
Aleatoric and Epistemic Discrimination: Fundamental Limits of Fairness Interventions

Hao Wang
MIT-IBM Watson AI Lab
hao@ibm.com

Luxi (Lucy) He
Harvard College
luxihe@college.harvard.edu

Rui Gao
The University of Texas at Austin
rui.gao@mccombs.utexas.edu

Flavio P. Calmon
Harvard University
flavio@seas.harvard.edu

Abstract

Machine learning (ML) models can underperform on certain population groups due to choices made during model development and bias inherent in the data. We categorize sources of discrimination in the ML pipeline into two classes: *aleatoric discrimination*, which is inherent in the data distribution, and *epistemic discrimination*, which is due to decisions made during model development. We quantify aleatoric discrimination by determining the performance limits of a model under fairness constraints, assuming perfect knowledge of the data distribution. We demonstrate how to characterize aleatoric discrimination by applying Blackwell’s results on comparing statistical experiments. We then quantify epistemic discrimination as the gap between a model’s accuracy when fairness constraints are applied and the limit posed by aleatoric discrimination. We apply this approach to benchmark existing fairness interventions and investigate fairness risks in data with missing values. Our results indicate that state-of-the-art fairness interventions are effective at removing epistemic discrimination on standard (overused) tabular datasets. However, when data has missing values, there is still significant room for improvement in handling aleatoric discrimination.

1 Introduction

Algorithmic discrimination may occur in different stages of the machine learning (ML) pipeline. For example, historical biases in the data-generating process can propagate to downstream tasks; human biases can influence a ML model through inductive bias; optimizing solely for accuracy can lead to disparate model performance across groups in the data (Suresh and Guttag, 2019; Mayson, 2019). The past years have seen a rapid increase in algorithmic interventions that aim to mitigate biases in ML models (see e.g., Zemel et al., 2013; Feldman et al., 2015; Calmon et al., 2017; Menon and Williamson, 2018; Zhang et al., 2018; Zafar et al., 2019; Friedler et al., 2019; Bellamy et al., 2019; Kim et al., 2019; Celis et al., 2019; Yang et al., 2020; Jiang and Nachum, 2020; Jiang et al., 2020; Martinez et al., 2020; Lowy et al., 2021; Alghamdi et al., 2022). A recent survey (Hort et al., 2022) found *nearly 400* fairness-intervention algorithms, including 123 pre-processing, 212 in-processing, and 56 post-processing algorithms introduced in the past decade.

Which sources of biases are (the hundreds of) existing fairness interventions trying to control? In order to create effective strategies for reducing algorithmic discrimination, it is critical to disentangle where biases in model performance originate. For instance, if a certain population group has significantly more missing features in training data, then it is more beneficial to collect data than selecting a more complex model class or training strategy. Conversely, if the model class does not accurately represent

the underlying distribution of a specific population group, then collecting more data for that group will not resolve performance disparities.

We divide algorithmic discrimination¹ into two categories: aleatoric and epistemic discrimination.² Aleatoric discrimination captures inherent biases in the data distribution that can lead to unfair decisions in downstream tasks. Epistemic discrimination, in turn, is due to algorithmic choices made during model development and lack of knowledge about the optimal “fair” predictive model.

In this paper, we provide methods for measuring aleatoric and epistemic discrimination in classification tasks for group fairness metrics. Since aleatoric discrimination only depends on properties of the data distribution and the fairness measure of choice, we quantify it by asking a fundamental question:

For a given data distribution, what is the best achievable performance (e.g., accuracy) under a set of group fairness constraints?

We refer to the answer as the *fairness Pareto frontier*. This frontier delineates the optimal performance achievable by a classifier when unlimited data and computing power are available. For a fixed data distribution, the fairness Pareto frontier represents the ultimate, information-theoretic limit for accuracy and group fairness beyond which no model can achieve. Characterizing this limit enables us to (i) separate sources of discrimination and create strategies to control them accordingly; (ii) evaluate the effectiveness of existing fairness interventions for reducing epistemic discrimination; and (iii) inform the development of data collection methods that promote fairness in downstream tasks.

At first, computing the fairness Pareto frontier can appear to be an intractable problem since it requires searching over all possible classifiers—even if the data distribution is known exactly. Our main technical contribution is to provide an upper bound estimate for this frontier by solving a sequence of optimization problems. The proof technique is based on Blackwell’s seminal results (Blackwell, 1953), which proposed the notion of comparisons of statistical experiments and inspired a line of works introducing alternative comparison criteria (see e.g., Shannon, 1958; Cam, 1964; Torgersen, 1991; Cohen et al., 1998; Raginsky, 2011). Here, we apply these results to develop an algorithm that iteratively refines the achievable fairness Pareto frontier. We also prove convergence guarantees for our algorithm and demonstrate how it can be used to benchmark existing fairness interventions.

We quantify epistemic discrimination by comparing a classifier’s performance with the information-theoretic optimal given by the fairness Pareto frontier. Our experiments indicate that given sufficient data, state-of-the-art (SOTA) group fairness interventions are effective at reducing epistemic discrimination as their gap to the information-theoretic limit is small (see Figure 1 and 2). Consequently, there are diminishing returns in benchmarking new fairness interventions on standard (overused) tabular datasets (e.g., UCI Adult and ProPublica COMPAS datasets). However, existing interventions *do not* eliminate aleatoric discrimination as this type of discrimination is not caused by choice of learning algorithm or model class, and is due to the data distribution. Factors such as data missing values can significantly contribute to aleatoric discrimination. We observe that when population groups have disparate missing patterns, aleatoric discrimination escalates, leading to a sharp decline in the effectiveness of fairness intervention algorithms (see Figure 3).

Related Work

There is significant work analyzing the tension between group fairness measures and model performance metrics (see e.g., Kleinberg et al., 2016; Chouldechova, 2017; Corbett-Davies et al., 2017; Chen et al., 2018; Wick et al., 2019; Dutta et al., 2020; Wang et al., 2021). For example, there is a growing body of work on omnipredictors (Gopalan et al., 2021; Hu et al., 2023; Globus-Harris et al., 2023) discussing how, and under which conditions, the fair Bayes optimal classifier can be derived using post-processing techniques from multicalibrated regressors. While previous studies (Hardt et al., 2016; Corbett-Davies et al., 2017; Menon and Williamson, 2018; Chzhen et al., 2019; Yang et al., 2020; Zeng et al., 2022a,b) have investigated the fairness Pareto frontier and fair Bayes

¹There are various measures to quantify algorithmic discrimination, and the choice should be based on the specific application of interest (see Blodgett et al., 2020; Varshney, 2021; Katzman et al., 2023, for a more detailed discussion). In this paper, we focus on group fairness measures (see Table 1 for some examples), which are crucial in contexts like hiring and recidivism prediction.

²We borrow this notion from ML uncertainty literature (see Hüllermeier and Waegeman, 2021, for a survey) and defer a detailed comparison in Appendix D.2.

optimal classifier, our approach differs from this prior work in the following aspects: our approach is applicable to *multiclass* classification problems with *multiple* protected groups; it *avoids disparate treatment* by not requiring the classifier to use group attributes as an input variable; and it can handle *multiple* fairness constraints simultaneously and produce fairness-accuracy trade-off curves (instead of a single point). Additionally, our proof techniques based on Blackwell's results on comparing statistical experiments are unique and may be of particular interest to fair ML and information theory communities. We present a detailed comparison with this line of work in Table 2 of Appendix E.

We recast the fairness Pareto frontier in terms of the conditional distribution $P_{\hat{Y}|Y,S}$ of predicted outcome \hat{Y} given true label Y and group attributes S . This conditional distribution is related to confusion matrices conditioned on each subgroup. In this regard, our work is related to [Verma and Rubin \(2018\)](#); [Alghamdi et al. \(2020\)](#); [Kim et al. \(2020\)](#); [Yang et al. \(2020\)](#); [Berk et al. \(2021\)](#), which observed that many group fairness metrics can be written in terms of the confusion matrices for each subgroup. Among them, the closest work to ours is [Kim et al. \(2020\)](#), which optimized accuracy and fairness objectives over these confusion matrices and proposed a post-processing technique for training fair classifiers. However, they only imposed marginal sum constraints for the confusion matrices. We demonstrate that the feasible region of confusion matrices can be much smaller (see Remark 2 for an example), leading to a tighter approximation of the fairness Pareto frontier.

Recently, many strategies have been proposed to reduce the tension between group fairness and model performance by investigating properties of the data distribution. For example, [Blum and Stangl \(2019\)](#); [Suresh and Guttag \(2019\)](#); [Fogliato et al. \(2020\)](#); [Wang et al. \(2020\)](#); [Mehrotra and Celis \(2021\)](#); [Fernando et al. \(2021\)](#); [Wang and Singh \(2021\)](#); [Zhang and Long \(2021\)](#); [Tomasev et al. \(2021\)](#); [Jacobs and Wallach \(2021\)](#); [Kallus et al. \(2022\)](#); [Jeong et al. \(2022\)](#) studied how noisy or missing data affect fairness and model accuracy. [Dwork et al. \(2018\)](#); [Ustun et al. \(2019\)](#); [Wang et al. \(2021\)](#) considered training a separate classifier for each subgroup when their data distributions are different. Another line of research introduces data pre-processing techniques that manipulate data distribution for reducing its bias (e.g., [Calmon et al., 2017](#); [Kamiran and Calders, 2012](#)). Among all these works, the closest one to ours is [Chen et al. \(2018\)](#), which decomposed group fairness measures into bias, variance, and noise (see their Theorem 1) and proposed strategies for reducing each term. Compared with [Chen et al. \(2018\)](#), the main difference is that we characterize a fairness Pareto frontier that depends on fairness metrics *and* a performance measure, giving a complete picture of how the data distribution influences fairness and accuracy.

2 Preliminaries

Next, we introduce notation, overview the key results in [Blackwell \(1953\)](#) on comparisons of experiments, and outline the fair classification setup considered in this paper.

Notation. For a positive integer n , let $[n] \triangleq \{1, \dots, n\}$. We denote all probability distributions on the set \mathcal{X} by $\mathcal{P}(\mathcal{X})$. Moreover, we define the probability simplex $\Delta_m \triangleq \mathcal{P}([m])$. When random variables A, X, Z form a Markov chain, we write $A - X - Z$. We write the mutual information between A, X as $I(A; X) \triangleq \mathbb{E}_{P_{A,X}} \left[\log \frac{P_{A,X}(A,X)}{P_A(A)P_X(X)} \right]$. Since $I(A; X)$ is determined by the marginal distribution P_A and the conditional distribution $P_{X|A}$, we also write $I(A; X)$ as $I(P_A; P_{X|A})$. When A, X are independent, we write $A \perp\!\!\!\perp X$.

If a random variable $A \in [n]$ has finite support, the conditional distribution $P_{X|A} : [n] \rightarrow \mathcal{P}(\mathcal{X})$ can be equivalently written as $\mathbf{P} \triangleq (P_1, \dots, P_n)$ where each $P_i = P_{X|A=i} \in \mathcal{P}(\mathcal{X})$. Additionally, if \mathcal{X} is a finite set $[m]$, then $P_{X|A}$ can be fully characterized by a transition matrix. We use $\mathcal{T}(m|n)$ to denote all transition matrices from $[n]$ to $[m]$: $\left\{ \mathbf{P} \in \mathbb{R}^{n \times m} \mid 0 \leq P_{i,j} \leq 1, \sum_{j=1}^m P_{i,j} = 1, \forall i \in [n] \right\}$.

Comparisons of Experiments

Given two statistical experiments (i.e., conditional distributions) \mathbf{P} and \mathbf{Q} , is there a way to decide which one is more informative? Here \mathbf{P} and \mathbf{Q} have the common input alphabet $[n]$ and potentially different output spaces. Blackwell gave an answer in his seminal work ([Blackwell, 1953](#)) from a decision-theoretic perspective. We review these results next.

FAIRNESS METRIC	ABBR.	DEFINITION EXPRESSION W.R.T. \mathbf{P}
Statistical Parity	$SP \leq \alpha_{SP}$	$ \Pr(\hat{Y} = \hat{y} S = s) - \Pr(\hat{Y} = \hat{y} S = s') \leq \alpha_{SP}$ $\left \sum_{y=1}^C \left(\frac{\mu_{s,y}}{\mu_s} P_{(s,y),\hat{y}} - \frac{\mu_{s',y}}{\mu_{s'}} P_{(s',y),\hat{y}} \right) \right \leq \alpha_{SP}$
Equalized Odds	$EO \leq \alpha_{EO}$	$ \Pr(\hat{Y} = \hat{y} S = s, Y = y) - \Pr(\hat{Y} = \hat{y} S = s', Y = y) \leq \alpha_{EO}$ $ P_{(s,y),\hat{y}} - P_{(s',y),\hat{y}} \leq \alpha_{EO}$
Overall Accuracy Equality	$OAE \leq \alpha_{OAE}$	$ \Pr(\hat{Y} = Y S = s) - \Pr(\hat{Y} = Y S = s') \leq \alpha_{OAE}$ $\left \sum_{y=1}^C \left(\frac{\mu_{s,y}}{\mu_s} P_{(s,y),y} - \frac{\mu_{s',y}}{\mu_{s'}} P_{(s',y),y} \right) \right \leq \alpha_{OAE}$

Table 1: Standard group fairness metrics under multi-group and multi-class classification tasks. Here $\alpha_{SP}, \alpha_{EO}, \alpha_{OAE} \in [0, 1]$ are threshold parameters, $\hat{y}, y \in [C]$, $s, s' \in [A]$, and $\mu_{s,y}, \mu_s$ are defined in Proposition 1. Our analysis can be extended to many other group fairness metrics (see e.g., Table 1 in [Kim et al., 2020](#)).

Let \mathcal{A} be a closed, bounded, convex subset of \mathbb{R}^n . A decision function $\mathbf{f}(x) = (a_1(x), \dots, a_n(x))$ is any mapping from \mathcal{X} to \mathcal{A} . It is associated with a loss vector:

$$\mathbf{v}(\mathbf{f}) = \left(\int a_1(x) dP_1(x), \dots, \int a_n(x) dP_n(x) \right). \quad (1)$$

The collection of all $\mathbf{v}(\mathbf{f})$ is denoted by $\mathcal{B}(\mathbf{P}, \mathcal{A})$. Blackwell defined that \mathbf{P} is more informative than \mathbf{Q} if for every \mathcal{A} , $\mathcal{B}(\mathbf{P}, \mathcal{A}) \supseteq \mathcal{B}(\mathbf{Q}, \mathcal{A})$. Intuitively, this result means any risk achievable with \mathbf{Q} is also achievable with \mathbf{P} . Moreover, Blackwell considered the standard measure P^* which is the probability distribution of $\mathbf{p}(\bar{X})$ where $\mathbf{p}(x) : \mathcal{X} \rightarrow \Delta_n$ is a function defined as

$$\left(\frac{dP_1}{dP_1 + \dots + dP_n}, \dots, \frac{dP_n}{dP_1 + \dots + dP_n} \right). \quad (2)$$

and \bar{X} follows the probability distribution $\frac{P_1 + \dots + P_n}{n}$. One of the most important findings by Blackwell in his paper is to discover the following equivalent conditions.

Lemma 1 ([Blackwell \(1951, 1953\)](#)). *The following three conditions are equivalent:*

- \mathbf{P} is more informative than \mathbf{Q} ;
- for any continuous and convex function $\phi : \Delta_n \rightarrow \mathbb{R}$, $\int \phi(\mathbf{p}) dP^*(\mathbf{p}) \geq \int \phi(\mathbf{p}) dQ^*(\mathbf{p})$;
- there is a stochastic transformation \mathbf{T} s.t. $\mathbf{T}P_i = Q_i$. In other words, there exists a Markov chain $\mathbf{A} - \mathbf{X} - \mathbf{Z}$ for any distributions on \mathbf{A} such that $\mathbf{P} = P_{X|\mathbf{A}}$ and $\mathbf{Q} = P_{Z|\mathbf{A}}$.

If $\mathbf{P} = P_{X|\mathbf{A}}$ is more informative than $\mathbf{Q} = P_{Z|\mathbf{A}}$, by the third condition of Lemma 1 and the data processing inequality, $I(P_{\mathbf{A}}; P_{X|\mathbf{A}}) \geq I(P_{\mathbf{A}}; P_{Z|\mathbf{A}})$ holds for any marginal distribution $P_{\mathbf{A}}$. However, the converse does not hold in general—even if the above inequality holds for any $P_{\mathbf{A}}$, \mathbf{P} is not necessarily more informative than \mathbf{Q} ([Rauh et al., 2017](#)). In this regard, Blackwell’s conditions are “stronger” than the mutual information based data processing inequality.

Group Fair Classification

Consider a multi-class classification task, where the goal is to train a probabilistic classifier $h : \mathcal{X} \rightarrow \Delta_C$ that uses input features \mathbf{X} to predict their true label $Y \in [C]$. Additionally, assume the classifier produces a predicted outcome $\hat{Y} \in [C]$ and let $S \in [A]$ represent group attributes (e.g., race and sex). Depending on the domain of interest, \mathbf{X} can either include or exclude S as an input to the classifier. Our framework can be easily extended to the setting where multiple subgroups overlap ([Kearns et al., 2018](#)). Throughout this paper, we focus on three standard group fairness measures: statistical parity (SP) ([Feldman et al., 2015](#)), equalized odds (EO) ([Hardt et al., 2016; Pleiss et al., 2017](#)), and overall accuracy equality (OAE) ([Berk et al., 2021](#)) (see Table 1 for their definitions) but our analysis can be extended to many other group fairness metrics, including the ones in Table 1 of [Kim et al. \(2020\)](#), as well as alternative performance measures beyond accuracy.

3 Fairness Pareto Frontier

In this section, we introduce our main concept—fairness Pareto frontier (FairFront). We use it to measure aleatoric discrimination and quantify epistemic discrimination by comparing a classifier’s performance to the FairFront. We recast FairFront in terms of the conditional distribution $P_{\hat{Y}|S,Y}$ and apply Blackwell’s conditions to characterize the feasible region of this conditional distribution. This effort converts a functional optimization problem into a convex program with a small number of variables. However, this convex program may involve infinitely many constraints. Hence, we introduce a greedy improvement algorithm that iteratively refines the approximation of FairFront and tightens the feasible region of $P_{\hat{Y}|S,Y}$. Finally, we establish a convergence guarantee for our algorithm.

Recall that we refer to aleatoric discrimination as the inherent biases of the data distribution that can lead to an unfair or inaccurate classifier. As its definition suggests, aleatoric discrimination only relies on properties of the data distribution and fairness metric of choice—it does not depend on the hypothesis class nor optimization method. Below we introduce FairFront that delineates a curve of optimal accuracy over all probabilistic classifiers under certain fairness constraints for a given data distribution $P_{S,X,Y}$. We use FairFront to quantify aleatoric discrimination.

Definition 1. For $\alpha_{SP}, \alpha_{EO}, \alpha_{OAE} \geq 0$ and a given $P_{S,X,Y}$, we define $\text{FairFront}(\alpha_{SP}, \alpha_{EO}, \alpha_{OAE})$ by

$$\text{FairFront}(\alpha_{SP}, \alpha_{EO}, \alpha_{OAE}) \triangleq \max_h \mathbb{E}[\mathbb{I}_{\hat{Y}=Y}] \quad (3a)$$

$$\text{s.t. } \text{SP} \leq \alpha_{SP}, \text{EO} \leq \alpha_{EO}, \text{OAE} \leq \alpha_{OAE} \quad (3b)$$

where \mathbb{I} is the indicator function; \hat{Y} is produced by applying the classifier h to X ; the maximum is taken over all measurable h ; and the definitions of SP, EO, and OAE are in Table 1. As a special case, if $\alpha_{SP}, \alpha_{EO}, \alpha_{OAE} \geq 1$, then $\text{FairFront}(\alpha_{SP}, \alpha_{EO}, \alpha_{OAE})$ is the accuracy of the Bayes optimal classifier.

Solving this functional optimization problem is difficult since it optimizes over all measurable classifiers. There is a line of works that proposed different fairness-intervention algorithms for training group-fair classifiers (see e.g., [Menon and Williamson, 2018](#); [Zhang et al., 2018](#); [Zafar et al., 2019](#); [Celis et al., 2019](#); [Yang et al., 2020](#); [Wei et al., 2021](#); [Alghamdi et al., 2022](#)). They restrict the model class and vary loss functions and optimizers to find classifiers that approach FairFront as close as possible. However, these algorithms only describe a lower bound for FairFront. They do not determine what is the *best* achievable accuracy for a given set of fairness constraints.

We circumvent the above-mentioned challenges by rewriting FairFront in terms of the conditional distribution $P_{\hat{Y}|S,Y}$. The caveat is that although each classifier yields a $P_{\hat{Y}|S,Y}$, not every conditional distribution corresponds to a valid classifier. Hence, we introduce the following definition which characterizes all feasible $P_{\hat{Y}|S,Y}$.

Definition 2. Given $P_{X|S,Y}$, we define \mathcal{C} as the set of all conditional distributions $P_{\hat{Y}|S,Y}$ where \hat{Y} is produced by some probabilistic classifier h . In other words,

$$\mathcal{C} \triangleq \{P_{\hat{Y}|S,Y} \mid (S, Y) - X - \hat{Y}\}. \quad (4)$$

Throughout this paper, we write $P_{\hat{Y}|S,Y}$ or its corresponding transition matrix $\mathbf{P} \in \mathcal{T}(C|AC)$ interchangeably. Specifically, the $(C(s-1) + y)$ -th row, \hat{y} -th column of \mathbf{P} represents $P_{\hat{Y}|S,Y}(\hat{y}|s, y)$ and we denote it by $P_{(s,y),\hat{y}}$.

Remark 1. We demonstrate the connection between the conditional distribution $P_{\hat{Y}|S,Y}$ and confusion matrices in the setting of binary classification with binary groups. We define $\hat{\mathcal{C}}$ as the counterpart of \mathcal{C} when we replace $P_{X|S,Y}$ with an empirical distribution $\hat{P}_{X|S,Y}$ computed from a dataset. The confusion matrix for group $s \in \{0, 1\}$ consists of four numbers: True Positive (TP_s), False Positive (FP_s), False Negative (FN_s), True Negative (TN_s). Assume that the number of positive-label data $n_s^+ = \text{TP}_s + \text{FN}_s$ and negative-label data $n_s^- = \text{TN}_s + \text{FP}_s$ are given—these numbers do not depend on the classifier. Then there is a one-to-one mapping from each element in $\hat{\mathcal{C}}$ to a confusion matrix:

$$\begin{aligned} \hat{P}_{\hat{Y}|S,Y}(1|s, 1) &= \frac{1}{n_s^+} \text{TP}_s, & \hat{P}_{\hat{Y}|S,Y}(1|s, 0) &= \frac{1}{n_s^-} \text{FP}_s, \\ \hat{P}_{\hat{Y}|S,Y}(0|s, 1) &= \frac{1}{n_s^+} \text{FN}_s, & \hat{P}_{\hat{Y}|S,Y}(0|s, 0) &= \frac{1}{n_s^-} \text{TN}_s. \end{aligned}$$

Hence, $\hat{\mathcal{C}}$ essentially characterizes all feasible confusion matrices and \mathcal{C} is the population counterpart of $\hat{\mathcal{C}}$. Note that \mathcal{C} is determined by the underlying data distribution while $\hat{\mathcal{C}}$ (and confusion matrices) are tailored to a specific dataset.

Proposition 1. *FairFront($\alpha_{SP}, \alpha_{EO}, \alpha_{OAE}$) in (3) is equal to the solution of the following convex optimization:*

$$\max_{\mathbf{P} \in \mathbb{R}^{AC \times C}} \sum_{s=1}^A \sum_{y=1}^C \mu_{s,y} P_{(s,y),y} \quad (5a)$$

$$\text{s.t. } \text{SP} \leq \alpha_{SP}, \text{EO} \leq \alpha_{EO}, \text{OAE} \leq \alpha_{OAE} \quad (5b)$$

$$\mathbf{P} \in \mathcal{C}. \quad (5c)$$

Here the constants $\mu_{s,y} \triangleq \Pr(S = s, Y = y)$ and $\mu_s \triangleq \Pr(S = s)$ for $s \in [A]$, $y \in [A]$ and $P_{(s,y),\hat{y}}$ denotes the $(C(s-1) + y)$ -th row, \hat{y} -th column of the transition matrix \mathbf{P} , which is $P_{\hat{Y}|S,Y}(\hat{y}|s, y)$.

For example, in binary classification with a binary group attribute, the above optimization only has 8 variables, 14 linear constraints + a single convex constraint $\mathbf{P} \in \mathcal{C}$. Hence, standard convex optimization solvers can directly compute its optimal value as long as we know how to characterize \mathcal{C} .

Remark 2. Note that [Kim et al. \(2020\)](#) investigated fairness Pareto frontiers via confusion matrices. The main difference is that Definition 1 in [Kim et al. \(2020\)](#) relaxed the constraint (5c) to $\mathbf{P} \in \mathcal{T}(C|AC)$ where $\mathcal{T}(C|AC)$ represents all transition matrices from $[AC]$ to $[C]$. This leads to a loose approximation of the frontier because \mathcal{C} is often a strict subset of $\mathcal{T}(C|AC)$. To demonstrate this point, consider the scenario where $X \perp\!\!\!\perp (S, Y)$. Then $\hat{Y} \perp\!\!\!\perp (S, Y)$ by data processing inequality so

$$\mathcal{C} = \{\mathbf{P} \in \mathcal{T}(C|AC) \mid \text{each row of } \mathbf{P} \text{ is the same}\}. \quad (6)$$

Optimizing over \mathcal{C} rather than $\mathcal{T}(C|AC)$ can significantly tighten the fairness Pareto frontier.

Before diving into the analysis, we first introduce a function $\mathbf{g} : \mathcal{X} \rightarrow \Delta_{AC}$ defined as $\mathbf{g}(x) = (P_{S,Y|X}(1, 1|x), \dots, P_{S,Y|X}(A, C|x))$. To obtain this function in practice, a common strategy among various post-processing fairness interventions (see e.g., [Menon and Williamson, 2018](#); [Alghamdi et al., 2022](#)) is to train a probabilistic classifier that uses input features X to predict (S, Y) . The output probability generated by this classifier is then utilized as an approximation of the function \mathbf{g} .

The following theorem is the main theoretical result in this paper. It provides a precise characterization of the set \mathcal{C} through a series of convex constraints.

Theorem 1. *The set \mathcal{C} is the collection of all transition matrices $\mathbf{P} \in \mathcal{T}(C|AC)$ such that the following condition holds:*

For any $k \in \mathcal{N}$ and any $\{\mathbf{a}_i \mid \mathbf{a}_i \in [-1, 1]^{AC}, i \in [k]\}$,

$$\sum_{\hat{y}=1}^C \max_{i \in [k]} \{\mathbf{a}_i^T \mathbf{\Lambda}_\mu \mathbf{p}_{\hat{y}}\} \leq \mathbb{E} \left[\max_{i \in [k]} \{\mathbf{a}_i^T \mathbf{g}(\mathbf{X})\} \right], \quad (7)$$

where $\mathbf{p}_{\hat{y}}$ is the \hat{y} -th column of \mathbf{P} and $\mathbf{\Lambda}_\mu = \text{diag}(\mu_{1,1}, \dots, \mu_{A,C})$.

Intuitively, (7) uses piecewise linear functions to approximate the boundary of the convex set \mathcal{C} where k represents the number of linear pieces. Unfortunately, replacing $\mathbf{P} \in \mathcal{C}$ with this series of constraints in (5) may result in an intractable problem since standard duality-based approaches will lead to infinitely many dual variables. To resolve this issue, we first fix k and let \mathcal{C}_k be the set of \mathbf{P} such that (7) holds under this fixed k . Accordingly, we define $\text{FairFront}_k(\alpha_{SP}, \alpha_{EO}, \alpha_{OAE})$ as the optimal value of (5) when replacing \mathcal{C} with \mathcal{C}_k . Since $\mathcal{C}_1 \supseteq \mathcal{C}_2 \supseteq \dots \supseteq \mathcal{C}$, we have $\text{FairFront}_1(\alpha_{SP}, \alpha_{EO}, \alpha_{OAE}) \geq \text{FairFront}_2(\alpha_{SP}, \alpha_{EO}, \alpha_{OAE}) \geq \dots \geq \text{FairFront}(\alpha_{SP}, \alpha_{EO}, \alpha_{OAE})$. However, computing $\text{FairFront}_k(\alpha_{SP}, \alpha_{EO}, \alpha_{OAE})$ still involves infinitely many constraints.

Next, we introduce a greedy improvement algorithm that consists of solving a sequence of tractable optimization problems for approximating $\text{FairFront}_k(\alpha_{SP}, \alpha_{EO}, \alpha_{OAE})$. We use \mathcal{A} to collect the constraints of \mathbf{P} and set $\mathcal{A} = \emptyset$ initially. At each iteration, our algorithm solves a convex program to find an optimal \mathbf{P} that maximizes the accuracy while satisfying the desired group fairness constraints and the constraints in \mathcal{A} ; then we verify if this \mathbf{P} is within the set \mathcal{C}_k by solving a DC (difference of

Algorithm 1 Approximate the fairness Pareto frontier.

Input: $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^N$, max number of iterations T ; max pieces k ; classifier $g(x)$; α_{SP} , α_{EO} , α_{OAE} .
Initialize: set $\mathcal{A} = \emptyset$; $\mu_{s,y} = \frac{|\{i|s_i=s, y_i=y\}|}{N}$; $t = 1$.

Repeat:

Solve a convex program:

$$\begin{aligned} \max_{\mathbf{P}} \quad & \sum_{s=1}^A \sum_{y=1}^C \mu_{s,y} P_{(s,y),y} \\ \text{s.t. } \mathbf{P} \in \mathcal{T}(C|AC), \text{SP} \leq \alpha_{\text{SP}}, \text{EO} \leq \alpha_{\text{EO}}, \text{OAE} \leq \alpha_{\text{OAE}} \\ & \sum_{\hat{y}=1}^C \max_{i \in [k]} \left\{ \mathbf{a}_i^T \mathbf{\Lambda}_\mu \mathbf{p}_{\hat{y}} \right\} \leq \mathbb{E} \left[\max_{i \in [k]} \{ \mathbf{a}_i^T \mathbf{g}(\mathbf{X}) \} \right] \quad \forall (\mathbf{a}_1, \dots, \mathbf{a}_k) \in \mathcal{A}. \end{aligned}$$

Let v^t and \mathbf{P}^t be the optimal value and optimal solution.

Solve a DC program:

$$\min_{\substack{\mathbf{a}_i \in [-1, 1]^{AC} \\ i \in [k]}} \mathbb{E} \left[\max_{i \in [k]} \{ \mathbf{a}_i^T \mathbf{g}(\mathbf{X}) \} \right] - \sum_{\hat{y}=1}^C \max_{i \in [k]} \left\{ \mathbf{a}_i^T \mathbf{\Lambda}_\mu \mathbf{p}_{\hat{y}}^t \right\}.$$

If the optimal value is ≥ 0 or $t = T$,

stop;

otherwise,

add the optimal $(\mathbf{a}_1, \dots, \mathbf{a}_k)$ to \mathcal{A} and $t = t + 1$.

return: v^t , \mathbf{P}^t , \mathcal{A} .

convex) program (Shen et al., 2016; Horst and Thoai, 1999). If $\mathbf{P} \in \mathcal{C}_k$, then the algorithm stops. Otherwise, the algorithm will find the constraint that is mostly violated by \mathbf{P} and add this constraint to \mathcal{A} . Specifically, we determine a piecewise linear function that divides the space into two distinct regions: one containing \mathbf{P} and the other containing \mathcal{C}_k . By ‘‘mostly violated’’, we mean the function is constructed to maximize the distance between \mathbf{P} and the boundary defined by the function. We describe our algorithm in Algorithm 1 and establish a convergence guarantee below.

Theorem 2. *Let $T = \infty$. If Algorithm 1 stops, its output \mathbf{P}^t is an optimal solution of $\text{FairFront}_k(\alpha_{\text{SP}}, \alpha_{\text{EO}}, \alpha_{\text{OAE}})$. Otherwise, any convergent sub-sequence of $\{\mathbf{P}^t\}_{t=1}^\infty$ converges to an optimal solution of $\text{FairFront}_k(\alpha_{\text{SP}}, \alpha_{\text{EO}}, \alpha_{\text{OAE}})$.*

Note that the output v^t from Algorithm 1 is always an *upper bound* for $\text{FairFront}(\alpha_{\text{SP}}, \alpha_{\text{EO}}, \alpha_{\text{OAE}})$, assuming the estimation error is sufficiently small. The tightness of this upper bound is determined by k (i.e., how well the piecewise linear functions approximate the boundary of \mathcal{C}), T (i.e., the total number of iterations). On the other hand, running off-the-shelf in-processing and post-processing fairness interventions can only yield *lower bounds* for $\text{FairFront}(\alpha_{\text{SP}}, \alpha_{\text{EO}}, \alpha_{\text{OAE}})$.

4 Numerical Experiments

In this section, we demonstrate the tightness of our upper bound approximation of FairFront, apply it to benchmark existing group fairness interventions, and show how data biases, specifically missing values, impact their effectiveness. We find that given sufficient data, SOTA fairness interventions are successful at reducing epistemic discrimination as their gap to (our upper bound estimate of) FairFront is small. However, we also discover that when different population groups have varying missing data patterns, aleatoric discrimination increases, which diminishes the performance of fairness intervention algorithms. Our numerical experiments are semi-synthetic since we apply fairness interventions to train classifiers using the *entire* dataset and resample from it as the test set. This setup enables us to eliminate the estimation error associated with Algorithm 1 (see Appendix E for a discussion). We provide additional experimental results and details in Appendix C.

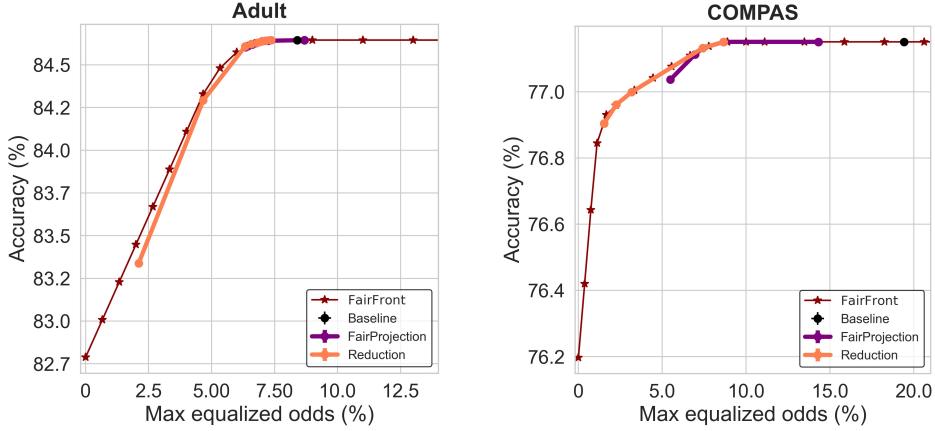


Figure 1: We compare Reduction and FairProjection with (our upper bound estimate of FairFront on the Adult (Left) and COMPAS (Right) datasets. We train a classifier that approximates the Bayes optimal and use it as a basis for Reduction and FairProjection. This result not only demonstrates the tightness of our approximation but also shows that SOTA fairness interventions have already achieved near-optimal fairness-accuracy curves.

4.1 Benchmark Fairness Interventions

Setup. We evaluate our results on the UCI Adult dataset (Bache and Lichman, 2013), the ProPublica COMPAS dataset (Angwin et al., 2016), the German Credit dataset (Bache and Lichman, 2013), and HSLS (High School Longitudinal Study) dataset (Ingels et al., 2011; Jeong et al., 2022). We recognize that Adult, COMPAS, and German Credit datasets are overused and acknowledge the recent calls to move away from them (see e.g., Ding et al., 2021). We adopt these datasets for benchmarking purposes only since most fairness interventions have available code for these datasets. The HSLS dataset is a new dataset that first appeared in the fair ML literature last year and captures a common use-case of ML in education (student performance prediction, see Jeong et al., 2022). It has multi-class labels and multiple protected groups. We apply existing (group) fairness interventions to these datasets and measure their fairness violations via *Max equalized odds*:

$$\max | \Pr(\hat{Y} = \hat{y} | S = s, Y = y) - \Pr(\hat{Y} = \hat{y} | S = s', Y = y) |$$

where the max is taken over y, \hat{y}, s, s' . We run Algorithm 1 with $k = 6$ pieces, 20 iterations, and varying α_{EO} to estimate FairFront on each dataset. We compute the expectations and the g function from the empirical distributions and solve the DC program by using the package in Shen et al. (2016). The details about how we pre-process these datasets and additional experimental results on the German Credit and HSLS datasets are deferred to Appendix C.

Group fairness interventions. We consider five existing fairness-intervention algorithms: Reduction (Agarwal et al., 2018), EqOdds (Hardt et al., 2016), CalEqOdds (Pleiss et al., 2017), LevEqOpp (Chzhen et al., 2019), and FairProjection Alghamdi et al. (2022). Among them, Reduction is an in-processing method and the rest are all post-processing methods. For the first three benchmarks, we use the implementations from IBM AIF360 library (Bellamy et al., 2018); for LevEqOpp and FairProjection, we use the Python implementations from the Github repo in Alghamdi et al. (2022). For Reduction and FairProjection, we can vary their tolerance of fairness violations to produce a fairness-accuracy curve; for EqOdds, CalEqOdds, and LevEqOpp, each of them produces a single point since they only allow hard equality constraint. We note that FairProjection is optimized for transforming probabilistic classifier outputs (see also Wei et al., 2021), but here we threshold the probabilistic outputs to generate binary predictions which may limit its performance. Finally, we train a random forest as the Baseline classifier.

Results. We observe that if we run Algorithm 1 for a single iteration, which is equivalent to solving Proposition 1 without (5c), its solution is very close to 1 for all α_{EO} . This demonstrates the benefits of incorporating Blackwell's conditions into the fairness Pareto frontier.

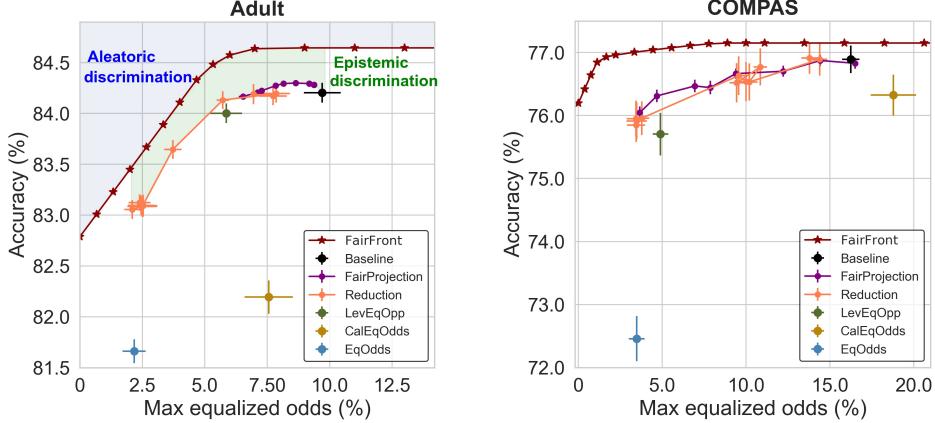


Figure 2: We benchmark existing fairness interventions using (our upper bound estimate of) FairFront. We use FairFront to quantify aleatoric discrimination and measure epistemic discrimination by comparing a classifier’s accuracy and fairness violation with FairFront. The results show that SOTA fairness interventions are effective at reducing epistemic discrimination.

We train a classifier that approximates the Bayes optimal and use it as a basis for both Reduction and FairProjection, which are SOTA fairness interventions. We then apply these two fairness interventions to the entire dataset and evaluate their performance on the same dataset. Figure 1 shows that in this infinite sample regime, the fairness-accuracy curves produced by Reduction and FairProjection can approach our upper bound estimate of FairFront. This result not only demonstrates the tightness of our approximation (recall that Algorithm 1 gives an upper bound of FairFront and existing fairness interventions give lower bounds) but also shows that SOTA fairness interventions have already achieved near-optimal fairness-accuracy curves.

Recall that we use FairFront to quantify aleatoric discrimination since it characterizes the highest achievable accuracy among all classifiers satisfying the desired fairness constraints. Additionally, we measure epistemic discrimination by comparing a classifier’s accuracy and fairness violation with FairFront. Given that our Algorithm 1 provides a tight approximation of FairFront, we use it to benchmark existing fairness interventions. Specifically, we first train a base classifier which may not achieve Bayes optimal accuracy. Then we use it as a basis for all existing fairness interventions. The results in Figure 2 show that SOTA fairness interventions remain effective at reducing epistemic discrimination. In what follows, we demonstrate how missing values in data can increase aleatoric discrimination and dramatically reduce the effectiveness of SOTA fairness interventions.

4.2 Fairness Risks in Missing Values

Real-world data often have missing values and the missing patterns can be different across different protected groups (see Jeong et al., 2022, for some examples). There is a growing line of research (see e.g., Jeong et al., 2022; Fernando et al., 2021; Wang and Singh, 2021; Subramonian et al., 2022; Caton et al., 2022; Zhang and Long, 2021; Schelter et al., 2019) studying the fairness risks of data with missing values. In this section, we apply our result to demonstrate how disparate missing patterns influence the fairness-accuracy curves.

Setup. We choose sex (group 0: female, group 1: male) as the group attribute for the Adult dataset, and race (group 0: African-American, group 1: Caucasian) for the COMPAS dataset. To investigate the impact of disparate missing patterns on aleatoric discrimination, we artificially generate missing values in both datasets. This is necessary as the datasets do not contain sufficient missing data. The missing values are generated according to different probabilities for different population groups. For each data point from group 0, we erase each input feature with a varying probability $p_0 \in \{10\%, 50\%, 70\%\}$, while for group 1, we erase each input feature with a fixed probability $p_1 = 10\%$. We then apply mode imputation to the missing values, replacing them with the mode of non-missing values for each feature. Finally, we apply Algorithm 1 along with Reduction and Baseline to the imputed data. The experimental results are shown in Figure 3.

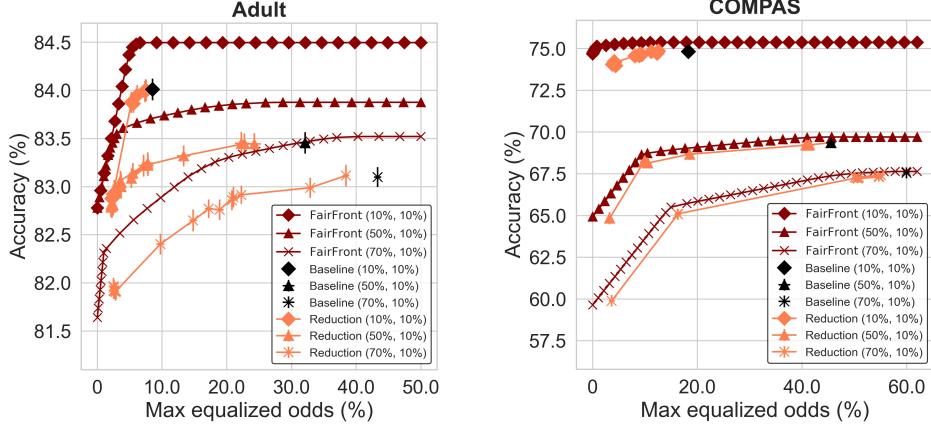


Figure 3: Fairness risks of disparate missing patterns. The missing probabilities of group 0 (female in Adult/African-American in COMPAS) and group 1 (male in Adult/Caucasian in COMPAS) are varying among $\{(10\%, 10\%), (50\%, 10\%), (70\%, 10\%)\}$. We apply Reduction and Baseline to the imputed data and plot their fairness-accuracy curves against FairFront. As shown, the effectiveness of fairness interventions substantially decrease with increasing disparate missing patterns in data.

Results. As we increase the missing probability of group 0, (our upper bound estimate of) FairFront decreases since it becomes more difficult to accurately predict outcomes for group 0. This in turn affects the overall model performance, since the fairness constraint requires that the model performs similarly for both groups. We also observe the fairness-accuracy curves of Reduction decrease as the missing data for group 0 become more prevalent. In other words, as the missing data for group 0 increase, it becomes more difficult to maintain both high accuracy and fairness in the model’s prediction.

5 Final Remarks

The past years have witnessed a growing line of research introducing various group fairness-intervention algorithms. Most of these interventions focus on optimizing model performance subject to group fairness constraints. Though comparing and benchmarking these methods on various datasets is valuable (e.g., see benchmarks in [Friedler et al., 2019](#); [Bellamy et al., 2019](#); [Wei et al., 2021](#)), this does not reveal if there is still room for improvement in their fairness-accuracy curves, or if existing methods approach the information-theoretic optimal limit when infinite data is available. Our results address this gap by introducing the fairness Pareto frontier, which measures the highest possible accuracy under a set of group fairness constraints. We precisely characterize the fairness Pareto frontier using Blackwell’s conditions and present a greedy improvement algorithm that approximates it from data. Our results show that the fairness-accuracy curves produced by SOTA fairness interventions are very close to the fairness Pareto frontier on standard datasets.

Additionally, we demonstrate that when data are biased due to missing values, the fairness Pareto frontier degrades. Although existing fairness interventions can still reduce performance disparities, they come at the cost of significantly lowering overall model accuracy. The methods we present for computing the fairness Pareto frontier can also be applied to analyze other sources of aleatoric discrimination, such as when individuals may misreport their data or when there are measurement errors. Overall, the fairness Pareto frontier can serve as a valuable framework for guiding data collection and cleaning. Our results indicate that existing fairness interventions can be effective in reducing epistemic discrimination, and there are diminishing returns in developing new fairness interventions focused solely on optimizing accuracy for a given group fairness constraint on pristine data. However, existing fairness interventions have yet to effectively provide both fair and accurate classification when additional sources of aleatoric discrimination are present (such as missing values in data). This suggests that there is still significant need for research on handling aleatoric sources of discrimination that appear throughout the data collection process.

We provide an in-depth discussion on future work in Appendix E.

Acknowledgement

This material is based upon work supported by the National Science Foundation under grants CAREER 1845852, CIF 2312667, FAI 2040880, CIF 1900750.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR.

Alghamdi, W., Asoodeh, S., Wang, H., Calmon, F. P., Wei, D., and Ramamurthy, K. N. (2020). Model projection: Theory and applications to fair machine learning. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2711–2716. IEEE.

Alghamdi, W., Hsu, H., Jeong, H., Wang, H., Michalak, P. W., Asoodeh, S., and Calmon, F. P. (2022). Beyond Adult and COMPAS: Fair multi-class prediction via information projection. In *Advances in Neural Information Processing Systems*.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica*.

Bache, K. and Lichman, M. (2013). UCI Machine Learning Repository.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al. (2019). Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1.

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

Blackwell, D. (1951). Comparison of experiments. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102.

Blackwell, D. (1953). Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of " bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Blum, A. and Stangl, K. (2019). Recovering from biased data: Can fairness constraints improve accuracy? *arXiv preprint arXiv:1912.01094*.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30.

Cam, L. L. (1964). Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, pages 1419–1455.

Caton, S., Malisetti, S., and Haas, C. (2022). Impact of imputation strategies on fairness in machine learning. *Journal of Artificial Intelligence Research*, 74:1011–1035.

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

Chen, I., Johansson, F. D., and Sontag, D. (2018). Why is my classifier discriminatory? *Advances in neural information processing systems*, 31.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2019). Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32.

Cohen, J., Kempermann, J. H., and Zbaganu, G. (1998). *Comparisons of stochastic matrices with applications in information theory, statistics, economics and population*. Springer Science & Business Media.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 797–806.

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

Dutta, S., Wei, D., Yueksel, H., Chen, P.-Y., Liu, S., and Varshney, K. (2020). Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*, pages 2803–2813. PMLR.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. (2018). Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pages 119–133. PMLR.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

Fernando, M.-P., Cèsar, F., David, N., and José, H.-O. (2021). Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 36(7):3217–3258.

Fogliato, R., Chouldechova, A., and G’Sell, M. (2020). Fairness evaluation in presence of biased noisy labels. In *International Conference on Artificial Intelligence and Statistics*, pages 2325–2336. PMLR.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338.

Globus-Harris, I., Gupta, V., Jung, C., Kearns, M., Morgenstern, J., and Roth, A. (2023). Multicalibrated regression for downstream fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 259–286.

Gopalan, P., Kalai, A. T., Reingold, O., Sharan, V., and Wieder, U. (2021). Omnipredictors. *arXiv preprint arXiv:2109.05389*.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29.

Horst, R. and Thoai, N. V. (1999). Dc programming: overview. *Journal of Optimization Theory and Applications*, 103(1):1–43.

Hort, M., Chen, Z., Zhang, J. M., Sarro, F., and Harman, M. (2022). Bia mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*.

Hu, L., Navon, I. R. L., Reingold, O., and Yang, C. (2023). Omnipredictors for constrained optimization. In *International Conference on Machine Learning*, pages 13497–13527. PMLR.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Ingels, S. J., Pratt, D. J., Herget, D. R., Burns, L. J., Dever, J. A., Ottem, R., Rogers, J. E., Jin, Y., and Leinwand, S. (2011). High school longitudinal study of 2009 (hsls: 09): Base-year data file documentation. nces 2011-328. *National Center for Education Statistics*.

Jacobs, A. Z. and Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 375–385.

Jeong, H., Wang, H., and Calmon, F. P. (2022). Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9558–9566.

Jiang, H. and Nachum, O. (2020). Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. (2020). Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pages 862–872. PMLR.

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

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

Katzman, J., Wang, A., Scheuerman, M., Blodgett, S. L., Laird, K., Wallach, H., and Barocas, S. (2023). Taxonomizing and measuring representational harms: A look at image tagging. *arXiv preprint arXiv:2305.01776*.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR.

Kim, J. S., Chen, J., and Talwalkar, A. (2020). Fact: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*, pages 5264–5274. PMLR.

Kim, M. P., Ghorbani, A., and Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Lowy, A., Pavan, R., Baharlouei, S., Razaviyayn, M., and Beirami, A. (2021). Fermi: Fair empirical risk minimization via exponential Rényi mutual information. *arXiv preprint arXiv:2102.12586*.

Martinez, N., Bertran, M., and Sapiro, G. (2020). Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR.

Mayson, S. G. (2019). Bias in, bias out. *The Yale Law Journal*, 128(8):2218–2300.

Mehrotra, A. and Celis, L. E. (2021). Mitigating bias in set selection with noisy protected attributes. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 237–248.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. *Advances in neural information processing systems*, 30.

Raginsky, M. (2011). Shannon meets blackwell and le cam: Channels, codes, and statistical experiments. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 1220–1224. IEEE.

Rauh, J., Banerjee, P. K., Olbrich, E., Jost, J., Bertschinger, N., and Wolpert, D. (2017). Coarse-graining and the blackwell order. *Entropy*, 19(10):527.

Schelter, S., He, Y., Khilnani, J., and Stoyanovich, J. (2019). Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. *arXiv preprint arXiv:1911.12587*.

Shannon, C. E. (1958). A note on a partial ordering for communication channels. *Information and control*, 1(4):390–397.

Shen, X., Diamond, S., Gu, Y., and Boyd, S. (2016). Disciplined convex-concave programming. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 1009–1014. IEEE.

Subramonian, A., Chang, K.-W., and Sun, Y. (2022). On the discrimination risk of mean aggregation feature imputation in graphs. *Advances in Neural Information Processing Systems*.

Suresh, H. and Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2:8.

Tomasev, N., McKee, K. R., Kay, J., and Mohamed, S. (2021). Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 254–265.

Torgersen, E. (1991). *Comparison of statistical experiments*, volume 36. Cambridge University Press.

Ustun, B., Liu, Y., and Parkes, D. (2019). Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382. PMLR.

Varshney, K. R. (2021). Trustworthy machine learning. *Chappaqua, NY*.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE.

Wang, H., Hsu, H., Diaz, M., and Calmon, F. P. (2021). To split or not to split: The impact of disparate treatment in classification. *IEEE Transactions on Information Theory*, 67(10):6733–6757.

Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M., and Jordan, M. (2020). Robust optimization for fairness with noisy protected groups. *Advances in Neural Information Processing Systems*, 33:5190–5203.

Wang, Y. and Singh, L. (2021). Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics*, 12(2):101–119.

Wei, D., Ramamurthy, K. N., and Calmon, F. P. (2021). Optimized score transformation for consistent fair classification. *J. Mach. Learn. Res.*, 22:258–1.

Wick, M., Tristan, J.-B., et al. (2019). Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32.

Yang, F., Cisse, M., and Koyejo, S. (2020). Fairness with overlapping groups; a probabilistic perspective. In *Advances in Neural Information Processing Systems*, volume 33, pages 4067–4078.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR.

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

Zeng, X., Dobriban, E., and Cheng, G. (2022b). Fair Bayes-optimal classifiers under predictive parity. In *Advances in Neural Information Processing Systems*.

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

Zhang, Y. and Long, Q. (2021). Assessing fairness in the presence of missing data. *Advances in neural information processing systems*, 34:16007–16019.

A Technical Background

In this section, we extend some results in [Blackwell \(1951, 1953\)](#) to our setting. For a random variable X , we denote its probability distribution by $\mathcal{L}(X)$. A conditional distribution $P_{X|A} : [n] \rightarrow \mathcal{P}(\mathcal{X})$ can be equivalently written as $\mathbf{P} \triangleq (P_1, \dots, P_n)$ where each $P_i = P_{X|A=i} \in \mathcal{P}(\mathcal{X})$. Let \mathcal{A} be a closed, bounded, convex subset of \mathbb{R}^n . A decision function is a mapping $\mathbf{f} : \mathcal{X} \rightarrow \mathcal{A}$, which can also be written as $\mathbf{f}(x) = (a_1(x), \dots, a_n(x))$. A decision function is associated a loss vector:

$$\mathbf{v}(\mathbf{f}) = \left(\int a_1(x) dP_1(x), \dots, \int a_n(x) dP_n(x) \right). \quad (9)$$

The collection of all $\mathbf{v}(\mathbf{f})$ is denoted by $\mathcal{B}(P_{X|A}, \mathcal{A})$ or $\mathcal{B}(\mathbf{P}, \mathcal{A})$.

For a vector $\lambda \in \Delta_n$ such that $\lambda > 0$, we define a function $\mathbf{p}_\lambda(x) : \mathcal{X} \rightarrow \Delta_n$:

$$\mathbf{p}_\lambda(x) = \left(\frac{\lambda_1 dP_1}{\lambda_1 dP_1 + \dots + \lambda_n dP_n}, \dots, \frac{\lambda_n dP_n}{\lambda_1 dP_1 + \dots + \lambda_n dP_n} \right). \quad (10)$$

Note that $\mathbf{p}_\lambda(X)$ is a sufficient statistic for X , considering A as the parameter (it can be proved by Fisher-Neyman factorization theorem). In other words, two Markov chains hold: $A - \mathbf{p}_\lambda(X) - X$ and $A - X - \mathbf{p}_\lambda(X)$ for any distribution on A .

Consider a new set of probability distributions $\mathbf{P}_\lambda^* \triangleq (\mathcal{L}(\mathbf{p}_\lambda(X_1)), \dots, \mathcal{L}(\mathbf{p}_\lambda(X_n)))$ where $\mathcal{L}(X_i) = P_i$. Here \mathbf{P}_λ^* can be viewed as a conditional distribution from $[n]$ to $\mathcal{P}(\Delta_n)$ since each $\mathcal{L}(\mathbf{p}_\lambda(X_i))$ is a probability distribution over Δ_n . The following lemma follows from the sufficiency of $\mathbf{p}_\lambda(X)$.

Lemma 2 (Adaptation of Theorem 3 in [Blackwell \(1951\)](#)). *For any \mathcal{A} , $\mathcal{B}(\mathbf{P}, \mathcal{A}) = \mathcal{B}(\mathbf{P}_\lambda^*, \mathcal{A})$.*

Proof. Suppose that $\mathbf{f}^*(\mathbf{p}) = (a_1^*(\mathbf{p}), \dots, a_n^*(\mathbf{p}))$ is a decision function for $(\mathbf{P}_\lambda^*, \mathcal{A})$. Accordingly, we define $\mathbf{f}(x) = (a_1^*(\mathbf{p}_\lambda(x)), \dots, a_n^*(\mathbf{p}_\lambda(x)))$ where the function \mathbf{p}_λ is defined in (10). Then it is clear that \mathbf{f} is a decision function for $(\mathbf{P}, \mathcal{A})$. By the law of unconscious statistician, we have

$$\int a_i^*(\mathbf{p}) dP_{\lambda,i}^*(\mathbf{p}) = \mathbb{E}[a_i^*(\mathbf{p}_\lambda(X_i))] = \int a_i^*(\mathbf{p}_\lambda(x)) dP_i(x). \quad (11)$$

Hence, $\mathbf{v}(\mathbf{f}^*) = \mathbf{v}(\mathbf{f})$, which implies $\mathcal{B}(\mathbf{P}_\lambda^*, \mathcal{A}) \subseteq \mathcal{B}(\mathbf{P}, \mathcal{A})$. For the other direction, suppose $\mathbf{f}(x) = (a_1(x), \dots, a_n(x))$ is a decision function for $(\mathbf{P}, \mathcal{A})$. Let $\mathbf{f}^*(\mathbf{p}) = (a_1^*(\mathbf{p}), \dots, a_n^*(\mathbf{p}))$ where $a_i^*(\mathbf{p}) \triangleq \mathbb{E}[a_i(X_i) | \mathbf{p}_\lambda(X_i) = \mathbf{p}]$. Since $\mathbf{p}_\lambda(X)$ is a sufficient statistic, for any $i \in [n]$

$$\mathcal{L}(X_i | \mathbf{p}_\lambda(X_i) = \mathbf{p}) = \mathcal{L}(X_1 | \mathbf{p}_\lambda(X_1) = \mathbf{p}). \quad (12)$$

Therefore, $\mathbf{f}^*(\mathbf{p}) = \mathbb{E}[\mathbf{f}(X_1) | \mathbf{p}_\lambda(X_1) = \mathbf{p}]$. Since \mathcal{A} is a convex set, \mathbf{f}^* is a decision function for $(\mathbf{P}^*, \mathcal{A})$. By the law of total expectation, we have

$$\int a_i^*(\mathbf{p}) dP_{\lambda,i}^*(\mathbf{p}) = \int a_i(x) dP_i(x). \quad (13)$$

Hence, $\mathbf{v}(\mathbf{f}) = \mathbf{v}(\mathbf{f}^*)$, which implies $\mathcal{B}(\mathbf{P}, \mathcal{A}) \subseteq \mathcal{B}(\mathbf{P}_\lambda^*, \mathcal{A})$. \square

For a vector $\lambda \in \Delta_n$ such that $\lambda > 0$, the condition distribution $P_{X|A}$ induces a weighted standard measure $P_\lambda^* \triangleq \mathcal{L}(\mathbf{p}_\lambda(\bar{X}))$ where $\mathcal{L}(\bar{X}) = \lambda_1 P_1 + \dots + \lambda_n P_n$.

Theorem 3 (Adaptation of Theorem 4 in [Blackwell \(1951\)](#)). *For any two conditional distributions $P_{X|A}$ and $Q_{Y|A}$, let P_λ^* and Q_λ^* be their weighted standard measures, respectively. Then $\mathcal{B}(P_{X|A}, \mathcal{A}) \supseteq \mathcal{B}(Q_{Y|A}, \mathcal{A})$ for any closed, bounded, convex set \mathcal{A} if and only if for any continuous convex $\phi : \Delta_n \rightarrow \mathbb{R}$, $\int \phi(\mathbf{p}) dP_\lambda^*(\mathbf{p}) \geq \int \phi(\mathbf{p}) dQ_\lambda^*(\mathbf{p})$*

Proof. First, by Lemma 2, we know $\mathcal{B}(P_{X|A}, \mathcal{A}) = \mathcal{B}(\mathbf{P}_\lambda^*, \mathcal{A})$ and $\mathcal{B}(Q_{Y|A}, \mathcal{A}) = \mathcal{B}(\mathbf{Q}_\lambda^*, \mathcal{A})$. We denote $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$. Consider any $\mathcal{A} = \text{conv}(\mathbf{a}_1, \dots, \mathbf{a}_k)$. Let

$$\mathbf{f}^*(\mathbf{p}) = \underset{\mathbf{a} \in \mathcal{A}}{\text{argmin}} \mathbf{p}^T \mathbf{\Lambda}^{-1} \mathbf{a}. \quad (14)$$

Note that $\mathbf{f}^*(\mathbf{p}) \in \{\mathbf{a}_1, \dots, \mathbf{a}_k\}$ since this set contains all the extreme points of \mathcal{A} .³ By definition, for any decision function w.r.t. $(\mathbf{P}_{\lambda}^*, \mathcal{A})$, we have

$$\mathbf{p}^T \mathbf{\Lambda}^{-1} \mathbf{f}(\mathbf{p}) \geq \mathbf{p}^T \mathbf{\Lambda}^{-1} \mathbf{f}^*(\mathbf{p}), \quad \forall \mathbf{p}. \quad (15)$$

Let $\mathbf{v} = \mathbf{v}(\mathbf{f})$. By the same reason with (11), we have

$$v_j = \int a_j(\mathbf{p}_{\lambda}(x)) dP_j(x) \quad (16)$$

$$= \frac{1}{\lambda_j} \int a_j(\mathbf{p}_{\lambda}(x)) \frac{\lambda_j dP_j}{\lambda_1 dP_1 + \dots + \lambda_n dP_n}(x) (\lambda_1 dP_1 + \dots + \lambda_n dP_n)(x) \quad (17)$$

$$= \frac{1}{\lambda_j} \int a_j(\mathbf{p}_{\lambda}(x)) [\mathbf{p}_{\lambda}(x)]_j (\lambda_1 dP_1 + \dots + \lambda_n dP_n)(x) \quad (18)$$

$$= \frac{1}{\lambda_j} \mathbb{E}[a_j(\mathbf{p}_{\lambda}(\bar{\mathbf{X}})) [\mathbf{p}_{\lambda}(\bar{\mathbf{X}})]_j] \quad (19)$$

$$= \frac{1}{\lambda_j} \int a_j(\mathbf{p}) p_j dP_{\lambda}^*(\mathbf{p}), \quad (20)$$

where the last step is due to the law of unconscious statistician. Therefore,

$$\sum_{j=1}^n v_j = \int \mathbf{p}^T \mathbf{\Lambda}^{-1} \mathbf{f}(\mathbf{p}) dP_{\lambda}^*(\mathbf{p}) \quad (21)$$

$$\geq \int \mathbf{p}^T \mathbf{\Lambda}^{-1} \mathbf{f}^*(\mathbf{p}) dP_{\lambda}^*(\mathbf{p}) \quad (22)$$

$$= \int \min_i \{\mathbf{p}^T \mathbf{\Lambda}^{-1} \mathbf{a}_i\} dP_{\lambda}^*(\mathbf{p}). \quad (23)$$

The equality is achieved by $\mathbf{v}(\mathbf{f}^*)$. Hence, for any $\mathcal{A} = \text{conv}(\mathbf{a}_1, \dots, \mathbf{a}_k)$

$$\min_{\mathbf{v} \in \mathcal{B}(P_{X|A}, \mathcal{A})} \sum_{j=1}^n v_j = \int \min_i \{\mathbf{a}_i^T \mathbf{\Lambda}^{-1} \mathbf{p}\} dP_{\lambda}^*(\mathbf{p}). \quad (24)$$

Recall that Theorem 2.(3) in [Blackwell \(1951\)](#) states

$$\begin{aligned} \mathcal{B}(P_{X|A}, \mathcal{A}) &\supseteq \mathcal{B}(P_{Y|A}, \mathcal{A}) \quad \text{for every closed, bounded, convex } \mathcal{A} \\ \Leftrightarrow \min_{\mathbf{v} \in \mathcal{B}(P_{X|A}, \mathcal{A})} \sum_{j=1}^n v_j &\leq \min_{\mathbf{v} \in \mathcal{B}(P_{Y|A}, \mathcal{A})} \sum_{j=1}^n v_j \quad \text{for every closed, bounded, convex } \mathcal{A}. \end{aligned}$$

By approximation theory, the second condition can be relaxed to any \mathcal{A} that is a convex hull of a finite set. By (24), this relaxed condition is equivalent to

$$\int \phi(\mathbf{p}) dP_{\lambda}^*(\mathbf{p}) \geq \int \phi(\mathbf{p}) dQ_{\lambda}^*(\mathbf{p}) \quad (25)$$

for all $\phi(\mathbf{p})$ that are the **maximum** of finitely many linear functions. By approximation theory again, the above condition is equivalent to the one holding for any continuous convex function ϕ . \square

B Omitted Proofs

B.1 Proof of Lemma 3

Proof. Clearly, \mathcal{C} is a subset of $\mathcal{T}(C|AC)$. Let $\lambda \in (0, 1)$ and $P_{\hat{Y}_0|S,Y}, P_{\hat{Y}_1|S,Y} \in \mathcal{C}$. Now we introduce a Bernoulli random variable B such that $\Pr(B = 0) = \lambda$. Finally, we define $\hat{Y}_{\lambda} = B\hat{Y}_1 + (1 - B)\hat{Y}_0$. By definition, we have $(S, Y) - X - \hat{Y}_{\lambda}$ so $P_{\hat{Y}_{\lambda}|S,Y} \in \mathcal{C}$. Moreover,

$$P_{\hat{Y}_{\lambda}|S,Y} = \lambda P_{\hat{Y}_0|S,Y} + (1 - \lambda) P_{\hat{Y}_1|S,Y}.$$

³If (14) has multiple optimal solutions, we always choose the one from $\{\mathbf{a}_1, \dots, \mathbf{a}_k\}$.

Hence, \mathcal{C} is convex.

Let $\lambda \in (0, 1)$. Assume \mathbf{P} and $\bar{\mathbf{P}}$ achieve the maximal values of Proposition 1 under $(\alpha_{\text{SP}}, \alpha_{\text{EO}}, \alpha_{\text{OAE}})$ and $(\bar{\alpha}_{\text{SP}}, \bar{\alpha}_{\text{EO}}, \bar{\alpha}_{\text{OAE}})$, respectively. We define $\mathbf{P}_\lambda = \lambda\mathbf{P} + (1 - \lambda)\bar{\mathbf{P}}$, which satisfies the constraints of Proposition 1 with thresholds $(\lambda\alpha_{\text{SP}} + (1 - \lambda)\bar{\alpha}_{\text{SP}}, \lambda\alpha_{\text{EO}} + (1 - \lambda)\bar{\alpha}_{\text{EO}}, \lambda\alpha_{\text{OAE}} + (1 - \lambda)\bar{\alpha}_{\text{OAE}})$. Finally, since the objective function of Proposition 1 is a linear function, it is equal to $\lambda\text{FairFront}(\alpha_{\text{SP}}, \alpha_{\text{EO}}, \alpha_{\text{OAE}}) + (1 - \lambda)\text{FairFront}(\bar{\alpha}_{\text{SP}}, \bar{\alpha}_{\text{EO}}, \bar{\alpha}_{\text{OAE}})$ under \mathbf{P}_λ . \square

B.2 Proof of Theorem 1

Proof. The proof relies on Theorem 3 and Lemma 1. For simplicity, we write the conditional $P_{\hat{Y}|S,Y}$ as its corresponding transition matrix \mathbf{P} . Let $\boldsymbol{\mu} = (\Pr(S=1, Y=1), \dots, \Pr(S=A, Y=C))$. The function (10) in our setting can be written as:

$$\mathbf{p}_{\boldsymbol{\mu}}(\hat{y}) = \left(\frac{\mu_{1,1}P_{(1,1),\hat{y}}}{\sum_{s,y}\mu_{s,y}P_{(s,y),\hat{y}}}, \dots, \frac{\mu_{A,C}P_{(A,C),\hat{y}}}{\sum_{s,y}\mu_{s,y}P_{(s,y),\hat{y}}} \right). \quad (26)$$

$$\mathbf{p}_{\boldsymbol{\mu}}(x) = \left(\frac{\mu_{1,1}dP_{X|S=1,Y=1}}{\sum_{s,y}\mu_{s,y}dP_{X|S=s,Y=y}}(x), \dots, \frac{\mu_{A,C}dP_{X|S=A,Y=C}}{\sum_{s,y}\mu_{s,y}dP_{X|S=s,Y=y}}(x) \right). \quad (27)$$

Note that $\mathbf{p}_{\boldsymbol{\mu}}(x) = \mathbf{g}(x)$ due to Bayes' rule. By Lemma 1, we can rewrite \mathcal{C} in Definition 2 as

$$\mathcal{C} = \left\{ \mathbf{P} \mid P_{\hat{X}|S,Y} \text{ is more informative than } \mathbf{P} \right\}. \quad (28)$$

By Lemma 1 and Theorem 3, the above set is further equivalent to all transition matrices $\mathbf{P} \in \mathcal{T}(C|AC)$ satisfying

$$\sum_{\hat{y}=1}^C \phi \left(\frac{\mu_{1,1}P_{(1,1),\hat{y}}}{\sum_{s,y}\mu_{s,y}P_{(s,y),\hat{y}}}, \dots, \frac{\mu_{A,C}P_{(A,C),\hat{y}}}{\sum_{s,y}\mu_{s,y}P_{(s,y),\hat{y}}} \right) \sum_{s,y} \mu_{s,y}P_{(s,y),\hat{y}} \leq \mathbb{E}[\phi(\mathbf{g}(\mathbf{X}))] \quad (29)$$

for any function $\phi : \Delta_{AC} \rightarrow \mathbb{R}$ which is the maximum of finitely many linear functions. Now we can write $\phi(\mathbf{p}) = \max_{i \in [k]} \{ \mathbf{a}_i^T \mathbf{p} \}$ —we ignore the bias term because $\mathbf{a}_i^T \mathbf{p} + b_i = (\mathbf{a}_i + b_i \mathbf{1})^T \mathbf{p}$. Then the inequality in (29) can be simplified as

$$\sum_{\hat{y}=1}^C \max_{i \in [k]} \{ \mathbf{a}_i^T \boldsymbol{\Lambda}_{\mu} \mathbf{p}_{\hat{y}} \} \leq \mathbb{E} \left[\max_{i \in [k]} \{ \mathbf{a}_i^T \mathbf{g}(\mathbf{X}) \} \right], \quad (30)$$

where $\mathbf{p}_{\hat{y}}$ is the \hat{y} -th column of \mathbf{P} and $\boldsymbol{\Lambda}_{\mu} = \text{diag}(\mu_{1,1}, \dots, \mu_{A,C})$. Finally, we can always normalize the above inequality so that each $\mathbf{a}_i \in [-1, 1]^{AC}$. \square

B.3 Proof of Theorem 2

Proof. We denote

$$\begin{aligned} f(\mathbf{P}) &\triangleq \sum_{s=1}^A \sum_{y=1}^C \mu_{s,y}P_{(s,y),y}, \\ g(\mathbf{P}; \mathbf{a}_1, \dots, \mathbf{a}_k) &\triangleq \sum_{\hat{y}=1}^C \max_{i \in [k]} \{ \mathbf{a}_i^T \boldsymbol{\Lambda}_{\mu} \mathbf{p}_{\hat{y}} \} - \mathbb{E} \left[\max_{i \in [k]} \{ \mathbf{a}_i^T \mathbf{g}(\mathbf{X}) \} \right], \\ \mathcal{F} &\triangleq \mathcal{C}_k \cap \{ \mathbf{P} \in \mathcal{T}(C|AC) \mid \text{SP} \leq \alpha_{\text{SP}}, \text{EO} \leq \alpha_{\text{EO}}, \text{OAE} \leq \alpha_{\text{OAE}} \}. \end{aligned}$$

Let \mathcal{F}^t be the constraint set of \mathbf{P} at the t -th iteration of our algorithm. Note that $\mathcal{F} \subseteq \mathcal{F}^t$ by definition. If the algorithm stops at the t -th iteration, then for any $\{ \mathbf{a}_i \mid \mathbf{a}_i \in [-1, 1]^{AC}, i \in [k] \}$, \mathbf{P}^t satisfies

$$g(\mathbf{P}^t; \mathbf{a}_1, \dots, \mathbf{a}_k) \leq 0,$$

which implies $\mathbf{P}^t \in \mathcal{F}$. Consequently,

$$f(\mathbf{P}^t) = \max_{\mathbf{P} \in \mathcal{F}^t} f(\mathbf{P}) \geq \max_{\mathbf{P} \in \mathcal{F}} f(\mathbf{P}) \geq f(\mathbf{P}^t).$$

As a result, $f(\mathbf{P}^t) = \max_{\mathbf{P} \in \mathcal{F}} f(\mathbf{P})$ so \mathbf{P}^t is an optimal solution of $\text{FairFront}_k(\alpha_{\text{SP}}, \alpha_{\text{EO}}, \alpha_{\text{OAE}})$.

If the algorithm never stops, consider any convergent sub-sequence of \mathbf{P}^t that converges to a limit point $\mathbf{P}^* \in \mathcal{T}(C|AC)$. To simplify our notation, we assume $\mathbf{P}^t \rightarrow \mathbf{P}^*$ as $t \rightarrow \infty$. Since $\{\mathcal{F}^t\}_{t \geq 1}$ is non-increasing and they all contain \mathcal{F} , there exists a set \mathcal{F}^* such that

$$\lim_{t \rightarrow \infty} \mathcal{F}^t = \mathcal{F}^*, \quad \mathcal{F} \subseteq \mathcal{F}^*.$$

Therefore, we have

$$f(\mathbf{P}^*) = \lim_{t \rightarrow \infty} f(\mathbf{P}^t) = \lim_{t \rightarrow \infty} \max_{\mathbf{P} \in \mathcal{F}^t} f(\mathbf{P}) = \max_{\mathbf{P} \in \mathcal{F}^*} f(\mathbf{P}).$$

Since $\mathcal{F} \subseteq \mathcal{F}^*$, we have

$$f(\mathbf{P}^*) = \max_{\mathbf{P} \in \mathcal{F}^*} f(\mathbf{P}) \geq \max_{\mathbf{P} \in \mathcal{F}} f(\mathbf{P}).$$

If $\mathbf{P}^* \notin \mathcal{F}$, then there exists a $(\bar{\mathbf{a}}_1, \dots, \bar{\mathbf{a}}_k)$, such that $g(\mathbf{P}^*; \bar{\mathbf{a}}_1, \dots, \bar{\mathbf{a}}_k) > 0$. Let $(\mathbf{a}_{1,t}, \dots, \mathbf{a}_{k,t})$ be the output of Step 2 at t -th iteration. Since $\mathbf{P}^* \in \mathcal{F}^t$ for all t , we have

$$g(\mathbf{P}^*; \mathbf{a}_{1,t}, \dots, \mathbf{a}_{k,t}) \leq 0. \quad (31)$$

By the optimality of $(\mathbf{a}_{1,t}, \dots, \mathbf{a}_{k,t})$, we have

$$g(\mathbf{P}^t; \mathbf{a}_{1,t}, \dots, \mathbf{a}_{k,t}) \geq g(\mathbf{P}^t; \bar{\mathbf{a}}_1, \dots, \bar{\mathbf{a}}_k). \quad (32)$$

Suppose that some sub-sequence of $(\mathbf{a}_{1,t}, \dots, \mathbf{a}_{k,t})$ converges to a vector $(\mathbf{a}_1^*, \dots, \mathbf{a}_k^*)$. For the sake of simplicity, we assume $(\mathbf{a}_{1,t}, \dots, \mathbf{a}_{k,t}) \rightarrow (\mathbf{a}_1^*, \dots, \mathbf{a}_k^*)$ as $t \rightarrow \infty$. On the one hand, taking limit of $t \rightarrow \infty$ on both sides of (32) leads to

$$g(\mathbf{P}^*; \mathbf{a}_1^*, \dots, \mathbf{a}_k^*) \geq g(\mathbf{P}^*; \bar{\mathbf{a}}_1, \dots, \bar{\mathbf{a}}_k).$$

On the other hand, taking limit of $t \rightarrow \infty$ on both sides of (31) leads to

$$g(\mathbf{P}^*; \mathbf{a}_1^*, \dots, \mathbf{a}_k^*) \leq 0.$$

Therefore,

$$0 \geq g(\mathbf{P}^*; \mathbf{a}_1^*, \dots, \mathbf{a}_k^*) \geq g(\mathbf{P}^*; \bar{\mathbf{a}}_1, \dots, \bar{\mathbf{a}}_k) > 0,$$

which is impossible. Therefore, $\mathbf{P}^* \in \mathcal{F}$ and, as a result, we have

$$f(\mathbf{P}^*) = \max_{\mathbf{P} \in \mathcal{F}^*} f(\mathbf{P}) \geq \max_{\mathbf{P} \in \mathcal{F}} f(\mathbf{P}) \geq f(\mathbf{P}^*) \implies \max_{\mathbf{P} \in \mathcal{F}} f(\mathbf{P}) = f(\mathbf{P}^*).$$

□

B.4 Additional Results

We establish basic properties of \mathcal{C} and $\text{FairFront}(\alpha_{\text{SP}}, \alpha_{\text{EO}}, \alpha_{\text{OAE}})$ in the following lemma.

Lemma 3. \mathcal{C} is a convex subset of $\mathcal{T}(C|AC)$ and $\text{FairFront}(\alpha_{\text{SP}}, \alpha_{\text{EO}}, \alpha_{\text{OAE}})$ is a concave function w.r.t. $\alpha_{\text{SP}}, \alpha_{\text{EO}}, \alpha_{\text{OAE}}$. Here the constants A and C denote the number of protected groups and the number of classes.

Next, we discuss a special case— X is discrete—under which \mathcal{C} has a simple characterization.

Remark 3. If X is a *discrete* variable with a *finite support* $[D]$, we can write $P_{X|S,Y}$ as a transition matrix $\Phi \in \mathcal{T}(D|AC)$. By introducing an auxiliary variable $\mathbf{M} \in \mathcal{T}(C|D)$, we can write $\mathbf{P} \in \mathcal{C}$ equivalently as linear constraints: $\mathbf{P} = \Phi \mathbf{M}$ by using the last condition of Lemma 1. Consequently, Proposition 1 boils down to a linear program. However, this characterization fails to generalize to continuous data because Φ and \mathbf{M} will have an infinite dimension; for categorical data, this characterization suffers from the curse of dimensionality since the support size of X grows exponentially fast w.r.t. the number of features.

C Details on the Experimental Results

C.1 Additional Experiments

In this section, we present additional experimental results to further support our findings. We reproduce our experimental results on the German Credit dataset (Bache and Lichman, 2013) and HSLS (High School Longitudinal Study) dataset (Ingels et al., 2011; Jeong et al., 2022) in Figure 4 and Figure 5. In particular, the HSLS dataset experiment is a multi-group, multi-label experiment. Our observation is consistent with those on the previous two datasets—the fairness-accuracy curves given by SOTA fairness interventions, such as Reduction and FairProjection, are close to the information-theoretic limit.

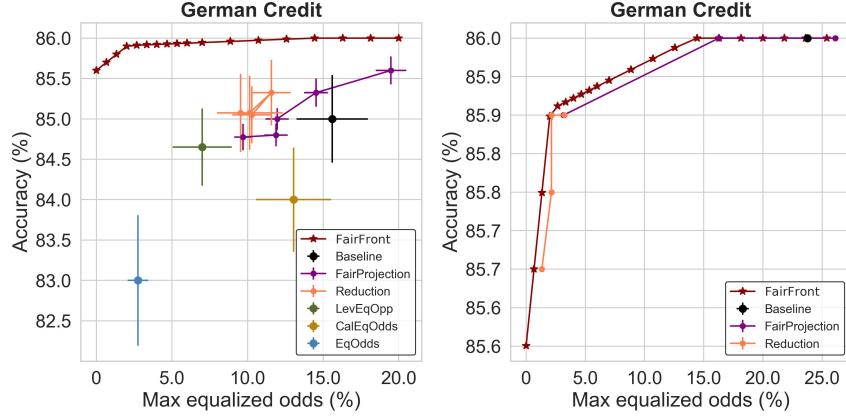


Figure 4: We reproduce our experiments on the German Credit dataset. Our observation is consistent with those on the previous two datasets—the fairness-accuracy curves given by SOTA fairness interventions, such as Reduction and FairProjection, are close to the information-theoretic limit.

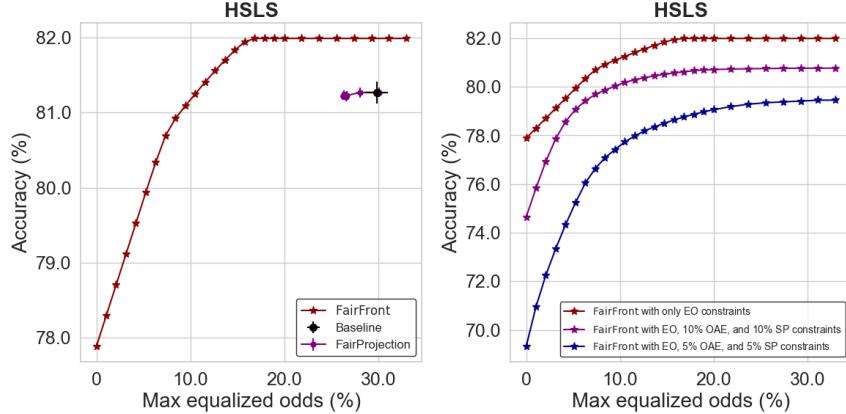


Figure 5: We reproduce our experiments on the HSLS dataset with multi-group and multi-label pre-processing. On the right, we also demonstrate that FairFront can take into account multiple fairness considerations at once. We show how the fairness-accuracy curve changes as we add new types of group fairness constraints (i.e., adding OAE and SP constraints in addition to EO).

C.2 Dataset

Adult. We use sex (female or male) as the group attribute and income ($> 50K$ or $\leq 50K$) as the target for prediction. We use sex, hours-per-week, education-num, age, marital status, relationship status (husband or wife) as the input features—we include the group attribute as an input feature. We group age into a total of 12 disjoint intervals: $[0, 20), [20, 25), \dots, [65, 70), [70, \infty)$; we group hours-per-week into a total of 14 disjoint intervals: $[0, 10), [10, 15), \dots, [65, 70), [70, \infty)$.

COMPAS. We use race (African-American or Caucasian) as the group attribute and is_recid (recid. or no recid.) as the target for prediction. We use race, age, c_charge_degree, sex, priors_count, c_jail_in, c_jail_out as the input features—we include the group attribute as an input feature. We use the last two features by taking their difference to be their length_of_stay. We remove entries where COMPAS case could not be found (is_recid = -1) and entries with inconsistent arrest information. We also binarize sex and remove traffic offenses. We quantize age the same way we do in the Adult dataset and quantize length_of_stay by every 30 days and let 0 be a separate category.

German Credit. We use age (below or above 25 years old) as the group attribute and the credit column, which represents whether the loan was a good decision, as the target for prediction. We use

loan duration in month, credit amount, age, number of existing credits at this bank, sex, credit history, savings, and length of present employment as input features. We include the group attribute age as an input feature. We group credit amount into three disjoint intervals: $[0, 5000]$, $[5000, 10000]$, $[10000, \infty)$. We group duration of loan into two categories: under 36 months and over 36 months.

HSLS. We use race as the group attribute and mathematics test score (number of questions answered correctly out of 72) as the target for prediction. This is a multi-group and multi-label dataset. The entire population is grouped by 4 categories: White, Asian, African American, and Others. We seek to predict the mathematics test performance from a set of attributes, including the scale of student’s mathematical identity, scale of student’s mathematics utility, scale of one’s mathematics self-efficacy, parent’s education, parent’s income, scale of student’s sense of school belonging, race, and sex. Note that we include the group attribute as an input feature. We group the target column (estimated number of questions answered correctly) into a total of 5 disjoint intervals: $[0, 30]$, $[30, 40]$, $[40, 50]$, $[50, 60]$, $[60, \infty)$; we group the scale of student’s mathematical identity, mathematics utility, mathematics self-efficacy, and sense of school belonging into a total of 4 disjoint intervals, characterized by standard deviations away from the mean: $(-\infty, -1)$, $[-1, 0)$, $[0, 1)$, $[1, \infty)$.

C.3 Benchmark

Each benchmark method’s hyper-parameter values are provided below. Each point in Figure 2 for Baseline, EqOdds, CalEqOdds, Reduction, LevEqOpp, and FairProjection is obtained by applying the obtained classifier to 10 different test sets. For the Adult dataset, we use Random Forest with $n_estimators=15$, $min_samples_leaf=3$, $criterion = \text{log_loss}$, $bootstrap = \text{False}$ as our baseline classifier; for the COMPAS dataset, we use Random Forest with $n_estimators = 17$ as our baseline classifier. For the German Credit dataset, we use Random Forest with $n_estimators=100, min_samples_split = 2, min_samples_leaf=1$ as our baseline classifier. They are all implemented by Scikit-learn ([Pedregosa et al., 2011](#)).

EqOdds ([Hardt et al., 2016](#)). We use AIF360 implementation of EqOddsPostprocessing and the default hyper-parameter setup.

CalEqOdds ([Pleiss et al., 2017](#)). We use AIF360 implementation of CalibratedEqOddsPostprocessing and the default hyper-parameter setup.

Reduction ([Agarwal et al., 2018](#)). We use AIF360 implementation of ExponentiatedGradientReduction. We vary the allowed fairness constraint violation $\epsilon \in \{0.001, 0.01, 0.2, 0.5, 1, 2, 5, 10, 15\}$ for Adult dataset and $\epsilon \in \{0.001, 0.01, 0.2, 0.5, 1, 2, 5, 10, 15\}$ for Adult with missing values. We vary $\epsilon \in \{0.001, 2, 5, 10, 15, 20, 25, 30, 35, 40\}$ for COMPAS to obtain a fairness-accuracy curve, and $\epsilon \in \{0.001, 0.1, 0.5, 1, 2, 7, 8, 10, 15, 20, 25, 30\}$ for COMPAS with 50% missing values in the minority group. We use $\epsilon \in \{20, 50, 80, 95\}$ for German Credit dataset and $\epsilon \in \{5, 8, 10, 20, 23\}$ when using Bayes Optimal classifier.

LevEqOpp ([Chzhen et al., 2019](#)). We use the Python implementation of LevEqopp from the Github repo in [Alghamdi et al. \(2022\)](#). We follow the same hyperparameters setup as in the original method.

FairProjection [Alghamdi et al. \(2022\)](#). We use the implementation from the Github repo in [Alghamdi et al. \(2022\)](#) and set `use_protected = True`. We use Random Forest with $n_estimators = 17$ as the baseline classifier to predict S from (X, Y) . We set the list of fairness violation tolerance to be $\{0.07, 0.075, 0.08, 0.085, 0.09, 0.095, 0.1, 0.5, 0.75, 1.0\}$ for Adult dataset and $\{0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.1, 0.5, 1.0\}$ for COMPAS dataset to obtain a fairness-accuracy curve. We set the list of fairness violation tolerance to be $\{0.005, 0.01, 0.02, 0.07, 0.1, 0.15\}$ on the German Credit dataset experiment, and $\{0.0001, 0.001, 0.005, 0.01, 0.015, 0.02, 0.05\}$ when using a Bayes optimal baseline classifier. For the HSLS dataset, the list of tolerance is $\{0.06, 0.065, 0.07, 0.08, 0.1, 0.09, 0.15, 0.2, 0.3\}$.

	Multiclass	Multigroup	Avoid disparate treatment	Multi-constraint	Curve
Hardt et al. (2016)	✗	✓	✗	✗	✗
Corbett-Davies et al. (2017)	✗	✓	✗	✗	✗
Menon and Williamson (2018)	✗	✗	✓	✗	✓
Chzhen et al. (2019)	✗	✗	✓	✗	✗
Yang et al. (2020)	✓	✓	✗	✓	✓
Zeng et al. (2022a)	✗	✓	✗	✗	✓
Zeng et al. (2022b)	✗	✓	✗	✗	✗
Our approach	✓	✓	✓	✓	✓

Table 2: Comparison with existing work that investigate the fairness Pareto frontier. **Multiclass/multigroup**: can handle multiclass classification problems with multiple protected groups; **Avoid disparate treatment**: not require the classifier to use group attributes as an input variable; **Multi-constraint**: can handle multiple (group) fairness constraints simultaneously; **Curve**: produce fairness-accuracy trade-off curves (instead of a single point).

D More on Related Work

We provide a detailed comparison with existing work on fairness Pareto frontier and ML uncertainty in this section.

D.1 Fairness Pareto Frontier

We present in Table 2 a detailed comparison of our approach with previous studies that have investigated the fairness Pareto frontier and fair Bayes optimal classifier. In short, our approach is different from this line of research as it simultaneously combines several important aspects: it is applicable to *multiclass* classification problems with *multiple* protected groups; it *avoids disparate treatment* by not requiring the classifier to use group attributes as an input variable; and it can handle *multiple* fairness constraints simultaneously and produce fairness-accuracy trade-off curves (instead of a single point).

D.2 Aleatoric and Epistemic Uncertainty

In this paper, we divide algorithmic discrimination into aleatoric and epistemic discrimination. We borrow this notion from ML uncertainty literature (see [Hüllermeier and Waegeman, 2021](#), for a survey). Here we provide a detailed comparison between them.

In terms of their definitions, epistemic uncertainty arises from a lack of knowledge about the best model, such as the Bayes predictor, while epistemic discrimination results from a lack of knowledge about the optimal “fair” predictive model. On the other hand, aleatoric uncertainty is the irreducible part of uncertainty caused by the random relationship between input features and label, while aleatoric discrimination is due to inherent biases in the data-generating distribution.

In terms of their characterization, epistemic uncertainty can in principle be reduced by including additional information (e.g., more data); epistemic discrimination can be reduced in a similar approach since a data scientist can choose a more effective fairness-intervention algorithm with access to more information.

Finally, in the infinite sample regime, a consistent learner will be able to remove all epistemic uncertainty, assuming the model class is large enough and there are no computational constraints. Analogously, we demonstrate in Figure 1 that when the underlying distribution is known, SOTA fairness interventions are able to eliminate epistemic discrimination as their fairness-accuracy curves are close to the fair front.

E More on Future Work

In this paper, we present an upper bound estimate for FairFront in Algorithm 1. It is important to note that this estimate may be subjected to errors originating from various sources. These include (i)

the approximation error of the function g , (ii) estimation errors from computing the expectation in (7) with a finite dataset, and (iii) the influence of hyperparameters, T (number of running iterations of Algorithm 1) and k (number of segments in the piece-wise linear functions). Regarding the dependence on T , our Theorem 2 ensures the algorithm's asymptotic convergence as $T \rightarrow \infty$. However, we have not established a proof for its behavior at a finite T . Regarding the dependence on k , we conjecture that $k = A * C$ should suffice, where A is the number of protected groups and C is the number of labels. While Blackwell proved this result for $k = 2$ in Theorem 10 of [Blackwell \(1953\)](#), an extension of this proof to a general value of k appears to remain an open problem.

We define aleatoric and epistemic discrimination with respect to the entire population. Investigating their per-instance counterparts and the relationship to individual fairness would be a compelling area of future research. Additionally, a more nuanced analysis of aleatoric and epistemic discrimination is desirable, further breaking them down into fine-grained components. For instance, epistemic discrimination may be attributed to various factors including limited training data, noisy observations of labels or sensitive attributes, and limitations of learning algorithms. Finally, investigating other criteria, such as scalability, generalization, and robustness in evaluating existing fairness interventions is a significant topic for future exploration.