Robust Optimization for Fairness with Noisy Protected Groups

Serena Wang [*1,2] **Wenshuo Guo** [*1] **Harikrishna Narasimhan** [2] **Andrew Cotter** [2] **Maya Gupta** [2]
Michael I. Jordan [1]

Abstract

Many existing fairness criteria for machine learning involve equalizing or achieving some metric across *protected groups* such as race or gender groups. However, practitioners trying to audit or enforce such group-based criteria can easily face the problem of noisy or biased protected group information. We study this important practical problem in two ways. First, we study the consequences of naïvely only relying on noisy protected groups: we provide an upper bound on the fairness violations on the true groups G when the fairness criteria are satisfied on noisy groups $\hat{G}$. Second, we introduce two new approaches using robust optimization that, unlike the naïve approach of only relying on $\hat{G}$, are guaranteed to satisfy fairness criteria on the true protected groups G while minimizing a training objective. We provide theoretical guarantees that one such approach converges to an optimal feasible solution. Using two case studies, we empirically show that the robust approaches achieve better true group fairness guarantees than the naïve approach.

1. Introduction

As machine learning becomes increasingly pervasive in real-world decision making, the question of ensuring *fairness* of ML models becomes increasingly important.

The definition of what it means to be "fair" is highly context dependent. Much work has been done on developing mathematical fairness criteria according to various societal and ethical notions of fairness, as well as methods for building machine-learning models that satisfy those fairness criteria (see, e.g., Dwork et al., 2012; Hardt et al., 2016; Russell et al., 2017; Kusner et al., 2017; Zafar et al., 2017; Cotter et al., 2019b; Friedler et al., 2019).

Many of these mathematical fairness criteria are *group-based*, where a given metric is equalized over subpopulations in the data, also known as *protected groups*. For example, the *equality of opportunity* criterion introduced by Hardt et al. (2016) specifies that the true positive rates for a binary classifier are equalized across protected groups.

One important practical question is whether or not these fairness notions can be reliably measured or enforced if the protected group information is noisy, missing, or unreliable. For example, survey participants may be incentivized to obfuscate their responses for fear of disclosure or discrimination, or may be subject to other forms of response bias. Social desirability response bias may affect participants' answers regarding religion, political affiliation, or sexual orientation (Krumpal, 2011). The collected data may also be outdated: census data collected ten years ago may not an accurate representation for measuring fairness today.

Another source of noise arises from estimating the labels of the protected groups. For various image recognition tasks (e.g., face detection), one may want to measure fairness across protected groups such as gender or race. However, many large image corpora do not include protected group labels, and one might instead use a separately trained classifier to estimate group labels, which is likely to be noisy (Buolamwini & Gebru, 2018). Similarly, zip codes can act as a noisy indicator for socioeconomic groups.

In this paper, we focus on the problem of training binary classifiers with fairness constraints when only noisy labels, $\hat{G} \in \{1, ..., \hat{m}\}$, are available for m true protected groups, $G \in \{1, ..., m\}$, of interest. We study two aspects: First, if one satisfies fairness constraints for noisy protected groups $\hat{G}$, what can one say with respect to those fairness constraints for the true groups G? Second, how can side information about the noise model between $\hat{G}$ and G be leveraged to better enforce fairness with respect to the true groups G?

Contributions: Our contributions are three-fold:

1. We provide a bound on the fairness violations with respect to the true groups G when the fairness criteria are satisfied on the noisy groups $\hat{G}$.

[*]Equal contribution [1]EECS, University of California, Berkeley [2]Google Research. Correspondence to: Wenshuo Guo <wsguo@berkeley.edu>, Serena Wang <serenalwang@berkeley.edu>.

Preliminary work.

2. We introduce two new robust-optimization methodologies that satisfy fairness criteria on the true protected groups G while minimizing a training objective. Each methodology differs in convergence properties, conservatism, and noise model specification.
3. We show empirically that unlike the naïve approach, our two proposed approaches are able to satisfy fairness criteria with respect to the true groups G on average.

The first approach we propose (Section 5) uses Distributionally Robust Optimization (DRO) (Duchi & Namkoong, 2018). Let p_j be the distribution of the data conditioned on the true groups being j, so $X, Y|G = j \sim p_j$, and let $\hat{p}_j$ be the equivalent but conditioned on the noisy groups. Given upper bounds $\gamma_j \geq TV(p_j, \hat{p}_j)$ for each $j \in \{1, ..., m\}$, we seek a model that satisfies the desired fairness criteria for *any* $\tilde{G}$ whose conditional distributions $\tilde{p}_j$ also fall within the bounds γ_i of $\hat{G}$ (a set which must include the unknown true G). Because it is based on the well-studied DRO setting, this approach has the advantage of being easy to analyze. However, tight bounds γ_j can be difficult to find, so the results may be overly conservative.

Our second robust optimization strategy (Section 6), uses a robust re-weighting of the data from soft protected group assignments, inspired by criteria proposed by Kallus et al. (2020) for auditing the fairness of ML models given imperfect group information. Extending their work, we *optimize* a constrained problem with their robust fairness criteria, and provide a *theoretically ideal* algorithm that is guaranteed to converge to an optimal feasible point, as well as an alternative *practical* version that is more computationally tractable. Compared to DRO, this second approach utilizes a more precise noise model, $P(\hat{G} = k|G = j)$, between $\hat{G}$ and G for all pairs of group labels j, k, that can be estimated from a small auxiliary dataset containing ground-truth labels for both G and $\hat{G}$. An advantage of this more detailed noise model is that a practitioner can incorporate knowledge of of any bias in the relationship between G and $\hat{G}$ (for instance, survey respondents favoring one socially preferable response over others), which causes it to be less likely than DRO to result in an overly-conservative model. Most notably, this approach does *not* require that $\hat{G}$ be a direct approximation of G—in fact, G and $\hat{G}$ can represent distinct (but related) groupings, or even groupings of different sizes, with the noise model tying them together. For example, if G represents "language spoken at home," then $\hat{G}$ could be a noisy estimate of "country of residence."

2. Related work

Group-based fair classification: Broadly, mathematical fairness notions proposed in the context of machine learning include notions of group-based fairness (Hardt et al., 2016; Friedler et al., 2019), individual fairness (Dwork et al., 2012), and causality/counterfactual fairness (Russell et al., 2017; Kusner et al., 2017). Group-based fairness notions involve achieving or equalizing some metric over protected groups such as race or gender groups. For example, *demographic parity* (Dwork et al., 2012) requires that a classifier's positive prediction rates are equal for all protected groups. Likewise, *equality of opportunity* (Hardt et al., 2016) requires that the classifier's true positive rates are equal for all protected groups (more listed by Cotter et al. (2019b)). Recent work has extended these group-based metrics to more general ranking and regression settings (Narasimhan et al., 2020).

Constrained optimization for training fair classifiers: The simplest techniques for enforcing group-based constraints involve the post-processing of an existing classifier. For example, one can enforce *equality of opportunity* by choosing different decision thresholds for an existing binary classifier for each protected group (Hardt et al., 2016). However, the classifiers resulting from these post-processing techniques may not necessarily be optimal in terms of accuracy. Thus, constrained optimization techniques have emerged to train machine-learning models that can more optimally satisfy the fairness constraints while minimizing a training objective (Cotter et al., 2019a;b; Zafar et al., 2017; Agarwal et al., 2018; Donini et al., 2018).

Fairness with noisy protected groups: Group-based fairness notions rely on the knowledge of *protected group* labels. However, practitioners may only have access to noisy or unreliable protected group information. One may naïvely try to enforce fairness constraints with respect to these noisy protected groups using the above constrained optimization techniques, but there is no guarantee that the resulting classifier will satisfy the fairness criteria with respect to the true protected groups (Gupta et al., 2018).

Under the conservative assumption that a practitioner has no information about the protected groups, Hashimoto et al. (2018) applied DRO in the context of fairness. Here, we do assume some knowledge of a noise model for the noisy protected groups, resulting in tighter results with DRO. We extend Hashimoto et al. (2018)'s work to allow for equality constraints, which may be reasonably desired in some practical applications (Kolodny, 2019). Further, by assuming knowledge of the noise model, we provide a practically meaningful maximum total variation distance bound to enforce in the DRO procedure, something that Hashimoto et al. (2018) note is difficult to provide under their more general assumptions.

Kallus et al. (2020) considered the problem of *auditing* fairness criteria given noisy groups. They propose a "robust" fairness criteria using soft group assignments and show that

if a given model satisfies those fairness criteria with respect to the noisy groups, then the model will satisfy the fairness criteria with respect to the true groups. Here, we build on that work by providing an algorithm for training a model that satisfies their robust fairness criteria while minimizing a training objective.

Lamy et al. (2019) showed that when there are only two protected groups, one need only tighten the "unfairness tolerance" when enforcing fairness with respect to the noisy groups. When there are more than two groups, and when the noisy groups are included as an input to the classifier, other robust optimization approaches may be necessary. When using post-processing instead of constrained optimization, Awasthi et al. (2019) showed that under certain conditional independence assumptions, post-processing using the noisy groups will not be worse in terms of fairness violations than not post-processing at all. In our work, we consider the problem of training the model subject to fairness constraints, rather than taking a trained model as given and only allowing post-processing, and we do not rely on conditional independence assumptions. Indeed, the model may include the noisy protected attribute as a feature.

Robust optimization: We utilize a minimax technique to control the uncertainty in our optimization problems. The minimax approach formulates this control in terms of a two-player game where the uncertainty is adversarial, and one minimizes a worst-case objective over a feasible set (Ben-Tal et al., 2009; Bertsimas et al., 2011); e.g., the noise is contained in a unit-norm ball around the input data. For example, a recent line of work on DRO assumes that the uncertain distribution of the data is known to belong to a certain class (Namkoong & Duchi, 2016; Duchi & Namkoong, 2018; Li et al., 2019).

3. Optimization problem setup

We start by setting up the training problem for enforcing group-based fairness criteria in a learning setting (Goh et al., 2016; Hardt et al., 2016; Donini et al., 2018; Agarwal et al., 2018; Cotter et al., 2019b).

Let $X \in \mathbb{R}^D$ be a random variable representing a feature vector, and let $Y \in \{0, 1\}$ be a random variable representing a binary target label. Every data point has its protected group label, which is represented by a random variable $G \in \mathcal{G} = \{1, ..., m\}$. Let $\hat{G} \in \hat{\mathcal{G}} = \{1, ..., \hat{m}\}$ be a random variable representing the noisy protected group label, which we assume to have access to during training. For simplicity, assume that $\hat{\mathcal{G}} = \mathcal{G}$ (and $\hat{m} = m$). Let $\phi(X; \theta)$ represent a binary classifier with parameters θ where $\phi(X; \theta) > 0$ indicates a positive classification.

We can write the optimization problem with fairness constraints in the following form:

$$\begin{aligned} \min_{\theta} \quad & f(\theta) \\ \text{s.t.} \quad & |g_j(\theta) - g(\theta)| \leq \alpha, \quad \forall j \in \mathcal{G}, \end{aligned} \tag{1}$$

where $f(\theta) = E[l_0(\theta, X, Y)]$, $g(\theta) = E[l_1(\theta, X, Y)]$, and $g_j(\theta) = E[l_1(\theta, X, Y)|G = j]$ for $j \in \mathcal{G}$.

$l_0(\theta, X, Y)$ represents the training loss function for a binary classifier (e.g., logistic loss, hinge loss), and $l_1(\theta, X, Y)$ represents some function we want to balance over different subgroups. For instance, when equalizing true positive rates for the *equality of opportunity* criterion, $l_1(\theta, X, Y) = \mathbb{1}\left(\phi(X; \theta) > 0, Y = 1\right)/P(Y = 1)$ (see Cotter et al. (2019b) for more examples). Setting $\alpha = 0$ achieves *equality of opportunity* by equalizing the true positive rates (TPR) across the groups, but one may also choose to allow some small slack of $\alpha > 0$.

When X, Y and G are all available without noise, the problem (1) can be solved using constrained optimization techniques (see, e.g., Eban et al., 2017; Agarwal et al., 2018; Cotter et al., 2019b).

4. Bounds for the naïve approach

When only given the noisy groups, one naïve approach to solving problem (1) is to simply redefine the constraints using the noisy groups (Gupta et al., 2018):

$$\begin{aligned} \min_{\theta} \quad & f(\theta) \\ \text{s.t.} \quad & |\hat{g}_j(\theta) - g(\theta)| \leq \alpha \ \ \forall j \in \mathcal{G}, \end{aligned} \tag{2}$$

where $\hat{g}_j(\theta) = E[l_1(\theta, X, Y)|\hat{G} = j], \ \ j \in \mathcal{G}$.

This introduces a practical question: if a model was constrained to satisfy fairness criteria on the noisy groups, how far would that model be from satisfying the constraints on the true groups? In the next section, we show that the fairness violations on the true groups G can at least be bounded when the fairness criteria are satisfied on the noisy groups $\hat{G}$, provided that $\hat{G}$ does not deviate too much from G.

4.1. Bounding fairness constraints using TV distance

Let p_j denote the conditional probability distribution $X, Y|G = j \sim p_j$ using the true groups, and let $\hat{p}_j$ denote the conditional probability distribution $X, Y|\hat{G} = j \sim \hat{p}_j$ using the noisy groups. We use the total variation (TV) distance $TV(p_j, \hat{p}_j)$ to measure distance between the probability distributions p_j and $\hat{p}_j$ (see Appendix A.1 and Villani (2009)).

Given a bound on $TV(p_j, \hat{p}_j)$, we obtain a bound on fairness violations for the true groups when naïvely solving the optimization problem (2) using only the noisy groups:

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

$$|\hat{g}_j(\theta) - g(\theta)| \leq \alpha \ \ \forall j \in \mathcal{G}.$$

Suppose $|l_1(\theta, x_1, y_1) - l_1(\theta, x_2, y_2)| \leq 1$ 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:

$$|g_j(\theta) - g(\theta)| \leq \alpha + \gamma_j \quad \forall j \in \mathcal{G}.$$

Proof in Appendix A.1.

Theorem 1 relies on the fact that $|l_1(\theta, x_1, y_1) - l_1(\theta, x_2, y_2)| \leq 1$. This condition holds for any fairness metrics based on rates such as equality of opportunity, where l_1 is simply some scaled combination of indicator functions. Cotter et al. (2019b) list many such rate-based fairness metrics. However, Theorem 1 can be generalized to functions l_1 that do not satisfy that criterion by looking beyond the TV distance to more general Wasserstein distances between p_j and $\hat{p}_j$. We show this in Appendix A.2, but for all known fairness metrics referenced in this work, formulating Theorem 1 with the TV distance is sufficient.

4.2. Estimating the TV distance bound in practice

Theorem 1 bounds the fairness violations of the naïve approach in terms of the TV distance between the conditional distributions p_j and $\hat{p}_j$, which is not always possible to estimate, since it assumes knowledge of the conditional joint distribution on $X, Y|G = j$. Instead, we can estimate an *upper bound* on $TV(p_j, \hat{p}_j)$ from metrics that are easier to obtain in practice. Specifically, the following lemma shows that under mild assumptions, $P(G \neq \hat{G}|G = j)$ directly translates into an upper bound on $TV(p_j, \hat{p}_j)$.

Lemma 1. Suppose $P(G = j) = P(\hat{G} = j)$ for a given $j \in \mathcal{G}$. Then $TV(p_j, \hat{p}_j) \leq P(G \neq \hat{G}|G = j)$.

Proof in Appendix A.1.

An estimate of $P(G \neq \hat{G}|G = j)$ may come from a variety of sources. As assumed by Kallus et al. (2020), a practitioner may have access to an *auxiliary* dataset containing G and $\hat{G}$, but no other features X or labels Y. Or, practitioners may have some prior estimate of $P(G \neq \hat{G}|G = j)$ based on the way they obtained $\hat{G}$: for example $\hat{G}$ could be estimated by mapping each example's zip code to the most common socioeconomic group for that zip code according to census data. The census data provides a prior for how often $\hat{G}$ produces an incorrect socioeconomic group.

By relating Theorem 1 to realistic noise models, Lemma 1 allows us to bound the fairness violations of the naïve approach using quantities that can be estimated empirically. However, in the next section we show that Lemma 1 can do more than just bound such violations; it can also be used to produce a *robust* approach that will actually guarantee full satisfaction of the fairness violations on the true groups G.

5. Robust Approach 1: Distributionally robust optimization (DRO)

While Theorem 1 provides an upper bound on the performance of the naïve approach, if any noise level γ_j exceeds the desired slack α, it would be infeasible to simply constrain the naïve approach with a tighter slack $\alpha - \gamma_j$. This may often be the case for fairness problems, where the desired α on the true groups can be arbitrarily small (approaching zero ideally). Therefore, it is important to find other ways to do better than the naïve optimization problem (2) in terms of satisfying the constraints on the true groups. In particular, suppose in practice we are able to assert that $P(G \neq \hat{G}|G = j) \leq \gamma_j$ for all groups $j \in \mathcal{G}$. Then Lemma 1 implies a bound on TV distance between the conditional distributions on the true groups and the noisy groups: $TV(p_j, \hat{p}_j) \leq \gamma_j$. Therefore, any feasible solution to the following constrained optimization problem is guaranteed to satisfy the fairness constraints on the true groups:

$$\begin{aligned} \min_{\theta, \tilde{p}_j} \quad & f(\theta) \\ \text{s.t.} \quad & \max_{\substack{\tilde{p}_j : TV(\tilde{p}_j, \hat{p}_j) \leq \gamma_j \\ \tilde{p}_j \ll p}} |\tilde{g}_j(\theta) - g(\theta)| \leq \delta \ \ \forall j \in \mathcal{G}, \end{aligned} \tag{3}$$

where $\tilde{g}_j(\theta) = E_{X,Y \sim \tilde{p}_j}[l_1(\theta, X, Y)]$, and $\tilde{p}_j \ll p$ denotes absolute continuity with respect to the full distribution $X, Y \sim p$. We show that this constrained optimization problem can be solved using techniques from DRO.

5.1. General DRO formulation

In general, given some initial distribution p, a DRO problem takes the following form (Duchi & Namkoong, 2018):

$$\min_{\theta \in \Theta} \max_{q : D(q,p) \leq \gamma} E_{X,Y \sim q}[l(\theta, X, Y)], \tag{4}$$

where D is some divergence metric between the distributions p and q, and $l : \Theta, \mathcal{X}, \mathcal{Y} \to \mathbb{R}$ is some function of random variables $X \in \mathcal{X}, Y \in \mathcal{Y}$ with parameters $\theta \in \Theta$. Much existing work on DRO focuses on how to solve the DRO problem for different divergence metrics D. Namkoong & Duchi (2016) provide methods for efficiently and optimally solving the DRO problem for f-divergences, and other work has provided methods for solving the DRO problem for Wasserstein distances (Li et al., 2019; Esfahani & Kuhn., 2018). Duchi & Namkoong (2018) further provide finite-sample convergence rates for the empirical version of the DRO problem.

An important and often difficult aspect of using DRO is specifying a divergence D and bound γ that are meaningful. In this case, Lemma 1 gives us the key to formulating a DRO problem that is guaranteed to satisfy the fairness criteria with respect to the true groups G.

5.2. Solving the DRO problem

The optimization problem (3) can be written in the form of a DRO problem (4) with total variation distance by using the Lagrangian formulation. Adapting a simplified version of a gradient-based algorithm provided by Namkoong & Duchi (2016), we are able to solve the empirical formulation of problem (4) efficiently. Details of the empirical Lagrangian formulation and pseudocode can be found in Appendix B.

6. Robust Approach 2: Soft group assignments

While any feasible solution to the distributionally robust constrained optimization problem (3) is guaranteed to satisfy the constraints on the true groups G, the bound in Lemma 1 is not tight, so choosing each $\gamma_j = P(G \neq \hat{G}|G = j)$ as an upper bound on $TV(p_j, \hat{p}_j)$ may be rather conservative. Therefore, as an alternative to the DRO constraints in (3), in this section we show how to optimize using the robust fairness criteria proposed by Kallus et al. (2020).

6.1. Soft group assignments

Given a trained binary predictor, $\hat{Y} = \mathbb{1}(\phi(\theta; X) > 0)$, Kallus et al. (2020) proposed a set of robust fairness criteria that can be used to audit the fairness of the given trained model with respect to the true groups $G \in \mathcal{G} = \{1, ..., m\}$ using the noisy groups $\hat{G} \in \hat{\mathcal{G}} = \{1, ..., \hat{m}\}$, where $\mathcal{G} = \hat{\mathcal{G}}$ is not required in general.

They assume access to a *main dataset* with the noisy groups $\hat{G}$, true labels Y, and the features X, as well as access to an *auxiliary dataset* containing both the noisy groups $\hat{G}$ and the true groups G. From the main dataset, one can obtain an estimate of the joint distribution of $(\hat{G}, Y, \hat{Y})$. From the auxiliary dataset, one can obtain an estimate of the joint distribution of $(\hat{G}, G)$ and a noise model $P(G = j|\hat{G} = k)$ for all $j \in \mathcal{G}, k \in \hat{\mathcal{G}}$.

These estimates will be used to associate each example with a vector of weights, where each weight is an estimated probability that the example belongs to the true group j. To motivate this, observe that, assuming $l_1(\theta, X, Y)$ only depends on X through $\hat{Y} = \mathbb{1}(\phi(\theta; X) > 0)$, we have[1]:

$$E[l_1(\theta, X, Y)|G = j] = \frac{E[l_1(\theta, X, Y)P(G = j|\hat{Y}, Y, \hat{G})]}{P(G = j)}.$$

Suppose that we have a function $w : \mathcal{G} \times \{0, 1\} \times \{0, 1\} \times \hat{\mathcal{G}} \to [0, 1]$, where $w(j \mid \hat{y}, y, k)$ estimates $P(G = j|\hat{Y} = \hat{y}, Y = y, \hat{G} = k)$. Rewriting in terms of w, we have:

$$E[l_1(\theta, X, Y)|G = j] \approx \frac{E[l_1(\theta, X, Y)w(j \mid \hat{Y}, Y, \hat{G})]}{P(G = j)}.$$

Unfortunately, even with the ability to estimate $P(\hat{Y} = \hat{y}, Y = y|\hat{G} = k)$ (from the main dataset) and $P(G = j|\hat{G} = k)$ (from the auxilary dataset), we do not have enough information to uniquely determine w. However, given any θ, we can construct the set $\mathcal{W}(\theta)$ of all w that are *consistent* with these quantities, and then require the desired constraints to hold for *all* elements of this set, robustly. Specifically, we define $\mathcal{W}(\theta)$ to be the set of all w satisfying two properties: first, that for all $\hat{y}, y \in \{0, 1\}$ and all $k \in \hat{\mathcal{G}}$, $w(\cdot \mid \hat{y}, y, k)$ forms a distribution over $\mathcal{G}$:

$$\sum_{j=1}^{m} w(j \mid \hat{y}, y, k) = 1, \quad w(j \mid \hat{y}, y, k) \geq 0 \ \ \forall j \in \mathcal{G}; \quad (5)$$

and second, for all $j \in \mathcal{G}$ and $k \in \hat{\mathcal{G}}$, $w(j \mid \cdot, \cdot, k)$ satisfies the law of total probability:

$$P(G = j|\hat{G} = k) = \sum_{\hat{y}, y \in \{0,1\}} w(j \mid \hat{y}, y, k)P(\hat{Y} = \hat{y}, Y = y|\hat{G} = k). \quad (6)$$

Notice that, since $\hat{Y}$ is a function of θ, $\mathcal{W}(\theta)$ depends on θ via this second property.

The robust fairness criteria can now be written in terms of $\mathcal{W}(\theta)$ as:

$$\max_{w \in \mathcal{W}(\theta)} |g_j^w(\theta) - g(\theta)| \leq \delta \ \ \forall j \in \mathcal{G}, \quad (7)$$

where $g_j^w(\theta) = E\big[l_1(\theta, X, Y)w(j \mid \hat{Y}, Y, \hat{G})\big]/P(G = j)$. As is shown by Kallus et al. (2020), any model $\hat{Y}$ that satisfies (7) also satisfies the fairness criteria for the *true* groups in the optimization problem (1).

6.2. Robust optimization with soft group assignments

We extend Kallus et al. (2020)'s work by formulating a robust optimization problem using soft group assignments.

[1] Explicit steps for this derivation can be found in Appendix C.

Combining the robust fairness criteria above with the training objective, we propose the following:

$$\begin{aligned} \min_{\theta\in\Theta} \quad & f(\theta) \\ \text{s.t.} \quad & \max_{w\in\mathcal{W}(\theta)} |g_j^w(\theta) - g(\theta)| \le \delta \;\; \forall j \in \mathcal{G}, \end{aligned} \tag{8}$$

where Θ denotes the space of model parameters. Any feasible solution is guaranteed to satisfy the original fairness criteria with respect to the true groups.

Using a Lagrangian, (8) can be rewritten as:

$$\min_{\theta\in\Theta}\max_{\lambda\in\Lambda} \underbrace{f(\theta) + \sum_{j=1}^{m} \lambda_j \max_{w\in\mathcal{W}(\theta)} f_j(\theta, w)}_{\mathcal{L}(\theta,\lambda)}, \tag{9}$$

where $f_j(\theta, w) = |g_j^w(\theta) - g(\theta)| - \delta$, $\Lambda \subseteq \mathbb{R}_+^m$, and $\mathcal{L}$ is the Lagrangian. When solving this optimization problem, we use the empirical finite-sample versions of each expectation.

As described in Proposition 9 of (Kallus et al., 2020), the inner maximization problem (7) over $w \in \mathcal{W}(\theta)$ can be solved as a linear program for a given fixed θ. However, the Lagrangian problem (9) is not as straightforward to optimize, since the feasible set $\mathcal{W}(\theta)$ depends on θ through $\hat{Y}$. While in general the pointwise maximum of convex functions is convex, the dependence of $\mathcal{W}(\theta)$ on θ means that even if $f_j(\theta, w)$ were convex, $\max_{w\in\mathcal{W}(\theta)} f_j(\theta, w)$ is not necessarily convex.

To solve problem (9), we first introduce a theoretically *ideal* algorithm that we prove converges to an optimal, feasible solution. This *ideal* algorithm relies on a minimization oracle which is not always computationally tractable. Therefore, we also provide a *practical* algorithm using gradient methods that mimics the *ideal* algorithm in structure. The *practical* algorithm is always computationally tractable but does not share the same convergence guarantees.

6.3. Ideal algorithm

The minimax problem in (9) can be interpreted as a zero-sum game between a player who minimizes the Lagrangian $\mathcal{L}$ over θ and another player who maximizes with respect to the Lagrange multipliers λ. In Algorithm 1, we provide an iterative procedure for solving (9), where at each step, the θ-player performs a full optimization, i.e., a *best response* over θ, and the λ-player responds with a gradient ascent update on λ.

Since $\mathcal{L}$ is linear in λ, the λ-updates can be performed efficiently. Indeed, for a fixed θ, the gradient of the Lagrangian $\mathcal{L}$ with respect to λ is given by $\partial\mathcal{L}(\theta,\lambda)/\partial\lambda_j = \max_{w\in\mathcal{W}(\theta)} f_j(\theta, w)$, which is a linear program in w. The challenging part, however, is the best response over θ; that is, finding a solution $\min_\theta \mathcal{L}(\theta, \lambda)$ for a given λ, as this involves a max over constraints $\mathcal{W}(\theta)$ which depend on θ. To implement this best response, we formulate a nested minimax problem that decouples this intricate dependence on θ, by introducing Lagrange multipliers for the constraints in $\mathcal{W}(\theta)$. We then solve this problem with an *oracle* that jointly minimizes over both θ and the newly introduced Lagrange multipliers. We provide the details in Algorithm 4 in Appendix D.

Algorithm 1 *Ideal* Algorithm

Require: learning rate $\eta_\lambda > 0$, estimates of $P(G = j|\hat{G} = k)$ to specify $\mathcal{W}(\theta)$, ρ, ρ'

1: **for** $t = 1, \ldots, T$ **do**
2: *Best response on* θ: run the oracle-based Algorithm 4 in Appendix D to find a distribution $\hat{\theta}^{(t)}$ over Θ s.t.:
3: *Estimate gradient* $\nabla_\lambda \mathcal{L}(\hat{\theta}^{(t)}, \lambda^{(t)})$*:* for each $j \in \mathcal{G}$, choose $\delta_j^{(t)}$ such that:
$\delta_j^{(t)} \le \mathbf{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\max_{w\in\mathcal{W}(\theta)} f_j(\theta, w)\right] \le \delta_j^{(t)} + \rho'$
4: *Ascent step on* λ:
$\tilde{\lambda}_j^{(t+1)} \leftarrow \lambda_j^{(t)} + \eta_\lambda\, \delta_j^{(t)} \;\; \forall j \in \mathcal{G}$
$\lambda^{(t+1)} \leftarrow \Pi_\Lambda(\tilde{\lambda}^{(t+1)})$,
5: **end for**
6: **return** $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \hat{\theta}^{(t)}$

The output of the best-response step is a stochastic classifier with a distribution $\hat{\theta}^{(t)}$ over a finite set of θs. Algorithm 1 then returns the average of these distributions, $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \hat{\theta}^t$, over T iterations. By extending recent results on constrained optimization (Cotter et al., 2019a), we show in Appendix D that the output $\bar{\theta}$ is near-optimal and near-feasible for the robust optimization problem in (8). That is, for a given $\epsilon > 0$, by picking T to be large enough, we have that the objective $\mathbf{E}_{\theta\sim\bar{\theta}}[f(\theta)] \le f(\theta^*) + \epsilon$, for any θ^* that is feasible, and the expected violations in the robust constraints are also no more than ϵ.

6.4. Practical algorithm

Algorithm 1 is guaranteed to converge to a near-optimal, near-feasible solution, but may be computationally intractable and impractical for the following reasons. First, the algorithm needs a nonconvex minimization oracle to compute a best response over θ. Second, there are multiple levels of nesting, making it difficult to scale the algorithm with mini-batch or stochastic updates. Third, the output is a distribution over multiple models, which can be be difficult to use in practice.

Therefore, we supplement Algorithm 1 with a *practical* al-

Algorithm 2 *Practical* Algorithm

Require: learning rates $\eta_\theta > 0$, $\eta_\lambda > 0$, estimates of $P(G = j|\hat{G} = k)$ to specify $\mathcal{W}(\theta)$

1: **for** $t = 1, \ldots, T$ **do**

2: Solve for w given θ using linear programming or a gradient method:
$w^{(t)} \leftarrow \max_{w \in \mathcal{W}(\theta^{(t)})} \sum_{j=1}^{m} \lambda_j^{(t)} f_j(\theta^{(t)}, w)$

3: *Descent step on* θ*:*
$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_\theta \delta_\theta^{(t)}$, where
$\delta_\theta^{(t)} = \nabla_\theta \left(f_0(\theta^{(t)}) + \sum_{j=1}^{m} \lambda_j^{(t)} f_j \left(\theta^{(t)}, w^{(t+1)}\right)\right)$

4: *Ascent step on* λ*:*
$\tilde{\lambda}_j^{(t+1)} \leftarrow \lambda_j^{(t)} + \eta_\lambda f_j \left(\theta^{(t+1)}, w^{(t+1)}\right) \quad \forall j \in \mathcal{G}$
$\lambda^{(t+1)} \leftarrow \Pi_\Lambda(\tilde{\lambda}^{(t+1)})$,

5: **end for**

6: **return** $\theta^{(t^*)}$ where t^* denotes the *best* iterate that satisfies the constraints in (8) with the lowest objective.

gorithm that is similar in structure, but approximates the inner best response routine with two simple steps: a maximization over $w \in \mathcal{W}(\theta^{(t)})$ using a linear program for the current iterate $\theta^{(t)}$, and a gradient step on θ at the maximizer $w^{(t)}$. Algorithm 2 outlines this practical algorithm, and leaves room for other practical modifications such as using stochastic gradients. We discuss the practical algorithm in further detail in Appendix E.

We show empirically in Section 7 that the *practical* algorithm can still achieve a solution that satisfies the original fairness constraints with respect to the true groups G.

7. Experiments

We compare the performance of the naïve approach and the two robust optimization approaches (DRO and soft group assignments) empirically using two case studies with *different* sources of noisy protected group labels.

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 assignments approach, we implement the *practical* algorithm (Algorithm 2).

For the optimization objective, we take l_0 to be the hinge loss. For both case studies, we enforce *equality of opportunity* by equalizing true positive rates (TPRs), i.e., $l_1(\theta, X, Y) = \mathbb{1}(\phi(X;\theta) > 0, Y = 1)/P(Y = 1)$. Specifically, we enforce that the TPR conditioned on each group is greater than or equal to the overall TPR on the full dataset with some slack α, which produces m true group-fairness criteria, $g_j(\theta) - g(\theta) \leq \alpha \ \ \forall j \in \mathcal{G}$. The fairness violations that we measure are $g_j(\theta) - g(\theta) - \alpha$. 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. (2010) and Eban et al. (2017) (details can be found in Appendix F).

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 (Awasthi et al., 2019). Aside from being used to estimate the noise model $P(G = k|\hat{G} = j)$ for the soft group assignments approach[2], the true groups G are never used in the training or validation process.

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[3]. All experiment code will be made available on Github.

7.1. Case study 1 (Adult): different noise levels

In this case study, we stress-test the performance of the different algorithms under different amounts of noise between the true groups G and the noisy groups $\hat{G}$. We use the Adult dataset from UCI (Dua & Graff, 2017), which has 48,842 examples and 14 features (details in Appendix F). The classification task is to determine whether an individual makes over \$50K per year. For the true groups, we use $m = 3$ race groups of "white," "black," and "other."

Generating noisy protected groups: Given the true race groups, we synthetically generate noisy protected groups by selecting a fraction γ of data uniformly at random. For each selected example, we perturb the group membership to a different group also selected uniformly at random

[2] 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.

[3] 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.

from the remaining race groups. This way, for a given γ, $P(\hat{G} \neq G) \approx P(\hat{G} \neq G | G = j) \approx \gamma$ for all groups $j, k \in \mathcal{G}$. We evaluate the performance of the different algorithms ranging from small to large amounts of noise: $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

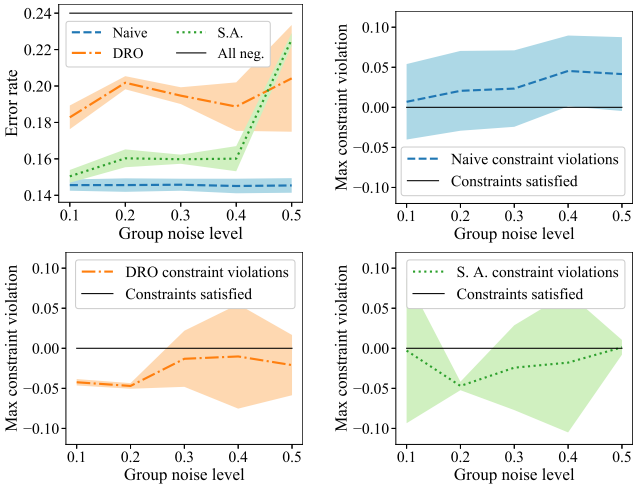

Figure 1. Case study 1: error rate and true group constraint violations on test for different group noise levels γ on the Adult dataest (over 10 train/val/test splits). The black solid line represents a max constraint violation of 0, as well the "all negatives" classifier (*top left*). The robust approaches DRO (*bottom left*) and soft group assignments (*bottom right*) satisfy the constraints on average for all noise levels. As the noise level increases, the naïve approach (*top right*) has increasingly higher true group constraint violations.

Results: The unconstrained model achieves an error rate of 0.1447 ± 0.0038 and a maximum constraint violation of 0.0234 ± 0.0518 on test with respect to the true groups. The model that assumes knowledge of the true groups achieves an error rate of 0.1459 ± 0.0038 and a maximum constraint violation of -0.0469 ± 0.0677 on test with respect to the true groups. As a sanity check, this demonstrates that when given access to the true groups, it is possible to satisfy the constraints on the test set with a reasonably low error rate.

Figure 1 shows that for all noise levels, the robust approaches are able to satisfy all constraints on average. As the noise level increases, the naïve approach no longer controls the fairness violations with respect to the true groups G, even though it does satisfy the constraints with respect to the noisy groups $\hat{G}$ (Figure 2). DRO generally suffers from a higher error rate compared to the soft assignments approach (Figure 1), even though both satisfy the constraints on average with respect to the true groups. This illustrates the conservativeness of the DRO approach and perhaps the looseness of the TV bound.

Both robust approaches come at a cost of a higher error rate than the naïve approach. As expected, the error rate of the naïve approach matches the error rate of the model optimized with constraints on the true groups G, regardless of the noise level γ.

7.2. Case study 2 (Boston): noisy groups from proxies

For this case study we consider a more realistic scenario where the noisy groups are generated from a proxy feature, combined with some prior knowledge. We use the Boston Stop-and-frisk dataset released by the Boston Police Department (BPD) (Analyze Boston, 2015), which includes data from individuals observed, interrogated, searched, or frisked by the BPD. The classification task is to predict whether an individual was either searched or frisked (as opposed to observed or interrogated). We use the $m = 3$ largest race groups of "white," "black," and "Hispanic" as true groups.

Generating noisy protected groups: The Boston dataset includes a "district" feature which can be used to create noisy race groups when combined with racial population percentages per district obtained from census data. Using census data reported from 2015 (Boston Planning & Development Agency, 2017), we assign each example a race group with a probability that matches the race percentages of the district to which the example belongs. This yields noisy race group assignments with noise levels $P(\hat{G} \neq G | G = j)$ of 0.55/0.37/0.74 conditioned on each of the true white/black/Hispanic groups, and an overall noise level $P(\hat{G} \neq G) = 0.54$. We use BPD data from 2014 and 2015 as well to match the census data. This yields a total of 40,666 examples, each with 9 features.

Results: As with the first case study, Table 1 shows that both of the robust approaches (DRO and soft group assignments) manage to satisfy fairness constraints with respect to the true groups G, even when the noise generated from the district proxy is not uniform across the true groups. The naïve approach, on the other hand, exhibited particularly bad constraint violations with respect to the true groups even though it satisfied the constraints with respect to the noisy groups (Appendix F.2), likely due to the protected group noise being particularly high for the Hispanic minority group. The DRO and the soft group assignments approaches still performed better than the degenerate classifier that always predicts negatives, as the label prior is $P(Y = 1) = 0.3213$.

8. Discussion

We explored the practical problem of enforcing group-based fairness for binary classification given noisy protected group information. In addition to providing novel theoretical analysis of the naïve approach of only enforcing fairness on the noisy groups, we also proposed two new

Table 1. Error rate and constraint violations on Boston test set

ALGORITHM	ERROR RATE	MAX FAIRNESS VIOL. ON TRUE G
UNCONST.	0.2780 ± 0.0043	0.0478 ± 0.0376
G KNOWN	0.2849 ± 0.0046	-0.0005 ± 0.0118
NAÏVE	0.2846 ± 0.0047	0.1438 ± 0.0453
DRO	0.3196 ± 0.0056	-0.0362 ± 0.0034
SOFT ASSIGN.	0.3158 ± 0.0034	-0.0191 ± 0.0055

robust approaches that guarantee satisfaction of the fairness criteria on the true groups. For the DRO approach, we gave a theoretical bound on the TV distance to use in the optimization problem using Lemma 1. For the soft group assignments approach, we provided a theoretically ideal algorithm and a practical alternative algorithm for satisfying the robust fairness criteria proposed by Kallus et al. (2020) while minimizing a training objective. We empirically showed that both of these approaches managed to satisfy the constraints with respect to the true groups, even under difficult noise models generated by realistic proxy features.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. M. A reductions approach to fair classification. In *ICML*, 2018.

Analyze Boston. BPD field interrogation and observation (FIO). 2015. URL https://data.boston.gov/dataset/boston-police-department-fio.

Awasthi, P., Kleindessner, M., and Morgenstern, J. Equalized odds postprocessing under imperfect group information. *arXiv preprint arXiv:1906.00285*, 2019.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

Bertsimas, D., Brown, D. B., and Caramanis, C. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

Boston Planning & Development Agency. Neighborhood profiles. 2017. URL http://www.bostonplans.org/getattachment/7987d9b4-193b-4749-8594-e41f1ae27719.

Buolamwini, J. and Gebru, T. Gender shades: Intersectiona accuracy disparities in commercial gender classification. In *Journal of Machine Learning Research*, 2018.

Cotter, A., Jiang, H., and Sridharan, K. Two-player games for efficient non-convex constrained optimization. In *ALT*, 2019a. URL https://arxiv.org/abs/1804.06500.

Cotter, A., Jiang, H., Wang, S., Narayan, T., You, S., Sridharan, K., and Gupta, M. R. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019b.

Davenport, M., Baraniuk, R. G., and Scott, C. D. Tuning support vector machines for minimax and Neyman-Pearson classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. Empirical risk minimization under fairness constraints. *NeurIPS*, 2018.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pp. 272–279, 2008.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proc. 3rd Innovations in Theoretical Computer Science*, pp. 214–226. ACM, 2012.

Eban, E., Schain, M., Mackey, A., Gordon, A., Saurous, R. A., and Elidan, G. Scalable learning of non-decomposable objectives. In *AISTATS*, 2017.

Esfahani, P. M. and Kuhn., D. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115166, 2018.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019.

Goh, G., Cotter, A., Gupta, M. R., and Friedlander, M. P. Satisfying real-world goals with dataset constraints. In *Neurips*, pp. 2415–2423, 2016.

Gupta, M., Cotter, A., Fard, M. M., and Wang, S. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *ICML*, 2018.

Kallus, N., Mao, X., and Zhou, A. Assessing algorithmic fairness with unobserved protected class using data combination. *FAT**, 2020.

Kolodny, N. Why equality of treatment and opportunity might matter. *Philosophical Studies*, 176:33573366, 2019.

Krumpal, I. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality and Quantity*, 47:20252047, 2011.

Kusner, M., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. *NeurIPS*, 2017.

Lamy, A., Zhong, Z., Menon, A. K., and Verma, N. Noise-tolerant fair classification. In *NeurIPS*, 2019.

Li, J., Huang, S., and So, A. M.-C. A first-order algorithmic framework for wasserstein distributionally robust logistic regression. *NeurIPS*, 2019.

Namkoong, H. and Duchi, J. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Neurips*, 2016.

Narasimhan, H., Cotter, A., Gupta, M., and Wang, S. Pairwise fairness for ranking and regression. In *AAAI*, 2020.

Russell, C., Kusner, M. J., Loftus, J. R., and Silva, R. When worlds collide: Integrating different counterfactual assumptions in fairness. In *NeurIPS*, 2017.

Villani, C. *Optimal Transport, Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, 2009.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017.

A. Proofs for Section 4

A.1. Proofs for TV distance

Definition 1. (Total variation 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

$$TV(p, q) = \inf_{\pi} E_{X,Y\sim\pi}[c(X, Y)]$$
$$\text{s.t.} \int \pi(x, y)dy = p(x), \int \pi(x, y)dx = q(y).$$

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

$$|\hat{g}_j(\theta) - g(\theta)| \leq \alpha \ \ \forall j \in \mathcal{G}.$$

Suppose $|l_1(\theta, x_1, y_1) - l_1(\theta, x_2, y_2)| \leq 1$ 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:

$$|g_j(\theta) - g(\theta)| \leq \alpha + \gamma_j \quad \forall j \in \mathcal{G}.$$

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

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

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

$$\begin{aligned}&|\hat{g}_j(\theta) - g_j(\theta)|\\ &= |E_{X,Y\sim\hat{p}_j}[l_1(\theta, X, Y)] - E_{X,Y\sim p_j}[l_1(\theta, X, Y)]|\\ &\leq TV(p_j, \hat{p}_j).\end{aligned}$$

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

Note that if p_j and $\hat{p}_j$ are discrete, then the total variation 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).*[4] *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*

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

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

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

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 total variation distance is given by

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

[4] Edwards, D.A. On the KantorovichRubinstein 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 total variation 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

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

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 loss functions l_1 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 $\hat{G}$:*

$$|\hat{g}_j(\theta) - g(\theta)| \leq \alpha \ \ \forall j \in \mathcal{G}.$$

Suppose the function l_1 satisfies $|l_1(\theta, x_1, y_1) - l_1(\theta, x_2, y_2)| \leq c((x_1, y_1), (x_2, y_2))$ for any $(x_1, y_1) \neq (x_2, y_2)$. 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:

$$|g_j(\theta) - g(\theta)| \leq \alpha + \gamma_j \quad \forall j \in \mathcal{G}.$$

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

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

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

$$\begin{aligned}|\hat{g}_j(\theta) - g_j(\theta)| &= |E_{X,Y\sim\hat{p}_j}[l_1(\theta, X, Y)] - E_{X,Y\sim p_j}[l_1(\theta, X, Y)]| \\ &\leq W_c(p_j, \hat{p}_j).\end{aligned}$$

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

B. Details on DRO formulation for TV distance

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:

$$\begin{aligned}\min_{\theta} \quad & \frac{1}{n}\sum_{i=1}^{n} l_0(\theta, X_i, Y_i) \\ \text{s.t.} \quad & \max_{\tilde{p}_j \in \mathbb{B}_{\gamma_j}(\hat{p}_j)} \left| \sum_{i=1}^{n} \tilde{p}_j^i l_1(\theta, X_i, Y_i) - \frac{1}{n}\sum_{i=1}^{n} l_1(\theta, X_i, Y_i) \right| \leq \alpha \quad \forall j \in \mathcal{G}\end{aligned} \tag{10}$$

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

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_0(\theta, X_i, Y_i)$$

$$f_j(\theta, \tilde{p}_j) = \left| \sum_{i=1}^{n} \tilde{p}_j^i l_1(\theta, X_i, Y_i) - \frac{1}{n}\sum_{i=1}^{n} l_1(\theta, X_i, Y_i) \right| - \alpha.$$

Then the Lagrangian of the empirical formulation (10) is

$$\mathcal{L}(\theta, \lambda) = f_0(\theta) + \sum_{j=1}^{m} \lambda_j \max_{\tilde{p}_j \in \mathbb{B}_{\gamma}(\hat{p}_j)} f_j(\theta, \tilde{p}_j)$$

and problem (10) 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:

$$\begin{aligned}\min_{\theta} \max_{\lambda \geq 0} \max_{\substack{\tilde{p}_j \in \mathbb{R}^n, \tilde{p}_j \geq 0, \\ j=1,...,m}} & f(\theta) + \sum_{j=1}^{m} \lambda_j f_j(\theta, \tilde{p}_j) \\ & \text{s.t. } ||\tilde{p}_j - \hat{p}_j||_1 \leq 2\gamma_j, \quad ||\tilde{p}_j||_1 = 1 \quad \forall j \in \{1, ..., m\}\end{aligned} \tag{11}$$

Since projections onto the ℓ_1-ball can be done efficiently (Duchi et al., 2008), we can solve problem (11) using a projected gradient descent ascent (GDA) algorithm. This is a simplified version of the algorithm introduced by Namkoong & Duchi (2016) for solving general classes of DRO problems. We provide pseudocode in Algorithm 3, as well as an actual implementation in the attached code.

B.2. Projected GDA Algorithm for DRO

Algorithm 3 Project GDA Algorithm

Require: learning rates $\eta_\theta > 0$, $\eta_\lambda > 0$, estimates of $P(G \neq \hat{G}|\hat{G} = j)$ to specify γ_j.

1: **for** $t = 1, \ldots, T$ **do**
2: *Descent step on* θ:
$\theta^{(t+1)} \leftarrow \theta^{(t)} - \nabla_\theta f(\theta^{(t)}) - \sum_{j=1}^m \lambda_j^{(t)} \nabla_\theta f_j(\theta^{(t)}, \tilde{p}_j^{(t)})$
3: *Ascent step on* λ:
$\lambda_j^{(t+1)} \leftarrow \lambda_j^{(t)} + f_j(\theta, \tilde{p}_j^{(t)})$
4: **for** $j = 1, ..., m$ **do**
5: *Ascent step on* $\tilde{p}_j$: $\tilde{p}_j^{(t+1)} \leftarrow \tilde{p}_j^{(t)} + \lambda_j^{(t)} \nabla_{\tilde{p}_j} f_j(\theta^{(t)}, \tilde{p}_j^{(t)})$
6: *Project* $\tilde{p}_j^{(t+1)}$ *onto* ℓ_1*-norm constraints*: $||\tilde{p}_j^{(t+1)} - \hat{p}_j||_1 \leq 2\gamma_j, ||\tilde{p}_j^{(t+1)}||_1 = 1$
7: **end for**
8: **end for**
9: **return** $\theta^{(t^*)}$ where t^* denotes the *best* iterate that satisfies the constraints in (3) with the lowest objective.

C. Further details for soft group assignments approach

Here we explicitly show that $E[l_1(\theta, X, Y)|G = j] = \frac{E[l_1(\theta,X,Y)P(G=j|\hat{Y},Y,\hat{G})]}{P(G=j)}$ using the tower property and the definition of conditional expectation. Using Tower property,

$$\begin{aligned}
&E[l_1(\theta, X, Y)|G = j] \\
&= \frac{E[l_1(\theta, X, Y)\, \mathbb{1}(G = j)]}{P(G = j)} \\
&= \frac{E[E[l_1(\theta, X, Y)\, \mathbb{1}(G = j)|\hat{Y}, Y, \hat{G}]]}{E[E[\mathbb{1}(G = j)|\hat{Y}, Y, \hat{G}]]} \\
&= \frac{E[l_1(\theta, X, Y)E[\mathbb{1}(G = j)|\hat{Y}, Y, \hat{G}]]}{E[E[\mathbb{1}(G = j)|\hat{Y}, Y, \hat{G}]]} \\
&= \frac{E[l_1(\theta, X, Y)P(G = j|\hat{Y}, Y, \hat{G})]}{E[P(G = j|\hat{Y}, Y, \hat{G})]} \\
&= \frac{E[l_1(\theta, X, Y)P(G = j|\hat{Y}, Y, \hat{G})]}{P(G = j)}
\end{aligned} \tag{12}$$

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 4.

Theorem 4 (Optimality and Feasibility for Algorithm 1). *Let* $\theta^* \in \Theta$ *be such that it satisfies the constraints* $\max_{w \in \mathcal{W}(\theta)} f_j(\theta^*, w) \leq 0,\ \forall j \in \mathcal{G}$ *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* $\bar{\theta}$ *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:*

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

and

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

Thus for any given $\epsilon > 0$, 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 $\bar{\lambda} = \frac{1}{T}\sum_{t=1}^{T} \lambda^{(t)}$. We will interpret the minimax problem in (9) 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:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\mathcal{L}(\theta, \lambda^{(t)})\right] &\leq \frac{1}{T}\sum_{t=1}^{T} \min_{\theta\in\Theta} \mathcal{L}(\theta, \lambda^{(t)}) + \epsilon \\
&\leq \min_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(\theta, \lambda^{(t)}) + \rho \\
&= \min_{\theta\in\Theta} \mathcal{L}(\theta, \bar{\lambda}) + \rho \\
&\leq \min_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathcal{L}(\theta, \lambda) + \rho \\
&\leq f(\theta^*) + \rho. \qquad (13)
\end{aligned}$$

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:

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

which further gives us:

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

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

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}_{\theta\sim\hat{\theta}^{(t)}}\left[\mathcal{L}(\theta, \lambda^{(t)})\right] \geq \max_{\lambda\in\Lambda} \left\{\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}_{\theta\sim\hat{\theta}^{(t)}}[\mathcal{L}(\theta, \lambda)] + \|\lambda\|_1 \rho'\right\} - \mathcal{O}\left(\frac{R}{\sqrt{T}}\right). \qquad (14)$$

Optimality. Now, substituting $\lambda = \mathbf{0}$ in (14) and combining with (13) completes the proof of the optimality guarantee:

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

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 (14) and combining with (13) gives us:

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

Algorithm 4 Best response on θ of Algorithm 1

Require: λ', learning rate $\eta_{\mathbf{w}} > 0$, estimates of $P(G = j|\hat{G} = k)$ to specify constraints $h_{g,\hat{g}}$'s, κ

1: **for** $q = 1, \ldots, Q$ **do**

2: *Best response on (θ, $\boldsymbol{\mu}$)*: use an oracle to find find $\theta^{(q)} \in \Theta$ and $\boldsymbol{\mu}^{(q)} \in \mathcal{M}^m$ such that:

$$\ell(\theta^{(q)}, \boldsymbol{\mu}^{(q)}, \mathbf{w}^{(q)}; \lambda') \leq \min_{\theta \in \Theta, \boldsymbol{\mu} \in \mathcal{M}^m} \ell(\theta, \boldsymbol{\mu}, \mathbf{w}^{(q)}; \lambda') + \kappa,$$

for a small slack $\kappa > 0$.

3: *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 ℓ w.r.t. w_j.

4: **end for**

5: **return** A uniform distribution $\hat{\theta}$ over $\theta^{(1)}, \ldots, \theta^{(Q)}$

which can be re-written as:

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

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.

Recall that the constraint set $\mathcal{W}(\theta)$ contains two sets of constraints, the total probability constraints (6) that depend on θ, and the simplex constraints (5) 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':

$$\min_{\theta \in \Theta} \mathcal{L}(\theta, \lambda') = \min_{\theta \in \Theta} f(\theta) + \sum_{j=1}^{m} \lambda'_j \max_{w_j \in \mathcal{W}(\theta)} f_j(\theta, w_j), \tag{15}$$

where we use w_j to denote the maximizer over $\mathcal{W}(\theta)$ for constraint f_j. We separate out the the simplex constraints (5) in $\mathcal{W}(\theta)$ 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 (6) in $\mathcal{W}(\theta)$ into a set of inequality constraints:

$$\begin{aligned} P(G = j|\hat{G} = k) - \sum_{\hat{y}, y \in \{0,1\}} w(j \mid \hat{y}, y, k) P(\hat{Y} = \hat{y}, Y = y|\hat{G} = k) - \tau &\leq 0 \\ \sum_{\hat{y}, y \in \{0,1\}} w(j \mid \hat{y}, y, k) P(\hat{Y} = \hat{y}, Y = y|\hat{G} = k) - P(G = j|\hat{G} = k) - \tau &\leq 0 \end{aligned}$$

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 $h_u(\theta, w) \leq 0$, with index u running from 1 to U. Note that each $h_u(\theta, w)$ is linear in w.

Introducing Lagrange multipliers μ for the relaxed total probability constraints, the optimization problem in (15) can be re-written equivalently as:

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

where note that each w_j is maximized over only the simplex constraints $\mathcal{W}_\Delta$ which are independent of θ, and $\mathcal{M} = \{\mu_j \in \mathbb{R}_+^{m \times \hat{m}} \mid \|\mu_j\|_1 \leq R'\}$, 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:

$$\min_{\theta \in \Theta} \max_{\mathbf{w} \in \mathcal{W}_\Delta^m} \min_{\boldsymbol{\mu} \in \mathcal{M}^m} \underbrace{f(\theta) + \sum_{j=1}^{m} \lambda'_j \Bigg(\underbrace{f_j(\theta, w_j) - \sum_{u=1}^{U} \mu_{j,u} \, h_u(\theta, w_j)}_{\omega(\theta, \mu_j, w_j)} \Bigg)}_{\ell(\theta, \boldsymbol{\mu}, \mathbf{w}; \lambda')}, \tag{16}$$

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 4 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 4.

Theorem 5 (**Optimality Guarantee for Algorithm 4**). *Suppose for every $\theta \in \Theta$, there exists a $\widetilde{w}_j \in \mathcal{W}_\Delta$ such that $h_u(\theta, \widetilde{w}_j) \leq -\gamma, \forall u \in [U]$, for some $\gamma > 0$. Let $0 \leq f_j(\theta, w_j) \leq B', \forall \theta \in \Theta, w_j \in \mathcal{W}_\Delta$. Let $B_{\mathbf{w}} \geq \max_q \|\nabla_{\mathbf{w}} \ell(\theta^{(q)}, \boldsymbol{\mu}^{(q)}, \mathbf{w}^{(q)}; \lambda'))\|_2$. Let $\hat{\theta}$ be the stochastic classifier returned by Algorithm 4 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:*

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

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

Lemma 2 (**Boundedness of Inner Lagrange Multipliers in (16)**). Suppose for every $\theta \in \Theta$, there exists a $\widetilde{w}_j \in \mathcal{W}$ such that $h_u(\theta, \widetilde{w}_j) \leq -\gamma, \forall u \in [U]$, for some $\gamma > 0$. Let $0 \leq f_j(\theta, w_j) \leq B', \forall \theta \in \Theta, w_j \in \mathcal{W}_\Delta$. Let $\mathcal{M} = \{\mu_j \in \mathbb{R}_+^K \mid \|\mu_j\|_1 \leq R'\}$ with the radius of the Lagrange multipliers $R' = B'/\gamma$. Then we have for all $j \in \mathcal{G}$:

$$\max_{w_j \in \mathcal{W}_\Delta} \min_{\mu_j \in \mathcal{M}} \omega\big(\theta, \mu_j, w_j\big) = \max_{w_j \in \mathcal{W}_\Delta : \, h_u(\theta, w_j) \leq 0, \, \forall u} f_j(\theta, w_j).$$

Proof. For a given $j \in \mathcal{G}$, let $w_j^* \in \operatorname*{argmax}_{w_j \in \mathcal{W}_\Delta : \, h_u(\theta, w_j) \leq 0, \, \forall u} f_j(\theta, w_j)$. Then:

$$f_j(\theta, w_j^*) \quad = \quad \max_{w_j \in \mathcal{W}_\Delta} \min_{\mu_j \in \mathbb{R}_+^K} \omega\big(\theta, \mu_j, w_j\big), \tag{17}$$

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:

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

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

$$f_j(\theta, w_j^*) \quad = \quad \max_{w_j \in \mathcal{W}} \omega\big(\theta, \mu_j^*, w_j\big)$$

$$
\begin{aligned}
&= \max_{w_j \in \mathcal{W}} \left\{ f_j(\theta, w_j) \,-\, \sum_{k=1}^{K} \mu^*_{j,k}\, h_k(\theta, w_j) \right\} \\
&\geq f_j(\theta, \widetilde{w}_j) \,-\, \|\mu^*_j\|_1 \max_{k \in [K]} h_k(\theta, \widetilde{w}_j) \\
&\geq f_j(\theta, w_j) \,+\, \|\mu^*_j\|_1 \gamma \;\geq\; \|\mu^*_j\|_1 \gamma.
\end{aligned}
$$

We further have:

$$
\|\mu^*_j\|_1 \;\leq\; f_j(\theta, w_j)/\gamma \;\leq\; B'/\gamma. \tag{18}
$$

Thus the minimizer $\mu^*_j \in \mathcal{M}$. So the minimization in (17) 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 $\bar{w}_j = \frac{1}{Q}\sum_{q=1}^{Q} w_j^{(q)}$. The best response on θ and μ gives us:

$$
\begin{aligned}
&\frac{1}{Q}\sum_{q=1}^{Q}\Big(f(\theta^{(q)}) \,+\, \sum_{j=1}^{m} \lambda'_j\, \omega\big(\theta^{(q)}, \mu_j^{(q)}, w_j^{(q)}\big)\Big) \\
&\quad\leq \frac{1}{Q}\sum_{q=1}^{Q} \min_{\theta\in\Theta,\, \boldsymbol{\mu}\in\mathcal{M}^m} \Big(f(\theta) \,+\, \sum_{j=1}^{m} \lambda'_j\, \omega\big(\theta, \mu_j, w_j^{(q)}\big)\Big) \,+\, \kappa \\
&\quad= \frac{1}{Q}\sum_{q=1}^{Q} \Big(\min_{\theta\in\Theta} f(\theta) \,+\, \sum_{j=1}^{m} \lambda'_j \min_{\mu_j\in\mathcal{M}} \omega\big(\theta, \mu_j, w_j^{(q)}\big)\Big) \,+\, \kappa \quad (j\text{-th summation term depends on } \mu_j \text{ alone}) \\
&\quad\leq \min_{\theta\in\Theta} \frac{1}{Q}\sum_{q=1}^{Q} \Big(f(\theta) \,+\, \sum_{j=1}^{m} \lambda'_j \min_{\mu_j\in\mathcal{M}} \omega\big(\theta, \mu_j, w_j^{(q)}\big)\Big) \,+\, \kappa \\
&\quad\leq \min_{\theta\in\Theta} \Big\{f(\theta) \,+\, \sum_{j=1}^{m} \lambda'_j \min_{\mu_j\in\mathcal{M}} \frac{1}{Q}\sum_{q=1}^{Q} \omega\big(\theta, \mu_j, w_j^{(q)}\big)\Big\} \,+\, \kappa \\
&\quad= \min_{\theta\in\Theta} \Big\{f(\theta) \,+\, \sum_{j=1}^{m} \lambda'_j \min_{\mu_j\in\mathcal{M}} \omega\big(\theta, \mu_j, \bar{w}_j\big)\Big\} \,+\, \kappa \\
&\quad\leq \min_{\theta\in\Theta} \Big\{f(\theta) \,+\, \sum_{j=1}^{m} \lambda'_j \max_{w_j\in\mathcal{W}} \min_{\mu_j\in\mathcal{M}} \omega\big(\theta, \mu_j, w_j\big)\Big\} \,+\, \kappa \quad (\text{by linearity of } \omega \text{ in } w_j) \\
&\quad= \min_{\theta\in\Theta} \Big\{f(\theta) \,+\, \sum_{j=1}^{m} \lambda'_j \max_{w_j:\, h_u(\theta,w_j)\leq 0,\, \forall u} f_j(\theta, w_j)\Big\} \,+\, \kappa \quad (\text{from Lemma 2}) \\
&\quad= \min_{\theta\in\Theta} \mathcal{L}(\theta, \lambda') \,+\, \kappa.
\end{aligned} \tag{19}
$$

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

$$
\begin{aligned}
&\frac{1}{Q}\sum_{q=1}^{Q}\Big(f(\theta^{(q)}) \,+\, \sum_{j=1}^{m} \lambda'_j\, \omega\big(\theta^{(q)}, \mu_j^{(q)}, w_j^{(q)}\big)\Big) \\
&\quad\geq \max_{\mathbf{w}\in\mathcal{W}_\Delta^m} \frac{1}{Q}\sum_{q=1}^{Q} \Big(f(\theta^{(q)}) \,+\, \sum_{j=1}^{m} \lambda'_j\, \omega\big(\theta^{(q)}, \mu_j^{(q)}, w_j\big)\Big) \,-\, \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) \\
&\quad= \frac{1}{Q}\sum_{q=1}^{Q} \Big(f(\theta^{(q)}) \,+\, \sum_{j=1}^{m} \lambda'_j \max_{w_j\in\mathcal{W}_\Delta} \omega\big(\theta^{(q)}, \mu_j^{(q)}, w_j\big)\Big) - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) \quad (j\text{-th summation term depends on } w_j \text{ alone})
\end{aligned}
$$

$$
\begin{aligned}
&\geq \frac{1}{Q}\sum_{q=1}^{Q}\left(f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j \max_{w_j\in\mathcal{W}_\Delta}\min_{\mu_j\in\mathcal{M}}\omega\big(\theta^{(q)},\mu_j,w_j\big)\right) - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) \quad \text{(by linearity of } \omega \text{ in } w_j \text{ and } \mu_j\text{)}\\
&= \mathbf{E}_{\theta\sim\hat{\theta}}\left[f(\theta) + \sum_{j=1}^{m}\lambda'_j \max_{w_j\in\mathcal{W}_\Delta}\min_{\mu_j\in\mathcal{M}}\omega\big(\theta,\mu_j,w_j\big)\right] - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right)\\
&= \mathbf{E}_{\theta\sim\hat{\theta}}\left[f(\theta^{(q)}) + \sum_{j=1}^{m}\lambda'_j \max_{w_j\in\mathcal{W}_\Delta:\, h_u(\theta,w_j)\leq 0,\,\forall u} f_j(\theta,w_j)\right] - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right) \quad \text{(from Lemma 2)}\\
&= \mathbf{E}_{\theta\sim\hat{\theta}}\left[\mathcal{L}(\theta,\lambda')\right] - \mathcal{O}\left(\frac{1}{\sqrt{Q}}\right). \qquad (20)
\end{aligned}
$$

Combining (19) and (20) completes the proof. □

E. Discussion on the *Practical* algorithm

We discuss further about how we arrive at the practical algorithm in Algorithm 2. Recall that in the minimax problem in (9), restated below, each of the m constraints contain a max over w:

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

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

Lemma 3. The minimax problem in (9) is equivalent to:

$$\min_{\theta\in\Theta}\max_{\lambda\in\Lambda}\max_{w\in\mathcal{W}(\theta)} f(\theta) + \sum_{j=1}^{m}\lambda_j f_j(\theta,w). \qquad (21)$$

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 $\beta\in[0,R]$:

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

which completes the proof. □

The practical algorithm outlined in Algorithm 2 seeks to solve the re-written minimax problem in (21), 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:

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

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:

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

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

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 code that we used for pre-processing the datasets from their publicly-downloadable versions is also provided.

F.1.1. Adult dataset

For the first case study, we used the Adult dataset from UCI (Dua & Graff, 2017), 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 (Dua & Graff, 2017). 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.

F.1.2. Boston stop-and-frisk (BPD) dataset

For the second case study, we used the Boston Stop-and-frisk dataset released by the Boston Police Department (BPD) (Analyze Boston, 2015). We used BPD data from 2014 and 2015 to match the available census data (Boston Planning & Development Agency, 2017). This yielded a total of 40,666 examples, each with 9 features. The features used were *sex*, *fio_date*, *priors*, *complexion*, *fiofs_reasons*, *age_at_fio_corrected*, *description*, *dist*, and *officer_id*. Detailed descriptions of what these features represent are provided by the BPD (Analyze Boston, 2015). The label was determined by the *fiofs_type* attribute, which provides whether an individual was frisked, searched, observed, or interrogated. A label value of 1 corresponds with the individual being frisked or searched, and a label value of 0 corresponds with the individual being observed or interrogated. The true protected groups were given by the *description* feature, which contained race labels. We filtered out all examples that did not contain either "B(Black)," "W(White)," or "H(Hispanic)" in the *description* feature. As with the Adult dataset, when training with the noisy group labels generated from the census data, we did not include the true *description* as a feature in the model, but included the noisy race label as a feature in the model instead.

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.

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 2.

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 (8). 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.

Table 2. 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.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 2 confirms that the naïve approach satisfied the fairness constraints for the noisy groups on the test set. The soft assignments approach exhibited higher variance for constraint violations on the noisy groups on the test set. This is likely due to the fact that the hyperparameters for the soft assignments approach were selected to satisfy the constraints in Equation (8) on the validation set, and not the fairness constraints with respect to the noisy groups. Interestingly, the DRO approach managed to satisfy the fairness constraints with respect to the noisy groups on the test set on average. This is

[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

probably a reflection of the fact that the robust constraints for DRO are more conservative than those of the soft assignments approach.

Figure 3 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 3 come from Equation (3), and for the soft assignments approach, the constraints measured in Figure 3 come from Equation (8).

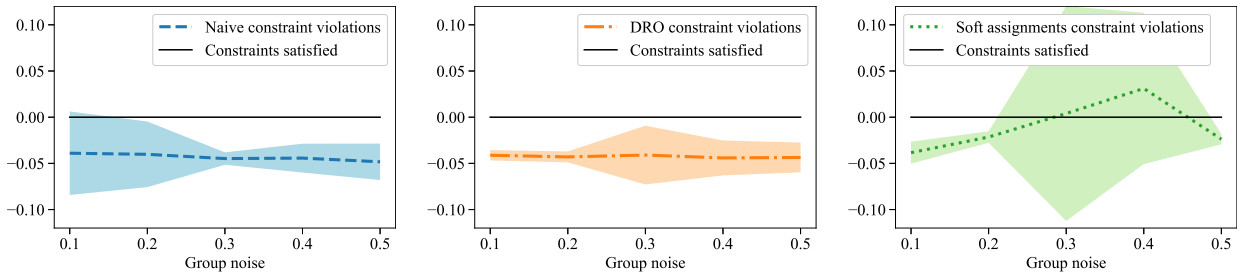

Figure 2. Maximum fairness constraint violations with respect to the noisy groups $\hat{G}$ on the test set for different group noise levels γ on the Adult dataset. For each noise level, we plot the mean and standard error over 10 random train/val/test splits. The black solid line illustrates a maximum 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 $\hat{G}$

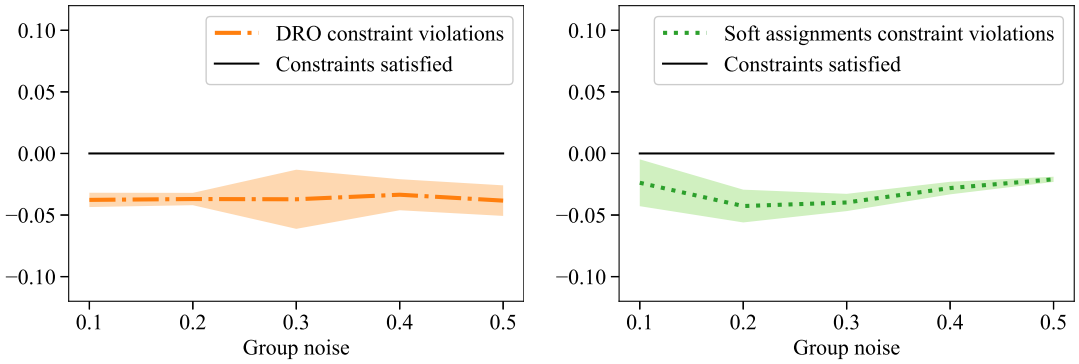

Figure 3. Maximum robust constraint violations on the test set for different group noise levels $P(\hat{G} \neq G)$ on the Adult dataset. For each noise level, we plot the mean and standard error over 10 random train/val/test splits. The black dotted line illustrates a maximum constraint violation of 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 3. Fairness constraint violations on the noisy $\hat{G}$ for the Boston dataset

ALGORITHM	MAX FAIRNESS VIOL. ON NOISY $\hat{G}$
NAÏVE	-0.0179 $\pm$ 0.0149
DRO	-0.0382 $\pm$ 0.0080
SOFT ASSIGNMENTS	-0.0198 $\pm$ 0.0016

Table 4. Robust constraint violations for the Boston dataset

ALGORITHM	ROBUST CONSTRAINT VIOLATIONS
DRO	-0.0178 $\pm$ 0.0052
SOFT ASSIGNMENTS	-0.0215 $\pm$ 0.0006

F.2.2. CASE STUDY 2 (BOSTON STOP-AND-FRISK)

This section provides additional experiment results on the Boston stop-and-frisk dataset, including fairness constraint violations with respect to the noisy groups and robust constraint violations. All results are reported on the test set.

Table 3 confirms that the naïve approach satisfied the fairness constraints on the noisy groups on the test set. Both DRO and the soft assignments approaches managed to satisfy the fairness constraints on the noisy groups as well. Table 4 confirms that both the DRO and soft assignments approaches managed to satisfy their respective robust constraints on the test set.