Journal of Machine Learning Research 27 (2026) 1-87 Submitted 2/24; Revised 8/25; Published 1/26

Bayes-Optimal Fair Classification with Linear Disparity Constraints via Pre-, In-, and Post-processing

Xianli Zeng xlzeng@xmu.edu.cn
WISE and Department of Statistics and Data Science, SOE, Xiamen University, Xiamen, China

Kevin Jiang kcjiang@wharton.upenn.edu
Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, USA

Guang Cheng guangcheng@ucla.edu
Department of Statistics, University of California, Los Angeles, Los Angeles, USA

Edgar Dobriban dobriban@wharton.upenn.edu
Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, USA

Editor: Maya Gupta

Abstract

Machine learning algorithms may have disparate impacts on protected groups. To address this, we develop methods for Bayes-optimal fair classification, aiming to minimize classification error subject to given group fairness constraints. We introduce the notion of *linear disparity measures*, which are linear functions of a probabilistic classifier; and *bilinear disparity measures*, which are also linear in the group-wise regression functions. We show that several popular disparity measures—the deviations from demographic parity, equality of opportunity, and predictive equality—are bilinear.

We find the form of Bayes-optimal fair classifiers under a single linear disparity measure, by uncovering a connection with the Neyman-Pearson lemma. For bilinear disparity measures, we are able to find the explicit form of Bayes-optimal fair classifiers as group-wise thresholding rules with explicitly characterized thresholds. We develop similar algorithms for when the protected attribute cannot be used at the prediction phase. Moreover, we obtain analogous theoretical characterizations of optimal classifiers for a multi-class protected attribute and for equalized odds.

Leveraging our theoretical results, we design methods that learn fair Bayes-optimal classifiers under bilinear disparity constraints. Our methods cover three popular approaches to fairness-aware classification, via pre-processing (Fair Up- and Down-Sampling), in-processing (Fair cost-sensitive Classification) and post-processing (a Fair Plug-In Rule). Our methods control disparity directly while achieving near-optimal fairness-accuracy tradeoffs. We show empirically that our methods have state-of-the-art performance compared to existing algorithms. In particular, our pre-processing method can reach a higher accuracy than prior pre-processing methods at low disparity levels.

Keywords: Algorithmic Bias, Fair Bayes-optimal Classifier, Newman-Pearson Lemma, Fair Pareto Frontier, Fair Algorithms.

1. Introduction

Machine learning is increasingly deployed in algorithmic decision-making systems in high-stakes domains deeply impacting human lives, including in education (Tsai et al., 2020),

©2026 Xianli Zeng, Kevin Jiang, Guang Cheng and Edgar Dobriban.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/24-0188.html.

finance (Ma et al., 2018), healthcare (Gupta and Mohammad, 2017), and judiciary systems (Angwin et al., 2016). While machine learning can improve decision-making efficiency, it has also surfaced many potential ethical risks.

One significant concern is algorithmic bias. Recent studies have revealed that without considering fairness, machine learning algorithms often make decisions that disadvantage vulnerable demographic groups, thereby exacerbating social injustice and potentially violating human rights (Angwin et al., 2016; Flores et al., 2016; Bhanot et al., 2021; Corbett-Davies et al., 2023).

In light of these risks, algorithmic fairness has attracted widespread attention from the public, government, and academia. Organizations like the White House (Executive Office of the President, 2016), and UNESCO (UNESCO, 2021) have all called for considering fairness when applying automated decision-making. In response, the emerging field of fair machine learning has seen significant development. An increasing amount of research focuses on defining quantitative fairness metrics for various applications (e.g., Calders et al., 2009; Dwork et al., 2012; Hardt et al., 2016; Zafar et al., 2017; Pleiss et al., 2017; Kusner et al., 2017; Zhang and Bareinboim, 2018; Narasimhan et al., 2020; Nandy et al., 2022, etc), designing algorithms that protect minority groups (e.g., Kamiran and Calders, 2012; Joseph et al., 2016; Zhang et al., 2018; Lahoti et al., 2019; Cho et al., 2020; Xian et al., 2023; Deng et al., 2023; Zhang et al., 2024, etc), providing public software to help practitioners assess and improve fairness (e.g., Bellamy et al., 2019; Weerts et al., 2023, etc), and studying the theoretical underpinnings of algorithmic fairness (e.g., Corbett-Davies et al., 2017; Menon and Williamson, 2018; Schreuder and Chzhen, 2021; Zeng et al., 2022b; Hou and Zhang, 2024, etc). We refer the readers to e.g., Mehrabi et al. (2021); Caton and Haas (2023), etc, for reviews of advances in fair machine learning.

In this paper, we develop methods for fair classification. In fair classification, the *fair Bayes-optimal classifier* is the method with the best possible accuracy under a given disparity constraint. This method serves as an ultimate "best possible" classifier. We consider the question of deriving Bayes-optimal classifiers under given group fairness constraints, which aim to equalize various quantities across protected groups.

Prior work by Hardt et al. (2016) showed that for oblivious fairness measures (which only depend on the joint distributions of the outcome, the protected group, and the class prediction), and for perfect fairness, the Bayes-optimal classifiers can always be expressed in terms of the group-wise regression functions; but without giving an explicit algorithm to find them. Further, given any predictor $\widehat{Y}$, they also show how to obtain a predictor that is a function of $\widehat{Y}$ and the protected attribute that achieves perfect equalized odds, but which may not be Bayes-optimal.

Further, Corbett-Davies et al. (2017) showed that, for exact fairness with some specific disparity measures, fair Bayes-optimal classifiers are group-wise thresholding rules, but did not identify their exact form (see Table 1 for a summary of, and comparison with, related work). For the specific fairness criteria of demographic parity and equality of opportunity (see Section 2.1), Menon and Williamson (2018) characterized fair Bayes-optimal classifiers by a connection to cost-sensitive risks. Explicit forms of fair Bayes-optimal classifiers were derived for perfect demographic parity in Chzhen et al. (2019) and for perfect equality of opportunity in Schreuder and Chzhen (2021); see also Hardt et al. (2016) for a previous characterization.

Table 1: Our contributions in comparison with prior theoretical work for Bayes-optimal classifiers: [1]: Corbett-Davies et al. (2017); [2]: Menon and Williamson (2018); [3]: Chzhen et al. (2019); [4]: Jiang et al. (2020); [5]: Schreuder and Chzhen (2021); [6]: Wei et al. (2021) [7]: Chen et al. (2023); [8]: Xu and Strohmer (2023). The comparison is made based on scope of the theoretical framework, the fairness metrics considered, and whether theoretically optimal algorithms are proposed. Within each category, several criteria are considered. Our methods satisfy all desired criteria, while the prior works satisfy only some of them.

References	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	This Work
Scope of Theoretical Framework									
Approximate Fairness		✓		✓		✓	✓	✓	✓
Explicit Form			✓	✓	✓			✓	✓
Multiple Constraints						✓	✓		✓
Pareto Analysis		✓						✓	✓
No Protected Attrib. A at test time		✓		✓		✓	✓		✓
Fairness Metrics Considered									
Demographic Parity	✓	✓		✓	✓	✓	✓	✓	✓
Equality of Opportunity	✓	✓	✓				✓		✓
Predictive Equality	✓								✓
Equalized Odds						✓	✓		✓
Theoretically Optimal Algorithms									
Pre-processing						✓		✓	✓
In-processing				✓					✓
Post-processing	✓	✓	✓	✓	✓	✓	✓	✓	✓

By considering the Wasserstein distance as the disparity measure, Jiang et al. (2020); Silvia et al. (2020); Xian et al. (2023) proved that the fair classification problem is equivalent to a Wasserstein-barycenter problem, and the perfectly fair Bayes-optimal classifier can be derived via optimal transport. Follow-up work by Xu and Strohmer (2023) further characterized the Pareto Frontier—i.e., the optimal fairness-accuracy tradeoff—using the Wasserstein geodesic. Rather than considering the fair Bayes-optimal classifier, Wei et al. (2021) considered the estimation of the Bayes-optimal score function and derived the optimal transformed score function that satisfies fairness constraints. In a selective classification framework, Rava et al. (2021) studied how to minimize outcomes without a decision, while equalizing the false selection rate among different protected groups. More recently, in closely related work, Chen et al. (2023) characterized Bayes-optimal classifiers as linear combinations of certain basis functions, without finding their explicit forms. We provide a detailed comparison with this work after Theorem 14.

See Table 1 for a quick overview of comparison with related work.

Despite this progress, there is no systematic approach to deriving explicit forms of Bayes-optimal fair classifiers for general disparity measures and arbitrary disparity levels. There is also no unified algorithmic framework to learn these classifiers from data.

In this paper, we propose a unified theoretical framework for deriving Bayes-optimal classifiers under various fairness measures, by leveraging a novel connection with the Neyman-Pearson argument for optimal hypothesis testing. We define a new notion of *linear disparity measures*, which are linear functions of a probabilistic classifier (See Definition 10), and show that finding the fair Bayes optimal classifier can then be recast as maximizing a linear functional subject to linear constraints. Since this is fundamentally similar to maximizing power subject to level constraints in optimal hypothesis testing, we can derive the explicit form of a fair Bayes-optimal classifier with any given level of disparity, for any given linear disparity measure.

While prior work has explored linear or bilinear representations of disparity, our framework differs in its precise definition of linearity, the choice of objective (e.g., misclassification error vs. other losses), and the classification setting (e.g., binary vs. multiclass). Agarwal et al. (2018) use a weaker notion of linearity that does not require linearity in the classifier itself; our results can thus be more explicit in terms of establishing the precise forms of the classifiers. Celis et al. (2019b) define a more general class of fractional linear constraints. In comparison, we highlight the bilinear case in the linear constraint setting, which enables us to establish the explicit form of the Bayes-optimal classifiers for equality of opportunity and predictive equality at a user-specified approximate level. Wei et al. (2021) use similar disparity constraints but minimize cross-entropy rather than misclassification error; and finally, Alghamdi et al. (2022) work in a multiclass setting with ratio-based disparity measures and optimize divergence from a base classifier rather than mis-classification error.

We further show that our general theoretical framework can be extended to a wide range of scenarios (see Section 4.3): handling multiple fairness constraints (such as equalized odds, ensuring both equality of opportunity and predictive equality); and for the common scenario when the protected attribute cannot be used at the prediction phase.

In addition to deriving fair Bayes-optimal classifiers, we also investigate the *fair Pareto frontier*, which characterizes the optimal tradeoff between fairness and accuracy. This Pareto frontier comprises a set of fair Bayes-optimal classifiers. We study the tradeoff function that characterizes the best achievable misclassification rate for a given disparity level. We prove that for linear disparity measures, the tradeoff function is convex, highlighting that the marginal cost of fairness increases for smaller disparity levels. Wang et al. (2023) study a different yet similar fair Pareto frontier, and we comment on their work and its relation to ours in Section 4.2.

In practice, Bayes-optimal classifiers need to be estimated from a finite dataset. For this, we develop methods covering three popular approaches to fairness-aware classification: pre-processing, in-processing, and post-processing (e.g., Caton and Haas, 2023).

- *Pre-processing methods* reduce biases implicit in the training data, and train classifiers on the debiased data. Examples include transformations (e.g., Feldman et al., 2015; Lum and Johndrow, 2016; Johndrow and Lum, 2019; Calmon et al., 2017, etc), fair representation learning (e.g., Zemel et al., 2013; Louizos et al., 2016; Creager et al., 2019, etc), and fair generative models (e.g., Xu et al., 2018; Sattigeri et al., 2019; Jang et al., 2021, etc). These methods are convenient to apply, as they do not change the training procedure. However, as argued for instance in Locatello et al. (2019), disparity could persist even after pre-processing.

- *In-processing algorithms* handle the fairness constraint during the training process. A commonly applied strategy is to incorporate fairness measures as a regularization term into the optimization objective (e.g., Goh et al., 2016; Zafar et al., 2019; Narasimhan, 2018; Celis et al., 2019a; Cotter et al., 2019; Oneto et al., 2020; Cho et al., 2020, etc). However, fairness measures are often non-convex and even non-differentiable with respect to model parameters, so that scalable training is challenging. In the alternative approach of adversarial learning (e.g., Zhang et al., 2018; Wadsworth et al., 2018; Xu et al., 2019a; Celis and Keswani, 2019, etc), the ability of the classifier to predict the protected attribute is minimized. Although this can achieve promising results through careful design, its training process can lack stability, as a min-max optimization problem is required to be solved (Cho et al., 2020). Other in-processing algorithms include domain-based training (Wang et al., 2020), where the protected attribute is explicitly encoded and its effect is mitigated.

- *Post-processing methods* fit classifiers to the training data in the standard way, and aim to mitigate disparities in the model output. The most frequently used post-processing algorithms include group-wise thresholding (e.g., Fish et al., 2016; Corbett-Davies et al., 2017; Valera et al., 2018; Menon and Williamson, 2018; Chzhen et al., 2019; Alabdulmohsin, 2020; Schreuder and Chzhen, 2021; Denis et al., 2021; Jang et al., 2022; Li et al., 2023, etc) and post-processing via optimal transport (e.g., Jiang et al., 2020; Silvia et al., 2020; Xian et al., 2023; Xu and Strohmer, 2023, etc). In group-wise thresholding, after estimating the conditional probability of $Y = 1$ for each protected group (i.e., the group-wise regression functions), the method classifies $\widehat{Y} = 1$ when these probabilities exceed certain group-specific thresholds. On the other hand, the optimal transport-based method transforms these conditional probabilities into a fair Bayes-optimal classifier by employing optimal transport algorithms.

Motivated by our formulas for fair Bayes-optimal classifiers, we design three algorithms that mitigate algorithmic bias, via pre-processing (Fair Up- and Down-Sampling or FUDS), in-processing (Fair cost-sensitive Classification or FCSC) and post-processing (Fair Plug-In Rule or FPIR).

For the pre-processing algorithm FUDS, we characterize the perturbed data distribution for which the unconstrained Bayes-optimal classifier equals the fair Bayes-optimal classifier on the original target distribution. We prove that this perturbation can be obtained by adjusting the proportion of each demographic group defined by the label and protected attribute. This enables us to design an easy-to-use Fair Up- and Down-Sampling pre-processing algorithm.

For the in-processing algorithm FCSC, we observe that the fair Bayes-optimal classifier adjusts thresholds for each protected group to mitigate disparities. Leveraging that cost-sensitive classification achieves a similar effect, we show that the fair Bayes-optimal classifier can be achieved through a carefully designed group-wise cost-sensitive risk, leading to Fair cost-sensitive Classification (FCSC).

Finally, for the post-processing algorithm FPIR, since we derived the explicit form of fair Bayes-optimal classifiers, we can design a two-stage plug-in rule to estimate it. In the first stage, we estimate the feature-conditional probability of a positive outcome $Y = 1$

for each protected group. In the second stage, the optimal fair decision boundary can be estimated by solving a one-dimensional search problem.

We summarize our contributions below. We also provide a detailed comparison of our contributions with prior theoretical work for fair Bayes-optimal classifiers in Table 1.

1. **Unifying framework for fair Bayes-optimal classifiers with linear disparity measures:** We provide a unified framework for deriving Bayes-optimal classifiers under group fairness constraints. We introduce the notions of linear and bilinear disparity measures (Definitions 10 and 11) and show that several popular disparity measures—the deviations from demographic parity, equality of opportunity, predictive equality, and equalized odds—are bilinear (Proposition 12). We characterize Bayes-optimal classifiers under linear disparity measures (Theorem 14), by uncovering a connection with the Neyman-Pearson lemma. For bilinear disparity measures, we are able to find the explicit form of Bayes-optimal fair classifiers as group-wise thresholding rules with explicitly characterized thresholds. In particular, this explicit form appears to be novel when a nonzero user-specified disparity level is specified.

2. **Extensions:** We illustrate that our approach can be extended to handle multiple fairness constraints (such as equalized odds (Theorem 21), which requires ensuring both equality of opportunity and predictive equality), and demographic parity with a multi-class protected attribute (Theorem 22). The result for equalized odds requires an intricate and lengthy argument, which unravels several fundamental properties of equality of opportunity and predictive equality. We also show how to handle cost-sensitive classification error (Corollary 23). Further, we derive Bayes-optimal fair classifiers for the common scenario when the protected attribute cannot be used at the prediction phase (Corollary 20).

   Finally, we give the form of Bayes-optimal fair classifiers under the general distributional assumptions that the features can belong to the decision boundary with nonzero probability; such as for discrete features. In this case, the optimal classifiers must be carefully randomized (Theorem 38).

3. **Pareto frontier for fair classification and fairness-accuracy tradeoff:** We characterize the fair Pareto frontier that represents the optimal trade-off between accuracy and fairness. We show that the tradeoff function between misclassification and disparity is convex for linear disparity measures, indicating that the marginal cost of fairness increases as the level of disparity decreases.

4. **Theoretically optimal fair classification via pre-, in-, and post-processing:** We propose pre-, in-, and post-processing algorithms aiming to recover the Bayes-optimal classifiers. Specifically we introduce fair up-/down-sampling as a pre-processing method (Section 5.1), fair cost-sensitive classification as an in-processing method (Section 5.2), and a fair plug-in rule as a post-processing method (Section 5.3). Collectively, these methods establish a cohesive methodological framework for fair classification with linear disparity constraints.

5. **Empirical evaluation:** We evaluate our methods in numerical simulations and experiments on empirical datasets (Section 6). We compare with several existing meth-

ods, such as the disparate impact remover (Feldman et al., 2015), FAWOS (Salazar et al., 2021), adaptive reweighting algorithm (Chai et al., 2022), KDE-based constrained optimization (Cho et al., 2020), adversarial training (Zhang et al., 2018), reductions (Agarwal et al., 2018), MinDiff regularization (Prost et al., 2019), post-processing through flipping (Chen et al., 2023), post-processing through optimal transport (Xian et al., 2023), and FRAPPÉ (Tifrea et al., 2023); on standard and more recent benchmark datasets. We show empirically that our methods have state-of-the-art performance compared to existing algorithms. In particular, our pre-processing method can a reach higher accuracy than prior pre-processing methods at low disparity levels. Our numerical results can be reproduced with the code provided at `https://github.com/XianliZeng/Bayes-Optimal-Fair-Classification`.

This paper significantly extends the results from our previous unpublished manuscript (Zeng et al., 2022a), introducing the notion of linear disparity measures, handling multiple constraints, and developing pre- and in-processing algorithms. Thus, this work supersedes Zeng et al. (2022a).

Paper organization. In Section 2.1, we review background concepts and notations for fair classification. Section 3 establishes the connection between the generalized Neyman-Pearson Lemma and fair Bayes-optimal classifiers. Our main results are presented in Section 4, providing explicit forms of fair Bayes-optimal classifiers under linear and bilinear group fairness measures. We characterize the fair Pareto frontier in Section 4.2. In Section 5, we propose methods for estimating fair Bayes-optimal classifiers through pre-, in- and post-processing methods. Simulation studies and empirical data analysis in Section 6 assess the finite sample performance of the proposed methods. We provide concluding remarks in Section 7. All proofs are in the supplementary material.

2. Classification with a Protected Attribute

In fair classification problems, two types of feature are observed: the usual feature $X \in \mathcal{X}$ for some feature space $\mathcal{X}$, and the protected feature[1] $A \in \mathcal{A} = \{0,1\}$ with respect to which we aim to be fair. Here, we consider a binary classification problem with labels in $\mathcal{Y} = \{0,1\}$. For example, in a credit lending setting, X may refer to common features such as education level and income, A may contain the race or gender of the individual and Y may correspond to the status of repayment or defaulting on a loan.

Notation and Conventions. For all $a \in \mathcal{A}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we denote $p_a := \mathbb{P}(A = a)$; $p_{a,y} := \mathbb{P}(A = a, Y = y)$; $\eta_a(x) := \mathbb{P}(Y = 1 \mid A = a, X = x)$. Further, for all $a \in \mathcal{A}$ and $y \in \mathcal{Y}$, we denote by $\mathbb{P}_X(\cdot)$, $\mathbb{P}_{X|A=a}(\cdot)$ and $\mathbb{P}_{X|A=a,Y=y}(\cdot)$ the marginal distribution function of X, the conditional distribution function of X given $A = a$, and the conditional distribution of X given $A = a, Y = y$, respectively.

For two scalars a, b, we denote their maximum by $\max\{a, b\}$ or $a \vee b$, and their minimum by $\min\{a, b\}$ or $a \wedge b$. All quantities considered will be measurable with respect to appropriate sigma-algebras; and measurability may not always be mentioned in what follows.

1. We consider a binary protected attribute here and extend our results to multi-class protected attributes in Section 4.3.

2.1 Preliminaries

A *randomized classifier* outputs a prediction $\widehat{Y} \in \{0,1\}$ with a certain probability based on the usual features X and the protected attribute A. Let $\mathrm{Bern}(p)$ be the Bernoulli distribution with success probability $p \in [0,1]$, and let $\mathcal{F}$ be the set of measurable functions $f : \mathcal{X} \times \mathcal{A} \to [0,1]$. We denote by $\widehat{Y}_f = \widehat{Y}_f(x,a) \in \{0,1\}$ the prediction induced by the classifier f, which can be a random variable.

Definition 1 (Randomized Classifier) *A randomized classifier $f \in \mathcal{F}$ gives, for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$, the probability $f(x,a)$ of predicting $\widehat{Y}_f = 1$ when observing $X = x$ and $A = a$, i.e., $\widehat{Y}_f \mid X = x, A = a \sim \mathrm{Bern}(f(x,a))$.*

Without a fairness constraint, a *Bayes-optimal classifier* minimizes the misclassification rate, and is defined as any randomized classifier $f^\star$ satisfying $f^\star \in \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{P}(Y \neq \widehat{Y}_f)$. The following classical result characterizes Bayes-optimal classifiers in terms of the class-conditional probability functions η_a for which[2] $\eta_a(x) = \mathbb{P}(Y = 1 \mid A = a, X = x)$ for all x, a (see e.g., Devroye et al., 1996, etc). We denote by $I(\cdot)$ the indicator function, which equals unity if its argument is true, and zero otherwise.

Proposition 2 *All Bayes-optimal classifiers $f^\star \in \mathcal{F}$ have the form $f^\star(x,a) = I(\eta_a(x) > 1/2) + \tau(x,a) I(\eta_a(x) = 1/2)$, for all $(x,a) \in \mathcal{X} \times \{0,1\}$, where $\tau : \mathcal{X} \times \mathcal{A} \to [0,1]$ is any measurable function.*

While Bayes-optimal classifiers are theoretically the best method, they depend on the usually unknown class-conditional probability functions η_a of the population. However, Proposition 2 is still useful, because it suggests an approach to classification, by estimating the functions η_a, for all a, and defining the classifier in terms of the level sets of the estimates, see e.g., Audibert and Tsybakov (2007).

The Bayes-optimal classifier does not take fairness into account. To mitigate unfairness, a number of notions of parity have been considered, and we list below the ones we consider in this paper.

Definition 3 (Demographic Parity (Calders et al., 2009)) *A classifier f satisfies demographic parity if its prediction $\widehat{Y}_f$ is probabilistically independent of the protected attribute A, so that $\mathbb{P}_{X|A=1}\left(\widehat{Y}_f = 1\right) = \mathbb{P}_{X|A=0}\left(\widehat{Y}_f = 1\right)$.*

Definition 4 (Equality of Opportunity (Hardt et al., 2016)) *A classifier f satisfies equality of opportunity if it achieves the same true positive rate among protected groups: $\mathbb{P}_{X|A=1,Y=1}\left(\widehat{Y}_f = 1\right) = \mathbb{P}_{X|A=0,Y=1}\left(\widehat{Y}_f = 1\right)$.*

Definition 5 (Predictive Equality (Corbett-Davies et al., 2017)) *A classifier f satisfies predictive equality if it achieves the same false positive rate among protected groups: $\mathbb{P}_{X|A=1,Y=0}\left(\widehat{Y}_f = 1\right) = \mathbb{P}_{X|A=0,Y=0}\left(\widehat{Y}_f = 1\right)$.*

2. We will sometimes write $\eta^Y_{A=a} := \eta_a$ for $a \in \mathcal{A}$.

Definition 6 (Equalized Odds (Hardt et al., 2016)) *A classifier f satisfies equalized odds if it satisfies both equality of opportunity and predictive equality, achieving the same true positive rate and false positive rate among protected groups:* $\mathbb{P}_{X|A=1,Y=1}\left(\widehat{Y}_f=1\right)=\mathbb{P}_{X|A=0,Y=1}\left(\widehat{Y}_f=1\right)$, *and* $\mathbb{P}_{X|A=1,Y=0}\left(\widehat{Y}_f=1\right)=\mathbb{P}_{X|A=0,Y=0}\left(\widehat{Y}_f=1\right)$.

In applications, perfect fairness may require a large sacrifice of accuracy, and a "limited" disparate impact could be preferred. Following, for instance, Cho et al. (2020), we use the difference in the quantities that are equalized under perfect parity to measure disparity:

Definition 7 (Disparity Measures) *We consider the following disparity measures: Demographic Disparity* (DD)*, Disparity of Opportunity* (DO)*, Predictive Disparity* (PD)*, defined for probabilistic classifiers f as follows:*

$$\begin{aligned}\mathrm{DD}(f)&=\mathbb{P}_{X|A=1}(\widehat{Y}_f=1)-\mathbb{P}_{X|A=0}\left(\widehat{Y}_f=1\right);\\ \mathrm{DO}(f)&=\mathbb{P}_{X|A=1,Y=1}(\widehat{Y}_f=1)-\mathbb{P}_{X|A=0,Y=1}\left(\widehat{Y}_f=1\right);\\ \mathrm{PD}(f)&=\mathbb{P}_{X|A=1,Y=0}\left(\widehat{Y}_f=1\right)-\mathbb{P}_{X|A=0,Y=0}\left(\widehat{Y}_f=1\right).\end{aligned}\tag{1}$$

Since equalized odds consists of two constraints, we will later use the maximum of the absolute values of the corresponding disparities as the measure of unfairness.

In the rest of this paper, we will use $\mathrm{Dis}:\mathcal{F}\to[-1,1]$ to refer to a generic disparity measure, such as any of the three measures from (1). Taking $K\geqslant 1$ disparity measures $\mathrm{Dis}_k:\mathcal{F}\to[-1,1],k=1,...,K$ into account, we say a classifier f satisfies δ-disparity for some $\delta\in[0,\infty)$ if $\max_{k=1}^K|\mathrm{Dis}_k(f)|\leqslant\delta$. We next define the most accurate—equivalently, least inaccurate—classifiers that satisfy δ-disparity.

Definition 8 (Fair Bayes-optimal Classifier) *Consider any $K\geqslant 1$ fairness measures $\mathrm{Dis}_k:\mathcal{F}\to[-1,1]$, $k=1,...,K$. Then, a δ-fair Bayes-optimal classifier $f^\star_{\mathrm{Dis},\delta}$ is defined by minimizing the misclassification error $R(f):=\mathbb{P}\left(Y\neq\widehat{Y}_f\right)$ over all classifiers that satisfy δ-disparity:*

$$f^\star_{\mathrm{Dis},\delta}\in\arg\min_{f\in\mathcal{F}}\left\{R(f):\max_{k=1}^K|\mathrm{Dis}_k(f)|\leqslant\delta\right\}.$$

Just like for the unconstrained case, fair Bayes-optimal classifiers are theoretically the best method, but they do not directly lead to a feasible method. To be able to use classifiers inspired by fair Bayes-optimal ones, it would be helpful to know if there is a characterization similar to that from the unconstrained case in Proposition 2. We now turn to establishing this, by making a connection with the Neyman-Pearson lemma.

3. Generalized Neyman-Pearson Lemma and Fair Classification

3.1 Generalized Neyman-Pearson Lemma

To derive Bayes-optimal classifiers under group fairness, we establish a connection with the Neyman-Pearson lemma (Neyman and Pearson, 1933); a theoretical result originally designed to derive most powerful tests with a given type I error. We use a slightly generalized version, which applies to optimization with linear constraints. For completeness and the reader's convenience, we present the full statement here; as this lemma is fundamental to our results.

Lemma 9 (Generalized Neyman-Pearson Lemma (Lehmann and Romano, 2005; Shao, 2003)) *Let $\phi_0, \phi_1, ..., \phi_m$ be $m+1$ real-valued functions defined on a Euclidean space $\mathcal{X}$. Assume they are ν-integrable for a σ-finite measure ν. Let $f^\star \in \mathcal{F}$ be any function of the form*

$$f^\star(x) = \begin{cases} 1, & \phi_0(x) > \sum_{i=1}^m c_i\phi_i(x); \\ \tau(x) & \phi_0(x) = \sum_{i=1}^m c_i\phi_i(x); \\ 0, & \phi_0(x) < \sum_{i=1}^m c_i\phi_i(x), \end{cases} \tag{2}$$

where $0 \leqslant \tau(x) \leqslant 1$ for all $x \in \mathcal{X}$. For given constants $t_1, ..., t_m \in \mathbb{R}$, let $\mathcal{F}_\leqslant$ be the class of measurable functions $f : \mathcal{X} \to \mathbb{R}$ satisfying

$$\int_{\mathcal{X}} f\phi_i d\nu \leqslant t_i, \ i \in \{1, 2, ..., m\}. \tag{3}$$

and $\mathcal{F}_=$ be the set of functions in $\mathcal{F}_\leqslant$ satisfying (3) *with all inequalities replaced by equalities.*

(1) If $f^\star \in \mathcal{F}_=$, then

$$f^\star \in \underset{f\in\mathcal{F}_=}{\arg\max} \int_{\mathcal{X}} f\phi_0 d\nu. \tag{4}$$

Moreover, if $\nu(\{x : \phi_0(x) = \sum_{i=1}^m c_i\phi_i(x)\}) = 0$, for all $f' \in \underset{f\in\mathcal{F}_=}{\arg\max} \int_{\mathcal{X}} f\phi_0 d\nu$, $f' = f^\star$ almost everywhere with respect to ν.

(2) Moreover, if $c_i \geqslant 0$ for all $i = 1, \dots, m$, then

$$f^\star \in \underset{f\in\mathcal{F}_\leqslant}{\arg\max} \int_{\mathcal{X}} f\phi_0 d\nu. \tag{5}$$

Moreover, if $\nu(\{x : \phi_0(x) = \sum_{i=1}^m c_i\phi_i(x)\}) = 0$, for all $f' \in \underset{f\in\mathcal{F}_\leqslant}{\arg\max} \int_{\mathcal{X}} f\phi_0 d\nu$, we have $f'(x) = f^\star(x)$ almost everywhere with respect to ν.

3.2 Fairness Measures with Linear and Bilinear Constraints

To characterize Bayes-optimal fair classifiers, we want to find classifiers with the highest accuracy given a disparity level. Recalling the class-conditional regression functions such that for $a \in \mathcal{A}$ and $x \in \mathcal{X}$, $\eta_a(x) = \mathbb{P}(Y = 1 \mid A = a, X = x)$, the misclassification rate can be expressed as a linear functional of the classifier f (see Lemma 27 for the argument), via

$$R(f) = \int_{\mathcal{A}} \int_{\mathcal{X}} f(x, a)(1 - 2\eta_a(x))d\mathbb{P}_{X,A}(x, a) + C_{\mathbb{P}},$$

with $C_{\mathbb{P}} = \int_{\mathcal{A}} \int_{\mathcal{X}} \eta_a(x) d\mathbb{P}_{X,A}(x,a)$. As $C_{\mathbb{P}}$ is independent of f, maximizing accuracy over f is equivalent to maximizing $\int_{\mathcal{A}} \int_{\mathcal{X}} f(x,a)(2\eta_a(x)-1) d\mathbb{P}_{X,A}(x,a)$. Therefore, a δ-fair Bayes-optimal classifier for a disparity measure Dis can be equivalently defined as

$$f^{\star}_{\mathrm{Dis},\delta} \in \underset{f\in\mathcal{F}}{\arg\max} \left\{ \int_{\mathcal{A}} \int_{\mathcal{X}} f(x,a)(2\eta_a(x)-1) d\mathbb{P}_{X,A}(x,a) : |\mathrm{Dis}(f)| \leqslant \delta \right\}. \tag{6}$$

Hence, the objective is linear in f. Specifically, taking $\nu = \mathbb{P}_{X,A}$ and $\phi_0(x,a) = 2\eta_a(x) - 1$ for all x, a in the objective from (5) in the Neyman-Peason lemma, we find that it reduces to the objective in (6), where the constraint set $\mathcal{F}_{\leqslant}$ from (3) remains to be specified.

As a result, the fair Bayes-optimal classifiers can be characterized by the generalized Neyman-Pearson lemma, *if the constraints are linear in the classifiers*. Motivated by this observation, we introduce the notion of *linear disparity measures*.

Definition 10 (Linear Disparity Measure) *We call a disparity measure* Dis : $\mathcal{F} \to [-1,1]$ linear *if for all* $\mathbb{P}$*, there is a weighting function* $w_{\mathrm{Dis},\mathbb{P}} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ *such that for all* $f \in \mathcal{F}$*,*

$$\mathrm{Dis}(f) = \int_{\mathcal{A}} \int_{\mathcal{X}} f(x,a) w_{\mathrm{Dis},\mathbb{P}}(x,a) d\mathbb{P}_{X,A}(x,a). \tag{7}$$

Furthermore, we call a linear disparity measure *bilinear* if its weighting function is linear in the group-wise regression functions η_a.

Definition 11 (Bilinear Disparity Measure) *A linear disparity measure* Dis : $\mathcal{F} \to [-1,1]$ *is* bilinear *if for all* $\mathbb{P}$*, there are* $s_{\mathrm{Dis},\mathbb{P},a}$ *and* $b_{\mathrm{Dis},\mathbb{P},a}$ *depending on* $a \in \mathcal{A}$ *such that for all* $x \in \mathcal{X}$*,* $w_{\mathrm{Dis},\mathbb{P}}(x,a) = s_{\mathrm{Dis},\mathbb{P},a}\eta_a(x) + b_{\mathrm{Dis},\mathbb{P},a}$.

Hereafter, we will omit the subscript $\mathbb{P}$ in our notation for simplicity. We next show that the disparity measures from (1) are bilinear, with specific weighting functions.

Proposition 12 (Classical Disparity Measures are Bilinear) *The disparity measures* DD*,* DO*, and* PD *from* (1) *are bilinear with weighting functions defined for all* x, a *by*

$$w_{\mathrm{DD}}(x,a) = \frac{(2a-1)}{p_a}; \quad w_{\mathrm{DO}}(x,a) = \frac{(2a-1)\eta_a(x)}{p_{a,1}}; \quad w_{\mathrm{PD}}(x,a) = \frac{(2a-1)(1-\eta_a(x))}{p_{a,0}} \tag{8}$$

where $p_a = \mathbb{P}(A = a)$ *and* $p_{a,y} = \mathbb{P}(A = a, Y = y)$ *for* $a \in \{0,1\}$ *and* $y \in \{0,1\}$*. Thus, we have, for instance* $s_{\mathrm{DD},a} = 0$ *and* $b_{\mathrm{DD},a} = (2a-1)/p_a$ *for all* a*; the quantities* $s_{\mathrm{Dis},a}$ *and* $b_{\mathrm{Dis},\mathbb{P},a}$ *for the other disparity measures can be similarly read off from* (8).

Some of our results apply to linear disparity measures, and some to bilinear ones; thus we use both definitions. We will use (7) and Proposition 12 to develop a unified framework for deriving fair Bayes-optimal classifiers. For ease of exposition, we discuss a binary protected attribute and a single fairness constraint in the main text. However, our theory can also handle other scenarios, such as multiple fairness constraints and in particular a multi-class protected attribute. We briefly discuss these extensions in Section 4.3, and provide details in Appendix B of the Appendix.

4. Fair Bayes-Optimal Classifiers and the Fair Pareto Frontier

4.1 Form of Fair Bayes-Optimal Classifiers

In this section, we derive the explicit form of fair Bayes-optimal classifiers based on our theoretical framework. For simplicity, we present the results for a single linear disparity measure as per Definition 10, and discuss cases with multiple constraints in Appendix B of the Appendix. Furthermore, in the main text, we consider the setting where both random variables $\eta_a(X)$ and $w_{\mathrm{Dis}}(X,a)$, for $a \in \{0,1\}$, have probability density functions over $\mathcal{X}$. This implies that the boundary sets of randomized classifiers are of measure zero, allowing us to only consider deterministic classifiers. We provide the—more involved—results applicable to the general case without this assumption in Appendix B.4 of the Appendix.

To build intuition for the explicit form of the δ-fair Bayes optimal classifier, consider a linear disparity measure Dis. For expositional simplicity, consider finding the Bayes-optimal classifier with respect to the one-sided constraint Dis $\leqslant \delta$; while later we will work with the two-sided constraint. Let $\phi_0(x,a) = 2\eta_a(x) - 1$ and $\phi_1(x,a) = w_{\mathrm{Dis}}(x,a)$ for all x and a in equations (3) and (4), respectively specify the accuracy and (one-sided) disparity constraint. The generalized Neyman-Pearson lemma, together with the implicit characterization of the δ-fair Bayes-optimal classifier in (6) and the form of the linear disparity constraints in (7), suggest that a fair Bayes-optimal classifier takes the deterministic form:

$$f_{\mathrm{Dis},t}(x,a) = I\left(\eta_a(x) > \frac{1}{2} + \frac{t}{2}\, w_{\mathrm{Dis}}(x,a)\right), \tag{9}$$

for all $x \in \mathcal{X}$, $a \in \{0,1\}$, and some $t \in \mathbb{R}$.

To make this derivation precise, it will be helpful to define a disparity function $D_{\mathrm{Dis}} : \mathbb{R} \to [-1,1]$ measuring the disparity level of $f_{\mathrm{Dis},t}$ as a function of t, such that for all $t \in \mathbb{R}$:

$$\begin{aligned} D_{\mathrm{Dis}}(t) :=& \mathrm{Dis}(f_{\mathrm{Dis},t}) = \int_{\mathcal{A}}\int_{\mathcal{X}} \left[w_{\mathrm{Dis}}(x,a) \cdot I\left(\eta_a(x) > \frac{1}{2} + \frac{t}{2} w_{\mathrm{Dis}}(x,a)\right)\right] d\mathbb{P}_{X,A}(x,a) \\ &= \sum_{a \in \{0,1\}} p_a \int_{\mathcal{X}} \left[w_{\mathrm{Dis}}(x,a) \cdot I\left(\eta_a(x) > \frac{1}{2} + \frac{t}{2} w_{\mathrm{Dis}}(x,a)\right)\right] d\mathbb{P}_{X|A=a}(x). \end{aligned} \tag{10}$$

The following proposition characterizes how the accuracy and disparity of $f_{\mathrm{Dis},t}$ depends on t, and will be a crucial stepping stone towards precisely characterizing the form of fair Bayes-optimal classifiers.

Proposition 13 (Properties of Risk and Disparity) *Let $f_{\mathrm{Dis},t}$ and D_{Dis} be defined in* (9) *and* (10)*, respectively. Then, as a function of t,*

(1) the disparity $D_{\mathrm{Dis}}(t)$ is monotone non-increasing;

(2) the misclassification error $R(f_{\mathrm{Dis},t})$ is monotone non-increasing on $(-\infty, 0)$ and monotone non-decreasing on $[0, \infty)$.

Now, consider finding a classifier $f_{\mathrm{Dis},t}$ that meets fairness constraints while minimizing the misclassification error. By Proposition 13, it is enough to minimize $|t|$; indeed this can be seen considering the cases of $|D_{\mathrm{Dis}}(0)| \leqslant \delta$ (in which case the optimal $t = 0$), $D_{\mathrm{Dis}}(0) < -\delta$ (in which case the optimal t is negative), and $D_{\mathrm{Dis}}(0) > \delta$ (in which case the optimal t is positive). This motivates us to define the function $t_{\mathrm{Dis}} : [0, \infty) \to \mathbb{R}$ as an "inverse function" of $|D_{\mathrm{Dis}}(t)|$ such that for all $\delta \geqslant 0$,

$$t_{\mathrm{Dis}}(\delta) = \arg\min_t \{|t| : |D_{\mathrm{Dis}}(t)| \leqslant \delta\}. \tag{11}$$

With these preparations, we can provide the form of fair Bayes-optimal classifiers for linear disparity measures.

Theorem 14 (Form of Fair Bayes-optimal Classifiers for Linear Disparity Measures) *For a given linear disparity measure* Dis *as per Definition 10, suppose that for $a \in \{0, 1\}$, both $\eta_a(X)$ and $w_{\mathrm{Dis}}(X, a)$ have probability density functions on $\mathcal{X}$. Recalling $f_{\mathrm{Dis},t}$ from (9) and $t_{\mathrm{Dis}} : [0, \infty) \to \mathbb{R}$ from (11), for any $\delta \geqslant 0$,* $f_{\mathrm{Dis},t_{\mathrm{Dis}}(\delta)}$ is a δ-fair Bayes-optimal classifier *as per Definition 8. This Bayes-optimal classifier $f^\star_{\mathrm{Dis},\delta}$ of the form*

$$f^\star_{\mathrm{Dis},\delta}(x, a) := f_{\mathrm{Dis},t_{\mathrm{Dis}}(\delta)}(x, a) = I\left(\eta_a(x) > \frac{1}{2} + \frac{t_{\mathrm{Dis}}(\delta)}{2} w_{\mathrm{Dis}}(x, a)\right), \tag{12}$$

for all x, a. Furthermore, when the disparity measure is bilinear as per Definition 11, the above δ-fair Bayes-optimal classifier simplifies to a group-wise thresholding rule, such that for all x, a,

$$f^\star_{\mathrm{Dis},\delta}(x, a) = I\left(\eta_a(x) > \frac{1 + b_{\mathrm{Dis},a} \cdot t_{\mathrm{Dis}}(\delta)}{2 - s_{\mathrm{Dis},a} \cdot t_{\mathrm{Dis}}(\delta)}\right). \tag{13}$$

Chen et al. (2023) similarly characterizes—and develops practical algorithms for—finding δ-fair Bayes optimal classifiers under a specific form of composite constraints, which refer to differences of probabilities of a flipped version of $\widehat{Y}$ given specific pairs of values of the protected attribute (which are a bit more restricted than our linear disparity measures). They characterize fair Bayes optimal classifiers in a different but equivalent way, using certain *instance-level bias scores* $s_k(x) = f_k(x)/\eta_a(x)$, where f_k are weighting functions involving protected class attribute membership probabilities, related to our w_{Dis}. They show that the optimal fair classifiers take the form $\kappa(x) = 1\left(\sum_k z_k s_k(x) > 1\right)$, for some unspecified scalars z_k.

In contrast, our result above concerns only one constraint, but it provides the explicit form of the optimal classifier, without unspecified scalars. Moreover, our monotonicity results enable developing efficient binary-search based algorithms, while Chen et al. (2023) are limited to essentially trying all possible values of the hyperparameter settings, which can be computationally expensive. Our results also inspire pre- and in-processing algorithms, which cannot be as easily gleaned from their characterization. Specifically, changing the threshold t equivalently leads to an optimal classifier for a certain cost-sensitive risk; and minimizing a cost-sensitive risk is in turn equivalent to minimizing a re-sampled/re-weighted dataset according to our label shift up-/down-sampling mechanism.

Similar to the unconstrained case from Proposition 2, we are able to find the explicit form of Bayes-optimal fair classifiers as group-wise thresholding rules with explicitly characterized thresholds. Clearly, the accuracy is maximized by predicting the more likely class in each group. Moreover, mitigating disparity over protected groups necessitates shifting the thresholds. This shift depends delicately on the disparity measure and the population distribution. By incorporating the expressions from Proposition 12, we conclude the following corollary for the common disparity measures. In particular, this explicit form appears to be novel when a nonzero user-specified disparity level is given.

Corollary 15 (Bayes-optimal Classifiers for Three Common Disparity Measures) *For any $\delta > 0$, the group-wise thresholding rules such that for all x, a,*

$$f^{\star}_{\mathrm{DD},\delta}(x,a) = I\left(\eta_a(x) > \frac{1}{2} + \frac{(2a-1)t_{\mathrm{DD}}(\delta)}{p_a}\right);$$
$$f^{\star}_{\mathrm{DO},\delta}(x,a) = I\left(\eta_a(x) > \frac{p_{a,1}}{2p_{a,1} - (2a-1)t_{\mathrm{DO}}(\delta)}\right);$$
$$f^{\star}_{\mathrm{PD},\delta}(x,a) = I\left(\eta_a(x) > \frac{p_{a,0} + (2a-1)t_{\mathrm{PD}}(\delta)}{2p_{a,0} + (2a-1)t_{\mathrm{PD}}(\delta)}\right),$$

with

$$t_{\mathrm{DD}}(\delta) = \arg\min_t \left\{ |t| : \left| \int_{\mathcal{X}} I\left(\eta_1(x) > \frac{1}{2} + \frac{t}{2p_1}\right) d\mathbb{P}_{X|A=1}(x) - \int_{\mathcal{X}} I\left(\eta_0(x) > \frac{1}{2} - \frac{t}{2p_0}\right) d\mathbb{P}_{X|A=0}(x) \right| \leqslant \delta \right\};$$
$$t_{\mathrm{DO}}(\delta) = \arg\min_t \left\{ |t| : \left| \int_{\mathcal{X}} I\left(\eta_1(x) > \frac{p_{1,1}}{2p_{1,1} - t}\right) d\mathbb{P}_{A=1,Y=1}(x) - \int_{\mathcal{X}} I\left(\eta_0(x) > \frac{p_{0,1}}{2p_{0,1} + t}\right) d\mathbb{P}_{A=0,Y=1}(x) \right| \leqslant \delta \right\};$$
$$t_{\mathrm{PD}}(\delta) = \arg\min_t \left\{ |t| : \left| \int_{\mathcal{X}} I\left(\eta_1(x) > \frac{p_{1,0} + t}{2p_{1,0} + t}\right) d\mathbb{P}_{A=1,Y=0}(x) - \int_{\mathcal{X}} I\left(\eta_0(x) > \frac{p_{0,0} - t}{2p_{0,0} - t}\right) d\mathbb{P}_{A=0,Y=0}(x) \right| \leqslant \delta \right\},$$

are δ-fair Bayes-optimal classifiers under DD, DO *and* PD *from* (1), *respectively.*

4.2 Fairness-Accuracy Tradeoff: the Fair Pareto Frontier

Theorem 14 specifies fair Bayes-optimal classifiers for a given disparity level δ. A core challenge is to set a suitable value for δ, balancing accuracy and fairness. Identifying a suitable balance among multiple objectives can be viewed through the Pareto frontier. In the context of fair classification with a disparity measure Dis, we consider the following ordering relation on classifiers. A classifier f_1 is *Pareto dominant* over another classifier f_2, if one of the following conditions is met:

(1) $R(f_1) < R(f_2)$ and $\mathrm{Dis}(f_1) \leqslant \mathrm{Dis}(f_2)$, or (2) $R(f_1) \leqslant R(f_2)$ and $\mathrm{Dis}(f_1) < \mathrm{Dis}(f_2)$.

A classifier f is considered *Pareto optimal* if it is not dominated by any other classifier. The collection of all such Pareto optimal classifiers constitutes the *fair Pareto frontier* (FPF). Formally, the FPF is the solution of the following vector objective optimization problem, with respect to the ordering defined above:

$$\arg\min_{f\in\mathcal{F}} \left(R(f), \mathrm{Dis}(f)\right). \tag{14}$$

The following proposition characterizes the fair Pareto frontier:

Proposition 16 (Fair Pareto Frontier) *Let* t_{Dis} *be defined in* (11)*, and let* $\underline{t}_0 = \min(0, t_{\mathrm{Dis}}(0))$ *and* $\bar{t}_0 = \max(0, t_{\mathrm{Dis}}(0))$*. Then the fair Pareto frontier* FPF *from* (14) *includes all classifiers* $f_{\mathrm{Dis},t}$ *for* $t \in [\underline{t}_0, \bar{t}_0]$*, i.e., we have*

$$\{f_{\mathrm{Dis},t} : t \in [\underline{t}_0, \bar{t}_0]\} \subset \mathrm{FPF}.$$

Moreover, for any $f_{\mathrm{FPF}} \in \mathrm{FPF}$*, there is* $t \in [\underline{t}_0, \bar{t}_0]$ *such that* $f_{\mathrm{Dis},t}$ *has the same classification error and disparity as* f_{FPF}*, i.e.,*

$$R(f_{\mathrm{Dis},t}) = R(f_{\mathrm{FPF}}) \quad \textit{and} \quad \mathrm{Dis}(f_{\mathrm{Dis},t}) = \mathrm{Dis}(f_{\mathrm{FPF}}).$$

Now, consider an equivalence relation "$\sim$" between classifiers $f_1, f_2 \in \mathcal{F}$ such that $f_1 \sim f_2$ if and only if $R(f_1) = R(f_2)$ and $\mathrm{Dis}(f_1) = \mathrm{Dis}(f_2)$. Then, for each equivalence class determined by "$\sim$", there is a minimizer of (14) over $\{f_{\mathrm{Dis},t} : t \in [\underline{t}_0, \bar{t}_0]\}$. Hence, the classifiers $f_{\mathrm{Dis},t}$ fully characterize the Pareto frontier.

We next study the cost of fairness as a function of the disparity level. We define the following trade-off function that measures the best achievable performance for a given disparity level.

Definition 17 (Tradeoff Function) *For a disparity measure* Dis *and any* $\delta \geqslant 0$*, let*

$$T(\delta) = \inf\{R(f) : \mathrm{Dis}(f) \leqslant \delta\}. \tag{15}$$

From Proposition 16, it follows that for all $\delta \geqslant 0$, $T(\delta) = R(f_{\mathrm{Dis},t_{\mathrm{Dis}}(\delta)})$. The tradeoff function plays a crucial role in characterizing the misclassification rate and disparity along the Pareto frontier. This function is monotone non-increasing by definition, as a more stringent fairness constraint cannot increase accuracy. The following proposition establishes that this tradeoff function is *convex* for linear disparity measures. Thus, the lower the disparity level, the larger the additional accuracy cost of fairness.

Proposition 18 *If the disparity measure* Dis *is linear, then the tradeoff function* T *is convex on* $[0, \mathrm{Dis}(0)]$.

A similar FPF has also been recently studied in Wang et al. (2023) which focuses on characterizing the Pareto frontier between accuracy and a related but different set of fairness constraints. While our approach primarily deals with a single linear or bilinear disparity constraint and provides an exact characterization of the fair Pareto frontier in Proposition 16, the work of Wang et al. (2023) simultaneously considers a set of constraints (i.e., demographic parity, equalized odds, and overall accuracy equality—e.g., Berk et al. (2021)) and derives an upper bound on the optimal fairness-accuracy tradeoff using Blackwell's theory of statistical experiments.

A further distinction between our works is that Wang et al. (2023) provide a numerical procedure to approximate the class of realizable conditional confusion matrices, whereas we exactly characterize a FPF. However, in practice, we restrict ourselves to a model hypotheses space for the regression functions η_a (e.g. logistic regression or neural networks), thereby—as any algorithm does—providing a model-dependent lower bound. In particular, our work enables finding these lower bounds in new ways, via new pre- and in-processing algorithms.

4.3 Extensions

In this section, we discuss extensions of our theoretical framework to four scenarios: (1) The protected attribute A is excluded from the predictive features; (2) Equalized odds; (3) Demographic parity with a multi-class protected attribute; (4) cost-sensitive loss; and We provide the insights and main results here and defer details to Appendix B in the Appendix.

4.3.1 Protected Attribute Not Available at the Prediction Phase

In the previous discussion, we assumed that the protected feature is available for training and prediction; and that it is allowed to use it in both phases. Although this is a common setting (e.g., Hardt et al. (2016); Corbett-Davies et al. (2017); Cho et al. (2020), etc.), in certain cases, there may be ethical or legal considerations that invalidate it. However, our framework can also be applied to fair classification when the protected attribute cannot be used at the prediction phase. When the protected attribute is not available for *training*, the problem is very different, and it may require inferring the unobserved protected attribute; this is beyond our scope.

When A is available for training but not for prediction/testing, the classifier we use at test-time must be defined on $\mathcal{X}$ rather than on $\mathcal{X} \times \mathcal{A}$. In other words, a classifier is a measurable function $f_X\colon \mathcal{X} \to [0,1]$ with $\widehat{Y}_{f_X}|X \sim \mathrm{Bern}(f_X(X))$. In this section, in addition to writing $\eta^Y_{A=a}(x) = \eta^Y_a(x) := \mathbb{P}(Y=1 \mid A=a, X=x)$ for all x, a, we will further denote $\eta^Y(x) = \mathbb{P}(Y=1 \mid X=x)$ and $\eta^A(x) = \mathbb{P}(A=1 \mid X=x)$ for all $x \in \mathcal{X}$, as the regression functions of Y and A on X, respectively.

We call a disparity measure Dis_X linear if (7) holds with (x,a) replaced by x, i.e., there is a weight function $w_{X,\mathrm{Dis}} : \mathcal{X} \to \mathbb{R}$ such that, for all $f_X : \mathcal{X} \to [0,1]$,

$$\mathrm{Dis}_X(f_X) = \int_{\mathcal{X}} f_X(x) w_{X,\mathrm{Dis}}(x) d\mathbb{P}_X(x). \tag{16}$$

We show that, when the protected attribute A is not used for prediction, the three common disparity measures from (1) are still linear.

Proposition 19 (Common Disparity Measures are Linear when Not Using A) *When the classifier depends only on X rather than on (X,A), the disparity measures* DD*,* DO*, and* PD *from* (1) *are linear with weighting functions defined for all x by*

$$w_{\mathrm{DD}}(x) = \frac{\eta^A(x)}{p_1} - \frac{1-\eta^A(x)}{p_0}; \qquad w_{\mathrm{DO}}(x) = \frac{\eta^Y_{A=1}(x)\eta^A(x)}{p_{1,1}} - \frac{\eta^Y_{A=0}(x)(1-\eta^A(x))}{p_{0,1}};$$

$$w_{\mathrm{PD}}(x) = \frac{(1-\eta^Y_{A=1}(x))\eta^A(x)}{p_{1,0}} - \frac{(1-\eta^Y_{A=0}(x))(1-\eta^A(x))}{p_{0,0}}.$$

For a general disparity measure Dis, a δ-fair Bayes-optimal classifier is defined as $f^\star_{X,\mathrm{Dis},\delta} \in \arg\min_{f_X:\mathcal{X}\to[0,1]}\{R(f_X) : |\mathrm{Dis}_X(f_X))| \leqslant \delta\}$. We observe that the analysis of linear disparity measures also applies to the scenario, leading to the following result.

Corollary 20 (Bayes-Optimal Classifiers for Linear Disparity Measures, Not Using A**)** *When A is not used for prediction, for a given linear disparity measure Dis_X as defined in (16), suppose that both $\eta^Y(X)$ and $w_{X,\mathrm{Dis}}(X)$ have probability density functions on $\mathcal{X}$. Let $D_{X,\mathrm{Dis}} : \mathbb{R} \to [-1,1]$ be defined for all t as*

$$D_{X,\mathrm{Dis}}(t) := \int_{\mathcal{X}} \left[w_{X,\mathrm{Dis}}(x) \cdot I\left(\eta^Y(x) > \frac{1}{2} + \frac{t}{2} w_{X,\mathrm{Dis}}(x)\right)\right] d\mathbb{P}_X(x).$$

Let $t_{X,\mathrm{Dis}} : [0,\infty) \to \mathbb{R}$ be defined for all $\delta \geqslant 0$ by

$$t_{X,\mathrm{Dis}}(\delta) = \arg\min_t\{|t| : |D_{X,\mathrm{Dis}}(t)| \leqslant \delta\}. \tag{17}$$

Then, for any $\delta \geqslant 0$, there exists a δ-fair Bayes-optimal classifier $f^\star_{X,\mathrm{Dis},\delta}$ taking the form, for all $x \in \mathcal{X}$,

$$f^\star_{X,\mathrm{Dis},\delta}(x) = I\left(\eta^Y(x) > \frac{1}{2} + \frac{t_{X,\mathrm{Dis}}(\delta)}{2} w_{X,\mathrm{Dis}}(x)\right).$$

This result can be used to develop similar algorithms to those for the attribute-aware setting. Theorems 35 and 36 provide the form of the pre- and in-processing algorithms (FUDS and FCSC) for common linear disparity measures (i.e., demographic parity, equality of opportunity and predictive equality). Pseudocode for the resulting algorithms is provided in Algorithms 4 and 5.

Our algorithms leverage the special form of these metrics and the resulting Bayes-optimal classifiers that we derived, and do not hold for arbitrary linear disparity measures. In particular, it is not clear to us how to leverage/extend the results of Chen et al. (2023) to reach the same conclusion and algorithms.

4.3.2 Equalized Odds

We continue with two examples that handle multiple fairness constraints. In many applications, achieving fairness is a multifaceted challenge, needing simultaneous consideration of different fairness metrics. Additionally, there can be multiple protected attributes, which are often multi-class rather than binary. As an example, the U.S. Equal Employment Opportunity Commission recognizes 11 discrimination types, mostly with multi-class attributes. In such scenarios, managing multiple fairness constraints becomes crucial. To illustrate how our method can be extended to handle multiple fairness constraints, we consider two examples: (1) equalized odds, which requires ensuring *both* equality of opportunity *and* predictive equality, and (2) demographic parity with a multi-class protected attribute.

Recalling disparity of opportunity (DO) and predictive disparity (PD) from (1), a classifier f satisfies the δ-parity constraint with respect to equalized odds from Definition 6 if

$$\max(|\mathrm{DO}(f)|, |\mathrm{PD}(f)|) \leqslant \delta.$$

The corresponding δ-fair Bayes-optimal classifiers are then defined as

$$f^\star_{\mathrm{DEO},\delta} \in \arg\min_{f\in\mathcal{F}} \left\{R(f) : \max\{|\mathrm{DO}(f)|, |\mathrm{PD}(f)|\} \leqslant \delta\right\}. \tag{18}$$

By taking ϕ_0, ϕ_1, ϕ_2 such that for all x, a, $\phi_0(x,a) = 2\eta_a(x)-1$, $\phi_1(x,a) = (2a-1)\eta_a(x)/p_{a,1}$, $\phi_2(x,a) = (2a-1)(1-\eta_a(x))/p_{a,0}$ and applying the generalized Neyman Pearson lemma, after some intricate calculations, we can derive the following result.

Theorem 21 (Bayes-optimal Classifiers for Equalized Odds) *For any $\delta > 0$, there is $t_{\mathrm{DEO},1}(\delta) \in (-p_{0,1}, p_{1,1})$ and $t_{\mathrm{DEO},2}(\delta) \in (-p_{1,0}, p_{0,0})$ such that a δ-fair Bayes-optimal classifier takes values, for all x, a,*

$$f^\star_{\mathrm{DEO}}(x,a) = I\left(\ \eta_a(x) > \frac{p_{a,1}p_{a,0} + (2a-1)t_{\mathrm{DEO},2}(\delta)p_{a,1}}{2p_{a,1}p_{a,0} + (2a-1)\left(t_{\mathrm{DEO},2}(\delta)p_{a,1} - t_{\mathrm{DEO},1}(\delta)p_{a,0}\right)}\right).$$

The values of $t_{\mathrm{DEO},1}(\delta)$ and $t_{\mathrm{DEO},2}(\delta)$ are provided in (42) in Appendix B.1.

Prior work such as Chen et al. (2023) showed that the optimal classifier belongs to a linear combination of certain scoring functions, but did not find its exact form.

The proof of Theorem 21 requires studying intricate properties of the DO and PD, as detailed in a series of lemmas (Proposition 30, Lemmas 31, 32, 33), as well as discussing the seven cases that are involved in the definitions of $t_{\mathrm{DEO},1}(\delta)$ and $t_{\mathrm{DEO},2}(\delta)$ in (42). Briefly, we define a candidate class of potentially Bayes-optimal classifiers, parametrized by two scalars t_1, t_2; and show that all Bayes-optimal classifiers belong to this class in Lemma 32. We show that the DO and PD of these classifiers is continuous and monotone non-increasing as a function of either of these arguments while holding the other one fixed (Proposition 30). This leads to the definitions of $t_{\mathrm{DEO},1}(\delta)$ and $t_{\mathrm{DEO},2}(\delta)$ in (42); where the claim that they are well-defined is proved in Lemma 33. Then, we show that the most efficient approach, in terms of minimizing changes to PD, to attain a particular level of DO involves solely varying the value of t_1. Conversely, when aiming to achieve a specific level of PD, the optimal strategy is to solely adjust the value of t_2 (Lemma 31). The proof of Theorem 21 follows by careful arguments leveraging all these properties.

Developing an algorithm here seems to be more challenging than in the basic single constraint setting, as the optimal classifier is determined by two non-linear equations, whose efficient solution (beyond grid search) may be nontrivial.

4.3.3 Demographic Parity with a Multi-class Protected Attribute

Next, we illustrate how our approach can be extended to handle demographic parity with a multi-class protected attribute. For a multi-class protected attribute $a \in \mathcal{A} = \{1, 2, ..., |\mathcal{A}|\}$, we say that a classifier satisfies *perfect* demographic parity if the $|\mathcal{A}|-1$ linear constraints

$$\mathbb{P}(\widehat{Y}_f = 1 \mid A = a) = \mathbb{P}(\widehat{Y}_f = 1 \mid A = 1) \quad \text{for} \quad a = 2, 3, ..., |\mathcal{A}|$$

hold. A fair Bayes-optimal classifier is defined as

$$f^\star_{\mathrm{DD},|\mathcal{A}|} \in \arg\min_{f\in\mathcal{F}} \left\{\mathbb{P}(Y \neq \widehat{Y}_f) : \sum_{a=2}^{|\mathcal{A}|} \left|\mathbb{P}(\widehat{Y}_f = 1 \mid A = a) = \mathbb{P}(\widehat{Y}_f = 1 \mid A = 1)\right| = 0\right\}.$$

By leveraging the generalized Neyman-Pearson lemma, we can show the following result:

Theorem 22 (Bayes-optimal Classifier for Demographic Parity with a Multi-class Protected Attribute) *Let for $a \in \mathcal{A}$, $p_a = \mathbb{P}(A = a)$. Then, there are constants $(t_{\mathrm{DD},a})_{a=1}^{\mathcal{A}}$ with $\sum_{a=1}^{|\mathcal{A}|} t_{\mathrm{DD},a} = 0$, such that, for all $a \in \{2, \ldots, |\mathcal{A}|\}$,*

$$\mathbb{P}_{X|A=a}\left(\eta_a(X) > \frac{1}{2} + \frac{t_{\mathrm{DD},a}}{2p_a}\right) = \mathbb{P}_{X|A=1}\left(\eta_1(x) > \frac{1}{2} + \frac{t_{\mathrm{DD},1}}{2p_1}\right). \tag{19}$$

Moreover, there is a fair Bayes-optimal classifier such that for all x, a,

$$f_{\mathrm{DD},t_{\mathrm{DD},1},\ldots,t_{\mathrm{DD},|\mathcal{A}|}}(x,a) = I\left(\eta_a(x) > \frac{1}{2} + \frac{t_{\mathrm{DD},a}}{2p_a}\right).$$

Extending our algorithmic procedures to multi-class protected attributes is non-trivial, since there are no analogous monotonicity results like Proposition 13 that relate the behavior of the disparity measures to the group-wise thresholds. To see this, Theorem 22 reveals the form of the δ-fair Bayes optimal classifier under demographic parity in the attribute-aware setting to be $I\left(\eta_{a_i}(x) > \frac{1}{2} + \frac{1}{2}\frac{\beta_{a_i}}{p_{a_i}}\right)$ where $\sum_{i=1}^{|\mathcal{A}|} \beta_{a_i} = 0$ (in the binary protected attribute setting $\mathcal{A} = \{0, 1\}$, we have $t = \beta_1 = -\beta_0$). For $|\mathcal{A}| > 2$, increasing β_1 may either increase or decrease the demographic disparity between groups a_i and a_j for $i \neq j \neq 1$, so a result about the monotonicity of disparity measures does not hold in general. Thus, developing algorithmic analogs of FUDS and FCSC is not straightforward, since those rely on the monotonicity properties of the disparity measure in the hyperparameter t.

4.3.4 Cost-sensitive Risk

Cost-sensitive classification is useful when the consequences of false negatives and positives are unequally important, such as in certain medical applications. For a given cost parameter $c \in [0, 1]$, the cost-sensitive classification error of a classifier f is defined as:

$$R_c(f) = c \cdot \mathbb{P}(\widehat{Y}_f = 1, Y = 0) + (1 - c) \cdot \mathbb{P}(\widehat{Y}_f = 0, Y = 1). \tag{20}$$

When $c = 1/2$, cost-sensitive risk reduces to the usual mis-classification error.

Taking fairness into account, let $\mathcal{F}_\delta = \{f \in \mathcal{F} : |\mathrm{Dis}(f)| \leqslant \delta\}$ be the set of δ-fair classifiers. A δ-fair Bayes-optimal classifier for the cost-sensitive risk is defined as $f^\star_{\mathrm{Dis},\delta,c} \in \mathrm{argmin}_{f\in\mathcal{F}_\delta} R_c(f)$, or equivalently $f^\star_{\mathrm{Dis},\delta,C} \in \arg\max_{f\in\mathcal{F}_\delta} \int_{\mathcal{A}}\int_{\mathcal{X}} f(x,a)(\eta_a(x) - c)d\mathbb{P}_{X,A}(x,a)$. Since this is a linear objective, our approach extends seamlessly to find fair Bayes-optimal classifiers.

Corollary 23 (Fair Bayes-optimal Classifiers for a cost-sensitive Risk) *For a given linear disparity measure* Dis*, suppose that for $a \in \{0, 1\}$, both $\eta_a(X)$ and $w_{\mathrm{Dis}}(X, a)$ have probability density functions on $\mathcal{X}$. Let $D_{\mathrm{Dis},c} : \mathbb{R} \to [-1, 1]$ be defined for all t as*

$$D_{\mathrm{Dis},c}(t) := \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} [w_{\mathrm{Dis}}(x,a) \cdot I(\eta_a(x) > c + tw_{\mathrm{Dis}}(x,a))]\, d\mathbb{P}_{X|A=a}(x),$$

and let $t_{\mathrm{Dis},c} : [0, \infty) \to \mathbb{R}$ be defined as, for all $\delta \geqslant 0$, $t_{\mathrm{Dis},c}(\delta) = \arg\min_t\{|t| : |D_{\mathrm{Dis},c}(t)| \leqslant \delta\}$. Then, for any $\delta > 0$, there exists a δ-fair Bayes-optimal classifier $f^\star_{\mathrm{Dis,c}}$ under cost-sensitive risk, such that for all x, a,

$$f^\star_{\mathrm{Dis},c}(x,a) = I\left(\eta_a(c) > c + t_{\mathrm{Dis},c} \cdot w_{\mathrm{Dis},c}(x,a)\right),$$

5. Bayes-Optimal Fair Classifiers via Pre-, In-, and Post-processing

In this section, we propose a comprehensive set of algorithms for Bayes-optimal fair classification, based on pre-, in-, and post-processing. These algorithms aim to ensure fairness at different stages of the training process. Each class of methods has strengths and limitations. Pre- and post-processing methods can be used without significant changes to standard classification methods, facilitating the use of widely-available software. However, these approaches might potentially clash with certain data protection regulations (Barocas and Selbst, 2016). On the other hand, in-processing methods are applicable whenever standard training workflows are used, but may require modifying the fitting methods or optimization objectives.

We develop methods of each type aiming to recover the Bayes-optimal classifiers. Our methods include pre-processing via fair up-/down-sampling, in-processing via group-based cost-sensitive classification, and post-processing via plug-in estimation. Each method inherits the strengths mentioned above.

In this section, we consider bilinear disparity measures as per Definition 11 for simplicity. However, our methods are adaptable to the more general scenario of linear disparity measures, as shown in case studies in Section B.3.1 of the Appendix. Given a bilinear disparity measure characterized by $s_{\mathrm{Dis},a}$ and $b_{\mathrm{Dis},a}$ for $a \in \mathcal{A}$, we denote for $t \in \mathbb{R}$ the threshold function

$$H_{\mathrm{Dis},a}(t) = \frac{1 + t \cdot b_{\mathrm{Dis},a}}{2 - t \cdot s_{\mathrm{Dis},a}}. \tag{21}$$

Then, the disparity function D from (10) has values, for all $t \in \mathbb{R}$,

$$D_{\mathrm{Dis}}(t) = \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left[(s_{\mathrm{Dis},a}\eta_a(x) + b_{\mathrm{Dis},a}) \cdot I\left(\eta_a(x) > H_{\mathrm{Dis},a}(t)\right)\right] d\mathbb{P}_{X|A=a}(x). \tag{22}$$

In the following, we let the observed data points $\mathcal{S}_n = \{(x_i, a_i, y_i)\}_{i=1}^n$ be an independent and identically distributed sample from $\mathbb{P}$ over the domain $\mathcal{X}\times\mathcal{A}\times\mathcal{Y}$. Moreover, we separate the data according to the protected feature, letting for $a \in \{0,1\}$, $\mathcal{S}_{n,a} = \{(x_i, a_i, y_i) \in \mathcal{S}_n, a_i = a\}$. For $a \in \{0,1\}$, the j-th element of $\mathcal{S}_{n,a}$ is denoted as $(x_{a,j}, a, y_{a,j})$, for $j \in [n_a]$. We further separate the dataset $S_{n,a}$ into $\mathcal{S}_{n,a,1}$ and $\mathcal{S}_{n,a,0}$, according to the label information. Moreover, we denote, for $a, y \in \{0,1\}^2$, $n_a = |\mathcal{S}_{n,a}|$ and $n_{a,y} = |\mathcal{S}_{n,a,y}|$.

5.1 Pre-processing: Fair Up- and Down-Sampling (FUDS)

Many pre-processing algorithms use heuristics to eliminate the impact of protected attributes on the common features X and the label Y (Lum and Johndrow, 2016; Xu et al., 2018). Nonetheless, pre-processing a dataset to satisfy a fairness criterion does not guarantee fair outcomes, see e.g., Eitan et al. (2022). To illustrate this, consider a transformed distribution $(\widetilde{X}, \widetilde{A}, \widetilde{Y})$, where $(\widetilde{X}, \widetilde{Y})$ is independent of A. A classifier f trained on a sample with the distribution $(\widetilde{X}, \widetilde{Y})$ would be independent of A. Yet, during the testing phase, the classifier's output $f(X)$ may still exhibit unfairness if X and A are dependent. This observation leads us to ask: *On what data distributions* $(\widetilde{X}, \widetilde{A}, \widetilde{Y}) \sim \widetilde{\mathbb{P}}$ *can we learn classifiers without a fairness constraint and recover a fair Bayes-optimal classifier for the original distribution* $(X, A, Y) \sim \mathbb{P}$*?*

We propose an answer by designing a joint distribution for $(\widetilde{A}, \widetilde{Y})$, while keeping the conditional distribution of $\widetilde{X}$ given $\widetilde{A}$ and $\widetilde{Y}$ unchanged from that of X given A and Y.

We will consider distributions $\widetilde{\mathbb{P}}^t$ depending on a scalar $t \in \mathbb{R}$. For the distribution $\widetilde{\mathbb{P}}^t$ of $(\widetilde{X}, \widetilde{A}, \widetilde{Y})$, we denote for all x, a, y, $\widetilde{p}^t_{a,y} = \widetilde{\mathbb{P}}^t(\widetilde{A} = a, \widetilde{Y} = y)$, denote $\widetilde{\eta}^t_a(x) = \widetilde{\mathbb{P}}^t(\widetilde{Y} = 1|\widetilde{A} = a, \widetilde{X} = x)$, and denote by $\widetilde{\mathbb{P}}^t_{\widetilde{X}|\widetilde{A}=a,\widetilde{Y}=y}(\cdot)$ the marginal distribution of $\widetilde{X}$ given $\widetilde{A} = a$ and $\widetilde{Y} = y$.

Theorem 24 (Bayes-Optimal Fair Up- and Down-Sampling) *Suppose that for $a \in \{0, 1\}$, both $\eta_a(X)$ and $w_{\mathrm{Dis}}(X, a)$ have probability density functions on $\mathcal{X}$. For $t \in \mathbb{R}$, let $\widetilde{\mathbb{P}}^t$ be the distribution satisfying that for all $(x, a, y) \in \mathcal{X} \times \{0, 1\}^2$, $\widetilde{\mathbb{P}}^t_{\widetilde{X}|\widetilde{A}=a,\widetilde{Y}=y}(x) = \mathbb{P}_{X|A=a,Y=y}(x)$, and, with $H_{\mathrm{Dis},a}$ from (21),*

$$\widetilde{p}^t_{a,y} = \widetilde{\mathbb{P}}^t\left(\widetilde{A} = a, \widetilde{Y} = y\right) = c_a[(1 - H_{\mathrm{Dis},a}(t))y + H_{\mathrm{Dis},a}(t)(1 - y)]p_{a,y}. \tag{23}$$

Here $c_1 > 0$ and $c_0 > 0$ are such that $\widetilde{p}^t_{11} + \widetilde{p}^t_{10} + \widetilde{p}^t_{01} + \widetilde{p}^t_{00} = 1$. Then, the unconstrained Bayes-optimal classifier for $\widetilde{\mathbb{P}}^t$ is a $|D_{\mathrm{Dis}}(t)|$-fair Bayes-optimal classifier for $\mathbb{P}$. Thus, for a given $\delta \geqslant 0$, any Bayes-optimal classifier for $\widetilde{\mathbb{P}}^{t_{\mathrm{Dis}}(\delta)}$ is a δ-fair Bayes-optimal classifier for $\mathbb{P}$.

Theorem 24 designs a fairness-inducing distribution by multiplying the probabilities of each group with a group-wise factor, as per (23). As a result, $(1-H_{\mathrm{Dis},a}(t))y+H_{\mathrm{Dis},a}(t)(1-y)$ can be viewed as a fairness-inducing factor for the group with $A = a$ and $Y = y$. As can be seen from (22), a large threshold $H_{\mathrm{Dis},a}(t)$ for the group with $A = a$ indicates that the fair Bayes-optimal classifier specifies a higher acceptance standard for group $A = a$. In other words, the unconstrained classifier has a higher acceptance rate on the subgroup with $A = a$. To avoid this disparity, we need to generate more data with $Y = 0$ and less data with $Y = 1$, for the group with $A = a$; reducing $\eta_a(x) = \mathbb{P}(Y = 1|X = x, A = a)$. For this reason, we call our approach Fair Up- and Down-Sampling (FUDS).

An extension. It is possible to show more generally that if the δ-fair Bayes-optimal classifier has the form $I(\eta_a(x) > H(a, \delta))$ for a weighting function $H : \{0, 1\} \times [0, \infty) \to \mathbb{R}$ that does not depend on x, then re-weighting for all $(x, a, y) \in \mathcal{X} \times \{0, 1\}^2$ by $\widetilde{\mathbb{P}}^t_{\widetilde{X}|\widetilde{A}=a,\widetilde{Y}=y}(x) = \mathbb{P}_{X|A=a,Y=y}(x)$, yields

$$\widetilde{p}^t_{a,y} = \widetilde{\mathbb{P}}^t\left(\widetilde{A} = a, \widetilde{Y} = y\right) = c_a[(1 - H(a, \delta))y + H(a, \delta)(1 - y)]p_{a,y}.$$

These re-weighted probabilities similarly specify a distribution whose unconstrained Bayes-optimal classifier is a δ-fair Bayes-optimal classifier for original distribution.

However, in the more general setting of linear disparity measures, which includes the attribute blind setting (Section 4.3), developing pre- and in-processing algorithms is delicate. Though we have shown how to do this for DD, DO, PD, this does not extend to general classifiers of the form $I(\eta^Y(x) > H(x; a, \delta))$, where the weighting function $H(x; a, \delta)$ may depend on x. Nevertheless, we show that such algorithms do exist for these common fairness criteria in the attribute-blind setting, and we formalize these findings in Theorems 35 and 36.

Resampling as a pre-processing method. In the broader fairness literature, both reweighting and resampling are widely used approaches for developing fairness aware pre-processing algorithms (e.g., Kamiran and Calders (2012), Chai et al. (2022), Calmon et al. (2017), and Hort et al. (2024)). Reweighting and resampling are equivalent in our case: assigning weights to data points implicitly defines a corresponding resampling distribution, and vice versa. While we presented the re-sampling version here, our methods are equally applicable to developing group-wise re-weighting methods.

Data augmentation vs. Resampling methods. There have been a number of recent studies that compare generative augmentation methods to resampling techniques on tabular datasets. For instance, Panagiotou et al. (2024) find that models trained on synthetic data from CTGAN (Xu et al., 2019b) and TVAE (Patki et al., 2016) outperform SMOTE (Chawla et al., 2002) by tracing a more favorable fair Pareto frontier. More recently, Hastings Blow et al. (2025) demonstrate that diffusion-based tabular data augmentation improves upon the work of Kamiran and Calders (2012).

However, the verdict of which fairness-aware approach is better remains context-dependent: modern resampling methods such as FairBatch (Roh et al., 2020), Adaptive Priority Reweighing (Hu et al., 2023) can have advantages, especially when generative models cannot reliably approximate the conditional distribution $\mathbb{P}(X, Y|A = a)$.

Ultimately, our method, FUDS, uses group-aware resampling of the original dataset, which differs fundamentally from generative data augmentation. A direct empirical comparison with data augmentation frameworks would require extensive implementation and domain-specific tuning. We view this as an important direction for future work, but it lies beyond the scope of the current work.

5.1.1 Algorithmic implementation

While the sampling probabilities from (23) depend on unknown population parameters, we can estimate them from the training data. Then, we can generate approximately fair datasets using up- and down-sampling; see Algorithm 1. According to Theorem 24, for each value of t, there is a fair Bayes-optimal classifier for *some* level—$|D_{\rm Dis}(t)|$—of disparity. Therefore, by adjusting t, we can create a range of data distributions, each leading to a different point on the fair Pareto frontier as defined in (14). When we need a dataset with a predetermined disparity level δ, we can estimate the threshold parameter $t_{\rm Dis}(\delta)$ using an iterative method for updating t, see Algorithm 1.

Specifically, for a given t, we estimate $f_t^{\rm FUDS}$ and its disparity level as follows:

(1) For all x, a, y, t, estimate $p_{a,y}$ by $n_{a,y}/n$ and estimate $w_{\rm Dis}(x, a)$ and $H_{{\rm Dis},a}(t)$ from (8) and (21) by plug-in estimation. Taking demographic disparity as an example, we let for all x, a and for all $t \in \mathbb{R}$,

$$\widehat{w}_{\rm DD}(x,a) = \frac{(2a-1)n}{n_{a,1}+n_{a,0}}, \quad \text{and} \quad \widehat{H}_{{\rm DD},a}(t) = \frac{1}{2} + \frac{(2a-1)nt}{n_{a,1}+n_{a,0}}.$$

(2) Estimate the group-wise proportions, for $(a, y) \in \{0, 1\}^2$,

$$\widehat{\widetilde{p}}^{\,t}_{a,y} = \widehat{c}_a[(1-\widehat{H}_{{\rm Dis},a}(t))y + \widehat{H}_{{\rm Dis},a}(t)(1-y)] \cdot n_{a,y}/n. \tag{24}$$

Here $\widehat{c}_1$ and $\widehat{c}_2$ are such that $\widehat{\widetilde{p}}^{\,t}_{11} + \widehat{\widetilde{p}}^{\,t}_{10} + \widehat{\widetilde{p}}^{\,t}_{01} + \widehat{\widetilde{p}}^{\,t}_{00} = 1$.

Algorithm 1: Fair Up-/Down-Sampling (FUDS)

Input: Step size $\alpha > 0$; Disparity Level $\delta \geqslant 0$; Error tolerance level $\varepsilon > 0$; Dataset $S = S_1 \cup S_0$ with $S = \{x_i, a_i, y_i\}_{i=1}^{n}$, $S_1 = \{x_{1,i}, y_{1,i}\}_{i=1}^{n_1}$ and $S_0 = \{x_{0,i}, y_{0,i}\}_{i=1}^{n_0}$; $n_{a,y} = \#\{a_i = a, y_i = y\}$.

Disparity Estimation sub-routine: Construct $\widehat{f}_t^{\mathbf{FUDS}}$ and estimate $\mathbf{Dis}(\widehat{f}_t^{\mathbf{FUDS}})$ for a given threshold parameter t:

1. For all a, estimate $\widehat{w}_{\mathrm{Dis}}$ and $\widehat{H}_{\mathrm{Dis},a}$ using the expressions in (8) and (21).
2. For all a, y, let $\widehat{\widetilde{p}}_{a,y}^{\,t}$ be as in (24).
3. Apply up- and down-sampling to generate a new dataset S_t with proportions $\widehat{\widetilde{p}}_{a,y}^{\,t}$ for all a, y.
4. Fit classifier $\widehat{f}_t^{\mathrm{FUDS}}$ on the dataset S_t using any method.
5. Estimate the disparity level of $\widehat{f}_t^{\mathrm{FUDS}}$ as in (25).

Estimate the δ-fair Bayes-optimal classifier:

Run Disparity Estimation sub-routine with $t = 0$.

if $\left|\widehat{\mathrm{Dis}}^{\mathrm{FUDS}}(0)\right| \leqslant \delta$ **then**
  $\widehat{t}_{\mathrm{Dis}}(\delta) = 0$.
else
  if $\widehat{\mathrm{Dis}}^{\mathrm{FUDS}}(0) > \delta$ **then**
    $\delta' = \delta$; $t_{\min} = 0$, $t_{\max} = 1$.
  else
    $\delta = -\delta$; $t_{\min} = -1$, $t_{\max} = 0$.
  end
  while $t_{\max} - t_{\min} > \varepsilon$ **do**
    $t = (t_{\max} - t_{\min})/2$;
    Run Disparity Estimation sub-routine with current t.
    if $\widehat{\mathrm{Dis}}^{\mathrm{FUDS}}(t) > \delta$ **then**
      $t_{\max} = t$
    else
      $t_{\min} = t$
    end
  end
  $\widehat{t}_{\mathrm{Dis}}(\delta) = t$.
end

Output: $\widehat{f}_{\mathrm{Dis},\delta} = \widehat{f}^{\mathrm{FUDS}}_{\widehat{t}_{\mathrm{Dis}}(\delta)}$.

(3) Apply up- and down- sampling to generate a new dataset $S_t = \{\widetilde{x}_i, \widetilde{a}_i, \widetilde{y}_i\}_{i=1}^n$ with $\widetilde{n}_{a,y}^t = \#\{\widetilde{a}_i = a, \widetilde{y}_i = y\} = \lfloor n \cdot \widehat{\widetilde{p}}^t{}_{a,y} \rfloor$. The precise sampling method is presented in the paragraphs below.

(4) Fit a classifier $\widehat{f}_t^{\mathrm{FUDS}}$ to the dataset S_t using any method, such as logistic regression or deep neural nets.

(5) Estimate the disparity level of $\widehat{f}_t^{\mathrm{FUDS}}$ for the given t:

$$\widehat{\mathrm{Dis}}^{\mathrm{FUDS}}(t) = \frac{1}{n_1}\sum_{j=1}^{n_1} \widehat{f}_t^{\mathrm{FUDS}}(x_{1,j}, 1)\widehat{w}_{\mathrm{Dis}}(x_{1,j}, 1) + \frac{1}{n_0}\sum_{j=1}^{n_0} \widehat{f}_t^{\mathrm{FUDS}}(x_{0,j}, 0)\widehat{w}_{\mathrm{Dis}}(x_{0,j}, 0). \quad (25)$$

Following this, t can be adjusted using the bisection method for solving $D_{\mathrm{Dis}}(t) = 0$, based on the estimating equation $\widehat{\mathrm{Dis}}^{\mathrm{FUDS}}(t) = 0$. This is grounded in our observation that, as per Proposition 13, at a population level, the disparity function $\mathrm{Dis}(\widehat{f}_t^{\mathrm{FUDS}})$ is monotone non-increasing in t.

Performing independent random resampling of the original dataset S_n in each iteration can lead to significant variability in the datapoints—and also in the induced classifiers. To address this issue, we propose a sampling strategy aimed at reducing variability.

Consider the generated dataset $\widetilde{\mathcal{S}}_n^{t'}$ for a given threshold t', and write it as $\widetilde{\mathcal{S}}_n^{t'} = \cup_{a,y\in\{0,1\}^2}\widetilde{\mathcal{S}}_{n,a,y}^{t'}$, where for all a, y, $\widetilde{\mathcal{S}}_{n,a,y}^{t'}$ represents the set of points with $\widetilde{a}_i = a$ and $\widetilde{y}_i = y$. Let t be the next threshold, and let $\widetilde{n}_{a,y}^{t'} = \lfloor n \cdot \widetilde{p}_{a,y}^{t'} \rfloor$ and $\widetilde{n}_{a,y}^{t} = \lfloor n \cdot \widetilde{p}_{a,y}^{t} \rfloor$ for all a, y. Our strategy considers the following two cases, for all a, y:

- If $\widetilde{n}_{a,y}^t \geqslant \widetilde{n}_{a,y}^{t'}$, we randomly select $\widetilde{n}_{a,y}^t - \widetilde{n}_{a,y}^{t'}$ data points from $S_{n,a}$ to add to $\widetilde{\mathcal{S}}_{n,a,y}^{t'}$, forming $\widetilde{\mathcal{S}}_{n,a,y}^{t}$.
- If $\widetilde{n}_{a,y}^t < \widetilde{n}_{a,y}^{t'}$, we randomly select $\widetilde{n}_{a,y}^t$ data points from $\widetilde{\mathcal{S}}_{n,a,y}^{t'}$ to form $\widetilde{\mathcal{S}}_{n,a,y}^{t}$.

Our simulation studies indicate that this strategy offers greater stability than independent random sampling in each iteration, leading to a more favorable fairness-accuracy tradeoff and reduced variance in the empirical performance.

Another challenge is posed by computational complexity. As discussed above, we fit a classifier at each iteration and then estimate its disparity level. For larger datasets and more complex models, this procedure can become time-consuming. A faster method is to adjust t while fitting each classifier, fine-tuning the optimal sampling ratio concurrently with model training. We only need to revise step (4): for instance, when training via an iterative optimization-based method, rather than finishing the entire training procedure, we only train for a small number of steps—or, epochs—in each iteration to obtain a classifier $\widehat{f}_t^{\mathrm{FUDS}}$. The classifier obtained in the each iteration is used as the starting point in the next iteration.

5.2 In-Processing: Fair cost-sensitive Classification (FCSC)

In this section, we propose an in-processing fair classification method. Our objective is to identify a risk function whose unconstrained minimizer is the fair Bayes-optimal classifier.

The key observation is that fair Bayes-optimal classifiers adjust thresholds for each protected group, which is also known to occur in cost-sensitive classification.

Recall the cost-sensitive risk R_c from (20). An unconstrained Bayes-optimal classifier for this cost-sensitive risk, denoted as $f_c^\star$, is a classifier that minimizes $R_c(f)$, i.e., $f_c^\star \in \arg\min_{f\in\mathcal{F}} R_c(f)$. All such classifiers can be represented as $f_c^\star(x,a) = I(\eta_a(x) > c)$ for all x, a, (see e.g., Elkan, 2001). Drawing inspiration from this connection, we propose a *group-wise* cost-sensitive classification approach. We aim to shift the thresholds used for the protected groups in opposite directions to reduce disparity.

Definition 25 (Fair cost-sensitive Risk) *Recall* $H_{\mathrm{Dis},a}$, $a \in \mathcal{A}$ *from* (21) *and let* $c_{a,y}(t) = (1-2y)H_{\mathrm{Dis},a}(t) + y$ *for all* t, a, y. *For* $t \in \mathbb{R}$, *define the following* fair cost-sensitive risk *of a classifier* f*:*

$$R_t^{\mathrm{FCSC}}(f) = \sum_{a\in\{0,1\}} \sum_{y\in\{0,1\}} c_{a,y}(t)\mathbb{P}(\widehat{Y}_f = 1-y, Y = y, A = a).$$

Fair cost-sensitive classification (FCSC) minimizes the fair cost-sensitive risk from Definition 25:

$$f_t^{\mathrm{FCSC}} = \arg\min_{f\in\mathcal{F}} R_t^{\mathrm{FCSC}}(f). \tag{26}$$

Our key result here is that fair cost-sensitive classification is Bayes-optimal.

Theorem 26 (Fair cost-sensitive Classification is Bayes-Optimal) *Letting* $t_0^\star = t_{\mathrm{Dis}}(0) = \arg\min_t\{|t| : |D_{\mathrm{Dis}}(t)| = 0\}$, *we have, for* $t \in [\min(t_0^\star, 0), \max(t_0^\star, 0)]$, *with* $D_{\mathrm{Dis}}(t)$ *defined in* (10), *that* f_t^{FCSC} *from* (26) *is a* $|D_{\mathrm{Dis}}(t)|$*-fair Bayes-optimal classifier for* $\mathbb{P}$. *In particular,* $f_{t_{\mathrm{Dis}}(\delta)}^{\mathrm{FCSC}}$ *is a* δ*-fair Bayes-optimal classifier for* $\mathbb{P}$.

Theorem 26 shows that carefully designed group-wise cost-sensitive classification can lead to a fair Bayes-optimal classifier. Similar to the fair sampling method from Section 5.1, one can generate a fairness-accuracy tradeoff curve by using FCSC and varying t. To achieve a specific disparity level δ, an updating method similar to that from Algorithm 1 can be used to estimate $t_{\mathrm{Dis}}(\delta)$; see Algorithm 2.

Similar to the FUDS algorithm, the computational efficiency of FCSC can be improved by adjusting the threshold parameter t and the associated loss function after a few training epochs (rather than after the whole training process in step 3).

5.3 Post-Processing: Fair Plug-in Thresholding Rule (FPIR)

Finally, we consider a post-processing algorithm that modifies the model output to control the disparity. Given the near-explicit form of Bayes-optimal classifiers from Theorem 14, after fitting any classifier $\widehat{f}$, we can estimate the group-wise thresholds. Such an approach has been proposed in a more limited case for demographic parity and equality of opportunity in Menon and Williamson (2018). In contrast, our framework is applicable to *any* bilinear disparity measure.

First, we can use any method to estimate the feature-conditional probability of $Y = 1$ for each protected group. Second, we estimate the threshold for each protected group. One can either vary the value of t to generate fairness-accuracy tradeoff curves or solve the estimated one-dimensional fairness equation for a specific disparity level.

Algorithm 2: Fair cost-sensitive Classification (FCSC)

Input: Step size $\alpha > 0$; Disparity Level $\delta \geqslant 0$, Error tolerance level $\varepsilon > 0$; Dataset $S = S_1 \cup S_0$ with $S = \{x_i, a_i, y_i\}_{i=1}^{n}$, $S_1 = \{x_{1,i}, y_{1,i}\}_{i=1}^{n_1}$ and $S_0 = \{x_{0,i}, y_{0,i}\}_{i=1}^{n_0}$; $n_{a,y} = \#\{a_i = a, y_i = y\}$; Step size $\alpha \geqslant 0$.

Disparity Estimation sub-routine: Construct $\widehat{f}_t^{\mathbf{FCSC}}$ and estimate $\mathbf{Dis}(\widehat{f}_t^{\mathbf{FCSC}})$ for a given threshold parameter t:

1. For all a, estimate $\widehat{w}_{\mathrm{Dis}}$ and $\widehat{H}_{\mathrm{Dis},a}$ by plug-in estimation using the expressions in (8) and (21).
2. Denote $\widehat{c}_{a,y}(t) = (1-2y)\widehat{H}_{\mathrm{Dis},a}(t) + y$, for all t, a, y.
3. Use any cost-sensitive classification method to fit $\widehat{f}_t^{\mathrm{FCSC}}(\cdot, 1)$ on S_1 and $\widehat{f}_t^{\mathrm{FCSC}}(\cdot, 0)$ on S_0.
4. Estimate the disparity level of $\widehat{f}_t^{\mathrm{FCSC}}$ as in (25) with $\widehat{f}_t^{\mathrm{FCSC}}$ instead of $\widehat{f}_t^{\mathrm{FUDS}}$.

Estimate the δ-fair Bayes-optimal classifier:

Run Disparity Estimation sub-routine with $t = 0$. **if** $\left|\widehat{\mathrm{Dis}}^{\mathrm{FCSC}}(0)\right| \leqslant \delta$ **then**
  $\widehat{t}_{\mathrm{Dis}}(\delta) = 0$.
else
  if $\widehat{\mathrm{Dis}}^{\mathrm{FCSC}}(0) > \delta$ **then**
    $\delta' = \delta$; $t_{\min} = 0$, $t_{\max} = 1$.
  else
    $\delta = -\delta$; $t_{\min} = -1$, $t_{\max} = 0$.
  end
  while $t_{\max} - t_{\min} > \varepsilon$ **do**
    $t = (t_{\max} - t_{\min})/2$;
    Run Disparity Estimation sub-routine with current t.
    if $\widehat{\mathrm{Dis}}^{\mathrm{FCSC}}(t) > \delta$ **then**
      $t_{\max} = t$
    else
      $t_{\min} = t$
    end
  end
  $\widehat{t}_{\mathrm{Dis}}(\delta) = t$.
end

Output: $\widehat{f}_{\mathrm{Dis},\delta} = \widehat{f}^{\mathrm{FUDS}}_{\widehat{t}_{\mathrm{Dis}}(\delta)}$.

Algorithm 3: Fair Plug-in Rule (FPIR)

Input: Step size $\alpha > 0$; Disparity level $\delta \geqslant 0$; Error tolerance level $\varepsilon > 0$; Dataset $S = S_1 \cup S_0$ with $S = \{x_i, a_i, y_i\}_{i=1}^{n}$, $S_1 = \{x_{1,i}, y_{1,i}\}_{i=1}^{n_1}$ and $S_0 = \{x_{0,i}, y_{0,i}\}_{i=1}^{n_0}$; $n_{a,y} = \#\{a_i = a, y_i = y\}$.

Step 1: Construct estimate $\widehat{\eta}_a$ of η_a using any approach, for all a.

Step 2: Estimate the the fair Pareto frontier or the δ-fair Bayes-optimal classifier:

Disparity Estimation sub-routine: Construct $\widehat{f}_t^{\mathbf{FPIR}}$ and estimate $\mathbf{Dis}(\widehat{f}_t^{\mathbf{FPIR}})$ with threshold parameter t:

1. For all a, estimate $\widehat{w}_{\mathrm{Dis}}$ and $\widehat{H}_{\mathrm{Dis},a}$ by plug-in estimation using the expressions in (8) and (21).
2. Let $\widehat{f}_t^{\mathrm{FPIR}}$ be defined by $\widehat{f}_t^{\mathrm{FPIR}}(x,a) = I(\widehat{\eta}_a(x) > \widehat{H}_{\mathrm{Dis},a}(t))$ for all x, a.
3. Evaluate the disparity level of $\widehat{f}_t^{\mathrm{FPIR}}$ as in (25) with $\widehat{f}_t^{\mathrm{FPIR}}$ instead of $\widehat{f}_t^{\mathrm{FUDS}}$.

Estimate the δ-fair Bayes-optimal classifier:

Run Disparity Estimation sub-routine with $t = 0$.

if $\left|\widehat{\mathrm{Dis}}^{\mathrm{FPIR}}(0)\right| \leqslant \delta$ **then**
  $\widehat{t}_{\mathrm{Dis}}(\delta) = 0.$
else
  if $\widehat{\mathrm{Dis}}^{\mathrm{FPIR}}(0) > \delta$ **then**
    $\delta' = \delta$; $t_{\min} = 0$, $t_{\max} = 1$.
  else
    $\delta = -\delta$; $t_{\min} = -1$, $t_{\max} = 0$.
  end
  while $t_{\max} - t_{\min} > \varepsilon$ **do**
    $t = (t_{\max} - t_{\min})/2$;
    Run Disparity Estimation sub-routine with current t.
    if $\widehat{\mathrm{Dis}}^{\mathrm{FPIR}}(t) > \delta$ **then**
      $t_{\max} = t$
    else
      $t_{\min} = t$
    end
  end
  $\widehat{t}_{\mathrm{Dis}}(\delta) = t.$
end

Output: $\widehat{f}_{\mathrm{Dis},\delta} = \widehat{f}^{\mathrm{FPIR}}_{\widehat{t}_{\mathrm{Dis}}(\delta)}$.

Specifically, let $\widehat{\eta}_a$ be an estimator of η_a, and consider the following plug-in rule for all x, a:

$$\widehat{f}_t^{\mathrm{FPIR}}(x,a) = I\left(\widehat{\eta}_a(x) > \widehat{H}_{\mathrm{Dis},a}\right),$$

where $\widehat{H}_{\mathrm{Dis},a}$ is estimated using the expression in (21). Now, our goal is to construct an estimate $\widehat{t}_\delta$ such that $\widehat{f}_{\widehat{t}_\delta}$ approximately satisfies the fairness constraint. We define $\widehat{\mathrm{Dis}}(t)$, an estimator of the disparity function $D_{\mathrm{Dis}}(t)$, as in (25) with $\widehat{f}_t^{\mathrm{FPIR}}$ instead of $\widehat{f}_t^{\mathrm{FUDS}}$, where $\widehat{w}_{\mathrm{Dis}}$ is estimated using the plug-in estimators $\widehat{p}_{a,y} = n_{a,y}/n$ in (8). We then follow the definition of $t_{\mathrm{Dis}}(\delta)$ in (11) to estimate it by

$$\widehat{t}_{\mathrm{Dis}}(\delta) = \arg\min_t \{|t| : |\widehat{\mathrm{Dis}}(t)| \leqslant \delta\}.$$

Since $\widehat{\mathrm{Dis}}(t)$ is monotone non-increasing as a function of t, $\widehat{t}_{\mathrm{Dis}}(\delta)$ can be estimated via the bisection method, as summarized in Algorithm 3. Our final FPIR estimator of the fair Bayes-optimal classifier is $\widehat{f}^{\mathrm{FPIR}}_{\widehat{t}_{\mathrm{Dis}}(\delta)}$.

We also note the recent findings of Cruz and Hardt (2024), which suggest that post-processing methods can achieve better fair Pareto frontiers compared to pre- and in-processing approaches. While this might imply that post-processing methods alone are sufficient, pre- and in-processing approaches play distinct and valuable roles, making relevant techniques such as resampling and reweighting essential components of the broader fairness toolbox. One of the main contributions of our work is to provide such tools via a unified framework for characterizing the δ-fair Bayes optimal classifier and for tracing the fair Pareto frontier.

6. Empirical Experiments

6.1 Synthetic Datasets

In this section, we conduct simulation studies to illustrate the numerical performance of our proposed methods. We consider a label-conditional normal distribution, for which the fair Bayes-optimal classifier has a simple closed form.

Data-generating process. For a positive integer dimension p, let $X = (X_1, X_2, ..., X_p)^\top \in \mathbb{R}^p$ be the usual features, $A \in \{0,1\}$ be the protected attribute and $Y \in \{0,1\}$ be the label. Recalling the notations from the end of Section 2, we generate A and Y according to the probabilities $p_{1,1} = 0.49$, $p_{1,0} = 0.21$ $p_{0,1} = 0.12$ and $p_{0,0} = 0.18$. Conditional on $A = a$ and $Y = y$, we generate X from a multivariate Gaussian distribution $\mathcal{N}(\mu_{a,y}, \sigma^2 I_p)$, where I_p is the p-dimensional identity covariance matrix, and σ^2 controls the variability of the feature entries. The entries of $\mu_{a,y}$ are sampled from $\mu_{ay,j} \sim \mathrm{Unif}(0,1)$, $j = 1, \ldots, p$, where $\mathrm{Unif}(0,1)$ is the uniform distribution over $[0,1]$. In this model, η_a has a closed form, and we can use it to find the fair Bayes-optimal classifier, see Appendix C. Intriguingly, under this model, group-wise thresholding rules are linear in x.

Experimental setting. To evaluate our FUDS, FCSC and FPIR algorithms, we randomly generate $10,000$ training data points and $5,000$ test data points. Since the group-wise thresholding rules are linear in x, we use logistic regression in our methods. All experiments

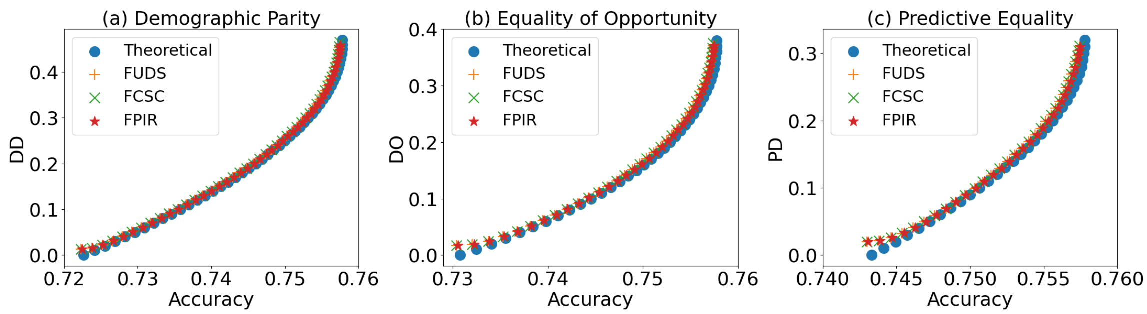

Figure 1: Numerical simulation results showing the fairness-accuracy tradeoff the fair Bayes-optimal classifier (fair Pareto frontier) and the three proposed methods (FUDS, FCSC, FPIR) using logistic regression.

are conducted in python and we use the scikit-learn package to train the logistic regression model. This package allows us to specify label weights for cost-sensitive classification, and thus works seamlessly for FCSC. In our experiments, we consider three fairness metrics: demographic parity, equality of opportunity and predictive equality.

We first evaluate the three algorithms with various pre-determined levels of disparity. We present the simulation results in Table 2 of the Appendix. We observe that all three methods control the disparity level at the pre-determined value, as desired. We further present the fairness-accuracy tradeoff curve of fair Bayes-optimal classifiers (the fair Pareto frontier from Section 4.2) and our three methods in Figure 1. Our classifiers closely track the fair Pareto frontier.

6.2 Empirical Data Analysis

To further illustrate our proposed methods, we compare our method with strong baseline methods on standard datasets.

Data Description. We consider four benchmark datasets: AdultCensus, COMPAS, LawSchool, and ACSIncome. For the first three datasets, we follow the pre-processing pipeline introduced in Cho et al. (2020), while for ACSIncome, we adopt the pipeline from Xian and Zhao (2024).

(1) AdultCensus: In the AdultCensus dataset, the target variable Y is whether the income of an individual is more than \$50,000. Age, marriage status, education level, and other related variables are included in X, and the protected attribute A refers to gender.

(2) COMPAS: In the COMPAS dataset, Y indicates whether or not a criminal will re-offend. Here X includes prior criminal records, age, and an indicator of misdemeanor. The protected attribute A is the race of an individual, “white-vs-non-white”.

(3) LawSchool: The target variable Y in the LawSchool dataset is whether a student gets admitted to law school. Here X includes the LSAT scores, undergraduate GPA and more. The protected attribute A we consider is the race, “white-vs-non-white”.

(4) ACSIncome: This dataset extends the AdultCensus dataset by incorporating more recent and comprehensive demographic and economic information from the American Community Survey (Ding et al. (2021)). The target variable Y indicates whether an individual’s total income exceeds \$50,000. The covariates X include age, education level, employment

status, hours worked per week, occupation, and marital status. The protected attribute A is gender.

Algorithms Considered. We select ten methods from the literature, comprising of 3 pre-processing, 4 in-processing, and 3 post-processing methods.

(1) Disparate Impact Remover (DIR, Feldman et al. (2015)):

DIR is a pre-processing method that aims to modify the value of Y to $\widetilde{Y}$, ensuring that the probability of $\widetilde{Y}=1$ is equal across different protected groups, specifically, $\mathbb{P}(\widetilde{Y}=1|A=1)=\mathbb{P}(\widetilde{Y}=1|A=0)$. The key idea of DIR is to align the cumulative distribution functions of $\eta^Y_{A=1}(X)$ and $\eta^Y_{A=0}(X)$ while maximizing similarity between the transformed and original datasets.

(2) Fairness-Aware Oversampling (FAWOS, Salazar et al. (2021)):

FAWOS is a pre-processing method that uses data augmentation to achieve fairness. Unlike our FUDS method that adjusts the sizes of all groups, FAWOS applies SMOTE (Chawla et al., 2002)——a popular synthetic data augmentation method for unbalanced classification problems—to increase the number of data points in the positive unprivileged group ($A=0, Y=1$). The number of data points generated is $N=\alpha_F\times(n_{11}n_{00}/n_{10}-n_{01})$, with α_F tuned to control the accuracy-fairness tradeoff.

(3) Adaptive Reweighing Algorithm (ARA, Chai et al. (2022))

Adaptive reweighing is a pre-processing technique that aims to induce fairness via finding weights for a cost-sensitive classification problem. It proceeds by iteratively (1) finding individual-level weights by solving a ℓ_2-regularized quadratic programming (QP) problem for each protected subgroup, and (2) fitting a classifier with the resulting weights via cost-sensitive classification. Although fairness constraints are not directly encoded in the iterative algorithm, fairness is encouraged by choosing an appropriate level of normalization through a hyperparameter in the QP step.

(4) KDE-based constrained optimization (KDE, Cho et al. (2020)):

KDE-based optimization is an in-processing technique that uses kernel density estimation (KDE) to approximate fairness constraints. The proposed estimator of disparity is a differentiable function with respect to the model parameters. KDE-based optimization uses this estimated disparity as a regularizer for empirical risk minimization.

(5) Adversarial training (ADV, Zhang et al. (2018)):

Adversarial training is an in-processing method that involves simultaneously training two models: (1) a classifier to predict the label and (2) an adversary that attempts to predict protected attributes from the classifier's outputs. This dual-training process encourages the classifier to learn predictable representations that are devoid of biases related to protected attributes, thereby promoting fairness in its predictions.

(6) Reductions (RED, Agarwal et al. (2018))

Reductions is an iterative in-processing algorithm. At each round, it trains a cost-sensitive classifier that balances error and fairness-violation penalties (using current Lagrange multipliers), and updates those multipliers using the observed constraint slack. Ultimately, it returns a randomized ensemble of the resulting classifiers.

(7) MinDiff (MD, Prost et al. (2019))

MinDiff is an in-processing algorithm that adds a fairness regularization term to the original loss function. The regularizer minimizes the Maximum Mean Discrepancy (Gretton et al. (2006)) between the distributions of score functions of the binary protected attribute subpopulations using an appropriate universal kernel (e.g. Gaussian or Laplace).

(8) Post-processing through flipping (PPF, Chen et al. (2023)):

Chen et al. (2023) showed that a Bayes-optimal fair classifier can be obtained by flipping the output of the unconstrained classifier and proved that the flipping probability satisfies an estimation equation. Based on these observations, they designed a post-processing method that first estimates the flipping probability of each output and then estimates the fair Bayes-optimal classifier.

(9) Post-processing through optimal transport (PPOT, Xian et al. (2023)):

Xian et al. (2023) proved that the fair classification problem with bounded demographic parity is equivalent to a Wasserstein-barycenter problem, and the fair Bayes-optimal classifier is given by the composition of the Bayes-optimal score function η_a and the optimal transport map from the Wasserstein-barycenter problem. They proposed a post-processing algorithm to estimate the fair Bayes-optimal classifier from the estimated score functions $\widehat{\eta}_a$.

(10) Group Fairness Framework for Post-processing Everything (FRAPPÉ, Tifrea et al. (2023))

FRAPPÉ is a post-processing method that converts any regularized in-processing fairness approach into a plug-in post-processing adjustment by learning a lightweight additive correction module. It is based on an optimization problem that minimizes a divergence between the base and adjusted model outputs on data without group attributes, while also incorporating the original in-processing fairness regularizer on protected-attribute data.

Experimental Setting: We follow the training settings from Cho et al. (2020) and Xian and Zhao (2024). With the exception of the pre-processing method ARA (Chai et al., 2022) across all datasets, and our in-processing method FCSC and the in-processing method RED (Agarwal et al., 2018) on the AdultCensus dataset, all methods across all datasets use a three-layer fully connected neural network and are trained with the Adam optimizer using $(\beta_1, \beta_2) = (0.9, 0.999)$, the default hyperparameters. For the AdultCensus, COMPAS, and LawSchool datasets, each hidden layer contains 32 neurons; for the ACSIncome dataset, we use the same optimizer configuration but adopt a larger architecture with hidden layers of sizes 500, 200, and 100. The ARA method is trained using a simple logistic regression model. The FCSC and RED methods on the AdultCensus dataset are trained using a histogram-based gradient boosting tree configured with 300 trees, a maximum depth of 4, a learning rate of 0.05, and an ℓ_2 regularization strength of 1.0.[3]

For adversarial training (Zhang et al., 2018), we use a linear classifier as the discriminator. For FRAPPÉ (Tifrea et al., 2023), the post-hoc module builds off the KDE-based fairness regularizer of Cho et al. (2020), using a single-layer perceptron with 128 hidden nodes for all datasets. The training batch size, training epochs, starting learning rate, and learning rate decay factor for the four datasets are summarized in Table 3.

Let n_{epoch} represent the total number of training epochs. For the FUDS and FCSC methods, we initially pre-train the model by training the neural network for $n_{\text{epoch}}/4$ epochs. Subsequently, we apply the bisection method as described in Algorithms 1 and 2 to update t after every $n_{\text{epoch}}/20$ epochs. Through this approach, we update t a total of 15 times, resulting in a tolerance level of $2^{-15} \sim 3 \times 10^{-5}$. For adversarial training, we first pre-train the classifier over $n_{\text{epoch}}/4$ epochs and then update the weights of both the classifier and discriminator in the later $3n_{\text{epoch}}/4$ epochs.

We repeat each experiment 50 times with different random seeds for the AdultCensus, COMPAS, and LawSchool datasets, and 10 times for ACSIncome. For each dataset and random seed, we randomly split the data into training and test sets (with a 70%-30% split). As a result, the randomness of the experiments arises from the stochasticity of the train-test set split, the neural network initialization, and the minibatch selection during the optimization.

Computational complexity: As outlined above, for our pre- and in-processing algorithms, we iteratively update the classifier rather than train an entirely new classifier at each iteration, so the total number of (stochastic) gradient steps used for estimating the δ-fair Bayes optimal classifier is

$$\mathcal{O}\left(\left(n_{\text{epoch}}^{\text{train}} \cdot \frac{N}{B}\right) + \left(\left(n_{\text{epoch}}^{\text{iter}} \cdot \frac{N}{B}\right) \cdot \log(1/\varepsilon) + N \log(1/\varepsilon)\right)\right)$$

where $n_{\text{epoch}}^{\text{train}}$ and $n_{\text{epoch}}^{\text{iter}}$ are the number of epochs used to train the unconstrained classifier and iteratively update the classifier, respectively. Here, N is the size of the dataset, B is the batch size and ε is the tolerance level. The computational complexity of the post-processing algorithm consists of only the latter term $\mathcal{O}(N \log(1/\varepsilon))$, since we assume access to a pre-trained classifier. This term specifies the complexity associated with repeatedly evaluating the disparity level for the trained classifier at each iteration of the bisection search. In particular, across all algorithms, we set $\varepsilon = 2^{-15}$, resulting in at most 15 updates to the parameter t via the bisection search.

These computational complexities are similar to those of state of the art methods. For example, the reductions method of Agarwal et al. (2018) has a runtime of $\mathcal{O}\left((n_{\text{epoch}}^{\text{iter}} \cdot N/B) \cdot 1/\nu^2\right)$ (where ν is the ν-approximate saddle point per their Theorem 1, ν typically ≈ 0.01). For post-processing, FPIR requires a computational runtime of $\mathcal{O}(N \log(1/\varepsilon))$ where for $\varepsilon = 2^{-15}$, $1/\varepsilon = 32768$. This is comparable to our sample sizes

3. The other in-processing baseline methods– ADV (Zhang et al., 2018), MinDiff (Prost et al., 2019), and KDE (Cho et al., 2020) – are not easily compatible with gradient boosting trees as base classifiers, since these approaches rely on adding a fairness regularization term to the loss function and require backpropagation through the classifier.

N, which are typically on the order of thousands to tens of thousands, so this complexity matches that of Chen et al. (2023).

Simulation Results for Our Methods: We first evaluate the FUDS, FCSC, and FPIR algorithms with various pre-determined levels of disparity. We present the simulation results in Table 4 . We observe that the proposed three methods control the disparity level at the pre-determined values, as desired. We then compare the fairness-accuracy tradeoff of the proposed methods to the baseline.

Experimental Setting for Controlling Disparity: In our proposed methods—FUDS, FCSC, and FPIR—along with the two other post-processing methods, the level of disparity is directly controlled. We set the range from zero to the empirical DD of the unconstrained classifier.

For the DIR method, as suggested by Feldman et al. (2015), a "partial remover" tool is employed. This tool features a parameter λ_{DIR}, which varies from zero (indicating the original dataset) to 1 (indicating perfect distribution alignment between $\eta^Y_{A=1}(X)$ and $\eta^Y_{A=0}(X)$). We adjust λ_{DIR} from 0.05 to 0.95 to explore the fairness-accuracy tradeoff.

In FAWOS, we vary α_{FAWOS} from zero to two. Higher values of α_F lead to an excess of data points in the positive unprivileged group, potentially causing reverse discrimination. We omit the result of this method on the ACSIncome dataset, as FAWOS requires computing nearest neighbors for each data point which was computationally infeasible in our experiments for the ACSIncome dataset ($\approx$1.66 million data points).

For ARA, we vary α_{ARA} from 0.40 to 0.50 on the COMPAS dataset. Higher values of α_{ARA} correspond to more uniform reweighting, which results in reduced fairness. For the other datasets, we were unable to meaningfully decrease demographic disparity.[4]

Both KDE-based constrained optimization and MinDiff balance fairness and accuracy by adjusting a tuning parameter that trades off empirical loss against fairness regularization. We vary the tuning parameter λ_{KDE} from 0.05 to 0.95, and λ_{MinDiff} from 0.0 to 10.0.

In adversarial training, the tradeoff is controlled by altering the parameter α_{ADV} that controls the gradient of the discriminator. We vary this parameter from zero to three, as we empirically find that, similar to FAWOS, a larger α_{ADV} would result in reverse discrimination.

Simulation Results for Controlling Disparity: Figure 2 presents the fairness-accuracy tradeoff with respect to Demographic Disparity (DD) evaluated across four datasets. In the plot, each point corresponds to a specific tuning parameter.

Pre-processing. Among the three pre-processing methods, our FUDS algorithm exhibits the most favorable fairness-accuracy tradeoff and allows for direct control of disparity.

In contrast, while the DIR method shows satisfactory accuracy, it falls short in achieving perfect fairness. This is because, although DIR ensures equal probabilities of positive outcomes across different groups in the transformed dataset—i.e., $\mathbb{P}(\tilde{Y}=1|A=1)=\mathbb{P}(\tilde{Y}=1|A=0)$—this does not automatically ensure that a fair model is learned. This shortcoming is readily apparent in the results of DIR on the ACSIncome dataset, where it fails to substantially reduce the disparity level. The third method, FAWOS, effectively manages disparity, but its heuristic-based sampling strategy strays from the ideal proportion we

4. To address this, we contacted the authors and explored several potential remedies, including alternative classifier model spaces and adjustments to hyperparameters within the QP subproblem, but still could not achieve a significant reduction in the disparity.

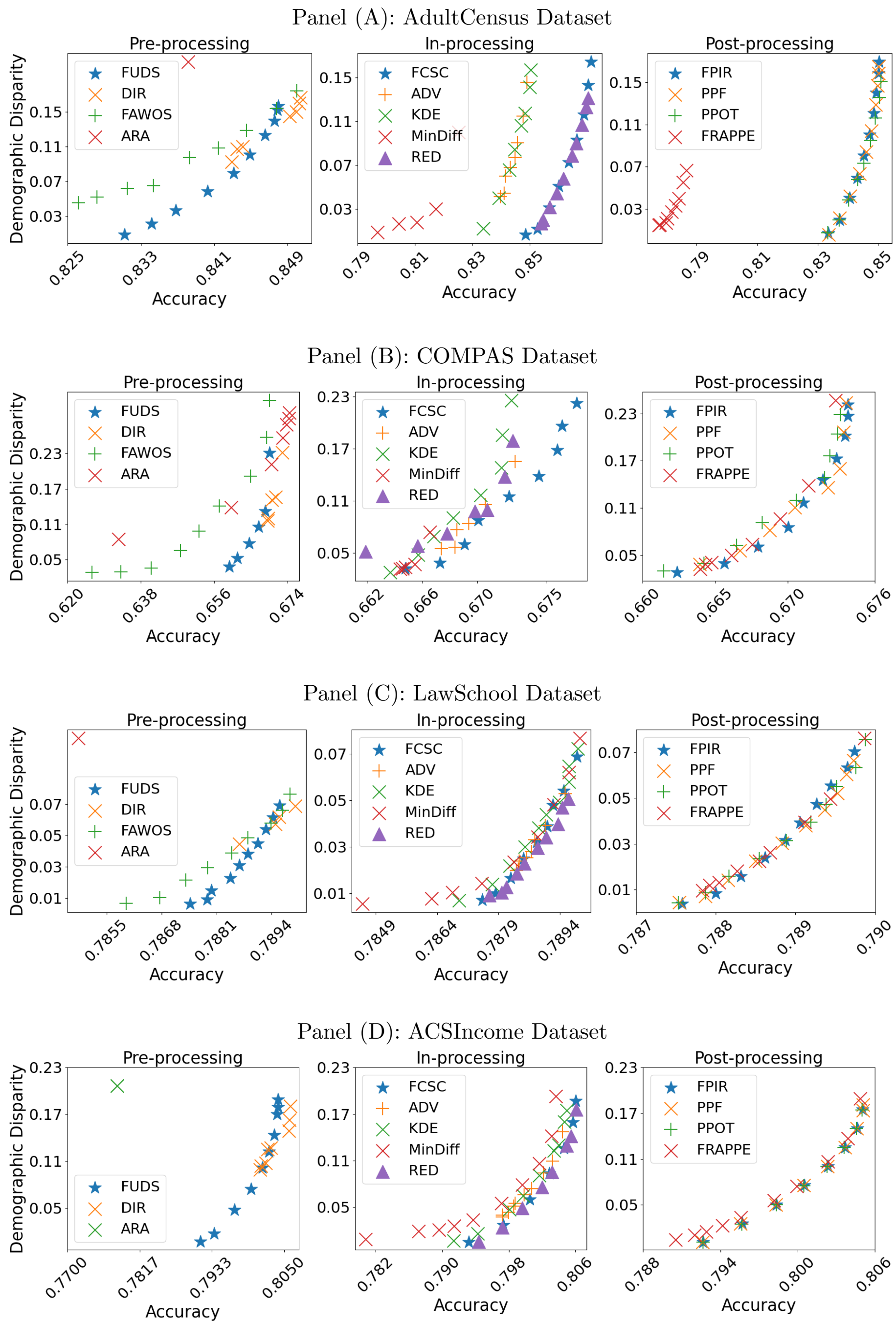

Figure 2: Fairness-accuracy tradeoff on the AdultCensus (Panel (A)), COMPAS (Panel (B)), LawSchool (Panel (C)), and ACSIncome (Panel (D)) datasets.

identified. As a result, compared to FUDS, FAWOS compromises more on accuracy at all levels of disparity.

In-processing. For the three in-processing methods, FCSC outperforms all baselines on the COMPAS dataset and demonstrates comparable or better results on the other three datasets. The MinDiff method of Prost et al. (2019) performs equally as well as FCSC and the other baselines (KDE and ADV) on the Lawschool and ACSIncome datasets, but has trouble tracing a reasonable Pareto frontier on AdultCensus.

The KDE method experiences a sudden drop in accuracy when the disparity level approaches zero. This issue may caused by its use of a Huber surrogate loss to address the non-differentiability of the absolute value function at $\delta = 0$. Compared with FCSC and KDE-based optimization, adversarial training exhibits inferior performance and fails to achieve near-perfect fairness. This shortcoming could be attributed to the challenges associated with minimax optimization, which often leads to unstable training processes.

Finally, the RED method traces a Pareto optimal frontier that is comparable with our FCSC method across all datasets, with the exception of COMPAS, for which it performs slightly worse across all levels of disparity. Similar to FCSC, RED can leverage a gradient boosting tree as its base classifier, thereby enabling strong performance on the AdultCensus dataset where such a model is more accurate.

Post-processing. Finally, with the exception of FRAPPÉ on the AdultCensus dataset, we observe that the four post-processing methods present comparable performance in balancing fairness and accuracy. This similarity arises because they all aim to estimate the fair Bayes-optimal classifier. However, FPIR can be viewed as a more direct approach.

Computational Runtimes. We also report runtimes measuring the average execution time of a single run per seed for the COMPAS and ACSIncome datasets (see Figure 3). All experiments are conducted on a computing cluster equipped with Intel Xeon Platinum 8375C CPUs @ 2.90GHz processors. For the COMPAS dataset, each run is allowed to use up to 10 CPU cores with 0.8 GB of RAM, while for ACSIncome, each run uses a single core with access of up to 32 GB of RAM.

The pre-processing method termed ARA of Chai et al. (2022) runs much quicker than its pre-processing counterparts since the classifier is an ordinary logistic regression estimator. However, ARA achieves a sub-optimal Pareto frontier on COMPAS and fails to trace any meaningful frontier for the other datasets we considered.

The post-processing methods do not include the runtime of the pre-trained classifiers which the post-hoc fairness corrections rely upon. The time required to obtain this unconstrained classifier—that does not impose any fairness constraints or fairness algorithms—is denoted as *BASE*.

Our method's runtime is either comparable or slightly better than that of competing methods with Pareto-optimal performance across all types of algorithms (pre-, in-, and post-processing).

In summary, our methods have state-of-the-art performance compared to existing algorithms. In particular, our pre-processing method can reach a higher accuracy than prior pre-processing methods at low disparity levels.

7. Summary and Discussion

In this paper, we study the statistical and methodological underpinnings of fair classification problems. We establish a unified framework for deriving fair Bayes-optimal classifiers. We introduce the notion of linear and bilinear disparity measures, and find Bayes-optimal classifiers under these constraints. Our framework can be applied in a wide range of scenarios, including various disparity measures, cost-sensitive loss functions, and multiple fairness constraints while allowing for direct control of the disparity level. Additionally, we study the Pareto frontier for fair classification and prove that the tradeoff function is convex for linear group fairness measures.

Based on our theoretical results, we further design pre-, in-, and post-processing methods that handle algorithmic bias. The proposed methods aim to recover the optimal fairness-accuracy tradeoff. Moreover, they allow for direct control of the disparity level.

Our work suggests several promising directions for further theoretical and algorithmic developments. From the theoretical perspective, the fair Bayes-optimal classifier only characterizes the best solution at the population level, and the finite-sample properties need to be characterized. From the algorithmic perspective, it is of interest to improve our algorithms. For example, for our pre-processing method, we only considered fair sampling. However, with the desired proportions in hand, we can generate fair synthetic data using more advanced conditional generative models, such as conditional GANs and conditional diffusion models.

Recent work by Xian and Zhao (2024) presents results for multi-class optimal fair classifiers via post-processing. It would be of interest, but it does not seem completely straightforward, to adapt our in- and pre-processing algorithms to this setting.

Furthermore, our analysis supposes the availability of protected attributes during training. However, the collection and use of protected information is often restricted. For instance, under the General Data Protection Regulations in the European Union, acquiring protected personal information requires consent from individuals. This motivates studying methods for protected attributes only partially available during training. Recently, Chai et al. (2022) proposed addressing this challenge through knowledge distillation, which generates soft labels for protected attributes. Exploring the integration of this technique into our framework presents an intriguing avenue for future research.

Acknowledgements

This work was supported in part by ARO W911NF-23-1-0296, NSF 2031895, NSF 2046874, NSF 2247795, ONR N00014-21-1-2843, and the Sloan Foundation. We are grateful to Junyi Chai and Xiaoqian Wang for insightful discussions regarding their work on the ARA method, and to Ruicheng Xian for his assistance with the implementation of the reductions approach.

References

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

I. Alabdulmohsin. Fair classification via unconstrained optimization. *arXiv preprint arXiv:2005.14621*, 2020.

W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P. Michalak, S. Asoodeh, and F. Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems*, 35:38747–38760, 2022.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, May 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

S. Barocas and A. D. Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016.

R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, 2019.

R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

K. Bhanot, M. Qi, J. S. Erickson, I. Guyon, and K. P. Bennett. The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165, 2021.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

S. Caton and C. Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2023.

L. E. Celis and V. Keswani. Improved adversarial learning for fair classification. *arXiv preprint arXiv:1901.10443*, 2019.

L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328. Association for Computing Machinery, 2019a.

L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019b.

J. Chai, T. Jang, and X. Wang. Fairness without demographics through knowledge distillation. In *Advances in Neural Information Processing Systems*, pages 19152–19164. Curran Associates, Inc., 2022.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002.

W. Chen, Y. Klochkov, and Y. Liu. Post-hoc bias scoring is optimal for fair classification. *arXiv preprint arXiv:2310.05725*, 2023.

J. Cho, G. Hwang, and C. Suh. A fair classifier using kernel density estimation. In *Advances in Neural Information Processing Systems*, pages 15088–15099. Curran Associates, Inc., 2020.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. Association for Computing Machinery, 2017.

S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312):1–117, 2023.

A. Cotter, H. Jiang, M. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019.

E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel. Flexibly fair representation learning by disentanglement. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1436–1445. PMLR, 2019.

A. Cruz and M. Hardt. Unprocessing seven years of algorithmic fairness. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jr03SfWsBS.

Z. Deng, C. Dwork, and L. Zhang. Happymap: A generalized multicalibration method. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, 2023.

C. Denis, R. Elie, M. Hebiri, and F. Hu. Fairness guarantee in multi-class classification. *arXiv preprint arXiv:2109.13642*, 2021.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer New York, NY, 1996.

F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, page 214–226. Association for Computing Machinery, 2012.

Y. Eitan, N. Cavaglione, M. Arbel, and S. Cohen. Fair synthetic data does not necessarily lead to fair models. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.

C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, pages 973–978. Morgan Kaufmann Publishers Inc., 2001.

Executive Office of the President. Big data: A report on algorithmic systems, opportunity, and civil rights. The White House, 2016. URL https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. Association for Computing Machinery, 2015.

B. Fish, J. Kun, and Á. D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.

A. W. Flores, K. Bechtel, and C. T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to Machine Bias: "There's software used across the country to predict future criminals. And it's biased against blacks". *Federal Probation*, 80(2):38–46, 2016.

G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

M. Gupta and Q. Mohammad. Advances in AI and ML are reshaping healthcare. TechCrunch, Mar. 2017. URL https://techcrunch.com/2017/03/16/advances-in-ai-and-ml-are-reshaping-healthcare/.

M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016.

C. Hastings Blow, L. Qian, C. Gibson, P. Obiomon, and X. Dong. Data augmentation via diffusion model to enhance ai fairness. *Frontiers in Artificial Intelligence*, 8:1530397, 2025.

M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2):1–52, 2024.

X. Hou and L. Zhang. Finite-sample and distribution-free fair classification: Optimal trade-off between excess risk and fairness, and the cost of group-blindness. *arXiv preprint arXiv:2410.16477*, 2024.

Z. Hu, Y. Xu, and X. Tian. Adaptive priority reweighing for generalizing fairness improvement. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE, 2023.

T. Jang, F. Zheng, and X. Wang. Constructing a fair classifier with generated fair data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7908–7916, 2021.

T. Jang, P. Shi, and X. Wang. Group-aware threshold adaptation for fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6988–6995, 2022.

R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, pages 862–872. PMLR, 2020.

J. E. Johndrow and K. Lum. An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 2019.

M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016.

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

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

P. Lahoti, K. P. Gummadi, and G. Weikum. iFair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345. IEEE, 2019.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer New York, NY, 3rd edition, 2005.

P. Li, J. Zou, and L. Zhang. FaiREE: fair classification with finite-sample and distribution-free guarantee. In *The Eleventh International Conference on Learning Representations*, 2023.

F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

C. Louizos, K. Swersky, Y. Li, M. Welling, and R. S. Zemel. The variational fair autoencoder. In *4th International Conference on Learning Representations (ICLR)*, 2016.

K. Lum and J. Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.

X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, 31:24–39, 2018.

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

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

P. Nandy, C. DiCiccio, D. Venugopalan, H. Logan, K. Basu, and N. El Karoui. Achieving fairness via post-processing in web-scale recommender systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, page 715–725. Association for Computing Machinery, 2022.

H. Narasimhan. Learning with complex loss functions and constraints. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 1646–1654. PMLR, 2018.

H. Narasimhan, A. Cotter, M. Gupta, and S. Wang. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5248–5255, 2020.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

L. Oneto, M. Donini, and M. Pontil. General fair empirical risk minimization. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

E. Panagiotou, A. Roy, and E. Ntoutsi. Synthetic tabular data generation for class imbalance and fairness: A comparative study. *arXiv preprint arXiv:2409.05215*, 2024.

N. Patki, R. Wedge, and K. Veeramachaneni. The synthetic data vault. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 399–410. IEEE, 2016.

G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

F. Prost, H. Qian, Q. Chen, E. H. Chi, J. Chen, and A. Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*, 2019.

B. Rava, W. Sun, G. M. James, and X. Tong. A burden shared is a burden halved: A fairness-adjusted approach to classification. *arXiv preprint arXiv:2110.05720*, 2021.

Y. Roh, K. Lee, S. E. Whang, and C. Suh. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*, 2020.

H. Royden and P. Fitzpatrick. *Real Analysis*. Pearson, 4th edition, 2010.

T. Salazar, M. S. Santos, H. Araújo, and P. H. Abreu. FAWOS: Fairness-aware oversampling algorithm based on distributions of sensitive attributes. *IEEE Access*, 9:81370–81379, 2021.

P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3:1–3:9, 2019.

N. Schreuder and E. Chzhen. Classification with abstention but without disparities. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 1227–1236. PMLR, 2021.

J. Shao. *Mathematical Statistics*. Springer New York, NY, 2nd edition, 2003.

C. Silvia, J. Ray, S. Tom, P. Aldo, J. Heinrich, and A. John. A general approach to fairness with optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3633–3640, 2020.

A. Tifrea, P. Lahoti, B. Packer, Y. Halpern, A. Beirami, and F. Prost. Frappé: A group fairness framework for post-processing everything. *arXiv preprint arXiv:2312.02592*, 2023.

S.-C. Tsai, C.-H. Chen, Y.-T. Shiao, J.-S. Ciou, and T.-N. Wu. Precision education with statistical learning and deep learning: A case study in Taiwan. *International Journal of Educational Technology in Higher Education*, 17(12):1–13, 2020.

UNESCO. Recommendation on the ethics of artificial intelligence. UNESCO, 2021. URL `https://unesdoc.unesco.org/ark:/48223/pf0000380455`.

I. Valera, A. Singla, and M. Gomez Rodriguez. Enhancing the accuracy and fairness of human decision making. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018.

C. Wadsworth, F. Vera, and C. Piech. Achieving fairness through adversarial learning: An application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018.

H. Wang, L. He, R. Gao, and F. Calmon. Aleatoric and epistemic discrimination: Fundamental limits of fairness interventions. *Advances in Neural Information Processing Systems*, 36:27040–27062, 2023.

Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8916–8925. IEEE, 2020.

H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio. Fairlearn: Assessing and improving fairness of AI systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023.

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

R. Xian and H. Zhao. A unified post-processing framework for group fairness in classification. *arXiv preprint arXiv:2405.04025*, 2024.

R. Xian, L. Yin, and H. Zhao. Fair and optimal classification via post-processing. In *Proceedings of the 40th International Conference on Machine Learning*, pages 37977–38012. PMLR, 2023.

D. Xu, S. Yuan, L. Zhang, and X. Wu. FairGAN: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, 2018.

D. Xu, Y. Wu, S. Yuan, L. Zhang, and X. Wu. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1452–1458. International Joint Conferences on Artificial Intelligence Organization, 2019a.

L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019b.

S. Xu and T. Strohmer. Fair data representation for machine learning at the Pareto frontier. *Journal of Machine Learning Research*, 24(331):1–63, 2023.

M. B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

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

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pages 325–333. PMLR, 2013.

X. Zeng, E. Dobriban, and G. Cheng. Bayes-optimal classifiers under group fairness. *arXiv preprint arXiv:2202.09724*, 2022a.

X. Zeng, E. Dobriban, and G. Cheng. Fair Bayes-optimal classifiers under predictive parity. In *Advances in Neural Information Processing Systems*, pages 27692–27705. Curran Associates, Inc., 2022b.

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. Association for Computing Machinery, 2018.

J. Zhang and E. Bareinboim. Fairness in decision-making - The causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

L. Zhang, A. Roth, and L. Zhang. Fair risk control: A generalized framework for calibrating multi-group fairness risks. In *Proceedings of the 41st International Conference on Machine Learning*, pages 59783–59805. PMLR, 2024.

Appendix

Additional notation and conventions. In this appendix, we use some additional notation. For a real-valued function f defined on $[a, b)$ for some $a < b$, we denote by $\lim_{x\to a^+} f(x)$ the limit from the right of f at a, if it exists. Similarly, if f is defined on $(b, a]$ for $b < a$, we denote by $\lim_{x\to a^-} f(x)$ the limit from the left of f at a, if it exists. For an interval $[a, b]$, and scalars $c \in \mathbb{R}$, $d > 0$, we denote $c+d[a, b] = [c+da, c+db]$. For a classifier f, we denote $\mathrm{Acc}(f) = 1 - R(f)$. When needed, we define $0/0 := 0$.

Appendix A. Proofs

In Sections A.1 to A.4, we present the proofs of our theoretical results from the main text, except for Theorem 21 and Theorem 22. We provide an independent Appendix B to discuss the interesting extension of our theoretical and methodological framework. We first introduce several technical lemmas that are essential for proving our theoretical results (Section A.1).

A.1 Additional Lemmas

Lemma 27 Risk and Accuracy as Linear Functionals of Classifiers *For any classifier $f : \mathcal{X} \times \mathcal{A} \to [0, 1]$, we have*

$$R(f) = \sum_{a\in\{0,1\}} p_a \int (1 - 2\eta_a(x))\, f(x, a) d\mathbb{P}_{X|A=a}(x) + \sum_{a\in\{0,1\}} p_a \int \eta_a(x) d\mathbb{P}_{X|A=a}(x), \quad (27)$$

and

$$\mathrm{Acc}(f) = \sum_{a\in\{0,1\}} p_a \int (2\eta_a(x) - 1)\, f(x, a) d\mathbb{P}_{X|A=a}(x) + \sum_{a\in\{0,1\}} p_a \int (1 - \eta_a(x))\, d\mathbb{P}_{X|A=a}(x), \quad (28)$$

Proof

By definition, $\widehat{Y}$ is conditionally independent of Y given X and A. Thus,

$$\mathbb{P}(\widehat{Y}_f = 1, Y = 0 \mid X = x, A = a) = f(x, a)(1 - \eta_a(x)),$$
$$\mathbb{P}(\widehat{Y}_f = 0, Y = 1 \mid X = x, A = a) = \eta_a(x)(1 - f(x, a)).$$

This implies that

$$R(f) = \mathbb{P}(Y \neq \widehat{Y}_f) = \sum_{a\in\mathcal{A}} p_a \mathbb{P}(Y \neq \widehat{Y}_f | A = a)$$
$$= \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \Big(\mathbb{P}(\widehat{Y}_f = 1, Y = 0 \mid X = x, A = a) + \mathbb{P}(\widehat{Y}_f = 0, Y = 1 \mid X = x, A = a)\Big)\, d\mathbb{P}_{X|A=a}(x)$$
$$= \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} (f(x, a)(1 - \eta_a(x)) + \eta_a(x)(1 - f(x, a)))\, d\mathbb{P}_{X|A=a}(x)$$

$$= \sum_{a\in\{0,1\}} p_a \int (1-2\eta_a(x))\, f(x,a) d\mathbb{P}_{X|A=a}(x) + \sum_{a\in\{0,1\}} p_a \int \eta_a(x) d\mathbb{P}_{X|A=a}(x).$$

Moreover,

$$\begin{aligned}\mathrm{Acc}(f) = 1 - R(f) = 1 - &\sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left(f(x,a)(1-\eta_a(x)) + \eta_a(x)(1-f(x,a))\right) d\mathbb{P}_{X|A=a}(x)\\ &= \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} 1 - \eta_a(x) - (1-2\eta_a(x)) f(x,a) d\mathbb{P}_{X|A=a}(x)\\ &= \sum_{a\in\{0,1\}} p_a \int (2\eta_a(x)-1)\, f(x,a) d\mathbb{P}_{X|A=a}(x) + \sum_{a\in\{0,1\}} p_a \int (1-\eta_a(x))\, d\mathbb{P}_{X|A=a}(x).\end{aligned}$$

■

Lemma 28 Characterizing Bayes-Optimal Classifiers for Linear Disparity Measures) *For $t \in \mathbb{R}$ and $\delta \geqslant 0$, let $f_{\mathrm{Dis},t}$, $D_{\mathrm{Dis}}(t)$ and $t_{\mathrm{Dis}}(\delta)$ be defined in* (9), (10) *and* (11), *respectively. Then, using the convention that $0/0 = 0$, for any fixed t, we have*

$$f_{\mathrm{Dis},t} = \arg\min_{f\in\mathcal{F}} \left\{ R(f) : \frac{t\cdot \mathrm{Dis}(f)}{|t|} \leqslant \frac{t\cdot D_{\mathrm{Dis}}(t)}{|t|} \right\}.$$

Moreover, for all classifiers $f' \in \arg\min_{f\in\mathcal{F}} \{R(f) : t\cdot\mathrm{Dis}(f)/|t| \leqslant t\cdot D_{\mathrm{Dis}}(t)/|t|\}$, $f' = f_{\mathrm{Dis},t}$ almost surely with respect to $\mathbb{P}_{X,A}$. In addition, if $t \in [\min(0, t_{\mathrm{Dis}}(0)), \max(0, t_{\mathrm{Dis}}(0))]$,

$$f_{\mathrm{Dis},t} = \arg\min_{f\in\mathcal{F}} \{R(f) : |\mathrm{Dis}(f)| \leqslant |D_{\mathrm{Dis}}(t)|\}.$$

Proof

This result is a consequence of the generalized Neyman-Pearson lemma. If $t = 0$, the result follows since $f_{\mathrm{Dis},0}$ is the unconstrained Bayes-optimal classifier. When $t \neq 0$, let, for all x, a, $\phi_0(x,a) = 2\eta^Y_{A=1}(x,a) - 1$, $\phi_1(x,a) = w_{\mathrm{Dis}}(x)$ and for all classifiers f, and $t \in \mathbb{R}$, $\overline{\mathrm{Dis}}_t(f) = t\cdot\mathrm{Dis}(f)/|t|$. We have, for all x, a,

$$f_{\mathrm{Dis},t}(x,a) = I\left(\phi_0(x,a) > t\phi_1(x,a)\right) = I\left(\phi_0(x,a) > |t|\frac{t\phi_1(x,a)}{|t|}\right).$$

Moreover, by Lemma 27 and (7), we can write $\mathrm{Acc}(f)$ and $\overline{\mathrm{Dis}}_t(f)$ as

$$\begin{aligned}\mathrm{Acc}(f) &= \int_{\mathcal{A}}\int_{\mathcal{X}} f(x,a)\phi_0(x,a) d\mathbb{P}_{X,A}(x,a) + \int_{\mathcal{A}}\int_{\mathcal{X}} (1-\eta_a(x)) d\mathbb{P}_{X,A}(x,a);\\ \overline{\mathrm{Dis}}_t(f) &= \frac{t}{|t|}\mathrm{Dis}(f) = \int_{\mathcal{X}\times\mathcal{A}} f(x,a)\frac{t\phi_1(x,a)}{|t|} d\mathbb{P}_{X,A}(x,a).\end{aligned}$$

Define:

$$\mathcal{F}_{t,=} = \left\{ f : \overline{\mathrm{Dis}}_t(f) = \frac{tD_{\mathrm{Dis}}(t)}{|t|} \right\}; \quad \mathcal{F}_{t,|\cdot|,\leqslant} = \left\{ f : \left|\overline{\mathrm{Dis}}_t(f)\right| \leqslant \frac{tD_{\mathrm{Dis}}(t)}{|t|} \right\};$$

$$\text{and} \quad \mathcal{F}_{t,\leqslant} = \left\{ f : \overline{\mathrm{Dis}}_t(f) \leqslant \frac{tD_{\mathrm{Dis}}(t)}{|t|} \right\}.$$

It is clear that $f_{\mathrm{Dis},t} \in \mathcal{F}_{t,=} \subset \mathcal{F}_{t,\leqslant}$. Since $|t| \geqslant 0$, by the generalized Neyman-Pearson lemma (Lemma 9),

$$f_{\mathrm{Dis},t} \in \underset{f \in \mathcal{F}_{t,\leqslant}}{\arg\max}\, \mathrm{Acc}(f),$$

and, since $\mathbb{P}_{X|A=a}(\eta_a(X) = 1/2 + t_{\mathrm{Dis}}(\delta) w_{\mathrm{Dis}}(X,a)/2) = 0$ when both $\eta_a(X)$ and $w_{\mathrm{Dis}}(X,a)$ are continuous random variables with respect to $\mathbb{P}_{X|A=a}$, for all $f' \in \arg\max_{f \in \mathcal{F}_{t,\leqslant}} \mathrm{Acc}(f)$, $f' = f_{\mathrm{Dis},t}$ almost surely with respect to $\mathbb{P}_{X,A}$.

Now, suppose further that $t \in [\min(0, t_{\mathrm{Dis}}(0)), \max(0, t_{\mathrm{Dis}}(0))]$. By result (1) of Proposition 13, $D_{\mathrm{Dis}}(t)$ is monotone non-increasing with respect to t. By the definition of $t_{\mathrm{Dis}}(\delta)$ in (11), we have $t_{\mathrm{Dis}}(0) \geqslant 0$ when $D_{\mathrm{Dis}}(0) \geqslant 0$ and $t_{\mathrm{Dis}}(0) \leqslant 0$ when $D_{\mathrm{Dis}}(0) \leqslant 0$. In both cases, we have $t_{\mathrm{Dis}}(0) \cdot D_{\mathrm{Dis}}(0) \geqslant 0$. This further implies $t \cdot D_{\mathrm{Dis}}(0) \geqslant 0$. Moreover, suppose now that $t_{\mathrm{Dis}}(\delta) \geqslant 0$ so that $t \in [0, t_{\mathrm{Dis}}(0))]$. Then, since D_{Dis} non-increasing, we have $D_{\mathrm{Dis}}(t) \geqslant D_{\mathrm{Dis}}(t_{\mathrm{Dis}}(0)) \geqslant 0$. Thus $t \cdot D_{\mathrm{Dis}}(t) \geqslant 0$ and $t \cdot D_{\mathrm{Dis}}(t)/|t| \geqslant 0$. The same claim holds when $t \leqslant 0$. Therefore, $\overline{\mathrm{Dis}}_t(f) = \frac{tD_{\mathrm{Dis}}(t)}{|t|}$ implies $\left|\overline{\mathrm{Dis}}_t(f)\right| \leqslant \frac{tD_{\mathrm{Dis}}(t)}{|t|}$. Consequently, $f_{\mathrm{Dis},t} \in \mathcal{F}_{t,=} \subset \mathcal{F}_{t,|\cdot|,\leqslant}$. Then,

$$\max_{f \in \mathcal{F}_{t,\leqslant}} \mathrm{Acc}(f) = \mathrm{Acc}(f_{\mathrm{Dis},t}) \leqslant \max_{f \in \mathcal{F}_{t,|\cdot|,\leqslant}} \mathrm{Acc}(f) \leqslant \max_{f \in \mathcal{F}_{t,\leqslant}} \mathrm{Acc}(f).$$

Thus, we can conclude that

$$f_t = \underset{f \in \mathcal{F}_{t,|\cdot|,\leqslant}}{\arg\max}\, \mathrm{Acc}(f) = \arg\min_{f \in \mathcal{F}} \left\{ R(f) : |\overline{\mathrm{Dis}}(f)| \leqslant \frac{t \cdot D_{\mathrm{Dis}}(t)}{|t|} \right\}$$

which shows the desired result since $t \cdot D_{\mathrm{Dis}}(t) \geqslant 0$ for $t \in [\min(0, t_{\mathrm{Dis}}(0)), \max(0, t_{\mathrm{Dis}}(0))]$. ■

Lemma 29 (Properties of the Tradeoff function) *For any $\delta \geqslant 0$, let $t_{\mathrm{Dis}}(\delta)$ and $T(\delta)$ be defined in* (10) *and* (15)*, respectively. Let for $a \in \{0,1\}$, $\eta_a(X)$ and $w_{\mathrm{Dis}}(X,a)$ be continuous random variables when $X \sim \mathbb{P}_{X|A=a}$, for $a \in \{0,1\}$. Then, for $\min(0, D_{\mathrm{Dis}}(0)) \leqslant \delta_1 < \delta_2 \leqslant \max(0, D_{\mathrm{Dis}}(0))$,*

$$t_{\mathrm{Dis}}(\delta_2)(\delta_2 - \delta_1) \leqslant T(\delta_1) - T(\delta_2) \leqslant t_{\mathrm{Dis}}(\delta_1)(\delta_2 - \delta_1). \tag{29}$$

Proof As $D_{\mathrm{Dis}}(t)$ is monotone non-increasing and continuous in $t \in \mathbb{R}$, it follows that $t_{\mathrm{Dis}}(\delta)$ is strictly decreasing in δ on $[0, |D_{\mathrm{Dis}}(0)|]$. Then, we have $t_{\mathrm{Dis}}(\delta_1) > t_{\mathrm{Dis}}(\delta_2)$. For $a \in \{0,1\}$, we define

$$\mathcal{T}_{a,+} = \left\{ x \in \mathcal{X},\, w_{\mathrm{Dis}}(x,a) > 0,\, t_{\mathrm{Dis}}(\delta_2) < \frac{2\eta_a(x) - 1}{w_{\mathrm{Dis}}(x,a)} \leqslant t_{\mathrm{Dis}}(\delta_1) \right\},$$

and

$$\mathcal{T}_{a,-} = \left\{ x \in \mathcal{X},\, w_{\mathrm{Dis}}(x,a) < 0,\, t_{\mathrm{Dis}}(\delta_2) \leqslant \frac{2\eta_a(x) - 1}{w_{\mathrm{Dis}}(x,a)} < t_{\mathrm{Dis}}(\delta_1) \right\}.$$

By definition,

$$\begin{aligned}
&f_{t_{\mathrm{Dis}}(\delta_1)}(x,a) - f_{t_{\mathrm{Dis}}(\delta_2)}(x,a)\\
&= I\left(\eta_a(x) > \frac{1}{2} + \frac{t_{\mathrm{Dis}}(\delta_1)}{2} w_{\mathrm{Dis}}(x,a)\right) - I\left(\eta_a(x) > \frac{1}{2} + \frac{t_{\mathrm{Dis}}(\delta_2)}{2} w_{\mathrm{Dis}}(x,a)\right)\\
&= \begin{cases} -I\left(t_{\mathrm{Dis}}(\delta_2) < \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t_{\mathrm{Dis}}(\delta_1)\right), & w_{\mathrm{Dis}}(x,a) > 0;\\ I\left(t_{\mathrm{Dis}}(\delta_2) \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_{\mathrm{Dis}}(\delta_1)\right), & w_{\mathrm{Dis}}(x,a) < 0;\\ 0, & w_{\mathrm{Dis}}(x,a) = 0, \end{cases}
\end{aligned}$$

In addition, we have, on $\mathcal{T}_{a,+}$,

$$t_{\mathrm{Dis}}(\delta_2)\cdot w_{\mathrm{Dis}}(x,a) \leqslant 2\eta_a(x) - 1 \leqslant t_{\mathrm{Dis}}(\delta_1)\cdot w_{\mathrm{Dis}}(x,a),$$

and on $\mathcal{T}_{a,-}$,

$$t_{\mathrm{Dis}}(\delta_1)\cdot w_{\mathrm{Dis}}(x,a) \leqslant 2\eta_a(x) - 1 \leqslant t_{\mathrm{Dis}}(\delta_2)\cdot w_{\mathrm{Dis}}(x,a).$$

Then, by (27) from Lemma 27,

$$\begin{aligned}
&T(\delta_1) - T(\delta_2) = R\left(f_{t_{\mathrm{Dis}}(\delta_1)}\right) - R\left(f_{t_{\mathrm{Dis}}(\delta_2)}\right)\\
&= \sum_{a\in\mathcal{A}} p_a \int_{\mathcal{X}} (1-2\eta_a(x))\left(f_{t_{\mathrm{Dis}}(\delta_1)}(x,a) - f_{t_{\mathrm{Dis}}(\delta_2)}(x,a)\right) d\mathbb{P}_{X|A=a}(x)\\
&= \sum_{a\in\mathcal{A}} p_a \int_{\{w_{\mathrm{Dis}}(x,a)>0\}} (2\eta_a(x)-1)\, I\left(t_{\mathrm{Dis}}(\delta_2) \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_{\mathrm{Dis}}(\delta_1)\right) d\mathbb{P}_{X|A=a}(x)\\
&\quad - \sum_{a\in\mathcal{A}} p_a \int_{\{w_{\mathrm{Dis}}(x,a)<0\}} (2\eta_a(x)-1)\, I\left(t_{\mathrm{Dis}}(\delta_2) \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_{\mathrm{Dis}}(\delta_1)\right) d\mathbb{P}_{X|A=a}(x)\\
&= \sum_{a\in\mathcal{A}} p_a \int_{\mathcal{T}_{a,+}} (2\eta_a(x)-1)\, d\mathbb{P}_{X|A=a}(x) - \sum_{a\in\mathcal{A}} p_a \int_{\mathcal{T}_{a,-}} (2\eta_a(x)-1)\, d\mathbb{P}_{X|A=a}(x)\\
&\leqslant t_{\mathrm{Dis}}(\delta_1) \sum_{a\in\mathcal{A}} p_a \int_{\mathcal{T}_{a,+}} w_{\mathrm{Dis}}(x,a) d\mathbb{P}_{X|A=a}(x) - t_{\mathrm{Dis}}(\delta_1) \sum_{a\in\mathcal{A}} p_a \int_{\mathcal{T}_{a,-}} w_{\mathrm{Dis}}(x,a) d\mathbb{P}_{X|A=a}(x)\\
&= -t_{\mathrm{Dis}}(\delta_1) \sum_{a\in\mathcal{A}} p_a \int_{\mathcal{X}\cap\{w_{\mathrm{Dis}}(x,a)>0\}} w_{\mathrm{Dis}}(x,a)\cdot\left(f_{t_{\mathrm{Dis}}(\delta_1)}(x,a) - f_{t_{\mathrm{Dis}}(\delta_2)}(x,a)\right) d\mathbb{P}_{X|A=a}(x)\\
&\quad - t_{\mathrm{Dis}}(\delta_1) \sum_{a\in\mathcal{A}} p_a \int_{\mathcal{X}\cap\{w_{\mathrm{Dis}}(x,a)<0\}} w_{\mathrm{Dis}}(x,a)\cdot\left(f_{t_{\mathrm{Dis}}(\delta_1)}(x,a) - f_{t_{\mathrm{Dis}}(\delta_2)}(x,a)\right) d\mathbb{P}_{X|A=a}(x)\\
&= -t_{\mathrm{Dis}}(\delta_1) \sum_{a\in\mathcal{A}} p_a \int_{\mathcal{X}} w_{\mathrm{Dis}}(x,a)\cdot\left(f_{t_{\mathrm{Dis}}(\delta_1)}(x,a) - f_{t_{\mathrm{Dis}}(\delta_2)}(x,a)\right) d\mathbb{P}_{X|A=a}(x)\\
&= -t_{\mathrm{Dis}}(\delta_1)\left(\mathrm{Dis}\left(f_{t_{\mathrm{Dis}}(\delta_1)}\right) - \mathrm{Dis}\left(f_{t_{\mathrm{Dis}}(\delta_2)}\right)\right) = t_{\mathrm{Dis}}(\delta_1)(\delta_2 - \delta_1).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&T(\delta_1) - T(\delta_2)\\
&= \sum_{a\in\mathcal{A}} p_a \int_{\mathcal{T}_{a,+}} (2\eta_a(x)-1)\, d\mathbb{P}_{X|A=a}(x) - \sum_{a\in\mathcal{A}} p_a \int_{\mathcal{T}_{a,-}} (2\eta_a(x)-1)\, d\mathbb{P}_{X|A=a}(x)
\end{aligned}$$

$$
\begin{aligned}
&\geqslant t_{\mathrm{Dis}}(\delta_2)\sum_{a\in\mathcal{A}} p_a \int_{\mathcal{T}_{a,+}} w_{\mathrm{Dis}}(x,a) d\mathbb{P}_{X|A=a}(x) - t_{\mathrm{Dis}}(\delta_2)\sum_{a\in\mathcal{A}} p_a \int_{\mathcal{T}_{a,-}} w_{\mathrm{Dis}}(x,a) d\mathbb{P}_{X|A=a}(x)\\
&= -t_{\mathrm{Dis}}(\delta_2)\sum_{a\in\mathcal{A}} p_a \int_{\mathcal{X}\cap\{w_{\mathrm{Dis}}(x,a)>0\}} w_{\mathrm{Dis}}(x,a)\cdot\left(f_{t_{\mathrm{Dis}}(\delta_1)}(x,a) - f_{t_{\mathrm{Dis}}(\delta_2)}(x,a)\right) d\mathbb{P}_{X|A=a}(x)\\
&\quad - t_{\mathrm{Dis}}(\delta_2)\sum_{a\in\mathcal{A}} p_a \int_{\mathcal{X}\cap\{w_{\mathrm{Dis}}(x,a)<0\}} w_{\mathrm{Dis}}(x,a)\cdot\left(f_{t_{\mathrm{Dis}}(\delta_1)}(x,a) - f_{t_{\mathrm{Dis}}(\delta_2)}(x,a)\right) d\mathbb{P}_{X|A=a}(x)\\
&= -t_{\mathrm{Dis}}(\delta_2)\sum_{a\in\mathcal{A}} p_a \int_{\mathcal{X}} w_{\mathrm{Dis}}(x,a)\cdot\left(f_{t_{\mathrm{Dis}}(\delta_1)}(x,a) - f_{t_{\mathrm{Dis}}(\delta_2)}(x,a)\right) d\mathbb{P}_{X|A=a}(x)\\
&= -t_{\mathrm{Dis}}(\delta_2)\left(\mathrm{Dis}\left(f^\star_{\delta_1}\right) - \mathrm{Dis}\left(f^\star_{\delta_2}\right)\right) = t_{\mathrm{Dis}}(\delta_2)(\delta_2-\delta_1).
\end{aligned}
$$

We conclude (29).

■

A.2 Proofs of Results from Section 3

A.2.1 Proof of Proposition 12

Let f be any classifier. Its DD can be expressed as

$$
\begin{aligned}
\mathrm{DD}(f) &= \mathbb{P}(\widehat{Y}_f = 1|A=1) - \mathbb{P}(\widehat{Y}_f = 1|A=0)\\
&= \int_{\mathcal{X}} f(x,1) d\mathbb{P}_{X|A=1}(x) - \int_{\mathcal{X}} f(x,0) d\mathbb{P}_{X|A=0}(x)\\
&= \sum_{a\in\{0,1\}} p_a \left(\int_{\mathcal{X}}\left[\frac{I(a=1)}{p_1} f(x,a) - \frac{I(a=0)}{p_0} f(x,a)\right] d\mathbb{P}_{X|A=a}(x)\right)\\
&= \int_{\mathcal{A}}\int_{\mathcal{X}} f(x,a)\left(\frac{2a-1}{p_a}\right) d\mathbb{P}_{X|A=a}(x) d\mathbb{P}_A(a)\\
&= \int_{\mathcal{A}}\int_{\mathcal{X}} f(x,a)(2a-1)\left(\frac{1}{p_a}\right) d\mathbb{P}_{X,A}(x,a).
\end{aligned} \tag{30}
$$

By Bayes' theorem, we can express the regression function η_a in terms of the conditional distributions of X given A, Y and A, as follows

$$
\eta_a(x) = \mathbb{P}\left(Y=1|X=x, A=a\right) = \frac{\mathbb{P}(A=a, Y=1) d\mathbb{P}_{X|A=a,Y=1}(x)}{\mathbb{P}(A=a) d\mathbb{P}_{X|A=a}(x)} = \frac{p_{a,1} d\mathbb{P}_{X|A=a,Y=1}(x)}{p_a d\mathbb{P}_{X|A=a}(x)}.
$$

Next, DO can be expressed as

$$
\begin{aligned}
\mathrm{DO}(f) &= \mathbb{P}(\widehat{Y}_f = 1|A=1, Y=1) - \mathbb{P}(\widehat{Y}_f = 1|A=0, =1)\\
&= \int_{\mathcal{X}} f(x,1) d\mathbb{P}_{X|A=1,Y=1}(x) - \int_{\mathcal{X}} f(x,0) d\mathbb{P}_{X|A=0,Y=1}(x)\\
&= \sum_{a\in\{0,1\}} p_a \left(\int_{\mathcal{X}}\left[\frac{I(a=1)}{p_1} f(x,a) - \frac{I(a=0)}{p_0} f(x,a)\right] d\mathbb{P}_{X|A=a,Y=1}(x)\right)
\end{aligned}
$$

$$= \int_{\mathcal{A}} \int_{\mathcal{X}} f(x,a) \left(\frac{2a-1}{p_a} \right) \frac{d\mathbb{P}_{X|A=a,Y=1}(x)}{d\mathbb{P}_{X|A=a}} d\mathbb{P}_{X|A=a}(x) d\mathbb{P}_A(a)$$
$$= \int_{\mathcal{X}\times\mathcal{A}} f(x,a)(2a-1) \left(\frac{\eta_a(x)}{p_{a,1}} \right) d\mathbb{P}_{X,A}(x,a). \tag{31}$$

Similarly to above

$$1-\eta_a(x) = \mathbb{P}\left(Y=0|X=x, A=a\right) = \frac{\mathbb{P}(A=a, Y=0) d\mathbb{P}_{X|A=a,Y=0}(x)}{\mathbb{P}(A=a) d\mathbb{P}_{X|A=a}(x)}$$
$$= \frac{p_{a,0} d\mathbb{P}_{X|A=a,Y=0}(x)}{p_a d\mathbb{P}_{X|A=a}(x)}.$$

Thus, PD can be expressed as

$$\begin{aligned}
\mathrm{PD}(f) &= \mathbb{P}(\widehat{Y}_f = 1|A=1, Y=0) - \mathbb{P}(\widehat{Y}_f = 1|A=0, Y=0) \\
&= \int_{\mathcal{X}} f(x,1) d\mathbb{P}_{X|A=1,Y=0}(x) - \int_{\mathcal{X}} f(x,0) d\mathbb{P}_{X|A=0,Y=0}(x) \\
&= \sum_{a\in\{0,1\}} p_a \left(\int_{\mathcal{X}} \left[\frac{I(a=1)}{p_1} f(x,a) - \frac{I(a=0)}{p_0} f(x,a) \right] d\mathbb{P}_{X|A=a,Y=0}(x) \right) \\
&= \int_{\mathcal{A}} \int_{\mathcal{X}} f(x,a) \left(\frac{2a-1}{p_a} \right) \frac{d\mathbb{P}_{X|A=a,Y=0}(x)}{d\mathbb{P}_{X|A=a}} d\mathbb{P}_{X|A=a}(x) d\mathbb{P}_A(a) \\
&= \int_{\mathcal{X}\times\mathcal{A}} f(x,a)(2a-1) \left(\frac{1-\eta_a(x)}{p_{a,0}} \right) d\mathbb{P}_{X,A}(x,a).
\end{aligned} \tag{32}$$

Thus, Proposition 12 follows.

A.3 Proofs of Results from Section 4

A.3.1 Proof of Proposition 13

For (1), since, for $a \in \{0,1\}$, both $\eta_a(X)$ and $w_{\mathrm{Dis}}(X,a)$ are continuous random variable given $A=a$, we have that for $a \in \{0,1\}$, $t \mapsto \mathbb{P}_{X|A=a}\left(\eta_a(X) > 1/2 + t w_{\mathrm{Dis}}(X,a)/2\right)$ is a continuous function. Thus, $t \mapsto D_{\mathrm{Dis}}(t)$ is continuous. Now, define, for $a \in \{0,1\}$,

$$\begin{aligned}
G_{a,+} &= \{x \in \mathcal{X}, w_{\mathrm{Dis}}(x,a) > 0\}; \\
G_{a,0} &= \{x \in \mathcal{X}, w_{\mathrm{Dis}}(x,a) = 0\}; \\
G_{a,-} &= \{x \in \mathcal{X}, w_{\mathrm{Dis}}(x,a) < 0\}.
\end{aligned}$$

Letting $t_1 < t_2$, we have for $a \in \{0,1\}$ and $x \in \mathcal{X}$,

$$f_{\mathrm{Dis},t_1}(x,a) - f_{\mathrm{Dis},t_2}(x,a) = \begin{cases} I\left(t_1 < \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t_2\right), & x \in G_{a,+}; \\ -I\left(t_1 \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_2\right), & x \in G_{a,-}; \\ 0, & \text{otherwise}. \end{cases} \tag{33}$$

It thus follows that,

$$
\begin{aligned}
& D_{\mathrm{Dis}}(t_1) - D_{\mathrm{Dis}}(t_2) = \mathrm{Dis}(f_{\mathrm{Dis},t_1}) - \mathrm{Dis}(f_{\mathrm{Dis},t_2}) \\
& = \int_{\mathcal{A}} \int_{\mathcal{X}} \left[f_{\mathrm{Dis},t_1}(x,a) - f_{\mathrm{Dis},t_2}(x,a)\right] w_{\mathrm{Dis}}(x,a) d\mathbb{P}_{X,A}(x,a) \\
& = \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left[f_{\mathrm{Dis},t_1}(x,a) - f_{\mathrm{Dis},t_2}(x,a)\right] w_{\mathrm{Dis}}(x,a) d\mathbb{P}_{X|A=a}(x) \\
& = \sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} I\left(t_1 < \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t_2\right) w_{\mathrm{Dis}}(x,a) d\mathbb{P}_{X|A=a}(x) \\
& \quad - \sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} I\left(t_1 \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_2\right) w_{\mathrm{Dis}}(x,a) d\mathbb{P}_{X|A=a}(x) \\
& \geqslant 0.
\end{aligned}
$$

The last inequality holds since the indicator function is non-negative, and $w_{\mathrm{Dis}}(x,a)$ is positive on $G_{a,+}$ and negative on $G_{a,-}$. Thus, $t \mapsto D_{\mathrm{Dis}}(t)$ is a monotone non-increasing function.

For (2), when $t_1 < t_2 < 0$, it follows that on $G_{a,+}$,

$$
(1-2\eta_a(x)) I\left(t_1 < \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t_2\right) \geqslant -t_2 w_{\mathrm{Dis}}(x,a) I\left(t_1 < \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t_2\right) \geqslant 0,
$$

and, on $G_{a,-}$,

$$
(2\eta_a(x)-1) I\left(t_1 \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_2\right) \geqslant t_2 w_{\mathrm{Dis}}(x,a) I\left(t_1 \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_2\right) \geqslant 0.
$$

Then, by Lemma 27 and (33),

$$
\begin{aligned}
& R(f_{\mathrm{Dis},t_1}) - R(f_{\mathrm{Dis},t_2}) = \int_{\mathcal{A}} \int_{\mathcal{X}} (1-2\eta_a(x)) \left[f_{\mathrm{Dis},t_1}(x,a) - f_{\mathrm{Dis},t_2}(x,a)\right] d\mathbb{P}_{X,A}(x,a) \\
& = \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} (1-2\eta_a(x)) \left[f_{\mathrm{Dis},t_1}(x,a) - f_{\mathrm{Dis},t_2}(x,a)\right] d\mathbb{P}_{X|A=a}(x) \\
& = \sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} (1-2\eta_a(x)) I\left(t_1 < \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t_2\right) d\mathbb{P}_{X|A=a}(x) \\
& \quad + \sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} (2\eta_a(x)-1) I\left(t_1 \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_2\right) d\mathbb{P}_{X|A=a}(x) \geqslant 0.
\end{aligned}
$$

Thus, $t \mapsto R(f_{\mathrm{Dis},t})$ is monotone non-increasing on $(-\infty, 0)$.

On the other hand, when $0 \leqslant t_1 < t_2$, we have on $G_{a,+}$ that

$$
(1-2\eta_a(x)) I\left(t_1 < \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t_2\right) \leqslant -t_1 w_{\mathrm{Dis}}(x,a) I\left(t_1 < \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t_2\right) \leqslant 0,
$$

and, on $G_{a,-}$ that

$$(2\eta_a(x)-1)I\left(t_1 \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_2\right) \leqslant t_1 w_{\mathrm{Dis}}(x,a) I\left(t_1 \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_2\right) \leqslant 0.$$

Then, by Lemma 27 and (33),

$$\begin{aligned}
R(f_{\mathrm{Dis},t_1}) - R(f_{\mathrm{Dis},t_2}) &= \int_{\mathcal{A}}\int_{\mathcal{X}} (1-2\eta_a(x))\,[f_{\mathrm{Dis},t_1}(x,a) - f_{\mathrm{Dis},t_2}(x,a)]\, d\mathbb{P}_{X,A}(x,a) \\
&= \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} (1-2\eta_a(x))\,[f_{\mathrm{Dis},t_1}(x,a) - f_{\mathrm{Dis},t_2}(x,a)]\, d\mathbb{P}_{X|A=a}(x) \\
&= \sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} (1-2\eta_a(x)) I\left(t_1 < \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t_2\right) d\mathbb{P}_{X|A=a}(x) \\
&\quad + \sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} (2\eta_a(x)-1) I\left(t_1 \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_2\right) d\mathbb{P}_{X|A=a}(x) \leqslant 0.
\end{aligned}$$

Thus, $t \mapsto R(f_{\mathrm{Dis},t})$ is monotone non-decreasing on $[0,\infty)$. This shows that $t \mapsto R(f_{\mathrm{Dis},t})$ is non-increasing in $|t|$.

A.3.2 Proof of Theorem 14

We analyze the following three cases separately: (1) $|\mathrm{Dis}(0)| \leqslant \delta$, (2) $\mathrm{Dis}(0) > \delta$ and (3) $\mathrm{Dis}(0) < -\delta$. Since the proof for case (3) is analogous to case (2), we omit the discussion of case (3).

Case (1): $|\mathrm{Dis}(0)| \leqslant \delta$. In this case, the unconstrained Bayes-optimal classifier satisfies the fairness constraint. As a result, we have $t_{\mathrm{Dis}}(\delta) = 0$ and a δ-fair Bayes-optimal classifier is given by $f^{\star}_{\mathrm{Dis},\delta}(x,a) = I(\eta_a(x) > 1/2)$. for all x, a.

Case (2): $\mathrm{Dis}(0) > \delta$. When, for $a \in \{0,1\}$, $\eta_a(X)$ has probability density function on $\mathcal{X}$, $D_{\mathrm{Dis}}(t)$ is a continuous non-increasing function on $\mathbb{R}$, and thus we have $D_{\mathrm{Dis}}(t_{\mathrm{Dis}}(\delta)) = \delta$. Moreover, $\mathrm{Dis}(0) > \delta$ indicates $t_{\mathrm{Dis}}(\delta) > 0$. Then, by Lemma 28, with $t = t_{\mathrm{Dis}}(\delta)$,

$$f_{\mathrm{Dis},t_{\mathrm{Dis}}(\delta)} = \arg\min_{f\in\mathcal{F}} \left\{ R(f) : |\mathrm{Dis}(f)| \leqslant \frac{t \cdot D_{\mathrm{Dis}}(t_{\mathrm{Dis}}(\delta))}{|t|} \right\} = \arg\min_{f\in\mathcal{F}} \{R(f) : |\mathrm{Dis}(f)| \leqslant \delta\}.$$

This finishes the proof.

A.3.3 Proof of Proposition 16

Without loss of generality, we assume $D_{\mathrm{Dis}}(0) \geqslant 0$. In this case, we have $\underline{t}_0 = 0$ and $\overline{t}_0 = t_{\mathrm{Dis}}(0)$. By Lemma 28, we have

$$\{f_{\mathrm{Dis},t} : t \in [0, t_{\mathrm{Dis}}(0)]\} \subset \mathrm{FPF}\}.$$

On the other hand, since $f_{\mathrm{Dis},0} \in \arg\min_{f\in\mathcal{F}}\{R(f)\}$, and by Lemma 28, $f_{\mathrm{Dis},t_{\mathrm{Dis}}(0)} \in \arg\min_{f\in\mathcal{F}}\{R(f) : \mathrm{Dis}(f) \leqslant 0\}$, we have, for f_1 with $\mathrm{Dis}(f_1) < 0$ and f_2 with $\mathrm{Dis}(f_2) > D_{\mathrm{Dis}}(0)$,

$$R(f_{\mathrm{Dis},t_{\mathrm{Dis}}(0)}) \leqslant R(f_1),\ |\mathrm{Dis}(f_{\mathrm{Dis},t_{\mathrm{Dis}}(0)})| = 0 < |\mathrm{Dis}(f_1)|,$$

and

$$R(f_{\mathrm{Dis},0}) \leqslant R(f_2),\ |\mathrm{Dis}(f_{\mathrm{Dis},0})| = D_{\mathrm{Dis}}(0) < |\mathrm{Dis}(f_2)|.$$

Thus, for any $f_{\mathrm{FPF}} \in \mathrm{FPF}$, we have

$$0 \leqslant \mathrm{Dis}(f_{\mathrm{FPF}}) \leqslant \mathrm{Dis}(0).$$

Moreover, by the definition of the fair Pareto frontier, we have

$$R(f_{\mathrm{FPF}}) = \min_{f\in\mathcal{F}}\{R(f) : \mathrm{Dis}(f) \leqslant \mathrm{Dis}(f_{\mathrm{FPF}})\}.$$

Since $t \mapsto D_{\mathrm{Dis}}(t)$ is a continuous monotone non-increasing function on $[0, t_{\mathrm{Dis}}(0)]$ with $D_{\mathrm{Dis}}(t_{\mathrm{Dis}}(0)) = 0$, there exists a $t \in [0, t_{\mathrm{Dis}}(0)]$ such that $\mathrm{Dis}(f_{\mathrm{Dis},t}) = D_{\mathrm{Dis}}(t) = \mathrm{Dis}(f_{\mathrm{FPF}})$. By Lemma 28,

$$R(f_{\mathrm{Dis},t}) = \min_{f\in\mathcal{F}}\{R(f) : \mathrm{Dis}(f) = \mathrm{Dis}(f_{\mathrm{FPF}})\}.$$

In conclusion, there exists a $t \in [0, t_{\mathrm{Dis}}(0)]$ such that,

$$R(f_{\mathrm{Dis},t}) = R(f_{\mathrm{FPF}}) \quad \text{and} \quad \mathrm{Dis}(f_{\mathrm{Dis},t}) = \mathrm{Dis}(f_{\mathrm{FPF}}).$$

A.3.4 Proof of Theorem 18

Let $0 \leqslant \delta_1 < \delta_2 \leqslant |D_{\mathrm{Dis}}(0)|$ and $\lambda \in [0, 1]$, and denote $\delta_\lambda = \lambda\delta_1 + (1-\lambda)\delta_2$. As $t_{\mathrm{Dis}}(\delta)$ is strictly decreasing on $[0, |D_{\mathrm{Dis}}(0)|]$, we have $t_{\mathrm{Dis}}(\delta_1) > t_{\mathrm{Dis}}(\delta_\lambda) > t_{\mathrm{Dis}}(\delta_2)$. By Lemma 29,

$$\begin{aligned}
&T(\delta_\lambda) = \lambda T(\delta_\lambda) + (1-\lambda)T(\delta_\lambda)\\
&= \lambda T(\delta_1) + (1-\lambda)T(\delta_2) + \lambda\left(T(\delta_\lambda) - T(\delta_1)\right) + (1-\lambda)\left(T(\delta_\lambda) - T(\delta_2)\right)\\
&\leqslant \lambda T(\delta_1) + (1-\lambda)T(\delta_2) - \lambda t_{\mathrm{Dis}}(\delta_\lambda)(\delta_\lambda - \delta_1) + (1-\lambda)t_{\mathrm{Dis}}(\delta_\lambda)(\delta_2 - \delta_\lambda)\\
&= \lambda T(\delta_1) + (1-\lambda)T(\delta_2).
\end{aligned}$$

The proof is thus completed.

A.3.5 Proof of Proposition 19

By Bayes' theorem, we have for all x, y, a that

$$\mathbb{P}\left(A = a \mid X = x\right) = \frac{p_a d\mathbb{P}_{X|A=a}(x)}{d\mathbb{P}_X(x)},$$

and

$$\mathbb{P}(Y = y \mid A = a, X = x)\mathbb{P}(A = a \mid X = x) = \mathbb{P}(A = a, Y = y \mid X = x) = \frac{p_{a,y} d\mathbb{P}_{X|A=a,Y=y}(x)}{d\mathbb{P}_X(x)}.$$

It then follows that

$$\begin{aligned}
\frac{d\mathbb{P}_{X|A=1}(x)}{d\mathbb{P}_X(x)} &= \frac{\mathbb{P}(A=1 \mid X=x)}{\mathbb{P}(A=1)} = \frac{\eta^A(x)}{p_1};\\
\frac{d\mathbb{P}_{X|A=0}(x)}{d\mathbb{P}_X(x)} &= \frac{\mathbb{P}(A=0 \mid X=x)}{\mathbb{P}(A=0)} = \frac{1-\eta^A(x)}{p_0};
\end{aligned}$$

$$
\begin{aligned}
\frac{d\mathbb{P}_{X|A=1,Y=1}(x)}{d\mathbb{P}_X(x)} &= \frac{\mathbb{P}(Y=1 \mid A=1, X=x)\mathbb{P}(A=1|X=x)}{p_{1,1}} = \frac{\eta^Y_{A=1}(x)\eta^A(x)}{p_{1,1}};\\
\frac{d\mathbb{P}_{X|A=1,Y=0}(x)}{d\mathbb{P}_X(x)} &= \frac{\mathbb{P}(Y=0 \mid A=1, X=x)\mathbb{P}(A=1|X=x)}{p_{1,0}} = \frac{(1-\eta^Y_{A=1}(x))\eta^A(x)}{p_{1,0}};\\
\frac{d\mathbb{P}_{X|A=0,Y=1}(x)}{d\mathbb{P}_X(x)} &= \frac{\mathbb{P}(Y=1 \mid A=0, X=x)\mathbb{P}(A=0|X=x)}{p_{0,1}} = \frac{\eta^Y_{A=0}(x)(1-\eta^A(x)}{p_{0,1}};\\
\frac{d\mathbb{P}_{X|A=0,Y=0}(x)}{d\mathbb{P}_X(x)} &= \frac{\mathbb{P}(Y=0 \mid A=0, X=x)\mathbb{P}(A=0|X=x)}{p_{0,0}} = \frac{(1-\eta^Y_{A=0}(x))(1-\eta^A(x))}{p_{0,0}}.
\end{aligned}
$$

Let f be any classifier. Its DD, DO and PD can be expressed in turn as

$$
\begin{aligned}
\mathrm{DD}(f) &= \mathbb{P}(\widehat{Y}_f = 1 \mid A=1) - \mathbb{P}(\widehat{Y}_f = 1 \mid A=0)\\
&= \int_{\mathcal{X}} f(x) d\mathbb{P}_{X|A=1}(x) - \int_{\mathcal{X}} f(x) d\mathbb{P}_{X|A=0}(x)\\
&= \int_{\mathcal{X}} \left[f(x) \left(\frac{d\mathbb{P}_{X|A=1}(x)}{d\mathbb{P}_X(x)} - \frac{d\mathbb{P}_{X|A=0}(x)}{d\mathbb{P}_X(x)} \right) \right] d\mathbb{P}_X(x)\\
&= \int_{\mathcal{X}} f(x,a) \left(\frac{\eta^A(x)}{p_1} - \frac{1-\eta^A(x)}{p_0} \right) d\mathbb{P}_X(x);\\
\mathrm{DO}(f) &= \mathbb{P}(\widehat{Y}_f = 1 \mid A=1, Y=1) - \mathbb{P}(\widehat{Y}_f = 1 \mid A=0, Y=1)\\
&= \int_{\mathcal{X}} f(x) d\mathbb{P}_{X|A=1,Y=1}(x) - \int_{\mathcal{X}} f(x) d\mathbb{P}_{X|A=0,Y=1}(x)\\
&= \int_{\mathcal{X}} \left[f(x) \left(\frac{d\mathbb{P}_{X|A=1,Y=1}(x)}{d\mathbb{P}_X(x)} - \frac{d\mathbb{P}_{X|A=0,Y=1}(x)}{d\mathbb{P}_X(x)} \right) \right] d\mathbb{P}_X(x)\\
&= \int_{\mathcal{X}} f(x,a) \left(\frac{\eta^Y_{A=1}(x)\eta^A(x)}{p_{1,1}} - \frac{\eta^Y_{A=0}(x)(1-\eta^A(x))}{p_{0,1}} \right) d\mathbb{P}_X(x);\\
\mathrm{PD}(f) &= \mathbb{P}(\widehat{Y}_f = 1 \mid A=1, Y=0) - \mathbb{P}(\widehat{Y}_f = 1 \mid A=0, Y=0)\\
&= \int_{\mathcal{X}} f(x) d\mathbb{P}_{X|A=1,Y=0}(x) - \int_{\mathcal{X}} f(x) d\mathbb{P}_{X|A=0,Y=0}(x)\\
&= \int_{\mathcal{X}} \left[f(x) \left(\frac{d\mathbb{P}_{X|A=1,Y=0}(x)}{d\mathbb{P}_X(x)} - \frac{d\mathbb{P}_{X|A=0,Y=0}(x)}{d\mathbb{P}_X(x)} \right) \right] d\mathbb{P}_X(x)\\
&= \int_{\mathcal{X}} f(x,a) \left(\frac{(1-\eta^Y_{A=1}(x))\eta^A(x)}{p_{1,0}} - \frac{(1-\eta^Y_{A=0}(x))(1-\eta^A(x))}{p_{0,0}} \right) d\mathbb{P}_X(x).
\end{aligned}
$$

Thus, Proposition 19 follows.

A.4 Proofs of Results from Section 5

A.4.1 Proof of Theorem 24

By Lemma 28, we only need to prove that, for $t \in [\underline{t}_0, \overline{t}_0]$, the unconstrained Bayes-optimal classifier for $\widetilde{\mathbb{P}}^t$ is the same as $f_{\mathrm{Dis},t}$. By Bayes' theorem, for all x, a,

$$
\widetilde{\eta}^t_a(x) = \frac{\widetilde{p}^t_{a,1} d\widetilde{\mathbb{P}}^t_{\widetilde{X}|\widetilde{A}=a,\widetilde{Y}=1}(x)}{\widetilde{p}^t_{a1} d\widetilde{\mathbb{P}}_{\widetilde{X}|\widetilde{A}=a,\widetilde{Y}=1}(x) + \widetilde{p}^t_{a0} d\widetilde{\mathbb{P}}_{\widetilde{X}|\widetilde{A}=a,\widetilde{Y}=0}(x)}.
$$

Then, the unconstrained Bayes-optimal classifier for $\widetilde{\mathbb{P}}^t$ is

$$\begin{aligned}
f_t^{\mathrm{FUDS}}(x,a) &= I\left(\widetilde{\eta}_a^t(x) > \frac{1}{2}\right) = I\left(\frac{\widetilde{p}_{a,1}^t d\widetilde{\mathbb{P}}^t_{\widetilde{X}|\widetilde{A}=a,\widetilde{Y}=1}(x)}{\widetilde{p}_{a1}^t d\widetilde{\mathbb{P}}^t_{\widetilde{X}|\widetilde{A}=a,\widetilde{Y}=1}(x) + \widetilde{p}_{a0}^t d\widetilde{\mathbb{P}}^t_{\widetilde{X}|\widetilde{A}=a,\widetilde{Y}=0}(x)} > \frac{1}{2}\right) \\
&= I\left(\frac{c_a(1-H_{\mathrm{Dis},a}(t))p_{a,y}d\mathbb{P}_{X|A=a,Y=1}(x)}{c_a(1-H_{\mathrm{Dis},a}(t))p_{a,y}d\mathbb{P}_{X|A=a,Y=1}(x) + c_a H_{\mathrm{Dis},a}(t)p_{a,y}d\mathbb{P}_{X|A=a,Y=0}(x)} > \frac{1}{2}\right) \\
&= I\left((1-H_{\mathrm{Dis},a}(t))p_{a,1}d\mathbb{P}_{X|A=a,Y=1}(x) > H_{\mathrm{Dis},a}(t)p_{a,0}d\mathbb{P}_{X|A=a,Y=0}(x)\right) \\
&= I\left(p_{a,1}d\mathbb{P}_{X|A=a,Y=1}(x) > H_{\mathrm{Dis},a}(t)\left[p_{a,1}d\mathbb{P}_{X|A=a,Y=1}(x) + p_{a,0}d\mathbb{P}_{X|A=a,Y=0}(x)\right)\right] \\
&= I\left(\eta_a(x) > H_{\mathrm{Dis},a}(t)\right) = I\left(\eta_a(x) > \frac{1}{2} + \frac{t}{2}w_{\mathrm{Dis}}(x,a)\right) = f_{\mathrm{Dis},t}(x,a).
\end{aligned}$$

This finishes the proof.

A.4.2 Proof of Theorem 26

Again, we only need to prove that for $t \in [\underline{t}_0, \bar{t}_0]$, $f_t^{\mathrm{FCSC}} = f_{\mathrm{Dis},t}$. By definition,

$$\begin{aligned}
&R_t^{\mathrm{FCSC}}(f) \\
&= \sum_{a\in\{0,1\}} \left[H_{\mathrm{Dis},a}(t)\cdot\mathbb{P}(\widehat{Y}_f = 1, Y=0, A=a) + (1-H_{\mathrm{Dis},a}(t))\cdot\mathbb{P}(\widehat{Y}_f = 0, Y=1, A=a)\right] \\
&= \sum_{a\in\{0,1\}} p_a\left[H_{\mathrm{Dis},a}(t)\cdot\mathbb{P}(\widehat{Y}_f = 1, Y=0|A=a) + (1-H_{\mathrm{Dis},a}(t))\cdot\mathbb{P}(\widehat{Y}_f = 0, Y=1|A=a)\right] \\
&= \sum_{a\in\{0,1\}} p_a H_{\mathrm{Dis},a}(t)\cdot\int_{\mathcal{X}}\mathbb{P}(\widehat{Y}_f = 1, Y=0|A=a, X=x)d\mathbb{P}_{X|A=a}(x) \\
&\quad + \sum_{a\in\{0,1\}} p_a(1-H_{\mathrm{Dis},a}(t))\cdot\int_{\mathcal{X}}\mathbb{P}(\widehat{Y}_f = 0, Y=1|X=x, A=a)d\mathbb{P}_{X|A=a}(x) \\
&= \sum_{a\in\{0,1\}} p_a\int_{\mathcal{X}}\left[H_{\mathrm{Dis},a}(t)f(x,a)(1-\eta_a(x)) + (1-H_{\mathrm{Dis},a}(t))(1-f(x,a))\eta_a(x)\right]d\mathbb{P}_{X|A=a}(x) \\
&= \sum_{a\in\{0,1\}} p_a\int_{\mathcal{X}}\left[H_{\mathrm{Dis},a}(t)f(x,a) + \eta_a(x) - H_{\mathrm{Dis},a}(t)\eta_a(x) - \eta_a(x)f(x,a)\right]d\mathbb{P}_{X|A=a}(x) \\
&= \sum_{a\in\{0,1\}} p_a\int_{\mathcal{X}}\left(H_{\mathrm{Dis},a}(t) - \eta_a(x)\right)f(x,a)d\mathbb{P}_{X|A=a}(x) \\
&\quad + \sum_{a\in\{0,1\}} p_a\int_{\mathcal{X}}\left(1-H_{\mathrm{Dis},a}(t)\right)\eta_a(x)d\mathbb{P}_{X|A=a}(x).
\end{aligned}$$

Note that the second term does not depend on f and that the first term is minimized by taking $f(x,a)=1$ if $\eta_a(x) > H_{\mathrm{Dis},a}(t)$ and $f(x,a)=0$ if $\eta_a(x) \leqslant H_{\mathrm{Dis},a}(t)$. We can thus conclude that for all x, a,

$$f_t^{\mathrm{FCSC}}(x,a) = I\left(\eta_a(x) > H_{\mathrm{Dis},a}(t)\right) = f_{\mathrm{Dis},t}(x,a).$$

This finishes the proof.

Appendix B. Extensions

In this section, we discuss some interesting extensions of our theoretical and methodological framework, including (1) fair Bayes-optimal classifier with equalized odds; (2) fair Bayes-optimal classifier with multi-class protected attribute; (3) fair Bayes-optimal classifier with no distributional assumptions; (4) FUDS and FCSC algorithms with linear disparities (protected attribute A is excluded from predictive attribute.)

B.1 Fair Bayes-optimal Classifier with Equalized Odds

In this section, we provide the detailed characterization of the fair Bayes-optimal classifier under equalized odds (EO), as defined in Theorem 21. Recall that, for any $\delta > 0$, a δ-fair Bayes-optimal classifier is defined as

$$f^{\star}_{\mathrm{DEO},\delta} = \arg\min \left\{ R(f) : \max\{|\mathrm{DO}(f)|, |\mathrm{PD}(f)|\} \leqslant \delta \right\},$$

with DO and PD defined in (1) and written more explicitly as

$$\mathrm{DO}(f) = \int_{\mathcal{X}\times\mathcal{A}} f(x,a)\frac{(2a-1)\eta_a(x)}{p_{a,1}} d\mathbb{P}_{X,A}(x,a); \tag{34}$$

$$\mathrm{PD}(f) = \int_{\mathcal{X}\times\mathcal{A}} f(x,a)\frac{(2a-1)(1-\eta_a(x))}{p_{a,0}} d\mathbb{P}_{X,A}(x,a). \tag{35}$$

The fairness constraint $\max\{|\mathrm{DO}(f)|, |\mathrm{PD}(f)|\} \leqslant \delta$ consists of four linear inequalities on f:

$$\mathrm{DO}(f) \leqslant \delta, \quad -\mathrm{DO}(f) \leqslant \delta, \quad \mathrm{PD}(f) \leqslant \delta \ \text{ and } \ -\mathrm{PD}(f) \leqslant \delta.$$

or

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathcal{A}} f(x,a)\frac{(2a-1)\eta_a(x)}{p_{a,1}} d\mathbb{P}_{X,A}(x,a) \leqslant \delta;\\
&\int_{\mathcal{X}\times\mathcal{A}} f(x,a)\frac{(1-2a)\eta_a(x)}{p_{a,1}} d\mathbb{P}_{X,A}(x,a) \leqslant \delta;\\
&\int_{\mathcal{X}\times\mathcal{A}} f(x,a)\frac{(2a-1)(1-\eta_a(x))}{p_{a,0}} d\mathbb{P}_{X,A}(x,a) \leqslant \delta;\\
&\int_{\mathcal{X}\times\mathcal{A}} f(x,a)\frac{(1-2a)(1-\eta_a(x))}{p_{a,0}} d\mathbb{P}_{X,A}(x,a) \leqslant \delta.
\end{aligned}$$

Then, the generalized Neyman-Pearson lemma motivates us to consider classifiers of the form, for $(c_1, c_2, c_3, c_4) \in \mathbb{R}^4$, $x \in \mathcal{X}$ and $a \in \{0, 1\}$,

$$\begin{aligned}
f_{c_1,c_2,c_3,c_4}(x,a) = &I\Bigg(2\eta_a(x) - 1 > c_1\frac{(2a-1)\eta_a(x)}{p_{a,1}} - c_2\frac{(2a-1)\eta_a(x)}{p_{a,1}}\\
&\qquad + c_3\frac{(2a-1)(1-\eta_a(x))}{p_{a,0}} - c_4\frac{(2a-1)(1-\eta_a(x))}{p_{a,0}}\Bigg).
\end{aligned} \tag{36}$$

Let $\mathcal{T} = [-p_{0,1}, p_{1,1}] \times [-p_{1,0}, p_{0,0}] \setminus \{(-p_{0,1}, p_{0,0}), (-p_{1,0}, p_{1,1})\})$. Take $t_1 = c_1 - c_2 \in \mathbb{R}$ and $t_2 = c_3 - c_4 \in \mathbb{R}$ and define, for $a \in \{0, 1\}$, $T_a : \mathcal{T} \to [0, 1]$ for all $(t_1, t_2) \in \mathcal{T}$ by

$$T_a(t_1, t_2) = \frac{p_{a,1}p_{a,0} + (2a-1)t_2 p_{a,1}}{2p_{a,1}p_{a,0} + (2a-1)\left(t_2 p_{a,1} - t_1 p_{a,0}\right)}. \tag{37}$$

Then, (36) can be simplified to, for all x, a,

$$\begin{aligned} f_{\mathrm{Dis},t_1,t_2}(x,a) =& I\left(2\eta_a(x)-1 > t_1\frac{(2a-1)\eta_a(x)}{p_{a,1}} + t_2\frac{(2a-1)(1-\eta_a(x))}{p_{a,0}}\right) \\ =& I\left(\eta_a(x) > T_a(t_1,t_2)\right). \end{aligned} \tag{38}$$

We then define the disparity functions D_{DO}: $\mathcal{T} \to [-1,1]$ and D_{PD}: $\mathcal{T} \to [-1,1]$ to measure the DO and PD of f_{Dis,t_1,t_2}, for all $t_1, t_2 \in \mathcal{T}$, as

$$D_{\mathrm{DO}}(t_1,t_2) = \mathrm{DO}(f_{\mathrm{Dis},t_1,t_2}) = \int_{\mathcal{X}\times\mathcal{A}} I\left(\eta_a(x) > T_a(t_1,t_2)\right)\frac{(1-2a)\eta_a(x)}{p_{a,1}} d\mathbb{P}_{X,A}(x,a); \tag{39}$$

$$D_{\mathrm{PD}}(t_1,t_2) = \mathrm{PD}(f_{\mathrm{Dis},t_1,t_2}) = \int_{\mathcal{X}\times\mathcal{A}} I\left(\eta_a(x) > T_a(t_1,t_2)\right)\frac{(2a-1)(1-\eta_a(x))}{p_{a,0}} d\mathbb{P}_{X,A}(x,a). \tag{40}$$

Proposition 30 (Continuity and Monotonicity of DO and PD)

(1) For fixed $t_2 \in (-p_{1,0}, p_{0,0})$, both $t \mapsto D_{\mathrm{DO}}(t,t_2)$ and $t \mapsto D_{\mathrm{PD}}(t,t_2)$ are continuous and monotone non-increasing. Moreover, $D_{\mathrm{DO}}(-p_{0,1},t_2) \geqslant 0$, $D_{\mathrm{DO}}(p_{1,1},t_2) \leqslant 0$, $D_{\mathrm{PD}}(-p_{0,1},t_2) \geqslant 0$ and $D_{\mathrm{PD}}(p_{1,1},t_2) \leqslant 0$.

(2) For fixed $t_1 \in (-p_{0,1}, p_{1,1})$, both $t \mapsto D_{\mathrm{DO}}(t_1,t)$ and $t \mapsto D_{\mathrm{PD}}(t_1,t)$ are continuous and monotone non-increasing. Moreover $D_{\mathrm{DO}}(t_1,-p_{1,0}) \geqslant 0$, $D_{\mathrm{DO}}(t_1,p_{0,0}) \leqslant 0$, $D_{\mathrm{PD}}(t_1,-p_{1,0}) \geqslant 0$ and $D_{\mathrm{PD}}(t_1,p_{0,0}) \leqslant 0$.

Next, we define the following quantities, which are useful in determining the optimal thresholds t_1, t_2 for the δ-fair Bayes-optimal classifier. For any $\delta > 0$, define: $\acute{t}_{\mathrm{DO}} : \mathbb{R}_+ \to [-p_{0,1}, p_{1,1}]$ and $\acute{t}_{\mathrm{PD}} : \mathbb{R}_+ \to [-p_{1,0}, p_{0,0}]$ by

$$\acute{t}_{\mathrm{DO}}(\delta) = \arg\min_t\{|t|, |D_{\mathrm{DO}}(t,0)| \leqslant \delta\}, \qquad \acute{t}_{\mathrm{PD}}(\delta) = \arg\min_t\{|t|, |D_{\mathrm{PD}}(0,t)| \leqslant \delta\}. \tag{41}$$

By Proposition 30 with $t_1 = t_2 = 0$, these quantities are well-defined.

Now, we define $t_{\mathrm{DEO},1}$: $[0,\infty) \to \mathbb{R}$ and $t_{\mathrm{DEO},2}$: $[0,\infty) \to \mathbb{R}$, the thresholds of the δ-fair Bayes-optimal classifiers in Theorem 21, for any $\delta \geqslant 0$, as follows; where the claim that they are well-defined is proved in Lemma 33:

$$\begin{aligned} &(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)) \\ =& \begin{cases} (0,0), & (1)\ |D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta))| < \delta \text{ and } |D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta),0)| < \delta; \\ (\acute{t}_{\mathrm{DO}}(\delta),0), & (2)\ |D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta))| \geqslant \delta \text{ and } |D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta),0)| < \delta; \\ (0,\acute{t}_{\mathrm{PD}}(\delta)), & (3)\ |D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta))| < \delta \text{ and } |D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta),0)| \geqslant \delta; \\ \mathrm{solve}\{(D_{\mathrm{DO}}(t_1,t_2), D_{\mathrm{PD}}(t_1,t_2)) = (\delta,\delta)\}, & (4)\ D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta)) \geqslant \delta \text{ and } D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta),0) \geqslant \delta; \\ \mathrm{solve}\{(D_{\mathrm{DO}}(t_1,t_2), D_{\mathrm{PD}}(t_1,t_2)) = (\delta,-\delta)\}, & (5)\ D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta)) \geqslant \delta \text{ and } D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta),0) \leqslant -\delta; \\ \mathrm{solve}\{(D_{\mathrm{DO}}(t_1,t_2), D_{\mathrm{PD}}(t_1,t_2)) = (-\delta,\delta)\}, & (6)\ D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta)) \leqslant -\delta \text{ and } D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta),0) \geqslant \delta; \\ \mathrm{solve}\{(D_{\mathrm{DO}}(t_1,t_2), D_{\mathrm{PD}}(t_1,t_2)) = (-\delta,-\delta)\}, & (7)\ D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta)) \leqslant -\delta \text{ and } D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta),0) \leqslant -\delta. \end{cases} \end{aligned} \tag{42}$$

The seven cases above are mutually exclusive and collectively exhaustive. The first three cases, where at least one of $t_{\mathrm{DEO},1}$ and $t_{\mathrm{DEO},2}$ equals zero, correspond to (1) the unconstrained Bayes-optimal classifier, (2) the δ-fair Bayes-optimal classifier under equality of opportunity, and (3) the δ-fair Bayes-optimal classifier under predictive equality. These satisfy the δ-parity constraint for equalized odds. For the remaining four cases, both $t_{\mathrm{DEO},1}$ and $t_{\mathrm{DEO},2}$ are carefully selected to satisfy the hard constraint that for $a \in \{\mathrm{DO}, \mathrm{PD}\}$, $D_a(t_{\mathrm{DEO},1}, t_{\mathrm{DEO},2})$ equals either δ or $-\delta$.

B.1.1 Proof of Proposition 30

We prove claim (1), and claim (2) follows by similar arguments. For any $0 \leqslant c_1 \leqslant c_2$, we have that $x \mapsto c_1/(c_2+x)$ is monotone non-increasing on $(-c_2, \infty)$ and $x \mapsto c_1/(c_2-x)$ is monotone non-decreasing on $(-\infty, c_2)$. Let, for $a \in \{0,1\}$, $0 \leqslant c_{1,a} = p_{a,1}p_{a,0} + t_2 p_{a,1} < 2p_{a,1}p_{a,0} + t_2 p_{a,1} = c_{2,a}$. Fixing $t_2 \in (-p_{1,0}, p_{0,0})$, on one hand, we have for $-p_{0,1} \leqslant t_{1,1} \leqslant t_{1,2} \leqslant p_{1,1}$ that $t_{1,1}p_{1,0} \leqslant t_{1,2}p_{1,0} \leqslant p_{1,1}p_{1,0} < 2p_{1,1}p_{1,0} + t_2 p_{1,1} = c_{2,1}$. Thus, since $t_{1,1} \leqslant t_{1,2}$,

$$T_1(t_{1,1}, t_2) = \frac{p_{1,1}p_{1,0} + t_2 p_{1,1}}{2p_{1,1}p_{1,0} + t_2 p_{1,1} - t_{1,1}p_{1,0}} \leqslant \frac{p_{1,1}p_{1,0} + t_2 p_{1,1}}{2p_{1,1}p_{1,0} + t_2 p_{1,1} - t_{1,2}p_{1,0}} = T_1(t_{1,2}, t_2).$$

On the other hand, $t_{1,2}p_{0,0} \geqslant t_{1,1}p_{0,0} \geqslant -p_{0,1}p_{0,0} > t_2 p_{0,1} - 2p_{0,1}p_{0,0} = -c_{2,0}$. Thus

$$T_0(t_{1,1}, t_2) = \frac{p_{0,1}p_{0,0} - t_2 p_{0,1}}{2p_{0,1}p_{0,0} - t_2 p_{0,1} + t_{1,1}p_{0,0}} \geqslant \frac{p_{0,1}p_{0,0} - t_2 p_{0,1}}{2p_{0,1}p_{0,0} - t_2 p_{0,1} + t_{1,2}p_{0,0}} = T_0(t_{1,2}, t_2).$$

Then, by definition, we have

$$\begin{aligned}
&D_{\mathrm{DO}}(t_{1,1}, t_2) = \mathbb{P}_{A=1,Y=1}\left(\eta^Y_{A=1}(X) > T_1(t_{1,1}, t_2)\right) - \mathbb{P}_{A=0,Y=1}\left(\eta^Y_{A=0}(X) > T_0(t_{1,1}, t_2)\right) \\
&\geqslant \mathbb{P}_{A=1,Y=1}\left(\eta^Y_{A=1}(X) > T_1(t_{1,2}, t_2)\right) - \mathbb{P}_{A=0,Y=1}\left(\eta^Y_{A=0}(X) > T_0(t_{1,2}, t_2)\right) = D_{\mathrm{DO}}(t_{1,2}, t_2);
\end{aligned}$$

and

$$\begin{aligned}
&D_{\mathrm{PD}}(t_{1,1}, t_2) = \mathbb{P}_{A=1,Y=0}\left(\eta^Y_{A=1}(X) > T_1(t_{1,1}, t_2)\right) - \mathbb{P}_{A=0,Y=0}\left(\eta^Y_{A=0}(X) > T_0(t_{1,1}, t_2)\right) \\
&\geqslant \mathbb{P}_{A=1,Y=0}\left(\eta^Y_{A=1}(X) > T_1(t_{1,2}, t_2)\right) - \mathbb{P}_{A=0,Y=0}\left(\eta^Y_{A=0}(X) > T_0(t_{1,2}, t_2)\right) = D_{\mathrm{PD}}(t_{1,2}, t_2).
\end{aligned}$$

Thus, for fixed $t_2 \in (-p_{1,0}, p_{0,0})$, both $t \mapsto D_{\mathrm{DO}}(t, t_2)$ and $t \mapsto D_{\mathrm{PD}}(t, t_2)$ are monotone non-increasing on $[-p_{0,1}, p_{1,1}]$. Continuity follows since, for $a \in \{0,1\}$, $t \mapsto T_a(t, t_2)$ is continuous and $\eta_a(X)$ has density function on $\mathcal{X}$.

Moreover, since for any fixed $t_2 \in (-p_{1,0}, p_{0,0})$, $T_1(p_{1,1}, t_2) = T_0(-p_{0,1}) = 1$, we have,

$$\begin{aligned}
&D_{\mathrm{DO}}(-p_{0,1}, t_2) = \mathbb{P}_{A=1,Y=1}\left(\eta^Y_{A=1}(X) > T_1(-p_{0,1}, t_2)\right) - \mathbb{P}_{A=0,Y=1}\left(\eta^Y_{A=0}(X) > 1\right) \\
&= \mathbb{P}_{A=1,Y=1}\left(\eta^Y_{A=1}(X) > T_1(-p_{0,1}, t_2)\right) \geqslant 0; \\
&D_{\mathrm{DO}}(p_{1,1}, t_2) = \mathbb{P}_{A=1,Y=1}\left(\eta^Y_{A=1}(X) > 1\right) - \mathbb{P}_{A=0,Y=1}\left(\eta^Y_{A=0}(X) > T_0(p_{1,1}, t_2)\right) \\
&= -\mathbb{P}_{A=0,Y=1}\left(\eta^Y_{A=0}(X) > T_0(p_{1,1}, t_2)\right) \leqslant 0; \\
&D_{\mathrm{PD}}(-p_{0,1}, t_2) = \mathbb{P}_{A=1,Y=0}\left(\eta^Y_{A=1}(X) > T_1(-p_{0,1}, t_2)\right) - \mathbb{P}_{A=0,Y=0}\left(\eta^Y_{A=0}(X) > 1\right) \\
&= \mathbb{P}_{A=1,Y=2-j}\left(\eta^Y_{A=1}(X) > T_1(-p_{0,1}, t_2)\right) \geqslant 0; \\
&D_{\mathrm{PD}}(p_{1,1}, t_2) = \mathbb{P}_{A=1,Y=0}\left(\eta^Y_{A=1}(X) > 1\right) - \mathbb{P}_{A=0,Y=0}\left(\eta^Y_{A=0}(X) > T_0(p_{1,1}, t_2)\right) \\
&= -\mathbb{P}_{A=0,Y=2-j}\left(\eta^Y_{A=0}(X) > T_0(p_{1,1}, t_2)\right) \leqslant 0.
\end{aligned}$$

B.1.2 Proof of Theorem 21

The following lemmas are useful in proving Theorem 21. The first result characterizes the relation between D_{DO} and D_{PD}.

Lemma 31 (Relation between $D_{\mathbf{DO}}$ and $D_{\mathbf{PD}}$) *Let t, t_1 and t_2 be three real numbers.*

(1) If $t_2 \geqslant 0$ and $D_{\mathrm{DO}}(t_1, t_2) = D_{\mathrm{DO}}(t, 0)$, we have $D_{\mathrm{PD}}(t_1, t_2) \leqslant D_{\mathrm{PD}}(t, 0)$, and the equality holds if and only if $f_{\mathrm{Dis},t_1,t_2} = f_{\mathrm{Dis},t,0}$ almost surely with respect to $\mathbb{P}_{X,A}$.

(2) If $t_2 \leqslant 0$ and $D_{\mathrm{DO}}(t_1, t_2) = D_{\mathrm{DO}}(t, 0)$, we have $D_{\mathrm{PD}}(t_1, t_2) \geqslant D_{\mathrm{PD}}(t, 0)$, and the equality holds if and only if $f_{\mathrm{Dis},t_1,t_2} = f_{\mathrm{Dis},t,0}$ almost surely with respect to $\mathbb{P}_{X,A}$.

(3) If $t_1 \geqslant 0$ and $D_{\mathrm{PD}}(t_1, t_2) = D_{\mathrm{PD}}(0, t)$, we have $D_{\mathrm{DO}}(t_1, t_2) \leqslant D_{\mathrm{DO}}(0, t)$, and the equality holds if and only if $f_{\mathrm{Dis},t_1,t_2} = f_{\mathrm{Dis},0,t}$ almost surely with respect to $\mathbb{P}_{X,A}$.

(4) If $t_1 \leqslant 0$ and $D_{\mathrm{PD}}(t_1, t_2) = D_{\mathrm{PD}}(0, t)$, we have $D_{\mathrm{DO}}(t_1, t_2) \geqslant D_{\mathrm{DO}}(0, t)$, and the equality holds if and only if $f_{\mathrm{Dis},t_1,t_2} = f_{\mathrm{Dis},0,t}$ almost surely with respect to $\mathbb{P}_{X,A}$.

Proof [Proof of Lemma 31] For conciseness, we show the proofs for claim (1) Claims (2) through (4) can be verified using the same approach as for claim (1).

Recall f_{Dis,t_1,t_2} and $f_{\mathrm{Dis},t,0}$ from (38). By Lemma 28 adapted to equality of opportunity, $f_{\mathrm{Dis},t,0}$ is Bayes-optimal with respect to DO, and as f_{Dis,t_1,t_2} is another classifier,

$$R(f_{\mathrm{Dis},t,0}) = \min_{f\in\mathcal{F}} \left\{ R_f : \frac{t \cdot \mathrm{DO}(f)}{|t|} \leqslant \frac{t \cdot D_{\mathrm{DO}}(t,0)}{|t|} \right\} \leqslant R(f_{\mathrm{Dis},t_1,t_2}).$$

Moreover, since $\mathbb{P}_{X|A=a}(\eta_a(X) = T_a(t_1,t_2)) = 0$ when $\eta_a(X)$ is a continuous random variable with respect to $\mathbb{P}_{X|A=a}$, for all $f' \in \arg\min_{f\in\mathcal{F}}\{R_f : t \cdot \mathrm{DO}(f)/|t| \leqslant t \cdot D_{\mathrm{DO}(t,0)}/|t|\}$, $f'(x,a) = f_{\mathrm{Dis},t,0}(x,a)$ almost surely with respect to $\mathbb{P}_{X,A}$.

We have two cases: (i) $R(f_{\mathrm{Dis},t,0}) = R(f_{\mathrm{Dis},t_1,t_2})$ and (ii) $R(f_{\mathrm{Dis},t,0}) < R(f_{\mathrm{Dis},t_1,t_2})$.

(i) When $R(f_{\mathrm{Dis},t,0}) = R(f_{\mathrm{Dis},t_1,t_2})$, we have $f_{\mathrm{Dis},t_1,t_2}(x,a) = f_{\mathrm{Dis},t,0}(x,a)$ almost surely with respect to $\mathbb{P}_{X,A}$. As a result,

$$\begin{aligned} &D_{\mathrm{PD}}(t_1,t_2) - D_{\mathrm{PD}}(t,0) \\ &= \int_{\mathcal{X}\times\mathcal{A}} (f_{\mathrm{Dis},t_1,t_2}(x,a) - f_{\mathrm{Dis},t,0}(x,a)) \frac{(2a-1)(1-\eta_a(x))}{p_{a,0}} d\mathbb{P}_{X,A}(x,a) = 0. \end{aligned}$$

(ii) When $R(f_{\mathrm{Dis},t,0}) < R(f_{\mathrm{Dis},t_1,t_2})$, we first notice that since $t_2 \geqslant 0$, by Lemma 32,

$$f_{\mathrm{Dis},t_1,t_2} \in \arg\min_{f\in\mathcal{F}} \left\{ R(f) : \frac{t_1 \cdot \mathrm{DO}(f)}{|t_1|} \leqslant \frac{t_1 \cdot D_{\mathrm{DO}}(t_1,t_2)}{|t_1|},\ \mathrm{PD}(f) \leqslant D_{\mathrm{PD}}(t_1,t_2) \right\}. \quad (43)$$

This implies $D_{\mathrm{PD}}(t_1,t_2) < D_{\mathrm{PD}}(t,0)$. Otherwise, $f_{\mathrm{Dis},t,0}$ satisfies $\mathrm{DO}(f_{\mathrm{Dis},t,0}) = D_{\mathrm{DO}}(t_1,t_2)$, $\mathrm{PD}(f_{\mathrm{Dis},t,0}) = D_{\mathrm{PD}}(t,0) \leqslant D_{\mathrm{PD}}(t_1,t_2)$ and $R(f_{\mathrm{Dis},t,0}) < R(f_{\mathrm{Dis},t_1,t_2})$, which contradicts (43).

Thus, we can conclude that $D_{\mathrm{PD}}(t_1, t_2) \leqslant D_{\mathrm{PD}}(t, 0)$, and the equality holds if and only if $f_{\mathrm{Dis},t_1,t_2} = f_{\mathrm{Dis},t,0}$ almost surely with respect to $\mathbb{P}_{X,A}$. ■

The next result characterizes Bayes-optimal classifiers under equalized odds.

Lemma 32 Characterizing f_{Dis,t_1,t_2} as Bayes-optimal Classifiers under Equalized Odds *For $(t_1, t_2) \in \mathbb{R}^2$ and $\delta \geqslant 0$, let f_{Dis,t_1,t_2}, $D_{\mathrm{DO}}(t)$, $D_{\mathrm{PD}}(t)$ be defined in* (38), (39) *and* (40)*, respectively. Then, recalling we use the convention that* $0/0 = 0$*,*

$$f_{\mathrm{Dis},t_1,t_2} \in \arg\min_{f\in\mathcal{F}} \left\{ R(f) : \frac{t_1 \cdot \mathrm{DO}(f)}{|t_1|} \leqslant \frac{t_1 \cdot D_{\mathrm{DO}}(t_1, t_2)}{|t_1|},\ \frac{t_2 \cdot \mathrm{PD}(f)}{|t_2|} \leqslant \frac{t_2 \cdot D_{\mathrm{PD}}(t_1, t_2)}{|t_2|} \right\}.$$

Moreover, for all

$$f' \in \arg\min_{f\in\mathcal{F}} \left\{ R(f) : \frac{t_1 \cdot \mathrm{DO}(f)}{|t_1|} \leqslant \frac{t_1 \cdot D_{\mathrm{DO}}(t_1, t_2)}{|t_1|},\ \frac{t_2 \cdot \mathrm{PD}(f)}{|t_2|} \leqslant \frac{t_2 \cdot D_{\mathrm{PD}}(t_1, t_2)}{|t_2|} \right\},$$

we have $f' = f_{\mathrm{Dis},t_1,t_2}$ almost surely with respect to $\mathbb{P}_{X,A}$. In particular, if $t_1 \cdot D_{\mathrm{DO}}(t_1, t_2) \geqslant 0$ and $t_2 \cdot D_{\mathrm{PD}}(t_1, t_2) \geqslant 0$,

$$f_{\mathrm{Dis},t_1,t_2} = \arg\min_{f\in\mathcal{F}} \{R(f) : |\mathrm{DO}(f)| \leqslant |D_{\mathrm{DO}}(t_1, t_2)|, |\mathrm{PD}(f)| \leqslant |D_{\mathrm{PD}}(t_1, t_2)|\}.$$

Proof [Proof of Lemma 32] Similar to Lemma 28, this result is a consequence of the generalized Neyman-Pearson lemma. If either $t_1 = 0$ or $t_2 = 0$, the result is an application of Lemma 28. Now, we assume $t_1 \neq 0$ and $t_2 \neq 0$. Let, for all x, a, $\phi_0(x, a) = 2\eta^Y_{A=1}(x, a) - 1$, $\phi_1(x, a) = (2a - 1)\eta_a(x)/p_{a,1}$ and $\phi_2(x, a) = (2a - 1)(1 - \eta_a(x))/p_{a,0}$. Also, let, for all classifiers f, and $t \in \mathbb{R}$, $\overline{\mathrm{DO}}_t(f) = t \cdot \mathrm{DO}(f)/|t|$ and $\overline{\mathrm{PD}}_t(f) = t \cdot \mathrm{PD}(f)/|t|$. We have, for $(t_1, t_2) \in \mathbb{R}^2$ and all x, a, that with f_{Dis,t_1,t_2} from (38),

$$\begin{aligned} f_{\mathrm{Dis},t_1,t_2}(x, a) =& I\left(\phi_0(x, a) > t_1\phi_1(x, a) + t_2\phi_2(x, a)\right) \\ =& I\left(\phi_0(x, a) > |t_1|\frac{t_1\phi_1(x, a)}{|t_1|} + |t_2|\frac{t_2\phi_2(x, a)}{|t_2|}\right). \end{aligned}$$

Moreover, by Lemma 27 and (34), we can write $\mathrm{Acc}(f)$, $\overline{\mathrm{DO}}_t(f)$ and $\overline{\mathrm{PD}}_t(f)$ as

$$\begin{aligned} \mathrm{Acc}(f) &= \int_{\mathcal{A}}\int_{\mathcal{X}} f(x, a)\phi_0(x, a)d\mathbb{P}_{X,A}(x, a) + \int_{\mathcal{A}}\int_{\mathcal{X}} (1 - \eta_a(x))d\mathbb{P}_{X,A}(x, a); \\ \overline{\mathrm{DO}}_{t_1}(f) &= \frac{t_1}{|t_1|}\mathrm{DO}(f) = \int_{\mathcal{X}\times\mathcal{A}} f(x, a)\frac{t_1\phi_1(x, a)}{|t_1|}d\mathbb{P}_{X,A}(x, a); \\ \overline{\mathrm{PD}}_{t_2}(f) &= \frac{t_2}{|t_2|}\mathrm{PD}(f) = \int_{\mathcal{X}\times\mathcal{A}} f(x, a)\frac{t_2\phi_2(x, a)}{|t_2|}d\mathbb{P}_{X,A}(x, a). \end{aligned}$$

Define the sets of classifiers

$$\begin{aligned} \mathcal{F}_{t_1,t_2,=} &= \left\{ f : \overline{\mathrm{DO}}_{t_1}(f) = \frac{t_1 \cdot D_{\mathrm{DO}}(t_1, t_2)}{|t_1|},\ \overline{\mathrm{PD}}_{t_2}(f) = \frac{t_2 \cdot D_{\mathrm{PD}}(t_1, t_2)}{|t_2|} \right\}; \\ \mathcal{F}_{t_1,t_2,|\cdot|,\leqslant} &= \left\{ f : |\overline{\mathrm{DO}}_{t_1}(f)| \leqslant |D_{\mathrm{DO}}(t_1, t_2)|, |\overline{\mathrm{PD}}_{t_2}(f)| \leqslant |D_{\mathrm{PD}}(t_1, t_2)| \right\}; \end{aligned}$$

$$\mathcal{F}_{t_1,t_2,\leqslant} = \left\{ f : \overline{\mathrm{DO}}_{t_1}(f) \leqslant \frac{t_1 \cdot D_{\mathrm{DO}}(t_1,t_2)}{|t_1|},\ \overline{\mathrm{PD}}_{t_2}(f) \leqslant \frac{t_2 \cdot D_{\mathrm{PD}}(t_1,t_2)}{|t_2|} \right\}.$$

As $f_{\mathrm{Dis},t_1,t_2} \in \mathcal{F}_{t_1,t_2,=}$, by the generalized Neyman-Pearson lemma (Lemma 9),

$$f_{\mathrm{Dis},t_1,t_2} \in \arg\max_{f\in\mathcal{F}_{t_1,t_2,=}} \mathrm{Acc}(f).$$

Moreover, since $|t_1| \geqslant 0$ and $|t_2| \geqslant 0$, we have

$$\begin{aligned} f_{\mathrm{Dis},t_1,t_2} \in & \arg\max_{f\in\mathcal{F}_{t_1,t_2,\leqslant}} \mathrm{Acc}(f) \\ =& \arg\min_{f\in\mathcal{F}} \left\{ R(f) : \frac{t_1 \mathrm{DO}(f)}{|t_1|} \leqslant \frac{t_1 D_{\mathrm{DO}}(t_1,t_2)}{|t_1|},\ \frac{t_2 \mathrm{PD}(f)}{|t_2|} \leqslant \frac{t_2 D_{\mathrm{DO}}(t_1,t_2)}{|t_2|} \right\}. \end{aligned}$$

In addition, since $\mathbb{P}_{X|A=a}(\eta_a(X) = T_a(t_1,t_2)) = 0$ when $\eta_a(X)$ is a continuous random variable with respect to $\mathbb{P}_{X|A=a}$, for all $f' \in \arg\max_{f\in\mathcal{F}_{t_1,t_2,\leqslant}} \mathrm{Acc}(f)$, $f' = f_{\mathrm{Dis},t_1,t_2}$ almost surely with respect to $\mathbb{P}_{X,A}$. Furthermore, if $t_1 \cdot D_{\mathrm{DO}}(t_1,t_2) \geqslant 0$ and $t_2 \cdot D_{\mathrm{PD}}(t_1,t_2) \geqslant 0$, we have $f_{\mathrm{Dis},t_1,t_2} \in \mathcal{F}_{t_1,t_2,|\cdot|,\leqslant} \subset \mathcal{F}_{t_1,t_2,\leqslant}$. Thus

$$f_{\mathrm{Dis},t_1,t_2} = \arg\min_{f\in\mathcal{F}} \{R(f) : |\mathrm{DO}(f)| \leqslant |D_{\mathrm{DO}}(t_1,t_2)|, |\mathrm{PD}(f)| \leqslant |D_{\mathrm{PD}}(t_1,t_2)|\}.$$

This finishes the proof. ■

Next, we show that $t_{\mathrm{DEO},1}(\delta)$ and $t_{\mathrm{DEO},2}(\delta)$ are well defined in cases (4) to (7). For conciseness, we consider case (4), and the arguments for the other cases are similar.

Lemma 33 (The Quantities $t_{\mathrm{DEO},1}(\delta)$ and $t_{\mathrm{DEO},2}(\delta)$ Are Well-Defined) *Let $\acute{t}_{\mathrm{DO}}(\delta)$ and $\acute{t}_{\mathrm{PD}}(\delta)$ be defined in* (41)*, Suppose that, for $a \in \{0,1\}$, $\eta_a(X)$ has a density function on $\mathcal{X}$. Then, for fixed $\delta \geqslant 0$, when $D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta)) > \delta$ and $D_{\mathrm{PD}}(0,\acute{t}_{\mathrm{DO}}(\delta)) > \delta$, we have $\acute{t}_{\mathrm{DO}}(\delta) > 0$, $\acute{t}_{\mathrm{PD}}(\delta) > 0$, and there exist $(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)) \in [0,\acute{t}_{\mathrm{DO}}(\delta)] \times [0,\acute{t}_{\mathrm{PD}}(\delta)]$ such that $(D_{\mathrm{DO}}(t_1,t_2), D_{\mathrm{PD}}(t_1,t_2)) = (\delta,\delta)$.*

Proof [Proof of Lemma 33] If $\acute{t}_{\mathrm{DO}}(\delta) \leqslant 0$, then by the definition of $\acute{t}_{\mathrm{DO}}(\delta)$ in (41), we have $D_{\mathrm{DO}}(0,0) \leqslant \delta$ and $\mathrm{DO}(\acute{t}_{\mathrm{DO}}(\delta),0) \leqslant \delta$. Moreover, by Proposition 30, $D_{\mathrm{PD}}(0,0) \geqslant D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta),0) > \delta$. Thus, it follows that $\acute{t}_{\mathrm{PD}}(\delta) > 0$. By applying Proposition 30 again, $D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta)) \leqslant D_{\mathrm{DO}}(0,0) \leqslant \delta$, which is a contradiction. As a result, we have $\acute{t}_{\mathrm{DO}}(\delta) > 0$ and $D_{\mathrm{DO}}(\acute{t}_{\mathrm{DO}}(\delta),0) = \delta$. On the other hand, since $D_{\mathrm{DO}}(0,p_{0,0}) \leqslant \delta$, we have $\acute{t}_{\mathrm{PD}}(\delta) < p_{0,0}$. Similarly, we can show $0 < \acute{t}_{\mathrm{DO}}(\delta) < p_{1,1}$ and $D_{\mathrm{PD}}(0,\acute{t}_{\mathrm{PD}}(\delta)) = \delta$.

Now, for any $t_2 \in (0,\acute{t}_{\mathrm{PD}}(\delta))$, we have $D_{\mathrm{DO}}(0,t_2) \geqslant D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta)) > \delta$ and $D_{\mathrm{DO}}(p_{1,1},t_2) \leqslant \delta$. By the continuity of $D_{\mathrm{DO}}(t,t_2)$ as a function of t, there exists $t_1 \in (0,p_{1,1})$ such that $D_{\mathrm{DO}}(t_1,t_2) = \delta$. We then define $Q_\delta : [0,\acute{t}_{\mathrm{PD}}(\delta)] \to (0,p_{1,1})$ for all $t_2 \in (0,p_{1,1})$ as

$$Q_\delta(t_2) = \inf_{t\in(0,p_{1,1})} \{D_{\mathrm{DO}}(t,t_2) < \delta\}.$$

Clearly, we have $Q_\delta(0) \leqslant \acute{t}_{\mathrm{DO}}(\delta)$ and, since $D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta)) > \delta$, we have $Q_\delta(\acute{t}_{\mathrm{PD}}(\delta)) > 0$.

Now, we consider the function P defined on $[0,\acute{t}_{\mathrm{PD}}(\delta)]$ $t_2 \mapsto \mathbb{P}(t_2) = D_{\mathrm{PD}}(Q_\delta(t_2),t_2)$. Then we have $\mathbb{P}(0) = D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta),0) > \delta$ and $\mathbb{P}(\acute{t}_{\mathrm{PD}}(\delta)) = \mathrm{PD}(Q_\delta(\acute{t}_{\mathrm{PD}}(\delta)),\acute{t}_{\mathrm{PD}}(\delta)) \leqslant$

$D_{\mathrm{PD}}(0, \acute{t}_{\mathrm{PD}}(\delta)) = \delta$. Thus, by the continuity of D_{PD}, there exists $t_{\mathrm{DEO},2}(\delta) \in [0, \acute{t}_{\mathrm{DO}}(\delta)]$ such that $D_{\mathrm{PD}}(Q_\delta(t_{\mathrm{DEO},2}(\delta)), t_{\mathrm{DEO},2}(\delta)) = \delta$. Setting $t_{\mathrm{DEO},1}(\delta) = Q_\delta(t_{\mathrm{DEO},2}(\delta))$, we have $D_{\mathrm{DO}}(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)) = D_{\mathrm{PD}}(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)) = \delta$. Moreover, since $t_{\mathrm{DEO},2}(\delta) \leqslant \acute{t}_{\mathrm{PD}}(\delta)$ and $D_{\mathrm{PD}}(0, \acute{t}_{\mathrm{PD}}(\delta)) = D_{\mathrm{PD}}(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)) = \delta$, we have $t_{\mathrm{DEO},1}(\delta) \geqslant 0$. On the other hand, since $t_{\mathrm{DEO},2}(\delta) \geqslant 0$ and $D_{\mathrm{DO}}(\acute{t}_{\mathrm{DO}}(\delta), 0) = D_{\mathrm{PD}}(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)) = \delta$, we have $t_{\mathrm{DEO},1}(\delta) \leqslant \acute{t}_{\mathrm{DO}}(\delta)$. This finishes the proof. ■

We proceed with the proof of Theorem 21.

Proof [Proof of Theorem 21] We demonstrate the result in cases (1), (2), (4), and (5). The results in other cases can be verified similarly: case (3) is analogous to case (2), case (6) is analogous to case (5), and case (7) is analogous to case (4).

The following facts follow directly from the definitions of $\acute{t}_{\mathrm{DO}}(\delta)$ and $\acute{t}_{\mathrm{PD}}(\delta)$ in (41), and Proposition 30:

Fact (1). For $j \in \{\mathrm{DO}, \mathrm{PD}\}$, $\delta \mapsto \acute{t}_j(\delta)$ is monotone non-increasing.

Fact (2). For $j \in \{\mathrm{DO}, \mathrm{PD}\}$, $\acute{t}_j > 0$ if and only if $D_j(0,0) > \delta$. Further, $\acute{t}_j < 0$ if and only if $D_j(0,0) < -\delta$;

Fact (3). If $D_{\mathrm{DO}}(0,0) > \delta$, then $D_{\mathrm{DO}}(\acute{t}_{\mathrm{PD}}(\delta), 0) = \delta$; If $D_{\mathrm{DO}}(0,0) < -\delta$, then $D_{\mathrm{DO}}(\acute{t}_{\mathrm{PD}}(\delta), 0) = -\delta$. Further, if $D_{\mathrm{PD}}(0,0) > \delta$, then $D_{\mathrm{PD}}(0, \acute{t}_{\mathrm{PD}}(\delta)) = \delta$. Finally, if $D_{\mathrm{PD}}(0,0) < -\delta$, then $D_{\mathrm{PD}}(0, \acute{t}_{\mathrm{PD}}(\delta)) = -\delta$;

We now study the cases mentioned for Theorem 21.

Case (1). $|D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta), 0)| < \delta$ and $|D_{\mathrm{DO}}(0, \acute{t}_{\mathrm{PD}}(\delta))| < \delta$.

In this scenario, we have $t_{\mathrm{DEO},1}(\delta) = t_{\mathrm{DEO},2}(\delta) = 0$ and, for all x, a, $f_{\mathrm{Dis},0,0}(x,a) = f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)}(x,a) = I\left(\eta_a(x) > 1/2\right)$ is the unconstrained Bayes-optimal classifier. As a result, we only need to prove $|D_{\mathrm{DO}}(0,0)| \leqslant \delta$ and $|D_{\mathrm{PD}}(0,0)| \leqslant \delta$.

If $\acute{t}_{\mathrm{DO}}(\delta) > 0$, then we have $D_{\mathrm{DO}}(0,0) > \delta$ and $D_{\mathrm{DO}}(\acute{t}_{\mathrm{DO}}(\delta), 0) = \delta$. By $|D_{\mathrm{DO}}(0, \acute{t}_{\mathrm{PD}}(\delta))| < \delta$ and Proposition (30), we conclude $\acute{t}_{\mathrm{PD}}(\delta) > 0$, which leads to $D_{\mathrm{PD}}(0,0) > \delta$ and $D_{\mathrm{PD}}(0, \acute{t}_{\mathrm{PD}}(\delta)) = \delta$. Now, let $\widetilde{t} = \arg\min_t\{t : D_{\mathrm{DO}}(t, 0) \leqslant D_{\mathrm{DO}}(0, \acute{t}_{\mathrm{PD}}(\delta))\}$. We have $D_{\mathrm{DO}}(\widetilde{t}, 0) = D_{\mathrm{DO}}(0, \acute{t}_{\mathrm{PD}}(\delta))$, which leads to $\widetilde{t} \geqslant \acute{t}_{\mathrm{DO}}(\delta) > 0$. Then, by result (1) of Lemma 31, $D_{\mathrm{PD}}(0, \acute{t}_{\mathrm{PD}}(\delta)) \leqslant D_{\mathrm{PD}}(\widetilde{t}, 0) \leqslant D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta), 0) < \delta$, which contradicts $|D_{\mathrm{PD}}(0, \acute{t}_{\mathrm{PD}}(\delta))| = \delta$. Thus, we can conclude that $\acute{t}_{\mathrm{DO}}(\delta) \leqslant 0$.

With analogous arguments, we can also demonstrate that $\acute{t}_{\mathrm{DO}}(\delta) \geqslant 0$. Then, we must have $\acute{t}_{\mathrm{DO}}(\delta) = 0$. As a result, $|D_{\mathrm{DO}}(0,0)| \leqslant \delta$ and $|D_{\mathrm{PD}}(0,0)| \leqslant \delta$.

Case (2). $|D_{\mathrm{DO}}(0, \acute{t}_{\mathrm{PD}}(\delta))| \geqslant \delta$ and $|D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta), 0)| < \delta$.

In this scenario, we have $t_{\mathrm{DEO},1} = \acute{t}_{\mathrm{DO}}(\delta) = \arg\min_t\{|t| : |\mathrm{DO}| \leqslant \delta\}$ and $t_{\mathrm{DEO},2} = 0$. On one hand, by Theorem 14 adapted for equality of opportunity, we have

$$R(f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)}) = R(f_{\mathrm{Dis},\acute{t}_{\mathrm{DO}}(\delta),0}) = \min_{f\in\mathcal{F}}\{R(f) : |\mathrm{DO}(f)| \leqslant \delta\}.$$

On the other hand, given that $|D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta), 0)| < \delta$, we have

$$f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)} = f_{\mathrm{Dis},\acute{t}_{\mathrm{DO}}(\delta),0}$$

$$\in \{f \in \mathcal{F} : \max(|\mathrm{DO}(f)|, |\mathrm{PD}(f))| \leqslant \delta\} \subset \{f \in \mathcal{F} : \mathrm{DO}(f) \leqslant \delta\}.$$

It follows that

$$\begin{aligned} R(f_{\mathrm{Dis}, t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)}) &\geqslant \min_{f \in \mathcal{F}} \{R(f) : \max(|\mathrm{DO}(f)|, |\mathrm{PD}(f)|) \leqslant \delta\} \\ &\geqslant \min_{f \in \mathcal{F}} \{R(f) : \mathrm{DO}(f) \leqslant \delta\}. \end{aligned}$$

Then, we can conclude that

$$f_{\mathrm{Dis}, t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)} \in \arg\min_{f \in \mathcal{F}} \{R(f) : \max(|\mathrm{DO}(f)|, |\mathrm{PD}(f))| \leqslant \delta\}.$$

Case (4) $D_{\mathrm{DO}}(0, \acute{t}_{\mathrm{PD}}(\delta)) \geqslant \delta$ and $D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta), 0) \geqslant \delta$

In this case, $(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta))$ is the solution of the equation $(D_{\mathrm{DO}}(t_1, t_2), D_{\mathrm{PD}}(t_1, t_2)) = (\delta, \delta)$. Following the proof of Lemma 33, we find that $\acute{t}_{\mathrm{DO}}(\delta) > 0$ and $\acute{t}_{\mathrm{PD}}(\delta) > 0$. This further implies $D_{\mathrm{DO}}(\acute{t}_{\mathrm{DO}}(\delta), 0) = D_{\mathrm{PD}}(0, \acute{t}_{\mathrm{DO}}(\delta)) = \delta$. We consider three sub-cases: (4.i) $t_{\mathrm{DEO},1} \leqslant 0$, (4.ii) $t_{\mathrm{DEO},1} > 0$ and $t_{\mathrm{DEO},2} \leqslant 0$ and (4.iii) $t_{\mathrm{DEO},1} > 0$ and $t_{\mathrm{DEO},2} > 0$.

(4.i) When $t_{\mathrm{DEO},1} \leqslant 0$, from result (4) of Lemma 31, since $D_{\mathrm{PD}}(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)) = \delta = D_{\mathrm{PD}}(0, \acute{t}_{\mathrm{PD}}(\delta))$, we have $D_{\mathrm{DO}}(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)) \geqslant D_{\mathrm{DO}}(0, \acute{t}_{\mathrm{PD}}(\delta)) \geqslant \delta$. As $D_{\mathrm{DO}}(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)) = \delta$, we have $D_{\mathrm{DO}}(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)) = D_{\mathrm{DO}}(0, \acute{t}_{\mathrm{PD}}(\delta)) = \delta$, which indicates $f_{\mathrm{Dis}, t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)} = f_{\mathrm{Dis}, 0, \acute{t}_{\mathrm{PD}}(\delta)}$ almost surely with respect to $\mathbb{P}_{X,A}$. On one hand, as $\acute{t}_{\mathrm{PD}}(\delta) > 0$, by Lemma 28 adapted for predictive equality, we have

$$\begin{aligned} R(f_{\mathrm{Dis}, t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)}) &= R(f_{\mathrm{Dis}, 0, \acute{t}_{\mathrm{PD}}(\delta)}) \\ &= \min_{f \in \mathcal{F}} \{R(f) : \mathrm{PD}(f) \leqslant D_{\mathrm{DP}}(0, \acute{t}_{\mathrm{PD}}(\delta)) = \delta\}. \end{aligned}$$

On the other hand, since $D_{\mathrm{DO}}(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)) = D_{\mathrm{PD}}(t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)) = \delta$, we have

$$\begin{aligned} f_{\mathrm{Dis}, t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)} &\in \{f \in \mathcal{F} : \max(|\mathrm{DO}(f)|, |\mathrm{PD}(f)|) \leqslant \delta\} \\ &\subset \{f \in \mathcal{F} : \mathrm{PD}(f) \leqslant \delta\}. \end{aligned}$$

It follows that

$$\begin{aligned} R(f_{\mathrm{Dis}, t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)}) &\geqslant \min_{f \in \mathcal{F}} \{R(f) : \max(|\mathrm{DO}(f)|, |\mathrm{PD}(f))| \leqslant \delta\} \\ &\geqslant \min_{f \in \mathcal{F}} \{R(f) : \mathrm{PD}(f) \leqslant \delta\}. \end{aligned}$$

Then, we can conclude that

$$f_{\mathrm{Dis}, t_{\mathrm{DEO},1}(\delta), t_{\mathrm{DEO},2}(\delta)} \in \arg\min_{f \in \mathcal{F}} \{R(f) : \max(|\mathrm{DO}(f)|, |\mathrm{PD}(f))| \leqslant \delta\}.$$

(4.ii) When $t_{\rm DEO,1} > 0$ and $t_{\rm DEO,2} \leqslant 0$, with a similar argument as in case (i), we can show that $f_{{\rm Dis},t_{\rm DEO,1}(\delta),t_{\rm DEO,2}(\delta)} = f_{{\rm Dis},\acute{t}_{\rm DO}(\delta),0}$ almost surely with respect to $\mathbb{P}_{X,A}$ and

$$f_{{\rm Dis},t_{\rm DEO,1}(\delta),t_{\rm DEO,2}(\delta)} \in \arg\min_{f\in\mathcal{F}}\{R(f) : \max(|{\rm DO}(f)|, |{\rm PD}(f))| \leqslant \delta\}.$$

(4.iii) When $t_{\rm DEO,1} > 0$ and $t_{\rm DEO,2} > 0$, we have $t_{\rm DEO,1}(\delta) \cdot D_{\rm DO}(t_{\rm DEO,1}, t_{\rm DEO,2}) = t_{\rm DEO,1}(\delta) \cdot \delta \geqslant 0$ and $t_{\rm DEO,2}(\delta) \cdot D_{\rm PD}(t_{\rm DEO,1}, t_{\rm DEO,2}) = t_{\rm DEO,2}(\delta) \cdot \delta \geqslant 0$. By Lemma 32,

$$f_{{\rm Dis},t_{\rm DEO,1}(\delta),t_{\rm DEO,2}(\delta)} \in \arg\min_{f\in\mathcal{F}}\{R(f) : \max(|{\rm DO}(f)|, |{\rm PD}(f))| \leqslant \delta\}.$$

Thus, (18) holds in all three sub-cases.

Case (5) $D_{\rm DO}(0, \acute{t}_{\rm PD}(\delta)) \geqslant \delta$ and $D_{\rm PD}(\acute{t}_{\rm DO}(\delta), 0) \leqslant -\delta$.

In this case, $(t_{\rm DEO,1}(\delta), t_{\rm DEO,2}(\delta))$ is the solution of the equation $(D_{\rm DO}(t_1, t_2), D_{\rm PD}(t_1, t_2)) = (\delta, -\delta)$.

We have the follow five sub-cases: (5.i) $\acute{t}_{\rm DO}(\delta) > 0$ and $t_{\rm DEO,1}(\delta) \leqslant 0$; (5.ii) $\acute{t}_{\rm DO}(\delta) > 0$, $t_{\rm DEO,1}(\delta) > 0$ and $t_{\rm DEO,2}(\delta) \geqslant 0$; (5.iii) $\acute{t}_{\rm DO}(\delta) \leqslant 0$ and $t_{\rm DEO,1}(\delta) \leqslant 0$; (5.iv) $\acute{t}_{\rm DO}(\delta) \leqslant 0$, $t_{\rm DEO,1}(\delta) > 0$ and $t_{\rm DEO,2}(\delta) \geqslant 0$; (5.v) $t_{\rm DEO,1}(\delta) > 0$ and $t_{\rm DEO,2}(\delta) < 0$.

We consider these five sub-cases in turn.

(5.i) When $\acute{t}_{\rm DO}(\delta) > 0$ and $t_{\rm DEO,1}(\delta) \leqslant 0$, we have $D_{\rm DO}(0, 0) > \delta$ and ${\rm DO}(\acute{t}_{\rm DO}(\delta), 0) = \delta$. Since $t_{\rm DEO,1}(\delta) \leqslant 0 < \acute{t}_{\rm DO}(\delta)$, by the monotonicity properties of $t \mapsto D_{\rm DO}$ from Proposition 30 and the fact that $D_{\rm DO}(\acute{t}_{\rm DO}(\delta), 0) = \delta = D_{\rm DO}(t_{\rm DEO,1}(\delta), t_{\rm DEO,2}(\delta))$, we have $t_{\rm DEO,2}(\delta) \geqslant 0$. Then, from result (1) of Lemma 31, since $D_{\rm DO}(t_{\rm DEO,1}(\delta), t_{\rm DEO,2}(\delta)) = \delta = D_{\rm DO}(\acute{t}_{\rm DO}(\delta), 0)$, we have $D_{\rm PD}(t_{\rm DEO,1}(\delta), t_{\rm DEO,2}(\delta)) \leqslant D_{\rm PD}(\acute{t}_{\rm DO}(\delta), 0) \leqslant -\delta$. As $D_{\rm PD}(t_{\rm DEO,1}(\delta), t_{\rm DEO,2}(\delta)) = -\delta$, we have $D_{\rm PD}(t_{\rm DEO,1}(\delta), t_{\rm DEO,2}(\delta)) = D_{\rm PD}(\acute{t}_{\rm DO}(\delta), 0) = -\delta$, which indicates $f_{{\rm Dis},t_{\rm DEO,1}(\delta),t_{\rm DEO,2}(\delta)} = f_{{\rm Dis},\acute{t}_{\rm DO}(\delta),0}$ almost surely with respect to $\mathbb{P}_{X,A}$. On one hand, as $\acute{t}_{\rm DO}(\delta) > 0$, by Lemma 28 adapted for equality of opportunity, we have

$$\begin{aligned} R(f_{{\rm Dis},t_{\rm DEO,1}(\delta),t_{\rm DEO,2}(\delta)}) &= R(f_{{\rm Dis},\acute{t}_{\rm DO}(\delta),0}) \\ &= \min_{f\in\mathcal{F}}\{R(f) : {\rm DO}(f) \leqslant D_{\rm DO}(\acute{t}_{\rm DO}(\delta), 0) = \delta\}. \end{aligned}$$

On the other hand, since $(D_{\rm DO}(t_{\rm DEO,1}(\delta), t_{\rm DEO,2}(\delta)), D_{\rm PD}(t_{\rm DEO,1}(\delta), t_{\rm DEO,2}(\delta))) = (\delta, -\delta)$, we have

$$\begin{aligned} f_{{\rm Dis},t_{\rm DEO,1}(\delta),t_{\rm DEO,2}(\delta)} &\in \{f \in \mathcal{F} : \max(|{\rm DO}(f)|, |{\rm PD}(f))| \leqslant \delta\} \\ &\subset \{f \in \mathcal{F} : {\rm DO}(f) \leqslant \delta\}. \end{aligned}$$

It follows that

$$\begin{aligned} R(f_{{\rm Dis},t_{\rm DEO,1}(\delta),t_{\rm DEO,2}(\delta)}) &\geqslant \min_{f\in\mathcal{F}}\{R(f) : \max(|{\rm DO}(f)|, |{\rm PD}(f))| \leqslant \delta\} \\ &\geqslant \min_{f\in\mathcal{F}}\{R(f) : {\rm DO}(f) \leqslant \delta\}. \end{aligned}$$

Then, we can conclude that

$$f_{{\rm Dis},t_{\rm DEO,1}(\delta),t_{\rm DEO,2}(\delta)} \in \arg\min_{f\in\mathcal{F}}\{R(f) : \max(|{\rm DO}(f)|, |{\rm PD}(f))| \leqslant \delta\}.$$

(5.ii) When $\acute{t}_{\mathrm{DO}}(\delta) > 0$, $t_{\mathrm{DEO},1}(\delta) > 0$ and $t_{\mathrm{DEO},1}(\delta) \geqslant 0$, from the arguments for case (5.ii), we can show that $f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)} = f_{\mathrm{Dis},\acute{t}_{\mathrm{DO}}(\delta),0}$ almost surely with respect to $\mathbb{P}_{X,A}$ and

$$f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)} \in \arg\min_{f\in\mathcal{F}}\{R(f) : \max(|\mathrm{DO}(f)|, |\mathrm{PD}(f))| \leqslant \delta\}.$$

(5.iii) When $\acute{t}_{\mathrm{DO}}(\delta) \leqslant 0$ and $t_{\mathrm{DEO},1}(\delta) \leqslant 0$, by Proposition 30, we have $D_{\mathrm{PD}}(0,0) \leqslant D_{\mathrm{PD}}(\acute{t}_{\mathrm{DO}}(\delta),0) \leqslant -\delta$, which implies $\acute{t}_{\mathrm{PD}}(\delta) \leqslant 0$ and $D_{\mathrm{PD}}(0,\acute{t}_{\mathrm{PD}}(\delta)) = -\delta$. Then, from result (4) of Lemma 31, since $D_{\mathrm{PD}}(t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)) = -\delta = D_{\mathrm{PD}}(0,\acute{t}_{\mathrm{PD}}(\delta))$, we have $D_{\mathrm{DO}}(t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)) \geqslant D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta)) \geqslant \delta$. As $D_{\mathrm{DO}}(t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)) = \delta$, we have $D_{\mathrm{DO}}(t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)) = D_{\mathrm{DO}}(0,\acute{t}_{\mathrm{PD}}(\delta)) = \delta$, which indicates $f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)} = f_{\mathrm{Dis},0,\acute{t}_{\mathrm{PD}}(\delta)}$ almost surely with respect to $\mathbb{P}_{X,A}$. On one hand, as $\acute{t}_{\mathrm{PD}}(\delta) \leqslant 0$, by Lemma 28 adapted for predictive equality, we have

$$\begin{aligned} R(f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)}) &= R(f_{\mathrm{Dis},0,\acute{t}_{\mathrm{PD}}(\delta)}) \\ &= \min_{f\in\mathcal{F}}\{R(f) : \mathrm{PD}(f) \geqslant D_{\mathrm{PD}}(0,\acute{t}_{\mathrm{PD}}(\delta)) = -\delta\}. \end{aligned}$$

On the other hand, since $(D_{\mathrm{DO}}(t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)), D_{\mathrm{PD}}(t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta))) = (\delta,-\delta)$, we have

$$\begin{aligned} f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)} &\in \{f\in\mathcal{F} : \max(|\mathrm{DO}(f)|, |\mathrm{PD}(f))| \leqslant \delta\} \\ &\subset \{f\in\mathcal{F} : \mathrm{PD}(f) \geqslant -\delta\}. \end{aligned}$$

It follows that

$$\begin{aligned} R(f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)}) &\geqslant \min_{f\in\mathcal{F}}\{R(f) : \max(|\mathrm{DO}(f)|, |\mathrm{PD}(f))| \leqslant \delta\} \\ &\geqslant \min_{f\in\mathcal{F}}\{R(f) : \mathrm{PD}(f) \geqslant -\delta\}. \end{aligned}$$

Then, we can conclude that

$$f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)} \in \arg\min_{f\in\mathcal{F}}\{R(f) : \max(|\mathrm{DO}(f)|, |\mathrm{PD}(f))| \leqslant \delta\}.$$

(5.iv) When $\acute{t}_{\mathrm{DO}}(\delta) \leqslant 0$, $t_{\mathrm{DEO},1}(\delta) > 0$ and $t_{\mathrm{DEO},2}(\delta) \geqslant 0$, by the monotonicity properties of $t \mapsto D_{\mathrm{PD}}$ from Proposition 30 and the fact that $D_{\mathrm{PD}}(0,\acute{t}_{\mathrm{PD}}(\delta)) = -\delta = D_{\mathrm{PD}}(t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta))$, we have $t_{\mathrm{DEO},1}(\delta) \leqslant 0$. Then, following the same argument as in case (5.iii), we can conclude that $f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)} = f_{\mathrm{Dis},0,\acute{t}_{\mathrm{PD}}(\delta)}$ almost surely with respect to $\mathbb{P}_{X,A}$ and

$$f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)} \in \arg\min_{f\in\mathcal{F}}\{R(f) : \max(|\mathrm{DO}(f)|, |\mathrm{PD}(f))| \leqslant \delta\}.$$

(5.v) When $t_{\mathrm{DEO},1}(\delta) > 0$ and $t_{\mathrm{DEO},2}(\delta) < 0$, we have $t_{\mathrm{DEO},1}(\delta)\cdot D_{\mathrm{DO}}(t_{\mathrm{DEO},1},t_{\mathrm{DEO},2}) = t_{\mathrm{DEO},1}(\delta)\cdot\delta \geqslant 0$ and $t_{\mathrm{DEO},2}(\delta)\cdot D_{\mathrm{PD}}(t_{\mathrm{DEO},1},t_{\mathrm{DEO},2}) = t_{\mathrm{DEO},2}(\delta)\cdot(-\delta) \geqslant 0$. By Lemma 32, we conclude that

$$f_{\mathrm{Dis},t_{\mathrm{DEO},1}(\delta),t_{\mathrm{DEO},2}(\delta)} \in \arg\min_{f\in\mathcal{F}}\{R(f) : \max(|\mathrm{DO}(f)|, |\mathrm{PD}(f))| \leqslant \delta\}.$$

Thus, (18) holds in all five sub-cases.

The proof of Theorem 21 is complete. ■

B.2 Fair Bayes-optimal Classifier with a Multi-class Protected Attribute

In this section, we consider demographic parity with a multi-class protected attribute. We assume that $A \in \mathcal{A} = \{1, 2, ..., |\mathcal{A}|\}$ for some integer $|\mathcal{A}| > 2$. Our theoretical results concern perfect fairness. Under perfect fairness, we have $|\mathcal{A}|-1$ equality constraints, and the fair Bayes-optimal classifier can be derived using the generalized Neyman-Pearson lemma. However, for approximate fairness for a multi-class protected attribute, it is unknown ahead of time how many constraints become equalities, and how many are strict inequalities. A careful analysis of these two types of constraints is required, and we leave this to future work. Recall that, a classifier satisfies demographic parity if

$$\mathbb{P}(\widehat{Y}_f = 1 \mid A = a) = \mathbb{P}\left(\widehat{Y}_f = 1 \mid A = 1\right), \quad \text{for} \quad a = 2, ..., \mathcal{A}.$$

Similar to the expression of DD in (30), these constraints can be expressed as:

$$\int_{\mathcal{A}} \int_{\mathcal{X}} f(x, \tilde{a}) \frac{I(\tilde{a} = a)}{p_a} - \frac{I(\tilde{a} = 1)}{p_1} d\mathbb{P}_{X,A}(x, \tilde{a}) \quad \text{for} \quad a = 2, ..., \mathcal{A}.$$

Take $\phi_0(x, \tilde{a}) = 2\eta_{\tilde{a}}(x) - 1$ and $\phi_a(x, \tilde{a}) = I(\tilde{a} = a)/p_a - I(\tilde{a} = 1)/p_1$ for $a = 2, ..., |\mathcal{A}|$, the generalized Neyman-Pearson lemma suggests us to consider the group-wise thresholding rule, for all $(x, \tilde{a}) \in \mathcal{X} \times \{0, 1\}$,

$$f(x, \tilde{a}, t_2, \ldots, t_{\mathcal{A}}) = I\left(\phi_0(x, \tilde{a}) > \sum_{a=2}^{|\mathcal{A}|} t_a \phi_a(x, \tilde{a})\right).$$

This further leads to our formal Theorem 22 in Appendix B.

B.2.1 Proof of Theorem 22

We first demonstrate the existence of thresholds $\{t_{\mathrm{Dis},a}\}_{a=1}^{|\mathcal{A}|}$ that satisfy the condition required in Theorem 22.

Lemma 34 (Existence of Thresholds) *Suppose that, for all $a \in \mathcal{A}$, $\eta_a(X)$ has a density function on $\mathcal{X}$. Then, there exist $\{t_{\mathrm{Dis},a}\}_{a=1}^{|\mathcal{A}|}$ such that $\sum_{a=1}^{|\mathcal{A}|} t_{\mathrm{Dis},a} = 0$, and for all $a \in \{2, \ldots, |\mathcal{A}|\}$, (19) holds.*

Proof [Proof of Lemma 34] Define the functions $\underline{Q}_a, \overline{Q}_a : [0, 1] \to \mathbb{R}$, such that for $s \in [0, 1]$,

$$\underline{Q}_a(s) = \sup\left\{t : \mathbb{P}_{X|A=a}\left(\eta_a(X) > \frac{1}{2} + \frac{t}{p_a}\right) > s\right\};$$
$$\overline{Q}_a(s) = \sup\left\{t : \mathbb{P}_{X|A=a}\left(\eta_a(X) > \frac{1}{2} + \frac{t}{p_a}\right) \geqslant s\right\}.$$

By definition, we have the following:

- Both $\underline{Q}_a$ and $\overline{Q}_a$ are strictly monotonically decreasing, as $\eta_a(X)$ has a density function;
- $\underline{Q}_a$ is right-continuous and $\overline{Q}_a$ is left-continuous;
- For all $s_0 \in [0,1]$, $\lim_{s\to s_0^-} \underline{Q}_a(s) = \lim_{s\to s_0^-} \overline{Q}_a(s)$ and $\lim_{s\to s_0^+} \underline{Q}_a(s) = \lim_{s\to s_0^+} \overline{Q}_a(s)$ with $\underline{Q}_a(s_0) \leqslant \overline{Q}_a(s_0)$;
- For all $t \in [\underline{Q}_a(s), \overline{Q}_a(s)]$, $\mathbb{P}_{X|A=a}(\eta_a(X) > 1/2 + \frac{t}{2p_a}) = s$.

Now, we set

$$s^\star = \sup\left\{s : \sum_{a=1}^{|\mathcal{A}|} \underline{Q}_a(s) > 0\right\} = \sup\left\{s : \sum_{a=1}^{|\mathcal{A}|} \overline{Q}_a(s) > 0\right\}.$$

By the right-continuity of $\sum_{a=1}^{|\mathcal{A}|} \underline{Q}_a(s)$ and the left-continuity of $\sum_{a=1}^{|\mathcal{A}|} \overline{Q}_a(s)$, we have

$$\sum_{a=1}^{|\mathcal{A}|} \underline{Q}_a(s^\star) \leqslant 0 \leqslant \sum_{a=1}^{|\mathcal{A}|} \overline{Q}_a(s^\star).$$

Letting

$$t_{\mathrm{Dis},a} = \frac{\sum_{a=1}^{|\mathcal{A}|} \overline{Q}_a(s^\star)}{\sum_{a=1}^{|\mathcal{A}|} \overline{Q}_a(s^\star) - \sum_{a=1}^{|\mathcal{A}|} \underline{Q}_a(s^\star)} \underline{Q}_a(s^\star) - \frac{\sum_{a=1}^{|\mathcal{A}|} \underline{Q}_a(s^\star)}{\sum_{a=1}^{|\mathcal{A}|} \overline{Q}_a(s^\star) - \sum_{a=1}^{|\mathcal{A}|} \underline{Q}_a(s^\star)} \overline{Q}_a(s^\star),$$

we have $\sum_{a=1}^{|\mathcal{A}|} t_{\mathrm{Dis},a} = 0$. Moreover, as $t_{\mathrm{Dis},a} \in [\underline{Q}_a(s), \overline{Q}_a(s)]$ for all $a \in \mathcal{A}$, we have, for all $a \in \mathcal{A}$,

$$\mathbb{P}_{X|A=a}\left(\eta_a(X) > 1/2 + \frac{t_{\mathrm{Dis},a}}{2p_a}\right) = s^\star.$$

This completes the proof. ■

We now proceed to the proof of Theorem 22.

Proof [Proof of Theorem 22] Recall that, for $(x,\tilde{a}) \in \mathcal{X}\times\{0,1\}$, we denote, for $a = 2, ..., |\mathcal{A}|$, $\phi_0(x,\widetilde{a}) = 2\eta_{\tilde{a}}(x) - 1$; and $\phi_a(x,\widetilde{a}) = I(\tilde{a} = a)/p_a - I(\tilde{a} = 1)/p_1$. Since $t_{\mathrm{Dis},1} = -\sum_{a=2}^{|\mathcal{A}|} t_{\mathrm{Dis},a}$, We have, for all $x, \tilde{a}$,

$$f^\star_{t_{\mathrm{Dis},1},\dots,t_{\mathrm{Dis},|\mathcal{A}|}}(x,\tilde{a}) = I\left(\eta_{\tilde{a}}(x) > \frac{1}{2} + \frac{t_{\mathrm{Dis},\tilde{a}}}{2p_{\tilde{a}}}\right) = I\left(2\eta_{\tilde{a}}(x) - 1 > \frac{t_{\tilde{a}}}{p_{\tilde{a}}}\right)$$
$$= I\left(2\eta_{\tilde{a}}(x) - 1 > \sum_{a=2}^{|\mathcal{A}|} t_a \left(\frac{I(\tilde{a}=a)}{p_a} - \frac{I(\tilde{a}=1)}{p_1}\right)\right) = I\left(\phi_0(x,\tilde{a}) > \sum_{t=2}^{|\mathcal{A}|} t_a \phi_a(x,\tilde{a})\right).$$

Moreover, for any classifier f, we can write its misclassification rate and corresponding disparity measures $(\mathbb{P}_{X|A=a}(\widehat{Y}_f = 1) - \mathbb{P}_{X|A=1}(\widehat{Y}_f = 1))$ as

$$R(f) = \mathbb{P}(\widehat{Y}_f \neq Y) = -\int_{\mathcal{A}}\int_{\mathcal{X}} f(x,\widetilde{a})\phi_0(x,\widetilde{a}) d\mathbb{P}_{X,A}(x,\tilde{a}) + \int_{\mathcal{A}}\int_{\mathcal{X}} \eta_{\tilde{a}}(x) d\mathbb{P}_{X,A}(x,\tilde{a})$$

and, for $a = 2, ..., |\mathcal{A}|$,

$$\mathbb{P}_{X|A=a}(\widehat{Y}_f = 1) - \mathbb{P}_{X|A=1}(\widehat{Y}_f = 1) = \int_{\mathcal{A}}\int_{\mathcal{X}} \phi_a(x,\widetilde{a}) f(x,\tilde{a}) d\mathbb{P}_{X,A}(x,\widetilde{a}). \tag{44}$$

Denote

$$\mathcal{F}_{D,=} = \left\{ f : \int_{\mathcal{A}} \int_{\mathcal{X}} f(x, \tilde{a}) \left(\frac{I(\tilde{a} = a)}{p_a} - \frac{I(\tilde{a} = 1)}{p_1} \right) d\mathbb{P}_{X,A}(x, \tilde{a}) = 0, \quad a = 2, 3, ..., |\mathcal{A}| \right\},$$

By the construction of $\{t_{\mathrm{Dis},a}\}_{a=1}^{|\mathcal{A}|}$, we have $f^\star_{t_{\mathrm{Dis},1},\ldots,t_{\mathrm{Dis},|\mathcal{A}|}} \in \mathcal{F}_{D,=}$. Then, by the generalized Neyman Pearson lemma,

$$f^\star_{t_{\mathrm{Dis},1},\ldots,t_{\mathrm{Dis},|\mathcal{A}|}} \in \arg\max_{f \in \mathcal{F}_{D,=}} \left\{ \int_{\mathcal{A}} \int_{\mathcal{X}} f(x, \widetilde{a}) \phi_0(x, \widetilde{a}) d\mathbb{P}_{X,A}(x, \widetilde{a}) \right\} = \arg\min_{f \in \mathcal{F}_{D,=}} R(f),$$

which completes the proof. ■

B.3 FUDS, FCSC and FPIR Algorithms for Extensions

B.3.1 Algorithms for Protected Attributes Unavailable at Test Time

In this section, we provide explicit algorithms for the case when the protected attribute, A, is unavailable during prediction. We focus on the common disparity measures of demographic parity, equality of opportunity, and predictive equality. As outlined in Proposition 19, this setting leads to a fairness constraint which is linear, but not bilinear. Based on Corollary 20 and Proposition 19, the general form of a δ-fair Bayes-optimal classifier with linear disparity measures in this setting is given by:

$$f^\star_{X,\mathrm{Dis},\delta}(x) = I\left(\eta^Y(x) > \frac{1}{2} + \frac{t_{X,\mathrm{Dis}}(\delta)}{2} w_{X,\mathrm{Dis}}(x) \right), \tag{45}$$

with

$$t_{X,\mathrm{Dis}}(\delta) = \arg\min_t \left\{ |t| : \left| \int_{\mathcal{X}} \left[w_{X,\mathrm{Dis}}(x) \cdot I\left(\eta^Y(x) > \frac{1}{2} + \frac{t}{2} w_{X,\mathrm{Dis}}(x) \right) \right] d\mathbb{P}_X(x) \right| \leqslant \delta \right\}. \tag{46}$$

Moreover, the specific form of the linear weighting function $w_{X,\mathrm{Dis}}(x)$ for our disparity measures of demographic parity, equality of opportunity, and predictive equality, respectively, is

$$w_{\mathrm{DD}}(x) = \frac{\eta^A(x)}{p_1} - \frac{1 - \eta^A(x)}{p_0}; \qquad w_{\mathrm{DO}}(x) = \frac{\eta^Y_{A=1}(x)\eta^A(x)}{p_{1,1}} - \frac{\eta^Y_{A=0}(x)(1 - \eta^A(x))}{p_{0,1}};$$
$$w_{\mathrm{PD}}(x) = \frac{(1 - \eta^Y_{A=1}(x))\eta^A(x)}{p_{1,0}} - \frac{(1 - \eta^Y_{A=0}(x))(1 - \eta^A(x))}{p_{0,0}}.$$

In particular, denote the δ-fair Bayes-optimal classifier under demographic parity as $f^\star_{X,\mathrm{DD},\delta}(x)$. By setting $w_{X,\mathrm{Dis}}(x) = w_{\mathrm{DD}}(x)$ in (45) and (46), respectively, we find it is of the form

$$f^\star_{X,\mathrm{DD},\delta}(x) = I\left(\eta^Y(x) > \frac{1}{2} + \frac{t_{X,\mathrm{DD}}(\delta)(\eta^A(x) - p_1)}{2p_1p_0} \right),$$

with

$$t_{X,\mathrm{DD}}(\delta) = \arg\min_t \left\{ |t| : \left| \int_{\mathcal{X}} \left[\left(\frac{\eta^A(x) - p_1}{p_1 p_0} \right) \cdot I\left(\eta^Y(x) > \frac{1}{2} + \frac{t(\eta^A(x) - p_1)}{2p_1 p_0} \right) \right] d\mathbb{P}_X(x) \right| \leqslant \delta \right\}.$$

Similarly, define the δ-fair Bayes-optimal classifier under equality of opportunity, $f^{\star}_{X,\mathrm{DO},\delta}(x)$; and predictive equality, $f^{\star}_{X,\mathrm{PD},\delta}(x)$; with associated thresholding parameters $t_{X,\mathrm{DO}}(\delta)$ and $t_{X,\mathrm{PD}}(\delta)$ by setting $w_{X,\mathrm{Dis}}(x) = w_{\mathrm{DO}}(x)$ and $w_{X,\mathrm{Dis}}(x) = w_{\mathrm{PD}}(x)$, respectively.

The following two theorems, proved below, parallel Theorems 24 and 26, provide a theoretical underpinning for our Fair Up-/Down-Sampling and Fair cost-sensitive Classification methods in the attribute-blind setting.

Theorem 35 (Bayes-Optimal FUDS for Common Disparity Measures with an Unavailable Protected Attribute) *For $\delta \geqslant 0$, let $\widetilde{\mathbb{P}}^{\delta}_{\mathrm{DD}}$ be the distribution satisfying $\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=a,\widetilde{Y}=y}(x) = \mathbb{P}_{X|A=a,Y=y}(x)$ for all x, and for $(a,y) \in \{0,1\}^2$,*

$$\widetilde{p}^{\delta}_{\mathrm{DD},a,y} = \widetilde{\mathbb{P}}^{\delta}_{\mathrm{DD}}\left(\widetilde{A} = a, \widetilde{Y} = y\right) = c_{\mathrm{DD}}(\delta)\left(1 + \frac{(2a-1)(1-2y)t_{X,\mathrm{DD}}(\delta)}{p_a}\right) p_{a,y},$$

where $c_{\mathrm{DD}}(\delta) > 0$ is such that $\sum_{a\in\{0,1\}}\sum_{y\in\{0,1\}} \widetilde{p}^{\delta}_{\mathrm{DD},a,y} = 1$.

Similarly, let $\widetilde{\mathbb{P}}^{\delta}_{\mathrm{DO}}$ be the distribution satisfying $\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=a,\widetilde{Y}=y}(x) = \mathbb{P}_{X|A=a,Y=y}(x)$ for all x, and for $(a,y) \in \{0,1\}^2$,

$$\widetilde{p}^{\delta}_{\mathrm{DO},a,y} = \widetilde{\mathbb{P}}^{\delta}_{\mathrm{DO}}\left(\widetilde{A} = a, \widetilde{Y} = y\right) = c_{\mathrm{DO}}(\delta)\left(1 + \frac{(1-2a)y t_{X,\mathrm{DO}}(\delta)}{p_{a,y}}\right) p_{a,y},$$

where $c_{\mathrm{DO}}(\delta) > 0$ is such that $\sum_{a\in\{0,1\}}\sum_{y\in\{0,1\}} \widetilde{p}^{\delta}_{\mathrm{DO},a,y} = 1$.

Lastly, let $\widetilde{\mathbb{P}}^{\delta}_{\mathrm{PD}}$ be the distribution satisfying $\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=a,\widetilde{Y}=y}(x) = \mathbb{P}_{X|A=a,Y=y}(x)$ for all x, and for $(a,y) \in \{0,1\}^2$,

$$\widetilde{p}^{\delta}_{\mathrm{PD},a,y} = \widetilde{\mathbb{P}}^{\delta}_{\mathrm{PD}}\left(\widetilde{A} = a, \widetilde{Y} = y\right) = c_{\mathrm{PD}}(\delta)\left(1 + \frac{(2a-1)(1-y)t_{X,\mathrm{PD}}(\delta)}{p_{a,y}}\right) p_{a,y}, \tag{47}$$

where $c_{\mathrm{PD}}(\delta) > 0$ is such that $\sum_{a\in\{0,1\}}\sum_{y\in\{0,1\}} \widetilde{p}^{\delta}_{\mathrm{PD},a,y} = 1$.

Then, the unconstrained Bayes-optimal classifiers for $\widetilde{\mathbb{P}}^{\delta}_{\mathrm{DD}}, \widetilde{\mathbb{P}}^{\delta}_{\mathrm{DO}}$, and $\widetilde{\mathbb{P}}^{\delta}_{\mathrm{PD}}$ are δ-fair Bayes-optimal classifiers under demographic parity, equality of opportunity, and predictive equality, respectively, for $\mathbb{P}$.

Next consider the following fair cost-sensitive risks:

$$R^{\mathrm{FCSC}}_{t_{X,\mathrm{DD}}(\delta)}(f) = \sum_{a\in\{0,1\}}\sum_{y\in\{0,1\}} \left(\frac{1}{2} + \frac{(2a-1)(1-2y)t_{X,\mathrm{DD}}(\delta)}{2p_a}\right) \mathbb{P}(\widehat{Y}_f = 1-y, Y = y, A = a)$$

$$R^{\mathrm{FCSC}}_{t_{X,\mathrm{DO}}(\delta)}(f) = \sum_{a\in\{0,1\}}\sum_{y\in\{0,1\}}\left(\frac{1}{2}+\frac{(1-2a)yt_{X,\mathrm{DO}}(\delta)}{2p_{a,y}}\right)\mathbb{P}(\widehat{Y}_f = 1-y, Y=y, A=a)$$
$$R^{\mathrm{FCSC}}_{t_{X,\mathrm{PD}}(\delta)}(f) = \sum_{a\in\{0,1\}}\sum_{y\in\{0,1\}}\left(\frac{1}{2}+\frac{(2a-1)(1-y)t_{X,\mathrm{PD}}(\delta)}{2p_{a,y}}\right)\mathbb{P}(\widehat{Y}_f = 1-y, Y=y, A=a), \tag{48}$$

and define the cost-sensitive classifiers

$$f^{\mathrm{FCSC}}_{X,t_{X,\mathrm{DD}}(\delta)} \in \arg\min R^{\mathrm{FCSC}}_{X,t_{X,\mathrm{DD}}(\delta)}(f); \quad f^{\mathrm{FCSC}}_{X,t_{X,\mathrm{DO}}(\delta)} \in \arg\min R^{\mathrm{FCSC}}_{X,t_{X,\mathrm{DO}}(\delta)}(f);$$
$$f^{\mathrm{FCSC}}_{X,t_{X,\mathrm{PD}}(\delta)} \in \arg\min R^{\mathrm{FCSC}}_{X,t_{X,\mathrm{PD}}(\delta)}(f).$$

Theorem 36 (Bayes-Optimal FCSC for Common Disparity Measures with an Unavailable Protected Attribute) *For any $\delta \geqslant 0$, we have that $f^{X,\mathrm{FCSC}}_{t_{X,\mathrm{DD}}(\delta)}, f^{X,\mathrm{FCSC}}_{t_{X,\mathrm{DO}}(\delta)}$ and $f^{X,\mathrm{FCSC}}_{t_{X,\mathrm{PD}}(\delta)}$ are δ-fair Bayes-optimal classifiers under demographic parity, equality of opportunity, and predictive equality, respectively, for $\mathbb{P}$.*

Based on these two theorems, pre- and in-processing algorithms analogous to those from Section 5 can be designed for the case that A is not used at test time. For the FUDS and FCSC algorithms, the main difference is that now we must fit a classifier without using the protected attribute A as a predictor on the up- and down-sampled dataset S_t for FUDS; and on the original dataset with a cost-sensitive risk for FSCS.

For the FPIR algorithm, we construct a plug-in estimator $\widehat{w}_{X,\mathrm{Dis}}$ of the weighting function $w_{X,\mathrm{Dis}}$ by constructing estimates of $\eta^Y(x) = \mathbb{P}(Y=1|X=x)$ and additional plug-in estimators of η^A for demographic parity, and $\eta^Y_{A=0}, \eta^Y_{A=1}$ for equality of opportunity and predictive equality.

For all methods, we also now estimate the disparity level $\widehat{f}^{\mathrm{FUDS,no}A}_t$ as

$$\widehat{\mathrm{Dis}}^{\mathrm{FUDS,\ no}A}(t) = \frac{1}{n_1}\sum_{j=1}^{n_1}\widehat{f}^{\mathrm{FUDS,no}A}_t(x_{1,j})\widehat{w}_{X,\mathrm{Dis}}(x_{1,j}) - \frac{1}{n_0}\sum_{j=1}^{n_0}\widehat{f}^{\mathrm{FUDS,no}A}_t(x_{0,j})\widehat{w}_{X,\mathrm{Dis}}(x_{0,j}) \tag{49}$$

where $\widehat{w}_{X,\mathrm{Dis}}(x_{a,j}) = \widehat{w}_{\mathrm{Dis}}(x_{a,j},a)$ for $j\in[n_a]$ and $a\in\{0,1\}$.

Proof [Proof of Theorem 35] We begin by proving the result for demographic parity. By Bayes' theorem, we have for all x that

$$\begin{aligned}
&\eta^Y(x) = \mathbb{P}\left(Y=1|X=x\right) = \mathbb{P}\left(Y=1, A=1|X=x\right) + \mathbb{P}\left(Y=1,A=0|X=x\right)\\
&= \frac{p_{1,1}d\mathbb{P}_{X|A=1,Y=1}(x)}{d\mathbb{P}_X(x)} + \frac{p_{0,1}d\mathbb{P}_{X|A=0,Y=1}(x)}{d\mathbb{P}_X(x)}\\
&= \frac{p_{1,1}d\mathbb{P}_{X|A=1,Y=1}(x) + p_{0,1}d\mathbb{P}_{X|A=0,Y=1}(x)}{p_{1,1}d\mathbb{P}_{X|A=1,Y=1}(x) + p_{1,0}d\mathbb{P}_{X|A=1,Y=0}(x) + p_{0,1}d\mathbb{P}_{X|A=0,Y=1}(x) + p_{0,0}d\mathbb{P}_{X|A=0,Y=0}(x)}.
\end{aligned}$$

Denoting $\widetilde{\mathbb{P}}^\delta_{\mathrm{DD}}$ as $\widetilde{\mathbb{P}}^\delta$ and $\widetilde{p}^\delta_{\mathrm{DD},a,y}$ as $\widetilde{p}^\delta_{a,y}$ for $(a,y)\in\{0,1\}^2$, we similarly have for all x that

$$\widetilde{\eta}^{Y,\delta}(x) = \widetilde{\mathbb{P}}^\delta\left(\widetilde{Y}=1|\widetilde{X}=x\right) = \widetilde{\mathbb{P}}^\delta\left(\widetilde{Y}=1,\widetilde{A}=1|\widetilde{X}=x\right) + \widetilde{\mathbb{P}}^\delta\left(\widetilde{Y}=1,\widetilde{A}=0|\widetilde{X}=x\right)$$

Algorithm 4: Fair Up-/Down-Sampling (FUDS) for Common Disparity Measures with an Unavailable Protected Attribute

Input: Disparity measure either Demographic parity, Equality of opportunity, Predictive equality; Step size $\alpha > 0$; Disparity Level $\delta \geqslant 0$; Error tolerance level $\varepsilon > 0$; Dataset $S = S_1 \cup S_0$ with $S = \{x_i, a_i, y_i\}_{i=1}^{n}$, $S_1 = \{x_{1,i}, y_{1,i}\}_{i=1}^{n_1}$ and $S_0 = \{x_{0,i}, y_{0,i}\}_{i=1}^{n_0}$; $n_{a,y} = \#\{a_i = a, y_i = y\}$.

Disparity Estimation sub-routine: Construct $\widehat{f}_t^{\textbf{FUDS,no}A}$ and estimate Dis$(\widehat{f}_t^{\textbf{FUDS,no}A})$ for a given threshold parameter t:

1. For all a, y and for all a, estimate $p_{a,y}$ and p_a by their empirical probabilities

$$\widehat{p}_{a,y} = n_{a,y}/n; \qquad \widehat{p}_a = (n_{a,0} + n_{a,1})/n.$$

2. For all a, y and for the specified disparity measure, let $\widehat{\widetilde{p}}_{a,y}^{\,t}$ be as in (47) using the empirical probabilities $\widehat{p}_a$ and $\widehat{p}_{a,y}$.
3. Apply up- and down-sampling to generate a new dataset S_t with proportions $\widehat{\widetilde{p}}_{a,y}^{\,t}$ for all a, y.
4. Fit classifier $\widehat{f}_t^{\mathrm{FUDS,no}A}$ on the dataset S_t using any method that does not use the protected attribute A as a predictor
5. Estimate the disparity level of $\widehat{f}_t^{\mathrm{FUDS,no}A}$ as in (49).

Estimate the δ-fair Bayes-optimal classifier:

Run Disparity Estimation sub-routine with $t = 0$.
if $\left|\widehat{\mathrm{Dis}}^{\mathrm{FUDS,no}A}(0)\right| \leqslant \delta$ **then**
  $\widehat{t}_{\mathrm{Dis}}(\delta) = 0$.
else
  if $\widehat{\mathrm{Dis}}^{\mathrm{FUDS,no}A}(0) > \delta$ **then**
    $\delta' = \delta$; $t_{\min} = 0$, $t_{\max} = 1$.
  else
    $\delta = -\delta$; $t_{\min} = -1$, $t_{\max} = 0$.
  end
  while $t_{\max} - t_{\min} > \varepsilon$ **do**
    $t = (t_{\max} - t_{\min})/2$;
    Run Disparity Estimation sub-routine with current t.
    if $\widehat{\mathrm{Dis}}^{\mathrm{FUDS,no}A}(t) > \delta$ **then**
      $t_{\max} = t$
    else
      $t_{\min} = t$
    end
  end
  $\widehat{t}_{\mathrm{Dis}}(\delta) = t$.
end
Output: $\widehat{f}_{\mathrm{Dis},\delta}^{\mathrm{no}A} = \widehat{f}_{\widehat{t}_{\mathrm{Dis}}(\delta)}^{\mathrm{FUDS,no}A}$.

Algorithm 5: Fair cost-sensitive Classification (FCSC) for Common Disparity Measures with an Unavailable Protected Attribute

Input: Disparity measure either Demographic parity, Equality of opportunity, Predictive equality; Step size $\alpha > 0$; Disparity Level $\delta \geqslant 0$, Error tolerance level $\varepsilon > 0$; Dataset $S = S_1 \cup S_0$ with $S = \{x_i, a_i, y_i\}_{i=1}^{n}$, $S_1 = \{x_{1,i}, y_{1,i}\}_{i=1}^{n_1}$ and $S_0 = \{x_{0,i}, y_{0,i}\}_{i=1}^{n_0}$; $n_{a,y} = \#\{a_i = a, y_i = y\}$; Step size $\alpha \geqslant 0$.

Disparity Estimation sub-routine: Construct $\widehat{f}_t^{\mathbf{FCSC,no}A}$ and estimate $\mathbf{Dis}(\widehat{f}_t^{\mathbf{FCSC,no}A})$ for a given threshold parameter t:

1. For all a, y and for all a, estimate $p_{a,y}$ and p_a by their empirical probabilities

$$\widehat{p}_{a,y} = n_{a,y}/n; \qquad \widehat{p}_a = (n_{a,0} + n_{a,1})/n.$$

2. For all a, y and for the specified disparity measure, read off the cost-sensitive misclassification risk $\widehat{c}_{a,y}(t)$ from (48).
3. Use any cost-sensitive classification method to fit $\widehat{f}_t^{\mathrm{FCSC,no}A}(\cdot)$ on S.
4. Estimate the disparity level of $\widehat{f}_t^{\mathrm{FCSC,no}A}$ as in (49) with $\widehat{f}_t^{\mathrm{FCSC,no}A}$ instead of $\widehat{f}_t^{\mathrm{FUDS,no}A}$.

Estimate the δ-fair Bayes-optimal classifier:

Run Disparity Estimation sub-routine with $t = 0$. **if** $\left|\widehat{\mathrm{Dis}}^{\mathrm{FCSC}}(0)\right| \leqslant \delta$ **then**
  $\widehat{t}_{\mathrm{Dis}}(\delta) = 0$.
else
  if $\widehat{\mathrm{Dis}}^{\mathrm{FCSC}}(0) > \delta$ **then**
    $\delta' = \delta$; $t_{\min} = 0$, $t_{\max} = 1$.
  else
    $\delta = -\delta$; $t_{\min} = -1$, $t_{\max} = 0$.
  end
  while $t_{\max} - t_{\min} > \varepsilon$ **do**
    $t = (t_{\max} - t_{\min})/2$;
    Run Disparity Estimation sub-routine with current t.
    if $\widehat{\mathrm{Dis}}^{\mathrm{FCSC}}(t) > \delta$ **then**
      $t_{\max} = t$
    else
      $t_{\min} = t$
    end
  end
  $\widehat{t}_{\mathrm{Dis}}(\delta) = t$.
end

Output: $\widehat{f}_{\mathrm{Dis},\delta}^{\mathrm{no}A} = \widehat{f}_{\widehat{t}_{\mathrm{Dis}}(\delta)}^{\mathrm{FCSC,no}A}$.

Algorithm 6: Fair Plug-in Rule (FPIR) for Common Disparity Measures with an Unavailable Protected Attribute

Input: Disparity measure either Demographic Parity, Equality of Opportunity, Predictive Equality; Step size $\alpha > 0$; Disparity level $\delta \geqslant 0$; Error tolerance level $\varepsilon > 0$; Dataset $S = S_1 \cup S_0$ with $S = \{x_i, a_i, y_i\}_{i=1}^{n}$, $S_1 = \{x_{1,i}, y_{1,i}\}_{i=1}^{n_1}$ and $S_0 = \{x_{0,i}, y_{0,i}\}_{i=1}^{n_0}$; $n_{a,y} = \#\{a_i = a, y_i = y\}$.

Step 1: Construct estimates $\widehat{\eta}^Y, \widehat{\eta}^A$ of η^Y, η^A, respectively, using any approach. If the disparity measure is Equality of Opportunity or Predictive Equality, further construct estimates $\widehat{\eta}_1, \widehat{\eta}_0$ for $\eta^Y_{A=1}, \eta^Y_{A=0}$, respectively, using any approach.

Step 2: Estimate the the fair Pareto frontier or the δ-fair Bayes-optimal classifier:

Disparity Estimation sub-routine: Construct $\widehat{f}_t^{\mathbf{FPIR,no}A}$ and estimate $\mathbf{Dis}(\widehat{f}_t^{\mathbf{FPIR,no}A})$ with threshold parameter t:

1. For all a, y and for all a, estimate $p_{a,y}$ and p_a by their empirical probabilities

$$\widehat{p}_{a,y} = n_{a,y}/n; \qquad \widehat{p}_a = (n_{a,0} + n_{a,1})/n.$$

2. Let $\widehat{f}_t^{\mathrm{FPIR,no}A}$ be defined by $\widehat{f}_t^{\mathrm{FPIR,no}A}(x,a) = I(\widehat{\eta}^Y(x) > \frac{1}{2} + \frac{t}{2}\widehat{w}_{X,\mathrm{Dis}}(x))$ for all x, a
3. Evaluate the disparity level of $\widehat{f}_t^{\mathrm{FPIR,no}A}$ as in (49) with $\widehat{f}_t^{\mathrm{FPIR,no}A}$ instead of $\widehat{f}_t^{\mathrm{FUDS,no}A}$.

Estimate the δ-fair Bayes-optimal classifier:

Run Disparity Estimation sub-routine with $t = 0$.
if $\left|\widehat{\mathrm{Dis}}^{\mathrm{FPIR}}(0)\right| \leqslant \delta$ **then**
  $\widehat{t}_{\mathrm{Dis}}(\delta) = 0$.
else
  if $\widehat{\mathrm{Dis}}^{\mathrm{FPIR}}(0) > \delta$ **then**
    $\delta' = \delta$; $t_{\min} = 0$, $t_{\max} = 1$.
  else
    $\delta = -\delta$; $t_{\min} = -1$, $t_{\max} = 0$.
  end
  while $t_{\max} - t_{\min} > \varepsilon$ **do**
    $t = (t_{\max} - t_{\min})/2$;
    Run Disparity Estimation sub-routine with current t.
    if $\widehat{\mathrm{Dis}}^{\mathrm{FPIR}}(t) > \delta$ **then**
      $t_{\max} = t$
    else
      $t_{\min} = t$
    end
  end
  $\widehat{t}_{\mathrm{Dis}}(\delta) = t$.
end

Output: $\widehat{f}_{\mathrm{Dis}^{\mathrm{no}A},\delta} = \widehat{f}^{\mathrm{FPIR,no}A}_{\widehat{t}_{\mathrm{Dis}}(\delta)}$.

$$= \frac{\widetilde{p}^{\delta}_{1,1} d\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=1,\widetilde{Y}=1}(x) + \widetilde{p}^{\delta}_{0,1} d\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=0,\widetilde{Y}=1}(x)}{\widetilde{p}^{\delta}_{1,1} d\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=1,\widetilde{Y}=1}(x) + \widetilde{p}^{\delta}_{1,0} d\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=1,\widetilde{Y}=0}(x) + \widetilde{p}^{\delta}_{0,1} d\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=0,\widetilde{Y}=1}(x) + \widetilde{p}^{\delta}_{0,0} d\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=0,\widetilde{Y}=0}(x)}.$$

Moreover,

$$\begin{aligned}
&\eta^A(x) = \mathbb{P}\left(A=1|X=x\right) = \mathbb{P}\left(A=1, Y=1|X=x\right) + \mathbb{P}\left(A=1, Y=0|X=x\right)\\
&= \frac{p_{1,1} d\mathbb{P}_{X|A=1,Y=1}(x)}{d\mathbb{P}_X(x)} + \frac{p_{1,0} d\mathbb{P}_{X|A=1,Y=0}(x)}{d\mathbb{P}_X(x)}\\
&= \frac{p_{1,1} d\mathbb{P}_{X|A=1,Y=1}(x) + p_{1,0} d\mathbb{P}_{X|A=1,Y=0}(x)}{p_{1,1} d\mathbb{P}_{X|A=1,Y=1}(x) + p_{1,0} d\mathbb{P}_{X|A=1,Y=0}(x) + p_{0,1} d\mathbb{P}_{X|A=0,Y=1}(x) + p_{0,0} d\mathbb{P}_{X|A=0,Y=0}(x)}.
\end{aligned}$$

Then, by our construction, for all x,

$$\begin{aligned}
f^{\mathrm{FUDS}}_{X,\mathrm{DD},t_{X,\mathrm{DD}}(\delta)}(x) &= I\left(\widetilde{\eta}^{Y,\delta}(x) > \frac{1}{2}\right)\\
&= I\Bigg(\widetilde{p}^{\delta}_{1,1} d\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=1,\widetilde{Y}=1}(x) + \widetilde{p}^{\delta}_{0,1} d\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=0,\widetilde{Y}=1}(x)\\
&\qquad > \widetilde{p}^{\delta}_{1,0} d\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=1,\widetilde{Y}=0}(x) + \widetilde{p}^{\delta}_{0,0} d\widetilde{\mathbb{P}}^{\delta}_{\widetilde{X}|\widetilde{A}=0,\widetilde{Y}=0}(x)\Bigg).
\end{aligned} \tag{50}$$

This also equals

$$\begin{aligned}
I\Bigg(&\left(1 - \frac{t_{X,\mathrm{DD}}(\delta)}{p_1}\right) p_{1,1}\mathbb{P}_{X|A=1,Y=1}(x) + \left(1 + \frac{t_{X,\mathrm{DD}}(\delta)}{p_0}\right) p_{0,1}\mathbb{P}_{X|A=0,Y=1}(x)\\
&> \left(1 + \frac{t_{X,\mathrm{DD}}(\delta)}{p_1}\right) p_{1,0}\mathbb{P}_{X|A=1,Y=0}(x) + \left(1 - \frac{t_{X,\mathrm{DD}}(\delta)}{p_0}\right) p_{0,0}\mathbb{P}_{X|A=0,Y=0}(x)\Bigg).
\end{aligned} \tag{51}$$

Denoting $g_{a,y}(x) = p_{a,y} d\mathbb{P}_{X|A=a,Y=y}(x)$, (51) further equals

$$\begin{aligned}
&I\Bigg(g_{1,1}(x) + g_{0,1}(x) - g_{1,0}(x) - g_{0,0}(x)\\
&\qquad > \frac{t_{X,\mathrm{DD}}(\delta)}{p_1}(g_{1,1}(x) + g_{1,0}(x)) - \frac{t_{X,\mathrm{DD}}(\delta)}{p_0}(g_{0,1}(x) + g_{0,0}(x))\Bigg)\\
=&I\Bigg(2(g_{1,1}(x) + g_{0,1}(x)) > \left(\frac{t_{X,\mathrm{DD}}(\delta)}{p_1} + \frac{t_{X,\mathrm{DD}}(\delta)}{p_0}\right)(g_{1,1}(x) + g_{1,0}(x))\\
&\qquad + \left(1 - \frac{t_{X,\mathrm{DD}}(\delta)}{p_0}\right)(g_{1,1}(x) + g_{1,0}(x) + g_{0,1}(x) + g_{0,0}(x))\Bigg).
\end{aligned}$$

Finally, this equals

$$I\Bigg(\frac{2(g_{1,1}(x) + g_{0,1}(x))}{g_{1,1}(x) + g_{0,1}(x) + g_{1,0}(x) + g_{0,0}(x)}$$

$$> \frac{t_{X,\mathrm{DD}}(\delta)(g_{1,1}(x)+g_{1,0}(x))}{p_1p_0\left(g_{1,1}(x)+g_{0,1}(x)+g_{1,0}(x)+g_{0,0}(x)\right)}+1-\frac{t_{X,\mathrm{DD}}(\delta)}{p_0}\Bigg)$$
$$=I\left(2\eta^Y(x)>\frac{t_{X,\mathrm{DD}}(\delta)}{p_1p_0}\eta^A(x)+1-\frac{t_{X,\mathrm{DD}}(\delta)}{p_0}\right)$$
$$=I\left(\eta^Y(x)>\frac{1}{2}+\frac{t_{X,\mathrm{DD}(\delta)}\left(\eta^A(x)-p_1\right)}{2p_1p_0}\right)=f^\star_{X,\mathrm{DD},\delta}(x).$$

This completes the proof for demographic parity. We show the result for equality of opportunity, as the proof for predictive equality is analogous to this setting. Denote $\widetilde{\mathbb{P}}^\delta_{\mathrm{DO}}$ as $\widetilde{\mathbb{P}}^\delta$ and $\widetilde{p}^\delta_{\mathrm{DO},a,y}$ as $\widetilde{p}^\delta_{a,y}$ for $(a,y)\in\{0,1\}^2$. Then by the exact same derivation for Bayes' rule, we arrive at (50). We then have,

$$f^{\mathrm{FUDS}}_{X,\mathrm{DO},t_{X,\mathrm{DO}}(\delta)}(x)$$
$$= I\Bigg(\widetilde{p}^\delta_{1,1}d\widetilde{\mathbb{P}}^\delta_{\widetilde{X}|\widetilde{A}=1,\widetilde{Y}=1}(x)+\widetilde{p}^\delta_{0,1}d\widetilde{\mathbb{P}}^\delta_{\widetilde{X}|\widetilde{A}=0,\widetilde{Y}=1}(x)$$
$$> \widetilde{p}^\delta_{1,0}d\widetilde{\mathbb{P}}^\delta_{\widetilde{X}|\widetilde{A}=1,\widetilde{Y}=0}(x)+\widetilde{p}^\delta_{0,0}d\widetilde{\mathbb{P}}^\delta_{\widetilde{X}|\widetilde{A}=0,\widetilde{Y}=0}(x)\Bigg)$$
$$= I\Bigg(\left(1-\frac{t_{X,\mathrm{DO}}(\delta)}{p_{1,1}}\right)p_{1,1}d\mathbb{P}_{X|A=1,Y=1}(x)+\left(1+\frac{t_{X,\mathrm{DO}}(\delta)}{p_{0,1}}\right)p_{0,1}d\mathbb{P}_{X|A=0,Y=1}(x)$$
$$> p_{1,0}d\mathbb{P}_{X|A=1,Y=0}(x)+p_{0,0}d\mathbb{P}_{X|A=0,Y=0}(x)\Bigg).$$

This also equals

$$I\left(\left(1-\frac{t_{X,\mathrm{DO}}(\delta)}{p_{1,1}}\right)g_{1,1}(x)+\left(1+\frac{t_{X,\mathrm{DO}}(\delta)}{p_{0,1}}\right)g_{0,1}(x)>g_{1,0}(x)+g_{0,0}(x)\right)$$
$$= I\left(\frac{1}{2}(g_{1,1}(x)+g_{0,1}(x))>\frac{1}{2}(g_{1,0}(x)+g_{0,0}(x))+\frac{1}{2}t_{X,\mathrm{DO}}(\delta)\left(\frac{g_{1,1}(x)}{p_{1,1}}-\frac{g_{0,1}(x)}{p_{0,1}}\right)\right)$$

Finally, this equals

$$= I\Bigg(\frac{g_{1,1}(x)+g_{0,1}(x)}{g_{1,1}(x)+g_{0,1}(x)+g_{1,0}(x)+g_{0,0}(x)}>\frac{1}{2}+\frac{t_{X,\mathrm{DO}}(\delta)}{2}\cdot$$
$$\left(\frac{g_{1,1}(x)}{p_{1,1}\left(g_{1,1}(x)+g_{0,1}(x)+g_{1,0}(x)+g_{0,0}(x)\right)}-\frac{g_{0,1}(x)}{p_{0,1}\left(g_{1,1}(x)+g_{0,1}(x)+g_{1,0}(x)+g_{0,0}(x)\right)}\right)\Bigg)$$
$$= I\left(\eta^Y(x)>\frac{1}{2}+\frac{t_{X,\mathrm{DO}}(\delta)}{2}\cdot\left(\frac{\eta^Y_{A=1}(x)\eta^A(x)}{p_{1,1}}-\frac{\eta^Y_{A=0}(x)(1-\eta^A(x))}{p_{0,1}}\right)\right).$$

This finishes the proof. ■

Proof [Proof of Theorem 36] We begin by proving the result for demographic parity. Denoting for all a,y, $c_{a,y}(\delta)=1/2+(2a-1)(1-2y)t_{X,\mathrm{DD}}(\delta)/(2p_a)$, we have,

$$R^{\mathrm{FCSC}}_{X,t_{X,\mathrm{DD}}(\delta)}(f)$$

$$
\begin{aligned}
&= \sum_{a\in\{0,1\}} \Big[c_{a,0}(\delta)\cdot \mathbb{P}(\widehat{Y}_f = 1, Y=0, A=a) + c_{a,1}(\delta)\cdot \mathbb{P}(\widehat{Y}_f = 0, Y=1, A=a)\Big] \\
&= \sum_{a\in\{0,1\}} p_a \Big[c_{a,0}(\delta)\cdot \mathbb{P}(\widehat{Y}_f = 1, Y=0|A=a) + c_{a,1}(\delta)\cdot \mathbb{P}(\widehat{Y}_f = 0, Y=1|A=a)\Big] \\
&= \sum_{a\in\{0,1\}} p_a c_{a,0}(\delta)\cdot \int_{\mathcal{X}} \mathbb{P}(\widehat{Y}_f = 1, Y=0|A=a, X=x) d\mathbb{P}_{X|A=a}(x) \\
&\quad + \sum_{a\in\{0,1\}} p_a c_{a,1}(\delta)\cdot \int_{\mathcal{X}} \mathbb{P}(\widehat{Y}_f = 0, Y=1|X=x, A=a) d\mathbb{P}_{X|A=a}(x).
\end{aligned}
$$

This further equals

$$
\begin{aligned}
&\sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left[c_{a,0}(\delta) f(x)(1-\eta^Y_{A=a}(x)) + c_{a,1}(\delta)(1-f(x))\eta^Y_{A=a}(x)\ \right] d\mathbb{P}_{X|A=a}(x) \\
&= \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left[c_{a,0}(\delta) f(x) + c_{a,1}(\delta)\eta^Y_{A=a}(x) - [c_{a,1}(\delta)+c_{a,0}(\delta)] f(x)\eta^Y_{A=a}(x)\right] d\mathbb{P}_{X|A=a}(x) \\
&= \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left(c_{a,0}(\delta) - (c_{a,1}(\delta)+c_{a,0}(\delta))\eta^Y_{A=a}(x)\right) f(x) d\mathbb{P}_{X|A=a}(x) \\
&\quad + \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} c_{a,1}(\delta)\eta^Y_{A=a}(x) d\mathbb{P}_{X|A=a}(x) \qquad (52) \\
&= \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left(c_{a,0}(\delta) - \eta^Y_{A=a}(x)\right) f(x) d\mathbb{P}_{X|A=a}(x) \\
&\quad + \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} c_{a,1}(\delta)\eta^Y_{A=a}(x) d\mathbb{P}_{X|A=a}(x). \qquad (53)
\end{aligned}
$$

This also equals

$$
\begin{aligned}
&\int_{\mathcal{X}} \left(\sum_{a\in\{0,1\}} p_a \left(c_{a,0}(\delta) - \eta^Y_{A=a}(x)\right) \frac{d\mathbb{P}_{X|A=a}(x)}{d\mathbb{P}_X(x)}\right) f(x) d\mathbb{P}_X(x) \\
&\quad + \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} c_{a,1}(\delta)\eta^Y_{A=a}(x) d\mathbb{P}_{X|A=a}(x) \\
&= \int_{\mathcal{X}} \left(\sum_{a\in\{0,1\}} \left(c_{a,0}(\delta) - \eta^Y_{A=a}(x)\right) \mathbb{P}(A=a|X=x)\right) f(x) d\mathbb{P}_X(x) \\
&\quad + \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} c_{a,1}(\delta)\eta^Y_{A=a}(x) d\mathbb{P}_{X|A=a}(x).
\end{aligned}
$$

The second term does not depend on f and the first term can be expressed as:

$$
= \int_{\mathcal{X}} \sum_{a\in\{0,1\}} \left[\left(c_{a,0}(\delta) - \eta^Y_{A=a}(x)\right) \mathbb{P}(A=a|X=x)\right] f(x) d\mathbb{P}_X(x)
$$

$$= \int_{\mathcal{X}} \sum_{a\in\{0,1\}} \left[(c_{a,0}(\delta)\mathbb{P}(A=a|X=x) - \mathbb{P}(Y=1, A=a|X=x))\right] f(x) d\mathbb{P}_X(x) \tag{54}$$
$$= \int_{\mathcal{X}} \left[\left(c_{1,0}(\delta)\eta^A(x) + c_{0,0}(\delta)(1-\eta^A(x))\right) - \eta^Y(x)\right] f(x) d\mathbb{P}_X(x)$$

In turn, this equals

$$= \int_{\mathcal{X}} \left[\left(\left(\frac{1}{2} + \frac{t_{X,\mathrm{DD}}(\delta)}{2p_1}\right)\eta^A(x) + \left(\frac{1}{2} - \frac{t_{X,\mathrm{DD}}(\delta)}{2p_0}\right)(1-\eta^A(x))\right) - \eta^Y(x)\right] f(x) d\mathbb{P}_X(x)$$
$$= \int_{\mathcal{X}} \left[\frac{1}{2} + \frac{t_{X,\mathrm{DD}}(\delta)\left(\eta^A(x) - p_1\right)}{2p_1p_0} - \eta^Y(x)\right] f(x) d\mathbb{P}_X(x).$$

Clearly, this quantity is minimized by taking

$$f(x) = \begin{cases} 1, & \eta^Y(x) > \frac{1}{2} + \frac{t_{\mathrm{X,DD}(\delta)}\left(\eta^A(x)-p_1\right)}{2p_1p_0}; \\ 0, & \eta^Y(x) \leqslant \frac{1}{2} + \frac{t_{X,\mathrm{DD}}\left(\eta^A(x)-p_1\right)}{2p_1p_0}. \end{cases}$$

We can thus conclude that for all x, a,

$$f^{\mathrm{FCSC}}_{X,t_{X,\mathrm{DD}}(\delta)}(x) = I\left(\eta^Y(x) > \frac{1}{2} + \frac{t_{X,\mathrm{DD}}\left(\eta^A(x) - p_1\right)}{2p_1p_0}\right) = f^{\star}_{X,\mathrm{DD},\delta}(x).$$

This finishes the proof for demographic parity. We show the result for equality of opportunity, as the proof for predictive equality is analogous to this setting. Denoting for all a, y, $c_{a,y}(\delta) = 1/2 + (1-2a)yt_{X,\mathrm{DO}}(\delta)/(2p_{a,y})$, then by a similar derivation leading up to (54) where the factor of $c_{a,1}(\delta) + c_{a,0}(\delta)$ no longer necessarily sums to 1 in (52), we have minimizing $R^{\mathrm{FCSC}}_{X,t_{X,\mathrm{DO}}(\delta)}(f)$ is equivalent to minimizing

$$\int_{\mathcal{X}} \sum_{a\in\{0,1\}} \left[(c_{a,0}(\delta)\mathbb{P}(A=a|X=x) - (c_{a,1}(\delta)+c_{a,0}(\delta))\mathbb{P}(Y=1, A=a|X=x))\right] f(x) d\mathbb{P}_X(x)$$
$$= \int_{\mathcal{X}} \left[\frac{1}{2} - \sum_{a\in\{0,1\}} \left[1 + \left(\frac{1-2a}{2p_{a,1}}\right) t_{X,\mathrm{DO}}(\delta)\mathbb{P}(Y=1, A=a|X=x)\right]\right] f(x) d\mathbb{P}_X(x)$$
$$= \int_{\mathcal{X}} \left[\frac{1}{2} - \eta^Y(x) - \sum_{a\in\{0,1\}} \left[\left(\frac{1-2a}{2p_{a,1}}\right) t_{X,\mathrm{DO}}(\delta)\mathbb{P}(Y=1, A=a|X=x)\right]\right] f(x) d\mathbb{P}_X(x)$$
$$= \int_{\mathcal{X}} \left[\frac{1}{2} - \eta^Y(x) + \frac{t_{X,\mathrm{DO}}(\delta)}{2}\left(\frac{\eta^Y_{A=1}(x)\eta^A(x)}{p_{1,1}} - \frac{\eta^Y_{A=0}(x)(1-\eta^A(x))}{p_{0,1}}\right)\right] f(x) d\mathbb{P}_X(x).$$

This quantity is minimized by taking

$$f(x) = \begin{cases} 1, & \eta^Y(x) > \frac{1}{2} + \frac{t_{X,\mathrm{DO}}(\delta)}{2}\left(\frac{\eta^Y_{A=1}(x)\eta^A(x)}{p_{1,1}} - \frac{\eta^Y_{A=0}(x)(1-\eta^A(x))}{p_{0,1}}\right); \\ 0, & \eta^Y(x) \leq \frac{1}{2} + \frac{t_{X,\mathrm{DO}}(\delta)}{2}\left(\frac{\eta^Y_{A=1}(x)\eta^A(x)}{p_{1,1}} - \frac{\eta^Y_{A=0}(x)(1-\eta^A(x))}{p_{0,1}}\right). \end{cases}$$

We thus conclude that for all x, a,

$$f^{\text{FCSC}}_{X,t_{X,\text{DO}}(\delta)}(x) = I\left(\eta^Y(x) > \frac{1}{2} + \frac{t_{X,\text{DO}}(\delta)}{2} \cdot \left(\frac{\eta^Y_{A=1}(x)\eta^A(x)}{p_{1,1}} - \frac{\eta^Y_{A=0}(x)(1-\eta^A(x))}{p_{0,1}}\right)\right)$$
$$= f^{\star}_{X,\text{DO},\delta}(x).$$

This finishes the proof. ■

B.4 Form of Fair Bayes-optimal Classifiers Without Distributional Assumptions

In our main text, we assume that, for $a \in \{0,1\}$, $\eta_a(X)$ has a probability density function on $\mathcal{X}$, In this section, we consider the more general case with no distributional assumptions on features. Without this assumption, the boundary case (where the features are exactly on the "fair" boundary) does not necessarily have zero probability. In this case, the optimal classifiers must be carefully randomized.

To handle this technical difficulty, we consider the randomized classifier. Let $t \in \mathbb{R}$ and $\tau : \mathcal{X} \times \mathcal{A} \to [0,1]$ be a measurable function. For $x \in \mathcal{X}$ and $a \in \{0,1\}$, we define $f_{\text{Dis},t,\tau}$ as,

$$f_{\text{Dis},t,\tau}(x,a) = I\left(\eta_a(x) > \frac{1}{2} + \frac{t}{2}w_{\text{Dis}}(x,a)\right) + \tau(x,a) I\left(\eta_a(x) = \frac{1}{2} + \frac{t}{2}w_{\text{Dis}}(x,a)\right). \quad (55)$$

We further define the functions $D_{\text{Dis},\min} : \mathbb{R} \to \mathbb{R}$ and $D_{\text{Dis},\max} : \mathbb{R} \to \mathbb{R}$ to measure the minimal and maximal disparities of $f_{\text{Dis},t,\tau}(x,a)$, such that for all $t \in \mathbb{R}$ and $\tau : \mathcal{X} \times \{0,1\} \to [0,1]$:

$$D_{\text{Dis},\min}(t) = \min_{\tau} \{\text{Dis}(f_{\text{Dis},t,\tau})\} \quad \text{and} \quad D_{\text{Dis},\max}(t) = \max_{\tau} \{\text{Dis}(f_{\text{Dis},t,\tau})\}.$$

We note that $D_{\text{Dis},\min}(0)$ and $D_{\text{Dis},\max}(0)$ are, respectively, the infimum and supremum of the disparity over all unconstrained Bayes-optimal classifiers. Moreover, $D_{\text{Dis},\min}$ and $D_{\text{Dis},\max}$ satisfy the following monotonicity properties.

Proposition 37 (Monotonicity of $D_{\text{Dis},\min}$ and $D_{\text{Dis},\max}$) *As functions of t, both $D_{\text{Dis},\min}$ and $D_{\text{Dis},\max}$ are monotone non-increasing.*

The proofs of all results are presented later in this section. For any $\delta > 0$, define the following quantity, which can be viewed as an "inverse" of the disparity functions:

$$t_{\text{Dis}}(\delta) = \arg\min_t \left\{|t| : \inf_{\tau} |\text{Dis}(f_{\text{Dis},t,\tau})| \leqslant \delta\right\}$$
$$= \begin{cases} \inf\{t : D_{\text{Dis},\min}(t) \leqslant \delta\}, & D_{\text{Dis},\min}(t) > \delta; \\ \sup\{t : D_{\text{Dis},\max}(0) \geqslant -\delta\}, & D_{\text{Dis},\max}(0) < -\delta; \\ 0, & \text{otherwise}. \end{cases} \quad (56)$$

Again, we want to find the t with minimal $|t|$ such that, for some τ, $f_{\text{Dis},t,\tau}$ satisfies the fairness constraint. Here, we identify three different cases: (1) $\delta \geqslant \max(D_{\text{Dis},\min}(t), -D_{\text{Dis},\max}(t))$, (2) $\delta < D_{\text{Dis},\min}(t)$, and (3) $\delta < -D_{\text{Dis},\max}(t)$. These three cases are disjoint and cover all possible scenarios. Clearly, either case (2) or case (3) is true when case (1) does not hold. Furthermore, since $\delta \geqslant 0$ and $D_{\text{Dis},\min}(t) \leqslant D_{\text{Dis},\max}(t)$ by definition, cases (2) and (3) cannot occur simultaneously. We obtain the following result.

Theorem 38 (Fair Bayes-optimal Classifiers Without Distributional Assumptions) *For any $\delta \geqslant 0$, there is a measurable function $\tau_{\mathrm{Dis},\delta} : \mathcal{X} \times \{0,1\} \to [0,1]$, such that $f^\star_{\mathrm{Dis},\delta} = f_{t_{\mathrm{Dis}}(\delta),\tau_{\mathrm{Dis},\delta}}$ is a δ-fair Bayes-optimal classifier. Here, $\tau_{\mathrm{Dis},\delta}$ is determined to satisfy the following constraints:*

(1). When $D_{\mathrm{Dis,min}}(t) > \delta$,

$$\mathrm{Dis}(f_{\mathrm{Dis},t_{\mathrm{Dis}}(\delta),\tau_{\mathrm{Dis},\delta}}) = \delta.$$

(2). When $D_{\mathrm{Dis,max}}(t) < -\delta$,

$$\mathrm{Dis}(f_{\mathrm{Dis},t_{\mathrm{Dis}}(\delta),\tau_{\mathrm{Dis},\delta}}) = -\delta.$$

(3). When $\delta \geqslant \max(D_{\mathrm{Dis,min}}(t), -D_{\mathrm{Dis,max}}(t))$,

$$|\mathrm{Dis}(f_{\mathrm{Dis},t_{\mathrm{Dis}}(\delta),\tau_{\mathrm{Dis},\delta}})| \leqslant \delta.$$

To specify a particular form for $\tau_{\mathrm{Dis},\delta}$ in Theorem 38, we first introduce the following proposition, which offers a more comprehensive understanding of $D_{\mathrm{Dis,min}}$ and $D_{\mathrm{Dis,max}}$. Recall that for any function $f : t \mapsto f(t)$, we denote its left-hand limit and right-hand limit at point $t \in \mathbb{R}$ as $\lim_{t' \to t^-} f(t')$ and $\lim_{t' \to t^+} f(t')$, respectively.

Proposition 39 *As functions of t,*

(1) For $G_{a,+}$ and $G_{a,-}$ in (57), we have, for all $t \in \mathbb{R}$, and all x, a

$$\tau_{\mathrm{Dis,min}}(x,a) := I((x,a) \in G_{a,-}) \in \arg\min_{\tau} \{\mathrm{Dis}(f_{\mathrm{Dis},t,\tau})\};$$
$$\tau_{\mathrm{Dis,max}}(x,a) := I((x,a) \in G_{a,+}) \in \arg\max_{\tau} \{\mathrm{Dis}(f_{\mathrm{Dis},t,\tau})\}.$$

Moreover,

$$\begin{aligned}
D_{\mathrm{Dis,min}}(t) = &\sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} > t \right) \right] d\mathbb{P}_{X|A=a}(x) \\
&+ \sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t \right) \right] d\mathbb{P}_{X|A=a}(x); \\
D_{\mathrm{Dis,max}}(t) = &\sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \geqslant t \right) \right] d\mathbb{P}_{X|A=a}(x) \\
&+ \sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t \right) \right] d\mathbb{P}_{X|A=a}(x).
\end{aligned}$$

(2) We have that $D_{\mathrm{Dis,min}}$ is right-continuous and $D_{\mathrm{Dis,min}}$ is left-continuous.

(3) We have, for $t \in \mathbb{R}$,

$$\lim_{t' \to t^-} D_{\mathrm{Dis,min}}(t') = D_{\mathrm{Dis,max}}(t).$$

Now, we provide a specific choice of $\tau_{\mathrm{Dis},\delta}$. Let τ_0 be the identically zero function with $\tau_0(x,a) := 0$, for all x, a, and for $t \in \mathbb{R}$, $D_{\mathrm{Dis},0}(t) = \mathrm{Dis}(f_{\mathrm{Dis},t,\tau_0})$. Clearly, for all t, $D_{\mathrm{Dis},\min}(t) \leqslant D_{\mathrm{Dis},0}(t) \leqslant D_{\mathrm{Dis},\max}(t)$. We consider three cases in order.

(1). When $D_{\mathrm{Dis},\min}(t) > \delta$, we have $t_{\mathrm{Dis}}(\delta) = \inf\{t : D_{\mathrm{Dis},\min}(t) \leqslant \delta\}$. Since $D_{\mathrm{Dis},\min}$ is right-continuous, we have $D_{\mathrm{Dis},\min}(t) \leqslant \delta \leqslant \lim_{t' \to t^-} D_{\mathrm{Dis},\min}(t') = D_{\mathrm{Dis},\max}(t')$. We can take, for all $x \in \mathcal{X}$ and $a \in \{0,1\}$,

$$\tau_{\mathrm{Dis},\delta}(x,a) = \begin{cases} \frac{[D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta))-\delta]\cdot I((x,a)\in G_{a,-})}{D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta))-D_{\mathrm{Dis},\min}(t_{\mathrm{Dis}}(\delta))}, & D_{\mathrm{Dis},\min}(t_{\mathrm{Dis}}(\delta)) \leqslant \delta \leqslant D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta)); \\ \frac{[\delta-D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta))]\cdot I((x,a)\in G_{a,+})}{D_{\mathrm{Dis},\max}(t_{\mathrm{Dis}}(\delta))-D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta))}, & D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta)) < \delta \leqslant D_{\mathrm{Dis},\max}(t_{\mathrm{Dis}}(\delta)). \end{cases}$$

(2). When $D_{\mathrm{Dis},\max}(t) < -\delta$, we have $t_{\mathrm{Dis}}(\delta) = \sup\{t : D_{\mathrm{Dis},\max}(t) \geqslant -\delta\}$. Since $D_{\mathrm{Dis},\max}$ is left-continuous, we have $D_{\mathrm{Dis},\min}(t) = \lim_{t' \to t^+} D_{\mathrm{Dis},\max}(t') \leqslant -\delta \leqslant D_{\mathrm{Dis},\max}(t)$. We can take, for all $x \in \mathcal{X}$ and $a \in \{0,1\}$,

$$\tau_{\mathrm{Dis},\delta}(x,a) = \begin{cases} \frac{[D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta))+\delta]\cdot I((x,a)\in G_{a,-})}{D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta))-D_{\mathrm{Dis},\min}(t_{\mathrm{Dis}}(\delta))}, & D_{\mathrm{Dis},\min}(t_{\mathrm{Dis}}(\delta)) \leqslant -\delta \leqslant D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta)); \\ \frac{[-\delta-D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta))]\cdot I((x,a)\in G_{a,+})}{D_{\mathrm{Dis},\max}(t_{\mathrm{Dis}}(\delta))-D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta))}, & D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta)) < -\delta \leqslant D_{\mathrm{Dis},\max}(t_{\mathrm{Dis}}(\delta)). \end{cases}$$

(3). When $\delta \geqslant \max(D_{\mathrm{Dis},\min}(t), -D_{\mathrm{Dis},\max}(t))$, since $D_{\mathrm{Dis},0} \leqslant D_{\mathrm{Dis},0} \leqslant D_{\mathrm{Dis},0}$, we have $D_{\mathrm{Dis},\min} \leqslant \delta$ when $\delta < D_{\mathrm{Dis},0}$ and $D_{\mathrm{Dis},\max} \geqslant -\delta$ when $\delta < -D_{\mathrm{Dis},0}$. Thus, we can take, for all $x \in \mathcal{X}$ and $a \in \{0,1\}$,

$$\tau_{\mathrm{Dis},\delta}(x,a) = \begin{cases} \frac{[D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta))-\delta]\cdot I((x,a)\in G_{a,-})}{D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta))-D_{\mathrm{Dis},\min}(t_{\mathrm{Dis}}(\delta))}, & \delta < D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta)); \\ \frac{[-\delta-D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta))]\cdot I((x,a)\in G_{a,+})}{D_{\mathrm{Dis},\max}(t_{\mathrm{Dis}}(\delta))-D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta))}, & \delta < -D_{\mathrm{Dis},0}(t_{\mathrm{Dis}}(\delta)); \\ 0; & \text{Otherwise.} \end{cases}$$

We now present the proofs of the results discussed in this section.
Proof [Proof of Proposition 37] We define, for $a \in \{0,1\}$,

$$\begin{aligned} G_{a,+} = \{x \in \mathcal{X}, w_{\mathrm{Dis}}(x,a) > 0\};\ G_{a,0} = \{x \in \mathcal{X}, w_{\mathrm{Dis}}(x,a) = 0\}; \\ \text{and} \quad G_{a,-} = \{x \in \mathcal{X}, w_{\mathrm{Dis}}(x,a) < 0\}. \end{aligned} \tag{57}$$

Let $t_1 < t_2$, as well as $\tau_1 : \mathcal{X} \times \{0,1\} \to [0,1]$ and $\tau_2 : \mathcal{X} \times \{0,1\} \to [0,1]$. For $x \in \mathcal{X}$ and $a \in \{0,1\}$ with $w_{\mathrm{Dis}}(x,a) \neq 0$, we let $q_{\mathrm{Dis}}(x,a) = (2\eta_a(x)-1)/w_{\mathrm{Dis}}(x,a)$. Then, we have, for all x, a,

$$\begin{aligned} & f_{\mathrm{Dis},t_1,\tau_{1,1},\tau_{0,1}}(x,a) - f_{\mathrm{Dis},t_2,\tau_{1,2},\tau_{0,2}}(x,a) \\ &= \begin{cases} \begin{array}{l} I\left(t_1 < q_{\mathrm{Dis}}(x,a) \leqslant t_2\right) \\ \qquad +\tau_{a,1}(x,a) I\left(q_{\mathrm{Dis}}(x,a) = t_1\right) - \tau_{a,2}(x,a) I\left(q_{\mathrm{Dis}}(x,a) = t_2\right) \end{array}, & x \in G_{a,+}; \\ \begin{array}{l} -I\left(t_1 \leqslant q_{\mathrm{Dis}}(x,a) < t_2\right) \\ \qquad +\tau_{a,1}(x,a) I\left(q_{\mathrm{Dis}}(x,a) = t_1\right) - \tau_{a,2}(x,a) I\left(q_{\mathrm{Dis}}(x,a) = t_2\right) \end{array}, & x \in G_{a,-}; \\ (\tau_{a,1}(x,a) - \tau_{a,2}(x,a)) I\left(\eta_a(x) = \frac{1}{2}\right), & x \in G_{a,0}; \end{cases} \end{aligned}$$

$$\geqslant \begin{cases} I\left(t_1 < \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_2\right), & x \in G_{a,+}; \\ -I\left(t_1 \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t_2\right), & x \in G_{a,-}; \\ -I\left(\eta_a(x) = \frac{1}{2}\right), & x \in G_{a,0}. \end{cases} \tag{58}$$

It thus follows that

$$\begin{aligned}
&D_{\mathrm{Dis},\min}(t_1) - D_{\mathrm{Dis},\min}(t_2) = \inf_{\tau_1} \mathrm{Dis}(f_{\mathrm{Dis},t_1,\tau_1}) - \inf_{\tau_2} \mathrm{Dis}(f_{\mathrm{Dis},t_2,\tau_2}) \\
&\geqslant \inf_{\tau_1,\tau_2} \int_{\mathcal{A}} \int_{\mathcal{X}} \left[f_{\mathrm{Dis},t_1,\tau_1}(x,a) - f_{\mathrm{Dis},t_2,\tau_2}(x,a)\right] w_{\mathrm{Dis}}(x,a) d\mathbb{P}_{X,A}(x,a) \\
&= \inf_{\tau_1,\tau_2} \left[\sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left[f_{\mathrm{Dis},t_1}(x,a) - f_{\mathrm{Dis},t_2}(x,a)\right] w_{\mathrm{Dis}}(x,a) d\mathbb{P}_{X|A=a}(x)\right] \\
&\geqslant \sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} I\left(t_1 < \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t_2\right) w_{\mathrm{Dis}}(x,a) d\mathbb{P}_{X|A=a}(x) \\
&\quad - \sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} I\left(t_1 \leqslant \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t_2\right) w_{\mathrm{Dis}}(x,a) d\mathbb{P}_{X|A=a}(x) \\
&\quad - \sum_{a\in\{0,1\}} p_a \int_{G_{a,0}} I\left(\eta_a(x) = \frac{1}{2}\right) \cdot 0 d\mathbb{P}_{X|A=a}(x) \geqslant 0.
\end{aligned}$$

The last inequality holds since the indicator function is non-negative, and w_{Dis} is positive on $G_{a,+}$ and negative on $G_{a,-}$. Using a similar argument, we can verify that $D_{\mathrm{Dis},\max}$ is also monotonically non-increasing. ∎

Proof [Proof of Theorem 38]

We analyze the following three cases: (1) $\delta \geqslant \max\{D_{\mathrm{Dis},\min}(0), -D_{\mathrm{Dis},\max}(0)\}$, (2) $D_{\mathrm{Dis},\min}(0) > \delta$ and (3) $D_{\mathrm{Dis},\max}(0) < -\delta$. Since the proof for case (3) is analogous to that for case (2), we omit the discussion of case (3).

Case 1: $\delta \geqslant \max\{D_{\mathrm{Dis},\min}(0), -D_{\mathrm{Dis},\max}(0)\}$. In this case, there exists at least one unconstrained Bayes-optimal classifier satisfying the fairness constraint. As a result, we have $t_{\mathrm{Dis}}(\delta) = 0$ and the value of function $\tau_{\mathrm{Dis},\delta}$ has no effect on the excess risk of the classifier. Thus, we only need to choose one such that

$$\left|\mathbb{P}\left(\widehat{Y}_{f_\delta^\star} = 1 \mid A = 1\right) - \mathbb{P}\left(\widehat{Y}_{f_\delta^\star} = 0 | A = 1\right)\right| \leqslant \delta.$$

Case 2: $D_{\mathrm{Dis},\min}(0) > \delta$. By the definition of $t_{\mathrm{Dis}}(\delta)$ in (56) and Proposition 37, we have $t_{\mathrm{Dis}}(\delta) > 0$. Let, for all x, a, $\phi_0(x,a) = 2\eta_{A=1}^Y(x,a) - 1$, $\phi_1(x,a) = w_{\mathrm{Dis}}(x)$. By Lemma 27 and (7), we can write $\mathrm{Acc}(f)$ and $D_{\mathrm{Dis}}(f)$ as

$$\begin{aligned}
\mathrm{Acc}(f) &= \int_{\mathcal{A}} \int_{\mathcal{X}} f(x,a)\phi_0(x,a) d\mathbb{P}_{X,A}(x,a) + \int_{\mathcal{A}} \int_{\mathcal{X}} (1 - \eta_a(x)) d\mathbb{P}_{X,A}(x,a); \\
\mathrm{Dis}(f) &= \int_{\mathcal{X}\times\mathcal{A}} f(x,a)\phi_1(x,a) d\mathbb{P}_{X,A}(x,a).
\end{aligned}$$

Define

$$\mathcal{F}_{=} = \{f : \mathrm{Dis}(f) = \delta\}\,; \mathcal{F}_{|\cdot|,\leqslant} = \{f : |\mathrm{Dis}(f)| \leqslant \delta\}\,; \text{ and } \mathcal{F}_{\leqslant} = \{f : \mathrm{Dis}(f) \leqslant \delta\}\,.$$

By our construction, we have $f_{\mathrm{Dis},t_{\mathrm{Dis}}(\delta),\tau_{\mathrm{Dis},\delta}} \in \mathcal{F}_{=} \subset \mathcal{F}_{|\cdot|,\leqslant} \subset \mathcal{F}_{\leqslant}$. Since $t_{\mathrm{Dis}}(\delta) \geqslant 0$, by the generalized Neyman-Pearson lemma (Lemma 9),

$$\max_{f\in\mathcal{F}_{\leqslant}} \mathrm{Acc}(f) = \mathrm{Acc}(f_{\mathrm{Dis},t_{\mathrm{Dis}}(\delta),\tau_{\mathrm{Dis},\delta}}) \leqslant \max_{f\in\mathcal{F}_{|\cdot|,\leqslant}} \mathrm{Acc}(f) \leqslant \max_{f\in\mathcal{F}_{\leqslant}} \mathrm{Acc}(f).$$

Thus, we can conclude that

$$f_{\mathrm{Dis},t_{\mathrm{Dis}}(\delta),\tau_{\mathrm{Dis},\delta}} = \arg\max_{f\in\mathcal{F}_{|\cdot|,\leqslant}} \mathrm{Acc}(f) = \arg\min_{f\in\mathcal{F}} \{R(f) : |\mathrm{Dis}(f)| \leqslant \delta\}\,.$$

The proof is thus completed. ∎

Proof [Proof of Proposition 39]

We only prove the results for $D_{\mathrm{Dis,min}}$ and similar arguments apply to the result for $D_{\mathrm{Dis,max}}(t)$. For (1), by definition, we have

$$\begin{aligned}
\mathrm{Dis}(f_{\mathrm{Dis},t,\tau}) &= \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} [w_{\mathrm{Dis}}(x,a) f_{\mathrm{Dis},t,\tau}(x,a)]\, d\mathbb{P}_{X|A=a}(x)\\
&= \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left[w_{\mathrm{Dis}}(x,a) I\left(\eta_a(x) > \frac{1}{2} + \frac{t}{2} w_{\mathrm{Dis}}(x,a)\right)\right] d\mathbb{P}_{X|A=a}(x)\\
&\quad + \sum_{a\in\{0,1\}} p_a \int_{\mathcal{X}} \left[w_{\mathrm{Dis}}(x,a)\tau(x,a) I\left(\eta_a(x) = \frac{1}{2} + \frac{t}{2} w_{\mathrm{Dis}}(x,a)\right)\right] d\mathbb{P}_{X|A=a}(x)\\
&= \sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} \left[w_{\mathrm{Dis}}(x,a) I\left(\eta_a(x) > \frac{1}{2} + \frac{t}{2} w_{\mathrm{Dis}}(x,a)\right)\right] d\mathbb{P}_{X|A=a}(x)\\
&\quad + \sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} \left[w_{\mathrm{Dis}}(x,a) I\left(\eta_a(x) > \frac{1}{2} + \frac{t}{2} w_{\mathrm{Dis}}(x,a)\right)\right] d\mathbb{P}_{X|A=a}(x)\\
&\quad + \sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} \left[w_{\mathrm{Dis}}(x,a)\tau(x,a) I\left(\eta_a(x) = \frac{1}{2} + \frac{t}{2} w_{\mathrm{Dis}}(x,a)\right)\right] d\mathbb{P}_{X|A=a}(x)\\
&\quad + \sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} \left[w_{\mathrm{Dis}}(x,a)\tau(x,a) I\left(\eta_a(x) = \frac{1}{2} + \frac{t}{2} w_{\mathrm{Dis}}(x,a)\right)\right] d\mathbb{P}_{X|A=a}(x). \qquad (59)
\end{aligned}$$

Since the indicator function is always non-negative, the infimum of (59) with respect to $\tau : \mathcal{X} \times \{0,1\} \to [0,1]$ is achieved by taking $\tau_{\mathrm{Dis,min}}(x,a) = I\left((x,a) \in G_{a,-}\right)$ for all x, a. Moreover, we have,

$$D_{\mathrm{Dis,min}}(t) = \sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} \left[w_{\mathrm{Dis}}(x,a) I\left(\eta_a(x) > \frac{1}{2} + \frac{t}{2} w_{\mathrm{Dis}}(x,a)\right)\right] d\mathbb{P}_{X|A=a}(x)$$

$$
\begin{aligned}
&+\sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} \left[w_{\mathrm{Dis}}(x,a) I\left(\eta_a(x) \geqslant \frac{1}{2}+\frac{t}{2} w_{\mathrm{Dis}}(x,a)\right)\right] d\mathbb{P}_{X|A=a}(x) \\
=&\sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} > t\right)\right] d\mathbb{P}_{X|A=a}(x) \\
&+\sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t\right)\right] d\mathbb{P}_{X|A=a}(x).
\end{aligned}
$$

For (2), denote

$$
\begin{aligned}
D_{\mathrm{Dis},+,>}(t) &= \sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} > t\right)\right] d\mathbb{P}_{X|A=a}(x) \\
D_{\mathrm{Dis},-,\leqslant}(t) &= \sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \leqslant t\right)\right] d\mathbb{P}_{X|A=a}(x).
\end{aligned}
$$

We study the left- and right-hand limit of $D_{\mathrm{Dis},+,>}$ at a point $t\in\mathbb{R}$ as follows. Let $(t_n)_{n=1}^{\infty}$ be monotone decreasing sequence tending to zero as $n\to\infty$. For fixed $t\in\mathbb{R}$ and $a\in\{0,1\}$, we consider the sets

$$
\begin{aligned}
J_{a,-,t} &= \left\{(x,a) : \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \geqslant t\right\}; \qquad & J_{n,a,-,t} &= \left\{(x,a) : \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} > t-t_n\right\}; \\
J_{a,+,t} &= \left\{(x,a) : \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} > t\right\}; \qquad & J_{n,a,+,t} &= \left\{(x,a) : \frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} > t+t_n\right\}.
\end{aligned}
$$

and define the following functions on $\mathcal{X}\times\{0,1\}$. For all x,a, let,

$$
\begin{aligned}
j_{n,a,-,t}(x,a) &= w_{\mathrm{Dis}}(x,a) I\left(G_{a,+}\cap J_{n,a,-,t}\right); \quad & j_{a,-,t}(x,a) &= w_{\mathrm{Dis}}(x,a) I\left(G_{a,+}\cap J_{a,-,t}\right); \\
j_{n,a,+,t}(x,a) &= w_{\mathrm{Dis}}(x,a) I\left(G_{a,+}\cap J_{n,a,+,t}\right); \quad & j_{a,+,t}(x,a) &= w_{\mathrm{Dis}}(x,a) I\left(G_{a,+}\cap J_{a,+,t}\right).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&j_{1,a,-,t}(x,a) \geqslant j_{2,a,-,t}(x,a) \geqslant j_{3,a,-,t}(x,a) \geqslant \dots \quad \text{with } \lim_{n\to\infty} j_{n,a,-,t}(x,a) = j_{a,-,t}(x,a); \\
&j_{1,a,+,t}(x,a) \leqslant j_{2,a,+,t}(x,a) \leqslant j_{3,a,+,t}(x,a) \leqslant \dots \quad \text{with } \lim_{n\to\infty} j_{n,a,+,t}(x,a) = j_{a,+,t}(x,a).
\end{aligned}
$$

Then, by the monotone convergence theorem (Royden and Fitzpatrick, 2010),

$$
\begin{aligned}
&\lim_{t'\to t-} D_{\mathrm{Dis},+,>}(t') = \lim_{n\to\infty} D_{\mathrm{Dis},+,>}(t-t_n) \\
=&\lim_{n\to\infty}\left[\sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} > t-t_n\right)\right] d\mathbb{P}_{X|A=a}(x)\right] \\
=&\sum_{a\in\{0,1\}}\left\{p_a \lim_{n\to\infty}\left[\left[\int_{\mathcal{X}} j_{n,a,-,t}(x,a) d\mathbb{P}_{X|A=a}(x)\right]\right]\right\} = \sum_{a\in\{0,1\}}\left\{p_a \int_{\mathcal{X}} j_{a,-,t}(x,a) d\mathbb{P}_{X|A=a}(x)\right\} \\
=&\sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \geqslant t\right)\right] d\mathbb{P}_{X|A=a}(x).
\end{aligned}
$$

Similarly, we can verify that

$$\begin{aligned}
&\lim_{t'\to t+} D_{\mathrm{Dis},+,>}(t') = \lim_{n\to\infty} D_{\mathrm{Dis},+,>}(t+t_n)\\
&= \sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} > t\right)\right] d\mathbb{P}_{X|A=a}(x) = D_{\mathrm{Dis},+,>}(t).
\end{aligned}$$

Thus, $D_{\mathrm{Dis},+,>}$ is right-continuous and for $t\in\mathbb{R}$,

$$\lim_{t'\to t^-} D_{\mathrm{Dis},+,>}(t') = \sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \geqslant t\right)\right] d\mathbb{P}_{X|A=a}(x).$$

With the similar arguments above, we can show that $D_{\mathrm{Dis},-,\leqslant}$ is also right-continuous, moreover, for $t\in\mathbb{R}$

$$\lim_{t'\to t^-} D_{\mathrm{Dis},-,\leqslant}(t') = \sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t\right)\right] d\mathbb{P}_{X|A=a}(x).$$

As a consequence, $D_{\mathrm{Dis},\min} = D_{\mathrm{Dis},+,>} + D_{\mathrm{Dis},-,\leqslant}$ is right-continuous.

Finally, (3) holds since, for $t\in\mathbb{R}$,

$$\begin{aligned}
&\lim_{t'\to t^-} D_{\mathrm{Dis},\min}(t') = \lim_{t'\to t^-} D_{\mathrm{Dis},+,>}(t') + \lim_{t'\to t^+} D_{\mathrm{Dis},-,\leqslant}(t')\\
&= \sum_{a\in\{0,1\}} p_a \int_{G_{a,+}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} \geqslant t\right)\right] d\mathbb{P}_{X|A=a}(x)\\
&\quad + \sum_{a\in\{0,1\}} p_a \int_{G_{a,-}} \left[w_{\mathrm{Dis}}(x,a) I\left(\frac{2\eta_a(x)-1}{w_{\mathrm{Dis}}(x,a)} < t\right)\right] d\mathbb{P}_{X|A=a}(x) = D_{\mathrm{Dis},\max}(t).
\end{aligned}$$

This finishes the proof. ■

Appendix C. Bayes-optimal Classifiers for Synthetic Data

In this section, we derive the δ-fair Bayes-optimal classifiers for our synthetic model used in Section 6.1. In particular, we consider the following data distribution of (X,A,Y): for $(a,y)\in\{0,1\}^2$, we let $\mathbb{P}(A=a,Y=y)=p_{a,y}$ and $X|A=a,Y=y\sim\mathcal{N}(\mu_{a,y},\sigma^2 I_p)$.

Denote, for all x, by $h_{a,y}(x) = (2\pi)^{-p/2}\sigma^{-p}\exp(-\|x-\mu_{a,y}\|^2/(2\sigma^2))$ the conditional probability density function of X given $A=a$ and $Y=y$. We have

$$\begin{aligned}
\eta_a(x) = \mathbb{P}(Y=1|X=x,A=a) &= \frac{p_{a,1}h_{a,1}(x)}{p_{a,1}h_{a,1}(x)+p_{a,o}h_{a,0}(x)}\\
&= \frac{p_{a,1}\exp(-\frac{1}{2\sigma^2}\|x-\mu_{a,1}\|^2)}{p_{a,1}\exp(-\frac{1}{2\sigma^2}\|x-\mu_{a,1}\|^2)+p_{a,0}\exp(-\frac{1}{2\sigma^2}\|x-\mu_{a,0}\|^2)}.
\end{aligned}$$

In the following, for $a\in\{0,1\}$, we denote $\Delta_{\mu,a}=\mu_{a,1}-\mu_{a,0}$. Moreover, for $t\in[0,1]$ and $a\in\{0,1\}$, we denote, $q_a(t) = tp_{a,0}/((1-t)p_{a,1})$. Then, we have, for $(a,y)\in\{0,1\}^2$,

$$\mathbb{P}_{X|A=a,Y=y}(\eta_a(X)>t) = \mathbb{P}_{X|A=a,Y=y}\left(\frac{p_{a,1}h_{a,1}(x)}{p_{a,1}h_{a,1}(x)+p_{a,0}h_{a,0}(x)} > t\right)$$

$$
\begin{aligned}
&= \mathbb{P}_{X|A=a,Y=y}\left((1-t)p_{a,1}h_{a,1}(x) > tp_{a,0}h_{a,0}(x)\right) = \mathbb{P}_{X|A=a,Y=y}\left(h_{a,1}(x) > q_a(t)h_{a,0}(x)\right)\\
&= \mathbb{P}_{X|A=a,Y=y}\left(\exp\left(-\frac{1}{2\sigma^2}\left\|x-\mu_{a,1}\right\|^2\right) > q_a(t)\cdot\exp\left(-\frac{1}{2\sigma^2}\left\|x-\mu_{a,0}\right\|^2\right)\right)\\
&= \mathbb{P}_{X|A=a,Y=y}\left(\left\|x-\mu_{a,0}\right\|^2 - \left\|x-\mu_{a,1}\right\|^2 > 2\sigma^2\log(q_a(t))\right)\\
&= \mathbb{P}_{X|A=a,Y=y}\left(2x^\top\Delta_{\mu,a} > 2\sigma^2\log(q_a(t)) + \|\mu_{a,1}\|^2 - \|\mu_{a,0}\|^2\right)\\
&= \mathbb{P}_{Z\sim\mathcal{N}(0,I_p)}\left(2(\sigma\cdot Z+\mu_{a,y})^\top\Delta_{\mu,a} > 2\sigma^2\log(q_a(t)) + \|\mu_{a,1}\|^2 - \|\mu_{a,0}\|^2\right).
\end{aligned}
$$

From the definition of the distribution of (X, A, Y), it follows that this further equals

$$
\begin{aligned}
&\mathbb{P}_{Z\sim\mathcal{N}(0,I_p)}\left(\sigma\cdot Z^\top\Delta_{\mu,a} > \sigma^2\log(q_a(t)) + \frac{\|\mu_{a,1}\|^2 - \|\mu_{a,0}\|^2 - 2\mu_{a,y}^\top\Delta_{\mu,a}}{2}\right)\\
&= \mathbb{P}_{Z\sim\mathcal{N}(0,I_p)}\left(Z^\top\Delta_{\mu,a} > \sigma\log(q_a(t)) + (1-2y)\frac{\|\Delta_{\mu,a}\|^2}{2\sigma}\right)\\
&= \mathbb{P}_{Z'\sim\mathcal{N}(0,1)}\left(Z' > \frac{\sigma\log(q_a(t))}{\|\Delta_{\mu,a}\|} + (1-2y)\frac{\|\Delta_{\mu,a}\|}{2\sigma}\right)\\
&= 1-\Phi\left(\frac{\sigma\log(q_a(t))}{\|\Delta_{\mu,a}\|} + (1-2y)\frac{\|\Delta_{\mu,a}\|}{2\sigma}\right),
\end{aligned}
$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Hereafter, we denote, for $t \in [0,1]$, and $(a, y) \in \{0,1\}^2$,

$$
\psi_{a,y}(t) = \Phi\left(\frac{\sigma\log(q_a(t))}{\|\Delta_{\mu,a}\|} + (1-2y)\frac{\|\Delta_{\mu,a}\|}{2\sigma}\right).
$$

Next, we derive δ-fair Bayes-optimal classifiers under demographic parity, equality of opportunity and predictive equality.

C.1 Demographic Parity

For demographic parity, the weight function is given by, for all (x, a), $w_{\mathrm{DD}}(x, a) = (2a-1)/p_a$. This time, the disparity function DD take values, for $t \in \mathcal{T}_{\mathrm{DD}} := [-\min(p_1, p_0), \min(p_1, p_0)]$,

$$
\begin{aligned}
\mathrm{DD}(t) &= \sum_{a\in\{0,1\}}(2a-1)\mathbb{P}_{X|A=a}\left(\eta_a(X) > \frac{1}{2} + \frac{(2a-1)t}{2p_a}\right)\\
&= \sum_{a\in\{0,1\}}\sum_{y\in\{0,1\}}\left[(2a-1)\mathbb{P}_{X|A=a}\left(Y=y, \eta_a(X) > \frac{1}{2} + \frac{(2a-1)t}{2p_a}\right)\right]\\
&= \sum_{a\in\{0,1\}}\sum_{y\in\{0,1\}}\left[(2a-1)\mathbb{P}(Y=y|A=a)\cdot\mathbb{P}_{X|A=a,Y=y}\left(\eta_a(X) > \frac{1}{2} + \frac{(2a-1)t}{2p_a}\right)\right]\\
&= \sum_{a\in\{0,1\}}\sum_{y\in\{0,1\}}\left[\frac{(2a-1)p_{a,y}}{p_a}\left(1-\psi_{a,y}\left(\frac{1}{2} + \frac{(2a-1)t}{2p_a}\right)\right)\right]
\end{aligned}
$$

$$= -\frac{p_{1,1}}{p_1}\psi_{1,1}\left(\frac{p_1+t}{2p_1}\right) - \frac{p_{1,0}}{p_1}\psi_{1,0}\left(\frac{p_1+t}{2p_1}\right) + \frac{p_{0,1}}{p_0}\psi_{0,1}\left(\frac{p_0-t}{2p_0}\right) + \frac{p_{0,0}}{p_0}\psi_{0,0}\left(\frac{p_0-t}{2p_0}\right).$$

By Corollary 15, a δ-fair Bayes-optimal classifier is, for all x, a,

$$f^\star_{\mathrm{DD},\delta}(x,a) = I\left(\eta_a(x) > \frac{1}{2} + \frac{(2a-1)t_{\mathrm{DD}}(\delta)}{2p_a}\right),$$

with $t_{\mathrm{DD}}(\delta) = \arg\min_{t\in\mathcal{T}_{\mathrm{DD}}}\{|t| : \mathrm{DD}(t) \leqslant \delta\}$. Moreover, we have,

$$\begin{aligned}
R(f^\star_{\mathrm{DD},\delta}) &= \mathbb{P}\left(\widehat{Y}_{f^\star_{\mathrm{DD},\delta}} \neq Y\right) = \sum_{y\in\{0,1\}} \mathbb{P}\left(\widehat{Y}_{f^\star_{\mathrm{DD},\delta}} = y, Y = 1-Y\right) \\
&= \sum_{a\in\{0,1\}}\sum_{y\in\{0,1\}} p_{a,y}\cdot \mathbb{P}_{X|A=a,Y=y}\left(\widehat{Y}_{f^\star_{\mathrm{DD},\delta}} = 1-y\right) \\
&= \sum_{a\in\{0,1\}}\left[p_{a,1}\cdot\mathbb{P}_{A=a,y=1}\left(\eta_a(X) \leqslant \frac{1}{2} + \frac{(2a-1)t}{2p_a}\right)\right] \\
&\quad + \sum_{a\in\{0,1\}}\left[p_{a,0}\cdot\mathbb{P}_{A=a,y=0}\left(\eta_a(X) > \frac{1}{2} + \frac{(2a-1)t}{2p_a}\right)\right] \\
&= p_{1,1}\psi_{1,1}\left(\frac{p_1+t}{2p_1}\right) - p_{1,0}\psi_{1,0}\left(\frac{p_1+t}{2p_1}\right) + p_{0,1}\psi_{0,1}\left(\frac{p_0-t}{2p_0}\right) \\
&\quad - p_{0,0}\psi_{0,0}\left(\frac{p_0-t}{2p_0}\right) + p_{1,0} + p_{0,0}.
\end{aligned}$$

C.2 Equality of Opportunity

For equality of opportunity, the weight function is given for all (x, a), by $w_{\mathrm{DO}}(x,a) = (2a-1)\eta_a(x)/p_{a,1}$. The disparity function DO take values, for $t \in \mathcal{T}_{\mathrm{DO}} := [-p_{0,1}, p_{1,1}]$,

$$\begin{aligned}
D_{\mathrm{DO}}(t) &= \sum_{a\in\{0,1\}}(2a-1)\mathbb{P}_{A=a,Y=1}\left(\eta_a(X) > \frac{1}{2} + \frac{(2a-1)t\eta_a(X)}{2p_{a,1}}\right) \\
&= \sum_{a\in\{0,1\}}\left[(2a-1)\mathbb{P}_{A=a,Y=1}\left(\eta_a(X) > \frac{p_{a,1}}{2p_{a,1}-(2a-1)t}\right)\right] \\
&= p_{1,1}\left(1-\psi_{1,1}\left(\frac{p_{1,1}}{2p_{1,1}-t}\right)\right) - p_{0,1}\left(1-\psi_{0,1}\left(\frac{p_{0,1}}{2p_{0,1}+t}\right)\right).
\end{aligned}$$

By Corollary 15, a δ-fair Bayes-optimal classifier is, for all x, a,

$$f^\star_{\mathrm{DO},\delta}(x,a) = I\left(\eta_a(x) > \frac{p_{a,1}}{2p_{a,1}-(2a-1)t_{\mathrm{DO}}(\delta)}\right),$$

with $t_{\mathrm{DO}}(\delta) = \arg\min_{t\in\mathcal{T}_{\mathrm{DO}}}\{|t| : D_{\mathrm{DO}}(t) \leqslant \delta\}$. Similarly, we have,

$$R(f^\star_{\mathrm{DO},\delta}) = \mathbb{P}\left(\widehat{Y}_{f^\star_{\mathrm{DO},\delta}} \neq Y\right) = \sum_{y\in\{0,1\}} \mathbb{P}\left(\widehat{Y}_{f^\star_{\mathrm{DO},\delta}} = y, Y = 1-Y\right)$$

$$= p_{1,1}\psi_{1,1}\left(\frac{p_{1,1}}{2p_{1,1}-t}\right) - p_{1,0}\psi_{1,0}\left(\frac{p_{1,1}}{2p_{1,1}-t}\right) + p_{0,1}\psi_{0,1}\left(\frac{p_{0,1}}{2p_{0,1}+t}\right)$$
$$- p_{0,0}\psi_{0,0}\left(\frac{p_{0,1}}{2p_{0,1}+t}\right) + p_{1,0} + p_{0,0}.$$

C.3 Predictive Equality

For predictive equality, the weight function is given by, for all (x,a), $w_{\rm PD}(x,a) = (1-\eta_a(x))/p_{a,0}$. This time, the disparity function PD take values, for $t \in \mathcal{T}_{\rm PD} := [-p_{1,0}, p_{0,0}]$,

$$D_{\rm PD}(t) = \sum_{a\in\{0,1\}} (2a-1)\mathbb{P}_{A=a,Y=0}\left(\eta_a(X) > \frac{1}{2} + \frac{(2a-1)t(1-\eta_a(X))}{2p_{a,0}}\right)$$
$$= \sum_{a\in\{0,1\}} \left[(2a-1)\mathbb{P}_{A=a,Y=0}\left(\eta_a(X) > \frac{p_{a,0}+(2a-1)t}{2p_{a,0}+(2a-1)t}\right)\right]$$
$$= p_{1,0}\left(1-\psi_{1,0}\left(\frac{p_{1,0}+t}{2p_{1,0}+t}\right)\right) - p_{0,0}\left(1-\psi_{0,0}\left(\frac{p_{0,0}-t}{2p_{0,0}-t}\right)\right).$$

By Corollary 15, a δ-fair Bayes-optimal classifier is, for all x, a,

$$f^\star_{{\rm PD},\delta}(x,a) = I\left(\eta_a(x) > \frac{p_{a,0}+(2a-1)t_{\rm PD}(\delta)}{2p_{a,0}+(2a-1)t_{\rm PD}(\delta)}\right),$$

with $t_{\rm PD}(\delta) = \arg\min_{t\in\mathcal{T}_{\rm PD}} \{|t| : D_{\rm PD}(t) \leqslant \delta\}$. Moreover,

$$R(f^\star_{{\rm PD},\delta}) = \mathbb{P}\left(\widehat{Y}_{f^\star_{{\rm PD},\delta}} \neq Y\right) = \sum_{y\in\{0,1\}} \mathbb{P}\left(\widehat{Y}_{f^\star_{{\rm PD},\delta}} = y, Y = 1-Y\right)$$
$$= p_{1,1}\psi_{1,1}\left(\frac{p_{1,0}+t}{2p_{1,0}+t}\right) - p_{1,0}\psi_{1,0}\left(\frac{p_{1,0}+t}{2p_{1,0}+t}\right) + p_{0,1}\psi_{0,1}\left(\frac{p_{0,0}-t}{2p_{0,0}-t}\right)$$
$$- p_{0,0}\psi_{0,0}\left(\frac{p_{0,0}-t}{2p_{0,0}-t}\right) + p_{1,0} + p_{0,0}.$$

Table 2: Numerical simulation results showing classification accuracies and levels of disparity of the true fair Bayes-optimal classifier and the three proposed methods (FUDS, FCSC, FPIR) using logistic regression. The reported results are averages over 5,000 test data points, with standard deviations shown in parentheses.

Demographic Parity							
Theoretical		FUDS		FCSC		FPIR	
δ	ACC	DD	ACC	DD	ACC	DD	ACC
0.00	0.727	0.013 (0.010)	0.722 (0.007)	0.013 (0.010)	0.722 (0.007)	0.013 (0.010)	0.722 (0.007)
0.10	0.736	0.100 (0.017)	0.735 (0.007)	0.100 (0.017)	0.736 (0.007)	0.100 (0.017)	0.736 (0.007)
0.20	0.746	0.200 (0.018)	0.746 (0.007)	0.200 (0.018)	0.746 (0.006)	0.200 (0.018)	0.746 (0.006)
0.30	0.753	0.300 (0.017)	0.753 (0.006)	0.300 (0.017)	0.753 (0.006)	0.300 (0.017)	0.753 (0.006)
Equality of Opportunity							
Theoretical		FUDS		FCSC		FPIR	
δ	ACC	DD	ACC	DD	ACC	DD	ACC
0.00	0.731	0.017 (0.013)	0.730 (0.007)	0.017 (0.013)	0.731 (0.007)	0.017 (0.013)	0.731 (0.007)
0.10	0.745	0.101 (0.022)	0.744 (0.007)	0.101 (0.022)	0.744 (0.007)	0.101 (0.023)	0.744 (0.007)
0.20	0.753	0.201 (0.024)	0.752 (0.006)	0.202 (0.024)	0.753 (0.006)	0.202 (0.024)	0.753 (0.006)
0.30	0.757	0.302 (0.025)	0.756 (0.006)	0.302 (0.025)	0.757 (0.006)	0.302 (0.025)	0.757 (0.006)
Predictive Equality							
Theoretical		FUDS		FCSC		FPIR	
δ	ACC	DD	ACC	DD	ACC	DD	ACC
0.00	0.743	0.020 (0.015)	0.743 (0.006)	0.020 (0.015)	0.743 (0.006)	0.020 (0.015)	0.743 (0.007)
0.10	0.751	0.099 (0.025)	0.750 (0.006)	0.099 (0.025)	0.750 (0.006)	0.099 (0.025)	0.750 (0.006)
0.20	0.756	0.199 (0.024)	0.755 (0.006)	0.199 (0.025)	0.755 (0.006)	0.199 (0.025)	0.755 (0.006)
0.30	0.758	0.297 (0.023)	0.757 (0.006)	0.296 (0.024)	0.757 (0.006)	0.297 (0.024)	0.757 (0.006)

Table 3: Training details for the empirical datasets.

Dataset	Batch size	Training Epochs	Starting Learning Rate	Learning Rate Decay Factor
AdultCensus	512	200	1e-1	0.98
COMPAS	2048	500	5e-4	None
LawSchool	2048	200	2e-4	None
ACSIncome	128	20	1e-3	None

Table 4: Empirical data results showing the DD of FUDS, FCSC and FPIR with predetermined unfairness levels.

AdultCensus						
Methods	δ	0.00	0.04	0.08	0.12	0.16
FUDS		0.009 (0.007)	0.037 (0.020)	0.080 (0.012)	0.123 (0.016)	0.152 (0.016)
FCSC	DD	0.017 (0.019)	0.037 (0.011)	0.081 (0.012)	0.123 (0.011)	0.159 (0.012)
FPIR		0.007 (0.006)	0.040 (0.010)	0.080 (0.019)	0.121 (0.020)	0.159 (0.015)
COMPAS						
Methods	δ	0.00	0.06	0.12	0.18	0.24
FUDS		0.039 (0.026)	0.053 (0.031)	0.106 (0.039)	0.166 (0.039)	0.218 (0.033)
FCSC	DD	0.033 (0.025)	0.060 (0.032)	0.115 (0.039)	0.169 (0.039)	0.223 (0.036)
FPIR		0.028 (0.023)	0.061 (0.028)	0.117 (0.038)	0.173 (0.037)	0.227 (0.033)
Lawschool						
Methods	δ	0.000	0.0016	0.032	0.048	0.064
FUDS		0.006 (0.006)	0.015 (0.007)	0.031 (0.007)	0.045 (0.008)	0.061 (0.006)
FCSC	DD	0.007 (0.005)	0.017 (0.007)	0.032 (0.008)	0.048 (0.007)	0.062 (0.007)
FPIR		0.004 (0.003)	0.016 (0.006)	0.032 (0.005)	0.047 (0.005)	0.063 (0.005)
ACSIncome						
Methods	δ	0.000	0.050	0.100	0.150	0.200
FUDS		0.007 (0.005)	0.047 (0.012)	0.102 (0.015)	0.143 (0.007)	0.179 (0.010)
FCSC	DD	0.012 (0.010)	0.050 (0.008)	0.100 (0.009)	0.148 (0.012)	0.185 (0.011)
FPIR		0.001 (0.001)	0.050 (0.001)	0.100 (0.001)	0.150 (0.001)	0.186 (0.010)

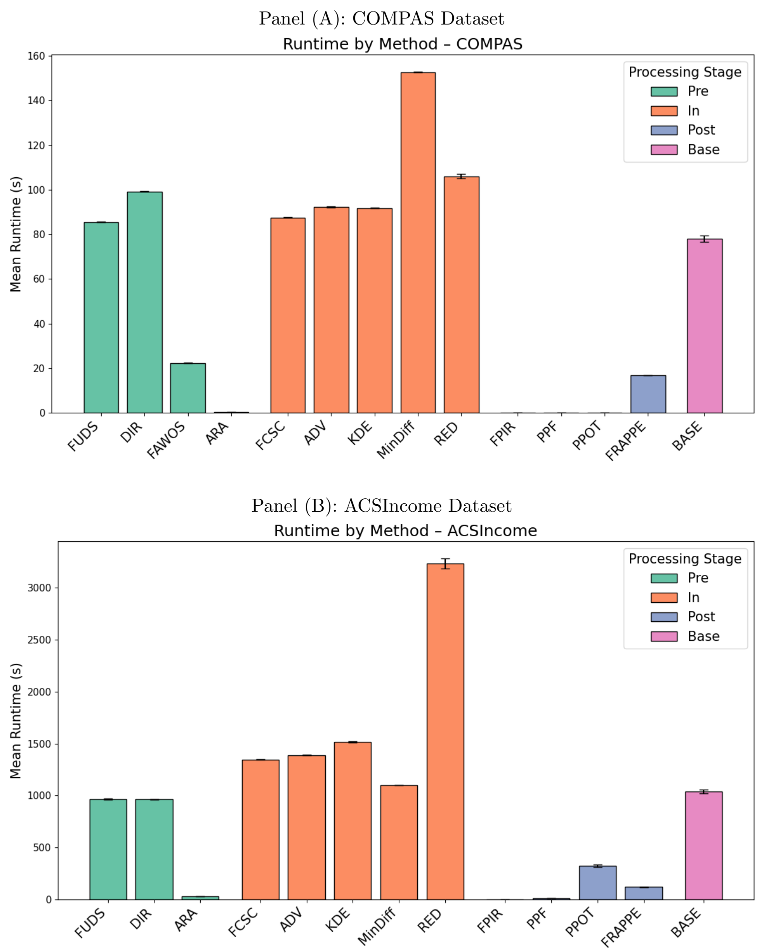

Figure 3: Computational runtimes on the COMPAS (Panel (A)) and ACSIncome (Panel (B)) datasets. *BASE* records the runtime of training an unconstrained classifier, i.e. one with no fairness constraints imposed or fairness algorithm implemented. This unconstrained classifier is used for post-processing.