DOUBLY-REGRESSING APPROACH FOR SUBGROUP FAIRNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Algorithmic fairness is a socially crucial topic in real-world applications of AI. Among many notions of fairness, subgroup fairness is widely studied when multiple sensitive attributes (e.g., gender, race, age) are present. However, as the number of sensitive attributes grows, the number of subgroups increases accordingly, creating heavy computational burdens and data sparsity problem (subgroups with too small sizes). In this paper, we develop a novel learning algorithm for subgroup fairness which resolves these issues by focusing on subgroups with sufficient sample sizes as well as marginal fairness (fairness for each sensitive attribute). To this end, we formalize a notion of subgroup-subset fairness and introduce a corresponding distributional fairness measure called the supremum Integral Probability Metric (supIPM). Building on this formulation, we propose the Doubly Regressing Adversarial learning for subgroup Fairness (DRAF) algorithm, which reduces a surrogate fairness gap for supIPM with much less computation than directly reducing supIPM. Theoretically, we prove that the proposed surrogate fairness gap is an upper bound of supIPM. Empirically, we show that the DRAF algorithm outperforms baseline methods in benchmark datasets, specifically when the number of sensitive attributes is large so that many subgroups are very small.

1 INTRODUCTION

Rapid deployments of AI models in socially consequential domains such as finance, hiring, and criminal justice have amplified the demand for fairness-aware predictions. Early definitions of algorithmic fairness predominantly focused on a single sensitive attribute, such as gender or race, requiring parity across these (marginal) protected groups. However, fairness with respect to a single attribute is not sufficient to protect against discrimination at the intersections of multiple attributes. In particular, the problem of *fairness gerrymandering*, where severe unfairness may remain on their intersections, even if fairness is satisfied on each marginal attribute, has been noticed (Kearns et al., 2018a;b). For instance, while a lending model may equalize approval rates between men and women, the subgroup defined by "female and minority race" may still experience significantly lower approval rates. This illustrates the necessity of *(intersectional) subgroup fairness*.

To state subgroup fairness formally, suppose that the i^{th} individual is specified by its q-dimensional sensitive attribute $s_i \in \{0,1\}^q$, where each coordinate (sensitive attribute) is binary. Then, there are 2^q subgroups, defined by

$$\mathcal{D}_v = \{i : s_i = v\}, \text{ for } v \in \{0,1\}^q.$$

Subgroup fairness requires that the distributions of prediction values be similar (i.e., distributional fairness) across all 2^q subgroups. However, when q is large, we may face two major challenges: (i) *data sparsity*, when certain subgroups contain very few samples, model estimation on such subgroups becomes unstable and inaccurate (Molina & Loiseau, 2023); (ii) *computational burden*, the number of fairness constraints scales exponentially in q.

Various learning algorithms for subgroup fairness have been proposed to resolve the aforementioned two problems (Foulds et al., 2019b; Molina & Loiseau, 2023; Foulds et al., 2019a; Shui et al., 2022; Maheshwari et al., 2023; Hu et al., 2024), but there are still several limitations. Existing algorithms either do not guarantee the marginal fairness (i.e., fairness on each sensitive attribute) which may lead to a socially unacceptable prediction model, or would be computationally demanding when an adversarial learning is required to measure fairness.

The aim of this paper is to develop a learning algorithm for subgroup fairness which resolves data sparsity and computational burden simultaneously. To avoid data sparsity, we simply focus on *active* subgroups, i.e., subgroups whose sample sizes are not too small. Considering only active subgroups is statistically sound since empirical fairness on non-active subgroups does not guarantee the fairness on the population level. A novel part of our proposed learning algorithm is to find a prediction model which achieves (active) subgroup fairness and marginal fairness simultaneously in the context of distributional fairness, the strongest notion of fairness (see Section 3.1 for definition), without resorting on heavy computational burden.

For this purpose, we define a *subgroup-subset* $W \subseteq \{0,1\}^q$ as a union of certain subgroups, and focus on $\mathcal{D}_W = \bigcup_{v \in W} \mathcal{D}_v$. Our approach enforces the distributional fairness over pre-selected subgroup-subsets whose sizes are not small. Then, we design a novel adversarial training strategy termed *doubly regressing adversarial learning* which learns a prediction model without heavy computational burden but guarantees the distributional fairness for all pre-selected subgroup-subsets. The doubly regressing adversarial learning algorithm requires only a single discriminator regardless of the number of pre-selected subgroup-subsets and so computational demand is practically acceptable even when q is large. By including all active subgroups and marginal subgroups (subgroups corresponding to each sensitive attribute) into the set of pre-selected subgroup-subsets, we can effectively achieve subgroup fairness and marginal fairness simultaneously.

The main contributions of this work can be summarized as follows:

1. We formalize *subgroup-subset fairness* and introduce a measure for the distributional subgroup-subset fairness called the supremum Integral Probability Metric (supIPM).
2. We propose a surrogate fairness measure for supIPM which requires only a single discriminator regardless of the number of subgroup-subsets, and develop an adversarial learning algorithm called *Doubly Regressing Adversarial learning for Fairness* (DRAF) algorithm to learn an accurate and subgroup-subset fair prediction model.
3. Theoretically, we prove that the proposed surrogate fairness measure becomes an upper bound of supIPM.
4. Empirically, we show that the DRAF algorithm outperforms baseline methods in benchmark datasets, with large margins when q is large so many subgroups are extremely small.

2 Related Works

Existing methods such as Kearns et al. (2018b); Agarwal et al. (2018) aim to reduce prediction-based subgroup disparities. In particular, Agarwal et al. (2018) proposed a reduction-based approach for marginal fairness by solving a min–max game using Lagrangian multipliers, while Kearns et al. (2018b) extended this framework to minimize the worst-case subgroup disparity. To mitigate data sparsity problem of tiny subgroups, Kearns et al. (2018b;a) employed weights proportional to the sample size of each subgroup.

However, as the number of intersectional subgroups grows, these methods can become computationally expensive. Moreover, they do not explicitly target distributional fairness (e.g., IPM-based criteria), which is the focus of our work. While other approaches focus on other fairness notions, for example, Lai & Guan (2025) designed an adversarial learning-based method for equalized odds, they also differ from our focus on the distributional fairness.

A Bayesian method is proposed to borrow information in large-size subgroups when estimating prediction models for small-sized subgroups (Foulds et al., 2019a). These approaches, however, do not guarantee the marginal fairness (i.e., fairness on each sensitive attribute) which makes it difficult to socially interpret the fairness of a prediction model. On the contrary, (Molina & Loiseau, 2023) consider only the marginal fairness but it could be vulnerable to fairness gerrymandering.

To resolve heavy computational burden, weak notions of fairness such as the DP (Demographic Parity) are employed in the fairness constraint (Kearns et al., 2018b;a) or post-processing techniques are used after learning a prediction model without fairness constraint (Hu et al., 2024). These methods, however, would yield suboptimal prediction models in view of other stronger fairness notions (e.g., distributional fairness) and/or prediction accuracy.

Our approach We propose an in-processing algorithm for distributional fairness on pre-selected subgroup-subsets whose sizes are not too small. We formalize *subgroup-subset fairness* and develop a computationally efficient adversarial algorithm to achieve the distributional fairness.

3 Subgroup-subset fairness

3.1 Problem setting

We consider data points (x_i, y_i, s_i) with input vectors $x_i \in \mathcal{X}$, output variables $y_i \in \mathcal{Y}$, and $s_i = (s_{i1}, \ldots, s_{iq})^\top \in \{0,1\}^q$ denoting the q binary sensitive attributes. Let $\mathbb{P}$ be the probability measure of $(X, Y, S) \in \mathcal{X} \times \mathcal{Y} \times \{0,1\}^q$ and $\mathbb{P}_n$ be its empirical counterpart. Let $\mathcal{F}$ be the set of prediction models $f : \mathcal{X} \times \{0,1\}^q \to \mathbb{R}^c$ for $c \geq 1$. Here, $c = 1$ for regression problems (i.e., $\mathcal{Y} = \mathbb{R}$), while c is the number of classes for classification problems (i.e., $\mathcal{Y} = \{1, \ldots, c\}$). For a given prediction model $f \in \mathcal{F}$ and $s \in \{0,1\}^q$, let $\mathbb{P}_{f,s}$ be the conditional distribution of $f(X)$ given on $S = s$.

We say that f is (perfectly) subgroup-fair if $\mathbb{P}_{f,s}, s \in \{0,1\}^q$ are all the same. To relax the perfect fairness, we define (ψ-distributional) subgroup fairness gap for a given deviance $\psi(\cdot,\cdot)$ between two probability measures as $\Delta_\psi(f) = \sup_{s \in \{0,1\}^q} \psi(\mathbb{P}_{f,s}, \mathbb{P}_{f,\cdot})$, where $\mathbb{P}_{f,\cdot}$ is the marginal distribution of $f(X)$. Then, we say f *is ψ-subgroup fair* with level $\delta > 0$ if $\Delta_\psi(f) \leq \delta$. The main goal of subgroup fair learning is to find an accurate prediction model among ψ-subgroup fair prediction models with level δ.

Various kinds of deviance have been used in fair AI. Examples are (i) the original DP when $\Delta_{\text{DP}}(f) = |\Pr(f(X,s) \geq \tau | S = s) - \Pr(f(X,\cdot) \geq \tau)|$ for a given threshold τ for binary classification (Agarwal et al., 2018), (ii) the mean DP when $\Delta_{\overline{\text{DP}}}(f) := |\mathbb{E}(f(X,s)|S = s) - \mathbb{E}(f(X,\cdot))|$ (Madras et al., 2018; Donini et al., 2018), (iii) the distributional DP when $\psi(\mathbb{P}_{f,s}, \mathbb{P}_{f,\cdot}) = 0$ implies $\mathbb{P}_{f,s} = \mathbb{P}_{f,\cdot}$ (Jiang et al., 2020a; Chzhen et al., 2020b; Silvia et al., 2020; Barata et al., 2021; Kim et al., 2025). Popularly used distributional DPs are Wasserstein distance, Maximum Mean Discrepancy (MMD), Kullback-Leibler divergence, and Kolmogorov-Smirnov distance, to name a few. Among these, distributional DP is the strongest one since it can imply other DPs. In the problem of subgroup fairness, the distributional DP has not been popularly used partly because its computation would be demanding when q is large.

There are large amounts of literature about subgroup fair learning algorithms (Kearns et al., 2018b;a; Úrsula Hébert-Johnson et al., 2018; Foulds et al., 2019b;a; Tian et al., 2025), which learn a prediction model by minimizing the empirical risk (e.g., the residual sum of squares or cross-entropy) subject to the constraint that empirical subgroup fairness gap $\Delta_{n,\psi}(f)$ is less than or equal to δ. Here, empirical subgroup fairness gap $\Delta_{n,\psi}(f)$ is the fairness gap on the empirical distributions.

Existing subgroup fair learning algorithms, however, are not easily applicable to the case of large q due to data sparsity and computational burden. Note that the number of subgroups grows exponentially in q and thus certain subgroups have too small amounts of data and so empirical subgroup fairness gap is not a good estimator of population subgroup fairness gap. With a limited amount of data, there is no hope to be able to guarantee the fairness of a given prediction model on all of subgroups. We could ignore subgroups having too small samples but this naive approach does not ensure the marginal fairness which would not be acceptable. In addition, 2^q many computations of the deviance ψ is required to calculate subgroup fairness gap, and so easy-to-compute ψs (e.g., mean DP) have been used. Furthermore, a subgroup-fair prediction model may not always satisfy the marginal fairness and thus would not be socially acceptable (see an example in Section B.7). Hence, rather than considering all subgroups, we focus only on subgroups whose sizes are sufficiently large and enforce fairness on such large subgroups. To do so, we introduce a new fairness concept called subgroup-subset fairness, in the next subsection. We also provide a more detailed discussion regarding the connection between marginal and subgroup fairness in Section B.7.

3.2 Definition of subgroup-subset fairness

To resolve the data sparsity problem, in this paper, we propose a new notion of subgroup fairness called *subgroup-subset fairness*. The main idea of subgroup-subset fairness is to guarantee fairness on two disjoint subsets of sensitive attributes. To be more specific, we call any subset W of $\{0,1\}^q$ as a *subgroup-subset* and let $\mathbb{P}_{f,W}$ be the distribution of $f(X)$ conditional on $S \in W$ and $\mathbb{P}^n_{f,W}$ be its empirical counterpart. For a given collection $\mathcal{W}$ of subgroup-subsets and a deviance ψ, let

$$\Delta_{\psi,\mathcal{W}}(f) = \sup_{W\in\mathcal{W}} \psi(\mathbb{P}_{f,W}, \mathbb{P}_{f,W^c}),$$

which we call the subgroup-subset fairness gap (with respect to $\mathcal{W}$). Then, we say f is *subgroup-subset fair* with level δ if $\Delta_{\psi,\mathcal{W}}(f) \leq \delta$.

Choice of $\mathcal{W}$ If $\mathcal{W}$ consists of all subgroups, subgroup-subset fairness is equal to subgroup fairness. To resolve data sparsity, we should only include large subgroups in $\mathcal{W}$. In turn, to simultaneously achieve the marginal fairness (i.e., fairness on each sensitive attribute), we add the marginal subgroups (i.e., $W_{j,s} = \{i : s_{ij} = s\}$ for $j \in [q]$ and $s \in \{0,1\}$) to $\mathcal{W}$. In general, we can guarantee fairness for any subgroup-subsets of interest in by adding those subgroup-subsets to $\mathcal{W}$. For example, we can guarantee the second-order marginal fairness (i.e., fairness on $W_{(j,k),(s_1,s_2)} = \{i : (s_{ij}, s_{ik}) = (s_1, s_2)\}$ for $j,k \in [q]$ and $(s_1,s_2) \in \{0,1\}^2$) by adding the corresponding subgroup-subsets. Similarly, we can consider the l^{th}-order marginal fairness for $l \in [q]$. Figure 17 in Section B.7 illustrates the hierarchical structure of subgroup-subsets (from marginal groups to intersectional subgroups).

However, one may worry that computation becomes difficult when $|\mathcal{W}|$ is too large. In Section 4.2, we develop a computationally efficient adversarial learning algorithm for subgroup-subset fairness, where only a single discriminator is used regardless of the size of $\mathcal{W}$. Furthermore, Table 4 in Section B.4 empirically supports the computational scalability of our proposed algorithm.

3.3 SUPREMUM IPM FOR SUBGROUP-SUBSET FAIRNESS GAP

Integral Probability Metric (IPM) In the group fairness problem with a single binary sensitive attribute (i.e., $q = 1$), the integral probability metric (IPM) (Müller, 1997; Sriperumbudur et al., 2009) has been popularly used as the deviation ψ (Chzhen et al., 2020a; Jiang et al., 2020b; Kim et al., 2022; 2025; Kong et al., 2025) to ensure the distributional fairness. For given two probability measures $\mathbb{P}_0$ and $\mathbb{P}_1$, the IPM with a given discriminator class $\mathcal{G} \subset \{g : \mathbb{R}^c \to \mathbb{R}\}$ is defined as

$$\text{IPM}_{\mathcal{G}}(\mathbb{P}_0, \mathbb{P}_1) = \sup_{g\in\mathcal{G}} \left| \int g(x)\mathbb{P}_0(dx) - \int g(x)\mathbb{P}_1(dx) \right|.$$

Various IPMs are obtained by selecting the discriminator class $\mathcal{G}$ accordingly. Popular examples for $\mathcal{G}$ are (i) the collection of 1-Lipschitz functions for Wasserstein distance (Villani, 2009); (ii) the unit ball of an RKHS for MMD (Gretton et al., 2012a); (iii) indicator functions over a VC-bounded family for Total Variation (Shorack, 2000).

Supremum IPM and its statistical property When ψ is $\text{IPM}_{\mathcal{G}}$, we call $\Delta_{\psi,\mathcal{W}}(\cdot)$ as the *supIPM*, and denote the supIPM and its empirical counterpart as $\Delta_{\mathcal{W},\mathcal{G}}(\cdot)$ and $\Delta_{n,\mathcal{W},\mathcal{G}}(\cdot)$, respectively. Theorem 3.1, whose proof is deferred to Section A.2, implies that the estimation error of $\Delta_{n,\mathcal{W},\mathcal{G}}(\cdot)$ does not depend heavily on the size of $\mathcal{W}$ but depends on $n_{\mathcal{W}} = \min_{W\in\mathcal{W}} \min\{n_W, n - n_W\}$, where $n_W = |\{i : s_i \in W\}|$. This result suggests that we can construct $\mathcal{W}$ as large as possible until $n_{\mathcal{W}}$ is sufficiently large.

Let $\mathcal{R}_m(\mathcal{H})$ denotes the empirical Rademacher complexity of a given function class $\mathcal{H}$ with m samples (see Definition A.1 for its detailed definition).

Theorem 3.1. *Let $\mathcal{W}$ be a collection of subgroup-subsets and $n_W := |\{i : s_i \in W\}|$ for $W \in \mathcal{W}$. Assume that $\|g\|_\infty \leq B, \forall g \in \mathcal{G}$. Then, we have for all $f \in \mathcal{F}$ that*

$$\Delta_{n,\mathcal{W},\mathcal{G}}(f) - \Delta_{\mathcal{W},\mathcal{G}}(f) \leq 4\mathcal{R}_{n_{\mathcal{W}}}(\mathcal{G}\circ\mathcal{F}) + 2B\sqrt{\frac{2\log(2n|\mathcal{W}|)}{n_{\mathcal{W}}}}, \tag{1}$$

with probability at least $1 - 1/n$, where $n_{\mathcal{W}} = \min_{W\in\mathcal{W}} \min\{n_W, n - n_W\}$.

In Section A.4, we show that $\mathcal{R}_{n_{\mathcal{W}}}(\mathcal{G}\circ\mathcal{F}) = \mathcal{O}(1/\sqrt{n_{\mathcal{W}}})$ for two cases of $\mathcal{G}$ and $\mathcal{F}$, which indicates that the estimation error of $\Delta_{n,\mathcal{W},\mathcal{G}}(f)$ is $O(\sqrt{\log(|\mathcal{W}|)/n_{\mathcal{W}}})$ ignoring $\log n$ term. This suggests that it would be reasonable to add only subgroup-subsets W with $|W| \geq \gamma n$ into $\mathcal{W}$ for some small $\gamma > 0$. Then, it is guaranteed that the population fairness level locates within the $O(\sqrt{\log(|\mathcal{W}|)/n})$ range of the empirical fairness level. See Section 5.1 how we choose γ in practice.

Challenges in using supIPM for subgroup-subset fairness A technical difficulty, however, exists in using supIPM since computation of supIPM could be very demanding when $|\mathcal{W}|$ is large. To be more specific, for given f and W, let $\hat{g}_{W,f} = \arg\max_{g\in\mathcal{G}} |\int g(z)\mathbb{P}_{f,W}(dz) - \int g(z)\mathbb{P}_{f,W^c}(dz)|$. To calculate supIPM, we should find $\hat{g}_{W,f}$ for all $W \in \mathcal{W}$, which is computationally demanding when $|\mathcal{W}|$ is large. We could avoid this problem by using the IPM which admits a closed-form calculation. An example is the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012b). However, computational cost of calculating supMMD is $O(|\mathcal{W}|n^2)$, which is still large when $|\mathcal{W}|$ and/or n is large. In addition, the choice of the kernel would not be easy.

In the following section, we propose a novel surrogate subgroup-subset fairness gap of supIPM which serves as an upper bound of supIPM and requires only a single discriminator to be computed.

4 Doubly Regressing Algorithm

4.1 A surrogate deviance for IPM

Fix $W \in \mathcal{W}$, and let $y_{W,i} := 2\mathbb{I}(s_i \in W) - 1$ be the indicator whether i^{th} observation belongs to W or not. To assess the fairness of a given prediction model f on W, a standard method is to investigate the error rate of a classifier learned with $f_i := f(x_i, s_i)$ being input and $y_{W,i}$ being the label, which is used for fair adversarial learning (Edwards & Storkey, 2016; Madras et al., 2018). That is, we look at $\sup_{g\in\mathcal{G}} \sum_{i=1}^n \mathbb{I}(y_{W,i} \times g(f_i) < 0)$. If f is fair on W, the distribution of f on W and W^c are similar and thus the misclassification error becomes large.

Instead of the misclassfication error, we could consider the Residual Sum of Squares (RSS) $\sup_{g\in\mathcal{G}} \sum_{i=1}^n (y_{W,i} - g(f_i))^2$ as the fairness measure. The RSS is mathematically more tractable than the misclassification error since the former is differentiable but the latter is not. This mathematical tractability plays an important role when we extend a surrogate measure of IPM for supIPM. A larger RSS implies a fairer f. An equivalent measure would be supSSR $:= \sup_{g\in\mathcal{G}} \left\{\sum_{i=1}^n (y_{W,i} - \bar{y}_W)^2 - \sum_{i=1}^n (y_{W,i} - g(f_i))^2\right\}$, which is an analogy of the Sum of Squares of Regression (SSR) used in the regression analysis. This measure becomes small when f is fair.

A related measure of supSSR is $\sup_{g\in\mathcal{G}} R^2(f, W, g)$, where

$$R^2(f, W, g) = 1 - \frac{\sum_{i=1}^n (y_{W,i} - g(f_i))^2}{\sum_{i=1}^n (y_{W,i} - \bar{y}_W)^2} = \frac{\sum_{i=1}^n (y_{W,i} - \bar{y}_W)^2 - \sum_{i=1}^n (y_{W,i} - g(f_i))^2}{\sum_{i=1}^n (y_{W,i} - \bar{y}_W)^2}, \quad (2)$$

which is an analogy of R^2 in the regression analysis. This measure becomes small when f is fair. A surprising result is that a slight modification of (2) is equal to IPM, which is stated in the following theorem. Refer to Section A.2 for its proof.

Theorem 4.1. *For given $f \in \mathcal{F}, W \subseteq \{0,1\}^q$ and $\mathcal{G}$, we have*

$$\text{IPM}_{\mathcal{G}}(\mathbb{P}_{f,W}, \mathbb{P}_{f,W^c}) = \sup_{g\in\mathcal{G}} |\tilde{R}^2(f, W, g)|,$$

where $\tilde{R}^2(f, W, g) = R^2(f, W, g) + \frac{\sum_{i=1}^n (g(f_i) - \bar{y}_W)^2}{\sum_{i=1}^n (y_{W,i} - \bar{y}_W)^2}$.

Suppose that $\mathcal{G}_{\text{obs}} = \{(g(x_1), \ldots, g(x_n))^\top : g \in \mathcal{G}\}$ is a linear space. Then $\hat{g} := \arg\min_{g\in\mathcal{G}} R^2(f, W, g)$ becomes the projection of $\{y_{W,i}\}$ onto $\mathcal{G}_{\text{obs}}$ and thus it can be shown that $R^2(f, W, \hat{g})$ is the squared correlation of $\{y_{W,i}\}$ and $\{\hat{g}_i\}$ and $\tilde{R}^2(f, W, \hat{g}) = 2R^2(f, W, \hat{g})$. In fact, the additional term in $\tilde{R}^2(f, W, g)$ is introduced for $\mathcal{G}_{\text{obs}}$ not being a linear space. An interesting new implication of Theorem 4.1 is that IPM is somehow related to the correlation between the class label and a discriminator.

4.2 A surrogate deviance for supIPM: Doubly Regressing R^2

Theorem 4.1 implies $\Delta_{n,\mathcal{W},\mathcal{G}}(f) = \sup_{W\in\mathcal{W}} \sup_{g\in\mathcal{G}} \tilde{R}^2(f, W, g)$, which is not easy to calculate since W is not a numerical quantity and so a gradient ascent algorithm cannot be applied. To resolve this computational problem, we introduce a smoother version of $\tilde{R}^2(f, W, g)$ so-called the *Doubly Regressing* R^2 (DR^2) as follows.

Suppose that $\mathcal{W} = \{W_1, \ldots, W_M\}$. For each $i \in [n]$, define $c_i \in \{-1, 1\}^M$ with $c_{im} = 2\mathbb{I}(s_i \in W_m) - 1$. Given a predictor f, discriminator g, and weight vector $\mathbf{v} \in \mathcal{S}^M$, we define

$$\mathrm{DR}^2(f, \mathbf{v}, g) := 1 - \frac{\left\{\sum_{i=1}^n (\mathbf{v}^\top c_i - g(f_i))^2 - \sum_{i=1}^n (g(f_i) - \mu_{\mathbf{v}})^2\right\}}{\sum_{i=1}^n (\mathbf{v}^\top c_i - \mu_{\mathbf{v}})^2},$$

where $\mu_{\mathbf{v}} = \frac{1}{n}\sum_{i=1}^n \mathbf{v}^\top c_i$ and $\mathcal{S}^M$ is the unit sphere on $\mathbb{R}^M$. The name 'Doubly Regressing' is used since we regress input $g(f_i)$ and output $\mathbf{v}^\top c_i$ simultaneously when calculating DR^2.

Note that $\mathrm{DR}^2(f, \mathbf{v}, g)$ is equal to $\tilde{R}^2(f, W_m, g)$ when $\mathbf{v} = \mathbf{e}_k$, where $\mathbf{e}_k$ is the vector whose entries are all 0 except the k^{th} entry being 1. Thus, it is obvious that the supIPM is bounded as:

$$\sup_{W \in \mathcal{W}} \mathrm{IPM}_{\mathcal{G}}(\mathbb{P}^n_{f,W}, \mathbb{P}^n_{f,W^c}) = \Delta_{n,\mathcal{W},\mathcal{G}}(f) \leq \sup_{g \in \mathcal{G}, \mathbf{v} \in \mathcal{S}^M} |\mathrm{DR}^2(f, \mathbf{v}, g)|. \quad (3)$$

Building on this inequality, our proposed surrogate subgroup-subset fairness gap for supIPM is $\mathrm{DR}_{n,\mathcal{W},\mathcal{G}}(f) := \sup_{g \in \mathcal{G}, \mathbf{v} \in \mathcal{S}^M} z\text{-}\mathrm{DR}^2(f, \mathbf{v}, g)$ where

$$z\text{-}\mathrm{DR}^2(f, \mathbf{v}, g) = \log\left(\frac{1 + |\mathrm{DR}^2(f, \mathbf{v}, g)|/2}{1 - |\mathrm{DR}^2(f, \mathbf{v}, g)|/2}\right). \quad (4)$$

We apply the Fisher's z-transformation to $|\mathrm{DR}^2|/2$ for numerical stability. We refer to $\mathrm{DR}_{n,\mathcal{W},\mathcal{G}}(f)$ as the *Doubly Regressing (DR)* subgroup-subset fairness gap. A smaller value of $\mathrm{DR}_{n,\mathcal{W},\mathcal{G}}(f)$ indicates a higher level of subgroup-subset fairness of f.

4.3 Algorithm: Doubly Regressing Adversarial learning for Fairness (DRAF)

Based on the DR gap, we introduce *DRAF (Doubly Regressing Adversarial learning for Fairness)* algorithm, which trains f by minimizing $\frac{1}{n}\sum_{i=1}^n l(y_i, f(x_i, s_i)) + \lambda \mathrm{DR}_{n,\mathcal{W},\mathcal{G}}(f)$, for a given loss l (e.g., cross-entropy) and Lagrangian multiplier λ. A key feature is that a single discriminator is used regardless of $\mathcal{W}$.

In the learning algorithm, we iteratively train the prediction model f and the pair of discriminator g and weight vector $\mathbf{v}$ iteratively. At each iteration, we (i) update f by applying a gradient descent algorithm to minimize $\frac{1}{n}\sum_{i=1}^n l(y_i, f(x_i, s_i)) + \lambda z\text{-}\mathrm{DR}^2(f, \mathbf{v}, g)$ while g and $\mathbf{v}$ are fixed, and then (ii) update g and $\mathbf{v}$ by applying a gradient ascent algorithm to maximize $z\text{-}\mathrm{DR}^2(f, \mathbf{v}, g)$ while f being fixed. Algorithm 1 in Section B.2 below provides the outline of our proposed algorithm.

5 Experiments

In this section, we empirically verify that DRAF can successfully achieve both marginal and subgroup fairness: (i) it shows competitive performance to baseline methods for datasets with less sparse subgroups; (ii) it outperforms baselines for datasets with extremely sparse subgroups. After that, we conduct analyses on the effect of managing $\mathcal{W}$ and the choice of discriminator.

5.1 Settings

Datasets We consider the following four benchmark datasets (three tabular datasets and a text dataset) popularly used in algorithmic fairness research. See Section B.1 for more details.

Adult (Tabular) (Becker & Kohavi, 1996): The class label is binary indicating the income above 50k$. The input features are several demographic census features. For the sensitive attributes, we consider sex, race, age, and marital-status, so that $q = 4$.

Dutch (Tabular) (van der Laan, 2000): The class label is binary indicating occupation level. The input features are several socio-economic features. For the sensitive attributes, we consider sex and age, so that $q = 2$.

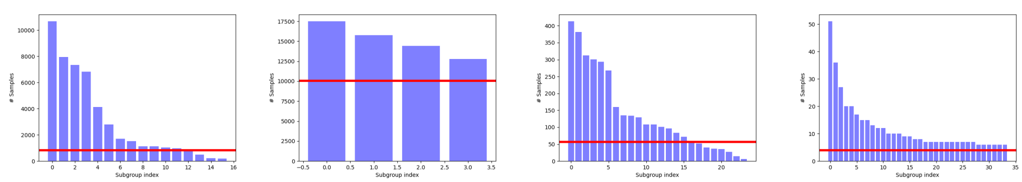
CivilComments (Text) (Borkan et al., 2019): The class label is binary, indicating comment toxicity. The input features are representations extracted from the pre-trained DistilBERT model (Sanh et al., 2019). For the sensitive attributes, we consider sex (male/female/other), race (black/white/asian/other), and religion (christian/other), which are non-binary, resulting in 24 subgroups.

COMMUNITIES (Tabular) (Redmond & Baveja, 2002): The class label is binary, indicating whether the violent crime rate is above a threshold. The input features are 122 community-level attributes covering demographics and economic indicators. For the sensitive attributes, we consider race, racial per-capita, and language/immigration variables so that $q = 18$.

Table 1 summarizes the basic statistics of the four datasets and Figure 1 presents the distribution of subgroup sizes for the datasets. These statistics highlight the severity of data sparsity: in particular, COMMUNITIES suffers from extreme sparsity with the vast majority of subgroups contain very few samples. We construct a 60/20/20 split for train, validation, and test, respectively for the datasets except COMMUNITIES. Due to the extreme sparsity of certain subgroups in COMMUNITIES dataset, ensuring sufficient samples within the test set would be important, so we use with 50/10/40 ratios. We repeat this procedure five times randomly and report the average performance.

Table 1: Summary of datasets. "# Subgroup" indicates the possible maximum number of subgroups ($= 2^q$). "# Actual Subgroup" indicates the actual number of subgroups in the datasets. "# Sparse subgroup" indicates the number of subgroups whose size is at most 1% of the total sample size n.

Dataset	n	q	# Subgroup	# Actual Subgroup	# Sparse subgroup
ADULT	48,842	4	16	16	2
DUTCH	60,420	2	4	4	0
CIVILCOMMENTS	3,365	3 (non-binary)	24	24	3
COMMUNITIES	1,994	18	262,144	1,180	1,175



(a) ADULT (b) DUTCH (c) CIVILCOMMENTS (d) COMMUNITIES

Figure 1: Histograms of subgroup sizes. The red horizontal line indicates γn used for the main analysis in Section 5.3. The subgroup indices are assigned by sorting the subgroup sizes.

Model and Performance measures Since the four datasets are for binary classification tasks, we only consider one dimensional prediction model f for which we consider a single-layered MLP and apply the sigmoid activation at the output layer to return the prediction score between $[0, 1]$. Recall that $f_i = f(x_i, s_i)$ and let $\hat{y}_i = 2\mathbb{I}(f_i \geq 1/2) - 1$. We consider the accuracy $\text{Acc}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i = \hat{y}_i)$ for prediction performance of f.

For fairness performance, we consider the l^{th}-order marginal fairness and subgroup level fairness. For distributional fairness, we use the Wasserstein distance, but only for the first-order marginal subgroup only, as calculating it for higher-orders would be unstable due to the lack of samples. To be more specific, let $\hat{p} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\hat{y}_i = 1)$ and $\hat{p}_s := \frac{1}{n_s}\sum_{i:s_i=s} \mathbb{I}(\hat{y}_i = 1), s \in \{0,1\}^q$ be the overall and subgroup-specific ratios of positive prediction, respectively. For a given order $l \in [q]$, consider $L \subseteq [q]$ such that $|L| = l$. Let $s_i[L] := (s_{ij})_{j \in L} \in \{0,1\}^l$ be the sensitive attribute subvector of the i^{th} individual. For a given $a \in \{0,1\}^l$, define $\hat{p}_L^{(a)} := \frac{1}{n_{L,a}}\sum_{i:s_i[L]=a} \mathbb{I}(\hat{y}_i = 1)$, where $n_L^{(a)} := \sum_{i=1}^{n} \mathbb{I}(s_i[L] = a)$. Let $\widehat{\mathbb{P}}_f(\cdot) := \frac{1}{n}\sum_{i=1}^{n} \delta_{f_i}(\cdot)$. For a given $j \in [q]$, define $\widehat{\mathbb{P}}_{f,j|a}(\cdot) := \frac{1}{n_j^{(a)}}\sum_{i:s_{ij}=a} \delta_{f_i}(\cdot)$, where $n_j^{(a)} := \sum_{i=1}^{n} \mathbb{I}(s_{ij} = a)$ for $a \in \{0,1\}$. Table 2 describes the fairness performance measures used in the experiments.

Implementation details and Baseline methods We sweep the Lagrangian multiplier λ from 0.01 to 10.0 to control the fairness level. For the discriminator $\mathcal{G}$, we use the discriminator class used in sIPM (Kim et al., 2022) (i.e., composition of sigmoid and a linear function). For $\mathcal{W}$, we include

Table 2: Fairness performance measures used in our experiments. MP, WMP, and SP are abbreviations of Marginal, Wasserstein Marginal, and Subgroup Parity, respectively. 'W_1' indicates the 1-Wasserstein distance between two probability measures on $\mathbb{R}$.

Name	Meaning	Formula
$\text{MP}^{(l)}$	l^{th}-order Marginal fairness	$\max_{L\subseteq[q],\lvert L\rvert=l}\sum_{a\in\{0,1\}^l}\frac{n_L^{(a)}}{n}\left\lvert\hat{p}_L^{(a)}-\hat{p}\right\rvert$
WMP	Distributional Marginal fairness	$\max_{j\in[q]}\max\left\{\frac{n_j^{(0)}}{n}\mathrm{W}_1(\widehat{\mathbb{P}}_{f,j\lvert 0},\widehat{\mathbb{P}}_f),\frac{n_j^{(1)}}{n}\mathrm{W}_1(\widehat{\mathbb{P}}_{f,j\lvert 1},\widehat{\mathbb{P}}_f)\right\}$
SP	Subgroup fairness	$\max_{s\in\{0,1\}^q}\frac{n_s}{n}\left\lvert\hat{p}_s-\hat{p}\right\rvert$

the first and second-order marginal subgroups as well as subgroups whose sizes are larger than γn. To find an optimal value of γ, we plot Pareto-front lines (for many γs) between Acc and SP using validation data, compute the area under the lines, and then choose the one with the largest area. As a result, we set γ to $0.01, 0.001, 0.3$, and 0.01 for ADULT, CIVILCOMMENTS, DUTCH, and COMMUNITIES, respectively. We consider four representative approaches as baselines: (i) Regularization (REG): this approach reduces the marginal disparities for q-many sensitive attributes; (ii) GerryFair (GF) (Kearns et al., 2018a;b): this approach reduces the (weighted) worst-case disparity $\max_{s\in\{0,1\}^q}\frac{n_s}{n}\left\lvert\hat{p}_s-\hat{p}\right\rvert$; (iii) Sequential (SEQ) (Hu et al., 2024): this approach sequentially maps the scores of a pre-trained fairness-agnostic model in each subgroup to a common barycenter; See Section B.2 for more details.

5.2 RELATIONSHIP BETWEEN DR GAP AND SUPIPM

As theoretically shown in Theorem 4.1 and Eq. (3), the DR gap (i.e., $\text{DR}_{n,\mathcal{W},\mathcal{G}}(f)$) and the supIPM (i.e., $\Delta_{n,\mathcal{W},\mathcal{G}}(f)$) are closely related, i.e., small DR gap $\implies$ small supIPM. To numerically confirm this, we provide plots between the DR gap and the supIPM in Figure 2, indicating that the DR gap is also a numerically valid surrogate quantity for supIPM (i.e., reducing DR results in reducing supIPM). Note that we use the Wasserstein distance for supIPM.

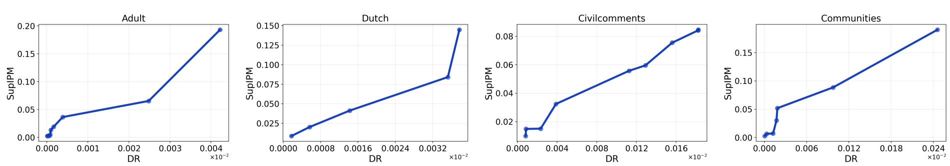


Figure 2: Empirical relationship between the DR gap and supIPM on ADULT, DUTCH, CIVILCOMMENTS, and COMMUNITIES datasets.

Moreover, Figure 5 in Section B.3 empirically shows that the DR gap and the supIPM are almost identical: after learning, the weight vector $\mathbf{v}$ becomes nearly a vertex of the simplex, so that $\text{DR}^2(f,\mathbf{v},g)\approx\tilde{R}^2(f,W_m,g)$ in practice. This finding empirically supports the claim that the DR gap is a valid surrogate for the supIPM.

5.3 PERFORMANCE COMPARISON

Trade-off between accuracy and fairness Figure 3 compares the trade-off between fairness levels (SP and $\text{MP}^{(1)}$) and accuracies of the five methods - DRAF, three baselines and unfair prediction model. Since the fairness level is not controllable for SEQ and unfair prediction model, their results are given as points instead of lines. Figure 6 in Section B.4 presents similar results for other fairness measures (WMP and $\text{MP}^{(2)}$). The main findings can be summarized as follows.

- Datasets with less sparse subgroups (ADULT, DUTCH and CIVILCOMMENTS): For ADULT and DUTCH, the three methods REG, GF, and DRAF perform similarly on both first-order marginal and subgroup fairness. Note that the slight better performance of REG on ADULT is due to a training-test data discrepancy: we observe that the three methods perform nearly the same on the training data. Specifically for CIVILCOMMENTS, REG underperforms GF and DRAF for SP, while GF slightly underperforms DRAF at small $\text{MP}^{(1)}$. These results recommend using DRAF

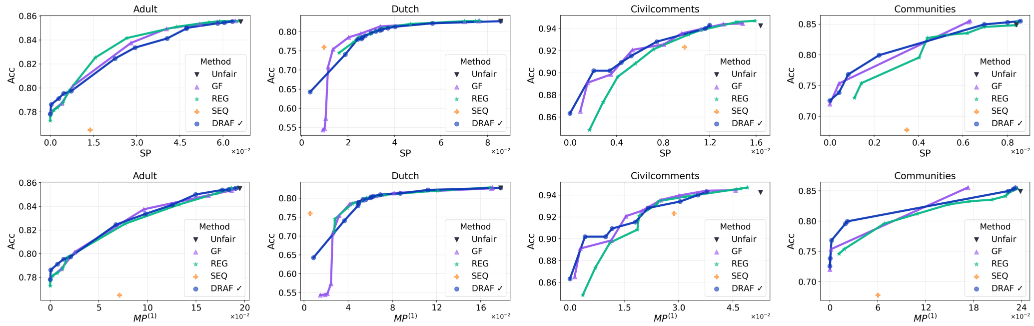


Figure 3: Trade-off between fairness level (top: SP, bottom: $\mathrm{MP}^{(1)}$) and accuracy.

for achieving both subgroup and first-order marginal fairness, even on datasets with less sparse subgroups. Similar results on an additional dataset (ACSINCOME (Ding et al., 2022)) are shown in Figure 8 in Section B.4.

- Datasets with sparse subgroups (COMMUNITIES): DRAF outperforms REG on both first-order marginal and subgroup fairness, and GF on first-order marginal fairness. These results suggest that reducing only first-order marginal fairness (REG) or only subgroup fairness (GF) would be suboptimal, and so DRAF is particularly effective when subgroups are sparse. See Table 3 in Section B.4 for similar results on a subsampled ADULT dataset with sparse subgroups.

Correlation between subgroup fairness and first-order marginal fairness We analyze the correlation between subgroup fairness and first-order marginal fairness, to investigate how a given algorithm can simultaneously control the both well. Figure 4 plots subgroup fairness (SP) versus first-order marginal fairness ($\mathrm{MP}^{(1)}$) for DRAF, GF, and REG. To quantify their correlation, we fit a linear regression and calculate the SSE (Sum of Squared Errors). The results show that the SSE for GF and REG is larger than that for DRAF in most cases, with a large margin for COMMUNITIES. It suggests that focusing solely on subgroup fairness (GF) or first-order marginal fairness (REG) does not guarantee the other, whereas DRAF can achieve both regardless of the sparsity. This highlights the benefit of DRAF: subgroup and first-order marginal fairness tend to behave together, so we can control both with a single λ without unexpected unfairness.

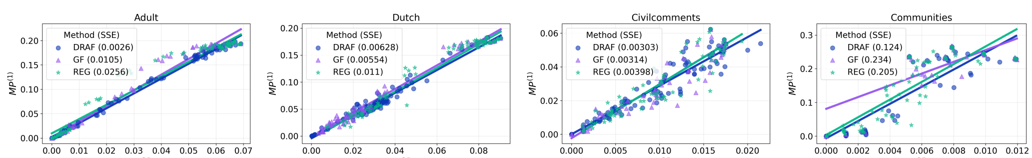


Figure 4: Scatter plots between SP and $\mathrm{MP}^{(1)}$ with linear regression lines, and SSE on ADULT, DUTCH, CIVILCOMMENTS, and COMMUNITIES datasets.

5.4 EXTENSIONS OF DRAF

Multi-class classification We empirically discuss that DRAF is not limited to binary classification but can be extended to multi-class classification problem. Following Denis et al. (2024), we use COMMUNITIES dataset with five label classes and apply DRAF to the dataset. As shown in Table 5 in Section B.5, DRAF achieves competitive performance compared to the baseline method of Denis et al. (2024).

Equalized Odds (EO) We also show that DRAF can be extended to EO, by a slight modification in $\mathrm{DR}^2(f, \mathbf{v}, g)$. We compare this modified DRAF for EO (DRAF-EO) with FairICP (Lai & Guan, 2025), which is an adversarial learning-based method for EO. Tables 7 and 8 in Section B.5 suggest that the DRAF-EO performs competitive to FairICP, in terms of both marginal and subgroup fairness.

5.5 Additional studies

Excluding the marginal subgroups from $\mathcal{W}$ We investigate how the marginal fairness is affected when $\mathcal{W}$ excludes the marginal subgroups. First, Figure 9 in Section B.6 shows that excluding first-order marginal subgroups could harm first-order marginal fairness even if subgroup fairness is satisfied. This result emphasizes the need to include the marginal subgroups in $\mathcal{W}$, to obtain socially acceptable subgroup fair models (i.e., as well as marginally fair). Similarly, we consider $\mathcal{W}$ without the second-order marginal subgroups. Figure 11 in Section B.6 shows that the second-order marginal fairness can be slightly worsen under such exclusion. Hence, we recommend including the second-order marginal subgroups in $\mathcal{W}$ as well, unless the optimization is numerical unstable.

Impact of γ Another simple way to manage $\mathcal{W}$ is to control the minimum sample size of $W \in \mathcal{W}$ (i.e., γ). As γ increases, the sizes of subgroup-subsets become larger, hence $\mathcal{W}$ excludes higher-order marginal subgroups as well as more subgroups. Suppose that we choose γ to be larger than the sizes of higher-order marginal subgroups but smaller than those of first-order marginal subgroups. Such a choice would achieve marginal fairness, but it may hamper higher-order marginal fairness. Section B.6 empirically supports the claim by comparing performance with various γs: Figures 12 and 13 show that a too large γ could degrade second-order marginal as well as subgroup fairness.

We further analyze the effect of γ by categorizing subgroups into large, medium, and small scales (using quantiles 0.3 and 0.6). Figure 14 in Section B.6 implies that (i) while large subgroups remain consistently fair, (ii) medium subgroups require a moderate γ to maintain fairness, (iii) and the fairness of small subgroups is not guaranteed even γ is small. Note that these results also support the implications of Theorem 3.1.

In conclusion, since guaranteeing fairness on tiny subgroups is difficult, and we advise selecting a moderate γ. This allows the model to focus on enforcing fairness for subgroups capable of generalization (i.e., those where fairness on the training data implies fairness on the test data).

Choice of $\mathcal{G}$ In the experiments, we consider the discriminator used in sIPM, which is used for fair representation learning (Kim et al., 2022). We also consider sIPM with ReLU IPM (RIPM, Park et al. (2025)) where discriminator functions are a composition of ReLU and linear functions, and Hölder IPM (HIPM, Wang et al. (2023)) which uses DNN discriminators. Figure 15 in Section B.6 shows that sIPM is generally the best and most stable.

Robustness under noisy sensitive attributes To investigate the robustness of DRAF under noisy sensitive attribute information, we conduct a controlled experiment on the Adult dataset by synthetically introducing missing values into the sensitive attribute at a certain rate (e.g., 1%). Existing methods (e.g., GF) typically require complete subgroup information and therefore discard samples with missing sensitive attributes. In contrast, DRAF can still be applied partially to such samples: for instances with missing sensitive attributes, we impose fairness constraints only on subgroup-subsets that can be formed using the observed attributes, and ignore subgroup-subsets that involve any missing attributes.

The results in Figure 16 in Section B.6 show that this partial DRAF approach is more robust than the baseline: it attains a better fairness–accuracy trade-off than GF, demonstrating the robustness of DRAF in the presence of noisy (missing) sensitive attributes.

6 Concluding Remarks

In this paper, we introduced a new notion of fairness called subgroup-subset fairness, and proposed a new adversarial learning algorithm for subgroup fairness. We empirically showed that the proposed algorithm works well in scenarios where the data contain sparse subgroups.

A possible future work is to decompose subgroup fairness into low-order marginal fairness (similar to ANOVA decomposition) and control fairness via these components. This approach would improve stability under sparse subgroups and interpretability. One could theoretically derive an upper bound of subgroup fairness in terms of low-order marginal fairness.

Ethics Statement We do not collect new human-subject datasets; all the datasets used in this paper are publicly available. The fairness notions we employ in this paper (i.e., subgroup fairness and marginal group fairness) are widely and popularly investigated in recent literature. Through these efforts, we believe this research helps mitigate potential discriminatory impacts, rather than introduce new ones, and can positively influence the responsible use of AI in practice.

Reproducibility Statement We made efforts to ensure the reproducibility of our main findings: (i) We provide full proofs and mathematical definitions used in the theorems in Appendix. (ii) We include implementation details throughout the paper (the main body and Appendix), and add source code in the supplementary materials.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

António Pereira Barata, Frank W. Takes, H. Jaap van den Herik, and Cor J. Veenman. Fair tree classifier using strong demographic parity, 2021.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7321–7331. Curran Associates, Inc., 2020a. URL `https://proceedings.neurips.cc/paper/2020/file/51cdbd2611e844ece5d80878eb770436-Paper.pdf`.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33: 7321–7331, 2020b.

Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. Fairness guarantees in multi-class classification with demographic parity. *J. Mach. Learn. Res.*, 25(1), January 2024. ISSN 1532-4435.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning, 2022. URL `https://arxiv.org/abs/2108.04884`.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference in Learning Representations (ICLR2016)*, pp. 1–14, May 2016. URL `https://iclr.cc/archive/www/doku.php%3Fid=iclr2016:main.html`. 4th International Conference on Learning Representations, ICLR 2016 ; Conference date: 02-05-2016 Through 04-05-2016.

James Foulds, Rashidul Islam, Kamrun Keya, and Shimei Pan. Bayesian modeling of intersectional fairness: The variance of bias, 2019a. URL `https://arxiv.org/abs/1811.07255`.

James Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness, 2019b. URL `https://arxiv.org/abs/1807.08362`.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012a. URL http://jmlr.org/papers/v13/gretton12a.html.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012b.

Francois Hu, Philipp Ratz, and Arthur Charpentier. A sequentially fair mechanism for multiple sensitive attributes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (11):12502–12510, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i11.29143. URL http://dx.doi.org/10.1609/aaai.v38i11.29143.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in artificial intelligence*, pp. 862–872. PMLR, 2020a.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 862–872. PMLR, 22–25 Jul 2020b. URL https://proceedings.mlr.press/v115/jiang20a.html.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning, 2018a. URL https://arxiv.org/abs/1808.08166.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018b.

Dongha Kim, Kunwoong Kim, Insung Kong, Ilsang Ohn, and Yongdai Kim. Learning fair representation with a parametric integral probability metric. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11074–11101. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/kim22b.html.

Kunwoong Kim, Insung Kong, Jongjin Lee, Minwoo Chae, Sangchul Park, and Yongdai Kim. Fairness through matching. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=dHljjaNHh1.

Insung Kong, Kunwoong Kim, and Yongdai Kim. Fair representation learning for continuous sensitive attributes using expectation of integral probability metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yuheng Lai and Leying Guan. FairICP: Encouraging equalized odds via inverse conditional permutation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32228–32245. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/lai25b.html.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

Gaurav Maheshwari, Aurélien Bellet, Pascal Denis, and Mikaela Keller. Fair without leveling down: A new intersectional fairness definition. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=RubWYFBZbG.

Mathieu Molina and Patrick Loiseau. Bounding and approximating intersectional fairness through marginal fairness, 2023. URL https://arxiv.org/abs/2206.05828.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. doi: 10.2307/1428011.

Yuha Park, Kunwoong Kim, Insung Kong, and Yongdai Kim. Relu integral probability metric and its applications, 2025. URL https://arxiv.org/abs/2504.18897.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3): 660–678, 2002.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Galen R. Shorack. *Probability for Statisticians*. Springer-Verlag, New York, 2000. ISBN 978-0-387-98945-6. doi: 10.1007/b98945.

Changjian Shui, Gezheng Xu, Qi CHEN, Jiaqi Li, Charles Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=YsRH6uVcx2l.

Chiappa Silvia, Jiang Ray, Stepleton Tom, Pacchiano Aldo, Jiang Heinrich, and Aslanides John. A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3633–3640, 2020.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On integral probability metrics, ϕ-divergences and binary classification, 2009. URL https://arxiv.org/abs/0901.2698.

Huan Tian, Bo Liu, Tianqing Zhu, Wanlei Zhou, and Philip S. Yu. Multifair: Model fairness with multiple sensitive attributes. *IEEE Transactions on Neural Networks and Learning Systems*, 36 (3):5654–5667, 2025. doi: 10.1109/TNNLS.2024.3384181.

Paul van der Laan. Integrating administrative registers and household surveys. *Netherlands Official Statistics*, 15(2):7–15, 2000.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9.

Jie Wang, Minshuo Chen, Tuo Zhao, Wenjing Liao, and Yao Xie. A manifold two-sample test study: integral probability metric with neural networks. *Information and Inference: A Journal of the IMA*, 12(3):1867–1897, 06 2023. ISSN 2049-8772. doi: 10.1093/imaiai/iaad018. URL https://doi.org/10.1093/imaiai/iaad018.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Calibration for the (computationally-identifiable) masses, 2018. URL https://arxiv.org/abs/1711.08513.

APPENDIX

A THEORETICAL STUDIES

A.1 OMITTED DEFINITIONS AND NOTATIONS

Definition A.1. Let $(\sigma_i)_{i=1}^m$ be the Rademacher random variables such that $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$, independently. Let $\mathcal{H}$ be a class of real-valued functions on a domain $\mathcal{Z}$, and let $S = (z_1, \dots, z_m) \in \mathcal{Z}^m$ be a fixed sample. The empirical Rademacher complexity of $\mathcal{H}$ on S is

$$\widehat{\mathcal{R}}_S(\mathcal{H}) := \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i \, h(z_i) \right], \tag{5}$$

where the expectation is with respect to the Rademacher variables $(\sigma_i)_{i=1}^m$.

Given a distribution P on $\mathcal{Z}$, the *Rademacher complexity* of $\mathcal{H}$ with sample size m is

$$\mathcal{R}_m(\mathcal{H}) := \mathbb{E}_{S \sim P^m} \left[\widehat{\mathcal{R}}_S(\mathcal{H}) \right] = \mathbb{E}_{z_1, \dots, z_m \sim P} \, \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i \, h(z_i) \right]. \tag{6}$$

A.2 PROOFS

Proof of Theorem 3.1. Fix $f \in \mathcal{F}$. By the definition of the supremum and the triangle inequality of IPMs,

$$\begin{aligned} \Delta_{\psi, \mathcal{W}}(f) - \Delta_{n, \psi, \mathcal{W}}(f) &\leq \sup_{W \in \mathcal{W}} \psi(\mathbb{P}_{f,W}, \mathbb{P}^n_{f,W}) - \sup_{W \in \mathcal{W}} \psi(\mathbb{P}_{f,W^c}, \mathbb{P}^n_{f,W^c}) \\ &\leq \sup_{W \in \mathcal{W}} \Big\{ \psi(\mathbb{P}_{f,W}, \mathbb{P}^n_{f,W}) + \psi(\mathbb{P}_{f,W^c}, \mathbb{P}^n_{f,W^c}) \Big\}. \end{aligned} \tag{7}$$

The first term in the right-hand-side can be re-written as

$$\begin{aligned} \psi(\mathbb{P}_{f,W}, \mathbb{P}^n_{f,W}) &= \sup_{g \in \mathcal{G}} \Big(\mathbb{E}_{\mathbb{P}_{f,W}}[g] - \mathbb{E}_{\mathbb{P}^n_{f,W}}[g] \Big) \\ &= \sup_{g \in \mathcal{G}} \left(\mathbb{E}\left[g \circ f(X, S) | S \in W \right] - \frac{1}{n_W} \sum_{i : s_i \in W} g \circ f(x_i, s_i) \right), \end{aligned} \tag{8}$$

where $n_W = |\{i : s_i \in W\}|$. Taking the supremum over $f \in \mathcal{F}$ and by Hoeffding's inequality combined with Rademacher symmetrization, we have with probability at least $1 - \delta_W$,

$$\sup_{f \in \mathcal{F}} \psi(\mathbb{P}_{f,W}, \mathbb{P}^n_{f,W}) \leq 2\mathcal{R}_{n_W}(\mathcal{G} \circ \mathcal{F}) + B \sqrt{\frac{2 \log(1/\delta_W)}{n_W}}$$

for any $\delta_W > 0$. An exactly same bound holds for W^c with n_{W^c} in place of n_W. Applying the union bound over all pairs $\{W, W^c\}$ with $\delta_W = \delta/(2|\mathcal{W}|)$ and using $n_W, n_{W^c} \geq n_{\mathcal{W}} = \min_{W \in \mathcal{W}} \{n_W, n_{W^c}\} = \min_{W \in \mathcal{W}} \{n_W, n - n_W\}$, we have

$$\sup_{f \in \mathcal{F}} \Big\{ \psi(\mathbb{P}_{f,W}, \mathbb{P}^n_{f,W}) + \psi(\mathbb{P}_{f,W^c}, \mathbb{P}^n_{f,W^c}) \Big\} \leq 4\mathcal{R}_{n_{\mathcal{W}}}\big(\mathcal{G} \circ \mathcal{F}\big) + 2B \sqrt{\frac{2 \log\big(2|\mathcal{W}|/\delta\big)}{n_{\mathcal{W}}}}$$

for all $W \in \mathcal{W}$. Taking $\delta = 1/n$ concludes the proof. □

Proof of Theorem 4.1. Let $f_i := f(x_i, s_i)$. Recall that $y_{W,i} = 2\mathbb{I}(s_i \in W) - 1 \in \{-1, 1\}, i \in [n]$. Then, we can rewrite

$$\text{IPM}_{\mathcal{G}}(\mathbb{P}_{f,W}, \mathbb{P}_{f,W^c}) := \sup_{g \in \mathcal{G}} \left| \frac{1}{|W|} \sum_{i : y_{W,i} = 1} g(f_i) - \frac{1}{|W^c|} \sum_{i : y_{W,i} = -1} g(f_i) \right|$$

and Lemma A.2 in the next subsection concludes

$$\text{IPM}_{\mathcal{G}}(\mathbb{P}_{f,W}, \mathbb{P}_{f,W^c}) = \sup_{g \in \mathcal{G}} \left| \frac{\sum_{i=1}^n (y_{W,i} - \bar{y}_W) g(f_i)}{\sum_{i=1}^n (y_{W,i} - \bar{y}_W)^2} \right| = \sup_{g \in \mathcal{G}} |\tilde{R}^2(f, W, g)|.$$

□

A.3 Technical Lemmas

Lemma A.2. *Fix $\mathcal{G} \subset \{g : \mathbb{R} \to \mathbb{R}\}$. Let W be a given subset of $\{0,1\}^q$ and $f_i = f(x_i, s_i), i \in [n]$. For a binary indicator $y_{W,i} = 2\mathbb{I}(s_i \in W) - 1 \in \{-1, 1\}, i \in [n]$, we have*

$$\frac{1}{|W|} \sum_{i:y_{W,i}=1} g(f_i) - \frac{1}{|W^c|} \sum_{i:y_{W,i}=-1} g(f_i) = \frac{\sum_{i=1}^n (y_{W,i} - \bar{y}_W) g(f_i)}{\sum_{i=1}^n (y_{W,i} - \bar{y}_W)^2}, \tag{9}$$

for any $g \in \mathcal{G}$, where $\bar{y}_W := \frac{1}{n}\sum_{i=1}^n y_{W,i}$ and $\bar{g} := \frac{1}{n}\sum_{i=1}^n g(f_i)$.

Proof. We begin by rewriting $\bar{g} = \frac{1+\bar{y}_W}{2}\left(\frac{1}{|W|}\sum_{i:y_{W,i}=1} g(f_i)\right) + \frac{1-\bar{y}_W}{2}\left(\frac{1}{|W^c|}\sum_{i:y_{W,i}=-1} g(f_i)\right)$ since $|W| = \sum_{i:y_{W,i}=1} 1 = \frac{n+\sum_{i=1}^n y_{W,i}}{2} = \frac{n(1+\bar{y}_W)}{2}$ and $|W^c| = \sum_{i:y_{W,i}=-1} 1 = \frac{n-\sum_{i=1}^n y_{W,i}}{2} = \frac{n(1-\bar{y}_W)}{2}$. Note that

$$\begin{aligned}
\sum_{i=1}^n \big(y_{W,i} - \bar{y}_W\big)^2 &= \sum_{i=1}^n y_{W,i}^2 - 2\bar{y}_W \sum_{i=1}^n y_{W,i} + n\bar{y}_W^2 \\
&= n - 2\bar{y}_W \left(|W| - |W^c|\right) + n\bar{y}_W^2 \qquad (\because y_{W,i}^2 = 1, \ \sum_i y_{W,i} = |W| - |W^c|) \\
&= n - \frac{(|W| - |W^c|)^2}{n} \qquad (\because \bar{y}_W = \frac{|W| - |W^c|}{n}) \\
&= \frac{4\,|W|\,|W^c|}{n} \qquad (\because 1 + \bar{y}_W = \frac{2|W|}{n}, 1 - \bar{y}_W = \frac{2|W^c|}{n}).
\end{aligned} \tag{10}$$

Then, we expand

$$\begin{aligned}
&\sum_{i=1}^n (y_{W,i} - \bar{y}_W)\big(g(f_i) - \bar{g}\big) = \sum_{i=1}^n y_{W,i} g(f_i) - \bar{g}\sum_{i=1}^n y_{W,i} - \bar{y}_W \sum_{i=1}^n g(f_i) + n\bar{y}_W\bar{g} \\
&= \sum_{i:\, y_{W,i}=1} g(f_i) - \sum_{i:\, y_{W,i}=-1} g(f_i) - \bar{g}\left(|W| - |W^c|\right) - \bar{y}_W\left(n\bar{g}\right) + n\bar{y}_W\bar{g} \\
&= \sum_{i:\, y_{W,i}=1} g(f_i) - \sum_{i:\, y_{W,i}=-1} g(f_i) - \bar{g}\left(|W| - |W^c|\right) \\
&= \sum_{i:\, y_{W,i}=1} g(f_i) - \sum_{i:\, y_{W,i}=-1} g(f_i) \\
&\quad - \left(|W| - |W^c|\right)\left(\frac{|W|}{n} \cdot \frac{1}{|W|}\sum_{i:y_{W,i}=1} g(f_i) + \frac{|W^c|}{n} \cdot \frac{1}{|W^c|}\sum_{i:y_{W,i}=-1} g(f_i)\right) \\
&= \left(1 - \frac{|W| - |W^c|}{n}\right)\sum_{i:\, y_{W,i}=1} g(f_i) - \left(1 + \frac{|W| - |W^c|}{n}\right)\sum_{i:y_{W,i}=-1} g(f_i) \\
&= \frac{2|W^c|}{n}\sum_{i:y_{W,i}=1} g(f_i) - \frac{2|W|}{n}\sum_{i:y_{W,i}=-1} g(f_i) \\
&= \frac{2\,|W|\,|W^c|}{n}\left(\frac{1}{|W|}\sum_{i:y_{W,i}=1} g(f_i) - \frac{1}{|W^c|}\sum_{i:\, y_{W,i}=-1} g(f_i)\right) \\
&= \frac{1}{2}\sum_{i=1}^n \big(y_{W,i} - \bar{y}_W\big)^2\left(\frac{1}{|W|}\sum_{i:y_{W,i}=1} g(f_i) - \frac{1}{|W^c|}\sum_{i:y_{W,i}=-1} g(f_i)\right),
\end{aligned} \tag{11}$$

where the last equality holds by Eq. (10). Dividing by $\sum_i (y_{W,i} - \bar{y}_W)^2$, we get

$$\frac{\sum_{i=1}^n (y_{W,i} - \bar{y}_W)\left(g(f_i) - \bar{g}\right)}{\sum_i (y_{W,i} - \bar{y}_W)^2} = \frac{1}{2}\left(\frac{1}{|W|}\sum_{i:y_{W,i}=1} g(f_i) - \frac{1}{|W^c|}\sum_{i:y_{W,i}=-1} g(f_i)\right). \tag{12}$$

Using the fact that

$$\frac{\sum_{i=1}^{n}(y_{W,i} - \bar{y}_W)\,(g(f_i) - \bar{g})}{\sum_i (y_{W,i} - \bar{y}_W)^2} = \frac{\sum_{i=1}^{n}(y_{W,i} - \bar{y}_W) g(f_i)}{\sum_i (y_{W,i} - \bar{y}_W)^2},$$

we conclude the proof. □

A.4 Examples of $\mathcal{F}$ and $\mathcal{G}$ in Theorem 3.1

We introduce two examples that yields small Rademacher complexities $\mathcal{R}_{n_\mathcal{W}}(\mathcal{G} \circ \mathcal{F}) = \mathcal{O}(1/\sqrt{n_\mathcal{W}})$ so the uniform population–empirical gap in Theorem 3.1 shrinks at a rate $\mathcal{O}(1/\sqrt{n_\mathcal{W}})$ up to a logarithm factor of n.

Example A.3 (Linear functions). Let $\mathcal{G} = \{g_u(z) = \langle u, z\rangle : \|u\|_2 \le 1\}$ and $\mathcal{F} = \{f_W(x,s) = Wz(x,s) : \|W\|_2 \le M\}$, where $z(x,s) \in \mathbb{R}^d$ are fixed features with $\|z(x,s)\|_2 \le B_z$ for all (x,s). Then for all $f_W \in \mathcal{F}$ and $g_u \in \mathcal{G}$, $|(g_u \circ f_W)(x,s)| = |\langle u, Wz(x,s)\rangle| \le \|u\|_2 \|W\|_2 \|z(x,s)\|_2 \le MB_z$, so the class is uniformly bounded by $R := MB_z$. Moreover,

$$\begin{aligned}\mathcal{R}_{n_\mathcal{W}}(\mathcal{G} \circ \mathcal{F}) &= \frac{1}{n_\mathcal{W}} \mathbb{E}_\sigma \sup_{\|u\| \le 1, \|W\| \le M} \sum_{i=1}^{n_\mathcal{W}} \sigma_i \langle u, Wz_i\rangle = \frac{1}{n_\mathcal{W}} \mathbb{E}_\sigma \sup_{\|W\| \le M} \Big\| \sum_{i=1}^{n_\mathcal{W}} \sigma_i W z_i \Big\|_2 \\ &\le \frac{1}{n_\mathcal{W}} \mathbb{E}_\sigma \sup_{\|W\| \le M} \|W\|_2 \Big\| \sum_{i=1}^{n_\mathcal{W}} \sigma_i z_i \Big\|_2 \le \frac{M}{n_\mathcal{W}} \mathbb{E}_\sigma \Big\| \sum_{i=1}^{n_\mathcal{W}} \sigma_i z_i \Big\|_2 \le \frac{MB_z}{\sqrt{n_\mathcal{W}}},\end{aligned}$$

since $\|z_i\|_2 \le B_z$. Consequently, for all $f \in \mathcal{F}$,

$$\Delta_{\psi,\mathcal{W}}(f) - \Delta_{n,\psi,\mathcal{W}}(f) \lesssim \frac{1}{\sqrt{n_\mathcal{W}}} \Big\{ 4MB_z + 2MB_z \sqrt{2\log(2n|\mathcal{W}|)} \Big\}.$$

Example A.4 (Deep Neural Networks). Suppose $\mathcal{G} \circ \mathcal{F}$ is a ReLU Deep Neural Network (e.g., $\mathcal{G}$ and $\mathcal{F}$ are both ReLU DNNs) with L-many layers and weight matrices A_ℓ of spectral norms $s_\ell = \|A_\ell\|_2$ and Frobenius norms $\|A_\ell\|_F$ for $\ell \in [L]$. Then, we have $\mathcal{R}_{n_\mathcal{W}}(\mathcal{G} \circ \mathcal{F}) \lesssim \frac{1}{\sqrt{n_\mathcal{W}}} (\prod_{\ell=1}^{L} s_\ell)(\sum_{\ell=1}^{L} \|A_\ell\|_F^2 / s_\ell^2)^{1/2}$ (Bartlett et al., 2017). Further, if $g \circ f$ is uniformly bounded, then we have

$$\Delta_{\psi,\mathcal{W}}(f) - \Delta_{n,\psi,\mathcal{W}}(f) \lesssim \frac{C}{\sqrt{n_\mathcal{W}}} + \frac{1}{\sqrt{n_\mathcal{W}}} 2R\sqrt{2\log(2n|\mathcal{W}|)}$$

for all $f \in \mathcal{F}$ and a constant C depending on the network parameters L and A_ℓ.

B EXPERIMENTS

B.1 DATASETS

- ADULT (Tabular) (Becker & Kohavi, 1996): We predict income (≥ 50K) from census features. For the sensitive attributes, we consider sex, race, age, and marital-status, so that $q = 4$.
- COMMUNITIES (Tabular) (Redmond & Baveja, 2002): The class label is binary, indicating whether the violent crime rate is above a threshold. For the sensitive attributes, we consider 4 variables regarding race (racepctwhite, racepctblack, racepctasian, racepcthisp), 6 racial per-capita variables (whitepercap, blackpercap, indianpercap, asianpercap, otherpercap, hisppercap), 8 language/immigration related-variables (pctnotspeakenglwell, pctforeignborn, pctimmigrecent, pctimmigrec5, pctimmigrec8, pctimmigrec10, pctrecentimmig, pctrecimmig5) so that $q = 18$.
- DUTCH (Tabular) (van der Laan, 2000): We predict occupation from socio-economic features. For the sensitive attributes, we consider sex and age, so that $q = 2$.
- CIVILCOMMENTS (Text) (Borkan et al., 2019): We predict toxicity from user-generated comments. For the sensitive attributes, we consider sex (male/female/other), race (black/white/asian/other), and religion (christian/other) so that $q = 3$.
- ACSINCOME (Tabular) (Ding et al., 2022): We predict income level using nationwide census features. Sensitive attributes include sex, race (White, Black, Asian, Other), and marital-status, yielding $q = 3$.

B.2 IMPLEMENTATION DETAILS

We run all algorithms over five random seeds and report the average performance.

DRAF algorithm To control the fairness level, the Lagrangian multiplier λ is swept over $\{0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 1.00, 2.00, 3.00, 4.00, 5.00, 10.00, 20.00\}$. The candidate set of γ is $\{0.001, 0.005, 0.01, 0.05, 0.10, 0.20, 0.30\}$, and we choose an optimal one using the Pareto-front lines, as mentioned in the main body. We run DRAF with a maximum of 200 epochs, and select the best model whose validation accuracy is the highest among the 200 epochs. Algorithm 1 outlines the the DRAF algorithm.

Baselines The fairness penalty of REG is the sum of group disparities: $\mathrm{DP}_{\mathrm{marg}}(f) := \sum_{l\in[q]} |\frac{1}{n}\sum_{i=1}^{n} f_i - \frac{1}{n_l}\sum_{i:s_{i,l}=1} f_i|$, where $s_{i,l}$ denotes the l^{th} component of s_i and $n_l = \sum_{i=1}^{n} \mathbb{I}(s_{i,l} = 1)$ for $l \in [q]$. The final objective is defined as $\frac{1}{n}\sum_{i=1}^{n} l(y_i, f(x_i, s_i)) + C_{\mathrm{REG}}\mathrm{DP}_{\mathrm{marg}}(f)$ for some $C_{\mathrm{REG}} \geq 0$. We sweep the regularization parameter C_{REG} over $\{0.001, 0.002, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0, 2.0, 5.0, 10.0, 20.0, 50.0, 100.0\}$ to control the fairness level. Similar to DRAF, we run REG with a maximum of 200 epochs, and select the best model whose validation accuracy is the highest among the epochs.

For GF, since the official code and the released AIF360 package[1] support only FP and FN (false positives and false negatives), we re-implement GF for the demographic parity setting targeted in this paper. For stable and fast optimization, we use a gradient-descent–based approach. The fairness penalty of GF is the (weighted) worst-group disparity: $\mathrm{DP}_{\max}(f) := \max_{s\in\{0,1\}^q} \frac{n_s}{n}\mathrm{DP}_s(f)$, where $\mathrm{DP}_s(f) := \left|\hat{p}_s - \hat{p}\right|, \hat{p} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\hat{y}_i = 1\}$, and $\hat{p}_s := \frac{1}{n_s}\sum_{i:\, s_i=s} \mathbb{I}\{\hat{y}_i = 1\}$. The final objective is then defined as $\frac{1}{n}\sum_{i=1}^{n} l(y_i, f(x_i, s_i)) + C_{\mathrm{GF}}\mathrm{DP}_{\max}(f)$. Here, we sweep the regularization parameter C_{GF} over $\{0.1, 0.5, 1.0, 5.0, 20.0, 50.0, 200.0, 500.0, 1000.0, 5000.0\}$ to control the fairness level. Note that, rather than taking maximum over s, we apply the softmax function to $\{\frac{n_s}{n}\mathrm{DP}_s(f)\}_{s\in\{0,1\}^q}$ to make the optimization stable. Similar to DRAF, we run GF with a maximum of 200 epochs, and select the best model whose validation accuracy is the highest among the epochs.

For SEQ, we re-implement the algorithm in the original paper (Hu et al., 2024). That is, we first learn a classifier without fairness constraints, then sequentially post-process the prediction scores

[1] https://aif360.readthedocs.io/en/v0.4.0/modules/generated/aif360.algorithms.inprocessing.GerryFairClassifier.html

Algorithm 1: DRAF algorithm

Input : Training data $\{(x_i, s_i, y_i)\}_{i=1}^n$, Learning rates $(\eta_{\text{cls}}, \eta_g, \eta_{\mathbf{v}})$, Number of iterations T, and Fairness Lagrangian multiplier λ

Output: Classifier parameters θ of $f = f_\theta$, Discriminator parameters ϕ of $g = g_\phi$, and Weight vector $\mathbf{v}$

Initialize: $\theta \leftarrow \theta_0$, $\phi \leftarrow \phi_0$, $\mathbf{v} \leftarrow \mathbf{v}_0$

do

  for $i = 1, \ldots, n$ **do**

    $\hat{y}_i \leftarrow f_\theta(x_i, s_i)$

  end

  Compute the classification loss: $L_{\text{cls}} = \frac{1}{n}\sum_{i=1}^n \text{CE}(\hat{y}_i, y_i)$

  Compute the fairness loss:

$$\widehat{\text{DR}} = \text{DR}_{n,\mathcal{W},\mathcal{G}}(f) := \sup_{g \in \mathcal{G}, \mathbf{v} \in \mathcal{S}^M} z\text{-DR}^2(f, \mathbf{v}, g) = \sup_{g \in \mathcal{G}, \mathbf{v} \in \mathcal{S}^M} \log\left(\frac{1 + |\text{DR}^2(f, \mathbf{v}, g)|/2}{1 - |\text{DR}^2(f, \mathbf{v}, g)|/2}\right)$$

  Update the discriminator and the subgroup weight by gradient ascending:

$$\phi \leftarrow \phi + \eta_g \nabla_\phi \widehat{\text{DR}}, \tilde{\mathbf{v}} \leftarrow \mathbf{v} + \eta_{\mathbf{v}} \nabla_{\mathbf{v}} \widehat{\text{DR}}, \mathbf{v} \leftarrow \text{Proj}_{\mathcal{S}^M}(\tilde{\mathbf{v}}), (\text{Proj}_{\mathcal{S}^M} = \text{unit sphere projection})$$

  Update the classifier:

$$\theta \leftarrow \theta - \eta_{\text{cls}} \nabla_\theta L_{\text{cls}} - \lambda \eta_{\text{cls}} \nabla_\theta \widehat{\text{DR}}$$

until *convergence or T iterations*;

Return $\theta, \phi, \mathbf{v}$

from each subgroups to a common barycenter. The learning rate used to learn the classifier is swept over $\{0.001, 0.005, 0.01, 0.05, 0.10\}$.

B.3 RELATIONSHIP BETWEEN DR GAP AND SUPIPM

As Eq. (3) shows that the supIPM is upper bounded by DR^2, and that DR^2 coincides with supIPM when the vector $\mathbf{v}$ lies on a vertex of the simplex (i.e., $\mathbf{v} = \mathbf{e}_k$ for some k). Thus, minimizing DR^2 reduces the upper bound on supIPM and recovers the supIPM when $\mathbf{v}$ is (approximately) a vertex (Theorem 4.1).

Furthermore, we empirically show that DR is almost equal to supIPM, by checking whether the learned $\mathbf{v}$ is near vertex. To quantify how close the learned weight vector $\mathbf{v}$ is to a simplex vertex, we compute the Euclidean distance $d(\mathbf{v}) := \min_{1 \le k \le m} \|\mathbf{v} - e_k\|_2$, where e_k denotes the k-th vertex of the m-dimensional probability simplex. We collect $\{d(\mathbf{v})\}$ over all runs and models, and plot their empirical distribution as a histogram. The resulting plots in Figure 5 show that $d(\mathbf{v})$ is highly concentrated near zero as training proceeds, indicating that $\mathbf{v}$ empirically converges to (or very close to) a simplex vertex well.

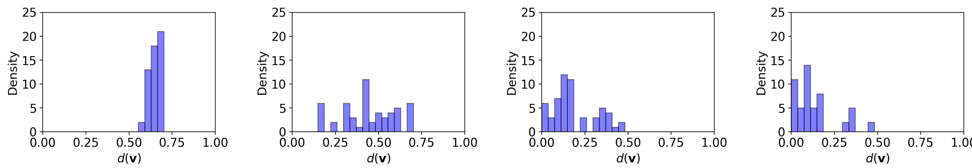
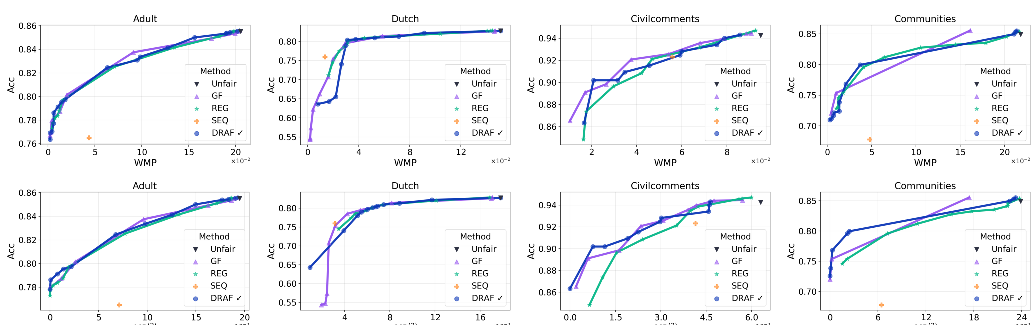


Figure 5: Histogram of $d(\mathbf{v})$ on CIVILCOMMENTS dataset for four time steps (left to right: 1st epoch, 5th epoch, 20th epoch, and 200th epoch), where the distribution becomes concentrated near 0 as training proceeds.

B.4 PERFORMANCE COMPARISON

Trade-off between accuracy and fairness Figure 6 compares the trade-off between the distributional first-order marginal and the second-order marginal fairness levels (i.e., WMP and $\text{MP}^{(2)}$) and accuracy. The results give the similar implications that we observe from Figure 3 in Section 5.3 of the main body. That is, compared to the baseline methods (GF, REG, and SEQ), DRAF performs comparable on ADULT, DUTCH, shows a slightly better performance on CIVILCOMMENTS, and outperforms on COMMUNITIES.



(a) ADULT (b) DUTCH (c) CIVILCOMMENTS (d) COMMUNITIES

Figure 6: Trade-off between fairness level and accuracy. (Top, Bottom) = WMP vs. Acc, $\text{MP}^{(2)}$ vs. Acc. We set γ to 0.2 for ADULT, 0.001 for COMMUNITIES, 0.2 for DUTCH, and 0.05 for CIVIL-COMMENTS, reflecting the sparsity of each dataset to determine the optimal value.

Analysis on additional datasets

1. **Synthetic ADULT dataset**: In addition to COMMUNITIES dataset, we conduct an additional study using a synthetic variant of ADULT dataset with sparse subgroups. We construct SPARSEADULT by selecting the five smallest subgroups (whose sizes are at least 192) from ADULT and randomly down-sampling them to smaller samples with sizes in $[40, 60]$ (see Table 3). We then evaluate five algorithms on SPARSEADULT and report the trade-off results in Figure 7. Similar to the

case for COMMUNITIES, it shows that DRAF preserves superior subgroup and marginal fairness performance, specifically for higher fairness range (e.g., small $\text{MP}^{(1)}$, WMP, and $\text{MP}^{(2)}$), on SPARSEADULT.

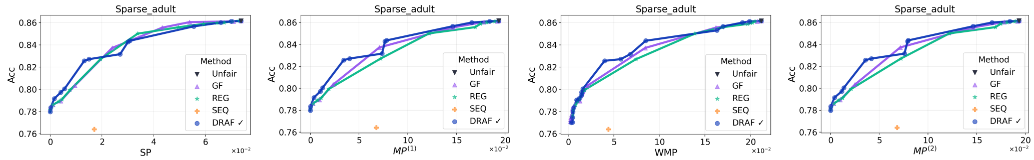


Figure 7: Trade-off between fairness level and accuracy on synthetic ADULT dataset. (Left to Right) $\{\text{SP}, \text{MP}^{(1)}, \text{WMP}, \text{MP}^{(2)}\}$ vs. $\texttt{Acc}$ on SPARSEADULT dataset. We set γ to 0.001.

Table 3: Subgroup sample counts of the original ADULT and SPARSEADULT datasets. Subgroup index starts at 1 with the smallest subgroup. The sizes for subgroups of index over 6 are the same.

Subgroup index	ADULT	SPARSEADULT
1	192	46
2	233	54
3	500	57
4	789	59
5	964	60

2. **ACSINCOME dataset:** We also conduct experiments on an additional tabular dataset (ACS INCOME, see Section B.1 for the details of this dataset). The results reported in Figure 8 show that DRAF exhibits similar behavior on these datasets as on the four datasets currently used, which further supports the empirical outperformance of DRAF.

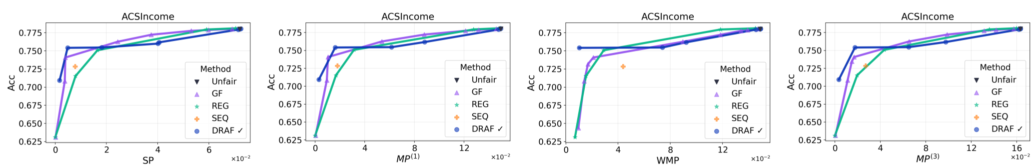


Figure 8: Trade-off between fairness level and accuracy on ACSINCOME dataset. (Left to Right) $\{\text{SP}, \text{MP}^{(1)}, \text{WMP}, \text{MP}^{(2)}\}$ vs. $\texttt{Acc}$ on ACSINCOME dataset. We set γ to 0.001.

Scalability To assess the scalability of DRAF with respect to the number of subgroups ($|\mathcal{S}| = 2^q$), we measure its running time as $|\mathcal{S}|$ increases. To do so, we construct a toy dataset with $n = 2^{16}$ and randomly assign sensitive attributes to the data. As shown in Table 4, the runtime of DRAF increases only moderately even though the number of subgroup-subsets grows exponentially (by 27% when $|\mathcal{S}|$ grows from 2^4 to 2^{14}). This suggests that DRAF remains scalable in terms of computation time.

Table 4: Comparison of runtime of DRAF for various cases of $|\mathcal{S}|$ (100% for $q = 4$).

$\|\mathcal{S}\| = 2^q$	2^4	2^6	2^8	2^{10}	2^{12}	2^{14}
Runtime	100%	101%	106%	105%	105%	127%

B.5 EXTENSIONS OF DRAF

Extension to multi-class classification Here, we show that DRAF is not limited to binary classification but can be extended to multi-class classification problem. We define the fairness measure in this setting as the maximum of the parity gaps across all classes, following Denis et al. (2024), and compare DRAF with the existing baseline method of Denis et al. (2024). Following Denis et al. (2024), we use COMMUNITIES dataset with five classes and binary sensitive attribute (percent of not speaking english well).

Table 5: Comparison of performance for multi-class classification problem.

	Acc	Max. of MP$^{(1)}$ over classes
Unfair	0.457	0.108
(Denis et al., 2024)	0.365	0.056
DRAF ✓	**0.380**	**0.054**

As shown in results in Table 5, DRAF achieves a competitive fairness level compared to the baseline, while attaining a higher accuracy.

Extension to equalized odds In addition to demographic parity notion, we further verify that DRAF can be extended to other group fairness notions, e.g., equalized odds (EO). We modify the DR penalty for EO as: $\mathrm{DR}^2_{\mathrm{TPR}}(f, \mathbf{v}, g) := 1 - \frac{\{\sum_{i=1}^n (\mathbf{v}^\top c_i - g(f_i))^2 - \sum_{i=1}^n (g(f_i) - \mu_{\mathbf{v}})^2\}}{\sum_{i=1}^n (\mathbf{v}^\top c_i - \mu_{\mathbf{v}})^2}, c_{im} = 2\mathbb{I}(s_i \in W_m, y_i = 1) - 1$ and $\mathrm{DR}^2_{\mathrm{FPR}}(f, \mathbf{v}, g) := 1 - \frac{\{\sum_{i=1}^n (\mathbf{v}^\top c_i - g(f_i))^2 - \sum_{i=1}^n (g(f_i) - \mu_{\mathbf{v}})^2\}}{\sum_{i=1}^n (\mathbf{v}^\top c_i - \mu_{\mathbf{v}})^2}, c_{im} = 2\mathbb{I}(s_i \in W_m, y_i = 0) - 1$, where the former is for TPR (True Positive Rate) and the latter is for FPR (False Positive Rate). Then, we minimize $\frac{1}{n}\sum_{i=1}^n l(y_i, f(x_i, s_i)) + \frac{\lambda}{2}\left(z\text{-}\mathrm{DR}^2_{\mathrm{TPR}}(f, \mathbf{v}, g) + z\text{-}\mathrm{DR}^2_{\mathrm{FPR}}(f, \mathbf{v}, g)\right)$, where $z\text{-}\mathrm{DR}^2_{\mathrm{TPR}}$ and $z\text{-}\mathrm{DR}^2_{\mathrm{FPR}}$ are the corresponding z-transformations of $\mathrm{DR}^2_{\mathrm{TPR}}(f, \mathbf{v}, g)$ and $\mathrm{DR}^2_{\mathrm{FPR}}(f, \mathbf{v}, g)$, respectively. We call this modified DRAF algorithm specially tailored for EO as DRAF-EO.

For fairness performance, we consider marginal fairness and subgroup fairness measures, similar to the main analysis on demographic parity. Let $n_+ = \sum_{i=1}^n \mathbb{I}(y_i = 1), n_- = \sum_{i=1}^n \mathbb{I}(y_i = 0), n_{+,s} = \sum_{i=1}^n \mathbb{I}(y_i = 1, s_i = s)$, and $n_{-,s} = \sum_{i=1}^n \mathbb{I}(y_i = 0, s_i = s)$. We also define the positive prediction ratios as $\hat{p}_+ := \frac{1}{n_+}\sum_{i:y_i=1} \mathbb{I}(\hat{y}_i = 1), \hat{p}_{+,s} := \frac{1}{n_{+,s}}\sum_{i:y_i=1,s_i=s} \mathbb{I}(\hat{y}_i = 1), \hat{p}_- := \frac{1}{n_-}\sum_{i:y_i=0} \mathbb{I}(\hat{y}_i = 1)$, and $\hat{p}_{-,s} := \frac{1}{n_{-,s}}\sum_{i:y_i=0,s_i=s} \mathbb{I}(\hat{y}_i = 1)$. Furthermore, $n_{+,L}^{(a)}, \hat{p}_{+,L}^{(a)}, n_{-,L}^{(a)}, \hat{p}_{-,L}^{(a)}$ and $\widehat{\mathbb{P}}_{+,f}, \widehat{\mathbb{P}}_{-,f}, \widehat{\mathbb{P}}_{+,f,j|a}, \widehat{\mathbb{P}}_{-,f,j|a}$ are defined similarly. Table 6 describes the fairness performance measures used in the experiments for EO.

Table 6: Fairness performance measures for EO.

Name	Meaning	Formula
TPR$^{(l)}$	l^{th}-order TPR	$\max_{L \subseteq [q], \lvert L \rvert = l} \sum_{a \in \{0,1\}^l} \frac{n_{+,L}^{(a)}}{n_+} \lvert \hat{p}_{+,L}^{(a)} - \hat{p}_+ \rvert$
FPR$^{(l)}$	l^{th}-order FPR	$\max_{L \subseteq [q], \lvert L \rvert = l} \sum_{a \in \{0,1\}^l} \frac{n_{-,L}^{(a)}}{n_-} \lvert \hat{p}_{-,L}^{(a)} - \hat{p}_- \rvert$
WTPR	Distributional TPR	$\max_{j \in [q]} \max \left\{ \frac{n_{+,j}^{(0)}}{n_+} \mathrm{W}_1(\widehat{\mathbb{P}}_{+,f,j\lvert 0}, \widehat{\mathbb{P}}_{+,f}), \frac{n_{+,j}^{(1)}}{n_+} \mathrm{W}_1(\widehat{\mathbb{P}}_{+,f,j\lvert 1}, \widehat{\mathbb{P}}_{+,f}) \right\}$
WFPR	Distributional FPR	$\max_{j \in [q]} \max \left\{ \frac{n_{-,j}^{(0)}}{n_-} \mathrm{W}_1(\widehat{\mathbb{P}}_{-,f,j\lvert 0}, \widehat{\mathbb{P}}_{-,f}), \frac{n_{-,j}^{(1)}}{n_-} \mathrm{W}_1(\widehat{\mathbb{P}}_{-,f,j\lvert 1}, \widehat{\mathbb{P}}_{-,f}) \right\}$
STPR	Subgroup TPR	$\max_{s \in \{0,1\}^q} \frac{n_{+,s}}{n_+} \lvert \hat{p}_{+,s} - \hat{p}_+ \rvert$
SFPR	Subgroup FPR	$\max_{s \in \{0,1\}^q} \frac{n_{-,s}}{n_-} \lvert \hat{p}_{-,s} - \hat{p}_- \rvert$

For a baseline method, we consider FairICP (Lai & Guan, 2025), which is a specially designed adversarial learning algorithm for EO in presence of subgroups. The results are given in Tables 7 and 8, which show that DRAF-EO performs competitive to FairICP, enhancing the empirical superiority and flexibility of our proposed doubly regressing approach.

Table 7: Comparison of marginal fairness on ADULT dataset.

		Prediction-based			Distribution-based		
Method	Acc	TPR$^{(1)}$	FPR$^{(1)}$	$\frac{\text{TPR}^{(1)}+\text{FPR}^{(1)}}{2}$	WTPR	WFPR	$\frac{\text{WTPR}+\text{WFPR}}{2}$
Unfair	0.855	0.0993	0.0886	0.0940	0.0687	0.1183	0.0928
FairICP	0.804	0.0342	0.0150	0.0238	0.0216	0.0214	0.0207
DRAF-EO ✓	0.804	**0.0155**	**0.0010**	**0.0082**	**0.0113**	**0.0087**	**0.0100**

Table 8: Comparison of subgroup fairness on ADULT dataset.

Method	Acc	STPR	SFPR	$\frac{\text{STPR}+\text{SFPR}}{2}$
Unfair	0.855	0.0565	0.0284	0.0422
FairICP	0.804	0.0154	0.0043	0.0091
DRAF-EO ✓	0.804	**0.0056**	**0.0004**	**0.0030**

B.6 Additional studies

Excluding the marginal subgroups from $\mathcal{W}$ Let DRAF$_{\text{-m}}$ denotes the DRAF variant whose $\mathcal{W}$ does not include the first-order marginal subgroups. Figure 9 shows that, excluding first-order marginal subgroups from $\mathcal{W}$ (i.e., DRAF$_{\text{-m}}$) can harm first-order marginal fairness, even subgroup fairness is satisfied. Moreover, on CivilComments dataset, DRAF$_{\text{-m}}$ and DRAF perform comparable in terms of MP$^{(1)}$, when MP$^{(1)}$ is not small, but DRAF significantly outperforms DRAF$_{\text{-m}}$ in view of WMP. This observation suggests that achieving prediction-based fairness (e.g., MP$^{(1)}$) does not necessarily guarantee distributional fairness (e.g., WMP), and it highlights the need to control distributional fairness as well, which DRAF aims at.

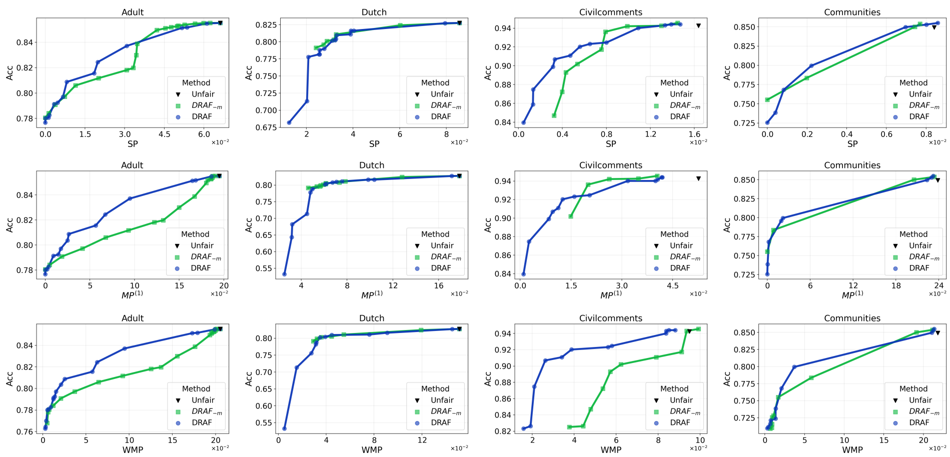


Figure 9: Comparison of DRAF$_{\text{-m}}$ and DRAF in terms of SP (top), MP$^{(1)}$ (center), and WMP (bottom). We set γ to $0.2, 0.001, 0.2$, and 0.05 for Adult, Dutch, CivilComments, and Communities dataset, respectively.

Note that, on Communities dataset, DRAF$_{\text{-m}}$ and DRAF may appear similar in terms of MP$^{(1)}$, however, it is because DRAF$_{\text{-m}}$ fails to achieve moderate fairness levels (e.g., $[0.02, 0.2]$), leaving no point on the Pareto-front line. See Figure 10 for evidence that controlling MP$^{(1)}$ is not numerically easy for DRAF$_{\text{-m}}$. That is, a large drop in MP$^{(1)}$ is occurred at $\lambda = 0.2$ and we observe that using $\lambda \in [0.2, 0.3]$ does not provide intermediate fairness levels.

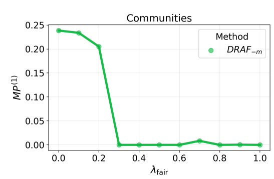


Figure 10: A plot between λ and MP$^{(1)}$ for DRAF$_{\text{-m}}$ on Communities dataset. We vary $\lambda \in [0.0, 0.1, \ldots, 1.0]$.

Similarly, we also consider $\mathcal{W}$ that excludes the second-order marginal subgroups. Let DRAF$_{\text{-m}^2}$ denotes the DRAF algorithm whose $\mathcal{W}$ does not include the second-order marginal subgroups. Figure 11 shows that the second-order marginal fairness can be slightly harmed when excluding the second-order marginal subgroups in $\mathcal{W}$. On the other hand, including the second-order marginal subgroups in $\mathcal{W}$ does not sacrifice first-order marginal or subgroup fairness, while can contribute to improving the second-order marginal fairness. Hence, we basically recommend building $\mathcal{W}$ to include all the first-order, the second-order, and subgroups.

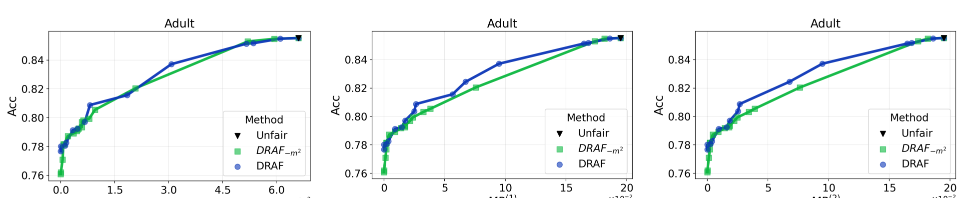


Figure 11: Comparison of DRAF$_{\text{-m}^2}$ and DRAF in terms of subgroup fairness (left: SP), first-order marginal fairness (center: MP$^{(1)}$), and the second-order marginal fairness (right: MP$^{(2)}$) on Adult dataset.

Impact of γ To support the claim in Section 5.5, we vary $\gamma \in \{0.001, 0.01, 0.1, 0.2, 0.3\}$ and compare the performance. The results in Figure 12 show that a larger γ (e.g., 0.3) degrades subgroup fairness performance compared to a small γ (e.g., 0.01). Conversely, since DRAF minimizes the worst disparity over subgroup-subsets in $\mathcal{W}$, a small γ may lead to slightly worse first-order marginal fairness than a large γ (e.g., 0.001 for CIVILCOMMENTS dataset), as it could focus on higher-order or subgroups rather than first-order marginal fairness for some cases.

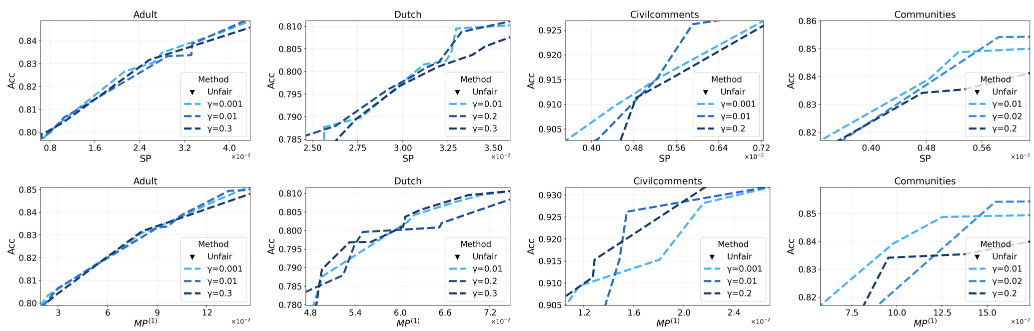


Figure 12: Impact of γ for DRAF in terms of subgroup fairness SP (top) and first-order marginal fairness $\mathrm{MP}^{(1)}$ (bottom).

Figure 13 provides similar results for (i) the distributional first-order marginal fairness WMP and (ii) the second-order marginal fairness $\mathrm{MP}^{(2)}$. Similar to Figure 12, a too small γ (e.g., 0.001) may lead to slightly worse first-order marginal fairness than a larger γ, while a too large γ (e.g., 0.2 in COMMUNITIES dataset) would harm the second-order marginal fairness.

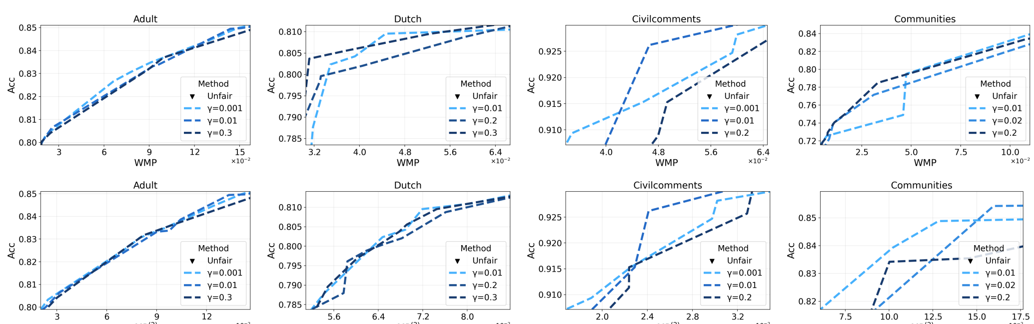


Figure 13: Impact of γ for DRAF in terms of distributional first-order marginal fairness WMP (top) and second-order marginal fairness $\mathrm{MP}^{(2)}$ (bottom).

To analyze the effect of γ in more detail, we categorize subgroups by size (large, medium, and small) using quantiles $0.3, 0.6$ and evaluate how their fairness on the test data changes as γ varies. This subgroup-wise analysis provides a clearer view of how γ influences fairness across different group scales. To this end, we train an unfair model (learning without any fairness constraint) and a fair model learned by DRAF for various values of γ. For every subgroup, we compute the disparity of the unfair model and that of the fair model, and then take their difference (unfair model - fair model). A positive difference means that the subgroup is treated more fairly by DRAF than by the unfair model, whereas a negative difference suggests the opposite. We then summarize these differences using boxplots, grouping subgroups into small, medium, and large categories by size.

The results in Figure 14 show that: (i) for large subgroups, the differences are consistently positive across all values of γ, indicating that DRAF reliably improves fairness for these subgroups; (ii) for medium-sized subgroups, the differences tend to be close to zero or negative when γ is large (e.g., $\gamma = 0.4$), but become positive as γ decreases, suggesting that a too large γ is undesirable for these groups; (iii) in contrast, for small subgroups, the differences are concentrated around zero or negative with large variations for all γ, indicating that we cannot reliably regard these tiny subgroups as being

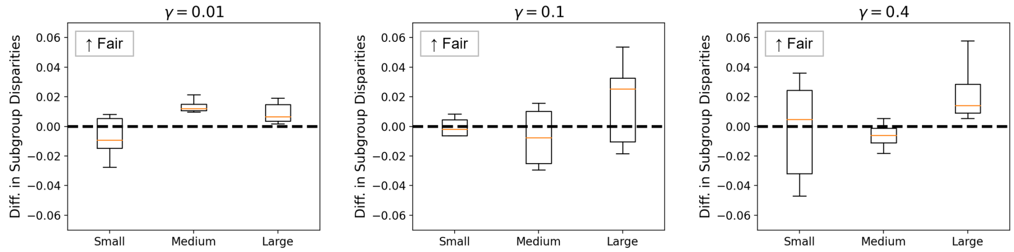
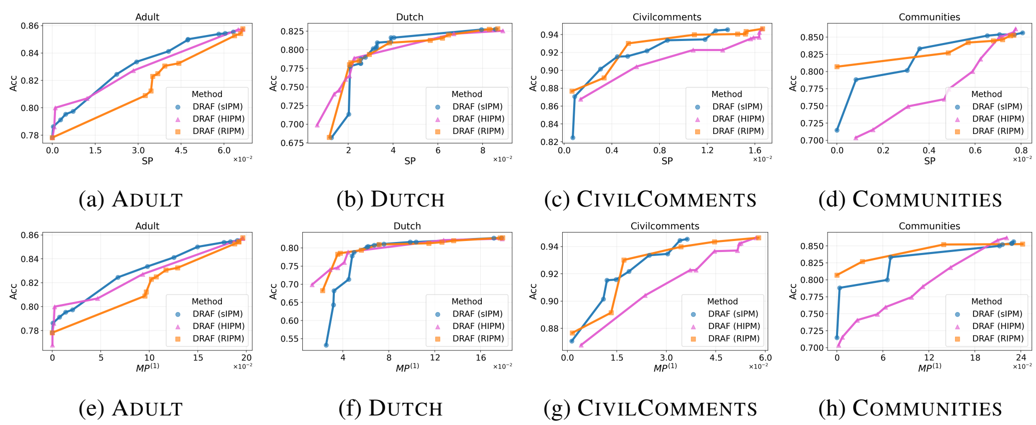


Figure 14: Comparison of differences in subgroup disparities (Diff. in Subgroup Disparities) between the unfair model and the fair model trained by DRAF, categorized by subgroup size (small, medium, and large).

treated fairly. This empirical result supports our theoretical finding in Theorem 3.1 that enforcing fairness on very small subgroups in the training data does not guarantee fairness for those subgroups on the test data.

Choice of $\mathcal{G}$ In this ablation study, we compare sIPM, RIPM, and HIPM for $\text{IPM}_{\mathcal{G}}$, in terms of the trade-off performance. See Figure 15 for the results on the four datasets. The key findings are: (i) sIPM (our default in the main analysis) performs best in most cases, though RIPM slightly outperforms sIPM on COMMUNITIES; (ii) RIPM performs similarly to sIPM overall except for ADULT dataset; (iii) HIPM underperforms both sIPM and RIPM in most cases.

Accordingly, we recommend using the more stable IPMs such as sIPM and RIPM rather than HIPM, whose more complex discriminator architecture often leads to less stable training and suboptimal models.



(a) ADULT (b) DUTCH (c) CIVILCOMMENTS (d) COMMUNITIES

(e) ADULT (f) DUTCH (g) CIVILCOMMENTS (h) COMMUNITIES

Figure 15: Trade-off between fairness level and accuracy with three IPMs (sIPM, RIPM, and HIPM) for $\mathcal{G}$. (Top, Bottom) = SP vs. $\texttt{Acc}$, $\text{MP}^{(1)}$ vs. $\texttt{Acc}$.

Robustness under noisy sensitive attributes To investigate the robustness of DRAF under the noisy sensitive attribute, we construct a controlled experiment on the ADULT dataset. We randomly introduce missing values into the sensitive attribute at rate 0.01. For the baseline GF, any column with at least one missing sensitive attribute is discarded, since GF requires complete subgroup to define its fair constraint. In contrast, DRAF can still be applied partially to samples with missing values: for samples with missing sensitive attributes, we impose fairness constraints only on subgroup-subsets that can be formed using the observed attributes, and ignore subgroup-subsets that involve any missing attributes.

For a sample with sensitive attribute vector $s_i = (s_{i1}, \dots, s_{iq}) \in \{0, 1, \text{NA}\}^q$, the indicator for a subgroup–subset $W \in \mathcal{W}$ is

$$c_{i,W} = \begin{cases} 1, & \text{if } s_i \in W \text{ and } s_{ij} \neq \text{NA}, \\ -1, & \text{if } s_i \notin W \text{ and } s_{ij} \neq \text{NA}, \\ 0, & \text{otherwise.} \end{cases}$$

Thus the full vector c is

$$c_i = \left(c_{i,W}\right)_{W \in \mathcal{W}},$$

where missing attributes selectively zero out only the corresponding components of c, and all subgroup-subsets built from observed sensitive attributes remain active in the c-vector. The results are shown in Figure 16, suggesting that this partial DRAF approach is more robust than the baseline: it achieves a better fairness-accuracy trade-off than GF, suggesting the robustness of DRAF in presence of noisy (missing) sensitive attributes.

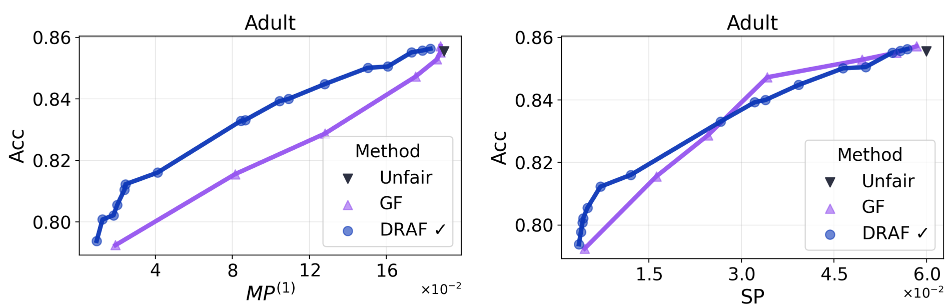


Figure 16: Trade-off between fairness level and accuracy when 1% of the sensitive attributes 'Married' in the ADULT dataset are randomly set to missing. (Left, Right) = $\text{MP}^{(1)}$ vs. Acc, SP vs. Acc.

B.7 Discussion on the relationship between marginal and subgroup fairness

Achieving marginal fairness alone does not guarantee subgroup fairness: even if each marginal group is treated fairly, certain intersections or unions of sensitive groups may still suffer from unfairness (fairness gerrymandering; (Kearns et al., 2018b)). Conversely, even if most subgroups are fairly treated, marginal fairness can still be violated when some groups are very sparse so that the fairness on the test data is not guaranteed (Theorem 3.1).

This observation motivates our choice to simultaneously monitor both marginal fairness and subgroup fairness. Furthermore, we believe that, from an ethical and policy perspective, a situation where most subgroups are fairly treated but a marginal group (e.g., 'all women') would be difficult to justify socially.

Figure 17 presents the overall hierarchy of subgroup-subsets, including marginal groups and subgroups, and highlights the subgroup-subsets that DRAF focuses on. We additionally illustrate that even if all subgroups are fairly treated, marginal fairness can still be violated when some subgroups are sparse.

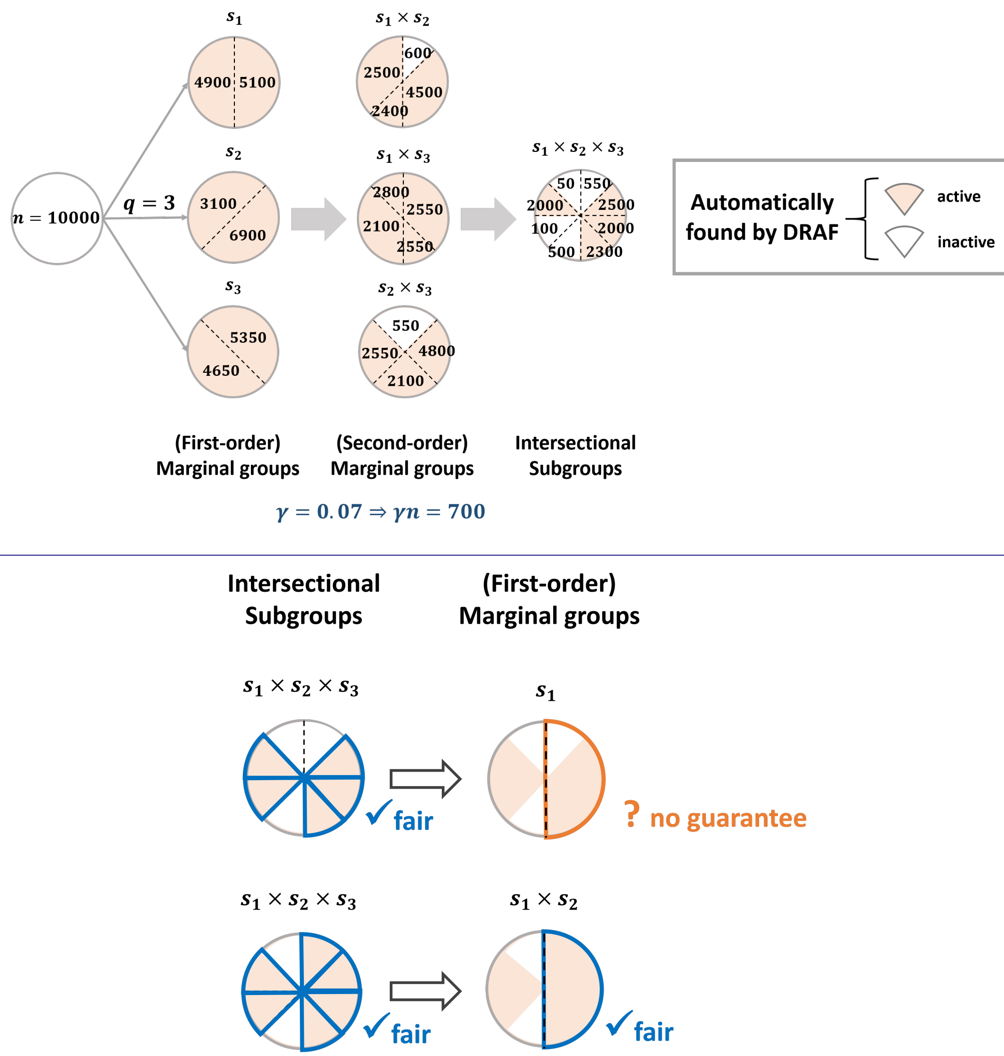


Figure 17: (Top) Hierarchy of marginal groups and intersectional subgroups, with active subgroups (whose sample sizes exceed a threshold) highlighted. Our proposed algorithm, DRAF, automatically identifies such active sets and enforces fairness on them. (Bottom) An visual illustration on the relationship between fairness guarantees from the intersectional subgroups to marginal groups.

Example: subgroup fairness does not always imply marginal fairness Suppose $q = 2$. Assume that we are given the following configuration of dataset and predictions for a given f. We write the two sensitive attributes as $a \in \{0, 1\}$ and $b \in \{0, 1\}$, for simplicity.

Subgroup (a, b)	# samples $n(a, b)$	# positive predictions $n_{pos}(a, b)$ by f	Positive rate $\hat{p}(a, b) = n_{pos}(a, b)/n$
$(0, 0)$	10	9	0.9
$(0, 1)$	10	9	0.9
$(1, 0)$	10	1	0.1
$(1, 1)$	10	1	0.1

The total sample size and positives are $n = \sum_{a,b} n(a, b) = 40$ and $n_{pos} = \sum_{a,b} n(a, b) = 20$, hence the overall rate of positive prediction is

$$\hat{p} = \frac{n_{pos}}{n} = \frac{20}{40} = 0.5.$$

Subgroup fairness measure of Kearns et al. (2018a) over the four intersectional subgroups is calculated as

$$\text{SP}(f) = \max_{(a,b) \in \{0,1\}^2} \frac{n(a, b)}{N} \big| \hat{p}(a, b) - \hat{p} \big| = 0.25 \times |0.9 - 0.5| = 0.1.$$

On the other hand, we have $n(0, 0) + n(0, 1) = 20, n_{pos}(0, 0) + n_{pos}(0, 1) = 18$ so that $\frac{n_{pos}(0,0)+n_{pos}(0,1)}{n(0,0)+n(0,1)} = 0.9$. Similarly, $n(1, 0) + n(1, 1) = 20, n_{pos}(1, 0) + n_{pos}(1, 1) = 2$ so that $\frac{n_{pos}(1,0)+n_{pos}(1,1)}{n(1,0)+n(1,1)} = 0.1$. Thus, the first-order marginal disparity for the sensitive attribute a is 0.4, which is relatively large.