Bayes-Optimal Fair Classification with Multiple Sensitive Features

Yi Yang[1], **Yinghui Huang**[*2], **Xiangyu Chang**[2, 3]

[1]Department of Information Systems, Arizona State University
[2]Department of Information Systems and Intelligent Business, Xi'an Jiaotong University
[3]SGIT AI Lab
yi.yang.10@asu.edu, yinghui.huang@xjtu.edu.cn, xiangyuchang@xjtu.edu.cn

Abstract

Existing theoretical work on Bayes-optimal fair classifiers usually considers a single (binary) sensitive feature. In practice, individuals are often defined by multiple sensitive features. In this paper, we characterize the Bayes-optimal fair classifier for multiple sensitive features under general approximate fairness measures, including *mean difference* (MD) and *mean ratio* (MR). We show that these approximate measures for existing group fairness notions, including Demographic Parity, Equal Opportunity, Predictive Equality, and Accuracy Parity, are linear transformations of selection rates for specific groups defined by both labels and sensitive features. We then characterize that Bayes-optimal fair classifiers for multiple sensitive features under both MD and MR become instance-dependent thresholding rules that rely on a weighted sum of these group membership probabilities. Our framework applies to both attribute-aware and attribute-blind settings and can accommodate composite fairness notions like Equalized Odds. Building on this, we propose two practical algorithms for Bayes-optimal fair classification via in-processing and post-processing. We show empirically that our methods compare favorably to existing methods.

Code — https://github.com/yhuangda/Bayes-Optimal-Fair-Classification-with-Multiple-Sensitive-Features

Extended version — https://arxiv.org/abs/2505.00631

1 Introduction

Machine learning (ML) models have become integral to decision-making processes in various high-stakes fields, such as credit scoring and criminal justice. However, a growing concern has emerged regarding the fairness of these models, particularly with respect to outputs that may disadvantage certain social groups defined by sensitive features such as race, gender, or socio-economic status (Barocas, Hardt, and Narayanan 2023). Therefore, addressing fairness issues in ML has garnered significant attention (Mehrabi et al. 2021; Caton and Haas 2024).

A considerable body of work has focused on fairness in classification settings, where specific groups may experience discrimination due to biased predictions. This has led to the formalization of several algorithmic fairness notions, such as *Demographi Parity* (Dwork et al. 2012), *Equal Opportunity* (Hardt, Price, and Srebro 2016), and *Accuracy Parity* (Zafar et al. 2017a). These notions aim to equalize various quantities across different groups. While *perfect fairness*–ensuring exactly identical quantities across groups–may entirely eliminate discrimination, it often incurs significant efficiency loss and may even be infeasible on finite data (Agarwal et al. 2018; Makhlouf, Zhioua, and Palamidessi 2021; Pinzón et al. 2022). Thus, *approximate fairness* is frequently adopted as a more practical alternative, where fairness level is quantified and limited using approximate measures such as *Mean Difference* (Chai and Wang 2022) and *Mean Ratio* (Menon and Williamson 2018) derived from fairness notions. See Section 3 for their definitions.

Researchers have developed various fair ML algorithms to operationalize these fairness notions, which are typically categorized into three groups: pre-processing, in-processing, or post-processing (Caton and Haas 2024). Pre-processing methods aim to reduce bias in the training data through techniques such as data cleaning or reweighting (Kamiran and Calders 2012; Calmon et al. 2017) before applying classical ML algorithms, but fairness in the training data does not always guarantee fairness in the resulting models. In-processing methods modify the model training objective by adding fairness regularizers or incorporating fairness constraints (Zemel et al. 2013; Agarwal et al. 2018; Zafar et al. 2017b; Yang, Cisse, and Koyejo 2020; Zhao et al. 2020). Post-processing (Menon and Williamson 2018; Gouic, Loubes, and Rigollet 2020; Xian, Yin, and Zhao 2023; Chen, Klochkov, and Liu 2024; Xian and Zhao 2024; Wei, Ramamurthy, and Calmon 2021; Cruz and Hardt 2024) remaps the model's outputs to satisfy fairness requirements.

Despite these advancements, foundational theoretical aspects of fair ML remain under-explored. One critical question concerns the characterization of Bayes-optimal classifiers in fair ML. A Bayes-optimal fair classifier minimizes classification risk while satisfying specific fairness constraints, serving as a theoretical benchmark or the "best possible" classifier for a given fairness-aware problem. Although Menon and Williamson (2018) and Chzhen et al. (2019) characterized Bayes-optimal fair classifiers, their analyses are limited to a single sensitive feature of binary values. This leaves more complex and realistic settings involving multiple sensitive

*Corresponsing author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

features[1] largely unaddressed. While several studies (Corbett-Davies et al. 2017; Schreuder and Chzhen 2021; Zeng, Cheng, and Dobriban 2024) have investigated the theoretical underpinnings of fair classification with multiple sensitive features, their works derive Bayes-optimal classifiers under the strict requirement of perfect fairness without addressing the practical requirement of approximate fairness. More recently, Chen, Klochkov, and Liu (2024) and Xian and Zhao (2024) extended the exploration of Bayes-optimal fair classifiers to approximate fairness settings with multiple sensitive features. However, their work focuses exclusively on post-processing algorithms for fair classification, and modifying model outputs in this manner may raise legal concerns (Caton and Haas 2024; Barocas and Selbst 2016). Their works are also restricted to the mean difference measure and fail to accommodate fairness notions like *accuracy parity*. For a summary and comparison with related work, see Table 1, and a more detailed discussion is provided in Appendix A of the supplementary material.

Therefore, it lacks a systematic approach for deriving Bayes-optimal fair classifiers, especially with multiple sensitive features under general approximate fairness measures. To this end, we explore their form while also explicitly addressing fairness notions such as accuracy parity. Our contributions are listed as follows:

- We characterize the form of Bayes-optimal fair classifiers for multiple sensitive features under both MD and MR measures, generalizing the framework of Menon and Williamson (2018). Their work can be viewed as a special case of our approach when restricted to a single (binary) sensitive feature.
- Our characterization accommodates fairness notions such as accuracy parity, whose Bayes-optimal fair classifier, to the best of our knowledge, has not been established before.
- Building on theoretical results, we propose both in-processing and post-processing algorithms to recover Bayes-optimal fair classifiers, offering flexibility in when to apply fairness interventions.

2 Background and Notation

2.1 Binary Classification

A binary classification problem is defined by a joint distribution $\mathcal{D}$ over input features $X \in \mathcal{X}$ and labels $Y \in \mathcal{Y} = \{0, 1\}$. The goal is to derive a measurable *randomized classifier* parametrized by $f : \mathcal{X} \to [0, 1]$, which outputs a prediction $\hat{Y}_f \in \{0, 1\}$ with a certain probability based on features X. Let $\mathrm{Bern}(p)$ be the Bernoulli distribution with success probability $p \in [0, 1]$, and let $\mathcal{F}$ denote the set of all such measurable functions f. Then, the randomized classifier $f \in \mathcal{F}$ specifies, for any $x \in \mathcal{X}$, the probability $f(x)$ of predicting $\hat{Y}_f = 1$ given $X = x$, i.e., $\hat{Y}_f \mid X = x \sim \mathrm{Bern}(f(x))$.

Typically, the quality of a classifier is evaluated using a statistical risk function $R(\cdot; \mathcal{D}) : \mathcal{F} \to \mathbb{R}_+$. A canonical risk is the cost-sensitive risk (Menon and Williamson 2018).

[1]Alternatively, a multi-class sensitive feature–see Section 2.2 for details on their connection.

Definition 1 (Cost-Sensitive Risk) *For a cost parameter $c \in [0, 1]$ and a classifier f, the cost-sensitive risk (of f) is given by:*

$$R_{cs}(f; c) = (1 - c) \cdot P(\hat{Y}_f = 0, Y = 1) + c \cdot P(\hat{Y}_f = 1, Y = 0). \quad (1)$$

The cost-sensitive risk allows for asymmetric penalization of false negatives and false positives, depending on the value of c. When $c = 0.5$, it reduces to the conventional error rate.

Bayes-Optimal Classifiers: For a given problem, the Bayes-optimal classifier is theoretically the best method, achieving the lowest possible average risk. For the cost-sensitive risk with parameter c, a Bayes-optimal classifier is defined as any minimizer $f^* \in \mathrm{argmin}_{f \in \mathcal{F}} R_{cs}(f; c)$. Let $\eta(x) \coloneqq P(Y = 1 \mid X = x)$ be the posterior probability of the positive class given x, and $\mathbb{1}[\cdot]$ denote the indicator function (equal to 1 if the argument is true and 0 otherwise). Then, Elkan (2001) characterizes Bayes-optimal classifiers as having the form of

$$f^*(x) = \mathbb{1}\left[H(x) > 0\right] + \alpha \cdot \mathbb{1}\left[H(x) = 0\right], \quad (2)$$

for all $x \in \mathcal{X}$, where $H(x) = \eta(x) - c$, and $\alpha \in [0, 1]$ is an arbitrary parameter. This shows that the Bayes-optimal classifier operates as a *thresholding rule* on the posterior class-probability of an instance. It makes predictions based on the threshold defined by the cost parameter c.

2.2 Fairness-Aware Learning in Binary Classification

Fairness-aware learning extends the conventional binary classification problem by incorporating sensitive features in addition to the target feature Y. Specifically, we assume the presence of sensitive features $A \in \mathcal{A}$ (e.g., gender and race) with respect to which we aim to ensure fairness. We note that X may or may not include the sensitive features A in practical applications.

Group Notation with Multiple Sensitive Features: In real applications, individuals might be coded with multiple sensitive features. We consider K sensitive features, where each feature is denoted by $A_k \in \mathcal{A}_k$ for $k \in [K]$.[2] For example, A_1 might correspond to race, A_2 to gender, and so on. However, the presence of multiple sensitive features (e.g., race and gender simultaneously) can lead to non-equivalent definitions of group fairness (Yang, Cisse, and Koyejo 2020):

- *Independent group fairness*: Fairness is evaluated separately for each sensitive feature, leading to overlapping subgroups (i.e., each sensitive feature defines its own set of groups independently).[3]
- *Intersectional group fairness*: Fairness is enforced on all subgroups defined by intersections of sensitive features, resulting in non-overlapping groups associated with all possible combinations of sensitive features.

[2]Here, $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_K$.

[3]See Appendix B.1 of the supplementary material for detailed explanations and examples.

References	Corbett-Davies et al. (2017)	Schreuder and Chzhen (2021)	Agarwal et al. (2018)	Chen, Klochkov, and Liu (2024)	Xian and Zhao (2024)	Zeng, Cheng, and Dobriban (2024)	Ours
SCOPE OF THEORETICAL FRAMEWORK							
Approximate Fairness (MD)			✓	✓	✓		✓
Approximate Fairness (MR)							✓
Attribute-Blind Setting[4]			✓	✓	✓	✓	✓
FAIRNESS METRICS CONSIDERED							
Demographic Parity	✓	✓	✓	✓	✓	✓	✓
Equal Opportunity	✓			✓	✓	✓	✓
Predictive Equality	✓					✓	✓
Accuracy Parity							✓
Equalized Odds			✓	✓	✓	✓	✓
THEORETICALLY OPTIMAL ALGORITHMS							
In-processing			✓			✓	✓
Post-processing	✓	✓		✓	✓	✓	✓

Table 1: Comparison with prior works on Bayes-optimal fair classifier with multiple sensitive features.

It is noteworthy that enforcing intersectional fairness inherently controls independent fairness, but the reverse does not always hold (Kearns et al. 2018). Thus, intersectional fairness is often considered ideal (Yang, Cisse, and Koyejo 2020). Consequently, we focus on intersectional fairness here when addressing multiple sensitive features and also extend our results to independent fairness in Appendix B.2 of the supplementary material.

To implement intersectional fairness for multiple sensitive features, a new composite sensitive feature S is constructed to represent all possible intersectional combinations of the existing sensitive features. Specifically, $S \in \mathcal{S} = \{1, \ldots, M\}$, where $M = \prod_{k=1}^{K} |\mathcal{A}_k|$, and $|\mathcal{A}_k|$ denotes the number of possible values for the k-th sensitive feature. Thus, S defines M non-overlapping subgroups, each corresponding to a unique combination of sensitive feature values. Note that this approach is equivalent to treating S as a single sensitive feature with multiple categorical values, enabling our results to be directly applicable to that scenario.

For all $m \in \mathcal{S}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, let $P_{S,Y}(m, y) := P(S = m, Y = y)$ denote the joint distribution of S and Y. Define $p^+ := P(Y = 1)$ and $p^- := P(Y = 0)$ as the marginal probabilities of the positive and negative classes. Let $P_S(\cdot)$ represent the marginal distribution of S, while $P_{S|Y=y}(\cdot)$ and $P_{Y|S=m}(\cdot)$ denote the conditional distributions of S given $Y = y$ and Y given $S = m$, respectively.

To address unfairness, various parity-based group fairness notions grounded in sensitive features have been proposed. Below are the key definitions considered in this paper.

Definition 2 (Demographic Parity (DP)) *(Dwork et al. 2012) A classifier f satisfies DP if its prediction $\hat{Y}_f$ is independent of the sensitive feature S: $P(\hat{Y}_f = 1) = P(\hat{Y}_f = 1 \mid S = m)$ for all $m \in [M]$.*

[4]Attribute-Blind Setting refers to the case where sensitive features cannot be used for prediction.

Definition 3 (Equal Opportunity (EO)) *(Hardt, Price, and Srebro 2016) A classifier f satisfies EO if it achieves the same true positive rate across all groups: $P(\hat{Y}_f = 1 \mid Y = 1) = P(\hat{Y}_f = 1 \mid S = m, Y = 1)$ for all $m \in [M]$.*

Definition 4 (Predictive Equality (PE)) *(Corbett-Davies et al. 2017) A classifier f satisfies PE if it achieves the same false positive rate across all groups: $P(\hat{Y}_f = 1 \mid Y = 0) = P(\hat{Y}_f = 1 \mid S = m, Y = 0)$ for all $m \in [M]$.*

Definition 5 (Accuracy Parity (AP)) *(Zafar et al. 2019) A classifier f satisfies AP if it achieves the same error rate across all groups: $P(\hat{Y}_f \neq Y) = P(\hat{Y}_f \neq Y \mid S = m)$ for all $m \in [M]$.*

In practice, perfect fairness (i.e., achieve equalities above) often leads to significant efficiency loss (i.e., higher expected risk) or even is infeasible (Makhlouf, Zhioua, and Palamidessi 2021). Thus, "approximate" fairness is usually more practical and preferable. Previous research typically quantifies fairness by measuring disparities in quantities that would be equalized under perfect fairness, focusing on optimizing risk while imposing constraints to limit these disparities (Zafar et al. 2017a).

3 General Approximate Fairness Measures

We focus on two general approximate fairness measures, *mean difference* and *mean ratio*, to quantify classifier disparity level. We begin by presenting the definitions of these measures and then demonstrate both of them are linear transformations of a classifier's selection rates for specific groups.

3.1 Mean Difference

For a composite sensitive feature $S \in \{1, \ldots, M\}$, the *mean difference* (MD) score (Chai and Wang 2022; Calders and

Notion	$\mathcal{G}(\hat{Y}_f)$	Z	z	a_m	b_m^y	c_m^y (MD)	c_m^y (MR)
DP	$\{\hat{Y}_f = 1\}$	$\mathbb{U}$[5]	$\mathbb{U}$	$P_S(m)$	$P_{Y\|S=m}(y)$	0	0
EO	$\{\hat{Y}_f = 1\}$	Y	1	$P_{S\|Y=1}(m)$	y	0	0
PE	$\{\hat{Y}_f = 1\}$	Y	0	$P_{S\|Y=0}(m)$	$1-y$	0	0
AP	$\{\hat{Y}_f \neq Y\}$	$\mathbb{U}$	$\mathbb{U}$	$P_S(m)$	$(1-2y)\cdot P_{Y\|S=m}(y)$	$(1-y)p^+ - yP_{Y\|S=m}(1)$	$(y-1)\delta p^+ + yP_{Y\|S=m}(1)$

Table 2: Recovering existing fairness criteria based on the choice of $\mathcal{G}(\cdot)$, Z, and z for MD and MR measures. For parameter values in Lemmas 1 and 2, MD and MR measures differ only in c_m^y for AP.

Verwer 2010) quantifies the fairness of a classifier f by calculating the difference in a specified outcome between the overall population and the subgroup defined by $S = m$.

Definition 6 (Mean Difference) *For $\forall m \in [M]$, the mean difference measure for group m is defined as:*

$$\mathrm{MD}_m(f) = P(\mathcal{G}(\hat{Y}_f) \mid Z = z) - P(\mathcal{G}(\hat{Y}_f) \mid Z = z, S = m),$$

where $\hat{Y}_f$ is the prediction of f, and the components $\mathcal{G}(\cdot)$, Z, and z depend on the fairness notion being considered.

The flexibility in the choice of $\mathcal{G}(\cdot)$, Z, and z allows Definition 6 to accommodate several commonly used group fairness notions, as shown in Table 2. Achieving perfect fairness indicates $\mathrm{MD}_m(f) = 0$ for all m. Usually, a limited level of disparity may be acceptable. To formalize this, we use the symmetrized version of the MD measure:

$$\mathrm{MD}(f) = \max_{m\in[M]} \max\left(\mathrm{MD}_m(f), \mathrm{MD}_m(1-f)\right) \leq \delta, \quad (3)$$

where δ is a pre-specified tolerance level for unfairness.

To simplify notation, we define $E_{y,m} = \{Y = y, S = m\}$ as the event where an individual has label $Y = y$ and belongs to group m, with its probability denoted by $P(E_{y,m}) = P(Y = y, S = m)$. Then, Lemma 1 shows that MD measures for these common group fairness notions are linear transformations of $P(\hat{Y}_f = 1 \mid E_{y,m})$. All proofs are deferred to Appendix C of the supplementary material.

Lemma 1 *For any randomized classifier f, any $\delta \in [0, 1]$, and the group fairness notions in Table 2, $\mathrm{MD}(f) \leq \delta \Leftrightarrow R_m^{\mathrm{MD}}(f) \in [-\delta, \delta]$ for all $m \in [M]$, where*

$$R_m^{\mathrm{MD}}(f) := \sum_{y\in\{0,1\}} \left\{ \Big[\sum_{m'=1}^{M} a_{m'} b_{m'}^y P(\hat{Y}_f = 1 \mid E_{y,m'}) \Big] - b_m^y P(\hat{Y}_f = 1 \mid E_{y,m}) + c_m^y \right\}.$$

Here, values of a_m, b_m^y, and c_m^y depend on the chosen fairness notion and are as defined in Table 2.

3.2 Mean Ratio

Approximate fairness can also be assessed using the *disparate impact* factor (Feldman et al. 2015; Menon and Williamson 2018), which is defined as the ratio of relevant probabilities. We refer to this as the *mean ratio* (MR) measure, shown below.

Definition 7 (Mean Ratio) *For $\forall m \in [M]$, the mean ratio measure for group m is defined as:*

$$\mathrm{MR}_m(f) = \frac{P(\mathcal{G}(\hat{Y}_f) \mid Z = z, S = m)}{P(\mathcal{G}(\hat{Y}_f) \mid Z = z)},$$

where $\hat{Y}_f$, $\mathcal{G}(\cdot)$, Z, and z are as defined in Definition 6.

Similarly, we consider the symmetrized version of the MR measure (Menon and Williamson 2018):

$$\mathrm{MR}(f) = \min_{m\in[M]} \min\left(\mathrm{MR}_m(f), \mathrm{MR}_m(1-f)\right) \geq \delta. \quad (4)$$

Then, Lemma 2 demonstrates that MR measures for common group fairness notions are also linear transformations of $P(\hat{Y}_f = 1 \mid E_{y,m})$.

Lemma 2 *For any randomized classifier f, any $\delta \in [0, 1]$, and the group fairness notions in Table 2, $\mathrm{MR}(f) \geq \delta \Leftrightarrow R_m^{\mathrm{MR}}(f) \in [\delta - 1, 0]$ for all $m \in [M]$, where*

$$R_m^{\mathrm{MR}}(f) := \sum_{y\in\{0,1\}} \left\{ \Big[\delta \sum_{m'=1}^{M} a_{m'} b_{m'}^y P(\hat{Y}_f = 1 \mid E_{y,m'}) \Big] - b_m^y P(\hat{Y}_f = 1 \mid E_{y,m}) + c_m^y \right\}.$$

Here, values of a_m, b_m^y, and c_m^y depend on the chosen fairness notion and are as defined in Table 2.

4 Bayes-Optimal Fair Classifiers

Given approximate fairness constraints in (3) or (4), our goal is to find a (randomized) fair classifier f_B^* optimizing the following problem: $\min_{f\in\mathcal{F}} \{R_{cs}(f; c) : \mathrm{MD}(f) \leq \delta\}$ for MD, or $\min_{f\in\mathcal{F}} \{R_{cs}(f; c) : \mathrm{MR}(f) \geq \delta\}$ for MR. Note that these constrained optimization problems can be further reduced to the following unconstrained problems via the Lagrangian principle and Lemmas 1 and 2.

Lemma 3 *For any $c \in [0, 1]$ and $\delta \in [0, 1]$, there exists $\boldsymbol{\lambda} \in \mathbb{R}^M$ such that:*

- *For MD:* $\min_{f\in\mathcal{F}} \{R_{cs}(f; c) : \mathrm{MD}(f) \leq \delta\} = \min_{f\in\mathcal{F}} \left(R_{cs}(f; c) - \sum_{m=1}^{M} \lambda_m \cdot R_m^{\mathrm{MD}}(f)\right)$.
- *For MR:* $\min_{f\in\mathcal{F}} \{R_{cs}(f; c) : \mathrm{MR}(f) \geq \delta\} = \min_{f\in\mathcal{F}} \left(R_{cs}(f; c) - \sum_{m=1}^{M} \lambda_m \cdot R_m^{\mathrm{MR}}(f)\right)$.

[5] $\mathbb{U}$ refers to the complete set.

Here, λ_m is the m-th component of $\boldsymbol{\lambda}$.

Lemma 3 shows that Bayes-optimal fair classifiers can be derived by solving an unconstrained optimization problem with a fairness regularizer incorporated into the objective. The trade-off parameter vector $\boldsymbol{\lambda}$ controls the balance between cost-sensitive risk (efficiency) and fairness. In fact, each of its component $\lambda_m \in \mathbb{R}$ corresponds to the difference in Lagrange multipliers for the two bounds associated with group m, and it can take negative values. With these foundations in place, we now present the form of Bayes-optimal fair classifiers for MD and MR measures.

4.1 Mean Difference

We begin with the explicit form of the Bayes-optimal fair classifier for MD measure. Recall that $\eta(x) := P(Y = 1 \mid X = x)$.

Theorem 1 (Bayes-Optimal Fair Classifier for MD) *For any $c \in [0,1]$ and $\delta \in [0,1]$, $\exists \boldsymbol{\lambda} \in \mathbb{R}^M$ such that the Bayes-optimal fair classifier $f_B^*(x) \in \operatorname{argmin}_{f\in\mathcal{F}} \{R(f) : \mathrm{MD}(f) \le \delta\}$ has the form of*

$$f_B^*(x) = \mathbb{1}\left[H_B^*(x) > 0\right] + \alpha \cdot \mathbb{1}\left[H_B^*(x) = 0\right], \quad (5)$$

where

$$H_B^*(x) = \eta(x) - c - \sum_{m=1}^{M} \sum_{y\in\{0,1\}} b_m^y \left(\lambda_m - \Lambda_M a_m\right) \gamma_m^y(x).$$

Here, λ_m is the m-th component of $\boldsymbol{\lambda}$, $\Lambda_M = \sum_{i=1}^M \lambda_m$, $\gamma_m^y(x) = \frac{P(E_{y,m}|X=x)}{P(E_{y,m})}$, and $\alpha \in [0,1]$ is an arbitrary parameter. The values of a_m and b_m^y depend on the fairness notion under consideration and are as shown in Table 2.

In (5), setting $\boldsymbol{\lambda} = \mathbf{0}$ results in the unconstrained Bayes-optimal classifiers for the cost-sensitive risk, as described in (2). For $\boldsymbol{\lambda} \neq \mathbf{0}$, the optimal classifier $f_B^*(x)$ adjusts the $\boldsymbol{\lambda} = \mathbf{0}$ solution by applying an instance-dependent threshold correction. This correction is determined by the weighted sum of $\gamma_m^y(x)$—the (normalized) probability that the individual x belongs to the group $\{Y = y, S = m\}$.

In the discussion above, we made no explicit assumption regarding whether the sensitive features are utilized during the prediction phase. Thus, the findings are applicable to the attribute-blind setting. If the sensitive features are available and allowed to be used for prediction,[6] the form of the Bayes-optimal fair classifier simplifies as follows:

Corollary 1 (Bayes-Optimal Fair Classifier for MD-S) *For any $c \in [0,1]$ and $\delta \in [0,1]$, $\exists \boldsymbol{\lambda} \in \mathbb{R}^M$ such that the Bayes-optimal fair classifier $f_B^*(x,s) \in \operatorname{argmin}_{f\in\mathcal{F}} \{R(f) : \mathrm{MD}(f) \le \delta\}$ has the form of*

$$f_B^*(x,s) = \mathbb{1}\left[H_B^*(x,s) > 0\right] + \alpha \cdot \mathbb{1}\left[H_B^*(x,s) = 0\right], \quad (6)$$

where

$$H_B^*(x,s) = \eta(x,s) - c - \sum_{y\in\{0,1\}} b_s^y \left(\lambda_s - \Lambda_M a_s\right) \gamma_s^y(x,s).$$

Here, $\eta(x,s) = P(Y = 1 \mid X = x, S = s)$, λ_s is the s-th component of $\boldsymbol{\lambda}$, $\Lambda_M = \sum_{i=1}^M \lambda_m$, and $\gamma_s^y(x,s) = \frac{P(Y=y|X=x,S=s)}{P(E_{y,s})}$. $\alpha \in [0,1]$ is an arbitrary parameter. The values of a_s and b_s^y depend on the selected fairness notion.

This result follows directly from Theorem 1, since for the data pair (x,s), we have $P(S = m \mid X = x, S = s) = \mathbb{1}[m = s]$ and $P(S = m, Y \mid X = x, S = s) = P(Y \mid X = x, S = s)\mathbb{1}[m = s]$. Note that in this case, (6) can further reduce to applying a group-wise constant threshold to the class probabilities $\eta(x,s)$ for each value of the sensitive feature. This simplification arises because $\gamma_s^y(x,s)$ is a linear function of $\eta(x,s)$ across all four fairness notions.

4.2 Mean Ratio

We now turn to the Bayes-optimal fair classifier for the MR measure. The result is analogous to Theorem 1, but it explicitly incorporates δ in the threshold correction.

Theorem 2 (Bayes-Optimal Fair Classifier for MR) *For any $c \in [0,1]$ and $\delta \in [0,1]$, $\exists \boldsymbol{\lambda} \in \mathbb{R}^M$ such that the Bayes-optimal fair classifier $f_B^*(x) \in \operatorname{argmin}_{f\in\mathcal{F}} \{R(f) : \mathrm{MR}(f) \ge \delta\}$ has the form of*

$$f_B^*(x) = \mathbb{1}\left[H_B^*(x) > 0\right] + \alpha \cdot \mathbb{1}\left[H_B^*(x) = 0\right], \quad (7)$$

where

$$H_B^*(x) = \eta(x) - c - \sum_{m=1}^{M} \sum_{y\in\{0,1\}} b_m^y \left(\lambda_m - \delta \cdot \Lambda_M a_m\right) \gamma_m^y(x).$$

Here, λ_m is the m-th component of $\boldsymbol{\lambda}$, $\Lambda_M = \sum_{i=1}^M \lambda_m$, $\gamma_m^y(x) = \frac{P(E_{y,m}|X=x)}{P(E_{y,m})}$, and $\alpha \in [0,1]$ is an arbitrary parameter. The values of a_m and b_m^y depend on the fairness notion under consideration and are as shown in Table 2.

When S is available for prediction, the Bayes-optimal fair classifier for MR, similar to the MD case, takes a simplified form as detailed in Corollary 2.

Corollary 2 (Bayes-Optimal Fair Classifier for MR-S) *For any $c \in [0,1]$ and $\delta \in [0,1]$, $\exists \boldsymbol{\lambda} \in \mathbb{R}^M$ such that the Bayes-optimal fair classifier $f_B^*(x,s) \in \operatorname{argmin}_{f\in\mathcal{F}} \{R(f) : \mathrm{MR}(f) \ge \delta\}$ has the form of*

$$f_B^*(x,s) = \mathbb{1}\left[H_B^*(x,s) > 0\right] + \alpha \cdot \mathbb{1}\left[H_B^*(x,s) = 0\right], \quad (8)$$

where

$$H_B^*(x,s) = \eta(x,s) - c - \sum_{y\in\{0,1\}} b_s^y \left(\lambda_s - \delta \cdot \Lambda_M a_s\right) \gamma_s^y(x,s).$$

Here, $\eta(x,s) = P(Y = 1 \mid X = x, S = s)$, λ_s is the s-th component of $\boldsymbol{\lambda}$, $\Lambda_M = \sum_{i=1}^M \lambda_m$, and $\gamma_s^y(x,s) = \frac{P(Y=y|X=x,S=s)}{P(E_{y,s})}$. $\alpha \in [0,1]$ is an arbitrary parameter. The values of a_s and b_s^y depend on the selected fairness notion.

Similar to the MD case, (8) can further reduce to applying a group-wise constant threshold to the class probabilities $\eta(x,s)$ for each value of the sensitive feature.

Remark 1 *In addition to fairness notions discussed in Table 2, our results can be directly extended to composite fairness notions like Equalized Odds (Hardt, Price, and Srebro 2016). See Appendix B.3 of the supplementary material for more details.*

[6] This refers to the Attribute-Aware Setting, i.e., A (and thus S) are included in X. In what follows, we slightly abuse notation by separating A (S) from X, with X representing only the non-sensitive features.

5 Algorithms

Section 4 establishes that Bayes-optimal fair classifiers for MD and MR measures apply instance-dependent threshold corrections to the unconstrained Bayes-optimal classifier. This insight facilitates practical training on finite data using in-processing and post-processing techniques.

5.1 In-Processing-Based Bayes-Optimal Fair Classification

We first introduce an in-processing method. Theorems 1 and 2 show that Bayes-optimal fair classifiers apply instance-dependent threshold adjustments. As cost-sensitive classification inherently accounts for such threshold adjustments, it forms the basis of our approach. We propose a fair cost-sensitive classification framework by first defining a fair cost-sensitive risk function and then demonstrating that minimizing this risk yields the Bayes-optimal fair classifier, as shown in the following theorem.

Theorem 3 *Let* $c_y^{\boldsymbol{\lambda}}(x) = (1-2y)\left[c + Q^{\boldsymbol{\lambda}}(x)\right] + y$, *where* $y \in \{0,1\}$, *and:*

$$Q^{\boldsymbol{\lambda}}_{\mathrm{MD}}(x) = \sum_{m=1}^{M} \sum_{y\in\{0,1\}} b_m^y \left(\lambda_m - \Lambda_M a_m\right) \gamma_m^y(x), \quad (9)$$

$$Q^{\boldsymbol{\lambda}}_{\mathrm{MR}}(x) = \sum_{m=1}^{M} \sum_{y\in\{0,1\}} b_m^y \left(\lambda_m - \delta\cdot\Lambda_M a_m\right) \gamma_m^y(x). \quad (10)$$

Here, $\boldsymbol{\lambda}$, λ_m, Λ_M, $\gamma_m^y(x)$, *and the values of* a_m *and* b_m^y *are as defined in Theorems 1.*

Define the fair cost-sensitive risk of a classifier f *as:*

$$R^{\boldsymbol{\lambda}}_{FCS}(f) = \sum_{y\in\{0,1\}} \Big\{ \int_{\mathcal{X}} c_y^{\boldsymbol{\lambda}}(x) \cdot P(\hat{Y}_f = 1-y, Y = y \mid X = x)\, dP_X(x) \Big\}. \quad (11)$$

Then, $f_B^{In}(x) = \operatorname{argmin}_{f\in\mathcal{F}} R^{\boldsymbol{\lambda}}_{FCS}(f)$ *is a Bayes-optimal fair classifier.*

Theorem 3 shows that a cost-sensitive classification with instance-dependent costs can yield a Bayes-optimal fair classifier. Building on this, Algorithm 1 outlines the proposed in-processing procedure.[7]

Remark 2 *On the line 1 of Algorithm 1, the group membership probabilities are estimated. This can be done directly by learning a multi-class predictor* $\hat{f}_{S,Y} : \mathcal{X} \to \Delta^{\mathcal{S}\times\mathcal{Y}}$. *Or one can break the problem down into learning two simple predictors,* $\hat{f}_Y : \mathcal{S}\times\mathcal{X} \to \Delta^{\mathcal{Y}}$ *and* $\hat{f}_S : \mathcal{X} \to \Delta^{\mathcal{S}}$, *then combining them into* $\hat{f}_{S,Y}(m,y) = \hat{f}_S(x)_m \cdot \hat{f}_Y(m,x)_y$.

When S is available for prediction, the construction of the Bayes-optimal fair cost-sensitive classifier can be further simplified. See Appendix C.8 of the supplementary material for details.

[7] We assume that S has already been constructed from A, and so have its observed values s_i.

Algorithm 1: Bayes-Optimal Fair Classification via In-Processing

Input: Cost parameter c, fairness tolerance level $\delta \geq 0$, dataset $D = \{x_i, s_i, y_i\}_{i=1}^N$, and trade-off parameter $\boldsymbol{\lambda}$.

1: Estimate the group membership probabilities $\tilde{P}(S, Y \mid X = x)$ using $\{(x_i, s_i, y_i)\}_{i=1}^N$, and calculate $\hat{\gamma}_m^y(x) = \frac{\tilde{P}(E_{y,m}|X=x)}{\tilde{P}(E_{y,m})}$ accordingly for all $m \in [M]$ and $y \in \{0,1\}$.

2: Estimate $\hat{Q}^{\boldsymbol{\lambda}}$ by plug-in estimation using $\hat{\gamma}_m^y$ and the expressions in (9) for MD or (10) for MR.

3: Denote $\hat{c}_y^{\boldsymbol{\lambda}} = (1-2y)\left[c + \hat{Q}^{\boldsymbol{\lambda}}\right] + y$ for all y.

4: Use any cost-sensitive classification method to train $\hat{f}_B^*(x)$ on $\{x_i, y_i\}_{i=1}^N$ by minimizing the empirical analogue of fair cost-sensitive risk in (11).

Output: $\hat{f}_B^*(x)$.

5.2 Post-Processing-Based Bayes-Optimal Fair Classification

Recall that Theorems 1 and 2 show that the Bayes-optimal fair classifiers $f_B^*(x)$ modify the unconstrained Bayes-optimal classifier by applying an instance-dependent threshold correction. This correction depends on the (normalized) group membership probabilities of the individual x, given by $\gamma_m^y(x)$. Motivated by this, we propose a post-processing algorithm that adopts a plugin approach for Bayes-optimal fair classification. Specifically, we construct a fair plugin classifier by separately estimating $\eta(x)$ and $\gamma_m^y(x)$, and then combining them according to (5) and (7). See Appendix E.3 for discussion regarding estimation probability calibration.

Algorithm 2 outlines the proposed post-processing plug-in approach. When S is available for prediction, we can estimate $\hat{\eta}(x,s) = \tilde{P}(Y = 1 \mid X = x, S = s)$ using any method applied to the dataset $\{(x_i, s_i, y_i)\}_{i=1}^N$ instead of $\{(x_i, y_i)\}_{i=1}^N$ as described in Line 1 of Algorithm 2.

Algorithm 2: Bayes-Optimal Fair Classification via Post-Processing

Input: Cost parameter c, fairness tolerance level $\delta \geq 0$, dataset $D = \{x_i, s_i, y_i\}_{i=1}^N$, and trade-off parameter $\boldsymbol{\lambda}$.

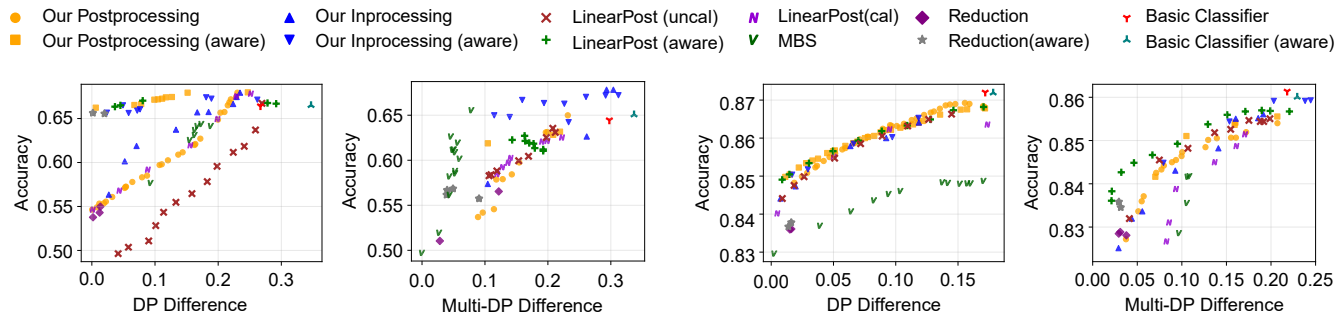
1: Estimate $\hat{\eta}(x) = \tilde{P}(Y = 1 \mid X = x)$ using any approach on $\{(x_i, y_i)\}_{i=1}^N$.

2: Estimate the group membership probabilities $\tilde{P}(S, Y \mid X = x)$ using $\{(x_i, s_i, y_i)\}_{i=1}^N$, and calculate $\hat{\gamma}_m^y(x) = \frac{\tilde{P}(E_{y,m}|X=x)}{\tilde{P}(E_{y,m})}$ accordingly for all $m \in [M]$ and $y \in \{0,1\}$.

3: Construct $\hat{f}_B^*(x)$ by plugging the estimates $\hat{\eta}(x)$ and $\hat{\gamma}_m^y(x)$ into the expression for the Bayes-optimal fair classifier: Use (5) for MD or (7) for MR.

Output: $\hat{f}_B^*(x)$.

Remark 3 *In Algorithms 1 and 2, the value of* $\boldsymbol{\lambda}$ *plays a critical role in balancing fairness and risk. It can be se-*



(a) COMPAS (b) Adult

Figure 1: Trade-offs between accuracy and fairness using MD. The prefix 'Multi-' represents the case of multiple sensitive features, and '(aware)' in classifier names indicates the attribute-aware setting. (uncal): uncalibrated; (cal): calibrated.

lected through various approaches. For example, a rough estimate can be obtained through grid search by ensuring that the achieved fairness level satisfies the pre-specified threshold (Menon and Williamson 2018; Chen, Klochkov, and Liu 2024). For a more accurate or efficient determination, optimization techniques based on the relationship between $\boldsymbol{\lambda}$ and the Lagrange multipliers can be employed, like solving a dual optimization problem (Xian and Zhao 2024) or using iterative updates guided by the fairness-accuracy trade-off (Zeng, Cheng, and Dobriban 2024). See Appendix F for more details on the dual update procedure.

Remark 4 *Our algorithms are estimator-agnostic, existing sparsity or calibration remedies can plug in directly when estimating group membership probability. When K is large, intersectional sparsity may occurs; see Appendix B.1 for discussion and its fixes. For many groups/high-dim X, scalable estimators (Zhang and Liu 2014; Liu et al. 2018) can help for efficiency.*

6 Experiments

Setup. We consider two real-world benchmark classification datasets that are widely used in the fair ML literature: Adult (Becker and Kohavi 1996) and COMPAS (Angwin et al. 2016). Both datasets contain demographic features such as *gender* and *race*. We use them to construct scenarios with a single sensitive feature as well as multiple sensitive features. We compare the proposed algorithms with established post-processing and in-processing fair classification algorithms, including *LinearPost* (Xian and Zhao 2024), *MBS* (Chen, Klochkov, and Liu 2024), and *Reduction* (Agarwal et al. 2018). We also include a basic baseline classifier without fairness constraints for comparison, including LR and XGBoost. We consider both attribute-blind and attribute-aware settings. Following Menon and Williamson (2018) and Chen, Klochkov, and Liu (2024), the values of the hyperparameter $\boldsymbol{\lambda} \in [-1, 1]^M$ are selected via grid search in our algorithms. The cost parameter is fixed at $c = 0.5$. Detailed dataset statistics and model setups are provided in the Appendix D of the supplementary material.

Evaluation Metrics. For various values of $\boldsymbol{\lambda}$, we report both accuracy and fairness levels, as they represent the key performance metrics of interest. All four fairness notions in Table 2 are considered. Approximate fairness is implemented using both MD and MR measures, and the achieved fairness level (i.e., values of $\mathrm{MD}(f)$ and $\mathrm{MR}(f)$) are reported.

Results. Figure 1 shows fairness-accuracy trade-offs for MD measure under DP across two datasets. It considers cases where *gender* or *race* is the only sensitive feature and where both of them are considered sensitive. Each point corresponds to a specific value of the tuning parameter $\boldsymbol{\lambda}$. Compared to the fair baselines, our methods achieve the more favorable fairness-accuracy trade-off in most cases, especially in the attribute-blind setting. Additionally, our in-processing method often outperforms our post-processing one in balancing fairness and accuracy, with a more significant advantage on the COMPAS dataset. We also evaluate our methods under other fairness notions and with respect to the MR measure. Due to space constraints, the complete results are reported in Appendix D.3 of the supplementary material, and Appendix E presents additional sensitivity and ablation analyses. All results suggest that our methods effectively navigates the trade-off between fairness and accuracy.

7 Conclusion

This work analyzes the fair classification problem with multiple sensitive features. We characterize that Bayes-optimal fair classifiers for approximate fairness–under both *mean difference* and *mean ratio* measures–can be represented by instance-dependent threshold corrections applied to the unconstrained Bayes-optimal classifier. The corrections are determined by a weighted sum of the probabilities that an individual belongs to specific groups. Our findings are applicable to both attribute-aware and attribute-blind settings and cover widely used fairness notions, including DP, EO, PE, and *accuracy parity* (AP). Notably, to the best of our knowledge, this is the first work to characterize the Bayes-optimal fair classifier under AP. Building on these insights, we proposed both in-processing and post-processing algorithms for learning Bayes-optimal fair classifiers from finite data. Empirical results show that our methods perform favorably compared to existing methods.

Acknowledgments

Dr. Yinghui Huang is supported by the Fundamental Research Funds for the Central Universities (No. SK2024006) and Natural Science Basic Research Program of Shaanxi (Program No.2025JC-QYCX-061). Dr. Xiangyu Chang is supported by the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001) and NSFC12326615.

References

Agarwal, A.; Beygelzimer, A.; Dudik, M.; Langford, J.; and Wallach, H. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 60–69. PMLR.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. ProPublica. Accessed: 2025-07-10.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Barocas, S.; and Selbst, A. D. 2016. Big Data's Disparate Impact. *California Law Review*, 104(3): 671–732.

Becker, B.; and Kohavi, R. 1996. Adult. UCI Machine Learning Repository.

Calders, T.; and Verwer, S. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21: 277–292.

Calmon, F.; Wei, D.; Vinzamuri, B.; Natesan Ramamurthy, K.; and Varshney, K. R. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Caton, S.; and Haas, C. 2024. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7): 1–38.

Chai, J.; and Wang, X. 2022. Fairness with Adaptive Weights. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, 2853–2866. PMLR.

Chen, W.; Klochkov, Y.; and Liu, Y. 2024. Post-hoc Bias Scoring is Optimal for Fair Classification. In *The Twelfth International Conference on Learning Representations*.

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2019. Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806.

Cruz, A. F.; and Hardt, M. 2024. Unprocessing Seven Years of Algorithmic Fairness. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Elkan, C. 2001. The Foundations of Cost-Sensitive Learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, volume 2, 973–978.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Gouic, T. L.; Loubes, J.-M.; and Rigollet, P. 2020. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3323–3331.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1): 1–33.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 2564–2572. PMLR.

Liu, L. Y.-F.; Liu, Y.; Zhu, H.; Initiative, A. D. N.; et al. 2018. SMAC: Spatial multi-category angle-based classifier for high-dimensional neuroimaging data. *NeuroImage*, 175: 230–245.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2021. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5): 102642.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Menon, A. K.; and Williamson, R. C. 2018. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, 107–118. PMLR.

Pinzón, C.; Palamidessi, C.; Piantanida, P.; and Valencia, F. 2022. On the Impossibility of Non-trivial Accuracy in Presence of Fairness Constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7): 7993–8000.

Schreuder, N.; and Chzhen, E. 2021. Classification with Abstention but Without Disparities. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161, 1227–1236. PMLR.

Wei, D.; Ramamurthy, K. N.; and Calmon, F. P. 2021. Optimized score transformation for consistent fair classification. *Journal of Machine Learning Research*, 22(258): 1–78.

Xian, R.; Yin, L.; and Zhao, H. 2023. Fair and Optimal Classification via Post-Processing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, 37977–38012. PMLR.

Xian, R.; and Zhao, H. 2024. A Unified Post-Processing Framework for Group Fairness in Classification. *arXiv preprint arXiv:2405.04025*.

Yang, F.; Cisse, M.; and Koyejo, S. 2020. Fairness with Overlapping Groups: a Probabilistic Perspective. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, 4067–4078.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017a. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2019. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75): 1–42.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017b. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, 962–970. PMLR.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, 325–333. PMLR.

Zeng, X.; Cheng, G.; and Dobriban, E. 2024. Bayes-optimal fair classification with linear disparity constraints via pre-, in-, and post-processing. *arXiv preprint arXiv:2402.02817*.

Zhang, C.; and Liu, Y. 2014. Multicategory angle-based large-margin classification. *Biometrika*, 101(3): 625–640.

Zhao, H.; Coston, A.; Adel, T.; and Gordon, G. J. 2020. Conditional Learning of Fair Representations. In *The 8th International Conference on Learning Representations*.