Estimating and Implementing Conventional Fairness Metrics With Probabilistic Protected Features

Anonymous authors
Paper under double-blind review

Abstract

The vast majority of techniques to train fair models require access to the protected attribute (e.g., race, gender), either at train time or in production. However, in many important applications this protected attribute is largely unavailable. In this paper, we develop methods for measuring and reducing fairness violations in a setting with limited access to protected attribute labels. Specifically, we assume access to protected attribute labels on a small subset of the dataset of interest, but only probabilistic estimates of protected attribute labels (e.g., via Bayesian Improved Surname Geocoding) for the rest of the dataset. With this setting in mind, we propose a method to estimate bounds on common fairness metrics for an existing model, as well as a method for training a model to limit fairness violations by solving a constrained non-convex optimization problem. Unlike similar existing approaches, our methods take advantage of contextual information – specifically, the relationships between a model's predictions and the probabilistic prediction of protected attributes, given the true protected attribute, and vice versa – to provide tighter bounds on the true disparity. We provide an empirical illustration of our methods using voting data. First, we show our measurement method can bound the true disparity up to 5.5x tighter than previous methods in these applications. Then, we demonstrate that our training technique effectively reduces disparity while incurring lesser fairness-accuracy trade-offs than other fair optimization methods with limited access to protected attributes.

1 Introduction

In both the private and public sectors, organizations are facing increased pressure to ensure they use equitable machine learning systems, whether through legal obligations or social norms ([22]; [21]; [43]; [26]; [25]). For instance, in 2022, Meta Platforms agreed to build a system for measuring and mitigating racial disparity in advertising to settle a lawsuit filed by the U.S. Department of Housing and Urban Development under the Fair Housing Act ([39]; [28]). Similarly, recent Executive Orders in the United States ([38]; [42]) direct government agencies to measure and mitigate disparity resulted from or exacerbated by their programs, including in the "design, develop[ment], acqui[sition], and us[e] [of] artificial intelligence and automated systems" ([42]).

Yet both companies ([3]) and government agencies ([38]) rarely collect or have access to individual-level data on race and other protected attributes on a comprehensive basis. Given that the majority of algorithmic fairness tools which could be used to monitor and mitigate racial bias require demographic attributes ([8]; [7]), the limited availability of protected attribute data represents a significant challenge in assessing algorithmic fairness and training fairness-constrained systems difficult.

In this paper, we address this problem by introducing methods for *1) measuring* fairness violations in, and *2) training* fair models on, data with limited access to protected attribute labels. We assume access to protected attribute labels on only a small subset of the dataset of interest, along with probabilistic estimates of protected attribute labels— for example, estimates generated using Bayesian Improved Surname Geocoding (BISG) ([27])—for the rest of the dataset.

We leverage this limited labeled data to establish whether certain relationships between the model's predictions, the probabilistic protected attributes, and the ground truth protected attributes hold. Given these conditions, our first main result (Theorem 1) shows that we can bound a range of common fairness metrics, from above and below, over the full dataset with easily computable (un)fairness estimators calculated using the *probabilistic* estimates of the protected attribute. We expound on these conditions, define the fairness estimators, and introduce this result in Section 2.

To train fair models, we leverage our results on measuring fairness violations to bound disparity during learning; we enforce the upper bound on unfairness *calculated with the probabilistic protected attribute* (measured on the full training set) as a surrogate fairness constraint, while also enforcing the conditions required to ensure the estimators accurately bound disparity in the model's predictions (calculated on the labeled subset), as constraints during training. We leverage recent work in constrained learning with non-convex losses (14) to ensure bounded fairness violations with near-optimal performance at prediction time.

We note that our data access setting is common across a variety of government and business contexts: first, estimating race using BISG is standard practice in government and industry (12; 23; 33; 40; 39). Although legal constraints or practical barriers often prevent collecting a full set of labels for protected attributes, companies and agencies can and do obtain protected attribute labels for subsets of their data. For example, companies such as Meta have started to roll out surveys asking for voluntary disclosure of demographic information to assess disparities (40). Another method for obtaining a subset of protected attribute data is to match data to publicly available administrative datasets containing protected attribute labels for a subset of records, as in, e.g. Elzayn et al. (20).

While our approach has stronger data requirements than recent work in similar domains (31; 46) in that a subset of it must have protected attribute labels, many important applications satisfy this requirement. The advantage to using this additional data is substantially tighter bounds on disparity: in our empirical applications, we find up to 5.5x tighter bounds for fairness metrics, and up to 5 percentage points less of an accuracy penalty when enforcing the same fairness bound during training.

In sum, we present the following contributions: *1)* We introduce a new method of bounding ground truth fairness violations across a wide range of fairness metrics in datasets with limited access to protected attribute data (Section 2); *2)* We introduce a new method of training models with near-optimal and near-feasible bounded unfairness with limited protected attribute data (Section 3); *3)* We show the utility of our approaches, including comparisons to a variety of baselines and other approaches, on various datasets relevant for assessing disparities in regulated contexts: we focus on voter registration data, commonly used to estimate racial disparities in voter turnout (1) (Section 4) with additional datasets presented in Appendix F.

2 Methodology for Measurement

In this section, we formally introduce our problem setting and notation, define the types of fairness metrics we can measure and enforce with our techniques, and define the *probabilistic* and *linear* estimators of disparity for these metrics. We then introduce our first main result: given certain relationships between the protected attribute, model predictions, and probabilistic estimates of protected attribute in the data, we can upper and lower bound the true fairness violation for a given metric using the linear and probabilistic estimators respectively.

2.1 Notation and Preliminaries

Setting and Datasets. We wish to learn a model of an outcome Y based on individuals' features X. Individuals have a special binary protected class feature $B \in \{0, 1\}$ which is usually unobserved, and *proxy variables* $Z \subset X$ which may be correlated with B. Our primary dataset, called the *learning dataset*, is $\mathscr{D} := \mathscr{D}_U \cup \mathscr{D}_L$, where $\mathscr{D}_U$ (the *unlabeled set*) consists of observations $\{(X_i, Y_i, Z_i)\}_{i=1}^{n_U}$ and $\mathscr{D}_L$ (the *labeled set*) additionally includes B and so consists of $\{(X_i, Y_i, Z_i, B_i)\}_{i=1}^{n_L}$. An *auxiliary dataset* $\{(Z, B)\}_{i=1}^{n_A}$ allows us to learn an estimate of $b_i := \Pr[B_i|Z_i]$; except where specified, we abstract away from the auxiliary dataset and assume access to b. When considering learning, we assume a *hypothesis class* of models $\mathcal{H}$ which map X either directly to Y or a superset (e.g. $[0, 1]$ rather than $\{0, 1\}$), and consider models parameterized by θ, i.e. $h_\theta \in \mathcal{H}$. An important random variable that we will use is the *conditional covariance* of random variables. In particular, for random variables Q, R, S, T, we write $C_{Q,R|S,T} := \mathrm{Cov}(Q, R|S, T)$.

Notation. For a given estimator θ and random variable X, we use $\hat{\theta}$ to denote the sample estimator and $\hat{X}$ to denote a prediction of X. We use $\bar{X}$ to indicate the sample average of a random variable taken over an appropriate dataset. In some contexts we use group-specific averages, which we indicate with a superscript. For example, we use $\bar{b}^{B_i}$ to denote the sample average of b among individuals who have protected class feature B equal to B_i. We will indicate a generic conditioning event using the symbol $\mathcal{E}$, and overloading it, we will write $\mathcal{E}_i$ as an indicator, i.e. 1 when $\mathcal{E}$ is true for individual i and 0 otherwise. In the learning setting, $\mathcal{E}_i$ will depend on our choice of model h; when we want to emphasize this, we write $\mathcal{E}_i(h)$. We will also use the $(\cdot)$ notation to emphasize dependence on context more generally, e.g. $C_{f,b|B}(h_\theta)$ is the covariance of f and b conditional on B under h_θ.

Fairness Metrics. In this paper, we focus on measuring and enforcing a group-level *fairness metric* that can be expressed as the difference across groups of some function of the outcome and the prediction, possibly conditioned on some event. More formally:

Definition 1. A *fairness metric* μ is an operator associated with a function f and an event $\mathcal{E}$ such that

$$\mu(\mathcal{D}) := \mathbb{E}_{\mathcal{D}}[f(\hat{Y}, Y)|\mathcal{E}, B = 1] - \mathbb{E}_{\mathcal{D}}[f(\hat{Y}, Y)|\mathcal{E}, B = 0],$$

where the distribution $\mathcal{D}$ corresponds to the process generating $(X, Y, \hat{Y})$.

Many common fairness metrics can be expressed in this form by defining an appropriate event $\mathcal{E}$ and function f. For instance, *demographic parity* in classification (10; 47; 50) corresponds to letting $\mathcal{E}$ be the generically true event and f be simply the indicator $\mathbf{1}[\hat{Y} = 1]$. False positive rate parity (15; 16) corresponds to letting $\mathcal{E}$ be the event that $Y = 0$ and letting $f(\hat{Y}, Y) = \mathbf{1}[\hat{Y} \neq Y]$. True positive rate parity (24) (also known as "equality of opportunity") corresponds to letting $\mathcal{E}$ be the event that $Y = 1$ and $f(\hat{Y}, Y) = \mathbf{1}[\hat{Y} \neq Y]$.

For simplicity, we have defined a fairness metric as a scalar and assume it is conditioned over a single event $\mathcal{E}$. It is easy to extend this definition to multiple events (e.g. for the fairness metric known as equalized odds) by considering a set of events $\{\mathcal{E}_j\}$ and keeping track of $\mathbb{E}_{\mathcal{D}}[f_j(\hat{Y}, Y)|\mathcal{E}_j, B]$ for each. For clarity, we demonstrate how many familiar notions of fairness can be written in the form of Definition 1 in Appendix A.5. There are other metrics that cannot be written in this form; we do not consider those here.

2.2 Fairness Metric Estimators

Our first main result is that we can bound fairness metrics of the form described above over a dataset with linear and probabilistic fairness estimates, given that certain conditions hold on the relationships between model predictions, predicted protected attribute, and the ground truth protected attribute. In order to understand this result, we define the *probabilistic* and *linear* estimators.

Intuitively, the probabilistic estimator is the population estimate of the given disparity metric weighted by each observation's probability of being in the relevant demographic group. Formally:

Definition 2 (Probabilistic Estimator)**.** For fairness metric μ with function f and event $\mathcal{E}$, the probabilistic estimator of μ for a dataset $\mathscr{D}$ is given by

$$\widehat{D}^P_\mu := \frac{\sum_{i \in \mathcal{E}} b_i f(\hat{Y}_i, Y_i)}{\sum_{i \in \mathcal{E}} b_i} - \frac{\sum_{i \in \mathcal{E}} (1 - b_i) f(\hat{Y}_i, Y_i)}{\sum_{i \in \mathcal{E}} (1 - b_i)}.$$

It is assumed that at least one observation in the dataset has had $\mathcal{E}$ occur.

Meanwhile, the linear disparity metric is the coefficient of the probabilistic estimate b in a linear regression of $f(\hat{Y}, Y)$ on b and a constant among individuals in $\mathcal{E}$. For example, in the case of demographic parity, where $f(\hat{Y}, Y) = \hat{Y}$, it is the coefficient on b in the linear regression of $\hat{Y}$ on b and a constant over the entire sample. Using the well-known form of the regression coefficient (see, e.g. (4), we define the linear estimator as:

Definition 3 (Linear Estimator)**.** For a fairness metric μ with function f and associated event $\mathcal{E}$, the linear estimator of μ for a dataset $\mathscr{D}$ is given by:

$$\widehat{D}^L_\mu := \frac{\sum_{i \in \mathcal{E}} \left(f(\hat{Y}_i, Y_i) - \overline{f(\hat{Y}, Y)} \right) (b_i - \bar{b})}{\sum_{i \in \mathcal{E}} (b_i - \bar{b})^2}$$

where $\bar{\cdot}$ represents the sample mean among event $\mathcal{E}$.

We define D^P_μ and D^L_μ to be the asymptotes of the probabilistic and linear estimators, respectively, as the identically and independently distributed sample grows large.

2.3 Bounding Fairness with Disparity Estimates

Our main result proves that when certain covariance conditions between model predictions, predicted demographic attributes, and true demographic attributes hold, we can guarantee that the linear and probabilistic estimators of disparity calculated with the *probabilistic* protected attribute serve as upper and lower bounds on *true*disparity. This result follows from the following proposition:

Proposition 1. Suppose that b is a probabilistic estimate of a demographic trait (e.g. race) given some observable characteristics Z and conditional on event $\mathcal{E}$, so that $b = \Pr[B = 1|Z, \mathcal{E}]$. Define D^P_μ as the asymptotic limit of the probabilistic disparity estimator, $\widehat{D}^P_\mu$, and D^L_μ as the asymptotic limit of the linear disparity estimator, $\widehat{D}^L_\mu$. Then:

$$D^P_\mu = D_\mu - \frac{\mathbb{E}[\text{Cov}(f(\hat{Y}, Y), B|b, \mathcal{E})]}{\text{Var}(B|\mathcal{E})} \tag{1}$$

and

$$D^L_\mu = D_\mu + \frac{\mathbb{E}[\text{Cov}(f(\hat{Y}, Y), b|B, \mathcal{E})]}{\text{Var}(b|\mathcal{E})}. \tag{2}$$

Since variance is always positive, the probabilistic and linear estimators serve as bounds on disparity when $C_{f,b|B,\mathcal{E}}$ and $C_{f,B|b,\mathcal{E}}$ are either both positive or both negative, since they are effectively separated from the true disparity by these values: if they are both positive, then D^L_μ serves as an upper bound and D^P_μ serves as a lower bound; if they are both negative, then D^P_μ serves as an upper bound and D^L_μ serves as a lower bound. Formally,

Theorem 1. Suppose that μ is a fairness measure with function f and conditioning event $\mathcal{E}$ as described above, and that $\mathbb{E}[\text{Cov}(f(\hat{Y}, Y), b|B, \mathcal{E})] > 0$ and $\mathbb{E}[\text{Cov}(f(\hat{Y}, Y), B|b, \mathcal{E}] > 0$. Then,

$$D^P_\mu \leq D_\mu \leq D^L_\mu.$$

Proposition 1 and Theorem 1, which we prove in Appendix A, subsume and generalize a result from (20). These results define the conditions under which D^L_μ and D^P_μ, easily computable quantities, serve as bounds on ground truth fairness violations — and as we show in Section 4.2, this allows us to bound the specified fairness metrics in practice when measuring predictions in existing models whenever these conditions hold. However, as we demonstrate in the next section, this also provides us with a simple method to bound fairness violations when training machine learning models.

3 Methodology for Training

We now combine our fairness estimators with existing constrained learning approaches to develop a methodology for training fair models when only a small subset labeled with ground true protected characteristics is available. The key idea to our approach is to enforce both an upper bound on the magnitude of fairness violations computed with the *probabilistic* protected attributes ($\widehat{D}^L_\mu$), while also leveraging the small labeled subset to enforce the *covariance constraints* referenced in Theorem 1. This way, as satisfaction of the covariance constraints guarantees that $\widehat{D}^L_\mu$ serves as a bound on unfairness, we ensure bounded fairness violations in models trained with probabilistic protected characteristic labels. Due to space constraints, we defer discussion of the mathematical framework underlying the ideas to Appendix B.

Problem Formulation In an ideal setting, given access to ground truth labels on the full dataset, we could simply minimize the expected risk subject to the constraint that - whichever fairness metric we have adopted - the magnitude of fairness violations do not exceed a given threshold α. However, in settings where we only have access to a small labeled subset of data, training a model by directly minimizing the expected risk subject to fairness constraints on the labeled subset may result in poor performance, particularly for complicated learning problems. Instead, we propose enforcing an upper bound on the disparity estimator as a *surrogate* fairness constraint. Recall that Theorem 1 describes conditions under which the linear estimator upper or lower bounds the true disparity; if we can *enforce* these conditions in our training process using the smaller *labeled* dataset, then our training process provides the fairness guarantees desired while leveraging the information in the full dataset.

To operationalize this idea, we recall that Theorem 1 characterizes two cases in which the linear estimator could serve as an upper bound in magnitude: in the first case, both residual covariance terms are positive, and $D_\mu \leq D^L_\mu$; in the second, both are negative, and $D^L_\mu \leq D_\mu$[1]. Minimizing risk while satisfying these constraints in each case separately gives the following two problems:

Problem 1.A.

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)] \text{ s.t. } D^L_\mu \leq \alpha; \mathbb{E}[C_{f,B|b,\mathcal{E}}] \geq 0; \mathbb{E}[C_{f,b|B,\mathcal{E}}] \geq 0$$

Problem 1.B.

$$\min_{h \in \mathcal{H}} \mathbb{E}[L(h(X), Y)] \text{ s.t. } -\alpha \leq D^L_\mu; \mathbb{E}[C_{f,B|b,\mathcal{E}}] \leq 0; \mathbb{E}[C_{f,b|B,\mathcal{E}}] \leq 0$$

To find the solution that minimizes the the fairness violation with the highest accuracy, we select:

$$h^* \in \text{argmin}_{h^*_{2a}, h^*_{2b}} \mathbb{E}[L(h(X), Y)].$$

By construction, h^* is feasible, and so satisfies $|D_\mu(h^*)| \leq \alpha$; moreover, while h^* may not be the lowest-loss predictor such that $|D_\mu| \leq \alpha$, it is the best predictor which admits the linear estimator as an upper bound on the magnitude of the disparity. In other words, it is the best model for which we can *guarantee* fairness using our measurement technique.

Remark. Note that the second covariance constraint (associated with the lower-bound, i.e. the probabilistic estimator) in each problem is necessary to rule out solution far below the desired range in the opposite sign; otherwise, a solution to Problem 1.A could have $D_\mu < -\alpha$ and to Problem 1.B $D_\mu > \alpha$, and the ultimate h^* selected could be infeasible with respect to the desired fairness constraint. (Note also that as a consequence, the probabilistic estimator will also serve as a *lower bound* for the magnitude of disparity under the selected model.)

Empirical Problem The problems above are over the full population, but in practice we usually only have samples. We thus now turn to the question of how we can solve the optimization problem with probabilistic fairness constraints empirically. We focus on the one-sided Problem 1.A for brevity but the other side follows similarly. The empirical analogue of Problem 1.A is the following:

Problem 2.A.

$$\min_{h_\theta \in \mathcal{H}} \frac{1}{n_{\mathscr{D}}} \sum_{i=1}^{n_{\mathscr{D}}} L(h_\theta(X_i), Y_i) \text{ s.t. } \widehat{D}^L_\mu(h_\theta) \leq \alpha; \widehat{C}_{f,b|B,\mathcal{E}}(h_\theta) \geq 0; \widehat{C}_{f,b|B,\mathcal{E}}(h_\theta) \geq 0$$

Solving the empirical problem. While Problem 2.A is a constrained optimization problem, it is not, except in special cases, a convex problem. Despite this, recent results (([13]),([14])) have shown that under relatively mild conditions, a primal-dual learning algorithm can be used to obtain approximate solutions with good performance guarantees.[2] In particular, if we define the *empirical Lagrangian* as:

$$\widehat{\mathcal{L}}(\theta, \vec{\mu}) = \frac{1}{n_{\mathscr{D}}} \sum_{i=1}^{n_{\mathscr{D}}} L(h_\theta(X_i), Y_i) + \mu_L \left(\widehat{D}^L_\mu(h_\theta) - \alpha\right) - \mu_{b|B}\widehat{C}_{f,b|B,\mathcal{E}} - \mu_{B|b}\widehat{C}_{f,B|b,\mathcal{E}} \qquad (3)$$

(where $\widehat{C}_{f,b|B,\mathcal{E}}$ and $\widehat{C}_{f,B|b,\mathcal{E}}$ are as in Problem 1.A), the optimization problem can be viewed as a min-max game between a primal (θ) and dual (μ) player where players are selecting θ and μ to $\max_\mu \min_\theta \widehat{\mathcal{L}}(\theta, \mu)$. Formally, Algorithm 1 in the appendix provides pseudocode for a primal-dual learner similar to ([14]; [17]), etc. specialized to our setting; adapting and applying Theorem 3 in ([14]), provides the following guarantee:

[1] Note that as a result of Proposition 1, when $C_{f,b|B,\mathcal{E}}$ and $C_{f,B|b,\mathcal{E}}$ are both positive, the true fairness metric is necessarily is forced to be positive, and symmetrically for for negative values.

[2] For the special case of linear regression with mean-squared error losses, we provide a closed-form solution to the primal problem. This can be used for a heuristic solution with appropriate dual weights.

Theorem 2. Let $\mathcal{H}$ have a VC-dimension d, be *decomposable*, and finely cover its convex hull. Assume that y takes on a finite number of values, the induced distribution $x|y$ is non-atomic for all y, and Problem 2.A has a feasible solution. Then if Algorithm 1 is run for T iterations, and $\tilde{\theta}$ is selected by uniformly drawing $t \in \{1...T\}$, the following holds with probability $1-\delta$: For each target constraint $\ell \in \{D_\mu^L, C_{f,b|B,\mathcal{E}}, C_{f,B|b,\mathcal{E}}\}$,

$$\mathbb{E}[\ell(h_{\tilde{\theta}})] \leq c_i + \mathcal{O}\left(\frac{d \log N}{\sqrt{N}}\right) + \mathcal{O}\left(\frac{1}{T}\right) \text{ and } \mathbb{E}[L(h_{\tilde{\theta}}, y)] \leq P^* + \mathcal{O}\left(\frac{d \log N}{\sqrt{N}}\right)$$

where P^* is the optimal value of Problem 2.A.

The theorem provides an *average-iterate* guarantee of approximate feasibility and optimality when a solution is drawn from the empirical distribution. Note that it is not a priori obvious whether our bounds remain informative over this empirical distribution, but we show in Appendix A that the covariance conditions holding on average imply that our bounds hold on average:

Proposition 2. Suppose $\tilde{\theta}$ is drawn from the empirical distribution produced by Algorithm 1. If:

$$\mathbb{E}\left[\mathbb{E}[\mathrm{Cov}(f(h_{\tilde{\theta}}(X), B))|\mathcal{E}, b]|\tilde{\theta}\right] \geq 0 \text{ and } \mathbb{E}\left[\mathbb{E}[\mathrm{Cov}(f(h_{\tilde{\theta}}(X), b))|\mathcal{E}, B]|\tilde{\theta}\right] \geq 0,$$

Then $\mathbb{E}D_\mu(h_{\tilde{\theta}}) \leq \mathbb{E}D_\mu^L(h_{\tilde{\theta}})$.

Remark. Combining Theorem 2 and Proposition 2 guarantees that a randomized classifier with parameters drawn according to the empirical distribution from Algorithm 1 will approximately meet our disparity bound goals *on average*. Without stronger assumptions, this is all that can be said; this is a general limitation of game-based empirical optimization methods, since they correspond equilibrium discovery, and only mixed-strategy equilibria are guaranteed to exit. In practice, however, researchers applying similar methods select the final or best feasible iterate of their model, and often find feasible good performance (17; 46); thus in our results section, we compare our best-iterate performance to other methods.

4 Empirical Evaluation

We now turn to experiments of our disparity measurement and fairness enforcing training methods on predicting voter turnout. We provide additional experiments on the COMPAS dataset (5), as well as on simulated data, in Appendix F and Appendix G, respectively.

4.1 Data

L2 Dataset. The L2 dataset provides demographic, voter, and consumer data from across the United States collected by the company L2. Here, we consider the task of predicting voter turnout for the general election in 2016 and measuring model fairness violations with respect to Black and non-Black voters. This application is particularly relevant since race/ethnicity information is often not fully available (27), and much of voting rights law hinges on determining whether there exists racially polarized voting and/or racial disparities in turnout (6). We focus on the six states with self-reported race labels (North Carolina, South Carolina, Florida, Georgia, Louisiana, and Alabama). We denote $\hat{Y} = 1$ if an individual votes in the 2016 election and $\hat{Y} = 0$ otherwise; refer to Appendix C.1 for a detailed description of this dataset.

Race Probabilities. The L2 dataset provides information on voters' first names, last names, and census block group, allowing the use of Bayesian Improved (Firstname and) Surname Geocoding Method (BISG/BIFSG) for estimating race probabilities (18; 19; 27). We obtain our priors through the decennial Census in 2010 on the census block group level. AUC for BISG/BIFSG across the six states we investigate in the L2 data ranges from 0.85-0.90. Further details on how we implement BISG/BIFSG for the L2 data and its performance can be found in Appendix C.2.

4.2 Measurement

In this section, we showcase our method of bounding true disparity when race is unobserved. Given *1)* model predictions on a dataset with probabilistic race labels and *2)* true race labels for a small

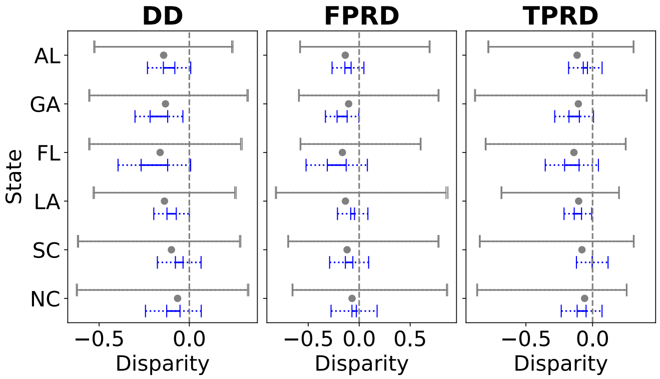

Figure 1: Comparison of our method of bounding true disparity (blue) to the method proposed in Kallus et al. ([31]) (grey), using a logistic regression model to predict voter turnout in six states. Only a small subset (here, $n = 1,500$, i.e. 1%) of the data contains information on true race. The grey dot represents true disparity. Both methods successfully bound true disparity within its 95% standard errors, but our estimators provide tighter bounds.

subset of that data, we attempt to obtain bounds on three disparity measures: demographic disparity (DD), false positive rate disparity (FPRD), and true positive rate disparity (TPRD).

4.2.1 Experimental Design and Comparisons.

Setup. To simulate measurement of fairness violations on predictions from a pre-trained model with limited access to protected attribute, we first train unconstrained logistic regression models with an 80/20 split of the available L2 data for each state. Then, in order to simulate realistic data access conditions, we measure fairness violations on a random subsample of the test set ($n = 150,000$), with 1% ($n = 1,500$) of this sample including ground truth race labels to constitute the labeled subset. We do this by first checking the covariance constraints on the labeled subset, and then calculating $\widehat{D}_L$ and $\widehat{D}_P$ on the entire set of $150,000$ examples sampled from the test set. We also compute standard errors for our estimators as specified by the procedure in Appendix Section B. To evaluate our method, we measure true fairness violations on the $150,000$ examples sampled from the test set, and check to see whether we do in fact bound the true fairness violations within standard error. Further information about our unconstrained models can be found in Appendix Section D.1. We present our results in Figure 1.

Comparisons. We compare our method of estimating fairness violations using probabilistic protected characteristic labels to the method described in Kallus et al. ([31]), which is one of the only comparable methods in the literature. We will refer to as KDC from here on. Details of KDC and our implementation can be found in Appendix Section D.2.

4.2.2 Results

Figure 1 compares our method of estimating disparity (blue) with KDC (grey) for the three disparity measures and all six states. This figure shows estimates when training a logistic regression model, and Figure 5 in the Appendix shows similar results for training random forests. Across all experiments, both KDC's and our estimators always bound true disparity. However, we observe two crucial differences: *1)* our bounds are markedly tighter (3.8x smaller on average, and as much as 5.5x smaller) than KDC, and as a result *2)* our bounds almost always indicate the direction of true disparity. When they do not, it is due to the standard error which shrinks with more data. By contrast, KDC's bounds consistently span[-0.5, 0.5], providing limited utility even for directional estimates.

4.3 Training

In this section, we demonstrate the efficacy of our approach to training fairness-constrained machine learning models. Following our algorithm in Section 3, we train models with both covariance conditions necessary for the fairness bounds to hold and also constrain the upper bound on absolute value of disparity, $\widehat{D}^L_\mu$, to be below some bound α. We find that our method *1)* results in lower true disparity on the test set than using the labeled subset alone, or using prior methods to bound disparity; *2)* more frequently reaches the target bound than other techniques; and *3)* often incurs less of an accuracy trade-off when enforcing the same bound on disparity compared to related techniques.

4.3.1 Experimental Design and Comparisons.

Experimental Design. We demonstrate our technique by training logistic regression models to make predictions with bounded DD, FPRD, and TPRD across a range of bounds. We include results for neural network models in Appendix Section E.7. We train these models on the data from Florida within the L2 dataset, as it has the largest unconstrained disparity among the six states, see Figure 1.We report the mean and standard deviations of our experimental results over ten trials. For each trial, we split our data ($n = 150,000$) into train and test sets, with a 80/20 split. From the training set, we subsample the labeled subset so that it is 1% of the total data ($n = 1,500$). To enforce fairness constraints during training, we solve the empirical problem 3A and its symmetric analogue, which enforces negative covariance conditions and $\widehat{D}_{\mu}^{L}$ as a (negative) lower bound. We use the labeled subset to enforce adherence to the covariance conditions during training. We use the remainder of the training data, as well as the labeled subset, to enforce the constraint on $\widehat{D}_{\mu}^{L}$ during training. As noted in Section 3, our method theoretically guarantees a near-optimal, near-feasible solution *on average* over $\theta^{(1)}...\theta^{(T)}$. However, following Wang et al. (46), for each of these sub-problems, we select the best iterate $\theta^{(t)}$ which satisfies the bound on $\widehat{D}_{\mu}^{L}$ on the training set, the covariance constraints on the labeled subset, and that achieves the lowest loss on the training set. We report our results on the solution between these two sub-problems that is feasible and has the lowest loss. We present the accuracy and resulting disparity of model predictions on the test set after constraining fairness violations during training for a range of metrics (DD, FPRD, TPRD), across a range of bounds (0.04, 0.06, 0.08, 0.10) for our method as well as three comparisons, described below, in Figure 2. Further details about the experimental setup can be found in Appendix Section E.1.

Comparisons. We compare our results for enforcing fairness constraints with probabilistic protected attribute labels to the following methods: *1)* A model trained *only* on the labeled subset with true race labels, enforcing a fairness constraint over those labels. This is to motivate the utility of using a larger dataset with noisy labels when a smaller dataset exists on the same distribution with true labels. To implement this method, we use the non-convex constrained optimization technique from Chamon et al. (14) to enforce bounds on fairness violations calculated directly on ground-truth race labels, as we describe in greater detail in Appendix E.2. *2)* We compare with a recent method by Wang et al. (46) for enforcing fairness constraints on data with noisy protected attributes and a labeled auxiliary set, which is based on an extension of Kallus et al. (31)'s disparity measurement method. This method guarantees that the relevant disparity metrics will be satisfied within the specified slack, which we take as a bound. However, their implementation does not consider DD – further details on this method can be found in Appendix Section E.3. *3)* We compare with a method for enforcing fairness with incomplete demographic labels introduced by Mozannar et al. (36), which essentially modifies Agarwal et al.'s (2) fair training approach to only enforce a fairness constraint on the available demographically labeled data. This method also guarantees that the relevant disparity metrics will be satisfied within specified slack, which we modify to be comparable to our bound. Details on this approach can be found in Appendix E.4. In Appendix Section E.6, we also compare to two other models: *1)* an "oracle" model trained to enforce a fairness constraint over the ground-truth race labels on the whole dataset; and *2)* a naive model which ignores label noise and enforces disparity constraints directly on the probabilistic race labels, thresholded to be in $(0, 1)$.

4.4 Results

We display our results in Figure 2, with additional results in Sections E and G of the Appendix. Looking at the top row of the figure, we find that our method, in all instances, reduces disparity further than training on the labeled subset alone (blue vs. orange bars in Figure 2), than using Wang et al. (46) (blue versus green bars in Figure 2), and than using Mozannar et al. (36) (blue versus pink bars in Figure 2). Second, our method satisfies the target fairness bound on the test set more often than the other methods (12 out of 12 times, as opposed to 0, 1, and 0 for labeled subset, Wang, and Mozannar respectively). In other words, the disparity bounds our method learns on the train set generalize better to the test set than the comparison methods. We note that deviations from the enforced bound on the test set, when they arise, are due to generalization error in enforcing constraints from the train to the test set, and because our training method guarantees *near*-feasible solutions.

The bottom row of the figure shows how our method performs with respect to accuracy in comparison to other methods. The results here are more variable; however, we note that this dataset seems to

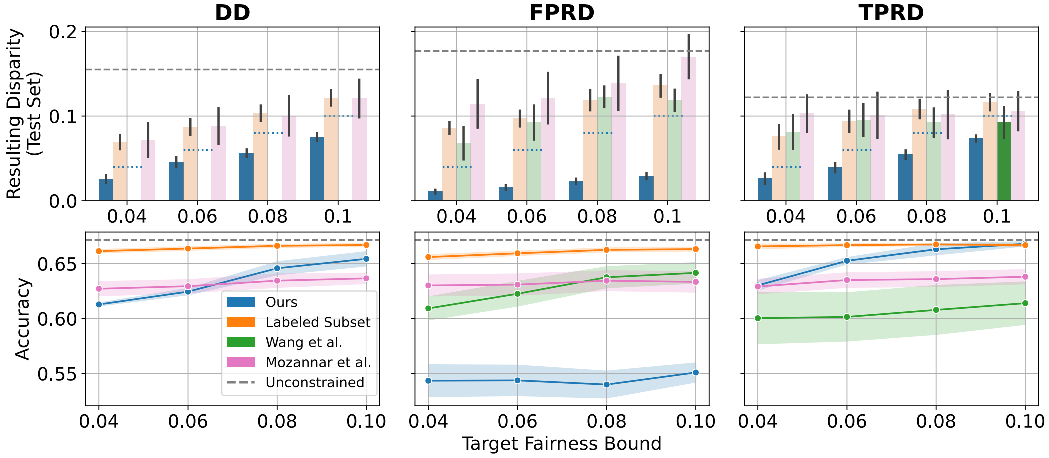

Figure 2: Mean and standard deviation of resulting disparity (top, y-axis) and accuracy (bottom, y-axis) on the test set after enforcing the target fairness bounds (x-axis) on our method (blue); only using the labeled subset with true labels (orange) and Wang et al. (46) (green) over ten trials. On the top row, we fade bars when the mean does not meet the desired bound, which is indicated by the dotted blue lines. The dashed grey line in all plots indicates disparity from the unconstrained model.

exhibit a steep fairness-accuracy tradeoff—yet despite our method reducing disparity much farther than all other methods (indeed, being the only metric that reliably bounds the resulting disparity in the test set), we often perform comparably or slightly better. For example, when mitigating TPRD, our method mitigates disparity much farther than Mozannar et al. (36) and Wang et al. (46), yet outperforms both with respect to accuracy. In the case of FPRD, while our method does exhibit worse accuracy, these sets of experiments also exhibit the largest difference in disparity reduction between our method and the other methods, which may make such an accuracy difference inevitable. Similarly, the accuracy discrepancy between the labeled subset method and our method is reasonable given the fairness-accuracy trade-off.

5 Related Work

While there are many methods available for training models with bounded fairness violations (2; 24; 7), the vast majority of them require access to the protected attribute at training or prediction time. While there are other works which assume access *only* to noisy protected attribute labels (46), and *no* protected attribute labels (34), or a even a labeled subset of protected attribute labels, but without an auxiliary set to generate probabilistic protected attribute estimates (30); very few works mirror our data access setting. One exception, from which we draw inspiration, is Elzayn et al. (20); that work studies in detail the policy-relevant question of whether Black U.S. taxpayers are audited at higher rates than non-Black taxpayers, and uses a special case of our Theorem 1 (for measurement *only*). In this paper, we formalize and extend their technique to bound a wide array of fairness constraints, and introduce methods to *train* fair models given this insight.

Within the set of techniques with a different data access paradigm, we differ from many in that we leverage information about the relationship between probabilistic protected attribute labels, ground truth protected attribute, and model predictions to measure and enforce our fairness bounds. Thus, while we do require the covariance conditions to hold in order to enforce our fairness bounds, we note that these are requirements we can *enforce* during training, unlike assumptions over noise models as in other approaches to bound true disparity with noisy labels (9; 29; 11). Intuitively, leveraging some labeled data can allow us to have less of an accuracy trade-off when training fair models, as demonstrated with our comparison to Wang et al. (46). In this case, using this data means we do not have to protect against every perturbation within a given distance to the distribution, as with distributionally robust optimization (DRO). Instead, need only to enforce constraints on optimization—in our experimental setting, we see that this can lead to a lower fairness-accuracy trade-off.

References

[1] Statutes enforced by the voting section, Accessed 2023.

[2] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *International Conference on Machine Learning* (2018), PMLR, pp. 60–69.

[3] Andrus, M., Spitzer, E., Brown, J., and Xiang, A. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (2021), pp. 249–260.

[4] Angrist, J. D., and Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

[5] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica* (2016).

[6] Barber, M., and Holbein, J. B. 400 million voting records show profound racial and geographic disparities in voter turnout in the united states. *Plos one 17*, 6 (2022), e0268134.

[7] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR abs/1810.01943* (2018).

[8] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. Fairlearn: A toolkit for assessing and improving fairness in AI. Tech. Rep. MSR-TR-2020-32, Microsoft, May 2020.

[9] Blum, A., and Stangl, K. Recovering from biased data: Can fairness constraints improve accuracy? *arXiv preprint arXiv:1912.01094* (2019).

[10] Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops* (2009), IEEE, pp. 13–18.

[11] Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning* (2021), PMLR, pp. 1349–1361.

[12] CFPB. Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment, 2014.

[13] Chamon, L., and Ribeiro, A. Probably approximately correct constrained learning. *Advances in Neural Information Processing Systems 33* (2020), 16722–16735.

[14] Chamon, L. F., Paternain, S., Calvo-Fullana, M., and Ribeiro, A. Constrained learning with non-convex losses. *IEEE Transactions on Information Theory* (2022).

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

[16] Corbett-Davies, S., and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[17] Cotter, A., Jiang, H., Gupta, M. R., Wang, S., Narayan, T., You, S., and Sridharan, K. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *J. Mach. Learn. Res. 20*, 172 (2019), 1–59.

[18] ELLIOTT, M. N., FREMONT, A., MORRISON, P. A., PANTOJA, P., AND LURIE, N. A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health services research 43*, 5p1 (2008), 1722–1736.

[19] ELLIOTT, M. N., MORRISON, P. A., FREMONT, A., MCCAFFREY, D. F., PANTOJA, P., AND LURIE, N. Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology 9* (2009), 69–83.

[20] ELZAYN, H., SMITH, E., HERTZ, T., RAMESH, A., FISHER, R., HO, D., AND GOLDIN, J. Measuring and mitigating racial disparities in tax audits. *Working Paper* (2023).

[21] Codified at 15 U.S.C. § 1691, et seq., 1974.

[22] Codified at 15 U.S.C. § 1681, et seq., 1970.

[23] FISCELLA, K., AND FREMONT, A. M. Use of geocoding and surname analysis to estimate race and ethnicity. *Health services research 41*, 4p1 (2006), 1482–1500.

[24] HARDT, M., PRICE, E., AND SREBRO, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (2016).

[25] HILL, K. Wrongfully accused by an algorithm. *The New York Times, June 24* (2020).

[26] HOUSE, W. Blueprint for an ai bill of rights: Making automated systems work for the american people, 2022.

[27] IMAI, K., AND KHANNA, K. Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis 24*, 2 (2016), 263–272.

[28] ISAAC, M. Meta agrees to alter ad technology in settlement with u.s., 2022.

[29] JIANG, H., AND NACHUM, O. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics* (2020), PMLR, pp. 702–712.

[30] JUNG, S., CHUN, S., AND MOON, T. Learning fair classifiers with partially annotated group labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10348–10357.

[31] KALLUS, N., MAO, X., AND ZHOU, A. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science 68*, 3 (2022), 1959–1981.

[32] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[33] KOH, H. K., GRAHAM, G., AND GLIED, S. A. Reducing racial and ethnic disparities: the action plan from the department of health and human services. *Health affairs 30*, 10 (2011), 1822–1829.

[34] LAHOTI, P., BEUTEL, A., CHEN, J., LEE, K., PROST, F., THAIN, N., WANG, X., AND CHI, E. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems 33* (2020), 728–740.

[35] MATSUSAKA, J. G., AND PALDA, F. Voter turnout: How much can we explain? *Public choice 98*, 3-4 (1999), 431–446.

[36] MOZANNAR, H., OHANNESSIAN, M., AND SREBRO, N. Fair learning with private demographic data. In *International Conference on Machine Learning* (2020), PMLR, pp. 7066–7075.

[37] NARAYANAN, A. Translation tutorial: 21 fairness definitions and their politics. In *Proc. conf. fairness accountability transp., new york, usa* (2018), vol. 1170, p. 3.

[38] PRESIDENT, U. Exec. order no. 13985 86 fed. reg. 7009, advancing racial equity and support for underserved communities through the federal government.

[39] ROY L. AUSTIN, J. Expanding our work on ads fairness, 2022.

[40] ROY L. AUSTIN, J. Race data measurement and meta's commitment to fair and inclusive products, 2022.

[41] SHALIZI, C. R. The truth about linear regression. *Online Manuscript. http:///www. stat. cmu. edu/~ cshalizi/TALR* (2015).

[42] U.S. EXECUITVE ORDER 14091. Exec. order no. 14091 88 fed. reg. 10825, further advancing racial equity and support for underserved communities through the federal government, 2023.

[43] U.S. EXECUTIVE ORDER 13985. Exec. order no. 13985 86 fed. reg. 7009, advancing racial equity and support for underserved communities through the federal government, 2021.

[44] VERMA, S., AND RUBIN, J. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)* (2018), IEEE, pp. 1–7.

[45] VOICU, I. Using first name information to improve race and ethnicity classification. *Statistics and Public Policy 5*, 1 (2018), 1–13.

[46] WANG, S., GUO, W., NARASIMHAN, H., COTTER, A., GUPTA, M., AND JORDAN, M. Robust optimization for fairness with noisy protected groups. *Advances in neural information processing systems 33* (2020), 5190–5203.

[47] ZAFAR, M. B., VALERA, I., ROGRIGUEZ, M. G., AND GUMMADI, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics* (2017), PMLR, pp. 962–970.

[48] ZHANG, Y. Assessing fair lending risks using race/ethnicity proxies. *Management Science 64*, 1 (2018), 178–197.

[49] ZHU, Z., YAO, Y., SUN, J., LI, H., AND LIU, Y. Weak proxies are sufficient and preferable for fairness with missing sensitive attributes.

[50] ŽLIOBAITĖ, I. On the relation between accuracy and fairness in binary classification. In *The 2nd Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2015) workshop at ICML* (2015), vol. 15.