tilde{y}_i, y_i)\}_{i=1}^n$ by sampling $y_i \sim P(Y \mid \mathbf{x}_i, s_i, \tilde{y}_i)$. This augmented dataset can then be used for fairness assessment of existing classifiers, as it would produce unbiased estimates of true positive and false positive rates w.r.t. the underlying distribution $P(\mathbf{X}, S, Y)$. Moreover, this method can be seen as a pre-processing step: the clean data can be passed to any fair classifier learning algorithm to enforce fairness constraints with respect to the inferred clean labels.

Note that due to sampling, above clean-up algorithm is inherently randomized, and thus multiple runs could output

[1] The classifier f may also use the sensitive attribute S in addition to features $\mathbf{X}$.

Table 1: Accuracy of inferring fair labels

$\texttt{Synth}_{10}$	$\texttt{Synth}_{20}$	$\texttt{Synth}_{30}$	COMPAS	Adult
0.9031	0.9395	0.9413	0.9787	0.9729

different datasets. We also provide a simpler deterministic alternative where we threshold the instance-specific fair label probability. That is, each example $(\mathbf{x}_i, s_i, \tilde{y}_i)$ is assigned a new label $y_i = \mathbb{1}[P(Y = 1 \mid \mathbf{x}_i, s_i, \tilde{y}_i) \geq T]$ for some threshold T (0.5 by default). While this no longer guarantees unbiased estimates of TPR and FPR, we empirically demonstrate its efficacy in retrieving the ground truth labels with high accuracy as well as in downstream fair learning.

3.1 LEARNING THE FAIR LABEL DISTRIBUTION

While our proposed data cleaning and importance reweighting (Section 4) do not constrain how the fair label probability $P(Y \mid \mathbf{X}, S, \widetilde{Y})$ is obtained, our implementation utilizes the probabilistic modeling framework FAIRPC [5] which we briefly describe here. FAIRPC faithfully learns a joint distribution $P(\mathbf{X}, S, \widetilde{Y}, Y)$ to best explain the observed data, with the assumption that Y is a fair label that is independent of the sensitive attribute and that $\widetilde{Y}$ is a biased version of it. In particular, the distribution is factorized as the following:

$$P(\mathbf{X}, S, \widetilde{Y}, Y) = P(\mathbf{X} \mid S, Y)P(\widetilde{Y} \mid S, Y)P(S)P(Y).$$

The distribution is represented by a *probabilistic circuit*, a type of probabilistic model that supports tractable inference [4]. In particular, we can compute the conditional probability $P(Y = 1 \mid \mathbf{x}, s, \tilde{y})$ in linear time in the size of the circuit for any arbitrary evidence $\mathbf{x}, s, \tilde{y}$. Moreover, this computation can easily be performed in parallel so that we can quickly obtain the corresponding probability for all observed data samples.

To see how effective FAIRPC is in inferring the fair label given the observed data $\mathbf{x}, s, \tilde{y}$, we evaluate the accuracy of probabilistic classifier $P(Y \mid \mathbf{x}, s, \tilde{y}) \geq 0.5$ on synthetic and real-world benchmark datasets (Table 1). We refer to Section 5 for details about the datasets. On synthetic datasets with $|\mathbf{X}| = 10, 20, 30$ where we can generate ground truth labels and the biased versions, FAIRPC trained on the biased observed data can predict the ground truth labels with test-set accuracy ranging from 90% to 94%. Moreover, in real-world datasets where hidden fair labels are not available, we compare inferred fair labels with observed labels (which may be biased) in order to evaluate whether inferred fair labels are still reasonably close to the given labels. We answer this in the affirmative, with the test-set accuracy of 98% and 97% for `COMPAS` and `Adult` datasets, respectively. Therefore, we can confidently use these inferred labels for downstream fair ML methods.

4 ESTIMATING AND ENFORCING EXPECTED GROUP FAIRNESS

The data cleaning approach has a strong benefit that it can be used with various fair classification learning or fairness auditing algorithms without any change to those downstream algorithms. However, it modifies the labels in the dataset, which may prohibit its use in some real-world applications due to data protection regulations.

Moreover, even though our data cleaning algorithm utilizes the conditional distribution $P(Y \mid \mathbf{X}, S, \widetilde{Y})$, we either threshold or sample from it, and the subsequent audit or learning methods will only see the binarized labels. Instead, we also introduce estimators (EST_{IR}) for the *expected* fairness violations (EO or EOp) and accuracy, by reweighting each sample with importance weights derived using the fair label probabilities.

Proposition 1. Consider a joint distribution $P(\mathbf{X}, S, \widetilde{Y}, Y)$ and i.i.d. data $\{(\mathbf{x}_i, s_i, \tilde{y}_i)\}_{i=1}^n$ drawn from its marginal distribution $P(\mathbf{X}, S, \widetilde{Y})$. For any function $g(\mathbf{X}, S, \widetilde{Y}, Y)$, the following is an unbiased estimate of $\mathbb{E}_P[g]$:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{y\in\{0,1\}} g(\mathbf{x}_i, s_i, \tilde{y}_i, y)P(y \mid \mathbf{x}_i, s_i, \tilde{y}_i).$$

For instance, the expected accuracy of a classifier f w.r.t. the hidden fair label Y can be estimated by setting $g(\mathbf{x}, y) = \mathbb{1}[f(\mathbf{x}) = y]$. For the expected fairness violation, we need to estimate the true positive and false positive rates which involve computing conditional expectations for each S and Y.

Proposition 2. Consider a joint distribution $P(\mathbf{X}, S, \widetilde{Y}, Y)$ and i.i.d. data $\{(\mathbf{x}_i, s_i, \tilde{y}_i)\}_{i=1}^n$ drawn from its marginal distribution $P(\mathbf{X}, S, \widetilde{Y})$. For a classifier $f : \mathcal{X} \to \{0, 1\}$, $s \in \mathcal{S}$, and $y \in \{0, 1\}$, the following is an unbiased estimate of $P(f(\mathbf{X}) = 1 \mid S = s, Y = y)$:

$$\frac{1}{n_s \cdot P(y \mid s)} \sum_{i:s_i=s} f(\mathbf{x}_i)P(y \mid \mathbf{x}_i, \tilde{y}_i, s_i) \qquad (2)$$

where $n_s = |\{i : s_i = s\}|$ is the number of samples whose sensitive attribute has the value s.

The detailed proofs of above propositions can be found in Appendix A. Briefly, they are based on viewing the observed data as samples from a distribution $Q(\mathbf{X}, S, \widetilde{Y}, Y) = P(\mathbf{X}, S, \widetilde{Y})Q(Y)$ in which Y is independent of other variables and can be completely random. Then we obtain estimates w.r.t. the target distribution $P(.)$ by using the importance weights $P(\mathbf{x}, s, \tilde{y}, y)/Q(\mathbf{x}, s, \tilde{y}, y)$.

Note that we can also derive a fair learning algorithm by optimizing the expected accuracy subject to the fairness constraints using the above estimators. We leave this as future work and in this paper focus on evaluating fair classification through data cleaning.

5 EXPERIMENTAL RESULTS

This section investigates the effectiveness of our framework for auditing and learning fair classifiers (complete results in Appendix B). We evaluate our methods and baselines on `COMPAS` [22] for recidivism prediction where the sensitive attribute is ethnicity, `Adult` [9] for income prediction with sex as sensitive attribute, and synthetic data [5] where we can generate ground truth labels and observed labels with group-dependent bias. We compare against three baselines: (1) logistic regression trained on the observed data (LR_{OBS}), (2) REDUCT. [1] which solves fair classification with equalized odds constraint by reducing it to cost-sensitive classification problems, and (3) REWGT. [17] which corrects bias by reweighting data points. On synthetic data, we additionally train logistic regression using ground truth labels (LR_{GT}). Each reported result is the average over 10 runs. All experiments were run on an Intel(R) Xeon(R) E5-2680 v4 CPU (2.4 GHz) with 128GB RAM.

Auditing under Label Bias Here we empirically demonstrate that fairness violation estimates using observed labels are not reliable and evaluate our proposed auditing methods. Table 2 summarizes the results on a synthetic dataset with $|\mathbf{X}| = 10$. Comparing EOp and EO estimates using the ground truth labels (TRUE) to those using observed labels (EST_{Obs}), we see that across all baselines EST_{Obs} underestimates fairness violations. This includes fair classification methods REDUCT. and REWGT. which still exhibit discrepancy when audited. On the other hand, our proposed approaches based on data cleaning (EST_{Fair} and $\text{EST}_{\text{Fair}\geq}$) and importance reweighting (EST_{IR}) follow the ground truth evaluation more closely, validating the utility of inferred fair labels as a reliable substitute for ground truth labels in scenarios where the latter is not available. EST_{Fair}which is the unbiased estimate of fair label is closest to ground truth. Both $\text{EST}_{\text{Fair}\geq}$and EST_{IR}result in similar evaluations, thus affirming our assertion that we can *estimate* true fairness violations of a classifier without replacing the actual labels.

Learning Fair Classifier Let us now turn our attention to learning fair classifiers from biased labels. We train the three baselines on pre-processed data by our threshold-based cleaning method. We compare them against the baselines trained directly on observed data; we evaluate using ground truth labels on synthetic dataset, and provide evaluations using both observed labels and importance reweighting on real-world datasets where ground truth is not available. Table 3 summarizes the results.

Impact of label bias during training is best quantified by comparing LR_{Obs} and LR_{Fair}, where we see significant improvement in fairness without loss of accuracy across all datasets. In fact, we observe that our pre-processing step results in overall improvement of fairness compared to the baseline counterparts. Moreover, we also see a significant increase in ground-truth accuracy on synthetic data, suggesting that our data cleaning indeed provide labels that are close to ground truth.

Table 2: Evaluation of auditing methods on synthetic dataset

	Eval	LR_{GT}	LR_{Obs}	REDUCT.	REWGT.
Acc.	TRUE	0.7208	0.6055	0.6620	0.6022
	EST_{Obs}	0.6133	0.6416	0.6351	0.6243
	EST_{Fair}	0.7112	0.6094	0.6638	0.6100
	$\text{EST}_{\text{Fair}\geq}$	0.7305	0.6222	0.6807	0.6211
	EST_{IR}	0.7098	0.6093	0.6637	0.6100
EOp	TRUE	0.1968	0.4937	0.0942	0.0520
	EST_{Obs}	0.0408	0.4447	0.0904	0.0196
	EST_{Fair}	0.1743	0.4903	0.0933	0.0467
	$\text{EST}_{\text{Fair}\geq}$	0.1949	0.4960	0.1036	0.0546
	EST_{IR}	0.1698	0.4876	0.0786	0.0430
EO	TRUE	0.2097	0.5073	0.1448	0.0553
	EST_{Obs}	0.1260	0.4746	0.1208	0.0312
	EST_{Fair}	0.1925	0.5095	0.1453	0.0542
	$\text{EST}_{\text{Fair}\geq}$	0.2206	0.5088	0.1552	0.0581
	EST_{IR}	0.1898	0.5060	0.1455	0.0660

Table 3: Evaluation of fair learning methods

		Synth_{10}	COMPAS		Adult	
	Method	TRUE	EST_{Obs}	EST_{IR}	EST_{Obs}	EST_{IR}
Accuracy	LR_{OBS}	0.6055	0.8822	0.8897	0.8364	0.8082
	REDUCT.	0.6620	0.8805	0.8764	0.8327	0.8202
	REWGT.	0.6022	0.8831	0.8921	0.8317	0.8250
	$\text{LR}_{\text{Fair}\geq}$	0.7204	0.8835	0.9012	0.8277	0.8230
	$\text{REDUCT.}_{\text{Fair}\geq}$	0.7085	0.8830	0.9009	0.8346	0.8204
	$\text{REWGT.}_{\text{Fair}\geq}$	0.7167	0.8833	0.9011	0.8331	0.8233
EO	LR_{OBS}	0.5073	0.2474	0.2513	0.2475	0.4636
	REDUCT.	0.1448	0.2981	0.2990	0.0953	0.1921
	REWGT.	0.0553	0.1632	0.1715	0.0599	0.2288
	$\text{LR}_{\text{Fair}\geq}$	0.2341	0.1500	0.1569	0.0797	0.2210
	$\text{REDUCT.}_{\text{Fair}\geq}$	0.0578	0.1264	0.1349	0.0525	0.2406
	$\text{REWGT.}_{\text{Fair}\geq}$	0.0930	0.1380	0.1447	0.0611	0.2546

6 CONCLUSION

This paper studied how label bias can make fairness evaluation challenging and demonstrated the need to explicitly correct such bias. We first proposed a data cleaning method that infers the hidden fair label for each data instance. This can be used to estimate the expected fairness violations and to learn fair classifiers using the clean labels rather than the biased ones. As this approach replaces the labels in data which may be problematic in certain domains, we also provide an importance reweighting approach that directly estimates the expected fairness w.r.t. the hidden labels without changing the data. Empirical evaluation showed that we are able to accurately estimate fairness metrics and better enforce them with respect to the hidden true labels.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69, 2018.

[2] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

[3] Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? In *1st Symposium on Foundations of Responsible Computing*, 2020.

[4] YooJung Choi, Antonio Vergari, and Guy Van den Broeck. Probabilistic circuits: A unifying framework for tractable probabilistic models. oct 2020. URL `http://starai.cs.ucla.edu/papers/ProbCirc20.pdf`.

[5] YooJung Choi, Meihua Dang, and Guy Van den Broeck. Group fairness by probabilistic modeling with latent fair decisions. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, Feb 2021.

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

[7] Danielle Keats Citron and Frank Pasquale. The scored society: Due process for automated predictions. *Wash. L. Rev.*, 89:1, 2014.

[8] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies*, 2015(1): 92–112, 2015.

[9] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[11] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[12] Riccardo Fogliato, Max G'Sell, and A. Chouldechova. Fairness evaluation in presence of biased noisy labels. In *AISTATS*, 2020.

[13] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5): 845–869, 2013.

[14] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

[15] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[16] Max Hort, Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. Bia mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*, 2022.

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

[18] Justin M Johnson and Taghi M Khoshgoftaar. A survey on classifying big data with label noise. *ACM Journal of Data and Information Quality*, 14(4):1–43, 2022.

[19] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

[20] Pasha Khosravi, YooJung Choi, Yitao Liang, Antonio Vergari, and Guy Van den Broeck. On tractable computation of expected predictions. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, dec 2019.

[21] Pasha Khosravi, Yitao Liang, YooJung Choi, and Guy Van den Broeck. What to expect of classifiers? reasoning about logistic regression with missing features. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, aug 2019.

[22] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm, 2016. URL `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algori`

[23] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[24] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 526–536, 2021.

[25] Songhua Wu, Mingming Gong, Bo Han, Yang Liu, and Tongliang Liu. Fair classification with instance-dependent label noise. In *Conference on Causal Learning and Reasoning*, pages 927–943. PMLR, 2022.

[26] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[27] Yixuan Zhang, Feng Zhou, Zhidong Li, Yang Wang, and Fang Chen. Fair representation learning with unreliable labels. In *International Conference on Artificial Intelligence and Statistics*, pages 4655–4667. PMLR, 2023.

A DETAILED PROOFS

A.1 INSTANCE-SPECIFIC FAIR LABEL PROBABILITY VS. NOISE RATE

As discussed in Section 1, existing fair ML methods addressing label bias often assume or estimate noise rates. For example, group-dependent noise rates can be characterized as $P(\widetilde{Y} \mid S, Y)$, whereas instance-dependent rates additionally condition on $\mathbf{X}$ [24, 25]. Instead, we focus on the instance-specific fair label probability $P(Y \mid \mathbf{X}, S, \widetilde{Y})$ which naturally led to a data cleaning method by simply inferring the fair label for each instance.

Moreover, it is crucial that we do not assume that noise rates are known, as it would be highly unlikely in real-world applications even for group-based rates, and especially so for the more finer-grained instance-based noise. Nevertheless, we can still provide the option to incorporate group-specific noise rates should they be available as background knowledge, as we have framed the problem of deriving fair labels as that of learning an interpretable probabilistic model with a latent variable.

More specifically, FAIRPC factorizes the joint distribution into smaller ones including the group-dependent noise $P(\widetilde{Y} \mid S, Y)$:

$$P(\mathbf{X}, S, \widetilde{Y}, Y) = P(\mathbf{X} \mid S, Y)P(\widetilde{Y} \mid S, Y)P(S)P(Y).$$

Thus, we could fix this distribution during the training of FAIRPC and learn the rest to best fit the observed data. The remaining algorithm stays the same, in which we use tractable conditional inference to compute $P(Y \mid \mathbf{X}, S, \widetilde{Y})$ for each instance.

This is in contrast to the surrogate fairness metric derived by Wang et al. [24], in which the empirical TPR and FPR are weighted by group-dependent noise rates to obtain estimates w.r.t. clean labels. For such group-based reweighting, an implicit independence assumption $\mathbf{X} \perp Y \mid S, \widetilde{Y}$ must hold. Our approach does not make such assumption and instead leaves the door open for additional constraints while learning the fair distribution for specialized use cases.

Let us now prove that a group-based reweighting of true positive and false positive rates requires the independence assumption $\mathbf{X} \perp Y \mid S, \widetilde{Y}$ to hold. Specifically, assuming some group-based noise rates $P(\widetilde{Y} \mid S, Y)$, Wang et al. [24] show that the true positive rate of a classifier $f(\mathbf{X})$ with respect to true label Y can be written in terms of TPR and FPR estimates w.r.t. the biased labels as follows:

$$\begin{aligned}
&P(f(\mathbf{X}){=}1 \mid S{=}s, Y{=}1)\\
&= P(f(\mathbf{X}){=}1 \mid S{=}s, Y{=}1, \widetilde{Y}{=}1)P(\widetilde{Y}{=}1 \mid S{=}s, Y{=}1)\\
&+ P(f(\mathbf{X}){=}1 \mid S{=}s, Y{=}1, \widetilde{Y}{=}0)P(\widetilde{Y}{=}0 \mid S{=}s, Y{=}1)\\
&\stackrel{?}{=} P(f(\mathbf{X}){=}1 \mid S{=}s, \widetilde{Y}{=}1)P(\widetilde{Y}{=}1 \mid S{=}s, Y{=}1)\\
&+ P(f(\mathbf{X}){=}1 \mid S{=}s, \widetilde{Y}{=}0)P(\widetilde{Y}{=}0 \mid S{=}s, Y{=}1).
\end{aligned}$$

Note that Equation 3 can be computed using the empirical TPR and FPR and the group-based noise rates. A similar derivation can be performed for true FPR by considering $Y = 0$ instead of $Y = 1$.

However, for Equation 3 to hold, we must have that

$$\begin{aligned}
&P(f(\mathbf{X}){=}1 \mid S{=}s, Y{=}y, \widetilde{Y} = \tilde{y})\\
&= P(f(\mathbf{X}){=}1 \mid S{=}s, \widetilde{Y}{=}\tilde{y})
\end{aligned}$$

for any assignment $s, y, \tilde{y}$ and arbitrary classifier $f(.)$. For instance, suppose $f(\mathbf{x}) = 1$ for a fixed assignment $\mathbf{x}$ and $f(\mathbf{x}') = 0$ for all $\mathbf{x}' \neq \mathbf{x}$. Then we need

$$\begin{aligned}
&P(\mathbf{X}{=}\mathbf{x} \mid S{=}s, Y{=}y, \widetilde{Y}{=}\tilde{y})\\
&= P(\mathbf{X}{=}\mathbf{x} \mid S{=}s, \widetilde{Y}{=}\tilde{y})
\end{aligned}$$

for arbitrary $\mathbf{x}, s, y, \tilde{y}$. This implies $\mathbf{X} \perp Y \mid S, \widetilde{Y}$, which may not hold in general. We remark that above result does not depend on which label was used to train f. That is, even if f was trained using $\widetilde{Y}$, additional information about Y could alter the distribution of features.

A.2 PROOFS OF PROPOSITIONS

This section provides detailed proofs of unbiasedness of our importance reweighted estimates.

At a high level, we construct a hypothetical distribution Q and show that our estimates correspond to using samples from Q with importance weights P/Q. We first define the distribution Q as the following:

$$\begin{aligned}
&Q(\mathbf{X}{=}\mathbf{x}, S{=}s, \widetilde{Y}{=}\tilde{y}, Y{=}y)\\
&= \frac{1}{2}P(\mathbf{X}{=}\mathbf{x}, S{=}s, \widetilde{Y}{=}\tilde{y}), \quad .
\end{aligned}$$

In particular, Q agrees with P in the marginal distribution of the observed data, and assumes that Y is uniformly distributed and independent of all other variables: $Q(y \mid \mathbf{x}, s, \tilde{y}) = Q(y) = 1/2$. Then for any assignment $\mathbf{x}, s, \tilde{y}, y$, we have the following:

$$\frac{P(\mathbf{x}, s, \tilde{y}, y)}{Q(\mathbf{x}, s, \tilde{y}, y)} = \frac{P(y \mid \mathbf{x}, s, \tilde{y})P(\mathbf{x}, s, \tilde{y})}{Q(y \mid \mathbf{x}, s, \tilde{y})Q(\mathbf{x}, s, \tilde{y})} = \frac{P(y \mid \mathbf{x}, s, \tilde{y})}{1/2}. \tag{3}$$

We are now ready to prove the propositions.

Proposition 1. Consider a joint distribution $P(\mathbf{X}, S, \widetilde{Y}, Y)$ and i.i.d. data $\{(\mathbf{x}_i, s_i, \tilde{y}_i)\}_{i=1}^n$ drawn from its marginal distribution $P(\mathbf{X}, S, \widetilde{Y})$. For any function $g(\mathbf{X}, S, \widetilde{Y}, Y)$, the following is an unbiased estimate of $\mathbb{E}_P[g]$:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{y\in\{0,1\}} g(\mathbf{x}_i, s_i, \tilde{y}_i, y)P(y \mid \mathbf{x}_i, s_i, \tilde{y}_i).$$

Proof. The dataset $\{(\mathbf{x}_i, s_i, \tilde{y}_i)\}_{i=1}^n$ consists of i.i.d. samples from $P(\mathbf{X}, S, \widetilde{Y})$ which is equivalent to $Q(\mathbf{X}, S, \widetilde{Y})$ by our construction of Q. Moreover, we can augment this dataset to $\{(\mathbf{x}_i, s_i, \tilde{y}_i, y_i)\}_{i=1}^n$ by sampling y_i uniformly at random. Then the expression in above proposition is simply a Monte Carlo estimator from $Q(\mathbf{X}, S, \widetilde{Y}, Y)$ whose expectation can be written as follows:

$$\begin{aligned}
&\mathbb{E}_{Q(\mathbf{X},S,\widetilde{Y},Y)}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{y\in\{0,1\}} g(\mathbf{x}_i, s_i, \tilde{y}_i, y)P(y \mid \mathbf{x}_i, s_i, \tilde{y}_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Q(\mathbf{X},S,\widetilde{Y},Y)}\left[\sum_{y\in\{0,1\}} g(\mathbf{x}, s, \tilde{y}, y)P(y \mid \mathbf{x}, s, \tilde{y})\right] \\
&= \mathbb{E}_{Q(\mathbf{X},S,\widetilde{Y},Y)}\left[\sum_{y\in\{0,1\}} g(\mathbf{x}, s, \tilde{y}, y)P(y \mid \mathbf{x}, s, \tilde{y})\right].
\end{aligned} \tag{4}$$

Considering the case $y = 0$ for $y \in \{0, 1\}$ in above summation, we have

$$\begin{aligned}
&\mathbb{E}_{Q(\mathbf{X},S,\widetilde{Y},Y)}\left[g(\mathbf{x}, s, \tilde{y}, Y=0)P(Y=0 \mid \mathbf{x}, s, \tilde{y})\right] \\
&= \mathbb{E}_{Q(\mathbf{X},S,\widetilde{Y},Y)}\left[g(\mathbf{x}, s, \tilde{y}, Y=0)\cdot\frac{P(\mathbf{x}, s, \tilde{y}, Y=0)}{2Q(\mathbf{x}, s, \tilde{y}, Y=0)}\right] && (5)\\
&= \frac{1}{2}\sum_{\mathbf{x},s,\tilde{y},y} g(\mathbf{x}, s, \tilde{y}, Y=0)\frac{P(\mathbf{x}, s, \tilde{y}, Y=0)}{Q(\mathbf{x}, s, \tilde{y}, Y=0)}Q(\mathbf{x}, s, \tilde{y}, y) && (6)\\
&= \frac{1}{2}\sum_{\mathbf{x},s,\tilde{y},y} g(\mathbf{x}, s, \tilde{y}, Y=0)P(\mathbf{x}, s, \tilde{y}, Y=0) && (7)\\
&= \frac{1}{2}\sum_{y\in\{0,1\}}\sum_{\mathbf{x},s,\tilde{y}} g(\mathbf{x}, s, \tilde{y}, Y=0)P(\mathbf{x}, s, \tilde{y}, Y=0) \\
&= \sum_{\mathbf{x},s,\tilde{y}} g(\mathbf{x}, s, \tilde{y}, Y=0)P(\mathbf{x}, s, \tilde{y}, Y=0). && (8)
\end{aligned}$$

Note that Equation 5 follows from the importance weight derived in Equation 3, and Equation 7 follows from Equation 6 because $Q(\mathbf{x}, s, \tilde{y}, Y=0) = Q(\mathbf{x}, s, \tilde{y}, Y=1)$. Similarly, we have that

$$\begin{aligned}
&\mathbb{E}_{Q(\mathbf{X},S,\widetilde{Y},Y)}\left[g(\mathbf{x}, s, \tilde{y}, Y=1)P(Y=1 \mid \mathbf{x}, s, \tilde{y})\right] && (9)\\
&= \sum_{\mathbf{x},s,\tilde{y}} g(\mathbf{x}, s, \tilde{y}, Y=1)P(\mathbf{x}, s, \tilde{y}, Y=1). && (10)
\end{aligned}$$

Combining Equations 8 and 10 with Equation 4, we conclude the following:

$$\begin{aligned}
&\mathbb{E}_{Q(\mathbf{X},S,\widetilde{Y},Y)}\left[\sum_{y\in\{0,1\}} g(\mathbf{x}, s, \tilde{y}, y)P(y \mid \mathbf{x}, s, \tilde{y})\right] \\
&= \sum_{\mathbf{x},s,\tilde{y}} g(\mathbf{x}, s, \tilde{y}, Y=0)P(\mathbf{x}, s, \tilde{y}, Y=0) \\
&\quad + \sum_{\mathbf{x},s,\tilde{y}} g(\mathbf{x}, s, \tilde{y}, Y=1)P(\mathbf{x}, s, \tilde{y}, Y=1) \\
&= \sum_{\mathbf{x},s,\tilde{y},y} g(\mathbf{x}, s, \tilde{y}, y)P(\mathbf{x}, s, \tilde{y}, y) = \mathbb{E}_P[g(\mathbf{X}, S, \widetilde{Y}, Y)].
\end{aligned}$$

□

Proposition 2. Consider a joint distribution $P(\mathbf{X}, S, \widetilde{Y}, Y)$ and i.i.d. data $\{(\mathbf{x}_i, s_i, \tilde{y}_i)\}_{i=1}^n$ drawn from its marginal distribution $P(\mathbf{X}, S, \widetilde{Y})$. For a classifier $f : \mathcal{X} \to \{0, 1\}$, $s \in \mathcal{S}$, and $y \in \{0, 1\}$, the following is an unbiased estimate of $P(f(\mathbf{X}) = 1 \mid S = s, Y = y)$:

$$\frac{1}{n_s \cdot P(y \mid s)}\sum_{i:s_i=s} f(\mathbf{x}_i)P(y \mid \mathbf{x}_i, \tilde{y}_i, s_i) \tag{11}$$

where $n_s = |\{i : s_i = s\}|$ is the number of samples whose sensitive attribute has the value s.

Proof. Restricting our dataset to those in which $s_i = s$, we effectively have i.i.d. samples from $Q(\mathbf{X}, \widetilde{Y} \mid S = s)$ which is equivalent to $Q(\mathbf{X}, \widetilde{Y} \mid S = s, Y = y)$ by construction (Y is independent of all other variables according to Q).

Let us first derive the importance weight given our target distribution $P(\mathbf{X}, \widetilde{Y} \mid s, y)$ and proposal distribution $Q(\mathbf{X}, \widetilde{Y} \mid s, y)$:

$$\begin{aligned}
\frac{P(\mathbf{x}, \tilde{y} \mid s, y)}{Q(\mathbf{x}, \tilde{y} \mid s, y)} &= \frac{P(\mathbf{x}, s, \tilde{y}, y)Q(s, y)}{Q(\mathbf{x}, s, \tilde{y}, y)P(s, y)} \\
= \frac{P(y \mid \mathbf{x}, s, \tilde{y})}{1/2}\frac{1/2 \cdot P(s)}{P(s, y)} &= \frac{P(y \mid \mathbf{x}, s, \tilde{y})}{P(y \mid s)}
\end{aligned}$$

Here, the second equality holds due to Equation 3 and $Q(SY) = Q(S)Q(Y) = \frac{1}{2}P(S)$.

We can then derive the expectation of our estimator:

$$\mathbb{E}_{Q(\mathbf{X},\widetilde{Y}|S=s,Y=y)}\left[\frac{1}{n_s P(y \mid s)} \sum_{i:s_i=s} f(\mathbf{x}_i) P(y \mid \mathbf{x}_i, \tilde{y}_i, s)\right]$$

$$= \frac{1}{n_s P(y \mid s)} \sum_{i:s_i=s} \mathbb{E}_{Q(\mathbf{X},\widetilde{Y}|S=s,Y=y)} \left[f(\mathbf{x}) P(y \mid \mathbf{x}, \tilde{y}, s)\right]$$

$$= \frac{1}{P(y \mid s)} \mathbb{E}_{Q(\mathbf{X},\widetilde{Y}|S=s,Y=y)} \left[f(\mathbf{x}) P(y \mid \mathbf{x}, \tilde{y}, s)\right]$$

$$= \frac{1}{P(y \mid s)} \sum_{\mathbf{x},\tilde{y}} f(\mathbf{x}) P(y \mid \mathbf{x}, \tilde{y}, s) Q(\mathbf{x}, \tilde{y} \mid s, y)$$

$$= \frac{1}{P(y \mid s)} \sum_{\mathbf{x},\tilde{y}} f(\mathbf{x}) \frac{P(\mathbf{x}, \tilde{y} \mid s, y) P(y \mid s)}{Q(\mathbf{x}, \tilde{y} \mid s, y)} Q(\mathbf{x}, \tilde{y} \mid s, y)$$

$$= \sum_{\mathbf{x},\tilde{y}} f(\mathbf{x}) P(\mathbf{x}, \tilde{y} \mid s, y)$$

$$= \sum_{\mathbf{x},\tilde{y}} \mathbb{1}[f(\mathbf{x}) = 1] P(\mathbf{x}, \tilde{y} \mid s, y) = P(f(\mathbf{x}) = 1 \mid s, y).$$

□

B ADDITIONAL EXPERIMENTS

Auditing under Label Bias Section 5 presented empirical evaluation of auditing methods on synthetic dataset with $|\mathbf{X}| = 10$. We additionally evaluate them on synthetic dataset with $|\mathbf{X}| = 20$ and $|\mathbf{X}| = 30$ and also provide standard deviation across 10 runs; results are shown in Tables 4–6. Again, audits are done using ground truth labels (TRUE), observed biased labels (EST_{Obs}), sampled fair labels (EST_{Fair}), thresholded fair labels ($\text{EST}_{\text{Fair}\geq}$), and importance reweighting (EST_{IR}). For EST_{Fair}, we sample 100 test label set across 10 folds and take average performance for evaluation metrics.

Tables 5 and 6 show reliability of using EST_{Fair} or $\text{EST}_{\text{Fair}\geq}$ for fairness audits in larger datasets. EST_{Obs} continues to underestimate fairness violations. Overall, we see reasonably low standard deviation in all our evaluation metrics.

We also provide the complete set of results for evaluation of auditing methods on real-world benchmark datasets in Table 7.

Learning Fair Classifier We provide the complete set of results for the evaluation of fair learning methods summarized in Table 3. In particular, we additionally include evaluation on synthetic datasets with $|\mathbf{X}| = 20, 30$ (Table 8) and an additional evaluation technique (EST_{Fair}, $\text{EST}_{\text{Fair}\geq}$) on the real-world benchmarks (Table 9). Methodology for generating evaluation sets from fair label distribution is same as previously described in B.

Similar trends in accuracy-fairness trade offs can be observed. While fairness violations differ depending on evaluation method, we remark that estimates using EST_{Fair} are theoretically unbiased with respect to the underlying distribution (see Section 3).

Table 4: Overview of auditing methods on synthetic dataset with $|\mathbf{X}| = 10$ with SD.

Metric	Eval	LR_{GT}	LR_{Obs}	REDUCT.	REWGT.
Accuracy (↑)	TRUE	0.7208 ± 0.0141	0.6055 ± 0.0162	0.6620 ± 0.0174	0.6022 ± 0.0139
	EST_{Obs}	0.6133 ± 0.0172	0.6416 ± 0.0144	0.6351 ± 0.0158	0.6243 ± 0.0175
	EST_{Fair}	0.7112 ± 0.0203	0.6094 ± 0.0163	0.6638 ± 0.0217	0.6100 ± 0.0161
	$\text{EST}_{\text{Fair}\geq}$	0.7305 ± 0.0183	0.6222 ± 0.0167	0.6807 ± 0.0214	0.6211 ± 0.0169
	EST_{IR}	0.7098 ± 0.0191	0.6093 ± 0.0144	0.6637 ± 0.0200	0.6100 ± 0.0134
EOp (↓)	TRUE	0.1968 ± 0.0504	0.4937 ± 0.0474	0.0942 ± 0.0753	0.0520 ± 0.0207
	EST_{Obs}	0.0408 ± 0.0285	0.4447 ± 0.0404	0.0904 ± 0.0516	0.0196 ± 0.0218
	EST_{Fair}	0.1743 ± 0.0465	0.4903 ± 0.0458	0.0933 ± 0.0713	0.0467 ± 0.0270
	$\text{EST}_{\text{Fair}\geq}$	0.1949 ± 0.0510	0.4960 ± 0.0448	0.1036 ± 0.0774	0.0546 ± 0.0300
	EST_{IR}	0.1698 ± 0.0506	0.4876 ± 0.0530	0.0786 ± 0.1023	0.0430 ± 0.0521
EO (↓)	TRUE	0.2097 ± 0.0350	0.5073 ± 0.0364	0.1448 ± 0.1087	0.0553 ± 0.0188
	EST_{Obs}	0.1260 ± 0.0456	0.4746 ± 0.0334	0.1208 ± 0.0937	0.0312 ± 0.0193
	EST_{Fair}	0.1925 ± 0.0426	0.5095 ± 0.0383	0.1453 ± 0.1056	0.0542 ± 0.0238
	$\text{EST}_{\text{Fair}\geq}$	0.2206 ± 0.0430	0.5088 ± 0.0360	0.1552 ± 0.1120	0.0581 ± 0.0267
	EST_{IR}	0.1898 ± 0.0444	0.5060 ± 0.0356	0.1455 ± 0.1118	0.0660 ± 0.0291

Table 5: Overview of auditing methods on synthetic dataset with $|\mathbf{X}| = 20$.

Metric	Eval	LR_{GT}	LR_{Obs}	REDUCT.	REWGT.
Accuracy (↑)	TRUE	0.7977 ± 0.0132	0.7219 ± 0.0189	0.7593 ± 0.0169	0.7088 ± 0.0149
	EST_{Obs}	0.6669 ± 0.0180	0.6849 ± 0.0173	0.6675 ± 0.0146	0.6630 ± 0.0134
	EST_{Fair}	0.7900 ± 0.0141	0.7234 ± 0.0170	0.7580 ± 0.0131	0.7118 ± 0.0122
	$\text{EST}_{\text{Fair}\geq}$	0.7992 ± 0.0159	0.7302 ± 0.0191	0.7650 ± 0.0123	0.7175 ± 0.0131
	EST_{IR}	0.7889 ± 0.0139	0.7232 ± 0.0168	0.7577 ± 0.0116	0.7112 ± 0.0112
EOp (↓)	TRUE	0.1069 ± 0.0540	0.3148 ± 0.0485	0.0837 ± 0.0291	0.1064 ± 0.0229
	EST_{Obs}	0.0419 ± 0.0483	0.2476 ± 0.0570	0.0340 ± 0.0250	0.0204 ± 0.0197
	EST_{Fair}	0.0854 ± 0.0436	0.3078 ± 0.0456	0.0730 ± 0.0312	0.0959 ± 0.0260
	$\text{EST}_{\text{Fair}\geq}$	0.0939 ± 0.0425	0.3082 ± 0.0485	0.0783 ± 0.0306	0.1017 ± 0.0235
	EST_{IR}	0.0935 ± 0.0415	0.3152 ± 0.0419	0.0815 ± 0.0447	0.1050 ± 0.0399
EO (↓)	TRUE	0.1256 ± 0.0380	0.4080 ± 0.0314	0.0895 ± 0.0317	0.1064 ± 0.0229
	EST_{Obs}	0.0881 ± 0.0437	0.3716 ± 0.0454	0.0658 ± 0.0370	0.0575 ± 0.0293
	EST_{Fair}	0.1056 ± 0.0325	0.4139 ± 0.0372	0.0901 ± 0.0279	0.0968 ± 0.0258
	$\text{EST}_{\text{Fair}\geq}$	0.1157 ± 0.0286	0.4096 ± 0.0407	0.0870 ± 0.0291	0.1022 ± 0.0234
	EST_{IR}	0.1068 ± 0.0374	0.4130 ± 0.0461	0.1018 ± 0.0269	0.1089 ± 0.0354

Table 6: Overview of auditing methods on synthetic dataset with $|\mathbf{X}| = 30$.

Metric	Eval	LR_{GT}	LR_{Obs}	REDUCT.	REWGT.
Accuracy (↑)	TRUE	0.8134 ± 0.0116	0.7189 ± 0.0183	0.7578 ± 0.0161	0.6995 ± 0.0161
	EST_{Obs}	0.6726 ± 0.0123	0.6969 ± 0.0124	0.6824 ± 0.0146	0.6603 ± 0.0066
	EST_{Fair}	0.8142 ± 0.0095	0.7186 ± 0.0172	0.7604 ± 0.0163	0.6993 ± 0.0164
	$\text{EST}_{\text{Fair}\geq}$	0.8275 ± 0.0097	0.7262 ± 0.0190	0.7701 ± 0.0162	0.7052 ± 0.0199
	EST_{IR}	$.8154 \pm 0.0076$	0.7180 ± 0.0162	0.7604 ± 0.0152	0.6988 ± 0.0155
EOp (↓)	TRUE	0.1276 ± 0.0456	0.3119 ± 0.0468	0.0946 ± 0.0532	0.0943 ± 0.0228
	EST_{Obs}	0.0523 ± 0.0494	0.2448 ± 0.0573	0.0674 ± 0.0292	0.0365 ± 0.0262
	EST_{Fair}	0.1306 ± 0.0416	0.3205 ± 0.0464	0.0988 ± 0.0467	0.1005 ± 0.0247
	$\text{EST}_{\text{Fair}\geq}$	0.1415 ± 0.0407	0.3208 ± 0.0506	0.1040 ± 0.0515	0.1037 ± 0.0261
	EST_{IR}	0.1273 ± 0.0890	0.3176 ± 0.0920	0.0962 ± 0.1048	0.0975 ± 0.0945
EO (↓)	TRUE	0.1492 ± 0.0265	0.4693 ± 0.0438	0.1246 ± 0.0751	0.0943 ± 0.0228
	EST_{Obs}	0.0934 ± 0.0456	0.4293 ± 0.0487	0.0877 ± 0.0392	0.0473 ± 0.0332
	EST_{Fair}	0.1463 ± 0.0245	0.4753 ± 0.0433	0.1247 ± 0.0677	0.1005 ± 0.0247
	$\text{EST}_{\text{Fair}\geq}$	0.1584 ± 0.0227	0.4692 ± 0.0442	0.1259 ± 0.0676	0.1038 ± 0.0261
	EST_{IR}	0.1501 ± 0.0624	0.4814 ± 0.0422	0.1521 ± 0.0763	0.1163 ± 0.0684

Table 7: Overview of auditing methods on real-world datasets

Dataset	Metric	Eval	LR_{Obs}	REDUCT.	REWGT.
COMPAS	Accuracy (↑)	EST_{Obs}	0.8822 ± 0.0044	0.8805 ± 0.0038	0.8831 ± 0.0042
		EST_{Fair}	0.8902 ± 0.0031	0.8769 ± 0.0108	0.8926 ± 0.0044
		$\text{EST}_{\text{Fair}\geq}$	0.8941 ± 0.0032	0.8782 ± 0.0145	0.8955 ± 0.0042
		EST_{IR}	0.8897 ± 0.0030	0.8764 ± 0.0108	0.8921 ± 0.0042
	EOp (↓)	EST_{Obs}	0.0094 ± 0.0065	0.0192 ± 0.0170	0.0325 ± 0.0244
		EST_{Fair}	0.0486 ± 0.0063	0.0364 ± 0.0125	0.0711 ± 0.0225
		$\text{EST}_{\text{Fair}\geq}$	0.0641 ± 0.0037	0.0439 ± 0.0145	0.0880 ± 0.0249
		EST_{IR}	0.0494 ± 0.0162	0.0371 ± 0.0215	0.0719 ± 0.0303
	EO (↓)	EST_{Obs}	0.2474 ± 0.0471	0.2981 ± 0.0800	0.1632 ± 0.0951
		EST_{Fair}	0.2543 ± 0.0372	0.3060 ± 0.0828	0.1758 ± 0.0880
		$\text{EST}_{\text{Fair}\geq}$	0.2844 ± 0.0455	0.3201 ± 0.0974	0.1962 ± 0.0984
		EST_{IR}	0.2513 ± 0.0394	0.2990 ± 0.1005	0.1715 ± 0.0832
Adult	Accuracy (↑)	EST_{Obs}	0.8364 ± 0.0070	0.8327 ± 0.0081	0.8317 ± 0.0075
		EST_{Fair}	0.8083 ± 0.0077	0.8203 ± 0.0060	0.8253 ± 0.0089
		$\text{EST}_{\text{Fair}\geq}$	0.8230 ± 0.0068	0.8350 ± 0.0059	0.8405 ± 0.0079
		EST_{IR}	0.8082 ± 0.0074	0.8202 ± 0.0052	0.8250 ± 0.0086
	EOp (↓)	EST_{Obs}	0.2475 ± 0.0631	0.0952 ± 0.0479	0.0542 ± 0.0229
		EST_{Fair}	0.3904 ± 0.0388	0.1021 ± 0.0806	0.1206 ± 0.0412
		$\text{EST}_{\text{Fair}\geq}$	0.3128 ± 0.0492	0.1013 ± 0.0487	0.0530 ± 0.0246
		EST_{IR}	0.4636 ± 0.0392	0.1921 ± 0.0730	0.2288 ± 0.0381
	EO (↓)	EST_{Obs}	0.2475 ± 0.0631	0.0953 ± 0.0477	0.0599 ± 0.0161
		EST_{Fair}	0.3904 ± 0.0388	0.1144 ± 0.0685	0.1247 ± 0.0372
		$\text{EST}_{\text{Fair}\geq}$	0.3128 ± 0.0493	0.1079 ± 0.0392	0.0949 ± 0.0102
		EST_{IR}	0.4636 ± 0.0392	0.1921 ± 0.0730	0.2288 ± 0.0381

Table 8: Evaluation of fair learning methods on synthetic datasets with $|\mathbf{X}| = 10, 20, 30$

		Synth_{10}	Synth_{20}	Synth_{30}
Metric	Method	TRUE	TRUE	TRUE
Accuracy (↑)	LR_{OBS}	0.6055 ± 0.0163	0.7219 ± 0.0189	0.7189 ± 0.0183
	REDUCT.	0.6620 ± 0.0173	0.7593 ± 0.0169	0.7578 ± 0.0161
	REWGT.	0.6022 ± 0.0139	0.7088 ± 0.0149	0.6995 ± 0.0161
	$\text{LR}_{\text{Fair}\geq}$	0.7204 ± 0.0116	0.7978 ± 0.0150	0.8137 ± 0.0128
	$\text{REDUCT.}_{\text{Fair}\geq}$	0.7085 ± 0.0151	0.7931 ± 0.0122	0.8087 ± 0.0120
	$\text{REWGT.}_{\text{Fair}\geq}$	0.7167 ± 0.0181	0.7943 ± 0.0122	0.8093 ± 0.0100
EO (↓)	LR_{OBS}	0.5073 ± 0.0364	0.4080 ± 0.0314	0.4693 ± 0.0438
	REDUCT.	0.1448 ± 0.1087	0.0895 ± 0.0317	0.1246 ± 0.0751
	REWGT.	0.0553 ± 0.0188	0.1064 ± 0.0229	0.0943 ± 0.0228
	$\text{LR}_{\text{Fair}\geq}$	0.2341 ± 0.0430	0.1356 ± 0.0353	0.1492 ± 0.0275
	$\text{REDUCT.}_{\text{Fair}\geq}$	0.0578 ± 0.0373	0.0552 ± 0.0196	0.0639 ± 0.0261
	$\text{REWGT.}_{\text{Fair}\geq}$	0.0930 ± 0.0345	0.0627 ± 0.0312	0.0576 ± 0.0295

Table 9: Evaluation of fair learning methods on real-world datasets

Dataset	Metric	Method	EST_{Obs}	EST_{Fair}	$\text{EST}_{\text{Fair}\geq}$	EST_{IR}
COMPAS	Accuracy (↑)	LR_{OBS}	0.8822 ± 0.0044	0.8902 ± 0.0031	0.8941 ± 0.0032	0.8897 ± 0.0030
		REDUCT.	0.8805 ± 0.0038	0.8769 ± 0.0108	0.8782 ± 0.0145	0.8764 ± 0.0108
		REWGT.	0.8831 ± 0.0042	0.8926 ± 0.0044	0.8955 ± 0.0042	0.8921 ± 0.0042
		$\text{LR}_{\text{Fair}\geq}$	0.8835 ± 0.0033	0.9014 ± 0.0031	0.9048 ± 0.0024	0.9012 ± 0.0027
		$\text{REDUCT.}_{\text{Fair}\geq}$	0.8830 ± 0.0038	0.9010 ± 0.0024	0.9041 ± 0.0020	0.9009 ± 0.0022
		$\text{REWGT.}_{\text{Fair}\geq}$	0.8833 ± 0.0020	0.9013 ± 0.0030	0.9046 ± 0.0028	0.9011 ± 0.0027
	EO (↓)	LR_{OBS}	0.2474 ± 0.0471	0.2543 ± 0.0372	0.2844 ± 0.0455	0.2513 ± 0.0394
		REDUCT.	0.2981 ± 0.0800	0.3060 ± 0.0828	0.3201 ± 0.0974	0.2990 ± 0.1005
		REWGT.	0.1962 ± 0.0951	0.1758 ± 0.0880	0.1962 ± 0.0984	0.1715 ± 0.0832
		$\text{LR}_{\text{Fair}\geq}$	0.1500 ± 0.0446	0.2543 ± 0.0372	0.1976 ± 0.0024	0.1569 ± 0.0234
		$\text{REDUCT.}_{\text{Fair}\geq}$	0.1264 ± 0.0339	0.1363 ± 0.0226	0.1730 ± 0.0261	0.1349 ± 0.0202
		$\text{REWGT.}_{\text{Fair}\geq}$	0.1380 ± 0.0405	0.1474 ± 0.0282	0.1856 ± 0.0024	0.1447 ± 0.0183
Adult	Accuracy (↑)	LR_{OBS}	0.8364 ± 0.0070	0.8083 ± 0.0077	0.8230 ± 0.0068	0.8082 ± 0.0074
		REDUCT.	0.8327 ± 0.0081	0.8203 ± 0.0060	0.8350 ± 0.0059	0.8202 ± 0.0052
		REWGT.	0.8317 ± 0.0075	0.8253 ± 0.0087	0.8405 ± 0.0079	0.8250 ± 0.0086
		$\text{LR}_{\text{Fair}\geq}$	0.8277 ± 0.0051	0.8234 ± 0.0073	0.8402 ± 0.0057	0.8230 ± 0.0069
		$\text{REDUCT.}_{\text{Fair}\geq}$	0.8346 ± 0.0060	0.8083 ± 0.0077	0.8353 ± 0.0069	0.8204 ± 0.0062
		$\text{REWGT.}_{\text{Fair}\geq}$	0.8331 ± 0.0075	0.8236 ± 0.0088	0.8388 ± 0.0076	0.8233 ± 0.0086
	EO (↓)	LR_{OBS}	0.2475 ± 0.0631	0.3905 ± 0.0388	0.3128 ± 0.0493	0.4636 ± 0.0392
		REDUCT.	0.0953 ± 0.0477	0.1144 ± 0.0685	0.1079 ± 0.0392	0.1921 ± 0.0730
		REWGT.	0.0599 ± 0.0161	0.1247 ± 0.0372	0.0949 ± 0.0102	0.2288 ± 0.0381
		$\text{LR}_{\text{Fair}\geq}$	0.0797 ± 0.0298	0.1341 ± 0.0355	0.1261 ± 0.0223	0.2210 ± 0.0475
		$\text{REDUCT.}_{\text{Fair}\geq}$	0.0525 ± 0.0187	0.1468 ± 0.0387	0.0862 ± 0.0197	0.2406 ± 0.0379
		$\text{REWGT.}_{\text{Fair}\geq}$	0.0611 ± 0.0233	0.1515 ± 0.0438	0.0975 ± 0.0109	0.2546 ± 0.0381

Table 10: Evaluation of inferring fair labels on synthetic and real-world datasets

Dataset	$\lvert\mathbf{X}\rvert = 10$	$\lvert\mathbf{X}\rvert = 20$	$\lvert\mathbf{X}\rvert = 30$	Adult	COMPAS
	Acc.	Acc.	Acc.	Acc.	Acc.
TRAIN	0.9002 ± 0.0044	0.9405 ± 0.0106	0.9411 ± 0.0028	0.9735 ± 0.0016	0.9790 ± 0.0010
VALID	0.9032 ± 0.0112	0.9443 ± 0.0078	0.9397 ± 0.0049	0.9736 ± 0.0016	0.9777 ± 0.0023
TEST	0.9031 ± 0.0082	0.9395 ± 0.0030	0.9413 ± 0.0098	0.9729 ± 0.0031	0.9787 ± 0.0027