A Main Proofs

A.1 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:

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

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:

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

by the strong law of large numbers. Similarly,

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

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

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

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

$$\xrightarrow{n_\mathcal{E} \to \infty} \frac{\mathbb{E}[b f(\hat{Y}, Y)|\mathcal{E}]}{\mathbb{E}[b|\mathcal{E}]} - \frac{\mathbb{E}[(1 - b) f(\hat{Y}, Y)|\mathcal{E}]}{\mathbb{E}[(1 - b)|\mathcal{E}]}$$

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

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

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.

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

2.

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

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:

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

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:

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

This can be expanded as:

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

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

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

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:

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

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

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

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

A.2 Proof of Proposition 2

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

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

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

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

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

A.3 Standard Errors

Here, we discuss the calculation of standard errors; these arguments are more general, but substantially similar, version of those made in (20). As shown in the proof of Theorem 1, $\widehat{D}^l_\mu$ and $\widehat{D}^p_\mu$ converge to their asymptotic limits, D^l_μ and D^p_μ, 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.(41) Note that, since as shown in Lemma 1

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

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:

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

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

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

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

A.4 Obtaining the probabilistic prediction

A.4.1 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 (23; 45; 48; 31) 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.

A.4.2 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 (20) 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.

A.5 Fairness Metrics

As noted, many fairness metrics can be written in the form required by our formulation. For concreteness, we provide a table based on (37; 44) 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 .

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	$\mathbf{1}[h \neq y]$	$\{y = 1\}$

B Mathematical Formulation of Fair Learning Problem

B.1 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.

B.1.1 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:

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

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.

B.1.2 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.2 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 (13) for a fuller and more detailed discussion of the constrained statistical learning relevant to our setting and (17) for more general discussion of non-convex optimization via primal-dual games.

B.2.1 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.

B.2.2 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.

$$\min_{h\in\mathcal{H}} \frac{1}{n} \sum_{i \in n_{\mathscr{D}}} L(h(X_i), Y_i) \text{ s.t. } \widehat{D}_\mu^L \leq \alpha$$

$$\text{and } 0 \leq -\frac{1}{n_{\mathscr{D}_L}} \sum_{i\in\mathscr{D}_L} \left[\left(f(h(X_i), Y_i) - \overline{f(h(X_i)), Y_i}^{B_i}\right)(b_i - \bar{b}^{B_i})\right]$$

$$\text{and } 0 \leq -\frac{1}{n_{\mathscr{D}_L}} \sum_{i\in\mathscr{D}_L} \left[\left(f(h(X_i), Y_i) - \overline{f(h(X_i)), Y_i}^{b_i}\right)(B_i - \bar{B}^{b_i})\right]$$

.

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.

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

B.3 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 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.

B.4 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.

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

As discussed in Section 2.1, 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

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

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,

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

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,

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

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

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

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

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

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

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

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

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

Solving for β we get the solution,

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

C Data

C.1 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 1 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.

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 1: 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.

C.2 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 Voicu (45) 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.

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

Table 2 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 3 shows that our predicted probabilities are generally well-calibrated to true probability of Black (although BIFSG tends to overestimate the probability of Black).

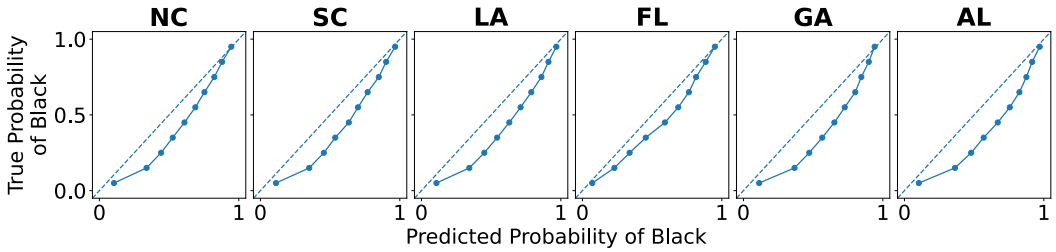

Figure 3: Calibration plots showing predicted probability of Black (x-axis) versus actual proportion of Black (y-axis).

D Details on Measurement Experiments

D.1 Voter Turnout Prediction Performance

Table 3 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" (35). Note again that our predictors rely solely on demographic factors of voters because those are the most reliable data L2 provides us.

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 3: 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).

D.2 The KDC Method

Kallus et al. (31) similarly propose a method of finding the tightest possible set of true disparity given probabilistic protected attributes. A subtle difference between KDC and our method is their assumptions around the auxiliary dataset. 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[3]; doing so

[3]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

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 also implement the KDC estimators as originally proposed in Figure 4 but the results do not change substantially[4].

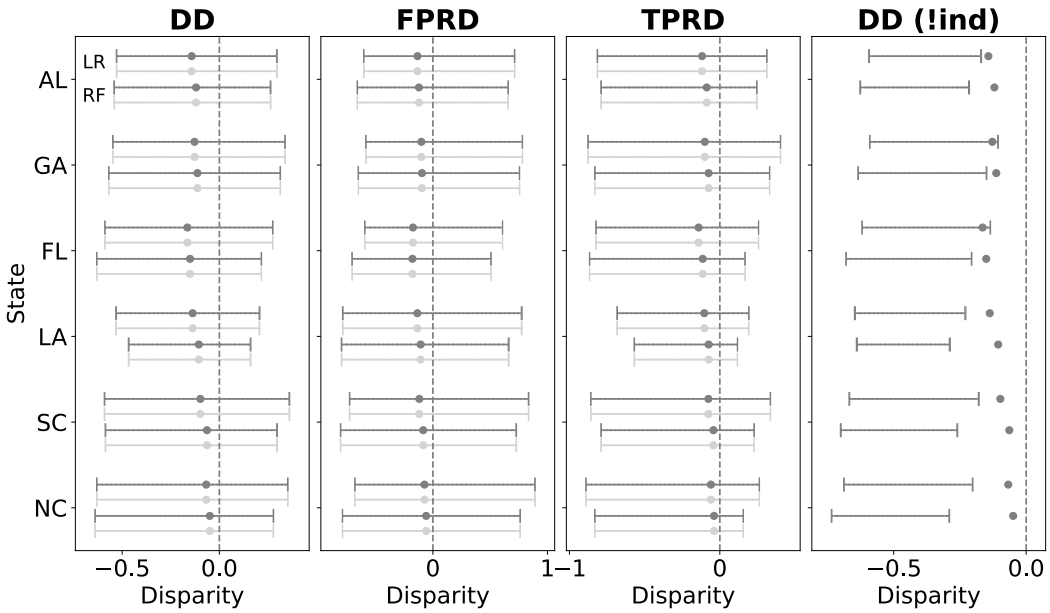

Figure 4: Comparison of different KDC implementations. In dark grey, we have our implementation that violates the independence assumption in Kallus et al. (31). In light grey, we have KDC's original implementation with the independence assumption – nothing substantively changed. The top and bottom pairs of each state correspond to the estimators from logistic regression (LR) and random forest (RF) models, respectively. Kallus et al. (31) 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.

D.3 Random Forest

We also run experiments on bounding disparity when voter turnout is predicted on random forest models, as seen in Figure 5. 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.

E Details on Training Experiments

E.1 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

in our calculation, $r = 1$, but this loosely goes against the assumption that r is closer to 0 in Section 7 of (31). 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.

[4] In Appendix A.5, (31) 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 4.

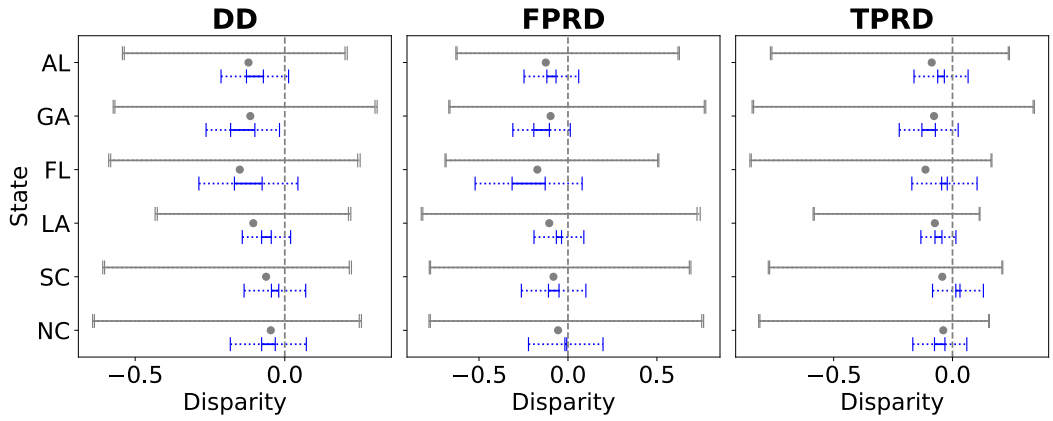

Figure 5: Comparison of our method of bounding true disparity (blue) to the method proposed in (31) (grey), using a random forest model to predict voter turnout on L2 data in six states. We evaluate three disparity measures: demographic disparity (DD), false positive rate disp. (FPRD), and true positive rate disp. (TPRD). The grey dot represents true disparity. Both methods always bound true disparity within their 95% standard errors.

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 Chamon et al. (14) method with two different model types under the hood, neural networks and logistic regression. Both types of models are implemented in Pytorch. Our neural network models consist of a single hidden layer of 8 nodes, with a ReLu activation.

E.2 CSL (Chamon et al.)

We implement our constrained problem using the official Pytorch implementation provided by (14).[5] for both a logistic regression model as well as a shallow neural network. We run the non-convex optimization problem for 1,000 iterations with a batch size of 1,024 and use Adam (32) 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 Chamon et al. (14) method in Appendix B above.

E.3 The Method of Wang et al.

Wang et al. (46) 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 Kallus et al. (31). We use code provided by Wang et al. (46).[6] 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 Wang et al. (46). 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 Wang et al. (46). In order to choose the best hyperparameters, we use the same data as outlined in Section 4.3.1 (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, Wang et al. (46) chooses the hyperparameter that results in the lowest loss while adhering to the fairness constraint with respect to **true race**. Since we

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

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

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.

E.4 The Method of Mozannar et al.

Mozannar et al (36) 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 moodel with incomplete demographic data. The authors do not discuss this in detail or provide the code for this extension, so we modify the code Mouzannar et. al. provide for their paper (36) 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 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) (2) 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.5 Satisfying Constraints and Bounding True Disparity on Training Set

In Table 4, we present a summarized description of the results of various training runs in terms of the satisfaction of covariance constraints on the labeled subset and meeting our desired bound α on $\widehat{D}^L_\mu$ on the training set, as well as whether or not we actually bound true disparity on the test set.Specifically, we show the rates at which we meet different constraints and the rate at which we actually bound disparity on the test set.

As we see in Table 4, we meet our desired covariance condition on the labeled subset, which we use to enforce these conditions, approximately 81% of the time, and we satisfy the condition that bounds $\widehat{D}^L_\mu$ by α approximately 75% of the time. This is a side effect of the near-feasibility of the Chamon et al. method (13).

Note that not satisfying the $\widehat{D}^L_\mu$ condition does *not* mean that we do not bound true disparity; it only means that we do not bound true disparity *within our desired bounds*. It is still possible that we bound true disparity, on a slightly larger bound—which we see is the case in almost 97% of instances on the test set.

We note that we mistakenly misrepresented our results in the main paper, and will fix the error to match our full results in this Appendix as soon as we have access to updating the main draft: we stated that we always meet the covariance constraint on the *training* set, when in fact the relevant test is the *labeled subset*, and we meet the bounds approximately 81% of the time. We also stated that we always bound the true disparity on the test set, when in fact we bound it approximately 97% of the time.

% Cov Match (aux)	% D_l (train) < Disp. bound	% Estimators bound true disp. (test)
80.83	75.0	96.67

Table 4: We present the rate of satisfaction of the covariance bounds in the labeled sunbset, as well as the rate at which we satisfy the bound on $\widehat{D}^L_\mu$, our linear estimate of disparity, on the training set, and the rate at which we bound true disparity within our error bounds on the test set. We note that lack of satisfaction of our bound on the training and labeled subset simply means that the Chamon et al. (14) method was only able to find a *near* feasible solution for certain rounds of certain problems.

E.6 Results on Oracle and Naive

In Figure 6, 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

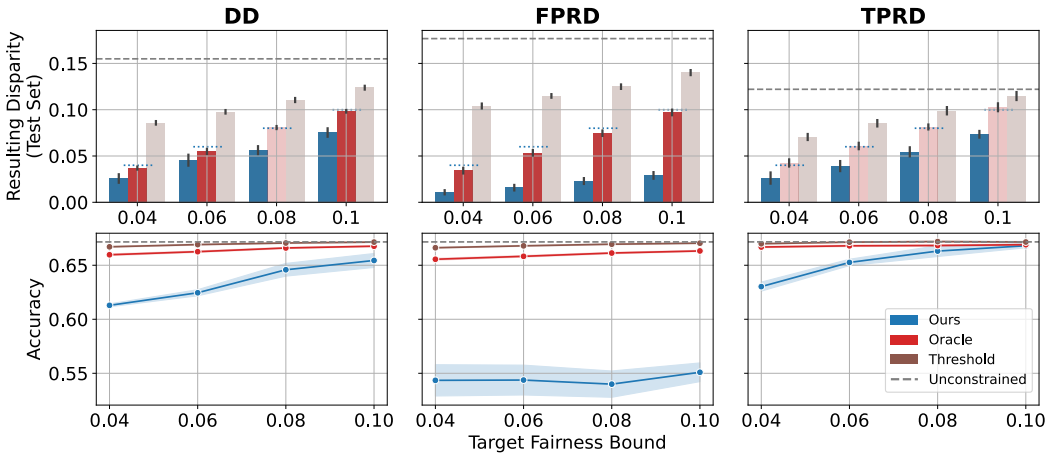

Figure 6: Mean and standard deviation of resulting disparity (top, y-axis) and accuracy (bottom, y-axis) on the test set after enforcing the target fairness bounds (x-axis) on our method (blue); using ground truth race on the entire data, i.e., "oracle" model (red); and using only the estimated race 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.

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.

E.7 Results on Neural Network Models

We describe the outcome of our shallow neural network experiments in Figure 7. We describe details on the optimization of the neural networks in Section E.1. We note that for these experiments, we do not compare to Wang et al. (46) as they do not provide a built-in way to work with neural networks in their code. Although we do not reach the desired disparity bounds as often when using neural networks, we consistently out-perform all methods except for the oracle on disparity reduction, as the guarantees on the labeled subset do not generalize, and enforcing constraints on the thresholded labels do not take the protected attribute label noise into account. In this case, we also outperform the labeled subset on accuracy. This may be because it takes a larger amount of data to effectively train a neural network than logistic regression models, so the accuracy does not saturate with the labeled subset.

F Additional Experiments: COMPAS

In this section, we present a suite of additional experiments we run on the COMPAS (5) 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 (5) 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.

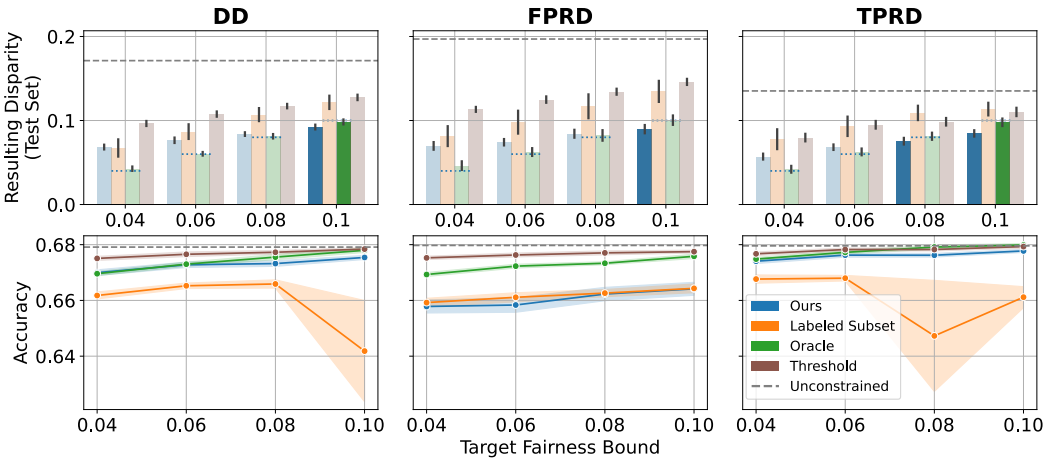

Figure 7: Mean and standard deviation of resulting disparity (top, y-axis) and accuracy (bottom, y-axis) on the test set after enforcing the target fairness bounds (x-axis) on our method (blue); only using the labeled subset with true labels (orange); using ground truth race on the entire data, i.e., "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.

F.1 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 Angwin et al. (5), resulting in $n = 6,128$ datapoints. Table 5 outlines the feature distribution of the dataset.

F.2 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 6 and Figure 8, 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.

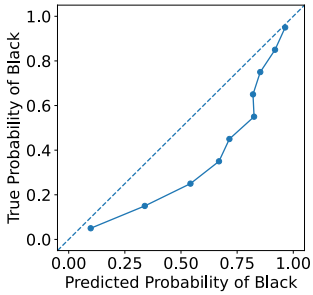

Figure 8: Calibration plot showing the predicted probability a person in the dataset is Black (x-axis) versus the actual proportion of Black people in the dataset (y-axis) for COMPAS.

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 5: 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
0.73	0.86	0.56	0.86

Table 6: Accuracy, precision, recall (thresholded on 0.5), and AUC for predicting probability a person is Black in the COMPAS dataset.

F.3 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 9. 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 7: Accuracy, precision, recall (thresholded on 0.5), and AUC for predicting two-year recidivism on the COMPAS dataset using a logistic regression model.

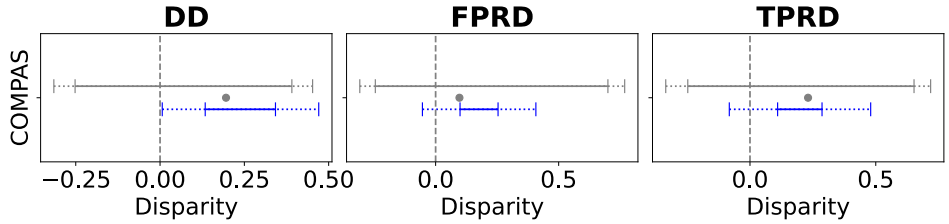

Figure 9: Comparison of our method of bounding true disparity (blue) to the method proposed in (31) (grey), using a logistic regression model to predict two-year recidivism on the COMPAS dataset. We compare results across three disparity measures: demographic disparity (DD), false positive rate disp. (FPRD), and true positive rate disp. (TPRD). The grey dot represents true disparity. Both methods always bound true disparity within the 95% standard errors.

F.4 Training Experiments

We compare our training method to Wang et al. (46), Mozannar et al. (36) 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 10. 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 4.3.1. 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.

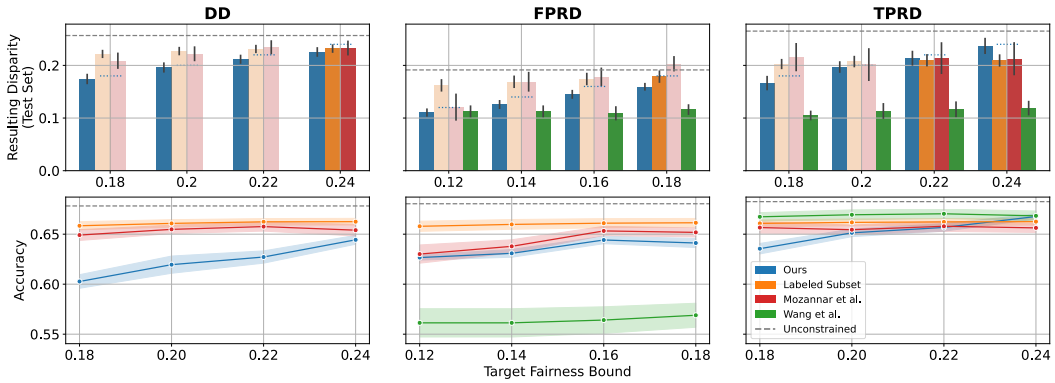

Figure 10: Mean and standard deviation of resulting disparity (top, y-axis) and accuracy (bottom, y-axis) on the test set after enforcing the target fairness bounds (x-axis) on our method (blue); Wang et al.'s method (green); Mozannar et al.'s method (red) and only using the labeled subset with true 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.

In Figure 10, 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.

G ADDITIONAL EXPERIMENTS: 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: data complexity, which we simulate by adjusting the number of features, and size of the labeled subset.

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 even when the size of the labeled subset is 10,000 samples, or 20% of the overall dataset.

G.1 SIMULATION DESIGN

In this section, we describe the design of our simulation used for additional experiments. 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. In particular, the model building blocks consist of:

- 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 0.1

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 8. Figure 11 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	$B \sim$ 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 8: 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.

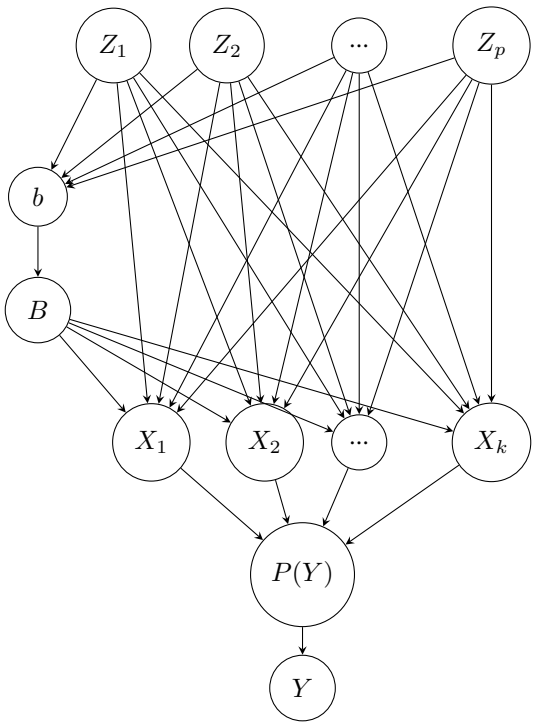

Figure 11: A heuristic depiction of the data generating process for our simulations. Nodes indicate random variables, and edges indicate (causal) relationships between nodes. Importantly, relationships are not necessarily linear.

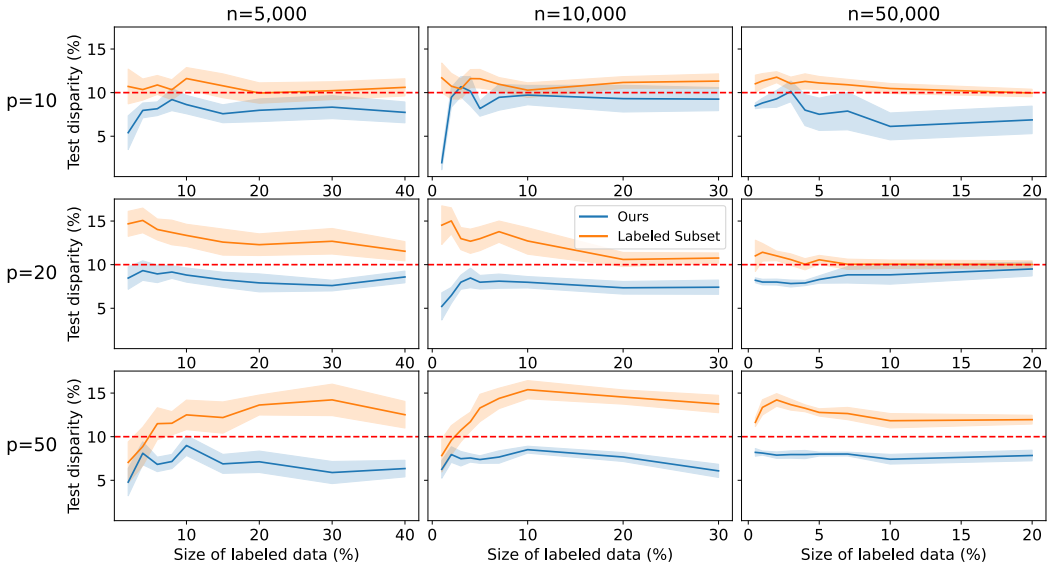

Figure 12: We present a three by three figure showing the test disparity of the our disparity reduction method when compared with relying on only the labeled subset to reduce disparity by directly enforcing a constraint on the protected attribute labels. The rows correspond to datasets of increasing 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 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.

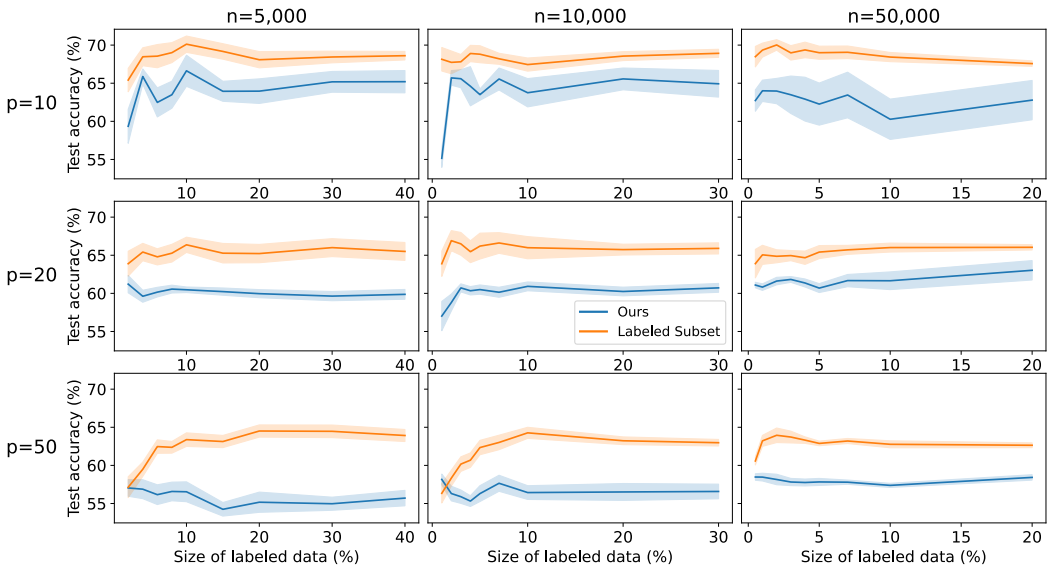

Figure 13: We present a three by three figure showing the test accuracy of the models created using our disparity reduction method when compared with relying on training models only on the labeled subset and reducing disparity by directly enforcing a constraint on the protected attribute labels. The 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.

G.2 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 9 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 9: 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 8.

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 10. 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.

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

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.

G.3 Results

We present our results in Figures 12 and 13. In Figure 12, 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.