UNIVERSITY OF CALIFORNIA
SANTA CRUZ

THE SYMPHONY OF ALIGNMENT: ENSURING FAIRNESS AND MITIGATING BIAS IN FOUNDATION MODELS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE AND ENGINEERING

by

Jialu Wang

December 2024

The Dissertation of Jialu Wang
is approved:

Professor Yang Liu, Chair

Professor Lise Getoor

Professor Xin Eric Wang

Professor Kai-Wei Chang

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

The Symphony of Alignment: Ensuring Fairness and Mitigating Bias in Foundation Models

by

Jialu Wang

Foundation models are poised to revolutionize decision-making across various domains, but their reliance on historical data can perpetuate and amplify existing biases. This risk of reinforcing societal stereotypes through biased outputs underscores the critical need to evaluate and mitigate biases in these models to ensure their responsible and ethical use. In this dissertation, we delve into three critical challenges in ensuring fairness and mitigating bias in foundation models and AI systems. It comprises three main contributions: (1) An exploration of fair learning under uncertainty, particularly when sensitive attributes are corrupted. The research proposes noise-resistant fair Empirical Risk Minimization approaches and a novel method for detecting groups with higher noise levels in labels. (2) An investigation into fairness and bias in multimodal applications of foundation models, including image search, multilingual text retrieval, and text-to-image generation. The study develops new intervention methods for mitigating gender bias in image search, reveals intrinsic trade-offs in multilingual fairness, and introduces association test in text-to-image generations. (3) The development of fairness influence functions to quantify the impact of individual data examples on model

fairness. This approach offers insights into machine unlearning, with efficient approximation techniques for large-scale applications. Ultimately, the thesis strives to advance the understanding of fairness in foundation models through the development of both theoretical frameworks and practical evaluations for responsible AI.

To the Ubermensch,

for sustaining endless nights of stress and solitude.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my wonderful advisor, Professor Yang Liu, whose guidance, patience, and unwavering support have been instrumental throughout my doctoral journey. His profound insights and constructive criticism have not only shaped this thesis but have fundamentally influenced my approach to research. I am equally indebted to my co-advisor, Professor Xin Eric Wang, whose expertise and encouragement helped me navigate the most challenging aspects of my research.

I extend my sincere appreciation to my thesis committee members, Professor Lise Getoor and Professor Kai-Wei Chang, for their valuable feedback and thought-provoking questions that have substantially improved the quality of this work. Their diverse perspectives have enriched my research in ways I could not have anticipated.

This journey would not have been possible without the support of my lab members. To every one in REAL lab, including Yatong Chen, Zhaowei Zhu, Jiaheng Wei, Reilly, Raab, Zonglin Di, Jinlong Pang, Chris Liu, and Yaxuan Wang, and every one in ERIC lab, including Yue Fan, Xuehai He, Kaizhi Zheng, Kaiwen Zhou, Minghao Liu, and Jing Gu, your friendship and intellectual companionship have made these years truly memorable.

I would also like to thank my friends outside UC Santa Cruz. Their endless trusts and cares gives me the strength to persist in my PhD journey. Especially, I

would like to thank Jingkang Wang, Ruijie Wang, Hongye Jin, Yifan Qiao for generously sharing their rooms during my travels.

Finally, I would like to thank my parents, who instilled in me the value of education and supported my dreams unconditionally. To my girlfriend Wenwen, who has been my moral support throughout the five years, thank you for your endless patience, understanding, and support. Their beliefs in me, especially during moments when I doubted myself, have been my greatest source of strength.

This achievement belongs not just to me, but to all of you who have been part of this remarkable journey.

Chapter 1

Introduction

1.1 Motivation

Foundation models, also known as large AI models, are pre-trained on vast amount of data and empower a wide range of applications. These large-scale artificial intelligence models are poised to revolutionize digital decision-making process across various domains, including recommender systems, credit scoring, and medical treatments. The effectiveness of foundation models stems from their analysis of historical data. However, this reliance on past information can be problematic when the training data contains biases or errors. In such cases, the models may inadvertently perpetuate and amplify these flaws, raising concerns about fairness and ethics. For example, when interacting with large language models, users may receive responses that reflect societal stereotypes related to gender, race, religion, and other sensitive topics. The dissemina-

tion of such biased information risks reinforcing these stereotypes in human's minds. In consequence, it is crucial to evaluate and mitigate such biases to ensure the responsible usage of foundation models.

Despite recent advances in related areas, the field of machine learning still faces numerous challenges on multiple fronts.

Ensuring Fairness in an Uncertain World. Many current machine learning methods assume that historical data is clean and accurate. However, in reality, model developers often only have access to biased datasets. It is therefore crucial to develop techniques that can mitigate biases stemming from unreliable or uninformative labels.

Bridging the Gap to Multi-Modal Foundation Models. Algorithmic fairness has been conceptualized and evaluated primarily in the context of classification tasks. However, with rapid development of AI models, this narrow focus is becoming increasingly inadequate. Modern foundation models, which serve as the basis for numerous AI applications, are trained on and capable of processing a diverse array of data modalities, including but not limited to text, images, audio, and video. This multi-modal nature of contemporary AI systems presents new challenges in ensuring and evaluating fairness across a broad spectrum of downstream applications.

Lacking the Interpretation from a Data-Centric Viewpoint. While recent machine learning literature has produced numerous in-processing methods for bias mitiga-

tion in AI systems, a critical gap remains in our understanding of these interventions from a data-centric perspective. The application of these techniques without a comprehensive grasp of how biased data samples influence model behavior can yield unreliable or even counterproductive results. This challenge is exacerbated by the emergence of foundation models, which introduce unprecedented complexities to the fairness landscape. These large-scale models pose a significant risk of memorizing training examples verbatim, including those tainted with biases. More alarmingly, due to their immense scale and training methodologies, foundation models may not merely reflect existing biases but potentially amplify them, creating a magnified echo chamber of unfairness. This underscores the urgent need for a deeper, data-centric approach to understanding and mitigating bias in AI systems, particularly navigating the specific corrupted data instances.

1.2 Overview of Results

This thesis aims to address the aforementioned challenges as follows:

In Chapter 2, we explore the challenge of fair learning in scenarios with imperfect information, specifically focusing on cases where sensitive attributes may be corrupted. In realistic applications, such as criminal justice and evaluating loan applications, labels are often contaminated by human biases against a certain protected group. Our research reveals, through both theoretical analysis and empirical evidence, that

blindly enforcing parity constraints without considering noisy labels can be detrimental. These findings underscore the critical importance of accounting for noise and bias when performing Empirical Risk Minimization (ERM) subject to fairness constraints. To address these issues, we developed two noise-resistant fair ERM approaches. The core concept involves constructing unbiased estimators for both loss functions and fairness constraints. Additionally, we prototype a method that strategically increases noise levels to balance disparities and further mitigate biases. We propose a novel detection method that identifies groups of labels likely to suffer from higher noise levels without relying on ground truth information. Our experimental results demonstrate that the aforementioned harms can indeed manifest in practice when using real datasets. The insights forewarn decision-makers that improperly mitigating unfairness might do harm on the clean groups. Our two fairness-aware solutions are an important step toward addressing this problem.

Chapter 3 explores fairness and bias in three key applications of foundation models: image search, multilingual text retrieval, and text-to-image generation. Our research aims to establish a unified framework for evaluating fairness and bias across both image and text modalities, extending traditional fairness criteria beyond their original classification context into these modern, multimodal domains. In addressing image search, we introduce two novel intervention methods: an in-processing approach and a post-processing technique, both designed to mitigate biases in search results. These methods demonstrate practical strategies for enhancing fairness in visual information

retrieval systems. Our investigation into multilingual text retrieval reveals a fundamental tension in fairness objectives. We demonstrate the impossibility of simultaneously achieving both individual fairness and group fairness in this context, highlighting the complex trade-offs inherent in multilingual AI systems. For text-to-image generation, our research focuses on state-of-the-art diffusion models. We provide evidence of bias amplification in these models, illustrating how existing biases can be exacerbated in the process of generating images from textual descriptions. This chapter represents a significant step towards understanding and addressing fairness issues in multimodal AI applications. Our findings not only shed light on the challenges of ensuring fairness across different modalities but also offer insights into potential solutions and areas requiring further research in the rapidly evolving field of multimodal AI.

Chapter 4 delves into the application of influence functions as a means to quantify the impact of individual data examples on model fairness. Our result introduces a novel family of fairness influence functions, designed to measure the change in fairness metrics when a specific training example is counterfactually removed from the training data. Influence function allows for a granular understanding of each data point's contribution to the overall fairness of the model. As the direct application, we show that such influence function can be used in coreset selection, noise detection, and machine unlearning. To address the computational complexity of influence functions, we develop efficient approximation techniques, making the approach feasible for large-scale datasets and complex foundation models. Recognizing that direct access to model parameters

is not always possible, we propose methods to estimate fairness influence functions in scenarios where only model predictions are available.

Chapter 2

Fair Classification with Imperfect Information

2.1 Motivation

Machine learning classifiers can perpetuate and amplify existing systemic injustices in society. Notable examples include discrepancies in allocation of medical care to patients on the basis of race [OPVM19] and significant disparities in predicting recidivism rates for African-American defendants [ALMK16a, Cho17a], and more [VBC18, PF16, BG18a]. A number of techniques have been developed in order to mitigate bias in machine learning classifiers [ZVRG17, FFM$^+$15a, HPS16a, ABD$^+$18a, MW18, CHKV19]. Typically, these methods consider populations with groups corresponding to a set of protected sensitive attributes, such as race or gender. The classifier is then re-

quired to exhibit similar behavior across all groups [ZVRG17, HPS16a, Cho17a, KC09]. This can be done by imposing equality of true positive rate or true negative rate conditioned on group membership. These are called "fairness," or parity constraints.

Many of these methods assume the availability of clean and accurate labels. However, this is often not the case. In fact, bias in data is particularly pertinent to label corruption. To make things worse, the accuracy of available labels is often strongly influenced by whether a person falls within a protected group, and these discrepancies can have significant and often life-altering outcomes. For example, it has been shown that labels for criminal activity generated via crowdsourcing are systematically biased against certain racial groups [DF18]. As another example, both women and lower-income individuals often receive significantly less accurate diagnoses for cancer and other ailments than men, due to imbalance in the sample population of medical trials [GTYS18], and due to bias from doctor treatment [Bra16]. Similar discrepancies arise in the accuracy of mathematical aptitude evaluations for males and females in primary school [LHPL10], and it has long been known that an employer's evaluation of a resume will be influenced by the perceived ethnic origin of an applicant's name [BM04]. Moreover, studies show that people of all races use and sell illegal drugs at remarkably similar rates, but in some states, black male have been admitted to prison on drug charges at rates twenty to fifty times greater than those of white men [Ale12]. The structure and magnitude of group-specific label noise can dramatically affect the performance *and* fairness of a classifier. To see this, we consider the following examples.

		Group A			Group B			Pooled		
(x_1, x_2),	y	$+1$	-1	f_A^*	$+1$	-1	f_B^*	$+1$	-1	f_{fair}^*
$(0, 0)$,	-1	0	25	-1	70	30	$+1$	70	55	$+1$
$(0, 1)$,	-1	0	25	-1	70	30	$+1$	70	55	-1
$(1, 0)$,	$+1$	25	0	$+1$	100	0	$+1$	125	0	$+1$
$(1, 1)$,	$+1$	25	0	$+1$	100	0	$+1$	125	0	-1

Example 2.1 Consider training classifiers using data from two groups $Z \in \{A, B\}$ with homogeneous data distributions $\Pr(Y = +1 \mid X = \boldsymbol{x}, Z = A) = \Pr(Y = +1 | X = \boldsymbol{x}, Z = B)$, where $\boldsymbol{x} = [x_1, x_2]$ is a 2-dimensional feature vector. In this setting, the Bayes-optimal classifiers for group A and B (denoted as f_A^* and f_B^* respectively) will obey any parity constraint. However, suppose group a has a set of clean labels, while group b has clean labels when the ground truth is $y = +1$ but there is a 70% chance that corrupting noise will cause the observed label to be flipped from the true value when $y = -1$. In this case, f_{fair}^* trained on both groups achieves perceived equal True Positive Rates (TPR) (50%) between the two groups and is the best one to do so - this indeed hurts group a's prediction performance (as opposed to 100% accuracy before), but the labels in group a are not affected by noise.

Example 2.2 Consider training classifiers using data from two groups $Z \in \{A, B\}$ with heterogeneous data distributions $\Pr(Y = +1 \mid X = \boldsymbol{x}, Z = A) = \Pr(Y = +1 \mid$

$X = \boldsymbol{x}, Z = B$). Suppose group A has a set of clean labels, while one quarter of group B's labels are incorrect. We denote the Bayes-optimal classifiers for A and B as f_A^* and f_B^* respectively and they obey any parity constraint. The classifier f_{fair}^* trained on the observed corrupted data is subject to equal TPR constraint for both groups. Note that f_{fair}^* on the pooled data output +1 for $(0, 1)$ and -1 for $(1, 0)$ because equal TPR constraint is enforced. In this case, the TPRs for both groups are 50%. If the classifier output -1 for $(0, 1)$ and +1 for $(1, 0)$ instead, the TPR for group A is 100% while the TPR for group B is only 50%, which violates the equal TPR constraint. However, f_{fair}^* has a higher TPR (2/3: 200 correct predictions out of 300 true +1 labels) on B than on A (1/2: 100 correct predictions out of 200 true +1 labels) when evaluated on the clean data.

	Group A			Group B			Pooled		
(x_1, x_2), (y_A, y_B)	$+1$	-1	f_A^*	$+1$	-1	f_B^*	$+1$	-1	f_{fair}^*
$(0,0)$, $(-1,-1)$	0	100	-1	75	225	-1	75	325	-1
$(0,1)$, $(-1,+1)$	0	100	-1	75	25	$+1$	75	125	$+1$
$(1,0)$, $(+1,+1)$	100	0	$+1$	75	25	$+1$	175	25	-1
$(1,1)$, $(+1,+1)$	100	0	$+1$	75	25	$+1$	175	25	$+1$

To sum up, the above examples has delivered two important messages:

1. Enforcing fairness constraints without accounting for group-specific label noise can

harm the accuracy of the classifier for the group whose labels have been accurate recorded. We remark that although [BS19] has considered the single-group noise setting and demonstrated that fairness interventions could aid in reducing the error caused by label bias, our observation demonstrates a special case where potential harm occurs.

2. A classifier may appear to achieve parity when it does not. Furthermore, imposing a parity constraint might actually make everyone worse off.

In this chapter, we look at the problem of fair classification from data whose labels are corrupted, such that the error rates of corruption are group-dependent. Several recent works deal with fair classification with noisy labels [JN19, LZMV19, BS19]. In particular, it has been shown that fairness constraints on the noisy training labels can be beneficial when the label noise is homogeneous across the different groups that are to be protected [BS19]. More recently, [FCG20] shows that how the true fairness rates, such as TPR, are related to observed quantities with respect to noise parameters. Our work complements these results: we show that enforcing fairness constraints when training on data with noisy labels produces a classifier that violates the fairness constraints as measured with respect to the clean data. We then provide a fair empirical risk minimization (ERM) framework that handles heterogenous label noise. Our framework uses an estimation procedure that infers the knowledge of group-dependent noise in the training data and applies this knowledge using bias removal techniques, thus

eliminating the effects of noisy labels in both the objective function and the fairness constraints in expectation.

2.2 Enforcing Fairness Constraints on Noisy Labels Can be Harmful

We start with a dataset with n examples $(\boldsymbol{x}_i, y_i, z_i)_{i=1}^n$, where each example consists of a *feature vector* $\boldsymbol{x}_i = (1, x_{i,1}, \ldots, x_{i,d}) \in \mathbb{R}^{d+1}$, a *label* $y_i \in \{+1, -1\}$, and a *group attribute* $z_i \in \mathcal{Z}$. We assume that there are $m = |\mathcal{Z}| \geq 2$ groups.

Example 2.1 illustrated how blindly imposing parity constraints on the corrupted labels could reduce the classifier's accuracy for unaffected groups. We aim to establish a theoretical basis for investigating the potential harms caused by label errors. Without loss of generality, we present our results in settings where we wish to train a classifier with equal true positive rates (TPR) across groups. Similar derivations hold for other related constraints (e.g., the ones as linear combinations of the entries in the confusion matrix), such as equal false positive rates (FPR), and equal balance error (BER). We use the following shorthand to denote different measures of performance, including TPR and FPR, computed for each group using the true labels y and the noisy

labels $\tilde{y}$, where $y, \tilde{y} \in \{+1, -1\}$:

$$\begin{aligned}
\mathrm{TPR}_z &\coloneqq \Pr(f(X) = +1 \mid Y = +1, Z = z) \\
\mathrm{FPR}_z &\coloneqq \Pr(f(X) = +1 \mid Y = -1, Z = z) \\
\widetilde{\mathrm{TPR}}_z &\coloneqq \Pr(f(X) = +1 \mid \widetilde{Y} = +1, Z = z) \\
\widetilde{\mathrm{FPR}}_z &\coloneqq \Pr(f(X) = +1 \mid \widetilde{Y} = -1, Z = z)
\end{aligned} \tag{2.1}$$

We consider a classification problem with two identical groups z and z' where samples from group z have uncorrupted labels while samples from group z' have noisy labels. A noisy label $\tilde{y}$ corresponds to a true label y that may have been flipped based on noise rate $0 \leq \epsilon_z^+ + \epsilon_z^- < 1$ (as a function of true label y). More precisely, we assume that the noise rates vary based on the true label y as well as the group attribute z:

$$\epsilon_z^+ = \Pr(\tilde{Y} = -1 \mid Y = +1, Z = z), \tag{2.2}$$

$$\epsilon_z^- = \Pr(\tilde{Y} = +1 \mid Y = -1, Z = z) \tag{2.3}$$

That is, the observed training labels are generated as:

$$\tilde{y}_i = \begin{cases} y_i & \text{w.p.} \quad 1 - \epsilon_{z_i}^{\mathrm{sign}(y_i)}, \\ -y_i & \text{w.p.} \quad \epsilon_{z_i}^{\mathrm{sign}(y_i)}. \end{cases}$$

We next show that the label noise presented in group z' can harm the clean group z when enforcing parity constraints.

Theorem 2.1. *Consider a setting with two identical groups $Z = z$ and $Z = z'$. Group z has clean labels, i.e., $\epsilon_z^+ = \epsilon_z^- = 0$. Group z' suffers from symmetric noise $\epsilon_{z'}^+ =$*

$\epsilon^{-}_{z'} = e > 0$. *In this setting, a classifier trained subject to the equal TPR constraint* $\widetilde{\text{TPR}}_z = \widetilde{\text{TPR}}_{z'}$ *leads to an uninformative classifier that* $\text{TPR}_z = \text{FPR}_z$.

Table 2.1: Label noise harms accuracy.

Metrics	Groups	f		f_{fair}
TPR	$female$	97.12%	$\Rightarrow$	96.44%
	$male$	92.40%	$\Rightarrow$	98.26%
FPR	$female$	53.35%	$\Rightarrow$	78.11%
	$male$	46.81%	$\Rightarrow$	84.32%
Accuracy	$female$	91.62%	$\Rightarrow$	88.32%
	$male$	80.39%	$\Rightarrow$	72.97%

We empirically examine the above observation on the Adult dataset from UCI Machine Learning repository [DG17a]. There are two sensitive groups, $Z = \{male, female\}$, in this data set. We inject symmetric noise $\epsilon^{+} = \epsilon^{-} = 0.3$ into labels for members of the $female$ group. Then, we train two classifiers: f, which is trained without any fairness constraints, and f_{fair}, which is trained with the imposition of equal TPR using the reduction method [ABD^{+}18a]. As is shown in Table 2.1, the empirical results mirror Theorem 2.1. When the difference between f_{fair}'s TPR for the two groups becomes small (less than 2%), f_{fair}'s TPR and FPR become close together, and the accuracy decreases significantly. The above trends hold even when we try to

equalize TPR and FPR together across groups. We notice that the two groups are not strictly identical in the Adult dataset, but our example implies that there exists dangerous cases where enforcing fairness constraints can harm classifier accuracy for the group with uncorrupted labels.

Our second message, as illustrated in Example 2.2, is that training fair classifiers using noisy labels may lead to a false impression of fairness. This arises when the fairness constraints are satisfied over the noisy labels while being violated over the clean labels. Before proceeding, we require extending Proposition 16 of [MVROW15] into the situation with group-dependent label noise. We note that a similar result appears in [SBH13].

Lemma 2.2. *For each group z we have that*

$$\mathrm{TPR}_z = (1 - \epsilon_z^+) \cdot \widetilde{\mathrm{TPR}}_z + \epsilon_z^+ \cdot \widetilde{\mathrm{FPR}}_z \tag{2.4}$$

$$\mathrm{FPR}_z = \epsilon_z^- \cdot \widetilde{\mathrm{TPR}}_z + (1 - \epsilon_z^-) \cdot \widetilde{\mathrm{FPR}}_z \tag{2.5}$$

We also note that, in the special case where all groups suffer from an identical rate of label corruption, the learner *can* be oblivious to the specific error rates:

Theorem 2.3. *Consider a classification problem with noisy labels where the noise rates are independent of group membership, so that $\epsilon_z^+ = \epsilon_{z'}^+$ and $\epsilon_z^- = \epsilon_{z'}^-$ $\forall z, z' \in Z$. Then it follows that* $\mathrm{TPR}_z = \mathrm{TPR}_{z'}$ *$\forall z, z' \in Z$, if equal odds (equalizing both* TPR *and* FPR*) on the noisy labels is imposed.*

The proof follows by applying the assumption of equal error rates and equal odds on the noisy labels with Lemma 2.2. However, things break down in the general case. If we impose equal odds across groups on a learner that is unaware of the labels' noisiness (i.e. whenever $\widetilde{\mathrm{TPR}}_z = \widetilde{\mathrm{TPR}}_{z'}$), then:

Theorem 2.4. *Assume that a classifier is subject to equal odds in the presence of group-dependent label noise. Then for any two groups $z, z' \in Z$, we have*

$$|\,\mathrm{TPR}_z - \mathrm{TPR}_{z'}\,| = |\widetilde{\mathrm{TPR}}_z - \widetilde{\mathrm{FPR}}_z| \cdot |\epsilon_z^+ - \epsilon_{z'}^+|,$$

$$|\,\mathrm{FPR}_z - \mathrm{FPR}_{z'}\,| = |\widetilde{\mathrm{TPR}}_z - \widetilde{\mathrm{FPR}}_z| \cdot |\epsilon_z^- - \epsilon_{z'}^-|.$$

Unless the classifier is random on the noisy training data, i.e., $\widetilde{\mathrm{TPR}}_z = \widetilde{\mathrm{FPR}}_z$, it is impossible to satisfy equal odds over the clean data whenever $\epsilon_z^+ \neq \epsilon_{z'}^+$ and $\epsilon_z^- \neq \epsilon_{z'}^-$.

The proof follows by a direct application of Lemma 2.2. Theorem 2.4 implies that the true fairness violation is proportional to the difference in error rates across the different sub-groups. We offer two remarks. First, if the error rates are systematically biased towards a particular group, then a perceived fair classifier will lead to unequal odds. Second, the above bias will be reinforced when the trained model is more accurate on noisy data; a more accurate model will lead to a larger difference in $|\widetilde{\mathrm{TPR}}_z - \widetilde{\mathrm{FPR}}_z|$.

Table 2.2: Surrogate constraints for surrogate loss.

Metric	$\widehat{F}_z(f)$
TPR	$(1-\epsilon_z^+)\cdot\widehat{\mathrm{TPR}}_z + \epsilon_z^+\cdot\widehat{\mathrm{FPR}}_z$
FPR	$\epsilon_z^-\cdot\widehat{\mathrm{TPR}}_z + (1-\epsilon_z^-)\cdot\widehat{\mathrm{FPR}}_z$
Equal Odds	both TPR and FPR

2.3 Fair ERM with Noisy Labels

In this section, we describe two noise-tolerant and fair ERM solutions that address the combined challenges of heterogeneous and group-dependent label noise. Both the surrogate loss and group-weighted peer loss approaches for handling noisy labels rely on estimations of the label noise.

2.3.1 A Surrogate Loss Approach

As we shall see, training an unmodified loss function using the noisy labels $\tilde{y}_i$ corrupts the model in a manner that cannot be addressed via post-hoc correction. Thus, a natural resolution is to modify the loss function itself. This modified loss function is called a *surrogate loss*.

Bias removal surrogate loss functions. Bias removal via a surrogate loss is a popular approach to handling label noise [NDRT13]. The original loss function $\ell(\cdot)$ is replaced with a surrogate loss function $\tilde{\ell}(\cdot)$ that "corrects" for noise in the labels in

expectation. Formally, the surrogate loss is chosen so that the cost of mis-classifying an element $\boldsymbol{x}_i$ with true label y_i is equivalent to the expected loss value that arises from using noisy label $\tilde{y}_i$. Thus, we want to find a surrogate loss $\tilde{\ell}$ such that:

$$\ell(f(\boldsymbol{x}), y) = \mathbb{E}_{\tilde{Y}}[\tilde{\ell}(f(\boldsymbol{x}), \tilde{Y}) \mid Y = y] \tag{2.6}$$

for all $\boldsymbol{x}$ and y. When the noise depends on the label value, the function given by

$$\tilde{\ell}(f(\boldsymbol{x}_i), \tilde{y}_i = +1) \coloneqq \frac{(1 - \epsilon_{z_i}^-)\ell(f(\boldsymbol{x}_i), +1) - \epsilon_{z_i}^+ \ell(f(\boldsymbol{x}_i), -1)}{1 - \epsilon_{z_i}^+ - \epsilon_{z_i}^-}, \tag{2.7}$$

$$\tilde{\ell}(f(\boldsymbol{x}_i), \tilde{y}_i = -1) \coloneqq \frac{(1 - \epsilon_{z_i}^+)\ell(f(\boldsymbol{x}_i), -1) - \epsilon_{z_i}^- \ell(f(\boldsymbol{x}_i), +1)}{1 - \epsilon_{z_i}^+ - \epsilon_{z_i}^+}. \tag{2.8}$$

satisfies the above property, as shown by Lemma 1 in [NDRT13]. A classifier f minimizing the surrogate loss on noisy data $\tilde{\ell}(X, \tilde{Y})$ will minimize the loss on clean data $\ell(X, Y)$ in expectation. This property allows us to perform model selection on a noisy validation set, and one could choose the model that performs better on the validation set to deploy.

Surrogate fairness constraints. We will also need to modify the fairness constraints to account for the effects of noise. Our method of doing so is inspired by the surrogate loss that we need to work with an unbiased estimate of the fairness constraints. For the case of binary classification, we can express the surrogate measures of group-based fairness constraints using Lemma 2.2.

We use Equation (2.7) and Equation (2.8) to define our surrogate loss functions $\tilde{\ell}_z(f(\boldsymbol{x}_i), \tilde{y}_i = +1)$, and $\tilde{\ell}_z(f(\boldsymbol{x}_i), \tilde{y}_i = -1)$. Furthermore, define the empirical TPR and

Table 2.3: Surrogate constraints for group weighted peer loss

Metric	$\widehat{F}_z(f)$
TPR	$\Pr(f(X)=+1\|Z=z) + \frac{\Delta_z}{2}(\widehat{\mathrm{TPR}}_z - \widehat{\mathrm{FPR}}_z)$
FPR	$\Pr(f(X)=+1\|Z=z) - \frac{\Delta_z}{2}(\widehat{\mathrm{TPR}}_z - \widehat{\mathrm{FPR}}_z)$
Equal Odds	both TPR and FPR

FPR over the noisy labels as follows:

$$\widehat{\mathrm{TPR}}_z(f) = \frac{\#(f(\boldsymbol{x}_i)=+1, \tilde{y}_i=+1, z_i=z)}{\#(\tilde{y}_i=+1, z_i=z)} \tag{2.9}$$

$$\widehat{\mathrm{FPR}}_z(f) = \frac{\#(f(\boldsymbol{x}_i)=+1, \tilde{y}_i=-1, z_i=z)}{\#(\tilde{y}_i=-1, z_i=z)} \tag{2.10}$$

We then define our surrogate fairness measures $\widehat{F}_z(f)$ using only noisy data, as detailed in Table 2.2. Our noise-resistant fair ERM states as follows:

$$\begin{aligned} &\min_{f\in\mathcal{H}} \sum_{i=1}^{n} \tilde{\ell}(f(\boldsymbol{x}_i), \tilde{y}_i) \\ &\text{s.t.} \quad |\widehat{F}_z(f) - \widehat{F}_{z'}(f)| \le \delta, \ \forall z, z'. \end{aligned} \tag{2.11}$$

2.3.2 Group Weighted Peer Loss Approach

The developed *peer loss* function partially circumvents the issue of noise rate estimation [LG20]. The peer loss requires less prior knowledge of the noise rates for each class. It is defined as:

$$\ell_{peer}(f(\boldsymbol{x}_i), \tilde{y}_i) \coloneqq \ell\left(f(\boldsymbol{x}_i), \tilde{y}_i\right) - \alpha \cdot \ell\left(f(\boldsymbol{x}_{i_1}), \tilde{y}_{i_2}\right), \tag{2.12}$$

where

$$\alpha = 1 - (1 - \epsilon^- - \epsilon^+) \cdot \frac{\Pr(Y = +1) - \Pr(Y = -1)}{\Pr(\tilde{Y} = +1) - \Pr(\tilde{Y} = -1)}$$

is a parameter to balance the instances for each label, and where i_1 and i_2 are uniformly and randomly selected samples from $I_z/\{i\}$ (i.e., "peer" samples which inspired the name peer loss as noted in [LG20]). Although the noise parameters explicitly appear in the definition of α, only the knowledge of $\Delta \coloneqq 1 - \epsilon^- - \epsilon^+$ is needed. In practice, we could tune α as a hyper-parameter during training. This loss function has the following important property, proven in Lemma 3 of [LG20]:

$$\mathbb{E}_{\tilde{\mathcal{D}}_z}[\ell_{peer}(f(X), \tilde{Y})] = \Delta_z \cdot \mathbb{E}_{\mathcal{D}_z}[\ell_{peer}(f(X), Y)], \tag{2.13}$$

where $\tilde{\mathcal{D}}_z$ denotes the noisy distribution for group z and $\Delta_z = 1 - \epsilon_z^- - \epsilon_z^+$. Adapting the peer loss function to labels with group dependent noise requires accounting for the differing values of Δ_z. We do so by re-weighting Equation (2.12) to obtain our *group-weighted peer loss* ℓ_{gp}:

$$\ell_{gp}(f(\boldsymbol{x}_i), \tilde{y}_i) \coloneqq \frac{1}{\Delta_{z_i}} \left(\ell(f(\boldsymbol{x}_i), \tilde{y}_i) - \alpha \cdot \ell\left(f(\boldsymbol{x}_{i_1}), \tilde{y}_{i_2}\right)\right). \tag{2.14}$$

When class is balanced for every group z, i.e., $\Pr_{Z=z}(Y = +1) = \Pr_{Z=z}(Y = -1) = \frac{1}{2}$, the parameter α is exactly 1. In this case, the expected group-weighted peer loss on the noisy distribution $\tilde{\mathcal{D}}$ is the same as the expected uncorrected loss ℓ on the true distribution $\mathcal{D}$. More precisely:

Theorem 2.5. *For all group dependent noise rates* ϵ_z^- *and* ϵ_z^+ *satisfying* $\epsilon_z^- + \epsilon_z^+ < 1$, *taking* $\ell(\cdot)$ *as the 0-1 loss* $\mathbb{1}(\cdot)$ *and when* $\Pr_{Z=z}(Y=+1) = \Pr_{Z=z}(Y=-1) = \frac{1}{2}$,

$$\mathbb{E}_{\tilde{\mathcal{D}}}[\ell_{gp}(f(X), \tilde{Y})] = \mathbb{E}_{\mathcal{D}}[\ell(f(X), Y] - \frac{1}{2}. \tag{2.15}$$

Peer-based surrogate fairness constraints. We acquire the following result in order to create group-aware surrogate constraints:

Lemma 2.6. *True* TPR *and* FPR *relate to* $\widetilde{\mathrm{TPR}}_z, \widetilde{\mathrm{FPR}}_z$ *defined on the noisy labels as follows:*

$$\mathrm{TPR}_z = \Pr(f(X)=+1|Z=z) + \Delta_z \cdot (\widetilde{\mathrm{TPR}}_z - \widetilde{\mathrm{FPR}}_z) \cdot \Pr(Y=-1|Z=z) \tag{2.16}$$

$$\mathrm{FPR}_z = \Pr(f(X)=+1|Z=z) - \Delta_z \cdot (\widetilde{\mathrm{TPR}}_z - \widetilde{\mathrm{FPR}}_z) \cdot \Pr(Y=+1|Z=z) \tag{2.17}$$

Lemma 2.6 allows us to derive the appropriate surrogate fairness constraints for the peer loss, displayed in Table 2.3. Note that we have assumed that the datasest is balanced for each group; i.e., $\forall z \in Z \quad \Pr(Y=+1 \mid Z=z) = \frac{1}{2}$. If the data is imbalanced, we will require knowing the marginal prior $\Pr(Y=+1 \mid Z=z)$.

We merely require knowledge of Δ_z for each z in order to define ℓ_{gp} and $\widehat{F}_z(f)$. This is a weaker requirement compared to knowing the error rates (which will carry estimation of two parameters for each group). We indeed see our group peer loss approach performs more stably as compared to the surrogate loss approach introduced in last subsection when using noisy estimates of the noise rates. With group-weighted peer loss function and surrogate fairness constraints, we are able to perform a fair ERM

as detailed in Equaltion (2.11) by replacing $\tilde{\ell}$ with ℓ_{gp} and the corresponding $\widehat{F}_z(f)$ term.

2.4 Error Rates Estimation and its Impact

We employ "confident learning" to perform noise rate estimation [NJC21]. The first step is to pre-train a classifier f_{pre} over the noisy labels directly and learn a noisy predicted probability

$$\hat{p}(y; \boldsymbol{x}, z) = \Pr(f_{\text{pre}}(\boldsymbol{x}) = y | Z = z).$$

Then, for each pair of classes $k, l \in \{+1, -1\}$, we define the subset of samples:

$$\widehat{X}_{\tilde{y}=k,z} \coloneqq \{\boldsymbol{x}_i \mid \tilde{y}_i = k, i \in I_z\},$$

$$\widehat{X}_{\tilde{y}=k,y=l,z} \coloneqq \{\boldsymbol{x}_i \mid \tilde{y}_i = k, \hat{p}(y = l; \boldsymbol{x}_i, z) \geq t_{l,z}, i \in I_z\},$$

where $t_{l,z} = \frac{1}{|\widehat{X}_{\hat{y}=l,z}|} \sum_{\boldsymbol{x} \in \widehat{X}_{\hat{y}=l,z}} \hat{p}(\hat{y} = l; \boldsymbol{x}, z)$ is the *expected self-confidence probability* for class l and group z. Using the above quantities, we estimate the group-aware joint probability $\widehat{Q}_{\tilde{y}=k,y=l,z} = \Pr(\widetilde{Y} = k, Y = l, Z = z)$ over the noisy labels $\tilde{y}$ and clean labels y with:

$$\widehat{Q}_{\tilde{y}=k,y=l,z} = \frac{\frac{|\widehat{X}_{\tilde{y}=k,y=l,z}|}{\sum_l |\widehat{X}_{\tilde{y}=k,y=l,z}|} \cdot |\widehat{X}_{\tilde{y}=k,z}|}{\sum_{k,l} \left(\frac{|\widehat{X}_{\tilde{y}=k,y=l,z}|}{\sum_l |\widehat{X}_{\tilde{y}=k,y=l,z}|} \cdot |\widehat{X}_{\tilde{y}=k,z}| \right)} \tag{2.18}$$

We use the marginals of estimated joint to compute the noise parameter estimates for each group z:

$$\begin{aligned}\hat{\epsilon}_z^+ &= \frac{\widehat{Q}_{\tilde{y}=-1,y=+1,z}}{\widehat{Q}_{\tilde{y}=-1,y=+1,z} + \widehat{Q}_{\tilde{y}=+1,y=+1,z}}, \\ \hat{\epsilon}_z^- &= \frac{\widehat{Q}_{\tilde{y}=+1,y=-1,z}}{\widehat{Q}_{\tilde{y}=+1,y=-1,z} + \widehat{Q}_{\tilde{y}=-1,y=-1,z}}\end{aligned} \tag{2.19}$$

To estimate Δ_z, we simply substitute $\hat{\epsilon}_z^-$ and $\hat{\epsilon}_z^+$ for ϵ_z^- and ϵ_z^+ in the equation for Δ_z. As a byproduct, we could estimate the marginal priors $\Pr(Y = +1|Z = z)$ by

$$\frac{\widehat{Q}_{\tilde{y}=+1,y=+1,z} + \widehat{Q}_{\tilde{y}=-1,y=+1,z}}{\widehat{Q}_{\tilde{y}=+1,y=+1,z} + \widehat{Q}_{\tilde{y}=-1,y=+1,z} + \widehat{Q}_{\tilde{y}=-1,y=+1,z} + \widehat{Q}_{\tilde{y}=-1,y=-1,z}} \tag{2.20}$$

Effects of noisy estimates. It is important to quantify the impact of the noise rate estimation error on the accuracy and fairness of the resulting classifier. We first note that, for any $\eta, \tau > 0$, the law of large numbers implies that taking sufficiently many samples from $\mathcal{D}$ will ensure that the following holds for all z with probability at least $1 - \eta$:

$$\begin{aligned}\max\Big\{ &\left|\hat{\epsilon}_z^+ - \epsilon_z^+\right|, \left|\tfrac{\hat{\epsilon}_z^+}{1-\hat{\epsilon}_z^+-\hat{\epsilon}_z^-} - \tfrac{\epsilon_z^+}{1-\epsilon_z^+-\epsilon_z^-}\right|, \\ &\left|\hat{\epsilon}_z^- - \epsilon_z^-\right|, \left|\tfrac{1-\hat{\epsilon}_z^-}{1-\hat{\epsilon}_z^+-\hat{\epsilon}_z^-} - \tfrac{1-\epsilon_z^-}{1-\epsilon_z^+-\epsilon_z^-}\right|\Big\} \leq \tau.\end{aligned} \tag{2.21}$$

Denote by $\hat{\ell}(\cdot)$ the surrogate loss function defined using the estimated noises $\{\hat{\epsilon}_z^+, \hat{\epsilon}_z^-\}$, and let

$$\hat{f}^* = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{N} \hat{\ell}(f(\boldsymbol{x}_i), \tilde{y}_i), \qquad \tilde{f}^* = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{N} \tilde{\ell}(f(\boldsymbol{x}_i), \tilde{y}_i) \tag{2.22}$$

We have the following result:

Theorem 2.7. *For every $\eta, \tau > 0$ there exists N such that*

$$\frac{1}{N} \cdot \sum_{i=1}^{N} \tilde{\ell}(\hat{f}^*(\boldsymbol{x}_i), \tilde{y}_i) - \frac{1}{N} \cdot \sum_{i=1}^{N} \tilde{\ell}(\tilde{f}^*(\boldsymbol{x}_i), \tilde{y}_i) \leq 4\tau \cdot \bar{\ell} \tag{2.23}$$

with probability at least $1 - \eta$, where $\bar{\ell} \coloneqq \max \ell$.

Because the fairness constraints $\widehat{F}_z(f)$ are linear in $\epsilon_z^+, \epsilon_z^-$s, the additional fairness violations incurred due to the noisy estimates of the error rates will also be linear in τ too. Similar observations hold when using the estimated $\tilde{\Delta}_z$ in the peer loss.

2.5 Equalizing Error Rates Improves Fairness Guarantee

Group-dependent label noise rates can exacerbate unfairness when fairness constraints are directly applied to noisy labels. Addressing these fairness issues requires knowledge of the label noise rates. While existing literature offers data cleaning solutions, learners typically need to estimate unknown noise rates through various procedures [ZSL21, NJC21, PRM+17]. It's important to note that misspecifying these noise rates can introduce additional learning errors, particularly when the label noise is asymmetric. Conversely, equalizing error rates by increasing the noise rate for the lower group, despite reducing the overall informativeness of training labels, is generally considered more manageable — one can always do so by randomly flipping a portion of labels. As shown in Theorem 2.3, loss correction procedures become unnecessary when the noise rates are balanced.

We assume that the error rates are balanced across classes:

$$\epsilon_z^+ = \epsilon_z^- = e_z \tag{2.24}$$

Resampling the noisy data examples such that $\Pr(\tilde{Y} = +1 \mid Z = z) = \Pr(\tilde{Y} = -1 \mid Z = z) = 0.5, z \in \{a, b\}$, we derive the following relationship:

Lemma 2.8. TPR_z *and* FPR_z *relate to* $\widetilde{\mathrm{TPR}}_z$, $\widetilde{\mathrm{FPR}}_z$ *as follows:*

$$\mathrm{TPR}_z = \frac{e_z \cdot \widetilde{\mathrm{TPR}}_z - (1 - e_z) \cdot \widetilde{\mathrm{FPR}}_z}{2e_z - 1} \tag{2.25}$$

$$\mathrm{FPR}_z = \frac{e_z \cdot \widetilde{\mathrm{FPR}}_z - (1 - e_z) \cdot \widetilde{\mathrm{TPR}}_z}{2e_z - 1} \tag{2.26}$$

Denote by $\tilde{e}_a, \tilde{e}_b$ (both < 0.5) the estimated noise rates of e_a, e_b that we have access to. Suppose we suffer from the following mis-specifications:

$$\texttt{err}_{\texttt{M}} \coloneqq \min\{\texttt{err}_a \coloneqq |\tilde{e}_a - e_a|, \texttt{err}_b \coloneqq |\tilde{e}_b - e_b|\}. \tag{2.27}$$

Denote the corrected TPR and FPR using $\widetilde{\mathrm{TPR}}$ and $\widetilde{\mathrm{FPR}}$ with estimated $\tilde{e}_a, \tilde{e}_b$ as

$$\mathrm{TPR}_z^c(h) = \frac{\tilde{e}_z \cdot \widetilde{\mathrm{TPR}}_z(h) - (1 - \tilde{e}_z) \cdot \widetilde{\mathrm{FPR}}_z(h)}{2\tilde{e}_z - 1} \tag{2.28}$$

$$\mathrm{FPR}_z^c(h) = \frac{\tilde{e}_z \cdot \widetilde{\mathrm{FPR}}_z(h) - (1 - \tilde{e}_z) \cdot \widetilde{\mathrm{TPR}}_z(h)}{2\tilde{e}_z - 1} \tag{2.29}$$

Theorem 2.9 establishes possible fairness violation due to noise rates mis-specification $\texttt{err}_{\texttt{M}}$:

Theorem 2.9. *Equalizing* $\mathrm{TPR}_z^c(h)$ *&* $\mathrm{FPR}_z^c(h)$ *for group* a, b *leads to following possible*

fairness violation:

$$|\operatorname{TPR}_a(h) - \operatorname{TPR}_b(h)| \geq \mathtt{err}_{\mathtt{M}} \cdot \left| \frac{\widetilde{\operatorname{TPR}}_a(h)}{(2e_a - 1)(2\tilde{e}_a - 1)} - \frac{\mathtt{err}_b}{\mathtt{err}_a} \frac{\widetilde{\operatorname{TPR}}_b(h)}{(2e_b - 1)(2\tilde{e}_b - 1)} \right|,$$

$$|\operatorname{FPR}_a(h) - \operatorname{FPR}_b(h)| \geq \mathtt{err}_{\mathtt{M}} \cdot \left| \frac{\widetilde{\operatorname{FPR}}_a(h)}{(2e_a - 1)(2\tilde{e}_a - 1)} - \frac{\mathtt{err}_b}{\mathtt{err}_a} \frac{\widetilde{\operatorname{FPR}}_b(h)}{(2e_b - 1)(2\tilde{e}_b - 1)} \right|.$$

But as a consequence of Lemma 2.8, we immediately know that

Theorem 2.10. *Whenever $e_a = e_b$, equalizing $\widetilde{\operatorname{TPR}}$ and $\widetilde{\operatorname{FPR}}$ suffices to equalizing the true* TPR *and* FPR*.*

The technical problem is that we would not know which protected group are suffering from higher noise, and again we would not have the ground truth labels to verify any possible hypothesis. We first present the following definition:

Definition 1 (Clusterability)**.** We say the dataset D satisfies 2-NN clusterability if each instance $\boldsymbol{x}$ shares the same true label class with its two nearest neighbors measured by $\|\boldsymbol{x} - \boldsymbol{x}'\|_2$.

For the rest of the section, we will assume that D satisfies 2-NN clusterability. 2-NN was similarly introduced in a recent work [ZSL21] and has been shown to be a requirement that is mild to satisfy. Now we define the following two quantities that are central to the development of our idea. For an arbitrary instance $\boldsymbol{x}_1$ with noisy label $\widetilde{y}_1$, denote the noisy labels for two nearest neighbor instances of $\boldsymbol{x}_1$ as $\widetilde{y}_2, \widetilde{y}_3$. Define the following agreement measures:

Definition 2 (2-NN Agreements)**.** Let $\widetilde{y}_1$ denote the noisy label for a randomly selected instance $\boldsymbol{x}_1$. $\widetilde{y}_2, \widetilde{y}_3$ are the noisy labels of $\boldsymbol{x}_1$'s 2-NN instances (measured by $||x - x'||_2$).

$$\text{Positive Agreements} \qquad \text{PA}_{\mathcal{D}} \coloneqq \Pr(\widetilde{y}_1 = \widetilde{y}_2 = \widetilde{y}_3 = +1 \mid z = a) \tag{2.30}$$

$$\text{Negative Agreements} \qquad \text{NA}_{\mathcal{D}} \coloneqq \Pr(\widetilde{y}_1 = \widetilde{y}_2 = \widetilde{y}_3 = +1 \mid z = b) \tag{2.31}$$

The agreement measures the likelihood of the neighbor data points "agreeing" on the same label. Now we will first sub-sample the noisy distribution and compute $\text{PA}_{\mathcal{D}}, \text{NA}_{\mathcal{D}}$:

- **Step 1**: Sample data examples such that the data examples are balanced across protected groups. Denote this resampled distribution as $\mathcal{D}^{\diamond}$.
- **Step 2**: Compute $\text{NA}_{\mathcal{D}^{\diamond}}$ and $\text{PA}_{\mathcal{D}^{\diamond}}$ by Definition 2.

Theorem 2.11. *When $e_a, e_b < 0.5$ and $\mathcal{D}$ satisfies 2-NN clusterability, $PA_{\mathcal{D}^{\diamond}}, NA_{\mathcal{D}^{\diamond}}$ relate to e_a, e_b as follows:*

$$PA_{\mathcal{D}^{\diamond}} - NA_{\mathcal{D}^{\diamond}} = 2(0.5 - e_a)(0.5 - e_b)(e_a - e_b). \tag{2.32}$$

Then if $PA_{\mathcal{D}^{\diamond}} > NA_{\mathcal{D}^{\diamond}}$, we know that $e_a < e_b$; otherwise $e_a > e_b$. If $PA_{\mathcal{D}^{\diamond}} = NA_{\mathcal{D}^{\diamond}}$, then $e_a = e_b$.

We show that randomly flipping $\widetilde{Y}$ from groups with the smaller noise rates by a small probability ϵ monotonically decreases the gap between noise rates $|e_a - e_b|$. Without the loss of generality suppose $e_a < e_b$, and we will only flip the labels from group a (but not flipping the ones from group b).

Proposition 2.12. *Denote by* $\hat{Y}$ *as the flipped version of* $\widetilde{Y}$*:* $\Pr(\hat{Y} \neq \widetilde{Y}) = \epsilon$*, and* $\hat{e}_a \coloneqq \Pr(\hat{Y} \neq | \ Z = a), \hat{e}_b \coloneqq \Pr(\hat{Y} \neq Y \mid Z = b)$*. We have:*

$$\hat{e}_a = (1 - e_a) \cdot \epsilon + e_a, \qquad \hat{e}_b = (1 - \epsilon) \cdot e_b \tag{2.33}$$

Further, the new gap between the noise rates of the flipped label $\hat{Y}$ *is a monotonic function of* ϵ*:*

$$\hat{e}_b - \hat{e}_a = e_b - e_a - (1 - e_a + e_b) \cdot \epsilon. \tag{2.34}$$

Since $1 - e_a + e_b > 0$, when ϵ is small, the above derivation shows the effectiveness in reducing the noise rate gap $e_b - e_a$ by randomly flipping the noisy labels that correspond to the class with lower noise rate. The only remaining question is how to find the optimal ϵ such that $\hat{e}_b - \hat{e}_a = 0$. Calling Theorem 9, we know

$$\mathrm{PA}_{\mathcal{D}^\diamond} - \mathrm{NA}_{\mathcal{D}^\diamond} = 2(0.5 - \hat{e}_+)(0.5 - \hat{e}_-)(\hat{e}_- - \hat{e}_+). \tag{2.35}$$

Denote by

$$f(\epsilon) \coloneqq 0.5 \cdot (\mathrm{PA}_{\mathcal{D}^\diamond} - \mathrm{NA}_{\mathcal{D}^\diamond}). \tag{2.36}$$

As shown in Figure 2.1, it is easy to derive the three solutions for $f(\epsilon) = 0$ (setting each of the terms to 0)

$$\epsilon_1 = 1 - \frac{0.5}{e_b} < 0 \tag{2.37}$$

$$\epsilon_2 = \frac{e_b - e_a}{1 - e_a + e_b} \tag{2.38}$$

$$\epsilon_3 = \frac{0.5 - e_a}{1 - e_a} \tag{2.39}$$

Note that $\epsilon_2 = \frac{e_b - e_a}{1 - e_a + e_b} < \frac{(e_b - e_a) + (1 - e_a - e_b)}{(1 - e_a + e_b) + (1 - e_a - e_b)} = \frac{1 - 2e_a}{2(1 - e_a)} = \epsilon_3$, and ϵ_3 will lead to an uninformative state where $\hat{e}_a = 0.5$. Therefore ϵ_2 is our target root.

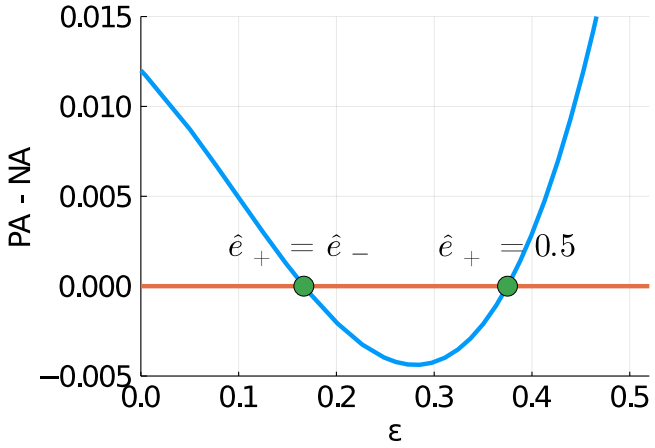


Figure 2.1: Agreement gap PA − NA varies for different ϵ. There are only two positive roots for PA − NA = 0. The less one results in $\hat{e}_a = \hat{e}_b$.

The monotonicity of $f(\epsilon)$ from 0 to ϵ_2 suggests using a binary search to find an appropriate ϵ. We start with two flipping parameters $\epsilon_l < \epsilon_r$, which induce synthetic datasets $\mathcal{D}_l$ and $\mathcal{D}_r$. The gaps of the counted agreements are $C_l = \text{PA}_{\mathcal{D}_l} - \text{NA}_{\mathcal{D}_l}$ and $C_r = \text{PA}_{\mathcal{D}_r} - \text{NA}_{\mathcal{D}_r}$, where $C_l > 0 > C_r$. In each iteration, we test a new flip parameter $\epsilon_{\text{mid}} = (\epsilon_l + \epsilon_r)/2$. If the resulting gap C_{mid} falls within a threshold γ, *i.e.*, $-\gamma \leq C_{\text{mid}} \leq \gamma$, we return the labels flipped by ϵ_{mid}. Otherwise, we update the values of ϵ_l and ϵ_r according to the sign of C_{mid}: if $C_{\text{mid}} < 0$, we set $\epsilon_r \leftarrow \epsilon_{\text{mid}}, \mathcal{D}_r \leftarrow \mathcal{D}_{\text{mid}}$ (reducing ϵ_r); otherwise $\epsilon_l \leftarrow \epsilon_{\text{mid}}, \mathcal{D}_l \leftarrow \mathcal{D}_{\text{mid}}$ (increasing ϵ_l). We propose NOISE+ in Algorithm 1. We initialize $\epsilon_r = 0.3$, which empirically works in various noise settings. If this fails, we can grid search different ϵ values (e.g., 0.1, 0.2) that satisfy $C_r < 0$ as the

initial ϵ_r. Note that Algorithm 1 assumes $e_+ < e_-$; the implementation is symmetric for $e_+ > e_-$.

Our algorithm is well-suited for loss functions that don't require prior knowledge of noise rates. While standard cross-entropy (CE) is certainly applicable, various robust loss functions are also excellent candidates. A particularly promising option is the recently introduced peer loss function [liu2019peer]. This function doesn't necessitate specifying noise rates and has been further adapted into a group-weighted peer loss for fairness constraints in Section 2.3.2. We believe this loss function aligns especially well with our algorithm's objectives.

2.6 Experiments

Due to the difficulty of acquiring real world datasets with known label corruption characteristics, we artificially synthesize the datasets with a noise generation step. These controlled experiments help us understand the robustness of our approaches under different noise scenarios.

2.6.1 Experimental Setup

Dataset We evaluate our methods as well as other baseline methods on five datasets:

- `Adult`, the Adult dataset from the UCI ML Repository with males and females as the protected groups [DG17a].

Table 2.4: Dataset statistic and parameters.

Dataset	Source	Number of data examples n	Fairness Tolerance δ	Sensitive Groups	Noise Rates ϵ^-	Noise Rates ϵ^+
`adult`	UCI [DG17a]	32561	2%	*female*	0.45	0.15
				male	0.35	0.55
`arrest`	COMPAS [ALMK16a]	6644	5%	*white*	0.40	0.30
				black	0.15	0.25
`arrest`	COMPAS [ALMK16a]	6644	5%	*white* male	0.45	0.10
				black male	0.10	0.35
				white female	0.35	0.45
				black female	0.55	0.25
`violent`	COMPAS [ALMK16a]	5278	5%	*white male*	0.45	0.10
				black male	0.10	0.35
				white female	0.35	0.45
				black female	0.55	0.25
`German`	UCI [DG17a]	1000	2%	*female*	0.45	0.15
				male	0.35	0.55
`law`	LSAC [Wig98]	18692	2%	*white*	0.45	0.15
				black	0.35	0.55

- `Arrest` and `Violent`, the COMPAS recidivism dataset for arrest and violent crime statistics, with race (restricted to white and black) and gender as the sensitive attributes [ALMK16a].

- `German`, the German credit dataset from UCI ML Repository with gender as the sensitive attribute [DG17a].

- `Law`, a subset of the original data set from LSAC with race (restricted to black and white) as the sensitive attribute [Wig98].

Table 2.4 describes the dataset statistics and parameters used in the experi-

ments. We chose to apply a diverse set of noise parameters to the different subgroups. The fairness tolerance δ and noise parameters ϵ for `Adult`, `German` and `Law` data sets are identical, but they are different from `Arrest` and `Violent` data sets because `Arrest` and `Violent` data sets contain more protected groups. We make this choice mainly for the baseline models to obtain meaningful results to compare with.

Noise generation We randomly split the clean dataset $\mathcal{D} = \{(x_i, y_i, z_i)\}_{i=1}^n$ into a training set and a test set in a ratio of 80 to 20. We add asymmetric label noises to the training dataset, and leave the test data untouched for verification purposes. For each sensitive group $z \in Z$, we randomly flip the clean label y with probability ϵ_z^- if its value is -1, and we flip the clean label with probability ϵ_z^+ if it's $+1$. After injecting this noise, we use the same training set and test set to benchmark all the methods.

Methods. For all of the methods above, we use logistic regression to perform classification and leverage the reduction approach as proposed in [ABD$^+$18a] for solving our constrained optimization problem. We evaluate the performance of several methods:

- `Clean`, in which the classifier is trained on the clean data subject to the equal odds constraint
- `Corrupt`, which directly trains the classifier on the corrupted data subject to the equal odds fairness constraint
- `Surrogate Loss`, which uses the surrogate loss approach described in Section 2.3.1

- `Group Peer Loss`, which uses the group weighted peer loss approach to train a fair classifier on the corrupted training set.

The `Corrupt` baseline gives us a sense about the harm caused by the unawareness of the labels' noise, and the `clean` baseline shows the biases contained in the datasets.

We set the same maximum fairness violation δ for all the methods on the same dataset during training. As there are more sensitive groups on `arrest` and `violent` datasets, we set $\delta = 5\%$ on these datasets and $\delta = 2\%$ on the other datasets. We report metrics for each of the above methods averaged over five runs.

Computing Infrastructure We conducted all the experiments on a 3 GHz 6-Core Intel Core i5 CPU. The running time for `Surrogate Loss` is about 10 minutes, while the running time for `Group Peer Loss` could be over 30 minutes.

Tuning α in Peer Loss The performance of our group weighted peer loss is highly influenced by the hyperparameter α. Recall that

$$\mathbb{E}_{\tilde{\mathcal{D}}_z}[\ell_{gp}(f(X), \tilde{Y})] = \mathbb{E}_{\mathcal{D}_z}[\ell_{gp}(f(X), Y)]$$

We split 10% of data examples in the train set for validation and found the optimal α using grid search. The range of α we searched varied between 0.0 to 2.0. We observed that both the accuracy and fairness violation on the validation set exhibit the same trends on the test set. In practice, the group weighted peer loss with $\alpha = 0.3$ achieves the best performance on the `Adult` dataset.

Table 2.5: Overview of group-based performance metrics for all methods on 5 data sets. We highlight the best values achieved for fairness violation and accuracy in green and the worst in red. m is the number of sensitive groups, $\bar{\epsilon}$ is the average of error rates over all the groups and all label classes $\epsilon_z^+, \epsilon_z^-$s. *true* indicates training with true noise parameters and *estimated* indicates training with estimated noise parameters. The values after $\pm$ are the standard deviation.

					Surrogate Loss		Group Peer Loss	
Dataset	**Metrics**	**Avg.** $\bar{\epsilon}$	Clean	Corrupt	*true*	*estimated*	*true*	*estimated*
Adult	*violation*	0.38	0.47%	$8.36 \pm 1.36\%$	$1.46 \pm 0.50\%$	$1.39 \pm 0.80\%$	$1.18 \pm 0.63\%$	$1.69 \pm 0.86\%$
$m = 2$	*accuracy*		83.76%	$76.08 \pm 2.49\%$	$81.16 \pm 3.41\%$	$75.99 \pm 7.45\%$	$77.00 \pm 2.52\%$	$75.13 \pm 5.15\%$
Arrest	*violation*	0.28	2.27%	$2.98 \pm 0.74\%$	$0.54 \pm 0.27\%$	$0.36 \pm 0.24\%$	$1.78 \pm 0.89\%$	$1.05 \pm 0.55\%$
$m = 2$	*accuracy*		65.16%	$60.72 \pm 0.66\%$	$61.7 \pm 3.23\%$	$62.3 \pm 5.30\%$	$63.81 \pm 3.35\%$	$65.31 \pm 3.41\%$
Arrest	*violation*	0.34	5.89%	$12.93 \pm 0.95\%$	$0.88 \pm 0.27\%$	$2.48 \pm 1.42\%$	$1.36 \pm 0.69\%$	$1.40 \pm 0.36\%$
$m = 4$	*accuracy*		66.0%	$53.7 \pm 1.82\%$	$65.7 \pm 2.92\%$	$58.8 \pm 4.96\%$	$60.27 \pm 2.90\%$	$57.56 \pm 2.96\%$
Violent	*violation*	0.34	0.37%	$7.16 \pm 0.80\%$	$4.81 \pm 0.70\%$	$7.76 \pm 1.02\%$	$2.06 \pm 0.81\%$	$0.68 \pm 0.28\%$
$m = 4$	*accuracy*		60.18%	$52.2 \pm 0.23\%$	$53.14 \pm 4.91\%$	$55.4 \pm 0.71\%$	$55.64 \pm 4.88\%$	$52.7 \pm 0.57\%$
German	*violation*	0.38	0.68%	$2.68 \pm 0.32\%$	$11.79 \pm 3.87\%$	$11.08 \pm 2.16\%$	$0.00 \pm 0.00\%$	$1.64 \pm 0.32\%$
$m = 2$	*accuracy*		74.5%	$70.5 \pm 0.00\%$	$68.5 \pm 4.27\%$	$71.5 \pm 2.53\%$	$70.0 \pm 0.71\%$	$70.5 \pm 2.53\%$
Law	*violation*	0.38	0.6%	$2.74 \pm 0.12\%$	$0.36 \pm 0.08\%$	$1.98 \pm 1.16\%$	$0.03 \pm 0.02\%$	$0.57 \pm 0.12\%$
$m = 2$	*accuracy*		90.67%	$90.16 \pm 0.79\%$	$90.26 \pm 0.48\%$	$89.92 \pm 2.86\%$	$90.32 \pm 0.10\%$	$90.29 \pm 0.20\%$

2.6.2 Results

We present an overview of the performance for each method on the test set in Table 2.5. We compare the two fair ERM approaches using both the true and estimated noise rates. The metrics we report include *violation*, the maximum difference in TPR and FPR between groups $z, z' \in Z$, and *accuracy*, the accuracy achieved on test set.

We make the following observations about our results. First, both of the two fair ERM approaches in Section 2.3 produce classifiers that are more effective at mitigating unfairness than a classifier that is naively trained on the corrupted data.

In particular, the group weighted peer loss approach achieves almost 0% violation on the `German` and `law` data sets, when given the true noise parameters. The only noticeable worse case arises when applying the surrogate loss approach to the `German` dataset. This may be due to the high variance of the `German` dataset, which has fewer than 1000 samples.

Second, as expected, models trained using our proposed fair ERM methods do not achieve the same level of accuracy as a model that is fit using clean labels. However, our models are typically more accurate than the model fit directly to the corrupted data. For example, on the `arrest` data set with four protected groups, the surrogate loss approach achieves a similar accuracy to the classifier trained on clean data while incurring an even smaller fairness violation. Third, Our methods perform similarly well when trained using both the true and with the estimated noise parameters, indicating that the noise estimation procedures are effective. On `arrest` and `violent` datasets, our methods with estimated noise parameters even perform better than those with true parameters. This is probably due to the biases and noise in these datasets. Finally, our fair ERM frameworks adapt well to multiple sensitive groups, as demonstrated by the good performance on the `Arrest` and `Violent` data sets.

Table 2.6: We show how different levels of symmetric noise $\epsilon^- = \epsilon^+ = \epsilon$ affect the classifiers' performance on `adult` dataset. SL: Surrogate Loss. GPL: Group Peer Loss. We highlight substantial improvement of fairness in green and sever violation in red.

Noise ϵ	Metric	Clean	Corrupt	SL	GPL
0.1	*violation*	0.47%	3.91%	5.15%	1.41%
	accuracy	83.76%	83.22%	82.73%	82.71%
0.2	*violation*	0.47%	3.83%	3.98%	1.49%
	accuracy	83.75%	82.08%	82.54%	82.16%
0.3	*violation*	0.47%	7.23%	3.63%	1.22%
	accuracy	83.76%	81.36%	82.01%	81.24%
0.4	*violation*	0.47%	5.14%	1.13%	3.1%
	accuracy	83.76%	79.58%	80.62%	80.21%

2.6.3 Impact of noise levels on classifier performance.

We present the results of varying noise rate on the `adult` data set (with two groups) in Table 2.6. We only add symmetric noise to $female$ group and keep the *male* group clean. ERM is generally robust to symmetric noises when a significant subset of the data is clean (one group in our example), so we do not expect significant accuracy improvement from our methods. We focus on how fairness violation reduces. Observe that, comparing to training with clean data, training on corrupted data substantially increases fairness violations, even for relatively low noise rates. The SL and GPL

columns show that our fair ERM approaches can effectively mitigate the biases. This holds true even when increasing the noise rate.

2.6.4 Insights on running on data directly, without adding additional noise

We evaluate our algorithm on the clean `adult` and `arrest` datasets as shown in Table 2.7. On the `arrest` dataset, our methods achieve a similar performance of accuracy compared with the Clean baseline, but we do observe a consistent drop of fairness violations on the `arrest` dataset. The fairness violation of our methods on `adult` dataset is not as good as that of Clean baseline. This fact may imply the possibility that the `arrest` dataset contains more human biases in labels than the `adult` dataset. The small drop in accuracy and (sometimes) in fairness is due to the additional noise estimation step, which introduces another layer of complication - this is the price we pay for dealing with potentially highly noisy labels.

2.7 Comparison to Related Works

A great deal of research has been devoted to fair classification in general, including fair classification under statistical constraints [ZVRG17, FFM$^+$15a, HPS16a, ABD$^+$18a], decoupled training with preference guarantees [ZVGRG17b, DIKL18, LMC18, ULP19, CHKV19], and preventing gerrymandering [KNRW17], among many others

Table 2.7: We examine the performance of our methods on the clean `adult` and `arrest` datasets. Clean: train a fair classifier directly with equal odds constraint. SL: Surrogate Loss with estimated noise parameters. GPL: Group Peer Loss with estimated noise parameters. The values after ± are the standard deviation.

	`adult`		`arrest`	
Method	*accuracy*	*violation*	*accuracy*	*violation*
Clean	83.76 ± 0.0	0.47 ± 0.0	65.46 ± 0.0	4.46 ± 0.0
SL	76.97 ± 0.24	3.51 ± 0.24	63.07 ± 0.44	2.90 ± 0.72
GPL	81.20 ± 0.19	3.76 ± 0.19	64.98 ± 0.40	1.85 ± 0.36

[MW18, CJS18].

In this work, we specifically focus on fairness in the presence of biased and group-dependent noisy training labels. Our work contributes to the fair classification literature by introducing robust methods for dealing with heterogeneous label noise. We also provide insight into the effects of noise being present in the labels. Our work parallels others' on fair classification with noisy labels [JN19, BS19]. Ours differs primarily in two main respects. First, existing works often assume knowledge of the noise generation process. Second, previous works have only considered noise rates that are homogeneous across different groups. We consider a more realistic setting, where different groups might suffer different levels of bias, and therefore reach very different conclusions. Mitigating bias is substantially more challenging in our setting. Nevertheless, our results could generalized prior work when the noise is assumed constant across groups, or only

one group is assumed to have noise.

Both of our fair ERM approaches extend the literature on learning with noisy data [AL88a, MS13, NDRT13, FV14, Sco15, MVROW15, LT16, PRM$^+$17, CLS19]. Our first uses surrogate loss functions based on [NDRT13] to create unbiased estimators of the fairness constraints. This first approach requires knowledge of the noise parameters. Our second approach relaxes this assumption by extending the work of [LG20] to account for both biases in the fairness constraints and for group specific label noise.

Recent work on fair classification with imperfect data shows how to emulate noiseless fair classification by appropriately re-scaling the fairness tolerance with the noise but is only restricted to class-conditional random noise without considering group difference [LZMV19]. Most of the reported results are for the cases with noisy sensitive attributes but not the labels (despite that the authors provided discussions to how the two problems are related). The surrogate fairness constraints in our paper could be viewed as an extension of their method. Nonetheless, our work is more general, as we consider the more sophisticated settings with group-dependent label noise. [GCFW18b] explores the use of proxy variables when the sensitive attributes are missing. Lastly, [FCG20] also provides some insights on correcting for observed predictive bias might further increase outcome disparities but is concerned with fairness evaluation rather than learning. In contrast with their work, we simplify the assumption on instance-dependent noise into group-dependent, and further develop two fair ERM approaches in terms of the unbiased estimators.

Algorithm 1 NOISE+: A binary search algorithm for balancing noise rates.

Require: $\gamma > 0$, $\epsilon_l = 0$, $\epsilon_r = 0.3$

Resample a balanced set $\mathcal{D}^\diamond$ from $\tilde{D}$;

Initialize $\mathcal{D}_l = \mathcal{D}^\diamond$, $\mathcal{D}_r = \mathsf{Flip}(\mathcal{D}^\diamond, \epsilon_r)$;

Estimate $\mathrm{PA}_{\mathcal{D}_l}$, $\mathrm{PA}_{\mathcal{D}_r}$, $\mathrm{NA}_{\mathcal{D}_l}$, $\mathrm{NA}_{\mathcal{D}_r}$.

while $\mathrm{PA}_{\mathcal{D}_l} - \mathrm{NA}_{\mathcal{D}_l} > \gamma$ and $\mathrm{PA}_{\mathcal{D}_r} - \mathrm{NA}_{\mathcal{D}_r} < -\gamma$ **do**

$\epsilon_{\mathrm{mid}} \leftarrow (\epsilon_l + \epsilon_r)/2, \mathcal{D}_{\mathrm{mid}} \leftarrow \mathsf{Flip}(\mathcal{D}^\diamond, \epsilon_{\mathrm{mid}})$;

Estimate $\mathrm{PA}_{\mathcal{D}_{\mathrm{mid}}}$ and $\mathrm{NA}_{\mathcal{D}_{\mathrm{mid}}}$;

if $\mathrm{PA}_{\mathcal{D}_{\mathrm{mid}}} - \mathrm{NA}_{\mathcal{D}_{\mathrm{mid}}} < -\gamma$ **then**

/* ϵ_{mid} is at the right of the root */

$\epsilon_r \leftarrow \epsilon_{\mathrm{mid}}, \mathcal{D}_r \leftarrow \mathcal{D}_{\mathrm{mid}}$;

$\mathrm{PA}_{\mathcal{D}_r} \leftarrow \mathrm{PA}_{\mathcal{D}_{\mathrm{mid}}}$, $\mathrm{NA}_{\mathcal{D}_r} \leftarrow \mathrm{NA}_{\mathcal{D}_{\mathrm{mid}}}$;

else if $\mathrm{PA}_{\mathcal{D}_{\mathrm{mid}}} - \mathrm{NA}_{\mathcal{D}_{\mathrm{mid}}} > \gamma$ **then**

/* ϵ_{mid} is at the left of the root */

$\epsilon_l \leftarrow \epsilon_{\mathrm{mid}}, \mathcal{D}_l \leftarrow \mathcal{D}_{\mathrm{mid}}$;

$\mathrm{PA}_{\mathcal{D}_l} \leftarrow \mathrm{PA}_{\mathcal{D}_{\mathrm{mid}}}$, $\mathrm{NA}_{\mathcal{D}_l} \leftarrow \mathrm{NA}_{\mathcal{D}_{\mathrm{mid}}}$;

else

return $\hat{D} = \mathsf{Flip}(\tilde{D}, (\epsilon_l + \epsilon_r)/2)$;

end if

end while

return unsuccessful;

Chapter 3

Fairness in Multi-modality

3.1 Mitigating Gender Bias in Image Search

3.1.1 Motivating Example

Internet information is shaping people's minds. The algorithmic processes behind modern search engines, with extensive use of machine learning, have great power to determine users' access to information [ERV+15]. These information systems are biased when results are systematically slanted in unfair discrimination against protected groups [FN96].

Gender bias is a severe fairness issue in image search. Figure 3.1 shows an example: given a gender-neutral natural language query "a person is cooking", only 2 out of 10 images retrieved by an image search model [RKH+21] depict females, while equalized exposure for male and female is expected. Such gender-biased search results

Figure 3.1: Gender bias in image search. We show the top-10 retrieved images for searching "a person is cooking" on the Flickr30K [YLHH14a] test set using a state-of-the-art model [RKH+21]. Despite the gender-neutral query, only 2 out of 10 images are depicting female cooking.

are harmful to society as they change people's cognition and worsen gender stereotypes [KMM15]. Mitigating gender bias in image search is imperative for social good.

In this section, we formally develop a framework for quantifying gender bias in image search results, where text queries in English are made gender-neutral, and gender-balanced search images are expected for models to retrieve. To evaluate model fairness, we use the normalized difference between masculine and feminine images in the retrieved results to represent gender bias. We diagnose the gender bias of two primary families of multimodal models for image search: (1) the specialized models that are often trained on in-domain datasets to perform text-image retrieval, and (2) the general-purpose representation models that are pre-trained on massive image and text data available online and can be applied to image search. Our analysis on MS-

COCO [LMB+14] and Flickr30K [YLHH14a] datasets reveals that both types of models lead to serious gender bias issues (e.g., nearly 70% of the retrieved images are masculine images).

To mitigate gender bias in image search, we propose two novel debiasing solutions for both model families. The specialized in-domain training methods such as SCAN [LCH+18] often adopt contrastive learning to enforce image-text matching by maximizing the margin between positive and negative image-text pairs. However, the gender distribution in the training data is typically imbalanced, which results in unfair model training. Thus we introduce a fair sampling (*FairSample*) method to alleviate the gender imbalance during training without modifying the training data.

Our second solution aims at debiasing the large, pre-trained multimodal representation models, which effectively learn pre-trained image and text representations to accomplish down-stream applications [BHB19, CRC+20a, CKNH20a, GCL+20, CLY+20, RKH+21]. We examine whether the representative CLIP model [RKH+21] embeds human biases into multimodal representations when they are applied to the task of image search. Furthermore, we propose a novel post-processing feature clipping approach, *clip*, that effectively prunes out features highly correlated with gender based on their mutual information to reduce the gender bias induced by multimodal representations. The *clip* method does not require any training and is compatible with various pre-trained models.

We evaluate both debiasing approaches on MS-COCO and Flickr30K and find that, on both benchmarks, the proposed approaches significantly reduce the gender bias

exhibited by SCAN and CLIP models when evaluated on the gender-neutral corpora, yielding fairer and more gender-balanced search results. In addition, we evaluate the similarity bias of the CLIP model in realistic image search results for occupations on the internet, and observe that the post-processing methods mitigate the discrepancy between gender groups by a large margin.

Our contributions are four-fold: (1) we diagnose a unique gender bias in image search, especially for gender-neutral text queries; (2) we introduce a fair sampling method to mitigate gender bias during model training; (3) we also propose a novel post-processing clip method to debias pre-trained multimodal representation models; (4) we conduct extensive experiments to analyze the prevalent bias in existing models and demonstrate the effectiveness of our debiasing methods.

3.1.2 Formulation

In an image search system, text queries may be either gender-neutral or gender-specific. Intuitively, when we search for a gender-neutral query like "a person is cooking", we expect a fair model returning approximately equal proportions of images depicting men and women. For gender-specific queries, an unbiased image search system is supposed to exclude images with mis-specified gender information. This intention aligns with seeking more accurate search results and would be much different from the scope of measuring gender bias in gender-neutral cases. Therefore, we focus on identifying and quantifying gender bias when only searching for gender-neutral text queries.

Given a text query provided by the users, the goal of an image search system is to retrieve the matching images from the curated images. In the domain of multi-modality, given the dataset $\{(v_n, c_n)\}_{n=1}^{N}$ with N image-text pairs, the task of image search aims at matching every image v based on the providing text c. We use $\mathcal{V} = \{v_n\}_{n=1}^{N}$ to denote the image set and $\mathcal{C} = \{c_n\}_{n=1}^{N}$ to denote the text set. Given a text query $c \in \mathcal{C}$ and an image $v \in \mathcal{V}$, image retrieval models often predict the similarity score $S(v, c)$ between the image and text. One general solution is to embed the image and text into a high-dimensional representation space and compute a proper distance metric, such as Euclidean distance or cosine similarity, between vectors [WSL+14]. We take cosine similarity for an example:

$$\begin{aligned} S(v, c) &= \frac{\vec{v} \cdot \vec{c}}{\|\vec{v}\| \|\vec{c}\|} \\ \text{s.t.} \quad \vec{v} &= \mathsf{image\_encoder}(v) \\ \vec{c} &= \mathsf{text\_encoder}(c) \end{aligned} \tag{3.1}$$

The image search system outputs a set of top-K retrieved images $\mathcal{R}_K(c)$ with the highest similarity scores. In this work, we assume that when evaluating on test data, $\forall c \in \mathcal{C}$, the text query c is written in gender-neutral language.

The situations of image search results are complex: there might be no people, one person, or more than one person in the images. Let $g(v) \in \{\text{male}, \text{female}, \text{neutral}\}$ represent the gender attribute of an image v. Note that in this study gender refers to biological sex [Lar17]. We use the following rules to determine $g(v)$: $g(v) = \text{male}$ when

there are only men in the image, $g(v) =$ female when there are only women in the image, otherwise $g(v) =$ neutral.

Portraits in image search results with different gender attributes often receive unequal exposure. Inspired by [KMM15] and [ZWY+17], we measure gender bias in image search by comparing the proportions of masculine and feminine images in search results. Given the set of retrieved images $\mathcal{R}_K(c)$, we count the images depicting males and females

$$N_{\text{male}} = \sum_{v \in \mathcal{R}_K(c)} \mathbb{1}[g(v) = \text{male}], \tag{3.2}$$

$$N_{\text{female}} = \sum_{v \in \mathcal{R}_K(c)} \mathbb{1}[g(v) = \text{female}], \tag{3.3}$$

and define the gender bias metric as:

$$\Delta_K(c) = \begin{cases} 0, & \text{if } N_{\text{male}} + N_{\text{female}} = 0 \\ \frac{N_{\text{male}} - N_{\text{female}}}{N_{\text{male}} + N_{\text{female}}}, & \text{otherwise} \end{cases} \tag{3.4}$$

We don't take absolute values for measuring the direction of skewness, i.e., if $\Delta_K(c) > 0$ it skews towards males. Note that a similar definition of gender bias $\frac{N_{\text{male}}}{N_{\text{male}} + N_{\text{female}}}$ in [ZWY+17] is equivalent to $(1 + \Delta(c))/2$. But our definition of gender bias considers the special case when none of the retrieved images are gender-specific, i.e., $N_{\text{male}} + N_{\text{female}} = 0$. For the whole test set, we measure the mean difference over all the text queries:

$$\text{Bias@}K = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \Delta_K(c) \tag{3.5}$$

3.1.3 Methodology

There are two fashions of multimodal models for the image search task. One is to build a specialized model that could embed image and text into representation vectors with measurable similarity scores. The other is to use general-purpose image-text representations pre-trained on sufficiently big data and compute a particular distance metric. We focus on two representative models, SCAN [LCH+18] and CLIP [RKH+21], for both fashions. For the first fashion, we propose an in-processing learning approach to ameliorate the unfairness caused by imbalanced gender distribution in training examples. This approach builds on contrastive learning but extends with a *fair sampling* step. The in-processing solution requires full training on in-domain data examples. For the second fashion, we propose a post-processing feature *clipping* technique to mitigate bias from an information-theoretical perspective. This approach is compatible with pre-trained models and is light to implement without repeating training steps.

3.1.3.1 In-processing Debiasing: Fair Sampling

Image search models in the first fashion are often trained under the contrastive learning framework [LHS20]. For our in-processing debiasing approach, we now explain the two primary components, contrastive learning and fair sampling, within our context.

Contrastive Learning We start by formally introducing the standard contrastive learning framework commonly used in previous works [LCH+18, CDL20] for image-

text retrieval. Given a batch of N image-text pairs $\mathcal{B} = \{(v_n, c_n)\}_{n=1}^{N}$, the model aims to maximize the similarity scores of matched image-text pairs (positive pairs) while minimizing that of mismatched pairs (negative pairs). The representative SCAN model [LCH+18], denoted as $S(v, c)$ outputting a similarity score between image and text, is optimized with a standard hinge-based triplet loss:

$$\mathcal{L}_{i-t} = \sum_{(v,c)\in\mathcal{B}} [\gamma - S(v, c) + S(v, \tilde{c})]_+ \tag{3.6}$$

$$\mathcal{L}_{t-i} = \sum_{(v,c)\in\mathcal{B}} [\gamma - S(v, c) + S(\tilde{v}, c)]_+ \tag{3.7}$$

where γ is the margin, $\tilde{v}$ and $\tilde{c}$ are negative examples, and $[\cdot]_+$ denotes the ramp function. $\mathcal{L}_{i-t}$ corresponds to image-to-text retrieval, while $\mathcal{L}_{t-i}$ corresponds to text-to-image retrieval (or image search). Common negative sampling strategy includes selecting all the negatives [HWW17], selecting hard negatives of highest similarity scores in the mini-batch [FFKF18], and selecting hard negatives from the whole training data [CDL20]. Minimizing the margin-based triplet loss will make positive image-text pairs closer to each other than other negative samples in the joint embedding space.

Fair Sampling One major issue in the contrastive learning framework is that the gender distribution in a batch of image-text pairs is typically imbalanced. Hence, the negative samples will slant towards the majority group, leading to systematic discrimination. To address this problem, we propose a fair sampling strategy. We split the batch of image-text pairs into masculine and feminine pairs based on the image's gen-

der attribute:

$$\mathcal{V}_{\text{male}} = \{v \mid g(v) = \text{male}, (v, c) \in \mathcal{B}\} \tag{3.8}$$

$$\mathcal{V}_{\text{female}} = \{v \mid g(v) = \text{female}, (v, c) \in \mathcal{B}\} \tag{3.9}$$

$$\mathcal{V}_{\text{neutral}} = \{v \mid g(v) = \text{neutral}, (v, c) \in \mathcal{B}\} \tag{3.10}$$

For every positive image and text pair $(v, c) \in \mathcal{B}$, we identify the gender information contained in the query c. If the natural language query is gender-neutral, we sample a negative image from the set of male and female images with probability $\frac{1}{2}$, respectively. Otherwise, we keep the primitive negative sampling selection strategy for keeping the model's generalization on gender-specific queries. Let $\mathcal{B}^*$ be the batch of gender-neutral image-text pairs, the image search loss with fair sampling is:

$$\mathcal{L}_{t-i}^{fair} = \sum_{(v,c)\in\mathcal{B}^*} (\frac{1}{2}\mathbb{E}_{\bar{v}\in\mathcal{V}_{\text{male}}}[\gamma - S(v, c) + S(\bar{v}, c)]_+ + \frac{1}{2}\mathbb{E}_{\bar{v}\in\mathcal{V}_{\text{female}}}[\gamma - S(v, c) + S(\bar{v}, c)]_+) + \sum_{(v,c)\in\mathcal{B}/\mathcal{B}^*} [\gamma - S(v, c) + S(\tilde{v}, c)]_+ \tag{3.11}$$

Empirically, we find that if we thoroughly apply the Fair Sampling strategy, the recall performance drops too much. To obtain a better tradeoff, we use a weight α to combine the objectives

$$\alpha\mathcal{L}_{t-i}^{fair} + (1 - \alpha)\mathcal{L}_{t-i} \tag{3.12}$$

as the final text-to-image loss function. We do not alter the sentence retrieval loss $\mathcal{L}_{i-t}$ during training for preserving generalization.

3.1.3.2 Post-processing Debiasing: Feature Clipping based on Mutual Information

Pre-training methods have shown promising zero-shot performance on extensive NLP and computer vision benchmarks. The recently introduced CLIP model [RKH+21] was pre-trained on an enormous amount of image-text pairs found across the internet to connect text and images. CLIP can encode image and text into d-dimensional embedding vectors, based on which we can use cosine similarity to quantify the similarity of image and text pairs. In this work, we find that the pre-trained CLIP model reaches the state-of-the-art performance but exhibits large gender bias due to training on uncurated image-text pairs collected from the internet. Although [RKH+21] released the pre-trained CLIP model, the training process is almost unreproducible due to limitations on computational costs and massive training data.

In order to avoid re-training of the CLIP model, we introduce a novel post-processing mechanism to mitigate the representation bias in the CLIP model. We propose to "clip" the dimensions of feature embeddings that are highly correlated with gender information. This idea is motivated by the fact that an unbiased retrieve implies the independence between the covariates (active features) and sensitive attributes (gender) [BHN19]. Clipping the highly correlating covariates will return us a relatively independent and neutral set of training data that does not encode hidden gender bias.

The proposed *clip* algorithm is demonstrated in Algorithm 2, and we explain

Algorithm 2 *clip* algorithm

Require: Index set $\Omega = \{1, ..., d\}$, number of clipped features $0 \leq m < d$

$\mathcal{Z} \leftarrow \emptyset$;

for $i = 1$ to d **do**

Estimate mutual information $I(V_i; g(V))$;

end for

for $j = 1$ to m **do**

$z \leftarrow \arg\max\{I(V_i; g(V)) : i \in \Omega/\mathcal{Z}\}$;

$\mathcal{Z} \leftarrow \mathcal{Z} \cup \{z\}$;

end for

return Index set of clipped features $\mathcal{Z}$

the key steps below. Let $\Omega = \{1, ..., d\}$ be the full index set. We use $V = V_\Omega = [V_1, V_2, ..., V_d]$ to represent the variable of d-dimensional encoding image vectors and $g(V) \in \{\text{male}, \text{female}, \text{neutral}\}$ to represent the corresponding gender attribute. The goal is to output the index set $\mathcal{Z}$ of clipped covariates that reduce the dependence between representations $V_{\Omega/\mathcal{Z}}$ and gender attributes $g(V)$. We measure the correlation between each dimension V_i and gender attribute $g(V)$ by estimating their mutual information $I(V_i; g(V))$ [GKOV17]:

$$I(V_I; g(V)) = D_{\text{KL}}(\Pr_{(V_i, g(V))} \| \Pr_{V_i} \otimes \Pr_{g(V)}) \tag{3.13}$$

where D_{KL} is the KL divergence [KL51], $\Pr_{(V_i, g(V))}$ indicates the joint distribution, $\Pr_{V_i}$

and $\Pr_{g(V)}$ indicate their marginals. Next, we greedily clip m covariates with highest mutual information, and construct $(d-m)$-dimensional embedding vectors $V_{\Omega/\mathcal{Z}}$. m is a hyper-parameter that we will experimentally find to best trade-off accuracy and the reduced gender bias, and we show how the selection of m affects the performance later. To project text representations, denoted by variable C, into the same embedding space, we also apply the index set $\mathcal{Z}$ to obtain clipped text embedding vectors $C_{\Omega/\mathcal{Z}}$.

The clipped image and text representations, denoted by $\vec{v}^*$ and $\vec{c}^*$, will have a relatively low correlation with gender attributes due to the "loss" of mutual information. Then we compute the cosine similarity between image and text by substituting $\vec{v}^*$ and $\vec{c}^*$ into Equation (3.1):

$$S(v,c) = \frac{\vec{v}^* \cdot \vec{c}^*}{\|\vec{v}^*\|\|\vec{c}^*\|} \tag{3.14}$$

Finally, we rank the images based on the cosine similarity between the clipped representations.

3.1.4 Experiments

3.1.4.1 Datasets

We evaluate our approaches on the standard MS-COCO [CFL+15] and Flickr30K [YLHH14a] datasets. Following [KFF17] and [FFKF18], we split MS-COCO captions dataset into 113,287 training images, 5,000 validation images and 5,000 test images.[1] Each image corresponds to 5 human-annotated captions. We report the results on the test set

[1]The data is available at `cocodataset.org`.

Before Pre-processing	After Pre-processing
A **man** with a red helmet on a small moped on a dirt road.	A **person** with a red helmet on a small moped on a dirt road.
A little **girl** is getting ready to blow out a candle on a small dessert.	A little **child** is getting ready to blow out a candle on a small dessert.
A **female** surfboarder dressed in black holding a white surfboard.	A surfboarder dressed in black holding a white surfboard.
A group of young **men and women** sitting at a table.	A group of young **people** sitting at a table.

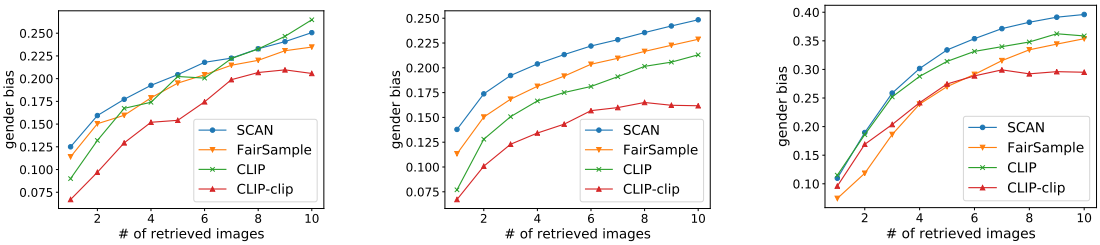
Table 3.1: Samples of the constructed gender-neutral captions. For evaluation, we convert gender-specific captions to gender-neutral ones by replacing or removing the gender-specific words.

by averaging over five folds of 1K test images or evaluating the full 5K test images. Flickr30K consists of 31,000 images collected from Flickr.[2] Following the same split of [KFF17, LCH+18], we select 1,000 images for validation, 1,000 images for testing, and the rest of the images for training.

Identifying Gender Attributes of Images Sensitive attributes such as gender are often not explicitly annotated in large-scale datasets such as MS-COCO and Flickr30K, but we observe that implicit gender attributes of images can be extracted from their associated human-annotated captions. Therefore, we pre-define a set of masculine words and a set of feminine words. Following [ZWY+17] and [BHDR18] we use the ground-

[2]The data is available at `http://bryanplummer.com/Flickr30kEntities/`.

truth annotated captions to identify the gender attributes of images. An image will be labeled as "male" if at least one of its captions contains masculine words and no captions include feminine words. Similarly, an image will be labeled as "female" if at least one of its captions contains feminine words and no captions include masculine words. Otherwise, the image will be labeled as "gender-neutral".



(a) MS-COCO 1K Test Set. (b) MS-COCO 5K Test Set. (c) Flick30K Test Set.

Figure 3.2: Gender bias analysis with different top-K results.

3.1.4.2 Models

We compare the fairness performance of the following approaches:

1. **SCAN** [LCH+18]: we use the official implementation for training and evaluation[3].

2. **FairSample**: we apply the fair sampling method proposed in Section 3.1.3.1 to the SCAN framework and adopt the same hyper-parameters suggested by [LCH+18] for training.

[3]The code is available at `https://github.com/kuanghuei/SCAN`.

3. **CLIP** [RKH+21]: we use the pre-trained CLIP model released by OpenAI.[4] The model uses a Vision Transformer [DBK+21a] as the image encoder and a masked self-attention Transformer [VSP+17] as the text encoder. The original model produces 500-dimensional image and text vectors.

4. **CLIP-clip**: we apply the feature pruning algorithm in Section 3.1.3.2 to the image and text features generated by the CLIP model. We set $m = 100$ and clip the image and text representations into 400-dimensional vectors.

Note that SCAN and FairSample are trained and tested on the in-domain MS-COCO and Flickr30K datasets, while the pre-trained CLIP model is directly tested on MS-COCO and Flickr30K test sets without fine-tuning on their training sets (same for CLIP-clip as it simply drops CLIP features).

3.1.4.3 Evaluation

Gender-Neutral Text Queries In this study, we focus on equalizing the search results of gender-neutral text queries. In addition to the existing gender-neutral captions in the test sets, we pre-process those gender-specific captions to construct a purely gender-neutral test corpus to guarantee a fair and large-scale evaluation. For every caption, we identify all these gender-specific words and remove or replace them with corresponding gender-neutral words. We show some pre-processing examples in Table 3.1.

[4]The pre-trained model is available at `https://github.com/openai/CLIP`.

Metrics We employ the fairness metric in Equation (3.5), Bias@K, to measure the gender bias among the top-K images. In addition, following standard practice, we measure the retrieval performance by Recall@K, defined as the fraction of queries for which the correct image is retrieved among the top-K images.

3.1.4.4 Main Results on MS-COCO & Flickr30K

We report the results comparing our debiasing methods and the baseline methods in Table 3.2.

Model Bias Although the pre-trained CLIP model is evaluated without fine-tuning, we observe that it achieves a comparable recall performance with the SCAN model on MS-COCO and dominates the Flickr30K dataset. However, both models suffer from severe gender bias. Especially, the Bias@10 of the SCAN model on Flickr30K is 0.3960, meaning nearly 70% of the retrieved gender-specific images portray men and only 30% portray women. Similarly, the CLIP model achieves 0.2648 gender bias on MS-COCO 1K test set, indicating about 6.4 out of 10 retrieved images portray men while about 3.6 out of 10 portray women. Given that all of the testing text queries are gender-neutral, this result shows that severe implicit gender bias exists in image search models.

Debiasing Effectiveness As shown in Table 3.2, both the in-processing sampling strategy *FairSample* and the post-processing feature pruning algorithm *clip* consistently mitigate the gender bias on test data. For instance, among the top-10 search images,

		Gender Bias↓			Recall↑		
Dataset	Method	Bias@1	Bias@5	Bias@10	Recall@1	Recall@5	Recall@10
COCO1K	SCAN	.1250	.2044	.2506	47.7	82.0	**91.0**
	FairSample	.1140	.1951	.2347	**49.7**	**82.5**	90.9
	CLIP	.0900	.2024	.2648	48.2	77.9	88.0
	CLIP-clip	**.0670**	**.1541**	**.2057**	46.1	75.2	86.0
COCO5K	SCAN	.1379	.2133	.2484	25.4	54.1	67.8
	FairSample	.1133	.1916	.2288	26.8	**55.3**	**68.5**
	CLIP	.0770	.1750	.2131	**28.7**	53.9	64.7
	CLIP-clip	**.0672**	**.1474**	**.1611**	27.3	50.8	62.0
Flickr30K	SCAN	.1098	.3341	.3960	41.4	69.9	79.1
	FairSample	**.0744**	**.2699**	.3537	35.8	67.5	77.7
	CLIP	.1150	.3150	.3586	**67.2**	**89.1**	**93.6**
	CLIP-clip	.0960	.2746	**.2951**	63.9	85.4	91.3

Table 3.2: Results on MS-COCO (1K and 5K) and Flickr30K test sets. We compare the baseline models (SCAN [LCH$^+$18] and CLIP [RKH$^+$21]) and our debiasing methods (FairSample and CLIP-clip) on both the gender bias metric Bias@K and the retrieval metric Recall@K.

SCAN with FairSample reduces gender bias from 0.3960 to 0.3537 (decreased by 10.7%) on Flickr30K. Using the clipped CLIP features for image search (CLIP-clip), the gender bias drops from 0.2648 to 0.2057 (22.3%) on MS-COCO 1K, from 0.2131 to 0.1611 (24.4%) on MS-COCO 5K, and from 0.3586 to 0.2951 (17.7%) on Flickr30K. For the

tradeoff, CLIP-clip sacrifices the recall performance slightly (from 93.6% Recall@10 to 91.3% on Flickr30K). On the other hand, SCAN with FairSample even achieves a comparable recall performance with SCAN.

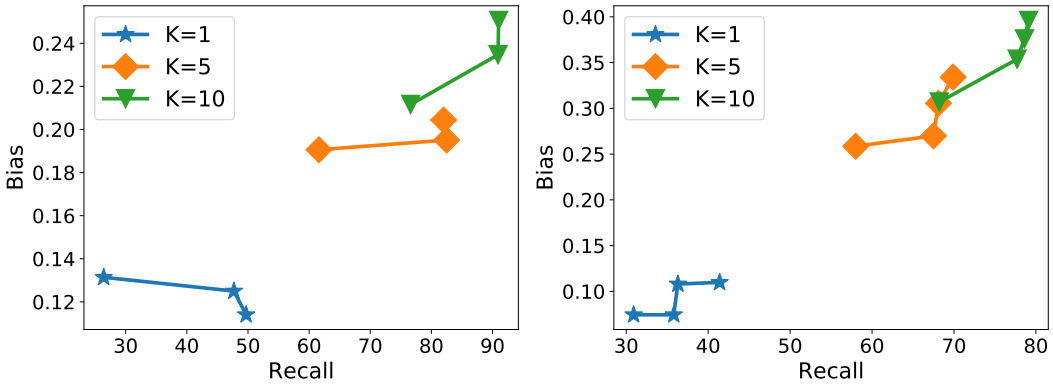
3.1.4.5 Gender Bias at Different Top-K Results

We plot how gender bias varies across different values of K (1-10) for all the compared methods in Figure 3.2. We observe that when $K < 5$, the gender bias has a higher variance due to the inadequate retrieved images. When $K \geq 5$, the curves tend to be flat. This result indicates that Bias@10 is more recommended than Bias@1 for measuring gender bias as it is more stable. It is also noticeable that CLIP-clip achieves the best fairness performance in terms of Bias@10 consistently on all three test sets compared to the other models.

3.1.4.6 Tradeoff between Recall and Bias

There is an inherent tradeoff between fairness and accuracy in fair machine learning [ZG19b]. To achieve the best recall-bias tradeoff in our methods, we further examine the effect of the controlling hyper-parameters: the weight α in FairSampling and the number of clipped dimensions m in CLIP-clip.

Figure 3.3 demonstrates the recall-bias curve with the fair sampling weight $\alpha \in [0, 1]$. Models of higher recall often suffer higher gender bias, but the fairness improvement outweighs the recall performance drop in FairSample models. For example,



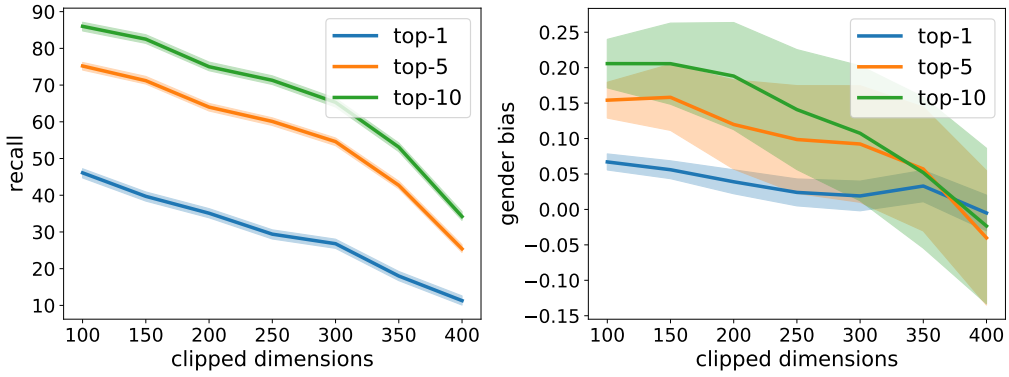
(a) MS-COCO 1K test set.

(b) Flickr30K test set.

Figure 3.3: The Pareto frontier of recall-bias tradeoff curve for FairSample on MS-COCO 1K and Flickr30K

the model fully trained with fair sampling ($\alpha = 1$) has the lowest bias and drops the recall performance the most—it relatively reduces 22.5% Bias@10 but only decreases 10.9% Recall@10 on Flickr30K. We choose $\alpha = 0.4$ for the final model, which has a better tradeoff in retaining the recall performance.

As shown in Figure 3.4, we set the range of the clipping dimension m between 100 and 400 on MS-COCO 1K. We find that clipping too many covariates (1) harms the expressiveness of image and text representations (Recall@1 drops from 46.1% to 11.3%, Recall@5 drops from 75.2% to 25.4%, and Recall@10 drops from 86.0% to 34.2%), and (2) causes high standard deviation in gender bias. In light of the harm on expressiveness, we select $m = 100$ for conventional use.



(a) recall

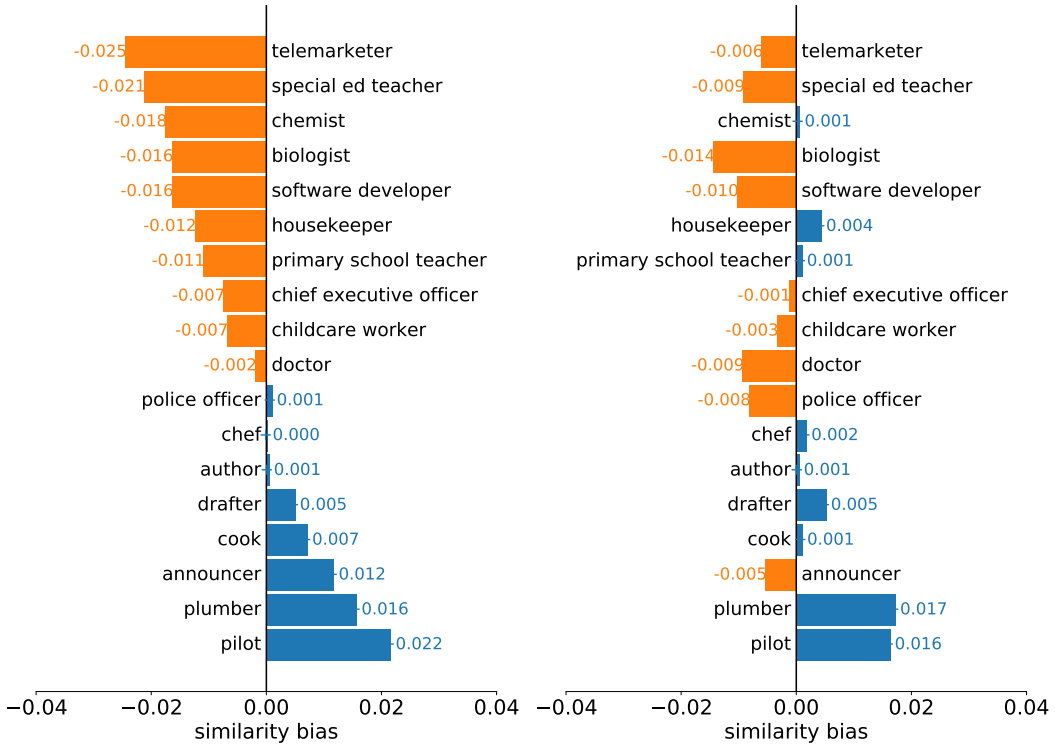
(b) gender bias

Figure 3.4: Effect of the number of clipped dimensions m on performance of recall and bias on MS-COCO 1K.

3.1.5 Evaluation on Internet Image Search

The aforementioned evaluation results on MS-COCO and Flickr30K datasets are limited that they rely on gender labels extracted from human captions. In this sense, it is important to measure the gender biases on a benchmark where the gender labels are identified by crowd annotators. To this end, we further evaluate on the `occupation` dataset [KMM15], which collects top 100 Google Image Search results for each gender-neutral occupation search term.[5] Each image is associated with the crowd-sourced gender attribute of the participant portrayed in the image. Inspired by [BHDR18] and [TDL+20], we measure the gender bias by computing the difference of expected cosine similarity between male and female occupational images. Given an occupation o, the

[5]The data is available at `https://github.com/mjskay/gender-in-image-search`.



(a) CLIP

(b) CLIP-clip

Figure 3.5: Gender bias evaluation of internet image search results on `occupations`. We visualize the similarity biases on 18 occupations. ■ indicates the occupation is biased towards males and ■ indicates it is biased towards females. The clip algorithm mitigates gender bias for a variety of occupations.

similarity bias is formulated as

$$\text{Bias} = \mathbb{E}_{v \in \mathcal{V}^o_{\text{male}}} S(v, o) - \mathbb{E}_{v \in \mathcal{V}^o_{\text{female}}} S(v, o) \tag{3.15}$$

where $\mathcal{V}^o_{\text{male}}$ and $\mathcal{V}^o_{\text{female}}$ are the sets of images for occupation o, labeled as "male" and "female".

Figure 3.5 demonstrates the absolute similarity bias of CLIP and CLIP-clip on the `occupation` dataset for 18 occupations. We observe that the CLIP model exhibits severe similarity discrepancy for some occupations, including telemarketer, chemist, and housekeeper, while the *clip* algorithm alleviates this problem effectively. Note that for doctor and police officer, the CLIP-clip model exaggerates the similarity discrepancy, but the similarity bias is still less than 0.01. In general, CLIP-clip is effective for mitigating similarity bias and obtains a 42.3% lower mean absolute bias of the 100 occupations than the CLIP model (0.0064 *vs.* 0.0111).

3.2 Multilingual Fairness

3.2.1 How Do We Assess Fairness for Pre-trained Multilingual and Multimodal Representations?

Recently pre-trained vision-and-language representations [LBPL19, TB19, SZC+20, LDF+20, CLY+20, LYL+20, GCL+20, YTY+21, DJ21, RKH+21, CLTB21] have received a surge of attention. Such pre-trained multimodal representations have shown great capabilities of bridging images and natural language on the downstream tasks,

including image captioning [LRN19], image retrieval [VJS+19], visual QA [ZPZ+20], text-to-image generation [RPG+21], etc. While it is commonly recognized that the multimodal representations trained on English corpora can be generalized to multilingualism by cross-lingual alignment [LC19, CKG+20], recent studies criticize that the multilingual textual representations *do not* learn equally high-quality representations for all the languages [WD20], especially for low-resource languages. [HRS+20] emphasize the need for general-purpose representations to seek equal performance across all languages. However, there is still a lack of a nuanced understanding of how multilingual representations fare on vision-and-language benchmarks.

This section provides a novel perspective for analyzing the principles of multilingual fairness in multimodal representations from two aspects. First, existing frameworks for measuring multilingual biases usually emulate text sources in different languages, which may have ambiguous meanings in varied contexts [GBH+20]. In contrast, we leverage visual grounding as the anchor to bridge text in different languages—text snippets in different languages but with similar semantics should be equitably relevant to the same images. Second, we equate a language as an aggregated group of individuals (e.g., French as a group of French sentences) in the terminology of fairness. As [CD21] has pointed out, "*each language has a distinct identity, defined by its vocabulary, syntactic structure, its typological features, amount of available resources, and so on.*" The notions of fairness, such as individual fairness [DHP+12a] and group fairness [ZWS+13, Cho17b, HPS16b, ZLL22], can be naturally adapted by comparing the

multimodal model's treatment across languages.

Therefore, we introduce two fairness notions: *multilingual individual fairness* presumes similar outcomes between similar language expressions grounding on the same images; *multilingual group fairness* postulates that multimodal models should induce similar predictive performance across different languages. These fairness notions are formalized to *compare the multimodal model's treatment of one language versus another* for either the individual target or the aggregated group.

3.2.2 Multilingual Individual Fairness

For an ideal multilingual vision-and-language model, text descriptions in different languages referring to similar semantic meanings should be equally similar or dissimilar to the same grounding images. We note that there are no language expressions that are perfectly identical to each other in real-world scenarios due to linguistic features. Nevertheless, at least in a normal vision-and-language task, multilingual models are desired to impose equal treatment to different languages. For instance, "this is a cat" (in English) and "das ist eine Katze" (in German) should be similarly related to an image of a cat in image-text retrieval. This intuition aligns with individual fairness in a multilingual manner. In this section, we investigate to what degree multilingual representations are individually fair.

Individual fairness requires that similar people should be treated similarly [DHP+12a]. In our multilingual setting, we require that the text snippets expressing

similar semantics in different languages should be similarly related to the same images. Taking the Euclidean distance function to measure the distance between text features, we can define α-multilingual individual fairness by:

Definition 3 (Multilingual Individual Fairness)**.** Given a set of image-text pairs $\{(I, T)\}$, a multimodal model M satisfies α-multilingual individual fairness if for all (I, T), for languages L and L':

$$|S(I, T^{(L)}) - S(I, T^{(L')})| \leq \alpha \|\boldsymbol{b} * t^{(L)} - \boldsymbol{b} * t^{(L')}\|$$

where $\boldsymbol{b} * t^{(L)}$ is the textual representation vector yielded by M in language L.

Here, α is a parameter to control the ratio of similarity gap to the text feature vectors' distance, and smaller α indicates the model is individually fairer. Note that the similarity gap is at most 2, because the range of cosine similarity is $[0, 1]$. In general settings, $S(I, T)$ is measured by the cosine similarity between the encoded visual vector $\boldsymbol{b} * v$ and textual vector $\boldsymbol{b} * t$.

Lemma 3.1. *Denote $\mathcal{O}_\rho(\boldsymbol{b} * t) = \{\boldsymbol{b} * x \mid \|\boldsymbol{b} * x - \boldsymbol{b} * t\| \leq \rho\}$ to be a closed ball of radius $\rho > 0$ and center $\boldsymbol{b} * t$. Then for any visual representation vector $\boldsymbol{b} * v$,*

$$\sup_{\substack{\boldsymbol{b}*t^{(L')} \in \mathcal{O}_\rho(\boldsymbol{b}*t^{(L)}) \\ 0 \leq \rho < \|\boldsymbol{b}*t^{(L)}\|}} |S(\boldsymbol{b}*v, \boldsymbol{b}*t^{(L')}) - S(\boldsymbol{b}*v, \boldsymbol{b}*t^{(L)})| \leq \sqrt{2(1 - \sqrt{1 - (\frac{\rho}{\|\boldsymbol{b} * t^{(L)}\|})^2})} \quad (3.16)$$

where $S(\cdot, \cdot)$ denotes the cosine similarity, $\boldsymbol{b} * t^{(L)}$ and $\boldsymbol{b} * t^{(L')}$ are textual representation vectors for languages L and L', respectively.

Lemma 3.1 implies that when the distance between multilingual textual representation vectors is bounded, the similarity with images can be bounded in terms of their distance. It is worth noting that the bounds are independent of the visual representation vectors. Nevertheless, the form of upper bound in Lemma !3.1 is a bit sophisticated, and can be simplified when $\rho \ll \|\boldsymbol{b} * t^{(L)}\|$.

Theorem 3.2. *When* $\|\boldsymbol{b} * t^{(L')} - \boldsymbol{b} * t^{(L)}\| \ll \|\boldsymbol{b} * t^{(L)}\|$,

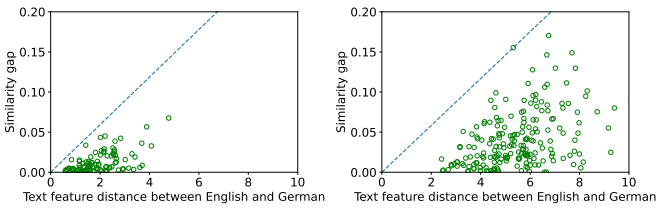
$$|S(\boldsymbol{b} * v, \boldsymbol{b} * t^{(L')}) - S(\boldsymbol{b} * v, \boldsymbol{b} * t^{(L)})| \lessapprox \frac{\|\boldsymbol{b} * t^{(L')} - \boldsymbol{b} * t^{(L)}\|}{\|\boldsymbol{b} * t^{(L)}\|}.$$

Theorem 3.2 is a direct application of Lemma 3.1 when the distance between multilingual vectors is small enough, and extends in many natural cases to approximate the multilingual individual fairness with $\alpha \approx \frac{1}{\|\boldsymbol{b} * t^{(L)}\|}$. Theorem 3.2 implicates to what degree the multimodal model satisfies individual fairness when text snippets are well aligned between different languages.

3.2.2.1 Evaluation

The theoretical analysis on multilingual individual fairness implies that the ratio of similarity difference to their text feature distance can be bounded by the reciprocal of the length of text feature vectors. To verify the implication, we conduct experiments on the Multi30K dataset [EFSS16].

Dataset. The Multi30K dataset [EFSS16] contains 31,014 Flickr30K [YLHH14b] images and composes the *translation* and the *independent* portions of English-German cap-



(a) Translation (b) Independent

	translation	independent
en	50.4	59.2
de	45.6	36.3
Gap	4.8	22.9

(c) Accuracy disparity

Figure 3.6: We empirically examine how does the multilingual CLIP fare on the translation and the independent portions. We also evaluate the accuracy disparity for image-text matching, and find out that the independent portion incurs huge accuracy disparity compared with the translation portion.

tion pairs. The German translations were collected from professional English-German translators by translating the English captions without seeing the images, one per image. The independent portion was independently annotated by German crowdworkers after seeing the images instead of English captions, five per image. Hence, the translated captions are strongly aligned in both languages, while the independent descriptions may have distinct context. We use 1,000 test images for our evaluation. For the independent portion, we select the first English caption and the first German caption of the five to pair with the image for a fair comparison.

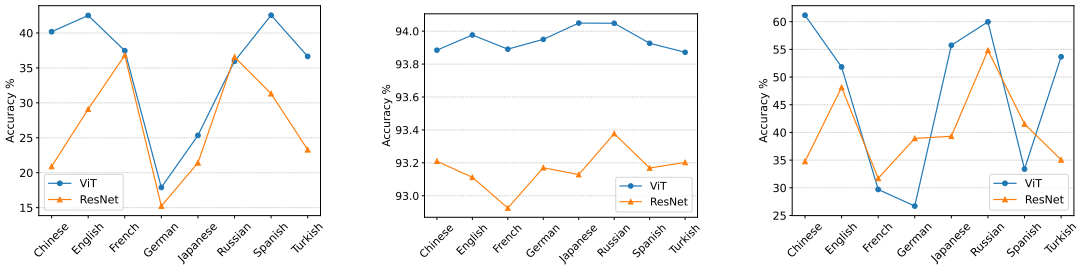
Results. We embed each English-German caption pair into textual representation vectors and the corresponding image into visual representation vectors. We compute the Euclidean distance between English-German text features, as well as the cosine similarity with respect to the image features. We plot their cross-lingual gap on the translation and the independent portions in Figure 3.6. For both portions, the blue

dashed lines represent the empirical upper bounds of the ratio between similarity gap and text feature distance.

Unsurprisingly, we find out that the English-German captions are more closely aligned on the translation portion (the average textual feature distance is 1.86) than the independent portion (average distance is 5.69). The similarity gaps regarding the translation portion are below 0.06 in general, and those regarding the independent portion are above 0.10 for many instances. The reason is apparent: translated captions have more similar semantics owning to the professional text-to-text translations, while independent captions have more diverse expressions of the same images, even if they might refer to the same content.

On the other hand, we observe that the slopes of blue dashed lines for translation and independent portions are approximate to each other, i.e., the empirical α for both portions are similar. This fact implies that the multilingual CLIP model evaluated on two different text corpora share a similar level of individual fairness, even though the cross-lingual similarity gaps are quite different. We also note that the empirical upper bound of α are much smaller than the theoretical upper bound $\frac{1}{\|\boldsymbol{b}*t^{(L)}\|}$ in Theorem 3.2.

Although we have verified that multilingual multimodal representations satisfy similar individual fairness, we demonstrate that they violate group fairness by evaluating their image-text matching accuracy. We find out that English captions dominate the Top-1 image-text matching accuracy over German captions, with 4.8% higher on the translation portion and 22.9% higher on the independent portion. This observation



(a) Race classification (b) Gender classification (c) Age classification

Figure 3.7: Race, gender, and age classification accuracy across different languages.

delivers an important message for researchers who are interested in learning fair representations [RBFV20]: individual fairness *does not* flatly prevent accuracy disparity among different languages [Bin20].

3.2.3 Multilingual Group Fairness

Distinct from individual fairness, multilingual group fairness appeals to the idea that multimodal models should achieve equivalent predictive performance across different languages. From the perspective of representations, it is hard to carry out this demand without well-defined tasks and metrics. Hence it is natural to ask how to define group fairness in this scenario properly? In this section, we shall answer this question by equating language as a unique dimension of group membership relating to the text modality. We formulate the criteria by equalizing the accuracy rates over different languages. We also observe that images are often connected to people in protected or unprotected groups. Given the image-text pairs, we consider the accuracy disparity

across different languages conditioning the subgroup of images.

3.2.3.1 Equality of Accuracy across Languages

Given a dataset $\mathcal{D}$ consisting of ground-truth image-text pairs $\{(I_i, T_i)\}$ and each text can be in different languages. The goal of a multimodal model M is to predict the similarity $S(I_i, T_j)$ for any image I_i and text T_j. Then the model matches $\hat{T}_i$ for images I_i by selecting the text with highest similarity scores, i.e., $\hat{T}_i = \arg\max_j S(I_i, T_j)$.

$$\texttt{Acc}(M) = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \mathbb{1}[\hat{T}_i = T_i] \tag{3.17}$$

We use the superscript (L) to indicate the accuracy $\texttt{Acc}^{(L)}$ is evaluated in language L. Next, we take language as group membership and define multilingual accuracy parity by equalizing accuracy across languages.

Definition 4 (multilingual accuracy parity)**.** A multimodal model M satisfies *multilingual accuracy parity* if $\texttt{Acc}^{(L)}(M) = \texttt{Acc}^{(L')}(M)$ for all languages L, L'.

In practice, it is impossible to achieve accuracy parity for all languages. Following [HRS+20], we use

$$\texttt{Gap}_M(L, L') = |\texttt{Acc}^{(L)}(M) - \texttt{Acc}^{(L')}(M)| \tag{3.18}$$

to represent the cross-lingual gap for model M.

3.2.3.2 When Language Meets Groups in Images

The above discussion on group fairness considers language as the sole group membership. In the real-world image and text applications, the people portrayed in the images are often associated with protected groups. For instance, the face attribute dataset [LLWT15a] contains sensitive attributes, such as race, age and gender. Let G denote the group membership of images and $\mathcal{D}_a$ denote the subset of data examples $\mathcal{D}$ given $G = a$. The accuracy of a multimodal model evaluated on the images of subgroup a is defined as

$$\texttt{Acc}_a(M) = \frac{1}{|\mathcal{D}_a|} \sum_{\mathcal{D}_a} \mathbb{1}[\hat{T}_i = T_i] \tag{3.19}$$

When language is connected to images of different groups, we can define accuracy disparity between group a and group b with respect to model M within language L as

$$\texttt{Disp}_M^{(L)}(a, b) = |\texttt{Acc}_a^{(L)}(M) - \texttt{Acc}_b^{(L)}(M)| \tag{3.20}$$

`Disp` represents the group rate gap in a single language. Mirroring *multilingual accuracy parity*, we can define the *multilingual group rate parity* as below.

Definition 5 (multilingual group rate parity)**.** A multi-modal model M satisfies multilingual group rate parity if $\texttt{Disp}_M^{(L)}(a, b) = \texttt{Disp}_M^{(L')}(a, b)$ with respect to groups a, b associated with images for all languages.

Definition 4 and Definition 5 evaluate the fairness of multilingual representations from diverse aspects. More broadly, we may be interested in the accuracy gap

between different combinations of languages and groups. A common case is that there are only two protected groups (e.g. female and male, young and old). Let $p_a = \frac{|\mathcal{D}_a|}{|\mathcal{D}|}$ and $p_b = \frac{|\mathcal{D}_b|}{|\mathcal{D}|}$ represent the population proportions of group a and group b respectively, satisfying $p_a + p_b = 1$. Then we can decompose the cross-lingual cross-group accuracy disparity as below:

Proposition 3.3. *When there are only two protected groups a and b, the following inequality holds for any two languages L and L'*

$$|\texttt{Acc}_a^{(L)} - \texttt{Acc}_b^{(L')}| \leq \texttt{Gap}(L, L') + p_b \cdot \texttt{Disp}^{(L)}(a, b) + p_a \cdot \texttt{Disp}^{(L')}(a, b) \tag{3.21}$$

Proposition 3.3 guarantees that the accuracy disparity between any combinations of languages and protected groups can be upper bounded by a variety of factors, and implicates that we only need to focus on cross-lingual gap and group rate gap measures to assess multilingual group fairness. In what follows, we will take a closer look at how the multilingual CLIP model performs with compositions of languages and protected groups under these fairness criteria.

3.2.3.3 Evaluation on Multilingual Accuracy Disparity

Dataset. FairFace [KJ21] is a face attribute dataset for the balanced race, gender, and age groups. It categorizes gender into two groups, including female and male, and race into seven groups, including White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. For ages, we categorize the raw labels into five groups: infants

(0–2), children and adolescents (2–19), adults (20–49), middle age adults (50–69), and seniors (more than 70). We follow their original data split and select the validation set consisting of 10,940 face images for evaluation.

Languages. We analyze the multilingual group fairness for 8 languages: Chinese (zh), English (en), French (fr), German (de), Japanese (ja), Russian (ru), Spanish (es), and Turkish (tr). We select English as the pivot language and write natural language prompts in English. Then we translate them into other languages: we first use Google Translate and then recruit native speakers to rate the prompts and fix any potential errors on Amazon Mechanical Turk. The rationale for only using English as the pivot language is that the multilingual CLIP [CE21] selects English as the pivot language for aligning multilingual text embeddings.

Text Prompts. Following [RKH+21], we construct the text prompt by the template "A photo of a {`label`} person". Concretely, for gender classification, we construct the text prompt "A photo of a woman" when the gender attribute is female, and construct "A photo of a man" otherwise. For race classification, we construct the text prompt by "A photo of a(n) {`race`} person". Note that Indian actually refers to South Asian ethnic groups in the Fairface race taxonomy [KJ21] but it can refer to Native Americans as well. To avoid ambiguity, we replace "Indian" by "South Eastern" to construct the prompts. For age classification, we notice that the age attributes in Fairface dataset

are numeric values and use the template "A photo of a person aged {`age`} years" to construct text prompts.

Results. We probe the multilingual accuracy disparity for race classification, gender classification, and age classification, as shown in Figure 3.7. We use two different pre-trained image encoders for extracting visual representation vectors, including Vision Transformer [DBK+21b] and ResNet-50 [HZRS16]. We observe that:

Cross-lingual gap varies across different protected groups. The predictive accuracy for gender classification is consistently higher than 90% across all the languages. In contrast, the multimodal model has relatively poor performance and more considerable variance for race and age classification. Furthermore, race classification yields 24.66% accuracy disparity and age classification yields 34.47% accuracy disparity for Vision Transformer. This implies that the huge disparity may result from the poor predictive performance of the model.

Visual representations affect accuracy disparity. For race classification, Vision Transformer features generally achieve higher accuracy across all languages than ResNet-50 (34.82% *vs.* 26.83% on average) except for Russian. The standard deviation of Vision Transformer is higher than ResNet-50 (8.18% *vs.* 7.34%). The maximal accuracy gap for Vision Transformer is 30.40% between German and Spanish, while the maximal accuracy gap for ResNet-50 is 23.12% between German and French. For gender classification, Vision Transformer dominantly achieves higher accuracy and incurs less

accuracy gap. For age classification, the accuracy is moderately low for all languages. However, Vision Transformer has 63.1% accuracy in Chinese while only 25.8% accuracy in German, exaggerating the accuracy gap between languages.

In Table 3.3, we present the complete results of Figure 3.8 by compositions of gender and race groups across different languages.

3.2.3.4 Evaluation on Multilingual Group Rate Disparity

We evaluate multilingual group rate disparity for gender classification on Fairface dataset. We follow the same setup as described in Section 3.2.3.3 and measure the gender gap given by Equation (3.20), where a is the composition of male and various race groups, b is the composition of female and various race groups. We try to answer the following research questions:

How do gender gaps differ across protected groups? We plot the gender accuracy gap across different languages and racial groups in Figure 3.8. It is clearly shown that Black and Southeast Asian groups dominantly exhibit larger gender gaps than other groups. We also observe that French has a similar performance with English. We conjecture this is because English and French share the same alphabet and similar syntactic structures. Besides, as shown in Table 3.3, English and French have the largest race inequality regarding gender gap—nearly zero gender gaps for White but near the maximal gaps for Black.

Are gender gaps amplified for different languages when compared

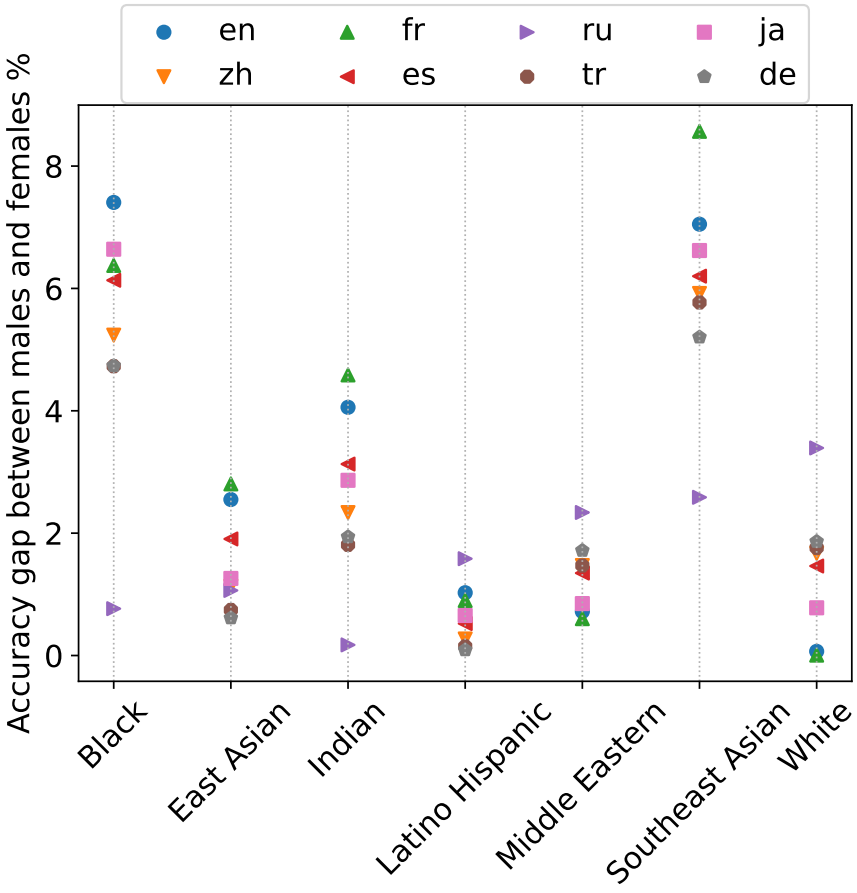


Figure 3.8: Gender accuracy gap across different languages and racial groups. Black and Southeast Asian people face significant larger gender gaps than other racial groups in most languages.

with English? We report the accuracy gap on gender classification of FairFace images by race groups across different languages in Table 3.3. We take English as the pivot language and examine whether the accuracy gaps by race groups are amplified for other languages. Compared with English, accuracy gaps for White and Middle Eastern groups are generally amplified for other languages. On the other hand, accuracy gaps are

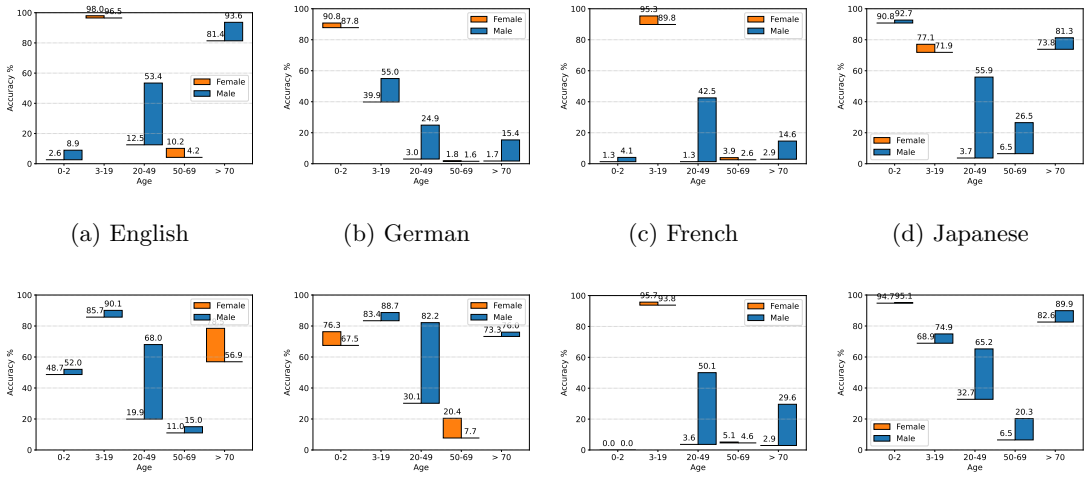
Table 3.3: Gender classification accuracy of FairFace images by race groups across different languages.

Language	Gender	White	Black	Indian	East Asian	Southeast Asian	Middle Eastern	Latino	Average
English	Female	95.1	90.9	94.5	95.2	96.0	96.0	94.2	94.6
	Male	95.2	83.5	90.4	92.7	89.0	96.7	93.2	91.5
	Disp	0.1	**7.4**	4.1	2.5	7.0	0.7	1.0	3.0
German	Female	93.8	90.1	94.0	94.2	95.0	95.5	93.9	93.8
	Male	95.6	85.4	92.0	93.6	89.8	97.2	93.9	92.5
	Disp	1.9	4.7	1.9	0.6	5.2	1.7	0.1	1.3
French	Female	95.0	90.4	94.6	95.0	96.3	95.7	94.2	94.5
	Male	95.0	84.0	90.0	92.1	87.8	96.3	93.3	91.2
	Disp	0.0	6.4	**4.6**	**2.8**	**8.6**	0.6	0.9	**3.2**
Japanese	Female	94.5	90.6	94.4	94.7	95.7	95.7	94.1	94.2
	Male	95.3	84.0	91.5	93.4	89.1	96.6	93.4	91.9
	Disp	0.8	6.6	2.9	1.3	6.6	0.8	0.7	2.3
Turkish	Female	93.9	90.0	93.8	94.6	95.3	95.5	94.1	93.9
	Male	95.6	85.2	92.0	93.8	89.5	96.9	93.9	92.4
	Disp	1.8	4.7	1.8	0.7	5.8	1.5	0.1	1.4
Russian	Female	93.0	88.4	93.1	93.4	94.6	95.2	93.4	93.0
	Male	96.4	87.6	93.2	94.5	92.0	97.5	95.0	93.7
	Disp	**3.4**	0.8	0.2	1.1	2.6	**2.3**	**1.6**	0.7
Spainish	Female	94.1	90.5	94.4	95.1	95.6	95.5	94.2	94.2
	Male	95.5	84.4	91.2	93.2	89.4	96.8	93.7	92.0
	Disp	1.5	6.1	3.1	1.9	6.2	1.3	0.5	2.2
Chinese	Female	93.9	90.1	94.1	94.8	95.4	95.5	94.2	94.0
	Male	95.5	84.9	91.8	93.7	89.5	96.9	93.9	92.3
	Disp	1.7	5.2	2.3	1.1	5.9	1.5	0.3	1.7

generally mitigated for groups including Black, Indian, East Asian, Southeast Asian, and Latino groups. The averaged cross-lingual gaps are mitigated for all the languages except for French.

We also evaluate multilingual group rate disparity for age classification. We composite gender and age as the group membership. We plot the age classification

accuracy by female and male groups across different languages in Figure 3.9. The blue bars indicate that the male group has higher accuracy than the female group, while the orange bars indicate that the female group has higher accuracy than the male group. The heights of bars represent the accuracy gaps between male and female groups. In general, the male group has higher accuracy than the female group. Especially, adults (20–49 years old) consistently suffer huge gender gaps across all the languages, with the largest gap 52.2% for Japanese. It is worth noting that the numerals to express ages are



(a) English (b) German (c) French (d) Japanese

(e) Turkish (f) Russian (g) Spanish (h) Chinese

Figure 3.9: Age classification accuracy across female and male groups for different languages. The blue bars indicate that the male group has higher accuracy than the female group, while orange bars indicate that the female group has higher accuracy. The heights of bars represent the accuracy gaps between male and female groups.

identical in text prompts for different languages, e.g., "a person aged 20 to 49 years" in English versus "eine Person im Alter von 20 bis 49 Jahren" in German. This controlled experiment helps us better understand whether the identical numeric digits have distinct

meanings in multilingual contexts. As shown in Figure 3.9, although text prompts in different languages share the same numerals of ages, the yielding accuracy exhibits significant disparity across languages. One prominent example is that the predictive accuracy for infants (0–2 years old) is 5.8% for English and 2.6% for French, but 89.4% for German and 91.6% for Japanese, implying the presence of significant cross-lingual accuracy gaps.

3.3 Text-to-Image Association Test

3.3.1 Motivating Example

Recent progress on generative image models has centered around utilizing text prompts to produce high quality images that closely align with the provided natural language descriptions [RDN+22, NDR+22, SCS+22, YXK+22, CZB+23]. Easy access to these models, notably the open-sourced Stable Diffusion model [RBL+22], has made it possible to develop them for a wide range of downstream applications at scale, such as generating stock photos [Rae22], and creating creative prototypes and digital assets [Ope22].

The success of text-to-image generation was enabled by the availability and accessibility of massive image-text paired datasets scraped from the web [SBV+22]. However, it has been shown that data obtained by these curations may contain human biases in various ways [BPK21]. Selection bias occurs when the data is not properly

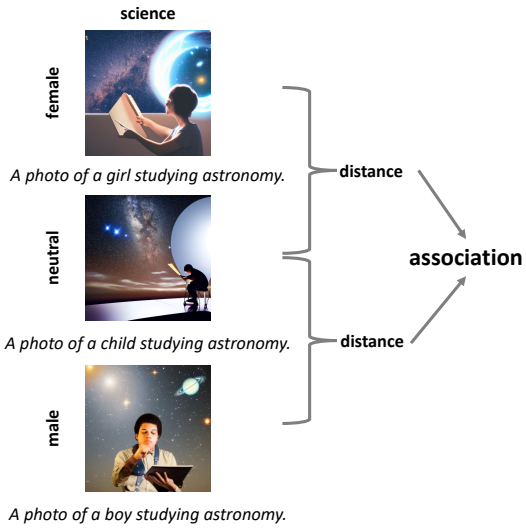


Figure 3.10: **Text-to-Image Association Test (T2IAT) procedure.** We instantiate the proposed bias test on Gender-Science. We use the text prompt "A photo of a child studying astronomy" to generate neutral images. Then we substitute "child" with feminine and masculine words and generate attribute-specific images. We calculate the average difference in the distance between the neutral and attribute-specific images as a measure of association.

collected from a diverse set of data sources, or the sources themselves do not properly represent groups of populations of interest. For example, it is reported that near half of the data samples of ImageNet came from the United States, while China and India, the two most populous countries in the world, were the contributors of only a small portion of the images [SHB+17]. It is important to be aware that the generative models trained on such datasets may replicate and perpetuate the biases in the generated images [WBC22].

Our work seeks to quantify the implicit human biases in text-to-image generative models. A large body of literature has identified the social biases pertaining to

gender and skin tone by analyzing the distribution of generated images across different social groups [BYMC22, CZB22]. These bias metrics build on the assumption that each generated image only associates with a single protected group of interest. However, in reality, the images might not belong to any of the protected groups when there is no discernible human subject or the appearances of the detectable human subjects are blurred and unclear. Moreover, the images may belong to multiple demographic groups when more than one human subjects are present in the image. Therefore, these bias measures can easily fail to detect the subtle differences between the visual concepts reified in the images and the attributes they are associated with.

Unlike previous studies, our work aims to provide a nuanced understanding on more complex stereotypical biases in image generations than the straightforward demographic biases. Examples of the complex stereotypes includes: there is a belief that boys are inherently more talented at math, while girls are more adept at language [NSS+09]; people with lighter skin tones are more likely to be appeared in home or hotel scenes, while people with dark skin tones are more likely to co-occur with object groups like vehicle [WNR20]. We investigate how these biases will be reified and quantified in machine generated images, with a special focus on valence (association with negative or unpleasant *vs*. positive or pleasant concepts) and stereotypical biases.

In this section, we propose the Text-to-Image Association Test (T2IAT), a systematic approach to measure the implicit biases of image generations between target concepts and attributes (see Figure 3.10). One benefit of our bias test procedure is

that it is not limited to specific demographic attributes. Rather, the bias test can be applied to a wide range of concepts and attributes, as long as the observed discrepancy between them can be justified as stereotyping biases by the model owners and users. For use cases, we conduct 8 image generation bias tests and the results of the tests exhibit various human-like biases at different significance levels as previously documented in social psychology.

We summarize our contribution as two-fold: first, we provide a generic test procedure to detect valence and stereotypical biases in image generation models. Second, we extensively conduct a variety of bias tests to provide evidence for the existence of such complex biases along with significance levels.

3.3.2 Approach

In this work, we adapt the Implicit Association Test (IAT) in social psychology to the task of text-to-image generation. We will first introduce the long history of association tests. But existing bias tests are primarily focusing on word embeddings. Therefore, we present the Text-to-Image Association Test (T2IAT), which quantifies the human biases in images generated by text-to-image generation models.

3.3.2.1 Implicit Association Test

In social psychology, the Implicit Association Test (IAT) introduced by [GMS98] is an assessment of implicit attitudes and stereotypes where the test subjects are held un-

consciously, such as associations between concepts (*e.g.* people in light/dark skin color) and evaluations (*e.g.* pleasant/unpleasant) or stereotypes. In general, IAT can be categorized into valence IATs, in which concepts are tested for association with positive or negative valence, and stereotype IATs, in which concepts are tested for association with stereotypical attributes (*e.g.* "male" *vs.* "female"). During a typical IAT test procedure, the participants will be presented with a series of stimuli (*e.g.*, pictures of black and white faces, words related to gay and straight people) and are asked to categorize them as quickly and accurately as possible using a set of response keys (e.g., "pleasant" or "unpleasant" for valence evaluations, "family" or "career" for stereotypes). The IAT score is interpreted based on the difference in response times for a series of categorization tasks with different stimuli and attributes, and higher scores indicate stronger implicit biases. For example, the Gender-Career IAT indicates that people are more likely to associate women with family and men with careers.

IAT was adapted to the field of natural language processing by measuring the associations between different words or concepts for language models [CBN17b]. Specifically, a systematic method, Word Embedding Association Test (WEAT), is proposed to measure a wide range of human-like biases by comparing the cosine similarity of word embeddings between verbal stimuli and attributes. More recently, WEAT was extended to compare the similarity between embedding vectors for text prompts instead of words [MWB$^+$19, BDC20, GC21b].

3.3.2.2 Text-to-Image Association Test

We borrow the terminology of association test from [CBN17b] to describe our proposed bias test procedure. Consider two sets of *target* concepts $\mathcal{X}$ and $\mathcal{Y}$ like science and art, and two sets of *attribute* concepts $\mathcal{A}$ and $\mathcal{B}$ like men and women. The null hypothesis is that, regardless of the attributes, there is no difference in the association between the sets of images generated with the target concepts. In the context of Gender-Science bias test, the null hypothesis is saying that no matter whether the text prompts describe science or arts, the generative models should output images that are equally associated with women and men. We note that in such a gender stereotype setting, a naïve way to measure association is to count the numbers of men and women who appeared in the generated images. This simplified measure reduces the fairness criteria to ensure that the image generation should contain the equal size of pictures depicting women and men, which has been adopted in many prior works [TSZ20, BYMC22].

To validate the significance of the null hypothesis, we design a standard statistical hypothesis test procedure, as shown in Figure 3.10. The key challenge is how to measure the association for one target concept $\mathcal{X}$ with the attributes $\mathcal{A}$ and $\mathcal{B}$, respectively. Our strategy is first to compose neutral text prompts about $\mathcal{X}$ that do not mention either $\mathcal{A}$ or $\mathcal{B}$. The idea is that the images generated with these neutral prompts should not be affected by the attributes but will be skewed towards them due to the possible implicit stereotyping biases in the generative model. We then include the

attributes in the prompts and generate attribute-guided images. The distance between the neutral and attribute-guided images can be used to measure the association between the concepts and the attributes.

More specifically, we construct text prompts that are based on the target concepts, with or without the attributes. Let X and Y denote the neutral prompts related to the target concepts $\mathcal{X}$ and $\mathcal{Y}$, respectively. Similarly, we use X^A to represent the set of text prompts that are created by editing X with a set of attribute modifiers A corresponding to the attribute $\mathcal{A}$. We feed these text prompts into the text-to-image generative model and use $\mathcal{G}(\cdot)$ to denote the set of generated images with input prompts. For ease of notation, we use lowercase letters to represent the image samples and those accented with right arrows to represent the vector representations of the images. We consider the following test statistics:

- **Differential association** measures the difference of the association between the target concepts with the attributes.

$$S(X, Y, A, B) = \mathbb{E}_{x \in \mathcal{G}(X)} \text{Asc}(x, X^A, X^B) - \mathbb{E}_{y \in \mathcal{G}(Y)} \text{Asc}(y, Y^A, Y^B) \tag{3.22}$$

Here $\text{Asc}(x, X^A, X^B)$ is the association for one sample image with the attributes, i.e.,

$$\text{Asc}(x, X^A, X^B) = \mathbb{E}_{a \in \mathcal{G}(X^A)} \cos(\boldsymbol{a}x, \boldsymbol{a}a) - \mathbb{E}_{b \in \mathcal{G}(X^B)} \cos(\boldsymbol{a}x, \boldsymbol{a}b) \tag{3.23}$$

In Eq (3.23), $\cos(\cdot, \cdot)$ is the distance measure between images. While there are several different methods for measuring the distance between images, we choose to

compute the cosine similarity between image embedding vectors that are generated with pre-trained vision encoders. During our experimental evaluation, we follow the fashion and use the vision encoder of CLIP model [RKH+21] for convenience.

- **p-value** is a measure of the likelihood that a random permutation of the target concepts would produce a greater difference than the sample means. To perform the permutation test, we randomly split the set $X \cup Y$ into two partitions $\widetilde{X}$ and $\widetilde{Y}$ of equal size. Note that the prompts in $\widetilde{X}$ might be related to concept $\mathcal{Y}$ and those in $\widetilde{Y}$ might be related to concept $\mathcal{X}$. The p-value of such a permutation test is given by

$$p = \Pr\Big(|S(\widetilde{X}, \widetilde{Y}, A, B)| > |S(X, Y, A, B)|\Big) \tag{3.24}$$

  The p-value represents the degree to which the differential association is statistically significant. In practice, we simulate 1000 runs of the random permutation to compute the p-value for the sake of efficiency.

- **Effect size** d is a normalized measure of how separated the distributions of the associations between two target concepts are. We adopt the Cohen's d to compute the effect size by

$$d = \frac{\mathbb{E}_x[\text{Asc}(x, X^A, X^B)] - \mathbb{E}_y[\text{Asc}(y, Y^A, Y^B)]}{s} \tag{3.25}$$

  where s is the pooled standard deviation for the samples of $\text{Asc}(x, X^A, X^B)$ and $\text{Asc}(y, Y^A, Y^B)$. According to Cohen, effect size is classified as small ($d = 0.2$), medium ($d = 0.5$), and large ($d \geq 0.8$).

We present the whole bias test procedure in Algorithm 3. The defined bias measures the degree to which the generations of the target concepts exhibit a preference towards one attribute over another. One qualitative example is provided in the first column of Figure 3.11. Although the prompt of those figures does not specify gender, almost all of the generated images for science and career are depicting boys.

Algorithm 3 Bias test procedure

Input: concepts X and Y, attributes A and B.

Output: $S(X, Y, A, B)$, p, d.

1: Construct a set of neutral prompts related to the concepts X and Y. Then construct attribute guided prompts for attributes A and B, respectively.

2: For $Z \in \{X, Y\}$, generate the sets of images $\mathcal{G}(Z)$, $\mathcal{G}(Z^A)$ and $\mathcal{G}(Z^B)$ from the text prompts.

3: Compute $S(X, Y, A, B)$ using Eq (3.22).

4: Run the permutation test to compute the p-value by Eq (3.24).

5: Compute the effect size d by Eq (3.25).

3.3.3 Association Test Results

3.3.3.1 Experimental Setup

Concepts and Text Prompts We replicate 8 bias tests for text-to-image generative models, including 6 valence tests: Flowers *vs*. Insects, Musical Instruments *vs*. Weapons, Judaism *vs*. Christianity, European American *vs*. African American, light skin *vs*. dark

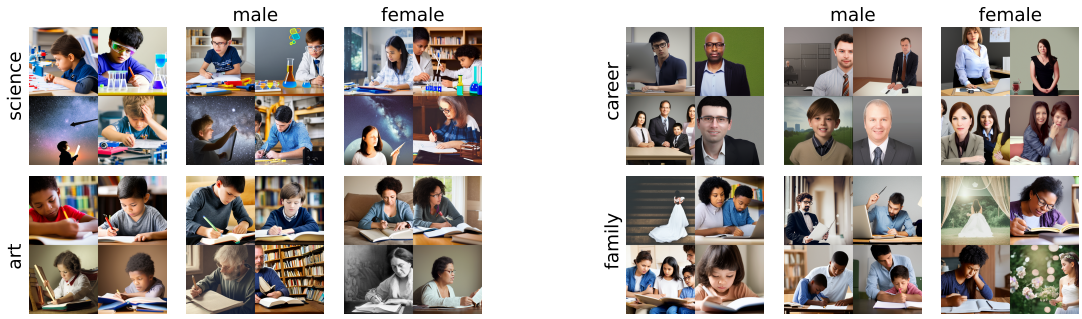
skin, and straight *vs*. gay; and 2 stereotype tests: science *vs*. arts and career *vs*. family. Each bias test includes two target concepts and two valence or stereotypical attributes. Following [GMS98], we adopt the same set of verbal stimuli for each of the concepts and attributes. For valence tests, the evaluation attributes are pleasant and unpleasant. For stereotype tests, the stereotyping attributes are male and female.

We systematically compose a set of representative text prompts with the collection of verbal stimuli for each pair of compared target concepts and attributes. The constructed text prompts will be fed into the diffusion model to generate images.

Generative Models For our initial evaluation, we use the Stable Diffusion model `stable-diffusion-2-1` [RBL+22]. We adopt the standard parameters as provided in the Huggingface's API to generate 10 images of size 512×512 for each text prompt, yielding hundreds of images for each concept. Through practical testing, we determined that this number of generations produces accurate estimates of the evaluated metrics with a high level of confidence. The number of denoising steps is set to 50 and the guidance scale is set to 7.5. The model uses `OpenCLIP-ViT/H` [RKH+21] to encode text descriptions.

3.3.3.2 Valence Tests

Flowers and Insects We begin by exploring the non-offensive stereotypes about flowers and insects, as these do not involve any demographic groups. The original IAT



(a) Science *vs*. Arts

(b) Careers *vs*. Family

Figure 3.11: Examples of generated images. Images in the first row are generated with the text prompts describing science or career, while images in the second row are generated with the text prompts describing arts or family. The first column of images are generated with neutral prompts, without adding any gender-specific words. The second and third columns of images are generated with gender-specific prompts by appending gendered words to the corresponding neutral prompts.

finding found that most people take less responding time to associate flowers with words that have pleasant meanings and insects with words that have unpleasant meanings [GMS98]. To replicate this test, we use the same set of verbal stimuli for flowers and insects categories that were used in the IAT test. We construct the text prompt "a photo of {`flower/insect`}" to generate images without any valence interventions. In parallel, we append the words expressing pleasant or unpleasant attitudes after the constructed prompt to generate the images with positive or negative valence. Examples of generated images can be seen in Figure 3.11. We report the evaluated differential association $S(X, Y, A, B)$, p-value, and effect size d in Table 3.4. To estimate the p-value, we perform the permutation test for 1,000 runs and find out that there is no other permutation of images that can yield a higher association score, indicating that

Concept X	Concept Y	Attribute A	Attribute B	Association Score	p-value	effect size d
Flower	Insect	Pleasant	Unpleasant	0.033	$< 1e^{-3}$	1.492
Musical Instrument	Weapon	Pleasant	Unpleasant	0.015	0.118	0.528
European American	African American	Pleasant	Unpleasant	0.011	0.270	0.323
Light skin	Dark skin	Pleasant	Unpleasant	-0.025	0.019	-1.237
Straight	Gay	Pleasant	Unpleasant	0.033	0.003	1.113
Judaism	Christianity	Pleasant	Unpleasant	-0.003	0.442	-0.099
Science	Arts	Male	Female	0.019	0.200	0.193
Careers	Family	Male	Female	0.026	$< 1e^{-3}$	0.639

Table 3.4: **Evaluated association scores, p-values, and effect size for 8 bias tests.** The larger absolute values of association score and effect size indicate a large bias. Smaller p-value indicates the test result is more significant.

the p-value is less than $1e^{-3}$. We note that an effect size of 0.8 generally indicates a strong association between concepts, and the effect size of 1.492 found in this test suggests that flowers are significantly more strongly associated with a positive valence, while insects are more strongly associated with a negative valence. Our observation demonstrates that human-like biases are universal in image generation models even when the concepts used are not associated with any social concerns.

Musical Instruments and Weapons To further understand the presence of implicit biases associated with text-prompt-generated images between non-offensive stereotypes, we perform the test on another set of non-offensive stereotypes of musical instruments and weapons by using the verbal stimuli for the original IAT test. Similar to our

test on flowers and insects, we first generated images only on the object itself, with the text prompt "a picture of {`musical instrument/weapon`}", then we modified the text prompts to include pleasant and unpleasant attitudes, and, finally, generated images with positive or negative valence. We report the evaluated differential association $S(X, Y, A, B)$, p-value, and effect size d in Table 3.4. The differential association score of 0.015 indicates that there is little difference in the association between our target concepts of musical instruments and weapons and the attributes of pleasant and unpleasant. We retrieved an effect size of 0.528, which implies that musical instruments have a much stronger association with a positive valence, and instead, weapons show a stronger association with a negative valence.

Judaism and Christianity We also perform the valence test on the concepts concerning religion, particularly Judaism and Christianity. Consistent with the tests on the previously mentioned concepts, we have two sets of text prompts constructed with the verbal stimuli that are used in the IAT test for Judaism and Christianity and for Pleasant and Unpleasant. The first set comes without valence intervention, only using the provided verbal stimuli for Judaism and Christianity. The second set of text prompts incorporates terms linked to pleasant and unpleasant attitudes. We derived images based on the different sets of prompts constructed. The valence test for this set of concepts yields a very small effect size, -0.099, suggesting that humans hold a rather neutral attitude towards Judaism and Christianity, only with a slight pleasant-

ness towards Christianity and a little unpleasantness towards Judaism. The differential association score of -0.003 demonstrates a tiny difference in the association between the two religions of Judaism and Christianity and the two social attitudes of pleasantness and unpleasantness. Our finding overturns the religion stereotype previously documented in IAT tests.

European American and African American In this valence test, we seek to explore the implicit racial stereotypes, besides non-harmful stereotypes, of European Americans and African Americans. From the original IAT paper, two sets of common European American and African American names are provided, and the result from our test shows that it is much easier to associate European American names with words that suggest a pleasant attitude and African American names with words that imply an unpleasant attitude. In our test, we continue to use the verbal stimuli for European American and African American names retrieved from [Tzi18] to construct our text prompts. For the text-prompt-generated images that are not valence-related, we use the text prompt "a portrait of {`European American name/African American name`}". Meanwhile, we create valence-related text prompt by including terms that embody pleasant and unpleasant attitudes. We recognize that there is an inconspicuous association between European American and pleasant terms and that between African American and unpleasant terms from the value of effect size of 0.323. The differential score of 0.011 shows a subtle association between the concepts of European American

and African American and the attributes of pleasant and unpleasant.

Light Skin and Dark Skin This valence test reveals the hostile biases towards humans with light skin and dark skin in the same racial group. We use the verbal stimuli collected by Project Implicit, a project initiated by [NSH+07], that aims to educate people on biases. Following the pattern of our purposed test, we create a set of text prompts without valence for both light skin and dark skin and another set of text prompts that consider the valence attributes of pleasant and unpleasant. We calculate the differential association $S(X, Y, A, B)$, p-value, and effect size d of the images generated based on the text prompts we constructed. We obtain a considerably large effect size of -1.237, indicating that light skin is much more closely associated with an unpleasant attribute, and dark skin, on the other hand, has a strong association with a pleasant attribute. In addition, we have a moderate p-value, 0.019, which way exceeds the statistically significant value of 0.05.

Straight and Gay We examine the implicit bias towards sexuality in this valence test that targets the concepts of straight and gay. Text prompts that do not contain the factor of valence are created, along with those composed with pleasant and unpleasant attitudes using the method as other valence tests. By running through text-to-image generative models, corresponding images are produced. We receive the effect size of 1.113, which is much bigger than the defined large effect size value of 0.8. It suggests

that the association between the concept of straight and the attribute of pleasant is significantly strong and that of gay and the attribute of unpleasant is tremendously strong as well. We also note that the p-value is 0.003, which is lower than 0.005.

The valence tests show that not only non-harmful human biases, but also hostile stereotypical biases such as inter-racial, intra-racial, and sexual biases exist in the text-to-image generative models.

3.3.3.3 Stereotype Tests

We conduct two gender-related stereotypical tests: gender-science and gender-career tests.

Science and Art We use the text prompt "a person studying {`science`/`art`}" for image generations. To generate images associated with male and female attributes, we modify the "person" with gender-specific words, such as "woman", "girl", "man", "boy", *etc*. The evaluated effect size of 0.193 is small, and demonstrates that the distribution of the association scores does not differ too much. In addition, the p-value of 0.200 is relatively large. This bias test demonstrates that the evaluated generative model does not contain bias towards science and art as is documented in human biases.

Career and Family The original IAT test has found that females are more associated with family and males with career [NBG02]. To replicate this test with image generations, we use the template of text prompts "a person focusing on {`career` / `family`}" to

generate images. We find that the effect size of 0.639 is relatively large and the p-value is less than $< 1e^{-3}$, indicating career is significantly more strongly associated with male than female.

3.3.3.4 Gender Stereotype in Occupations

Prior work has demonstrated that text prompts pertaining to occupations may lead the model to reconstruct social disparities regarding gender and racial groups, even though they make no mention of such demographic attributes [BKD+22]. We are also interested in how the generated images are skewed towards women and men, assessed by their association scores with gender.

We collect the list of common occupation titles from the U.S. Bureau of Labor Statistics[6]. For each occupation title, we construct the gender-neutral text prompt "A photo of a {`occupation`}", and gender-specific versions by amending gendered descriptions. For each occupation, we use Stable Diffusion to generate 100 gender-neutral images, 100 masculine images, and 100 feminine images, respectively. We use Eq. (3.23) to calculate the association score between occupation and gender attributes.

We plot the distribution of association scores, and the quartiles, for eight different occupations in Figure 3.12. The figure shows that the 0.75 quantiles of association scores for computer programmers and pharmacists are higher than the others by a large margin, indicating that these occupations are more strongly associated with

[6] `https://www.bls.gov/oes/current/oes_stru.htm`

men. Conversely, the mean association scores for elementary school teachers, librarians, announcers, and chemists are negative, indicating that these occupations are more strongly associated with women. The association score for chef and police is neutral, suggesting that there is insufficient evidence to establish a stereotype.

Concept	Attribute	Score
Flowers	Pleasant *vs*. Unpleasant	1.00
Insects	Pleasant *vs*. Unpleasant	0.15
Musical Instrument	Pleasant *vs*. Unpleasant	0.90
Weapon	Pleasant *vs*. Unpleasant	0.05
Science	Male *vs*. Female	0.75
Arts	Male *vs*. Female	0.30
Careers	Male *vs*. Female	0.75
Family	Male *vs*. Female	0.40

Table 3.5: **Human evaluation results.** For each pair of concept and attributes, we report the fraction of images that are chosen as being more closely associated with pleasant or male attributes. We find out that the machine-rated association scores can properly represent human's perceptions.

Stereotype Amplification Do images generated by the diffusion model amplify the implicit stereotypes in the textual representations used to guide image generation? Specifically, we examine occupational images and calculate the association scores be-

tween the text prompts by substituting the text embeddings of CLIP into Eq. (3.23) and Eq. (3.22). We then compare these associations for text prompts to the associations for the generated images to investigate whether the biases are amplified.

Figure 3.13 demonstrates the stereotype amplification between text prompts and generated images. For each occupation, we use an arrow to represent the change of associations on the axis of gender. We observe that the associations are amplified on a large scale for most occupations. In particular, the textual association between a computer programmer and gender is only -0.0039 but enlarged to 0.0186 for images. Similar amplifications are observed for elementary school teachers, librarians, and chemists. For the occupation of chef, the association of text prompts is skewed towards females, while the association of images is skewed towards males.

Comparison to Human Evaluation We recruit university students to evaluate the generated images and compare how the perceptions of human differ with the machine-evaluated association scores. Specifically, for each set of concepts, we ask three student participants to view 20 images generated with neutral prompts and choose which valence or stereotypical attribute is more closely associated. We report the fraction of images that are chosen as being more closely associated with pleasant or male attributes. As shown in Table 3.5, the human's preference of association aligns with the strength of our association scores. For flowers *vs.* insects and musical instruments *vs.* weapon, humans mostly prefer to associate flowers and musical instruments with pleasant while

insects and weapons with unpleasant. For science *vs*. arts and career *vs*. family, we find that the significance of the bias is reduced. The Kendall's τ coefficient between the machine-evaluated and human-rated scores is 0.55, indicating that the association scores can properly represent human's perceptions.

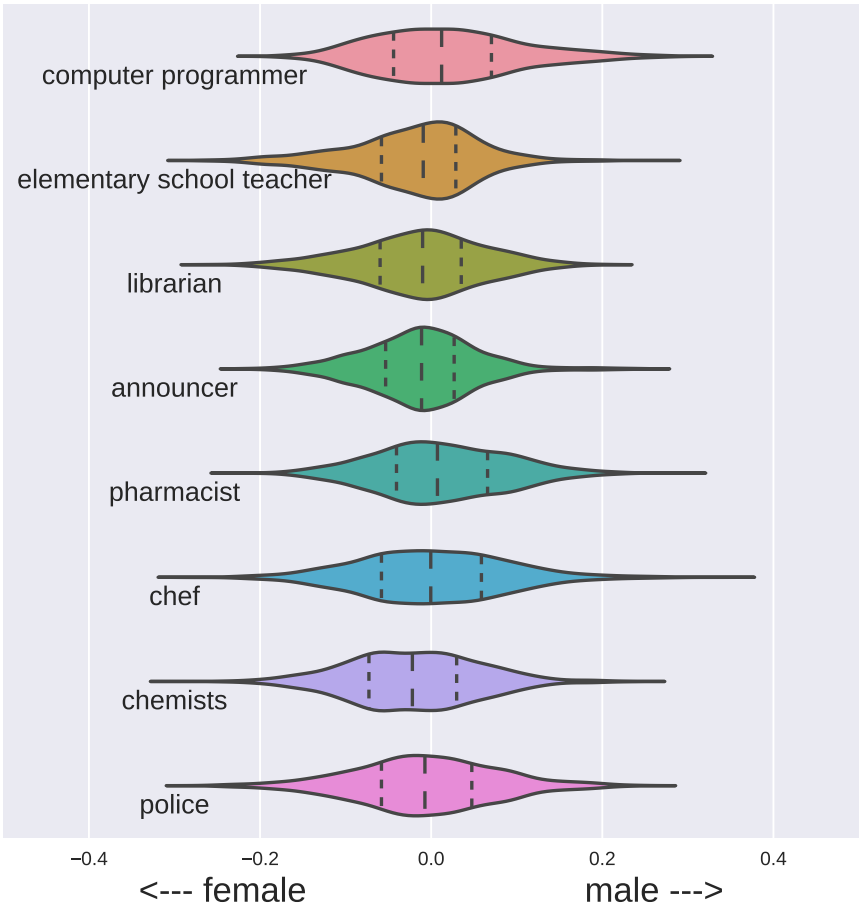


Figure 3.12: **Gender stereotype in occupation.** For each occupation, we compare the association score with gender and plot their distribution. The x-axis represents the extent to which the generated images are associated with male or female. Our analysis suggests that computer programmers and pharmacists are more strongly associated with man, while elementary school teachers, librarians and announcers are more strongly associated with woman.

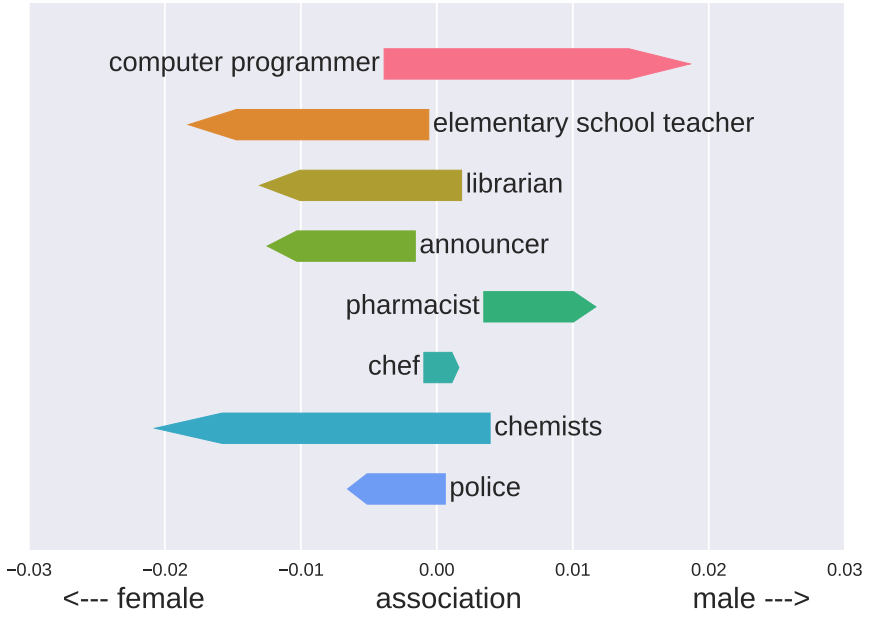


Figure 3.13: **Stereotype amplification.** For each occupation, we compare the association scores for generated images to the association scores for the text prompts. The association scores for the text prompts are represented by the tails of the arrows, and the association scores for the images are represented by the heads of the arrows.

Chapter 4

Fairness Influence Function

4.1 Motivation

Despite the successes of algorithmic treatments proposed in the fairness community, the question of *why* a particular "fair" training process leads to a more fair model remains less addressed. The explanation for the above *why* question is essential in improving user trustworthiness in the models and often regulated by legal requirements [Nin17]. There has been a recent surge of interest in explaining algorithmic fairness. Much of the work chose to quantify the importance of the input feature variables used to make fair decisions [LL17, SN20, MOS21]. This line of research makes explanations on the population level, as the importance measures are quantified statistically over the entire subset of instances.

Nevertheless, the impact of fairness constraints on individual instances is rarely

discussed. Our central inquiry is how each individual training instance influences the model decisions when a fairness constraint is imposed. Demystifying and characterizing the influence of individual instances subject to fairness constraints is important and opens up the possibility of auditing a machine learning model at the instance level. Among other potentials, we believe that such understanding might help with developing preprocessing solutions to mitigate bias by emphasizing more on instances that have a high influence on fairness.

To this end, we borrow the idea from recent literature on *influence function* [STY17], which has largely focused on approximating the effect of training examples in prediction accuracy rather than fairness constraints. Concretely, an influence function characterizes the change of model predictions compared to the counterfactual that one training example is removed. We instantiate the change, due to the penalty of disparity, on prominent fairness criteria that have been widely applied in the community. We illustrate that the influence scores can be potentially applied to mitigate the unfairness by pruning less influential examples on a synthetic setting. We implement this idea on different domains including tabular data, images and natural language.

4.2 Influence Function

We will consider the problem of predicting a target binary label y based on its corresponding feature vector x under fairness constraints with respect to sensitive

Table 4.1: Examples of fairness measures.

Fairness Criteria	Measure $\psi(f)$
Demographic Parity	$\sum_{g\in\mathcal{Z}} \lvert \Pr(f(x)=+1 \mid z=g) - \Pr(f(x)=+1)\rvert$
Equal TPR	$\sum_{a\in\mathcal{Z}} \lvert \Pr(f(x)=+1 \mid z=a, y=+1) - \Pr(f(x)=+1 \mid y=+1)\rvert$
Equal FPR	$\sum_{a\in\mathcal{Z}} \lvert \Pr(f(x)=+1 \mid z=a, y=-1) - \Pr(f(x)=+1 \mid y=-1)\rvert$
Equal Odds	$\sum_{a\in\mathcal{Z}} \sum_{b\in\mathcal{Y}} \lvert \Pr(f(x)=+1 \mid z=a, y=b) - \Pr(f(x)=+1 \mid y=b)\rvert$

attributes z. We assume that the data points (x, y, z) are drawn from an unknown underlying distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. $\mathcal{X} \in \mathbb{R}^d$ is d-dimensional instance space, $\mathcal{Y} \in \{-1, +1\}$ is the label space, and $\mathcal{Z} \in \{0, 1, \ldots, m-1\}$ is the (sensitive) attribute space. Here we assume that sensitive attribute is a categorical variable regarding m sensitive groups. The goal of fair classification is to find a classifier $f : \mathcal{X} \to \mathbb{R}$ with the property that it minimizes expected true loss $\mathrm{err}(f)$ while mitigating a certain measure of fairness violation $\psi(f)$. We assume that the model f is parameterized by a vector $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_p]$ of size p. Thereby $\mathrm{err}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x; \boldsymbol{\theta}), y)]$, where the expectation is respect to the true underlying distribution $\mathcal{D}$ and $\ell(\cdot, \cdot)$ is the loss function. We show exemplary fairness metrics $\psi(\cdot)$ in Table 4.1, including demographic parity [Cho17a, JHF+22], equality of opportunity [HPS16c], among many others. Without loss of generality, $f(x)$ induces the prediction rule $2 \cdot \mathbb{1}[f(x) \geq 0] - 1$, where $\mathbb{1}[\cdot]$ is the indicator function. Denote by $\mathcal{F}$ the family of classifiers, we can express the objective

of the learning problem as

$$\min_{f \in \mathcal{F}} \ \mathrm{err}(f), \quad \text{s.t. } \psi(f) \leq \mu, \tag{4.1}$$

where μ is a tolerance parameter for fairness violations. Let $D = \{(x_i, y_i, z_i)\}_{i=1}^n$ denote n data examples sampled from true data distribution $\mathcal{D}$. In this case, the empirical loss is $\widehat{\mathrm{err}}(f) = \frac{1}{n} \sum_{(x_i,y_i,z_i) \in D} \ell(f(x_i), y_i)$. Due to the fact that $\psi(f)$ is non-convex and non-differentiable in general, practically we will use a surrogate $\phi(f)$ to approximate it. Let $\hat{\phi}(\cdot)$ denote the empirical version of $\phi(\cdot)$, then the *Empirical Risk Minimization* (ERM) problem is defined as

$$\min_{f \in \mathcal{F}} \ \widehat{\mathrm{err}}(f), \quad \text{s.t. } \hat{\phi}(f) \leq \mu. \tag{4.2}$$

This work aims to discuss the influence of a certain training example (x_i, y_i, z_i) on a target example (x_j, y_j, z_j), when fairness constraints are imposed to the classifier f. Let f^D represent the model f trained over the whole dataset D and $f^{D/\{i\}}$ represent the counterfactual model f trained over the dataset D by excluding the training example (x_i, y_i, z_i). The influence function with respect to the output of classifier f is defined as

$$\mathrm{infl}_f(D, i, j) \coloneqq f^{D/\{i\}}(x_j) - f^D(x_j) \tag{4.3}$$

Note that j may be either a training point $j \in D$ or a test point outside D.

4.2.1 Influence Function in Unconstrained Learning

We start by considering the unconstrained classification setting when parity constraints are not imposed in the learning objective. Recall that the standard Empirical

Risk Minimization (ERM) problem is

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \boldsymbol{\theta}), y_i) \tag{4.4}$$

where $\boldsymbol{\theta}$ is the parameters of model f. We assume that $\boldsymbol{\theta}$ evolves through the following gradient flow along time t:

$$\frac{\partial \boldsymbol{\theta}}{\partial t} = -\frac{1}{n} \nabla \ell(f(x_i; \boldsymbol{\theta}), y_i) \tag{4.5}$$

Let $\boldsymbol{\theta}_0$ denote the final parameter of classifier f trained on the whole set D. To track the influence of an observed instance i, we hypothesis the update of parameter $\boldsymbol{\theta}$ with respect to instance i is recovered by one counterfactual step of gradient descent with a weight of $-\frac{1}{n}$ and a learning rate of η. This process can also be regarded as inverting the gradient flow of Equation 4.5 with a small time step $\Delta t = \eta$. Next, to compute the output of model f on the target example j, we may Taylor expand f around $\boldsymbol{\theta}_0$

$$\begin{aligned}
f(x_j; \boldsymbol{\theta}) - f(x_j; \boldsymbol{\theta}_0) \approx & \frac{\partial f(x_j; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) && \text{(by Taylor series expansion)} \\
= & \frac{\partial f(x_j; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \left(-\eta \left. \frac{\partial \boldsymbol{\theta}}{\partial t} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right) && \text{(by inverting gradient flow)} \\
= & \frac{\eta}{n} \frac{\partial f(x_j; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \nabla \ell(f(x_i; \boldsymbol{\theta}_0), y_i) && \text{(by substituting Equation 4.5)} \\
= & \frac{\eta}{n} \frac{\partial f(x_j; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial \ell(f(x_i; \boldsymbol{\theta}_0), y_i)}{\partial f} \frac{\partial f(x_i; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} && \text{(by chain rule)} \\
= & \frac{\eta}{n} \frac{\partial f(x_j; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial f(x_i; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \left. \frac{\partial \ell(w, y_i)}{\partial w} \right|_{w=f(x_i; \boldsymbol{\theta}_0)} && (4.6)
\end{aligned}$$

In the language of kernel methods, the product of $\partial f(x_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\partial f(x_j;\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is named Neural Tangent Kernel (NTK) [JGH18]

$$\Theta(x_i, x_j;\boldsymbol{\theta}) = \frac{\partial f(x_j;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial f(x_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum_p \frac{\partial f(x_j;\boldsymbol{\theta})}{\partial\theta_p}\frac{\partial f(x_i;\boldsymbol{\theta})}{\partial\theta_p} \tag{4.7}$$

NTK describes the evolution of deep neural networks during the learning dynamics. Substituting the NTK in Equation 4.7 into Equation 4.6 and combining Equation 4.3, we obtain the following close-form statement:

Lemma 4.1. *In unconstrained learning, the influence function of training example i subject to the prediction of f on the target example j is*

$$\mathit{infl}_f(D, i, j) \approx \frac{\eta}{n}\Theta(x_i, x_j;\boldsymbol{\theta}_0)\left.\frac{\partial\ell(w, y_i)}{\partial w}\right|_{w=f(x_i;\boldsymbol{\theta}_0)} \tag{4.8}$$

Equation 4.8 mimics the first-order approximation in [PLKS20] with a focus on tracking the change on model output instead of the change on loss.

4.2.2 Influence Function in Constrained Learning

In classification problems, the outcome of an algorithm may be skewed towards certain protected groups, such as gender and ethnicity. While the definitions of fairness are controversial, researchers commonly impose the parity constraints like demographic parity [Cho17a] and equal opportunity [HPS16c] for fairness-aware learning. A large number of approaches have been well studied to mitigate the disparity, which in general can be categorized into pre-processing, in-processing, and post-processing algorithms.

Pre-processing algorithms [KC12, FFM+15b, CWV+17] usually reweigh the training instances, resulting in the influence scores will also be scaled by a instance-dependent weight factor. Post-Processing algorithms [HPS16c] will not alter the learning objective, thus the influence function of training examples stays unchanged.

In this work, we primarily focus on the influence function in the in-processing treatment frameworks [CJW+19, ZVGRG17a, WGOS17, ABD+18b, Nar18a, SKG+19, KAS11]. In such fashion, the fair classification problem are generally formulated as a constrained optimization problem as Equation 4.1. The common solution is to impose the penalty of fairness violations $\psi(f)$ as a regularization term. The constrained risk minimization problem thus becomes

$$\min_{f \in \mathcal{F}} \ \mathrm{err}(f) + \lambda \psi(f) \tag{4.9}$$

where in above λ is a regularizer that controls the trade-off between fairness and accuracy. Note λ is not necessary static, e.g., in some game-theoretic approaches [ABD+18b, Nar18a, CJW+19, CJS19], the value of λ will be dynamically chosen. We notice that while the empirical $\psi(f)$ is often involving the rates related to indicator function, it might be infeasible to solve the constrained ERM problem. For instance, demographic parity, as mentioned in Table 4.1, requires that different protected groups have an equal acceptance rate. The acceptance rate for group $a \in \mathcal{Z}$ is given by

$$\Pr(f(x) \geq 0 \mid z = a) = \frac{\sum_i \mathbb{1}[f(x_i) \geq 0, z_i = a]}{\sum_i \mathbb{1}[z_i = a]} \tag{4.10}$$

Since non-differentiable indicator function cannot be directly optimized by gradient-

based algorithms, researchers often substitute the direct fairness measure $\psi(f)$ by a differentiable surrogate $\phi(f)$. In consequence, the constrained ERM problem is

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; \boldsymbol{\theta}), y_i) + \lambda \hat{\phi}(f) \tag{4.11}$$

We make the following decomposability assumption:

Assumption 1 (Decomposability)**.** The empirical surrogate of fairness measure $\hat{\phi}(f)$ can be decomposed into

$$\hat{\phi}(f) = \frac{1}{n} \sum_{i=1}^{n} \hat{\phi}(f, i), \tag{4.12}$$

where in above each $\hat{\phi}(f, i)$ is only related to the instance i and independent of other instances $j \neq i$.

Assumption 1 guarantees that the influence of one training example i will not be entangled with the influence of another training example j. Following an analogous derivation to Equation 4.6, we obtain the kernelized influence function

Lemma 4.2. *When the empirical fairness measure $\hat{\phi}(\cdot)$ satisfies the decomposability assumption, the influence function of training example i with respect to the prediction of f on the target j can be expressed as*

$$\mathit{infl}_f(D, i, j) \approx \underbrace{\frac{\eta}{n} \Theta(x_i, x_j; \boldsymbol{\theta}_0) \left. \frac{\partial \ell(w, y_i)}{\partial w} \right|_{w=f(x_i;\boldsymbol{\theta}_0)}}_{\textit{influence of loss}} + \underbrace{\lambda \frac{\eta}{n} \Theta(x_i, x_j; \boldsymbol{\theta}_0) \left. \frac{\partial \hat{\phi}(f, i)}{\partial f} \right|_{f(x_i;\boldsymbol{\theta}_0)}}_{\textit{influence of fairness constraint}}$$

Lemma 4.2 presents that the general expression of influence function can be decoupled by the influence subject to accuracy (the first term) and that subject to parity constraint (the second term).

However, the situation will be more complicated when we are concerned about the influence function subject to loss. In this constrained setting,

$$\begin{aligned}\text{infl}_\ell(D,i,j) &= \ell(f(x_j;\boldsymbol{\theta},y_j) - \ell(f(x_j;\boldsymbol{\theta}_0),y_j) \\ &\approx \left.\frac{\partial \ell(w,y_j)}{\partial w}\right|_{w=f(x_j;\boldsymbol{\theta}_0)} (f(x_j;\boldsymbol{\theta}) - f(x_j;\boldsymbol{\theta}_0)) \\ &\approx \frac{\eta}{n} \left.\frac{\partial \ell(w,y_j)}{\partial w}\right|_{w=f(x_j;\boldsymbol{\theta}_0)} \Theta(x_i,x_j;\boldsymbol{\theta}_0) \left(\left.\frac{\partial \ell(w,y_i)}{\partial w}\right|_{w=f(x_i;\boldsymbol{\theta}_0)} + \lambda \left.\frac{\partial \hat{\phi}(f,i)}{\partial f}\right|_{f(x_i;\boldsymbol{\theta}_0)} \right)\end{aligned} \tag{4.13}$$

Equation 4.13 implies the intrinsic tension between accuracy and fairness — when the signs of $\partial \ell(f(x_i),y_i)/\partial f$ and $\partial \hat{\phi}(f,i) \Big/ \partial f$ are opposite, the influence of parity constraint will contradict with that of loss.

4.3 Influence Function through Smooth Approximation

In this section, we will take a closer look at the specific influence functions for several commonly used surrogate constraints. Since the influence induced by loss is independent of the expressions for fairness constraint, we will ignore the first term in Equation 4.2 and focus on the second term throughout this section. We define the pairwise influence score subject to fairness constraint as

$$S(i,j) \coloneqq \lambda \frac{\eta}{n} \Theta(x_i,x_j;\boldsymbol{\theta}_0) \left.\frac{\partial \hat{\phi}(f,i)}{\partial f}\right|_{f(x_i;\boldsymbol{\theta}_0)} \tag{4.14}$$

In what follows, we will instantiate $S(i,j)$ on three regularized fairness constraints.

4.3.1 Relaxed Constraint

Throughout this part, we assume that the sensitive attribute is binary, *i.e.*, $\mathcal{Z} \in \{0, 1\}$. The technique of relaxing fairness constraints was introduced in [MCPZ18]. We will analyse the influence of relaxed constraints, including demographic parity and equality of opportunity as below.

Demographic Parity. [MCPZ18] propose to replace the demographic parity metric

$$\psi(f) = |\Pr(f(x;\boldsymbol{\theta}) \geq 0 \mid z = 1) - \Pr(f(x;\boldsymbol{\theta}) \geq 0 \mid z = 0)| \tag{4.15}$$

by a relaxed measure

$$\phi(f) = |\mathbb{E}[f(x;\boldsymbol{\theta}) \cdot \mathbb{1}[z = 1]] - \mathbb{E}[f(x;\boldsymbol{\theta}) \cdot \mathbb{1}[z = 0]]| \tag{4.16}$$

Without loss of generality, we assume that the group $z = 1$ is more favorable than the group $z = 0$ such that $\mathbb{E}[f(x;\boldsymbol{\theta})\mathbb{1}[z = 1]] \geq \mathbb{E}[f(x;\boldsymbol{\theta})\mathbb{1}[z = 0]]$ during the last step of optimization. We construct a group-dependent factor $\alpha_z \coloneqq \mathbb{1}[z = 1] - \mathbb{1}[z = 0]$ by assigning $\alpha_0 = -1$ and $\alpha_1 = +1$. Then we can eliminate the absolute value notation in the $\hat{\phi}(f)$ as follows:

$$\begin{aligned}\hat{\phi}(f) &= \frac{1}{n}\sum_{i=1}^{n} f(x_i;\boldsymbol{\theta})(\mathbb{1}[z_i = 1] - \mathbb{1}[z_i = 0]) \\ &= \frac{1}{n}\sum_{i=1}^{n} \alpha_{z_i} f(x_i;\boldsymbol{\theta})\end{aligned} \tag{4.17}$$

Equation 4.17 is saying the relaxed demographic parity constraint satisfies the decomposability assumption with

$$\hat{\phi}(f, i) = \alpha_{z_i} f(x_i; \boldsymbol{\theta}) \tag{4.18}$$

Applying Lemma 4.2, we obtain the influence of demographic parity constraint for a training example i on the target example j

$$S_{\mathsf{DP}}(i, j) = \lambda \frac{\eta}{n} \alpha_{z_i} \Theta(x_i, x_j; \boldsymbol{\theta}_0) \tag{4.19}$$

The above derivation presumes that the quantity inside the absolute value notation in Equation 4.16 is non-negative. In the opposite scenario where group $z = 0$ is more favorable, we only need to reverse the sign of α_z to apply Equation 4.19. We note that in some cases, the sign of the quantity will flip after one-step optimization, violating this assumption.

Equality of Opportunity. For ease of notation, we define the utilities of True Positive Rate (TPR) and False Positive Rate (FPR) for each group $z \in \mathcal{Z}$ as

$$\mathrm{TPR}_z \coloneqq \Pr(f(x) \geq 0 \mid z = z, y = 1) \tag{4.20}$$

$$\mathrm{FPR}_z \coloneqq \Pr(f(x) \geq 0 \mid z = z, y = 0) \tag{4.21}$$

For the equal TPR measure $\psi(f) = |\,\mathrm{TPR}_1 - \mathrm{TPR}_0\,|$, we may relax it by

$$\phi(f) = |\mathbb{E}[f(x; \boldsymbol{\theta}) \cdot \mathbb{1}[z = 1, y = 1]] - \mathbb{E}[f(x; \boldsymbol{\theta}) \cdot \mathbb{1}[z = 0, y = 1]]| \tag{4.22}$$

Without loss of generality, we assume that the group $z = 1$ has a higher utility such that the quantity within the absolute value notation is positive. We may construct the group-dependent factor

$$\alpha_{z,y} \coloneqq \mathbb{1}[z = 1, y = 1] - \mathbb{1}[z = 0, y = 1]$$

by assigning $\alpha_{0,1} = -1$, $\alpha_{1,1} = +1$, and $\alpha_{z,-1} = 0$ for $z \in \{0, 1\}$. Then we may decompose $\hat{\phi}(f)$ into

$$\hat{\phi}(f) = \frac{1}{n} \sum_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \mathbb{1}[y_i = 1](\mathbb{1}[z_i = 1] - \mathbb{1}[z_i = 0]) \tag{4.23}$$

The above equation satisfies the decomposability assumption with

$$\hat{\phi}(f, i) = \alpha_{z_i, y_i} f(x_i; \boldsymbol{\theta}).$$

Applying Lemma 4.2 again, we obtain the influence of equal TPR constraint

$$S_{\mathsf{TPR}}(i, j) = \lambda \frac{\eta}{n} \alpha_{z_i, y_i} \Theta(x_i, x_j; \boldsymbol{\theta}_0) \tag{4.24}$$

For the equal FPR measure $\psi(f) = |\operatorname{FPR}_1 - \operatorname{FPR}_0|$, we may relax it by

$$\phi(f) = |\mathbb{E}[f(x; \boldsymbol{\theta}) \cdot \mathbb{1}[z = 1, y = -1]] - \mathbb{E}[f(x; \boldsymbol{\theta}) \cdot \mathbb{1}[z = 0, y = -1]]| \tag{4.25}$$

Likewise, we still assume the group $z = 1$ has a higher utility. We construct the factor

$$\tilde{\alpha}_{z,y} \coloneqq \mathbb{1}[z = 1, y = -1] - \mathbb{1}[z = 0, y = -1]$$

by assigning $\tilde{\alpha}_{0,-1} = -1$, $\tilde{\alpha}_{1,-1} = +1$, and $\tilde{\alpha}_{z,+1} = 0$ for $z \in \{0, 1\}$. Following the similar deduction, we may verify the relaxed equal FPR measure satisfies the decomposability

assumption. Then we can obtain the influence of equal FPR constraint as

$$S_{\mathsf{FPR}}(i,j) = \lambda \frac{\eta}{n} \tilde{\alpha}_{z_i,y_i} \Theta(x_i, x_j; \boldsymbol{\theta}_0) \tag{4.26}$$

In the opposite scenario when group $z = 0$ has a higher utility of either TPR or FPR, we may reverse the sign of $\alpha_{z,y}$ or $\tilde{\alpha}_{z,y}$, respectively. Finally, imposing equal odds constraint is identical to imposing equal TPR and equal FPR simultaneously, implying the following equality holds:

$$S_{\mathsf{EO}} = S_{\mathsf{TPR}} + S_{\mathsf{FPR}} \tag{4.27}$$

Corollary 4.3. *When one group has higher utilities (TPR and FPR) than the other group, the influence of imposing equal odds $S_{\mathsf{EO}}(i,j)$ is equivalent to that of imposing demographic parity $S_{\mathsf{DP}}(i,j)$.*

4.3.2 Covariance as Constraint

Another common approach is to reduce the covariance between the group membership z and the encoded feature $f(x;\boldsymbol{\theta})$ [ZVGRG17a, WGOS17]. Formally, the covariance is defined by

$$\mathtt{Cov}(z, f(x)) = \mathbb{E}[z \cdot f(x;\boldsymbol{\theta})] - \mathbb{E}[z] \cdot \mathbb{E}[f(x;\boldsymbol{\theta})] \tag{4.28}$$

Then the empirical fairness measure is the absolute value of covariance

$$\hat{\phi}(f) = \left| \frac{1}{n} \sum_{i=1}^{n} z_i f(x_i;\boldsymbol{\theta}) - (\frac{1}{n} \sum_{i=1}^{n} z_i) \cdot (\frac{1}{n} \sum_{i=1}^{n} f(x_i;\boldsymbol{\theta})) \right| \tag{4.29}$$

Since we can observe the whole training set, the mean value of group membership can be calculated by $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$. As a result, we can decompose the covariance as follows:

$$\begin{aligned}\hat{\phi}(f) &= |\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})f(x_i;\boldsymbol{\theta})| \\ &= \frac{1}{n}\sum_{i=1}^{n}\beta_i(z_i - \bar{z})f(x_i;\boldsymbol{\theta}) \end{aligned} \tag{4.30}$$

where $\beta_i \in \{-1, +1\}$ is an instance-dependent parameter. Then the covariance constraint satisfies the decomposability assumption by taking

$$\hat{\phi}(f, i) = \beta_i(z_i - \bar{z})f(x_i;\boldsymbol{\theta}), \quad \beta_i \in \{-1, +1\}. \tag{4.31}$$

Finally, the influence score induced by the covariance constraint in Equation 4.29 is

$$S_{\mathsf{cov}}(i, j) = \lambda\frac{\eta}{n}\beta_i(z_i - \bar{z})\Theta(x_i, x_j;\boldsymbol{\theta}_0) \tag{4.32}$$

In this kernelized expression, the pairwise influence score is neatly represented as NTK scaled by an instance weight $\beta_i(z_i - \bar{z})$.

Connection to Relaxed Constraint. We may connect the influence function of the covariance approach to that of the relaxation approach in a popular situation where there are only two sensitive groups.

Corollary 4.4. *When sensitive attribute z is binary, the influence score of covariance is half of the influence of relaxed demographic parity.*

$$|\mathcal{Z}| = 2 \implies S_{\mathsf{cov}}(i, j) = \frac{1}{2}S_{\mathsf{DP}}(i, j)$$

4.3.3 Information Theoretic Algorithms

The demographic parity constraint can be interpreted as the independence of prediction $f(x)$ and group membership z. Denoted by $I(f(x); z)$ the mutual information between $f(x)$ and z, the independence condition $f(x) \perp\!\!\!\perp z$ implies $I(f(x); z) = 0$. In consequence, a number of algorithms [SKG+19, GFDS21, BNBR20] propose to adopt the bounds of mutual information $I(f(x); z)$ as the empirical fairness measure. We consider approximating mutual information by MINE [BBR+18a, vdOLV18] as an example.

$$\hat{\phi}(f) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp g(f(x_i), z_i)}{\frac{1}{n}\sum_{k=1}^{n} \exp g(f(x_i), z_k)} \tag{4.33}$$

where the function $g(\cdot, \cdot)$ is parameterized by a neural network. In this case, $\hat{\phi}(f)$ satisfies the decomposability assumption by straightly taking

$$\hat{\phi}(f, i) = \log \frac{\exp g(f(x_i), z_i)}{\frac{1}{n}\sum_{k=1}^{n} \exp g(f(x_i), z_k)} \tag{4.34}$$

Although the denominator inside the logarithm in Equation 4.34 contains the sum over all the z_k in the training set, we can always calculated the sum when we know the prior distribution of the categorical variable z. Taking the derivative of $\hat{\phi}(f, i)$, the influence of MINE constraint is given by

$$\begin{aligned} S_{\mathsf{MINE}}(i, j) &= \lambda \frac{\eta}{n} \Theta(x_i, x_j; \boldsymbol{\theta}_0) \cdot \left.\frac{\partial \mathbf{G}}{\partial w}\right|_{w=f(x_i;\boldsymbol{\theta})} \\ \text{where} \quad \mathbf{G} &= g(w, z_i) - \frac{1}{n}\sum_{k=1}^{n} g(w, z_k) \end{aligned} \tag{4.35}$$

Connection to Covariance. In a special case when $g(f(x; \boldsymbol{\theta}), z) = z f(x; \boldsymbol{\theta})$, we have $\partial_f g(f(x; \boldsymbol{\theta}), z) = z$. Substitute the partial derivative back into Equation 4.35, the influ-

ence of MINE reduces to the influence of covariance measure $\lambda \frac{\eta}{n} \alpha_i (z_i - \bar{z}) \Theta(x_i, x_j; \boldsymbol{\theta}_0)$. However, it is very likely that the influence scores of MINE and covariance are much different, due to the fact that the unknown function $g(f(x), z)$ is parameterized by neural networks in more generic applications.

4.4 Influence Function through Zeroth-Order Approximation

Our previous discussion explores how smooth approximation can address the non-differentiability of fairness constraints. As an alternative approach to circumvent direct gradient computation, we propose using a zeroth-order gradient estimator [FKM05, Sha17, GJZ18]. For ease of notations, we use $\boldsymbol{z} = (\boldsymbol{x}, y, z)$ to represent the data examples. Following [NS17], the gradient estimate of $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z})$ can be obtained by

$$\mathbf{g}(\boldsymbol{\theta}, \boldsymbol{z}, \mu) = d \frac{\mathcal{L}(\boldsymbol{\theta} + \mu \boldsymbol{u}, \boldsymbol{z}) - \mathcal{L}(\boldsymbol{\theta} - \mu \boldsymbol{u}, \boldsymbol{z})}{2\mu} \boldsymbol{u}, \quad \boldsymbol{u} \sim \varphi \tag{4.36}$$

In above equation, d denotes the dimension of the model parameters $\boldsymbol{\theta}$ used for high-dimensional unbiased estimation, $\mu > 0$ is a smoothing radius (hyper-parameter), and φ is spherically symmetric with $\mathbb{E}_{\boldsymbol{u} \sim \varphi}[\|\boldsymbol{u}\|] = 1$, where $\boldsymbol{u}$ is a random unit vector $\boldsymbol{u} \in \mathbb{R}^D$. Thus, $\boldsymbol{u} \sim \varphi$ means some random direction sampled from the unit space φ. For example, if $\boldsymbol{\theta}$ is scalar ($d = 1$), then $\boldsymbol{u} \in \{-1, +1\}$ with equal probability. Typically, we employ an *unbiased* estimate of the gradient over a random pertubation $\boldsymbol{u} \sim \varphi$ with smoothing

radius μ

$$\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}) = \mathbb{E}_{\boldsymbol{u}\sim\varphi}[\mathbf{g}(\boldsymbol{\theta}, \boldsymbol{z}, \mu)] \tag{4.37}$$

For succinctness, let $\phi(\boldsymbol{\theta}, \boldsymbol{z})$ denote the fairness constraint of a single training example $\boldsymbol{z}_n$. Denote the total loss for any example z as $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}) = \ell(\boldsymbol{\theta}, \boldsymbol{z}) + \lambda\phi(\boldsymbol{\theta}, \boldsymbol{z})$. When considering the second-order terms in the gradient flow, the kernel function incorporates an inverse Hessian matrix term. Combined with a two-point random gradient estimator, the influence function of training example $\boldsymbol{z}$ subject to the prediction of f on the target example $\boldsymbol{z}_{\text{test}}$ can be formulated as

$$\widehat{\text{infl}}(\mathcal{D}, \boldsymbol{z}, \boldsymbol{z}_{\text{test}}) \approx -\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}_{\text{test}})^{\top}\widehat{\mathbf{H}}^{-1}\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}) \tag{4.38}$$

4.4.1 Approximating Inverse Hessian Matrix

Computing the inverse Hessian matrix requires $\mathcal{O}(d^3)$ operations for a model with d parameters, suggesting a significant computational burden. To address this challenge, we employ the WoodFisher approximation [SA20], which provides an efficient alternative by approximating the Hessian structure using the Sherman-Morrison formula.

Fisher Information Matrix. In a probalistic view, the Fisher Information Matrix (FIM), denoted as $\mathbf{F}$, serves as a way of measuring the amount of information about a negative log-likelihood for an underlying joint distribution $p(\boldsymbol{x}, y|\boldsymbol{\theta})$ parameterized by

model parameters $\boldsymbol{\theta}$. The Fisher Information Matrix is defined as:

$$\mathbf{F}(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{p(\boldsymbol{x},y|\boldsymbol{\theta})}\left[\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{x},y|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{x},y|\boldsymbol{\theta})^{\top}\right] \tag{4.39}$$

Typically, under the assumption of maximum likelihood estimation (MLE) or when the loss function is a negative log-likelihood function, it can be easily proved that the observed Fisher information matrix is equivalent to the negative Hessian matrix for the model prediction, *i.e.*,

$$\mathbb{E}\left[\mathbf{H}(\boldsymbol{\theta})\right] = -\mathbf{F}(\boldsymbol{\theta}) \tag{4.40}$$

In practice, it's a common approach to estimate the Fisher matrix by employing its empirical counterpart. The formulation of this empirical Fisher matrix can be presented as

$$\begin{aligned}\hat{\mathbf{F}} &= \frac{1}{N}\sum_{n=1}^{N} \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{z}_n)\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{z}_n)^{\top} \\ &\approx \frac{1}{N}\sum_{n=1}^{N} \widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\theta}_n)\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\theta}_n)^{\top}\end{aligned} \tag{4.41}$$

where the first equality holds due to the assumption that the loss function is a negative log-likelihood function. When there is no confusion, we use the Fisher matrix $\mathbf{F}$ and the Hessian matrix $\mathbf{H}$ interchangeably.

WoodFisher Approximation. We define

$$\hat{\boldsymbol{g}}_n = \widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{z}_n) \tag{4.42}$$

Leveraging the structure of the empirical Fisher matrix, we can derive the following recursive equation to estimate the inverse of the Hessian matrix following [SA20]:

$$\begin{aligned}\hat{\mathbf{H}}_{n+1} &= \hat{\mathbf{H}}_n + \frac{1}{N}\hat{\boldsymbol{g}}_{n+1}\hat{\boldsymbol{g}}_{n+1}^\top, \\ \widehat{\mathbf{H}}_{n+1}^{-1} &= \widehat{\mathbf{H}}_n^{-1} - \frac{\widehat{\mathbf{H}}_n^{-1}\hat{\boldsymbol{g}}_{n+1}\hat{\boldsymbol{g}}_{n+1}^\top\widehat{\mathbf{H}}_n^{-1}}{N + \hat{\boldsymbol{g}}_{n+1}^\top\widehat{\mathbf{H}}_n^{-1}\hat{\boldsymbol{g}}_{n+1}},\end{aligned} \tag{4.43}$$

where $\widehat{\mathbf{H}}_0^{-1} = \kappa^{-1}\boldsymbol{I}_d$ and κ denotes the dampening term. Eq (4.43) mainly follows the Sherman-Morrison formula, which provides an efficient way to update the inverse of a matrix when a rank-one modification is applied, avoiding the need to recompute the entire inverse matrix.

4.4.2 Proposed Algorithm

Now we are ready to present Algorithm 4 for approximating the influence function using zeroth-order gradient estimation, where explicit gradient computations are infeasible. The core of Algorithm 4 revolves around iteratively updating the inverse Hessian through the Sherman-Morisson formula. A crucial aspect of this method is its reliance solely on first-order gradients of the loss function $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \cdot)$. To estimate the gradients without direct access to model parameters, we employ a two-point random gradient estimator in the ZOGradEstimator procedure to estimate gradients. This subroutine executes multiple iterations, each involving random perturbations to the model parameters. By utilizing finite differences, it approximates the gradient direction, thereby enabling the estimation of influence without requiring direct gradient

calculations.

4.5 Estimating the Aggregated Influence Score

In this section, we intend to discuss the expected influence of a training example on the whole data distribution. We will focus on the changes of empirical fairness constraints $\hat{\phi}(f)$ when a data point (x_i, y_i, z_i) is excluded in the training set or not. Suppose that $\hat{\phi}(f)$ satisfies the decomposability assumption. We define the realized influence score of a training example i aggregated over the whole data distribution $\mathcal{D}$ as

$$\mathcal{S}(i) := \int_{(x_j,y_j,z_j)\in\mathcal{D}} \frac{\partial\hat{\phi}(f,j)}{\partial f} S(i,j)\, d\Pr(x_j, y_j, z_j) \tag{4.44}$$

$\mathcal{S}(i)$ takes into account the change on $\hat{\phi}(f)$ by applying the first-order approximation again for each test point j. In practice, the model f can only observe finite data examples in D that are drawn from the underlying distribution $\mathcal{D}$. We estimate the influence score of a training example i over the training set D.

$$S(i) := \frac{1}{n}\sum_{i=1}^{n} \frac{\partial\hat{\phi}(f,j)}{\partial f} S(i,j) \tag{4.45}$$

We wonder how the measure of $S(i)$ deviates from $\mathcal{S}(i)$.

Theorem 4.5 (Generalization Bound)**.** *With probability at least* $1-\epsilon$*,*

$$\mathcal{S}(i) - S(i) \leq \mathcal{O}\left(\sqrt{\frac{\log\frac{1}{\epsilon}}{2n}}\right) \tag{4.46}$$

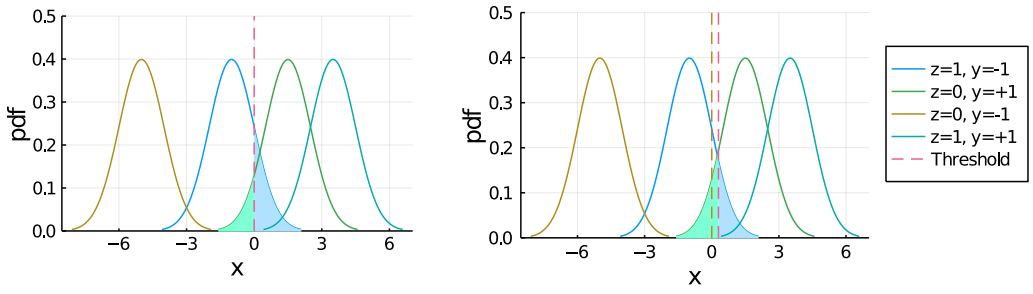


Figure 4.1: A toy example to interpret the influence scores of fairness. Left: The optimal classifier is $\mathbb{1}[x \geq 0]$. Four curves in different colors represent the distributions for each (z, y) combination. The blue area and green area represent the violation of demographic parity. The data examples with low influence scores of fairness constraints are around $x = 0$. Right: When we down-weight the examples around $x = 0$, the optimal classifier will be perturbed towards right. Since the sum of blue area and green area decreases, the violation of demographic parity is mitigated.

Interpreting Influence Scores on Synthetic Data. We consider a synthetic example as visualized in Figure 4.1 to illustrate why our influence scores help with identifying instances that affect the fairness. We assume that the individual examples are independently drawn from an underlying normal distribution corresponding to the label $y \in \{-1, +1\}$ and group membership $z \in \{0, 1\}$, *i.e.*, $x_{z,y} \sim N(\mu_{z,y}, \sigma)$. We assume that $\mu_{0,-1} < \mu_{1,-1} < 0 < \mu_{0,+1} < \mu_{1,+1}$. Suppose that we train a linear model $f(x) = w \cdot x + b$, and the obtained classifier is $\mathbb{1}[f(x) >= 0]$ which reduces to $\mathbb{1}[x >= 0]$ for our toy example. Then we have the following proposition:

Proposition 4.6. *In our considered setting, if we down-weight the training examples with smaller absolute fairness influence scores, the model will tend to mitigate the violation of demographic parity.*

Proposition 4.6 informs us that we can mitigate the unfairness by up-weighting the data instances with higher influence scores, or equivalently by removing some low-influence training points.

4.6 Empirical Evaluations

In this section, we examine the influence score subject to parity constraints on three different application domains: tabular data, images and natural language.

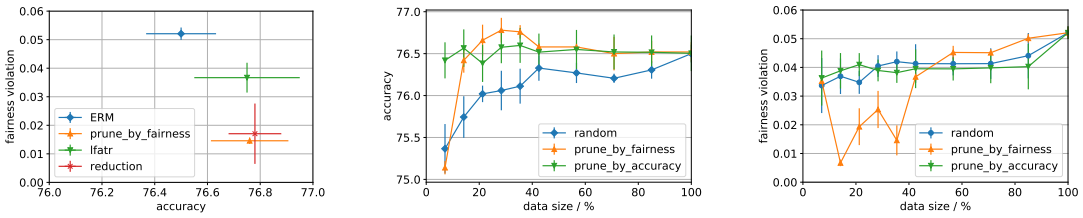


Figure 4.2: Data pruning results on Adult dataset.

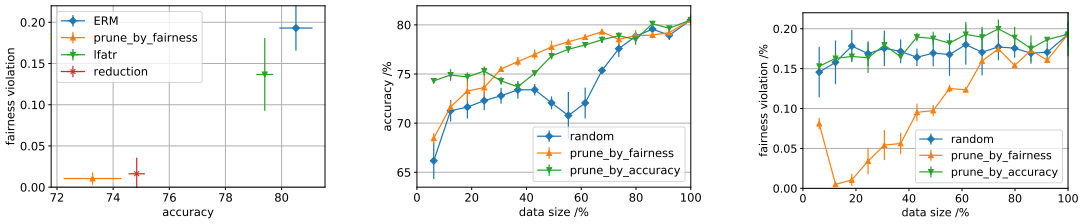


Figure 4.3: Data pruning results on CelebA dataset.

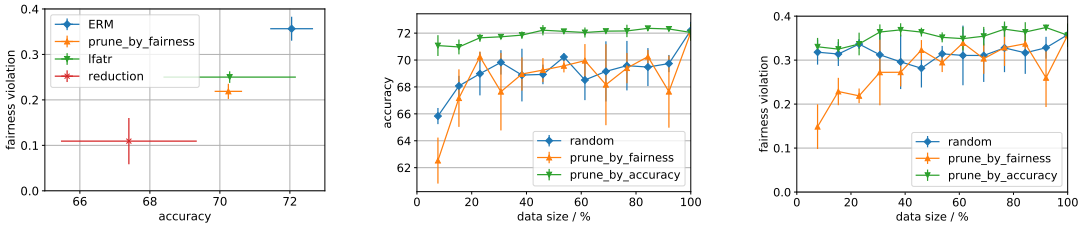


Figure 4.4: Data pruning results on Jigsaw dataset.

4.6.1 Setup

We adopt the evaluation protocol that has been widely used by the previous papers on interpreting the impact of data examples. The basic idea is to train a model on a subset of training data by removing the less influential examples [PGD21]. In particular, we assess the performance of three data prune strategies: (1) random, which randomly selects a subset of training examples; (2) prune by fairness, which removes the data examples in the ascending order of the absolute values of aggregated influence scores subject to fairness as described in Equation 4.45; (3) prune by accuracy, which removes the data examples in the descending order of absolute influence scores in terms of the loss. The influence score is equivalent to the first order approximate proposed in [PLKS20]. For the three strategies above, a model pre-trained on the whole training set will be used to estimate the influence scores of training examples with the direct application of Equation 4.45. We then execute the data prune procedure and impose the relaxed demographic parity in Equation 4.16 to train a fair model.

We compare the performance of pruning by influence scores with following optimization algorithms that regularize the constraint:

- ERM, which trains the model directly without imposing fairness constraints.
- lfatr [MCPZ18], which regularizes the model with relaxed constraint as given in Equation 4.17.
- reduction [ABD$^+$18b], which reduces the constrained optimization to a cost-sensitive learning problem.

We used the Adam optimizer with a learning rate of 0.001 to train all the models. We used $\gamma = 1$ for models requiring the regularizer parameter of fairness constraints. Any other hyperparameters keep the same among the compared methods. We report two metrics: the accuracy evaluated on test set and the difference of acceptance rates between groups as fairness violation.

4.6.2 Result on Tabular Data

Firstly, we work with multi-layer perceptron (MLP) trained on the Adult dataset [DG17b]. We select sex, including female and male, as the sensitive attribute. We resample the dataset to balance the class and group membership. The MLP model is a two-layer ReLU network with hidden size 64. We train the model 5 times with different random seeds to report the mean and standard deviation of accuracy and fairness metrics. In each trial, the dataset is randomly split into a training and a test set in a

ratio of 80 to 20. We compare the performance of prune by fairness in Figure 4.2 and find it has a similar fairness-accuracy trade-off with the reduction approach. To gain further insights, we further plot how the size of pruned training examples affects the accuracy and fairness metrics for three prune strategies. Not surprisingly, the random baseline remains a high fairness violation with a large accuracy drop when the data size decreases. In contrast, prune by fairness has a similar accuracy with pruning by accuracy when the data size is greater than 20% and mitigates the fairness violation by a large margin when the data size is less than 40%. We also notice that prune by fairness anomalously has a high fairness violation when the data size is less than 10%. We conjecture such a small size of training data does not contain sufficient information, leading to the significant performance degradation. These observations suggest that we may obtain the best trade-off with a subset of only 20%–40% of training data.

4.6.3 Result on Images

Next, we train a ResNet-18 network [HZRS15] on the CelebA face attribute dataset [LLWT15a]. We select smiling as binary classification target and gender as the sensitive attribute. The left figure in Figure 4.3 shows the trade-off between accuracy and fairness violation for each baseline method. We explore how the size of pruned training examples affects the accuracy and fairness metrics in right two figures in Figure 4.3. Again, the accuracy of prune by fairness has a similar trend with that of prune by accuracy when data size is larger than 20%, but drops strikingly with much smaller data

size. On the other hand, prune by fairness mitigates the fairness violation straightly when data size decreases.

4.6.4 Result on Natural Language

Lastly, we consider Jigsaw Comment Toxicity Classification [Jig18] with text data. We select race as the sensitive attribute in our evaluation. We use pre-trained BERT [DCLT19b] to encode each raw comment text into a 768-dimensional textual representation vector and train a two-layer neural network to perform classification. We report the experimental result in Figure 4.4. The left figure shows that prune by fairness has a mimic performance with lfatr while preserving smaller standard deviation. The middle figure shows that prune by accuracy keeps a relatively high accuracy when a large subset of training examples are removed. In comparison, both prune by fairness and random prune failed to make informative prediction when the data size is below 20%. The right figure implies that prune by fairness is capable of mitigating bias. This result cautions that we need to carefully account for the price of a fair classifier, particularly in this application domain.

Algorithm 4 Approximating Influence Function with Zeroth-Order Gradient Estimation

Input: data samples $\mathcal{D} = \{\boldsymbol{z}_n\}_{n=1}^N$, target training example $\boldsymbol{z}$, test example $\boldsymbol{z}_{\text{test}}$, loss function $\mathcal{L}$, model parameter $\boldsymbol{\theta}$ at size d, perturbation scale μ, dampening term κ.

Initialize $\widehat{\mathbf{H}}^{-1} \leftarrow \kappa^{-1}\boldsymbol{I}_d$.

for all $\boldsymbol{z}_n \in \mathcal{D}$ **do**

  $\boldsymbol{g} \leftarrow \text{ZOGradEstimator}(\boldsymbol{\theta}, \boldsymbol{z}_n)$

  $\widehat{\mathbf{H}}^{-1} \leftarrow \widehat{\mathbf{H}}^{-1} - \widehat{\mathbf{H}}^{-1}\boldsymbol{g}\boldsymbol{g}^\top\widehat{\mathbf{H}}^{-1}/(N + \boldsymbol{g}^\top\widehat{\mathbf{H}}^{-1}\boldsymbol{g})$

end for

$\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}_{\text{test}}) \leftarrow \text{ZOGradEstimator}(\boldsymbol{\theta}, \boldsymbol{z}_{\text{test}})$

$\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}) \leftarrow \text{ZOGradEstimator}(\boldsymbol{\theta}, \boldsymbol{z})$

$\text{infl}(\mathcal{D}, \boldsymbol{z}, \boldsymbol{z}_{\text{test}}) \leftarrow -\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}_{\text{test}})^\top\widehat{\mathbf{H}}_N^{-1}\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z})$

return $\text{infl}(\mathcal{D}, \boldsymbol{z}, \boldsymbol{z}_{\text{test}})$

procedure ZOGradEstimator($\boldsymbol{\theta}, \boldsymbol{z}$)

  for $t = \{1, \ldots, T\}$ **do**

    Randomly sample $\boldsymbol{u}_t \sim \mathcal{N}(0, 1)$ at size d.

    $\mathbf{g}_t \leftarrow \boldsymbol{u}_t \cdot [\mathcal{L}(\boldsymbol{z}, \boldsymbol{\theta} + \mu\boldsymbol{u}_t) - \mathcal{L}(\boldsymbol{z}, \boldsymbol{\theta} - \mu\boldsymbol{u}_t)]/2\mu$

  end for

  $\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}) \leftarrow \frac{1}{T}\sum_{t=1}^T \mathbf{g}_t$

  return $\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z})$

end procedure

Chapter 5

Conclusion and Future Works

5.1 Conclusion

In this dissertation, we explore three crucial aspects of promoting fairness in foundation models. First, we examine scenarios where certain subgroups have less accurate feature and label information compared to others. Our research reveals that applying standard bias mitigation techniques without considering these data quality disparities can inadvertently increase outcome inequality. We propose novel methods to address unfairness while accounting for data uncertainty. Second, we extend current state-of-the-art fairness techniques, typically focused on classification problems, to a broader range of multi-modal AI applications. We propose new fairness metrics to measure and quantify the biases when images are connected to text. Third, we highlight the critical importance of detecting and addressing potentially harmful instances within

training datasets to improve overall model fairness.

From a high-level perspective, the second part of our work establishes methodologies for measuring fairness in foundation models across both image and text modalities. The first part presents approaches to mitigate unfairness and bias, with particular emphasis on handling uncertainty. The last part of our work on fairness influence function provides a robust framework for auditing the quality of dataset. Notably, the proposed zeroth-order approximation technique offers a computationally efficient approach for large-scale foundation models by eliminating the need for direct gradient computation.

This dissertation provides valuable insights for machine learning practitioners, researchers, and policymakers considering the implementation of foundation models. The proposed technical solutions and approaches are particularly relevant when dealing with training data that exhibits significant disparity. By addressing these challenges, the research provides practical tools and methodologies to enhance the equitable performance of AI systems, even when confronted with biased or problematic observations. Ultimately, it serves as a valuable resource to harness the power of AI while upholding principles of fairness and equality; it contributes to the ongoing effort to create more just and inclusive systems for responsible AI; it aligns technological advancements with ethical considerations and societal values of humans.

5.2 Future Works

We separate the potential future works on the relevant chapters in this thesis:

From Chapter 2 Our current work primarily focuses on static environments, without considering the dynamic interplay between human subjects and policy makers. In real-world applications, machine learning policies and populations continuously adapt to each other, resulting in shifts in the underlying data distribution and evolving decision landscapes. This mutual adaptation presents significant challenges for maintaining fairness in the long run. Future research includes exploring how to achieve long-term fairness without compromising model utility, particularly in these dynamic environments where both decision policies and populations evolve. This research direction is highly related with the literature of performative prediction, dynamic fairness or long-term fairness.

From Chapter 3 Our current multimodal fairness measures are task-specific. In contrast, future research should aim to develop a unified framework capable of evaluating fairness across diverse tasks. Moreover, the current fairness measures heavily relied on human-crafted tasks or text prompts. We envision developing automatic benchmarks that can comprehensively evaluate the fairness for flexible group definitions. One possible approach is to apply generative AI models to enhance the magnitude and dimension of fairness benchmarks.

From Chapter 4 While our current influence function framework mainly focuses on classification models, the rise of generative AI necessitates extending these analytical techniques to generative models. Two particularly promising directions are understanding the impacts of training data on diffusion models and large language models. The key theoretical extension involves shifting the counterfactual analysis from the conditional probability $\Pr(Y \mid X)$ to the joint probability distribution $\Pr(X, Y)$. Through analogous mathematical analysis, we can derive explicit expressions for influence functions in generative models. Another interesting topic is, the current influence function focuses on the impact of individual training instances. For large language models, the inputs are sequence of text tokens. Then a natural research question arises: how do we measure the influence of these sequence-level tokens on the large language models' outputs? Addressing this research question will benefit understanding how changing the text prompts may affect the model's outputs. We emphasize that influence functions serve as powerful tools for multiple critical applications: protecting from data poisoning attacks, enhancing data privacy and security, analyzing the memorization effects of specific training examples, and *etc*.

Bibliography

[ABD+18a] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In Dy and Krause [DK18], pages 60–69.

[ABD+18b] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In Dy and Krause [DK18], pages 60–69.

[AFSV16] Ifeoma Ajunwa, Sorelle A. Friedler, Carlos Eduardo Scheidegger, and Suresh Venkatasubramanian. Hiring by algorithm: Predicting and preventing disparate impact. In *the Yale Law School Information Society Project conference Unlocking the Black Box: The Promise and Limits of Algorithmic Accountability in the Professions*, April 2016.

[AG07] Galen Andrew and Jianfeng Gao. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40, 2007.

[AL88a] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[AL88b] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, Apr 1988.

[Ale12] Michelle Alexander. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. New Press, New York, 2012.

[ALMK16a] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias, 2016.

[ALMK16b] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

[Ame83] American Psychological Association. *Publications Manual*. American Psychological Association, Washington, DC, 1983.

[AU72] Alfred V. Aho and Jeffrey D. Ullman. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ, 1972.

[AZ05] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, December 2005.

[BAHAZ19] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and

Richard Zemel. Understanding the origins of bias in word embeddings. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR, 09–15 Jun 2019.

[Ban22] Rajas Bansal. A survey on bias and fairness in natural language processing. *arXiv preprint arXiv:2204.09591*, 2022.

[BBR+18a] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR, 10–15 Jul 2018.

[BBR+18b] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 530–539. PMLR, 2018.

[BCD+19] Alex Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Zhe Zhao, L. Hong, Ed Huai hsin Chi, and Cristos Goodrow. Fairness in recommendation ranking through pairwise comparisons. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[BCZ+16] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc.

[BDC20] Rishi Bommasani, Kelly Davis, and Claire Cardie. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online, July 2020. Association for Computational Linguistics.

[BDNP19] Maria-Florina F Balcan, Travis Dick, Ritesh Noothigattu, and Ariel D Procaccia. Envy-free classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[BF21] Emily Black and Matt Fredrikson. Leave-one-out unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 285–295, New York, NY, USA, 2021. Association for Computing Machinery.

[BG18a] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency (FAT)*, pages 77–91. ACM, 2018.

[BG18b] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.

[BGW18] Asia J. Biega, K. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018.

[BHB19] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Gar-

nett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 15535–15545. Curran Associates, Inc., 2019.

[BHDR18] Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018.

[BHN19] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[Bin20] Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 514–524, New York, NY, USA, 2020. Association for Computing Machinery.

[BJM06] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[BKD+22] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv preprint arXiv:2211.03759*, 2022.

[BKW+20] Andrea Burns, Donghyun Kim, D. Wijaya, Kate Saenko, and Bryan A. Plummer. Learning to scale multilingual representations for vision-language tasks. *ArXiv*, abs/2004.04312, 2020.

[BM04] Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013, 2004.

[BNBR20] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *International Conference on Learning Representations*, 2020.

[BPF21] Samyadeep Basu, Phil Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*, 2021.

[BPK21] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *CoRR*, abs/2110.01963, 2021.

[BR18] Miranda Bogen and Aaron Rieke. Help wanted: an examination of hiring algorithms, equity, and bias. Technical report, Upturn, 2018.

[Bra16] Brain Tumour Charity. Finding myself in your arms: The reality of brain tumour treatment and care, 2016.

[BS19] Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy?, 2019.

[BYF20] Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 715–724. PMLR, 2020.

[BYMC22] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? *ArXiv*, abs/2210.15230, 2022.

[CA18] Damian Clifford and Jef Ausloos. Data protection and the role of fairness. *Yearbook of European Law*, 37:130–187, 2018.

[CBN17a] Aylin Caliskan, J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183 – 186, 2017.

[CBN17b] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics

derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017.

[CD21] Monojit Choudhury and Amit Deshpande. How linguistically fair are multilingual pre-trained language models? In *AAAI-21*. AAAI, AAAI, February 2021.

[CDL20] T. Chen, Jiajun Deng, and Jiebo Luo. Adaptive offline quintuplet loss for image-text matching. *ArXiv*, abs/2003.03669, 2020.

[CDPF+17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz 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, 2017.

[CE20] Nicholas Carrara and Jesse Ernst. On the estimation of mutual information. *arXiv: Data Analysis, Statistics and Probability*, 33:31, 2020.

[CE21] Fredrik Carlsson and Ariel Ekgren. Pre-trained multilingual-clip encoders. `https://github.com/FreddeFrallan/Multilingual-CLIP`, 2021.

[CFL+15] Xinlei Chen, H. Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325, 2015.

[CHKV19] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 319–328, New York, NY, USA, 2019. Association for Computing Machinery.

[Cho17a] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5 2:153–163, 2017.

[Cho17b] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5 2:153–163, 2017.

[CJS18] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3539–3550. Curran Associates, Inc., 2018.

[CJS19] Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. In *ALT*, 2019.

[CJW+19] Andrew Cotter, Heinrich Jiang, Serena Wang, Taman Narayan, M. Gupta, S. You, and K. Sridharan. Optimization with non-

differentiable constraints with applications to fairness, recall, churn, and other goals. *ArXiv*, abs/1809.04198, 2019.

[CK11] Imre Csiszár and János Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2 edition, 2011.

[CKG+20] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *ACL*, 2020.

[CKNH20a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

[CKNH20b] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[CKS81] Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. Alterna-

tion. *Journal of the Association for Computing Machinery*, 28(1):114–133, 1981.

[CLS19] Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In *International Conference on Machine Learning*, pages 961–970, 2019.

[CLTB21] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR, 18–24 Jul 2021.

[CLY+20] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

[CM21] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021.

[CMJ+19] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In Kamalika Chaudhuri and

Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[CRC+20a] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020.

[CRC+20b] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020.

[CW80] R. Dennis Cook and Sanford Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22:495–508, 1980.

[CWV+17] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in*

Neural Information Processing Systems 30, pages 3992–4001. Curran Associates, Inc., 2017.

[CZB22] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *CoRR*, abs/2202.04053, 2022.

[CZB+23] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming Yang, Kevin P. Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. *ArXiv*, abs/2301.00704, 2023.

[DBK+21a] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[DBK+21b] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image

recognition at scale. In *International Conference on Learning Representations*, 2021.

[DCLT19a] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[DCLT19b] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[DF18] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018.

[DG17a] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[DG17b] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[DG17c] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[DHP+12a] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[DHP+12b] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[DHP+12c] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[DIKL18] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of

Proceedings of Machine Learning Research, pages 119–133, New York, NY, USA, 23–24 Feb 2018. PMLR.

[DJ21] Karan Desai and Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR*, 2021.

[DK18] Jennifer G. Dy and Andreas Krause, editors. *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 2018.

[DKW+23] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks?, 2023.

[DMB16] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales : Demonstrating accuracy equity and predictive parity performance of the compas risk scales in broward county. Technical report, Northpointe Inc., July 2016.

[EFSS16] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, August 2016. Association for Computational Linguistics.

[ERV+15] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein

Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. "i always assumed that i wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. In *CHI 2015 - Proceedings of the 33rd Annual CHI Conference on Human Factors in Computing Systems*, Conference on Human Factors in Computing Systems - Proceedings, pages 153–162. Association for Computing Machinery, April 2015. 33rd Annual CHI Conference on Human Factors in Computing Systems, CHI 2015 ; Conference date: 18-04-2015 Through 23-04-2015.

[FCG20] Riccardo Fogliato, Alexandra Chouldechova, and Max G'Sell. Fairness evaluation in presence of biased noisy labels. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2325–2336. PMLR, 2020.

[FCSS05] Michal Feldman, John Chuang, Ion Stoica, and Scott Shenker. Hidden-action in multi-hop routing. In *Proceedings of the 6th ACM conference on Electronic commerce*, EC '05, pages 117–126, New York, NY, USA, 2005. ACM.

[Fel20] Vitaly Feldman. Does learning require memorization? a short tale about

a long tail. *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, 2020.

[FFKF18] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 12. BMVA Press, 2018.

[FFM+15a] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Conference on Knowledge Discovery and Data Mining (KDD)*, pages 259–268. ACM, 2015.

[FFM+15b] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery.

[FKM05] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, page 385–394, USA, 2005. Society for Industrial and Applied Mathematics.

[FN96] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347, July 1996.

[FSV16] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[FV14] Benoit Frenay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

[FZ20] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2881–2891. Curran Associates, Inc., 2020.

[GBH+20] Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, Online, November 2020. Association for Computational Linguistics.

[GC21a] Ben Green and Yiling Chen. Algorithmic risk assessments can alter hu-

man decision-making processes in high-stakes government contexts. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021.

[GC21b] Wei Guo and Aylin Caliskan. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA, 2021. Association for Computing Machinery.

[GCFW18a] M. Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *ArXiv*, abs/1806.11212, 2018.

[GCFW18b] Maya R. Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *CoRR*, abs/1806.11212, 2018.

[GCL+20] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020.

[GFDS21] Umang Gupta, Aaron Ferber, Bistra N. Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *AAAI*, 2021.

[GJZ18] Xiang Gao, Bo Jiang, and Shuzhong Zhang. On the information-adaptive

variants of the admm: An iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363, July 2018.

[GKOV17] Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5986–5997. Curran Associates, Inc., 2017.

[GMS98] Anthony G Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74 6:1464–80, 1998.

[GPAM+20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[Gre20] B. Green. The false promise of risk assessments: epistemic reform and the limits of fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[GTYS18] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178(11):1544, 2018.

[Gus97] Dan Gusfield. *Algorithms on Strings, Trees and Sequences.* Cambridge University Press, Cambridge, UK, 1997.

[GZGW18] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 51–60. AAAI Press, 2018.

[HDF13] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25:1445–1459, 2013.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[HMPW16] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122. ACM, 2016.

[HP23] Hailong Hu and Jun Pang. Membership inference of diffusion models, 2023.

[HPS16a] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.

[HPS16b] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.

[HPS16c] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.

[HRS+20] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Fi-

rat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR, 13–18 Jul 2020.

[HWW17] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal LSTM. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7254–7262. IEEE Computer Society, 2017.

[HXDP20] Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France, May 2020. European Language Resources Association.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[HZRS16] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[JGH18] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[JHF+22] Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. Generalized demographic parity for group fairness. In *International Conference on Learning Representations*, 2022.

[Jig18] Jigsaw. Jigsaw unintended bias in toxicity classification, 2018.

[JN19] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. *CoRR*, abs/1901.04966, 2019.

[KAS11] Toshihiro Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650, 2011.

[KBSH21] Eugenia Kim, De'Aira Bryant, Deepak Srikanth, and Ayanna Howard. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 638–644, 2021.

[KC09] F. Kamiran and T. Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6, 2009.

[KC12] KamiranFaisal and CaldersToon. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 2012.

[KFF17] A. Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676, 2017.

[KJ21] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[KL51] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951.

[KL17] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*,

volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR, 06–11 Aug 2017.

[KL19] Fereshte Khani and Percy Liang. Noise induces loss discrepancy across groups for linear regression. *ArXiv*, abs/1911.09876, 2019.

[KLRS17a] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[KLRS17b] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[KLRS17c] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017.

[KMM15] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. *Unequal Representation and Gender Stereotypes in Image Search Results for Occupations*, page 3819–3828. Association for Computing Machinery, New York, NY, USA, 2015.

[KMZ20] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[KNRW17] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

[KNRW18] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. *arXiv preprint arXiv:1808.08166*, 2018.

[KRCP+17] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 656–666, Red Hook, NY, USA, 2017. Curran Associates Inc.

[KRR18] M. P. Kim, O. Reingold, and G. N. Rothblum. Fairness through computationally-bounded awareness. *ArXiv*, abs/1803.03239, 2018.

[KSG04] A. Kraskov, Harald Stögbauer, and P. Grassberger. Estimating mutual information. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69 6 Pt 2:066138, 2004.

[KSL18] Pang Wei Koh, Jacob Steinhardt, and Percy Liang. Stronger data poisoning attacks break data sanitization defenses. *CoRR*, abs/1811.00741, 2018.

[Lar17] Brian Larson. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain, April 2017. Association for Computational Linguistics.

[LBPL19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[LC19] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. In *NeurIPS*, 2019.

[LCH+18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong

He. Stacked cross attention for image-text matching. *arXiv preprint arXiv:1803.08024*, 2018.

[LDF+20] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and M. Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020.

[LG20] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates, 2020.

[LHPL10] Sara M. Lindberg, Janet Shibley Hyde, Jennifer L. Petersen, and Marcia C. Linn. New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, 136(6):1123–1135, 2010.

[LHS20] P. H. Le-Khac, G. Healy, and A. F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.

[LJJvdM17] Ang Li, A. Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4193–4202, 2017.

[LL17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neu-*

ral Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[LLWT15a] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[LLWT15b] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[LMB+14] Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[LMC18] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8135–8145. Curran Associates, Inc., 2018.

[LRN19] Iro Laina, C. Rupprecht, and N. Navab. Towards unsupervised image captioning with shared multimodal embeddings. *2019 IEEE/CVF In-*

ternational Conference on Computer Vision (ICCV), pages 7413–7423, 2019.

[LT16] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.

[LW21] Yang Liu and Jialu Wang. Can less be more? when increasing-to-balancing label noise rates considered beneficial. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[LYL+20] Xiujun Li, Xi Yin, C. Li, X. Hu, Pengchuan Zhang, Lei Zhang, Longguang Wang, H. Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[LZMV19] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. Noise-tolerant fair classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 294–306. Curran Associates, Inc., 2019.

[May19] Sandra Gabriel Mayson. Bias in, bias out. *Yale Law Journal*, 128(8):2218, June 2019.

[MCK19] Jérémie Mary, Clément Calauzènes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4382–4391. PMLR, 2019.

[MCPZ18] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. *CoRR*, abs/1802.06309, 2018.

[MKK21] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *J. Artif. Int. Res.*, 71:1183–1317, sep 2021.

[MOS21] Masayoshi Mase, Art B. Owen, and Benjamin B. Seiler. Cohort shapley value for algorithmic fairness. *ArXiv*, abs/2105.07168, 2021.

[MPDR21] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *arXiv preprint arXiv:2110.08527*, 2021.

[MS13] N. Manwani and P. S. Sastry. Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics*, 43(3):1146–1151, 2013.

[MSD19] Varun Manjunatha, Nirat Saini, and Larry S. Davis. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[MSHJ20] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. *Controlling Fairness and Bias in Dynamic Learning-to-Rank*, page 429–438. Association for Computing Machinery, New York, NY, USA, 2020.

[MVROW15] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pages 125–134, 2015.

[MW18] Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In Sorelle A. Friedler and Christo Wilson, editors, *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118. PMLR, 2018.

[MWB+19] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[MWZ+19a] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery.

[MWZ+19b] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery.

[MYCBT19] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing

multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[Nar18a] Harikrishna Narasimhan. Learning with complex loss functions and constraints. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1646–1654. PMLR, 09–11 Apr 2018.

[Nar18b] Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp.*, 2018.

[NBG02] Brian A. Nosek, Mahzarin R. Banaji, and Anthony G Greenwald. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6:101–115, 2002.

[NBR20] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[NDR+22] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav

Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 17–23 Jul 2022.

[NDRT13] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1196–1204. Curran Associates, Inc., 2013.

[Nin17] Ninth Circuit Jury Instructions Committee. *Manual of Model Civil Jury Instructions for the Ninth Circuit*, chapter 11. St. Paul, Minn. :West Publishing, 2017.

[NJC21] Curtis G Northcutt, Lu Jiang, and Isaac L Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 2021.

[NS17] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimiza-

tion of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

[NSH+07] Brian A. Nosek, Frederick L. Smyth, Jeffrey Jay Hansen, Thierry Devos, Nicole M. Lindner, Kate A Ranganath, Colin Tucker Smith, Kristina R. Olson, and Dolly Chugh. Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18:36 – 88, 2007.

[NSS+09] Brian A. Nosek, Frederick L. Smyth, Natarajan Sriram, Nicole M. Lindner, Thierry Devos, Alfonso Ayala, Yoav Bar-Anan, Robin Bergh, Huajian Cai, Karen Gonsalkorale, Selin Kesebir, Norbert Maliszewski, Félix Neto, Eero Olli, Jaihyun Park, Konrad Schnabel, Kimihiro Shiomura, Bogdan Tudor Tulbure, Reinout W. Wiers, Mónika Somogyi, Nazar Akrami, Bo Ekehammar, Michelangelo Vianello, Mahzarin R. Banaji, and Anthony G Greenwald. National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, 106:10593 – 10597, 2009.

[NVBB20] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*, 2020.

[OBC17] Jahna Otterbacher, Jo Bates, and Paul Clough. *Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Re-*

sults, page 6620–6631. Association for Computing Machinery, New York, NY, USA, 2017.

[Ope22] OpenAI. https://openai.com/blog/dall-e-2-extending-creativity/, July 2022.

[OPVM19] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[PF16] Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature News*, 538(7624):161, 2016.

[PFS21] Stephen R. Pfohl, Agata Foryciarz, and Nigam Haresh Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, page 103621, 2021.

[PGD21] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[PLKS20] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, ed-

itors, *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930. Curran Associates, Inc., 2020.

[PMSY21] Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. Post-processing for individual fairness. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[PRM+17] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2233–2241. IEEE Computer Society, 2017.

[PRW+17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5684–5693, Red Hook, NY, USA, 2017. Curran Associates Inc.

[PSG19] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[PW17] Yury Polyanskiy and Y. Wu. Strong data-processing inequalities for channels and bayesian networks. *arXiv: Information Theory*, pages 211–249, 2017.

[Rae22] Nina Raemont. Adobe stock to allow ai-generated images on its service, December 2022.

[RBFV20] Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. Learning certified individually fair representations. In *Advances in Neural Information Processing Systems 33*, 2020.

[RBKL20] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 469–481, New York, NY, USA, 2020. Association for Computing Machinery.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[RDN+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark

Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[RKB21] Candace Ross, B. Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *NAACL*, 2021.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[RPG+21] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021.

[RT15] Mohammad Sadegh Rasooli and Joel R. Tetreault. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733, 2015. version 2.

[SA20] Sidak Pal Singh and Dan Alistarh. Woodfisher: Efficient second-order approximation for neural network compression. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18098–18109. Curran Associates, Inc., 2020.

[SB21] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. *ArXiv*, abs/2104.08666, 2021.

[SBDK23] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22522–22531. IEEE, 2023.

[SBH13] Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. *ArXiv*, abs/1303.1208, 2013.

[SBV+22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia

Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022.

[SC21a] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*, New York, NY, USA, 23–24 Feb 2021.

[SC21b] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 701–713, 2021.

[SCIF20] Vivek K. Singh, Mary Chayko, Raj Inamdar, and Diana Floegel. Female librarians and male computer programmers? gender bias in occupational images on digital media platforms. *J. Assoc. Inf. Sci. Technol.*, 71(11):1281–1294, 2020.

[Sco15] Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015.

[SCS+22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[SGP+22] Prasanna Sattigeri, Soumya Ghosh, Inkit Padhi, Pierre Dognin, and Kush R. Varshney. Fair infinitesimal jackknife: Mitigating the influence of biased training data points without refitting. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[Sha17] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *J. Mach. Learn. Res.*, 18(1):1703–1713, January 2017.

[SHB+17] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv: Machine Learning*, 2017.

[SHF+23] Lukas Struppek, Dominik Hintersdorf, Felix Friedrich, Manuel Brack,

Patrick Schramowski, and Kristian Kersting. Exploiting cultural biases via homoglyphs in text-to-image synthesis, 2023.

[Sid05] Naeem Siddiqi. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Wiley, September 2005.

[SKG+19] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and S. Ermon. Learning controllable fair representations. In *AISTATS*, 2019.

[SL16] J. Skeem and Christopher T. Lowenkamp. Risk, race, recidivism: Predictive bias and disparate impact. *Political Economy: Structure & Scope of Government eJournal*, 2016.

[SLYZ23] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models, 2023.

[SN20] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278. PMLR, 13–18 Jul 2020.

[STY17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors,

Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017.

[SZC+20] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

[TB19] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[TDL+20] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, and X. Hu. Mitigating gender bias in captioning systems. *ArXiv*, abs/2006.08315, 2020.

[TDL+21] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, WWW '21, page 633–645, New York, NY, USA, 2021. Association for Computing Machinery.

[TNKB20] Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and José H. Blanchet. A distributionally robust approach to fair classification. *ArXiv*, abs/2007.09530, 2020.

[TSZ20] Shuhan Tan, Yujun Shen, and Bolei Zhou. Improving the fairness of deep

generative models without retraining. *arXiv preprint arXiv:2012.04842*, 2020.

[Tzi18] Konstantinos Tzioumis. Demographic aspects of first names. *Scientific Data*, 5, 2018.

[ULP19] Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6373–6382, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[USL19] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 10–19, New York, NY, USA, 2019. Association for Computing Machinery.

[VBC18] Effy Vayena, Alessandro Blasimme, and I Glenn Cohen. Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, 15(11):e1002689, 2018.

[vdOLV18] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

[VDOV+17] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[VJS+19] Nam S. Vo, Lu Jiang, C. Sun, K. Murphy, L. Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6432–6441, 2019.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.

[WBC22] Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan. Evidence for hypodescent in visual semantic ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1293–1304, New York, NY, USA, 2022. Association for Computing Machinery.

[WD20] Shijie Wu and Mark Dredze. Are all languages created equal in multilingual bert? In *REPL4NLP*, 2020.

[WGN+20] S. Wang, Wenshuo Guo, H. Narasimhan, Andrew Cotter, M. Gupta, and

Michael I. Jordan. Robust optimization for fairness with noisy protected groups. *ArXiv*, abs/2002.09343, 2020.

[WGOS17] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1920–1953. PMLR, 07–10 Jul 2017.

[WHM19] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019.

[Wig98] Linda F. Wightman. Lsac national longitudinal bar passage study. lsac research report series., 1998.

[WLL21] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 526–536, New York, NY, USA, 2021. Association for Computing Machinery.

[WLW21a] Jialu Wang, Yang Liu, and Xin Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Pro-*

cessing, pages 1995–2008, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[WLW21b] Jialu Wang, Yang Liu, and Xin Eric Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *EMNLP*, 2021.

[WLW22] Jialu Wang, Yang Liu, and Xin Wang. Assessing multilingual fairness in pre-trained multimodal representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2681–2695, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[WLZ+22] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. REVISE: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision (IJCV)*, 2022.

[WNR20] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130:1790 – 1810, 2020.

[WSL+14] J. Wang, Yang Song, Thomas Leung, C. Rosenberg, J. Philbin, Bo Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking.

2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 1386–1393, 2014.

[WWL22] Jialu Wang, Xin Eric Wang, and Yang Liu. Understanding instance-level impact of fairness constraints. In *International Conference on Machine Learning*, pages 23114–23130. PMLR, 2022.

[WYL+22] Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models, 2022.

[XCR+20] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer, 2020.

[YLHH14a] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[YLHH14b] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[YTY+21] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng

Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3208–3216. AAAI Press, 2021.

[YXK+22] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. Featured Certification.

[ZC22] Miao Zhang and Rumi Chunara. Fair contrastive pre-training for geographic images, 2022.

[ZDL22] Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022.

[ZG19a] H. Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. In *NeurIPS*, 2019.

[ZG19b] Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[ZHF+21] Yuyin Zhou, Shih-Cheng Huang, Jason Alan Fries, Alaa Youssef, Timothy J. Amrhein, Marcello Chang, Imon Banerjee, Daniel Rubin, Lei Xing, Nigam Shah, and Matthew P. Lungren. Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr, 2021.

[ZLL22] Zhaowei Zhu, Tianyi Luo, and Yang Liu. The rich get richer: Disparate impact of semi-supervised learning. In *International Conference on Learning Representations*, 2022.

[ZMH+20] Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online, July 2020. Association for Computational Linguistics.

[ZPZ+20] L. Zhou, H. Palangi, Lei Zhang, H. Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2020.

[ZSL21] Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[ZVGRG17a] M. Zafar, I. Valera, M. Gomez-Rodriguez, and K. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017.

[ZVGRG17b] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[ZVRG15] M. Zafar, I. Valera, M. G. Rodriguez, and K. Gummadi. Learning fair classifiers. *arXiv: Machine Learning*, 2015.

[ZVRG17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 962–970. Proceedings of Machine Learning Research, 2017.

[ZWS+13] R. Zemel, Ledell Yu Wu, Kevin Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML*, 2013.

[ZWY+17] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[ZWY+18] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Appendix A

Proofs

A.1 Proofs for Chapter 2

A.1.1 Proof of Theorem 2.1

Proof. Equality of TPR on the noisy data implies

$$\Pr(f(X) = +1 \mid Y = +1, Z = z) = \Pr(f(X) = +1 \mid \widetilde{Y} = +1, Z = z'). \tag{A.1}$$

Since the two groups are drawn from the identical distribution, we have for $y \in \{-1, +1\}$,

$$\Pr(f(X) = +1 \mid Y = y, Z = z) = \Pr(f(X) = +1 \mid Y = y, Z = z'). \tag{A.2}$$

Expanding $\Pr(f(X) = +1 \mid Y = +1, Z = z')$ using law of total probability we have

$$\begin{aligned}
&\Pr(f(X) = +1 | Y = +1, Z = z') \\
=\ &\Pr(f(X) = +1 | \mid \widetilde{Y} = +1, Y = +1, Z = z') \cdot \Pr(\widetilde{Y} = +1 \mid Y = +1, Z = z') \\
&+ \Pr(f(X) = +1 \mid \widetilde{Y} = -1, Y = +1, Z = z') \cdot \Pr(\widetilde{Y} = -1 \mid Y = +1, Z = z') \\
=\ &\Pr(f(X) = +1 \mid \widetilde{Y} = +1, Z = z') \cdot (1 - e) + \Pr(f(X) = +1 \mid \widetilde{Y} = -1, Z = z') \cdot e
\end{aligned} \tag{A.3}$$

Combining Equation (A.1) with the above,

$$\begin{aligned}
&\Pr(f(X) = +1 \mid \widetilde{Y} = +1, Z = z') \\
=\ &\Pr(f(X) = +1 \mid Y = +1, Z = z) && \text{(by Equation (A.1))} \\
=\ &\Pr(f(X) = +1 \mid \widetilde{Y} = +1, Z = z') \cdot (1 - e) + \Pr(f(X) = +1 \mid \widetilde{Y} = -1, Z = z') \cdot e \\
& && \text{(by Equation (A.3))} \\
\Leftrightarrow\ &\Pr(f(X) = +1 \mid \widetilde{Y} = +1, Z = z') \cdot e \\
=\ &\Pr(f(X) = +1 \mid \widetilde{Y} = -1, Z = z') \cdot e && \text{(A.4)} \\
\Leftrightarrow\ &\Pr(f(X) = +1 \mid \widetilde{Y} = +1, Z = z') = \Pr(f(X) = +1 \mid \widetilde{Y} = -1, Z = z') && \text{(A.5)}
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\Pr(f(X) = +1 \mid Y = -1, Z = z') \\
=\ &\Pr(f(X) = +1 \mid \widetilde{Y} = +1, Z = z') \cdot e \\
&+ \Pr(f(X) = +1 \mid \widetilde{Y} = -1, Z = z') \cdot (1 - e) \\
=\ &\Pr(f(X) = +1 \mid \widetilde{Y} = -1, Z = z') \quad \text{(by Equation (A.5))} \qquad \text{(A.6)}
\end{aligned}$$

Equation (A.5) and (A.6) jointly imply

$$\begin{aligned}
&\Pr(f(X) = +1 \mid Y = -1, Z = z) \\
=&\Pr(f(X) = +1 \mid Y = -1, Z = z') && \text{(by Equation (A.2))} \\
=&\Pr(f(X) = +1 \mid \widetilde{Y} = -1, Z = z') && \text{(by Equation A.6)} \\
=&\Pr(f(X) = +1 \mid \widetilde{Y} = +1, Z = z') && \text{(by Equation A.5)} \\
=&\Pr(f(X) = +1 \mid Y = +1, Z = z), && \text{(by Equation (A.1))}
\end{aligned}$$

thus completing the proof. □

A.1.2 Proof of Lemma 2.2

Proof. Expanding $\Pr(f(X) = +1 \mid Y = +1, Z = z)$ using law of total probability, we have

$$\begin{aligned}
\mathrm{TPR}_z &= \Pr(f(X) = +1 \mid Y = +1, Z = z) \\
&= \Pr(f(X) = +1, \tilde{Y} = +1 \mid Y = +1, Z = z) \\
&\quad + \Pr(f(X) = +1, \tilde{Y} = -1 \mid Y = +1, Z = z) \\
&= \Pr(\tilde{Y} = +1 | Y = +1, Z = z) \cdot \Pr(f(X) = +1 \mid \tilde{Y} = +1, Y = +1, Z = z) \\
&\quad + \Pr(\tilde{Y} = -1 | Y = +1, Z = z) \cdot \Pr(f(X) = +1 \mid \tilde{Y} = -1, Y = +1, Z = z) \\
&= \Pr(\tilde{Y} = +1 | Y = +1, Z = z) \cdot \Pr(f(X) = +1 \mid \tilde{Y} = +1, Z = z) \\
&\quad + \Pr(\tilde{Y} = -1 | Y = +1, Z = z) \cdot \Pr(f(X) = +1 \mid \tilde{Y} = -1, Z = z) \\
&= (1 - \epsilon_z^+) \cdot \widetilde{\mathrm{TPR}}_z + \epsilon_z^+ \cdot \widetilde{\mathrm{FPR}}_z \qquad \text{(A.7)}
\end{aligned}$$

Note in the above we drop the dependence on Y when conditioning on $\tilde{Y}$. This is because f is trained purely on the noisy labels, and $\tilde{Y}$ encodes all the information f has about Y. A similar derivation holds for FPR_z. □

A.1.3 Proof of Theorem 2.4

Proof. Noticing that $\widetilde{\mathrm{TPR}}_z = \widetilde{\mathrm{TPR}}_{z'}$ and $\widetilde{\mathrm{FPR}}_z = \widetilde{\mathrm{FPR}}_{z'}$ (equalizing fairness metrics on the noisy data) and applying Lemma 2.2, we obtain

$$\begin{aligned}
|\,\mathrm{TPR}_z - \mathrm{TPR}_{z'}\,| &= |((1-\epsilon_z^+)\cdot\widetilde{\mathrm{TPR}}_z + \epsilon_z^+\cdot\widetilde{\mathrm{FPR}}_z) - ((1-\epsilon_{z'}^+)\cdot\widetilde{\mathrm{TPR}}_{z'} + \epsilon_{z'}^+\cdot\widetilde{\mathrm{FPR}}_{z'})| \\
&= |\epsilon_z^+\cdot(\widetilde{\mathrm{FPR}}_z - \widetilde{\mathrm{TPR}}_z) - \epsilon_{z'}^+\cdot(\widetilde{\mathrm{FPR}}_z - \widetilde{\mathrm{TPR}}_z)| \\
&= |(\epsilon_z^+ - \epsilon_{z'}^+)\cdot(\widetilde{\mathrm{FPR}}_z - \widetilde{\mathrm{TPR}}_z)| \\
&= |\widetilde{\mathrm{TPR}}_z - \widetilde{\mathrm{FPR}}_z|\cdot|\epsilon_z^+ - \epsilon_{z'}^+|
\end{aligned}$$

The argument for FPR is symmetrical:

$$\begin{aligned}
|\,\mathrm{FPR}_z - \mathrm{FPR}_{z'}\,| &= |(\epsilon_z^-\cdot\widetilde{\mathrm{TPR}}_z + (1-\epsilon_z^-)\cdot\widetilde{\mathrm{FPR}}_z) - (\epsilon_{z'}^-\cdot\widetilde{\mathrm{TPR}}_{z'} + (1-\epsilon_{z'}^+)\cdot\widetilde{\mathrm{FPR}}_{z'})| \\
&= |\epsilon_z^-\cdot(\widetilde{\mathrm{TPR}}_z - \widetilde{\mathrm{FPR}}_z) - \epsilon_{z'}^-\cdot(\widetilde{\mathrm{TPR}}_z - \widetilde{\mathrm{FPR}}_z)| \\
&= |(\epsilon_z^- - \epsilon_{z'}^-)\cdot(\widetilde{\mathrm{TPR}}_z - \widetilde{\mathrm{FPR}}_z)| \\
&= |\widetilde{\mathrm{TPR}}_z - \widetilde{\mathrm{FPR}}_z|\cdot|\epsilon_z^- - \epsilon_{z'}^-|
\end{aligned}$$

Therefore

$$|\,\mathrm{TPR}_z - \mathrm{TPR}_{z'}\,| > 0, |\,\mathrm{FPR}_z - \mathrm{FPR}_{z'}\,| > 0,$$

when $\widetilde{\mathrm{TPR}}_z \neq \widetilde{\mathrm{FPR}}_z$, $\epsilon_z^+ \neq \epsilon_{z'}^+, \epsilon_z^- \neq \epsilon_{z'}^-$. □

A.1.4 Proof of Theorem 2.5

Proof. Observe that

$$\ell_{gp}(f(\boldsymbol{x}_i), \tilde{y}) = \frac{1}{\Delta_{z_i}}\ell_{peer}(f(\boldsymbol{x}_i), \tilde{y})$$

Taking expectations over noisy data, we have

$$
\begin{aligned}
&\mathbb{E}_{\tilde{\mathcal{D}}}[\ell_{gp}(f(X), \tilde{Y})] \\
&= \frac{1}{|I|} \cdot \sum_{z \in Z} |I_z| \cdot \mathbb{E}_{\tilde{\mathcal{D}}_z}[\ell_{gp}(f(X_z), \tilde{Y}_z)] \\
&= \frac{1}{|I|} \cdot \sum_{z \in Z} \frac{|I_z|}{\Delta_z} \cdot \mathbb{E}_{\tilde{\mathcal{D}}_z}[\ell_{peer}(f(X_z), \tilde{Y}_z)] \\
&= \frac{1}{|I|} \cdot \sum_{z \in Z} \frac{|I_z|}{\Delta_z} \cdot \Delta_z \mathbb{E}_{\mathcal{D}_z}[\ell_{peer}(f(X_z), Y_z)] \qquad \text{(by Equation (2.13))} \\
&= \mathbb{E}_{\mathcal{D}}[\ell_{peer}(f(X), Y)] \qquad \text{(A.8)}
\end{aligned}
$$

Notice that $\alpha = 1$ when $\Pr(Y = +1) = \Pr(Y = -1) = \frac{1}{2}$, the definition of peer loss function gives

$$
\mathbb{E}_{X,Y}[\ell_{peer}(f(X), Y)] = \mathbb{E}_{X,Y}[\ell(f(X), Y)] - \mathbb{E}_X \mathbb{E}_Y[\ell(f(X), Y)] \qquad \text{(A.9)}
$$

Using the assumption that $\Pr(Y = +1) = \Pr(Y = -1) = \frac{1}{2}$ and the fact that ℓ is 0-1 loss function,

$$
\begin{aligned}
\mathbb{E}_X \mathbb{E}_Y[\ell(f(X), Y)] &= \Pr(Y = +1) \cdot \mathbb{E}_X[\ell(f(X), +1)] + \Pr(Y = -1) \cdot \mathbb{E}_X[\ell(f(X), -1)] \\
&= \frac{1}{2} \cdot \ell(f(X), +1) + \frac{1}{2} \cdot \ell(f(X), +1) \\
&= \frac{1}{2} \cdot \mathbb{1}(f(X) \neq +1) + \frac{1}{2} \cdot \mathbb{1}(f(X) \neq -1) \\
&= \frac{1}{2} \Pr(f(X) = -1) + \frac{1}{2} \cdot \Pr(f(X) = +1) \\
&= \frac{1}{2} \qquad \text{(A.10)}
\end{aligned}
$$

Combining Equation (A.8), (A.9) and (A.10), we complete the proof

$$\mathbb{E}_{\tilde{\mathcal{D}}}[\ell_{gp}(f(X), \tilde{Y})] = \mathbb{E}_{\mathcal{D}}[\ell(f(X), Y] - \frac{1}{2} \tag{A.11}$$

□

A.1.5 Proof of Lemma 2.6

Proof. Following Lemma 2.2 we have,

$$\mathrm{TPR}_z - \mathrm{FPR}_z = (1 - \epsilon_z^+ - \epsilon_z^-)(\widetilde{\mathrm{TPR}} - \widetilde{\mathrm{FPR}}) = \Delta_z \cdot (\widetilde{\mathrm{TPR}} - \widetilde{\mathrm{FPR}})$$

Notice that

$$\begin{aligned}
\Pr(f(X) = +1 \mid Z = z) &= \Pr(Y = +1 \mid Z = z) \cdot \Pr(f(X) = +1 \mid Y = +1, Z = z) \\
&\quad + \Pr(Y = -1 | Z = z) \cdot \Pr(f(X) = +1 \mid Y = -1, Z = z) \\
&= \Pr(Y = +1 | Z = z) \cdot \mathrm{TPR}_z + \Pr(Y = -1 | Z = z) \cdot \mathrm{FPR}_z
\end{aligned}$$

Solving the two equations above we complete the proof. □

A.1.6 Proof of Theorem 2.7

Proof. Define the following risk measures

$$\tilde{R}(f) := \frac{1}{N} \sum_{i=1}^{N} \tilde{\ell}(f(x_i), \tilde{y}_i)$$

$$\hat{R}(f) := \frac{1}{N} \sum_{i=1}^{N} \hat{\ell}(f(x_i), \tilde{y}_i)$$

First, because of Equation (2.21), we have

$$
\begin{aligned}
\tilde{R}(f) &= \frac{1}{N}\sum_{i=1}^{N}\tilde{\ell}(f(x_i),\tilde{y}_i) \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{(1-\epsilon_{z_i}^{sgn(-\tilde{y}_i)})\ell(f(\boldsymbol{x}_i),\tilde{y}_i)-\epsilon_{z_i}^{sgn(\tilde{y}_i)}\ell(f(\boldsymbol{x}_i),-\tilde{y}_i)}{1-\epsilon_{z_i}^{+}-\epsilon_{z_i}^{-}} \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{(1-\hat{\epsilon}_{z_i}^{sgn(-\tilde{y}_i)})\ell(f(\boldsymbol{x}_i),\tilde{y}_i)-\hat{\epsilon}_{z_i}^{sgn(\tilde{y}_i)}\ell(f(\boldsymbol{x}_i),-\tilde{y}_i)}{1-\hat{\epsilon}_{z_i}^{+}-\hat{\epsilon}_{z_i}^{-}} \\
&+ \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1-\epsilon_{z_i}^{sgn(-\tilde{y}_i)}}{1-\epsilon_z^{+}-\epsilon_z^{-}}-\frac{1-\hat{\epsilon}_{z_i}^{sgn(-\tilde{y}_i)}}{1-\hat{\epsilon}_z^{+}-\hat{\epsilon}_z^{-}}\right)\ell(f(\boldsymbol{x}_i),\tilde{y}_i) \\
&+ \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\epsilon^{sgn(\tilde{y}_i)}}{1-\epsilon_z^{+}-\epsilon_z^{-}}-\frac{\hat{\epsilon}^{sgn(\tilde{y}_i)}}{1-\hat{\epsilon}_z^{+}-\hat{\epsilon}_z^{-}}\right)\ell(f(\boldsymbol{x}_i),-\tilde{y}_i) \\
&= \hat{R}(f) \qquad \text{(by definition of } \hat{R}(f)\text{)} \\
&+ \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1-\epsilon_{z_i}^{sgn(-\tilde{y}_i)}}{1-\epsilon_z^{+}-\epsilon_z^{-}}-\frac{1-\hat{\epsilon}_{z_i}^{sgn(-\tilde{y}_i)}}{1-\hat{\epsilon}_z^{+}-\hat{\epsilon}_z^{-}}\right)\ell(f(\boldsymbol{x}_i),\tilde{y}_i) \\
&+ \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\epsilon^{sgn(\tilde{y}_i)}}{1-\epsilon_z^{+}-\epsilon_z^{-}}-\frac{\hat{\epsilon}^{sgn(\tilde{y}_i)}}{1-\hat{\epsilon}_z^{+}-\hat{\epsilon}_z^{-}}\right)\ell(f(\boldsymbol{x}_i),-\tilde{y}_i).
\end{aligned}
$$

Using the error bound in Equation (2.21), we have

$$
\begin{aligned}
&\left|\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1-\epsilon_{z_i}^{sgn(-\tilde{y}_i)}}{1-\epsilon_z^{+}-\epsilon_z^{-}}-\frac{1-\hat{\epsilon}_{z_i}^{sgn(-\tilde{y}_i)}}{1-\hat{\epsilon}_z^{+}-\hat{\epsilon}_z^{-}}\right)\ell(f(\boldsymbol{x}_i),\tilde{y}_i)+\left(\frac{\epsilon^{sgn(\tilde{y}_i)}}{1-\epsilon_z^{+}-\epsilon_z^{-}}-\frac{\hat{\epsilon}^{sgn(\tilde{y}_i)}}{1-\hat{\epsilon}_z^{+}-\hat{\epsilon}_z^{-}}\right)\ell(f(\boldsymbol{x}_i),-\tilde{y}_i)\right| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left|\left(\frac{1-\epsilon_{z_i}^{sgn(-\tilde{y}_i)}}{1-\epsilon_z^{+}-\epsilon_z^{-}}-\frac{1-\hat{\epsilon}_{z_i}^{sgn(-\tilde{y}_i)}}{1-\hat{\epsilon}_z^{+}-\hat{\epsilon}_z^{-}}\right)\right|\ell(f(\boldsymbol{x}_i),\tilde{y}_i) \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\left|\left(\frac{\epsilon^{sgn(\tilde{y}_i)}}{1-\epsilon_z^{+}-\epsilon_z^{-}}-\frac{\hat{\epsilon}^{sgn(\tilde{y}_i)}}{1-\hat{\epsilon}_z^{+}-\hat{\epsilon}_z^{-}}\right)\right|\ell(f(\boldsymbol{x}_i),-\tilde{y}_i) \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\tau\bar{\ell}+\frac{1}{N}\sum_{i=1}^{N}\tau\bar{\ell} \\
&= 2\tau\bar{\ell}
\end{aligned}
$$

Then we conclude that $\forall f$

$$|\tilde{R}(f) - \hat{R}(f)| \leq 2\tau\bar{\ell} \tag{A.12}$$

This enables us to obtain the following bound

$$\begin{aligned}
\tilde{R}(\hat{f}^*) - \tilde{R}(\tilde{f}^*) &\leq \hat{R}(\hat{f}^*) + 2\tau\bar{\ell} - \tilde{R}(\tilde{f}^*) && \text{(by Equation (A.12))} \\
&\leq \hat{R}(\tilde{f}^*) - \tilde{R}(\tilde{f}^*) + 2\tau\bar{\ell} \\
&\leq 2\tau\bar{\ell} + 2\tau\bar{\ell} && \text{(by Equation (A.12))} \\
&= 4\tau\bar{\ell}.
\end{aligned}$$

where in above the 2nd inequality is due to the optimality $\hat{f}^*$ wrt $\hat{R}(f)$. □

A.1.7 Proof for Lemma 2.8

Proof.

$$\mathbb{P}(h(X) = +1|\tilde{Y} = +1, Z = a) = \frac{\mathbb{P}(h(X) = +1, \tilde{Y} = +1|Z = a)}{\mathbb{P}(\tilde{Y} = +1|Z = a)} \tag{A.13}$$

Again we do the trick of sampling $\mathbb{P}(\tilde{Y} = +1|Z = a)$ to be 0.5, which allows us to focus

on the numerator.

$$\begin{aligned}
&\mathbb{P}(h(X)=+1,\tilde{Y}=+1|Z=a)\\
=\ &\mathbb{P}(h(X)=+1,\tilde{Y}=+1,Y=+1|Z=a)\\
&+\mathbb{P}(h(X)=+1,\tilde{Y}=+1,Y=-1|Z=a)\\
=\ &\mathbb{P}(h(X)=+1,\tilde{Y}=+1|Y=+1,Z=a)\cdot\mathbb{P}(Y=+1|Z=a)\\
&+\mathbb{P}(h(X)=+1,\tilde{Y}=+1|Y=-1,Z=a)\cdot\mathbb{P}(Y=-1|Z=a)\\
=\ &\mathbb{P}(h(X)=+1|Y=+1,Z=a)\cdot(1-e_a)\cdot\mathbb{P}(Y=+1|Z=a)\\
&+\mathbb{P}(h(X)=+1|Y=-1,Z=a)\cdot e_a\cdot\mathbb{P}(Y=-1|Z=a)
\end{aligned}$$

(Independence of X and $\tilde{Y}$ given Y)

That is

$$0.5\cdot\widetilde{\mathrm{TPR}}_a(h)=$$

$$\mathrm{TPR}_a(h)\cdot(1-e_a)\cdot\mathbb{P}(Y=+1|Z=a)+\mathrm{FPR}_a(h)\cdot e_a\cdot\mathbb{P}(Y=-1|Z=a) \quad \text{(A.14)}$$

Similarly for FPR we have

$$\mathbb{P}(h(X)=+1|\tilde{Y}=-1,Z=a)=\frac{\mathbb{P}(h(X)=+1,\tilde{Y}=-1|Z=a)}{\mathbb{P}(\tilde{Y}=-1|Z=a)} \quad \text{(A.15)}$$

Following similar steps as above, the numerator further derives as

$$\begin{aligned}
&\mathbb{P}(h(X)=+1,\tilde{Y}=+1|Z=a)\\
=\ &\mathbb{P}(h(X)=+1|Y=-1,Z=a)\cdot(1-e_a)\cdot\mathbb{P}(Y=+1|Z=a)\\
&+\mathbb{P}(h(X)=+1|Y=+1,Z=a)\cdot e_a\cdot\mathbb{P}(Y=-1|Z=a)
\end{aligned}$$

That is

$$0.5 \cdot \widetilde{\mathrm{FPR}}_a(h) =$$

$$\mathrm{FPR}_a(h) \cdot (1 - e_a) \cdot \mathbb{P}(Y = +1|Z = a) + \mathrm{TPR}_a(h) \cdot e_a \cdot \mathbb{P}(Y = -1|Z = a) \quad \text{(A.16)}$$

When $\mathbb{P}(\widetilde{Y} = +1|Z = a) = \mathbb{P}(\widetilde{Y} = +1|Z = b) = 0.5$, we will also have

$$0.5 = \mathbb{P}(\widetilde{Y} = +1|Z = a) = \mathbb{P}(Y = +1|Z = a)(1 - e_a) + \mathbb{P}(Y = -1|Z = a)e_a \quad \text{(A.17)}$$

which returns us that $\mathbb{P}(Y = +1|Z = a) = \frac{0.5-e_a}{1-2e_a} := p = 0.5$. Using this knowledge and solving the linear equations defined by Eqn. (A.14) and (A.16) we have

$$\mathrm{TPR}_a(h) = \frac{C_{a,1} \cdot \widetilde{\mathrm{TPR}}_a(h) - C_{a,2} \cdot \widetilde{\mathrm{FPR}}_a(h)}{e_a - 0.5} \quad \text{(A.18)}$$

$$\mathrm{FPR}_a(h) = \frac{C_{a,1} \cdot \widetilde{\mathrm{FPR}}_a(h) - C_{a,2} \cdot \widetilde{\mathrm{TPR}}_a(h)}{e_a - 0.5} \quad \text{(A.19)}$$

□

A.1.8 Proof of Theorem 2.9

Proof. Combining Equation (2.25) and (2.28) we have

$$\begin{aligned} & |\,\mathrm{TPR}_z(h) - \mathrm{TPR}_z^c(h)| \\ = \; & \left| \frac{0.5 \cdot e_z \cdot \widetilde{\mathrm{TPR}}_z(h) - 0.5(1 - e_z) \cdot \widetilde{\mathrm{FPR}}_z(h)}{e_z - 0.5} - \right. \quad \text{(A.20)} \\ & \left. \frac{0.5 \cdot \widetilde{e}_z \cdot \widetilde{\mathrm{TPR}}_z(h) - 0.5(1 - \widetilde{e}_z) \cdot \widetilde{\mathrm{FPR}}_z(h)}{\tilde{e}_z - 0.5} \right| \\ = \; & \frac{|\tilde{e}_z - e_z| \cdot \widetilde{\mathrm{TPR}}_z(h)}{(2e_z - 1)(2\tilde{e}_z - 1)} \\ = \; & \frac{\mathtt{err}_z \cdot \widetilde{\mathrm{TPR}}_z(h)}{(2e_z - 1)(2\tilde{e}_z - 1)} \; . \quad \text{(A.21)} \end{aligned}$$

Recall $\texttt{err}_z = |\tilde{e}_z - e_z|$. The second equality is algebraic - we simply unify the denominator of both quantities and rearrange terms. Then equalizing TPR that $\text{TPR}_a^c(h) = \text{TPR}_b^c(h)$ returns us

$$
\begin{aligned}
&|\,\text{TPR}_a(h) - \text{TPR}_b(h)| \\
=&|\,\text{TPR}_a(h) - \text{TPR}_a^c(h) + \text{TPR}_b^c(h) - \text{TPR}_b(h)| \\
\geq&\,||\,\text{TPR}_a(h) - \text{TPR}_a^c(h)| - |\,\text{TPR}_b^c(h) - \text{TPR}_b(h)|| \\
=&\left|\frac{\texttt{err}_a \cdot \widetilde{\text{TPR}}_a(h)}{(2e_a-1)(2\tilde{e}_a-1)} - \frac{\texttt{err}_b \cdot \widetilde{\text{TPR}}_b(h)}{(2e_b-1)(2\tilde{e}_b-1)}\right|,
\end{aligned}
$$

where the last equality is an application of Eqn. (A.21). Then

$$
\begin{aligned}
&\left|\frac{\texttt{err}_a \cdot \widetilde{\text{TPR}}_a(h)}{(2e_a-1)(2\tilde{e}_a-1)} - \frac{\texttt{err}_b \cdot \widetilde{\text{TPR}}_b(h)}{(2e_b-1)(2\tilde{e}_b-1)}\right| \\
=&\texttt{err}_a \cdot \left|\frac{\widetilde{\text{TPR}}_a(h)}{(2e_a-1)(2\tilde{e}_a-1)} - \frac{\texttt{err}_b}{\texttt{err}_a}\frac{\widetilde{\text{TPR}}_b(h)}{(2e_b-1)(2\tilde{e}_b-1)}\right| \\
\geq&\texttt{err}_{\texttt{M}} \cdot \left|\frac{\widetilde{\text{TPR}}_a(h)}{(2e_a-1)(2\tilde{e}_a-1)} - \frac{\texttt{err}_b}{\texttt{err}_a}\frac{\widetilde{\text{TPR}}_b(h)}{(2e_b-1)(2\tilde{e}_b-1)}\right|
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&|\,\text{FPR}_z(h) - \text{FPR}_z^c(h)| \\
=&\left|\frac{0.5 \cdot e_z \cdot \widetilde{\text{FPR}}_z(h) - 0.5(1-e_z) \cdot \widetilde{\text{TPR}}_z(h)}{e_z - 0.5}\right. \\
&\qquad \left. - \frac{0.5 \cdot \widetilde{e}_z \cdot \widetilde{\text{FPR}}_z(h) - 0.5(1-\widetilde{e}_z) \cdot \widetilde{\text{TPR}}_z(h)}{\tilde{e}_z - 0.5}\right| \\
=&\frac{|\tilde{e}_z - e_z| \cdot \widetilde{\text{FPR}}_z(h)}{(2e_z-1)(2\tilde{e}_z-1)}. \\
=&\frac{\texttt{err}_z \cdot \widetilde{\text{FPR}}_z(h)}{(2e_z-1)(2\tilde{e}_z-1)}.
\end{aligned}
$$

Then equalizing FPR that $\mathrm{FPR}_a^c(h) = \mathrm{FPR}_b^c(h)$ we have

$$
\begin{aligned}
&|\,\mathrm{FPR}_a(h) - \mathrm{FPR}_b(h)| \\
=&|\,\mathrm{FPR}_a(h) - \mathrm{FPR}_a^c(h) + \mathrm{FPR}_b^c(h) - \mathrm{FPR}_b(h)| \\
\geq&\,||\,\mathrm{FPR}_a(h) - \mathrm{FPR}_a^c(h)| - |\,\mathrm{FPR}_b^c(h) - \mathrm{FPR}_b(h)|| \\
\geq&\left|\frac{\mathtt{err}_a \cdot \widetilde{\mathrm{FPR}}_a(h)}{(2e_a-1)(2\tilde{e}_a-1)} - \frac{\mathtt{err}_b \cdot \widetilde{\mathrm{FPR}}_b(h)}{(2e_b-1)(2\tilde{e}_b-1)}\right| \\
\geq&\mathtt{err}_{\mathtt{M}} \cdot \left|\frac{\widetilde{\mathrm{FPR}}_a(h)}{(2e_a-1)(2\tilde{e}_a-1)} - \frac{\mathtt{err}_b}{\mathtt{err}_a}\frac{\widetilde{\mathrm{FPR}}_b(h)}{(2e_b-1)(2\tilde{e}_b-1)}\right|.
\end{aligned}
$$

□

A.1.9 Proof of Theorem 2.10

Proof. Easy to show that when $e_a = e_b$, $C_{a,1} = C_{b,1}$ and $C_{a,2} = C_{b,2}$. Therefore, from Eqn. (2.25) we know equalizing

$$\widetilde{\mathrm{TPR}}_a(h) = \widetilde{\mathrm{TPR}}_b(h), \quad \widetilde{\mathrm{FPR}}_a(h) = \widetilde{\mathrm{FPR}}_b(h) \tag{A.22}$$

will also return us

$$\mathrm{TPR}_a(h) = \mathrm{TPR}_b(h), \quad \mathrm{FPR}_a(h) = \mathrm{FPR}_b(h) \tag{A.23}$$

□

A.1.10 Proof of Theorem 2.11

Proof. We start with deriving $\text{PA}_{\mathcal{D}^\diamond}$:

$$\text{PA}_{\mathcal{D}^\diamond} = \mathbb{P}(\widetilde{Y}_2 = \widetilde{Y}_3 = +1 \mid \widetilde{Y}_1 = +1) = \frac{\mathbb{P}(\widetilde{Y}_1 = \widetilde{Y}_2 = \widetilde{Y}_3 = +1)}{\mathbb{P}(\widetilde{Y}_1 = +1)}$$

Due to the sampling step, we have $\mathbb{P}(\widetilde{Y}_1 = +1) = 0.5$ - this allows us to focus on the denominator:

$$\begin{aligned}
& \mathbb{P}(\widetilde{Y}_1 = \widetilde{Y}_2 = \widetilde{Y}_3 = +1) \\
\overset{(1)}{=} \; & \mathbb{P}(Y = +1) \prod_{i=1}^{3} \mathbb{P}(\widetilde{Y}_i = +1 | Y = +1) + \mathbb{P}(Y = -1) \prod_{i=1}^{3} \mathbb{P}(\widetilde{Y}_i = +1 | Y = -1) \\
\overset{(2)}{=} \; & \mathbb{P}(Y = +1) \cdot (1 - e_+)^3 + \mathbb{P}(Y = -1) \cdot e_-^3
\end{aligned}$$

where in above, (1) uses the 2-NN clusterability of D, and (2) uses the conditional independence between the noisy labels. Similarly for $\text{NA}_{\mathcal{D}^\diamond}$ we have:

$$\mathbb{P}(\widetilde{Y}_2 = \widetilde{Y}_3 = -1 | \widetilde{Y}_1 = -1) = \frac{\mathbb{P}(\widetilde{Y}_1 = \widetilde{Y}_2 = \widetilde{Y}_3 = -1)}{\mathbb{P}(\widetilde{Y}_1 = -1)}$$

Again we have that $\mathbb{P}(\widetilde{Y}_1 = -1) = 0.5$, and the numerator derives as

$$\begin{aligned}
& \mathbb{P}(\widetilde{Y}_1 = \widetilde{Y}_2 = \widetilde{Y}_3 = -1) \\
= \; & \mathbb{P}(Y = +1) \prod_{i=1}^{3} \mathbb{P}(\widetilde{Y}_i = -1 | Y = +1) + \mathbb{P}(Y = -1) \prod_{i=1}^{3} \mathbb{P}(\widetilde{Y}_i = -1 | Y = -1) \\
= \; & \mathbb{P}(Y = +1) \cdot e_+^3 + \mathbb{P}(Y = -1) \cdot (1 - e_-)^3
\end{aligned}$$

Taking the difference (and normalize by 0.5) we have

$$\begin{aligned} & 0.5 \cdot (\mathrm{PA}_{\mathcal{D}^\diamond} - \mathrm{NA}_{\mathcal{D}^\diamond}) \\ = \; & \mathbb{P}(\widetilde{Y}_2 = \widetilde{Y}_3 = +1 | \widetilde{Y}_1 = +1) - \mathbb{P}(\widetilde{Y}_2 = \widetilde{Y}_3 = -1 | \widetilde{Y}_1 = -1) \\ = \; & \mathbb{P}(Y = +1) \left((1 - e_+)^3 - e_+^3 \right) + \mathbb{P}(Y = -1) \left(e_-^3 - (1 - e_-)^3 \right) \end{aligned} \tag{A.24}$$

Notice two facts: first we can derive that

$$(1 - e_+)^3 - e_+^3 = (1 - 2e_+)(e_+^2 - e_+ + 1), \;\; e_-^3 - (1 - e_-)^3 = -(1 - 2e_-)(e_-^2 - e_- + 1)$$

Second, we will use the following fact:

$$0.5 = \mathbb{P}(\widetilde{Y} = +1) = \mathbb{P}(Y = +1)(1 - e_+) + \mathbb{P}(Y = -1)e_- \tag{A.25}$$

from which we solve that $\mathbb{P}(Y = +1) = \frac{0.5 - e_-}{1 - e_+ - e_-}$. Symmetrically, $\mathbb{P}(Y = -1) = \frac{0.5 - e_+}{1 - e_+ - e_-}$.

Return the above two facts back into Eq (A.24), we have

$$\begin{aligned} & \mathbb{P}(Y = +1)((1 - e_+)^3 - e_+^3) + \mathbb{P}(Y = -1)(e_-^3 - (1 - e_-)^3) \\ = \; & 2 \cdot \frac{(0.5 - e_+)(0.5 - e_-)}{1 - e_+ - e_-} \left((e_+^2 - e_+ + 1) - (e_-^2 - e_- + 1) \right) \\ = \; & 2 \cdot (0.5 - e_+) \cdot (0.5 - e_-) \cdot (e_- - e_+) \end{aligned}$$

completing the proof when $e_+, e_- < 0.5$. □

A.1.11 Proof of Proposition 2.12

Proof. Expanding $\mathbb{P}(\hat{Y} = -1|Y = +1)$ using the law of total probability we have

$$\begin{aligned}
\hat{e}_+ &= \mathbb{P}(\hat{Y} = -1|Y = +1) \\
&= \mathbb{P}(\hat{Y} = -1, \widetilde{Y} = +1|Y = +1) + \mathbb{P}(\hat{Y} = -1, \widetilde{Y} = -1|Y = +1) \\
&= \mathbb{P}(\hat{Y} = -1|\widetilde{Y} = +1, Y = +1) \cdot \mathbb{P}(\widetilde{Y} = +1|Y = +1) \\
&\quad + \mathbb{P}(\hat{Y} = -1|\widetilde{Y} = -1, Y = +1) \cdot \mathbb{P}(\widetilde{Y} = -1|Y = +1) \\
&= \epsilon \cdot (1 - e_+) + 1 \cdot e_+ \qquad \text{(Independence between } \hat{Y} \text{ and } Y \text{ given } \widetilde{Y}) \\
&= (1 - e_+) \cdot \epsilon + e_+
\end{aligned}$$

Similarly,

$$\begin{aligned}
\hat{e}_- &= \mathbb{P}(\hat{Y} = +1|Y = -1) \\
&= \mathbb{P}(\hat{Y} = +1, \widetilde{Y} = +1|Y = -1) + \mathbb{P}(\hat{Y} = +1, \widetilde{Y} = -1|Y = -1) \\
&= \mathbb{P}(\hat{Y} = +1|\widetilde{Y} = +1, Y = -1) \cdot \mathbb{P}(\widetilde{Y} = +1|Y = -1) \\
&\quad + \mathbb{P}(\hat{Y} = +1|\widetilde{Y} = -1, Y = -1) \cdot \mathbb{P}(\widetilde{Y} = -1|Y = -1) \\
&= (1 - \epsilon) \cdot e_-.
\end{aligned}$$

The last equality is again due to the independence between $\hat{Y}$ and Y given $\widetilde{Y}$, as well as the fact that we do not flip the $\widetilde{Y} = -1$ labels so $\mathbb{P}(\hat{Y} = +1|\widetilde{Y} = -1, Y = -1) = 0$. Taking the difference we finish the proof. □

A.2 Proofs for Chapter 3

A.2.1 Proof of Lemma 3.1

Proof. Given by the definition of cosine similarity, we have

$$\begin{aligned} & |S(\boldsymbol{b} * v, \boldsymbol{b} * t^{(L)}) - S(\boldsymbol{b} * v, \boldsymbol{b} * t^{(L')})| \\ = & \ |\frac{\boldsymbol{b} * v \cdot \boldsymbol{b} * t^{(L)}}{\|\boldsymbol{b} * v\| \|\boldsymbol{b} * t^{(L)}\|} - \frac{\boldsymbol{b} * v \cdot \boldsymbol{b} * t^{(L')}}{\|\boldsymbol{b} * v\| \|\boldsymbol{b} * t^{(L')}\|}| \\ = & \ \frac{|\boldsymbol{b} * v \cdot (\|\boldsymbol{b} * t^{(L')}\| \boldsymbol{b} * t^{(L)} - \|\boldsymbol{b} * t^{(L)}\| \boldsymbol{b} * t^{(L')})|}{\|\boldsymbol{b} * v\| \|\boldsymbol{b} * t^{(L)}\| \|\boldsymbol{b} * t^{(L')}\|} \end{aligned} \tag{A.26}$$

From the definition of dot product,

$$|\boldsymbol{b} * v \cdot (\|\boldsymbol{b} * t^{(L')}\| \boldsymbol{b} * t^{(L)} - \|\boldsymbol{b} * t^{(L)}\| \boldsymbol{b} * t^{(L')})| \leq$$

$$\|\boldsymbol{b} * v\| \cdot \|(\|\boldsymbol{b} * t^{(L')}\| \boldsymbol{b} * t^{(L)} - \|\boldsymbol{b} * t^{(L)}\| \boldsymbol{b} * t^{(L')})\| \tag{A.27}$$

We plug Equation (A.27) into Equation (A.26) and eliminate the variable $\boldsymbol{b} * v$

$$|S(\boldsymbol{b} * v, \boldsymbol{b} * t^{(L)}) - S(\boldsymbol{b} * v, \boldsymbol{b} * t^{(L')})| \leq \frac{\|(\|\boldsymbol{b} * t^{(L')}\| \boldsymbol{b} * t^{(L)} - \|\boldsymbol{b} * t^{(L)}\| \boldsymbol{b} * t^{(L')})\|}{\|\boldsymbol{b} * t^{(L)}\| \|\boldsymbol{b} * t^{(L')}\|} \tag{A.28}$$

Let θ denote the angle between $\boldsymbol{b} * t^{(L)}$ and $\boldsymbol{b} * t^{(L')}$, i.e.,

$$\cos\theta = \frac{\boldsymbol{b} * t^{(L)} \cdot \boldsymbol{b} * t^{(L')}}{\|\boldsymbol{b} * t^{(L)}\| \|\boldsymbol{b} * t^{(L')}\|}, \tag{A.29}$$

the square of numerator in Equation (A.28) expands as

$$(\|\boldsymbol{b} * t^{(L')}\| \boldsymbol{b} * t^{(L)} - \|\boldsymbol{b} * t^{(L)}\| \boldsymbol{b} * t^{(L')})^2 = 2\|\boldsymbol{b} * t^{(L)}\|^2 \|\boldsymbol{b} * t^{(L')}\|^2 (1 - \cos\theta) \tag{A.30}$$

Substituting the square root of Equation (A.30) into Equation (A.28), we eliminate the denominator and obtain

$$|S(\boldsymbol{b} * v, \boldsymbol{b} * t^{(L)}) - S(\boldsymbol{b} * v, \boldsymbol{b} * t^{(L')})| \leq \sqrt{2(1 - \cos\theta)} \tag{A.31}$$

Recall that $\boldsymbol{b} * t^{(L')} \in \mathcal{O}_\rho(\boldsymbol{b} * t^{(L)})$, we can bound θ by the law of sines

$$\sup_{\theta} |\sin\theta| = \sup_{\boldsymbol{b}*t^{(L')}} \frac{\|\boldsymbol{b} * t^{(L')} - \boldsymbol{b} * t^{(L)}\|}{\|\boldsymbol{b} * t^{(L)}\|} = \frac{\rho}{\|\boldsymbol{b} * t^{(L)}\|} \tag{A.32}$$

Taking supremums on both sides of Equation (A.31) and combining Equation (A.32), we complete the proof

$$\begin{aligned} & \sup_{\substack{\boldsymbol{b}*t^{(L')} \in \mathcal{O}_\rho(\boldsymbol{b}*t^{(L)}) \\ 0 \le \rho < \|\boldsymbol{b}*t^{(L)}\|}} |S(\boldsymbol{b} * v, \boldsymbol{b} * t^{(L')}) - S(\boldsymbol{b} * v, \boldsymbol{b} * t^{(L)})| \\ \le & \sup_{\theta} \sqrt{2(1 - \sqrt{1 - sin^2\theta})} \\ = & \sqrt{2(1 - \sqrt{1 - (\frac{\rho}{\|\boldsymbol{b} * t^{(L)}\|})^2})} \end{aligned}$$

□

A.2.2 Proof of Theorem 3.2

Proof. Due to Half-Angle Identities, Equation (A.31) derives as

$$|S(\boldsymbol{b} * v, \boldsymbol{b} * t^{(L')}) - S(\boldsymbol{b} * v, \boldsymbol{b} * t^{(L)})| \le 2|\sin\frac{\theta}{2}| \tag{A.33}$$

For sufficiently small θ, i.e., $\|\boldsymbol{b} * t^{(L')} - \boldsymbol{b} * t^{(L)}\| \ll \|\boldsymbol{b} * t^{(L)}\|$, we take the first-order Taylor approximation

$$2|\sin\frac{\theta}{2}| \approx |\theta| \approx |\sin\theta| = \frac{\|\boldsymbol{b} * t^{(L')} - \boldsymbol{b} * t^{(L)}\|}{\|\boldsymbol{b} * t^{(L)}\|} \tag{A.34}$$

Combining Equation (A.33) and Equation (A.34) we complete the proof. □

A.2.3 Proof of Proposition 3.3

Proof. Expanding $|\mathtt{Acc}_a^{(L)} - \mathtt{Acc}_b^{(L')}|$ by triangle inequality we have

$$\begin{aligned} |\mathtt{Acc}_a^{(L)} - \mathtt{Acc}_b^{(L')}| = \; & |\mathtt{Acc}_a^{(L)} - \mathtt{Acc}^{(L)} + \mathtt{Acc}^{(L)} - \mathtt{Acc}^{(L')} + \mathtt{Acc}^{(L')} - \mathtt{Acc}_b^{(L')}| \\ \leq \; & |\mathtt{Acc}_a^{(L)} - \mathtt{Acc}^{(L)}| + |\mathtt{Acc}^{(L)} - \mathtt{Acc}^{(L')}| + |\mathtt{Acc}^{(L')} - \mathtt{Acc}_b^{(L')}| \end{aligned} \tag{A.35}$$

Noticing that $\mathtt{Acc}^{(L)} = p_a \cdot \mathtt{Acc}_a^{(L)} + p_b \cdot \mathtt{Acc}_b^{(L)}$ and $p_a + p_b = 1$, we have

$$\begin{aligned} |\mathtt{Acc}_a^{(L)} - \mathtt{Acc}^{(L)}| = \; & p_b \cdot |\mathtt{Acc}_a^{(L)} - \mathtt{Acc}_b^{(L)}| \\ = \; & p_b \cdot \mathtt{Disp}^{(L)}(a, b) \end{aligned} \tag{A.36}$$

Similarly,

$$\begin{aligned} |\mathtt{Acc}^{(L')} - \mathtt{Acc}_b^{(L')}| = \; & p_a \cdot |\mathtt{Acc}_a^{(L')} - \mathtt{Acc}_b^{(L')}| \\ = \; & p_a \cdot \mathtt{Disp}^{(L)}(a, b) \end{aligned} \tag{A.37}$$

Substituting Equation (3.18), (A.36), and (A.37) into Equation (A.35) we complete the proof. □

A.3 Proofs for Chapter 4

A.3.1 Proof of Corollary 4.3

Proof. Without loss of generality, we assume group $z = 1$ has higher utilities than group $z = 0$, i.e.,

$$\mathbb{E}[f(x;\boldsymbol{\theta})\mathbb{1}[z=1,y=1]] \geq \mathbb{E}[f(x;\boldsymbol{\theta})\mathbb{1}[z=0,y=1]]$$

$$\mathbb{E}[f(x;\boldsymbol{\theta})\mathbb{1}[z=1,y=0]] \geq \mathbb{E}[f(x;\boldsymbol{\theta})\mathbb{1}[z=0,y=0]]$$

Equal odds indicates equal TPR and equal FPR constraints will be imposed simultaneously. Thereby

$$\begin{aligned}
S_{\mathsf{EO}} &= S_{\mathsf{TPR}} + S_{\mathsf{FPR}} \\
&= \lambda\frac{\eta}{n}\alpha_{z_i,y_i}\Theta(x_i,x_j;\boldsymbol{\theta}_0) + \lambda\frac{\eta}{n}\tilde{\alpha}_{z_i,y_i}\Theta(x_i,x_j;\boldsymbol{\theta}_0) \\
&= \lambda\frac{\eta}{n}\alpha_{z_i}\Theta(x_i,x_j;\boldsymbol{\theta}_0) && (\text{by } \alpha_z = \alpha_{z,y} + \tilde{\alpha}_{z,y}) \\
&= S_{\mathsf{DP}}
\end{aligned}$$

The third equality is due to $\alpha_z = \mathbb{1}[z=1] - \mathbb{1}[z=0] = (\mathbb{1}[z=1,y=+1] + \mathbb{1}[z=1,y=-1]) - (\mathbb{1}[z=0,y=+1] + \mathbb{1}[z=0,y=-1]) = \alpha_{z,y} + \tilde{\alpha}_{z,y}$. □

A.3.2 Proof of Corollary 4.4:

Proof. When there are only two groups, the covariance measure in Equation 4.29 reduces to

$$\hat{\phi}(f) = |\frac{1}{n}\sum_{i=1}^{n}(z_i - \frac{1}{2})f(x_i;\boldsymbol{\theta})|$$

Again, we assume group $z = 1$ is more favorable than group $z = 0$ such that

$$\mathbb{E}[f(x;\boldsymbol{\theta})\mathbb{1}[z=1]] \geq \mathbb{E}[f(x;\boldsymbol{\theta})\mathbb{1}[z=0]].$$

Then we can rewrite the above equation as

$$\hat{\phi}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}f(x_i;\boldsymbol{\theta})\mathbb{1}[z_i=1] - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}f(x_i;\boldsymbol{\theta})\mathbb{1}[z_i=0]) \geq 0$$

The above $\hat{\phi}(f)$ is saying the covariance between z and $f(x)$ is non-negative per se, so we do not need to take the absolute value of it. In other words, $\forall i, \beta_i = 1$. The final influence score of covariance thus becomes

$$S_{\mathsf{cov}}(i,j) = \lambda\frac{\eta}{2n}\Theta(x_i, x_j; \boldsymbol{\theta}_0)(\mathbb{1}[z_i=1] - \mathbb{1}[z_i=0])$$

Recall that $\alpha_i = \mathbb{1}[z_i=1] - \mathbb{1}[z_i=0]$, we conclude $S_{\mathsf{cov}}(i,j) = \frac{1}{2}S_{\mathsf{DP}}(i,j)$. We note that the connection builds upon the common assumption that group $z = 1$ has a higher utility. We can reach the same conclusion in the symmetric situation where $\mathbb{E}[f(x;\boldsymbol{\theta})\mathbb{1}[z=1]] < \mathbb{E}[f(x;\boldsymbol{\theta})\mathbb{1}[z=0]]$.

We remark, the coefficient $\frac{1}{2}$ arises from encoding the categorical sensitive variable z into $\{0,1\}$ and does not have physical meanings. If z is encoded by $\{-1,+1\}$

instead, the coefficient will be 1 such that $S_{\mathsf{DP}} = S_{\mathsf{cov}}$. This property suggests that the covariance is not a perfect measure of independence, and using mutual information is a more plausible approach. □

A.3.3 Proof of Theorem 4.5

Proof. For any t and any $\delta > 0$,

$$\begin{aligned}
\Pr(\mathcal{S}(f,j) - S(f,j) > \delta) &= \Pr(\exp\{nt\left(\mathcal{S}(f,j) - S(f,j)\right)\} > \exp\{nt\delta\}) \\
&\leq \frac{\mathbb{E}[\exp\{nt(\mathcal{S}(f,j) - S(f,j))\}]}{\exp\{nt\delta\}} \quad \text{(by Markov's inequality)} \\
&\leq \exp\left\{\frac{1}{8}nt^2C^2 - nt\delta\right\} \quad \text{(by Hoeffding's inequality)}
\end{aligned}$$

In above C is some constant. Since $\frac{1}{8}nC^2t^2 - n\delta t$ is a quadratic function regarding t, we may minimize it by taking

$$\frac{\partial}{\partial t}\left(\frac{1}{8}nC^2t^2 - n\delta t\right) = 0 \implies \frac{1}{4}nC^2t - n\delta t = 0$$

Solving the above equation, we know the quadratic function takes the minimum value at $t = \frac{\delta}{4C^2}$. Therefore,

$$\Pr\left(\mathcal{S}(f,j) - S(f,j) > \delta\right) \leq \exp\left\{-\frac{2n\delta^2}{C^2}\right\}$$

Let $\epsilon = \exp\left\{-\frac{2n\delta^2}{C^2}\right\}$, we complete the proof by substituting δ with ϵ

$$\begin{aligned}
& \Pr\left(\mathcal{S}(f,j) - S(f,j) > C\sqrt{\frac{\log\frac{1}{\epsilon}}{2n}}\right) \leq \epsilon \\
\implies \quad & \Pr\left(\mathcal{S}(f,j) - S(f,j) \leq C\sqrt{\frac{\log\frac{1}{\epsilon}}{2n}}\right) > 1 - \epsilon
\end{aligned}$$

□

A.3.4 Proof of Proposition 4.6

Proof. We visualize the considered example in Figure 4.1. The area in blue represents the false positive examples from group $z = 1$, while the area in green represents the false negative examples from group $z = 0$. The sum of blue area and green area is exactly representing the acceptance rate difference between group $z = 1$ and $z = 0$.

Recall that the model is $f(x) = w \cdot x + b$, the influence function subject to relaxed fairness constraint can be computed by Equation 4.19, i.e., $S(i, j) = k(z_i - \bar{z}) \cdot (x_i \cdot x_j + 1)$ where k is a constant coefficient corresponding to learning rate η, data size n and regularizer λ. For each individual example x_i, the overall influence score $S(i)$ consists of two components. The first component

$$\int_{x_j \in (-\infty, +\infty)} k(z_j - \bar{z}) \cdot (z_i - \bar{z}) \cdot x_i \cdot x_j \, d\Pr(x_j) = k(z_i - \bar{z}) x_i \cdot \int_{x_j \in (-\infty, +\infty)} (z_j - \bar{z}) \cdot x_j \, d\Pr(x_j)$$

is proportional to $(z_i - \bar{z})x_i$ since the integral can be treated as a constant. The second component

$$\begin{aligned} & \int_{x_j \in (-\infty, +\infty)} k(z_j - \bar{z}) \cdot (z_i - \bar{z}) \, d\Pr(x_j) \\ = \quad & k(z_i - \bar{z}) \cdot \int_{x_j \in (-\infty, +\infty)} (z_j - \bar{z}) \, d\Pr(x_j) = k(z_i - \bar{z}) \mathbb{E}[z_j - \bar{z}] \end{aligned}$$

becomes 0 due to $\mathbb{E}[z_j] = \bar{z}$. $|S(i)|$ is then proportional to $|x_i|$, thus the data examples around $x = 0$ will have smaller absolute values of influence scores.

Then we consider the classifier trained by down-weighting the data examples around $x = 0$. We show the case when $|\mu_{1,-1}| < |\mu_{0,+1}|$ in the right figure in Figure 4.1. In this case, the down-weighted negative examples from group $z = 1$ dominates the down-weighted positive examples from group $z = 0$. In consequence, the decision threshold will be perturbed towards right. Coloring the mis-classified examples again, we find out the sum of blue and green area has decreased. The case for $|\mu_{1,-1}| < |\mu_{0,+1}|$ will be symmetric. In conclusion, we demonstrate that removing training examples with smaller absolute influence scores is capable of mitigating the fairness violation. $\square$