Estimating and Implementing Conventional Fairness Metrics With Probabilistic Protected Features

***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 practically important applications this protected attribute is largely unavailable. Still, AI systems used in sensitive business and government applications—such as housing ad delivery and credit underwriting—are increasingly legally required to measure and mitigate their bias. 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 as well as the COMPAS dataset. 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 in comparison to an unconstrained model while often incurring lesser fairness-accuracy trade-offs than other fair optimization methods with limited access to protected attributes.**

***Index Terms*—algorithmic fairness, fair machine learning, anti-discrimination, disparity reduction, probabilistic protected attribute**

I. Introduction

In both the private and public sectors, organizations are facing increasing pressure to ensure they use equitable machine learning systems, whether through legal obligations or social norms [1, 2, 3, 4, 5]. 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 [6, 7]. Similarly, recent Executive Orders in the United States [3, 8] direct government agencies to measure and mitigate disparity resulting from or exacerbated by their programs, including in the "design, develop[ment], acqui[sition], and us[e] [of] artificial intelligence and automated systems" [8].

Yet both companies [9] and government agencies [3] 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 [10, 11], the limited availability of protected attribute data represents a significant challenge in assessing algorithmic fairness and makes 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 the rest of the dataset. These probabilistic estimates may be generated using Bayesian Improved Surname Geocoding (BISG) [12, 13] or any predictive model which can output probabilistic predictions.

We leverage this limited labeled data to establish (or ensure, in the case of training) whether a certain condition holds regarding the relationship between the model's predictions, the probabilistic protected attributes, and the ground truth protected attributes hold. In particular, this condition is that two residual correlations – the residual correlation between the probabilistic proxy and the outcome of interest conditioned on ground truth race, and the residual correlation between ground truth race status and the outcome conditional on the proxy – share the same sign. Given this condition, 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 II.

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 [6, 15, 16, 17, 18]. 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 [18]. 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. [19].

While our approach has stronger data requirements than recent work in similar domains [20, 21] 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 II);
2) We introduce a new method of training models with near-optimal and near-feasible bounded unfairness with limited protected attribute data (Section III);
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 [22], and also demonstrate our results on COMPAS data [23], a common dataset used in related work (Section IV) In addition, we present some experiments on synthetic data which outline the conditions under which our technique is the most effective: relatively complex problems with little access to labeled data.

The rest of this paper proceeds as follows: in the remainder of this section (Section I-A), we describe in greater detail two examples of real-world settings in which our approach may be applicable. Following this, in Section II, we describe our method of measuring disparities in data regimes with limited access to protected attribute labels, then in Section III we leverage our measurement results to develop our training techniques which bound unfairness in the resulting model. We display our experimental evaluation of our method in Section IV, including comparisons to related bias measurement [20] and fair training techniques [21, 24]. Finally, we end our paper with our review of the related work (Section V) and Conclusion (Section VI).

A. Correspondence to real-world Settings

We now highlight two real-world examples which correspond to our setting. First, consider the example of Meta Platforms ("Meta"). Meta is the parent company of Facebook, a social media platform with a large advertising business. Meta uses machine learning to identify users likely to interact with particular ads [25]. The Department of Housing and Urban Development brought a lawsuit [26] under the Fair Housing Act alleging algorithmic discrimination by Meta. As part of a settlement resolving the suit [7], Meta agreed to build software called the *Variance Reduction System* (VRS) [6] which uses a differentially-private version of BISG to estimate deviation of delivery rates by group relative to an underlying eligible audience [27]. In accordance with the recommendations of civil rights groups [28], Meta also began to work with third-party survey administrator YouGov to prompt users to provide individual race off-platform (with privacy protection via tools from secure multi-party computation) [18, 29].

Second, consider the example of government agencies such as the Internal Revenue Service (IRS). IRS, like many other government agencies, does not collect taxpayer data on race [30], yet recent executive orders have required equity (disparity) assessments [3] and consideration of protections from "algorithmic discrimination" [8]. A paper by academic and government researchers, [19], combines BISG for the population of taxpayers with a publicly available administrative dataset (voter registration data) that does contain ground truth and can be matched to a subset of taxpayers and uses this combined dataset to assess audit rate disparity.

In both these examples, disparity estimation is an important goal hindered by a lack of individual-race data, yet probabilistic estimates of race via BISG are available and race data can be obtained for small subset of individuals. The key features thus correspond to the setting we describe formally in Section II-A These prominent examples are likely representative of scenarios faced by many other private and public sector actors; hence, our approach is likely to be broadly useful. Indeed, while these instances are some of the first legally required investigations of disparities arising from algorithmic systems [31], they are unlikely to be the last: along with recent executive orders [8, 32] and the Blueprint for an AI Bill of Rights [4], a recent advanced notice of proposed rulemaking (ANPR) from the Federal Trade Comission (FTC) suggests the possibility of stricter rules around the deployment of discriminatory systems [33]. Increased regulation of algorithmic decision systems requires the development of bias measurement and mitigation techniques which correspond to the realities of data access, and legal scrutiny, that exist on the ground.

II. 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.

A. 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. the *unlabeled set*, $\mathscr{D}_U$, consists of observations $\{(X_i, Y_i, Z_i)\}_{i=1}^{n_U}$ and the *labeled set*, $\mathscr{D}_L$, 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 $b_i := \Pr[B_i|Z_i]$. All three datasets are assumed to be independent and drawn from the same underlying population. 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 [34, 35, 36] 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 [37, 38] 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 [39] (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 Table II-A. There are other metrics that cannot be written in this form; we do not consider those here.

Metric	$\mathbf{f}(\mathbf{h}(\mathbf{X}), \mathbf{Y})$	$\mathcal{E}$
Accuracy	$\mathbf{1}[h \neq y]$	{true}
Demographic Parity	$\mathbf{1}[h = 1]$	{true}
True Positive Rate Parity	$\mathbf{1}[h \neq y]$	$\{y = 1\}$
False Positive Rate Parity	$\mathbf{1}[h \neq y]$	$\{y = 0\}$
True Negative Rate Parity	$\mathbf{1}[h \neq y]$	$\{y = 0\}$
False Negative Rate Parity	$\mathbf{1}[h \neq y]$	$\{y = 1\}$

TABLE I: Many fairness metrics can be written in the form required by our formulation. For concreteness, we provide a table based on [40, 41] summarizing the choice of f and $\mathcal{E}$ that correspond to the many of the most prominent definitions that can be written in our formulation.

B. 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

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

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. [42], 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:

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

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

We define D_{μ}^{P} and D_{μ}^{L} to be the asymptotes of the probabilistic and linear estimators, respectively, as the identically and independently distributed sample grows large.

C. 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_{μ}^{P} as the asymptotic limit of the probabilistic disparity estimator, $\widehat{D}_{\mu}^{P}$, and D_{μ}^{L} as the asymptotic limit of the linear disparity estimator, $\widehat{D}_{\mu}^{L}$. Then:

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

and

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

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_{μ}^{L} serves as an upper bound and D_{μ}^{P} serves as a lower bound; if they are both negative, then D_{μ}^{P} serves as an upper bound and D_{μ}^{L} 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 $\mathbb{E}[\mathrm{Cov}(f(\hat{Y}, Y), b|B, \mathcal{E})] > 0$ and $\mathbb{E}[\mathrm{Cov}(f(\hat{Y}, Y), B|b, \mathcal{E}] > 0$. Then,

$$D_{\mu}^{P} \leq D_{\mu} \leq D_{\mu}^{L}.$$

Proposition 1 and Theorem 1, which we prove in Appendix A, subsume and generalize a result from [19]. These results define the conditions under which D_{μ}^{L} and D_{μ}^{P} serve as bounds on ground truth fairness violations; since we can use $\widehat{D}_{\mu}^{P}$ and $\widehat{D}_{\mu}^{L}$ to estimate these quantities from data (up to sampling uncertainty[1]) Theorem 1 thus provides a path to bound fairness metrics as long as the assumed conditions hold. We demonstrate the efficacy of this method for measuring fairness metrics of existing models in practice in Section IV-B. 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.

[1]We show how to compute these standard errors in Appendix A-C, and then take the extremes of the confidence intervals as our bounds.

III. 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 ($\widehat{D}_{\mu}^{L}$), 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 $\widehat{D}_{\mu}^{L}$ 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 $D_{\mu} \leq D_{\mu}^{L}$; in the second, both are negative, and $D_{\mu}^{L} \leq D_{\mu}$[2]. Minimizing risk while satisfying these constraints in each case separately gives the following two problems:

Problem 1.A.

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

[2]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.

Problem 1.B.

$$\begin{aligned}\min_{h\in\mathcal{H}} \mathbb{E}[L(h(X),Y)]\\ \text{s.t. } -\alpha \le D_\mu^L\\ \mathbb{E}[C_{f,B|b,\mathcal{E}}] \le 0\\ \mathbb{E}[C_{f,b|B,\mathcal{E}}] \le 0\end{aligned}$$

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

$$h^* \in \text{argmin}_{h_{1a}^*, h_{1b}^*} \mathbb{E}[L(h(X),Y)],$$

where h_{1a}^*, h_{1b}^* are the solutions to Problems 1.A and 1.B. By construction, h^* is feasible, and so satisfies $|D_\mu(h^*)| \le \alpha$; moreover, while h^* may not be the lowest-loss predictor such that $|D_\mu| \le \alpha$, 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.

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

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 [14, 43] have shown that under relatively mild conditions, a primal-dual learning algorithm can be used to obtain approximate solutions with good performance guarantees.[3] In particular, if we define the *empirical Lagrangian* as:

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

(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], [44], etc. specialized to our setting; adapting and applying Theorem 3 in [14], provides the following guarantee:

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 $\ell \in \{D_\mu^L, C_{f,b|B,\mathcal{E}}, C_{f,B|b,\mathcal{E}}\}$,

$$\mathbb{E}[\ell(h_{\tilde{\theta}})] \le c_i + \mathcal{O}\left(\frac{d\log N}{\sqrt{N}}\right) + \mathcal{O}\left(\frac{1}{T}\right)$$

and

$$\mathbb{E}[L(h_{\tilde{\theta}},y)] \le P^* + \mathcal{O}\left(\frac{d\log N}{\sqrt{N}}\right)$$

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:

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

and

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

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

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

[3] 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.

feasible iterate of their model, and often find feasible good performance [21, 44]; thus in our results section, we compare our best-iterate performance to other methods.

IV. Empirical Evaluation

We now turn to experiments of our disparity measurement and fairness enforcing training methods on predicting voter turnout as well as on the COMPAS dataset [45]. In addition, we provide experiments on simulated data in order to outline the conditions under which our method is most successful, and in particular, outperforms relying on training a model with the labeled subset alone, which we expand upon in Appendix G.

A. Data

We perform experiments over two datasets: the L2 dataset [46] and the COMPAS dataset [23]. In both of these datasets, the demographic attribute to which we pay attention is race.

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 [13], and much of voting rights law hinges on determining whether there exists racially polarized voting and/or racial disparities in turnout [47]. 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-A for a detailed description of this dataset. We select seven features as predictors in our model based on data completeness and predictive value: gender, age, estimated household income, estimated area median household income, estimated home value, area median education, and estimated area median housing value. Information on our selection process, pre-processing, and distribution of these features are presented in Appendix Section C-A. We denote $\hat{Y} = 1$ if a voter shows up to vote for the 2016 election and $\hat{Y} = 0$ otherwise. The baseline rates of voter turnout range between 52-63% across the six states (see more information in Section C-A in the Appendix).

L2 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 [12, 13, 48]. 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-B.

COMPAS Dataset. We also evaluate our measurement and training methods on models trained on the COMPAS [45] dataset. The COMPAS algorithm is used by parole officers and judges across the United States to determine a criminal's risk of recidivism, or re-committing the same crime. In 2016, ProPublica released a seminal article [45] detailing how the algorithm is systematically biased against Black defendants. The dataset used to train the algorithm has since been widely used as benchmarks in the fair machine learning literature. We use the eight features used in previous analyses of the dataset as predictors in our model: the decile of the COMPAS score, the decile of the predicted COMPAS score, the number of prior crimes committed, the number of days before screening arrest, the number of days spent in jail, an indicator for whether the crime committed was a felony, age split into categories, and the score in categorical form. Further information about our preparation of the COMPAS dataset can be found in Section F of the Appendix.

COMPAS Race Probabilities. In the COMPAS dataset, we generate estimates of race (Black vs. non-Black) based on first name and last name using a LSTM model used in Zhu et al. [49] that was trained on voter rolls from Florida. Accuracy of these models is 73% while the AUC is 86%. Further detail can be found in Appendix F.

B. Fairness 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 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).

1) Experimental Design: 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 data: in the case of L2, this is state by state. Then, in order to simulate realistic data access conditions, we measure fairness violations on a random subsample of the test set, with a percentage of this sample including ground truth race labels to constitute the labeled subset which we use to calculate the covariance constraints. In the case of the L2 data, the random subsample over which we measure fairness violations has $n = 150,000$, with 1% ($n = 1,500$) of this sample including ground truth race labels to constitute the labeled subset. In the case of the COMPAS dataset, which is much smaller, we use the entire test set, with $n = 1,226$, and we construct the labeled subset by sampling 50% of the test set ($n = 613$).

We first check the covariance constraints on the labeled subset, and then calculate $\widehat{D}_L$ and $\widehat{D}_P$ on the entire set of 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 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-A. We present our results in Figure 1, which shows the results

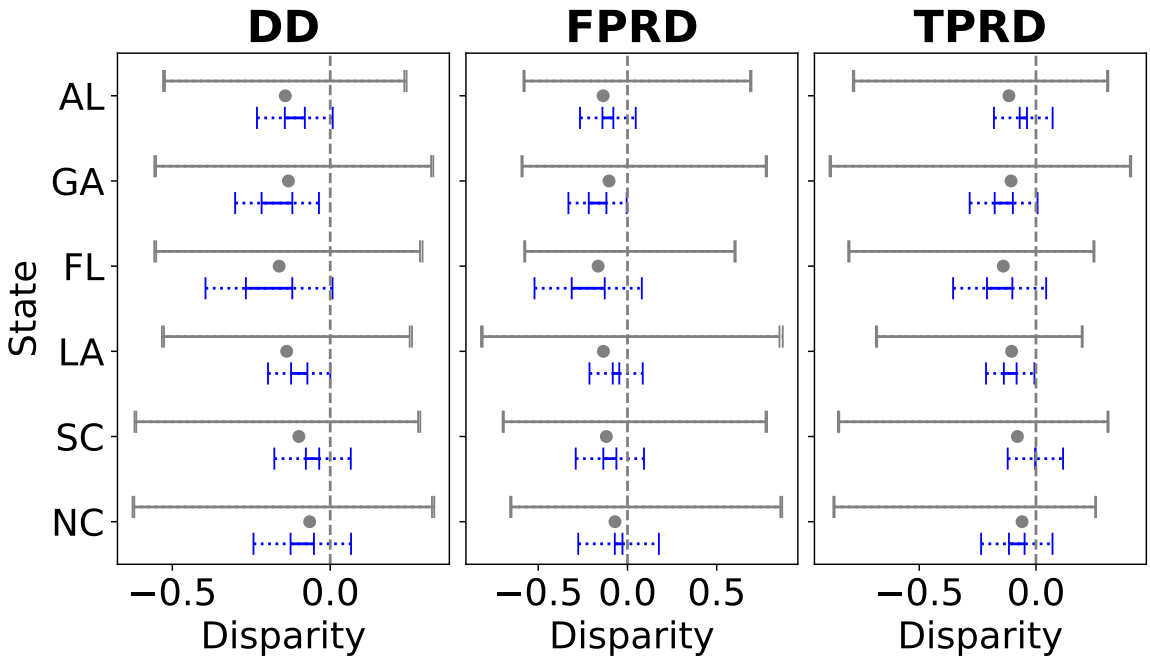


Fig. 1: (**Bounding Disparity in L2 Data**) Comparison of our method of bounding true disparity (blue) to the method proposed in Kallus et al. [20] (grey), using a logistic regression model to predict voter turnout in six states. We compare results across three disparity measures: demographic disparity (DD), false positive rate disp. (FPRD), and true positive rate disp. (TPRD). 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. The dashed lines represent 95% confidence intervals. Both methods successfully bound true disparity within its 95% standard errors, but our estimators provide much tighter bounds.

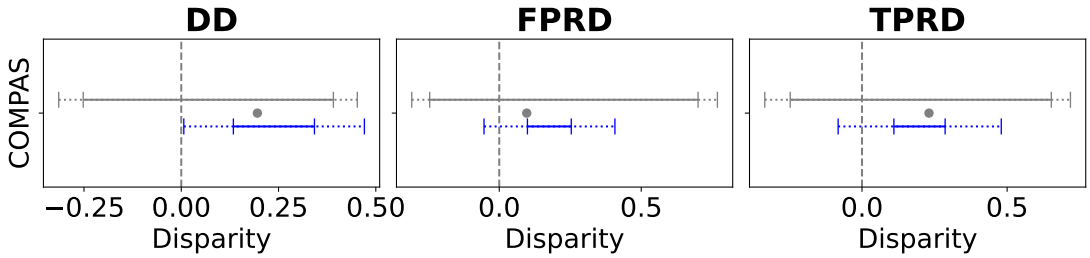


Fig. 2: (**Bounding Disparity in COMPAS Data**) Comparison of our method of bounding true disparity (blue) to the method proposed in Kallus et al. [20] (grey), using a logistic regression model to predict two-year recidivism on the COMPAS dataset. We access disparity over the same measures as in Figure 1. The grey dot represents true disparity. The dashed lines represent 95% confidence intervals. Both methods always bound true disparity within the 95% standard errors, but our method provides tighter bounds.

over the L2 data, and Figure 2, which shows the results over the COMPAS data.

2) Comparisons: We compare our method of estimating fairness violations using probabilistic protected characteristic labels to the method described in Kallus et al. [20], 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-B.

3) Results: We first analyze our results on voter data. Figure 1 compares our method of estimating disparity (blue) with KDC (grey) for the three disparity measures on the six states we consider. This figure shows estimates when training a logistic regression model, and Figure 8 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.

We now turn to the COMPAS data. Similar to the L2 data, our bounds are consistently tighter than KDC, albeit to a lesser extent in this case since the COMPAS dataset is significantly smaller (1.69x on average, and up to 2.04x smaller). We emphasize that, unlike KDC, our estimators are always within the same sign as the true disparity, barring the standard errors which shrink as the data grows larger.

C. Fairness-constrained Training

In this section, we demonstrate the efficacy of our approach to training fairness-constrained machine learning models. Following our algorithm in Section III, 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. We also demonstrate via our simulation study that there exist regimes in which our approach meets the goal of keeping disparity below the desired threshold whereas training on the small labeled subset alone does not.

1) 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, on both the L2 dataset and the COMPAS dataset. We use logistic regression as a proof-of-concept, but because our method builds upon the algorithm proposed in [14], it can be extended to any gradient-based machine learning method, including e.g. neural networks. Within the L2 dataset, we train these models on the data from Florida, 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 ($n = 150,000$ for L2 data, $n = 6,128$ for COMPAS data) 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 ($n = 1,500$) for the L2 data, and 10% of the total data for the COMPAS dataset, since it is much smaller (around $n = 613$). To enforce fairness constraints during training, we solve the empirical problem 3A and its symmetric analogue, which enforces negative covariance conditions and $\widehat{D}^L_\mu$ 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 $\widehat{D}^L_\mu$ during training. As noted in Section III, our method theoretically guarantees a near-optimal, near-feasible solution *on average* over $\theta^{(1)}...\theta^{(T)}$. However, following Wang et al. [21], for each of these sub-problems, we select the best iterate $\theta^{(t)}$ which satisfies the bound on $\widehat{D}^L_\mu$ 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 for our method as well as three comparisons, described below, over L2 data and COMPAS data, in Figure 3 and Figure 4 respectively. We note that the resulting disparities for the unconstrained model differ among the three fairness metrics. On DD and TPRD, the unconstrained model resulted in a 0.28-0.29 disparity, but it drops to 0.21 for FPRD. We adjusted our target fairness bounds accordingly. Further details about the experimental setup can be found in Appendix Section E-A. Our experimental design for our experiments on synthetic data differ, and we outline our setup and results in Section IV-D.

2) Comparisons: We compare our results for enforcing fairness constraints with probabilistic protected attribute labels to the following methods:

(a) 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-B.

(b) We compare with a recent method by Wang et al. [21] 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. [20]'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-C.

(c) We compare with a method for enforcing fairness with incomplete demographic labels introduced by Mozannar et al. [24], which essentially modifies Agarwal et al. [50]'s fair training approach to optimize accuracy on the entire available data, but 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-D.

In Appendix Section E-F, 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 $(0, 1)$.

3) Results: We first analyze our results on the L2 data. We display our results in Figure 3. 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 3), than using Wang et al. [21] (blue versus green bars in Figure 3), and than using Mozannar er

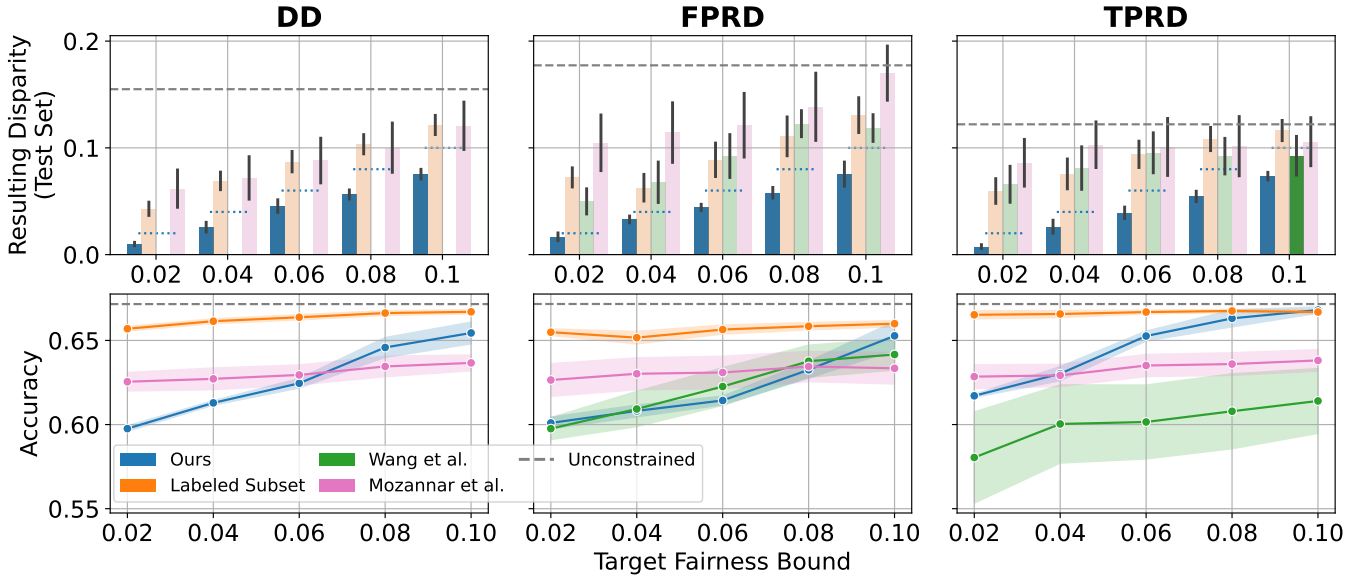


Fig. 3: (**Satisfying fairness constraints in L2 Data**) Mean and standard deviation of resulting disparity (top, y-axis) and accuracy (bottom, y-axis) on the test set after enforcing the target fairness bounds (x-axis) on our method (blue); only using the labeled subset with true labels (orange) and Wang et al. [21] (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.

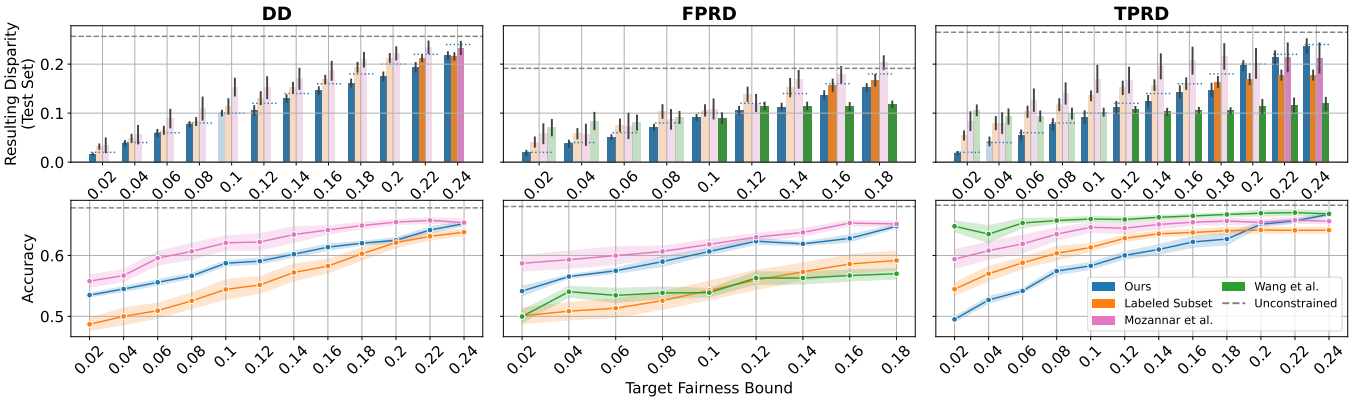


Fig. 4: (**Satisfying fairness constraints in COMPAS Data**) Mean and standard deviation of resulting disparity (top, y-axis) and accuracy (bottom, y-axis) on the test set after enforcing the target fairness bounds (x-axis) on our method (blue); Wang et al.'s method (green); Mozannar et al.'s method (red) and only using the labeled subset with true labels (orange). 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.

al. [24] (blue versus pink bars in Figure 3). Second, our method satisfies the target fairness bound on the test set more often than the other methods (12 out of 12 experiments, 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 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 more than Mozannar et al. [24] and Wang et al. [21], yet generally outperforms both with respect to accuracy. In the case of FPRD our method exhibits accuracy comparable to to Wang et al.

while consistently satisfying the target fairness constraint.

Next, we turn to our results on the COMPAS [45] dataset in Figure 4, which is set up identically to Figure 3, with disparity results on top and accuracy results on the bottom. We see that our method again is able to meet the desired disparity bound for all but two disparity bound values (i.e. for all but 2 out of 36 experiments) across the three different metrics, even for small target disparity values, while achieving accuracy comparable to the baseline methods. In the cases where our method's accuracy is lower than the comparison methods, it is the only method that consistently satisfies the target disparity constraint. While Mozannar et al. (red) has the highest accuracy across different target disparity values for DD and FPRD, it satisfies the target disparity bound in only 3 of the 36 experiments and particularly fails to satisfy the target disparity constraint for small disparity values. Wang et al. (green) has the highest accuracy for the TPRD experiments, however, it only satisfies the disparity constraint for FPRD and TPRD for disparity values greater than 0.1. Finally, the labeled subset baseline (orange) is only able to satisfy the target disparity constraint for large disparity values and typically has lower accuracy than the other comparison methods.

D. Simulation Study

We note that the utility of our method is often dependent upon the size of the subset of the data labeled with the protected attribute—if this subset is relatively large, then (depending on the complexity of the learning problem) it may be sufficient to train a model using the available labeled data. Symmetrically, if the labeled subset is exceedingly small, the enforcement of the covariance constraints during training may not generalize to the larger dataset. To characterize the regimes under which our method may be likely to perform well relative to others, we empirically study simulations that capture the essence of the situation. We study the utility of our method in comparison to only relying on the labeled subset to train a model along two axes: *1)* size of the labeled subset and *2)* data complexity, which we simulate by adjusting the number of features. While stylized, our simulation has the advantage that we can vary key features of the setting like the dimensionality and distribution of the data, the size of the labeled and unlabeled datasets, the complexity of the relationship between the features and the outcome, and so on. To be useful, however, we must be able to ensure that the key conditions of our method are met by the data-generating process.

To ensure this while also allowing for the tuneability and flexibility we require, we settle on a hierarchical model specified by parameterized components that are individually simple but can serve as building blocks; mapping out these relationships via the language of causal diagrams gives us intuition about the conditional covariance terms. See Appendix G, including Figure 15, for visualization and further discussion.

At a high-level, the model can be described as follows. Individuals have a set of "primary" features denoted which are drawn randomly from some distribution. The probability that they the individual is Black is a function of these primary features, and their status as Black or non-Black is simply a Bernoulli random variable with mean of said probability. There are then "secondary" features which each are functions of all the primary features. A score is generated as a function of these secondary features, and the outcome of interest is generated by thresholding this score and randomly perturbing it with small probability.

Using this high-level structure, we can generate a family of data-generating processes by choosing different functions representing the links between the features. In particular, we will use polynomials with randomly selected coefficients. This allows us to vary the model by increasing the number of features or degree of the polynomials without directly selecting all the constants involved. We provide further details, including specific functional forms and assumed distributions, in Section G.

Given the family of data-generating processes, we consider three different levels of complexity - cubic polynomials of 10, 20, or 50 features - and draw datasets of 5,000, 10,000, or 50,000 observations; of these, we vary the percentage with labels revealed to the learner ranging from 0.5 to 40%, depending on the size of the dataset. We then compare our method to simply training on a fair model on the true labels of the labeled subset. Figure 5 shows disparity for both methods across the scenarios. Overall, we find that there exists a regime, even in simple problems, where there is insufficient data for the labeled subset to effectively bound disparity to the desired threshold. We find that the more complex the data is, the larger this regime is—with the most complex setting in our simulations (50 features) suggesting that the labeled subset technique does not converge to the desired disparity bounds even when the size of the labeled subset is 10,000 samples, or 20% of the overall dataset.

V. Related Work

Kallus et al. [20] propose a method for measuring fairness violations in data with limited access to protected attribute labels. Their method involves finding the tightest possible set of true disparity given probabilistic protected attributes. An important difference between Kallus et al. and our method relates to their assumptions around the auxiliary dataset. The core difference is that Kallus et al. considers settings where the auxiliary and test sets are independent data sets while our method considers the case where the test set subsumes the auxiliary data. We explain this difference in further detail in Appendix D-B

With regards to bias mitigation, while there are many methods available for training models with bounded fairness violations [11, 39, 50], 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 [21], and *no* protected attribute labels [51], or a even a labeled subset of protected attribute labels, but without an auxiliary set to generate probabilistic protected attribute estimates [52]; very few works mirror our data access setting. One exception, from which we draw inspiration, is

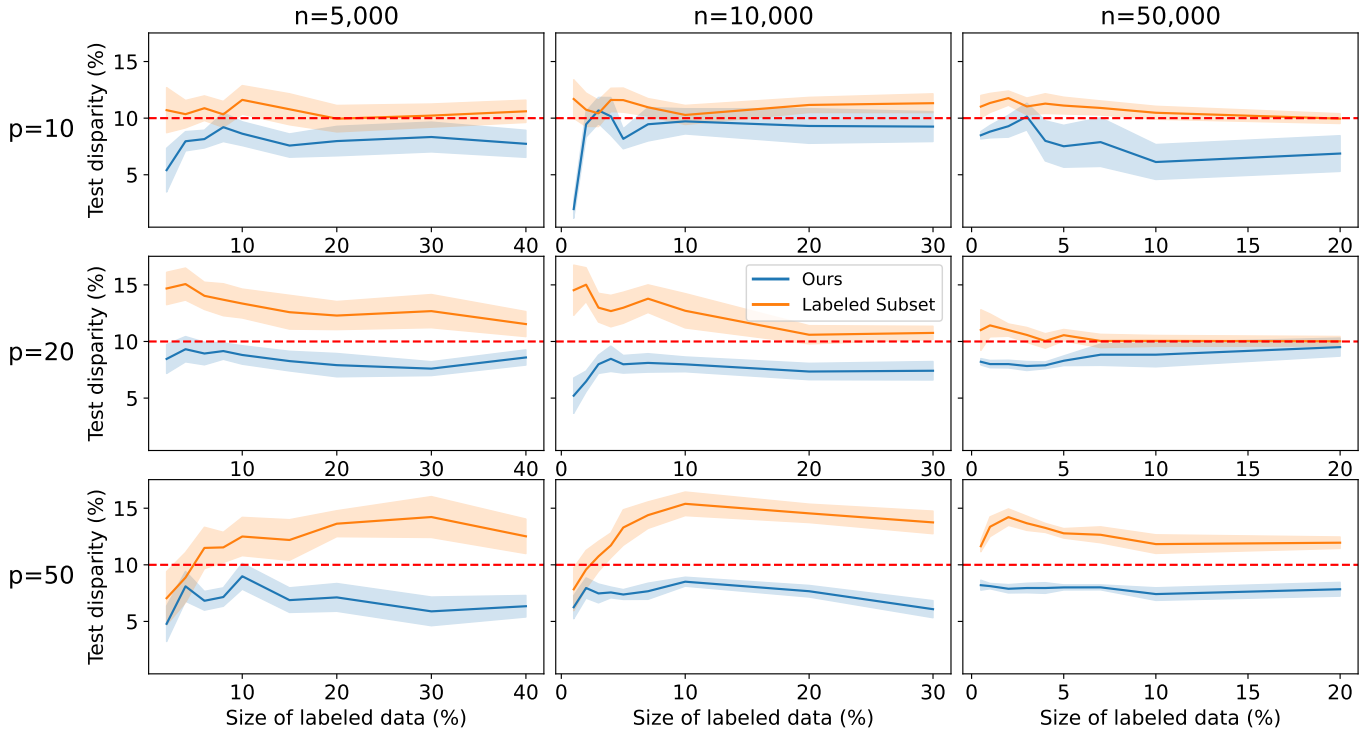


Fig. 5: (**Simulation varying size of labeled subset**) We present a three by three figure showing the test disparity of the our disparity reduction method when compared with relying on only the labeled subset to reduce disparity by directly enforcing a constraint on the protected attribute labels. The rows correspond to datasets of increasing sizes (number of features from 10 to 50), indicating problems of increasing complexity. The columns correspond to the size of the overall dataset, ranging from 5,000 to 50,000 samples. The x-axis shows the percentage of the total dataset is dedicated to the labeled subset, and the y-axis denotes the percentage disparity between the two groups calculated on the test set. The blue graphs correspond to our method, and the orange to the labeled subset method. The red dashed line is the desired disparity bound.

Elzayn et al. [19]; 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.

Another exception, which we compare to in Section IV-C, is that of Mozannar et al. [24]. While Mozannar et. al largely focus on the problem of training *private* fair models, thus employing very strong conditional independence assumptions on the protected attribute proxy which are infesible in our setting, the authors do propose an extension of their method to handle a the case of limited protected attributes without considering privacy, which mirrors our data access assumptions. This extension is essentially a re-purposing of Agarwal et al. [50] fair training approach, modified such that the model is trained with all available data, but the fairness bounds are only enforced during training on the small subset of training points with protected attribute labels. It is this extension that we compare to in which we compare to in Section IV-C, and find that our method often outperforms theirs on disparity reduction and performs comparably on accuracy.

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 [53, 54, 55]. 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. [21]. 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.

VI. Discussion

In this work, we introduce a technique for measuring and reducing fairness violations in a setting with limited access to protected attribute data by leveraging probabilistic proxies (e.g.

based on name and geolocation). These techniques may help private and public actors better measure algorithmic disparity and fulfill legal and moral obligations to ensure that algorithmic decision-making does not disparately impact disadvantaged or protected groups. However, the collection and use of protected attribute information is inherently sensitive and brings up privacy concerns. Additionally, building a probabilistic model to estimate protected attributes raises important ethical and practical questions as well, such as who has access to these models and what are the protocols for its responsible deployment. Moreover, the approach requires committing to a particular notion of groups to measure and mitigate fairness with respect to, an exercise which in itself can be fraught. Given the increasing stakes of algorithmic deployment as well as increasing regulatory and public pressure, we believe that the benefit of being able to more effectively measure and reduce unfairness in model predictions outweighs these risks, but practitioners applying our method must carefully consider these concerns in the wider context in which they work.

We note several avenues for future work. First, while our framework can be applied iteratively to handle multiple sensitive groups, generalizing our framework to account for them directly, and additionally to handle intersectional groups, would be preferable. Second, while binary classification is perhaps the most common task in machine learning, handling more general tasks, such as multi-label classification or regression, would extend the applicability of results. Finally, in the proposed method it is important that the probabilistic predictions are representative of the population of interest; in practice this means either assuming that the dataset from which probabilistic predictions are learned is drawn from the same population, or that reweighting techniques can be used to construct a representative sample. In the future, it would be useful to use techniques from sensitivity analysis to bound the impact of selection bias on measurement error and robust learning to train low-disparity models under worst-case selection bias.

REFERENCES

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

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

[3] 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.

[4] T. W. House, "Blueprint for an ai bill of rights: Making automated systems work for the american people," 2022.

[5] K. Hill, "Wrongfully accused by an algorithm," *The New York Times, June*, vol. 24, 2020.

[6] R. L. Austin, Jr., "Expanding our work on ads fairness," https://about.fb.com/news/2022/06/expanding-our-work-on-ads-fairness/, 2022.

[7] M. Isaac, "Meta agrees to alter ad technology in settlement with u.s." 2022. [Online]. Available: https://www.nytimes.com/2022/06/21/technology/meta-ad-targeting-settlement.html

[8] U.S. E.O. 14091, "Exec. order no. 14091 88 fed. reg. 10825, Further Advancing Racial Equity and Support for Underserved Communities Through The Federal Government," 2023.

[9] M. Andrus, E. Spitzer, J. Brown, and A. Xiang, "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.

[10] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft, Tech. Rep. MSR-TR-2020-32, May 2020. [Online]. Available: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[11] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *CoRR*, vol. abs/1810.01943, 2018. [Online]. Available: http://arxiv.bs/1810.0194

[12] M. N. Elliott, P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, and N. Lurie, "Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities," *Health Services and Outcomes Research Methodology*, vol. 9, pp. 69–83, 2009.

[13] K. Imai and K. Khanna, "Improving ecological inference by predicting individual ethnicity from voter registration records," *Political Analysis*, vol. 24, no. 2, pp. 263–272, 2016.

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

[15] C. F. P. Bureau, "Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment," 2014. [Online]. Available: https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

[16] K. Fiscella and A. M. Fremont, "Use of geocoding and surname analysis to estimate race and ethnicity," *Health services research*, vol. 41, no. 4p1, pp. 1482–1500, 2006.

[17] H. K. Koh, G. Graham, and S. A. Glied, "Reducing racial and ethnic disparities: the action plan from the department of health and human services," *Health affairs*, vol. 30, no. 10, pp. 1822–1829, 2011.

[18] R. L. Austin, Jr., "Race data measurement and meta's commitment to fair and inclusive products," https://about.fb.com/news/2021/11/inclusive-products-through-race-data-measurement/, 2022.

[19] H. Elzayn, E. Smith, T. Hertz, A. Ramesh, R. Fisher, D. Ho, and J. Goldin, "Measuring and mitigating racial disparities in tax audits," *Working Paper*, 2023.

[20] N. Kallus, X. Mao, and A. Zhou, "Assessing algorithmic fairness with unobserved protected class using data combination," *Management Science*, vol. 68, no. 3, pp. 1959–1981, 2022.

[21] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. Jordan, "Robust optimization for fairness with noisy protected groups," *Advances in neural information processing systems*, vol. 33, pp. 5190–5203, 2020.

[22] U.S. DOJ, "Statutes enforced by the voting section," 2023. [Online]. Available: https://www.justice.gov/crt/statutes-enforced-voting-section

[23] Equivant, "Practitioner's guide to COMPAS core," http://quivant.com/wp-content/uploas/Practitioners-Guide-to-COMPAS-Core-040419.pdf, 2019.

[24] H. Mozannar, M. Ohannessian, and N. Srebro, "Fair learning with private demographic data," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7066–7075.

[25] "Good questions, real answers: How does facebook use machine learning to deliver ads?" 2020. [Online]. Available: https://www.facebook.com/business/news/good-questions-real-answers-how-does-facebook-use-machine-learning-to-deliver-ads

[26] "United states of america, plaintiff, v. meta platforms, inc., f/k/a facebook, inc." 2022.

[27] A. S. Timmaraju, M. Mashayekhi, M. Chen, Q. Zeng, Q. Fettes, W. Cheung, Y. Xiao, M. R. Kannadasan, P. Tripathi, S. Gahagan *et al.*, "Towards fairness in personalized ads using impression variance aware reinforcement learning," *arXiv preprint arXiv:2306.03293*, 2023.

[28] "Facebookś Civil Rights Audit – Final Report," 2020.

[29] R. Alao, M. Bogen, J. Miao, I. Mironov, and J. Tannen, "How meta is working to assess fairness in relation to race in the u.s. across its products and systems," 2021.

[30] D. I. Werfel, "Werfel letter on audit selection," 2023.

[31] D. of Justice, "Justice department secures groundbreaking settlement agreement with meta platforms, formerly known as facebook, to resolve allegations of discriminatory advertising," 2022. [Online]. Available: https://www.justice.gov/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known

[32] U.S. Executive Order 13985, "Exec. order no. 13985 86 fed. reg. 7009, executive order on the safe, secure, and trustworthy development and use of artificial intelligence," 2021.

[33] F. T. Commission, "Trade regulation rule on commercial surveillance and data security," Proposed 08/22/2022.

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

[35] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 962–970.

[36] I. Žliobaitė, "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*, vol. 15, 2015.

[37] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

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

[39] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016.

[40] A. Narayanan, "Translation tutorial: 21 fairness definitions and their politics," in *Proc. conf. fairness accountability transp., new york, usa*, vol. 1170, 2018, p. 3.

[41] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018, pp. 1–7.

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

[43] L. Chamon and A. Ribeiro, "Probably approximately correct constrained learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 722–16 735, 2020.

[44] A. Cotter, H. Jiang, M. R. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan, "Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals." *J. Mach. Learn. Res.*, vol. 20, no. 172, pp. 1–59, 2019.

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

[46] "L2 voter and demographic dataset," 2000. [Online]. Available: https://purl.stanford.edu/pf141bs6760

[47] M. Barber and J. B. Holbein, "400 million voting records show profound racial and geographic disparities in voter turnout in the united states," *Plos one*, vol. 17, no. 6, p. e0268134, 2022.

[48] M. N. Elliott, A. Fremont, P. A. Morrison, P. Pantoja, and N. Lurie, "A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity," *Health services research*, vol. 43, no. 5p1, pp. 1722–1736, 2008.

[49] Z. Zhu, Y. Yao, J. Sun, H. Li, and Y. Liu, "Weak proxies are sufficient and preferable for fairness with missing sensitive attributes," in *Proceedings of Machine Learning Research*, 2023.

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

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

[52] S. Jung, S. Chun, and T. Moon, "Learning fair classifiers with partially annotated group labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 348–10 357.

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

[54] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 702–712.

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

[56] C. R. Shalizi, "The truth about linear regression," *Online Manuscript. http://www. stat. cmu. edu/˜ cshalizi/TALR*, 2015.

[57] I. Voicu, "Using first name information to improve race and ethnicity classification," *Statistics and Public Policy*, vol. 5, no. 1, pp. 1–13, 2018.

[58] Y. Zhang, "Assessing fair lending risks using race/ethnicity proxies," *Management Science*, vol. 64, no. 1, pp. 178–197, 2018.

[59] J. G. Matsusaka and F. Palda, "Voter turnout: How much can we explain?" *Public choice*, vol. 98, no. 3-4, pp. 431–446, 1999.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[61] D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford, "Large-scale methods for distributionally robust optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8847–8860, 2020.

[62] E. Diana, W. Gill, M. Kearns, K. Kenthapadi, A. Roth, and S. Sharifi-Malvajerdi, "Multiaccurate proxies for downstream fairness," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1207–1239.

APPENDIX A
MAIN PROOFS

A. Proof of Theorem 1

First, we demonstrate the following lemma:

Lemma 1. Suppose that $0 < b < 1$ almost surely and $\mathbb{E}|f(\hat{Y}, y)|\mathcal{E}|$ is finite. Under the assumption of independent and identically distributed data with $\mathcal{E}$ having strictly positive probability, the asymptotic limits D_μ^P and D_μ^L satisfy:

$$D_\mu^P = \frac{\text{Cov}\left[b, f(\hat{Y}, Y)|\mathcal{E}\right]}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])} \text{ and } D_\mu^L = \frac{\text{Cov}\left[b, f(\hat{Y}, Y)|\mathcal{E}\right]}{\text{Var}[b|\mathcal{E}]},$$

and thus

$$D_\mu^P = D_\mu^L \cdot \frac{\text{Var}[b|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])}.$$

Proof. We note that:

$$\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} b_i \overset{n_\mathcal{E} \to \infty}{\longrightarrow} \mathbb{E}[b|\mathcal{E}]$$

and

$$\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} b_i f(\hat{Y}, Y) \overset{n_\mathcal{E} \to \infty}{\longrightarrow} \mathbb{E}[b \cdot f(\hat{Y}, Y)|\mathcal{E}]$$

by the strong law of large numbers. Similarly,

$$\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} (1 - b_i) f(\hat{Y}, Y) \overset{n_\mathcal{E} \to \infty}{\longrightarrow} \mathbb{E}[(1 - b) \cdot f(\hat{Y}, Y)|\mathcal{E}]$$

$$\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} (1 - b_i) \overset{n_\mathcal{E} \to \infty}{\longrightarrow} \mathbb{E}[1 - b|\mathcal{E}]$$

Then diving numerators and denominators in the definition of the empirical estimator gives that:

$$\widehat{D}_\mu^P = \frac{\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} b_i f(\hat{Y}_i, Y_i)}{\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} b_i} - \frac{\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} (1 - b_i) f(\hat{Y}_i, Y_i)}{\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} (1 - b_i)}$$
$$\overset{n_\mathcal{E} \to \infty}{\longrightarrow} \frac{\mathbb{E}[b f(\hat{Y}, Y)|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}]} - \frac{\mathbb{E}[(1 - b) f(\hat{Y}, Y)|\mathcal{E}]}{\mathbb{E}[(1 - b)|\mathcal{E}]}$$

Combining terms and expanding out the algebra, the last term is:

$$\frac{\mathbb{E}[b f(\hat{Y}, Y)|\mathcal{E}] - \mathbb{E}[b|\mathcal{E}]\mathbb{E}[f(\hat{Y}, Y)|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])} = \frac{\text{Cov}\left[b, f(\hat{Y}, Y)|\mathcal{E}\right]}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])}.$$

On the other hand, the linear estimator converges asymptotically to

$$\widehat{D}_\mu^L \overset{n_\mathcal{E} \to \infty}{\longrightarrow} \frac{\text{Cov}\left[b, f(\hat{Y}, Y)|\mathcal{E}\right]}{\text{Var}[b|\mathcal{E}]}.$$

This result can be seen by conditioning on $\mathcal{E}$ and then making the standard arguments for the asymptotic convergence of the OLS estimator. Comparing forms of the limits gives the final result. □

Our key theorem follows as a corollary from the following proposition, (Proposition 1 in the main text):

Proposition. Suppose that b is a prediction of an individual's protected attribute (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_μ^P as the asymptotic limit of the probabilistic disparity estimator, $\widehat{D}_p$, and D_l as the asymptotic limit of the linear disparity estimator, $\widehat{D}_l$. Then:

1)

$$D_\mu^P = D_\mu - \frac{\mathbb{E}[\text{Cov}(f(\hat{Y}, Y), B|b, \mathcal{E})]}{\text{Var}(B|\mathcal{E})} \quad \textbf{(1.1)}$$

2)

$$D_\mu^L = D_\mu + \frac{\mathbb{E}[\text{Cov}(f(\hat{Y}, Y), b|B, \mathcal{E})]}{\text{Var}(b|\mathcal{E})} \quad \textbf{(1.2)}$$

We'll split things into separate proofs for **(1.1)** and **(1.2)**. We'll also first separately highlight that disparity is simply the dummy coefficient on race in a(n appropriately conditioned) regression model. This fact may be known by some readers in the context of regression analysis (especially without conditioning on a given event), but we provide proof of the general case.

Lemma 2. Let D_μ be the disparity with function f and event $\mathcal{E}$. Then D_μ can be written as:

$$D_\mu = \frac{\text{Cov}(f(\hat{Y}, Y), B|\mathcal{E})}{\text{Var}(B|\mathcal{E})}.$$

Proof. Note that by definition:

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

If the right hand side of the equation in the statement of the lemma can be written this way, we are done. But note that:

$$\frac{\text{Cov}(f(\hat{Y}, Y), B|\mathcal{E})}{\text{Var}(B|\mathcal{E})} = \frac{\mathbb{E}[f(\hat{Y}, Y)B|\mathcal{E}] - \mathbb{E}[f(\hat{Y}, Y)|\mathcal{E}]\mathbb{E}[B|\mathcal{E}]}{\mathbb{E}[B|\mathcal{E}](1 - \mathbb{E}[B|\mathcal{E}])}.$$

Now using the law of iterated expectations and simplifying:

$$\begin{aligned}
\mathbb{E}[f(\hat{Y}, Y)B|\mathcal{E}] &= \mathbb{E}[\mathbb{E}[f(\hat{Y}, Y)B|\mathcal{E}, B] \\
&= \mathbb{E}[f(\hat{Y}, Y)B|B = 1, \mathcal{E}] \Pr[B = 1|\mathcal{E}] \\
&+ \mathbb{E}[f(\hat{Y}, Y)B|B = 0, \mathcal{E}] \Pr[B = 0|\mathcal{E}] \\
&= \mathbb{E}[f(\hat{Y}, Y)|B = 1, \mathcal{E}] \Pr[B = 1|\mathcal{E}] \\
&+ \mathbb{E}[0] \Pr[B = 0|\mathcal{E}] \\
&= \mathbb{E}[f(\hat{Y}, Y)|B = 1, \mathcal{E}] \Pr[B = 1|\mathcal{E}]
\end{aligned}$$

Moreover, since B is a Bernoulli random variable, $\Pr[B = 1|\mathcal{E}] = \mathbb{E}[B|\mathcal{E}]$ and

$$\text{Var}(B|\mathcal{E}) = \mathbb{E}[B|\mathcal{E}](1 - \mathbb{E}[B|\mathcal{E}])$$

Combining these, we can write:

$$\frac{\mathbb{E}[f(\hat{Y}, Y)B|\mathcal{E}]\mathbb{E}[B|\mathcal{E}] - \mathbb{E}[f(\hat{Y}, Y)|\mathcal{E}]\mathbb{E}[B|\mathcal{E}]}{\mathbb{E}[B|\mathcal{E}](1 - \mathbb{E}[B|\mathcal{E}])}$$
$$= \frac{\mathbb{E}[f(\hat{Y}, Y)|B = 1, \mathcal{E}] - \mathbb{E}[f(\hat{Y}, Y)|\mathcal{E}]\mathbb{E}[B|\mathcal{E}]}{(1 - \mathbb{E}[B|\mathcal{E}])}$$

This can be expanded as:

$$\begin{aligned}
&\frac{\mathbb{E}[f(\hat{Y},Y)|B=1,\mathcal{E}]}{(1-\mathbb{E}[B|\mathcal{E}])} \\
&- \frac{\mathbb{E}[f(\hat{Y},Y)|B=1,\mathcal{E}]\Pr[B=1|\mathcal{E}]}{(1-\mathbb{E}[B|\mathcal{E}])} \\
&- \frac{\mathbb{E}[f(\hat{Y},Y)|B=0,\mathcal{E}]\Pr[B=0|\mathcal{E}]}{(1-\mathbb{E}[B|\mathcal{E}])} \\
&= \frac{\mathbb{E}[f(\hat{Y},Y)|B=1,\mathcal{E}](1-\Pr[B=1|\mathcal{E}])}{(1-\Pr[B=1|\mathcal{E}])} \\
&\frac{-\mathbb{E}[f(\hat{Y},Y)|B=0,\mathcal{E}](1-\Pr[B=1|E])}{(1-\Pr[B=1|\mathcal{E}])} \\
&= \mathbb{E}[f(\hat{Y},Y)|B=1,\mathcal{E}] - \mathbb{E}[f(\hat{Y},Y)|B=0,\mathcal{E}]
\end{aligned}$$

as desired. □

Note that the familiar interpretation of demographic disparity being the dummy coefficient falls out from this lemma by letting $\mathcal{E}$ be the event "always true" and $f(\hat{Y},Y)=Y$.

Now we can turn to proving (**1.1**). Recall first that, by assumption:

$$\begin{aligned}
b = \Pr[B=1|Z,\mathcal{E}] &= \mathbb{E}[\mathbb{1}[B=1]|Z,\mathcal{E}] \\
&\implies b = \mathbb{E}[B|Z,\mathcal{E}]\ \forall Z \\
&\implies \mathbb{E}[b|\mathcal{E}] = \mathbb{E}[\mathbb{E}[B|Z,\mathcal{E}]] = \mathbb{E}[B|\mathcal{E}]
\end{aligned}$$

by the law of iterated expectations. Moreover, if we define ϵ as $B-b$, then:

$$\mathbb{E}[\epsilon|Z,\mathcal{E}] = \mathbb{E}[B|Z,\mathcal{E}] - \mathbb{E}[b|Z,\mathcal{E}] = 0$$

Proof of (**1.1**). Note that by Lemmas 1 and 2:

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

Since $\mathbb{E}[b|\mathcal{E}] = \mathbb{E}[B|\mathcal{E}]$ and $\text{Var}[B|\mathcal{E}] = \mathbb{E}[B|\mathcal{E}](1-\mathbb{E}[B|\mathcal{E}]) = \mathbb{E}[b|\mathcal{E}](1-\mathbb{E}[b|\mathcal{E}])$, the denominators are the same and be collected as $\text{Var}(B|\mathcal{E})$. As for the numerators, we note that

$$\begin{aligned}
&\text{Cov}\left[f(\hat{Y},Y),B|\mathcal{E}\right] - \text{Cov}\left[f(\hat{Y},Y),b|\mathcal{E}\right] \\
&= \text{Cov}\left[f(\hat{Y},Y),B-b|\mathcal{E}\right]
\end{aligned}$$

by the distributive property of covariance. Recall that the law of total covariance allows us to break up the covariance of random variables into two parts when conditioned on a third. Applying this to $f(\hat{Y},Y)$ and $B-b$, with the conditioning variable being b, we have that:

$$\begin{aligned}
\text{Cov}\left[f(\hat{Y},Y),B-b|\mathcal{E}\right] &= \mathbb{E}\left[\text{Cov}\left(f(\hat{Y},Y),B-b\right)|\mathcal{E},b\right] \\
&+ \text{Cov}\left(\mathbb{E}[f(\hat{Y},Y)|\mathcal{E},b],\mathbb{E}[B-b|\mathcal{E},b]\right) \\
&= \mathbb{E}\left[\text{Cov}\left(f(\hat{Y},Y),B-b\right)|\mathcal{E},b\right] \\
&= \mathbb{E}\left[\text{Cov}\left(f(\hat{Y},Y),B\right)|\mathcal{E},b\right]
\end{aligned}$$

where the second equality follows because $b=\mathbb{E}[B|Z,\mathcal{E}] \implies \mathbb{E}[B|b,\mathcal{E}]=b$ and the third because b is trivially a constant given b. Combining these together, we have that:

$$\begin{aligned}
D_\mu - D_\mu^P &= \frac{\mathbb{E}\left[\text{Cov}\left(f(\hat{Y},Y),B\right)|\mathcal{E},b\right]}{\text{Var}[B|\mathcal{E}]} \\
&\implies D_\mu^P = D_\mu - \frac{\mathbb{E}\left[\text{Cov}\left(f(\hat{Y},Y),B\right)|\mathcal{E},b\right]}{\text{Var}[B|\mathcal{E}]},
\end{aligned}$$

as desired. □

Let's do (**1.2**).

Proof of (**1.2**). First, consider the linear projection of $f(\hat{Y},Y)$ onto B given that $\mathcal{E}$ occurs. We can write this as:

$$f(\hat{Y},Y) = \alpha + \gamma\cdot B + \nu,$$

where it is understood that the equation holds given $\mathcal{E}$. Now, by the definition of linear projection,

$$\gamma = \frac{\text{Cov}(f(\hat{Y},Y),B|\mathcal{E})}{\text{Var}(B|\mathcal{E})} = D_\mu$$

where the last equality follows by Lemma 2, and by the definition of linear projection, $\text{Cov}(B,\nu|\mathcal{E})=0$.

Now, consider the linear projection of $f(\hat{Y},Y)$ onto b given $\mathcal{E}$. Again we can write the equation:

$$f(\hat{Y},Y) = \alpha' + \beta b + \eta$$

and similarly

$$\beta = \frac{\text{Cov}(f(\hat{Y},Y),b|\mathcal{E})}{\text{Var}(b|\mathcal{E})} = D_\mu^L$$

and $\text{Cov}(b,\eta|\mathcal{E})=0$.

Now, by applying the Law of Total Covariance to the equation above, we have:

$$\begin{aligned}
\beta\text{Var}(b|\mathcal{E}) &= \text{Cov}(f(\hat{Y},Y),b|\mathcal{E}) \\
&= \mathbb{E}[\text{Cov}(f(\hat{Y},Y),b|\mathcal{E},B] \\
&+ \text{Cov}(\mathbb{E}[f(\hat{Y},Y)|\mathcal{E},B],\mathbb{E}[b|\mathcal{E},B]).
\end{aligned}$$

We'll focus for now on the latter term. Note that by replacing $f(\hat{Y},Y)$ by $\alpha+\gamma B+\nu$, we can obtain:

$$\begin{aligned}
\text{Cov}(\mathbb{E}[f(\hat{Y},Y)|B,\mathcal{E}],\mathbb{E}[b|B,\mathcal{E}]) &= \text{Cov}(\gamma B + \mathbb{E}[\nu|B], \\
&\qquad B - \mathbb{E}[\epsilon|B]\big|\mathcal{E})
\end{aligned}$$

where we've moved out the event $\mathcal{E}$ and used the fact that α is a constant and B is a constant conditional on B to remove them from the inner expectations. We can expand as

$$\text{Cov}\left(\gamma B + \mathbb{E}[\nu|B,\mathcal{E}], B - \mathbb{E}[\epsilon|B]\big|\mathcal{E}\right).$$

We can further expand this covariance term to be

$$\begin{aligned}
&= \gamma\text{Var}(B|\mathcal{E}) - \gamma\,\text{Cov}(B,\mathbb{E}(\epsilon|B)\big|\mathcal{E}) \\
&+ \text{Cov}(\mathbb{E}(\nu|B),B\big|\mathcal{E}) - \text{Cov}(\mathbb{E}(\nu|B),\mathbb{E}(\epsilon|B)\big|\mathcal{E}) \\
&= \gamma\text{Var}(B|\mathcal{E}) - \gamma\,\text{Cov}(B,\mathbb{E}(\epsilon|B)\big|\mathcal{E}),
\end{aligned}$$

where the last equality is due to the fact that B is binary so the covariance between B and ν equals zero.

Next we show that the term $\mathrm{Cov}(B, \mathbb{E}(\epsilon|B)\big|\mathcal{E})$ can be written in terms of b and ϵ,

$$\begin{aligned}\mathrm{Cov}(B, \mathbb{E}(\epsilon|B)\big|\mathcal{E}) &= \mathbb{E}[B\mathbb{E}[\epsilon|B]] - \mathbb{E}[B]\mathbb{E}[\mathbb{E}[\epsilon|B]] \\ &= \mathbb{E}[\mathbb{E}[B\epsilon|B]\big|\mathcal{E}] - \mathbb{E}[B|\mathcal{E}]\mathbb{E}[\mathbb{E}[\epsilon|B]|\mathcal{E}] \\ &= \mathbb{E}[B\epsilon|\mathcal{E}] - \mathbb{E}[B|\mathcal{E}]\mathbb{E}[\epsilon|\mathcal{E}] \\ &= \mathrm{Cov}(B, \epsilon\big|\mathcal{E}) \\ &= \mathrm{Cov}(b+\epsilon, \epsilon\big|\mathcal{E}) \\ &= \mathrm{Cov}(b, \epsilon\big|\mathcal{E}) + \mathrm{Var}(\epsilon|\mathcal{E}).\end{aligned}$$

Plugging these results back into the original equation and using the fact that $B = b + \epsilon$, we have

$$\begin{aligned}\beta\mathrm{Var}(b|\mathcal{E}) &= \mathbb{E}[\mathrm{Cov}(f(\hat{Y}, Y), b|\mathcal{E}, B] \\ &\quad + \gamma\mathrm{Var}(B|\mathcal{E}) - \gamma\mathrm{Var}(\epsilon|\mathcal{E}) - \gamma\,\mathrm{Cov}(b, \epsilon\big|\mathcal{E}) \\ &= \gamma[\mathrm{Var}(b|\mathcal{E}) + \mathrm{Cov}(b, \epsilon\big|\mathcal{E})] \\ &\quad + \mathbb{E}[\mathrm{Cov}(f(\hat{Y}, Y), b|\mathcal{E}, B] \\ &= \gamma\mathrm{Var}(b|\mathcal{E}) + \mathbb{E}[\mathrm{Cov}(f(\hat{Y}, Y), b|\mathcal{E}, B],\end{aligned}$$

where the last equality is due to the fact that $\mathbb{E}[\epsilon|Z, \mathcal{E}] = 0$. □

B. Proof of Proposition 2

Proof. For a *fixed* $\tilde{\theta}$, we can apply Theorem 1 to write that:

$$D_\mu^p(h_{\tilde{\theta}}) = D_\mu(h_{\tilde{\theta}}) - \frac{\mathbb{E}[\mathrm{Cov}(f(h_{\tilde{\theta}}, Y), B|b, \mathcal{E}]}{\mathrm{Var}[B|\mathcal{E}]},$$

where the expectation in the numerator is over the distribution of the data. Now, if $\tilde{\theta}$ is drawn from a distribution $\boldsymbol{\theta}$ (in particular, $\boldsymbol{\theta}$ corresponding to θ_t with t being drawn from $1...T$) that is independent of the data, we can treat the quantities as random variables drawn from a two step data-generating process. In our setting (as in classical, but not all, learning settings), the distribution of future data is assumed not to depend on our selected model. Then by the linearity of expectations, we have that

$$\begin{aligned}&\mathbb{E}_{\tilde{\theta}\sim\boldsymbol{\theta}}\left[D_\mu^p(h_{\tilde{\theta}})\right] - \mathbb{E}_{\tilde{\theta}\sim\boldsymbol{\theta}}\left[D_\mu(h_{\tilde{\theta}})\right] \\ &= \mathbb{E}_{\tilde{\theta}\sim\boldsymbol{\theta}}\left[\frac{\mathbb{E}[\mathrm{Cov}(f(h_{\tilde{\theta}}, Y), B|b, \mathcal{E}]}{\mathrm{Var}[B|\mathcal{E}]}\right].\end{aligned}$$

A similar statement can be made for the relationship between $\mathbb{E}_{\tilde{\theta}\sim\boldsymbol{\theta}_T}\left[D_\mu^p(h_{\tilde{\theta}})\right]$ and $\mathbb{E}_{\tilde{\theta}\sim\boldsymbol{\theta}_T}\left[D_\mu(h_{\tilde{\theta}})\right]$. □

C. Standard Errors

Here, we discuss the calculation of standard errors; these arguments are more general, but substantially similar, version of those made in [19]. As shown in the proof of Theorem 1, $\widehat{D}_\mu^l$ and $\widehat{D}_\mu^p$ converge to their asymptotic limits, D_μ^l and D_μ^p, respectively; however, given that we observe only a finite sample, our estimates $\widehat{D}_\mu^l$ and $\widehat{D}^p$ are subject to uncertainty whose magnitude depends on the sample size of the data.

Since the $\widehat{D}_\mu^l$ is simply the linear regression coefficient, its distribution is well-studied and well known. In particular, under the classical ordinary least squares (OLS) assumptions of normally distributed error, $\widehat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{n s_b^2}\right)$ where s_b^2 is the sample variance of b; under mild technical conditions, central limit theorems can be invoked to show that as the size of data increases, $\widehat{\beta}$ follows a distribution that is increasingly well-approximated by said normal distribution [56]. Note that, since as shown in Lemma 1

$$D_\mu^L = \frac{\mathrm{Cov}(f(\hat{Y}, Y), b|\mathcal{E})}{\mathrm{Var}[b|\mathcal{E}}$$

and

$$D_\mu^P = \frac{\mathrm{Cov}(f(\hat{Y}, Y), b|\mathcal{E})}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])},$$

it follows that

$$D_\mu^P = D_\mu^L \cdot \frac{\mathrm{Var}[b|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|E])};$$

analogously, by expanding the definitions of the sample estimators, we can easily see that:

$$\widehat{D}_\mu^P = \widehat{D}_\mu^L = \frac{\frac{1}{n_\mathcal{E}}\sum_{i\in\mathcal{E}}(b_i - \bar{b}^\mathcal{E})^2}{\bar{b}^\mathcal{E}(1 - \bar{b}^\mathcal{E})}.$$

Then by Slutsky's theorem, we can state that:

$$\widehat{D}_\mu^P \overset{n\to\infty}{\longrightarrow} \widehat{D}_\mu^L \frac{\mathrm{Var}[b|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|E])}.$$

As a consequence, the distribution of $\widehat{D}_\mu^P$ is a scaled version of the distribution of $\widehat{D}_\mu^L$, and in particular

$$\frac{\widehat{D}_\mu^P - D_\mu^P}{\mathrm{Var}\widehat{D}_\mu^L \sqrt{\frac{\mathrm{Var}[b|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}](1-\mathbb{E}[b|E])}}} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0, 1).$$

Thus, in practice, we can estimate the variance of $\widehat{D}_\mu^L$ as if it were the usual OLS estimator and then estimate $\mathrm{Var}[b|\mathcal{E}]$ and $\mathbb{E}[b|\mathcal{E}]$ to scale it appropriately.

D. Obtaining the probabilistic prediction

1) BIFSG: Recall that conceptually, b functions as a probabilistic confidence score we have that an individual has $B = 1$. A perfectly calibrated b will thus have $\mathbb{E}[B|b] = b$, and our main theorems assume that we have access to this. In practice, however, b must be estimated; in this work, we focus on the commonly used [16, 20, 57, 58] Bayesian Imputations with First Names, Surnames, and Geography (BIFSG). In BIFSG, we make the naive conditional independence assumption that the proxy features are independent conditional on the protected characteristic. In the case of BIFSG, this amounts to assume that:

$$\Pr[F, S, G|B] = \Pr[F|B]\Pr[S|B]\Pr[G|B],$$

where the random variable F is first name, S is surname, and G is geography . By applying Bayes' rules to this assumption,

we can obtain that:

$$\Pr[B|F,S,G] = \frac{\Pr[F,S,G|B]}{\Pr[F,S,G]} = \frac{\Pr[F|B]\Pr[S|B]\Pr[G|B]}{\Pr[F,S,G]}.$$

The right-hand side of this equation is fairly easy to estimate because it requires knowing only marginals rather than joint distributions (the denominator can be normalized away by noting that we must have that $\Pr[B = 1|F,S,G]$ and $\Pr[B = 0|F,S,G]$ must sum to 1), and these marginals are often obtainable in the form of publicly available datasets. Note that, BIFSG can be written in multiple forms by applying Bayes' rule again to the individual factors (e.g. replacing $\Pr[F|B]$ with $\Pr[B|F]\Pr[F]/\Pr[B]$, which may be convenient depending on the form of auxiliary data available.

For our setting, we leverage the census and home mortgage disclosure act (HMDA) data, as mentioned, to estimate b from publicly available data. We provide quantitative details on our estimates in Appendix C. We note also that since b is continuous, we will discretize into equally sized bins whenever we need to compute quantities conditional on b.

2) Impact of Miscalibration: Throughout the theoretical work, we have assumed that we have $b = \Pr[B = 1|Z]$ - i.e. that b is *perfectly calibrated*. In reality, this is a quantity that is estimated, and will thus contain some uncertainty, including bias due to the fact that the dataset which it is estimated on (e.g. the census for the U.S. as a whole) may not be fully representative of the relevant distribution (i.e. the distribution of individuals to whom the model will be applied, which may be a particular subset). This could result in *miscalibration*; when this happens, it could be that applying our method with our miscalibrated b results in failing to bound disparity (both in measuring alone, and in training).

Ultimately, miscalibration is only a real problem insofar as it causes the method to fail. For small amounts of miscalibration, the method tends to succeed anyway – e.g. in our setting, we do observe that our estimates are not perfectly calibrated, but we still achieve good results. For larger, or unknown, miscalibration, there are two paths that can be taken. The first is to conduct a "recalibration" exercise, and obtain a modified b that more closely matches the distribution of interest; this can be as simple as fitting a linear regression of B on b in the labeled dataset and replace b with the predictions of this regression. Alternatively, given an assumed bound on the magnitude of the miscalibration, Theorem 1 can be extended to incorporate its effect. In practice, recalibration is more straightforward to do empirically, but the theoretical method can also be used for sensitivity analysis; see [19] for their discussion of the recalibration approach as well as the effect on their special-case bounds.

Note also that, in settings where $\mathcal{E}$ is affected by the modeling choice h - i.e. when the fairness metric involves conditioning on model predictions, as in the case of positive predictive value (PPV) - it may be the case that a perfect or well-calibrated b for one model may be poorly-calibrated for another. That is, it may be that among observations, we find that that our estimate $|b(Z) - \Pr[B|Z,\mathcal{E}(h_\theta)]|$ is small while our estimate of $|b(Z) - \Pr[B|Z,\mathcal{E}(h_{\theta'})|$ is large. In this case, we can introduce a recalibration step in-between iterations, although this deviates from the theoretical assumptions that ensure convergence. Note that a sufficiently expressive model over a sufficiently powerful set of proxy features should be able obtain good calibration overall events $\mathcal{E}$; this suggests that another path forward in such a setting may be in investing in alternative, more powerful (e.g. machine-learned) models of b.

APPENDIX B
MATHEMATICAL FORMULATION OF FAIR LEARNING PROBLEM

A. Theoretical Problem

We begin by discussing the *theoretical* problems - i.e. abstracting away from the sample of data and considering the problems we are trying to solve.

1) One-sided bound: We first consider the case of imposing a one-sided bound on disparity, i.e. requiring that $D_\mu \leq \alpha$ but allowing $D_\mu < -\alpha$; certainly this will not be desirable in all situations, but we can use it as a building block for the two-sided bound as well.

We begin by formalizing the ideal problem - that is, the problem we would solve if we had access to ground truth protected class. This is simply to minimize the expected risk subject to the constraint that - whichever disparity metric we have adopted - disparity is not "too high". This is the:

Problem 3 (Ideal Problem)**.** Given individual features X, labels Y, a loss function L, a model class $\mathcal{H}$, a disparity metric μ, and a desired bound on disparity α, find an h to:

$$\min_{h\in\mathcal{H}} \mathbb{E}[L(h(X),Y)] \text{ s.t. } D_\mu(h) \leq \alpha,$$

where $D_\mu(h)$ is the μ-disparity obtained by h.

The ideal problem is not something we can solve because we cannot directly calculate D_μ over the dataset, since it requires the ground truth protected class label B. But the Theorem 1 suggests an alternative and feasible approach: using the linear estimate of disparity as a proxy bound. That is, if the linear estimator is an upper bound on the disparity, and the linear estimator is below α, then disparity is below α too.

Formally, we would solve following problem:

Problem 4 (Bounded Problem Direct)**.** Given individual features X, labels Y, a loss function L, a model class $\mathcal{H}$, a disparity metric μ, a desired on disparity α, and a predicted protected attribute proxy b, find an h to:

$$\min_{h\in\mathcal{H}} \mathbb{E}[L(h(X),Y)] \text{ s.t. } D^L_\mu \leq \alpha \text{ and } D_\mu \leq D^L_\mu$$

Notice that any feasible solution to Problem 4 must satisfy the constraints of Problem 3, i.e. we must have that $D_\mu(h) \leq \alpha$. The gap between the performance of these two solutions can be regarded as a "price of uncertainty"; it captures the loss we incur

by being forced to use our proxy to bound disparity implicitly rather than being able to bound it directly. We explore this price by comparing to an "oracle" which can observe the ground truth on the full dataset and perform constrained statistical learning.

As in Problem 2, we cannot directly observe D_μ, so the second constraint is not one that we can directly attempt to satisfy. But we know that it holds exactly in the conditions under which Theorem 1 applies. Therefore, we can replace that constraint with the covariance conditions:

Problem 5 (Fair Problem - Indirect)**.** Given individual features X, labels Y, a loss function L, a model class $\mathcal{H}$, a disparity metric μ (with associated event $\mathcal{E}$ and function $f(h(X), Y)$), a desired maximum disparity α, and a predicted proxy b, find an h to:

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

And indeed, these problems are equivalent:

Proposition 3. Problems 5 and 4 are equivalent.

Proof. Theorem 1 directly says that $D_\mu^L \geq D_\mu \iff \mathbb{E}[\mathrm{Cov}(f(h(X),Y), b|B,\mathcal{E})] \geq 0$. Hence if h satisfies the constraints of Problem 5 iff it satisfies those of Problem 4. Since the objectives are also the same, the problems are equivalent. □

As written, Problem 5 is still using the population distributions; we will discuss its empirical analogue below.

2) Two-sided bound: The two-sided bound requires that $|D_\mu| \leq \alpha$; this may be more common in practice. Again, we begin by considering the ideal problem:

Problem 6 (Ideal Symmetric Problem)**.** Given individual features X, labels Y, a loss function L, a model class $\mathcal{H}$, a disparity metric μ, and a desired bound on disparity α, find an h to:

$$\min_{h\in\mathcal{H}} \mathbb{E}[L(h(X),Y)] \text{ s.t. } |D_\mu(h)| \leq \alpha,$$

where $D_\mu(h)$ is the μ-disparity obtained by h.

As with Problem 4, we cannot directly bound disparity, since we do not have it, but we do have the disparity estimator. This leads to the following problem:

Problem 7 (Symmetric Problem Direct)**.** Given individual features X, labels Y, a loss function L, a model class $\mathcal{H}$, a disparity metric μ, a desired on disparity α, and a predicted protected attribute proxy b, find an h to:

$$\min_{h\in\mathcal{H}} \mathbb{E}[L(h(X),Y)] \text{ s.t. } |D_\mu^L| \leq |\alpha| \\ \text{and } |D_\mu| \leq |D_\mu^L|$$

Unfortunately, we don't have any theory about putting an absolute value bound on disparity, and indeed, because the weighted and linear disparity estimators are positive scalar multiples of one another, we cannot hope to use one as a positive upper bound and the other as a negative lower bound. But notice that if we were to find the best solution when $D_\mu^L \in [0,\alpha]$, and the best solution when $D_\mu^L \in [-\alpha, 0]$, then we would cover the same range as $[-\alpha,\alpha]$.

One attempt to apply this principle would be to solve the following two subproblems:

Problem 6.A.

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

Problem 6.B.

$$\min_{h\in\mathcal{H}} \mathbb{E}[L(h(X),Y)] \text{ s.t. } -\alpha \leq D_\mu^L \\ \text{and } \mathbb{E}[\mathrm{Cov}(f(h(X),Y), b|B,\mathcal{E})] \geq 0$$

And take:

$$h_5^* = \text{argmin}_{h_{6a}^*, h_{6b}^*} \mathbb{E}[L(h(X),Y)].$$

But this does not even guarantee a *feasible*, let alone optimal, solution to Problem 7. To see this, note that there is nothing prevent h_{6a}^* to be not simply $\leq \alpha$, but in fact $< -\alpha$, and vice versa. In particular, what went wrong is that we did not find the two best solutions over $[-\alpha, 0]$ and $[0,\alpha]$, but rather the two best over $[-\infty,\alpha]$ and $[-\alpha,\infty]$, which is no constraint at all.

To get around, this, though, we can solve the following two problems instead:

Problem 7.A.

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

Problem 7.B.

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

Why are these different? Notice that imposing both covariance constraints in 1.A enforces that $D_\mu^p \leq D_\mu \leq D_\mu^L$; since $D_\mu^p = D_\mu^L \frac{\mathrm{Var}b}{\mathcal{E}[b](1-\mathcal{E}[b])}$ – i.e. D_μ^p is always an attenuated version of D_μ^l – this can *only* be the case if all three terms are nonnegative. Similarly, 1.B enforces that $\mathcal{D}_\mu^p \geq D_\mu \geq D_\mu^l$; this similarly ensures that all three terms are nonpositive. Since these terms also include the bound on the linear estimator, they thus ensure that if we take:

$$h \in \text{argmin}_{h_{7a}^*, h_{7b}^*} \mathbb{E}[L(h(X),Y)],$$

we will indeed obtain a feasible solution to Problem 7. As in Problem 5, there may again be a suboptimality gap since we have effectively imposed more constraints to the original problem.

B. Solving the Empirical Problems

In this section, we use recent results in constrained statistical learning to formulate and motivate empirical problems that we can solve which obtain approximately feasible and performant solutions to the problems above. We summarize here the conceptual basis at a high level, providing a discussion of the rationale behind Theorem 2 in the main text, drawing heavily on [14], and refer interested readers to said work as well as [43] for a fuller and more detailed discussion of the constrained statistical learning relevant to our setting and [44] for more general discussion of non-convex optimization via primal-dual games.

1) Relating our Formulation: We begin by describing the relationship between our problem of interest and that considered in [14]. The (parameterized version of the) problem in [14] is the following:

Problem 8 (Parameterized Constrained Statistical Learning (P-CSL) from [14])**.**

$$P^* = \min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_0} \left[\ell_0(f_\theta(x), y)\right]$$
$$\text{s.t. } \mathbb{E}_{(x,y)\sim\mathcal{D}_i} \left[\ell_i(f_\theta(x,y)\right] \leq c_i,\ i = 1...m$$

That is, they aim to minimize some expected loss subject to some constrained on other expected losses, with loss functions that may vary and be over different distributions. Our problem, i.e. Problem 5 can be seen as a special case of this, though our framing is different. To see the correspondence, consider applying the following to Problem 8:

1) Take $\mathcal{D}_i$ to be the restriction of $\mathcal{D}$ to $\mathcal{E}$
2) Take ℓ_0 to be the loss function of interest, e.g. $\mathbf{1}[h \neq y]$ for accuracy
3) Take $\ell_1 = f(h(X), Y)$ and c_1 as α
4) Take $\ell_2 = f(h(X), Y) \cdot B - \overline{f(h(X), Y)}^B \bar{b}^B$ and $c_2 = 0$
5) Take $\ell_3 = f(h(X), Y) \cdot b - \overline{f(h(X), Y)}^b \bar{B}^b$ and $c_3 = 0$

Then we arrive at Problem 5.

2) Moving to the empirical problem: The problems described above relate to the population distribution, but we only have samples from this distribution. This is, of course, the standard feature of machine learning situations; the natural strategy in such a setting is to simply solve the empirical analogue - i.e. to replace expectations over a distribution with a sample average over the realized data. Instantiating this and focusing on Problem 1.A (since the other problems can be solved analogously and/or using it as a subproblem) we could write the following empirical problem.

Problem 9.

$$\min_{h\in\mathcal{H}} \frac{1}{n} \sum_{i \in n_{\mathscr{D}}} L(h(X_i), Y_i) \text{ s.t. } \widehat{D}_\mu^L \leq \alpha$$
$$0 \leq -\frac{1}{n_{\mathscr{D}_L}} \sum_{i \in \mathscr{D}_L} \left[\left(f(h(X_i), Y_i) - \overline{f(h(X_i)), Y_i}^{B_i}\right)(b_i - \bar{b}^{B_i})\right]$$
$$0 \leq -\frac{1}{n_{\mathscr{D}_L}} \sum_{i \in \mathscr{D}_L} \left[\left(f(h(X_i), Y_i) - \overline{f(h(X_i)), Y_i}^{b_i}\right)(B_i - \bar{B}^{b_i})\right]$$

.

Problem 9 is not, in general, a convex optimization problem; if it were, the standard machinery and solutions of convex optimization, i.e. formulating the dual problem and recovering from it a primal solution via strong duality, could be applied. However, as shown in [14], under some conditions, there exists a solution to the empirical dual problem that obtains nearly the same objective value as the primal population problem. In other words, rather than applying strong duality as a consequence of problem convexity, [14] directly prove a relationship between the primal and the dual under some conditions. These conditions are that:

1) The losses $\ell_i(\cdot, y)$ are Lipschitz continuous for all y
2) Existence of a family of funtions $\zeta_i(N, \delta) \geq 0$ that decreases monontically in N and bounds the difference between the sample average and population expectatoin for each loss function
3) There is a $\nu \geq 0$ so that for each Φ in the closed convex hull of $\mathcal{H}$, there is a θ such that
4) The problem is feasible

We briefly discussing these conditions. For 1), we note that Lipschitz continuity requires existence of scalar such that $|f(x) - f(x')| \leq M|x - y|$, which will be true for bounded features when using sample averages. 2) simply requires that we are in a situation where more data is better, and is implied by the stronger condition we assume of $\mathcal{H}$ being of finite VC-dimension. 3) asks that our hypothesis class is rich enough to cover the space finely enough (how fine will determine the quality of the solution), which is met for reasonable model classes. 4), is simply a technical requirement ensuring that there exists at least some solution, is analogous to Slater's criterion in numerical optimization.

Thus, we can leverage the described guarantees to assert that solving the empirical dual would Yet this initial result, while positive, is one of existence; to actually find a solution requires a solution. To do so, one can construct an empirical Lagrangian from the constrained empirical problem, and this can be solved by running a game between primal player, who selects a model to minimize loss, and a dual player, who selects dual parameters in an attempt to maximize it. If we construct this empirical dual in our settings, it is as in Equation 3; Algorithm 1 provides a primal-dual learner that instantiates this idea of a game.

C. Theoretical Guarantees

If either all of the losses are convex, or:

6) The outcome of interest Y takes values in a finite set
7) The conditional random variables $X|Y$ is are non-atomic
8) The closed convex hull of $\mathcal{H}$ is *decomposable*

Then the primal-dual algorithm 1 performs well. In the classification setting, which we focus on, Item 5) is trivially true. Item 6) asks that it not be the case that any of the distribution over which losses are measured induce an outcomes induce an atomic distribution; this mild regularity condition prevents pathological cases that would be impossible to satisfy. For 7) *Decomposability* is a technical condition stating that for a

Algorithm 1: Primal-dual algorithm for probabilistic fairness

Input : Labeled subset $\mathscr{D}_L$, unlabeled data $\mathscr{D}_U$, θ-oracle, number of iterations $T \in \mathbb{N}$, step size $\eta > 0$

Define : $h_{\theta^{(t)}}$ as the model parameterized by $\theta^{(t)}$

Initialize : $\mu_L^{(1)} \leftarrow 0$; $\mu_{b|B}^{(1)} \leftarrow 0$; $\mu_{B|b}^{(1)} \leftarrow 0$

for $t = 1 \dots T$ **do**

  $\theta^{(t)} \leftarrow \text{argmin}_\theta \widehat{\mathcal{L}}(\theta, \mu^{(t)})$

  $\mu_{b|B}^{(t+1)} \leftarrow \mu_{b|B}^{(t)} + \eta \widehat{C}_{f,b|B}(h_{\theta^{(t)}})$; $\mu_{B|b}^{(t+1)} \leftarrow \mu_{B|b}^{(t)} + \eta \widehat{C}_{f,B|b}(h_{\theta^{(t)}})$

  $\mu_L^{(t+1)} \leftarrow \mu_L^{(t)} + \eta \left(\widehat{D}_L(h_{\theta^{(t)}} - \alpha \right)$

end

return $< \theta^{(1)}, \dots, \theta^{(T)} >$

given function space, it closed in a particular sense: for any two function Φ, Φ' and any measurable set χ, the function that is Φ on χ and Φ' on its complement is also in the function space; many machine learning methods can be viewed from a functional analysis as optimizing over decomposable function space.

As we have shown that our problem can be written as a case of the CSL problem, and Algorithm 1 is a specialization of the primal-dual learner analyzed in [14], Theorem 3 in the same applies, again with appropriate translation. In particular, the promise is that when an iterate is drawn uniformly at random, the expected losses (over the distribution of the data and this draw) for the constraints are bounded by the constraint limit c_i plus the family of functions at the datasize mention in Assumption 2, plus $2C/(\eta T)$, where T is number of iterations, η is the learning rate, and C is a constant; at the same time, the expected loss (again over both the data and drawing the iterate) is bounded by the value of primal plus several problem-specific constants that capture the difficult of the learning problem and meeting the constraints, as well as said monotonically decreasing function of the data capturing the rate of convergence. Our Theorem 2 can be obtained by applying a standard result from statistical learning theory and collecting/re-arrange/hide problem-specific constants.

In this section, we discuss our approach to learning a fair model using the probabilistic proxies and a small subset of labeled data. To do so, we leverage recent results in constrained statistical learning.

D. Handling Imperfect Calibration

In general, it may be that we do not have access to $b = \Pr[B = 1|Z = z]$, but instead have access to some *imperfectly calibrated* $\hat{b}$. In this case, we can write $\hat{b} = b + \varepsilon$, where ε by definition is $\hat{b} - b$. We could apply $\hat{D}_\mu^P$ and $\hat{D}_\mu^L$ using $\hat{b}$ instead, but Theorem 1 assumes access to b, and so does not directly apply. To overcome this, we can obtain a *recalibrated* b^*. As a first step, we know that for a general b, the linear and probabilistic estimators converge to:

$$D_\mu^L \longrightarrow D_\mu(1 + \frac{\text{Cov}(b, \varepsilon|\mathcal{E})}{\text{Var}[b|\mathcal{E}]}) + \frac{\mathbb{E}[\text{Cov}(f(\hat{Y}, Y), b|B]}{\text{Var}[b|\mathcal{E}]}$$

and

$$D_\mu^P \longrightarrow \frac{D_\mu \text{Var}[B|\mathcal{E}] - D_\mu^L \text{Cov}(b, \varepsilon)|\mathcal{E}}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])} - \frac{\mathbb{E}[\text{Cov}(Y, B|b, \mathcal{E}) + \mu}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])},$$

respectively; $\varepsilon := B - b$; and $\mu := \text{Cov}(\mathbb{E}(\eta|b, \mathcal{E}), \mathbb{E}[\varepsilon|b, \mathcal{E}]|\mathcal{E})$.

Now, with this form, we can see the following. First, for general b, as long as $\text{Cov}(b, \varepsilon|\mathcal{E}) = 0$ - that is, as long as miscalibration error ε is not correlated with the predictor itself - then we will have exactly the same equation as in 1.1. But we can obtain such a predictor simply by regressing B on b among $\mathcal{E}$; that is, if we run the linear regression

$$B = \alpha + \beta b + \varepsilon,$$

and define b^* as the $\hat{\alpha} + \hat{\beta} b$, then $\varepsilon^* = B - b^*$ by construction satisfies $\text{Cov}(b^*, \varepsilon^*) = 0$.

Then, in that case, we define:

$$D_\mu^{L,*} = D_\mu + \frac{\mathbb{E}[\text{Cov}(f(\hat{Y}, Y)), b^*|B]}{\text{Var}[b^*|\mathcal{E}]},$$

and we can now solve an empirical version of the one-sided problem (i.e. Problem 6.A using b^* instead of b, and all the math discussed above follows directly. However, to solve 7.A, we of course must handle the probabilistic estimator as well.

Here, again we can use $\text{Cov}(b^*, \varepsilon^*|\mathcal{E}) = 0$ and also observe that by construction:

$$\begin{aligned} \mathbb{E}[b^*|E] = \mathbb{E}[B|\mathcal{E}] \implies & \mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}]) \\ & = \mathbb{E}[B|\mathcal{E}](1 - \mathbb{E}[B|\mathcal{E}]) \end{aligned}$$

to simplify the first term in D_μ^{P*}, and so overall write:

$$\mathcal{D}_\mu^{P*} \longrightarrow D_\mu^P - \frac{\mathbb{E}[\text{Cov}(Y, B|b^*, \mathcal{E})] + \text{Cov}(\mathbb{E}[\eta^*|b^*], \mathbb{E}[\varepsilon^*|b^*]\mathcal{E})}{\text{Var}[B|\mathcal{E}]}$$

So to ensure that the lower bound holds, we must now incorporate the second term of the numerator into the optimization problem. But this can be done in a similar manner as before, as the residuals η^* and ε^* can again be expressed as algebraic sample averages.

E. Closed-form Solution to Fair Learning Problem for Regression Setting

In this appendix we provide a closed-form solution to the primal problem Problem 2.A for the special case of linear regression with mean-squared error losses and demographic parity as the disparity metric. We express the constraints in matrix notation and show that the constraints are linear in the parameter β. Thus, we are able to find a unique, closed-form solution for β by solving the first-order conditions. Given a choice of dual variables, it can be interpreted as a regularized heuristic problem with particular weights; while there are no guarantees that this will produce a performant or even feasible

solution, it may be useful when applying the method in its entirety is computationally prohibitive.

We define the following notation for our derivation. Let n denote the number of observations and p the number of features in our dataset. Then let $X \in \mathbb{R}^{n\times p}, y \in \mathbb{R}^{n\times 1}, \beta \in \mathbb{R}^{p\times 1}, b \in \mathbb{R}^{n\times 1}$, and $B \in \{0,1\}^{n\times 1}$. For $j = 0, 1$, let $B_j = \{i : B_i = j\}$ and $n_j = |B_j|$ denote the set of observations for which the observed protected feature $B = j$ and the size of the corresponding set, respectively. Since we consider demographic parity as the disparity metric of interest, we denote the disparity metric as $f(\hat{Y}, Y) = \hat{Y}$.

For ease of exposition, we restate the empirical version of the constrained optimization problem for linear regression and demographic parity.

Problem 9.A.

$$\begin{aligned}&\min_{\beta}\ (y - X\beta)^\top (y - X\beta)\\ &\text{s.t. } \widehat{D}_\mu^L \le \alpha,\\ &\mathbb{E}[\mathrm{Cov}(\widehat{Y}, b|B)] \ge 0,\\ &\mathbb{E}[\mathrm{Cov}(\widehat{Y}, B|b)] \ge 0\end{aligned}$$

As discussed in Section II-A, the linear disparity metric $\widehat{D}_\mu^L$ is the coefficient of the probabilistic attribute b in a linear regression of $\hat{Y}$ on b. Thus, $\widehat{D}_\mu^L$ can be expressed as

$$\widehat{D}_\mu^L = (b^\top b)^{-1}(b^\top X\beta).$$

The covariance of $\hat{Y}$ and b conditional on B can be written as

$$\mathrm{Cov}(\hat{Y}, b|B) = \mathbb{E}(b^\top X\beta|B) - \mathbb{E}(X\beta|B)\mathbb{E}(b|B) \quad (4)$$

We expand the first term on the right-hand side of Equation 4, considering the case where $B = 1$.

$$\begin{aligned}\mathbb{E}(b^\top X\beta|B=1) &= \frac{1}{n_1}\sum_{i\in B_1} b_i X_i\beta\\ &= \frac{1}{n_1}\sum_{i\in B_1}\sum_{j=1}^{p} b_i X_{ij}\beta_j\\ &= \frac{1}{n_1}\sum_{j=1}^{p}\sum_{i\in B_1} b_i X_{ij}\beta_j\\ &= \frac{1}{n_1}\sum_{j=1}^{p}\beta_j\sum_{i\in B_1} b_i X_{ij}.\end{aligned}$$

Collecting the second summation as the vector $v_{1j} = \frac{1}{n_1}\sum_{i\in B_1} b_i X_{ij}$, we can write the expression for $\mathbb{E}(b^\top X\beta|B=1)$ as

$$\mathbb{E}(b^\top X\beta|B=1) = \sum_{j=1}^{p}\beta_j v_{1j} = \beta^\top v_1,$$

where $v_1 = (v_{1j})_{j=1}^p$.

For the second term on the right-hand side of Equation 4 we can rewrite the summation in a similar manner. Again focusing on the case where $B = 1$,

$$\begin{aligned}\mathbb{E}(X\beta|B)\mathbb{E}(b|B) &= \left(\frac{1}{n_1}\sum_{i\in B_1} X_i\beta\right)\left(\frac{1}{n_1}\sum_{i\in B_1} b_i\right)\\ &= \left(\frac{1}{n_1}\sum_{i\in B_1}\sum_{j=1}^{p} X_{ij}\beta_j\right)\left(\frac{1}{n_1}\sum_{i\in B_1} b_i\right)\\ &= \bar{b}_1\frac{1}{n_1}\sum_{i\in B_1}\sum_{j=1}^{p} X_{ij}\beta_j.\end{aligned}$$

We again collect the second summation and write it as $w_{1j} = \frac{1}{n_1}\sum_{i\in B_1} X_{ij}$ and then we can write $\mathbb{E}(X\beta|B)\mathbb{E}(b|B)$ as

$$\mathbb{E}(X\beta|B)\mathbb{E}(b|B) = \bar{b}_1\beta^\top w_1,$$

where $w_1 = (w_{1j})_{j=1}^p$.

Now we can write Equation 4 in matrix notation as,

$$\mathrm{Cov}(\hat{Y}, b|B) = \beta^\top v_1 - \bar{b}_1\beta^\top w_1 + \beta^\top v_0 - \bar{b}_0\beta^\top w_0, \quad (5)$$

where v_0, w_0 and $\bar{b}_0$ are defined equivalently for the set B_0. Finally we take the expectation of this covariance term to get,

$$\begin{aligned}\mathbb{E}(\mathrm{Cov}(\hat{Y}, b|B)) =& \frac{n_1}{n}\left(\beta^\top v_1 - \bar{b}_1\beta^\top w_1\right)\\ &+ \frac{n_0}{n}\left(\beta^\top v_0 - \bar{b}_0\beta^\top w_0\right)\end{aligned} \quad (6)$$

We now consider the covariance of $\hat{Y}$ and B conditional on b which can be written as

$$\mathrm{Cov}(\hat{Y}, B|b) = \mathbb{E}(B^\top X\beta|B) - \mathbb{E}(X\beta|b)\mathbb{E}(B|b). \quad (7)$$

The steps for expressing this conditional covariance in matrix notation are similar to the first covariance term, however, now we are summing over the continuous-valued variable b. Let $k \in [0, 1]$ denote the value of b we are conditioning on and let $G_k = \{i : b_i = k\}$, $n_k = |G_k|$ denote the set of observations with $b = k$ and the size of the set, respectively.

Once again we expand the first term on the right-hand side of Equation 7, this time considering the general case where $b = k$,

$$\mathbb{E}(B^\top X\beta|B) = \frac{1}{n_k}\sum_{j=1}^{p}\beta_j\sum_{i\in G_k} B_i X_{ij} = \beta^\top v_k.$$

Here we define $v_k = (v_{kj})_{j=1}^p$ and $v_{kj} = \frac{1}{n_k}\sum_{i\in G_k} B_i X_{ij}$. Following a similar process for the second term, we can express the term as

$$\mathbb{E}(X\beta|b)\mathbb{E}(B|b) = \bar{B}_k\beta^\top w_k,$$

where $w_k = (w_{kj})_{j=1}^p$ and $w_{kj} = \frac{1}{n_k}\sum_{i\in G_k} X_i j$. Combining the two terms together we write Equation 7 as

$$\mathrm{Cov}(\hat{Y}, B|b) = \sum_k \beta^\top v_k - \bar{B}_k\beta^\top w_k. \quad (8)$$

For the last step we take the expectation of the conditional covariance term to get,

$$\mathbb{E}(\mathrm{Cov}(\hat{Y}, B|b)) = \sum_k \frac{n_k}{n}\left(\beta^\top v_k - \bar{B}_k\beta^\top w_k\right). \quad (9)$$

Now we can write the empirical Lagrangian of Problem 9.A as

$$\widehat{\mathcal{L}}(\beta, \vec{\mu}) = (y - X\beta)^\top (y - X\beta) - \mu_L \left((b^\top b)^{-1} (b^\top X\beta) \right) + \mu_{b|B} \left(\frac{n_1}{n} \left(\beta^\top v_1 - \bar{b}_1 \beta^\top w_1 \right) + \frac{n_0}{n} \left(\beta^\top v_0 - \bar{b}_0 \beta^\top w_0 \right) \right) + \mu_{B|b} \left(\sum_k \frac{n_k}{n} \left(\beta^\top v_k - \bar{B}_k \beta^\top w_k \right) \right).$$

Solving for β we get the solution,

$$\beta^* = \frac{1}{2} (X^\top X)^{-1} \Big[2X^\top y + \mu_L \left((b^\top b)^{-1} (b^\top X) \right) - \mu_{b|B} \left(\frac{n_1}{n} \left(v_1 - \bar{b}_1 w_1 \right) + \frac{n_0}{n} \left(v_0 - \bar{b}_0 w_0 \right) \right) - \mu_{B|b} \left(\sum_k \frac{n_k}{n} \left(v_k - \bar{B}_k w_k \right) \right) \Big].$$

APPENDIX C
DATA

A. L2 Data Description

We select seven features as predictors in our model based on data completeness and predictive value: gender, age, estimated household income, estimated area median household income, estimated home value, area median education, and estimated area median housing value. While L2 provides a handful of other variables that point to political participation (e.g., interest in current events or number of political contributions), these features suffer from issues of data quality and completeness. For instance, only 15% of voters have a non-null value for interest in current events. We winsorize voters with an estimated household income of greater than \$250,000 (4%) of the dataset. Table II shows the distribution of these characteristics, as well as the number of datapoints, for each of the states we consider. In general, across the six states, a little more than half of voters are female, and the average age hovers at around 50. There is high variance across income indicators, though the mean education level attained in all states is just longer than 12 years (a little past high school). Voting rates range from 53% in Georgia to 62% in North Carolina, while Black voters comprise a minority of all voters in each state, anywhere from 16% in Florida to 35% in Louisiana and Georgia.

B. Race Probabilities

The decennial Census in 2010 provides the probabilities of race given common surnames, as well as the probabilities of geography (at the census block group level) given race. In order to incorporate BIFSG, we also use the dataset provided by [57] which has the probabilities of common first names given race.

We default to using BIFSG for all voters but use BISG when a voter's first name is rare since we do not have data for them. Consequently, we only use geography instead of BISG when both one's first name and surname are rare. On average, around 70% of people's race across the six states were predicted using BIFSG, 10% using BISG, and 18% using just geography; $< 2\%$ of observations were dropped because we could not infer race probabilities using any of the three options.

Table III shows results for our BI(FS)G procedure with respect to true race. Accuracy and precision range from 80-90%, but recall is much lower at around 30-50%. Note, however, that we evaluate these metrics by binarizing race probabilities; in our estimators, we use raw probabilities instead, which provide a decent signal to true race. For instance, AUC hovers at 85-90%, while Figure 6 shows that our predicted probabilities are generally well-calibrated to true probability of Black (although BIFSG tends to overestimate the probability of Black).

APPENDIX D
DETAILS ON MEASUREMENT EXPERIMENTS

A. Voter Turnout Prediction Performance

Table IV shows results for voter turnout prediction on logistic regression and random forest models. In general, predicting voter turnout with the features given in L2 is a difficult task. Accuracy and precision hovers at around 70% throughout all experiments, while recall for logistic regression ranges from 71-82% and random forests perform slightly better at 80-90%. This result is in line with previous literature on predicting turnout, which suggest that "whether or not a person votes is to a large degree random" [59]. Note again that our predictors rely solely on demographic factors of voters because those are the most reliable data L2 provides us.

B. The KDC Method

In this section we expand on the different assumptions the KDC method and our method make related to the auxiliary data set. While we consider the case where the test set (with predicted outcomes and race probabilities) subsumes the auxiliary data (which contains true race), KDC mainly considers settings where the marginal distributions $\mathbb{P}(B, Z)$ and $\mathbb{P}(Y, \hat{Y}, Z)$ are learned from two completely independent datasets – in particular, to estimate $\mathbb{P}(B|Z)$ and $\mathbb{P}(\hat{Y}, Y|Z)$. Therefore, in order to produce a fairer comparison between the two methods, we instead reconfigure KDC to incorporate all the data available by treating the auxiliary data as a subset of our test set[4]; doing so only strengthens KDC because we pass in more information to learn both marginal distributions. However, their main method does not leverage information on $\mathbb{P}(Y, Z|B)$, as we do, so their bounds are notably wider. We

[4]Note that a component in calculating the variance of the KDC estimators is r, the proportion of datapoints from the marginal distribution $\mathbb{P}(Y, \hat{Y}, Z)$ to the entire data. Without considering this independence assumption in our calculation, $r = 1$, but this loosely goes against the assumption that r is closer to 0 in Section 7 of [20]. For simplicity, we attenuate the multiplicative terms in the variance calculations of Equations 25 and 26 to give KDC the tightest bounds possible. However, as will be seen in Figure 1, KDC's incredibly large bounds are mostly attributed to its point estimates rather than their variances, which are quite small.

Feature	NC (n=6,305,309)	SC (n=3,191,254)	LA (n=2,678,258)	GA (n=6,686,846)	AL (n=3,197,735)	FL (n=13,703,026)
Gender (F)	0.54 (0.5)	0.54 (0.5)	0.55 (0.5)	0.53 (0.5)	0.54 (0.5)	0.53 (0.5)
Age	49.62 (18.76)	52.2 (18.69)	50.16 (18.29)	48.24 (18.07)	50.27 (18.44)	52.17 (18.89)
Est. Household (HH) Income	89,788.54 (56,880.78)	82,172.22 (53,886.64)	80,770.79 (54,579.77)	90,622.61 (57,699.76)	79,919.66 (52,237.42)	90,145.4 (56,786.94)
Est. Area Median HH Income	76,424.55 (32,239.45)	69,666.4 (25,911.0)	68,068.86 (29,779.93)	78,377.2 (35,941.68)	69,070.63 (27,226.34)	74,547.99 (29,820.33)
Est. Home Value	300,802.36 (202,634.22)	233,354.36 (155,221.32)	199,286.06 (123,564.26)	273,424.9 (176,273.9)	201,901.9 (126,255.0)	360,533.81 (243,854.1)
Area Median Education Year	12.83 (1.13)	12.64 (0.98)	12.36 (0.92)	12.72 (1.12)	12.51 (0.99)	12.65 (0.97)
Area Median Housing Value	206,312.82 (106,274.59)	193,172.13 (107,225.93)	170,521.45 (81,184.86)	206,253.25 (112,142.54)	162,925.8 (81,467.58)	237,245.18 (118,270.22)
Black	0.22	0.26	0.32	0.33	0.27	0.14
Vote in 2016	0.61	0.57	0.63	0.52	0.55	0.57

TABLE II: Distribution of features used for L2 across all six states: from left to right, North Carolina, South Carolina, Louisiana, Georgia, Alabama, and Florida. Each cell shows the mean of each feature and the standard deviation in parentheses. The last two rows show the proportion of observations that are black, and voted in the 2016 General Election.

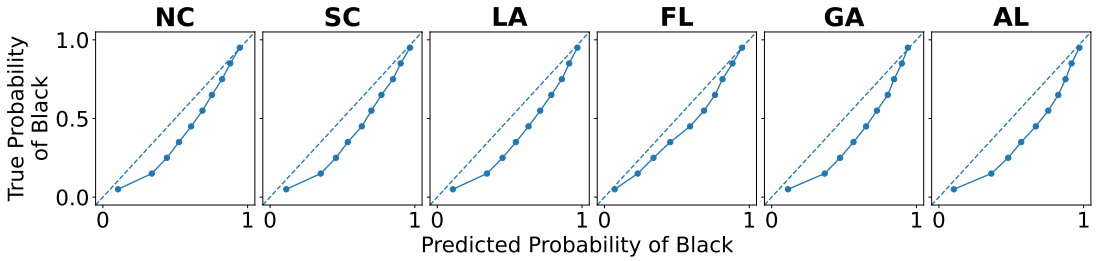


Fig. 6: Calibration plots showing predicted probability of Black (x-axis) versus actual proportion of Black (y-axis).

State	Accuracy	Precision	Recall	AUC
NC	0.83	0.77	0.30	0.85
SC	0.81	0.83	0.35	0.86
LA	0.82	0.87	0.52	0.89
GA	0.80	0.85	0.49	0.88
AL	0.84	0.89	0.45	0.90
FL	0.89	0.80	0.33	0.86

TABLE III: Accuracy, precision, recall (thresholded on 0.5), and AUC for BI(FS)G for all six states considered in L2.

also implement the KDC estimators as originally proposed in Figure 7 but the results do not change substantially[5].

C. Random Forest

We also run experiments on bounding disparity when voter turnout is predicted on random forest models, as seen in Figure 8. We observe similar results to logistic regression in that our methods always bound true disparity within 95% confidence intervals, and with bounds that are markedly tighter than KDC's. While our bounds are always within 5 p.p. and the same sign as true disparity, KDC is ranges from -0.5 to 0.5.

[5]In Appendix A.5, [20] do in fact propose an estimator where the independence assumption is violated (i.e., precisely the setting we consider where we have race probabilities in our entire data), but it suffers from two key limitations: *a)* we are only provided estimators for DD and none other disparity measure, and *b)* we implemented the DD estimator and it failed to bound true disparity in both applications we consider – see Figure 7.

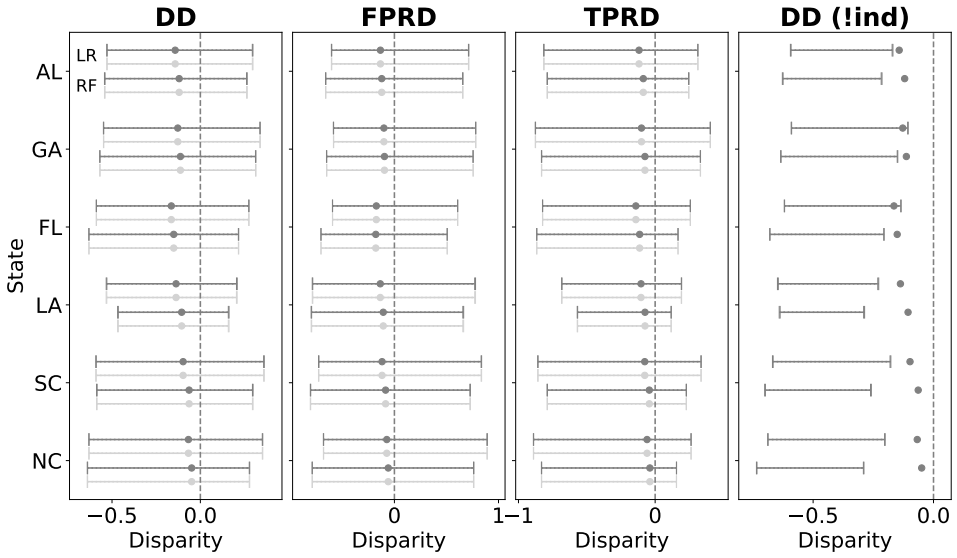


Fig. 7: Comparison of different KDC implementations. In dark grey, we have our implementation that violates the independence assumption in [20]. In light grey, we have KDC's original implementation with the independence assumption – nothing substantively changed. The top and bottom pairs of each state correspond to the estimators from logistic regression (LR) and random forest (RF) models, respectively. [20] additionally proposes estimators for estimating DD where the independence assumption is violated but they rarely bound true disparity (right subfigure), so we omit these results in our main experiments.

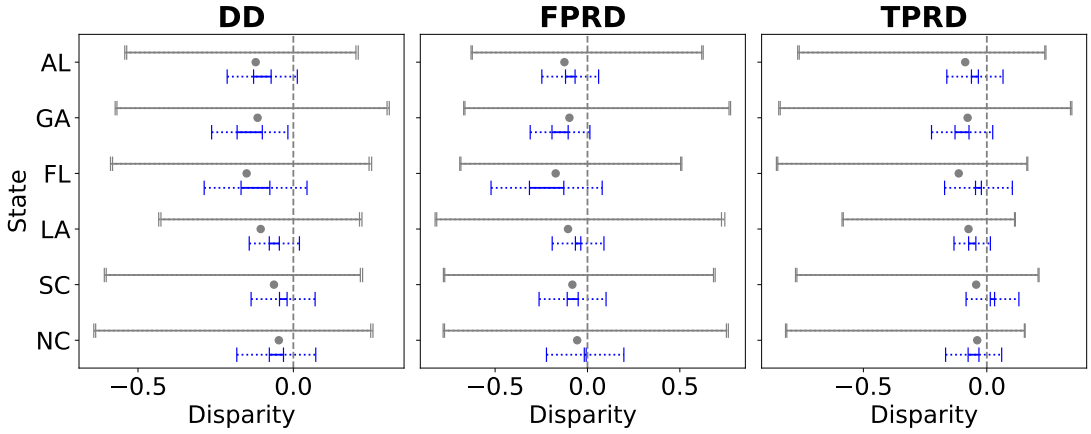


Fig. 8: Comparison of our method of bounding true disparity (blue) to the method proposed in [20] (grey), using a random forest model to predict voter turnout on L2 data in six states. We evaluate three disparity measures: demographic disparity (DD), false positive rate disp. (FPRD), and true positive rate disp. (TPRD). The grey dot represents true disparity. Both methods always bound true disparity within their 95% standard errors.

APPENDIX E
DETAILS ON TRAINING EXPERIMENTS

A. Experimental Setup

As noted in the main text, to enforce fairness constraints during training, we solve the empirical version of Problem 1.A and its symmetric analogue, which enforces negative covariance conditions and $\widehat{D}_{\mu}^{L}$ as a (negative) lower bound. For both of these problems we run the primal-dual algorithm described in Algorithm 1 for T iterations and then select the iteration from these two problems with the lowest loss on the training data while satisfying the constraints on the training and labeled subset.

We use the [14] method with logistic regression models under the hood.

B. CSL (Chamon et al.)

We implement our constrained problem using the official Pytorch implementation provided by [14][6] for a logistic regression model We run the non-convex optimization problem

[6]https://github.com/lfochamon/csl

State	Model	Accuracy	Precision	Recall	AUC
NC	LR	0.72	0.75	0.81	0.75
	RF	0.72	0.72	0.89	0.76
SC	LR	0.67	0.69	0.77	0.71
	RF	0.67	0.67	0.86	0.71
LA	LR	0.70	0.73	0.84	0.72
	RF	0.70	0.71	0.91	0.73
GA	LR	0.69	0.70	0.71	0.75
	RF	0.69	0.68	0.78	0.75
AL	LR	0.67	0.69	0.74	0.72
	RF	0.67	0.67	0.80	0.72
FL	LR	0.67	0.69	0.76	0.71
	RF	0.67	0.67	0.85	0.72

TABLE IV: Accuracy, precision, recall, and AUC for voter turnout prediction for all six states considered in L2. We evaluate two different model performances for turnout prediction: logistic regression (LR) and random forests (RF).

for 1,000 iterations with a batch size of 1,024 and use Adam [60] for the gradient updates of the primal and dual problems with learning rates 0.001 and 0.005, respectively. We provide further explanation of the mathematical background to the [14] method in Appendix B above.

C. The Method of Wang et al.

[21] propose two methods to impose fairness with noisy labels: *1)* a distributionally robust optimization approach and *2)* another optimization approach using robust fairness constraints, which is based on [20]. We use code provided by [21][7] to implement only the second method because it directly utilizes the protected attribute probabilities and yields better results.

We tune the following hyperparameters: $\eta_\theta \in \{0.001, 0.01, 0.1\}$ and $\eta_\lambda \in \{0.25, 0.5, 1, 2\}$, which correspond to the descent step for θ and the ascent step for λ in a zero-sum game between the θ-player and λ-player, see Algorithm 1 and 4 of [21]. Finally, we also tune $\eta_w \in \{0.001, 0.01, 0.1\}$, which is the ascent step for w (a component in the robust fairness criteria), see Algorithm 3 of [21]. In order to choose the best hyperparameters, we use the same data as outlined in Section IV-C1 (80/20 train/test split), but use a validation set on 30% of the training data (i.e., 24% of the entire data). Note that as implemented in the codebase, [21] chooses the hyperparameter that results in the lowest loss while adhering to the fairness constraint with respect to **true race**. Since we assume access to true race on a small subset (1%) of the data, we only evaluate the fairness constraint on 1% of the validation set.

D. The Method of Mozannar et al.

[24] primarily focus on the setting of training a fair model with differentially private demograpghic data, which poses assumptions which are infeasible for our setting—however, the authors do propose a potential extension of their method to handle a case that matches ours: training a fair model with incomplete demographic data. The authors do not discuss this in detail or provide the code for this extension, so we modify the code [24] provide for their paper to implement the extension of their approach, detailed in Section 6 of their paper that is relevant for our setting. This involves using Fairlearn's[8] exponentiated gradient method changed so that it will only update for its fairness-related loss on data points in the labeled subset, but allows classification loss to be calculated over the entire training set.

We note that Mozannar's method guarantees fairness violation 2(epsilon + best gap) [50] on their test set where epsilon is set by the user, but gives no method of approximating best_gap. Thus, we set epsilon = $\alpha/2$ (i.e. assume best gap=0) in our experiments in order to come as close as possible to their method providing similar fairness bounds to ours on the test set.

E. Pareto-Frontier of Accuracy vs. Disparity

In Figure 9 through 12, we show the fairness-accuracy Pareto frontiers for the L2 and COMPAS datasets enforcing demographic parity (DD), false positive rate parity (FPRD), and true positive rate parity (TPRD). We first note that the full benefit of using our method is not fully captured by comparison along Pareto frontiers. This is because the core aim of our method is to ensure that the disparity does not go over a particular bound input by the user, so the relationship between the exact amount of disparity observed on the test data to the bound set by the user is important beyond the fairness-accuracy tradeoff itself; even if another method were to appear better in terms of a fairness accuracy tradeoff, it cannot make the guarantees to the user about meeting the bound that ours can. We highlight the difference between the desired bound and the disparity demonstrated on the test set by noting particular points in the pareto frontier with symbols indicating the specified bound (for example, in Figure 9, a circle indicates a a bound or α value of 0.04). We note the specified bounds as dashed lines parallel to the y axis. As we can see from all of the graphs, our method is the only method which consistently meets the desired fairness bound, and thus fully explores the disparity regimes targeted.

In terms of dominance on the accuracy-fairness Pareto frontier, we note that we do not count the oracle (the red line) against our method as that is a model with complete knowledge of the protected attributes of the dataset, where as we only have protected attributes for a small subset. For the L2 experiments, our method strictly dominates Mozannar et al. and Wang et al. methods when available for comparison for DD, FPRD, and TPRD. As expected, the oracle dominates our method. For the labeled subset method, our method dominates this approach nearly everywhere in the FPRD and TPRD plots. The labeled subset method dominates in the middle fairness values of the accuracy-fairness frontier for DD on L2 data.

[7]https://github.com/wenshuoguo/robust-fairness-code

[8]https://fairlearn.org/

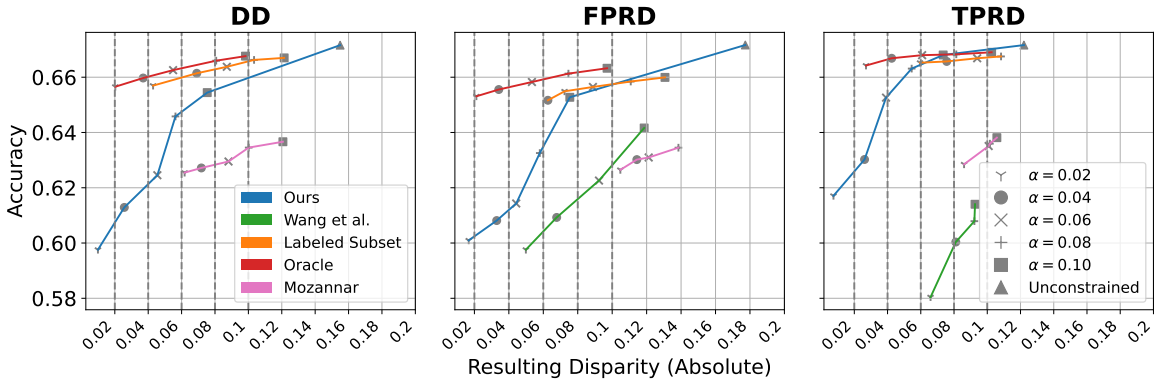


Fig. 9: Resulting disparity (x-axis) and accuracy (y-axis) trade-off for L2 Florida data. Each point corresponds to the average result over 10 seeds on a given target disparity α, which map to different marker styles (e.g., circle points are experiments with target disparity of 0.04). For ease of interpretation, each of the target disparities are marked in dashed vertical lines; e.g., any circle point to the left of 0.04 satisfies the desired target disparity.

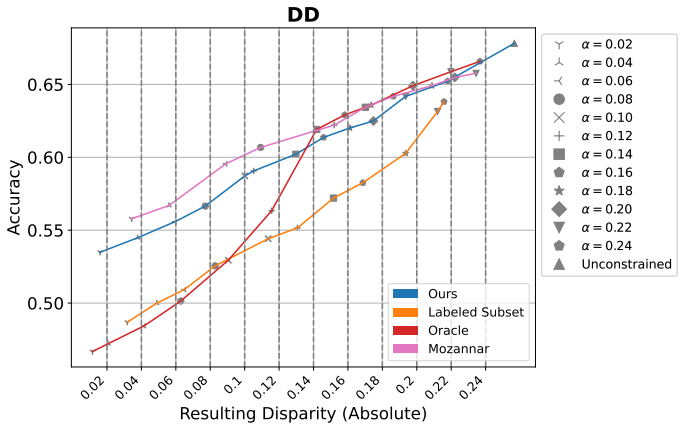


Fig. 10: Resulting demographic disparity (x-axis) and accuracy (y-axis) trade-off for COMPAS data. Each point corresponds to the average result over 10 seeds on a given target disparity α, which map to different marker styles (e.g., square points are experiments with target disparity of 0.14). For ease of interpretation, each of the target disparities are marked in dashed vertical lines; e.g., any circle point to the left of 0.14 satisfies the desired target disparity.

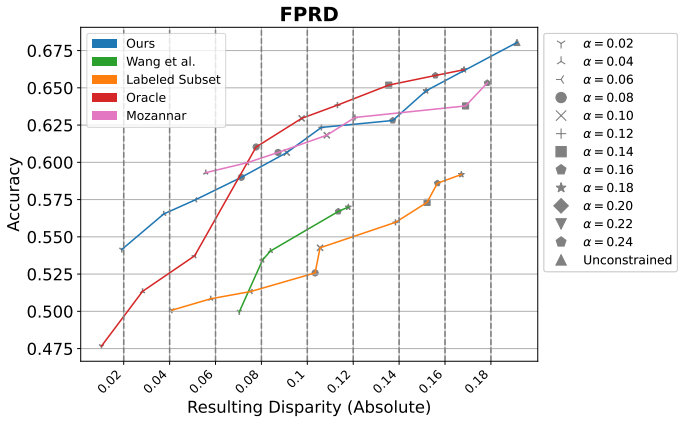


Fig. 11: Resulting false positive rate disparity (x-axis) and accuracy (y-axis) trade-off for COMPAS data. Each point corresponds to the average result over 10 seeds on a given target disparity α, which map to different marker styles (e.g., circle points are experiments with target disparity of 0.08). For ease of interpretation, each of the target disparities are marked in dashed vertical lines; e.g., any circle point to the left of 0.08 satisfies the desired target disparity.

Metric	State	Method	Lower Bound (95% CI)	True Disparity	Upper Bound (95% CI)
DP	AL	KDC	-0.52 ± 0.01	-0.14	0.23 ± 0.01
		Ours	-0.14 ± 0.09	-0.14	-0.08 ± 0.09
	FL	KDC	-0.55 ± 0.01	-0.16	0.28 ± 0.01
		Ours	-0.27 ± 0.13	-0.16	-0.12 ± 0.13
	GA	KDC	-0.55 ± 0.01	-0.13	0.32 ± 0.01
		Ours	-0.22 ± 0.08	-0.13	-0.12 ± 0.08
	LA	KDC	-0.53 ± 0.01	-0.14	0.25 ± 0.01
		Ours	-0.12 ± 0.07	-0.14	-0.07 ± 0.07
	NC	KDC	-0.62 ± 0.01	-0.07	0.32 ± 0.01
		Ours	-0.13 ± 0.12	-0.07	-0.05 ± 0.12
	SC	KDC	-0.61 ± 0.01	-0.1	0.28 ± 0.01
		Ours	-0.08 ± 0.1	-0.1	-0.03 ± 0.1
FPR	AL	KDC	-0.58 ± 0.01	-0.14	0.69 ± 0.01
		Ours	-0.14 ± 0.13	-0.14	-0.08 ± 0.13
	FL	KDC	-0.57 ± 0.01	-0.16	0.6 ± 0.01
		Ours	-0.31 ± 0.21	-0.16	-0.13 ± 0.21
	GA	KDC	-0.59 ± 0.01	-0.1	0.77 ± 0.01
		Ours	-0.22 ± 0.11	-0.1	-0.12 ± 0.11
	LA	KDC	-0.81 ± 0.01	-0.13	0.85 ± 0.02
		Ours	-0.08 ± 0.13	-0.13	-0.05 ± 0.13
	NC	KDC	-0.65 ± 0.01	-0.07	0.86 ± 0.01
		Ours	-0.07 ± 0.21	-0.07	-0.03 ± 0.2
	SC	KDC	-0.69 ± 0.01	-0.12	0.77 ± 0.01
		Ours	-0.14 ± 0.15	-0.12	-0.06 ± 0.15
TPR	AL	KDC	-0.78 ± 0.01	-0.12	0.3 ± 0.01
		Ours	-0.07 ± 0.11	-0.12	-0.04 ± 0.11
	FL	KDC	-0.8 ± 0.01	-0.14	0.25 ± 0.0
		Ours	-0.21 ± 0.15	-0.14	-0.1 ± 0.15
	GA	KDC	-0.88 ± 0.01	-0.11	0.4 ± 0.01
		Ours	-0.18 ± 0.11	-0.11	-0.1 ± 0.11
	LA	KDC	-0.68 ± 0.01	-0.1	0.2 ± 0.0
		Ours	-0.14 ± 0.08	-0.1	-0.08 ± 0.08
	NC	KDC	-0.86 ± 0.01	-0.06	0.25 ± 0.0
		Ours	-0.12 ± 0.12	-0.06	-0.05 ± 0.12
	SC	KDC	-0.84 ± 0.01	-0.08	0.31 ± 0.0
		Ours	-0.0 ± 0.12	-0.08	-0.0 ± 0.12

TABLE V: Companion table to Figure 1.

However, again we note that the labeled subset method was not able to meet the desired fairness bounds on *any* experiment across the L2 and COMPAS datasets, so there are other reasons why this method is undesirable in situations where a reliable bound is needed. For FPRD on the COMPAS dataset, with a few exceptions, our method dominates all other methods (except the oracle, as expected). For TPRD, besides the oracle, a few points in the middle of the range (0.16, 0.14, 0.12, 0.1, 0.8) are dominated by either Mozannar et al. (0.1, 0.14), Wang et al. (0.08, 0.12, 0.14) or labeled subset (0.16). However, our method dominates the most consistently (7 out of 12 points) and noticeably in the lower unfairness regime. For DD, Mozannar et al. lead to a comparable but lower fairness-accuracy tradeoff for much of the space, but again we note that the Mozannar et al. method cannot meet the desired fairness bounds for 33 out of 36 experiments, suggesting it is not preferable in situations where a bound is necessary.

F. Results on Oracle and Naive

In Figure 13, we present the mean and standard deviation of the resulting disparity and on the test set, as well as classifier accuracy on the test set, of experiments with our method compared to an oracle model, that has access to ground truth race on the *whole* dataset and uses these to enforce a constraint directly on ground truth disparity during training, as well as a naive model which simply enforces a constrained directly on the observed disparity of the noisy labels, without any

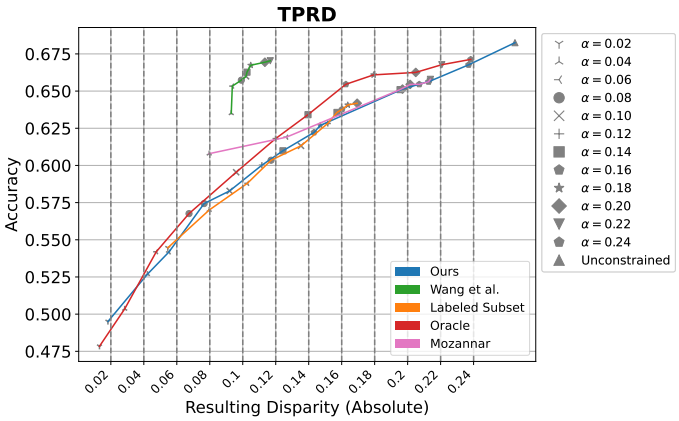


Fig. 12: Resulting true positive rate disparity (x-axis) and accuracy (y-axis) trade-off for COMPAS data. Each point corresponds to the average result over 10 seeds on a given target disparity α, which map to different marker styles (e.g., square points are experiments with target disparity of 0.14). For ease of interpretation, each of the target disparities are marked in dashed vertical lines; e.g., any circle point to the left of 0.14 satisfies the desired target disparity.

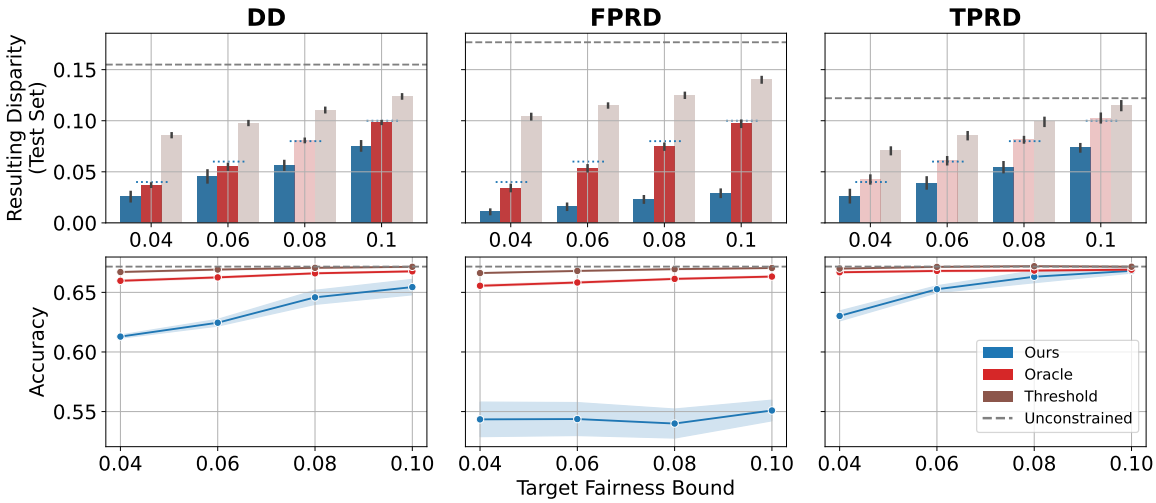


Fig. 13: Mean and standard deviation of resulting disparity (top, y-axis) and accuracy (bottom, y-axis) on the test set after enforcing the target fairness bounds (x-axis) on our method (blue); using ground truth race on the entire data, i.e., "oracle" model (red); and using only the estimated race probabilities, thresholded to be binary (brown) 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.

correction. (Namely, in this technique, we simply threshold the probabilistic predictions of race on 0.5 to make them binary, and use as race labels.) As a whole, we perform relatively comparably to the oracle, except on FPRD. We always outperform the naive method in terms of reducing disparity, which is to be expected. We typically perform within 2 percentage points of accuracy from the oracle, (except for the 0.04 and 0.06 bounds on DD and the 0.04 bound on TPRD). We suggest the accuracy results in this figure show the fairness-accuracy trade-off in this setting: when we dip below the oracle in terms of accuracy, it is most often because we are bounding disparity lower than the oracle is (e.g., on the 0.04 bounds in DD or TPRD). And, while we do not outperform the naive method in terms of accuracy, we consistently out-perform it in terms of disparity.

Appendix F
Additional Experiments: COMPAS

In this section, we present a suite of additional experiments we run on the COMPAS [45] dataset. The COMPAS algorithm is used by parole officers and judges across the United States to determine a criminal's risk of recidivism, or re-committing the same crime. In 2016, ProPublica released a seminal article [45] detailing how the algorithm is systematically biased against Black defendants. The dataset used to train the algorithm has since been widely used as benchmarks in the fair machine learning literature.

A. Data Description

We use the eight features used in previous analyses of the dataset as predictors in our model: the decile of the COMPAS score, the decile of the predicted COMPAS score, the number of prior crimes committed, the number of days before screening arrest, the number of days spent in jail, an indicator for whether the crime committed was a felony, age split into categories, and the score in categorical form. We process the data following [45], resulting in $n = 6,128$ data points. Table VI outlines the feature distribution of the dataset.

Feature	COMPAS (n=6,128)
Decile Score	4.41 (2.84)
Predited Decile Score	3.64 (2.49)
# of Priors	3.23 (4.72)
# of Days Before Screening Arrest	-1.75 (5.05)
Length of Stay in Jail (Hours)	361.26 (1,118.60)
Crime is a Felony	0.64 (0.48)
Age Category	0.65 (0.82)
Risk Score in 3 Levels	1.08 (0.66)
Black	0.51
Two Year Recidivism	0.45

TABLE VI: Distribution of features used for COMPAS. Each cell shows the mean of each feature and the standard deviation in parentheses. The last two rows show the proportion of observations that are Black and who recidivized within two years.

B. Race Probabilities

We generate estimates of race (Black vs. non-Black) based on first name and last name using a LSTM model used in [49] that was trained on voter rolls from Florida. The predictive performance and calibration of these estimates is displayed in Table VII and Figure 14, respectively. In general, the results are quite reasonable; accuracy is at 73% while the AUC is 86%. The probabilities are somewhat calibrated, although the LSTM model tends to overestimate the probability of Black.

Accuracy	Precision	Recall	AUC
0.73	0.86	0.56	0.86

TABLE VII: Accuracy, precision, recall (thresholded on 0.5), and AUC for predicting probability a person is Black in the COMPAS dataset.

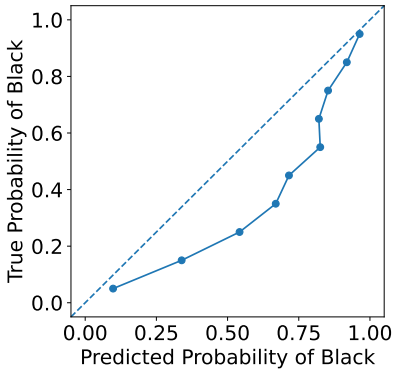


Fig. 14: Calibration plot showing the predicted probability a person in the dataset is Black (x-axis) versus the actual proportion of Black people in the dataset (y-axis) for COMPAS.

C. Measurement Experiments

We first compare our method of bounding disparity to that of KDC. We train an unconstrained logistic regression model with a 80/20 split on the data, i.e., $n = 1,226$ in the test set. Then, we construct the labeled subset by sampling 50% of the test set ($n = 613$) and use that to check out covariance constraints. We also compute $\hat{D}_L$ and $\hat{D}_P$ with standard errors on the entire test set, as specified by the procedure in Appendix Section D.

Our main results are displayed in Figure 2. Similar to the L2 data, our bounds are consistently tighter than KDC, albeit to a lesser extent in this case since the COMPAS dataset is significantly smaller. Despite this fact, we emphasize that, unlike KDC, our estimators are always within the same sign as the true disparity, barring the standard errors which shrink as the data grows larger.

Accuracy	Precision	Recall	AUC
0.69	0.69	0.57	0.74

TABLE VIII: Accuracy, precision, recall (thresholded on 0.5), and AUC for predicting two-year recidivism on the COMPAS dataset using a logistic regression model.

D. Training Experiments Details

We compare our training method to [21], [24] and a baseline where we directly enforce disparity constraints on only the labeled subset. We run 10 trials – each corresponding to different seeds – and report the mean and standard deviation of

the accuracy and disparity on the test set in Figure 4. For each trial, we split our data ($n = 6,128$) into train and test sets, with a 80/20 split. From the training set, we subsample the labeled subset so that it is 10% of the total data (around $n = 613$). We chose a higher proportion of the data compared to L2 to adjust for the smaller dataset. The remaining details are as described in Section IV-C1. Note that the resulting disparities for the unconstrained model differ among the three fairness metrics. On DD and TPRD, the unconstrained model resulted in a 0.28-0.29 disparity, but it drops to 0.21 for FPRD. We adjusted our target fairness bounds accordingly.

APPENDIX G
SIMULATIONS

A. Simulation Design

In this section, we describe the design of our simulation used for additional experiments.

- Primitive features $Z_1, ..., Z_m$
- Conditional probability b of being Black a function of $Z_1...Z_m$
- Realized status as Black or not B drawn from Bernoulli(b)
- Downstream features $X_1, ...X_p$, a function of $Z_1, ..., Z_m$ and B
- Score for outcome $P(Y)$, a function of downstream features $X_1...X_p$
- Outcome Y,which is an indicator of $P(Y)$ at threshold τ with some noise probability of being flipped $0 \leftrightarrow 1$

The primitive features $Z_1, ..., Z_p$ intuitively represent the variables that correspond to proxies in BIFSG, e.g. geographic locations. They serve a dual role: first, as in BIFSG, they give rise to the probability that an individual is Black. Second, since the secondary features X are a function of Z, they affect the distribution of these features; thus downstream, they affect $P(Y)$ and ultimately Y, but do not directly enter into $P(Y)$ or Y themselves. This corresponds to how geography and other variables which are correlated to race may also be correlated to many learning-relevant features, even when not directly entering causing the outcome of interest themselves. Note that in addition to primitives affecting $P(Y)$ through each X, we allow for B to affect $P(Y)$. This corresponds to how there may be associations between group membership and features which affect the outcome of the interest via the downstream features even if the group status is not directly relevant tot he outcome of interest.

These relationships are not fully specified by the description in the text above, of course, so we provide details of the selected functional forms in Table IX. Figure 12 also summarizes the features and their associative relationships visually. This visualization, along with the language of directed acyclic graphs (DAGs), allows us to more easily reason about whether the covariance conditions are likely to be satisfied in our model, at least for the underlying outcome.

B. Experimental Setup

Following the notation above, we have p to be the number of features X in our data, and let n be the number of

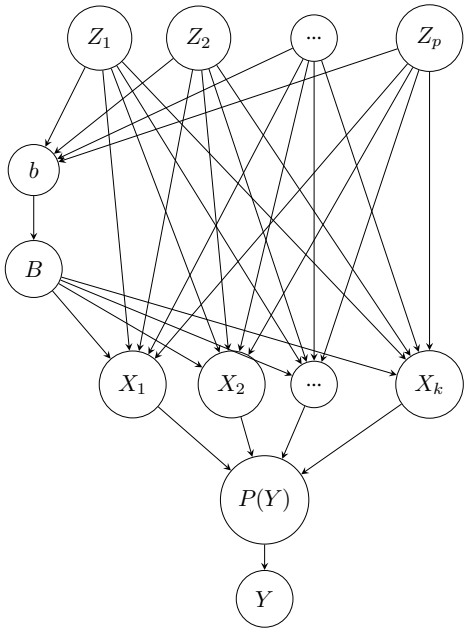


Fig. 15: A heuristic depiction of the data generating process for our simulations. Nodes indicate random variables, and edges indicate (causal) relationships between nodes. Importantly, relationships are not necessarily linear.

datapoints. We run experiments for $p \in \{10, 20, 50\}$ and $n \in \{5000, 10000, 50000\}$. For each p, we fix the parameters in the data generation process and realize 50,000 datapoints. Refer to Table X for a list of parameter values, which differ slightly for each p to control demographic disparity on the dataset at around 0.25-0.28. For experiments $n \in \{5000, 10000\}$, we simply randomly subsample from the 50,000 dataset.

The last dimension we tune is the size of the labeled subset (measured by the percentage of n), which from hereon we refer to as e. For each n, we specified slightly different e as outlined in Table XI. This is to account for the fact that, for instance, one might need 40% of 5,000 datapoints with protected attribute labels to learn a predictor that reaches the target disparity bound. On the other hand, using 20% of 50,000 datapoints might be more than enough, especially considering the exponentially higher costs to query thousands of people's protected attributes.

We prototype these simulation experiments on demographic parity. For each experiment, we split the data 80/20 into train/test data, then repeat 10 times with different seeds. We run both our method and the labeled subset method, evaluating disparity and accuracy on the test set.

C. Results

We present our results in Figures 5 and 16. In Figure 5, we see that while increasing the size of the labeled subset can sometimes lead to a regime where training on the labeled subset alone can produce a model which comes close to (or in one case–$n = 50,000$, $p = 10$, reaches) the desired disparity

Feature	Interpretation	Functional Form
Z_j	Primitive Feature	$Z_j \sim U[0,1],\ j = 1, ...m$
X_i	Secondary Feature	$X_i = \sum_{k=1}^{h_k} c_i X_i^k, i = 1, ...p$
h_k	Degree	$h_k \sim U\{0,1,2,3\}$
c_i	Coefficients	$c_i \sim U[0,1],\ i = 1, ...p$
b	Probability Black	$b = \max\{0, \min\{1, \tilde{b}\}\},$
		$\tilde{b} \sim \begin{cases} \mathcal{N}(0.1, .04) & \frac{1}{m}\sum_{j=1}^{m} Z_j \leq \tau_b \\ \mathcal{N}(0.9, .04) & \frac{1}{m}\sum_{j=1}^{m} Z_j > \tau_b \end{cases}$
τ_b	Threshold on b (based Irwin-Hall distribution)	$\frac{1}{2} + 1.2\sqrt{1/(12m)}$
B	Indicator for Black	$B \sim \text{Bernoulli}(b)$
$\tilde{P}(Y)$	Score of Outcome	$\tilde{P}(Y) = \sum_i \left[d_i X_i^k + d_{iB} B\right]$
$P(Y)$	Normalized Score of Outcome	$P(Y) = \frac{\tilde{P}(Y) - \min(\tilde{P}(Y))}{\max(\tilde{P}(Y)) - \min(\tilde{P}(Y))}$
Y	Realized Outcome	$Y \sim \begin{cases} \text{Bernoulli}(0.1) & P(Y) \leq \tau \\ \text{Bernoulli}(0.9) & P(Y) > \tau \end{cases}$
d_i	Coefficients for features X	$d_i \sim U[0,1]$
d_{iB}	Coefficients for indicator for Black	$d_{iB} \sim U[0, u_B]$

TABLE IX: Description of several variables we use in our simulation study and their functional forms. For ease of notation, we omit the index denoting individuals in the dataset. Unspecified constants were selected by inspection to match key indicators across scenario and are specified in Table 8.

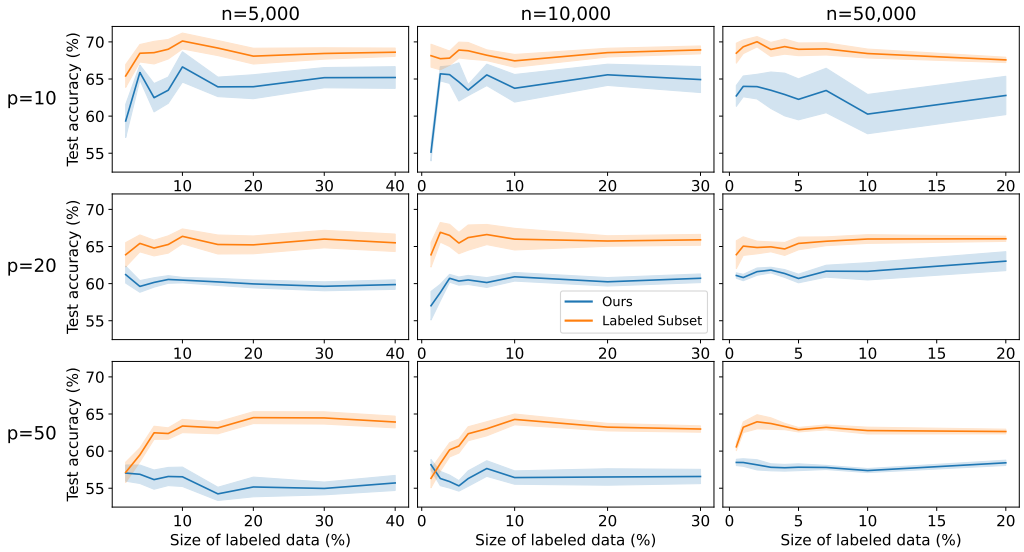


Fig. 16: We present a three by three figure showing the test accuracy of the models created using our disparity reduction method when compared with relying on training models only on the labeled subset and reducing disparity by directly enforcing a constraint on the protected attribute labels. The rows correspond to datasets of increasing sizes (number of features from 10 to 50), indicating problems of increasing complexity. The columns correspond to the size of the overall dataset, ranging from 5,000 to 50,000 samples. The x-axis shows the percentage of the total dataset is dedicated to the labeled subset, and the y-axis denotes the test accuracy of the models. The blue graphs correspond to our method, and the orange to the labeled subset method.

p	m	τ	u_B
10	4	0.4	0.05
20	5	0.4	0.1
50	10	0.425	0.2

TABLE X: List of parameters in the data generation process for each p, the number of secondary features X in the data. m corresponds to the number of primitive features Z, τ is the threshold for $P(Y)$, while u_B is the upper bound for the uniform distribution to generate d_{iB}, see Table IX.

n	e
5,000	$\{2, 4, 6, 8, 10, 15, 20, 30, 40\}$
10,000	$\{1, 2, 3, 4, 5, 7, 10, 20, 30\}$
50,000	$\{0.5, 1, 2, 3, 4, 5, 7, 10, 20\}$

TABLE XI: Suite of experiments varying percentage of the data taken as labeled subset (e) by the size of the full dataset (n).

bound, for the most part, even with a large labeled subset, the mean of the disparity over 10 trials is above the desired disparity threshold. Meanwhile, our method stays below the desired disparity threshold across all nine experiments.

As we can see by looking at the rows from top to bottom, the complex (i.e., more features in the data) the problem is, the more data is necessary for the labeled subset to get close to the desired disparity bound. Thus, our simulation experiment sheds light on the fact that model applications with small amounts of labeled data, and more complex data, are particularly well-suited for our method.

APPENDIX H
PAST REVIEWS AND RESPONSES

We enclose our unedited reviews and responses from a prior submission. We provide here a summary of our additions and changes from our initial version of the paper:

- **Additional Datasets**: We provide experiments on the COMPAS dataset, using a non-BIFSG proxy as our probabilistic estimate. We also provide experiments on synthetic data.
- **Additional Methods**: We incorporate a comparison to Mozannar et al. [24], which, to our knowledge, is the only paper that addresses a setting that is close to ours. We also discuss why other methods are inapplicable to our setting.
- **Simulation Study:** We conduct a study on simulated data where we can vary both dataset size and underlying complexity of the data-generating process to understand under what regimes our method is preferable to existing methods.

We also add a variety of clarification measures, including adding Table II-A to the main paper; adding a section to the introduction which clarified the real-world situations where our technique may be used, and setting some of the stakes for our paper, and others.

Additional Datasets: In response to reviewer concerns surrounding the number of datasets considered, we provide experimental results of our method over COMPAS data, which had not originally been in the paper. We also note that while the paper focused on the application of voter turnout, the 2 states for which we run learning experiments and the 7 states for which we perform measurement experiments all have significantly different data distributions, providing insight to different data settings.

Additional Baselines: In response to one reviewer, we added a comparison to Mozannar et al. [24] in our main set of results. As we show in the paper, we almost always outperform their method on disparity reduction and accuracy. We address the particular comparisons the reviewer brought up, and why the Mozannar paper was the only comparison which shed additional light to the paper, here:

Our setting differs substantially from the papers cited by the reviewer and the literature more broadly. In particular, while some papers assume noisy, perturbed, or nonexistent protected attribute data, we consider an empirically common "mixed" case, in which the learner has access to a probabilistic proxy for the protected attribute over all their data, and ground truth over a small subset only, spurring the question of how best to combine the two data sources with fairness and performance in mind.

With respect to the works the reviewer specifically highlighted: Mozannar et. al [24], primarily focuses on a setting with a very strong conditional independence assumption on the protected attribute proxy, which is unlikely to be met besides the (infeasible, for our setting) case in which the learner can differentially privately generate noisy labels using the ground truth. However, as you point out, the authors do propose a potential extension of their method to handle a case that matches ours. The authors do not discuss this in detail or provide the code for this extension, but we have modified their original code to implement the changes described and include these results in our updated draft. We note that Mozannar et. al.'s method almost never meets the disparity bounds, and does worse in accuracy on L2; see Figures 1-2 attached. Levy et. al [61] proposes and analyzes distributionally robust optimization (DRO) algorithms. At its core, the idea of DRO is to optimize for the worst-case distribution that is "nearby" according to some metrics. This has applications to fairness, as the reviewer points out. In particular, Wang et al, to which we have already provided a comparison, applies DRO to precisely the problem of fair classification with noisy sensitive features. Like other methods that aim for worst-case guarantees, the method is robust but sacrifices performance by not using all available information, as can be seen in Figure 2 in our main results. Diana et. al. [62] aims at an upstream and somewhat orthogonal problem to that of ours. Our paper focuses on training a fair model using an imperfect proxy (e.g. BIFSG), where as Diana et al. study how to learn a model to create proxy features with which to learn a fair model. They do not focus on the downstream task of training the fair model, and the methods are hence not directly comparable. We are happy to include discussion of the above works in the related work section of our paper.

Investigating Questions Around the Size of the Labeled Subset: Prompted by the response of several reviewers, we added experiments on the relationship between the size of the labeled subset, dataset complexity, and the efficacy of our method which we pointed to as future work in our limitations section. Our results are in Section IV-D of the paper. Overall, our method is consistently able to bound disparity across varying sizes of the labeled subset, whereas training on the labeled subset alone only comes close to bounding disparity in low-complexity regimes.

We now enclose the unedited reviews and our responses.

Summary:
The paper proposes method of evaluating and mitigating unfairness with probabilistic estimate of sensitive attribute (e.g. predicting race from zip code). The key insight is to derive upper and lower bound on the fairness measure, and then optimize the bound-related terms for mitigation.

Soundness: 2 fair
Presentation: 2 fair
Contribution: 2 fair
Strengths:
1. A practical and important problem

Weaknesses:
1. My main concern is the experiments. It only tried on one dataset, this is rare in machine learning. In addition, only one baseline (Wang et. al) is compared. Fairness under noisy sensitive attribute is a growing area with many more recent baselines should be compared to, e.g.

[1] Levy, Daniel, et al. "Large-scale methods for distributionally robust optimization." Advances in Neural Information Processing Systems 33 (2020): 8847-8860.

[2] Mozannar, Hussein, Mesrob Ohannessian, and Nathan Srebro. "Fair learning with private demographic data." International Conference on Machine Learning. PMLR, 2020.

[3] Diana, Emily, et al. "Multiaccurate proxies for downstream fairness." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.

There are more. I am not suggesting authors to compare to all of them. But since many papers in this area target the problem in a similar way as this paper, i.e. deriving bounds on fairness measure and optimizing it as proxy, I think the experimental comparison is important.

2. I think notation can and should be simplified, e.g. the notation $C_{f,B|b,\mathcal{E}}$ is clumsy. Consider simplifying by not defining event $\mathcal{E}$ and writing it out explicitly. And write b as $B^{+}_{|Z,\mathcal{E}}$ more explicitly.

Questions:
1. Can authors give some examples of what the event $\mathcal{E}$ can be?

Limitations:
See Weakness.

Flag For Ethics Review: No ethics review needed.
Rating: 3: Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility and incompletely addressed ethical considerations.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Fig. 17

Rebuttal:
Thank you for your review. We have made the following updates, which we will describe in detail below:

- **We have added experiments on the COMPAS dataset, where our method performs better at decreasing disparity than the labeled subset method and Mozannar et al, which we include as a comparison.** We also note that the 2 states for which we run learning and the 7 states for which we perform measurement all have significantly different data distributions.
- **We have added a comparison to Mozannar et al. in our main results,** though there are substantial differences between our method and theirs, **where our method outperforms on bounding disparity in all experiments and accuracy in most.** We describe our differences from the other references pointed to below.

Additional Datasets

The experiments in the current draft of the paper are focused on one application, but over different populations: we consider data from 7 different states for measurement and 2 states for our predictive task. As we detail in appendices C and D, the states have significant variation in key metrics like the fraction Black and non-Black, turnout, and the ease of the underlying prediction problem, which provide informative variation over the space of possible problem settings.

We have selected this application because it is one with easily available public data in which individuals' names and addresses are linked to their data, allowing for the application of Bayesian Improved Firstname Surname Geocoding (BIFSG), which is likely the most common race proxy used for fairness in industry [2] and government [1]. The measurement of disparity in voting is also an important question of both academic interest and legal importance. However, as our method does not require the use of BIFSG to estimate a protected feature, we have added experiments on an additional dataset, COMPAS; for this application, we use an LSTM model used in [5] to estimate race (Black vs. non-Black) based on first name and last name instead of BIFSG. Our results are attached in Figure 2 of the global response. We show that our method is able to train a classifier that makes predictions which satisfy the target disparity threshold for several different threshold values, unlike the comparison methods which do not satisfy the target threshold except for the largest threshold value.

Additional Benchmarks

Our setting differs substantially from the papers cited by the reviewer and the literature more broadly. In particular, while some papers assume noisy, perturbed, or nonexistent protected attribute data, we consider an empirically common "mixed" case, in which the learner has access to a probabilistic proxy for the protected attribute over all their data, and ground truth over a small subset only, spurring the question of how best to combine the two data sources with fairness and performance in mind.

With respect to the works you specifically highlighted: Mozannar et. al, primarily focuses on a setting with a very strong conditional independence assumption on the protected attribute proxy, which is unlikely to be met besides the (infeasible, for our setting) case in which the learner can differentially privately generate noisy labels using the ground truth. However, as you point out, the authors do propose a potential extension of their method to handle a case that matches ours. The authors do not discuss this in detail or provide the code for this extension, but we have modified their original code to implement the changes described and include these results in our updated draft. We note that Mozannar et. al.'s method almost never meets the disparity bounds, and does worse in accuracy on L2; see Figures 1-2 attached. Levy et. al proposes and analyzes distributionally robust optimization (DRO) algorithms. At its core, the idea of DRO is to optimize for the worst-case distribution that is "nearby" according to some metrics. This has applications to fairness, as the reviewer points out. In particular, Wang et al, to which we have already provided a comparison, applies DRO to precisely the problem of fair classification with noisy sensitive features. Like other methods that aim for worst-case guarantees, the method is robust but sacrifices performance by not using all available information, as can be seen in Figure 2 in our main results.
Diana et. al. aims at an upstream and somewhat orthogonal problem to that of ours. Our paper focuses on training a fair model using an imperfect proxy (e.g. BIFSG), where as Diana et al. study how to learn a model to create proxy features with which to learn a fair model. They do not focus on the downstream task of training the fair model, and the methods are hence not directly comparable. We are happy to include discussion of the above works in the related work section of our paper.

Notation and Event

For examples of the event $\mathcal{E}$, we direct you to appendix A5, where we have a table of functions $f(h(X), y)$ and corresponding epsilons that refer to various common fairness definitions. For example, to enforce equalized false positive rate parity, $f(h(X), y) := 1[h \neq y]$ and $\mathcal{E} = 1[y = 0]$. We are happy to include further examples and a more detailed demonstration of their equivalence in our updated version. We acknowledge that the notation is somewhat unwieldy; however, we do believe that the tracking of the event $\mathcal{E}$ is helpful to illustrate the generality of the methods. We will revisit this notation and experiment with other options (including with the reviewer's proposed modification).

[1] Elzayn, Hadi et al. Measuring and mitigating racial disparities in tax audits. SIEPR, 2023.

[2] Austin, Roy L. Expanding Our Work on Ads Fairness. 21 June 2022.

[3] Agarwal, Alekh et al. "A reductions approach to fair classification." PMLR, 2018.

[4] Awasthi, Pranjal, et al. "Equalized odds postprocessing under imperfect group information." PMLR, 2020.

Fig. 18

Summary:
This paper develops methods for measuring & reducing fairness violations when the protected attribute is private. Access to protected attribute labels for a small subset of the data is assumed, but only probabilistic estimates are available for the remainder of the data. A method for measuring common fairness metrics is proposed for this setting. A theoretical bound on a range of common fairness metrics is given. The utility of the approach is shown on various datasets.

Soundness: 4 excellent
Presentation: 4 excellent
Contribution: 3 good
Strengths:
This paper has numerous strengths. The paper is well motivated, with convincing practical examples of why this data setting and problem is relevant. The paper is also very well written, the notation is clean and I found it easy to follow.

The technical results are strong. Theorem 1 is neat, drawing on a nice simple relationship between the bias in what the paper labels the probabilistic estimator and the linear estimator (or more precisely, their asymptotic limit). The result is then applied to help solve the fair learning problem in a principled way.

The description of the experiments is thorough and precise. The results are well describes and are convincing evidence in favor of the proposed approach. The figures are well designed and easy to understand.

Weaknesses:
There is generally not much to criticize about this paper. However it would be useful if the paper could elaborate on how well this approach works on small sample sizes when there is presumably considerable variability in the probabilistic/linear estimates.

It would also be useful to better understand (either theoretically or via the experiments) how much public data is needed for this approach to give good performance (this is actually mentioned as a limitation/extension).

Some discussion of how computationally demanding this approach is when compared to alternatives could be beneficial.

It would also be valuable to extend this to situations where the protected variable has more than two categories (perhaps this approach still applies, but that was not clear to me). I agree this approach is still valuable in this setting but clearly generalization would be valuable.

Questions:
Why does the empirical problem not account for variability/uncertainty in the estimators?

How robust is the method to incorrectly calibrated probabilities for the protected features? In particular, what happens if there's disparity in these too?

Limitations:
The authors have addressed limitations adequately and noted certain privacy concerns that arise in this setting.

Flag For Ethics Review: No ethics review needed.
Rating: 8: Strong Accept: Technically strong paper, with novel ideas, excellent impact on at least one area, or high-to-excellent impact on multiple areas, with excellent evaluation, resources, and reproducibility, and no unaddressed ethical considerations.
Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Code Of Conduct: Yes

Fig. 19

Rebuttal:
Thank you for your review. We address your questions and concerns below:

Variability and dataset size: Regarding your questions on variability in probabilistic/linear estimates in small sample sizes, and questions about how much labeled subset data is required for good performance, we have attached a study between our method and only training on labeled subset data that explores this question by training with different percentages of labeled data on a synthetic dataset in Figure 3 in the global response PDF.

As we mentioned in limitations, it's not just labeled subset size that impacts how our method may function, but complexity of the dataset as well—as the more complex the data is, the more data will be necessary to successfully bound disparity as well as predict accurately.

To conduct this study, we implement a data-generating process using analogues of the key features of our setting, as well as a tunable complexity parameter that we may vary (or hold fixed) with sample size. Figure 3 shows the true disparity and accuracy obtained by our method (y-axes) on a test set over several different sizes of labeled subset data (from 50-2000 points) (x-axes). The four graphs comprise a range of models on data generated from a 3rd degree polynomial with random coefficients over 10 features (less complex data, top row) and 20 features (more complex data, bottom row), with an overall dataset of 5k (left) and 10k (right). The blue method is labeled subset, the orange method is our method, and the dotted line is the disparity of an unconstrained classifier.

Results: **Figure 3 in the attached PDF of the global response shows that our method consistently meets disparity goals even with small amounts of labeled data in both the low and high complexity regimes.** By contrast, the labeled subset method converges from above towards the desired disparity bound as the size of the labeled set increases; it comes close to meeting the bound in the low complexity regime but does not meet it even with a large amount of data in the complex regime. Thus, this study suggests that our method is relatively more valuable in high complexity regimes. We will add these results and discussion to the paper and discuss the simulation in detail in the appendix. We are happy to expand these experiments in the final version should the paper be accepted.

Computation time: We would be happy to record the timing of the different methods in the appendix pending acceptance. As a rough estimate, the methods take approximately the following amount of time: L2, 150k observations: Ours: ~7.5 minutes Mozannar: ~3 seconds Wang: ~24 minutes

Extension beyond two demographic groups: Our method is designed to measure/mitigate disparity between two demographic groups, but there is no reason why one would not be able to use these bounds in tandem for any non-overlapping groups: in the measurement setting, one could simply compute the disparities across whatever comparisons are of interest, and extra constraints could theoretically be added to the training setting (e.g. to enforce covariance constraints and upper bounds on multiple groups). However, with additional groups, additional data would be needed to correct for multiple comparisons and the increased complexity of the learning problem.

Variability in empirical problem: We understand your concern to refer to why sample variation is not addressed directly in the covariance conditions or the linear estimator in the empirical problem outlined in Problem 2A. We note that the generalization bounds in Theorem 2 provides guarantees that the bounds on all of the relevant terms in the empirical problem— the covariance constraints as well as the bound on the disparity estimator—will translate to bounds on the ground truth data up to an error term which decreases with larger amounts of data. Thus, once there is sufficient data, the instability in the samples of the covariance terms and disparity estimates should not influence the results. We're of course happy to clarify the presentation of this result.

Calibration: Thank you for your question. It is true that a miscalibrated (i.e. systematically biased) probabilistic proxy will result in inaccurate bounds; to the extent this is known, this can be accounted for in the bounds at the cost of some additional algebra (which we can add to the Appendix). But in practice, imperfect calibration does not seem to be a problem if it is not extreme or pathological. We also note that the proxy can also be recalibrated for the specific population of interest, either via flexible machine learned models [2,3] or simple linear regression. This latter technique is particularly compelling given recent work which suggests that additional features tend to make marginal improvements for accuracy after the key factors in BIFSG are taken into account [1].

Thank you for taking the time to read and review our paper, and please let us know if you have any further questions.

[1] Cheng, Lingwei, et al. "How Redundant are Redundant Encodings? Blindness in the Wild and Racial Disparity when Race is Unobserved." FACCT, 2023.

[2] Pleiss, Geoff, et al. "On fairness and calibration." Neurips, 2017.

[3] Elzayn, Hadi, et al. 2023. (16 in original paper)

Fig. 20

Summary:
The paper proposes a technique for measuring and mitigating fairness disparities when most of the true attributes are protected and only probabilistic estimates of protected attribute labels (e.g., via BISG) are accessible. The proposed method takes advantage of contextual information, i.e., the relationship between a model's predictions and the probabilistic prediction of attributes, to provide tighter bounds on the true disparity.

Soundness: 3 good
Presentation: 3 good
Contribution: 2 fair
Strengths:

1. The paper considers a critical and practical problem, i.e., the sensitive attributes may be protected.
2. The proposed method is validated by both theoretical analyses and experiments.
3. The proposed method can give a significantly tighter bound.

Weaknesses:

1. The main concern is that the assumption that $b = Pr[B = 1|Z, \epsilon]$ is too strong. In practice, the probabilistic estimates of attributes are likely to be biased. If b is unbiased, can we just estimate the disparity by letting $\hat{B} = \mathbb{1}_{b>0.5}$? How would this method compare to the proposed method?
2. Should the labeled attributes be iid with the unlabeled ones? In practice, iid is hard to guarantee, and requiring a small set of sensitive attributes may violate privacy regulations, e.g., differential privacy.
3. The compared baselines are limited in Figure 1. There are five more baselines:

- The method in weakness 1.
- Evaluate fairness disparity only with the labeled attributes.
- The method in [R1--R3]

[R1] P. Awasthi et al. Evaluating fairness of machine learning models under uncertain and incomplete information. FAccT, 2021.

[R2] J. Chen et al. Fairness under unawareness: Assessing disparity when protected class is unobserved. FAccT, 2019.

[R3] Z. Zhu et al. Weak proxies are sufficient and preferable for fairness with missing sensitive attributes. ICML, 2023.

Questions:
Here are several questions in addition to the questions mentioned in weakness:

1. What would happen if the true disparity in Figure 1 is large? Would the method still work?
2. In Line 300, does the "labeled subset with true race labels" refer to the set with 1,500 examples? If true, does this setting correspond to the orange curve in Figure 2 (bottom)? If also true, it seems that this setting achieves the best performance, rather than the proposed method.

Limitations:
NA

Flag For Ethics Review: No ethics review needed.
Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Fig. 21

Rebuttal:
We thank the reviewer for their feedback. We first respond to your questions, and then the overall concerns:

Questions:

Larger true disparity: Yes, our method works in settings with larger disparity. In fact, the larger the disparity is, the smaller the variance in the estimators and thus the more precisely they can be estimated with the same amount of data. We additionally highlight Figure 1, in which D_P and D_L do in fact bound the true disparity, despite the relatively high disparity of an unconstrained classifier at 15%. This is further supported by our new COMPAS experiments in which the disparity of an unconstrained classifier is around 26% yet we are able to successfully train our classifier to make predictions with a disparity below 18%.

Clarification of Figure 2 re: labeled subset with true race labels: The orange line does correspond to using a model that enforces the fairness constraint on the set of 1,500 with true race labels. This baseline has higher accuracy, however, our target is not to maximize accuracy alone; it is to maximize accuracy subject to disparity constraints, and the labeled subset method does not satisfy this constraint. The drop in accuracy of our method corresponds to a fairness-accuracy tradeoff: for huge decreases in disparity (e.g. from 15% to below 4% in the upper left-hand graph, for demographic disparity), there is a noticeable difference in accuracy between our method and the labeled subset—however, when we bound demographic disparity at 10% instead of 4% in the same graph, the accuracies of the two methods are comparable while our method is far more successful at reducing disparity.

The method and results speak to the ongoing debate about the extent of the fairness/accuracy tradeoff [1]. While many works have shown that it is possible to reduce disparity by some amount without noticeable accuracy tradeoffs, our work adds to this debate by showing that one cannot reduce disparity by an arbitrary amount—e.g. to a specific threshold—with no repercussions on accuracy. Using the labeled subset alone maintains higher accuracy primarily by failing to decrease disparity to the desired threshold. As we see Figure 2 in the original paper and Figures 1 and 2 in the additional PDF, as we relax the fairness constraints, the accuracy improves with it linearly.

Concerns:

Bias in the probabilistic estimate b:

- We understand that miscalibration of b is a concern, and direct the reviewer to appendix A.4.2 where we discuss this issue. Overall, while miscalibration of b will affect the bounds, the method still works with some miscalibration (indeed, as we see in appendix C.2, the race probabilities are not perfectly calibrated).
- Regarding your suggestion to directly use b thresholded at 0.5, we have included this experiment in Appendix Figure 6 (the bars labeled “Threshold”). As you can see, this method of thresholding the BISG estimates does not ever effectively bound disparity to the desired level. This is consistent with Chen et. al.’s findings that thresholded estimators will under- or over-estimate disparity depending on fundamental parameters of the problems.

I.I.D. Samples: As in many ML settings, our theoretical guarantees require that the labeled subset is drawn from the same underlying distribution as the unlabeled dataset. But, also like many machine learning settings, deviation from this assumption will degrade results smoothly rather than catastrophically. And in settings of interest - e.g. healthcare, tax audits, recidivism prediction, etc. - existing empirical evidence as well as (setting-specific) theoretical arguments often suggest that satisfaction of the covariance conditions is driven by societal-level factors like historical discrimination and socioeconomic differences which are very likely to generalize even if the labeled subset is not perfectly representative. Of course, the gold standard remains a perfectly representative subset. Note that one can also conduct sensitivity analyses to quantify the degree to which the labeled subset must differ in order to change the qualitative impact of measurements or the presumption of improved fairness via our model, but this is beyond the scope of the paper.

Other comparisons for disparity measurement: Chen et. al. analyze the probabilistic estimator (which is in fact well-known in the literature, dating at least as far back as 1953 [2]) and the thresholded estimator. Both estimators are biased (even with a perfectly calibrated probabilistic proxy), which Chen et. al. highlight as an impediment to their usage. But we note that we do incorporate the probabilistic estimator and take advantage of its bias in our framework. As mentioned above, the thresholded estimator does not bound disparity well in our experiments. We will add a discussion as to these points and the citations to our related works section appropriately.

More generally, the disparity estimation methods pointed to in the review are all point estimators. By contrast, our approach recovers upper and lower bounds on the disparity. These approaches are fundamentally different: whereas the bounds approach tries to capture all parameter values that could have generated the data without attaching special significance to any one of them, the point estimation approach tries instead to obtain one parameter. Kallus et. al., to which we do compare, is the only alternative method for disparity estimation we are aware of that also obtains upper and lower bounds.

We are happy to add a discussion on these papers and the differences from our approach to the related work.

[1] Rodolfa, Kit T., et al. "Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy." 2021.

[2] Duncan, Otis Dudley, and Beverly Davis. "An alternative to ecological correlation." 1953.

Fig. 22

Summary:
The paper proposes to regress the label's dependence on the sensitive attribute b assuming access to sensitive labels for a small subset of the data, and using BISG for the rest of the samples. This estimate is then used to bound fairness violations in the form of what authors call probabilistic constraints.

This process decouples the sensitive attribute estimation from the concept learning, supposedly reducing estimation errors of a typical fairness-constrained optimization problem. The method is for voter turnout data. However, despite the additional burden of sensitive attribute regression, the method does not bound worst-case fairness violations.

Soundness: 2 fair
Presentation: 2 fair
Contribution: 2 fair
Strengths:
The two-step training procedure is an interesting approach.

Weaknesses:
- Compared to [16], the novelty of theorem 1 is questionable.
- The empirical section is somewhat sparse. In particular, the method is tested on one dataset. I suspect this is a limitation of the particular data regime the paper adopts where BISG applicability is a concern. Otherwise, the methods and baseline comparisons are sensible. Given the lack of novelty (see previous point), I think the empirical section should have been more substantial.
  In particular, several ablation studies are in order. First, given that BISG estimation comes from a separate data domain, the paper needs an ablation study to show how much the method improves upon just using a normal fairness-constrained optimization problem using BISG labels.
  Yet another ablation study should be on the size of the labeled dataset. In other words, the paper should provide a concrete answer to this observation:
  > We note that the utility of our method is dependent upon the size360 of the subset of the data labeled with the protected attribute—if this subset is relatively large, then361(depending on the complexity of the learning problem) it may be sufficient to train a model using the362 available labeled data. Symmetrically, if the labeled subset is exceedingly small, the enforcement of363 the covariance constraints during training may not generalize to the larger data set
- Despite the additional burden of sensitive attribute regression, the method does not bound worst-case fairness violations, not even asymptotically. I am afraid the generalization bound in Theorem 2 does not alleviate these concerns, as it looks pretty much like a normal generalization bound. What is the extra constraint doing here? I may be missing something, and invite authors to explain their exact contribution here.
- Overall, between the fact that the sensitive attribute regression does not help bound worse-case violations despite the additional hassle (and BISG dependence) and the limited empirical evidence, I am not convinced of the method's general usefulness.
- Some factual issues in the text:
  - > demographic parity101 in classification (6; 35 ; 36) corresponds to letting E be the generically true event and f be simply102 Y = 1 . False positive rate parity (11 ; 12) corresponds to letting E be the event that Y = 0 and the103 function f (ˆY , Y) = 1[ˆY 6 = Y] .
    - Demographic parity does not take into account the ground truth.
  - > well-known form of the regression coefficient,
    - This lacks a reference. It seems the paper assumes L2 loss for regression.
  - Certain less-common terms are not defined before use. For instance, "near-feasibility." The reader should not be left guessing here.
- Criticism but not ground for reduced score:
  - Predicting sensitive attributes is problematic at best and illegal in many contexts where fairness is a concern. I also take issue with the systematic use of sensitive data prediction. We should not be promoting cross-matching of data, beyond the reason they were collected for.

Questions:
- Can you expand on the novelty of your analysis in Theorem 1 compared to prior work [16]?
- Figure 8 (middle), it seems none of the methods can bound FPRD. How can you claim this is near-feasible?

Limitations:
The limitations are well-addressed in the main paper. Maybe authors could expand on it using the collective reviews.

Flag For Ethics Review: No ethics review needed.
Rating: 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Fig. 23

Rebuttal:
We thank the reviewer for their feedback. We note that:

- **We have performed the two ablation studies mentioned which we describe in detail further below**—we point to in Appendix F for the naive comparison of our method to thresholded BIFSG estimates, and the study about dataset size we present in Figure 3 in the PDF attached to the global response.
- **We also include new versions of our false positive rate disparity (FPRD) results in response to your questions on near-feasibility.**

We first provide some clarifications:

- It is true that in our setting, the problem of learning a sensitive attribute proxy is decoupled from the problem of learning the outcome of interest. But our learning problem does depend on the proxy: we learn a fair model by measuring and constraining the relationship between the outcome and the proxy (namely, constraining the covariance conditions $C_{f,b|B}$ and $C_{f,B|b}$ to be the same sign in order to guarantee that the linear estimator D_L will serve as a bound on disparity, and then constraining D_L to the desired upper bound on disparity) in the training process.
- The proxy we use–BIFSG–is used over the entire dataset. The labeled subset is used only to constrain the relationship between the proxy and the outcome of interest.

Worst-case violations and Theorem 2: If we understand the point correctly, the reviewer is concerned about the "worst case" in the sense of "worst case over distributions", i.e. that the method will not work if the covariance conditions are not met. In our technique for training a fair model, it is precisely for this reason that we add the constraints on the covariance terms. Thus, we are limiting our search to models for which we can guarantee that our measurement method works and thus that we can reliably bound disparity; in doing so, we trade off some performance for the security of bounding disparity. Of course, we cannot work with the population covariance terms directly, but instead work with their sample analogues. As in any empirical optimization problem, working with the sample introduces some noise and approximation error. Theorem 2 is useful, because it provides a formal guarantee and quantification of the intuition that (with high probability) these errors will become negligible with enough data and iterations under mild conditions.

Size of the labeled subset: We agree that this is an important question; we have conducted studies of this and display our results in Figure 3 of the attached one-page addition. Due to space constraints, please see our response to reviewer Vf4j for a discussion of this experiment.

Comparing to fairness-constrained optimization with BISG labels: In Appendix F, we provide an ablation study in which we train a fairness-constrained optimization method based on thresholded versions of the BIFSG labels for all experiments in Figure 2 of our original paper. Our study shows that while the thresholded approach has higher accuracy than our method, it consistently fails to control disparity below the specified threshold. We will highlight this more prominently in the text.

Novelty of Theorem 1: Theorem 1 is similar to the result of [16], but Theorem 1 generalizes the result beyond demographic parity to a very broad class of fairness definitions, a generalization that was not obvious from [16]. In any case, we do not view this particular theorem as our primary contribution, but rather as a rigorous justification of and basis for our methods as applied to fairness metrics more generally.

Near-feasibility: We thank the reviewer for pointing out this imprecision. Near feasibility refers to a solution produced which may violate constraints, but that this violation can be made arbitrarily small (in particular, with enough data and training iterations), as described in Theorem 2. Near feasibility occurs widely in settings where constraints are formulated over distributions but only sample data is available; in practice, researchers will either fix a constant below which constraint violation is considered negligible, or fix a number of iterations and dataset size based on data availability and observe the constraint violations. See, e.g. [1]. We will clarify this point in the paper.

Regarding the FPRD results, we provide improved results displayed in Figure 2b of the attached PDF. In our initial presentation, for consistency, we used the same hyperparameters for each of the disparity metric experiments. By tuning for the FPRD problem specifically, we see greatly improved results. We will update the paper to reflect the problem-specific hyperparameter optimization approach.

Relationship between linear estimator and linear regression coefficient: We thank the reviewer for pointing out how our statement could be misinterpreted. We will change the sentence to emphasize its relationship to the ordinary least squares regression coefficient.

Typos: We thank the reviewer for pointing out the typos, which we will fix.

Finally, we share the reviewer's concern that sensitive data be protected. But it is well-established in the literature that fairness cannot be achieved through unawareness; hence, in settings without labeled data, if we don't use proxies to measure and mitigate unfairness, our options are to either remain ignorant about potential unfairness, or use distributionally-robust optimization approaches, which may come at a cost to performance (see discussion in our response to Reviewer ZFXj). In some cases, these solutions may be desirable, but we believe enough high-stakes settings (e.g. [1], [2], and 3]) exist that developing methods to measure and mitigate unfairness based on proxies is worth the potential risks.

[1] Cotter et al, 2019. (13 in original paper)

[2] Elzayn, et al. 2023. (16 in original paper)

[3] Obermeyer, et al. "Dissecting racial bias in an algorithm used to manage the health of populations." Science, 2019.

[4] Executive Order 13985

Fig. 24

Estimating and Implementing Conventional Fairness Metrics With Probabilistic Protected Features

***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 practically important applications this protected attribute is largely unavailable. Still, AI systems used in sensitive business and government applications—such as housing ad delivery and credit underwriting—are increasingly legally required to measure and mitigate their bias. 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 as well as the COMPAS dataset. 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 in comparison to an unconstrained model while often incurring lesser fairness-accuracy trade-offs than other fair optimization methods with limited access to protected attributes.**

***Index Terms*—algorithmic fairness, fair machine learning, anti-discrimination, disparity reduction, probabilistic protected attribute**

I. Introduction

In both the private and public sectors, organizations are facing increasing pressure to ensure they use equitable machine learning systems, whether through legal obligations or social norms [1, 2, 3, 4, 5]. 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 [6, 7]. Similarly, recent Executive Orders in the United States [3, 8] direct government agencies to measure and mitigate disparity resulting from or exacerbated by their programs, including in the "design, develop[ment], acqui[sition], and us[e] [of] artificial intelligence and automated systems" [8].

Yet both companies [9] and government agencies [3] 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 [10, 11], the limited availability of protected attribute data represents a significant challenge in assessing algorithmic fairness and makes 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 the rest of the dataset. These probabilistic estimates may be generated using Bayesian Improved Surname Geocoding (BISG) [12, 13] or any predictive model which can output probabilistic predictions.

We leverage this limited labeled data to establish (or ensure, in the case of training) whether a certain condition holds regarding the relationship between the model's predictions, the probabilistic protected attributes, and the ground truth protected attributes hold. In particular, this condition is that two residual correlations – the residual correlation between the probabilistic proxy and the outcome of interest conditioned on ground truth race, and the residual correlation between ground truth race status and the outcome conditional on the proxy – share the same sign. Given this condition, 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 II.

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 [6, 15, 16, 17, 18]. 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 [18]. 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. [19].

While our approach has stronger data requirements than recent work in similar domains [20, 21] 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 II);
2) We introduce a new method of training models with near-optimal and near-feasible bounded unfairness with limited protected attribute data (Section III);
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 [22], and also demonstrate our results on COMPAS data [23], a common dataset used in related work (Section IV) In addition, we present some experiments on synthetic data which outline the conditions under which our technique is the most effective: relatively complex problems with little access to labeled data.

The rest of this paper proceeds as follows: in the remainder of this section (Section I-A), we describe in greater detail two examples of real-world settings in which our approach may be applicable. Following this, in Section II, we describe our method of measuring disparities in data regimes with limited access to protected attribute labels, then in Section III we leverage our measurement results to develop our training techniques which bound unfairness in the resulting model. We display our experimental evaluation of our method in Section IV, including comparisons to related bias measurement [20] and fair training techniques [21, 24]. Finally, we end our paper with our review of the related work (Section V) and Conclusion (Section VI).

A. Correspondence to real-world Settings

We now highlight two real-world examples which correspond to our setting. First, consider the example of Meta Platforms ("Meta"). Meta is the parent company of Facebook, a social media platform with a large advertising business. Meta uses machine learning to identify users likely to interact with particular ads [25]. The Department of Housing and Urban Development brought a lawsuit [26] under the Fair Housing Act alleging algorithmic discrimination by Meta. As part of a settlement resolving the suit [7], Meta agreed to build software called the *Variance Reduction System* (VRS) [6] which uses a differentially-private version of BISG to estimate deviation of delivery rates by group relative to an underlying eligible audience [27]. In accordance with the recommendations of civil rights groups [28], Meta also began to work with third-party survey administrator YouGov to prompt users to provide individual race off-platform (with privacy protection via tools from secure multi-party computation) [18, 29].

Second, consider the example of government agencies such as the Internal Revenue Service (IRS). IRS, like many other government agencies, does not collect taxpayer data on race [30], yet recent executive orders have required equity (disparity) assessments [3] and consideration of protections from "algorithmic discrimination" [8]. A paper by academic and government researchers, [19], combines BISG for the population of taxpayers with a publicly available administrative dataset (voter registration data) that does contain ground truth and can be matched to a subset of taxpayers and uses this combined dataset to assess audit rate disparity.

In both these examples, disparity estimation is an important goal hindered by a lack of individual-race data, yet probabilistic estimates of race via BISG are available and race data can be obtained for small subset of individuals. The key features thus correspond to the setting we describe formally in Section II-A. These prominent examples are likely representative of scenarios faced by many other private and public sector actors; hence, our approach is likely to be broadly useful. Indeed, while these instances are some of the first legally required investigations of disparities arising from algorithmic systems [31], they are unlikely to be the last: along with recent executive orders [8, 32] and the Blueprint for an AI Bill of Rights [4], a recent advanced notice of proposed rulemaking (ANPR) from the Federal Trade Comission (FTC) suggests the possibility of stricter rules around the deployment of discriminatory systems [33]. Increased regulation of algorithmic decision systems requires the development of bias measurement and mitigation techniques which correspond to the realities of data access, and legal scrutiny, that exist on the ground.

II. 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.

A. 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. the *unlabeled set*, $\mathscr{D}_U$, consists of observations $\{(X_i, Y_i, Z_i)\}_{i=1}^{n_U}$ and the *labeled set*, $\mathscr{D}_L$, 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 $b_i := \Pr[B_i|Z_i]$. All three datasets are assumed to be independent and drawn from the same underlying population. 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 [34, 35, 36] 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 [37, 38] 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 [39] (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 Table II-A. There are other metrics that cannot be written in this form; we do not consider those here.

Metric	$\mathbf{f}(\mathbf{h}(\mathbf{X}), \mathbf{Y})$	$\mathcal{E}$
Accuracy	$\mathbf{1}[h \neq y]$	$\{\text{true}\}$
Demographic Parity	$\mathbf{1}[h = 1]$	$\{\text{true}\}$
True Positive Rate Parity	$\mathbf{1}[h \neq y]$	$\{y = 1\}$
False Positive Rate Parity	$\mathbf{1}[h \neq y]$	$\{y = 0\}$
True Negative Rate Parity	$\mathbf{1}[h \neq y]$	$\{y = 0\}$
False Negative Rate Parity	$\mathbf{1}[h \neq y]$	$\{y = 1\}$

TABLE I: Many fairness metrics can be written in the form required by our formulation. For concreteness, we provide a table based on [40, 41] summarizing the choice of f and $\mathcal{E}$ that correspond to the many of the most prominent definitions that can be written in our formulation.

B. 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

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

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. [42], 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:

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

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

We define D_{μ}^{P} and D_{μ}^{L} to be the asymptotes of the probabilistic and linear estimators, respectively, as the identically and independently distributed sample grows large.

C. 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_{μ}^{P} as the asymptotic limit of the probabilistic disparity estimator, $\widehat{D}_{\mu}^{P}$, and D_{μ}^{L} as the asymptotic limit of the linear disparity estimator, $\widehat{D}_{\mu}^{L}$. Then:

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

and

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

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_{μ}^{L} serves as an upper bound and D_{μ}^{P} serves as a lower bound; if they are both negative, then D_{μ}^{P} serves as an upper bound and D_{μ}^{L} 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 $\mathbb{E}[\mathrm{Cov}(f(\hat{Y}, Y), b|B, \mathcal{E})] > 0$ and $\mathbb{E}[\mathrm{Cov}(f(\hat{Y}, Y), B|b, \mathcal{E}] > 0$. Then,

$$D_{\mu}^{P} \leq D_{\mu} \leq D_{\mu}^{L}.$$

Proposition 1 and Theorem 1, which we prove in Appendix A, subsume and generalize a result from [19]. These results define the conditions under which D_{μ}^{L} and D_{μ}^{P} serve as bounds on ground truth fairness violations; since we can use $\widehat{D}_{\mu}^{P}$ and $\widehat{D}_{\mu}^{L}$ to estimate these quantities from data (up to sampling uncertainty[1]) Theorem 1 thus provides a path to bound fairness metrics as long as the assumed conditions hold. We demonstrate the efficacy of this method for measuring fairness metrics of existing models in practice in Section IV-B. 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.

[1]We show how to compute these standard errors in Appendix A-C, and then take the extremes of the confidence intervals as our bounds.

III. 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 ($\widehat{D}_{\mu}^{L}$), 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 $\widehat{D}_{\mu}^{L}$ 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 $D_{\mu} \leq D_{\mu}^{L}$; in the second, both are negative, and $D_{\mu}^{L} \leq D_{\mu}$[2]. Minimizing risk while satisfying these constraints in each case separately gives the following two problems:

Problem 1.A.

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

[2]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.

Problem 1.B.

$$\begin{aligned}\min_{h\in\mathcal{H}} \mathbb{E}[L(h(X),Y)]\\ \text{s.t. } -\alpha \le D_\mu^L\\ \mathbb{E}[C_{f,B|b,\mathcal{E}}] \le 0\\ \mathbb{E}[C_{f,b|B,\mathcal{E}}] \le 0\end{aligned}$$

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

$$h^* \in \text{argmin}_{h_{1a}^*, h_{1b}^*} \mathbb{E}[L(h(X),Y)],$$

where h_{1a}^*, h_{1b}^* are the solutions to Problems 1.A and 1.B. By construction, h^* is feasible, and so satisfies $|D_\mu(h^*)| \le \alpha$; moreover, while h^* may not be the lowest-loss predictor such that $|D_\mu| \le \alpha$, 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.

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

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 [14, 43] have shown that under relatively mild conditions, a primal-dual learning algorithm can be used to obtain approximate solutions with good performance guarantees.[3] In particular, if we define the *empirical Lagrangian* as:

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

(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], [44], etc. specialized to our setting; adapting and applying Theorem 3 in [14], provides the following guarantee:

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 $\ell \in \{D_\mu^L, C_{f,b|B,\mathcal{E}}, C_{f,B|b,\mathcal{E}}\}$,

$$\mathbb{E}[\ell(h_{\tilde{\theta}})] \le c_i + \mathcal{O}\left(\frac{d\log N}{\sqrt{N}}\right) + \mathcal{O}\left(\frac{1}{T}\right)$$

and

$$\mathbb{E}[L(h_{\tilde{\theta}},y)] \le P^* + \mathcal{O}\left(\frac{d\log N}{\sqrt{N}}\right)$$

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:

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

and

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

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

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

[3] 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.

feasible iterate of their model, and often find feasible good performance [21, 44]; thus in our results section, we compare our best-iterate performance to other methods.

IV. Empirical Evaluation

We now turn to experiments of our disparity measurement and fairness enforcing training methods on predicting voter turnout as well as on the COMPAS dataset [45]. In addition, we provide experiments on simulated data in order to outline the conditions under which our method is most successful, and in particular, outperforms relying on training a model with the labeled subset alone, which we expand upon in Appendix G.

A. Data

We perform experiments over two datasets: the L2 dataset [46] and the COMPAS dataset [23]. In both of these datasets, the demographic attribute to which we pay attention is race.

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 [13], and much of voting rights law hinges on determining whether there exists racially polarized voting and/or racial disparities in turnout [47]. 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-A for a detailed description of this dataset. We select seven features as predictors in our model based on data completeness and predictive value: gender, age, estimated household income, estimated area median household income, estimated home value, area median education, and estimated area median housing value. Information on our selection process, pre-processing, and distribution of these features are presented in Appendix Section C-A. We denote $\hat{Y} = 1$ if a voter shows up to vote for the 2016 election and $\hat{Y} = 0$ otherwise. The baseline rates of voter turnout range between 52-63% across the six states (see more information in Section C-A in the Appendix).

L2 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 [12, 13, 48]. 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-B.

COMPAS Dataset. We also evaluate our measurement and training methods on models trained on the COMPAS [45] dataset. The COMPAS algorithm is used by parole officers and judges across the United States to determine a criminal's risk of recidivism, or re-committing the same crime. In 2016, ProPublica released a seminal article [45] detailing how the algorithm is systematically biased against Black defendants. The dataset used to train the algorithm has since been widely used as benchmarks in the fair machine learning literature. We use the eight features used in previous analyses of the dataset as predictors in our model: the decile of the COMPAS score, the decile of the predicted COMPAS score, the number of prior crimes committed, the number of days before screening arrest, the number of days spent in jail, an indicator for whether the crime committed was a felony, age split into categories, and the score in categorical form. Further information about our preparation of the COMPAS dataset can be found in Section F of the Appendix.

COMPAS Race Probabilities. In the COMPAS dataset, we generate estimates of race (Black vs. non-Black) based on first name and last name using a LSTM model used in Zhu et al. [49] that was trained on voter rolls from Florida. Accuracy of these models is 73% while the AUC is 86%. Further detail can be found in Appendix F.

B. Fairness 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 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).

1) Experimental Design: 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 data: in the case of L2, this is state by state. Then, in order to simulate realistic data access conditions, we measure fairness violations on a random subsample of the test set, with a percentage of this sample including ground truth race labels to constitute the labeled subset which we use to calculate the covariance constraints. In the case of the L2 data, the random subsample over which we measure fairness violations has $n = 150,000$, with 1% ($n = 1,500$) of this sample including ground truth race labels to constitute the labeled subset. In the case of the COMPAS dataset, which is much smaller, we use the entire test set, with $n = 1,226$, and we construct the labeled subset by sampling 50% of the test set ($n = 613$).

We first check the covariance constraints on the labeled subset, and then calculate $\widehat{D}_L$ and $\widehat{D}_P$ on the entire set of 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 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-A. We present our results in Figure 1, which shows the results

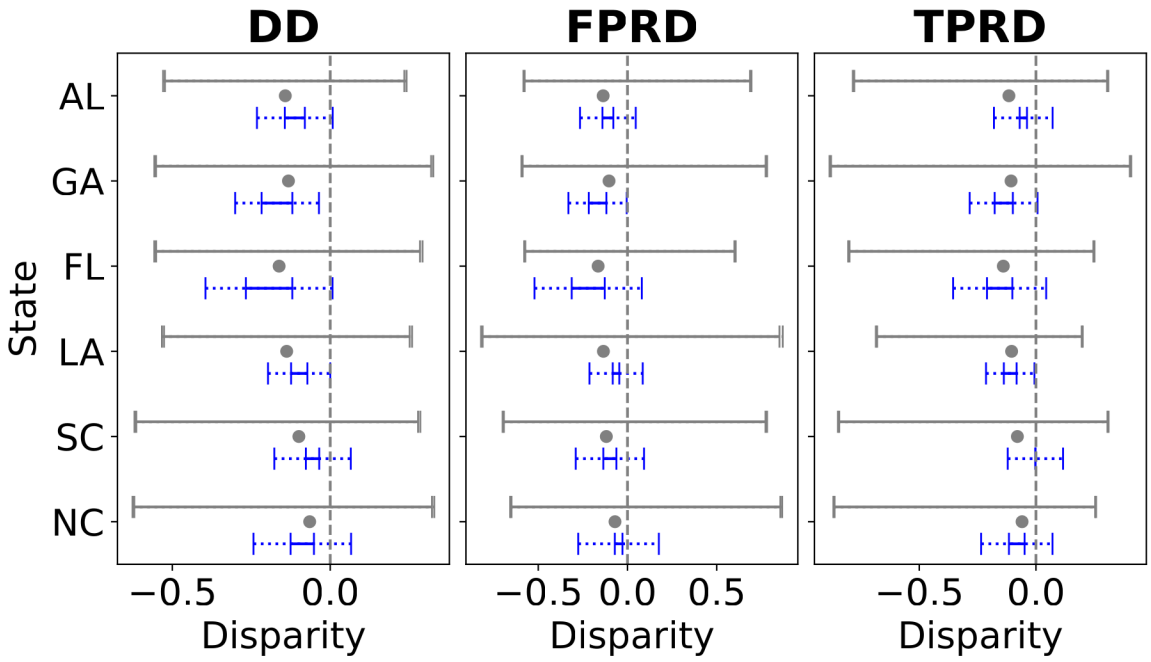


Fig. 1: **Bounding Disparity in L2 Data:** Comparison of our method of bounding true disparity (blue) to the method proposed in Kallus et al. [20] (grey), using a logistic regression model to predict voter turnout in six states. We compare results across three disparity measures: demographic disparity (DD), false positive rate disp. (FPRD), and true positive rate disp. (TPRD). 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. The dashed lines represent 95% confidence intervals. Both methods successfully bound true disparity within its 95% standard errors, but our estimators provide much tighter bounds.

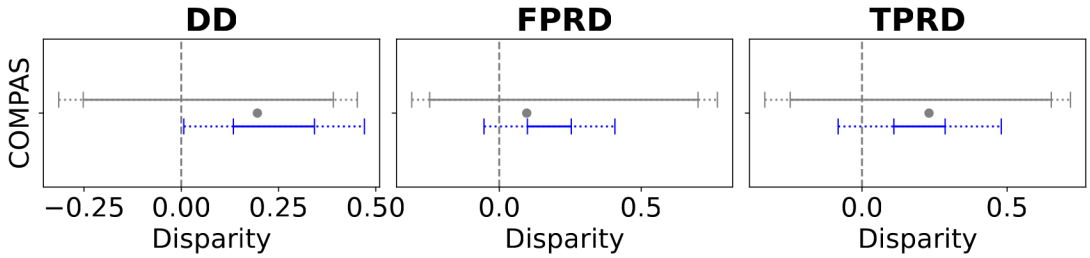


Fig. 2: **Bounding Disparity in COMPAS Data:** Comparison of our method of bounding true disparity (blue) to the method proposed in Kallus et al. [20] (grey), using a logistic regression model to predict two-year recidivism on the COMPAS dataset. We access disparity over the same measures as in Figure 1. The grey dot represents true disparity. The dashed lines represent 95% confidence intervals. Both methods always bound true disparity within the 95% standard errors, but our method provides tighter bounds.

over the L2 data, and Figure 2, which shows the results over the COMPAS data.

2) Comparisons: We compare our method of estimating fairness violations using probabilistic protected characteristic labels to the method described in Kallus et al. [20], 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-B.

3) Results: We first analyze our results on voter data. Figure 1 compares our method of estimating disparity (blue) with KDC (grey) for the three disparity measures on the six states we consider. This figure shows estimates when training a logistic regression model, and Figure 8 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.

We now turn to the COMPAS data. Similar to the L2 data, our bounds are consistently tighter than KDC, albeit to a lesser extent in this case since the COMPAS dataset is significantly smaller (1.69x on average, and up to 2.04x smaller). We emphasize that, unlike KDC, our estimators are always within the same sign as the true disparity, barring the standard errors which shrink as the data grows larger.

C. Fairness-constrained Training

In this section, we demonstrate the efficacy of our approach to training fairness-constrained machine learning models. Following our algorithm in Section III, 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. We also demonstrate via our simulation study that there exist regimes in which our approach meets the goal of keeping disparity below the desired threshold whereas training on the small labeled subset alone does not.

1) 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, on both the L2 dataset and the COMPAS dataset. We use logistic regression as a proof-of-concept, but because our method builds upon the algorithm proposed in [14], it can be extended to any gradient-based machine learning method, including e.g. neural networks. Within the L2 dataset, we train these models on the data from Florida, 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 ($n = 150,000$ for L2 data, $n = 6,128$ for COMPAS data) 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 ($n = 1,500$) for the L2 data, and 10% of the total data for the COMPAS dataset, since it is much smaller (around $n = 613$). To enforce fairness constraints during training, we solve the empirical problem 3A and its symmetric analogue, which enforces negative covariance conditions and $\widehat{D}^L_\mu$ 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 $\widehat{D}^L_\mu$ during training. As noted in Section III, our method theoretically guarantees a near-optimal, near-feasible solution *on average* over $\theta^{(1)}...\theta^{(T)}$. However, following Wang et al. [21], for each of these sub-problems, we select the best iterate $\theta^{(t)}$ which satisfies the bound on $\widehat{D}^L_\mu$ 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 for our method as well as three comparisons, described below, over L2 data and COMPAS data, in Figure 3 and Figure 4 respectively. We note that the resulting disparities for the unconstrained model differ among the three fairness metrics. On DD and TPRD, the unconstrained model resulted in a 0.28-0.29 disparity, but it drops to 0.21 for FPRD. We adjusted our target fairness bounds accordingly. Further details about the experimental setup can be found in Appendix Section E-A. Our experimental design for our experiments on synthetic data differ, and we outline our setup and results in Section IV-D.

2) Comparisons: We compare our results for enforcing fairness constraints with probabilistic protected attribute labels to the following methods:

(a) 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-B.

(b) We compare with a recent method by Wang et al. [21] 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. [20]'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-C.

(c) We compare with a method for enforcing fairness with incomplete demographic labels introduced by Mozannar et al. [24], which essentially modifies Agarwal et al. [50]'s fair training approach to optimize accuracy on the entire available data, but 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-D.

In Appendix Section E-F, 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 $(0, 1)$.

3) Results: We first analyze our results on the L2 data. We display our results in Figure 3. 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 3), than using Wang et al. [21] (blue versus green bars in Figure 3), and than using Mozannar er

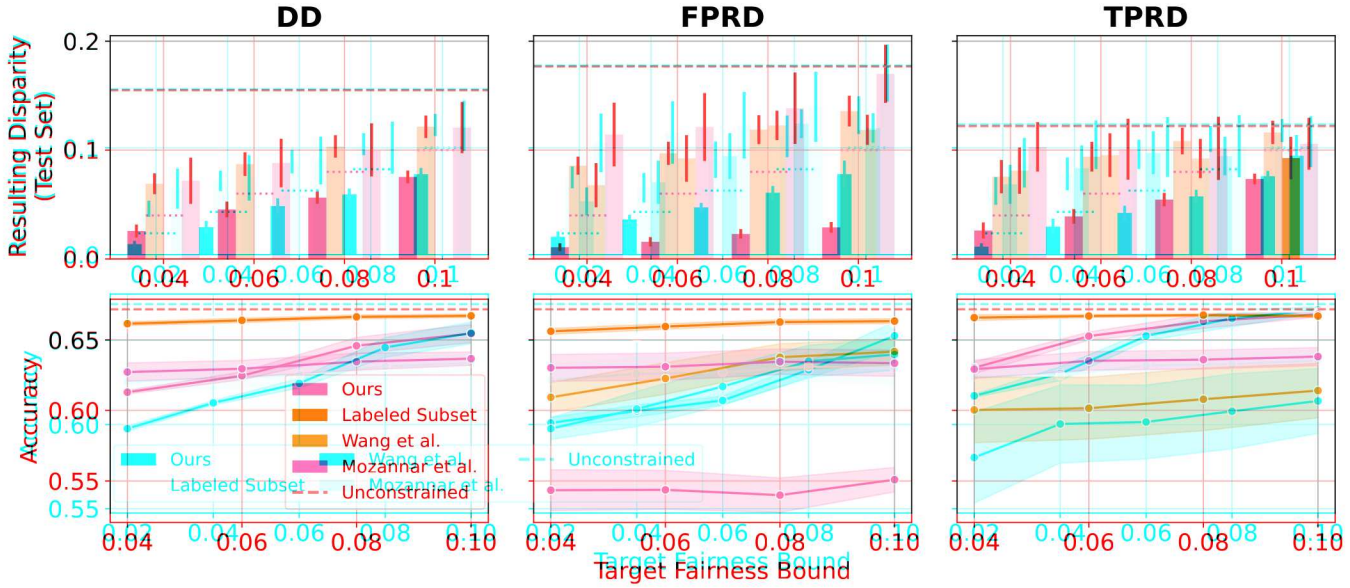


Fig. 3: Mean and standard deviation of resulting disparity (top, y-axis) and accuracy (bottom, y-axis) on the test set after enforcing the target fairness bounds (x-axis) on our method (blue); only using the labeled subset with true labels (orange) and Wang et al. [21] (green) over ten trials. On the 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.

al. [24] (blue versus pink bars in Figure 3). 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 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 more than Mozannar et al. [24] and Wang et al. [21], 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.

Next, we turn to our results on the COMPAS [45] dataset in Figure 4, which is set up identically to Figure 3, with disparity resuts on top and accuracy results on the bottom. We see that our method again is able to consistently meet the desired disparity bound across all experiments, as opposed to the Mozannar et al. method (red) or the labeled subset method (orange), which only meet the constraint 3 out of 12 times each. While the Wang et al. method does meet the disparity bound at each experiment where the comparison is possible (i.e., excluding DD), in the case of FPRD, there is a steep accuracy cost. In the case of DD, our method has worse accuracy bounds likely due to actually meeting the disparity bounds (the accuracy is comparable in the experiment where all three method reach the DD constraint, i.e. 0.24). In TPDR an FPRD, our method performs largely comparably to the other methods, with the exception of the low accuracy of Wang et al. in FPRD.

D. Simulation Study

We note that the utility of our method is often dependent upon the size of the subset of the data labeled with the protected attribute—if this subset is relatively large, then (depending on the complexity of the learning problem) it may be sufficient to train a model using the available labeled data. Symmetrically, if the labeled subset is exceedingly small, the enforcement of the covariance constraints during training may not generalize to the larger dataset. To characterize the regimes under which our method may be likely to perform well relative to others, we empirically study simulations that capture the essence of the situation. We study the utility of our method in comparison to only relying on the labeled subset to train a model along two axes: *1)* size of the labeled subset and *2)* data complexity, which we simulate by adjusting the number of features. While stylized, our simulation has the advantage that we can vary key features of the setting like the dimensionality and distribution

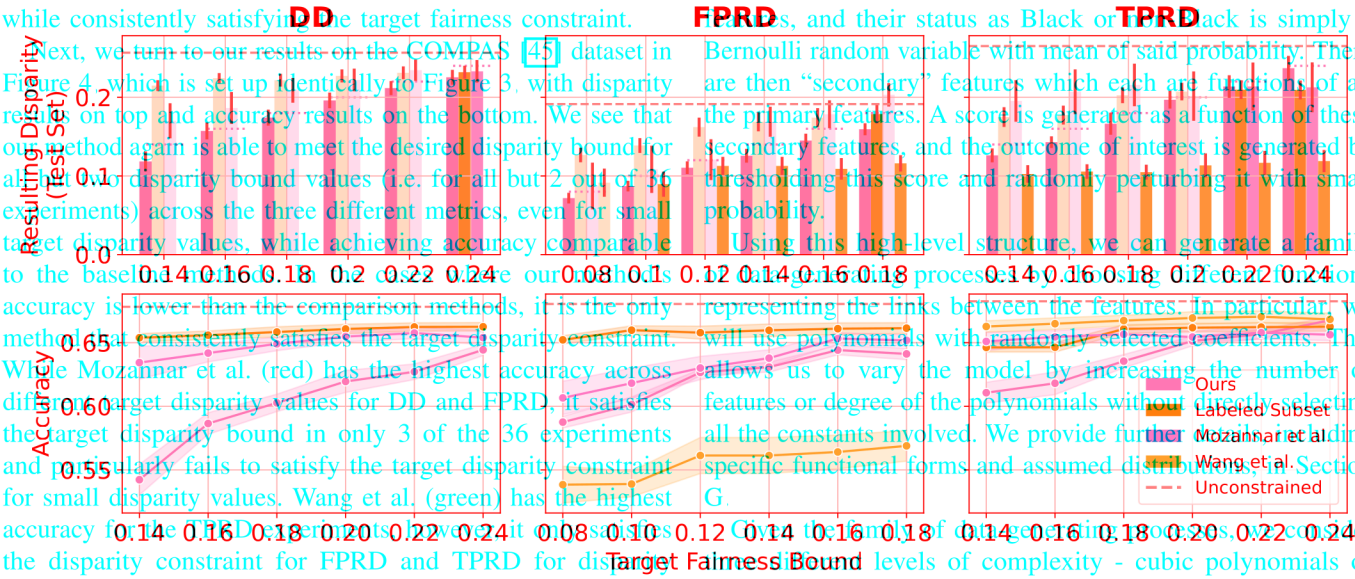


Fig. 4: Mean and standard deviation of resulting disparity (top, y-axis) and accuracy (bottom, y-axis) on the test set after enforcing the target fairness bounds (x-axis) on our method (blue), Wang et al.'s method (green), Mozannar et al.'s method (red) and only using the labeled subset with true labels (orange). 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.

We note that the utility of our method is often dependent upon the size of the subset of the data labeled with the protected attribute, the size of the labeled and unlabeled datasets, the complexity of the relationship between the features and the outcome, and so on. To be useful, however, we must be able to ensure that the key conditions of our method are met by the data-generating process.

To ensure this while also allowing for the tuneability and flexibility we require, we settle on a hierarchical model specified by parameterized components that are individually simple but can serve as building blocks; mapping out these relationships via the language of causal diagrams gives us intuition about the conditional covariance terms. See Appendix G including Figure 15 for visualization and further discussion.

At a high-level, the model can be described as follows. Individuals have a set of "primary" features denoted which are drawn randomly from some distribution. The probability that they the individual is Black is a function of these primary features, and their status as Black or non-Black is simply a Bernoulli random variable with mean of said probability. There are then "secondary" features which each are functions of all the primary features. A score is generated as a function of these secondary features, and the outcome of interest is generated by thresholding this score and randomly perturbing it with small probability.

Using this high-level structure, we can generate a family of data-generating processes by choosing different functions representing the links between the features. In particular, we will use polynomials with randomly selected coefficients. This allows us to vary the model by increasing the number of features or degree of the polynomials without directly selecting all the constants involved. We provide further details, including specific functional forms and assumed distributions, in Section G.

Given the family of data-generating processes, we consider three different levels of complexity - cubic polynomials of 10, 20, or 50 features - and draw datasets of 5,000, 10,000, or 50,000 observations. Of these, we vary the percentage with labels revealed to the learner ranging from 0.5 to 40%, depending on the size of the dataset. We then compare our method to simply training on a fair model on the true labels of the labeled subset. Figure 5 shows disparity for both methods across the scenarios. Overall, we find that there exists a regime, even in simple problems, where there is insufficient data for the labeled subset to effectively bound disparity to the desired threshold. We find that the more complex the data is, the larger this regime is—with the most complex setting in our simulations (50 features) suggesting that the labeled subset technique does not converge to the desired disparity bounds even when the size of the labeled subset is 10,000 samples, or 20% of the overall dataset.

V. Related Work

Kallus et al. [20] propose a method for measuring fairness violations in data with limited access to protected attribute labels. Their method involves finding the tightest possible set of true disparity given probabilistic protected attributes. An important difference between Kallus et al. and our method relates to their assumptions around the auxiliary dataset. The core difference is that Kallus et al. considers settings where the auxiliary and test sets are independent data sets while our method considers the case where the test set subsumes the

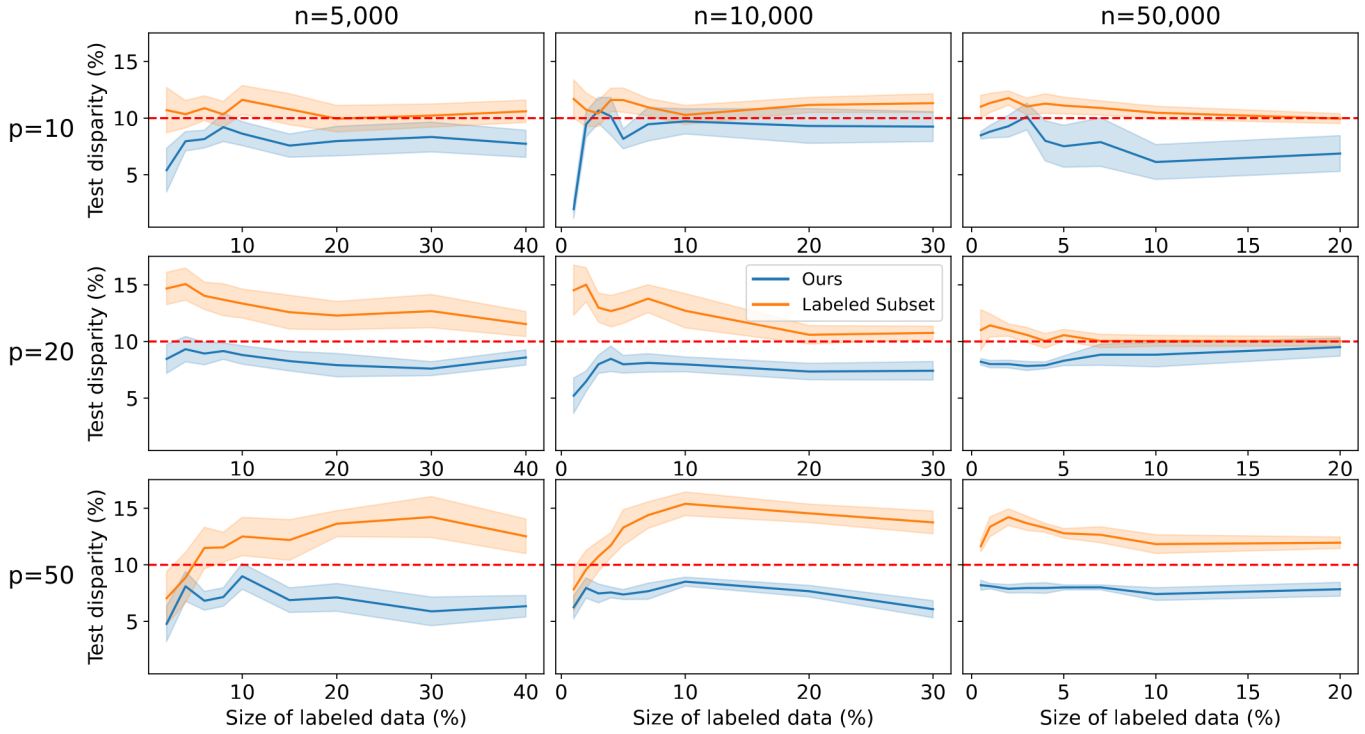


Fig. 5: We present three by three figure showing the test disparity of the our disparity reduction method when compared with relying on only the labeled subset to reduce disparity by directly enforcing a constraint on the protected attribute labels. The rows correspond to datasets of increasing sizes (number of features from 10 to 50), indicating problems of increasing complexity. The columns correspond to the size of the overall dataset, ranging from 5,000 to 50,000 samples. The x-axis shows the percentage of the total dataset is decicated to the labeled subset, and the y-axis denotes the percentage disparity between the two groups calculated on the test set. The blue graphs correspond to our method, and the orange to the labeled subset method. The red dashed line is the desired disparity bound.

auxiliary data. We explain this difference in further detail in Appendix D-B.

With regards to bias mitigation, while there are many methods available for training models with bounded fairness violations [11, 39, 50], 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 [21], and *no* protected attribute labels [51], or a even a labeled subset of protected attribute labels, but without an auxiliary set to generate probabilistic protected attribute estimates [52]; very few works mirror our data access setting. One exception, from which we draw inspiration, is Elzayn et al. [19]; 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.

Another exception, which we compare to in Section IV-C, is that of Mozannar et al. [24]. While Mozannar et. al largely focus on the problem of training *private* fair models, thus employing very strong conditional independence assumptions on the protected attribute proxy which are infesible in our setting, the authors do propose an extension of their method to handle a the case of limited protected attributes without considering privacy, which mirrors our data access assumptions. This extension is essentially a re-purposing of Agarwal et al. [50] fair training approach, modified such that the model is trained with all available data, but the fairness bounds are only enforced during training on the small subset of training points with protected attribute labels. It is this extension that we compare to in which we compare to in Section IV-C, and find that our method often outperforms theirs on disparity reduction and performs comparably on accuracy.

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 [53, 54, 55]. Intuitively, leveraging some labeled data can allow us to have less of an accuracy trade-off when training

VI. Discussion

fair models, as demonstrated with our comparison to Wang et al. [21]. 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.

VI. Discussion

In this work, we introduce a technique for measuring and reducing fairness violations in a setting with limited access to protected attribute data by leveraging probabilistic proxies (e.g. based on name and geolocation). These techniques may help private and public actors better measure algorithmic disparity and fulfill legal and moral obligations to ensure that algorithmic decision-making does not disparately impact disadvantaged or protected groups. However, the collection and use of protected attribute information is inherently sensitive and brings up privacy concerns. Additionally, building a probabilistic model to estimate protected attributes raises important ethical and practical questions as well, such as who has access to these models and what are the protocols for its responsible deployment. Moreover, the approach requires committing to a particular notion of groups to measure and mitigate fairness with respect to, an exercise which in itself can be fraught. Given the increasing stakes of algorithmic deployment as well as increasing regulatory and public pressure, we believe that the benefit of being able to more effectively measure and reduce unfairness in model predictions outweighs these risks, but practitioners applying our method must carefully consider these concerns in the wider context in which they work.

We note several avenues for future work. First, while our framework can be applied iteratively to handle multiple sensitive groups, generalizing our framework to account for them directly, and additionally to handle intersectional groups, would be preferable. Second, while binary classification is perhaps the most common task in machine learning, handling more general tasks, such as multi-label classification or regression, would extend the applicability of results. Finally, in the proposed method it is important that the probabilistic predictions are representative of the population of interest; in practice this means either assuming that the dataset from which probabilistic predictions are learned is drawn from the same population, or that reweighting techniques can be used to construct a representative sample. In the future, it would be useful to use techniques from sensitivity analysis to bound the impact of selection bias on measurement error and robust learning to train low-disparity models under worst-case selection bias.

References

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

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

[3] 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.

[4] T. W. House, "Blueprint for an ai bill of rights: Making automated systems work for the american people," 2022.

[5] K. Hill, "Wrongfully accused by an algorithm," *The New York Times, June*, vol. 24, 2020.

[6] R. L. Austin, Jr., "Expanding our work on ads fairness," https://about.fb.com/news/2022/06/expanding-our-work-on-ads-fairness/ 2022.

[7] M. Isaac, "Meta agrees to alter ad technology in settlement with u.s." 2022. [Online]. Available: https://www.nytimes.com/2022/06/21/technology/meta-ad-targeting-settlement.html

[8] U.S. E.O. 14091, "Exec. order no. 14091 88 fed. reg. 10825, Further Advancing Racial Equity and Support for Underserved Communities Through The Federal Government," 2023.

[9] M. Andrus, E. Spitzer, J. Brown, and A. Xiang, "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.

[10] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft, Tech. Rep. MSR-TR-2020-32, May 2020. [Online]. Available: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[11] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *CoRR*, vol. abs/1810.01943, 2018. [Online]. Available: http://arxiv.bs/1810.0194

[12] M. N. Elliott, P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, and N. Lurie, "Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities," *Health Services and Outcomes Research Methodology*, vol. 9, pp. 69–83, 2009.

[13] K. Imai and K. Khanna, "Improving ecological inference by predicting individual ethnicity from voter registration records," *Political Analysis*, vol. 24, no. 2, pp. 263–272, 2016.

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

[15] C. F. P. Bureau, "Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment," 2014. [Online]. Available: https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

[16] K. Fiscella and A. M. Fremont, "Use of geocoding and surname analysis to estimate race and ethnicity," *Health services research*, vol. 41, no. 4p1, pp. 1482–1500, 2006.

[17] H. K. Koh, G. Graham, and S. A. Glied, "Reducing racial and ethnic disparities: the action plan from the department of health and human services," *Health affairs*, vol. 30, no. 10, pp. 1822–1829, 2011.

[18] R. L. Austin, Jr., "Race data measurement and meta's commitment to fair and inclusive products," https://about.fb.com/news/2021/11/inclusive-products-through-race-data-measurement/ 2022.

[19] H. Elzayn, E. Smith, T. Hertz, A. Ramesh, R. Fisher, D. Ho, and J. Goldin, "Measuring and mitigating racial disparities in tax audits," *Working Paper*, 2023.

[20] N. Kallus, X. Mao, and A. Zhou, "Assessing algorithmic fairness with unobserved protected class using data combination," *Management Science*, vol. 68, no. 3, pp. 1959–1981, 2022.

[21] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. Jordan, "Robust optimization for fairness with noisy protected groups," *Advances in neural information processing systems*, vol. 33, pp. 5190–5203, 2020.

[22] U.S. DOJ, "Statutes enforced by the voting section," 2023. [Online]. Available: https://www.justice.gov/crt/statutes-enforced-voting-section

[23] Equivant, "Practitioner's guide to COMPAS core," http://quivant.com/wp-content/uploas/Practitioners-Guide-to-COMPAS-Core-040419.pdf, 2019.

[24] H. Mozannar, M. Ohannessian, and N. Srebro, "Fair learning with private demographic data," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7066–7075.

[25] "Good questions, real answers: How does facebook use machine learning to deliver ads?" 2020. [Online]. Available: https://www.facebook.com/business/news/good-questions-real-answers-how-does-facebook-use-machine-learning-to-deliver-ads

[26] "United states of america, plaintiff, v. meta platforms, inc., f/k/a facebook, inc." 2022.

[27] A. S. Timmaraju, M. Mashayekhi, M. Chen, Q. Zeng, Q. Fettes, W. Cheung, Y. Xiao, M. R. Kannadasan, P. Tripathi, S. Gahagan *et al.*, "Towards fairness in personalized ads using impression variance aware reinforcement learning," *arXiv preprint arXiv:2306.03293*, 2023.

[28] "Facebookś Civil Rights Audit – Final Report," 2020.

[29] R. Alao, M. Bogen, J. Miao, I. Mironov, and J. Tannen, "How meta is working to assess fairness in relation to race in the u.s. across its products and systems," 2021.

[30] D. I. Werfel, "Werfel letter on audit selection," 2023.

[31] D. of Justice, "Justice department secures groundbreaking settlement agreement with meta platforms, formerly known as facebook, to resolve allegations of discriminatory advertising," 2022. [Online]. Available: https://www.justice.gov/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known

[32] U.S. Executive Order 13985, "Exec. order no. 13985 86 fed. reg. 7009, executive order on the safe, secure, and trustworthy development and use of artificial intelligence," 2021.

[33] F. T. Commission, "Trade regulation rule on commercial surveillance and data security," Proposed 08/22/2022.

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

[35] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 962–970.

[36] I. Žliobaitė, "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*, vol. 15, 2015.

[37] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

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

[39] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016.

[40] A. Narayanan, "Translation tutorial: 21 fairness definitions and their politics," in *Proc. conf. fairness accountability transp., new york, usa*, vol. 1170, 2018, p. 3.

[41] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018, pp. 1–7.

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

[43] L. Chamon and A. Ribeiro, "Probably approximately correct constrained learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 722–16 735, 2020.

[44] A. Cotter, H. Jiang, M. R. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan, "Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals." *J. Mach. Learn. Res.*, vol. 20, no. 172, pp. 1–59, 2019.

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

[46] "L2 voter and demographic dataset," 2000. [Online]. Available: https://purl.stanford.edu/pf141bs6760

[47] M. Barber and J. B. Holbein, "400 million voting records show profound racial and geographic disparities in voter turnout in the united states," *Plos one*, vol. 17, no. 6, p. e0268134, 2022.

[48] M. N. Elliott, A. Fremont, P. A. Morrison, P. Pantoja, and N. Lurie, "A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity," *Health services research*, vol. 43, no. 5p1, pp. 1722–1736, 2008.

[49] Z. Zhu, Y. Yao, J. Sun, H. Li, and Y. Liu, "Weak proxies are sufficient and preferable for fairness with missing sensitive attributes," in *Proceedings of Machine Learning Research*, 2023.

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

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

[52] S. Jung, S. Chun, and T. Moon, "Learning fair classifiers with partially annotated group labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 348–10 357.

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

[54] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 702–712.

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

[56] C. R. Shalizi, "The truth about linear regression," *Online Manuscript. http://www. stat. cmu. edu/˜ cshalizi/TALR*, 2015.

[57] I. Voicu, "Using first name information to improve race and ethnicity classification," *Statistics and Public Policy*, vol. 5, no. 1, pp. 1–13, 2018.

[58] Y. Zhang, "Assessing fair lending risks using race/ethnicity proxies," *Management Science*, vol. 64, no. 1, pp. 178–197, 2018.

[59] J. G. Matsusaka and F. Palda, "Voter turnout: How much can we explain?" *Public choice*, vol. 98, no. 3-4, pp. 431–446, 1999.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[61] D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford, "Large-scale methods for distributionally robust optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8847–8860, 2020.

[62] E. Diana, W. Gill, M. Kearns, K. Kenthapadi, A. Roth, and S. Sharifi-Malvajerdi, "Multiaccurate proxies for downstream fairness," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1207–1239.

APPENDIX A
MAIN PROOFS

A. Proof of Theorem 1

First, we demonstrate the following lemma:

Lemma 1. Suppose that $0 < b < 1$ almost surely and $\mathbb{E}|f(\hat{Y}, y)|\mathcal{E}|$ is finite. Under the assumption of independent and identically distributed data with $\mathcal{E}$ having strictly positive probability, the asymptotic limits D_μ^P and D_μ^L satisfy:

$$D_\mu^P = \frac{\text{Cov}\left[b, f(\hat{Y}, Y)|\mathcal{E}\right]}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])} \text{ and } D_\mu^L = \frac{\text{Cov}\left[b, f(\hat{Y}, Y)|\mathcal{E}\right]}{\text{Var}[b|\mathcal{E}]},$$

and thus

$$D_\mu^P = D_\mu^L \cdot \frac{\text{Var}[b|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])}.$$

Proof. We note that:

$$\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} b_i \overset{n_\mathcal{E} \to \infty}{\longrightarrow} \mathbb{E}[b|\mathcal{E}]$$

and

$$\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} b_i f(\hat{Y}, Y) \overset{n_\mathcal{E} \to \infty}{\longrightarrow} \mathbb{E}[b \cdot f(\hat{Y}, Y)|\mathcal{E}]$$

by the strong law of large numbers. Similarly,

$$\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} (1 - b_i) f(\hat{Y}, Y) \overset{n_\mathcal{E} \to \infty}{\longrightarrow} \mathbb{E}[(1 - b) \cdot f(\hat{Y}, Y)|\mathcal{E}]$$

$$\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} (1 - b_i) \overset{n_\mathcal{E} \to \infty}{\longrightarrow} \mathbb{E}[1 - b|\mathcal{E}]$$

Then diving numerators and denominators in the definition of the empirical estimator gives that:

$$\widehat{D}_\mu^P = \frac{\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} b_i f(\hat{Y}_i, Y_i)}{\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} b_i} - \frac{\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} (1 - b_i) f(\hat{Y}_i, Y_i)}{\frac{1}{n_\mathcal{E}} \sum_{i \in \mathcal{E}} (1 - b_i)}$$
$$\overset{n_\mathcal{E} \to \infty}{\longrightarrow} \frac{\mathbb{E}[b f(\hat{Y}, Y)|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}]} - \frac{\mathbb{E}[(1 - b) f(\hat{Y}, Y)|\mathcal{E}]}{\mathbb{E}[(1 - b)|\mathcal{E}]}$$

Combining terms and expanding out the algebra, the last term is:

$$\frac{\mathbb{E}[b f(\hat{Y}, Y)|\mathcal{E}] - \mathbb{E}[b|\mathcal{E}]\mathbb{E}[f(\hat{Y}, Y)|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])} = \frac{\text{Cov}\left[b, f(\hat{Y}, Y)|\mathcal{E}\right]}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])}.$$

On the other hand, the linear estimator converges asymptotically to

$$\widehat{D}_\mu^L \overset{n_\mathcal{E} \to \infty}{\longrightarrow} \frac{\text{Cov}\left[b, f(\hat{Y}, Y)|\mathcal{E}\right]}{\text{Var}[b|\mathcal{E}]}.$$

This result can be seen by conditioning on $\mathcal{E}$ and then making the standard arguments for the asymptotic convergence of the OLS estimator. Comparing forms of the limits gives the final result. □

Our key theorem follows as a corollary from the following proposition, (Proposition 1 in the main text):

Proposition. Suppose that b is a prediction of an individual's protected attribute (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_μ^P as the asymptotic limit of the probabilistic disparity estimator, $\widehat{D}_p$, and D_l as the asymptotic limit of the linear disparity estimator, $\widehat{D}_l$. Then:

1)

$$D_\mu^P = D_\mu - \frac{\mathbb{E}[\text{Cov}(f(\hat{Y}, Y), B|b, \mathcal{E})]}{\text{Var}(B|\mathcal{E})} \quad \textbf{(1.1)}$$

2)

$$D_\mu^L = D_\mu + \frac{\mathbb{E}[\text{Cov}(f(\hat{Y}, Y), b|B, \mathcal{E})]}{\text{Var}(b|\mathcal{E})} \quad \textbf{(1.2)}$$

We'll split things into separate proofs for (1.1) and (1.2). We'll also first separately highlight that disparity is simply the dummy coefficient on race in a(n appropriately conditioned) regression model. This fact may be known by some readers in the context of regression analysis (especially without conditioning on a given event), but we provide proof of the general case.

Lemma 2. Let D_μ be the disparity with function f and event $\mathcal{E}$. Then D_μ can be written as:

$$D_\mu = \frac{\text{Cov}(f(\hat{Y}, Y), B|\mathcal{E})}{\text{Var}(B|\mathcal{E})}.$$

Proof. Note that by definition:

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

If the right hand side of the equation in the statement of the lemma can be written this way, we are done. But note that:

$$\frac{\text{Cov}(f(\hat{Y}, Y), B|\mathcal{E})}{\text{Var}(B|\mathcal{E})} = \frac{\mathbb{E}[f(\hat{Y}, Y)B|\mathcal{E}] - \mathbb{E}[f(\hat{Y}, Y)|\mathcal{E}]\mathbb{E}[B|\mathcal{E}]}{\mathbb{E}[B|\mathcal{E}](1 - \mathbb{E}[B|\mathcal{E}])}.$$

Now using the law of iterated expectations and simplifying:

$$\begin{aligned}
\mathbb{E}[f(\hat{Y}, Y)B|\mathcal{E}] &= \mathbb{E}[\mathbb{E}[f(\hat{Y}, Y)B|\mathcal{E}, B] \\
&= \mathbb{E}[f(\hat{Y}, Y)B|B = 1, \mathcal{E}] \Pr[B = 1|\mathcal{E}] \\
&+ \mathbb{E}[f(\hat{Y}, Y)B|B = 0, \mathcal{E}] \Pr[B = 0|\mathcal{E}] \\
&= \mathbb{E}[f(\hat{Y}, Y)|B = 1, \mathcal{E}] \Pr[B = 1|\mathcal{E}] \\
&+ \mathbb{E}[0] \Pr[B = 0|\mathcal{E}] \\
&= \mathbb{E}[f(\hat{Y}, Y)|B = 1, \mathcal{E}] \Pr[B = 1|\mathcal{E}]
\end{aligned}$$

Moreover, since B is a Bernoulli random variable, $\Pr[B = 1|\mathcal{E}] = \mathbb{E}[B|\mathcal{E}]$ and

$$\text{Var}(B|\mathcal{E}) = \mathbb{E}[B|\mathcal{E}](1 - \mathbb{E}[B|\mathcal{E}])$$

Combining these, we can write:

$$\frac{\mathbb{E}[f(\hat{Y}, Y)B|\mathcal{E}]\mathbb{E}[B|\mathcal{E}] - \mathbb{E}[f(\hat{Y}, Y)|\mathcal{E}]\mathbb{E}[B|\mathcal{E}]}{\mathbb{E}[B|\mathcal{E}](1 - \mathbb{E}[B|\mathcal{E}])}$$
$$= \frac{\mathbb{E}[f(\hat{Y}, Y)|B = 1, \mathcal{E}] - \mathbb{E}[f(\hat{Y}, Y)|\mathcal{E}]\mathbb{E}[B|\mathcal{E}]}{(1 - \mathbb{E}[B|\mathcal{E}])}$$

This can be expanded as:

$$\begin{aligned}
&\frac{\mathbb{E}[f(\hat{Y},Y)|B=1,\mathcal{E}]}{(1-\mathbb{E}[B|\mathcal{E}])} \\
&- \frac{\mathbb{E}[f(\hat{Y},Y)|B=1,\mathcal{E}]\Pr[B=1|\mathcal{E}]}{(1-\mathbb{E}[B|\mathcal{E}])} \\
&- \frac{\mathbb{E}[f(\hat{Y},Y)|B=0,\mathcal{E}]\Pr[B=0|\mathcal{E}]}{(1-\mathbb{E}[B|\mathcal{E}])} \\
&= \frac{\mathbb{E}[f(\hat{Y},Y)|B=1,\mathcal{E}](1-\Pr[B=1|\mathcal{E}])}{(1-\Pr[B=1|\mathcal{E}])} \\
&\frac{-\mathbb{E}[f(\hat{Y},Y)|B=0,\mathcal{E}](1-\Pr[B=1|E])}{(1-\Pr[B=1|\mathcal{E}])} \\
&= \mathbb{E}[f(\hat{Y},Y)|B=1,\mathcal{E}] - \mathbb{E}[f(\hat{Y},Y)|B=0,\mathcal{E}]
\end{aligned}$$

as desired. □

Note that the familiar interpretation of demographic disparity being the dummy coefficient falls out from this lemma by letting $\mathcal{E}$ be the event "always true" and $f(\hat{Y},Y)=Y$.

Now we can turn to proving (1.1). Recall first that, by assumption:

$$\begin{aligned}
b = \Pr[B=1|Z,\mathcal{E}] &= \mathbb{E}[\mathbb{1}[B=1]|Z,\mathcal{E}] \\
\implies b &= \mathbb{E}[B|Z,\mathcal{E}]\ \forall Z \\
\implies \mathbb{E}[b|\mathcal{E}] &= \mathbb{E}[\mathbb{E}[B|Z,\mathcal{E}]] = \mathbb{E}[B|\mathcal{E}]
\end{aligned}$$

by the law of iterated expectations. Moreover, if we define ϵ as $B-b$, then:

$$\mathbb{E}[\epsilon|Z,\mathcal{E}] = \mathbb{E}[B|Z,\mathcal{E}] - \mathbb{E}[b|Z,\mathcal{E}] = 0$$

Proof of (1.1). Note that by Lemmas 1 and 2:

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

Since $\mathbb{E}[b|\mathcal{E}] = \mathbb{E}[B|\mathcal{E}]$ and $\text{Var}[B|\mathcal{E}] = \mathbb{E}[B|\mathcal{E}](1-\mathbb{E}[B|\mathcal{E}]) = \mathbb{E}[b|\mathcal{E}](1-\mathbb{E}[b|\mathcal{E}])$, the denominators are the same and be collected as $\text{Var}(B|\mathcal{E})$. As for the numerators, we note that

$$\begin{aligned}
&\text{Cov}\left[f(\hat{Y},Y),B|\mathcal{E}\right] - \text{Cov}\left[f(\hat{Y},Y),b|\mathcal{E}\right] \\
&= \text{Cov}\left[f(\hat{Y},Y),B-b|\mathcal{E}\right]
\end{aligned}$$

by the distributive property of covariance. Recall that the law of total covariance allows us to break up the covariance of random variables into two parts when conditioned on a third. Applying this to $f(\hat{Y},Y)$ and $B-b$, with the conditioning variable being b, we have that:

$$\begin{aligned}
\text{Cov}\left[f(\hat{Y},Y),B-b|\mathcal{E}\right] &= \mathbb{E}\left[\text{Cov}\left(f(\hat{Y},Y),B-b\right)|\mathcal{E},b\right] \\
&+ \text{Cov}\left(\mathbb{E}[f(\hat{Y},Y)|\mathcal{E},b],\mathbb{E}[B-b|\mathcal{E},b]\right) \\
&= \mathbb{E}\left[\text{Cov}\left(f(\hat{Y},Y),B-b\right)|\mathcal{E},b\right] \\
&= \mathbb{E}\left[\text{Cov}\left(f(\hat{Y},Y),B\right)|\mathcal{E},b\right]
\end{aligned}$$

where the second equality follows because $b=\mathbb{E}[B|Z,\mathcal{E}] \implies \mathbb{E}[B|b,\mathcal{E}]=b$ and the third because b is trivially a constant given b. Combining these together, we have that:

$$\begin{aligned}
D_\mu - D_\mu^P &= \frac{\mathbb{E}\left[\text{Cov}\left(f(\hat{Y},Y),B\right)|\mathcal{E},b\right]}{\text{Var}[B|\mathcal{E}]} \\
\implies D_\mu^P &= D_\mu - \frac{\mathbb{E}\left[\text{Cov}\left(f(\hat{Y},Y),B\right)|\mathcal{E},b\right]}{\text{Var}[B|\mathcal{E}]},
\end{aligned}$$

as desired. □

Let's do (1.2).

Proof of (1.2). First, consider the linear projection of $f(\hat{Y},Y)$ onto B given that $\mathcal{E}$ occurs. We can write this as:

$$f(\hat{Y},Y) = \alpha + \gamma\cdot B + \nu,$$

where it is understood that the equation holds given $\mathcal{E}$. Now, by the definition of linear projection,

$$\gamma = \frac{\text{Cov}(f(\hat{Y},Y),B|\mathcal{E})}{\text{Var}(B|\mathcal{E})} = D_\mu$$

where the last equality follows by Lemma 2, and by the definition of linear projection, $\text{Cov}(B,\nu|\mathcal{E})=0$.

Now, consider the linear projection of $f(\hat{Y},Y)$ onto b given $\mathcal{E}$. Again we can write the equation:

$$f(\hat{Y},Y) = \alpha' + \beta b + \eta$$

and similarly

$$\beta = \frac{\text{Cov}(f(\hat{Y},Y),b|\mathcal{E})}{\text{Var}(b|\mathcal{E})} = D_\mu^L$$

and $\text{Cov}(b,\eta|\mathcal{E})=0$.

Now, by applying the Law of Total Covariance to the equation above, we have:

$$\begin{aligned}
\beta\text{Var}(b|\mathcal{E}) &= \text{Cov}(f(\hat{Y},Y),b|\mathcal{E}) \\
&= \mathbb{E}[\text{Cov}(f(\hat{Y},Y),b|\mathcal{E},B] \\
&+ \text{Cov}(\mathbb{E}[f(\hat{Y},Y)|\mathcal{E},B],\mathbb{E}[b|\mathcal{E},B]).
\end{aligned}$$

We'll focus for now on the latter term. Note that by replacing $f(\hat{Y},Y)$ by $\alpha+\gamma B+\nu$, we can obtain:

$$\begin{aligned}
\text{Cov}(\mathbb{E}[f(\hat{Y},Y)|B,\mathcal{E}],\mathbb{E}[b|B,\mathcal{E}]) &= \text{Cov}(\gamma B + \mathbb{E}[\nu|B], \\
&\quad B - \mathbb{E}[\epsilon|B]\big|\mathcal{E})
\end{aligned}$$

where we've moved out the event $\mathcal{E}$ and used the fact that α is a constant and B is a constant conditional on B to remove them from the inner expectations. We can expand as

$$\text{Cov}\left(\gamma B + \mathbb{E}[\nu|B,\mathcal{E}], B - \mathbb{E}[\epsilon|B]\big|\mathcal{E}\right).$$

We can further expand this covariance term to be

$$\begin{aligned}
&= \gamma\text{Var}(B|\mathcal{E}) - \gamma\,\text{Cov}(B,\mathbb{E}(\epsilon|B)\big|\mathcal{E}) \\
&+ \text{Cov}(\mathbb{E}(\nu|B),B\big|\mathcal{E}) - \text{Cov}(\mathbb{E}(\nu|B),\mathbb{E}(\epsilon|B)\big|\mathcal{E}) \\
&= \gamma\text{Var}(B|\mathcal{E}) - \gamma\,\text{Cov}(B,\mathbb{E}(\epsilon|B)\big|\mathcal{E}),
\end{aligned}$$

where the last equality is due to the fact that B is binary so the covariance between B and ν equals zero.

Next we show that the term $\text{Cov}(B, \mathbb{E}(\epsilon|B)\big|\mathcal{E})$ can be written in terms of b and ϵ,

$$\begin{aligned}\text{Cov}(B, \mathbb{E}(\epsilon|B)\big|\mathcal{E}) &= \mathbb{E}[B\mathbb{E}[\epsilon|B]] - \mathbb{E}[B]\mathbb{E}[\mathbb{E}[\epsilon|B]] \\ &= \mathbb{E}[\mathbb{E}[B\epsilon|B]\big|\mathcal{E}] - \mathbb{E}[B|\mathcal{E}]\mathbb{E}[\mathbb{E}[\epsilon|B]|\mathcal{E}] \\ &= \mathbb{E}[B\epsilon|\mathcal{E}] - \mathbb{E}[B|\mathcal{E}]\mathbb{E}[\epsilon|\mathcal{E}] \\ &= \text{Cov}(B, \epsilon\big|\mathcal{E}) \\ &= \text{Cov}(b+\epsilon, \epsilon\big|\mathcal{E}) \\ &= \text{Cov}(b, \epsilon\big|\mathcal{E}) + \text{Var}(\epsilon|\mathcal{E}).\end{aligned}$$

Plugging these results back into the original equation and using the fact that $B = b + \epsilon$, we have

$$\begin{aligned}\beta\text{Var}(b|\mathcal{E}) &= \mathbb{E}[\text{Cov}(f(\hat{Y}, Y), b|\mathcal{E}, B] \\ &\quad + \gamma\text{Var}(B|\mathcal{E}) - \gamma\text{Var}(\epsilon|\mathcal{E}) - \gamma\,\text{Cov}(b, \epsilon\big|\mathcal{E}) \\ &= \gamma[\text{Var}(b|\mathcal{E}) + \text{Cov}(b, \epsilon\big|\mathcal{E})] \\ &\quad + \mathbb{E}[\text{Cov}(f(\hat{Y}, Y), b|\mathcal{E}, B] \\ &= \gamma\text{Var}(b|\mathcal{E}) + \mathbb{E}[\text{Cov}(f(\hat{Y}, Y), b|\mathcal{E}, B],\end{aligned}$$

where the last equality is due to the fact that $\mathbb{E}[\epsilon|Z, \mathcal{E}] = 0$. □

B. Proof of Proposition 2

Proof. For a *fixed* $\tilde{\theta}$, we can apply Theorem 1 to write that:

$$D_\mu^p(h_{\tilde{\theta}}) = D_\mu(h_{\tilde{\theta}}) - \frac{\mathbb{E}[\text{Cov}(f(h_{\tilde{\theta}}, Y), B|b, \mathcal{E}]}{\text{Var}[B|\mathcal{E}]},$$

where the expectation in the numerator is over the distribution of the data. Now, if $\tilde{\theta}$ is drawn from a distribution $\boldsymbol{\theta}$ (in particular, $\boldsymbol{\theta}$ corresponding to θ_t with t being drawn from $1...T$) that is independent of the data, we can treat the quantities as random variables drawn from a two step data-generating process. In our setting (as in classical, but not all, learning settings), the distribution of future data is assumed not to depend on our selected model. Then by the linearity of expectations, we have that

$$\begin{aligned}&\mathbb{E}_{\tilde{\theta}\sim\boldsymbol{\theta}}\left[D_\mu^p(h_{\tilde{\theta}})\right] - \mathbb{E}_{\tilde{\theta}\sim\boldsymbol{\theta}}\left[D_\mu(h_{\tilde{\theta}})\right] \\ &= \mathbb{E}_{\tilde{\theta}\sim\boldsymbol{\theta}}\left[\frac{\mathbb{E}[\text{Cov}(f(h_{\tilde{\theta}}, Y), B|b, \mathcal{E}]}{\text{Var}[B|\mathcal{E}]}\right].\end{aligned}$$

A similar statement can be made for the relationship between $\mathbb{E}_{\tilde{\theta}\sim\boldsymbol{\theta}_T}\left[D_\mu^p(h_{\tilde{\theta}})\right]$ and $\mathbb{E}_{\tilde{\theta}\sim\boldsymbol{\theta}_T}\left[D_\mu(h_{\tilde{\theta}})\right]$. □

C. Standard Errors

Here, we discuss the calculation of standard errors; these arguments are more general, but substantially similar, version of those made in [19]. As shown in the proof of Theorem 1, $\widehat{D}_\mu^l$ and $\widehat{D}_\mu^p$ converge to their asymptotic limits, D_μ^l and D_μ^p, respectively; however, given that we observe only a finite sample, our estimates $\widehat{D}_\mu^l$ and $\widehat{D}^p$ are subject to uncertainty whose magnitude depends on the sample size of the data.

Since the $\widehat{D}_\mu^l$ is simply the linear regression coefficient, its distribution is well-studied and well known. In particular, under the classical ordinary least squares (OLS) assumptions of normally distributed error, $\widehat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{ns_b^2}\right)$ where s_b^2 is the sample variance of b; under mild technical conditions, central limit theorems can be invoked to show that as the size of data increases, $\widehat{\beta}$ follows a distribution that is increasingly well-approximated by said normal distribution [56]. Note that, since as shown in Lemma 1

$$D_\mu^L = \frac{\text{Cov}(f(\hat{Y}, Y), b|\mathcal{E})}{\text{Var}[b|\mathcal{E}}$$

and

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

it follows that

$$D_\mu^P = D_\mu^L \cdot \frac{\text{Var}[b|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|E])};$$

analogously, by expanding the definitions of the sample estimators, we can easily see that:

$$\widehat{D}_\mu^P = \widehat{D}_\mu^L = \frac{\frac{1}{n_\mathcal{E}}\sum_{i\in\mathcal{E}}(b_i - \bar{b}^\mathcal{E})^2}{\bar{b}^\mathcal{E}(1 - \bar{b}^\mathcal{E})}.$$

Then by Slutsky's theorem, we can state that:

$$\widehat{D}_\mu^P \xrightarrow{n\to\infty} \widehat{D}_\mu^L \frac{\text{Var}[b|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|E])}.$$

As a consequence, the distribution of $\widehat{D}_\mu^P$ is a scaled version of the distribution of $\widehat{D}_\mu^L$, and in particular

$$\frac{\widehat{D}_\mu^P - D_\mu^P}{\text{Var}\widehat{D}_\mu^L \sqrt{\frac{\text{Var}[b|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}](1-\mathbb{E}[b|E])}}} \xrightarrow{n\to\infty} \mathcal{N}(0, 1).$$

Thus, in practice, we can estimate the variance of $\widehat{D}_\mu^L$ as if it were the usual OLS estimator and then estimate $\text{Var}[b|\mathcal{E}]$ and $\mathbb{E}[b|\mathcal{E}]$ to scale it appropriately.

D. Obtaining the probabilistic prediction

1) BIFSG: Recall that conceptually, b functions as a probabilistic confidence score we have that an individual has $B = 1$. A perfectly calibrated b will thus have $\mathbb{E}[B|b] = b$, and our main theorems assume that we have access to this. In practice, however, b must be estimated; in this work, we focus on the commonly used [16, 20, 57, 58] Bayesian Imputations with First Names, Surnames, and Geography (BIFSG). In BIFSG, we make the naive conditional independence assumption that the proxy features are independent conditional on the protected characteristic. In the case of BIFSG, this amounts to assume that:

$$\Pr[F, S, G|B] = \Pr[F|B]\Pr[S|B]\Pr[G|B],$$

where the random variable F is first name, S is surname, and G is geography . By applying Bayes' rules to this assumption,

we can obtain that:

$$\Pr[B|F,S,G] = \frac{\Pr[F,S,G|B]}{\Pr[F,S,G]} = \frac{\Pr[F|B]\Pr[S|B]\Pr[G|B]}{\Pr[F,S,G]}.$$

The right-hand side of this equation is fairly easy to estimate because it requires knowing only marginals rather than joint distributions (the denominator can be normalized away by noting that we must have that $\Pr[B = 1|F,S,G]$ and $\Pr[B = 0|F,S,G]$ must sum to 1), and these marginals are often obtainable in the form of publicly available datasets. Note that, BIFSG can be written in multiple forms by applying Bayes' rule again to the individual factors (e.g. replacing $\Pr[F|B]$ with $\Pr[B|F]\Pr[F]/\Pr[B]$, which may be convenient depending on the form of auxiliary data available.

For our setting, we leverage the census and home mortgage disclosure act (HMDA) data, as mentioned, to estimate b from publicly available data. We provide quantitative details on our estimates in Appendix C. We note also that since b is continuous, we will discretize into equally sized bins whenever we need to compute quantities conditional on b.

2) Impact of Miscalibration: Throughout the theoretical work, we have assumed that we have $b = \Pr[B = 1|Z]$ - i.e. that b is *perfectly calibrated*. In reality, this is a quantity that is estimated, and will thus contain some uncertainty, including bias due to the fact that the dataset which it is estimated on (e.g. the census for the U.S. as a whole) may not be fully representative of the relevant distribution (i.e. the distribution of individuals to whom the model will be applied, which may be a particular subset). This could result in *miscalibration*; when this happens, it could be that applying our method with our miscalibrated b results in failing to bound disparity (both in measuring alone, and in training).

Ultimately, miscalibration is only a real problem insofar as it causes the method to fail. For small amounts of miscalibration, the method tends to succeed anyway – e.g. in our setting, we do observe that our estimates are not perfectly calibrated, but we still achieve good results. For larger, or unknown, miscalibration, there are two paths that can be taken. The first is to conduct a "recalibration" exercise, and obtain a modified b that more closely matches the distribution of interest; this can be as simple as fitting a linear regression of B on b in the labeled dataset and replace b with the predictions of this regression. Alternatively, given an assumed bound on the magnitude of the miscalibration, Theorem 1 can be extended to incorporate its effect. In practice, recalibration is more straightforward to do empirically, but the theoretical method can also be used for sensitivity analysis; see [19] for their discussion of the recalibration approach as well as the effect on their special-case bounds.

Note also that, in settings where $\mathcal{E}$ is affected by the modeling choice h - i.e. when the fairness metric involves conditioning on model predictions, as in the case of positive predictive value (PPV) - it may be the case that a perfect or well-calibrated b for one model may be poorly-calibrated for another. That is, it may be that among observations, we find that that our estimate $|b(Z) - \Pr[B|Z,\mathcal{E}(h_\theta)]|$ is small while our estimate of $|b(Z) - \Pr[B|Z,\mathcal{E}(h_{\theta'})|$ is large. In this case, we can introduce a recalibration step in-between iterations, although this deviates from the theoretical assumptions that ensure convergence. Note that a sufficiently expressive model over a sufficiently powerful set of proxy features should be able obtain good calibration overall events $\mathcal{E}$; this suggests that another path forward in such a setting may be in investing in alternative, more powerful (e.g. machine-learned) models of b.

Appendix B
Mathematical Formulation of Fair Learning Problem

A. Theoretical Problem

We begin by discussing the *theoretical* problems - i.e. abstracting away from the sample of data and considering the problems we are trying to solve.

1) One-sided bound: We first consider the case of imposing a one-sided bound on disparity, i.e. requiring that $D_\mu \leq \alpha$ but allowing $D_\mu < -\alpha$; certainly this will not be desirable in all situations, but we can use it as a building block for the two-sided bound as well.

We begin by formalizing the ideal problem - that is, the problem we would solve if we had access to ground truth protected class. This is simply to minimize the expected risk subject to the constraint that - whichever disparity metric we have adopted - disparity is not "too high". This is the:

Problem 3 (Ideal Problem)**.** Given individual features X, labels Y, a loss function L, a model class $\mathcal{H}$, a disparity metric μ, and a desired bound on disparity α, find an h to:

$$\min_{h\in\mathcal{H}} \mathbb{E}[L(h(X),Y)] \text{ s.t. } D_\mu(h) \leq \alpha,$$

where $D_\mu(h)$ is the μ-disparity obtained by h.

The ideal problem is not something we can solve because we cannot directly calculate D_μ over the dataset, since it requires the ground truth protected class label B. But the Theorem 1 suggests an alternative and feasible approach: using the linear estimate of disparity as a proxy bound. That is, if the linear estimator is an upper bound on the disparity, and the linear estimator is below α, then disparity is below α too.

Formally, we would solve following problem:

Problem 4 (Bounded Problem Direct)**.** Given individual features X, labels Y, a loss function L, a model class $\mathcal{H}$, a disparity metric μ, a desired on disparity α, and a predicted protected attribute proxy b, find an h to:

$$\min_{h\in\mathcal{H}} \mathbb{E}[L(h(X),Y)] \text{ s.t. } D^L_\mu \leq \alpha \text{ and } D_\mu \leq D^L_\mu$$

Notice that any feasible solution to Problem 4 must satisfy the constraints of Problem 3, i.e. we must have that $D_\mu(h) \leq \alpha$. The gap between the performance of these two solutions can be regarded as a "price of uncertainty"; it captures the loss we incur

by being forced to use our proxy to bound disparity implicitly rather than being able to bound it directly. We explore this price by comparing to an "oracle" which can observe the ground truth on the full dataset and perform constrained statistical learning.

As in Problem 2, we cannot directly observe D_μ, so the second constraint is not one that we can directly attempt to satisfy. But we know that it holds exactly in the conditions under which Theorem 1 applies. Therefore, we can replace that constraint with the covariance conditions:

Problem 5 (Fair Problem - Indirect)**.** Given individual features X, labels Y, a loss function L, a model class $\mathcal{H}$, a disparity metric μ (with associated event $\mathcal{E}$ and function $f(h(X), Y)$), a desired maximum disparity α, and a predicted proxy b, find an h to:

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

And indeed, these problems are equivalent:

Proposition 3. Problems 5 and 4 are equivalent.

Proof. Theorem 1 directly says that $D_\mu^L \geq D_\mu \iff \mathbb{E}[\text{Cov}(f(h(X),Y), b|B,\mathcal{E})] \geq 0$. Hence if h satisfies the constraints of Problem 5 iff it satisfies those of Problem 4. Since the objectives are also the same, the problems are equivalent. □

As written, Problem 5 is still using the population distributions; we will discuss its empirical analogue below.

2) Two-sided bound: The two-sided bound requires that $|D_\mu| \leq \alpha$; this may be more common in practice. Again, we begin by considering the ideal problem:

Problem 6 (Ideal Symmetric Problem)**.** Given individual features X, labels Y, a loss function L, a model class $\mathcal{H}$, a disparity metric μ, and a desired bound on disparity α, find an h to:

$$\min_{h\in\mathcal{H}} \mathbb{E}[L(h(X),Y)] \text{ s.t. } |D_\mu(h)| \leq \alpha,$$

where $D_\mu(h)$ is the μ-disparity obtained by h.

As with Problem 4, we cannot directly bound disparity, since we do not have it, but we do have the disparity estimator. This leads to the following problem:

Problem 7 (Symmetric Problem Direct)**.** Given individual features X, labels Y, a loss function L, a model class $\mathcal{H}$, a disparity metric μ, a desired on disparity α, and a predicted protected attribute proxy b, find an h to:

$$\min_{h\in\mathcal{H}} \mathbb{E}[L(h(X),Y)] \text{ s.t. } |D_\mu^L| \leq |\alpha| \\ \text{and } |D_\mu| \leq |D_\mu^L|$$

Unfortunately, we don't have any theory about putting an absolute value bound on disparity, and indeed, because the weighted and linear disparity estimators are positive scalar multiples of one another, we cannot hope to use one as a positive upper bound and the other as a negative lower bound. But notice that if we were to find the best solution when $D_\mu^L \in [0,\alpha]$, and the best solution when $D_\mu^L \in [-\alpha, 0]$, then we would cover the same range as $[-\alpha,\alpha]$.

One attempt to apply this principle would be to solve the following two subproblems:

Problem 6.A.

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

Problem 6.B.

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

And take:

$$h_5^* = \text{argmin}_{h_{6a}^*, h_{6b}^*} \mathbb{E}[L(h(X),Y)].$$

But this does not even guarantee a *feasible*, let alone optimal, solution to Problem 7. To see this, note that there is nothing prevent h_{6a}^* to be not simply $\leq \alpha$, but in fact $< -\alpha$, and vice versa. In particular, what went wrong is that we did not find the two best solutions over $[-\alpha, 0]$ and $[0,\alpha]$, but rather the two best over $[-\infty,\alpha]$ and $[-\alpha,\infty]$, which is no constraint at all.

To get around, this, though, we can solve the following two problems instead:

Problem 7.A.

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

Problem 7.B.

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

Why are these different? Notice that imposing both covariance constraints in 1.A enforces that $D_\mu^p \leq D_\mu \leq D_\mu^L$; since $D_\mu^p = D_\mu^L \frac{\text{Var}b}{\mathcal{E}[b](1-\mathcal{E}[b])}$ – i.e. D_μ^p is always an attenuated version of D_μ^l – this can *only* be the case if all three terms are nonnegative. Similarly, 1.B enforces that $\mathcal{D}_\mu^p \geq D_\mu \geq D_\mu^l$; this similarly ensures that all three terms are nonpositive. Since these terms also include the bound on the linear estimator, they thus ensure that if we take:

$$h \in \text{argmin}_{h_{7a}^*, h_{7b}^*} \mathbb{E}[L(h(X),Y)],$$

we will indeed obtain a feasible solution to Problem 7. As in Problem 5, there may again be a suboptimality gap since we have effectively imposed more constraints to the original problem.

B. Solving the Empirical Problems

In this section, we use recent results in constrained statistical learning to formulate and motivate empirical problems that we can solve which obtain approximately feasible and performant solutions to the problems above. We summarize here the conceptual basis at a high level, providing a discussion of the rationale behind Theorem 2 in the main text, drawing heavily on [14], and refer interested readers to said work as well as [43] for a fuller and more detailed discussion of the constrained statistical learning relevant to our setting and [44] for more general discussion of non-convex optimization via primal-dual games.

1) Relating our Formulation: We begin by describing the relationship between our problem of interest and that considered in [14]. The (parameterized version of the) problem in [14] is the following:

Problem 8 (Parameterized Constrained Statistical Learning (P-CSL) from [14])**.**

$$P^* = \min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_0}\left[\ell_0(f_\theta(x), y)\right]$$
$$\text{s.t. } \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\ell_i(f_\theta(x,y)\right] \leq c_i,\ i = 1...m$$

That is, they aim to minimize some expected loss subject to some constrained on other expected losses, with loss functions that may vary and be over different distributions. Our problem, i.e. Problem 5 can be seen as a special case of this, though our framing is different. To see the correspondence, consider applying the following to Problem 8:

1) Take $\mathcal{D}_i$ to be the restriction of $\mathcal{D}$ to $\mathcal{E}$
2) Take ℓ_0 to be the loss function of interest, e.g. $\mathbf{1}[h \neq y]$ for accuracy
3) Take $\ell_1 = f(h(X), Y)$ and c_1 as α
4) Take $\ell_2 = f(h(X), Y) \cdot B - \overline{f(h(X), Y)}^B \bar{b}^B$ and $c_2 = 0$
5) Take $\ell_3 = f(h(X), Y) \cdot b - \overline{f(h(X), Y)}^b \bar{B}^b$ and $c_3 = 0$

Then we arrive at Problem 5.

2) Moving to the empirical problem: The problems described above relate to the population distribution, but we only have samples from this distribution. This is, of course, the standard feature of machine learning situations; the natural strategy in such a setting is to simply solve the empirical analogue - i.e. to replace expectations over a distribution with a sample average over the realized data. Instantiating this and focusing on Problem 1.A (since the other problems can be solved analogously and/or using it as a subproblem) we could write the following empirical problem.

Problem 9.

$$\min_{h\in\mathcal{H}} \frac{1}{n} \sum_{i \in n_{\mathscr{D}}} L(h(X_i), Y_i) \text{ s.t. } \widehat{D}_\mu^L \leq \alpha$$
$$0 \leq -\frac{1}{n_{\mathscr{D}_L}} \sum_{i\in\mathscr{D}_L} \left[\left(f(h(X_i), Y_i) - \overline{f(h(X_i)), Y_i}^{B_i}\right)(b_i - \bar{b}^{B_i})\right]$$
$$0 \leq -\frac{1}{n_{\mathscr{D}_L}} \sum_{i\in\mathscr{D}_L} \left[\left(f(h(X_i), Y_i) - \overline{f(h(X_i)), Y_i}^{b_i}\right)(B_i - \bar{B}^{b_i})\right]$$

.

Problem 9 is not, in general, a convex optimization problem; if it were, the standard machinery and solutions of convex optimization, i.e. formulating the dual problem and recovering from it a primal solution via strong duality, could be applied. However, as shown in [14], under some conditions, there exists a solution to the empirical dual problem that obtains nearly the same objective value as the primal population problem. In other words, rather than applying strong duality as a consequence of problem convexity, [14] directly prove a relationship between the primal and the dual under some conditions. These conditions are that:

1) The losses $\ell_i(\cdot, y)$ are Lipschitz continuous for all y
2) Existence of a family of funtions $\zeta_i(N, \delta) \geq 0$ that decreases monontically in N and bounds the difference between the sample average and population expectatoin for each loss function
3) There is a $\nu \geq 0$ so that for each Φ in the closed convex hull of $\mathcal{H}$, there is a θ such that
4) The problem is feasible

We briefly discussing these conditions. For 1), we note that Lipschitz continuity requires existence of scalar such that $|f(x) - f(x')| \leq M|x - y|$, which will be true for bounded features when using sample averages. 2) simply requires that we are in a situation where more data is better, and is implied by the stronger condition we assume of $\mathcal{H}$ being of finite VC-dimension. 3) asks that our hypothesis class is rich enough to cover the space finely enough (how fine will determine the quality of the solution), which is met for reasonable model classes. 4), is simply a technical requirement ensuring that there exists at least some solution, is analogous to Slater's criterion in numerical optimization.

Thus, we can leverage the described guarantees to assert that solving the empirical dual would Yet this initial result, while positive, is one of existence; to actually find a solution requires a solution. To do so, one can construct an empirical Lagrangian from the constrained empirical problem, and this can be solved by running a game between primal player, who selects a model to minimize loss, and a dual player, who selects dual parameters in an attempt to maximize it. If we construct this empirical dual in our settings, it is as in Equation 3; Algorithm 1 provides a primal-dual learner that instantiates this idea of a game.

C. Theoretical Guarantees

If either all of the losses are convex, or:

6) The outcome of interest Y takes values in a finite set
7) The conditional random variables $X|Y$ is are non-atomic
8) The closed convex hull of $\mathcal{H}$ is *decomposable*

Then the primal-dual algorithm 1 performs well. In the classification setting, which we focus on, Item 5) is trivially true. Item 6) asks that it not be the case that any of the distribution over which losses are measured induce an outcomes induce an atomic distribution; this mild regularity condition prevents pathological cases that would be impossible to satisfy. For 7) *Decomposability* is a technical condition stating that for a

Algorithm 1: Primal-dual algorithm for probabilistic fairness

Input : Labeled subset $\mathscr{D}_L$, unlabeled data $\mathscr{D}_U$, θ-oracle, number of iterations $T \in \mathbb{N}$, step size $\eta > 0$

Define : $h_{\theta^{(t)}}$ as the model parameterized by $\theta^{(t)}$

Initialize : $\mu_L^{(1)} \leftarrow 0$; $\mu_{b|B}^{(1)} \leftarrow 0$; $\mu_{B|b}^{(1)} \leftarrow 0$

for $t = 1 \dots T$ **do**

  $\theta^{(t)} \leftarrow \text{argmin}_\theta \widehat{\mathcal{L}}(\theta, \mu^{(t)})$

  $\mu_{b|B}^{(t+1)} \leftarrow \mu_{b|B}^{(t)} + \eta \widehat{C}_{f,b|B}(h_{\theta^{(t)}})$;
  $\mu_{B|b}^{(t+1)} \leftarrow \mu_{B|b}^{(t)} + \eta \widehat{C}_{f,B|b}(h_{\theta^{(t)}})$

  $\mu_L^{(t+1)} \leftarrow \mu_L^{(t)} + \eta \left(\widehat{D}_L(h_{\theta^{(t)}} - \alpha \right)$

end

return $< \theta^{(1)}, \dots, \theta^{(T)} >$

given function space, it closed in a particular sense: for any two function Φ, Φ' and any measurable set χ, the function that is Φ on χ and Φ' on its complement is also in the function space; many machine learning methods can be viewed from a functional analysis as optimizing over decomposable function space.

As we have shown that our problem can be written as a case of the CSL problem, and Algorithm 1 is a specialization of the primal-dual learner analyzed in [14], Theorem 3 in the same applies, again with appropriate translation. In particular, the promise is that when an iterate is drawn uniformly at random, the expected losses (over the distribution of the data and this draw) for the constraints are bounded by the constraint limit c_i plus the family of functions at the datasize mention in Assumption 2, plus $2C/(\eta T)$, where T is number of iterations, η is the learning rate, and C is a constant; at the same time, the expected loss (again over both the data and drawing the iterate) is bounded by the value of primal plus several problem-specific constants that capture the difficult of the learning problem and meeting the constraints, as well as said monotonically decreasing function of the data capturing the rate of convergence. Our Theorem 2 can be obtained by applying a standard result from statistical learning theory and collecting/re-arrange/hide problem-specific constants.

In this section, we discuss our approach to learning a fair model using the probabilistic proxies and a small subset of labeled data. To do so, we leverage recent results in constrained statistical learning.

D. Handling Imperfect Calibration

In general, it may be that we do not have access to $b = \Pr[B = 1|Z = z]$, but instead have access to some *imperfectly calibrated* $\hat{b}$. In this case, we can write $\hat{b} = b + \varepsilon$, where ε by definition is $\hat{b} - b$. We could apply $\hat{D}_\mu^P$ and $\hat{D}_\mu^L$ using $\hat{b}$ instead, but Theorem 1 assumes access to b, and so does not directly apply. To overcome this, we can obtain a *recalibrated* b^*. As a first step, we know that for a general b, the linear and probabilistic estimators converge to:

$$D_\mu^L \longrightarrow D_\mu(1 + \frac{\text{Cov}(b, \varepsilon|\mathcal{E})}{\text{Var}[b|\mathcal{E}]}) + \frac{\mathbb{E}[\text{Cov}(f(\hat{Y}, Y), b|B]}{\text{Var}[b|\mathcal{E}]}$$

and

$$D_\mu^P \longrightarrow \frac{D_\mu \text{Var}[B|\mathcal{E}] - D_\mu^L \text{Cov}(b, \varepsilon)|\mathcal{E}}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])} - \frac{\mathbb{E}[\text{Cov}(Y, B|b, \mathcal{E}) + \mu}{\mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}])},$$

respectively; $\varepsilon := B - b$; and $\mu := \text{Cov}(\mathbb{E}(\eta|b, \mathcal{E}), \mathbb{E}[\varepsilon|b, \mathcal{E}]|\mathcal{E})$.

Now, with this form, we can see the following. First, for general b, as long as $\text{Cov}(b, \varepsilon|\mathcal{E}) = 0$ - that is, as long as miscalibration error ε is not correlated with the predictor itself - then we will have exactly the same equation as in 1.1. But we can obtain such a predictor simply by regressing B on b among $\mathcal{E}$; that is, if we run the linear regression

$$B = \alpha + \beta b + \varepsilon,$$

and define b^* as the $\hat{\alpha} + \hat{\beta} b$, then $\varepsilon^* = B - b^*$ by construction satisfies $\text{Cov}(b^*, \varepsilon^*) = 0$.

Then, in that case, we define:

$$D_\mu^{L,*} = D_\mu + \frac{\mathbb{E}[\text{Cov}(f(\hat{Y}, Y)), b^*|B]}{\text{Var}[b^*|\mathcal{E}]},$$

and we can now solve an empirical version of the one-sided problem (i.e. Problem 6.A using b^* instead of b, and all the math discussed above follows directly. However, to solve 7.A, we of course must handle the probabilistic estimator as well.

Here, again we can use $\text{Cov}(b^*, \varepsilon^*|\mathcal{E}) = 0$ and also observe that by construction:

$$\begin{aligned} \mathbb{E}[b^*|E] = \mathbb{E}[B|\mathcal{E}] \implies \mathbb{E}[b|\mathcal{E}](1 - \mathbb{E}[b|\mathcal{E}]) \\ = \mathbb{E}[B|\mathcal{E}](1 - \mathbb{E}[B|\mathcal{E}]) \end{aligned}$$

to simplify the first term in D_μ^{P*}, and so overall write:

$$\mathcal{D}_\mu^{P*} \longrightarrow D_\mu^P - \frac{\mathbb{E}[\text{Cov}(Y, B|b^*, \mathcal{E})] + \text{Cov}(\mathbb{E}[\eta^*|b^*], \mathbb{E}[\varepsilon^*|b^*]\mathcal{E})}{\text{Var}[B|\mathcal{E}]}$$

So to ensure that the lower bound holds, we must now incorporate the second term of the numerator into the optimization problem. But this can be done in a similar manner as before, as the residuals η^* and ε^* can again be expressed as algebraic sample averages.

E. Closed-form Solution to Fair Learning Problem for Regression Setting

In this appendix we provide a closed-form solution to the primal problem Problem 2.A for the special case of linear regression with mean-squared error losses and demographic parity as the disparity metric. We express the constraints in matrix notation and show that the constraints are linear in the parameter β. Thus, we are able to find a unique, closed-form solution for β by solving the first-order conditions. Given a choice of dual variables, it can be interpreted as a regularized heuristic problem with particular weights; while there are no guarantees that this will produce a performant or even feasible

solution, it may be useful when applying the method in its entirety is computationally prohibitive.

We define the following notation for our derivation. Let n denote the number of observations and p the number of features in our dataset. Then let $X \in \mathbb{R}^{n\times p}, y \in \mathbb{R}^{n\times 1}, \beta \in \mathbb{R}^{p\times 1}, b \in \mathbb{R}^{n\times 1}$, and $B \in \{0,1\}^{n\times 1}$. For $j = 0, 1$, let $B_j = \{i : B_i = j\}$ and $n_j = |B_j|$ denote the set of observations for which the observed protected feature $B = j$ and the size of the corresponding set, respectively. Since we consider demographic parity as the disparity metric of interest, we denote the disparity metric as $f(\hat{Y}, Y) = \hat{Y}$.

For ease of exposition, we restate the empirical version of the constrained optimization problem for linear regression and demographic parity.

Problem 9.A.

$$\begin{aligned}
&\min_{\beta} \ (y - X\beta)^\top (y - X\beta) \\
&\text{s.t. } \widehat{D}_\mu^L \le \alpha, \\
&\mathbb{E}[\mathrm{Cov}(\widehat{Y}, b|B)] \ge 0, \\
&\mathbb{E}[\mathrm{Cov}(\widehat{Y}, B|b)] \ge 0
\end{aligned}$$

As discussed in Section II-A, the linear disparity metric $\widehat{D}_\mu^L$ is the coefficient of the probabilistic attribute b in a linear regression of $\hat{Y}$ on b. Thus, $\widehat{D}_\mu^L$ can be expressed as

$$\widehat{D}_\mu^L = (b^\top b)^{-1}(b^\top X\beta).$$

The covariance of $\hat{Y}$ and b conditional on B can be written as

$$\mathrm{Cov}(\hat{Y}, b|B) = \mathbb{E}(b^\top X\beta|B) - \mathbb{E}(X\beta|B)\mathbb{E}(b|B) \quad (4)$$

We expand the first term on the right-hand side of Equation 4, considering the case where $B = 1$.

$$\begin{aligned}
\mathbb{E}(b^\top X\beta|B=1) &= \frac{1}{n_1}\sum_{i\in B_1} b_i X_i \beta \\
&= \frac{1}{n_1}\sum_{i\in B_1}\sum_{j=1}^{p} b_i X_{ij}\beta_j \\
&= \frac{1}{n_1}\sum_{j=1}^{p}\sum_{i\in B_1} b_i X_{ij}\beta_j \\
&= \frac{1}{n_1}\sum_{j=1}^{p}\beta_j \sum_{i\in B_1} b_i X_{ij}.
\end{aligned}$$

Collecting the second summation as the vector $v_{1j} = \frac{1}{n_1}\sum_{i\in B_1} b_i X_{ij}$, we can write the expression for $\mathbb{E}(b^\top X\beta|B=1)$ as

$$\mathbb{E}(b^\top X\beta|B=1) = \sum_{j=1}^{p}\beta_j v_{1j} = \beta^\top v_1,$$

where $v_1 = (v_{1j})_{j=1}^p$.

For the second term on the right-hand side of Equation 4 we can rewrite the summation in a similar manner. Again focusing on the case where $B = 1$,

$$\begin{aligned}
\mathbb{E}(X\beta|B)\mathbb{E}(b|B) &= \left(\frac{1}{n_1}\sum_{i\in B_1} X_i\beta\right)\left(\frac{1}{n_1}\sum_{i\in B_1} b_i\right) \\
&= \left(\frac{1}{n_1}\sum_{i\in B_1}\sum_{j=1}^{p} X_{ij}\beta_j\right)\left(\frac{1}{n_1}\sum_{i\in B_1} b_i\right) \\
&= \bar{b}_1 \frac{1}{n_1}\sum_{i\in B_1}\sum_{j=1}^{p} X_{ij}\beta_j.
\end{aligned}$$

We again collect the second summation and write it as $w_{1j} = \frac{1}{n_1}\sum_{i\in B_1} X_{ij}$ and then we can write $\mathbb{E}(X\beta|B)\mathbb{E}(b|B)$ as

$$\mathbb{E}(X\beta|B)\mathbb{E}(b|B) = \bar{b}_1\beta^\top w_1,$$

where $w_1 = (w_{1j})_{j=1}^p$.

Now we can write Equation 4 in matrix notation as,

$$\mathrm{Cov}(\hat{Y}, b|B) = \beta^\top v_1 - \bar{b}_1\beta^\top w_1 + \beta^\top v_0 - \bar{b}_0\beta^\top w_0, \quad (5)$$

where v_0, w_0 and $\bar{b}_0$ are defined equivalently for the set B_0. Finally we take the expectation of this covariance term to get,

$$\begin{aligned}
\mathbb{E}(\mathrm{Cov}(\hat{Y}, b|B)) = &\frac{n_1}{n}\left(\beta^\top v_1 - \bar{b}_1\beta^\top w_1\right) \\
&+ \frac{n_0}{n}\left(\beta^\top v_0 - \bar{b}_0\beta^\top w_0\right)
\end{aligned} \quad (6)$$

We now consider the covariance of $\hat{Y}$ and B conditional on b which can be written as

$$\mathrm{Cov}(\hat{Y}, B|b) = \mathbb{E}(B^\top X\beta|B) - \mathbb{E}(X\beta|b)\mathbb{E}(B|b). \quad (7)$$

The steps for expressing this conditional covariance in matrix notation are similar to the first covariance term, however, now we are summing over the continuous-valued variable b. Let $k \in [0, 1]$ denote the value of b we are conditioning on and let $G_k = \{i : b_i = k\}$, $n_k = |G_k|$ denote the set of observations with $b = k$ and the size of the set, respectively.

Once again we expand the first term on the right-hand side of Equation 7, this time considering the general case where $b = k$,

$$\mathbb{E}(B^\top X\beta|B) = \frac{1}{n_k}\sum_{j=1}^{p}\beta_j \sum_{i\in G_k} B_i X_{ij} = \beta^\top v_k.$$

Here we define $v_k = (v_{kj})_{j=1}^p$ and $v_{kj} = \frac{1}{n_k}\sum_{i\in G_k} B_i X_{ij}$. Following a similar process for the second term, we can express the term as

$$\mathbb{E}(X\beta|b)\mathbb{E}(B|b) = \bar{B}_k\beta^\top w_k,$$

where $w_k = (w_{kj})_{j=1}^p$ and $w_{kj} = \frac{1}{n_k}\sum_{i\in G_k} X_i j$. Combining the two terms together we write Equation 7 as

$$\mathrm{Cov}(\hat{Y}, B|b) = \sum_k \beta^\top v_k - \bar{B}_k\beta^\top w_k. \quad (8)$$

For the last step we take the expectation of the conditional covariance term to get,

$$\mathbb{E}(\mathrm{Cov}(\hat{Y}, B|b)) = \sum_k \frac{n_k}{n}\left(\beta^\top v_k - \bar{B}_k\beta^\top w_k\right). \quad (9)$$

Now we can write the empirical Lagrangian of Problem 9.A as

$$\widehat{\mathcal{L}}(\beta,\vec{\mu}) = (y-X\beta)^\top(y-X\beta) - \mu_L\left((b^\top b)^{-1}(b^\top X\beta)\right) + \mu_{b|B}\left(\frac{n_1}{n}\left(\beta^\top v_1 - \bar{b}_1\beta^\top w_1\right) + \frac{n_0}{n}\left(\beta^\top v_0 - \bar{b}_0\beta^\top w_0\right)\right) + \mu_{B|b}\left(\sum_k \frac{n_k}{n}\left(\beta^\top v_k - \bar{B}_k\beta^\top w_k\right)\right).$$

Solving for β we get the solution,

$$\beta^* = \frac{1}{2}(X^\top X)^{-1}\Big[2X^\top y + \mu_L\left((b^\top b)^{-1}(b^\top X)\right) - \mu_{b|B}\left(\frac{n_1}{n}\left(v_1 - \bar{b}_1 w_1\right) + \frac{n_0}{n}\left(v_0 - \bar{b}_0 w_0\right)\right) - \mu_{B|b}\left(\sum_k \frac{n_k}{n}\left(v_k - \bar{B}_k w_k\right)\right)\Big].$$

APPENDIX C
DATA

A. L2 Data Description

We select seven features as predictors in our model based on data completeness and predictive value: gender, age, estimated household income, estimated area median household income, estimated home value, area median education, and estimated area median housing value. While L2 provides a handful of other variables that point to political participation (e.g., interest in current events or number of political contributions), these features suffer from issues of data quality and completeness. For instance, only 15% of voters have a non-null value for interest in current events. We winsorize voters with an estimated household income of greater than $250,000 (4%) of the dataset. Table II shows the distribution of these characteristics, as well as the number of datapoints, for each of the states we consider. In general, across the six states, a little more than half of voters are female, and the average age hovers at around 50. There is high variance across income indicators, though the mean education level attained in all states is just longer than 12 years (a little past high school). Voting rates range from 53% in Georgia to 62% in North Carolina, while Black voters comprise a minority of all voters in each state, anywhere from 16% in Florida to 35% in Louisiana and Georgia.

B. Race Probabilities

The decennial Census in 2010 provides the probabilities of race given common surnames, as well as the probabilities of geography (at the census block group level) given race. In order to incorporate BIFSG, we also use the dataset provided by [57] which has the probabilities of common first names given race.

We default to using BIFSG for all voters but use BISG when a voter's first name is rare since we do not have data for them. Consequently, we only use geography instead of BISG when both one's first name and surname are rare. On average, around 70% of people's race across the six states were predicted using BIFSG, 10% using BISG, and 18% using just geography; $< 2\%$ of observations were dropped because we could not infer race probabilities using any of the three options.

Table III shows results for our BI(FS)G procedure with respect to true race. Accuracy and precision range from 80-90%, but recall is much lower at around 30-50%. Note, however, that we evaluate these metrics by binarizing race probabilities; in our estimators, we use raw probabilities instead, which provide a decent signal to true race. For instance, AUC hovers at 85-90%, while Figure 6 shows that our predicted probabilities are generally well-calibrated to true probability of Black (although BIFSG tends to overestimate the probability of Black).

APPENDIX D
DETAILS ON MEASUREMENT EXPERIMENTS

A. Voter Turnout Prediction Performance

Table IV shows results for voter turnout prediction on logistic regression and random forest models. In general, predicting voter turnout with the features given in L2 is a difficult task. Accuracy and precision hovers at around 70% throughout all experiments, while recall for logistic regression ranges from 71-82% and random forests perform slightly better at 80-90%. This result is in line with previous literature on predicting turnout, which suggest that "whether or not a person votes is to a large degree random" [59]. Note again that our predictors rely solely on demographic factors of voters because those are the most reliable data L2 provides us.

B. The KDC Method

In this section we expand on the different assumptions the KDC method and our method make related to the auxiliary data set. While we consider the case where the test set (with predicted outcomes and race probabilities) subsumes the auxiliary data (which contains true race), KDC mainly considers settings where the marginal distributions $\mathbb{P}(B,Z)$ and $\mathbb{P}(Y,\hat{Y},Z)$ are learned from two completely independent datasets – in particular, to estimate $\mathbb{P}(B|Z)$ and $\mathbb{P}(\hat{Y},Y|Z)$. Therefore, in order to produce a fairer comparison between the two methods, we instead reconfigure KDC to incorporate all the data available by treating the auxiliary data as a subset of our test set[4]; doing so only strengthens KDC because we pass in more information to learn both marginal distributions. However, their main method does not leverage information on $\mathbb{P}(Y,Z|B)$, as we do, so their bounds are notably wider. We

[4]Note that a component in calculating the variance of the KDC estimators is r, the proportion of datapoints from the marginal distribution $\mathbb{P}(Y,\hat{Y},Z)$ to the entire data. Without considering this independence assumption in our calculation, $r = 1$, but this loosely goes against the assumption that r is closer to 0 in Section 7 of [20]. For simplicity, we attenuate the multiplicative terms in the variance calculations of Equations 25 and 26 to give KDC the tightest bounds possible. However, as will be seen in Figure 1, KDC's incredibly large bounds are mostly attributed to its point estimates rather than their variances, which are quite small.

Feature	NC (n=6,305,309)	SC (n=3,191,254)	LA (n=2,678,258)	GA (n=6,686,846)	AL (n=3,197,735)	FL (n=13,703,026)
Gender (F)	0.54 (0.5)	0.54 (0.5)	0.55 (0.5)	0.53 (0.5)	0.54 (0.5)	0.53 (0.5)
Age	49.62 (18.76)	52.2 (18.69)	50.16 (18.29)	48.24 (18.07)	50.27 (18.44)	52.17 (18.89)
Est. Household (HH) Income	89,788.54 (56,880.78)	82,172.22 (53,886.64)	80,770.79 (54,579.77)	90,622.61 (57,699.76)	79,919.66 (52,237.42)	90,145.4 (56,786.94)
Est. Area Median HH Income	76,424.55 (32,239.45)	69,666.4 (25,911.0)	68,068.86 (29,779.93)	78,377.2 (35,941.68)	69,070.63 (27,226.34)	74,547.99 (29,820.33)
Est. Home Value	300,802.36 (202,634.22)	233,354.36 (155,221.32)	199,286.06 (123,564.26)	273,424.9 (176,273.9)	201,901.9 (126,255.0)	360,533.81 (243,854.1)
Area Median Education Year	12.83 (1.13)	12.64 (0.98)	12.36 (0.92)	12.72 (1.12)	12.51 (0.99)	12.65 (0.97)
Area Median Housing Value	206,312.82 (106,274.59)	193,172.13 (107,225.93)	170,521.45 (81,184.86)	206,253.25 (112,142.54)	162,925.8 (81,467.58)	237,245.18 (118,270.22)
Black	0.22	0.26	0.32	0.33	0.27	0.14
Vote in 2016	0.61	0.57	0.63	0.52	0.55	0.57

TABLE II: Distribution of features used for L2 across all six states: from left to right, North Carolina, South Carolina, Louisiana, Georgia, Alabama, and Florida. Each cell shows the mean of each feature and the standard deviation in parentheses. The last two rows show the proportion of observations that are black, and voted in the 2016 General Election.

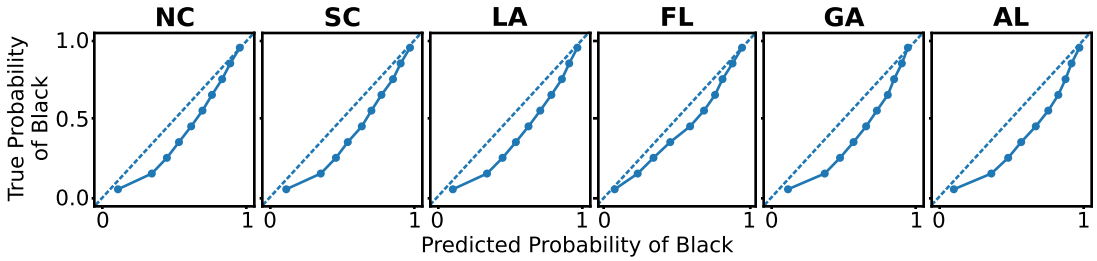


Fig. 6: Calibration plots showing predicted probability of Black (x-axis) versus actual proportion of Black (y-axis).

State	Accuracy	Precision	Recall	AUC
NC	0.83	0.77	0.30	0.85
SC	0.81	0.83	0.35	0.86
LA	0.82	0.87	0.52	0.89
GA	0.80	0.85	0.49	0.88
AL	0.84	0.89	0.45	0.90
FL	0.89	0.80	0.33	0.86

TABLE III: Accuracy, precision, recall (thresholded on 0.5), and AUC for BI(FS)G for all six states considered in L2.

also implement the KDC estimators as originally proposed in Figure 7 but the results do not change substantially[5].

C. Random Forest

We also run experiments on bounding disparity when voter turnout is predicted on random forest models, as seen in Figure 8. We observe similar results to logistic regression in that our methods always bound true disparity within 95% confidence intervals, and with bounds that are markedly tighter than KDC's. While our bounds are always within 5 p.p. and the same sign as true disparity, KDC is ranges from -0.5 to 0.5.

[5]In Appendix A.5, [20] do in fact propose an estimator where the independence assumption is violated (i.e., precisely the setting we consider where we have race probabilities in our entire data), but it suffers from two key limitations: *a)* we are only provided estimators for DD and none other disparity measure, and *b)* we implemented the DD estimator and it failed to bound true disparity in both applications we consider – see Figure 7.

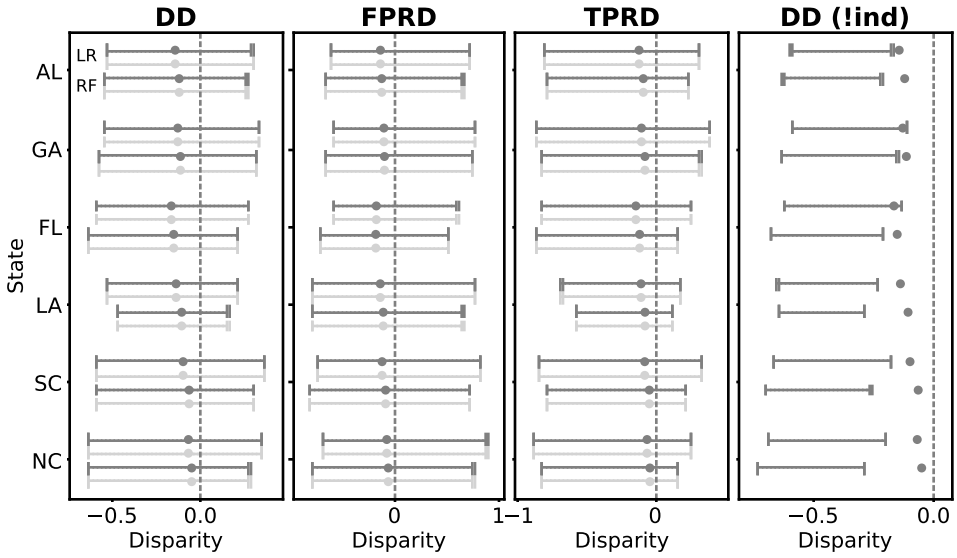


Fig. 7: Comparison of different KDC implementations. In dark grey, we have our implementation that violates the independence assumption in [20]. In light grey, we have KDC's original implementation with the independence assumption – nothing substantively changed. The top and bottom pairs of each state correspond to the estimators from logistic regression (LR) and random forest (RF) models, respectively. [20] additionally proposes estimators for estimating DD where the independence assumption is violated but they rarely bound true disparity (right subfigure), so we omit these results in our main experiments.

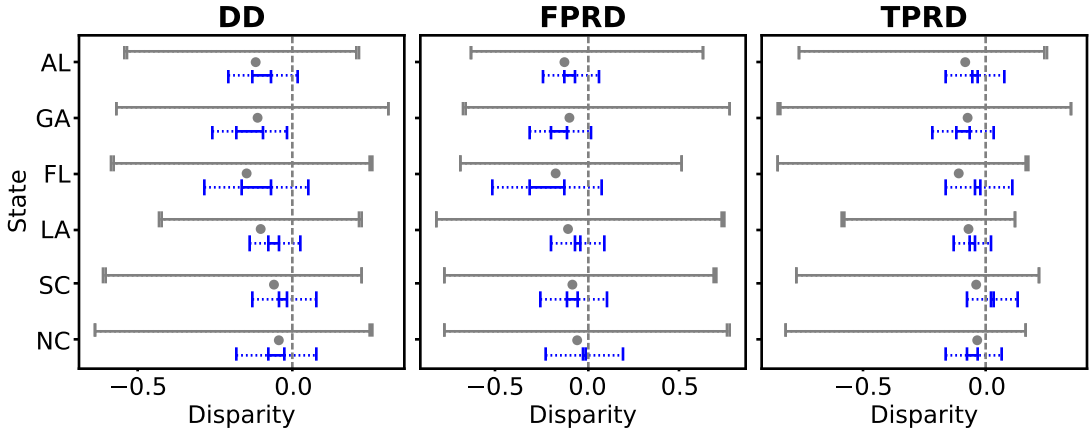


Fig. 8: Comparison of our method of bounding true disparity (blue) to the method proposed in [20] (grey), using a random forest model to predict voter turnout on L2 data in six states. We evaluate three disparity measures: demographic disparity (DD), false positive rate disp. (FPRD), and true positive rate disp. (TPRD). The grey dot represents true disparity. Both methods always bound true disparity within their 95% standard errors.

APPENDIX E
DETAILS ON TRAINING EXPERIMENTS

A. Experimental Setup

As noted in the main text, to enforce fairness constraints during training, we solve the empirical version of Problem 1.A and its symmetric analogue, which enforces negative covariance conditions and $\widehat{D}_{\mu}^{L}$ as a (negative) lower bound. For both of these problems we run the primal-dual algorithm described in Algorithm 1 for T iterations and then select the iteration from these two problems with the lowest loss on the training data while satisfying the constraints on the training and labeled subset.

We use the [14] method with logistic regression models under the hood.

B. CSL (Chamon et al.)

We implement our constrained problem using the official Pytorch implementation provided by [14].[6] for a logistic regression model We run the non-convex optimization problem

[6] https://github.com/lfochamon/csl

State	Model	Accuracy	Precision	Recall	AUC
NC	LR	0.72	0.75	0.81	0.75
	RF	0.72	0.72	0.89	0.76
SC	LR	0.67	0.69	0.77	0.71
	RF	0.67	0.67	0.86	0.71
LA	LR	0.70	0.73	0.84	0.72
	RF	0.70	0.71	0.91	0.73
GA	LR	0.69	0.70	0.71	0.75
	RF	0.69	0.68	0.78	0.75
AL	LR	0.67	0.69	0.74	0.72
	RF	0.67	0.67	0.80	0.72
FL	LR	0.67	0.69	0.76	0.71
	RF	0.67	0.67	0.85	0.72

TABLE IV: Accuracy, precision, recall, and AUC for voter turnout prediction for all six states considered in L2. We evaluate two different model performances for turnout prediction: logistic regression (LR) and random forests (RF).

for 1,000 iterations with a batch size of 1,024 and use Adam [60] for the gradient updates of the primal and dual problems with learning rates 0.001 and 0.005, respectively. We provide further explanation of the mathematical background to the [14] method in Appendix B above.

C. The Method of Wang et al.

[21] propose two methods to impose fairness with noisy labels: *1)* a distributionally robust optimization approach and *2)* another optimization approach using robust fairness constraints, which is based on [20]. We use code provided by [21] [7] to implement only the second method because it directly utilizes the protected attribute probabilities and yields better results.

We tune the following hyperparameters: $\eta_\theta \in \{0.001, 0.01, 0.1\}$ and $\eta_\lambda \in \{0.25, 0.5, 1, 2\}$, which correspond to the descent step for θ and the ascent step for λ in a zero-sum game between the θ-player and λ-player, see Algorithm 1 and 4 of [21]. Finally, we also tune $\eta_w \in \{0.001, 0.01, 0.1\}$, which is the ascent step for w (a component in the robust fairness criteria), see Algorithm 3 of [21]. In order to choose the best hyperparameters, we use the same data as outlined in Section IV-C1 (80/20 train/test split), but use a validation set on 30% of the training data (i.e., 24% of the entire data). Note that as implemented in the codebase, [21] chooses the hyperparameter that results in the lowest loss while adhering to the fairness constraint with respect to **true race**. Since we assume access to true race on a small subset (1%) of the data, we only evaluate the fairness constraint on 1% of the validation set.

D. The Method of Mozannar et al.

[24] primarily focus on the setting of training a fair model with differentially private demograpghic data, which poses assumptions which are infeasible for our setting—however, the authors do propose a potential extension of their method to handle a case that matches ours: training a fair model with incomplete demographic data. The authors do not discuss this in detail or provide the code for this extension, so we modify the code [24] provide for their paper to implement the extension of their approach, detailed in Section 6 of their paper that is relevant for our setting. This involves using Fairlearn's [8] exponentiated gradient method changed so that it will only update for its fairness-related loss on data points in the labeled subset, but allows classification loss to be calculated over the entire training set.

We note that Mozannar's method guarantees fairness violation 2(epsilon + best gap) [50] on their test set where epsilon is set by the user, but gives no method of approximating best_gap. Thus, we set epsilon = $\alpha/2$ (i.e. assume best gap=0) in our experiments in order to come as close as possible to their method providing similar fairness bounds to ours on the test set.

E. Pareto-Frontier of Accuracy vs. Disparity

F. Results on Oracle and Naive

In Figure 13 we present the mean and standard deviation of the resulting disparity and on the test set, as well as classifier accuracy on the test set, of experiments with our method compared to an oracle model, that has access to ground truth race on the *whole* dataset and uses these to enforce a constraint directly on ground truth disparity during training, as well as a naive model which simply enforces a constrained directly on the observed disparity of the noisy labels, without any correction. (Namely, in this technique, we simply threshold the probabilistic predictions of race on 0.5 to make them binary, and use as race labels.) As a whole, we perform relatively comparably to the oracle, except on FPRD. We always outperform the naive method in terms of reducing disparity, which is to be expected. We typically perform within 2 percentage points of accuracy from the oracle, (except for the 0.04 and 0.06 bounds on DD and the 0.04 bound on TPRD). We suggest the accuracy results in this figure show the fairness-accuracy trade-off in this setting: when we dip below the oracle in terms of accuracy, it is most often because we are bounding disparity lower than the oracle is (e.g., on the 0.04 bounds in DD or TPRD). And, while we do not outperform the naive method in terms of accuracy, we consistently out-perform it in terms of disparity.

APPENDIX F
ADDITIONAL EXPERIMENTS: COMPAS

In this section, we present a suite of additional experiments we run on the COMPAS [45] dataset. The COMPAS algorithm is used by parole officers and judges across the United States to determine a criminal's risk of recidivism, or re-committing the same crime. In 2016, ProPublica released a seminal article [45] detailing how the algorithm is systematically biased against Black defendants. The dataset used to train the algorithm has

[7] https://github.com/wenshuoguo/robust-fairness-code

[8] https://fairlearn.org/

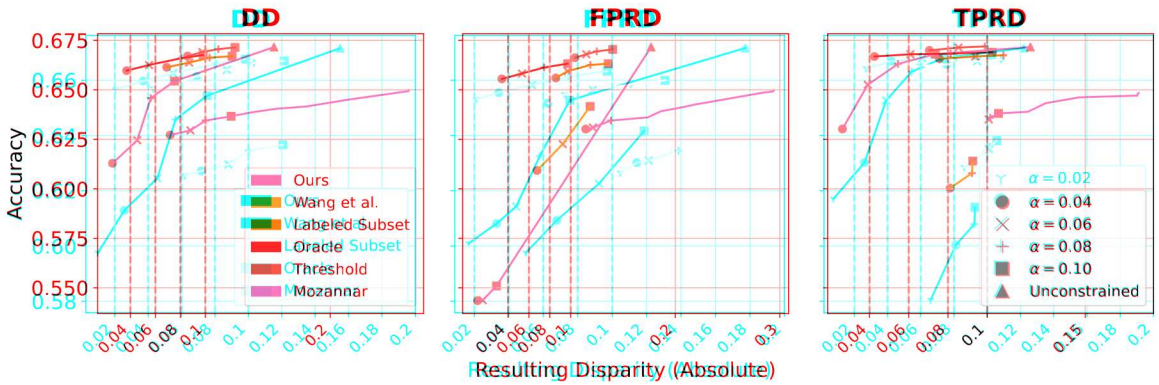


Fig. 9: Resulting disparity (x-axis) and accuracy (y-axis) trade-off for L2 Florida data. Each point corresponds to the average result over 10 seeds on a given target disparity α, which map to different marker styles (e.g., circle points are experiments with target disparity of 0.04). For ease of interpretation, the four target disparities are marked in dashed vertical lines; e.g., any circle point to the left of 0.04 satisfies the desired target disparity.

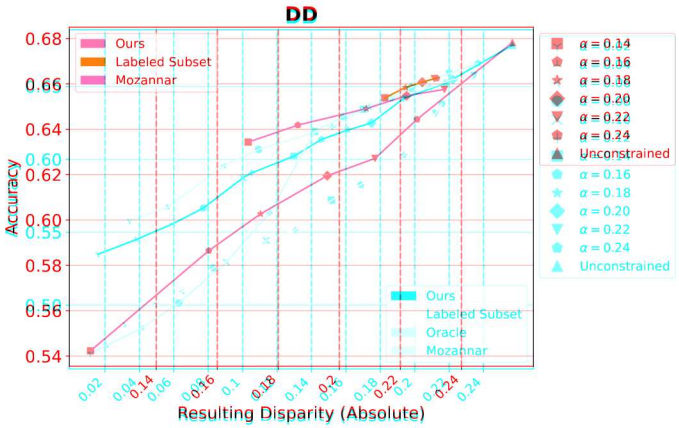


Fig. 10: Resulting demographic disparity (x-axis) and accuracy (y-axis) trade-off for COMPAS data. Each point corresponds to the average result over 10 seeds on a given target disparity α, which map to different marker styles (e.g., square points are experiments with target disparity of 0.14). For ease of interpretation, the four target disparities are marked in dashed vertical lines; e.g., any circle point to the left of 0.14 satisfies the desired target disparity.

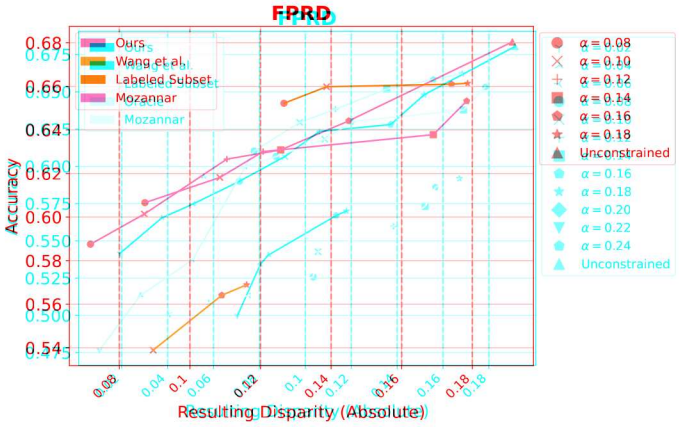


Fig. 11: Resulting false positive rate disparity (x-axis) and accuracy (y-axis) trade-off for COMPAS data. Each point corresponds to the average result over 10 seeds on a given target disparity α, which map to different marker styles (e.g., circle points are experiments with target disparity of 0.08). For ease of interpretation, the four target disparities are marked in dashed vertical lines; e.g., any circle point to the left of 0.08 satisfies the desired target disparity.

Metric	State	Method	Lower Bound (95% CI)	True Disparity	Upper Bound (95% CI)
DP	AL	KDC	-0.52 ± 0.01	-0.14	0.23 ± 0.01
		Ours	-0.14 ± 0.09	-0.14	-0.08 ± 0.09
	FL	KDC	-0.55 ± 0.01	-0.16	0.28 ± 0.01
		Ours	-0.27 ± 0.13	-0.16	-0.12 ± 0.13
	GA	KDC	-0.55 ± 0.01	-0.13	0.32 ± 0.01
		Ours	-0.22 ± 0.08	-0.13	-0.12 ± 0.08
	LA	KDC	-0.53 ± 0.01	-0.14	0.25 ± 0.01
		Ours	-0.12 ± 0.07	-0.14	-0.07 ± 0.07
	NC	KDC	-0.62 ± 0.01	-0.07	0.32 ± 0.01
		Ours	-0.13 ± 0.12	-0.07	-0.05 ± 0.12
	SC	KDC	-0.61 ± 0.01	-0.1	0.28 ± 0.01
		Ours	-0.08 ± 0.1	-0.1	-0.03 ± 0.1
FPR	AL	KDC	-0.58 ± 0.01	-0.14	0.69 ± 0.01
		Ours	-0.14 ± 0.13	-0.14	-0.08 ± 0.13
	FL	KDC	-0.57 ± 0.01	-0.16	0.6 ± 0.01
		Ours	-0.31 ± 0.21	-0.16	-0.13 ± 0.21
	GA	KDC	-0.59 ± 0.01	-0.1	0.77 ± 0.01
		Ours	-0.22 ± 0.11	-0.1	-0.12 ± 0.11
	LA	KDC	-0.81 ± 0.01	-0.13	0.85 ± 0.02
		Ours	-0.08 ± 0.13	-0.13	-0.05 ± 0.13
	NC	KDC	-0.65 ± 0.01	-0.07	0.86 ± 0.01
		Ours	-0.07 ± 0.21	-0.07	-0.03 ± 0.2
	SC	KDC	-0.69 ± 0.01	-0.12	0.77 ± 0.01
		Ours	-0.14 ± 0.15	-0.12	-0.06 ± 0.15
TPR	AL	KDC	-0.78 ± 0.01	-0.12	0.3 ± 0.01
		Ours	-0.07 ± 0.11	-0.12	-0.04 ± 0.11
	FL	KDC	-0.8 ± 0.01	-0.14	0.25 ± 0.0
		Ours	-0.21 ± 0.15	-0.14	-0.1 ± 0.15
	GA	KDC	-0.88 ± 0.01	-0.11	0.4 ± 0.01
		Ours	-0.18 ± 0.11	-0.11	-0.1 ± 0.11
	LA	KDC	-0.68 ± 0.01	-0.1	0.2 ± 0.0
		Ours	-0.14 ± 0.08	-0.1	-0.08 ± 0.08
	NC	KDC	-0.86 ± 0.01	-0.06	0.25 ± 0.0
		Ours	-0.12 ± 0.12	-0.06	-0.05 ± 0.12
	SC	KDC	-0.84 ± 0.01	-0.08	0.31 ± 0.0
		Ours	-0.0 ± 0.12	-0.08	-0.0 ± 0.12

TABLE V: Companion table to Figure 1.

since been widely used as benchmarks in the fair machine learning literature.

A. Data Description

We use the eight features used in previous analyses of the dataset as predictors in our model: the decile of the COMPAS score, the decile of the predicted COMPAS score, the number of prior crimes committed, the number of days before screening arrest, the number of days spent in jail, an indicator for whether the crime committed was a felony, age split into categories, and the score in categorical form. We process the data following [45], resulting in $n = 6,128$ datapoints. Table VI outlines the feature distribution of the dataset.

B. Race Probabilities

We generate estimates of race (Black vs. non-Black) based on first name and last name using a LSTM model used in [49] that was trained on voter rolls from Florida. The predictive performance and calibration of these estimates is displayed in Table VII and Figure 14, respectively. In general, the results are quite reasonable; accuracy is at 73% while the AUC is 86%. The probabilities are somewhat calibrated, although the LSTM model tends to overestimate the probability of Black.

C. Measurement Experiments

We first compare our method of bounding disparity to that of KDC. We train an unconstrained logistic regression model

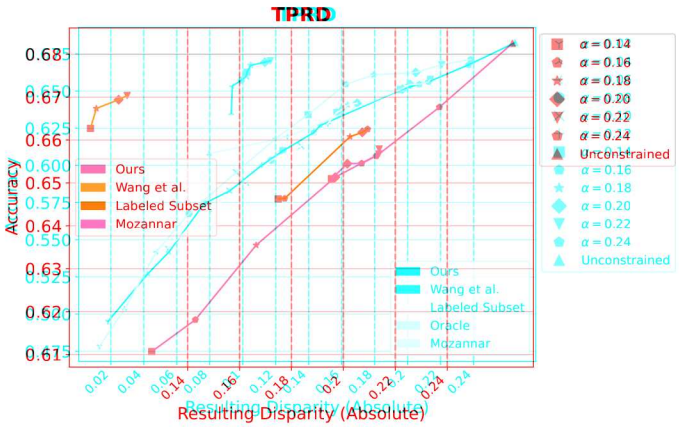


Fig. 12: Resulting true positive rate disparity (x-axis) and accuracy (y-axis) trade-off for COMPAS data. Each point corresponds to the average result over 10 seeds on a given target disparity α, which map to different marker styles (e.g., square points are experiments with target disparity of 0.14). For ease of interpretation, each of the target disparities are marked in dashed vertical lines; e.g., any circle point to the left of 0.14 satisfies the desired target disparity.

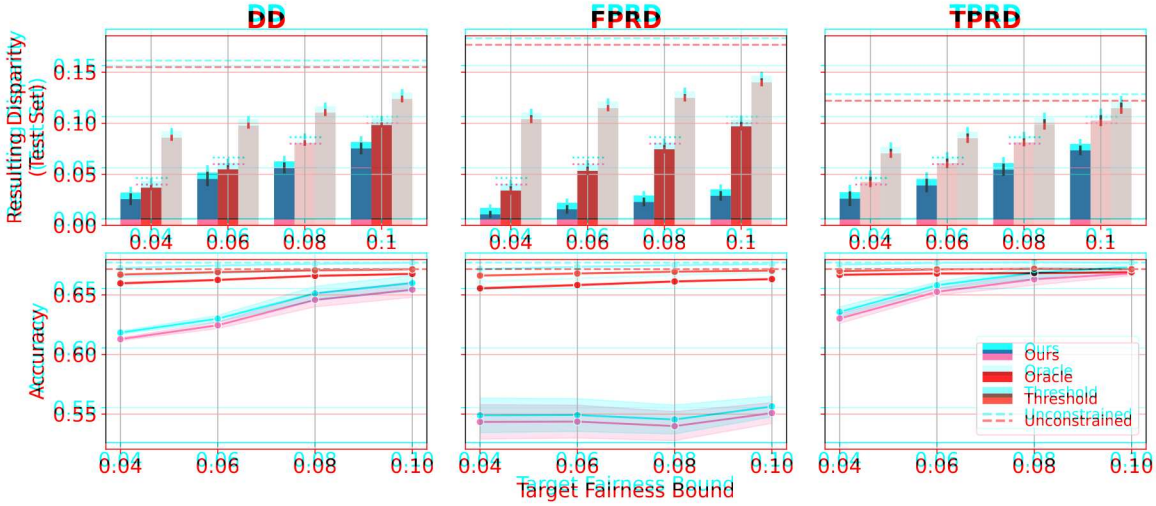


Fig. 13: Mean and standard deviation of resulting disparity (top, y-axis) and accuracy (bottom, y-axis) on the test set after enforcing the target fairness bounds (x-axis) on our method (blue); using ground truth race on the entire data, i.e., "oracle" model (red); and using only the estimated race probabilities, thresholded to be binary (brown) 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.

with a 80/20 split on the data, i.e., $n = 1,226$ in the test set. Then, we construct the labeled subset by sampling 50% of the test set ($n = 613$) and use that to check out covariance constraints. We also compute $\hat{D}_L$ and $\hat{D}_P$ with standard errors on the entire test set, as specified by the procedure in Appendix Section D.

Our main results are displayed in Figure 2. Similar to the L2 data, our bounds are consistently tighter than KDC, albeit to a lesser extent in this case since the COMPAS dataset is significantly smaller. Despite this fact, we emphasize that, unlike KDC, our estimators are always within the same sign as the true disparity, barring the standard errors which shrink as the data grows larger.

D. Training Experiments Details

We compare our training method to [21], [24] and a baseline where we directly enforce disparity constraints on only the labeled subset. We run 10 trials – each corresponding to different seeds – and report the mean and standard deviation of the accuracy and disparity on the test set in Figure 4. For each trial, we split our data ($n = 6,128$) into train and test sets, with a 80/20 split. From the training set, we subsample the labeled subset so that it is 10% of the total data (around $n = 613$). We chose a higher proportion of the data compared to L2 to adjust for the smaller dataset. The remaining details are as described in Section IV-C1. Note that the resulting disparities for the unconstrained model differ among the three fairness

Feature	COMPAS (n=6,128)
Decile Score	4.41 (2.84)
Predited Decile Score	3.64 (2.49)
# of Priors	3.23 (4.72)
# of Days Before Screening Arrest	-1.75 (5.05)
Length of Stay in Jail (Hours)	361.26 (1,118.60)
Crime is a Felony	0.64 (0.48)
Age Category	0.65 (0.82)
Risk Score in 3 Levels	1.08 (0.66)
Black	0.51
Two Year Recidivism	0.45

TABLE VI: Distribution of features used for COMPAS. Each cell shows the mean of each feature and the standard deviation in parentheses. The last two rows show the proportion of observations that are Black and who recidivized within two years.

	Accuracy	Precision	Recall	AUC
Black	0.73	0.86	0.56	0.86

TABLE VII: Accuracy, precision, recall (thresholded on 0.5), and AUC for predicting probability a person is Black in the COMPAS dataset.

metrics. On DD and TPRD, the unconstrained model resulted in a 0.28-0.29 disparity, but it drops to 0.21 for FPRD. We adjusted our target fairness bounds accordingly.

Appendix G
Simulations

A. Simulation Design

In this section, we describe the design of our simulation used for additional experiments.

- Primitive features $Z_1, \ldots, Z_m$
- Conditional probability b of being Black a function of $Z_1 \ldots Z_m$
- Realized status as Black or not B drawn from Bernoulli(b)
- Downstream features $X_1, \ldots X_p$, a function of $Z_1, \ldots, Z_m$ and B

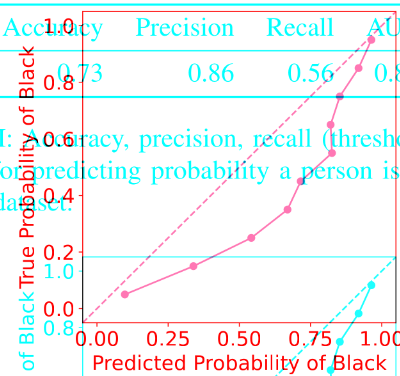


Fig. 14: Calibration plot showing the predicted probability a person in the dataset is Black (x-axis) versus the actual proportion of Black people in the dataset (y-axis) for COMPAS.

Accuracy	Precision	Recall	AUC
0.69	0.69	0.57	0.74

TABLE VIII: Accuracy, precision, recall (thresholded on 0.5), and AUC for predicting two-year recidivism on the COMPAS dataset using a logistic regression model.

- Score for outcome $P(Y)$, a function of downstream features $X_1, \ldots X_p$
- Outcome Y, which is an indicator of $P(Y)$ at threshold τ with some noise probability of being flipped $0 \leftrightarrow 1$

The primitive features $Z_1, \ldots, Z_p$ intuitively represent the variables that correspond to proxies in BIFSG, e.g. geographic locations. They serve a dual role: first, as in BIFSG, they give rise to the probability that an individual is Black. Second, since the secondary features X are a function of Z, they affect the distribution of these features; thus downstream, they affect $P(Y)$ and ultimately Y, but do not directly enter into $P(Y)$ or Y themselves. This corresponds to how geography and other variables which are correlated to race may also be correlated to many learning-relevant features, even when not directly entering causing the outcome of interest themselves. Note that in addition to primitives affecting $P(Y)$ through each X, we allow for B to affect $P(Y)$. This corresponds to how there may be associations between group membership and features which affect the outcome of the interest via the downstream features even if the group status is not directly relevant to the outcome of interest.

These relationships are not fully specified by the description in the text above, of course, so we provide details of the selected functional forms in Table IX. Figure 12 also summarizes the features and their associative relationships visually. This visualization, along with the language of directed acyclic graphs (DAGs), allows us to more easily reason about whether the covariance conditions are likely to be satisfied in our model, at least for the underlying outcome.

Feature	Interpretation	Functional Form
Z_j	Primitive Feature	$Z_j \sim U[0,1],\ j = 1,...m$
X_i	Secondary Feature	$X_i = \sum_{k=1}^{h_k} c_i X_i^k,\ i = 1,...p$
h_k	Degree	$h_k \sim U\{0,1,2,3\}$
c_i	Coefficients	$c_i \sim U[0,1],\ i = 1,...p$
b	Probability Black	$b = \max\{0, \min\{1, \tilde{b}\}\},$ $\tilde{b} \sim \begin{cases} \mathcal{N}(0.1, .04) & \frac{1}{m}\sum_{j=1}^{m} Z_j \leq \tau_b \\ \mathcal{N}(0.9, .04) & \frac{1}{m}\sum_{j=1}^{m} Z_j > \tau_b \end{cases}$
τ_b	Threshold on b (based Irwin-Hall distribution)	$\frac{1}{2} + 1.2\sqrt{1/(12m)}$
B	Indicator for Black	$B \sim \text{Bernoulli}(b)$
$\tilde{P}(Y)$	Score of Outcome	$\tilde{P}(Y) = \sum_i^X [d_i X_i^k + d_{iB} B]$
$P(Y)$	Normalized Score of Outcome	$P(Y) = \frac{\tilde{P}(Y) - \min(\tilde{P}(Y))}{\max(\tilde{P}(Y)) - \min(\tilde{P}(Y))}$
Y	Realized Outcome	$Y \sim \begin{cases} \text{Bernoulli}(0.1) & P(Y) \leq \tau \\ \text{Bernoulli}(0.9) & P(Y) > \tau \end{cases}$
d_i	Coefficients for features X	$d_i \sim U[0,1]$
d_{iB}	Coefficients for indicator for Black	$d_{iB} \sim U[0, u_B]$

TABLE IX: Description of several variables we use in our simulation study and their functional forms. For ease of notation, we omit the index denoting individuals in the dataset. Unspecified constants were selected by inspection to match key indicators across scenario and are specified in Table 8.

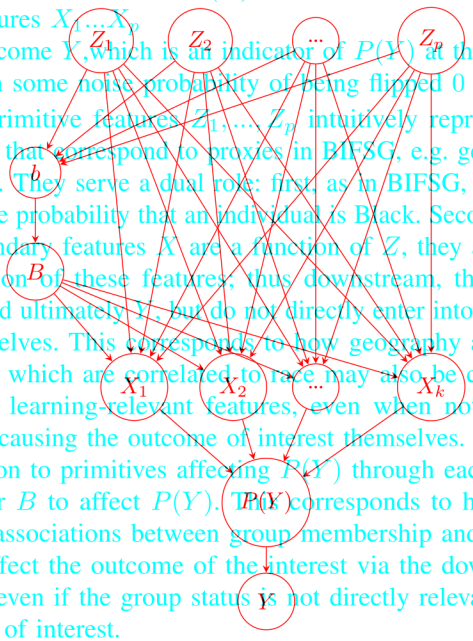


Fig. 15: A heuristic depiction of the data generating process for our simulations. Nodes indicate random variables, and edges indicate (causal) relationships between nodes. Importantly, relationships are not necessarily linear.

APPENDIX G
SIMULATIONS

A. Simulation Design

In this section, we describe the design of our simulation used for additional experiments.

- Primitive features $Z_1, ..., Z_m$
- Conditional probability b of being Black a function of $Z_1, ..., Z_m$
- Realized status as Black or not B drawn from Bernoulli(b)
- Downstream features $X_1, ..., X_p$, function of $Z_1, ..., Z_m$ and B
- Score for outcome $P(Y)$, a function of downstream features $X_1 ... X_p$
- Outcome Y,which is an indicator of $P(Y)$ at threshold τ with some noise probability of being flipped $0 \leftrightarrow 1$

The primitive features $Z_1, ..., Z_p$ intuitively represent the variables that correspond to proxies in BIFSG, e.g. geographic locations. They serve a dual role: first, as in BIFSG, they give rise to the probability that an individual is Black. Second, since the secondary features X are a function of Z, they affect the distribution of these features, thus downstream, they affect $P(Y)$ and ultimately Y, but do not directly enter into $P(Y)$ or Y themselves. This corresponds to how geography and other variables which are correlated to race may also be correlated to many learning-relevant features, even when not directly entering causing the outcome of interest themselves. Note that in addition to primitives affecting $P(Y)$ through each X, we allow for B to affect $P(Y)$. This corresponds to how there may be associations between group membership and features which affect the outcome of the interest via the downstream features even if the group status is not directly relevant tot he outcome of interest.

These relationships are not fully specified by the description in the text above, of course, so we provide details of the selected functional forms in Table IX. Figure 12 also summarizes the features and their associative relationships visually. This visualization, along with the language of directed acyclic graphs (DAGs), allows us to more easily reason about whether the covariance conditions are likely to be satisfied in our model, at least for the underlying outcome.

B. Experimental Setup

Following the notation above, we have p to be the number of features X in our data, and let n be the number of datapoints. We run experiments for $p \in \{10, 20, 50\}$ and $n \in \{5000, 10000, 50000\}$. For each p, we fix the parameters in the data generation process and realize 50,000 datapoints. Refer to Table X for a list of parameter values, which differ slightly for each p to control demographic disparity on the dataset at around 0.25-0.28. For experiments $n \in \{5000, 10000\}$, we simply randomly subsample from the 50,000 dataset.

p	m	τ	u_B
10	4	0.4	0.05
20	5	0.4	0.1
50	10	0.425	0.2

TABLE X: List of parameters in the data generation process for each p, the number of secondary features X in the data. m corresponds to the number of primitive features Z, τ is the threshold for $P(Y)$, while u_B is the upper bound for the uniform distribution to generate d_{iB}, see Table IX.

The last dimension we tune is the size of the labeled subset (measured by the percentage of n), which from hereon we refer to as ϵ. For each n, we specified slightly different ϵ as outlined in Table XI. This is to account for the fact that, for instance, one might need 40% of 5,000 datapoints with protected attribute labels to learn a predictor that reaches the target disparity bound. On the other hand, using 20% of 50,000 datapoints might be more than enough, especially considering the exponentially higher costs to query thousands of people's protected attributes.

We prototype these simulation experiments on demographic parity. For each experiment, we split the data 80/20 into train/test data, then repeat 10 times with different seeds. We run both our method and the labeled subset method, evaluating

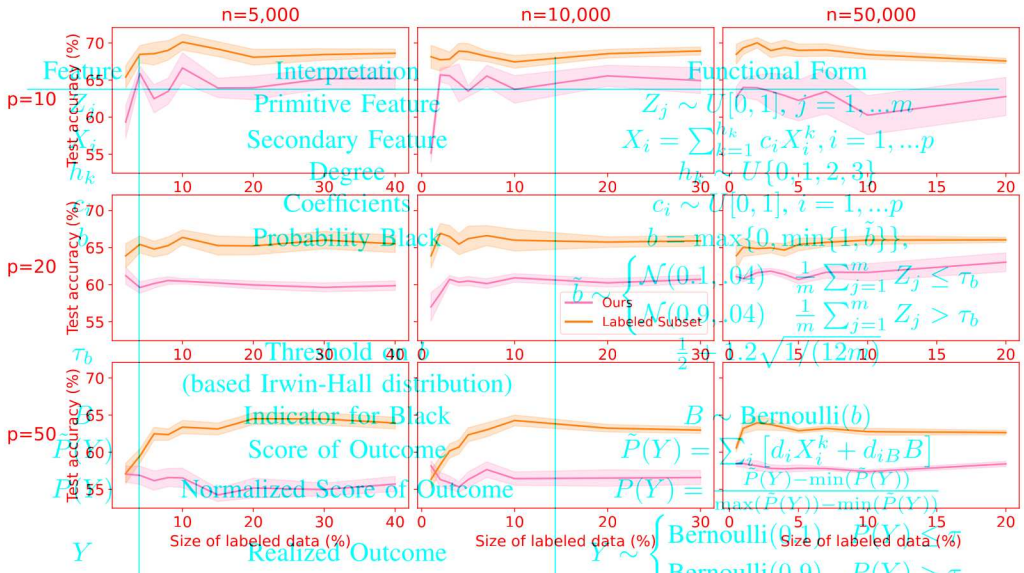


Fig. 16: We present a three by three figure showing the test accuracy of the models created using our disparity reduction method when compared with relying on training models only on the labeled subset and reducing disparity by directly enforcing a constraint on the protected attribute labels. The rows correspond to datasets of increasing sizes (number of features from 10 to 50), indicating problems of increasing complexity. The columns correspond to the size of the overall dataset, ranging from 5,000 to 50,000 samples. The x-axis shows the percentage of the total dataset is dedicated to the labeled subset, and the y-axis denotes the test accuracy of the models. The blue graphs correspond to our method, and the orange to the labeled subset method.

n	e
5,000	$\{2, 4, 6, 8, 10, 15, 20, 30, 40\}$
10,000	$\{1, 2, 3, 4, 5, 7, 10, 20, 30\}$
50,000	$\{0.5, 1, 2, 3, 4, 5, 7, 10, 20\}$

TABLE XI: Suite of experiments varying percentage of the data taken as labeled subset (e) by the size of the full dataset (n).

disparity and accuracy on the test set.

C. Results

We present our results in Figures 5 and 16. In Figure 5, we see that while increasing the size of the labeled subset can sometimes lead to a regime where training on the labeled subset alone can produce a model which comes close to (or in one case–$n = 50,000$, $p = 10$, reaches) the desired disparity bound, for the most part, even with a large labeled subset, the mean of the disparity over 10 trials is above the desired disparity threshold. Meanwhile, our method stays below the desired disparity threshold across all nine experiments.

As we can see by looking at the rows from top to bottom, the complex (i.e., more features in the data) the problem is, the more data is necessary for the labeled subset to get close to the desired disparity bound. Thus, our simulation experiment sheds light on the fact that model applications with small amounts of labeled data, and more complex data, are particularly well-suited for our method.

APPENDIX H
PAST REVIEWS AND RESPONSES

We enclose our unedited reviews and responses from a prior submission. We provide here a summary of our additions and changes from our initial version of the paper:

- **Additional Datasets**: We provide experiments on the COMPAS dataset, using a non-BIFSG proxy as our probabilistic estimate. We also provide experiments on synthetic data.
- **Additional Methods**: We incorporate a comparison to Mozannar et al. [24], which, to our knowledge, is the only paper that addresses a setting that is close to ours. We also discuss why other methods are inapplicable to our setting.
- **Simulation Study:** We conduct a study on simulated data where we can vary both dataset size and underlying complexity of the data-generating process to understand under what regimes our method is preferable to existing methods.

We also add a variety of clarification measures, including adding Table II-A to the main paper; adding a section to the introduction which clarified the real-world situations where our technique may be used, and setting some of the stakes for our paper, and others.

Additional Datasets: In response to reviewer concerns surrounding the number of datasets considered, we provide experimental results of our method over COMPAS data, which had not originally been in the paper. We also note that while the paper focused on the application of voter turnout, the 2 states for which we run learning experiments and the 7 states for which we perform measurement experiments all have significantly different data distributions, providing insight to different data settings.

Additional Baselines: In response to one reviewer, we added a comparison to Mozannar et al. [24] in our main set of results. As we show in the paper, we almost always outperform their method on disparity reduction and accuracy. We address the particular comparisons the reviewer brought up, and why the Mozannar paper was the only comparison which shed additional light to the paper, here:

Our setting differs substantially from the papers cited by the reviewer and the literature more broadly. In particular, while some papers assume noisy, perturbed, or nonexistent protected attribute data, we consider an empirically common "mixed" case, in which the learner has access to a probabilistic proxy for the protected attribute over all their data, and ground truth over a small subset only, spurring the question of how best to combine the two data sources with fairness and performance in mind.

With respect to the works the reviewer specifically highlighted: Mozannar et. al [24], primarily focuses on a setting with a very strong conditional independence assumption on the protected attribute proxy, which is unlikely to be met besides the (infeasible, for our setting) case in which the learner can differentially privately generate noisy labels using the ground truth. However, as you point out, the authors do propose a potential extension of their method to handle a case that matches ours. The authors do not discuss this in detail or provide the code for this extension, but we have modified their original code to implement the changes described and include these results in our updated draft. We note that Mozannar et. al.'s method almost never meets the disparity bounds, and does worse in accuracy on L2; see Figures 1-2 attached. Levy et. al [61] proposes and analyzes distributionally robust optimization (DRO) algorithms. At its core, the idea of DRO is to optimize for the worst-case distribution that is "nearby" according to some metrics. This has applications to fairness, as the reviewer points out. In particular, Wang et al, to which we have already provided a comparison, applies DRO to precisely the problem of fair classification with noisy sensitive features. Like other methods that aim for worst-case guarantees, the method is robust but sacrifices performance by not using all available information, as can be seen in Figure 2 in our main results. Diana et. al. [62] aims at an upstream and somewhat orthogonal problem to that of ours. Our paper focuses on training a fair model using an imperfect proxy (e.g. BIFSG), where as Diana et al. study how to learn a model to create proxy features with which to learn a fair model. They do not focus on the downstream task of training the fair model, and the methods are hence not directly comparable. We are happy to include discussion of the above works in the related work section of our paper.

Investigating Questions Around the Size of the Labeled Subset: Prompted by the response of several reviewers, we added experiments on the relationship between the size of the labeled subset, dataset complexity, and the efficacy of our method which we pointed to as future work in our limitations section. Our results are in Section IV-D of the paper. Overall, our method is consistently able to bound disparity across varying sizes of the labeled subset, whereas training on the labeled subset alone only comes close to bounding disparity in low-complexity regimes.

We now enclose the unedited reviews and our responses.

APPENDIX H
PAST REVIEWS AND RESPONSES

We enclose our unedited reviews and responses from a prior submission. We provide here a summary of our additions and changes from our initial version of the paper:

- **Additional Datasets**: We provide experiments on the COMPAS dataset, using a non-BIFSG proxy as our probabilistic estimate. We also provide experiments on synthetic data.
- **Additional Methods**: We incorporate a comparison to Mozannar et al. [24], which, to our knowledge, is the only paper that addresses a setting that is close to ours. We also discuss why other methods are inapplicable to our setting.
- **Simulation Study:** We conduct a study on simulated data where we can vary both dataset size and underlying complexity of the data-generating process to understand under what regimes our method is preferable to existing methods.

We also add a variety of clarification measures, including adding Table II-A to the main paper; adding a section to the introduction which clarified the real-world situations where our technique may be used, and setting some of the stakes for our paper, and others.

Additional Datasets: In response to reviewer concerns surrounding the number of datasets considered, we provide experimental results of our method over COMPAS data, which had not originally been in the paper. We also note that while the paper focused on the application of voter turnout, the 2 states for which we run learning experiments and the 7 states for which we perform measurement experiments all have significantly different data distributions, providing insight to different data settings.

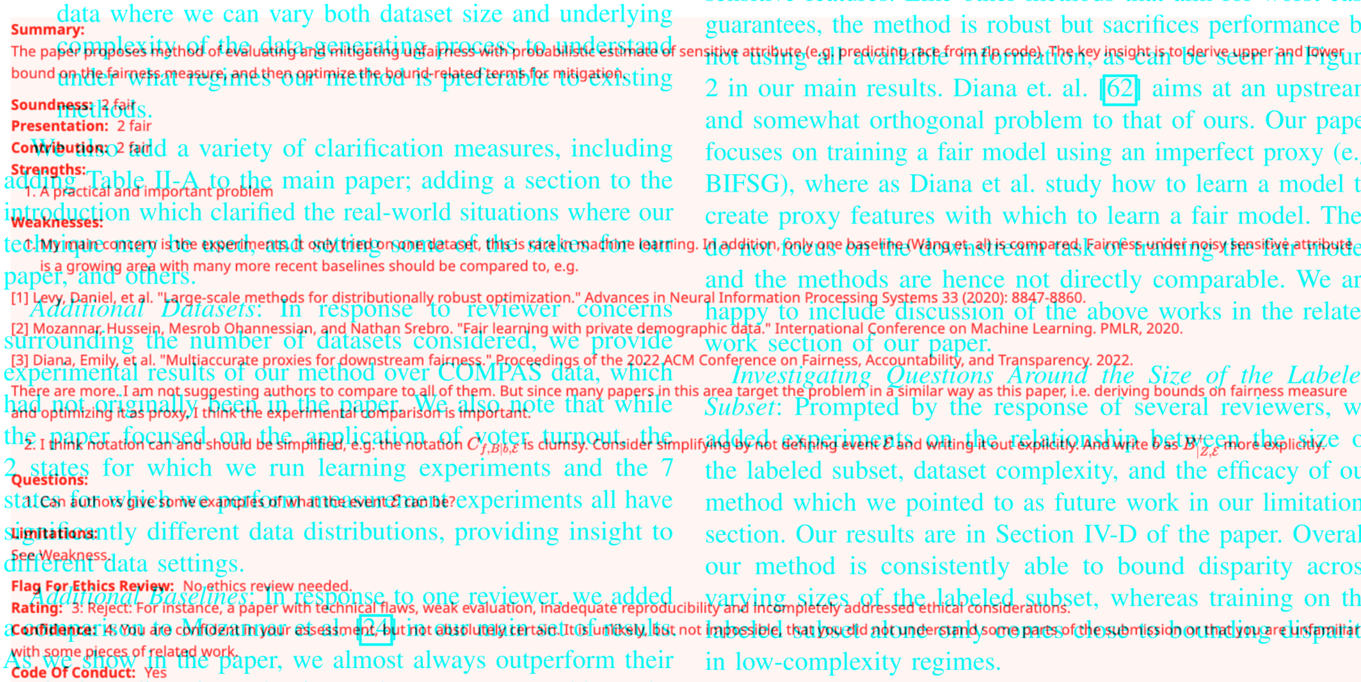
Additional Baselines: In response to one reviewer, we added a comparison to Mozannar et al. [24] in our main set of results. As we show in the paper, we almost always outperform their method on disparity reduction and accuracy. We address the particular comparisons the reviewer brought up, and why the Mozannar paper was the only comparison which shed additional light to the paper, here:

Our setting differs substantially from the papers cited by the reviewer and the literature more broadly. In particular, while some papers assume noisy, perturbed, or nonexistent protected attribute data, we consider an empirically common "mixed" case, in which the learner has access to a probabilistic proxy for the protected attribute over all their data, and ground truth over a small subset only, spurring the question of how best to combine the two data sources with fairness and performance in mind.

With respect to the works the reviewer specifically highlighted: Mozannar et. al [24], primarily focuses on a setting with a very strong conditional independence assumption on the protected attribute proxy, which is unlikely to be met besides the (infeasible, for our setting) case in which the learner can differentially privately generate noisy labels using the ground truth. However, as you point out, the authors do propose a potential extension of their method to handle a case that matches ours. The authors do not discuss this in detail or provide the code for this extension, but we have modified their original code to implement the changes described and include these results in our updated draft. We note that Mozannar et. al.'s method almost never meets the disparity bounds, and does worse in accuracy on L2; see Figures 1-2 attached. Levy et. al [61] proposes and analyzes distributionally robust optimization (DRO) algorithms. At its core, the idea of DRO is to optimize for the worst-case distribution that is "nearby" according to some metrics. This has applications to fairness, as the reviewer points out. In particular, Wang et al, to which we have already provided a comparison, applies DRO to precisely the problem of fair classification with noisy sensitive features. Like other methods that aim for worst-case guarantees, the method is robust but sacrifices performance by not using all available information, as can be seen in Figure 2 in our main results. Diana et. al. [62] aims at an upstream and somewhat orthogonal problem to that of ours. Our paper focuses on training a fair model using an imperfect proxy (e.g. BIFSG), where as Diana et al. study how to learn a model to create proxy features with which to learn a fair model. They do not focus on the downstream task of training the fair model, and the methods are hence not directly comparable. We are happy to include discussion of the above works in the related work section of our paper.

Investigating Questions Around the Size of the Labeled Subset: Prompted by the response of several reviewers, we added experiments on the relationship between the size of the labeled subset, dataset complexity, and the efficacy of our method which we pointed to as future work in our limitations section. Our results are in Section IV-D of the paper. Overall, our method is consistently able to bound disparity across varying sizes of the labeled subset, whereas training on the labeled subset alone only comes close to bounding disparity in low-complexity regimes.

We now enclose the unedited reviews and our responses.



Summary:
The paper proposes method of evaluating and mitigating unfairness with probabilistic estimate of sensitive attribute (e.g. predicting race from zip code). The key insight is to derive upper and lower bound on the fairness measure, and then optimize the bound-related terms for mitigation.

Soundness: 2 fair
Presentation: 2 fair
Contribution: 2 fair
Strengths:

1. A practical and important problem

Weaknesses:

1. My main concern is the experiments. It only tried on one dataset, this is rare in machine learning. In addition, only one baseline (Wang et. al) is compared. Fairness under noisy sensitive attribute is a growing area with many more recent baselines should be compared to, e.g.

[1] Levy, Daniel, et al. "Large-scale methods for distributionally robust optimization." Advances in Neural Information Processing Systems 33 (2020): 8847-8860.
[2] Mozannar, Hussein, Mesrob Ohannessian, and Nathan Srebro. "Fair learning with private demographic data." International Conference on Machine Learning. PMLR, 2020.
[3] Diana, Emily, et al. "Multiaccurate proxies for downstream fairness." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.
There are more..I am not suggesting authors to compare to all of them. But since many papers in this area target the problem in a similar way as this paper, i.e. deriving bounds on fairness measure and optimizing it as proxy, I think the experimental comparison is important.

2. I think notation can and should be simplified, e.g. the notation $C_{f,B|b,\mathcal{E}}$ is clumsy. Consider simplifying by not defining event $\mathcal{E}$ and writing it out explicitly. And write b as $B'_{|Z,\mathcal{E}}$ more explicitly.

Questions:
1. Can authors give some examples of what the event $\mathcal{E}$ can be?

Limitations:
See Weakness.

Flag For Ethics Review: No ethics review needed.
Rating: 3: Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility and incompletely addressed ethical considerations.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Fig. 17

Rebuttal:

Thank you for your review. We have made the following updates, which we will describe in detail below:

- **We have added experiments on the COMPAS dataset, where our method performs better at decreasing disparity than the labeled subset method and Mozannar et al, which we include as a comparison.** We also note that the 2 states for which we run learning and the 7 states for which we perform measurement all have significantly different data distributions.
- **We have added a comparison to Mozannar et al. in our main results,** though there are substantial differences between our method and theirs, **where our method outperforms on bounding disparity in all experiments and accuracy in most.** We describe our differences from the other references pointed to below.

Additional Datasets

The experiments in the current draft of the paper are focused on one application, but over different populations: we consider data from 7 different states for measurement and 2 states for our predictive task. As we detail in appendices C and D, the states have significant variation in key metrics like the fraction Black and non-Black, turnout, and the ease of the underlying prediction problem, which provide informative variation over the space of possible problem settings.

We have selected this application because it is one with easily available public data in which individuals' names and addresses are linked to their data, allowing for the application of Bayesian Improved Firstname Surname Geocoding (BIFSG), which is likely the most common race proxy used for fairness in industry [2] and government [1]. The measurement of disparity in voting is also an important question of both academic interest and legal importance. However, as our method does not require the use of BIFSG to estimate a protected feature, we have added experiments on an additional dataset, COMPAS; for this application, we use an LSTM model used in [5] to estimate race (Black vs. non-Black) based on first name and last name instead of BIFSG. Our results are attached in Figure 2 of the global response. We show that our method is able to train a classifier that makes predictions which satisfy the target disparity threshold for several different threshold values, unlike the comparison methods which do not satisfy the target threshold except for the largest threshold value.

Additional Benchmarks

Our setting differs substantially from the papers cited by the reviewer and the literature more broadly. In particular, while some papers assume noisy, perturbed, or nonexistent protected attribute data, we consider an empirically common "mixed" case, in which the learner has access to a probabilistic proxy for the protected attribute over all their data, and ground truth over a small subset only, spurring the question of how best to combine the two data sources with fairness and performance in mind.

With respect to the works you specifically highlighted: Mozannar et. al, primarily focuses on a setting with a very strong conditional independence assumption on the protected attribute proxy, which is unlikely to be met besides the (infeasible, for our setting) case in which the learner can differentially privately generate noisy labels using the ground truth. However, as you point out, the authors do propose a potential extension of their method to handle a case that matches ours. The authors do not discuss this in detail or provide the code for this extension, but we have modified their original code to implement the changes described and include these results in our updated draft. We note that Mozannar et. al.'s method almost never meets the disparity bounds, and does worse in accuracy on L2; see Figures 1-2 attached. Levy et. al proposes and analyzes distributionally robust optimization (DRO) algorithms. At its core, the idea of DRO is to optimize for the worst-case distribution that is "nearby" according to some metrics. This has applications to fairness, as the reviewer points out. In particular, Wang et al, to which we have already provided a comparison, applies DRO to precisely the problem of fair classification with noisy sensitive features. Like other methods that aim for worst-case guarantees, the method is robust but sacrifices performance by not using all available information, as can be seen in Figure 2 in our main results.

Diana et. al. aims at an upstream and somewhat orthogonal problem to that of ours. Our paper focuses on training a fair model using an imperfect proxy (e.g. BIFSG), where as Diana et al. study how to learn a model to create proxy features with which to learn a fair model. They do not focus on the downstream task of training the fair model, and the methods are hence not directly comparable. We are happy to include discussion of the above works in the related work section of our paper.

Notation and Event

For examples of the event $\mathcal{E}$, we direct you to appendix A5, where we have a table of functions $f(h(X), y)$ and corresponding epsilons that refer to various common fairness definitions. For example, to enforce equalized false positive rate parity, $f(h(X), y) := 1[h \neq y]$ and $\mathcal{E} = 1[y = 0]$. We are happy to include further examples and a more detailed demonstration of their equivalence in our updated version. We acknowledge that the notation is somewhat unwieldy; however, we do believe that the tracking of the event $\mathcal{E}$ is helpful to illustrate the generality of the methods. We will revisit this notation and experiment with other options (including with the reviewer's proposed modification).

[1] Elzayn, Hadi et al. Measuring and mitigating racial disparities in tax audits. SIEPR, 2023.

[2] Austin, Roy L. Expanding Our Work on Ads Fairness. 21 June 2022.

[3] Agarwal, Alekh et al. "A reductions approach to fair classification." PMLR, 2018.

[4] Awasthi, Pranjal, et al. "Equalized odds postprocessing under imperfect group information." PMLR, 2020.

Fig. 18

Summary:
This paper develops methods for measuring & reducing fairness violations when the protected attribute is private. Access to protected attribute labels for a small subset of the data is assumed, but only probabilistic estimates are available for the remainder of the data. A method for measuring common fairness metrics is proposed for this setting. A theoretical bound on a range of common fairness metrics is given. The utility of the approach is shown on various datasets.

Soundness: 4 excellent

Presentation: 4 excellent

Contribution: 3 good

Strengths:
This paper has numerous strengths. The paper is well motivated, with convincing practical examples of why this data setting and problem is relevant. The paper is also very well written, the notation is clean and I found it easy to follow.

The technical results are strong. Theorem 1 is neat, drawing on a nice simple relationship between the bias in what the paper labels the probabilistic estimator and the linear estimator (or more precisely, their asymptotic limit). The result is then applied to help solve the fair learning problem in a principled way.

The description of the experiments is thorough and precise. The results are well describes and are convincing evidence in favor of the proposed approach. The figures are well designed and easy to understand.

Weaknesses:
There is generally not much to criticize about this paper. However it would be useful if the paper could elaborate on how well this approach works on small sample sizes when there is presumably considerable variability in the probabilistic/linear estimates.

It would also be useful to better understand (either theoretically or via the experiments) how much public data is needed for this approach to give good performance (this is actually mentioned as a limitation/extension).

Some discussion of how computationally demanding this approach is when compared to alternatives could be beneficial.

It would also be valuable to extend this to situations where the protected variable has more than two categories (perhaps this approach still applies, but that was not clear to me). I agree this approach is still valuable in this setting but clearly generalization would be valuable.

Questions:
Why does the empirical problem not account for variability/uncertainty in the estimators?

How robust is the method to incorrectly calibrated probabilities for the protected features? In particular, what happens if there's disparity in these too?

Limitations:
The authors have addressed limitations adequately and noted certain privacy concerns that arise in this setting.

Flag For Ethics Review: No ethics review needed.

Rating: 8: Strong Accept: Technically strong paper, with novel ideas, excellent impact on at least one area, or high-to-excellent impact on multiple areas, with excellent evaluation, resources, and reproducibility, and no unaddressed ethical considerations.

Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

Code Of Conduct: Yes

Fig. 19

Fig. 18

Rebuttal:
Thank you for your review. We address your questions and concerns below:

Variability and dataset size: Regarding your questions on variability in probabilistic/linear estimates in small sample sizes, and questions about how much labeled subset data is required for good performance, we have attached a study between our method and only training on labeled subset data that explores this question by training with different percentages of labeled data on a synthetic dataset in Figure 3 in the global response PDF.

As we mentioned in limitations, it's not just labeled subset size that impacts how our method may function, but complexity of the dataset as well—as the more complex the data is, the more data will be necessary to successfully bound disparity as well as predict accurately.

To conduct this study, we implement a data-generating process using analogues of the key features of our setting, as well as a tunable complexity parameter that we may vary (or hold fixed) with sample size. Figure 3 shows the true disparity and accuracy obtained by our method (y-axes) on a test set over several different sizes of labeled subset data (from 50-2000 points) (x-axes). The four graphs comprise a range of models on data generated from a 3rd degree polynomial with random coefficients over 10 features (less complex data, top row) and 20 features (more complex data, bottom row), with an overall dataset of 5k (left) and 10k (right). The blue method is labeled subset, the orange method is our method, and the dotted line is the disparity of an unconstrained classifier.

Results: **Figure 3 in the attached PDF of the global response shows that our method consistently meets disparity goals even with small amounts of labeled data in both the low and high complexity regimes.** By contrast, the labeled subset method converges from above towards the desired disparity bound as the size of the labeled set increases; it comes close to meeting the bound in the low complexity regime but does not meet it even with a large amount of data in the complex regime. Thus, this study suggests that our method is relatively more valuable in high complexity regimes. We will add these results and discussion to the paper and discuss the simulation in detail in the appendix. We are happy to expand these experiments in the final version should the paper be accepted.

Computation time: We would be happy to record the timing of the different methods in the appendix pending acceptance. As a rough estimate, the methods take approximately the following amount of time: L2, 150k observations: Ours: ~7.5 minutes Mozannar: ~3 seconds Wang: ~24 minutes

Extension beyond two demographic groups: Our method is designed to measure/mitigate disparity between two demographic groups, but there is no reason why one would not be able to use these bounds in tandem for any non-overlapping groups: in the measurement setting, one could simply compute the disparities across whatever comparisons are of interest, and extra constraints could theoretically be added to the training setting (e.g. to enforce covariance constraints and upper bounds on multiple groups). However, with additional groups, additional data would be needed to correct for multiple comparisons and the increased complexity of the learning problem.

Variability in empirical problem: We understand your concern to refer to why sample variation is not addressed directly in the covariance conditions or the linear estimator in the empirical problem outlined in Problem 2A. We note that the generalization bounds in Theorem 2 provides guarantees that the bounds on all of the relevant terms in the empirical problem— the covariance constraints as well as the bound on the disparity estimator—will translate to bounds on the ground truth data up to an error term which decreases with larger amounts of data. Thus, once there is sufficient data, the instability in the samples of the covariance terms and disparity estimates should not influence the results. We're of course happy to clarify the presentation of this result.

Calibration: Thank you for your question. It is true that a miscalibrated (i.e. systematically biased) probabilistic proxy will result in inaccurate bounds; to the extent this is known, this can be accounted for in the bounds at the cost of some additional algebra (which we can add to the Appendix). But in practice, imperfect calibration does not seem to be a problem if it is not extreme or pathological. We also note that the proxy can also be recalibrated for the specific population of interest, either via flexible machine learned models [2,3] or simple linear regression. This latter technique is particularly compelling given recent work which suggests that additional features tend to make marginal improvements for accuracy after the key factors in BIFSG are taken into account [1].

Thank you for taking the time to read and review our paper, and please let us know if you have any further questions.

[1] Cheng, Lingwei, et al. "How Redundant are Redundant Encodings? Blindness in the Wild and Racial Disparity when Race is Unobserved." FACCT, 2023.

[2] Pleiss, Geoff, et al. "On fairness and calibration." Neurips, 2017.

[3] Elzayn, Hadi, et al. 2023. (16 in original paper)

Fig. 19

Fig. 20

Summary: The paper proposes a technique for measuring and mitigating fairness disparities when most of the true attributes are protected and only probabilistic estimates of protected attribute labels (e.g., via BISG) are accessible. The proposed method takes advantage of contextual information, i.e., the relationship between a model's predictions and the probabilistic prediction of attributes, to provide tighter bounds on the true disparity.

Soundness: 3 good

Presentation: 3 good

Contribution: 2 fair

Strengths:

1. The paper considers a critical and practical problem, i.e., the sensitive attributes may be protected.
2. The proposed method is validated by both theoretical analyses and experiments.
3. The proposed method can give a significantly tighter bound.

Weaknesses:

1. The main concern is that the assumption that $b = \Pr[B = 1|Z, \epsilon]$ is too strong. In practice, the probabilistic estimates of attributes are likely to be biased. If b is unbiased, can we just estimate the disparity by letting $\hat{B} \leftarrow 1_{b>0.5}$? How would this method compare to the proposed method?
2. Should the labeled attributes be iid with the unlabeled ones? In practice, iid is hard to guarantee, and requiring a small set of sensitive attributes may violate privacy regulations, e.g., differential privacy.
3. The compared baselines are limited in Figure 1. There are five more baselines:
   - The method in weakness 1.
   - Evaluate fairness disparity only with the labeled attributes.
   - The method in [R1–R5]

[R1] P. Awasthi et al. Evaluating fairness of machine learning models under uncertain and incomplete information. FAccT, 2021.

[R2] J. Chen et al. Fairness under unawareness: Assessing disparity when protected class is unobserved. FAccT, 2019.

[R3] Z. Zhu et al. Weak proxies are sufficient and preferable for fairness with missing sensitive attributes. ICML, 2023.

Questions:

Here are several questions in addition to the questions mentioned in weakness:

1. What would happen if the true disparity in Figure 1 is large? Would the method still work?
2. In Line 300, does the "labeled subset with true race labels" refer to the set with 1,500 examples? If true, does this setting correspond to the orange curve in Figure 2 (bottom)? If also true, it seems that this setting achieves the best performance, rather than the proposed method.

Limitations:

NA

Flag For Ethics Review: No ethics review needed.

Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes

Fig. 21

Fig. 20

Rebuttal:

We thank the reviewer for their feedback. We first respond to your questions, and then the overall concerns:

Questions:

Larger true disparity: Yes, our method works in settings with larger disparity. In fact, the larger the disparity is, the smaller the variance in the estimators and thus the more precisely they can be estimated with the same amount of data. We additionally highlight Figure 1, in which D_B and D_U do in fact bound the true disparity, despite the relatively high disparity of an unconstrained classifier at 15%. This is further supported by our new COMPAS experiments in which the disparity of an unconstrained classifier is around 26% yet we are able to successfully train our classifier to make predictions with a disparity below 18%.

Clarification of Figure 2 re: labeled subset with true race labels: The orange line does correspond to using a model that enforces the fairness constraint on the set of 1,500 with true race labels. This baseline has higher accuracy, however, our target is not to maximize accuracy alone; it is to maximize accuracy subject to disparity constraints, and the labeled subset method does not satisfy this constraint. The drop in accuracy of our method corresponds to a fairness-accuracy tradeoff: for huge decreases in disparity (e.g. from 15% to below 4% in the upper left-hand graph, for demographic disparity), there is a noticeable difference in accuracy between our method and the labeled subset—however, when we bound demographic disparity at 10% instead of 4% in the same graph, the accuracies of the two methods are comparable while our method is far more successful at reducing disparity.

The method and results speak to the ongoing debate about the extent of the fairness/accuracy tradeoff [1]. While many works have shown that it is possible to reduce disparity by some amount without noticeable accuracy tradeoffs, our work adds to this debate by showing that one cannot reduce disparity by an arbitrary amount—e.g. to a specific threshold—with no repercussions on accuracy. Using the labeled subset alone maintains higher accuracy primarily by failing to decrease disparity to the desired threshold. As we see Figure 2 in the original paper and Figures 1 and 2 in the additional PDF, as we relax the fairness constraints, the accuracy improves with it linearly.

Concerns:

Bias in the probabilistic estimate b:

- We understand that miscalibration of b is a concern, and direct the reviewer to appendix A.4.2 where we discuss this issue. Overall, while miscalibration of b will affect the bounds, the method still works with some miscalibration (indeed, as we see in appendix C.2, the race probabilities are not perfectly calibrated).
- Regarding your suggestion to directly use b thresholded at 0.5, we have included this experiment in Appendix Figure 6 (the bars labeled "Threshold"). As you can see, this method of thresholding the BISG estimates does not ever effectively bound disparity to the desired level. This is consistent with Chen et. al.'s findings that thresholded estimators will under- or over-estimate disparity depending on fundamental parameters of the problems.

I.I.D. Samples: As in many ML settings, our theoretical guarantees require that the labeled subset is drawn from the same underlying distribution as the unlabeled dataset. But, also like many machine learning settings, deviation from this assumption will degrade results smoothly rather than catastrophically. And in settings of interest - e.g. healthcare, tax audits, recidivism prediction, etc. - existing empirical evidence as well as (setting-specific) theoretical arguments often suggest that satisfaction of the covariance conditions is driven by societal-level factors like historical discrimination and socioeconomic differences which are very likely to generalize even if the labeled subset is not perfectly representative. Of course, the gold standard remains a perfectly representative subset. Note that one can also conduct sensitivity analyses to quantify the degree to which the labeled subset must differ in order to change the qualitative impact of measurements or the presumption of improved fairness via our model, but this is beyond the scope of the paper.

Other comparisons for disparity measurement: Chen et. al. analyze the probabilistic estimator (which is in fact well-known in the literature, dating at least as far back as 1953 [2]) and the thresholded estimator. Both estimators are biased (even with a perfectly calibrated probabilistic proxy), which Chen et. al. highlight as an impediment to their usage. But we note that we do incorporate the probabilistic estimator and take advantage of its bias in our framework. As mentioned above, the thresholded estimator does not bound disparity well in our experiments.

We will add a discussion as to these points and the citations to our related works section appropriately.

More generally, the disparity estimation methods pointed to in the review are all point estimators. By contrast, our approach recovers upper and lower bounds on the disparity. These approaches are fundamentally different: whereas the bounds approach tries to capture all parameter values that could have generated the data without attaching special significance to any one of them, the point estimation approach tries instead to obtain one parameter. Kallus et. al., to which we do compare, is the only alternative method for disparity estimation we are aware of that also obtains upper and lower bounds.

We are happy to add a discussion on these papers and the differences from our approach to the related work.

[1] Rodolfa, Kit T., et al. "Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy." 2021.

[2] Duncan, Otis Dudley, and Beverly Davis. "An alternative to ecological correlation." 1953.

Fig. 22

Summary:
The paper proposes to regress the label's dependence on the sensitive attribute b assuming access to sensitive labels for a small subset of the data, and using BISG for the rest of the samples. This estimate is then used to bound fairness violations in the form of what authors call probabilistic constraints.

This process decouples the sensitive attribute estimation from the concept learning, supposedly reducing estimation errors of a typical fairness-constrained optimization problem. The method is for voter turnout data. However, despite the additional burden of sensitive attribute regression, the method does not bound worst-case fairness violations.

Soundness: 2 fair
Presentation: 2 fair
Contribution: 2 fair

Strengths:
The two-step training procedure is an interesting approach.

Weaknesses:
- Compared to [16], the novelty of theorem 1 is questionable.
- The empirical section is somewhat sparse. In particular, the method is tested on one dataset. I suspect this is a limitation of the particular data regime the paper adopts where BISG applicability is a concern. Otherwise, the methods and baseline comparisons are sensible. Given the lack of novelty (see previous point), I think the empirical section should have been more substantial. In particular, several ablation studies are in order. First, given that BISG estimation comes from a separate data domain, the paper needs an ablation study to show how much the method improves upon just using a normal fairness-constrained optimization problem using BISG labels.
  - Yet another ablation study should be on the size of the labeled dataset. In other words, the paper should provide a concrete answer to this observation:
    We note that the utility of our method is dependent upon the size360 of the subset of the data labeled with the protected attribute—if this subset is relatively large, then361(depending on the complexity of the learning problem) it may be sufficient to train a model using the362 available labeled data. Symmetrically, if the labeled subset is exceedingly small, the enforcement of363 the covariance constraints during training may not generalize to the larger data set
- Despite the additional burden of sensitive attribute regression, the method does not bound worst-case fairness violations, not even asymptotically. I am afraid the generalization bound in Theorem 2 does not alleviate these concerns, as it looks pretty much like a normal generalization bound. What is the extra constraint doing here? I may be missing something, and invite authors to explain their exact contribution here.
- Overall, between the fact that the sensitive attribute regression does not help bound worse-case violations despite the additional hassle (and BISG dependence) and the limited empirical evidence, I am not convinced of the method's general usefulness.
- Some factual issues in the text:
  - demographic parity101 in classification (6; 35 ; 36) corresponds to letting E be the generically true event and f be simply102 Y = 1 . False positive rate parity (11 ; 12) corresponds to letting E be the event that Y = 0 and the103 function f (^Y , Y) = 1[^Y 6 = Y] .
    - Demographic parity does not take into account the ground truth,
  - well-known form of the regression coefficient,
    - This lacks a reference. It seems the paper assumes L2 loss for regression.
  - Certain less-common terms are not defined before use. For instance, "near-feasibility." The reader should not be left guessing here.
- Criticism but not ground for reduced score:
  - Predicting sensitive attributes is problematic at best and illegal in many contexts where fairness is a concern. I also take issue with the systematic use of sensitive data prediction. We should not be promoting cross-matching of data, beyond the reason they were collected for.

Questions:
- Can you expand on the novelty of your analysis in Theorem 1 compared to prior work [16]?
- Figure 8 (middle), it seems none of the methods can bound FPRD. How can you claim this is near-feasible?

Limitations:
The limitations are well-addressed in the main paper. Maybe authors could expand on it using the collective reviews.

Flag For Ethics Review: No ethics review needed.

Rating: 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Fig. 22
Fig. 23

Rebuttal:

We thank the reviewer for their feedback. We note that:

- **We have performed the two ablation studies mentioned which we describe in detail further below**—we point to in Appendix F for the naive comparison of our method to thresholded BIFSG estimates, and the study about dataset size we present in Figure 3 in the PDF attached to the global response.
- **We also include new versions of our false positive rate disparity (FPRD) results in response to your questions on near-feasibility.**

We first provide some clarifications:

- It is true that in our setting, the problem of learning a sensitive attribute proxy is decoupled from the problem of learning the outcome of interest. But our learning problem does depend on the proxy: we learn a fair model by measuring and constraining the relationship between the outcome and the proxy (namely, constraining the covariance conditions $C_{f,b|B}$ and $C_{f,B|b}$ to be the same sign in order to guarantee that the linear estimator D_L will serve as a bound on disparity, and then constraining D_L to the desired upper bound on disparity) in the training process.
- The proxy we use—BIFSG—is used over the entire dataset. The labeled subset is used only to constrain the relationship between the proxy and the outcome of interest.

Worst-case violations and Theorem 2: If we understand the point correctly, the reviewer is concerned about the "worst case" in the sense of "worst case over distributions", i.e. that the method will not work if the covariance conditions are not met. In our technique for training a fair model, it is precisely for this reason that we add the constraints on the covariance terms. Thus, we are limiting our search to models for which we can guarantee that our measurement method works and thus that we can reliably bound disparity; in doing so, we trade off some performance for the security of bounding disparity. Of course, we cannot work with the population covariance terms directly, but instead work with their sample analogues. As in any empirical optimization problem, working with the sample introduces some noise and approximation error. Theorem 2 is useful, because it provides a formal guarantee and quantification of the intuition that (with high probability) these errors will become negligible with enough data and iterations under mild conditions.

Size of the labeled subset: We agree that this is an important question; we have conducted studies of this and display our results in Figure 3 of the attached one-page addition. Due to space constraints, please see our response to reviewer Vf4j for a discussion of this experiment.

Comparing to fairness-constrained optimization with BISG labels: In Appendix F, we provide an ablation study in which we train a fairness-constrained optimization method based on thresholded versions of the BIFSG labels for all experiments in Figure 2 of our original paper. Our study shows that while the thresholded approach has higher accuracy than our method, it consistently fails to control disparity below the specified threshold. We will highlight this more prominently in the text.

Novelty of Theorem 1: Theorem 1 is similar to the result of [16], but Theorem 1 generalizes the result beyond demographic parity to a very broad class of fairness definitions, a generalization that was not obvious from [16]. In any case, we do not view this particular theorem as our primary contribution, but rather as a rigorous justification of and basis for our methods as applied to fairness metrics more generally.

Near-feasibility: We thank the reviewer for pointing out this imprecision. Near feasibility refers to a solution produced which may violate constraints, but that this violation can be made arbitrarily small (in particular, with enough data and training iterations), as described in Theorem 2. Near feasibility occurs widely in settings where constraints are formulated over distributions but only sample data is available; in practice, researchers will either fix a constant below which constraint violation is considered negligible, or fix a number of iterations and dataset size based on data availability and observe the constraint violations. See, e.g. [1]. We will clarify this point in the paper.

Regarding the FPRD results, we provide improved results displayed in Figure 2b of the attached PDF. In our initial presentation, for consistency, we used the same hyperparameters for each of the disparity metric experiments. By tuning for the FPRD problem specifically, we see greatly improved results. We will update the paper to reflect the problem-specific hyperparameter optimization approach.

Relationship between linear estimator and linear regression coefficient: We thank the reviewer for pointing out how our statement could be misinterpreted. We will change the sentence to emphasize its relationship to the ordinary least squares regression coefficient.

Typos: We thank the reviewer for pointing out the typos, which we will fix.

Finally, we share the reviewer's concern that sensitive data be protected. But it is well-established in the literature that fairness cannot be achieved through unawareness; hence, in settings without labeled data, if we don't use proxies to measure and mitigate unfairness, our options are to either remain ignorant about potential unfairness, or use distributionally-robust optimization approaches, which may come at a cost to performance (see discussion in our response to Reviewer ZFXj). In some cases, these solutions may be desirable, but we believe enough high-stakes settings (e.g. [1], [2], and 3]) exist that developing methods to measure and mitigate unfairness based on proxies is worth the potential risks.

[1] Cotter et al, 2019. (13 in original paper)

[2] Elzayn, et al. 2023. (16 in original paper)

[3] Obermeyer, et al. "Dissecting racial bias in an algorithm used to manage the health of populations." Science, 2019.

[4] Executive Order 13985

Fig. 23

Fig. 24

Rebuttal:
We thank the reviewer for their feedback. We note that:

- **We have performed the two ablation studies mentioned which we describe in detail further below**—we point to in Appendix F for the naive comparison of our method to thresholded BIFSG estimates, and the study about dataset size we present in Figure 3 in the PDF attached to the global response.
- **We also include new versions of our false positive rate disparity (FPRD) results in response to your questions on near-feasibility.**

We first provide some clarifications:

- It is true that in our setting, the problem of learning a sensitive attribute proxy is decoupled from the problem of learning the outcome of interest. But our learning problem does depend on the proxy: we learn a fair model by measuring and constraining the relationship between the outcome and the proxy (namely, constraining the covariance conditions $C_{f,b|B}$ and $C_{f,B|b}$ to be the same sign in order to guarantee that the linear estimator D_L will serve as a bound on disparity, and then constraining D_L to the desired upper bound on disparity) in the training process.
- The proxy we use–BIFSG–is used over the entire dataset. The labeled subset is used only to constrain the relationship between the proxy and the outcome of interest.

Worst-case violations and Theorem 2: If we understand the point correctly, the reviewer is concerned about the "worst case" in the sense of "worst case over distributions", i.e. that the method will not work if the covariance conditions are not met. In our technique for training a fair model, it is precisely for this reason that we add the constraints on the covariance terms. Thus, we are limiting our search to models for which we can guarantee that our measurement method works and thus that we can reliably bound disparity; in doing so, we trade off some performance for the security of bounding disparity. Of course, we cannot work with the population covariance terms directly, but instead work with their sample analogues. As in any empirical optimization problem, working with the sample introduces some noise and approximation error. Theorem 2 is useful, because it provides a formal guarantee and quantification of the intuition that (with high probability) these errors will become negligible with enough data and iterations under mild conditions.

Size of the labeled subset: We agree that this is an important question; we have conducted studies of this and display our results in Figure 3 of the attached one-page addition. Due to space constraints, please see our response to reviewer Vf4j for a discussion of this experiment.

Comparing to fairness-constrained optimization with BISG labels: In Appendix F, we provide an ablation study in which we train a fairness-constrained optimization method based on thresholded versions of the BIFSG labels for all experiments in Figure 2 of our original paper. Our study shows that while the thresholded approach has higher accuracy than our method, it consistently fails to control disparity below the specified threshold. We will highlight this more prominently in the text.

Novelty of Theorem 1: Theorem 1 is similar to the result of [16], but Theorem 1 generalizes the result beyond demographic parity to a very broad class of fairness definitions, a generalization that was not obvious from [16]. In any case, we do not view this particular theorem as our primary contribution, but rather as a rigorous justification of and basis for our methods as applied to fairness metrics more generally.

Near-feasibility: We thank the reviewer for pointing out this imprecision. Near feasibility refers to a solution produced which may violate constraints, but that this violation can be made arbitrarily small (in particular, with enough data and training iterations), as described in Theorem 2. Near feasibility occurs widely in settings where constraints are formulated over distributions but only sample data is available; in practice, researchers will either fix a constant below which constraint violation is considered negligible, or fix a number of iterations and dataset size based on data availability and observe the constraint violations. See, e.g. [1]. We will clarify this point in the paper.

Regarding the FPRD results, we provide improved results displayed in Figure 2b of the attached PDF. In our initial presentation, for consistency, we used the same hyperparameters for each of the disparity metric experiments. By tuning for the FPRD problem specifically, we see greatly improved results. We will update the paper to reflect the problem-specific hyperparameter optimization approach.

Relationship between linear estimator and linear regression coefficient: We thank the reviewer for pointing out how our statement could be misinterpreted. We will change the sentence to emphasize its relationship to the ordinary least squares regression coefficient.

Typos: We thank the reviewer for pointing out the typos, which we will fix.

Finally, we share the reviewer's concern that sensitive data be protected. But it is well-established in the literature that fairness cannot be achieved through unawareness; hence, in settings without labeled data, if we don't use proxies to measure and mitigate unfairness, our options are to either remain ignorant about potential unfairness, or use distributionally-robust optimization approaches, which may come at a cost to performance (see discussion in our response to Reviewer ZFXj). In some cases, these solutions may be desirable, but we believe enough high-stakes settings (e.g. [1], [2], and 3]) exist that developing methods to measure and mitigate unfairness based on proxies is worth the potential risks.

[1] Cotter et al, 2019. (13 in original paper)

[2] Elzayn, et al. 2023. (16 in original paper)

[3] Obermeyer, et al. "Dissecting racial bias in an algorithm used to manage the health of populations." Science, 2019.

[4] Executive Order 13985

Fig. 24