Decentering Imputation: Fair Learning at the Margins of Demographics

Evan Dong and Cyrus Cousins

July 2022

Abstract

Defining demographic categories for fair learning frequently requires both private information collection at the individual level and expert judgment in the construction of relevant categories, which is often removed from common task data collection. In this work, we present a new approach to empowering data owners and model developers with the flexibility to move toward more nuanced construction of identity in given correlates. In particular, our contributions are as follows: (1) an analysis of the ethics and politics of imputing demographic data: the double-bind of imputing sensitive attributes, the flaws of such approaches in theory, practice, and interpretation, and the technical implications of these politics; (2) a mathematical and conceptual framework for analyzing/understanding uncertain group identities; and (3) a correspondingly reformulated objective and adversarial minimax training algorithm for fair learning, with provable accuracy and training efficiency guarantees.

1. Introduction

As machine learning techniques proliferate in more varied and consequential decisions, considerations of group-level fairness become increasingly critical. While a burgeoning literature (Barocas et al., 2017; Chouldechova and Roth, 2018; Cousins, 2021; 2022; Cousins et al., 2022; Friedler et al., 2019; Mehrabi et al., 2021) has developed to explore and satisfy various group-based fairness objectives, by nature most such techniques require access to relevant demographic data. However, not all datasets may include such information, due to data collection limitations or privacy concerns. Moreover, the definitions and categories of demographics subject to fairness concerns is complex and changing. Data schemas and collection methods are political choices, which then can be a point of contention and site of erasure for different identities. For example, binary gender or sex options, whether in government documentation or customer surveys, are tautologically limiting and erasive. This excludes a breadth of intersex, non-binary, two-spirit, and gender-nonconforming people and erases the materially different experiences of transgender people that do identify within a binary (Keyes, 2018). These problems are not limited to gender; for example, many Southwest-Asian and North-African ethnicities are counted as white, despite facing discrimination and structural barriers (Maghbouleh et al., 2022). Delineating the intersections and spectra of identities into discrete groups is both contextual and dynamic (Benthall and Haynes, 2019), while also overlapping with structural questions of data ownership, privacy, and power.

This work seeks to address these challenges. While the relevant dimensions and specifications for fairness depend on domain expertise, being able to project well-defined granular data or distribution specifications defined by such experts is an important technical step for operationalizing these considerations into learning tasks. This paper introduces a method to create a group-labeling adversary from auxiliary demographic data, which can be applied to a sufficiently similar task dataset with partial, noisy, or missing sensitive attribute information for fair training. To do this, we calculate statistics on a small true-group-labeled dataset, and probabilistically bound the difference of these statistics calculated on training data. This generates a *feasible set* encapsulating our partial knowledge of training group labels, within which the true training data demographics fall with high probability.

While this framework is broad, we focus on mean statistics for their ease of interpretability, well-studied and strong ε-δ *additive error tail bounds*. These bounds constrain a linear program adversary, which finds a worst-case demographic distribution given model parameters. We then use this adversary to (provably and efficiently) train a fair model via the *projected subgradient method*, by optimizing model parameters for Rawlsian fairness (i.e., minimizing worst-case risk over all protected groups) against *adversarially selected* feasible group labelings over the task dataset.

1.1. Related Work

Fair Machine Learning and Minimax Learning Fair machine learning includes a plethora of different definitions (Mehrabi et al., 2021), criteria (Corbett-Davies and Goel, 2018), and approaches. Formulation of fairness as a minimax problem (Abernethy et al., 2020; Cousins, 2021; 2022; Diana et al., 2021; Mohri et al., 2019) over per-group risk (expected loss) have become common. These methods often dovetail nicely with interpretations of John Rawls' political philosophy — specifically, the *Difference Principle* (Rawls, 2001), to define a notion of *Rawlsian Maximin* (or minimax) fairness. This is mathematically equivalent to the setting of *Group Distributionally Robust Optimization* (Group DRO) (Hu et al., 2018; Oren et al., 2019; Sagawa et al., 2019), which considers different protected groups as a family of distributions, where minimizing worst-case empriical risk can handle spurious correlations and under- or over-parametrized settings. As Rawlsian fairness can provide a convex objective (exact formulation outlined in (4) of section 4.1), drawing upon minimax optimization techniques offers computational feasibility guarantees.

Minimax Learning for Uncertainty Minimax techniques are also used to navigate uncertainty in fairness-unrelated domains. The ALL framework (Arachie and Huang, 2019; 2021), and in particular the setup of Mazzetto et al. (2021) is quite similar to ours, except their uncertainty is over *unknown labels* in classification (rather than *unknown groups*), and they simply seek to minimize the unweighted risk over the entire population (rather than a Rawlsian objective). In particular, they have the empirical objective

$$\tilde{\theta} = \operatorname*{argmin}_{\theta} \max_{\boldsymbol{y} \in \boldsymbol{Y}} \sum_{i=1}^{m} \ell(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i) \ . \tag{1}$$

We note that, despite their surface similarity, significant novel technical hurdles arise in this work. In particular, observe that (1) has convex-concave structure in θ and $\boldsymbol{y}$, making it directly amenable to standard minimax optimization techniques. In contrast, due to the Rawlsian objective not being concave in the unknown group identities $\boldsymbol{z}$, the same does not happen in our work, thus complicating the construction of the adversary.

Fairness Without Demographics These techniques converge when approaching fair learning without access to demographic data. Identifying fairness disparities requires some source of information to infer possible relevant groups. Hashimoto et al. (2018) uses a form of DRO for fair learning over unknown attributes in repeated, dynamic decision-making with fairness disparities being amplified over time. *Adversarially Reweighted Learning* (ARL) (Lahoti et al., 2020), in contrast, draws upon a notion of computational identifiability to infer demographic groups based on within-data covariance.

However, identifying potential groups is not perfect; by construction, these techniques may fall prey to a tendency to focus on borderline examples, at a cost to efficacy across defined, interpretable groups of concern. Similar to both of these techniques, we focus our efforts on a Rawlsian definition of fairness, but instead explicitly define and model protected groups of interest to avoid this problem. Our work differs in setting by incorporating not only auxiliary data and *a priori* knowledge about discrimination, but also reframing training from problems of group estimation to bounded group possibilities. While techniques utilizing additional data or prior information can be quite common, they often fall into a separate framework — which we believe our work straddles the boundary between.

Auxiliary Data and Proxies The use of auxiliary data in many other challenging machine learning domains is well-established; approaches in semi-supervised, self-supervised, zero-shot, and other paradigms of learning with limited labeled data often leverage access to other data sources — sometimes lacking labels or drawn from a different distribution. In the realm of fairness, other work has explored utilizing auxiliary data to both assess fairness disparities and inform training, with some caveats.

A common approach is to estimate fairness-relevant information. Kallus et al. (2022) use the notion of auxiliary data to produce a set of possible fairness-related values consistent with training data — finding the impossibility of calculating exact values of fairness disaprities. Chen et al. (2019) dub the usage of models that enable the imputation of demographic data, primarily for the purposes of fairness assessment, as proxy models. The *Bayesian Improved Surname Geocoding* model (Elliott et al., 2009) uses information such as name and geographic location, which regulators have applied with real-world impacts (Bureau, 2014). Similarly, developing proxy groups based on correlates (Gupta et al., 2018) is another approach built toward fair model training. More recently, Diana et al. (2022) use a notion of multiaccuracy to define a robust proxy estimate for a given dataset, which downstream learners can train fairly over. These techniques have their own room for improvement; Chen et al. (2019) note that most proxy estimates have statistical bias, Diana et al. (2022) note some challenges with efficiently learnable estimates, and Gupta et al. (2018) advise caution in selecting proxy variables.

2. Considerations Around Imputation

Motivations for Imputation The driving motivation behind this approach is understandably nuanced in scope, degree, and application. The contexts around data collection, imputation, and ownership are varied. We acknowledge that there are circumstances where making assumptions and estimations are acceptable, and also that refusing to work with particular demographic data whatsoever may be well-justified by in context. In some cases, imputation is a harm-minimizing choice. The US Census Bureau argued in *Utah v. Evans* that refusing to impute household size data would be equivalent to imputing a degenerate value of 0 — assuming empty homes (Cantwell et al., 2004). In the US Census, this means systematic undercounting by geographic region. Within machine learning, ignoring fairness considerations when demographic labels are missing may not only be an ineffective approach of fairness through unawareness, but also frequently a *de facto* assumption of dominant hegemonic identities within the relevant population.

Individual Predictions and Ecological Fallacy At the same time, there are discomforting flaws about imputation at its most aggressive — particularly when predicting entirely new values. A conventional model, however grounded in true statistical relationship, ultimately assigns predictions and labels to individuals — particularly when built on historically injustices — in what some would deem algorithmic stereotyping. By construction, inferring predictions about individuals due to aggregate statistics about groups they belong to leads to aggregation bias; combined with confounding causal factors, this becomes the *ecological fallacy* (Freedman, 1999). Not everyone living in an affluent majority-white suburb is white and wealthy, and certainly not every person with a commonly gendered legal name lives and identifies the same way — and such underlying mechanisms are certainly more complex. By their nature, these model predictions encode stereotypes.

Implications for Structural Inequality Benjamin (2019) critiques the use of demographic prediction, albeit in the realm of advertising segmentation. Her perspective contextualizes these algorithms in histories of discrimination: "By fixing group identities as stable features of the social landscape," she argues, "[d]ifference is... now codified beyond the law, in the digital structures of everyday life." While not all applications of proxy models necessarily exploit race and ethnicity for the sake of profit, they are nonetheless built on — and are effective because of — historical injustices. Understanding predictive models as a mode of knowledge (re)production means that every application replicates the act of racialization. Incorporating a critical race methodology of fairness (Hanna et al., 2020) means situating the algorithmic act of classification in the racial project of (re)constructing race upon marginalized people. Anti-racism must critically navigate constructions of race without reinforcing them, and our work seeks to accomplish this by utilizing a flexible concept of race to prevent discrimination, rather than to provide differential treatment.

In the LGBTQ community, these problems are readily apparent in the case of trans-exclusive automatic gender recognition (Keyes, 2018), which is almost always built upon a binary. While most discourse focuses on prediction with direct physiological traits, especially facial recognition, the gendered nature of everything from occupation to marriage status and name superimpose historical norms upon us. Beyond the magnified privacy concerns for marginalized individuals (Hamidi et al., 2018), any attempt to promote fairness based on such proxy methods is built on foundations that already exclude the most marginalized — often to the point of literal poor model performance at identity intersections (Buolamwini and Gebru, 2018). This is particularly problematic for the notion of Rawlsian fairness; the least well-off are quite literally erased from consideration to begin with. Imputed data that misgenders transgender women will lead to models that present only an incorrect and superficial notion of fairness — a *mechanical TERF*, if you will. Those on the structural margins, as articulated in Black feminist theory (hooks, 2000), often end up at statistical margins.

Refusing Data Collection We take care to distinguish this problem setting of poor or impractical data collection from strategic collective or individual choices to not be studied and have their data collected. Intentional, politically motivated choices to refuse research serve as a way of claiming sovereignty over particular forms or pieces of knowledge, as documented and argued by Tuck and Yang (2014). When situated in a decolonial agenda, this serves to draw boundaries against research as an imperial project and highlight invisibilized limitations of colonial academia.

Similarly, even when data may be collected, data sovereignty, governance, and control are necessary considerations before seeking out auxiliary data — or predicting data over communities that have refused to be categorized. Even beyond individual rights to privacy and refusal, the collective impact on and control over knowledge held by marginalized communities outweighs hegemonic notions of fairness enforced from within the imperial metropole. For example, te reo Māori data collected by Te Hiku Media (2018) is used to develop language models within their community, motivating and contextualizing their — and their communities' — refusal to share their language

data in rights of ownership, leadership, and control. Our work ought not justify colonial acquisition of data without both informed individual consent and a collectively accountable distribution of power.

Navigating Needs However, there remain plenty of cases where a lack of data is both unintentional — or not refused on specifically so motivated grounds — and harmful to fairness. By far is it more likely that most binary gender or sex categories on demographic surveys come from a desire for convenience in data cleaning (or, more bluntly, erasure of TGNC and intersex people) than more complex, community-engaged motivations. Our goal then is to satisfy the motivating shortfalls behind imputation and proxy models without falling into their pitfalls in both theory and practice. We believe our proposed technique answers the above challenges; we avoid aggregation bias and its compounding effects both on defining worst-case fairness and reproducing structural inequality. Moreover, we weaken the data access necessary to seek fairness, allowing for the exportation a fairness-defining adversary derived from a private dataset without handing over individual records. In other words, a trusted data owner need only provide group-level mean statistics and relevant calculated bounds to implement a fair training adversary for a given task.

3. Approach

In this section, we outline the high-level framework of our approach, formally defining our problem setting in section 3.1, the notion of statistically constrained possible group labels in section 3.2, and the advantages and practical interpretation of this technique in sections 3.3 and 3.4.

3.1. Preliminaries

In conventional group-based fair learning, with access to group membership data at training time, our data points are triplets $(x, y, z) \in (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})$, where $\mathcal{Z} \doteq \{1, 2, \ldots, g\}$ is a (WLOG finite) space of g *protected groups*. Instead, in this paper, we have a task-specific dataset of m_y training data points in features $\boldsymbol{x_y}$ and labels $\boldsymbol{y}$, with limited knowledge of memberships in $\mathcal{Z}$. Additionally, we have an auxiliary demographically rich dataset $\boldsymbol{x_z}$ and group membership data $\boldsymbol{z}'$ of size m_z points. Our approach uses this auxiliary data to construct a feasible set $\boldsymbol{Z}$ of possible $\boldsymbol{z}$ memberships across all data points, which contains the true distribution $\boldsymbol{z}^*$ across the task dataset $\boldsymbol{x_y}$ as well as the worst-case (with respect to a fairness metric) distribution of group memberships given model parameters θ.

For the sake of computational tractability, we relax $\mathcal{Z}$ such that $\boldsymbol{Z}$ is a *convex set*, i.e., we work with the convex relaxation space $\mathcal{Z}_\diamond^m$ for any m data points. In particular, $\mathcal{Z}_\diamond^m$ is a *simplicial product space*, thus for each $\boldsymbol{z} \in \boldsymbol{Z}$ and index $i \in 1, \ldots, m$, the values $\boldsymbol{z}_{i,\cdot}$ — like discrete one-hots — sum to 1. Each $\boldsymbol{z} \in \boldsymbol{Z}$ is a right-stochastic matrix $\boldsymbol{z} \in [0, 1]^{m \times g}$, in the simplex $\triangle_g^m$, such that $\boldsymbol{z}_{i,\cdot}$ (row i) is thus a *distribution* (weighting) over groups. While this could potentially interpreted as a sense of partial membership, we note that the nature of such concepts for many relevant axes of identity are much more complex in dimension, scope, and dynamic than captured here.

Assuming that $\boldsymbol{x_z}$ and $\boldsymbol{x_y}$ are drawn from the same distribution, we can establish this feasible set using statistical tail bounds on a set of the group-level statistics of our training data $\boldsymbol{x_y}$. We restrict our analysis in section 3.2 and section 4.4 to mean statistics but this is generalizable to other statistics, e.g., *variance* or *CDF* statistics, given the appropriate tail bounds.

No Information	**Partial Information**	**Full Information**
Any $\boldsymbol{z}$ is feasible $\boldsymbol{Z} = \triangle_g^m$	Some $\boldsymbol{z}$ are feasible $\boldsymbol{z}^* \in \boldsymbol{Z} \subset \triangle_g^m$	Only $\boldsymbol{z}^*$ is feasible $\boldsymbol{Z} = \{\boldsymbol{z}^*\}$
Unconstrained DRO	This work, η-DRO ARL	Group DRO Minimax Fair Learning Egalitarian Malfare Min.

Table 1. Comparison of various information settings in group-fair learning.

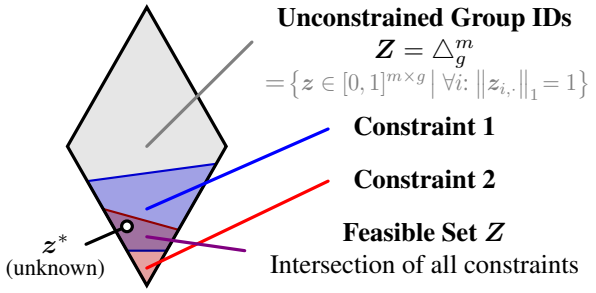

Figure 1. A feasible set $\boldsymbol{Z}$ of group labelings is generated by the intersection of multiple constraints on the simplicial space of possible per-instance group labelings. Constraint one is *bidirectional*, constraint two is *unidirectional*, and their intersection forms the feasible set $\boldsymbol{Z}$, which contains the true $\boldsymbol{z}^*$ (with high probability, assuming these are valid statistical constraints, as described in section 4.4).

3.2. Constraining the Feasible Set

We now discuss the statistical mechanisms for deriving constraints on the feasible set. Similar to Kallus et al. (2022), we utilize statistical knowledge from the auxiliary data distribution to bound the range on an unknown quantity. In this case, however, instead of a given dispar-

ity measure, we bound the space of demographic labels directly, much like how (Mazzetto et al., 2021) constrain feasible labelings. As emphasized, the crux of this technique depends on the definition of the feasible set $\boldsymbol{Z}$ through a set of statistics U. Our analysis and implementation focus on mean statistics, where we have k functions $u \in U$ on individual data points and their corresponding mean functions $\bar{u}(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{m}\sum_{i=1}^{m} u(\boldsymbol{x}_i, \boldsymbol{z}_i)$. As some basic examples, this might be the mean proportion of a given demographic attribute "60% of this population self-identifies as white," or a conditional or joint mean ("70% of people in this income bracket are men," or "5% of this entire population is both queer and unmarried"). We focus on statistics relevant to our protected groups — capturing the joint distribution(s) across $\boldsymbol{z}'$ and $\boldsymbol{x}_{\boldsymbol{z}}$.

For each u, the sample mean over the auxiliary dataset $\bar{u}(\boldsymbol{x}_{\boldsymbol{z}}, \boldsymbol{z}')$ allows us to estimate the true mean $\mu = \mathbb{E}_{x,z}[u(x, z)]$ of the underlying distribution, using tail bounds in the same spirit as a confidence interval (elaborated in detail in section 3.3). This, in turn, can — with high probability — bound the value of the empirical mean $\bar{u}(\boldsymbol{x}_{\boldsymbol{y}}, \boldsymbol{z}^*)$ on another sample (namely, the task dataset), producing constraints. We describe specific bounds and techniques in section 4.4.

We cannot calculate $\bar{u}(\boldsymbol{x}_{\boldsymbol{y}}, \boldsymbol{z}^*)$ on the task dataset without knowing $\boldsymbol{z}^*$. However, if it is confined within some range (i.e., must be within ε of μ) by using the bounds described above, it becomes possible to work backwards and limit the domain of possible $\boldsymbol{z} \in \boldsymbol{Z}$ through constraints on the range $\bar{u}(\boldsymbol{x}_{\boldsymbol{y}}, \boldsymbol{z}) \in [\mu_i - \varepsilon, \mu_i + \varepsilon]$ for all $u \in U$, forming the basis of our adversary in section 4.2.

3.3. Philosophical Interpretations

Our technique differs from imputation-based methods. Instead of making statistical predictions on a singular most likely distribution of protected attribute labels, we instead bound a statistically feasible set of possible distributions. In an analogy to regression analysis, instead of interpreting a singular identity prediction as a mean conditioned on covariates (e.g., "70% chance that any given person who has changed their name is transgender"), our feasible set is instead a confidence interval ("with high probability, 6–8 of these 10 people are in some way gender nonconforming"). This explicitly avoids the ecological fallacy outlined earlier; we avoid applying stereotypes to individuals, as we work only with *statistics across multiple people*. This not only guarantees direct transparency with respect to grounding the demographics used for fair training, but also provides a less presumptive approach in interpreting a chosen $\boldsymbol{z}$.

3.4. Example

To demonstrate this interpretation, we present a simplified example. For example, say that 7% of our target population self-identifies as LGBTQ in our auxiliary data, as per Jones (2022). Even if we have little or no information about LGBTQ identity in our training data, we can bound the probable proportion of LGBTQ individuals in our task dataset within some confidence interval (say, between 6% and 8% with probability at least 95%). This rules out statistically improbable distributions of demographics (e.g., the unlikely case that every person in the task sample is gay). This single example alone is insufficient to ground fair training, but the addition of more statistics further constraints the feasible set of distributions. If according to our trusted data, LGBTQ people are roughly evenly distributed by city district, our adversary is barred from proposing a world where every queer person in Philadelphia lives exclusively in the Gayborhood; similarly, if demographic trends show as such, it can reject distributions that erase geographic concentrations around queer communities. As the feasible set of possible worlds shrinks, the power of the adversary wanes and it becomes possible for a model to minimize risk for a group with respect to any distribution within the bounds, allowing a fair training regime.

4. Model

This section describes the choices we use to apply and implement our framework to a given problem, defining a convex learning objective in section 4.1, a tractable adversary applying the feasible set in section 4.2, convergence guarantees in section 4.3, and statistical techniques for deriving effective constraints in section 4.4.

4.1. Objective

We adapt our notion of fairness to this problem space; similarly to ARL and DRO, we focus on Rawlsian Min-Max fairness and utilize an adversarial approach to solve the optimization problem. While many — sometimes mutually exclusive — definitions of fairness exist, an optimally fair hypothesis θ^* is here defined as

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \max_{j \in \mathcal{Z}} \mathop{\mathbb{E}}_{x,y} [\ell(h_\theta(x), y) | Z = j] \qquad (2)$$

for a loss function ℓ. Working in our continuous relaxation of $\mathcal{Z}$ means that calculating the empirical risk conditioned on each protected group is less straightforward. Instead of simply calculating the average empirical risk on a partitioned subset of the dataset for each group, we generalize this idea to a weighting of the empirical risk calculation by the fractional membership for each data point. On a

sample, this gives

$$\mathbb{E}[\ell(h_\theta(x), y)|Z = j] \approx \frac{\sum_{i=1}^m \boldsymbol{z}_{i,j}\ell\big(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i\big)}{\sum_{i=1}^m \boldsymbol{z}_{i,j}} \quad (3)$$

for each j in $\mathcal{Z}$. This is the conditional empirical risk for each group. Note that, with discrete z, as with one-hot vectors, this cleanly reduces to the conventional conditional empirical risk.

The empirical Rawlsian risk minimizer

$$\tilde{\theta}_{z^*} = \operatorname*{argmin}_{\theta\in\Theta} \max_{j\in\mathcal{Z}} \frac{\sum_{i=1}^m \boldsymbol{z}^*_{i,j}\ell\big(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i\big)}{\sum_{i=1}^m \boldsymbol{z}^*_{i,j}} \quad (4)$$

serves as practical approximation to θ^* in the same vein as (3). As we do not know the ground-truth group memberships $\boldsymbol{z}^*$, we instead seek the $\tilde{\theta}_{\boldsymbol{Z}}$ that minimizes its worst statistically-feasible Rawlsian empirical risk. This comes through an additional maximization term over the feasible set $\boldsymbol{Z}$, which incorporates the varying restrictions on $\boldsymbol{z}$, i.e.,

$$\tilde{\theta}_{\boldsymbol{Z}} = \operatorname*{argmin}_{\theta\in\Theta} \max_{\boldsymbol{z}\in\boldsymbol{Z}} \max_{j\in\mathcal{Z}} \frac{\sum_{i=1}^m \boldsymbol{z}_{i,j}\ell\big(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i\big)}{\sum_{i=1}^m \boldsymbol{z}_{i,j}} \ . \quad (5)$$

Each of these solution concepts are theoretical constructs and, like most solutions to minimization problems, generally intractable to compute exactly. In section 4.3 we provide a training algorithm that defines a $\hat{\theta}$ that ε-approximates the objective $\tilde{\theta}_{\boldsymbol{Z}}$.

4.2. Defining the Adversary

Formulation as Linear Program Put plainly, our adversary calculates the maximum terms of (5) for a given value of θ, i.e.,

$$\max_{\boldsymbol{z}\in\boldsymbol{Z}} \max_{j\in\mathcal{Z}} \frac{\sum_{i=1}^m \boldsymbol{z}_{i,j}\ell\big(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i\big)}{\sum_{i=1}^m \boldsymbol{z}_{i,j}} \quad (6)$$

$$= \max_{j\in\mathcal{Z}} \max_{\boldsymbol{z}\in\boldsymbol{Z}} \frac{\sum_{i=1}^m \boldsymbol{z}_{i,j}\ell\big(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i\big)}{\sum_{i=1}^m \boldsymbol{z}_{i,j}} \ . \quad (7)$$

Note that while both the numerator and denominator of this term are linear in $\boldsymbol{z}$, their ratio is neither concave nor convex in $\boldsymbol{z}$. While, in the general case, maximization over non-concave functions is difficult, (6) can be made tractable. While $\boldsymbol{Z}$ is infinite, making enumeration impossible, $\mathcal{Z}$ is not, so commuting (6) to (7) allows us to only solve a small number of g instances of the inner maximization problem.

Now, all that remains is to solve the inner maximization problem in (7). Let the numerator, the weighted empirical risk $\sum_{i=1}^m \boldsymbol{z}_{i,j}\ell\big(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i\big)$, be $L^{(j)}(\boldsymbol{z})$, and the denominator $\sum_{i=1}^m \boldsymbol{z}_{i,j}$ be $W^{(j)}(\boldsymbol{z})$ for any $j \in \mathcal{Z}$. For a given θ, let $\mathrm{R}^{(j)}(\theta, \boldsymbol{z}) = \frac{L^{(j)}}{W^{(j)}}$ for a particular group j. We assign partial group labels to data point such that it maximizes the empirical risk of a given individual group. In other words, it describes the worst-case group membership for fairness with respect to the minimax criterion within the feasible set. This can be constructed as a maximization over a set of optimization problems. Each of the mg variables represents $\boldsymbol{z}_{i,j}$ for some $i = 1, ..., m$ and $j \in \mathcal{Z}$, with the matrix of constraints $\boldsymbol{C}$ constructed from both statistics in U and m simplex restrictions (i.e., variables for each i must sum to 1). Let $\vec{z}$ be the vector-form unrolling of $\boldsymbol{z}$, and $\vec{b}$ likewise a vector of bounds corresponding to rows in $\boldsymbol{C}$.

Note that, for each $j \in \mathcal{Z}$, this is a linear fractional program

$$\max \frac{L^{(j)}(\boldsymbol{z})}{W^{(j)}(\boldsymbol{z})} : \boldsymbol{C}\vec{z} \preceq \vec{b} \ , \quad (8)$$

which can be transformed into a standard linear program

$$\max L^{(j)}(\boldsymbol{v}) : \boldsymbol{C}\vec{v} \preceq \vec{b}t, W^{(j)}(\boldsymbol{v}) = 1, t \geq 0 \ , \quad (9)$$

with $\boldsymbol{v} = \frac{\boldsymbol{z}}{W^{(j)}(\boldsymbol{z})}$, and additional variable $t = \frac{1}{W^{(j)}(\boldsymbol{z})}$, using the Charnes-Cooper (1962) transformation. We calculate this for each group $j \in \mathcal{Z}$, storing values of $\boldsymbol{z}_j = \frac{\boldsymbol{v}^{(j)}}{t}$ for each j, as well as the objective value $\mathrm{R}^{(j)}(\theta, \boldsymbol{z}^{(j)})$. The output of our adversary is then the maximum (and corresponding $\boldsymbol{z}^{(j)}$) across these g linear programs, $\Lambda(\theta, \boldsymbol{z}) = \max_{j\in\mathcal{Z}} \mathrm{R}^{(j)}(\theta, \boldsymbol{z}^{(j)})$ — the Rawlsian fairness empirical risk for any given weights θ and their corresponding worst-case demographic distribution. Note that each of these linear programs can be efficiently solved in polynomial time (Jiang et al., 2021; Karmarkar, 1984), meaning our overall adversary is likewise tractable.

Efficiency Concerns To minimize computational costs, we note that individual small gradient updates will change little in each component linear program. Therefore, the prior solution serves as a warm-start initialization point for solving methods. Additionally, remembering prior solutions as bounds can help avoid solving every linear program every time. If the loss function ℓ is λ-Lipschitz, then we can bound the increase in the conditional loss for each group, allowing us to determine that a group's risk is not maximal without solving the corresponding linear program.

4.3. Fair Learning Given a Feasible Set

We provide a polynomial-time training procedure algorithm 1, interweaving adversary computations between each gradient step the projected subgradient method of Shor (2012) to get sets of group label weightings.

Practically speaking, gradient descent will apply to most situations while offering faster calculation. If a function

is convex, at least one subgradient exists for each point within the domain; where the function is differentiable, the gradient is a unique subgradient. Within this maximization problem, there will only be multiple subgradients at non-differentiable points — where multiple terms in the maximization, and therefore, worst-case empirical risk for multiple groups, are equal. This means our model for any task gradient descent-compatible when agnostic to fairness can be initially more simplistically trained to get a warm-start set of parameters $\theta^{(0)}$.

Even without assuming a practical warm start, if our subgradient calculations and updates are efficient, training a learning model with convex loss function to optimize Rawlsian fairness is achievable in polynomial time.

Theorem 4.1 (Efficient Adversarial Training)**.** Suppose the convex parameter set $\Theta \subset \mathbb{R}^d$ is bounded s.t. (Euclidean) $\text{Diam}_2(\Theta) \leq R$ and $\text{Proj}_\Theta(\vec{q}) \; \forall \; \vec{q} \in \mathbb{R}^d$ is computable in $\text{Poly}(d)$ time. Suppose also that the hypothesis $h_\theta(x)$ and the (sub)gradient $\nabla_\theta h_\theta(x)$ can be evaluated in $\text{Poly}(d)$ time, and furthermore that the loss function $\ell : \mathcal{Y}' \times \mathcal{Y} \mapsto \mathbb{R}$ s.t. $\forall x \in \mathcal{X}, y \in \mathcal{Y} : \theta \mapsto \ell(h_\theta(x), y)$ is convex and λ-Lipschitz continuous. Finally, suppose g groups, m training data points, convex feasible set polytope $\boldsymbol{Z} \subseteq \triangle_g^m \subset [0,1]^{m \times g}$ defined by c linear constraints, and additive error tolerance $\epsilon > 0$. Then algorithm 1 terminates in $\text{Poly}(g, m, d, c, R, \lambda, \frac{1}{\varepsilon})$ time and returns a $\hat{\theta} \in \Theta$ s.t.

$$\max_{\boldsymbol{z} \in \boldsymbol{Z}} \text{M}(\hat{\theta}, \boldsymbol{z}) - \max_{\boldsymbol{z} \in \boldsymbol{Z}} \text{M}(\tilde{\theta}_{\boldsymbol{Z}}, \boldsymbol{z}) \leq \varepsilon \;\; .$$

Proof. Our Rawlsian fairness objective

$$\underset{\theta \in \Theta}{\text{argmin}} \max_{\boldsymbol{z} \in \boldsymbol{Z}} \max_{j \in \mathcal{Z}} \frac{\sum_{i=1}^m \boldsymbol{z}_{i,j} \ell\big(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i\big)}{\sum_{i=1}^m \boldsymbol{z}_{i,j}}$$

is a maximum over conditional empirical risk. While this conditional is not concave in $\boldsymbol{z}$, the relevant composition remains convex in θ. This objective is then likewise convex — allowing for application of convex optimization techniques. However, as the inner term uses is a maximum over disjoint terms, this objective is not differentiable everywhere, and so gradient-based methods (e.g., stochastic gradient descent) theoretically do not apply. Nonetheless, convexity guarantees the existence of a subgradient, which generalizes the gradient for convex functions. Shor (2012), assuming only a convex Lipschitz-continuous objective function on a bounded domain $f(\theta)$, proves that repeated weight updates from of projected subgradient yields a set of weights $\hat{\theta}$ that evaluate the objective with at most additive ε error from that of minimum $\tilde{\theta}_{\boldsymbol{Z}}$. Concretely, given a choice of step size α and number of iterations T,

$$f(\hat{\theta}) - f(\tilde{\theta}_{\boldsymbol{Z}}) \leq \frac{R^2 + \lambda^2 \alpha^2 T}{2\alpha T} \;\; . \tag{10}$$

Consequently, applying the projected subgradient method in algorithm 1 yields an ε-optimal solution.

We now bound the time complexity of algorithm 1. From the error guarantee Shor (2012) provides for the projected subgradient method, it follows that additive error ε can be achieved with choice of step size $\alpha = \frac{\varepsilon}{\lambda^2}$ and iterations $T = \frac{R^2 \lambda^2}{\varepsilon^2}$ (line 4). Calculating $\text{M}(\theta, \boldsymbol{z})$ and the subgradient take $\text{Poly}(g, m, d)$ time with polynomial evaluation of $h_\theta(\boldsymbol{x}_i)$ (line 9), and projected subgradient updates take $\text{Poly}(d)$ time (line 10). Given a feasible set $\boldsymbol{Z}$ in the form of a convex H-polytope, the adversarial maximization (line 11) essentially consists of solving g linear programs with $\mathbf{O}(mg)$ variables and $\mathbf{O}(m + c)$ constraints (see section 4.2) and is thus solvable in $\text{Poly}(g, m, c)$ time, as each linear program takes worst-case polynomial time. Then, combined with the $\mathbf{O}(\frac{R^2 \lambda^2}{\epsilon^2})$ number of training steps, the total number of computational steps taken by algorithm 1 is

$$\mathbf{O}(\tfrac{R^2 \lambda^2}{\epsilon^2}) \cdot \text{Poly}(g, m, d, c) \subset \text{Poly}(g, m, d, c, R, \lambda, \tfrac{1}{\varepsilon}) \;\; .$$

□

These guarantees in error and computability provide practical value. Without correctness guarantees, a fair learning model may be unreliable on highly consequential or biased tasks. As a technique born of concern for understudied or highly private populations, the value of this guarantee compounds where downstream or *post-hoc* fairness assessment may be less feasible. Similarly, error bounds articulate a defined tolerance to replace the unreliability of imputation-based tools, which have magnified risk for those multiply marginalized. To promise fairness, error at every junction must be limited.

The modular nature of our statistical bounds and auxiliary data sources situates our work in contexts of grassroots data collection and community-engaged algorithm deployment. We "democratize" fair learning for easy implementation by keeping computational requirements accessible. Without polynomial-time tractability, the computing resource requirements for implementation would prevent practical application by many relevant under-resourced and marginalized stakeholders.

4.4. Calculating Practical Constraints

Tight bounds are critical to the efficacy of our technique. Assuming that data points are drawn independently from the same distribution, the absolute difference of an empirically calculated statistic from its expectation is bounded by additive error ε with probability at least $1 - \delta$. Applying this over many statistics, we can bound how much statistics from a small, well-collected dataset with group

Algorithm 1 Fair Adversarial Training

procedure FAIRTRAINING($\ell, \boldsymbol{x_y}, \boldsymbol{y}, \Theta, R, \theta^{(0)}, \boldsymbol{Z}, \lambda, \varepsilon) \to \hat{\theta}$
  input: loss function $\ell \in \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}$, data points $\boldsymbol{x_y}$, labels $\boldsymbol{y}$, parameter space Θ, parameter space Euclidean diameter R, initial warm-start model parameters $\theta^{(0)}$, constrained feasible set $\boldsymbol{Z}$, loss function Lipschitz bound λ, chosen error bound ε
  output: worst-group empirical risk minimizing hypothesis $\hat{\theta}$
  $\alpha \leftarrow \frac{\varepsilon}{\lambda^2}; T \leftarrow \frac{R^2\lambda^2}{\varepsilon^2}$ ▷ calculate sufficient step size α, number of iterations T
  $\Lambda(\theta, \boldsymbol{z}) \doteq \max_{j \in \mathcal{Z}} \frac{\sum_{i=1}^{m_y} \boldsymbol{z}_{i,j} \ell\big(h_\theta(\boldsymbol{x}_i), \boldsymbol{y}_i\big)}{\sum_{i=1}^{m_y} \boldsymbol{z}_{i,j}}$ ▷ define Rawlsian objective of h_θ for group labeling $\boldsymbol{z}$
  $\boldsymbol{z}^{(0)} \leftarrow \operatorname{argmax}_{\boldsymbol{z} \in \boldsymbol{Z}} \Lambda(\theta^{(0)}, \boldsymbol{z})$ ▷ adversarially choose groups labelings $\boldsymbol{z}^{(0)}$ against $\theta^{(0)}$
  $\hat{\theta} \leftarrow \theta^{(0)}; \Lambda_{\min} \leftarrow \Lambda(\theta^{(0)}, \boldsymbol{z}^{(0)})$ ▷ initialize running best solution to initial solution
  for $t \in 1, ..., T$ **do**
    $g(\theta^{(t-1)}) \leftarrow \nabla_\theta \Lambda(\theta^{(t-1)}, \boldsymbol{z}^{(t-1)})$ ▷ subgradient calculation
    $\theta^{(t)} \leftarrow \operatorname{argmin}_{\theta \in \Theta} \left\| \theta - \left(\theta^{(t-1)} - \alpha g(\theta^{(t-1)})\right) \right\|_2$ ▷ subgradient update, projected onto domain Θ
    $\boldsymbol{z}^{(t)} \leftarrow \operatorname{argmax}_{\boldsymbol{z} \in \boldsymbol{Z}} \Lambda(\theta^{(t)}, \boldsymbol{z})$ ▷ adversary, as described in section 4.2
    if $\Lambda(\theta^{(t)}, \boldsymbol{z}^{(t)}) < \Lambda_{\min}$ **then**
      $\hat{\theta} \leftarrow \theta^{(t)}$ ▷ update running best parameters
    end if
  end for
  return $\hat{\theta}$
end procedure

memberships predict the true distributional value, and in turn how much the unlabeled dataset can feasibly differ. Through this, we establish the feasible set of possible group labels $\boldsymbol{Z}$.

We create a set U of statistics calculated over $\boldsymbol{z}'$ and $\boldsymbol{x_z}$, focusing on covariates likely implicated in bias (with the help of expert knowledge). For example, a mean statistic may be the fraction of data points that are in both a certain income bracket and protected group (e.g., race). Calculating $\bar{u}_k(\boldsymbol{x_z}, \boldsymbol{z}')$ gives us a confidence interval on the true distributional expectation μ_k. In turn, we can guarantee with some probability δ a limit on the absolute deviation of the distributional expectation μ_k from empirical average statistic over the task data, $\bar{u}_k(\boldsymbol{x_y}, \boldsymbol{z}^*)$, as some ε. Combining these, we have probability bounds in the form

$$\mathbb{P}\left(\left|\bar{u}(\boldsymbol{x_z}, \boldsymbol{z}') - \bar{u}(\boldsymbol{x_y}, \boldsymbol{z}^*)\right| > \varepsilon\right) \leq \delta \ , \tag{11}$$

and therefore when defining our adversary,

$$\bar{u}(\boldsymbol{x_z}, \boldsymbol{z}') - \varepsilon \leq \bar{u}(\boldsymbol{x_y}, \boldsymbol{z}) \leq \bar{u}(\boldsymbol{x_z}, \boldsymbol{z}') + \varepsilon \tag{12}$$

for all $\boldsymbol{z} \in \boldsymbol{Z}$.

In the simplest approach, we create *unconditional expectation* constraints for fixed probability threshold δ and some ε dependent on δ and other factors. In conventional simple uniform convergence bounds, we use Hoeffding's (1963) inequality for each u, and sum failure probabilities with the union bound. However, especially if we have a small m, as may be the case for auxiliary dataset size m_z, this bound is relatively loose. Hoeffding bounds are not sharp for low-variance functions, and the union bound is likewise weak for correlated variables (Mazzetto et al., 2021). If we calculate the empirical Rademacher average for the family of statistics U over the data, we get a tighter bound. We can further strengthen this by incorporating the wimpy variance, $\sup_{u \in U} \frac{1}{m} \sum u(\boldsymbol{x}_i, \boldsymbol{z}_i)^2$, as well as minimizing this variance by using the empirical centralization of the statistic function; for computational efficiency, this empirically centralized Rademacher average can also be Monte-Carlo estimated (Cousins and Riondato, 2020).

These sharp bounds provide an initial ε_1 and δ_1 dependent *only on the auxiliary data*, and have a direct interpretation as the bounds on the true population mean μ. In contrast, while the distribution of sensitive attribute variables across the task dataset may be unknown, we require no additional assumptions on the size of the task dataset m_y as the uniform convergence bound for PAC learnability relies on the same number — the error ε_2 and associated probability δ_2 on the feasible set here is bounded by the same variables and at the same rate as the overall learnability of the problem. The error bound ε, then, is composed of two separate components $\varepsilon_1, \varepsilon_2$, one of which is parametrized

by the auxiliary data and the other by the task dataset. This allows for the modularized approach of exporting an adversary without the full auxiliary data, with the individual ε_2 portion of the bound depending on a particular task dataset.

5. Discussion

Our work establishes a new technique that recasts the notion of fairness over uncertainty. Motivated by arguments from Black feminist authors and theories of decolonization, we move beyond the flawed basis of imputation and shift power toward holders of sensitive data. Beyond the advantages of this conceptual framework, including increased transparency in data interpretation, our implementation of linear program- and subgradient-based adversarial training provides provable performance and efficiency guarantees. We maximize the efficacy of existing data by allowing the adversary's constraints to come from a variety of statistical sources, including sophisticated, sharp bounds.

Our work holds plenty of room for further development and requires both experimental validation as well as further development in theoretical proofs and connections — bounds on other sources of error, efficient verification of warm-start efficacy, and minimal sizes for group representation. We aim to explore definitive privacy guarantees for data owners and adversary producers, both for the auxiliary dataset, as well as for the implications of projecting adversarial possibilities onto the task dataset. For example, investigating whether adversary constraints and expected statistic values can be made differentially private, to protect from auxiliary database reconstruction attacks, would offer a calculable privacy guarantee.

Nonetheless, our work so far articulates both statistical and theoretical flaws with preexisting proxy methods and other approaches to fairness without direct demographic data. Our technique explicitly addresses these to repair a meaningful notion of Rawlsian fairness, avoid biases within protected groups, and structurally give collectors of sensitive demographic data more range to provide fairness-applicable information to individual task developers, without compromising values of privacy or critical politics.

References

Jacob Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. Active sampling for min-max fairness. *arXiv preprint arXiv:2006.06879*, 2020.

Chidubem Arachie and Bert Huang. Adversarial label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3183–3190, 2019.

Chidubem Arachie and Bert Huang. A general framework for adversarial label learning. *J. Mach. Learn. Res.*, 22:118–1, 2021.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS tutorial*, 1:2, 2017.

Ruha Benjamin. *Race after technology: Abolitionist tools for the New Jim Code*. Polity, Oxford, 2019.

Sebastian Benthall and Bruce D Haynes. Racial categories in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 289–298, 2019.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Consumer Financial Protection Bureau. Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment. *Washington, DC: CFPB, Summer*, 2014.

Patrick J Cantwell, Howard Hogan, and Kathleen M Styles. The use of statistical methods in the US Census: Utah v. Evans. *The American Statistician*, 58(3):203–212, 2004.

A. Charnes and W. W. Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9 (3-4):181–186, 1962.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348, 2019.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Cyrus Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. In *Advances in Neural Information Processing Systems*, 2021.

Cyrus Cousins. Uncertainty and the social planner's problem: Why sample complexity matters. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Cyrus Cousins and Matteo Riondato. Sharp uniform convergence bounds through empirical centralization. In *Advances in Neural Information Processing Systems*, 2020.

Cyrus Cousins, Kavosh Asadi, and Michael L. Littman. Fair E^3: Efficient welfare-centric fair reinforcement learning. In *5th Multidisciplinary Conference on Reinforcement Learning and Decision Making*. RLDM, 2022.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajerdi. Multiaccurate proxies for downstream fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1207–1239, 2022.

Marc N Elliott, Peter A Morrison, Allen Fremont, Daniel F McCaffrey, Philip Pantoja, and Nicole Lurie. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9(2):69–83, 2009.

David A Freedman. Ecological inference and the ecological fallacy. *International Encyclopedia of the social & behavioral sciences*, 6(4027-4030):1–7, 1999.

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 *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

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

Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–13, 2018.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512, 2020.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

bell hooks. *Feminist theory: From margin to center*. Pluto Press, 2000.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

Shunhua Jiang, Zhao Song, Omri Weinstein, and Hengjie Zhang. A faster algorithm for solving general LPs. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 823–832, 2021.

Jeffrey M Jones. LGBT identification in US ticks up to 7.1%. *Gallup News*, 2022.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311, 1984.

Os Keyes. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22, 2018.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33: 728–740, 2020.

Neda Maghbouleh, Ariela Schachter, and René D Flores. Middle Eastern and North African Americans may not be perceived, nor perceive themselves, to be white. *Proceedings of the National Academy of Sciences*, 119(7):e2117940119, 2022.

Alessio Mazzetto, Cyrus Cousins, Dylan Sam, Stephen H Bach, and Eli Upfal. Adversarial multi class learning under weak supervision with performance guarantees. In *International Conference on Machine Learning*, pages 7534–7543. PMLR, 2021.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54 (6):1–35, 2021.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.

John Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Naum Zuselevich Shor. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012.

Te Hiku Media. Indigenous data theft, Aug 2018. URL https://tehiku.nz/te-hiku-tv/haukainga/8037/indigenous-data-theft.

Eve Tuck and K Wayne Yang. R-words: Refusing research. *Humanizing research: Decolonizing qualitative inquiry with youth and communities*, 223:248, 2014.