UNIVERSITY OF CALIFORNIA

Los Angeles

Fairness–Preserving Empirical Risk Minimization

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Electrical and Computer Engineering

by

Guanqun Yang

2019

ABSTRACT OF THE THESIS

Fairness–Preserving Empirical Risk Minimization

by

Guanqun Yang
Master of Science in Electrical and Computer Engineering
University of California, Los Angeles, 2019
Professor Vwani P. Roychowdhury, Chair

The concerns regarding ramifications of societal bias targeted at a particular identity group (for example, gender or race) residing in algorithmic decision-making systems have been ever-growing in the past decade. It is a common practice of machine learning models' participation in these systems through empirical risk minimization (ERM) principle, which is often the cause of unfairness by trading off underrepresented groups for overall performance. Despite the importance of preserving fairness in such systems, there is hardly consensus in defining unified fairness metrics, designing widely-applicable bias-mitigation algorithms, and delivering interpretable models abiding by the ERM principle. The situation is made even more grievous when *non-structural* data, including text, image, and audio, is involved in these systems due to the unavailability of the well-defined identity attribute. Current approaches attempt to tackle algorithmic bias in non-structural settings from data itself and intermediate representation together with the inference component within models. In this thesis, we propose to unify all three bias-mitigation operations into one streamlined machine learning pipeline. At the same time, to provide interpretable results, the explorations will be made while carrying out debiasing procedures, and theoretical justifications will be provided accordingly. By ameliorating different bias-mitigation strategies through synergistic effects and addressing model transparency issues by investigating internal representations, we show that the proposed pipeline could provide interpretable machine learning models that embody fairness across different identity groups in numerous non-structural data settings.

The thesis of Guanqun Yang is approved.

Quanquan Gu

Junghoo Cho

Lin Yang

Vwani P. Roychowdhury, Committee Chair

University of California, Los Angeles

2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction and Literature Review

1.1 Background

The knowledge discovery process is increasingly more relying on data-driven approaches. As a result, algorithmic decision-making systems have cast significant influence on people's life due to their quantitative nature. With the rise of mobile networks, the availability of colossal dataset and ever-increasing computation power, all of a sudden, from something as small as how to distribute free admissions to a movie to more consequential ones like deciding whether to grant an individual mortgage loan, a large number of decisions are automated through data-driven protocols without even users' awareness.

However, despite the reduced manual labor, improved efficiency and user experiences, some societal concerns, including fairness and privacy with regard to these systems and corresponding decisions, arise. To name a few, the algorithms used by Goldman Sachs to approve applications for Apple cards show alarming gender inequality [Vig19]. The resume tracking system used by Amazon shows bias against the female, resulting in potentially lower admission rates when male and female applicants are both qualified [Das18]. Industrial facial recognition systems developed by Microsoft, Face++, and IBM *all* show a significant disparity in accuracy across different races, where darker skin colors could cause up to 30% performance degradation [BG18]. This list could go on while sharing the same core concern: the machine learning algorithms we interact with on a daily basis to automate decision-making *prove* to be unfair.

At the same time, as is noted in [KR19], the machine learning algorithms and underlying empirical risk minimization (ERM) principle could probably *not* satisfy requirements such as privacy and fairness for free. As the artifacts of human inventions, despite their capacity to work without being explicitly programmed, machine learning algorithms bearing empirical risk minimization (ERM) principle could *not* resolve fairness concerns without appending additional constraints other than minimizing empirical risk, or equivalently, maximizing predictive accuracy.

The machine learning community, researchers from law, economics and social sciences, and other stakeholders from the government and private sectors have recognized this challenge and initiated a series of works to tackle this problem.

1.2 Related Work

Starting from the trailblazing work by Dwork et al. in formalizing the fair machine learning problem [DHP12], researchers have previously spent extensive efforts in characterizing bias in data and predictive results, mitigating bias in machine learning models and applying these algorithms under multiple settings, which include classification, regression, and recommendation.

1.2.1 Fairness Metrics

The requisite to enforce fairness requirement to any algorithmic decision-making process is to pinpoint the component to which we believe to be critical. As is observed in [KR19], the data, learning algorithms, models given by such algorithms, and ultimate decisions (predictions) could all be questioned for potential fairness issues. However, it is not clear how learning algorithms like gradient descent could encode human bias into the learning procedure. Therefore, the categorization of fairness metrics goes into two streams.

- **Bias in learned models**
  Zhao et al. propose to use co-occurrence frequency statistics to characterize bias amplification before and after training predictive models [ZWY17]. Wang et al. refine and generalize this definition of bias amplification to a broader setting by introducing

the notion of information leakage, where they claim any model embodies bias amplification whenever leakage of sensitive attributes' information is larger than the leakage by change under similar performance measure [WZY]. These works do not specifically focus on *bias* in the machine learning pipeline but rather consider *bias amplification*. This emphasis on bias amplification shifts the responsibility of fair machine learning to the data gathering process and may better capture the dynamics of prediction.

- **Bias in data and decisions**

  Compared to the work to characterize bias in learned models, the work towards the definition of bias (or fairness) in data and decisions are much more abundant. As is noted in [CR18], these definitions could be defined statistically and individually, with the former one being practical. Some frequently used statistical definitions of fairness include demographic parity (DP), equalized odds (EO), and statistical parity [BHJ18, FSV19]. Despite empirical success in some applications [ABD18], the interpretation of these metrics could be distorted when data is imbalanced. What makes things worse, these metrics could be mutually exclusive except for degenerate cases [KMR16].

1.2.2 Bias Mitigation Strategy

Similar to the discussion of bias (fairness) metrics, current literature also focuses on mitigating bias in different components in the machine learning pipeline, which are called preprocessing, in-processing, and postprocessing, respectively [BDH18].

- **Preprocessing**

  Preprocessing algorithms is closely related to the notion of *fair representation* [CR18]. They try to apply a transformation to input data so that most of the task-specific information is maintained while removing sensitive information. For example, Louizos et al. add maximum mean discrepancy regularization to variational auto-encoder (VAE) to obtain posterior distribution that is invariant to latent variables, which preserves the utility of resulting representation while attaining some degree of fairness [LSL15].

- **In-processing**

In-processing algorithms add additional fairness constraints to existing algorithms so that the learned models could simultaneously satisfy some performance measure (for example, accuracy) and fairness requirements. The works mentioned in Section 1.2.1 both adopt this approach [WZY, ZWY17].

- **Postprocessing**

  Postprocessing maintains fairness by adjusting the predictive labels given by potentially unfair models so that the adjusted label distribution could satisfy fairness requirements. One foundational work is done by Hardt et al.. They try to adjust the output label distribution using probabilities given by the solution of a linear program, which in turn satisfies some fairness requirements [HPS16].

1.2.3 Application

- **Classification**

  Most of the current works regarding fair machine learning systems focus on the setting of classification because of the better-defined problem setup and already abundant source of literature.

  Zhang et al. try to remove or adjust the discriminatory features in the original dataset through the insights given by the causal graph. By removing or adjusting one set of nodes (features) within causal graphs, the preprocessed data empirically shows fairness under the fairness measure [ZWW17]. Zheng et al. consider the sampling bias in electronic medical records (EMR) for predicting patients' health conditions. For example, patients only have medical tests when they feel sick and go to hospitals, which could lead to misled interpretation for some chronic diseases. They apply the transformation of the original EMR time series using the HMM model and acquire unbiased data, which shows high accuracy for predicting patients' actual health conditions [ZGN17].

- **Regression**

  Most efforts of algorithmic fairness are put in the classification setting, and relatively less attention is on regression. One foundational work done by Berk et al. provides formulation and corresponding metrics for fair regression, including the explicit measure of the tradeoff between fairness and performance measure called the price of fairness (PoF). They also give benchmarks on multiple datasets, which is conducive to future research of fairness in regression [BHJ17].

- **Recommendation and matching**

  Due to the pervasive use of recommendation and matching systems, the fairness issues associated with such systems also raise researchers' awareness.

  In order to provide (mostly) gender-representative search results in talent acquisition, Geyik et al. deliver a system on LinkedIn by re-ranking the query results by recruiters to satisfy some fairness constraints without statistically harming business metrics [GAK19]. Fairness is also considered in the setting of matching submitted papers and reviewers, Kobren et al. observe that many papers are reviewed by people without sufficient qualification using the allocation given by conference organizers through the maximization of some utility. By exploiting scores assigned to each reviewer to reflect their expertise, they add additional constraints to the original optimization problem to make sure each paper is reviewed by people with sufficient relevant experiences [KSM19].

1.3 Overview of this Work

In this work, we propose to apply the adversarial training method previously used in domain adaptation to the mitigation of bias when non-structural data like images and texts come into play. Importantly, we quantitatively select the particular component of the model to apply the adversarial branch to for maximum utility. As is outlined in Figure 1.1, our main contribution is to quantitatively select the component of interest to append the adversarial branch in the original model T_θ so that a fairness-preserving model T_θ^{adv} could be attained.

Throughout the text, $\{(\mathbf{x}_i, y_i, A_i)\}_{i=1}^N$ is used for a batch of N data, where $\mathbf{x}_i$ is the feature, y_i is the associated label, and A_i stands for underlying sensitive attribute that

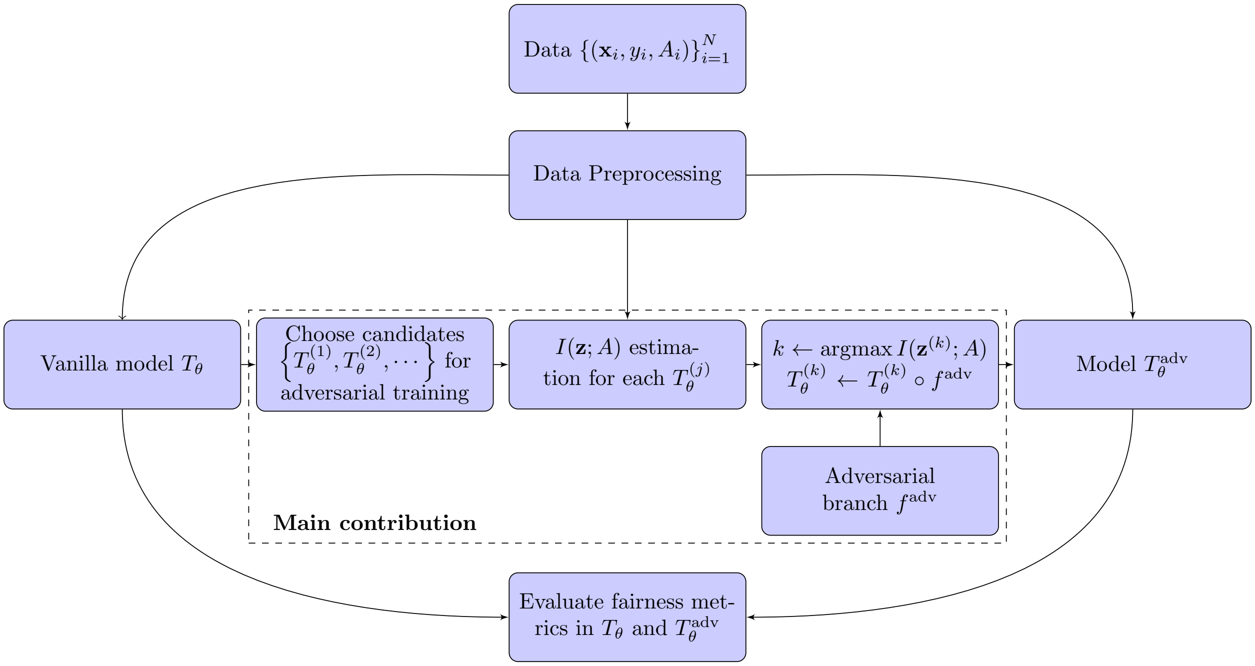

Data $\{(\mathbf{x}_i, y_i, A_i)\}_{i=1}^{N}$
Data Preprocessing
Vanilla model T_θ
Choose candidates $\left\{T_\theta^{(1)}, T_\theta^{(2)}, \cdots\right\}$ for adversarial training
$I(\mathbf{z}; A)$ estimation for each $T_\theta^{(j)}$
$k \leftarrow \text{argmax}\, I(\mathbf{z}^{(k)}; A)$
$T_\theta^{(k)} \leftarrow T_\theta^{(k)} \circ f^{\text{adv}}$
Model T_θ^{adv}
Adversarial branch f^{adv}
Main contribution
Evaluate fairness metrics in T_θ and T_θ^{adv}

Figure 1.1: Organization of the system

correspond to $\mathbf{x}_i$. When $\mathbf{x}_i$ is fed into network, vanilla network T_θ or the network T_θ^{adv} with adversarial branch f^{adv}, a representation $\mathbf{z}$ is attained in the intermediate layer. When we have representation $\mathbf{z}_i$ and attribute A_i in hand, their mutual information $I(\mathbf{z}_i; A_i)$ is computable.

This remainder of this work is organized as follows. Chapter 2 describes the metrics used in this work to objectively compare network performance in terms of fairness. Chapter 3 presents the technique used to capture the dynamics of bias during the training process. Chapter 4 discusses the bias mitigation algorithm used in this work. Chapter 5 provides comprehensive experimental evaluations in both text and image datasets in the classification setting. Finally, Chapter 6 summarizes this work and provides some pointers for further research.

CHAPTER 2

Characterizing Bias in Machine Learning Algorithms

As is noted in Chapter 1, there is no consensus in defining a unified fairness measure because of the complexity of the machine learning pipeline and many of these metrics' mutually exclusive nature. This chapter presents a novel fairness metric that empirically shows appealing properties other metrics do not have.

2.1 Demographic Parity (DP) and Equalized Odds (EO)

Demographic parity (DP) and equalized odds (EO) are two frequently used metrics in measuring the fairness of the predictive results. They both ask for the parity of different identity groups. Formally

- **Demographic Parity (DP)**

  A classifier T_θ is said to satisfy demographic parity (DP) metric when its prediction $\hat{y} = \mathbb{1}(T_\theta(\mathbf{x}) \geq 0.5)$ is independent of the attribute A. Equivalently, the value of sensitive attribute A is uncorrelated with decision $\hat{y}$.

$$\Pr\left[\hat{y} = 1 | A = 1\right] = \Pr\left[\hat{y} = 1 | A = 0\right]$$

- **Equalized Odds (EO)**

  A classifier T_θ is said to satisfy equalized odds (EO) metric when its prediction $\hat{y} = \mathbb{1}(T_\theta(\mathbf{x}) \geq 0.5)$ is conditionally independent of attribute A given the label y. This is equivalent to equal true positive rate (TPR) and false positive rate (FPR) across demographics $A \in \{0, 1\}$.

$$\Pr\left[\hat{y} = 1 | A = 1, y\right] = \Pr\left[\hat{y} = 1 | A = 0, y\right],\ y \in \{0, 1\}$$

We could see that the equalized odds (EO) metric provides a more refined formulation than demographic parity (DP) by additionally accessing the label of data. Importantly, this enables equalized odds (EO) to depend on A through y, while previously demographic parity (DP) forbids this dependence. This refinement provides ways to circumvent two major issues associated with demographic parity (DP) [HPS16]. Specifically, consider the setting of job applicants selection, where $y, \hat{y} \in \{0, 1\}$ and they stand for rejection and admission for 0 and 1, suppose binary sensitive attribute $A \in \{0, 1\}$, and feature $\mathbf{x}$ characterizes the applicant's qualification for the job, then the demographic parity (DP) shows

- **Unfairness for imbalanced demographics**
  When the distributions of $A = 0$ and $A = 1$ are skewed towards one party, demographic parity (DP) requires the company to either reject qualified applicants or admit unqualified applicants, which is not considered fair for all applicants as a whole.
- **Rejection to the optimal classifier**
  If the optimal classifier T_θ^{opt} is attained, it evaluates the applicant's qualification $\mathbf{x}$ correctly with probability 1 and gives the corresponding admission decision $\hat{y}$. It is not clear how classifier T_θ^{opt} is unfair under the demographic parity (DP) metric.

However, despite its merits over demographic parity (DP), as the following example shows, equalized odds (EO) metric is not without its flaws.

Proceeding with the setting of job applicant selection, when evaluating the fairness of predictive results, one could compute the violation of individual metric.

$$\nu_{\text{DP}} = \Big| \Pr\big[\hat{y} = 1 | A = 1\big] - \Pr\big[\hat{y} = 1 | A = 0\big] \Big|$$

$$\nu_{\text{EO}} = \sum_{y \in \{0,1\}} \Big| \Pr\big[\hat{y} = 1 | A = 1, y\big] - \Pr\big[\hat{y} | A = 0, y\big] \Big|$$

Suppose the classifier T_θ returns probability following the distribution $\mathcal{N}(\mu, 0.03^2)$ with base $\mu_1 = 0.2, \mu_2 = 0.6$ and potential shifts depicted in Figure 2.1. An accurate classifier could correctly distribute scores above and below threshold, which is often set to 0.5 by default. As is shown in Figure 2.1, given labels of $\mathbf{x}$ (denoted ”positive” and ”negative” in the figures), when probability $T_\theta(\mathbf{x})$ is distributed differently, the corresponding fairness violation ν_{DP} and ν_{EO} could *not* fully capture the fairness of output. Specifically,

- **Failure when optimal classifier exists**

  When there exists a threshold (potentially differing from 0.5) that could correctly classify all samples, the optimal classifier T_θ^{opt} is attained. However, both demographic parity (DP) and equalized odds (EO) signal unfairness.

- **Failure to account for different distributions**

  The fairness violations for distributions except ”low data separability below threshold” are very close for both demographic parity (DP) and equalized odds (EO). A preferred metric should provide more distribution-specific information other than fairness violation.

Note for distribution ”inseparable data below threshold”, even though both fairness violation is 0, this does not show the flaws of either of the two metrics since they should be 0 by definition.

However, except for the two issues mentioned above, additional subtleties arise when there are additional complications associated with data.

- **Failure to take missing value into account**

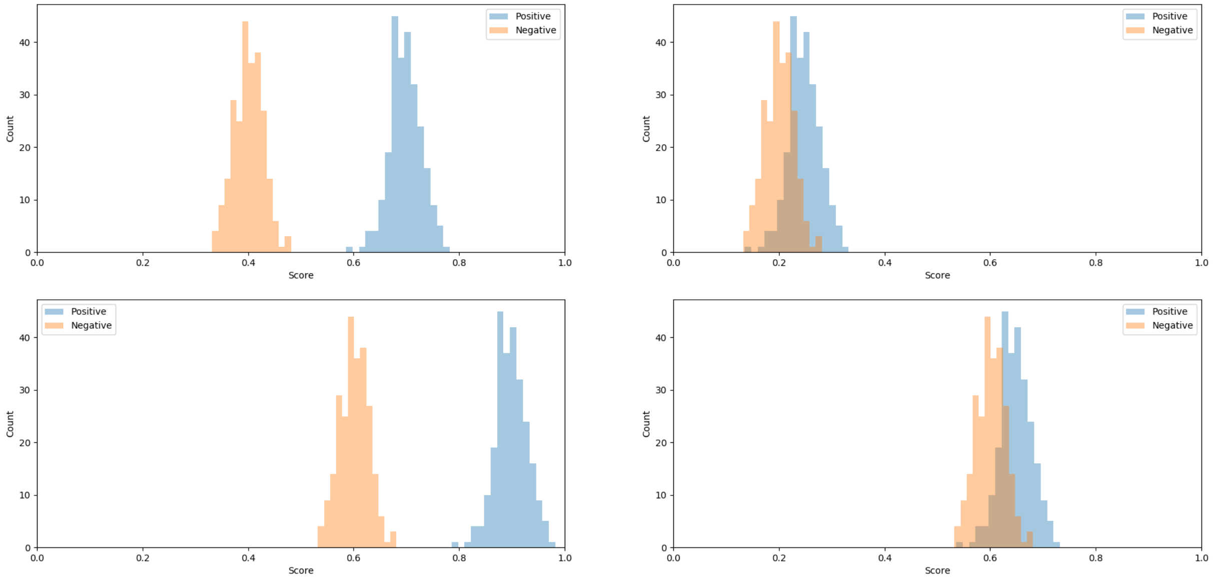
  Suppose we would like to train a fair text classifier with respect to attribute A. We could not anticipate each of the gathered texts includes A, which makes A a missing value. At the same time, both the computations of demographic parity (DP) and equalized odds (EO) require access to A. In this case, the unavailability of A makes neither of them computable.

- **Failure to extend to the case where multiple A's are involved**

  The definitions of both demographic parity (DP) and equalized odds (EO) assume that only one A needs to be addressed in the dataset. However, it is no surprise that some datasets could include gender, race, and age simultaneously (for example, adult income dataset shown in [Koh96]). It is not clear how definitions of demographic parity (DP) and equalized odds (EO) could be extended to this case.

Table 2.1: Violation of demographic parity (DP) and equalized odds (EO)

	Small right shift ($\mu_1 = 0.4,\ \mu_2 = 0.7$)	Large right shift ($\mu_1 = 0.6,\ \mu_2 = 0.9$)	Inseparable data below threshold ($\mu_1 = 0.2,\ \mu_2 = 0.25$)	Inseparable data above threshold ($\mu_1 = 0.6,\ \mu_2 = 0.65$)
ν_{DP}	0.06800	0.06400	0.00000	0.06400
ν_{EO}	0.06800	0.07200	0.00000	0.07200



(a) Data distribution with subgroup shift (up: small shift, down: large shift)

(b) Data distribution with inseparable subgroup (up: inseparable subgroup above 0.5, down: inseparable subgroup below 0.5)

Figure 2.1: Four typical data distributions without missing value

2.2 AUC-based Metric

In hope of circumventing issues presented in Section 2.1, a novel AUC-based metric is proposed in [BDS19], which borrows some aspects of previous metrics while maintaining its validity when the dataset shows skewness and missing values. At the same time, the extension of this metric is natural when multiple A's come into play.

For a binary classification problem, suppose there are I identity groups in the dataset D where each group is denoted as $g_i, i \in [I]$, then for one particular identity group m, define

- **BPSN AUC** (Background Positive Subgroup Negative AUC)

$$\text{BPSN AUC} = \text{AUC}(D^+ + D^-_{g_m}),\ D^+ = D \backslash D^+_{g_m}$$

- **BNSP AUC** (Background Negative Subgroup Positive AUC)

$$\text{BNSP AUC} = \text{AUC}(D^- + D^+_{g_m}),\ D^- = D \backslash D^-_{g_m}$$

- **Subgroup AUC**

$$\text{Subgroup AUC} = \text{AUC}(D^+_{g_m} + D^-_{g_m})$$

- **Overall AUC**

$$\text{Overall AUC} = \text{AUC}(D^+ + D^-)$$

Since there are I identity groups, in order to integrate metrics from all subgroups, Borkan et.al. use geometric mean $(\frac{1}{I}\sum_{i=1}^{I} m_i^p)^{\frac{1}{p}}$ to all metrics except overall AUC [BDS19]. Formally, the metric we use to evaluate our system is defined as

$$\frac{1}{4}\left[\text{Overall AUC} + (\frac{1}{I}\sum_{i=1}^{I}\text{BPSN AUC}_i)^{\frac{1}{p}} + (\frac{1}{I}\sum_{i=1}^{I}\text{BNSP AUC}_i)^{\frac{1}{p}} + (\frac{1}{I}\sum_{i=1}^{I}\text{Subgroup AUC}_i)^{\frac{1}{p}}\right]$$

The distributions in Figure 2.2 follow the setting in Section 2.1, the difference is that the base $\mu_1 = 0.2, \mu_2 = 0.6$ here correspond to background distributions, where A is missing. The subgroups are samples with A but shift with respect to background distributions. As could be seen in Table 2.2, the limitations of demographic parity (DP) and equalized odds (EO) are mostly overcome. Specifically,

- **Threshold adaptability**

  When there exists a decision threshold that could make the correct classification (for distribution ”small right shift” and ”large right shift”), the AUC-based metric shows high confidence in providing correct classification.

- **Natural extension with missing value**

  As is shown in Figure 2.2, the AUC-based metric is designed to account for missing values through background distributions D^+ and D^-. However, it is worthwhile to note that the AUC-based metric is still valid when there are no missing values. This makes it comparable with demographic parity (DP) and equalized odds (EO) metrics discussed in Section 2.1.

- **Distribution awareness**

  Contrary to fairness violation shown in Table 2.1, the results shown in Table 2.2 show disparity across different distributions, which confirm the awareness of AUC-based metric to distributions.

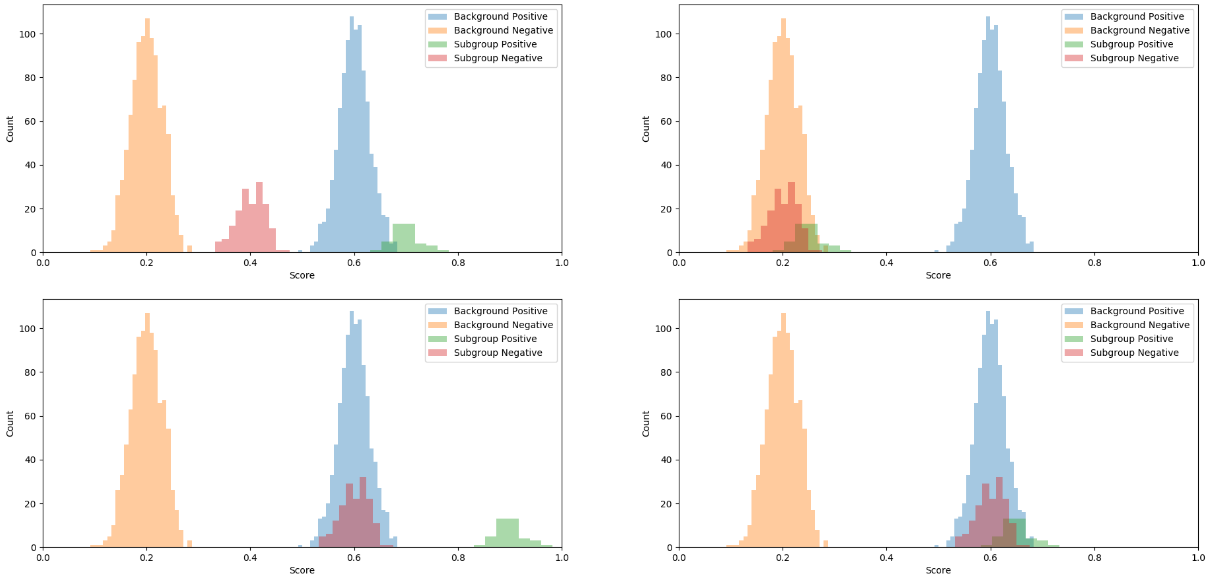
- **Direct extensibility with multiple A's**

  According to the formulation, the AUC-based metric could incorporate any number of A's into computation through g_i.

2.3 Case Study: Empirical Comparison of Fairness Metrics

In order to demonstrate the ability to capture unfairness in real-world settings, this section provides an empirical comparison of these metrics on adult income dataset [Koh96], a dataset with multiple sensitive attributes A's that is frequently used in the research of algorithmic fairness.

The goal of the adult income dataset is to predict whether an individual could have an annual income of more than 50 thousand dollars based on features like occupation, education, investment decision, and many others. Importantly, this dataset includes multiple sensitive attributes, including gender, race, and age. The prediction made from this dataset



(a) Data distribution with subgroup shift (up: small shift, down: large shift)

(b) Data distribution with inseparable subgroup (up: inseparable subgroup above 0.5, down: inseparable subgroup below 0.5)

Figure 2.2: Four typical data distributions with missing values

Table 2.2: AUC-based metric under different distributions

Distribution	Metric
Small right shift $(\mu_1 = 0.4,\ \mu_2 = 0.7)$	0.99963
Large right shift $(\mu_1 = 0.6,\ \mu_2 = 0.9)$	0.83886
Inseparable data below threshold $(\mu_1 = 0.2,\ \mu_2 = 0.25)$	0.86492
Inseparable data above threshold $(\mu_1 = 0.6,\ \mu_2 = 0.65)$	0.83886

Table 2.3: Predictive results on adult income dataset

Label	$y = 1$				$y = 0$			
Prediction	$\hat{y} = 1$		$\hat{y} = 0$		$\hat{y} = 1$		$\hat{y} = 0$	
Gender	Male	Female	Male	Female	Male	Female	Male	Female
Count	1202	165	771	151	437	60	3749	2510
Accuracy	$\frac{1202+165+3749+2510}{9045} = 84.31\%$							

Table 2.4: Statistics of adult income dataset with respect to gender

(a) Full dataset

Annul income	< 50K	≥ 50K
Male	20988	9539
Female	13026	1669

(b) Validation dataset

Annul income	< 50K	≥ 50K
Male	4189	1973
Female	2570	316

may be used in applications from seemingly innocuous targeted advertising to a more consequential approving mortgage loan, thus causing unfairness when classifiers fail to attain 100% accuracy.

The statistics of the label with respect to gender is shown in Table 2.4. When a Logistic regression classifier is trained with 80% of the data, the prediction results on the 20% hold-out set achieve 84.31% accuracy (see Table 2.3). However, as is shown in Table 2.5, the prediction result shows unfairness from high DP violation, EO violation, and low AUC-based metric. Furthermore, compared to DP violation and EO violation, the AUC-based metric provides a more nuanced view about this unfairness. For example, for the ”Female” subgroup, the lower BPSN AUC and BNSP AUC than the ”Male” subgroup show that the classifier has difficulty classifying samples with different gender attributes.

Table 2.5: Comparison of fairness metrics of predictions made on adult income dataset

(a) Comparison of metrics

Metric	Value
DP violation ν_{DP}	0.156339
EO violation ν_{EO}	0.156339
AUC-based metric	0.210551

(b) Nuanced statistics of AUC-based metric

Subgroup	Male	Female
BPSN AUC	0.326572	0.178786
BNSP AUC	0.365101	0.161898
Subgroup AUC	0.161898	0.365101
Overall AUC	0.268112	
Final metric	0.210551	

CHAPTER 3

Measuring Bias Dynamics through Mutual Information

Chapter 2 discuses the measurement of bias in the final predictive results. However, the metrics introduced in Chapter 2 could not provide insights into the expression of bias during the training process. This chapter shows that this dynamic could be captured using mutual information.

3.1 Mutual Information

Mutual information captures the amount of information one random variable has about another random variable or equivalently, the reduction of uncertainty of one random variable when another random variable is given. As is shown below, zero mutual information between two random variables indicates independence and vice versa.

$$I(X;Y) = H(X) - H(X|Y) = 0 \Leftrightarrow H(X) = H(X|Y) \Leftrightarrow X \perp\!\!\!\perp Y$$

The use of mutual information as a measure to characterize bias dynamic naturally arise from the demographic parity (DP) introduced from Chapter 2. It requires statistical independence between sensitive attribute A and predictive results.

$$\Pr\left[\hat{y} = 1 | A = 1\right] = \Pr\left[\hat{y} = 1 | A = 0\right]$$

However, as is noted in Chapter 2, there are flaws with this notion of fairness as it might not guarantee fairness under various settings. We instead ask for the reduction of information leakage about sensitive attribute A through the access to intermediate representation $\mathbf{z}$ given by machine learning models. In the case of the neural network, the intermediate representation $\mathbf{z}$ comes from each layer.

3.2 Mutual Information Neural Estimation

The estimation of mutual information between intermediate representation $\mathbf{z}$ and sensitive attribute A involves (potentially) high dimensional vector. The accuracy of this estimation could not be guaranteed since many of the estimation algorithms are designed for relatively low dimensional input.

As is noted in [BBR], estimating mutual information between high dimensional random vectors X and Z could be reduced to the maximization of the KL divergence lower bound between joint distribution P_{XZ} and produce of their marginal distributions $P_X \otimes P_Z$

$$\left.\begin{aligned} &I(X;Z) = D_{KL}(P_{XZ}||P_X \otimes P_Z) \\ &D_{KL}(P||Q) \geq \sup_{T \in \mathcal{F}} \mathbb{E}_P[T] - \log \mathbb{E}_Q[e^T] \end{aligned}\right\} \rightarrow I(X;Z) \geq \underbrace{\sup_{\theta \in \Theta} \mathbb{E}_{P_{XZ}}[T_\theta] - \log \mathbb{E}_{P_X \otimes P_Z}[e^{T_\theta}]}_{\text{maximization of lower bound}}$$

where the function class $\mathcal{F}$ is replaced with neural network T parametrized by $\theta \in \Theta$ in this work.

In practice, P_{XZ} and $P_X \otimes P_Z$ are replaced with their empirical counterparts $P_{XZ}^{(n)}$ and $P_X^{(n)} \otimes P_Z^{(n)}$, which are constructed by mini-batch samples $\left\{(\mathbf{z}_i, A_i, y_i)\right\}_{i=1}^N$. Given optimal parameter $\hat{\theta}$, the mutual information between X and Z is

$$I(X;Z) \approx \widehat{I(X;Z)} = \mathbb{E}_{P_{XZ}^{(n)}}[T_{\hat{\theta}}] - \log \mathbb{E}_{P_X^{(n)} \otimes P_Z^{(n)}}[e^{T_{\hat{\theta}}}]$$

The architecture used to compute $\widehat{I(X;Z)}$, namely `StatisticsNetwork` shown in Table 3.1, is chosen to be consistent with the one presented in [BBR]. The Protocol 1 describes the estimation process using `StatisticsNetwork`. Specifically,

1. **Preprocessing**

   As the input tensor $\mathbf{z}$ greatly increase in dimension after it is flattened. We apply PCA to decrease the dimension of flattened tensor $\tilde{\mathbf{z}}$ while keeping at 95% of its information. Empirically, an image of $(3, 224, 224)$ transformed by a convolution layer of 128 filters is turned into a tensor of shape $(128, 112, 112)$. The flattened tensor is more than 1.6 million dimensional while the same vector after reduction is only several hundred dimensional. This operation largely improves the runtime of mutual information evaluation with marginal performance tradeoff.

2. **Estimation loop**

   A batch of N samples constitute empirical marginal distributions of $\{\mathbf{z_i}\}_{i=1}^N$ and $\{A_i\}_{i=1}^N$ individually and an empirical joint distribution of $\{\mathbf{z}_i, A_i\}_{i=1}^N$. In i-th of total M estimation epochs, b samples are drawn of three aforementioned distributions, and they are used to approximate the lower bound

$$\frac{1}{b}\sum_{i=1}^{b} T_\theta(\tilde{\mathbf{z}}^{(i)}, A^{(i)}) - \log \frac{1}{b}\sum_{i=1}^{b} e^{T_\theta(\bar{\mathbf{z}}^{(i)}\bar{A}^{(i)})}$$

   which is recorded in $\mathbf{v}_i$. The standard gradient *ascent* procedure is then used to improve this lower bound. After sufficiently many epochs (by setting M a large number like 10000), the moving average of $\mathbf{v}$ is output as final mutual information estimation.

In order to verify the correctness of our implementation, we utilize the Protocol 1 to compute mutual information between Gaussian random variable x_1 and x_2 in $\mathbf{x} = \left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right] \sim \mathcal{N}(\left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right], \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$. As is shown in Figure 3.1, the estimation given by our approach is consistent with theoretical value $-\frac{1}{2}\log(1-\rho^2)$ and results given by the KSG estimator, which is another mutual information estimation algorithm commonly used in low dimensional setting [GOV].

Table 3.1: StatisticsNetwork architecture

Operation	Output dimension	Activation
Input $[\tilde{\mathbf{z}}; A]$	(N, k)	-
FC	$(N, 512)$	ELU
FC	$(N, 512)$	ELU
FC	$(N, 1)$	-

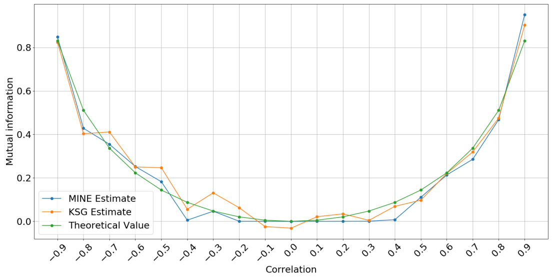


Figure 3.1: Mutual information estimate between two dimensional correlated Gaussian with varying correlation ρ

Protocol 1 Mutual Information Neural Estimation (MINE)

```
procedure MINE($\big\{(\mathbf{z}_i, A_i)\big\}_{i=1}^{N}, M, b$)
    $\tilde{\mathbf{z}} \leftarrow \text{Flatten}(\mathbf{z}_{i=1}^{N})$        ▷ $(N, d_1, d_2, \cdots) \rightarrow (N, \prod_i d_i)$
    $\tilde{\mathbf{z}} \leftarrow \text{PCA}(\tilde{\mathbf{z}})$        ▷ $(N, \prod_i d_i) \rightarrow (N, k)$
        ▷ reduce $\tilde{\mathbf{z}}$ to the dimension $k$ that attains 95% explained variance
        ▷ generally $\prod_i d_i >> k$
    $\mathbf{v} \leftarrow \mathbf{0} \in \mathbb{R}^M$
    for $i \leftarrow 1 : M$ do
        $N$ mini-batch samples constitute empirical distribution $P_{ZA}^{(N)}$, $P_{Z}^{(N)}$ and $P_{A}^{(N)}$.
        Draw $b$ samples out of $N$ mini-batch samples (empirical distributions).
            $(\tilde{\mathbf{z}}^{(1)}, A^{(1)}), \cdots, (\tilde{\mathbf{z}}^{(b)}, A^{(b)}) \sim P_{ZA}^{(N)}$
            $\bar{\mathbf{z}}^{(1)}, \cdots, \bar{\mathbf{z}}^{(b)} \sim P_{Z}^{(N)}, \bar{A}^{(1)}, \cdots, \bar{A}^{(b)} \sim P_{A}^{(N)}$
        Record the mutual information estimation at iteration $i$ in $\mathbf{v}$
            $\mathbf{v}_i \leftarrow T_\theta([\tilde{\mathbf{z}}; A])$
        Evaluate the lower bound and its gradient
            $\mathcal{V}(\theta) \leftarrow \frac{1}{b}\sum_{i=1}^{b} T_\theta(\tilde{\mathbf{z}}^{(i)}, A^{(i)}) - \log \frac{1}{b}\sum_{i=1}^{b} e^{T_\theta(\bar{\mathbf{z}}^{(i)}\bar{A}^{(i)})}$
            $\nabla_\theta \mathcal{V}(\theta)$
        Apply bias correction to gradient
            $\tilde{\nabla}_\theta \mathcal{V}(\theta) \leftarrow \nabla_\theta \mathcal{V}(\theta)$
        Update parameter for network $T_\theta$
            $\theta \leftarrow \theta + \tilde{\nabla}_\theta \mathcal{V}(\theta)$
    end for
    return MovingAverage($\mathbf{v}$)
end procedure
```

CHAPTER 4

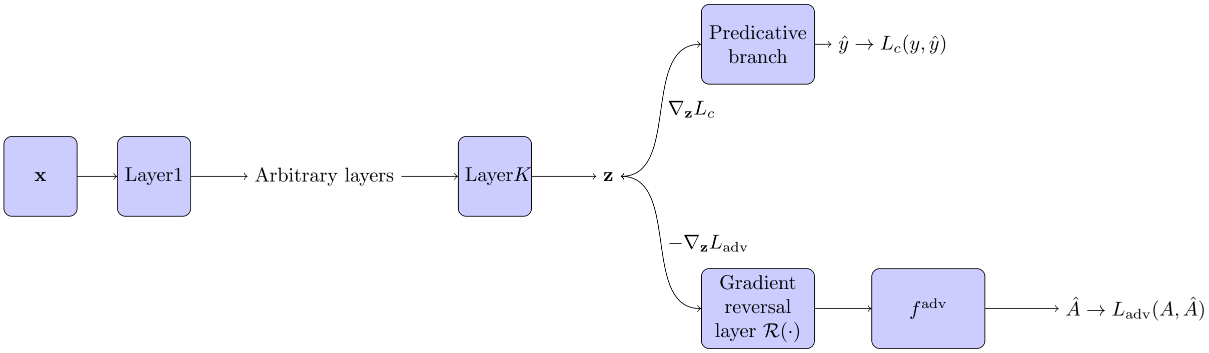
Mitigating Bias through Adversarial Empirical Risk Minimization

Chapter 2 and 3 provide tools to measure the bias in predictive results and their dynamics during the training process. In the hope of improving the eventual bias measure, this chapter shows the method to mitigate bias led by the insights of bias dynamics.

4.1 Adversarial Training of Neural Networks

Adversarial empirical risk minimization is first proposed in the study of domain adaptation as a way to approximate $\mathcal{H}$-divergence, which characterizes the amount of difference between the source data distribution $\mathcal{D}_S^X$ and target data distribution $\mathcal{D}_T^X$ captured by a particular hypothesis class $\mathcal{H}$. Formally, it is defined as

$$d_{\mathcal{H}} = 2 \sup_{h \in \mathcal{H}} \left| \Pr_{\mathbf{x} \sim \mathcal{D}_S^X} \big[h(\mathbf{x} = 1)\big] - \Pr_{\mathbf{x} \sim \mathcal{D}_T^X} \big[h(\mathbf{x}) = 1\big] \right|$$



$\mathbf{x}$
Layer1
Arbitrary layers
Layer$K$
$\mathbf{z}$
$\nabla_{\mathbf{z}} L_c$
$-\nabla_{\mathbf{z}} L_{\text{adv}}$
Predicative branch
$\hat{y} \rightarrow L_c(y, \hat{y})$
Gradient reversal layer $\mathcal{R}(\cdot)$
f^{adv}
$\hat{A} \rightarrow L_{\text{adv}}(A, \hat{A})$

Figure 4.1: Adversarial training scheme of neural network

In order for a model trained on the source domain to generalize well on the target domain, the $d_{\mathcal{H}}$ has to be small [BBC07]. Ganin et al. argue that, by applying adversarial training of the neural network, it could learn the representation that is indicative of both source and target domain while being indiscriminate across domains [GUA]. This adversarial training scheme provides empirically pleasing results and has successful applications in machine translation, text classification, and image classification [WXZ17], where distributional shift and sparsity of available data are often the concerns. As is shown in Figure 4.1, the backbone of this scheme is the novel gradient reversal layer appended at the beginning of the auxiliary branch that predicts the auxiliary label. This architecture tries to acquire a representation that is simultaneously

- Indicative to both source and target domain.

  Since the gradient flow within the auxiliary branch follows the conventional backward propagation protocol. After the negated gradient from the adversarial branch meets and joins gradient flow in the predictive branch by addition. The learned representation could still be informative to the target domain. With the combined signal transmitted through the entire network through backward propagation, this weakens the learning signal in the predictive branch for predicting domain labels while empowering the overall predictive capability for target labels in the network as a whole.

- Indiscriminate across domains.

  The adversarial empirical risk minimization could be seen as the modification of multitask learning [Rud17], where labels of related tasks (for example, sentiment analysis and POS tagging in natural language processing) are jointly predicted on different branches to achieve synergistic effects for individual task's performance. With the only difference in the gradient reversal layer, the adversarial branch reverses this synergy and tries to reduce the discriminative power of learned representation.

Formally, the forward and backward gradient reversal layer $\mathcal{R}(\cdot)$ is defined as mutually incompatible operations

$$\mathcal{R}(\mathbf{z}) = \mathbf{z},\ \nabla_{\mathbf{z}}\mathcal{R} = -\mathbf{I}$$

where conventional $\nabla_{\mathbf{z}}\mathcal{R}$ should be $\mathbf{I}$ rather than $-\mathbf{I}$.

4.2 Adversarial Empirical Risk Minimization Inspired by Mutual Information

Recently, many researchers notice the potential of adversarial training in mitigating bias in numerous tasks [AZP19, BCZ17, WZY, ZLM, AVG19]. However, most of these works follow the original setup proposed in [GUA] and locate the adversarial branch (often a multilayer perceptron) at the output layer. From the information-theoretic point of view, this default choice is not backed by strong theoretical justification. Furthermore, as is evidenced by observation in [DBC], the information expressed in the neural network does not follow a monotonic fashion. At the same time, there is the concern of vanishing gradient in the deep neural network, which motivates multiple auxiliary branches in Inception architecture [SVI16].

Therefore, we propose to learn from insights given by the mutual information between intermediate representation $\mathbf{z}$ and attribute A (and target y). It is then possible to avoid the choice of adversarial branch's location in an ad-hoc fashion. In order to quantitatively compare the performance of our optimal location and default one, we follow the original setup in [GUA] by choosing a simple multilayer perceptron as the architecture of the adversarial branch (details of this architecture could be found in Chapter 5).

As is shown in Protocol 2, for each individual task, every component of interest $T_\theta^{(j)}$ in vanilla model T_θ is explored for the mutual information between representation $\mathbf{z} = T_\theta^{(j)}(\mathbf{x})$ and attribute A. Then the component with maximum mutual information is chosen as the candidate to append the adversarial branch. The resulting network T_θ^{adv} is chosen to perform specific task.

Protocol 2 Training Protocol

```
**Hyperparameters**: learning rate $\alpha$, batch size $N$, epoch $K$
procedure TRAIN($\left\{(\mathbf{x}_i, A_i, y_i)\right\}_{i=1}^{S}$)
    Initialize network without adversarial branch $T_\theta$
    repeat
        Invoke Protocol 1 to each $T_\theta^{(j)}$ in $T_\theta$ and acquire representation $\mathbf{z}$.
        return $\widehat{I(\mathbf{z}; A)}$
    until all components $\left\{T_\theta^{(1)}, T_\theta^{(2)} \cdots\right\}$ of interests in $T_\theta$ are evaluated
    Choose the component $T_\theta^{(j)}$ to apply adversarial branch to based on $\widehat{I(\mathbf{z}; A)}$

    Initialize network with adversarial branch $T_\theta^{\text{adv}}$
    for $1 : K$ do
        for $1 : N$ do
            Forward propagation
    $\{\hat{y}_i\}_{i=1}^{N}, \left\{\hat{A}_i\right\}_{i=1}^{N} \leftarrow T_\theta^{\text{adv}}(\{\mathbf{x}_i\}_{i=1}^{N}) \begin{cases} L_c(y, \hat{y}) \leftarrow \text{CrossEntropyLoss}(\{y_i\}_{i=1}^{N}, \{\hat{y}_i\}_{i=1}^{N}) \\ L_{\text{adv}}(A, \hat{A}) \leftarrow \text{CrossEntropyLoss}(\{A_i\}_{i=1}^{N}, \left\{\hat{A}_i\right\}_{i=1}^{N}) \end{cases}$
            Backward propagation
                $\theta \leftarrow \text{Adam}(L_c, L_{\text{adv}}, \theta; \alpha)$
        end for
    end for
    return Fairness-preserving network $T_\theta^{\text{adv}}$
end procedure
```

CHAPTER 5

Experiments

We evaluate our adversarial empirical risk minimization (ERM) approach with text and image datasets under the classification setting. The architecture for text and image datasets are chosen differently, which aims to show the wide applicability of our method.

5.1 Experiment Setup

5.1.1 Datasets

The text dataset we use is Jigsaw Toxic Comment Dataset [DLS18]. Applied in the environment of toxicity comment detection in online forums, the collection of this dataset is motivated by the observation that the vanilla predictive model could incorrectly associate high toxicity level with particular identities that are historically underrepresented. For example, as is shown in Table 5.1a, the *non-toxic* comments that share the body ”I am a _____ man/woman” are assigned wide range of toxicity scores, revealing the propensity of discriminating identity groups that could be characterized as ”black”, ”female”, and ”gay”.

The statistics of this dataset and some text samples are shown in Table 5.1b and Table 5.1d. For each individual identity, there are consistently more non-toxic samples than toxic ones in the dataset.

The image dataset we use is the SCUT-FBP5500 dataset [LLJ18]. This choice is motivated by the report that many researchers have observed that the error of commercial face recognition systems is strongly correlated with the darkness of skin color [BG18]. We instead seek to extend this study and explore a similar application in evaluating facial attractiveness

Table 5.1: Overview of dataset

(a) Predictive toxicity of vanilla model

Sentence	Toxicity Score
I am a man	20%
I am a woman	41%
I am a white man	66%
I am a white woman	77%
I am a black man	80%
I am a black woman	85%
I am a gay man	57%
I am a gay woman	66%
I am a gay white man	78%
I am a gay white woman	80%
I am a gay black man	82%
I am a gay black woman	87%

(b) Statistics of Jigsaw dataset

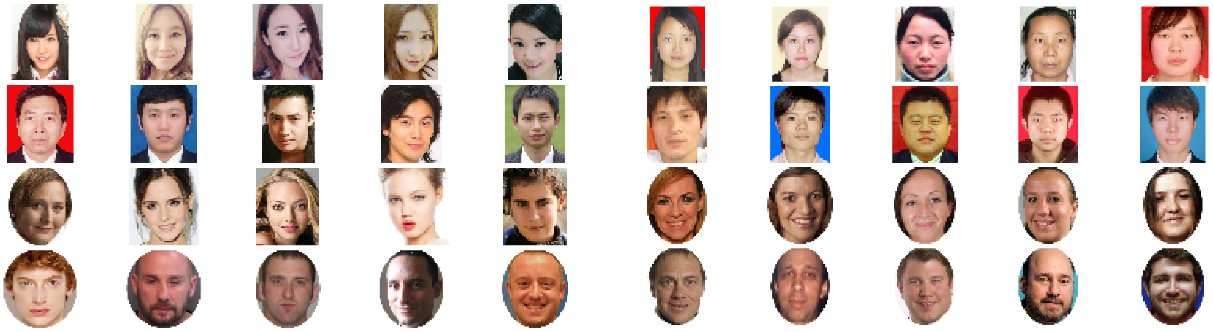
Identity	Non-Toxic	Toxic
Male	37799	6685
Female	46118	7311
Black	10223	4678
White	18044	7038

(c) Statistics of SCUT-FBP5500 dataset

Identity	Non-Attractive	Attractive
Asian male	1000	1000
Asian female	1000	1000
Caucasian male	375	375
Caucasian female	375	375

(d) Sample of some toxic/non-toxic comment texts

Toxic	Non-toxic
Corrupt hypocrites throughout the government. Of the money, by the money and for the money. Brought to you by the jesus freaks cause birds of a feather flock together.	Jeff Sessions is another one of Trump's Orwellian choices. He believes and has believed his entire career the exact opposite of what the position requires.
If it walks like a duck, and quacks like a duck....	That's already been happening, Carl, it's called Fake News.
Fool.	Did Mark Shore lose his job? I have not seen his guff for quite a while now.



(a) Positive samples

(b) Negative samples

Figure 5.1: Samples of SCUT-FBP5500 dataset

and the system performance with respect to both gender and race. The statistics of this dataset are shown in Table 5.1c.

5.1.2 Models

The network architecture in text classification is a three layer convolutional neural network, which is outlined in Table 5.3. The architecture used for image classification is ResNet18 [HZR16], which is described in Table 5.4.

5.2 Results

5.2.1 Jigsaw Toxic Comment Dataset

As is shown in Figure 5.4, the $I(\mathbf{z}_i; A),\ i \in \{1, 2, 3\}$ fluctuates above 0 and no consistent pattern could be observed from the relationship between training progress and mutual information. However, there is indeed tendency that mutual information estimate converge to particular value with increasing number of epochs.

The change of mutual information across layers in the last training epoch is shown in Table 5.5 and Figure 5.2. Because of their maximum mutual information across all layers, the layer two and layer three will be chosen to append adversarial branch f^{adv} to attain model T_θ^{adv} for gender and race, respectively.

Table 5.2: The network architecture for adversarial branch

(a) Image classification adversarial branch

Operation	Output dimension	Activation
Input $\mathbf{z}$	(N, C, H, W)	-
AdaptiveAvgPool	$(N, C, 7, 7)$	-
Flatten	$(N, 49C)$	-
Linear	$(N, 100)$	LeakyReLU
Linear	$(N, 100)$	LeakyReLU
Linear	$(N, 2)$	-

(b) Text classification adversarial branch

Operation	Output dimension	Activation
Input $\mathbf{z}$	(N, L, C)	-
Flatten	(N, LC)	-
Linear	$(N, 100)$	LeakyReLU
Linear	$(N, 100)$	LeakyReLU
Linear	$(N, 2)$	-

Table 5.3: Text classification network architecture.

Operation	Output dimension	Activation
Input **x**	(N, L)	-
Embedding	$(N, L, 50)$	-
Layer1		
CONV, POOL, BN	$(N, 128, 50)$	ReLU
Layer2		
CONV, POOL, BN	$(N, 128, 50)$	ReLU
Layer3		
CONV, POOL, BN	$(N, 128, 50)$	ReLU
Flatten	$(N, 128 \times 50)$	-
Linear	$(N, 128)$	ReLU
Dropout	$(N, 128)$	-
Linear	$(N, 2)$	-

Table 5.4: Image classification network architecture

(a) ResNet18

Operation	Output dimension	Activation
Input $\mathbf{x}$	$(N, 3, 224, 224)$	-
Preprocessing		
CONV	$(N, 64, 112, 112)$	-
BN	$(N, 64, 112, 112)$	ReLU
MaxPool	$(N, 64, 56, 56)$	-
Layer1		
ResidualBlock×2	$(N, 64, 56, 56)$	-
Layer2		
ResidualBlock×2	$(N, 128, 28, 28)$	-
Layer3		
ResidualBlock×2	$(N, 256, 14, 14)$	-
Layer4		
ResidualBlock×2	$(N, 512, 7, 7)$	-
Output		
AdaptiveAvgPool	$(N, 512, 1, 1)$	ReLU
Flatten	$(N, 512)$	-
FC	$(N, 2)$	-

(b) Residual block

Input $\mathbf{z}$	
Shortcut	CONV
	BN
	ReLU
	CONV
	BN
Output $\mathbf{z} + f(\mathbf{z})$	

The results of T_θ and T_θ^{adv} are shown in Table 5.7. The improvements in the final metric after applying adversarial training in gender and race are 2.64% and 2.93%, respectively. Additionally, we could observe the following

- **Improvement of individual metric after applying adversarial training**

  When applying adversarial training over gender, there is *consistent* improvement for identity ”Male” and ”Female” for all individual metrics, including BPSN AUC, BNSP AUC, and subgroup AUC. Comparatively, when adversarial training is applied to race, even though the improvement is not consistent for all cases, there are indeed improvements for individual metrics except BPSN AUC for identity ”White” and BNSP AUC for identity ”Black”.

- **Overall performance is not traded off for fairness**

  After applying adversarial training, there is *no* drop in overall AUC, which shows that performance degradation is not associated with applied adversarial training branch.

Table 5.5: Change of $I(\mathbf{z}; A)$ with respect to network depth in text classification

Layer	Target	Gender	Race
1	0.01037	0.00734	0.00277
2	0.00000	0.00964	0.00293
3	0.00777	0.00798	0.00403

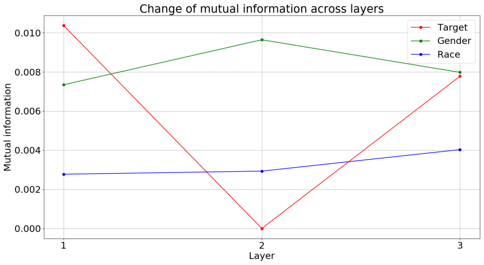


Figure 5.2: Change of $I(\mathbf{z}; A)$ with respect to network depth in text classification

5.2.2 SCUT-FBP5500 Dataset

The Figure 5.4 shows the changes of mutual information $I(\mathbf{z}_i; A)$ $(i = 1, 2, 3, 4)$, with respect to number of epochs. Unlike previous experiment with text dataset, patterns are evident for us to draw several insights.

- **Non-monotonicity**

None of the mutual information change in a monotonic fashion. This confirms the result reported in [DBC], where Dhar et al. also find similar non-monotonic changes of mutual information.

- **Synchronicity and asynchronicity**

  The mutual information estimates between target and representation across layers change in an almost synchronous fashion, while for latent variables (race and gender), this synchronicity is not evident. This disparity shows that representation $\mathbf{z}_i$ is more informative about the label than latent variables.

- **Disparity of representation level for latent variables across layers**

  The mutual information between $\mathbf{z}_i$ and gender is generally larger than that of race, and it fluctuates in a much smoother way. This indicates that not all latent variables are equivalently expressed in the neural network.

Besides these insights, we could also conclude that the layer we apply adversarial branch f^{adv} is layer three for race and layer two for gender.

The results of T_θ and T_θ^{adv} are shown in Table 5.7. The improvement of the final metric after applying adversarial training over gender and race is 5.00% and 7.43%. Note that since the dataset is fully balanced (see Table 5.1c), then the statistics (Table 5.8) given by prediction results also show symmetricity. For example, the BPSN AUC for identity ”Male” is equal to BNSP AUC for identity ”Female”.

Besides the fact that these results still follow the observations in text classification, we could also find

- **Marked improvement for the individual metric for some identities**

  When we train adversarially against gender and race, the BPSN AUC shows 24.74% and 47.79% improvement for identity ”Caucasian”. The same goes with BNSP AUC for identity ”Asian” because of the symmetricity noted before. This significant improvement shows the validity of our approach against the stochasticity of neural network training.

- **Coupled effects of latent variables**

  When applying adversarial training to one particular attribute A, the resulting effects do not just specialize in the attribute we apply adversarial training to, the other attribute is also influenced. For example, when adversarial training is over gender, the BPSN AUC for identity ”Male” shows marginal improvement (2.66%). Yet, the improvement for identity ”Caucasian” is as high as 24.47%, which indicates the correlation between latent variables A’s.

Table 5.6: Change of $I(\mathbf{z}; A)$ with respect to network depth in image classification

Layer	Target	Gender	Race
1	0.47357	0.60357	0.43204
2	0.48108	0.59015	0.48911
3	0.52364	0.65615	0.42024
4	0.52149	0.65487	0.42874

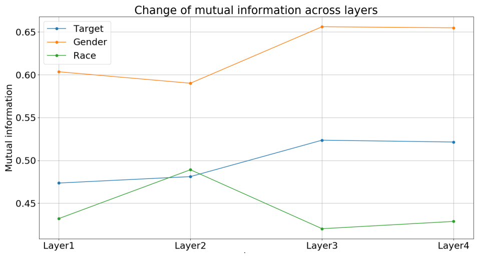
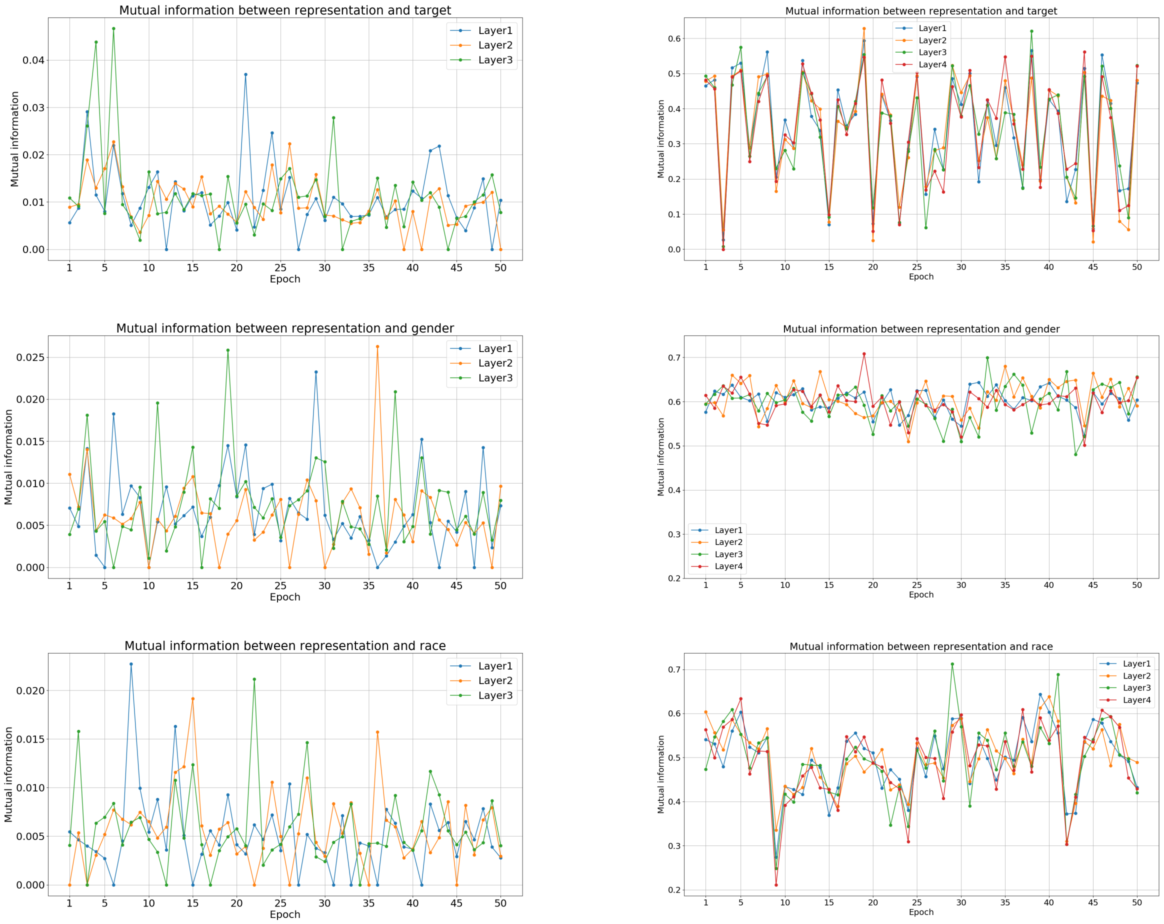


Figure 5.3: Change of $I(\mathbf{z}; A)$ with respect to network depth in image classification



(a) Text classification

(b) Image classification

Figure 5.4: Mutual information between representation $\mathbf{z}$ and attribute A for network T_θ

Table 5.7: Final metric for text classification

	No Adv	Adv @ Gender	Adv @ Race
BPSN AUC			
Male	0.564598	0.572683	0.557217
Female	0.502665	0.511772	0.500215
White	0.523946	0.517052	0.484918
Black	0.506660	0.523065	0.526015
BNSP AUC			
Male	0.489427	0.506296	0.505935
Female	0.512227	0.530751	0.526784
White	0.516937	0.543278	0.541647
Black	0.504677	0.516467	0.498857
Subgroup AUC			
Male	0.484545	0.498750	0.491830
Female	0.530688	0.548902	0.533397
White	0.490080	0.505871	0.505752
Black	0.520062	0.535509	0.523264
Overall AUC	0.503984	0.517261	0.509989
Final metric	0.508647	**0.522088**	**0.523554**

Table 5.8: Final metric for image classification

	No Adv	Adv @ Gender	Adv @ Race
	BPSN AUC		
Male	0.787993	0.808955	0.820446
Female	0.594667	0.571372	0.578742
Caucasian	0.491775	0.613415	0.726789
Asian	0.875663	0.789727	0.721067
	BNSP AUC		
Male	0.594667	0.571372	0.578742
Female	0.787993	0.808955	0.820446
Caucasian	0.875663	0.789727	0.721067
Asian	0.491775	0.613415	0.726789
	Subgroup AUC		
Male	0.682254	0.671642	0.675794
Female	0.702465	0.723348	0.739000
Caucasian	0.702936	0.726024	0.748832
Asian	0.696561	0.695870	0.702294
Overall AUC	0.700596	0.704080	0.713733
Final metric	0.646979	**0.677289**	**0.695079**

CHAPTER 6

Conclusion and Future Work

In this work, we investigate the use of the adversarial training strategy in preserving neural networks' fairness under the classification setting. Our conceptual contribution is to introduce mutual information as a measure to quantify the amount of information expressed through different layers of the neural network, and then objectively select the layer to apply adversarial training. The results show as much as 7.43% of improvement under our fairness measure.

In the future, the following directions are worth exploring

- **Disentangled representation of latent variables**

  As is shown in our experiments, the adversarial training does not just result in the improvement in the attribute we applied, which ascertains the entanglement of related latent variables. In future work, we aim to work towards the disentanglement of these latent variables and further improve the system performance.

- **Partial information decomposition of intermediate representation**

  In order for the principled understanding of the neural network predictive process, it is vital to understand how the information contained in each attribute propagate towards the end of the output. There has already been preliminary work for partial information decomposition to simple neural network architecture [TMS17]. We would like to extend previous work to our setting.

- **Theoretical understanding of adversarial training**

  Even though the adversarial training scheme receives empirical success, there is no theoretical analysis with regards to capacity, generalization, and convergence of this scheme as a result of the non-conventional gradient reversal layer. We believe that follow-up work should address these concerns.

REFERENCES

[ABD18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. "A reductions approach to fair classification." *arXiv preprint arXiv:1803.02453*, 2018.

[AVG19] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. "One-network adversarial fairness." In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[AZP19] Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M Pohl. "Bias-resilient neural network." *arXiv preprint arXiv:1910.03676*, 2019.

[BBC07] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. "Analysis of Representations for Domain Adaptation." In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pp. 137–144. MIT Press, 2007.

[BBR] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. "MINE: Mutual Information Neural Estimation.".

[BCZ17] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. "Data decisions and theoretical implications when adversarially learning fair representations." *arXiv preprint arXiv:1707.00075*, 2017.

[BDH18] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. "AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias." *arXiv preprint arXiv:1810.01943*, 2018.

[BDS19] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. "Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification." In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500. ACM, 2019.

[BG18] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, accountability and transparency*, pp. 77–91, 2018.

[BHJ17] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. "A convex framework for fair regression." *arXiv preprint arXiv:1706.02409*, 2017.

[BHJ18] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. “Fairness in criminal justice risk assessments: The state of the art.” *Sociological Methods & Research*, p. 0049124118782533, 2018.

[CR18] Alexandra Chouldechova and Aaron Roth. “The frontiers of fairness in machine learning.” *arXiv preprint arXiv:1810.08810*, 2018.

[Das18] Jeffrey Dastin. “Amazon scraps secret AI recruiting tool that showed bias against women.” *San Fransico, CA: Reuters. Retrieved on October*, **9**:2018, 2018.

[DBC] Prithviraj Dhar, Ankan Bansal, Carlos D. Castillo, Joshua Gleason, P. Jonathon Phillips, and Rama Chellappa. “How are attributes expressed in face DCNNs?”.

[DHP12] 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*, pp. 214–226. ACM, 2012.

[DLS18] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. “Measuring and mitigating unintended bias in text classification.” In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73. ACM, 2018.

[FSV19] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. “A comparative study of fairness-enhancing interventions in machine learning.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 329–338. ACM, 2019.

[GAK19] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. “Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search.” *arXiv preprint arXiv:1905.01989*, 2019.

[GOV] Weihao Gao, Sewoong Oh, and Pramod Viswanath. “Demystifying Fixed k-Nearest Neighbor Information Estimators.”.

[GUA] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. “Domain-Adversarial Training of Neural Networks.”.

[HPS16] Moritz Hardt, Eric Price, Nati Srebro, et al. “Equality of opportunity in supervised learning.” In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

[HZR16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition.” In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[KMR16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. “Inherent trade-offs in the fair determination of risk scores.” *arXiv preprint arXiv:1609.05807*, 2016.

[Koh96] Ron Kohavi. "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." In *Kdd*, volume 96, pp. 202–207. Citeseer, 1996.

[KR19] Michael Kearns and Aaron Roth. *The ethical algorithm: the science of socially aware algorithm design*. Oxford University Press, 2019.

[KSM19] Ari Kobren, Barna Saha, and Andrew McCallum. "Paper Matching with Local Fairness Constraints." *arXiv preprint arXiv:1905.11924*, 2019.

[LLJ18] Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li. "SCUT-FBP5500: A Diverse Benchmark Dataset for Multi-Paradigm Facial Beauty Prediction." 2018.

[LSL15] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. "The variational fair autoencoder." *arXiv preprint arXiv:1511.00830*, 2015.

[Rud17] Sebastian Ruder. "An overview of multi-task learning in deep neural networks." *arXiv preprint arXiv:1706.05098*, 2017.

[SVI16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[TMS17] Tycho Tax, Pedro Mediano, and Murray Shanahan. "The partial information decomposition of generative neural network models." *Entropy*, **19**(9):474, 2017.

[Vig19] Neil Vigdor. "Apple Card Investigated After Gender Discrimination Complaints.", November 2019.

[WXZ17] Lijun Wu, Yingce Xia, Li Zhao, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. "Adversarial neural machine translation." *arXiv preprint arXiv:1704.06933*, 2017.

[WZY] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. "Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations.".

[ZGN17] Kaiping Zheng, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, and Wei Luen James Yip. "Resolving the bias in electronic medical records." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2171–2180. ACM, 2017.

[ZLM] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating Unwanted Biases with Adversarial Learning.".

[ZWW17] Lu Zhang, Yongkai Wu, and Xintao Wu. "Achieving non-discrimination in data release." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1335–1344. ACM, 2017.

[ZWY17] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. "Men also like shopping: Reducing gender bias amplification using corpus-level constraints." *arXiv preprint arXiv:1707.09457*, 2017.