A Proofs for Section 4

This section provides proofs and definitions details for the theorems and lemmas presented in Section 4.

A.1 Proofs for TV distance

Definition 1. (TV distance) Let $c(x, y) = \mathbb{1}(x \neq y)$ be a metric, and let π be a coupling between probability distributions p and q. Define the total variation (TV) distance between two distributions p, q as

$$\begin{aligned} TV(p, q) = &\inf_{\pi} \mathbb{E}_{X,Y \sim \pi}[c(X, Y)] \\ &\text{s.t.} \int \pi(x, y)dy = p(x), \int \pi(x, y)dx = q(y). \end{aligned}$$

Theorem 1. *Suppose a model with parameters θ satisfies fairness criteria with respect to the noisy groups $\hat{G}$:*

$$\hat{g}_j(\theta) \leq 0 \;\; \forall j \in \mathcal{G}.$$

Suppose $|h(\theta, x_1, y_1) - h(\theta, x_2, y_2)| \leq 1$ for any $(x_1, y_1) \neq (x_2, y_2)$. If $TV(p_j, \hat{p}_j) \leq \gamma_j$ for all $j \in \mathcal{G}$, then the fairness criteria with respect to the true groups G will be satisfied within slacks γ_j for each group:

$$g_j(\theta) \leq \gamma_j \quad \forall j \in \mathcal{G}.$$

Proof. For any group label j,

$$g_j(\theta) = g_j(\theta) - \hat{g}_j(\theta) + \hat{g}_j(\theta) \leq |g_j(\theta) - \hat{g}_j(\theta)| + \hat{g}_j(\theta).$$

By Kantorovich-Rubenstein theorem (provided here as Theorem 2), we also have

$$|\hat{g}_j(\theta) - g_j(\theta)| = |\mathbb{E}_{X,Y \sim \hat{p}_j}[h(\theta, X, Y)] - \mathbb{E}_{X,Y \sim p_j}[h(\theta, X, Y)]| \leq TV(p_j, \hat{p}_j).$$

By assumption that θ satisifes fairness constraints with respect to the noisy groups $\hat{G}$, $\hat{g}_j(\theta) \leq 0$. Thus, we have the desired result that $g_j(\theta) \leq TV(p_j, \hat{p}_j) \leq \gamma_j$.

Note that if p_j and $\hat{p}_j$ are discrete, then the TV distance $TV(p_j, \hat{p}_j)$ could be very large. In that case, the bound would still hold, but would be loose. □

Theorem 2. *(Kantorovich-Rubinstein).*[2] *Call a function f Lipschitz in c if $|f(x) - f(y)| \leq c(x, y)$ for all x, y, and let $\mathcal{L}(c)$ denote the space of such functions. If c is a metric, then we have*

$$W_c(p, q) = \sup_{f \in \mathcal{L}(c)} \mathbb{E}_{X \sim p}[f(X)] - \mathbb{E}_{X \sim q}[f(X)].$$

As a special case, take $c(x, y) = \mathbb{I}(x \neq y)$ (corresponding to TV distance). Then $f \in \mathcal{L}(c)$ if and only if $|f(x) - f(y)| \leq 1$ for all $x \neq y$. By translating f, we can equivalently take the supremum over all f mapping to $[0, 1]$. This says that

$$TV(p, q) = \sup_{f:\mathcal{X} \to [0,1]} \mathbb{E}_{X \sim p}[f(X)] - \mathbb{E}_{X \sim q}[f(X)]$$

Lemma 1. *Suppose $P(G = i) = P(\hat{G} = i)$ for a given $i \in \{1, 2, ..., m\}$. Then $TV(p_i, \hat{p}_i) \leq P(G \neq \hat{G} | G = i)$.*

Proof. For probability measures p_i and $\hat{p}_i$, the TV distance is given by

$$TV(p_i, \hat{p}_i) = \sup\{|p_i(A) - \hat{p}_i(A)| : A \text{ is a measurable event}\}.$$

[2] Edwards, D.A. On the Kantorovich–Rubinstein theorem. *Expositiones Mathematicae*, 20(4):387-398, 2011.

Fix A to be any measurable event for both p_i and $\hat{p}_i$. This means that A is also a measurable event for p, the distribution of the random variables X, Y. By definition of p_i, $p_i(A) = P(A|G = i)$. Then

$$\begin{aligned}
|p_i(A) - \hat{p}_i(A)| &= |P(A|G = i) - P(A|\hat{G} = i)| \\
&= |P(A|G = i, \hat{G} = i)P(\hat{G} = i|G = i) \\
&\quad + P(A|G = i, \hat{G} \neq i)P(\hat{G} \neq i|G = i) \\
&\quad - P(A|\hat{G} = i, G = i)P(G = i|\hat{G} = i) \\
&\quad - P(A|\hat{G} = i, G \neq i)P(G \neq i|\hat{G} = i)| \\
&= |P(A|G = i, \hat{G} = i) \Big(P(\hat{G} = i|G = i) - P(G = i|\hat{G} = i)\Big) \\
&\quad - P(\hat{G} \neq G|G = i) \Big(P(A|G = i, \hat{G} \neq i) - P(A|\hat{G} = i, G \neq i)\Big)| \\
&= |0 - P(\hat{G} \neq G|G = i) \Big(P(A|G = i, \hat{G} \neq i) - P(A|\hat{G} = i, G \neq i)\Big)| \\
&\leq P(\hat{G} \neq G|G = i)
\end{aligned}$$

The second equality follows from the law of total probability. The third and the fourth equalities follow from the assumption that $P(G = i) = P(\hat{G} = i)$, which implies that $P(\hat{G} = G|G = i) = P(G = \hat{G}|\hat{G} = i)$ since

$$P(G = \hat{G}|G = i) = \frac{P(G = \hat{G}, G = i)}{P(G = i)} = \frac{P(G = \hat{G}, \hat{G} = i)}{P(\hat{G} = i)} = P(G = \hat{G}|\hat{G} = i).$$

This further implies that $P(\hat{G} \neq i|G = i) = P(G \neq i|\hat{G} = i)$.

Since $|p_i(A) - \hat{p}_i(A)| \leq P(\hat{G} \neq G|G = i)$ for any measurable event A, the supremum over all events A is also bounded by $P(\hat{G} \neq G|G = i)$. This gives the desired bound on the TV distance. □

A.2 Generalization to Wasserstein distances

Theorem 1 can be directly extended to loss functions that are Lipschitz in other metrics. To do so, we first provide a more general definition of Wasserstein distances:

Definition 2. (Wasserstein distance) Let $c(x, y)$ be a metric, and let π be a coupling between p and q. Define the Wasserstein distance between two distributions p, q as

$$\begin{aligned}
W_c(p, q) = \inf_{\pi} \;& \mathbb{E}_{X,Y \sim \pi}[c(X, Y)] \\
& \text{s.t.} \int \pi(x, y)dy = p(x), \int \pi(x, y)dx = q(y).
\end{aligned}$$

As a familiar example, if $c(x, y) = ||x - y||_2$, then W_c is the earth-mover distance, and $\mathcal{L}(c)$ is the class of 1-Lipschitz functions. Using the Wasserstein distance W_c under different metrics c, we can bound the fairness violations for constraint functions h beyond those specified for the TV distance in Theorem 1.

Theorem 3. *Suppose a model with parameters θ satisfies fairness criteria with respect to the noisy groups $\hat{G}$:*

$$\hat{g}_j(\theta) \leq 0 \;\; \forall j \in \mathcal{G}.$$

Suppose the function h satisfies $|h(\theta, x_1, y_1) - h(\theta, x_2, y_2)| \leq c((x_1, y_1), (x_2, y_2))$ for any $(x_1, y_1) \neq (x_2, y_2)$ w.r.t a metric c. If $W_c(p_j, \hat{p}_j) \leq \gamma_j$ for all $j \in \mathcal{G}$, then the fairness criteria with respect to the true groups G will be satisfied within slacks γ_j for each group:

$$g_j(\theta) \leq \gamma_j \quad \forall j \in \mathcal{G}.$$

Proof. By the triangle inequality, for any group label j,

$$|g_j(\theta) - g(\theta)| \leq |g_j(\theta) - \hat{g}_j(\theta)| + \hat{g}_j(\theta)$$

By Kantorovich-Rubenstein theorem (provided here as Theorem 2), we also have

$$\begin{aligned}|\hat{g}_j(\theta) - g_j(\theta)| &= |\mathbb{E}_{X,Y\sim\hat{p}_j}[h(\theta, X, Y)] - \mathbb{E}_{X,Y\sim p_j}[h(\theta, X, Y)]| \\ &\leq W_c(p_j, \hat{p}_j).\end{aligned}$$

By the assumption that θ satisifes fairness constraints with respect to the noisy groups $\hat{G}$, $\hat{g}_j(\theta) \leq 0$. Therefore, combining these with the triangle inequality, we get the desired result. □

B Details on DRO formulation for TV distance

Here we describe the details on solving the DRO problem (3) with TV distance using the empirical Lagrangian formulation. We also provide the pseudocode we used for the projected gradient-based algorithm to solve it.

B.1 Empirical Lagrangian Formulation

We rewrite the constrained optimization problem (3) as a minimax problem using the Lagrangian formulation. We also convert all expectations into expectations over empirical distributions given a dataset of n samples $(X_1, Y_1, G_1), ..., (X_n, Y_n, G_n)$.

Let n_j denote the number of samples that belong to a true group $G = j$. Let the empirical distribution $\hat{p}_j \in \mathbb{R}^n$ be a vector with i-th entry $\hat{p}_j^i = \frac{1}{n_j}$ if the i-th example has a noisy group membership $\hat{G}_i = j$, and 0 otherwise. Replacing all expectations with expectations over the appropriate empirical distributions, the empirical form of (3) can be written as:

$$\begin{aligned}&\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n} l(\theta, X_i, Y_i) \\ &\text{s.t.} \quad \max_{\tilde{p}_j \in \mathbb{B}_{\gamma_j}(\hat{p}_j)} \sum_{i=1}^{n} \tilde{p}_j^i h(\theta, X_i, Y_i) \leq 0 \quad \forall j \in \mathcal{G}\end{aligned} \tag{9}$$

where $\mathbb{B}_{\gamma_j}(\hat{p}_j) = \{\tilde{p}_j \in \mathbb{R}^n : \frac{1}{2}\sum_{i=1}^{n} |\tilde{p}_j^i - \hat{p}_j^i| \leq \gamma_j, \sum_{i=1}^{n} \tilde{p}_j^i = 1, \tilde{p}_j^i \geq 0 \quad \forall i = 1, ..., n\}$.

For ease of notation, for $j \in \{1, 2, ..., m\}$, let

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(\theta, X_i, Y_i)$$

$$f_j(\theta, \tilde{p}_j) = \sum_{i=1}^{n} \tilde{p}_j^i h(\theta, X_i, Y_i).$$

Then the Lagrangian of the empirical formulation (9) is

$$\mathcal{L}(\theta, \lambda) = f(\theta) + \sum_{j=1}^{m} \lambda_j \max_{\tilde{p}_j \in \mathbb{B}_{\gamma}(\hat{p}_j)} f_j(\theta, \tilde{p}_j)$$

and problem (9) can be rewritten as

$$\min_{\theta} \max_{\lambda \geq 0} f(\theta) + \sum_{j=1}^{m} \lambda_j \max_{\tilde{p}_j \in \mathbb{B}_{\gamma}(\hat{p}_j)} f_j(\theta, \tilde{p}_j)$$

Moving the inner max out of the sum and rewriting the constraints as ℓ_1-norm constraints:

$$\begin{aligned}&\min_{\theta} \max_{\lambda \geq 0} \max_{\substack{\tilde{p}_j \in \mathbb{R}^n, \tilde{p}_j \geq 0, \\ j=1,...,m}} f(\theta) + \sum_{j=1}^{m} \lambda_j f_j(\theta, \tilde{p}_j) \\ &\text{s.t.} \ ||\tilde{p}_j - \hat{p}_j||_1 \leq 2\gamma_j, \ \ ||\tilde{p}_j||_1 = 1 \quad \forall j \in \{1, ..., m\}\end{aligned} \tag{10}$$

Since projections onto the ℓ_1-ball can be done efficiently [20], we can solve problem (10) using a projected gradient descent ascent (GDA) algorithm. This is a simplified version of the algorithm introduced by Namkoong and Duchi [46] for solving general classes of DRO problems. We provide pseudocode in Algorithm 2, as well as an actual implementation in the attached code.

B.2 Projected GDA Algorithm for DRO

Algorithm 2 Project GDA Algorithm

```
Require: learning rates $\eta_\theta > 0$, $\eta_\lambda > 0$, $\eta_p > 0$, estimates of $P(G \neq \hat{G} | \hat{G} = j)$ to specify $\gamma_j$.
for $t = 1, \ldots, T$ do
    Descent step on $\theta$:
    $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_\theta \nabla_\theta f(\theta^{(t)}) - \eta_\theta \sum_{j=1}^{m} \lambda_j^{(t)} \nabla_\theta f_j(\theta^{(t)}, \tilde{p}_j^{(t)})$
    Ascent step on $\lambda$:
    $\lambda_j^{(t+1)} \leftarrow \lambda_j^{(t)} + \eta_\lambda f_j(\theta, \tilde{p}_j^{(t)})$
    for $j = 1, ..., m$ do
        Ascent step on $\tilde{p}_j$: $\tilde{p}_j^{(t+1)} \leftarrow \tilde{p}_j^{(t)} + \eta_p \lambda_j^{(t)} \nabla_{\tilde{p}_j} f_j(\theta^{(t)}, \tilde{p}_j^{(t)})$
        Project $\tilde{p}_j^{(t+1)}$ onto $\ell_1$-norm constraints: $||\tilde{p}_j^{(t+1)} - \hat{p}_j||_1 \leq 2\gamma_j$, $||\tilde{p}_j^{(t+1)}||_1 = 1$
    end for
end for
return $\theta^{(t^*)}$ where $t^*$ denotes the best iterate that satisfies the constraints in (3) with the lowest objective.
```

B.3 Equalizing TPRs and FPRs using DRO

In the two case studies in Section 7, we enforce *equality of opportunity* and *equalized odds* [32] by equalizing true positive rates (TPRs) and/or false positive rates (FPRs) within some slack α. In this section, we describe in detail the implementation of the constraints for equalizing TPRs and FPRs under the DRO approach.

To equalize TPRs with slack α under the DRO approach, we set

$$\tilde{g}_j^{\text{TPR}}(\theta) = \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]} - \frac{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=1)]} - \alpha. \tag{11}$$

The first term corresponds to the TPR for the full population. The second term estimates the TPR for group j. Setting $\alpha = 0$ exactly equalizes true positive rates.

To equalize FPRs with slack α under the DRO approach, we set

$$\tilde{g}_j^{\text{FPR}}(\theta) = \frac{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=0)]} - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]} - \alpha. \tag{12}$$

The first term estimates the FPR for group j. The second term corresponds to the FPR for the full population. Setting $\alpha = 0$ exactly equalizes false positive rates.

To equalize TPRs for Case Study 1, we apply m constraints, $\left\{\max_{\tilde{p}_j : TV(\tilde{p}_j, \hat{p}_j) \leq \gamma_j, \tilde{p}_j \ll p} \tilde{g}_j^{\text{TPR}}(\theta) \leq 0\right\}\ \forall j \in \mathcal{G}$.

To equalize both TPRs and FPRs simultaneously for Case Study 2, we apply $2m$ constraints, $\left\{\max_{\tilde{p}_j : TV(\tilde{p}_j, \hat{p}_j) \leq \gamma_j, \tilde{p}_j \ll p} \tilde{g}_j^{\text{TPR}}(\theta) \leq 0, \max_{\tilde{p}_j : TV(\tilde{p}_j, \hat{p}_j) \leq \gamma_j, \tilde{p}_j \ll p} \tilde{g}_j^{\text{FPR}}(\theta) \leq 0\right\}\ \forall j \in \mathcal{G}$.

B.3.1 $h(\theta, X, Y)$ for equalizing TPRs and FPRs

Since the notation in Section 5 and in the rest of the paper uses generic functions h to express the group-specific constraints, we show in Lemma 2 that the constraint using $\tilde{g}_j^{\text{TPR}}(\theta)$ in Equation (11) can also be written as an equivalent constraint in the form of Equation (3), as

$$\tilde{g}_j^{\text{TPR}}(\theta) = \mathbb{E}_{X,Y\sim \tilde{p}_j}[h^{\text{TPR}}(\theta, X, Y)]$$

for some function $h^{\text{TPR}} : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Lemma 2. Denote $\hat{Y}$ as $\mathbb{1}(\phi(X;\theta) > 0)$. Let $h^{\text{TPR}}(\theta, X, Y)$ be given by

$$h^{\mathsf{TPR}}(\theta, X, Y) = \frac{1}{2}\left(-\mathbb{1}(\hat{Y}=1, Y=1) - \mathbb{1}(Y=1)\left(\alpha - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]}\right)\right).$$

Then

$$\begin{aligned}\frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]} - \frac{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=1)]} - \alpha \leq 0 \\ \iff \mathbb{E}_{X,Y\sim\tilde{p}_j}[h^{\mathsf{TPR}}(\theta,X,Y)] \leq 0.\end{aligned}$$

Proof. Substituting the given function $h^{\mathsf{TPR}}(\theta, X, Y)$, and using the fact that $\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=1)] \geq 0$:

$$\begin{aligned}&\mathbb{E}_{X,Y\sim\tilde{p}_j}[h^{\mathsf{TPR}}(\theta,X,Y)] \leq 0 \\ &\iff \mathbb{E}_{X,Y\sim\tilde{p}_j}\left[\frac{1}{2}\left(-\mathbb{1}(\hat{Y}=1, Y=1) - \mathbb{1}(Y=1)\left(\alpha - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]}\right)\right)\right] \leq 0 \\ &\iff -\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(\hat{Y}=1,Y=1)] - \mathbb{E}_{X,Y\sim\tilde{p}_j}\left[\mathbb{1}(Y=1)\left(\alpha - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]}\right)\right] \leq 0 \\ &\iff -\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(\hat{Y}=1,Y=1)] - \alpha\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=1)] \\ &\qquad + \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]}\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=1)] \leq 0 \\ &\iff \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=1)]} - \frac{\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(\hat{Y}=1,Y=1)]}{\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=1)]} - \alpha \leq 0\end{aligned}$$

□

By similar proof, we also show in Lemma 3 that the constraint using $\tilde{g}_j^{\mathsf{FPR}}(\theta)$ in Equation (12) can also be written as an equivalent constraint in the form of Equation (3), as

$$\tilde{g}_j^{\mathsf{FPR}}(\theta) = \mathbb{E}_{X,Y\sim\tilde{p}_j}[h^{\mathsf{FPR}}(\theta,X,Y)]$$

for some function $h^{\mathsf{FPR}} : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Lemma 3. Denote $\hat{Y}$ as $\mathbb{1}(\phi(X;\theta) > 0)$. Let $h^{\mathsf{FPR}}(\theta, X, Y)$ be given by

$$h^{\mathsf{FPR}}(\theta, X, Y) = \frac{1}{2}\left(\mathbb{1}(\hat{Y}=1, Y=0) - \mathbb{1}(Y=0)\left(\alpha + \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]}\right)\right).$$

Then

$$\begin{aligned}\frac{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim \tilde{p}_j}[\mathbb{1}(Y=0)]} - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]} - \alpha \leq 0 \\ \iff \mathbb{E}_{X,Y\sim\tilde{p}_j}[h^{\mathsf{FPR}}(\theta,X,Y)] \leq 0.\end{aligned}$$

Proof. Substituting the given function $h^{\text{FPR}}(\theta, X, Y)$, and using the fact that $\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=0)] \geq 0$:

$$
\begin{aligned}
&\mathbb{E}_{X,Y\sim\tilde{p}_j}[h^{\text{FPR}}(\theta, X, Y)] \leq 0 \\
\iff\; &\mathbb{E}_{X,Y\sim\tilde{p}_j}\left[\frac{1}{2}\left(\mathbb{1}(\hat{Y}=1, Y=0) - \mathbb{1}(Y=0)\left(\alpha + \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]}\right)\right)\right] \leq 0 \\
\iff\; &\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(\hat{Y}=1, Y=0)] - \mathbb{E}_{X,Y\sim\tilde{p}_j}\left[\mathbb{1}(Y=0)\left(\alpha + \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]}\right)\right] \leq 0 \\
\iff\; &\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(\hat{Y}=1, Y=0)] - \alpha\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=0)] \\
&\quad - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]}\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=0)] \leq 0 \\
\iff\; &\frac{\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(\hat{Y}=1, Y=0)]}{\mathbb{E}_{X,Y\sim\tilde{p}_j}[\mathbb{1}(Y=0)]} - \frac{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0,\hat{Y}=1)]}{\mathbb{E}_{X,Y\sim p}[\mathbb{1}(Y=0)]} - \alpha \leq 0
\end{aligned}
$$

□

B.4 DRO when $\hat{G}$ and G have different dimensionalities

The soft assignments approach is naturally formulated to be able to handle $G \in \mathcal{G} = \{1, ..., m\}$ and $\hat{G} \in \hat{\mathcal{G}} = \{1, ..., \hat{m}\}$ when $\hat{m} \neq m$. The DRO approach can be extended to handle this case by generalizing Lemma 1 to $TV(p_j, \hat{p}_i) \leq P(\hat{G} \neq i | G = j), j \in \mathcal{G}, i \in \hat{\mathcal{G}}$, and generalizing the DRO formulation to have the true group distribution p_j bounded in a TV distance ball centered at $\hat{p}_i$. Empirically comparing this generalized DRO approach to the soft group assignments approach when $\hat{m} \neq m$ is an interesting avenue of future work.

C Further details for soft group assignments approach

Here we provide additional technical details regarding the soft group assignments approach introduced in Section 7.

C.1 Derivation for $\mathbb{E}[h(\theta, X, Y)|G = j]$

Here we show $\mathbb{E}[h(\theta, X, Y)|G = j] = \frac{\mathbb{E}[h(\theta,X,Y)P(G=j|\hat{Y},Y,\hat{G})]}{P(G=j)}$, assuming that $h(\theta, X, Y)$ depends on X through $\hat{Y}$, i.e. $\hat{Y} = \mathbb{1}(\phi(\theta, X) > 0)$. Using the tower property and the definition of conditional expectation:

$$
\begin{aligned}
\mathbb{E}[h(\theta, X, Y)|G = j] &= \frac{\mathbb{E}[h(\theta, X, Y)\,\mathbb{1}(G = j)]}{P(G = j)} \\
&= \frac{\mathbb{E}[\mathbb{E}[h(\theta, X, Y)\,\mathbb{1}(G = j)|\hat{Y}, Y, \hat{G}]]}{P(G = j)} \\
&= \frac{\mathbb{E}[h(\theta, X, Y)\mathbb{E}[\mathbb{1}(G = j)|\hat{Y}, Y, \hat{G}]]}{P(G = j)} \\
&= \frac{\mathbb{E}[h(\theta, X, Y)P(G = j|\hat{Y}, Y, \hat{G})]}{P(G = j)}
\end{aligned}
\tag{13}
$$

C.2 Equalizing TPRs and FPRs using soft group assignments

In the two case studies in Section 7, we enforce *equality of opportunity* and *equalized odds* [32] by equalizing true positive rates (TPRs) and/or false positive rates (FPRs) within some slack α. In this section, we describe in detail the implementation of the constraints for equalizing TPRs and FPRs under the soft group assignments approach.

To equalize TPRs with slack α under the soft group assignments approach, we set

$$g_j^{\text{TPR}}(\theta, w) = \frac{\mathbb{E}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=1)]} - \frac{\mathbb{E}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)w(j|\hat{Y},Y,\hat{G})]}{\mathbb{E}[\mathbb{1}(Y=1)w(j|\hat{Y},Y,\hat{G})]} - \alpha. \tag{14}$$

The first term corresponds to the TPR for the full population. The second term estimates the TPR for group j as done by Kallus et al. [37] in Equation (5) and Proposition 8. Setting $\alpha = 0$ exactly equalizes true positive rates.

To equalize FPRs with slack α under the soft group assignments approach, we set

$$g_j^{\text{FPR}}(\theta, w) = \frac{\mathbb{E}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)w(j|\hat{Y},Y,\hat{G})]}{\mathbb{E}[\mathbb{1}(Y=0)w(j|\hat{Y},Y,\hat{G})]} - \frac{\mathbb{E}[\mathbb{1}(Y=0)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=0)]} - \alpha. \tag{15}$$

The first term estimates the FPR for group j as done previously for the TPR. The second term corresponds to the FPR for the full population. Setting $\alpha = 0$ exactly equalizes false positive rates.

To equalize TPRs for Case Study 1, we apply m constraints, $\left\{\max_{w\in\mathcal{W}(\theta)} g_j^{\text{TPR}}(\theta, w) \leq 0\right\}$ $\forall j \in \mathcal{G}$. To equalize both TPRs and FPRs simultaneously for Case Study 2, we apply $2m$ constraints, $\left\{\max_{w\in\mathcal{W}(\theta)} g_j^{\text{TPR}}(\theta, w) \leq 0, \max_{w\in\mathcal{W}(\theta)} g_j^{\text{FPR}}(\theta, w) \leq 0\right\}$ $\forall j \in \mathcal{G}$.

C.2.1 $h(\theta, X, Y)$ for equalizing TPRs and FPRs

Since the notation in Section 6 and in the rest of the paper uses generic functions h to express the group-specific constraints, we show in Lemma 4 that the constraint using $g_j^{\text{TPR}}(\theta, w)$ in Equation (14) can also be written as an equivalent constraint in the form of Equation (6), as

$$g_j^{\text{TPR}}(\theta, w) = \frac{\mathbb{E}[h^{\text{TPR}}(\theta, X, Y)w(j|\hat{Y},Y,\hat{G})]}{P(G=j)}$$

for some function $h^{\text{TPR}} : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Lemma 4. Denote $\hat{Y}$ as $\mathbb{1}(\phi(X;\theta) > 0)$. Let $h^{\text{TPR}}(\theta, X, Y)$ be given by

$$h^{\text{TPR}}(\theta, X, Y) = \frac{1}{2}\left(-\,\mathbb{1}(\hat{Y}=1, Y=1) - \mathbb{1}(Y=1)\left(\alpha - \frac{\mathbb{E}[\mathbb{1}(Y=1, \hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=1)]}\right)\right).$$

Then

$$\begin{aligned}\frac{\mathbb{E}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)]}{\mathbb{E}[\mathbb{1}(Y=1)]} - \frac{\mathbb{E}[\mathbb{1}(Y=1)\,\mathbb{1}(\hat{Y}=1)w(j|\hat{Y},Y,\hat{G})]}{\mathbb{E}[\mathbb{1}(Y=1)w(j|\hat{Y},Y,\hat{G})]} - \alpha \leq 0 \\ \iff \frac{\mathbb{E}[h^{\text{TPR}}(\theta, X, Y)w(j|\hat{Y},Y,\hat{G})]}{P(G=j)} \leq 0.\end{aligned}$$

for all $j \in \mathcal{G}, P(G=j) > 0$.

Proof. Substituting the given function $h^{\text{TPR}}(\theta, X, Y)$, and using the fact that $P(G = j) > 0$ and $\mathbb{E}[\mathbb{1}(Y = 1)w(j|\hat{Y}, Y, \hat{G})] \geq 0$:

$$\begin{aligned}
&\frac{\mathbb{E}[h^{\text{TPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})]}{P(G = j)} \leq 0 \\
&\iff \mathbb{E}[h^{\text{TPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})] \leq 0 \\
&\iff \mathbb{E}\left[\frac{1}{2}\left(-\mathbb{1}(\hat{Y} = 1, Y = 1) - \mathbb{1}(Y = 1)\left(\alpha - \frac{\mathbb{E}[\mathbb{1}(Y = 1, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 1)]}\right)\right) w(j|\hat{Y}, Y, \hat{G})\right] \leq 0 \\
&\iff -\mathbb{E}[\mathbb{1}(\hat{Y} = 1, Y = 1)w(j|\hat{Y}, Y, \hat{G})] \\
&\qquad - \mathbb{E}\left[\mathbb{1}(Y = 1)\left(\alpha - \frac{\mathbb{E}[\mathbb{1}(Y = 1, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 1)]}\right) w(j|\hat{Y}, Y, \hat{G})\right] \leq 0 \\
&\iff -\mathbb{E}[\mathbb{1}(\hat{Y} = 1, Y = 1)w(j|\hat{Y}, Y, \hat{G})] - \alpha\mathbb{E}[\mathbb{1}(Y = 1)w(j|\hat{Y}, Y, \hat{G})] \\
&\qquad + \frac{\mathbb{E}[\mathbb{1}(Y = 1, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 1)]}\mathbb{E}[\mathbb{1}(Y = 1)w(j|\hat{Y}, Y, \hat{G})] \leq 0 \\
&\iff \frac{\mathbb{E}[\mathbb{1}(Y = 1, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 1)]} - \frac{\mathbb{E}[\mathbb{1}(\hat{Y} = 1, Y = 1)w(j|\hat{Y}, Y, \hat{G})]}{\mathbb{E}[\mathbb{1}(Y = 1)w(j|\hat{Y}, Y, \hat{G})]} - \alpha \leq 0
\end{aligned}$$

□

By similar proof, we also show in Lemma 5 that the constraint using $g_j^{\text{FPR}}(\theta, w)$ in Equation (15) can also be written as an equivalent constraint in the form of Equation (6), as

$$g_j^{\text{FPR}}(\theta, w) = \frac{\mathbb{E}[h^{\text{FPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})]}{P(G = j)}$$

for some function $h^{\text{FPR}} : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Lemma 5. *Denote $\hat{Y}$ as $\mathbb{1}(\phi(X; \theta) > 0)$. Let $h^{\text{FPR}}(\theta, X, Y)$ be given by*

$$h^{\text{FPR}}(\theta, X, Y) = \frac{1}{2}\left(\mathbb{1}(\hat{Y} = 1, Y = 0) - \mathbb{1}(Y = 0)\left(\alpha + \frac{\mathbb{E}[\mathbb{1}(Y = 0, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 0)]}\right)\right).$$

Then

$$\begin{aligned}
&\frac{\mathbb{E}[\mathbb{1}(Y = 0)\,\mathbb{1}(\hat{Y} = 1)w(j|\hat{Y}, Y, \hat{G})]}{\mathbb{E}[\mathbb{1}(Y = 0)w(j|\hat{Y}, Y, \hat{G})]} - \frac{\mathbb{E}[\mathbb{1}(Y = 0)\,\mathbb{1}(\hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 0)]} - \alpha \leq 0 \\
&\qquad\qquad\qquad\qquad \iff \frac{\mathbb{E}[h^{\text{FPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})]}{P(G = j)} \leq 0.
\end{aligned}$$

for all $j \in \mathcal{G}, P(G = j) > 0$.

Proof. Substituting the given function $h^{\text{FPR}}(\theta, X, Y)$, and using the fact that $P(G = j) > 0$ and $\mathbb{E}[\mathbb{1}(Y = 0)w(j|\hat{Y}, Y, \hat{G})] \geq 0$:

$$
\begin{aligned}
&\frac{\mathbb{E}[h^{\text{FPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})]}{P(G = j)} \leq 0 \\
&\iff \mathbb{E}[h^{\text{FPR}}(\theta, X, Y)w(j|\hat{Y}, Y, \hat{G})] \leq 0 \\
&\iff \mathbb{E}\left[\frac{1}{2}\left(\mathbb{1}(\hat{Y} = 1, Y = 0) - \mathbb{1}(Y = 0)\left(\alpha + \frac{\mathbb{E}[\mathbb{1}(Y = 0, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 0)]}\right)\right) w(j|\hat{Y}, Y, \hat{G})\right] \leq 0 \\
&\iff \mathbb{E}[\mathbb{1}(\hat{Y} = 1, Y = 0)w(j|\hat{Y}, Y, \hat{G})] \\
&\qquad - \mathbb{E}\left[\mathbb{1}(Y = 0)\left(\alpha + \frac{\mathbb{E}[\mathbb{1}(Y = 0, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 0)]}\right) w(j|\hat{Y}, Y, \hat{G})\right] \leq 0 \\
&\iff \mathbb{E}[\mathbb{1}(\hat{Y} = 1, Y = 0)w(j|\hat{Y}, Y, \hat{G})] - \alpha\mathbb{E}[\mathbb{1}(Y = 0)w(j|\hat{Y}, Y, \hat{G})] \\
&\qquad - \frac{\mathbb{E}[\mathbb{1}(Y = 0, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 0)]}\mathbb{E}[\mathbb{1}(Y = 0)w(j|\hat{Y}, Y, \hat{G})] \leq 0 \\
&\iff \frac{\mathbb{E}[\mathbb{1}(\hat{Y} = 1, Y = 0)w(j|\hat{Y}, Y, \hat{G})]}{\mathbb{E}[\mathbb{1}(Y = 0)w(j|\hat{Y}, Y, \hat{G})]} - \frac{\mathbb{E}[\mathbb{1}(Y = 0, \hat{Y} = 1)]}{\mathbb{E}[\mathbb{1}(Y = 0)]} - \alpha \leq 0
\end{aligned}
$$

□

D Optimality and feasibility for the *Ideal* algorithm

D.1 Optimality and feasibility guarantees

We provide optimality and feasibility guarantees for Algorithm 1 and optimality guarantees for Algorithm 3.

Theorem 4 (Optimality and Feasibility for Algorithm 1). *Let $\theta^* \in \Theta$ be such that it satisfies the constraints $\max\limits_{w \in \mathcal{W}(\theta)} g_j(\theta^*, w) \leq 0$, $\forall j \in \mathcal{G}$ and $f_0(\theta^*) \leq f(\theta)$ for every $\theta \in \Theta$ that satisfies the same constraints. Let $0 \leq f_0(\theta) \leq B, \forall \theta \in \Theta$. Let the space of Lagrange multipliers be defined as $\Lambda = \{\lambda \in \mathbb{R}_+^m \mid \|\lambda\|_1 \leq R\}$, for $R > 0$. Let $B_\lambda \geq \max_t \|\nabla_\lambda \mathcal{L}(\theta^{(t)}, \lambda^{(t)})\|_2$. Let $\overline{\theta}$ be the stochastic classifier returned by Algorithm 1 when run for T iterations, with the radius of the Lagrange multipliers $R = T^{1/4}$ and learning rate $\eta_\lambda = \frac{R}{B_\lambda\sqrt{T}}$ Then:*

$$
\mathbf{E}_{\theta \sim \overline{\theta}}\left[f(\theta)\right] \leq f(\theta^*) + \mathcal{O}\left(\frac{1}{T^{1/4}}\right) + \rho
$$

and

$$
\mathbf{E}_{\theta \sim \overline{\theta}}\left[\max_{w \in \mathcal{W}(\theta)} g_j(\theta, w)\right] \leq \mathcal{O}\left(\frac{1}{T^{1/4}}\right) + \rho'
$$

Thus for any given $\varepsilon > 0$, by solving Steps 2 and 4 of Algorithm 1 to sufficiently small errors ρ, ρ', and by running the algorithm for a sufficiently large number of steps T, we can guarantee that the returned stochastic model is ε-optimal and ε-feasible.

Proof. Let $\overline{\lambda} = \frac{1}{T}\sum_{t=1}^{T} \lambda^{(t)}$. We will interpret the minimax problem in (8) as a zero-sum between the θ-player who optimizes $\mathcal{L}$ over θ, and the λ-player who optimizes $\mathcal{L}$ over λ. We first bound the average regret incurred by the players over T steps. The best response computation in Step 2 of Algorithm 1 gives us:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\theta \sim \hat{\theta}^{(t)}}\left[\mathcal{L}(\theta, \lambda^{(t)})\right] &\leq \frac{1}{T}\sum_{t=1}^{T} \min_{\theta \in \Theta} \mathcal{L}(\theta, \lambda^{(t)}) + \varepsilon \\
&\leq \min_{\theta \in \Theta} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(\theta, \lambda^{(t)}) + \rho
\end{aligned}
$$

$$
\begin{aligned}
&= \min_{\theta\in\Theta} \mathcal{L}(\theta, \overline{\lambda}) \,+\, \rho \\
&\leq \min_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathcal{L}(\theta, \lambda) \,+\, \rho \\
&\leq f(\theta^*) + \rho. \qquad (16)
\end{aligned}
$$

We then apply standard gradient ascent analysis for the projected gradient updates to λ in Step 4 of the algorithm, and get:

$$\max_{\lambda\in\Lambda} \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\lambda_j\delta_j^{(t)} \;\geq\; \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\lambda_j^{(t)}\delta_j^{(t)} \,-\, \mathcal{O}\left(\frac{R}{\sqrt{T}}\right).$$

We then plug the upper and lower bounds for the gradient estimates $\delta_j^{(t)}$'s from Step 3 of the Algorithm 1 into the above inequality:

$$
\begin{aligned}
&\max_{\lambda\in\Lambda} \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\lambda_j\left(\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)\right] \,+\, \rho'\right) \\
&\geq \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{m}\lambda_j^{(t)}\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)\right] \,-\, \mathcal{O}\left(\frac{R}{\sqrt{T}}\right).
\end{aligned}
$$

which further gives us:

$$
\begin{aligned}
&\max_{\lambda\in\Lambda}\left\{\sum_{j=1}^{m}\lambda_j\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)\right] \,+\, \|\lambda\|_1\rho'\right\} \\
&\geq \sum_{j=1}^{m}\lambda_j^{(t)}\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)\right] \,-\, \mathcal{O}\left(\frac{R}{\sqrt{T}}\right).
\end{aligned}
$$

Adding $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[f(\theta)\right]$ to both sides of the above inequality, we finally get:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\mathcal{L}(\theta, \lambda^{(t)})\right] \;\geq\; \max_{\lambda\in\Lambda}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\mathcal{L}(\theta, \lambda)\right] \,+\, \|\lambda\|_1\rho'\right\} \,-\, \mathcal{O}\left(\frac{R}{\sqrt{T}}\right). \quad (17)$$

Optimality. Now, substituting $\lambda = \mathbf{0}$ in (17) and combining with (16) completes the proof of the optimality guarantee:

$$\mathbb{E}_{\theta\sim\overline{\theta}}\left[f(\theta)\right] \;\leq\; f_0(\theta^*) \,+\, \mathcal{O}\left(\frac{R}{\sqrt{T}}\right) \,+\, \rho$$

Feasibility. To show feasibility, we fix a constraint index $j \in \mathcal{G}$. Now substituting $\lambda_j = R$ and $\lambda_{j'} = 0, \forall j' \neq j$ in (17) and combining with (16) gives us:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\theta\sim\hat{\theta}^{(t)}}\left[f(\theta) + R\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)\right] \;\leq\; f(\theta^*) \,+\, \mathcal{O}\left(\frac{R}{\sqrt{T}}\right) \,+\, \rho \,+\, R\rho'.$$

which can be re-written as:

$$
\begin{aligned}
\mathbb{E}_{\theta\sim\overline{\theta}}\left[\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)\right] \;&\leq\; \frac{f(\theta^*) \,-\, \mathbb{E}_{\theta\sim\overline{\theta}}\left[f(\theta)\right]}{R} \,+\, \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) \,+\, \frac{\rho}{R} \,+\, \rho'. \\
&\leq\; \frac{B}{R} \,+\, \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) \,+\, \frac{\rho}{R} \,+\, \rho',
\end{aligned}
$$

which is our feasibility guarantee. Setting $R = \mathcal{O}(T^{1/4})$ then completes the proof. □

D.2 Best Response over θ

We next describe our procedure for computing a best response over θ in Step 2 of Algorithm 1. We will consider a slightly relaxed version of the best response problem where the equality constraints in $\mathcal{W}(\theta)$ are replaced with closely-approximating inequality constraints.

Algorithm 3 Best response on θ of Algorithm 1

```
Require: $\lambda'$, learning rate $\eta_{\mathbf{w}} > 0$, estimates of $P(G = j|\hat{G} = k)$ to specify constraints $r_{g,\hat{g}}$'s, $\kappa$
for $q = 1, \ldots, Q$ do
    Best response on ($\theta$, $\boldsymbol{\mu}$): use an oracle to find find $\theta^{(q)} \in \Theta$ and $\boldsymbol{\mu}^{(q)} \in \mathcal{M}^m$ such that:
        $\ell(\theta^{(q)}, \boldsymbol{\mu}^{(q)}, \mathbf{w}^{(q)}; \lambda') \le \min_{\theta \in \Theta, \, \boldsymbol{\mu} \in \mathcal{M}^m} \ell(\theta, \boldsymbol{\mu}, \mathbf{w}^{(q)}; \lambda') + \kappa,$
    for a small slack $\kappa > 0$.
    Ascent step on $\mathbf{w}$:
        $w_j^{(q+1)} \leftarrow \Pi_{\mathcal{W}_\Delta}\left(w_j^{(q)} + \eta_{\mathbf{w}} \nabla_{w_j} \ell(\theta^{(q)}, \boldsymbol{\mu}^{(q)}, \mathbf{w}^{(q)}; \lambda')\right),$
    where $\nabla_{w_j} \ell(\cdot)$ is a sub-gradient of $\ell$ w.r.t. $w_j$.
end for
return A uniform distribution $\hat{\theta}$ over $\theta^{(1)}, \ldots, \theta^{(Q)}$
```

Recall that the constraint set $\mathcal{W}(\theta)$ contains two sets of constraints (5), the total probability constraints that depend on θ, and the simplex constraints that do not depend on θ. So to decouple these constraint sets from θ, we introduce Lagrange multipliers μ for the total probability constraints to make them a part of the objective, and obtain a nested *minimax* problem over θ, μ, and w, where w is constrained to satisfy the simplex constraints alone. We then jointly minimize the inner Lagrangian over θ and μ, and perform gradient ascent updates on w with projections onto the simplex constraints. The joint-minimization over θ and μ is not necessarily convex and is solved using a minimization oracle.

We begin by writing out the best-response problem over θ for a fixed λ':

$$\min_{\theta \in \Theta} \mathcal{L}(\theta, \lambda') \;=\; \min_{\theta \in \Theta} f(\theta) + \sum_{j=1}^{m} \lambda'_j \max_{w_j \in \mathcal{W}(\theta)} g_j(\theta, w_j), \tag{18}$$

where we use w_j to denote the maximizer over $\mathcal{W}(\theta)$ for constraint g_j explicitly. We separate out the the simplex constraints in $\mathcal{W}(\theta)$ (5) and denote them by:

$$\mathcal{W}_\Delta = \left\{ w \in \mathbb{R}_+^{\mathcal{G} \times \{0,1\}^2 \times \hat{\mathcal{G}}} \;\middle|\; \sum_{j=1}^{m} w(j \mid \hat{y}, y, k) = 1, \; \forall k \in \hat{\mathcal{G}}, y, \hat{y} \in \{0,1\} \right\},$$

where we represent each w as a vector of values $w(i|\hat{y}, y, k)$ for each $j \in \mathcal{G}, \hat{y} \in \{0,1\}, y \in \{0,1\}$, and $k \in \hat{\mathcal{G}}$. We then relax the total probability constraints in $\mathcal{W}(\theta)$ into a set of inequality constraints:

$$P(G = j|\hat{G} = k) \;-\; \sum_{\hat{y}, y \in \{0,1\}} w(j \mid \hat{y}, y, k) P(\hat{Y}(\theta) = \hat{y}, Y = y|\hat{G} = k) \;-\; \tau \quad \le \quad 0$$

$$\sum_{\hat{y}, y \in \{0,1\}} w(j \mid \hat{y}, y, k) P(\hat{Y}(\theta) = \hat{y}, Y = y|\hat{G} = k) \;-\; P(G = j|\hat{G} = k) \;-\; \tau \quad \le \quad 0$$

for some small $\tau > 0$. We have a total of $U = 2 \times m \times \hat{m}$ relaxed inequality constraints, and will denote each of them as $r_u(\theta, w) \le 0$, with index u running from 1 to U. Note that each $r_u(\theta, w)$ is linear in w.

Introducing Lagrange multipliers μ for the relaxed total probability constraints, the optimization problem in (18) can be re-written equivalently as:

$$\min_{\theta \in \Theta} f(\theta) \;+\; \sum_{j=1}^{m} \lambda'_j \max_{w_j \in \mathcal{W}_\Delta} \min_{\mu_j \in \mathcal{M}} \left\{ g_j(\theta, w_j) - \sum_{u=1}^{U} \mu_{j,u} \, r_u(\theta, w_j) \right\},$$

where note that each w_j is maximized over only the simplex constraints $\mathcal{W}_\Delta$ which are independent of θ, and $\mathcal{M} = \{\mu_j \in \mathbb{R}_+^{m \times \hat{m}} \mid \|\mu_j\|_1 \le R'\}$, for some constant $R' > 0$. Because each w_j and μ_j appears only in the j-th term in the summation, we can pull out the max and min, and equivalently rewrite the above problem as:

$$\min_{\theta\in\Theta}\;\max_{\mathbf{w}\in\mathcal{W}_\Delta^m}\;\min_{\boldsymbol{\mu}\in\mathcal{M}^m}\;\underbrace{f(\theta)+\sum_{j=1}^{m}\lambda'_j\bigg(\underbrace{g_j(\theta,w_j)-\sum_{u=1}^{U}\mu_{j,u}\,r_u(\theta,w_j)}_{\omega(\theta,\mu_j,w_j)}\bigg)}_{\ell(\theta,\boldsymbol{\mu},\mathbf{w};\lambda')},\tag{19}$$

where $\mathbf{w}=(w_1,\ldots,w_m)$ and $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_m)$. We then solve this nested minimax problem in Algorithm 3 by using an minimization *oracle* to perform a full optimization of ℓ over (θ, μ), and carrying out gradient ascent updates on ℓ over w_j.

We now proceed to show an optimality guarantee for Algorithm 3.

Theorem 5 (**Optimality Guarantee for Algorithm 3**). *Suppose for every $\theta\in\Theta$, there exists a $\widetilde{w}_j\in\mathcal{W}_\Delta$ such that $r_u(\theta,\widetilde{w}_j)\le-\gamma$, $\forall u\in[U]$, for some $\gamma>0$. Let $0\le g_j(\theta,w_j)\le B'$, $\forall\theta\in\Theta, w_j\in\mathcal{W}_\Delta$. Let $B_{\mathbf{w}}\ge\max_q\|\nabla_{\mathbf{w}}\,\ell(\theta^{(q)},\boldsymbol{\mu}^{(q)},\mathbf{w}^{(q)};\lambda'))\|_2$. Let $\hat{\theta}$ be the stochastic classifier returned by Algorithm 3 when run for a given λ' for Q iterations, with the radius of the Lagrange multipliers $R'=B'/\gamma$ and learning rate $\eta_{\mathbf{w}}=\frac{R'}{B_{\mathbf{w}}\sqrt{T}}$. Then:*

$$\mathbb{E}_{\theta\sim\hat{\theta}}\left[\mathcal{L}(\theta,\lambda')\right]\;\le\;\min_{\theta\in\Theta}\,\mathcal{L}(\theta,\lambda')\,+\,\mathcal{O}\left(\frac{1}{\sqrt{Q}}\right)+\kappa.$$

Before proving Theorem 5, we will find it useful to state the following lemma.

Lemma 6 (**Boundedness of Inner Lagrange Multipliers in** (19))**.** Suppose for every $\theta\in\Theta$, there exists a $\widetilde{w}_j\in\mathcal{W}$ such that $r_u(\theta,\widetilde{w}_j)\le-\gamma$, $\forall u\in[U]$, for some $\gamma>0$. Let $0\le g_j(\theta,w_j)\le B'$, $\forall\theta\in\Theta, w_j\in\mathcal{W}_\Delta$. Let $\mathcal{M}=\{\mu_j\in\mathbb{R}_+^K\,|\,\|\mu_j\|_1\le R'\}$ with the radius of the Lagrange multipliers $R'=B'/\gamma$. Then we have for all $j\in\mathcal{G}$:

$$\max_{w_j\in\mathcal{W}_\Delta}\;\min_{\mu_j\in\mathcal{M}}\;\omega\big(\theta,\mu_j,w_j\big)=\max_{w_j\in\mathcal{W}_\Delta:\,r_u(\theta,w_j)\le0,\,\forall u}g_j(\theta,w_j).$$

Proof. For a given $j\in\mathcal{G}$, let $w_j^*\in\operatorname*{argmax}_{w_j\in\mathcal{W}_\Delta:\,r_u(\theta,w_j)\le0,\,\forall u}g_j(\theta,w_j)$. Then:

$$g_j(\theta,w_j^*)\;\;=\;\;\max_{w_j\in\mathcal{W}_\Delta}\;\min_{\mu_j\in\mathbb{R}_+^K}\;\omega\big(\theta,\mu_j,w_j\big),\tag{20}$$

where note that μ_j is minimized over all non-negative values. Since the ω is linear in both μ_j and w_j, we can interchange the min and max:

$$g_j(\theta,w_j^*)\;\;=\;\;\min_{\mu_j\in\mathbb{R}_+^K}\;\max_{w_j\in\mathcal{W}_\Delta}\;\omega\big(\theta,\mu_j,w_j\big).$$

We show below that the minimizer μ^* in the above problem is in fact bounded and present in $\mathcal{M}$.

$$\begin{aligned}
g_j(\theta,w_j^*)\;&=\;\max_{w_j\in\mathcal{W}}\;\omega\big(\theta,\mu_j^*,w_j\big)\\
&=\;\max_{w_j\in\mathcal{W}}\;\left\{g_j(\theta,w_j)\;-\;\sum_{k=1}^{K}\mu_{j,k}^*\,r_k(\theta,w_j)\right\}\\
&\ge\;g_j(\theta,\widetilde{w}_j)\;-\;\|\mu_j^*\|_1\;\max_{k\in[K]}\;r_k(\theta,\widetilde{w}_j)\\
&\ge\;g_j(\theta,w_j)\;+\;\|\mu_j^*\|_1\gamma\;\ge\;\|\mu_j^*\|_1\gamma.
\end{aligned}$$

We further have:

$$\|\mu_j^*\|_1\;\le\;g_j(\theta,w_j)/\gamma\;\le\;B'/\gamma.\tag{21}$$

Thus the minimizer $\mu_j^*\in\mathcal{M}$. So the minimization in (20) can be performed over only $\mathcal{M}$, which completes the proof of the lemma. □

Equipped with the above result, we are now ready to prove Theorem 5.

Proof of Theorem 5. Let $\overline{w}_j = \frac{1}{Q}\sum_{q=1}^{Q} w_j^{(q)}$. The best response on θ and μ gives us:

$$
\begin{aligned}
&\frac{1}{Q}\sum_{q=1}^{Q}\Big(f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j\,\omega\big(\theta^{(q)},\mu_j^{(q)},w_j^{(q)}\big)\Big)\\
&\quad\leq\ \frac{1}{Q}\sum_{q=1}^{Q}\min_{\theta\in\Theta,\,\boldsymbol{\mu}\in\mathcal{M}^m}\Big(f(\theta) + \sum_{j=1}^{m}\lambda'_j\,\omega\big(\theta,\mu_j,w_j^{(q)}\big)\Big) + \kappa\\
&\quad=\ \frac{1}{Q}\sum_{q=1}^{Q}\Big(\min_{\theta\in\Theta} f(\theta) + \sum_{j=1}^{m}\lambda'_j\min_{\mu_j\in\mathcal{M}}\omega\big(\theta,\mu_j,w_j^{(q)}\big)\Big) + \kappa \quad (j\text{-th summation term depends on } \mu_j \text{ alone})\\
&\quad\leq\ \min_{\theta\in\Theta}\frac{1}{Q}\sum_{q=1}^{Q}\Big(f(\theta) + \sum_{j=1}^{m}\lambda'_j\min_{\mu_j\in\mathcal{M}}\omega\big(\theta,\mu_j,w_j^{(q)}\big)\Big) + \kappa\\
&\quad\leq\ \min_{\theta\in\Theta}\Big\{f(\theta) + \sum_{j=1}^{m}\lambda'_j\min_{\mu_j\in\mathcal{M}}\frac{1}{Q}\sum_{q=1}^{Q}\omega\big(\theta,\mu_j,w_j^{(q)}\big)\Big\} + \kappa\\
&\quad=\ \min_{\theta\in\Theta}\Big\{f(\theta) + \sum_{j=1}^{m}\lambda'_j\min_{\mu_j\in\mathcal{M}}\ \omega\big(\theta,\mu_j,\overline{w}_j\big)\Big\} + \kappa\\
&\quad\leq\ \min_{\theta\in\Theta}\Big\{f(\theta) + \sum_{j=1}^{m}\lambda'_j\max_{w_j\in\mathcal{W}}\min_{\mu_j\in\mathcal{M}}\ \omega\big(\theta,\mu_j,w_j\big)\Big\} + \kappa \quad (\text{by linearity of } \omega \text{ in } w_j)\\
&\quad=\ \min_{\theta\in\Theta}\Big\{f(\theta) + \sum_{j=1}^{m}\lambda'_j\max_{w_j:\,r_u(\theta,w_j)\leq 0,\,\forall u} g_j(\theta,w_j)\Big\} + \kappa \quad (\text{from Lemma 6})\\
&\quad=\ \min_{\theta\in\Theta}\mathcal{L}(\theta,\lambda') + \kappa.
\end{aligned}
\qquad (22)
$$

Applying standard gradient ascent analysis to the gradient ascent steps on $\mathbf{w}$ (using the fact that ω is linear in $\mathbf{w}$)

$$
\begin{aligned}
&\frac{1}{Q}\sum_{q=1}^{Q}\Big(f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j\,\omega\big(\theta^{(q)},\mu_j^{(q)},w_j^{(q)}\big)\Big)\\
&\quad\geq\ \max_{\mathbf{w}\in\mathcal{W}_\Delta^m}\frac{1}{Q}\sum_{q=1}^{Q}\Big(f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j\,\omega\big(\theta^{(q)},\mu_j^{(q)},w_j\big)\Big) - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right)\\
&\quad=\ \frac{1}{Q}\sum_{q=1}^{Q}\Big(f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j\max_{w_j\in\mathcal{W}_\Delta}\omega\big(\theta^{(q)},\mu_j^{(q)},w_j\big)\Big) - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) \quad (j\text{-th summation term depends on } w_j \text{ alone})\\
&\quad\geq\ \frac{1}{Q}\sum_{q=1}^{Q}\Big(f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j\max_{w_j\in\mathcal{W}_\Delta}\min_{\mu_j\in\mathcal{M}}\omega\big(\theta^{(q)},\mu_j,w_j\big)\Big) - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) \quad (\text{by linearity of } \omega \text{ in } w_j \text{ and } \mu_j)\\
&\quad=\ \mathbb{E}_{\theta\sim\hat{\theta}}\left[f(\theta) + \sum_{j=1}^{m}\lambda'_j\max_{w_j\in\mathcal{W}_\Delta}\min_{\mu_j\in\mathcal{M}}\omega\big(\theta,\mu_j,w_j\big)\right] - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right)\\
&\quad=\ \mathbb{E}_{\theta\sim\hat{\theta}}\left[f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j\max_{w_j\in\mathcal{W}_\Delta:\,r_u(\theta,w_j)\leq 0,\,\forall u} g_j(\theta,w_j)\right] - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) \quad (\text{from Lemma 6})\\
&\quad=\ \mathbb{E}_{\theta\sim\hat{\theta}}\left[\mathcal{L}(\theta,\lambda')\right] - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right).
\end{aligned}
\qquad (23)
$$

Combining (22) and (23) completes the proof. □

Algorithm 4 *Practical* Algorithm

```
Require: learning rates $\eta_\theta > 0$, $\eta_\lambda > 0$, estimates of
    $P(G = j|\hat{G} = k)$ to specify $\mathcal{W}(\theta)$
for $t = 1, \ldots, T$ do
    Solve for $w$ given $\theta$ using linear programming or a gradient method:
    $w^{(t)} \leftarrow \max_{w\in\mathcal{W}(\theta^{(t)})} \sum_{j=1}^{m} \lambda_j^{(t)} g_j(\theta^{(t)}, w)$
    Descent step on $\theta$:
    $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_\theta \delta_\theta^{(t)}$, where
    $\delta_\theta^{(t)} = \nabla_\theta \left( f_0(\theta^{(t)}) + \sum_{j=1}^{m} \lambda_j^{(t)} g_j \left(\theta^{(t)}, w^{(t+1)}\right)\right)$
    Ascent step on $\lambda$:
    $\tilde{\lambda}_j^{(t+1)} \leftarrow \lambda_j^{(t)} + \eta_\lambda g_j \left(\theta^{(t+1)}, w^{(t+1)}\right) \quad \forall j \in \mathcal{G}$
    $\lambda^{(t+1)} \leftarrow \Pi_\Lambda(\tilde{\lambda}^{(t+1)})$,
end for
return $\theta^{(t^*)}$ where $t^*$ denotes the best iterate that satisfies the constraints in (7) with the lowest
    objective.
```

E Discussion on the *Practical* algorithm

Here we provide the details of the *practical* Algorithm 4 to solve problem (8). We also further discuss how we arrive at Algorithm 4. Recall that in the minimax problem in (8), restated below, each of the m constraints contain a max over w:

$$\min_{\theta\in\Theta} \max_{\lambda\in\Lambda} f(\theta) + \sum_{j=1}^{m} \lambda_j \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w).$$

We show below that this is equivalent to a minimax problem where the sum over j and max over w are swapped:

Lemma 7. The minimax problem in (8) is equivalent to:

$$\min_{\theta\in\Theta} \max_{\lambda\in\Lambda} \max_{w\in\mathcal{W}(\theta)} f(\theta) + \sum_{j=1}^{m} \lambda_j g_j(\theta, w). \tag{24}$$

Proof. Recall that the space of Lagrange multipliers $\Lambda = \{\lambda \in \mathbb{R}_+^m \mid \|\lambda\|_1 \leq R\}$, for $R > 0$. So the above maximization over Λ can be re-written in terms of a maximization over the m-dimensional simplex Δ_m and a scalar $\beta \in [0, R]$:

$$\begin{aligned}
&\min_{\theta\in\Theta} \max_{\beta\in[0,R],\, \nu\in\Delta_m} f(\theta) + \beta \sum_{j=1}^{m} \nu_j \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w) \\
&= \min_{\theta\in\Theta} \max_{\beta\in[0,R]} f(\theta) + \beta \max_{\nu\in\Delta_m} \sum_{j=1}^{m} \nu_j \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w) \\
&= \min_{\theta\in\Theta} \max_{\beta\in[0,R]} f(\theta) + \beta \max_{j\in\mathcal{G}} \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w) \\
&= \min_{\theta\in\Theta} \max_{\beta\in[0,R]} f(\theta) + \beta \max_{w\in\mathcal{W}(\theta)} \max_{j\in\mathcal{G}} g_j(\theta, w) \\
&= \min_{\theta\in\Theta} \max_{\beta\in[0,R]} f(\theta) + \beta \max_{w\in\mathcal{W}(\theta)} \max_{\nu\in\Delta_m} \sum_{j=1}^{m} \nu_j g_j(\theta, w) \\
&= \min_{\theta\in\Theta} f(\theta) + \max_{\beta\in[0,R],\, \nu\in\Delta_m} \max_{w\in\mathcal{W}(\theta)} \sum_{j=1}^{m} \beta\nu_j g_j(\theta, w) \\
&= \min_{\theta\in\Theta} f(\theta) + \max_{\lambda\in\Lambda} \max_{w\in\mathcal{W}(\theta)} \sum_{j=1}^{m} \lambda_j g_j(\theta, w),
\end{aligned}$$

which completes the proof. □

The practical algorithm outlined in Algorithm 4 seeks to solve the re-written minimax problem in (24), and is similar in structure to the ideal algorithm in Algorithm 1, in that it has two high-level steps: an approximate best response over θ and gradient ascent updates on λ. However, the algorithm works with deterministic classifiers $\theta^{(t)}$, and uses a simple heuristic to approximate the best response step. Specifically, for the best response step, the algorithm finds the maximizer of the Lagrangian over w for a fixed $\theta^{(t)}$ by e.g. using linear programming:

$$w^{(t)} \leftarrow \max_{w \in \mathcal{W}(\theta^{(t)})} \sum_{j=1}^{m} \lambda_j^{(t)} g_j(\theta^{(t)}, w),$$

uses the maximizer $w^{(t)}$ to approximate the gradient of the Lagrangian at $\theta^{(t)}$:

$$\delta_\theta^{(t)} = \nabla_\theta \Big(f_0(\theta^{(t)}) + \sum_{j=1}^{m} \lambda_j^{(t)} f_j \left(\theta^{(t)}, w^{(t+1)} \right) \Big)$$

and performs a single gradient update on θ:

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_\theta \delta_\theta^{(t)}.$$

The gradient ascent step on λ is the same as the ideal algorithm, except that it is simpler to implement as the iterates $\theta^{(t)}$ are deterministic:

$$\tilde{\lambda}_j^{(t+1)} \leftarrow \lambda_j^{(t)} + \eta_\lambda f_j \left(\theta^{(t+1)}, w^{(t+1)} \right) \ \ \forall j \in \mathcal{G};$$

$$\lambda^{(t+1)} \leftarrow \Pi_\Lambda(\tilde{\lambda}^{(t+1)}).$$

F Additional experiment details and results

We provide more details on the experimental setup as well as further results.

F.1 Additional experimental setup details

This section contains further details on the experimental setup, including the datasets used and hyperparameters tuned. All categorical features in each dataset were binarized into one-hot vectors. All numerical features were bucketized into 4 quantiles, and further binarized into one-hot vectors. All code that we used for pre-processing the datasets from their publicly-downloadable versions can be found at `https://github.com/wenshuoguo/robust-fairness-code`.

For the naïve approach, we solve the constrained optimization problem (2) with respect to the noisy groups $\hat{G}$. For comparison, we also report the results of the unconstrained optimization problem and the constrained optimization problem (1) when the true groups G are known. For the DRO problem (3), we estimate the bound $\gamma_j = P(\hat{G} \neq G | G = j)$ in each case study. For the soft group assignments approach, we implement the *practical* algorithm (Algorithm 4).

In the experiments, we replace all expectations in the objective and constraints with finite-sample empirical versions. So that the constraints will be convex and differentiable, we replace all indicator functions with hinge upper bounds, as in Davenport et al. [16] and Eban et al. [22]. We use a linear model: $\phi(X;\theta) = \theta^T X$. The noisy protected groups $\hat{G}$ are included as a feature in the model, demonstrating that conditional independence between $\hat{G}$ and the model $\phi(X;\theta)$ is not required here, unlike some prior work [4]. Aside from being used to estimate the noise model $P(G = k | \hat{G} = j)$ for the soft group assignments approach[3], the true groups G are never used in the training or validation process.

[3] If $P(G = k | \hat{G} = j)$ is estimated from an auxiliary dataset with a different distribution than test, this could lead to generalization issues for satisfying the true group constraints on test. In our experiments, we lump those generalization issues in with any distributional differences between train and test.

Each dataset was split into train/validation/test sets with proportions 0.6/0.2/0.2. For each algorithm, we chose the *best* iterate $\theta^{(t^*)}$ out of T iterates on the train set, where we define *best* as the iterate that achieves the lowest objective value while satisfying all constraints. We select the hyperparameters that achieve the best performance on the validation set (details in Appendix F). We repeat this procedure for ten random train/validation/test splits and record the mean and standard errors for all metrics[4].

F.1.1 Adult dataset

For the first case study, we used the Adult dataset from UCI [18], which includes 48,842 examples. The features used were *age*, *workclass*, *fnlwgt*, *education*, *education_num*, *marital_status*, *occupation*, *relationship*, *race*, *gender*, *capital_gain*, *capital_loss*, *hours_per_week*, and *native_country*. Detailed descriptions of what these features represent are provided by UCI [18]. The label was whether or not *income_bracket* was above $50,000. The true protected groups were given by the *race* feature, and we combined all examples with race other than "white" or "black" into a group of race "other." When training with the noisy group labels, we did *not* include the true *race* as a feature in the model, but included the noisy race labels as a feature in the model instead. We set $\alpha = 0.05$ as the constraint slack.

The constraint violation that we report in Figure 1 is taken over a test dataset with n examples $(X_1, Y_1, G_1), ..., (X_n, Y_n, G_n)$, and is given by:

$$\max_{j \in \mathcal{G}} \quad \frac{\sum_{i=1}^n \mathbb{1}(\hat{Y}(\theta)_i = 1, Y_i = 1)}{\sum_{i=1}^n \mathbb{1}(Y_i = 1)} - \frac{\sum_{i=1}^n \mathbb{1}(\hat{Y}(\theta)_i = 1, Y_i = 1, G_i = j)}{\sum_{i=1}^n \mathbb{1}(Y_i = 1, G_i = j)} - \alpha,$$

where $\hat{Y}(\theta)_i = \mathbb{1}(\phi(\theta; X_i) > 0)$.

Section C.2 shows how we specifically enforce equality of opportunity using the soft assignments approach, and Section B.3 shows how we enforce equality of opportunity using DRO.

F.1.2 Credit dataset

For the second case study, we used default of credit card clients dataset from UCI [18] collected by a company in Taiwan [53], which contains 30000 examples and 24 features. The features used were *amount_of_the_given_credit*, *gender*, *education*, *education*, *marital_status*, *age*, *history_of_past_payment*, *amount_of_bill_statement*, *amount_of_previous_payment*. Detailed descriptions of what these features represent are provided by UCI [18]. The label was whether or not *default* was true. The true protected groups were given by the *education* feature, and we combined all examples with education level other than "graduate school" or "university" into a group of education level "high school and others". When training with the noisy group labels, we did *not* include the true *education* as a feature in the model, but included the noisy education level labels as a feature in the model instead. We set $\alpha = 0.03$ as the constraint slack.

The constraint violation that we report in Figure 1 is taken over a test dataset with n examples $(X_1, Y_1, G_1), ..., (X_n, Y_n, G_n)$, and is given by:

$$\max_{j \in \mathcal{G}} \quad \max(\Delta_j^{\text{TPR}}, \Delta_j^{\text{FPR}})$$

where

$$\Delta_j^{\text{TPR}} = \frac{\sum_{i=1}^n \mathbb{1}(\hat{Y}(\theta)_i = 1, Y_i = 1)}{\sum_{i=1}^n \mathbb{1}(Y_i = 1)} - \frac{\sum_{i=1}^n \mathbb{1}(\hat{Y}(\theta)_i = 1, Y_i = 1, G_i = j)}{\sum_{i=1}^n \mathbb{1}(Y_i = 1, G_i = j)} - \alpha$$

and

$$\Delta_j^{\text{FPR}} = \frac{\sum_{i=1}^n \mathbb{1}(\hat{Y}(\theta)_i = 1, Y_i = 0, G_i = j)}{\sum_{i=1}^n \mathbb{1}(Y_i = 0, G_i = j)} - \frac{\sum_{i=1}^n \mathbb{1}(\hat{Y}(\theta)_i = 1, Y_i = 0)}{\sum_{i=1}^n \mathbb{1}(Y_i = 0)} - \alpha$$

and $\hat{Y}(\theta)_i = \mathbb{1}(\phi(\theta; X_i) > 0)$.

Section C.2 shows how we specifically enforce equalized odds using the soft assignments approach, and Section B.3 shows how we enforce equalized odds using DRO.

[4] When we report the "maximum" constraint violation, we use the mean and standard error of the constraint violation for the group j with the maximum mean constraint violation.

F.1.3 Optimization code

For all case studies, we performed experiments comparing the naïve approach, the DRO approach (Section 5) and the soft group assignments approach (Section 6). We also compared these to the baselines of optimizing without constraints and optimizing with constraints with respect to the true groups. All optimization code was written in Python and TensorFlow [5]. All gradient steps were implemented using TensorFlow's Adam optimizer [6], though all experiments can also be reproduced using simple gradient descent without momentum. We computed full gradients over all datasets, but minibatching can also be used for very large datasets. Implementations for all approaches are included in the attached code. Training time was less than 10 minutes per model.

Table 1: Hyperparameters tuned for each approach

HPARAM	VALUES TRIED	RELEVANT APPROACHES	DESCRIPTION
η_θ	$\{0.001,0.01,0.1\}$	ALL APPROACHES	LEARNING RATE FOR θ
η_λ	$\{0.25,0.5,1.0,2.0\}$	ALL EXCEPT UNCONSTRAINED	LEARNING RATE FOR λ
$\eta_{\tilde{p}_j}$	$\{0.001, 0.01, 0.1\}$	DRO	LEARNING RATE FOR $\tilde{p}_j$
η_w	$\{0.001, 0.01, 0.1\}$	SOFT ASSIGNMENTS	LEARNING RATE USING GRADIENT METHODS FOR w

F.1.4 Hyperparameters

The hyperparameters for each approach were chosen to achieve the best performance on the validation set on average over 10 random train/validation/test splits, where "best" is defined as the set of hyperparameters that achieved the lowest error rate while satisfying all constraints relevant to the approach. The final hyperparameter values selected for each method were neither the largest nor smallest of all values tried. A list of all hyperparameters tuned and the values tried is given in Table 1.

For the naïve approach, the constraints used when selecting the hyperparameter values on the validation set were the constraints with respect to the noisy group labels given in Equation (2). For the DRO approach and the soft group assignments approach, the respective robust constraints were used when selecting hyperparameter values on the validation set. Specifically, for the DRO approach, the constraints used were those defined in Equation (3), and for the soft group assignments approach, the constraints used were those defined in Equation (7). For the unconstrained baseline, no constraints were taken into account when selecting the best hyperparameter values. For the baseline constrained with access to the true group labels, the true group constraints were used when selecting the best hyperparameter values.

Hinge relaxations of all constraints were used during training to achieve convexity. Since the hinge relaxation is an upper bound on the real constraints, the hinge-relaxed constraints may require some additional slack to maintain feasibility. This positive slack β was added to the original slack α when training with the hinge-relaxed constraints, and the amount of slack β was chosen so that the relevant hinge-relaxed constraints were satisfied on the training set.

All approaches ran for 750 iterations over the full dataset.

F.2 Additional experiment results

This section provides additional experiment results. All results reported here and in the main paper are on the test set (averaged over 10 random train/validation/test splits).

F.2.1 Case study 1 (Adult)

This section provides additional experiment results for case study 1 on the Adult dataset.

Figure 4 that the naïve approach, DRO approach, and soft assignments approaches all satisfied the fairness constraints for the noisy groups on the test set.

[5] Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. tensorflow.org.

[6] `https://www.tensorflow.org/api_docs/python/tf/compat/v1/train/AdamOptimizer`

Figure 5 confirms that the DRO approach and the soft assignments approaches both managed to satisfy their respective robust constraints on the test set on average. For the DRO approach, the constraints measured in Figure 5 come from Equation (3), and for the soft assignments approach, the constraints measured in Figure 5 come from Equation (7). We provide the exact error rate values and maximum violations on the true groups for the Adult dataset in Table 2.

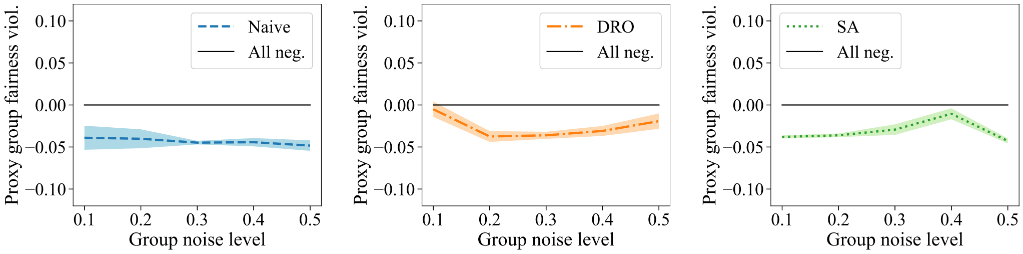

Figure 4: Maximum fairness constraint violations with respect to the noisy groups $\hat{G}$ on the test set for different group noise levels γ on the Adult dataset. For each noise level, we plot the mean and standard error over 10 random train/val/test splits. The black solid line illustrates a maximum constraint violation of 0. While the naïve approach (*left*) has increasingly higher fairness constraints with respect to the true groups as the noise increases, it always manages to satisfy the constraints with respect to the noisy groups $\hat{G}$

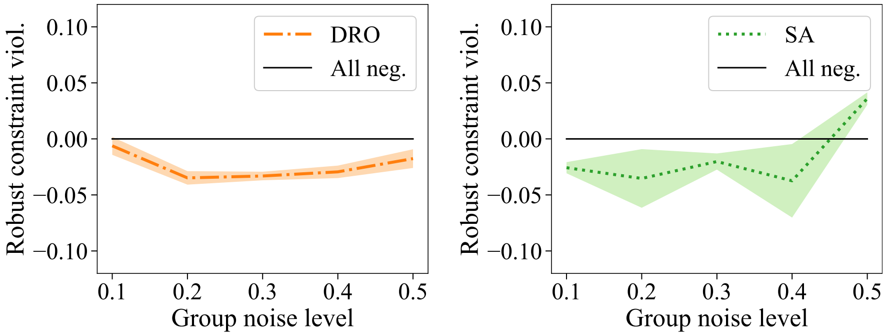

Figure 5: Maximum robust constraint violations on the test set for different group noise levels $P(\hat{G} \neq G)$ on the Adult dataset. For each noise level, we plot the mean and standard error over 10 random train/val/test splits. The black dotted line illustrates a maximum constraint violation of 0. Both the DRO approach (*left*) and the soft group assignments approach (*right*) managed to satisfy their respective robust constraints on the test set on average for all noise levels.

Table 2: Error rate and fairness constraint violations on the true groups for the Adult dataset (mean and standard error over 10 train/test/splits).

	DRO		Soft Assignments	
Noise	Error rate	Max G Viol.	Error rate	Max G Viol.
0.1	0.152 ± 0.001	0.002 ± 0.019	0.148 ± 0.001	-0.048 ± 0.002
0.2	0.200 ± 0.002	-0.045 ± 0.003	0.157 ± 0.003	-0.048 ± 0.002
0.3	0.216 ± 0.010	-0.044 ± 0.004	0.158 ± 0.005	0.002 ± 0.030
0.4	0.209 ± 0.006	-0.019 ± 0.031	0.188 ± 0.003	-0.016 ± 0.016
0.5	0.219 ± 0.012	-0.030 ± 0.032	0.218 ± 0.002	0.004 ± 0.006

F.2.2 Case study 2 (Credit)

This section provides additional experiment results for case study 2 on the Credit dataset.

Figure 6 shows the constraint violations with respect to the true groups on test separated into TPR violations and FPR violations. For all noise levels, there were higher TPR violations than FPR violations. However, this does not mean that the FPR constraint was meaningless – the FPR constraint still ensured that the TPR constraints weren't satisfied by simply adding false positives.

Figure 7 confirms that the naïve approach, DRO approach, and soft assignments approaches all satisfied the fairness constraints for the noisy groups on the test set.

Figure 8 confirms that the DRO approach and the soft assignments approaches both managed to satisfy their respective robust constraints on the test set on average. For the DRO approach, the constraints measured in Figure 8 come from Equation (3), and for the soft assignments approach, the constraints measured in Figure 8 come from Equation (7).

We provide the exact error rate values and maximum violations on the true groups for the Credit dataset in Table 3.

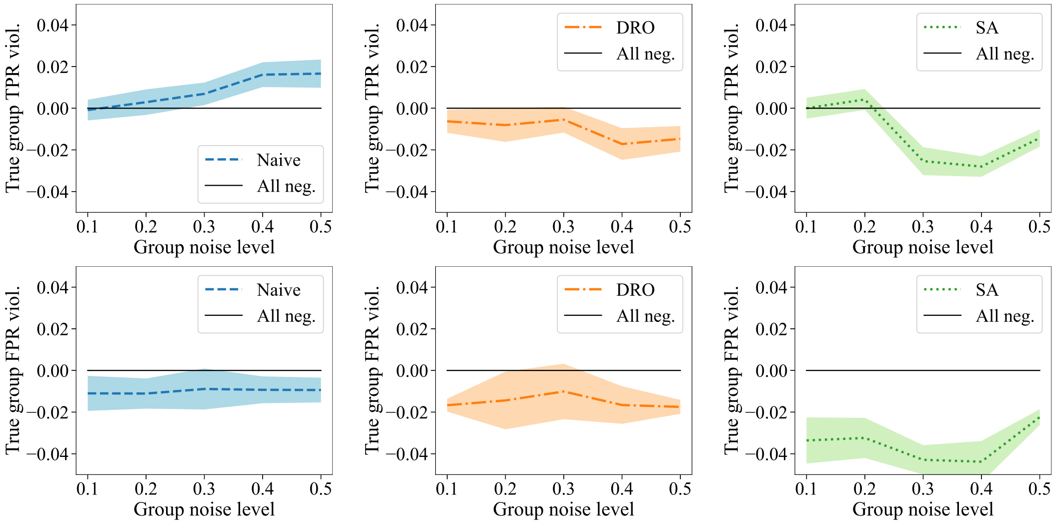

Figure 6: Case study 2 (Credit): Maximum true group TPR (top) and FPR (bottom) constraint violations for the Naive, DRO, and soft assignments (SA) approaches on test set for different group noise levels γ on the Credit dataset (mean and standard error over 10 train/val/test splits). The black solid line represents the performance of the trivial "all negatives" classifier, which has constraint violations of 0. A negative violation indicates satisfaction of the fairness constraints on the true groups.

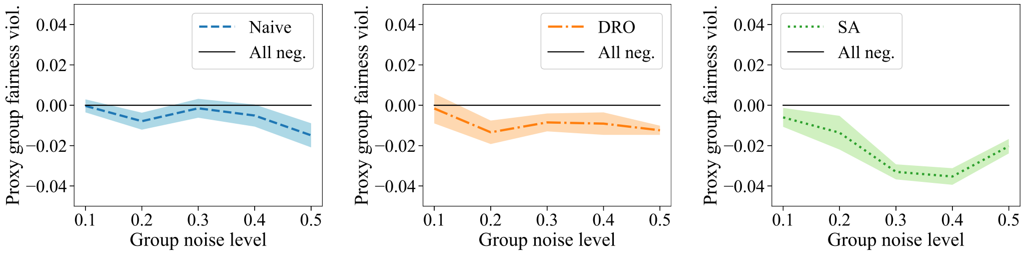

Figure 7: Maximum fairness constraint violations with respect to the noisy groups $\hat{G}$ on the test set for different group noise levels γ on the Credit dataset. For each noise level, we plot the mean and standard error over 10 random train/val/test splits. The black solid line illustrates a maximum constraint violation of 0. While the naïve approach (*left*) has increasingly higher fairness constraints with respect to the true groups as the noise increases, it always manages to satisfy the constraints with respect to the noisy groups $\hat{G}$

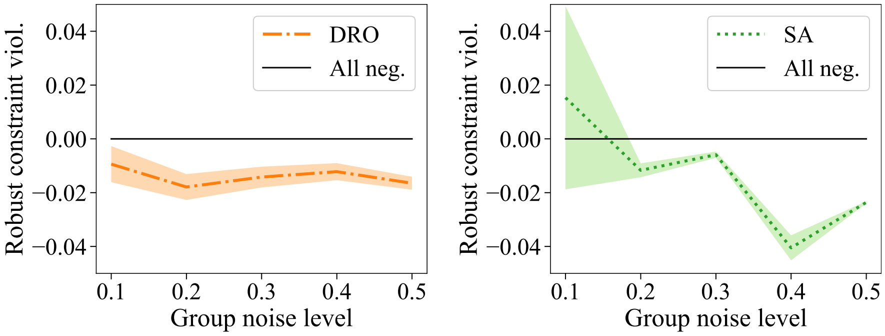

Figure 8: Maximum robust constraint violations on the test set for different group noise levels $P(\hat{G} \neq G)$ on the Credit dataset. For each noise level, we plot the mean and standard error over 10 random train/val/test splits. The black dotted line illustrates a maximum constraint violation of 0. Both the DRO approach (*left*) and the soft group assignments approach (*right*) managed to satisfy their respective robust constraints on the test set on average for all noise levels.

Table 3: Error rate and fairness constraint violations on the true groups for the Credit dataset (mean and standard error over 10 train/test/splits).

	DRO		Soft Assignments	
Noise	Error rate	Max G Viol.	Error rate	Max G Viol.
0.1	0.206 ± 0.003	-0.006 ± 0.006	0.182 ± 0.002	0.000 ± 0.005
0.2	0.209 ± 0.002	-0.008 ± 0.008	0.182 ± 0.001	0.004 ± 0.005
0.3	0.212 ± 0.002	-0.006 ± 0.006	0.198 ± 0.001	-0.025 ± 0.007
0.4	0.210 ± 0.002	-0.017 ± 0.008	0.213 ± 0.001	-0.028 ± 0.005
0.5	0.211 ± 0.003	-0.015 ± 0.006	0.211 ± 0.001	-0.014 ± 0.004