Robust Optimization for Fairness with Noisy Protected Groups

Serena Wang *
UC Berkeley
Google Research
serenalwang@berkeley.edu

Wenshuo Guo *
UC Berkeley
wsguo@berkeley.edu

Harikrishna Narasimhan
Google Research
hnarasimhan@google.com

Andrew Cotter
Google Research
acotter@google.com

Maya Gupta
Google Research
mayagupta@google.com

Michael I. Jordan
UC Berkeley
jordan@berkeley.edu

Abstract

Many existing fairness criteria for machine learning involve equalizing some metric across *protected groups* such as race or gender. However, practitioners trying to audit or enforce such group-based criteria can easily face the problem of noisy or biased protected group information. First, we study the consequences of naïvely relying on noisy protected group labels: 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 show empirically 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., 21, 32, 49, 41, 54, 14, 25, 51].

Many of these mathematical fairness criteria are *group-based*, where a target metric is equalized or enforced over subpopulations in the data, also known as *protected groups*. For example, the *equality of opportunity* criterion introduced by Hardt et al. [32] specifies that the true positive rates for a binary classifier are equalized across protected groups. The *demographic parity* [21] criterion requires that a classifier's positive prediction rates are equal for all 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

*First two authors have equal contribution

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

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 [40]. 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 [12]. 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 for the noisy groups $\hat{G}$.
2. We introduce two new robust-optimization methodologies that satisfy fairness criteria on the true protected groups G while minimizing a training objective. These methodologies differ 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) is based on distributionally robust optimization (DRO) [19, 8]. Let p denotes the full distribution of the data $X, Y \sim p$. 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 $\hat{p}_j$ be the distribution of X, Y conditioned on the noisy groups. Given an upper bound on the total variation (TV) distance $\gamma_j \geq TV(p_j, \hat{p}_j)$ for each $j \in \{1, ..., m\}$, we define $\tilde{p}_j$ such that the conditional distributions $(X, Y|\tilde{G} = j \sim \tilde{p}_j)$ fall within the bounds γ_i with respect to $\hat{G}$. Therefore, the set of all such $\tilde{p}_j$ is guaranteed to include the unknown true group distribution $p_j, \forall j \in \mathcal{G}$. Because it is based on the well-studied DRO setting, this approach has the advantage of being easy to analyze. However, the results may be overly conservative unless tight bounds $\{\gamma_j\}_{j=1}^m$ can be given.

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. [37] for auditing the fairness of ML models given imperfect group information. Extending their work, we *optimize* a constrained problem to achieve 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 uses 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 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. 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

Constrained optimization for group-based fairness metrics: The simplest techniques for enforcing group-based constraints apply a post-hoc correction of an existing classifier [32, 52]. For example, one can enforce *equality of opportunity* by choosing different decision thresholds for an existing

binary classifier for each protected group [32]. 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 [27, 13, 14, 54, 2, 17].

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 [30].

Under the conservative assumption that a practitioner has no information about the protected groups, Hashimoto et al. [33] applied DRO in the context of fairness. In contrast, here we assume some knowledge of a noise model for the noisy protected groups, and are thus able to provide tighter results with DRO: we provide a practically meaningful maximum total variation distance bound to enforce in the DRO procedure. We further extend Hashimoto et al. [33]'s work by applying DRO to problems equalizing fairness metrics over groups, which may be desired in some practical applications [39].

Kallus et al. [37] 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. [42] 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. [4] 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 use a minimax set-up of a two-player game where the uncertainty is adversarial, and one minimizes a worst-case objective over a feasible set [7, 11]; e.g., the noise is contained in a unit-norm ball around the input data. As one such approach, we apply a recent line of work on DRO which assumes that the uncertain distributions of the data are constrained to belong to a certain set [46, 19, 44].

3 Optimization problem setup

We begin with the training problem for incorporating group-based fairness criteria in a learning setting [27, 32, 17, 2, 14]. Let $X \in \mathcal{X} \subseteq \mathbb{R}^D$ be a random variable representing a feature vector, with a random binary label $Y \in \mathcal{Y} = \{0, 1\}$ and random protected group membership $G \in \mathcal{G} = \{1, ..., m\}$. In addition, let $\hat{G} \in \hat{\mathcal{G}} = \{1, ..., \hat{m}\}$ be a random variable representing the noisy protected group label for each (X, Y), which we assume we 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 $\theta \in \Theta$ where $\phi(X; \theta) > 0$ indicates a positive classification.

Then, training with fairness constraints [27, 32, 17, 2, 14] is:

$$\min_{\theta} \quad f(\theta) \quad \text{s.t.} \quad g_j(\theta) \leq 0, \forall j \in \mathcal{G}, \tag{1}$$

The objective function $f(\theta) = \mathbb{E}[l(\theta, X, Y)]$, where $l(\theta, X, Y)$ is any standard binary classifier training loss. The constraint functions $g_j(\theta) = \mathbb{E}[h(\theta, X, Y)|G = j]$ for $j \in \mathcal{G}$, where $h(\theta, X, Y)$ is the target fairness metric, e.g. $h(\theta, X, Y) = \mathbb{1}\left(\phi(X; \theta) > 0\right) - \mathbb{E}[\mathbb{1}\left(\phi(X; \theta) > 0\right)]$ when equalizing positive rates for the *demographic parity* [21] criterion (see [14] for more examples). Algorithms have been studied for problem (1) when the true protected group labels G are given [see, e.g., 22, 2, 14].

4 Bounds for the naïve approach

When only given the noisy groups $\hat{G}$, one naïve approach to solving problem (1) is to simply re-define the constraints using the noisy groups [30]:

$$\min_{\theta} \quad f(\theta) \quad \text{s.t.} \quad \hat{g}_j(\theta) \leq 0, \ \ \forall j \in \mathcal{G}, \tag{2}$$

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

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? 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

Recall that $X, Y|G = j \sim p_j$ and $X, Y|\hat{G} = j \sim \hat{p}_j$. We use the TV distance $TV(p_j, \hat{p}_j)$ to measure the distance between the probability distributions p_j and $\hat{p}_j$ (see Appendix A.1 and Villani [50]). 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. *(proof in Appendix A.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: $g_j(\theta) \leq \gamma_j, \ \ \forall j \in \mathcal{G}$.*

Theorem 1 is tight for the family of functions h that satisfy $|h(\theta, x_1, y_1) - h(\theta, x_2, y_2)| \leq 1$ for any $(x_1, y_1) \neq (x_2, y_2)$. This condition holds for any fairness metrics based on rates such as demographic parity, where h is simply some scaled combination of indicator functions. Cotter et al. [14] list many such rate-based fairness metrics. Theorem 1 can be generalized to functions h whose differences are not bounded by 1 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 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 assumes knowledge of p_j and is not always possible to estimate. 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 shows that if the prior on class j is unaffected by the noise, $P(G \neq \hat{G}|G = j)$ directly translates into an upper bound on $TV(p_j, \hat{p}_j)$.

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

In practice, an estimate of $P(G \neq \hat{G}|G = j)$ may come from a variety of sources. As assumed by Kallus et al. [37], a practitioner may have access to an *auxiliary* dataset containing G and $\hat{G}$, but not X or Y. Or, practitioners may have some prior estimate of $P(G \neq \hat{G}|G = j)$: if $\hat{G}$ is estimated by mapping zip codes to the most common socioeconomic group for that zip code, then 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. In the next section we show that Lemma 1 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, it fails to provide a guarantee that the constraints on the true groups are satisfied, i.e. $g_j(\theta) \leq 0$. Thus, 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:

$$\min_{\theta \in \Theta} \quad f(\theta) \quad \text{s.t.} \quad \max_{\substack{\tilde{p}_j : TV(\tilde{p}_j, \hat{p}_j) \leq \gamma_j \\ \tilde{p}_j \ll p}} \tilde{g}_j(\theta) \leq 0, \quad \forall j \in \mathcal{G}, \tag{3}$$

where $\tilde{g}_j(\theta) = \mathbb{E}_{X,Y \sim \tilde{p}_j}[h(\theta, X, Y)]$, and $\tilde{p}_j \ll p$ denotes absolute continuity.

5.1 General DRO formulation

A DRO problem is a minimax optimization [19]:

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

where D is some divergence metric between the distributions p and q, and $l : \Theta \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Much existing work on DRO focuses on how to solve the DRO problem for different divergence metrics D. Namkoong and Duchi [46] 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 [44, 23]. Duchi and Namkoong [19] further provide finite-sample convergence rates for the empirical version of the DRO problem.

5.2 Solving 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.

The optimization problem (3) can be written in the form of a DRO problem (4) with TV distance by using the Lagrangian formulation. Adapting a simplified version of a gradient-based algorithm provided by Namkoong and Duchi [46], we are able to solve the empirical formulation of problem (4) efficiently. Details of our empirical Lagrangian formulation and pseudocode are 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, 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. [37].

6.1 Constraints with soft group assignments

Given a trained binary predictor, $\hat{Y}(\theta) = \mathbb{1}(\phi(\theta; X) > 0)$, Kallus et al. [37] 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 $G \in \mathcal{G}$ using the noisy groups $\hat{G}$, 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 estimates of the joint distributions $(\hat{Y}(\theta), Y, \hat{G})$; from the auxiliary dataset, one can obtain estimates of the joint distributions $(\hat{G}, G)$ and a noise model $P(G = j|\hat{G} = k)$ for all $j \in \mathcal{G}, k \in \hat{\mathcal{G}}$.

These estimates are 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. Specifically, 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

$P(G = j|\hat{Y}(\theta) = \hat{y}, Y = y, \hat{G} = k)$. We rewrite the fairness constraint $E[h(\theta, X, Y)|G = j] = \frac{E[h(\theta,X,Y)P(G=j|\hat{Y}(\theta),Y,\hat{G})]}{P(G=j)}$ (derivation in Appendix C.1), and estimate this using w. We also show how h can be adapted to the *equality of opportunity* setting in Appendix C.2.

Given the main dataset and auxiliary dataset, we limit the possible values of the function $w(j \mid \hat{y}, y, k)$ using the law of total probability (as in [37]). The set of possible functions w is given by:

$$\mathcal{W}(\theta) = \left\{ w : \begin{array}{c} \sum_{\hat{y},y\in\{0,1\}} w(j|\hat{y},y,k)P(\hat{Y}(\theta)=\hat{y},Y=y|\hat{G}=k)=P(G=j|\hat{G}=k), \\ \sum_{j=1}^{m} w(j|\hat{y},y,k)=1, w(j|\hat{y},y,k)\geq 0 \quad \forall \hat{y},y\in\{0,1\},j\in\mathcal{G},k\in\hat{\mathcal{G}} \end{array} \right\}. \tag{5}$$

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

$$\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w) \leq 0, \quad \forall j \in \mathcal{G} \quad \text{where} \quad g_j(\theta, w) = \frac{\mathbb{E}[h(\theta, X, Y)w(j|\hat{Y}(\theta), Y, \hat{G})]}{P(G = j)}. \tag{6}$$

6.2 Robust optimization with soft group assignments

We extend Kallus et al. [37]'s work by formulating a robust optimization problem using soft group assignments. Combining the robust fairness criteria above with the training objective, we propose:

$$\min_{\theta\in\Theta} \quad f(\theta) \quad \text{s.t.} \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w) \leq 0, \quad \forall j \in \mathcal{G}, \tag{7}$$

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, problem (7) can be rewritten as:

$$\min_{\theta\in\Theta} \max_{\lambda\in\Lambda} \mathcal{L}(\theta, \lambda) \tag{8}$$

where the Lagrangian $\mathcal{L}(\theta, \lambda) = f(\theta) + \sum_{j=1}^{m} \lambda_j \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)$, and $\Lambda \subseteq \mathbb{R}_+^m$.

When solving this optimization problem, we use the empirical finite-sample versions of each expectation. As described in Proposition 9 of Kallus et al. [37], the inner maximization (6) over $w \in \mathcal{W}(\theta)$ can be solved as a linear program for a given fixed θ. However, the Lagrangian problem (8) 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 $g_j(\theta, w)$ were convex, $\max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)$ is not necessarily convex. 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 further provide a practical algorithm using gradient methods that mimics the ideal algorithm in structure and computationally tractable, but does not share the same convergence guarantees.

6.3 Ideal algorithm

The minimax problem in (8) can be interpreted as a zero-sum game between the θ-player and λ-player. In Algorithm 1, we provide an iterative procedure for solving (8), 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.

For a fixed θ, the gradient of the Lagrangian $\mathcal{L}$ with respect to λ is given by $\partial\mathcal{L}(\theta, \lambda)/\partial\lambda_j = \max_{w\in\mathcal{W}(\theta)} g_j(\theta, w)$, 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 3 in Appendix D.

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, $\overline{\theta} = \frac{1}{T}\sum_{t=1}^{T} \hat{\theta}^t$, over T iterations. By extending recent results on constrained optimization [13], we show in Appendix D that the output $\overline{\theta}$ is near-optimal and near-feasible for the robust optimization problem in (7). That is, for a given $\varepsilon > 0$, by picking T to be large enough, we have that the objective $\mathbb{E}_{\theta\sim\overline{\theta}}[f(\theta)] \leq f(\theta^*) + \varepsilon$, for any θ^* that is feasible, and the expected violations in the robust constraints are also no more than ε.

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 3 to find a distribution $\hat{\theta}^{(t)}$ over Θ s.t. $\mathbb{E}_{\theta \sim \hat{\theta}^{(t)}} \left[\mathcal{L}(\theta, \lambda^{(t)})\right] \leq \min_{\theta \in \Theta} \mathcal{L}(\theta, \lambda^{(t)}) + \rho$.
3: *Estimate gradient* $\nabla_\lambda \mathcal{L}(\hat{\theta}^{(t)}, \lambda^{(t)})$: for each $j \in \mathcal{G}$, choose $\delta_j^{(t)}$ s.t. $\delta_j^{(t)} \leq \mathbb{E}_{\theta \sim \hat{\theta}^{(t)}} \left[\max_{w \in \mathcal{W}(\theta)} g_j(\theta, w)\right] \leq \delta_j^{(t)} + \rho'$
4: *Ascent step on* λ: $\tilde{\lambda}_j^{(t+1)} \leftarrow \lambda_j^{(t)} + \eta_\lambda \, \delta_j^{(t)}, \;\; \forall j \in \mathcal{G}; \quad \lambda^{(t+1)} \leftarrow \Pi_\Lambda(\tilde{\lambda}^{(t+1)})$
5: **end for**
6: **return** $\overline{\theta} = \frac{1}{T} \sum_{t=1}^{T} \hat{\theta}^{(t)}$

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 [47].

Therefore, we supplement Algorithm 1 with a practical algorithm, Algorithm 4 (see Appendix E) 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 4 leaves room for other practical modifications such as using stochastic gradients. We provide further discussion in Appendix E.

7 Experiments

We compare the performance of the naïve approach and the two robust optimization approaches (DRO and soft group assignments) empirically using two datasets from UCI [18] with different constraints. For both datasets, 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 take l to be the hinge loss. The specific constraint violations measured and additional training details can be found in Appendix F.1. All experiment code is available on GitHub at `https://github.com/wenshuoguo/robust-fairness-code`.

Generating noisy protected groups: Given the true protected 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 from the remaining 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: $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

7.1 Case study 1 (Adult): equality of opportunity

We use the Adult dataset from UCI [18] collected from 1994 US Census, 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." As done by [14, 25, 55], we enforce *equality of opportunity* by equalizing true positive rates (TPRs). 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, $\{g_j^{\text{TPR}}(\theta) \leq 0\} \;\; \forall j \in \mathcal{G}$ (details about the constraint function h in Appendix B.3 and C.2).

7.2 Case study 2 (Credit): equalized odds

We consider another application of group-based fairness constraints to credit default prediction. Fourcade and Healy [24] provide an in depth study of the effect of credit scoring techniques on

the credit market, showing that this scoring system can perpetuate inequity. Enforcing group-based fairness with credit default predictions has been considered in a variety of prior works [32, 10, 51, 3, 9, 28, 25, 6]. Following Hardt et al. [32] and Grari et al. [28], we enforce *equalized odds* [32] by equalizing both true positive rates (TPRs) and false positive rates (FPRs) across groups.

We use the "default of credit card clients" dataset from UCI [18] collected by a company in Taiwan [53], which contains 30,000 examples and 24 features (details in Appendix F). The classification task is to determine whether an individual defaulted on a loan. We use $m = 3$ groups based on education levels: "graduate school," "university," and "high school/other" (the use of education in credit lending has previously been studied in the algorithmic fairness and economics literature [26, 9, 43]). We constrain the TPR conditioned on each group to be greater than or equal to the overall TPR on the full dataset with a slack α, and the FPR conditioned on each group to be less than or equal to the overall FPR on the full dataset. This produces $2m$ true group-fairness criteria, $\{g_j^{\text{TPR}}(\theta) \leq 0, g_j^{\text{FPR}}(\theta) \leq 0\} \quad \forall j \in \mathcal{G}$ (details about the constraint functions h can be found in Appendix B.3 and C.2).

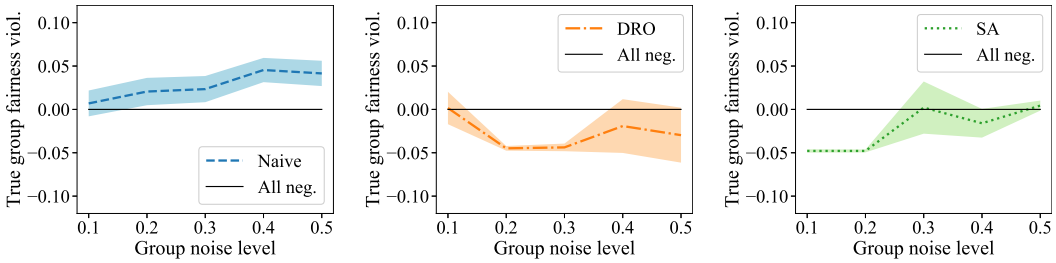

Figure 1: Case study 1 (Adult): maximum true group constraint violations on test set for the Naive, DRO, and soft assignments (SA) approaches for different group noise levels γ on the Adult dataset (mean and standard error over 10 train/val/test splits). The black solid line represents the performance 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.

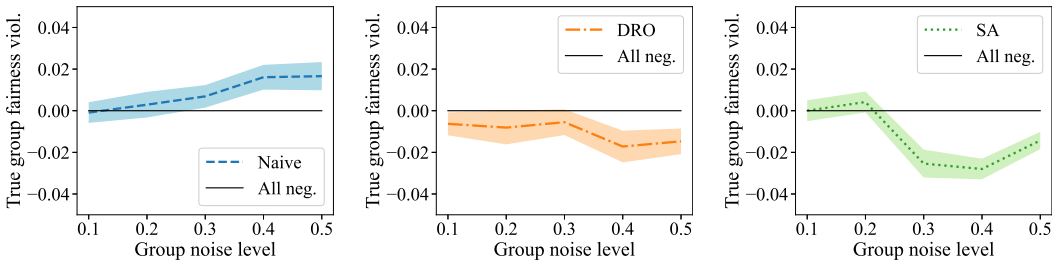

Figure 2: Case study 2 (Credit): maximum true group constraint violations on test set for the Naive, DRO, and soft assignments (SA) approaches for different group noise levels γ on the Credit dataset (mean and standard error over 10 train/val/test splits). This figure shows the max constraint violation over all TPR and FPR constraints, and Figure 6 in Appendix F.2 shows the breakdown of these constraint violations into the max TPR and the max FPR constraint violations.

7.3 Results

In case study 1 (Adult), the unconstrained model achieves an error rate of 0.1447 ± 0.0012 (mean and standard error over 10 splits) and a maximum constraint violation of 0.0234 ± 0.0164 on test set with respect to the true groups. The model that assumes knowledge of the true groups achieves an error rate of 0.1459 ± 0.0012 and a maximum constraint violation of -0.0469 ± 0.0068 on test set 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.

In case study 2 (Credit), the unconstrained model achieves an error rate of 0.1797 ± 0.0013 (mean and standard error over 10 splits) and a maximum constraint violation of 0.0264 ± 0.0071 on

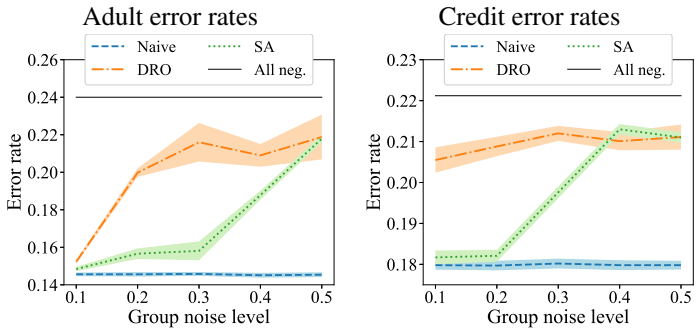

Figure 3: Error rates on test set for different group noise levels γ on the Adult dataset (*left*) and the Credit dataset (*right*) (mean and standard error over 10 train/val/test splits). The black solid line represents the performance of the trivial "all negatives" classifier, which has error rate 0.239. The soft assignments (SA) approach achieves lower error rates than DRO, and as the noise level increases, the gap in error rate between the naive approach and each robust approach increases.

the test set with respect to the true groups. The constrained model that assumes knowledge of the true groups achieves an error rate of 0.1796 ± 0.0011 and a maximum constraint violation of -0.0105 ± 0.0070 on the test set with respect to the true groups. For this dataset, it was possible to satisfy the constraints with approximately the same error rate on test as the unconstrained model. Note that the unconstrained model achieved a lower error rate on the train set than the constrained model (0.1792 ± 0.0015 unconstrained vs. 0.1798 ± 0.0024 constrained).

For both case studies, results in Figure 1 and 2 show that the robust approaches DRO (*center*) and soft group assignments (SA) (*right*) satisfy the constraints on average for all noise levels. As the noise level increases, the naïve approach (*left*) has increasingly higher true group constraint violations. The DRO and SA approaches come at a cost of a higher error rate than the naïve approach (Figure 3). The error rate of the naïve approach is close to the model optimized with constraints on the true groups G, regardless of the noise level γ. However, as the noise increases, the naïve approach no longer controls the fairness violations on the true groups G, even though it does satisfy the constraints on the noisy groups $\hat{G}$ (see Figure 4, Figure 7 in Appendix F.2). DRO generally suffers from a higher error rate compared to SA (Figure 1 and 2). This illustrates the conservativeness of the DRO approach and perhaps the looseness of the TV bound.

8 Conclusion

We explored the practical problem of enforcing group-based fairness for binary classification given noisy protected group information. In addition to providing new theoretical analysis of the naïve approach of only enforcing fairness on the noisy groups, we also proposed two new 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. [37] 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.

One avenue of future work would be to empirically compare the robust approaches when the noisy groups have different dimensionality from the true groups. We discuss this setup in Appendix B.4. Second, we note that the looseness of the bound in Lemma 1 can lead to over-conservativeness of the DRO approach, and future work would benefit from methods to better calibrate the DRO neighborhood. Finally, further study of the impact of distribution mismatch between the main dataset and the auxiliary dataset would be valuable future work.

Broader Impact

As machine learning is increasingly employed in high stakes environments, any potential application has to be scrutinized to ensure that it will not perpetuate, exacerbate, or create new injustices. Aiming to make machine learning algorithms themselves intrinsically fairer, more inclusive, and more equitable plays an important role in achieving that goal. Group-based fairness [32, 25] is a popular approach that the machine learning community has used to define and evaluate fair machine learning algorithms. Until recently, such work has generally assumed access to clean, correct protected group labels in the data. Our work addresses the technical challenge of enforcing group-based fairness criteria under noisy, unreliable, or outdated group information. However, we emphasize that this technical improvement alone does not necessarily lead to an algorithm having positive societal impact, for reasons that we now delineate.

Choice of fairness criteria

First, our work does not address the choice of the group-based fairness criteria. Many different algorithmic fairness criteria have been proposed, with varying connections to prior sociopolitical framing [48, 35]. From an algorithmic standpoint, these different choices of fairness criteria have been shown to lead to very different prediction outcomes and tradeoffs [25]. Furthermore, even if a mathematical criterion may seem reasonable (e.g., equalizing positive prediction rates with *demographic parity*), Liu et al. [45] show that the long-term impacts may not always be desirable, and the choice of criteria should be heavily influenced by domain experts, along with awareness of tradeoffs.

Choice of protected groups

In addition to the specification of fairness criteria, our work also assumes that the true protected group labels have been pre-defined by the practitioner. However, in real applications, the selection of appropriate true protected group labels is itself a nontrivial issue.

First, the measurement and delineation of these protected groups should not be overlooked, as "the process of drawing boundaries around distinct social groups for fairness research is fraught; the construction of categories has a long history of political struggle and legal argumentation" [31]. Important considerations include the context in which the group labels were collected, who chose and collected them, and what implicit assumptions are being made by choosing these group labels. One example is the operationalization of race in the context of algorithmic fairness. Hanna et al. [31] critiques "treating race as an attribute, rather than a structural, institutional, and relational phenomenon." The choice of categories surrounding gender identity and sexual orientation have strong implications and consequences as well [29], with entire fields dedicated to critiquing these constructs. Jacobs and Wallach [36] provide a general framework for understanding measurement issues for these sensitive attributes in the machine-learning setting, building on foundational work from the social sciences [5].

Another key consideration when defining protected groups is problems of *intersectionality* [15, 34]. Group-based fairness criteria inherently do not consider within-group inequality [38]. Even if we are able to enforce fairness criteria robustly for a given set of groups, the intersections of groups may still suffer [12].

Domain specific considerations

Finally, we emphasize that group-based fairness criteria simply may not be sufficient to mitigate problems of significant background injustice in certain domains. Abebe et al. [1] argue that computational methods have mixed roles in addressing social problems, where they can serve as *diagnostics*, *formalizers*, and *rebuttals*, and also that "computing acts as synecdoche when it makes long-standing social problems newly salient in the public eye." Moreover, the use of the algorithm itself may perpetuate inequity, and in the case of credit scoring, create stratifying effects of economic classifications that shape life-chances [24]. We emphasize the importance of domain specific considerations ahead of time before applying any algorithmic solutions (even "fair" ones) in sensitive and impactful settings.

Acknowledgments and Disclosure of Funding

We thank Rediet Abebe for many useful pointers into the broader socio-technical literature that provides essential context for this work, and Jacob Steinhardt for helpful technical discussions. We

also thank Stefania Albanesi and Domonkos Vámossy for an inspiring early discussion of practical scenarios when noisy protected groups occur. Finally, we thank Kush Bhatia, Collin Burns, Mihaela Curmei, Sara Fridovich-Keil, Frances Ding, Preetum Nakkiran, Adam Sealfon, and Alex Zhao for their helpful feedback on an early version of the paper. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1752814.

References

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*, 2018.

[3] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 1418–1426, 2019.

[4] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[5] Deborah L. Bandalos. *Measurement Theory and Applications for the Social Sciences*. Guilford Publications, 2017.

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[7] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

[8] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[9] Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4954–4965, 2019.

[10] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

[11] Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

[12] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Journal of Machine Learning Research (JMLR)*, 2018.

[13] A. Cotter, H. Jiang, and K. Sridharan. Two-player games for efficient non-convex constrained optimization. In *International Conference on Algorithmic Learning Theory (ALT)*, 2019.

[14] Andrew Cotter, Heinrich Jiang, Serena Wang, Taman Narayan, Seungil You, Karthik Sridharan, and Maya R. Gupta. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research (JMLR)*, 20 (172):1–59, 2019.

[15] Kimberle Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43:1241, 1990.

[16] Mark A. Davenport, Richard G. Baraniuk, and Clayton D. Scott. Tuning support vector machines for minimax and Neyman-Pearson classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[17] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[18] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[19] John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

[20] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *International Conference on Machine Learning (ICML)*, 2008.

[21] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pages 214–226. ACM, 2012.

[22] Elad Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Rif A. Saurous, and Gal Elidan. Scalable learning of non-decomposable objectives. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[23] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018.

[24] Marion Fourcade and Kieran Healy. Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society*, 38(8):559–572, 2013.

[25] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.

[26] Talia B Gillis. False dreams of algorithmic fairness: The case of credit pricing. *Available at SSRN 3571266*, 2020.

[27] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[28] Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. Achieving fairness with decision trees: An adversarial approach. *Data Science and Engineering*, 5(2):99–110, 2020.

[29] The GenIUSS Group. *Best Practices for Asking Questions to Identify Transgender and Other Gender Minority Respondents on Population-Based Surveys*. The Williams Institute, 2014.

[30] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

[31] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

[32] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[33] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018.

[34] Bell Hooks. *Yearning: Race, gender, and cultural politics*. 1992.

[35] Ben Hutchinson and M. Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.

[36] Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. *arXiv preprint arXiv:1912.05511*, 2019.

[37] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

[38] Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. *ICML Workshop on Participatory Approaches to Machine Learning*, 2020.

[39] Niko Kolodny. Why equality of treatment and opportunity might matter. *Philosophical Studies*, 176:3357–3366, 2019.

[40] Ivar Krumpal. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality and Quantity*, 47:2025–2047, 2011.

[41] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[42] Alexandre Lamy, Ziyuan Zhong, Aditya Krishna Menon, and Nakul Verma. Noise-tolerant fair classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[43] Claire Lazar and Suhas Vijaykumar. A resolution in algorithmic fairness: Calibrated scores for fair classifications. *arXiv preprint arXiv:2002.07676*, 2020.

[44] Jiajin Li, Sen Huang, and Anthony Man-Cho So. A first-order algorithmic framework for Wasserstein distributionally robust logistic regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[45] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning (ICML)*, 2018.

[46] Hongseok Namkoong and John Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[47] Harikrishna Narasimhan, Andrew Cotter, and Maya R. Gupta. On making stochastic classifiers deterministic. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[48] Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2018.

[49] Chris Russell, Matt J. Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[50] Cédric Villani. *Optimal Transport, Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, 2009.

[51] Serena Wang and Maya R. Gupta. Deontological ethics by monotonicity shape constraints. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[52] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory (COLT)*, pages 1920–1953, 2017.

[53] I-Cheng Yeh and Che hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36: 2473–2480, 2009.

[54] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[55] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research (JMLR)*, 20:1–42, 2019.