DOUBLY-REGRESSING: CLOSING THE SPARSITY GAP IN SUBGROUP FAIRNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

The abstract paragraph should be indented 1/2 inch (3 picas) on both left and right-hand margins. Use 10 point type, with a vertical spacing of 11 points. The word ABSTRACT must be centered, in small caps, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one Gohar & Cheng (2023) paragraph.

1 INTRODUCTION

The rapid deployment of machine learning systems in socially consequential domains such as finance, hiring, and criminal justice has amplified the demand for fairness-aware predictions. Early definitions of algorithmic fairness predominantly focused on single sensitive attributes, such as gender or race, requiring parity across these marginal groups. However, fairness with respect to a single attribute is not sufficient to protect against discrimination at the intersections of multiple attributes. 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* (Foulds et al., 2019b; Úrsula Hébert-Johnson et al., 2018).

This extension arises from the problem of *fairness gerrymandering* (Kearns et al., 2018b;a). That is, even when fairness is satisfied on each marginal attribute, severe unfairness may remain on their intersections. Formally, suppose that each ith individual is specified by its sensitive attribute vector $s_i \in \{0, 1\}^q$, where each coordinate is binary. Then, there are 2^q number of subgroups, defined by

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

Subgroup fairness requires that predictive distributions be similar across all 2^q subgroups. However, this induces two major challenges: (i) the *computational burden*, as the number of constraints scales exponentially in q; and (ii) the *data sparsity problem*, when many subgroups contain only a few of samples, making fairness estimation unstable (Molina & Loiseau, 2023). Among these, sparsity is particularly detrimental because it undermines fairness protection for precisely those minority subgroups most in need of safeguards.

Recent works have attempted to mitigate these challenges. Hu et al. (2024) proposed a sequential Wasserstein barycenter mechanism to progressively debias multiple attributes. Molina & Loiseau (2023) established probabilistic bounds connecting marginal and intersectional fairness. Tian et al. (2025) developed *MultiFair*, a mixup-based data fusion approach to neutralize subgroup information. Foulds et al. (2019b;a) introduced *differential fairness*, a privacy-inspired criterion with theoretical guarantees. Yet, these approaches have notable limitations: sequential correction accumulates error, marginal-based bounds fail to protect minorities, mixup methods risk prediction degradation, and differential fairness is computationally intensive. None fully resolve the dual obstacles of exponential computation and subgroup sparsity.

To overcome these limitations, we propose a novel fairness notion called *partial subgroup fairness*. We first define a *subgroup-subset* $S' \subseteq \{0, 1\}^q$ as the union of multiple subgroups: $\mathcal{D}_{S'} = \bigcup_{v \in S'} \mathcal{D}_v$. Then, we enforce fairness only on subgroup-subsets with sufficient sample size, to mitigate the sparsity problem. Even if a single subgroup $\mathcal{D}_v$ is too small, its union combined with other subgroups can form a stable subset with sufficient size. Further, instead of directly enforcing fairness across all possible subgroups, our approach enforces fairness over certain carefully chosen subgroup-subsets, which substantially reduces computational cost. Furthermore, we design a new adversarial training strategy, termed *doubly regressing*, that employs a single discriminator with a weight vector over subgroup-subsets.

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

1. We introduce the **partial subgroup fairness framework**, a relaxation of full subgroup fairness that enforces fairness only on statistically significant subgroup-subsets, measured via the novel supIPM criterion.
2. We propose the **doubly regressing adversarial mechanism**, which achieves partial subgroup fairness with a single discriminator and weight vector, thereby resolving the computational inefficiency of supIPM.
3. We provide **theoretical guarantees and empirical validation**, proving equivalence bounds between $DR(f)$ and $\mathrm{supIPM}(f)$, and demonstrating superior fairness–accuracy and fairness–efficiency trade-offs on benchmark datasets compared to prior baselines.

2 Related Work

Category	**Method**
Extension of Marginal → Subgroup Fairness	Sequential correction Probabilistic bounds
Addressing Subgroup Data Sparsity	Data augmentation New model design Post-processing calibration

Table 1: Categorization of fairness research.

Recent work on algorithmic fairness has progressed along two main axes: (i) extending marginal fairness to intersectional or subgroup fairness, and (ii) addressing the subgroup data sparsity that inevitably arises in such extensions. Hu et al. Hu et al. (2024) propose a sequential framework that applies multi-marginal Wasserstein barycenters across multiple sensitive attributes, while Molina and Loiseau Molina & Loiseau (2023) demonstrate that intersectional fairness can be approximately upper bounded and controlled via marginal constraints. However, the former retains the structural limitation of "attribute-wise sequential debiasing," and the latter, despite the utility of approximate upper bounds, remains confined to marginal-based extensions.

Approaches that directly tackle sparsity fall into three categories. First, *data augmentation*: Tian et al. Tian et al. (2025) introduce *MultiFair*, which leverages mixup operations to interpolate across multiple attributes, synthesizing subgroup samples to alleviate data sparsity. Yet, such linear combinations often fail to correspond to realistic subgroup instances, raising interpretability concerns.

Second, *new model design*: Kearns et al. (2018b;a) propose rich subgroup fairness algorithms that repeatedly identify the most violated subgroup and update the classifier accordingly. While this mitigates fairness gerrymandering, such iterative subgroup-focused updates can destabilize fairness across previously satisfied groups. To address this instability, Foulds et al. Foulds et al. (2019b) redefine the fairness criterion through *differential fairness (DF)*, which constrains outcome-probability ratios across all intersectional subgroup pairs within $e^{\pm\varepsilon}$. This global ratio constraint structurally prevents fairness oscillations and ensures intersectionality preservation, such that protecting intersectional groups implies protection at the marginal level.

Third, *Post-processing calibration*: *Multicalibration* Úrsula Hébert-Johnson et al. (2018) iteratively aligns predictive probabilities with empirical frequencies across all subgroups. This "iterative boosting" perspective emphasizes that error-prone subgroups receive increased corrective weight, analogous to boosting hard-to-classify samples. In parallel, *Bayesian smoothing* Foulds et al. (2019a) stabilizes estimates in extremely sparse subgroups via prior distributions (e.g., Dirichlet) and soft counts, thereby improving estimation stability. However, such corrections, while effective for calibration, do not guarantee that the learned distribution embodies a normatively fair allocation.

To summarize, existing research falls into three strands: (i) marginal-to-intersectional relaxations with approximate guarantees (Molina & Loiseau, 2023), (ii) iterative subgroup repair (Kearns et al., 2018b;a), and (iii) sparsity mitigation through augmentation and calibrated smoothing (Tian et al., 2025; Úrsula Hébert-Johnson et al., 2018; Foulds et al., 2019a). Nevertheless, these approaches face persistent limitations, including subgroup instability, reduced interpretability of synthetic data, and the semantic gap of post-hoc calibrated probabilities.

Our approach By coupling fairness guarantees of minority subgroups with sufficiently-sampled subgroups, and by jointly enforcing *all* subgroup constraints during training, we overcome the sequential and iterative limitations of prior methods. This yields stable, intersectional fairness guarantees that remain robust even in high-dimensional subgroup spaces Shui et al. (2022); Gohar & Cheng (2023).

3 Methodology

3.1 Notation

*

We consider data points (x_i, y_i, s_i) with $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, and $s_i = (s_{i1}, \ldots, s_{iq})^\top \in \{0,1\}^q$ denoting q binary sensitive attributes. Let $S = \{0,1\}^q$ be the set of all sensitive attribute combinations, and for each $v \in S$ define

$$\mathcal{D}_v = \{i : s_i = v\}.$$

The number of subgroups is 2^q, which becomes too large so certain subgroups contain only a few samples when q is large. To mitigate this sparsity problem, we restrict attention to a collection of subgroup-subsets $\mathcal{V} = \{S_1, \ldots, S_M\}$, where each $S_m \subseteq S$ is predefined. To measure subgroup fairness, prior work has proposed distributional similarity metrics. When subgroup sample sizes are sufficiently large, the Expectation of Integral Probability Metrics (EIPM) from Kong et al. (2025) can be a natural choice, which is defined by:

$$\mathrm{EIPM}(f) = \sum_{s \in S} \pi_s \, \mathrm{IPM}(\mathbb{P}_{f,s}, \mathbb{P}_{f,\cdot}), \quad \pi_s = \frac{|\mathcal{D}_s|}{n},$$

where $\mathbb{P}_{f,s}$ and $\mathbb{P}_{f,\cdot}$ denote the empirical distribution of predictions in subgroup s and the overall distribution, respectively.

3.2 Supremum IPM for subgroup fairness

Let $\mathcal{S}_1, \ldots, \mathcal{S}_M$ be prespecified subsets of $\mathcal{S}$. Given $\mathcal{S}_1, \ldots, \mathcal{S}_M$, partial subgroup fairness requires that $\mathrm{pred}(f)_{\mathcal{S}_m} \overset{d}{\approx} \mathrm{pred}(f)_{\mathcal{S}_m^c}$ for all $m \in [M]$, where $\mathrm{pred}(f)_{\mathcal{S}'}$ for $\mathcal{S}' \subset \mathcal{S}$ is defined as $\mathrm{pred}(f)_{\mathcal{S}'} = \cup_{s \in \mathcal{S}'} \mathrm{pred}(f)_s$. Choosing $\mathcal{S}_1, \ldots, \mathcal{S}_M$ carefully so that $|\mathcal{D}_{\mathcal{S}_m}|, m \in [M]$ are sufficiently large, we can avoid over-penalization of the fairness constraint (that harms prediction accuracy) and preserve statistical guarantee of the fairness level.

An example of $\mathcal{S}_m$ is $\mathcal{S}_m = \{\mathbf{s} : s_m = 1\}$ for $m = 1 \ldots, q$. By doing so, we can control the fairness level for each sensitive attribute. Another example is the collection of all subsets of $\mathcal{S}$ whose cardinalities are greater than a prespecified positive integer n_0.

For the fairness constraint, we consider the *supIPM* which is defined as

$$\mathrm{supIPM}_d(f) = \sup_{m \in [M]} d(\mathbb{P}_{f,\mathcal{S}_m}, \mathbb{P}_{f,\mathcal{S}_m^c}) \tag{1}$$

for a given distance d between two probability measures. Examples of d are the symmetric KL divergence (reference), IPM (reference) and so....

A problem of using $s$upIPM is that computation is demanding in particular when M is large. When the symmetric KL divergence or IPM are used for d, M many discriminators should be learned to calculate $supIPM$. When d is MMD, the computational complexity becomes Mn^2 (??).

3.3 Doubly regressing approach

Doubly regressing (DR) approach is a computationally efficient adversarial learning algorithm for partial subgroup fairness. As we mentioned, using $supIPM$ would computationally demanding when M is large because M-many discriminators should be learned at each iteration of the adversarial training step. DR approach requires learning only one discriminator but guarantees partial subgroup fairness.

For a given collection of subsets $\mathcal{S}_M$, a given $\mathbf{s} \in \mathcal{S}$, and each i, define $c_i \in \{0,1\}^M$ with $c_{im} = \mathbb{I}(s_i \in \mathcal{S}_m)$. Given a predictor f, a discriminator g, and weights $\mathbf{w} \in \mathbb{R}^M$, we define

$$R^2(\mathbf{w}, g; f) = 1 - \frac{\sum_{i=1}^n \left(\mathbf{w}^\top c_i - g(f(x_i, s_i))\right)^2}{\sum_{i=1}^n \left(\mathbf{w}^\top c_i - \mu_w\right)^2}, \quad \mu_{\mathbf{w}} = \frac{1}{n}\sum_{i=1}^n \mathbf{w}^\top c_i.$$

Building on this quantity, our proposed unfairness measure is

$$DR(f) = \sup_{g \in \mathcal{G},\, \mathbf{w} \in \mathcal{W}} R^2(\mathbf{w}, g; f),$$

where $\mathcal{W}$ is the probability simplex in $\mathbb{R}^M$.

We refer to $DR(f)$ as the Doubly Regressing (DR) deviance, since it simultaneously regresses the subgroup indicator vectors c_i against the discriminator outputs $g(f(x_i, s_i))$. A larger value of $DR(f)$ indicates that the predictor f better aligns with subgroup constraints and thus satisfies partial subgroup fairness to a greater extent. In practice, we train f by minimizing the prediction loss while encouraging $DR(f)$ to be large, thereby controlling the trade-off between accuracy and fairness.

3.4 Theoretical guarantee of the fairness levels

Measure for subgroup fairness As previously discussed, we say a given model $f : \mathcal{X} \times \{0,1\}^q \to \mathbb{R}$ is subgroup-fair if $\text{pred}(f)_s = \{f(\mathbf{x}_i, \mathbf{s}_i) : i \in \mathcal{D}_s\}, s \in \{0,1\}^q$ are similar in distribution. To quantify the fairness (or unfairness) in binary classification task, we define the subgroup fairness level as follows.

Definition 3.1 (Subgroup fairness level)**.** Given a distance d between two probability measures, the subgroup fairness level of a given model $f : \mathcal{X} \times \{0,1\}^q \to \mathbb{R}$ is defined by

$$\Delta(f; d) := \sup_{s \in \{0,1\}^q} d(\mathbb{P}_f, \mathbb{P}_{f,s}). \tag{2}$$

Remark 3.2. If d is an IPM given a discriminator class $\mathcal{H}$, then $\Delta(f;d)$ is equivalent to $\sup_{h \in \mathcal{H}} \sup_{s \in \{0,1\}^q} \left|\mathbb{E}\left(h \circ f(X,S)\right) - \mathbb{E}\left(h \circ f(X,s) | S = s\right)\right|$. If $\mathcal{H} = \{\mathbb{I}(\cdot > 0)\}$, then $\Delta_{\mathcal{G}}(f)$ becomes the conventional demographic parity. If $\mathcal{H}$ is a collection of all 1-Lipschitz functions, then $\Delta_{\mathcal{G}}(f)$ becomes the Wasserstein distance.

We consider IPMs for d in this paper.

Theorem 3.3. *Fix $\epsilon \geq 0$. Let d be an IPM based on a given discriminator class $\mathcal{H}$. Then, we have*

$$\text{supIPM}_d(f) \leq \epsilon \implies \Delta(f; d) \leq C\epsilon + REMAINDER \tag{3}$$

for some constant $C > 0$ depending on M and the choice of $\{\mathcal{S}_m\}_{m=1}^M$.

Proof of Theorem 3.3. Recall that

$$\begin{aligned} \text{supIPM}_d(f) &= \sup_{m \in [M]} d(\mathbb{P}_{f,\mathcal{S}_m}, \mathbb{P}_{f,\mathcal{S}_m^c}) \\ &= \sup_{h \in \mathcal{H}} \sup_{m \in [M]} \left|\mathbb{E}(h \circ f(X,s) | s \in \mathcal{S}_m) - \mathbb{E}(h \circ f(X,s) | s \in \mathcal{S}_m^c)\right| \end{aligned} \tag{4}$$

and

$$\begin{aligned} \Delta(f; d) &= \sup_{s \in \{0,1\}^q} d(\mathbb{P}_f, \mathbb{P}_{f,s}) \\ &= \sup_{h \in \mathcal{H}} \sup_{s \in \{0,1\}^q} \left|\mathbb{E}\left(h \circ f(X,S)\right) - \mathbb{E}\left(h \circ f(X,s) | S = s\right)\right|. \end{aligned} \tag{5}$$

Hence, ...

□

For each $m \in [M]$, define $C_m := \mathbb{I}(S \in \mathcal{S}_m)$, $p_m := \mathbb{P}(S \in \mathcal{S}_m)$, and $V_m := \mathrm{Var}(C_m) = p_m(1 - p_m)$. Write $V^* := \max_m V_m$ and $p_* := \min_m V_m$ (so $0 < p_* \leq V_m \leq V^* \leq \frac{1}{4}$).

Theorem 3.4. *If $DR_{\mathcal{G}}(f) \geq \varepsilon$ for some $\varepsilon \in (0, V^*]$, then we have*

$$\mathrm{supIPM}_d(f) \leq \frac{B}{p_*}\sqrt{V^* - \varepsilon + \mathcal{E}_*^2}.$$

Proof of Theorem 3.4. Let $Z = f(X, S)$. For each m, define the Bayes estimator $\eta_m(Z) := \mathbb{P}(C_m = 1 \mid Z)$ and the error of $\mathcal{G}$ by

$$\mathcal{E}_m(\mathcal{G}) := \inf_{g \in \mathcal{G}} \|\eta_m(Z) - g(Z)\|^2, \mathcal{E}_* := \sup_{m \in [M]} \mathcal{E}_m(\mathcal{G}).$$

Since $w = e_m$ is admissible in the simplex $\mathcal{W}$, we have

$$DR_{\mathcal{G}}(f) \geq \varepsilon \implies \alpha_m^{\mathcal{G}}(f) := \inf_{g \in \mathcal{G}} \mathbb{E}(C_m - g(Z))^2 \geq \varepsilon \quad (\forall m \in [M]),$$

by the definition of the DR deviance. For any $g \in \mathcal{G}$, we can decompose

$$\mathbb{E}(C_m - g(Z))^2 = \underbrace{\mathbb{E}\left[\mathrm{Var}(C_m \mid Z)\right]}_{\text{conditional noise}} + \mathbb{E}(\eta_m(Z) - g(Z))^2$$

and hence $\alpha_m^{\mathcal{G}}(f) = \mathbb{E}\left[\mathrm{Var}(C_m \mid Z)\right] + \mathcal{E}_m(\mathcal{G})$. From $\alpha_m^{\mathcal{G}}(f) \geq \varepsilon$, we get $\mathbb{E}[\mathrm{Var}(C_m \mid Z)] \geq \varepsilon - \mathcal{E}_m(\mathcal{G})$ and thus $\mathrm{Var}(\eta_m(Z)) = V_m - \mathbb{E}\left[\mathrm{Var}(C_m \mid Z)\right] \leq V_m - \varepsilon + \mathcal{E}_m(\mathcal{G})$.

For any $h \in \mathcal{H}$,

$$\mathbb{E}[h(Z) \mid C_m = 1] - \mathbb{E}[h(Z) \mid C_m = 0] = \frac{\mathrm{Cov}(C_m, h(Z))}{V_m} = \frac{\mathrm{Cov}(\eta_m(Z), h(Z))}{V_m}.$$

By Cauchy–Schwarz and $\|h\|_{L_2} \leq B$,

$$|\,\mathrm{Cov}(\eta_m(Z), h(Z))| \leq \sqrt{\mathrm{Var}(\eta_m(Z))\ \mathrm{Var}(h(Z))} \leq B\sqrt{\mathrm{Var}(\eta_m(Z))}.$$

Therefore

$$\sup_{h \in \mathcal{H}} |\mathbb{E}[h(Z) \mid C_m = 1] - \mathbb{E}[h(Z) \mid C_m = 0]| = \sup_{h \in \mathcal{H}} \frac{|\,\mathrm{Cov}(\eta_m(Z), h(Z))|}{V_m} \leq \frac{B}{V_m}\sqrt{\mathrm{Var}(\eta_m(Z))}.$$

Plugging $(\star)$ yields

$$\sup_{h \in \mathcal{H}} |\mathbb{E}[h(Z) \mid C_m = 1] - \mathbb{E}[h(Z) \mid C_m = 0]| \leq \frac{B}{V_m}\sqrt{V_m - \varepsilon + \mathcal{E}_m(\mathcal{G})},$$

and taking $\sup_m$ together with $V_m \geq p_*$, $V_m \leq V^*$ gives the claim. □

4 Algorithm

The proposed algorithm jointly trains a prediction model, a discriminator, and subgroup weights to simultaneously achieve predictive accuracy and subgroup fairness. Specifically, the prediction model f maps features and sensitive attributes (x_i, s_i) to outcome predictions $\hat{y}_i$. The discriminator g is a function that aims to predict linear combinations of subgroup indicators, while the subgroup weight vector w represents the coefficients of these linear combinations. At each iteration, the prediction model minimizes the cross-entropy loss while simultaneously optimizing a doubly regressing loss that prevents the discriminator from distinguishing subgroup-specific linear combinations from the model outputs. The weight vector w is projected onto the probability simplex to resolve identifiability issues, and the discriminator is updated adversarially in the opposite direction, minimizing fairness violations. This iterative process yields a final predictor that balances accuracy with partial subgroup fairness guarantees.

Algorithm 1: Training with Gradient Descent for Doubly Regressing Fairness

```
Input  : Training data $\{(x_i, s_i, c_i, y_i)\}_{i=1}^n$, learning rates $(\eta_{\text{cls}}, \eta_g, \eta_w)$, number of iterations $T$,
         trade–off hyperparameter $\lambda$
Output: Classifier parameters $\theta$, discriminator $\phi$, weight vector $w$
Initialize: $\theta \leftarrow \theta_0, \phi \leftarrow \phi_0, w \leftarrow w_0$
do
    for $i = 1, \dots, n$ do
        $\hat{y}_i \leftarrow f_\theta(x_i, s_i)$
    end
    Compute the classification loss:
        $L_{\text{cls}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{CE}(\hat{y}_i, y_i)$
    Compute the fairness loss:
        $\widehat{\mathrm{DR}} = \frac{1}{n} \sum_{i=1}^{n} \big(w^\top c_i - g_\phi(\hat{y}_i)\big)^2$
    Update the discriminator:
        $\phi \leftarrow \phi + \eta_g \nabla_\phi \widehat{\mathrm{DR}}$
    Update the weights:
        $\tilde{w} \leftarrow w + \eta_w \nabla_w \widehat{\mathrm{DR}}, \quad w \leftarrow \mathrm{Proj}_{\mathcal{W}}(\tilde{w})$
    Update the classifier:
        $\theta \leftarrow \theta - \eta_{\text{cls}} \nabla_\theta L_{\text{cls}} - \lambda \eta_{\text{cls}} \nabla_\theta \widehat{\mathrm{DR}}$
until convergence or $T$ iterations;
Return $\theta, \phi, w$
```

Dataset	Type	# Features	# Sensitive attributes
UCI Adult	Tabular	14	up to 8
Communities & Crime	Tabular	128	up to 3
Law School Admission	Tabular	10	up to 4
Student Performance	Tabular	30	up to 5
CelebA	Image		up to 39

Table 2: Datasets used in experiments

5 EXPERIMENT

5.1 DATASETS

We evaluate our approach on five benchmarks spanning tabular and image domains. To study subgroup sparsity, we vary the number of sensitive attributes from two up to the maximum available categorical features in each dataset. **UCI Adult** is a census benchmark for income prediction, where we extend the usual `sex` and `race` attributes to include up to eight binarized features (e.g., education, marital status). **Communities and Crime** contains community-level demographic and socio-economic statistics for crime prediction, with race-related features as sensitive attributes. **Law School Admission** tracks student outcomes with up to four protected attributes (gender, race, income, age). **Student Performance** records secondary school data for grade prediction, with gender, age, and relationship status as sensitive features. **CelebA** is a large-scale face dataset with forty attributes, where we use binary features such as gender, age, and skin tone to assess fairness in high-dimensional image settings.

5.2 Baselines

5.3 Results

5.4 Ablation Study

6 Conclusions

References

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`.

Usman Gohar and Lu Cheng. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI-2023, pp. 6619–6627. International Joint Conferences on Artificial Intelligence Organization, August 2023. doi: 10.24963/ijcai.2023/742. URL `http://dx.doi.org/10.24963/ijcai.2023/742`.

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`.

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, 2018b. URL `https://arxiv.org/abs/1711.05144`.

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*, 47(5):3784–3795, 2025. doi: 10.1109/TPAMI.2025.3538915.

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

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, 2022. URL `https://arxiv.org/abs/2210.10837`.

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.

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`.

A APPENDIX

You may include other additional sections here.