THE RICH GET RICHER: DISPARATE IMPACT OF SEMI-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Semi-supervised learning (SSL) has demonstrated its potential to improve the model accuracy for a variety of learning tasks when the high-quality supervised data is severely limited. Although it is often established that the average accuracy for the entire population of data is improved, it is unclear how SSL fares with different sub-populations. Understanding the above question has substantial fairness implications when these different sub-populations are defined by the demographic groups we aim to treat fairly. In this paper, we reveal the disparate impacts of deploying SSL: the sub-population who has a higher baseline accuracy without using SSL (the "rich" sub-population) tends to benefit more from SSL; while the sub-population who suffers from a low baseline accuracy (the "poor" sub-population) might even observe a performance drop after adding the SSL module. We theoretically and empirically establish the above observation for a broad family of SSL algorithms, which either explicitly or implicitly use an auxiliary "pseudo-label". Our experiments on a set of image and text classification tasks confirm our claims. We discuss how this disparate impact can be mitigated and hope that our paper will alarm the potential pitfall of using SSL and encourage a multifaceted evaluation of future SSL algorithms.

1 INTRODUCTION

The success of deep neural networks benefits from large-size datasets with high-quality supervisions. The collection of them though can often be difficult due to the high cost of data labeling process e.g., medical diagnosis (Agarwal et al., 2016). Practically, a much cheaper and easier solution is to obtain a large number of unlabeled dataset, and apply semi-supervised learning (SSL), which is one of the most promising paradigms, to leverage the information in the unlabeled dataset. Recent SSL methods based on consistency regularization, such as MixMatch (Berthelot et al., 2019b), UDA (Xie et al., 2019), and MixText (Chen et al., 2020), have achieved significant improvements on the model accuracy in a variety of tasks of computer vision and natural language processing.

Although the global model performance for the entire population of data is almost always improved by SSL, it is unclear how the improvements fare for different sub-populations. Analyzing and understanding the above question will have substantial fairness implications especially when these sub-populations are defined by the demographic groups e.g., race and gender.

Motivating Example We start from an example illustrated in Figure 1, which shows the change of test accuracy for each class during SSL. Each class denotes one sub-population and representative sub-populations are highlighted. Figure 1 shows that, no matter the size of each sub-population is equal or not, test accuracies of different sub-populations are improved differently during SSL. The *"rich"* sub-population, such as *automobile* that has a high baseline accuracy at the beginning of SSL, tends to be consistently benefiting during SSL. But the *"poor"* sub-population, such as *dog* that has a low baseline accuracy, will remain a low-level performance as Figure 1(a) or even get worse performance as Figure 1(b). This example shows disparate impact of accuracies for different sub-populations are common in SSL. In extreme cases such as Figure 1(b), we observe the Matthew effect: the rich get richer and the poor get poorer.

In this paper, we aim to understand the disparate impact of SSL from both theoretical and empirical aspects, and propose to evaluate SSL from a different dimension. Specifically, based on classifications tasks, we study the disparate impact of model accuracies with respect to different

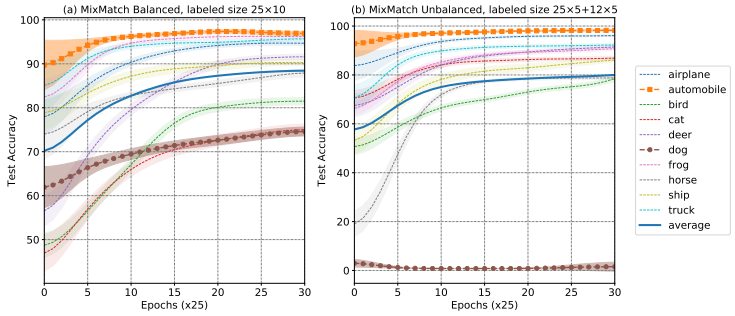


Figure 1: Disparate impacts in the model accuracy of SSL. MixMatch (Berthelot et al., 2019b) is applied on CIFAR-10 with 250 clean labels in the balanced case (25 labeled instances for each class) and 185 clean labels in the unbalanced case (25 labeled instances in each of the first 5 classes, 12 labeled instances in each of the remaining classes). Other instances are used as unlabeled data.

sub-populations (such as label classes, feature groups and demographic groups) after applying SSL. To this end, we first theoretically analyze why and how disparate impact are generated. The theoretical results motivate us to further propose a new metric, *benefit ratio*, which evaluates the normalized improvement of model accuracy using SSL methods for different sub-populations. The benefit ratio helps us uncover the *"Matthew effect"* of SSL: a high baseline accuracy tends to reach a high benefit ratio that may even be larger than 1 (the rich get richer), and a sufficiently low baseline accuracy tends to return a negative benefit ratio (the poor get poorer). We then numerically demonstrate using benefit ratios to evaluate SSL and discuss how two possible treatment methods, i.e., balancing the data and collecting more labeled data, might mitigate the disparate impact. We hope our analyses and discussions could encourage future contributions to promote the fairness of SSL.

Our main contributions and findings are summarized as follows: 1) We propose an analytical framework that unifies a broad family of SSL algorithms, which either explicitly or implicitly use an auxiliary "pseudo-label". The framework provides a novel perspective for analyzing SSL by connecting consistency regularization to the problem of learning with noisy labels. 2) Based on our framework, we further prove an upper bound for the generalization error of SSL and theoretically show the heterogeneous error of learning with the smaller scale supervised dataset is one primary reason for disparate impacts. This observation reveals a "Matthew effect" of SSL. 3) We contribute a novel metric called *benefit ratio* to measure the disparate impact in SSL, the effectiveness of which is theoretically guaranteed by our proved generalization bounds. 4) We conduct experiments on both image and text classification tasks to demonstrate the ubiquitous disparate impacts in SSL. 5) We also discuss how the disparate impact could be mitigated.

1.1 Related Works

Semi-supervised learning SSL is popular in both computer vision and natural language processing communities (Dai & Le, 2015; Howard & Ruder, 2018; Clark et al., 2018; Yang et al., 2019; Sachan et al., 2019; Guo et al., 2020; Wang et al., 2021; Luo et al., 2021). We briefly review recent advances in SSL. See comprehensive overviews by (Chapelle et al., 2006; Zhu et al., 2003) for traditional SSL methods. A major line of recent works focus on assigning pseudo-labels generated by the supervised model to unlabeled dataset (Lee et al., 2013; Iscen et al., 2019; Berthelot et al., 2019b;a), where the pseudo-labels are often confident or with low-entropy (Sohn et al., 2020; Zhou, 2018; Meng et al., 2018). There are also many works on minimizing entropy of predictions on unsupervised data (Grandvalet et al., 2005) or regularizing the model consistency on the same feature with different data augmentations (Tarvainen & Valpola, 2017; Miyato et al., 2018; Sajjadi et al., 2016; Clark et al., 2018), including mixup (Zhang et al., 2018), adversarial noise (Miyato et al., 2016), and other advanced augmentations (Xie et al., 2019). In addition to network inputs, augmentations can also be applied on some hidden layers (Chen et al., 2020). Besides, some works conduct pre-training on the unlabeled dataset first and then fine-tune the model on the labeled dataset (Peters et al., 2018; Devlin et al., 2018; Gururangan et al., 2019; Chen et al., 2019; Yang et al., 2017), or use ladder networks to combine unsupervised learning with supervised learning (Rasmus et al., 2015).

Disparate impact Over the past few years, the issue of fairness in machine learning has been extensively studied. Even models developed with the best intentions may introduce discriminatory

biases. Researchers in various fields have found the unfairness issues such as model compression (Bagdasaryan et al., 2019), differential privacy (Hooker et al., 2019; 2020), recommendation system (Gómez et al., 2021), information retrieval (Gao & Shah, 2021), machine translation (Khan et al., 2021), and learning with noisy labels (Liu, 2021). To our best knowledge, the unfairness issue in SSL methods has not been sufficiently explored.

2 PRELIMINARIES

We summarize the key concepts and notations as follows.

2.1 SUPERVISED CLASSIFICATION TASKS

Consider a K-class classification task given a set of N_L labeled training examples denoted by $D_L := \{(x_n, y_n)\}_{n\in[N_L]}$, where $[N_L] := \{1, 2, \cdots, N_L\}$ is the set of indices, x_n is a high-dimension feature, $y_n \in [K]$ stands for a particular label class. The corresponding clean data distribution with full supervision is denoted by $\mathcal{D}$. Examples (x_n, y_n) are drawn according to random variables $(X, Y) \sim \mathcal{D}$. The classification task aims to identify a classifier f that maps X to Y accurately. Let $\mathbb{1}\{\cdot\}$ be the indicator function taking value 1 when the specified condition is satisfied and 0 otherwise. Define the *0-1 loss* as $\mathbb{1}(f(X), Y) := \mathbb{1}\{f(X) \neq Y\}$. The optimal f is denoted by the Bayes classifier $f^* = \arg\min_f \mathbb{E}_{\mathcal{D}}[\mathbb{1}(f(X), Y)]$. One common choice is training a deep neural network (DNN) by minimizing the empirical risk: $\hat{f} = \arg\min_f \frac{1}{N}\sum_{n=1}^{N} \ell(f(x_n), y_n)$. Notation $\ell(\cdot)$ stands for the cross-entropy (CE) loss $\ell(f(x), y) := -\ln(\boldsymbol{f}_x[y]), y \in [K]$, where $\boldsymbol{f}_x[y]$ denotes the y-th component of $\boldsymbol{f}(x)$. Notations f and $\boldsymbol{f}$ stand for the same model but different outputs. Specifically, vector $\boldsymbol{f}(x)$ denotes the probability of each class that model f predicts given feature x. The predicted label $f(x)$ is the class with maximal probability, i.e., $f(x) := \arg\max_{i\in[K]} \boldsymbol{f}_x[i]$. We use notation f if we only refer to a model.

2.2 SEMI-SUPERVISED CLASSIFICATION TASKS

In the semi-supervised learning (SSL) task, there is also an unlabeled (a.k.a. unsupervised) dataset $D_U := \{(x_{n+N_L}, \cdot)\}_{n\in[N_U]}$ drawn from $\mathcal{D}$, while the labels are missing or unobservable. Let $N := N_L + N_U$. Denote the corresponding unobservable supervised dataset by $D := \{(x_n, y_n)\}_{n\in[N]}$. Compared with the supervised learning tasks, it is critical to leverage the unsupervised information in semi-supervised learning. To improve the model generalization ability, many recent SSL methods rely on building consistency regularization terms with unlabeled data to ensure that the model output remains unchanged with randomly augmented inputs (Berthelot et al., 2019b; Xie et al., 2019; Berthelot et al., 2019a; Sohn et al., 2020; Xu et al., 2021). To better formulate the consistency regularization in these popular methods, we introduce soft/pseudo-labels $\boldsymbol{y}$ as follows.

Soft-labels To motivate the definitions, we first consider a special case $\boldsymbol{y}_n$, i.e., the one-hot encoding of y_n, where each element writes as $\boldsymbol{y}_n[i] = \mathbb{1}\{i = y_n\}$. More generally, we can extend the one-hot encoding to soft labels by requiring each element $\boldsymbol{y}[i] \in [0, 1]$ and $\sum_{i\in[K]} \boldsymbol{y}[i] = 1$. The CE loss with soft $\boldsymbol{y}$ writes as $\ell(\boldsymbol{f}(x), \boldsymbol{y}) := -\sum_{i\in[K]} \boldsymbol{y}[i] \ln(\boldsymbol{f}_x[i])$. If we interpret $\boldsymbol{y}[i] = \mathbb{P}(Y = i)$ as a probability and denote by $\mathcal{D}_{\boldsymbol{y}}$ the corresponding label distribution, the above CE loss with soft labels can be interpreted as the expected loss with respect to a stochastic label $\widetilde{Y}$, i.e.,

$$\ell(\boldsymbol{f}(x), \boldsymbol{y}) := \sum_{i\in[K]} \mathbb{P}(\widetilde{Y} = i)\ell(f(x), i) = \mathbb{E}_{\widetilde{Y}\sim\mathcal{D}_{\boldsymbol{y}}}\left[\ell(f(x), \widetilde{Y})\right]. \tag{1}$$

Pseudo-labels In consistency regularization, by using model predictions, the unlabeled data will be assigned pseudo-labels either explicitly (Berthelot et al., 2019b) or implicitly (Xie et al., 2019), where the pseudo-labels can be modeled as soft-labels. In the following, we review both the explicit and the implicit approaches and unify them in our analytical framework.

Consistency regularization with explicit pseudo-labels For each unlabeled feature x_n, pseudo-labels can be explicitly generated based on model predictions (Berthelot et al., 2019b). The pseudo-label is later used to evaluate the model predictions. To avoid trivial solutions where model predictions and pseudo-labels are always identical, independent data augmentations of feature x_n are

often generated for M rounds. The augmented feature is denoted by $x'_{n,m} := \text{Augment}(x_n), m \in [M]$. Then the pseudo-label $\boldsymbol{y}_n$ in epoch-t can be determined based on M model predictions as $\boldsymbol{y}_n^{(t)} = \text{Sharpen}\left(\frac{1}{M}\sum_{m=1}^{M} \bar{\boldsymbol{f}}^{(t)}(x'_{n,m})\right)$, where model $\bar{f}^{(t)}$ is a copy of the DNN at the beginning of epoch-t but without gradients. The function Sharpen$(\cdot)$ reduces the entropy of a pseudo-label, e.g., setting to one-hot encoding $\boldsymbol{e}_j, j = \arg\max_{i\in[K]} \boldsymbol{y}_n^{(t)}$ (Sohn et al., 2020). In epoch-t, with some consistency regularization loss $\ell_{\text{CR}}(\cdot)$, the empirical risk using pseudo-labels is:

$$L_1(f, D_L, D_U) = \frac{1}{N_L}\sum_{n=1}^{N_L} \ell(f(x_n), y_n) + \frac{1}{N_U}\sum_{n=N_L+1}^{N_L+N_U} \ell_{\text{CR}}(\boldsymbol{f}(x_n), \boldsymbol{y}_n^{(t)}).$$

Consistency regularization with implicit pseudo-labels Consistency regularization can also be applied without specifying particular pseudo-labels, where a divergence metric between predictions on the original feature and the augmented feature is minimized to make predictions consistent. For example, the KL-divergence could be applied and the data augmentation could be domain-specific (Xie et al., 2019) or adversarial (Miyato et al., 2018). In epoch-t, the total loss is:

$$L_2(f, D_L, D_U) = \frac{1}{N_L}\sum_{n=1}^{N_L} \ell(f(x_n), y_n) + \lambda \cdot \frac{1}{N_U}\sum_{n=N_L+1}^{N_L+N_U} \ell_{\text{CR}}(\boldsymbol{f}(x_n), \bar{\boldsymbol{f}}^{(t)}(x'_n)),$$

where λ balances the supervised loss and the unsupervised loss, $x'_n := \text{Augment}(x_n)$ stands for one-round data augmentation ($m = 1$ following the previous notation $x'_{n,m}$). Without loss of generality, we use $\lambda = 1$ in our analytical framework.

Consistency regularization loss $\ell_{\text{CR}}(\cdot)$ In the above two lines of works, there are different choices of $\ell_{\text{CR}}(\cdot)$, such as mean-squared error $\ell_{\text{CR}}(\boldsymbol{f}(x), \boldsymbol{y}) := \|\boldsymbol{f}(x) - \boldsymbol{y}\|_2^2/K$ (Berthelot et al., 2019b; Laine & Aila, 2016) or CE loss (Miyato et al., 2018; Xie et al., 2019) defined in Eq. (1). For a clean analytical framework, we consider the case when both supervised loss and unsupervised loss are the same, i.e., $\ell_{\text{CR}}(\boldsymbol{f}(x), \boldsymbol{y}) = \ell(\boldsymbol{f}(x), \boldsymbol{y})$. Note L_2 implies the entropy minimization (Grandvalet et al., 2005) when both loss functions are CE and there is no augmentation, i.e., $x'_n = x_n$.

2.3 Analytical Framework

We propose the following analytical framework to unify both the explicit and the implicit approaches. With Eqn. (1), the unsupervised loss term in the above two methods can be respectively written as:

$$\frac{1}{N_U}\sum_{n=N_L+1}^{N_L+N_U} \mathbb{E}_{\widetilde{Y}\sim\mathcal{D}_{\boldsymbol{y}_n^{(t)}}}\left[\ell(f(x_n), \widetilde{Y})\right] \quad \text{and} \quad \frac{1}{N_U}\sum_{n=N_L+1}^{N_L+N_U} \mathbb{E}_{\widetilde{Y}\sim\mathcal{D}_{\bar{\boldsymbol{f}}^{(t)}(x'_n)}}\left[\ell(f(x_n), \widetilde{Y})\right].$$

Both unsupervised loss terms inform us that, for each feature x_n, we compute the loss with respect to the reference label $\widetilde{Y}$, which is a random variable following distribution $\mathcal{D}_{\boldsymbol{y}_n^{(t)}}$ or $\mathcal{D}_{\bar{\boldsymbol{f}}^{(t)}(x'_n)}$. Compared with the original clean label Y, the unsupervised reference label $\widetilde{Y}$ contains label noise, and the noise transition depends on feature x_n. Specifically, we have

$$\mathbb{P}(\widetilde{Y} = i|X = x_n) = \boldsymbol{y}_n^{(t)}[i] \text{ (Explicit)} \quad \text{or} \quad \mathbb{P}(\widetilde{Y} = i|X = x_n) = \bar{\boldsymbol{f}}_{x'_n}^{(t)}[i] \text{ (Implicit)}.$$

Then with model $\bar{f}^{(t)}$, we can define a new dataset with instance-dependent noisy reference labels: $\widetilde{D} = \{(x_n, \tilde{\boldsymbol{y}}_n)\}_{n\in[N]}$, where $\tilde{\boldsymbol{y}}_n = \boldsymbol{y}_n, \forall n \in [N_L]$, and $\tilde{\boldsymbol{y}}_n = \boldsymbol{y}_n^{(t)}$ or $\bar{\boldsymbol{f}}_{x'_n}^{(t)}, \forall n \in \{N_L+1, \cdots, N\}$. The unified loss is:

$$L(f, \widetilde{D}) = \frac{1}{N}\sum_{n=1}^{N} \ell(\boldsymbol{f}(x_n), \tilde{\boldsymbol{y}}_n) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{\widetilde{Y}\sim\mathcal{D}_{\tilde{\boldsymbol{y}}_n}}\left[\ell(f(x_n), \widetilde{Y})\right]. \tag{2}$$

Therefore, with the pseudo-label as a bridge, we can model the popular SSL solution with consistency regularization as a problem of learning with noisy labels (in the challenging instance-dependent label noise setting).

3 DISPARATE IMPACTS OF SSL

To understand the disparate impact of SSL, we study how supervised learning with D_L affects the performance of SSL with both D_L and D_U.

Intuition The rich sub-population with a higher baseline accuracy (from supervised learning with D_L) will have higher-quality pseudo-labels for consistency regularization, which helps to further improve the performance. In contrast, the poorer sub-population with a lower baseline accuracy (again from supervised learning with D_L) can only have lower-quality pseudo-labels to regularize the consistency of unsupervised features. With a wrong regularization direction, the unsupervised feature will have its augmented copies reach consensus on a wrong label class, which leads to a performance drop. Therefore, when the baseline accuracy is getting worse, there will be more and more unsupervised features that are wrongly regularized, resulting in disparate impacts of model accuracies as shown in Figure 1.

In this section, we first analyze the generalization error for supervised learning with labeled data, then extend the analyses to semi-supervised learning with the help of pseudo-labels. Without loss of generality, we consider minimizing 0-1 loss $\mathbb{1}(f(X), Y)$ with infinite search space. Our analyses can be generalized to bounded loss $\ell(\cdot)$ and finite function space $\mathcal{F}$ following the generalization bounds that can be introduced using Rademacher complexity (Bartlett & Mendelson, 2002).

3.1 LEARNING WITH CLEAN DATA

Denote the expected error rate of classifier f on distribution $\mathcal{D}$ by $R_{\mathcal{D}}(f) := \mathbb{E}_{\mathcal{D}}[\mathbb{1}(f(X), Y)]$. Let $\hat{f}_D$ denote the classifier trained by minimizing 0-1 loss with clean dataset D, i.e., $\hat{f}_D := \arg\min_f \hat{R}_D(f)$, where $\hat{R}_D(f) := \frac{1}{N}\sum_{n\in[N]} \mathbb{1}(f(x_n), y_n)$. Denote by $Y^*|X := \arg\max_{i\in[K]} \mathbb{P}(Y|X)$ the Bayes optimal label on clean distribution $\mathcal{D}$. Theorem 1 shows the generalization bound in the clean case. Replacing D and N with D_L and N_L yields bounds for D_L.

Theorem 1 (Supervised error)**.** *With probability at least $1-\delta$, the generalization error of supervised learning on clean dataset D is upper-bounded by $R_{\mathcal{D}}(\hat{f}_D) \leq \sqrt{\frac{2\log(2/\delta)}{N}} + \mathbb{P}(Y^* \neq Y)$.*

3.2 LEARNING WITH SEMI-SUPERVISED DATA

We further derive generalization bounds for learning with semi-supervised data. Following our analytical framework in Section 2.3, in each epoch-t, we can transform the semi-supervised data to supervised data with noisy supervisions by assigning pseudo-labels, where the semi-supervised dataset $D_L \cup D_U$ is converted to the dataset $\widetilde{D}$ given the model learned from the previous epoch.

Two-iteration scenario In our theoretical analyses, to find a clean-structured performance bound for learning with semi-supervised data and highlight the impact of the model learned from supervised data, we consider a particular *two-iteration scenario* where the model is first trained to convergence with the small labeled dataset D_L and get $\hat{f}_{D_L}$, then trained on the pseudo noisy dataset $\widetilde{D}$ labeled by $\hat{f}_{D_L}$. Noting assigning pseudo labels iteratively may improve the performance as suggested by most SSL algorithms (Berthelot et al., 2019b; Xie et al., 2019), our considered two-iteration scenario can be seen as a worst case for an SSL algorithm.

Independence of samples in $\widetilde{D}$ Recall x'_n denotes the augmented copy of x_n. The N instances in $\widetilde{D}$ may not be independent since pseudo-label $\widetilde{\boldsymbol{y}}_n$ comes from $\hat{f}_{D_L}(x'_n)$, which depends on x_n. In other words, the number of independent instances N' should be in the range of $[N_L, N]$. Intuitively, with appropriate noise injection or data augmentation (Xie et al., 2019; Miyato et al., 2018) to x_n such that x'_n could be treated as independent of x_n, the number of independent samples in $\widetilde{D}$ could be improved to N. We consider the ideal case where all N instances are *i.i.d.* in the following analyses.

By minimizing the unified loss defined in Eq. (2), we can get classifier $\hat{f}_{\widetilde{D}} := \arg\min_f \hat{R}_{\widetilde{D}}(f)$, where $\hat{R}_{\widetilde{D}}(f) := \frac{1}{N}\sum_{n\in[N]} \left(\sum_{i\in[K]} \tilde{\boldsymbol{y}}[i] \cdot \mathbb{1}(f(x_n), i)\right)$. The expected error given classifier f is

denoted by $R_{\widetilde{\mathcal{D}}}(f) := \mathbb{E}_{\widetilde{\mathcal{D}}}[\mathbb{1}(f(X), \widetilde{Y})]$, where the probability density function of distribution $\widetilde{\mathcal{D}}$ can be defined as $\mathbb{P}_{(X,\widetilde{Y})\sim\widetilde{\mathcal{D}}}(X = x_n, \widetilde{Y} = i) = \mathbb{P}_{(X,Y)\sim\mathcal{D}}(X = x_n) \cdot \tilde{\boldsymbol{y}}_n[i]$.

Decomposition With the above definitions, the generalization error (on the clean distribution) of classifier $\hat{f}_{\widetilde{D}}$ could be decomposed as $R_{\mathcal{D}}(\hat{f}_{\widetilde{D}}) = \underbrace{(R_{\mathcal{D}}(\hat{f}_{\widetilde{D}}) - R_{\widetilde{\mathcal{D}}}(\hat{f}_{\widetilde{D}}))}_{\text{Term-1}} + \underbrace{R_{\widetilde{\mathcal{D}}}(\hat{f}_{\widetilde{D}})}_{\text{Term-2}}$, where **Term-1** transforms the evaluation of $\hat{f}_{\widetilde{D}}$ from clean distribution $\mathcal{D}$ to the pseudo noisy distribution $\widetilde{\mathcal{D}}$. **Term-2** is similar to the generalization error in Theorem 1 but the model is trained and evaluated on noisy distribution $\widetilde{\mathcal{D}}$. Both terms are analyzed as follows.

Upper and Lower Bounds for Term-1 Let $\eta(X) := \frac{1}{2}\sum_{i\in[K]} |\mathbb{P}(\widetilde{Y} = i|X) - \mathbb{P}(Y = i|X)|$, $e(X) := \mathbb{P}(Y \neq \widetilde{Y}|X)$ be the feature-dependent error rate, $\widetilde{A}_f(X) := \mathbb{P}(f(X) = \widetilde{Y}|X)$ be the accuracy of prediction $f(X)$ on noisy dataset $\widetilde{D}$. Denote their expectations (over X) by $\bar{\eta} := \mathbb{E}_X[\eta(X)]$, $\bar{e} := \mathbb{E}_X[e(X)]$, $\widetilde{A}_f = \mathbb{E}_X[\widetilde{A}_f(X)]$. To highlight that $\bar{\eta}$ and $\bar{e}$ depends on the noisy dataset $\widetilde{D}$ labeled by $\hat{f}_{D_L}$, we denote them as $\bar{\eta}(\hat{f}_{D_L})$ and $\bar{e}(\hat{f}_{D_L})$. Then we have:

Lemma 1 (Bounds for Term-1)**.** $(2\widetilde{A}_{\hat{f}_{\widetilde{D}}} - 1)\bar{e}(\hat{f}_{D_L}) \leq R_{\mathcal{D}}(\hat{f}_{\widetilde{D}}) - R_{\widetilde{\mathcal{D}}}(\hat{f}_{\widetilde{D}}) \leq \bar{\eta}(\hat{f}_{D_L})$.

Note the upper bound builds on $\bar{\eta}(\hat{f}_{D_L})$ while the lower bound relates to $\bar{e}(\hat{f}_{D_L})$. To compare two bounds and show the tightness, we consider the case where $Y|X$ is *confident*, i.e., each feature X belongs to one particular true class Y with probability 1, which is generally held in classification problems (Liu & Tao, 2015). Lemma 2 shows $\eta(X) = e(X)$ in this particular case[1].

Lemma 2 (η vs. e)**.** *For any feature X, if $Y|X$ is confident, $\eta(X)$ is the error rate of the model prediction on X, i.e., $\exists i \in [K] : \mathbb{P}(Y = i|X) = 1 \Rightarrow \eta(X) = \mathbb{P}(\widetilde{Y} \neq Y|X) = e(X)$.*

Upper bound for Term-2 Denote by $\widetilde{Y}^*|X := \arg\max_{i\in[K]} \mathbb{P}(\widetilde{Y} = i|X)$ the Bayes optimal label on noisy distribution $\widetilde{\mathcal{D}}$. Following the proof for Theorem 1, we have:

Lemma 3 (Bound for Term-2)**.** *W. p. at least $1-\delta$, $R_{\widetilde{\mathcal{D}}}(\hat{f}_{\widetilde{D}})(f_{\widetilde{\mathcal{D}}}) \leq \sqrt{\frac{2\log(2/\delta)}{N}} + \mathbb{P}(\widetilde{Y} \neq \widetilde{Y}^*)$.*

Wrap-up Lemma 1 shows Term-1 is in the range of $[(2\widetilde{A}_{\hat{f}_{\widetilde{D}}} - 1)\bar{e}(\hat{f}_{D_L}), \bar{\eta}(\hat{f}_{D_L})]$. Lemma 2 informs us $\bar{\eta}(\hat{f}_{D_L}) = \bar{e}(\hat{f}_{D_L})$ in classification problems where $Y|X, \forall X$ are confident. With a *well-trained* model $\hat{f}_{\widetilde{D}}$ that learns the noisy distribution $\widetilde{\mathcal{D}}$, we have $\widetilde{A}_{\hat{f}_{\widetilde{D}}} = 1 - \epsilon$ and $\epsilon \to 0_+$, thus Term-1 is in the range of $[1 - 2\epsilon, 1]$, indicating *our bounds for Term-1 are tight*. Besides, noting Lemma 3 is derived following the same techniques as Theorem 1, we know both bounds have similar tightness. Therefore, by adding upper bounds for Term-1 and Term-2, we can upper bound the error of semi-supervised learning in Theorem 2, which has similar tightness to that in Theorem 1.

Theorem 2 (Semi-supervised learning error)**.** *Suppose the model trained with only D_L has generalization error $\bar{\eta}'(\hat{f}_{D_L})$. When $Y|X$ is confident, with probability at least $1-\delta$, the generalization error of semi-supervised learning on datasets $D_L \cup D_U$ is upper-bounded by*

$$R_{\mathcal{D}}(\hat{f}_{D_L\cup D_U}) \leq \underbrace{\bar{\eta}(\hat{f}_{D_L})}_{\textit{Disparity due to baseline}} + \underbrace{\mathbb{P}(\widetilde{Y} \neq \widetilde{Y}^*)}_{\textit{Sharpness of pseudo labels}} + \underbrace{\sqrt{\frac{2\log(2/\delta)}{N}}}_{\textit{Data dependency}},$$

where $\bar{\eta}(\hat{f}_{D_L}) := \bar{\eta}'(\hat{f}_{D_L}) \cdot N_U/N$ is the expected label error in the pseudo noisy dataset $\widetilde{D}$.

Takeaways Theorem 2 explains how disparate impacts in SSL are generated.

- Supervised error $\bar{\eta}'(\hat{f}_{D_L})$: the major source of disparity. The sub-population that generalizes well before SSL tends to have a lower SSL error. Namely, the rich get richer.
- Sharpness of noisy labels $\mathbb{P}(\widetilde{Y} \neq \widetilde{Y}^*)$: a minor source of disparity depends on how one processes pseudo-labels. This term is negligible if we sharpen the pseudo-label.
- Sample complexity $\sqrt{2\log(2/\delta)/N}$: a minor source of disparity depends on the number of instances N and their independence. Recall we assume ideal data augmentations to get N in this term. There will be much less than N *i.i.d.* instances with poor data augmentations.

[1] We can also prove $\eta(X) = e(X)$ when $\widetilde{Y}|X$ is confident. See more details in the proof.

4 BENEFIT RATIO: A EVALUATION METRIC

Figure 1 demonstrates that SSL leads to disparate impacts of different sub-populations' accuracies, but it is still not clear how much that SSL benefits a certain sub-population. For example, a sub-population with a low test accuracy after SSL may be benefited much less from SSL since it is very hard to learn. On the other hand, a sub-population with a high test accuracy after SSL may not get benefits from SSL since it is easy to learn and the baseline accuracy is high enough. To highlight the benefit and evaluate the disparate impacts of SSL, we propose a new metric *benefit ratio*.

Benefit Ratio We now introduce the benefit ratio, $\mathsf{BR}(\mathcal{P})$, which depends on three classifiers: (baseline) supervised learning only with labeled data $\hat{f}_{D_L}$, (ideal) fully-supervised learning $\hat{f}_D$, and SSL $\hat{f}_{D_L \cup D_U}$, to capture the relative accuracy improvement of sub-population $\mathcal{P}$ after SSL. Note $\mathsf{BR}(\mathcal{P})$ are defined with three learning results: 1) the baseline accuracy from supervised learning with only a small labeled dataset, denoted by $a_{\text{baseline}}(\mathcal{P})$; 2) the accuracy from semi-supervised learning with both labeled dataset and unsupervised dataset, denoted by $a_{\text{semi}}(\mathcal{P})$; 3) the accuracy from fully-supervised learning with the ground truth labeled dataset, denoted by $a_{\text{ideal}}(\mathcal{P})$. Then:

$$\mathsf{BR}(\mathcal{P}) = \frac{a_{\text{semi}}(\mathcal{P}) - a_{\text{baseline}}(\mathcal{P})}{a_{\text{ideal}}(\mathcal{P}) - a_{\text{baseline}}(\mathcal{P})}.$$

Intuitively, a larger benefit ratio indicates more benefits from SSL. We have $\mathsf{BR}(\mathcal{P}) = 1$ when SSL performs as well as the corresponding fully-supervised learning. A negative benefit ratio indicates the poor population is hurt by SSL such that $a_{\text{semi}}(\mathcal{P}) < a_{\text{baseline}}(\mathcal{P})$, i.e., the poor get poorer as shown in Figure 1(b) (sub-population *dog*). Therefore, compared with the test accuracy, benefit ratio shows more advantages in evaluating the disparate impacts of SSL.

Theoretical Explanation Recall we have error upper bounds for both supervised learning (Theorem 1) and semi-supervised learning (Theorem 2). Both bounds have similar tightness thus we can compare them and get a proxy of benefit ratio as $\widehat{\mathsf{BR}}(\mathcal{P}) := \frac{\sup(R_{\mathcal{D}}(\hat{f}_{D_L \cup D_U|\mathcal{P}})) - \sup(R_{\mathcal{D}}(\hat{f}_{D_L|\mathcal{P}}))}{\sup(R_{\mathcal{D}}(\hat{f}_{D_L \cup D_U|\mathcal{P}})) - \sup(R_{\mathcal{D}}(\hat{f}_{D|\mathcal{P}}))}$, where $\sup(\cdot)$ denotes the upper bound derived in Theorem 1 and Theorem 2, $\mathcal{P}$ is a sub-population, and $D|\mathcal{P}$ denotes the set of *i.i.d.* instances in D that affect model generalization on $\mathcal{P}$. By assuming $\mathbb{P}(Y \neq Y^*) = \mathbb{P}(\widetilde{Y} \neq \widetilde{Y}^*)$ (both distributions have the same sharpness), we have:

Corollary 1. *The benefit ratio proxy for $\mathcal{P}$ is $\widehat{\mathsf{BR}}(\mathcal{P}) = 1 - \frac{\bar{\eta}(\hat{f}_{D_L|\mathcal{P}})}{\Delta(N_{\mathcal{P}}, N_{\mathcal{P}_L})}$, where $\Delta(N_{\mathcal{P}}, N_{\mathcal{P}_L}) = \sqrt{\frac{2\log(2/\delta)}{N_{\mathcal{P}_L}}} - \sqrt{\frac{2\log(2/\delta)}{N_{\mathcal{P}}}}$, $N_{\mathcal{P}}$ and $N_{\mathcal{P}_L}$ are the numbers of instances in $D|\mathcal{P}$ and $D_L|\mathcal{P}$.*

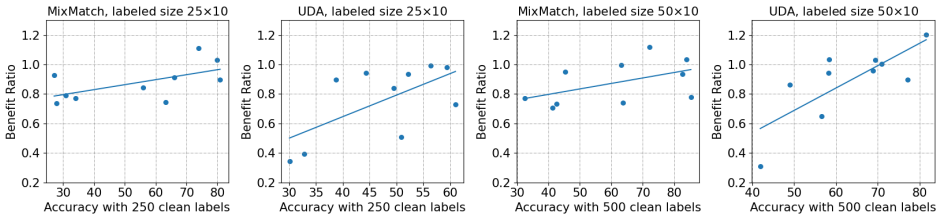
Corollary 1 shows the benefit ratio is negatively related to the error rate of baseline models and positively related to the number of *i.i.d.* instances after in SSL, which is consistent with our take-aways from Section 3. It also informs us that SSL may have a negative effect on sub-population $\mathcal{P}$ if $\frac{\eta}{\Delta(N_{\mathcal{P}}, N_{\mathcal{P}_L})} > 1$, i.e., the benefit from getting more effective *i.i.d.* instances is less than the harm from wrong pseudo-labels.

5 EXPERIMENTS

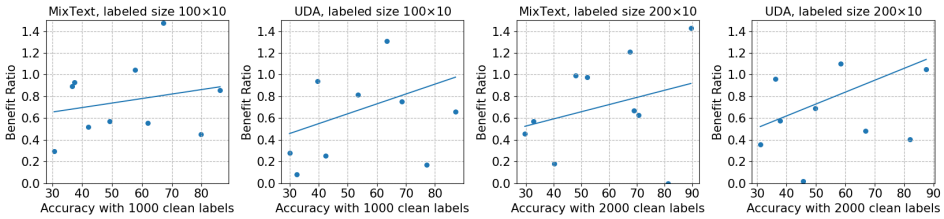
In this section, we first show the existence of disparate impact on several representative SSL methods and then discuss the possible treatment methods to mitigate this disparate impact.

Settings Three representative SSL methods, i.e., MixMatch (Berthelot et al., 2019b), UDA (Xie et al., 2019), and MixText (Chen et al., 2020), are tested on several image and text classification tasks. For image classification, we experiment on CIFAR-10 and CIFAR-100 datasets (Krizhevsky et al., 2009), where we randomly sample a class-balanced subset as the supervised data and treat features of the remainder as unlabeled data. We adopt the *coarse labels* (20 classes) in CIFAR-100 for training and test the performance for each *fine label* (100 classes). Thus our training on CIFAR-100 is a 20-class classification task and each coarse class contains 5 sub-populations. For text classification, we employ three datasets Yahoo! Answers (Chang et al., 2008), AG News (Zhang et al., 2015) and Jigsaw Toxicity from Kaggle[2]. It should be noticed that Jigsaw Toxicity dataset

[2]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification



(a) Benefit ratios (y-axis) versus baseline accuracies before SSL (x-axis) on CIFAR-10



(b) Benefit ratios (y-axis) versus baseline accuracies before SSL (x-axis) on Yahoo! Answers

Figure 2: Benefit ratios across explicit sub-populations. Dot: Result of each label class. Line: Best linear fit of dots.

contains not only classification labels (one text comment is toxic or non-toxic), but also the labels of a variety of identity attributes (e.g., race and gender information), representing the identities that are mentioned in the text comment. Therefore, we can demonstrate the disparate impact with respect to demographic groups in the Jigsaw Toxicity dataset and show the disparate impact on the classification labels for other tasks. We defer more experimental details in Appendix B.

5.1 Disparate Impact Exists in Popular SSL Methods

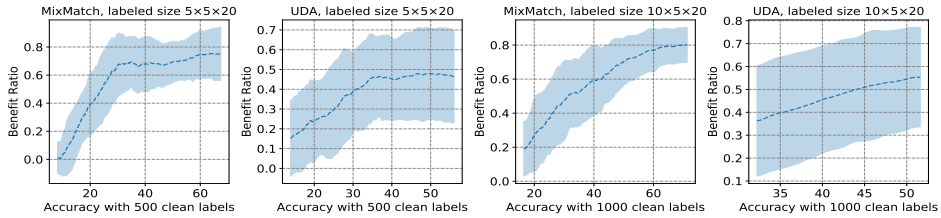
We show disparate impacts on the model accuracy of different sub-populations, i.e., 1) *explicit sub-populations* such as classification labels in Figure 2 (and Figure 4 in Appendix B.2), and 2) *implicit sub-populations* such as fine-categories under coarse classification labels in Figure 3a, demographic groups including race in Figure 3b and gender (Figure 5 in Appendix B.3).

Disparate impact across explicit sub-populations In this part, we show the disparate impact on model accuracy across different classification labels on CIFAR-10, Yahoo! Answers, and AG News datasets. Detailed results on AG News can be found in Appendix B.2. In Figure 2a and 2b, we show the benefit ratios (y-axis) versus baseline accuracies before SSL (x-axis). From left to right, we show results with different sizes of labeled data: 25 per class to 50 on CIFAR-10 and 100 per class to 200 on Yahoo! Answers. Figure 2a and 2b utilized two SSL methods respectively (CIFAR-10: MixMatch & UDA; Yahoo! Answers: MixText & UDA).We consistently observe that the class labels with higher baseline accuracies have higher benefit ratios on both CIFAR-10 and Yahoo! Answers datasets. It means that "richer" classes benefit more from applying SSL methods than the "poorer" ones. We also observe that for some models with low baseline accuracy (left side of the x-axis), applying SSL results in rather low benefit ratios that are close to 0.

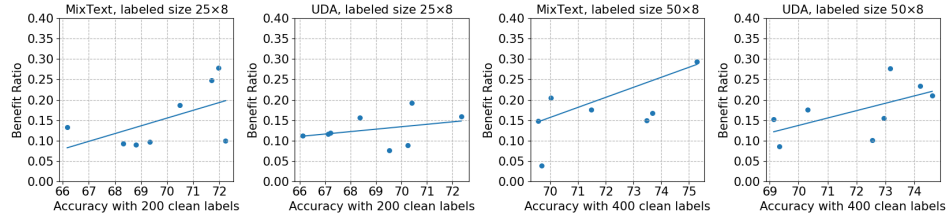
Disparate impact across implicit sub-populations We demonstrate the disparate impacts on model accuracy across different sub-populations on CIFAR-100 (fine-labels) and Jigsaw (race and gender) datasets. In Figure 3, we can consistently observe the disparate impact across different sub-populations on both datasets for three baseline SSL methods. Detailed results on Jigsaw with gender are shown in Appendix B.3. We again observe very similar disparate improvements as presented in Figure 2a and 2b - for some classes in CIFAR-100, this ratio can even go negative. Note the disparate impact on the demographic groups in Jigsaw raises fairness concerns in real-world applications.

5.2 Mitigating Disparate Impact

Our analyses in Section 3 and Section 4 indicate the disparate impacts may be mitigated by balancing the supervised error $\bar{\eta}(\hat{f}_{D_L|\mathcal{P}})$ and the number of effective *i.i.d.* instances for different populations.



(a) Benefit ratios of fine-labels (y-axis) versus baseline accuracies before SSL (x-axis) on CIFAR-100



(b) Benefit ratios of different race attributes (y-axis) versus baseline accuracies before SSL on Jigsaw.

Figure 3: Benefit ratios across implicit sub-populations.

Table 1: Comparison on the benefit ratio (%) of all race identities & standard deviation (%) between settings of unbalanced (400 labeled), balance labeled only (400 labeled), balance both labeled & unlabeled (400 labeled), and balance both labeled & unlabeled (800 labeled) on the Jigsaw dataset. t_asian means the benefit ratio of asian identity in toxic class and nt_asian means the one in non-toxic class. *SD* stands for standard deviation. The negtive benefit ratios are highlighted with red colors.

Settings	t_asian	t_black	t_latin	t_white	nt_asian	nt_black	nt_latin	nt_white	*SD*
Unbalanced (400)	24.71	21.76	28.21	**-9.57**	27.27	**-8.33**	**-13.82**	29.47	19.28
Balance labeled (400)	28.18	20.00	25.23	33.93	14.33	**-11.29**	**-10.37**	3.70	17.29
Balance both (400)	28.57	0.00	11.11	40.63	15.38	14.29	6.90	**-7.41**	15.29
Balance both (800)	13.04	25.00	7.69	24.00	3.85	10.71	16.67	20.00	7.64

We perform preliminary tests to check the effectiveness of the above findings in this subsection. We hope our analyses and experiments could inspire future contributions to disparity mitigation in SSL.

Balancing and Collecting more labeled data To balance the supervised error and effective instances for different sub-populations, an intuitive method is to balance the labeled data by resampling, e.g., if the size of two sub-populations is 1:2, we simply resample instances from two sub-populations with a probability ratio of 2:1 during training. Table 1 shows a detailed comparison on the benefit ratio of each race identity & standard deviation. The first three rows show three balancing settings with 400 labeled instances on the Jigsaw dataset. We can observe that both the standard deviation and the number of negative benefit ratios become lower with more balanced settings, which demonstrates the effectiveness of balancing treatment strategy. Although there still exists a sub-population that has a negative benefit ratio, indicating an unfair learning result. To further improve fairness, as suggested in our analyses, we "collect" (rather add) another 400 *i.i.d.* instances (800 in total) on the Jigsaw dataset, and show the result after balancing both labeled and unlabeled data in the last row of Table 3. Both the standard deviation and the number of negative benefit ratios can be further reduced with more labeled data. More experimental results are on CIFAR-100 and Jigsaw (gender) datasets shown in Appendix B.4 (balancing) and Appendix B.5 (more data).

6 Conclusions

We have theoretically and empirically shown that the disparate impact (the "richer" sub-populations get richer and the poor ones get poorer) exists universally for a broad family of SSL algorithms. We have also proposed and studied a new metric *benefit ratio* to facilitate the evaluation of SSL. We hope our work could inspire more efforts towards mitigating the disparate impact and encourage a multifaceted evaluation of existing and future SSL algorithms.

7 ETHICS STATEMENT

In recent years, machine learning methods are widely applied in the real world across a broad variety of fields such as face recognition, recommendation system, information retrieval, medical diagnosis, and marketing decision. The trend is likely to continue growing. However, it is noted that these machine learning methods are susceptible to introducing unfair treatments and can lead to systematic discrimination against some demographic groups. Our paper echoes this effort and looks into the disparate model improvement resulted from deploying a semi-supervised learning (SSL) algorithm.

Our work uncovers the unfairness issues in semi-supervised learning methods. More specifically, we theoretically and empirically show the disparate impact exists in the recent SSL methods on several public image and text classification datasets. Our results provide understandings of the potential disparate impacts of an SSL algorithm and help raise awareness from a fairness perspective when deploying an SSL model in real-world applications. Our newly proposed metric benefit ratio contributes to the literature a new measure to evaluate the fairness of SSL. Furthermore, we discuss two possible treatment methods: balancing and collecting more labeled data to mitigate the disparate impact in SSL methods. We have carefully studied and presented their effectiveness in mitigating unfairness in the paper. Our experimental results, such as Table 1, demonstrate the disparate impacts across demographic groups (e.g., gender, race) naturally exist in SSL. We are not aware of the misuse of our proposals, but we are open to discussions.

REFERENCES

Vibhu Agarwal, Tanya Podchiyska, Juan M Banda, Veena Goel, Tiffany I Leung, Evan P Minty, Timothy E Sweeney, Elsie Gyang, and Nigam H Shah. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association*, 23 (6):1166–1173, 2016.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 32:15479–15488, 2019.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019a.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. Importance of semantic representation: Dataless classification. In *Aaai*, volume 2, pp. 830–835, 2008.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. *MIT Press*, 2006.

Jiaao Chen, Zichao Yang, and Diyi Yang. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *arXiv preprint arXiv:2004.12239*, 2020.

Mingda Chen, Qingming Tang, Karen Livescu, and Kevin Gimpel. Variational sequential labelers for semi-supervised learning. *arXiv preprint arXiv:1906.09535*, 2019.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc V Le. Semi-supervised sequence modeling with cross-view training. *arXiv preprint arXiv:1809.08370*, 2018.

Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. *Advances in neural information processing systems*, 28:3079–3087, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ruoyuan Gao and Chirag Shah. Addressing bias and fairness in search systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2643–2646, 2021.

Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. Disparate impact in item recommendation: A case of geographic imbalance. In *European Conference on Information Retrieval*, pp. 190–206. Springer, 2021.

Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. *CAP*, 367:281–296, 2005.

Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International Conference on Machine Learning*, pp. 3897–3906. PMLR, 2020.

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A Smith. Variational pretraining for semi-supervised text classification. *arXiv preprint arXiv:1906.02242*, 2019.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5070–5079, 2019.

Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. Translation tutorial: Fairness and friends. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

Yang Liu. The importance of understanding instance-level noisy labels, 2021.

Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, 2020.

Huixiang Luo, Hao Cheng, Yuting Gao, Ke Li, Mengdan Zhang, Fanxu Meng, Xiaowei Guo, Feiyue Huang, and Xing Sun. On the consistency training for open-set semi-supervised learning. *arXiv preprint arXiv:2101.08237*, 2021.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised neural text classification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 983–992, 2018.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder networks. *arXiv preprint arXiv:1507.02672*, 2015.

Devendra Singh Sachan, Manzil Zaheer, and Ruslan Salakhutdinov. Revisiting lstm networks for semi-supervised text classification via mixed objective function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6940–6948, 2019.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171, 2016.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.

Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Changxin Gao, and Nong Sang. Self-supervised learning for semi-supervised temporal action proposal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1905–1914, 2021.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation. *arXiv preprint arXiv:1904.12848*, 2019.

Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pp. 11525–11536. PMLR, 2021.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. 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)*, pp. 3620–3630, 2019.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *International conference on machine learning*, pp. 3881–3890. PMLR, 2017.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1Ddp1-Rb`.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657, 2015.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1): 44–53, 2018.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pp. 912–919, 2003.

Appendix

The Appendix is organized as follows.

- Section A presents the detailed derivations for generalization bounds.
  - Section A.1 proves the upper bound for Term-1.
  - Section A.2 proves the lower bound for Term-1.
  - Section A.2 proves the upper bound for Term-2 (also for Theorem 1).
  - Section A.4 proves Lemma 2.
  - Section A.5 proves Corollary 1.
- Section B presents more experimental settings and results.

A Theoretical Results

A.1 Term-1 Upper Bound

$$
\begin{aligned}
& R_{\mathcal{D}}(f) - R_{\widetilde{\mathcal{D}}}(f) \\
=& \int_X \Big(\mathbb{P}(f(X) \neq Y|X) - \mathbb{P}(f(X) \neq \widetilde{Y}|X)\Big)\,\mathbb{P}(X) \\
=& \int_X \Big(\mathbb{P}(f(X) = \widetilde{Y}|X) - \mathbb{P}(f(X) = Y|X)\Big)\,\mathbb{P}(X) \\
\leq& \frac{1}{2}\int_X \Big(\Big|\mathbb{P}(f(X) = \widetilde{Y}|X) - \mathbb{P}(f(X) = Y|X)\Big| + \Big|\mathbb{P}(f(X) \neq \widetilde{Y}|X) - \mathbb{P}(f(X) \neq Y|X)\Big|\Big)\,\mathbb{P}(X) \\
\overset{(a)}{=}& \int_X \text{TD}(\widetilde{Y}_f(X)||Y_f(X))\mathbb{P}(X) \\
\overset{(b)}{\leq}& \int_X \text{TD}(\widetilde{Y}(X)||Y(X))\mathbb{P}(X) \\
=& \frac{1}{2}\int_X \sum_{i\in[K]} \Big|\mathbb{P}(\widetilde{Y} = i|X) - \mathbb{P}(Y = i|X)\Big|\,\mathbb{P}(X) \\
=& \int_X \eta(X)\mathbb{P}(X) \\
=& \eta
\end{aligned}
$$

where (a) holds due to the definition of total variation of two distributions, i.e.,

$$
\text{TD}(P||Q) := \frac{1}{2}\int |dP - dQ|.
$$

Inequality (b) holds due to the data processing inequality. $\eta(X)$ is the accuracy of labels on feature X (in some cases).

A.2 Term-1 Lower Bound

Let $e(X) := \mathbb{P}(Y \neq \widetilde{Y}|X)$ be the feature-dependent error rate, $\widetilde{A}_f(X) := \mathbb{P}(f(X) = \widetilde{Y}|X)$ be the accuracy of prediction $f(X)$ on noisy dataset $\widetilde{D}$. Denote their expectations (over X) by $\bar{e} := \mathbb{E}_X[e(X)]$, $\widetilde{A}_f = \mathbb{E}_X[\widetilde{A}_f(X)]$. We have:

$$
\begin{aligned}
& R_{\mathcal{D}}(f) - R_{\widetilde{\mathcal{D}}}(f) \\
=& \int_X \Big(\mathbb{P}(f(X) \neq Y|X) - \mathbb{P}(f(X) \neq \widetilde{Y}|X)\Big)\,\mathbb{P}(X) \\
=& \int_X \Big(\mathbb{P}(f(X) = \widetilde{Y}|X) - \mathbb{P}(f(X) = Y|X)\Big)\,\mathbb{P}(X) \\
=& \int_X \Big(\mathbb{P}(f(X) = \widetilde{Y}|f(X) = \widetilde{Y}, X) - \mathbb{P}(f(X) = Y|f(X) = \widetilde{Y}, X)\Big)\,\mathbb{P}(f(X) = \widetilde{Y}|X)\mathbb{P}(X) \\
&+ \int_X \Big(\mathbb{P}(f(X) = \widetilde{Y}|f(X) \neq \widetilde{Y}, X) - \mathbb{P}(f(X) = Y|f(X) \neq \widetilde{Y}, X)\Big)\,\mathbb{P}(f(X) \neq \widetilde{Y}|X)\mathbb{P}(X) \\
=& \int_X \Big(1 - \mathbb{P}(\widetilde{Y} = Y|X)\Big)\,\mathbb{P}(f(X) = \widetilde{Y}|X)\mathbb{P}(X) \\
&+ \int_X \Big(0 - \mathbb{P}(f(X) = Y|f(X) \neq \widetilde{Y}, X)\Big)\,\mathbb{P}(f(X) \neq \widetilde{Y}|X)\mathbb{P}(X) \\
=& \int_X e(X)\widetilde{A}_f(X)\mathbb{P}(X) \\
&- \int_X \Big(\mathbb{P}(f(X) = Y|f(X) \neq \widetilde{Y}, Y = \widetilde{Y}, X)\mathbb{P}(Y = \widetilde{Y}|f(X) \neq \widetilde{Y}, X)\Big)\,\mathbb{P}(f(X) \neq \widetilde{Y}|X)\mathbb{P}(X) \\
&- \int_X \Big(\mathbb{P}(f(X) = Y|f(X) \neq \widetilde{Y}, Y \neq \widetilde{Y}, X)\mathbb{P}(Y \neq \widetilde{Y}|f(X) \neq \widetilde{Y}, X)\Big)\,\mathbb{P}(f(X) \neq \widetilde{Y}|X)\mathbb{P}(X) \\
\geq& \int_X e(X)\widetilde{A}_f(X)\mathbb{P}(X) \\
&- \int_X \Big(0 \cdot \mathbb{P}(Y = \widetilde{Y}|f(X) \neq \widetilde{Y}, X)\Big)\,\mathbb{P}(f(X) \neq \widetilde{Y}|X)\mathbb{P}(X) \\
&- \int_X \Big(1 \cdot \mathbb{P}(Y \neq \widetilde{Y}|f(X) \neq \widetilde{Y}, X)\Big)\,\mathbb{P}(f(X) \neq \widetilde{Y}|X)\mathbb{P}(X) \\
=& \int_X e(X)\widetilde{A}_f(X)\mathbb{P}(X) - \int_X \mathbb{P}(Y \neq \widetilde{Y}|X)\mathbb{P}(f(X) \neq \widetilde{Y}|X)\mathbb{P}(X) \\
=& \int_X e(X)(2\widetilde{A}_f(X) - 1)\mathbb{P}(X) \\
=& (2\widetilde{A}_f - 1)e.
\end{aligned}
$$

A.3 Term-2 Upper Bound

We adopt the same technique to prove Theorem 1 and Lemma 3. The proof follows Liu & Guo (2020). We prove Theorem 1 as follows.

Denote the expected error rate of classifier f on distribution $\mathcal{D}$ by

$$R_{\mathcal{D}}(f) := \mathbb{E}_{\mathcal{D}}[\mathbb{1}(f(X), Y)].$$

Let $\hat{f}_D$ denote the classifier trained by minimizing 0-1 loss with clean dataset D, i.e.,

$$\hat{f}_D := \arg\min_f \ \hat{R}_D(f),$$

where

$$\hat{R}_D(f) := \frac{1}{N} \sum_{n \in [N]} \mathbb{1}(f(x_n), y_n).$$

The Bayes optimal classifier is denoted by

$$f^*_{\mathcal{D}} := \arg\min_f R_{\mathcal{D}}(f).$$

The expected error given by $f^*_{\mathcal{D}}$ is written as

$$R^* := R_{\mathcal{D}}(f^*_{\mathcal{D}}).$$

Denote by $Y^*|X := \arg\max_{i\in[K]} \mathbb{P}(Y|X)$ the Bayes optimal label on clean distribution $\mathcal{D}$. With probability at least $1-\delta$, we have:

$$\begin{aligned}
&R_{\mathcal{D}}(\hat{f}_D) - R_{\mathcal{D}}(f^*_{\mathcal{D}}) \\
=&\hat{R}_D(\hat{f}_D) - \hat{R}_D(f^*_{\mathcal{D}}) + \Big(R_{\mathcal{D}}(\hat{f}_D) - \hat{R}_D(\hat{f}_D)\Big) + \Big(\hat{R}_D(f^*_{\mathcal{D}}) - R_{\mathcal{D}}(f^*_{\mathcal{D}})\Big) \\
\leq&0 + 2\max_{f\in\mathcal{F}} |\hat{R}_D(f) - R_{\mathcal{D}}(f)| \\
\leq&\sqrt{\frac{2\log(2/\delta)}{2N}},
\end{aligned}$$

where the last inequality holds due to the Hoeffding's inequality. Noting $R_{\mathcal{D}}(f^*_{\mathcal{D}}) = \mathbb{P}(Y \neq Y^*)$, we can prove Theorem 1. But substituting $\mathcal{D}$ for $\widetilde{\mathcal{D}}$, we can also prove Lemma 3.

A.4 Proof for Lemma 2

Proof. Note

$$\mathbb{P}(\widetilde{Y} \neq Y|X) = 1 - \mathbb{P}(\widetilde{Y} = Y|X) = 1 - \sum_{i\in[K]} \mathbb{P}(\widetilde{Y} = i|X)\mathbb{P}(Y = i|X).$$

and

$$\eta(X) := \frac{1}{2}\sum_{i\in[K]} |\mathbb{P}(\widetilde{Y} = i|X) - \mathbb{P}(Y = i|X)|.$$

Assume $\mathbb{P}(Y = i'|X) = 1$. We have $\mathbb{P}(\widetilde{Y} \neq Y|X) = 1 - \mathbb{P}(\widetilde{Y} = i'|X)$ and

$$\eta(X) := \frac{1}{2}\left(1 - \mathbb{P}(\widetilde{Y} = i'|X) + \sum_{i\in[K], i\neq i'} \mathbb{P}(\widetilde{Y} = i|X)\right) = 1 - \mathbb{P}(\widetilde{Y} = i'|X).$$

□

A.5 Proof for Corollary 1

$$\begin{aligned}
\widehat{\mathsf{BR}}(\mathcal{P}) &= \frac{\eta + \sqrt{\frac{2\log(2/\delta)}{N_{\mathcal{P}}}} + \mathbb{P}(\widetilde{Y} \neq \widetilde{Y}^*) - \sqrt{\frac{2\log(2/\delta)}{N_{\mathcal{P}_L}}} - \mathbb{P}(Y \neq Y^*)}{\sqrt{\frac{2\log(2/\delta)}{N_{\mathcal{P}}}} - \sqrt{\frac{2\log(2/\delta)}{N_{\mathcal{P}_L}}}} \\
&= \frac{\Delta(N_{\mathcal{P}}, N_{\mathcal{P}_L}) - \eta}{\Delta(N_{\mathcal{P}}, N_{\mathcal{P}_L})}.
\end{aligned}$$

B More Experiments

B.1 Experiment Settings

Explicit sub-populations: classification labels In this section, we show the statistics of datasets where we conducted the experiments to demonstrate the disparate impact on the explicit sub-populations (classification labels) in Table 2. In addition, the train/valid/test splitting in the image datasets (CIFAR-10 and CIFAR-100) is 45000:5000:10000. As for the splitting in the text datsets (Yahoo! Answers and AG News), we follow the setting in (Chen et al., 2020).

Table 2: Dataset statistics and train/test splitting. The number of train and test samples means the number of data per class.

Datasets	*Label Type*	*Number of classes*	*Number of train samples*	*Number of test samples*
CIFAR-10	Image	10	5000	1000
CIFAR-100	Image	100	500	100
Yahoo! Answer	Question Answering	10	140000	5000
AG News	News	4	30000	1900

Implicit sub-populations: fine-categories under coarse classification labels In this section, we provide a description about coarse labels in CIFAR-100 dataset when we conducted the experiments to demonstrate the disparate impact on the implicit sub-populations: fine-categories under coarse classification labels. In CIFAR-100, its 100 classes can be categorized into 20 superclasses. Each image can have a "fine" label (the original class) and a "coarse" label (the superclass). The list of superclasses in the CIFAR-100 is as follows: aquatic mammals, fish , flowers orchids, food containers bottles, fruit and vegetables, household electrical devices, household furniture bed, large carnivores, large man-made outdoor things, large natural outdoor scenes, large omnivores and herbivores, medium-sized mammals, non-insect invertebrates, people, reptiles, small mammals, trees, vehicles 1, and vehicles 2.

Implicit sub-populations: demographic groups (race & gender) In this section, we describe the detailed data distribution before and after balancing methods about implicit sub-populations: demographic groups (race & gender on Jigsaw dataset) . The identities in the race sub-population includes 'asian', 'black', 'latin', and 'white'. In each identity, we consider different classification labels (toxic and non-toxic). Therefore, we have eight identities in the race sub-population. The numbers of samples on these eight race identities in the unbalanced Jigsaw dataset are 'asian (toxic): 466; black (toxic): 1673; latin (toxic): 282; white (toxic): 3254; asian (non-toxic): 633; black (non-toxic): 1879; latin (non-toxic): 511; white (non-toxic): 2652'. The numbers of samples on these eight race identities in the balanced Jigsaw dataset are 'asian (toxic): 1419; black (toxic): 1419; latin (toxic): 1419; white (toxic): 1419; asian (non-toxic): 1419; black (non-toxic): 1419; latin (non-toxic): 1419; white (non-toxic): 1419'. Similar to race, we also have eight identities in the gender sub-population by considering four different gender types which are 'male', 'female', 'male femal' and 'transgender' and two classification labels (toxic and non-toxic). The numbers of samples on these eight gender identities in the unbalanced Jigsaw dataset are 'male (toxic): 5532; female (toxic): 4852; male female (toxic): 1739; transgender (toxic): 405; male (non-toxic): 5167; female (non-toxic): 4985; male female (non-toxic): 2002; transgender (non-toxic): 374'. The numbers of samples on these eight gender identities in the balanced Jigsaw dataset are 'male (toxic): 3142; female (toxic): 3142; male female (toxic): 3142; transgender (toxic): 3142; male (non-toxic): 3142; female (non-toxic): 3142; male female (non-toxic): 3142; transgender (non-toxic): 3142'. In either race or gender, Jigsaw dataset contains more than four identities, we set the number of identities to 4 due to the relatively small number of other identities. In addition, the ratio of 'train:valid:test' on either race or gender sub-population case is '8:1:1'.

B.2 More Experimental Results for Disparate Impact across Different Classes on AG News

In this work, we also conducted experiments on the AG News text classification dataset. Figure 4 demonstrates the disparate impact on model accuracy across different classification labels on the AG News dataset.

B.3 More Experimental Results for Disparate Impact across Different Sub-populations on Jigsaw (gender)

In this work, we also conducted experiments on the Jigsaw Toxicity text classification dataset with gender sub-populations. Figure 5 shows the disparate impact on model accuracy across different sub-populations (gender) on the Jigsaw Toxicity dataset.

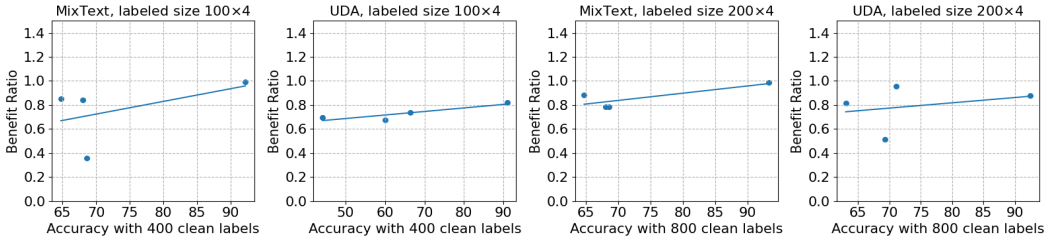


Figure 4: Benefit ratios of different class labels (y-axis) versus baseline accuracies before SSL (x-axis) on AG News

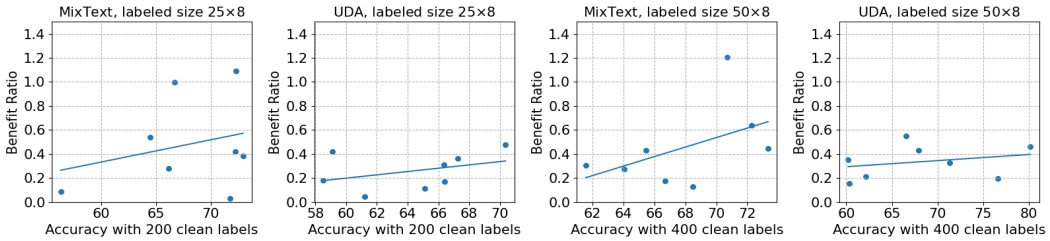


Figure 5: Benefit ratios of different gender attributes (y-axis) versus baseline accuracies before SSL (x-axis) on Jigsaw.

B.4 More Experimental Results of "Balancing" Treatment to Disparate Impact

In this section, we show more experimental results of "Balancing" treatment to disparate impact described in Section 5.2. In Table 3, we compare the changes of mean and standard deviation on the model accuracy and our proposed benefit ratio in three settings: (a) SSL with unbalanced data $\rightarrow$ (b) SSL with balanced labeled datasets only (by resampling) $\rightarrow$ (c) SSL with both balanced labeled & unlabeled datasets. Setting (c) is not practical due to the lack of labels in unlabeled data but it shows the best performance that this simple resampling can achieve.

Table 3: Balance samples with resampling. We sample unbalanced CIFAR-100 (1:2 means 40 instances per class for the first 50 fine classes, 20 instances per class for the next 50 fine classes). Jigsaw is naturally unbalanced. For race, we consider asian, black, latin, and white identities. For gender, we consider male, female, male female, and transgender identities. The rounded ratios on the number of samples between different identities in race and gender are listed in the table.

Datasets	*Mean*	*Standard Deviation*
CIFAR-100 1:2 Accuracy	69.91 $\rightarrow$ 69.98 $\rightarrow$ **69.99**	22.26 $\rightarrow$ 21.48 $\rightarrow$ **21.46**
CIFAR-100 1:2 Benefit Ratio	55.16 $\rightarrow$ 55.19 $\rightarrow$ **56.09**	20.85 $\rightarrow$ 20.16 $\rightarrow$ **19.17**
CIFAR-100 1:5 Accuracy	61.90 $\rightarrow$ 62.76 $\rightarrow$ 63.21	29.50 $\rightarrow$ 27.73 $\rightarrow$ 26.97
CIFAR-100 1:5 Benefit Ratio	44.67 $\rightarrow$ 49.47 $\rightarrow$ 47.40	37.07 $\rightarrow$ 33.82 $\rightarrow$ 32.18
Jigsaw (race) (1.4:4.5:1:7.5) Accuracy	66.71 $\rightarrow$ 67.25 $\rightarrow$ **67.88**	25.64 $\rightarrow$ 21.33 $\rightarrow$ 17.84
Jigsaw (race) (1.4:4.5:1:7.5) Benefit Ratio	12.46 $\rightarrow$ 13.09 $\rightarrow$ **13.67**	19.28 $\rightarrow$ 17.29 $\rightarrow$ **15.29**
Jigsaw (gender) (13.7:12.6:4.8:1) Accuracy	66.03 $\rightarrow$ 66.74 $\rightarrow$ 67.17	12.90 $\rightarrow$ 9.75 $\rightarrow$ **8.68**
Jigsaw (gender) (13.7:12.6:4.8:1) Benefit Ratio	11.90 $\rightarrow$ 12.18 $\rightarrow$ 12.54	88.52 $\rightarrow$ 50.67 $\rightarrow$ 34.24

B.5 More Experimental Results of "Collecting More Labeled Data" Treatment to Disparate Impact

In this subsection, we show more experimental results of "Collecting more labeled data" treatment to disparate impact described on Section 5.2. Table 4 demonstrates the changes of mean and standard deviation on model accuracy and benefit ratio with the varying number (grow exponentially from left to right) of labeled data on CIFAR-100 and Jigsaw (race & gender). As shown in Table 4, the

mean accuracy tends to be higher and the standard deviation becomes lower when more labeled data is utilized to train the model, which shows the effectiveness of collecting more labeled data.

Table 4: Collect more data. $a \rightarrow b \rightarrow c \rightarrow d$: stands for the change of mean or standard deviation with different sizes of labeled dataset. For CIFAR-100, the sizes are $5\times100, 10\times100, 20\times100, 40\times100$. For Jigsaw (both race and gender), the sizes are $25 \times 8, 50 \times 8, 100 \times 8, 150 \times 8$.

Datasets	*Mean*	*Standard Deviation*
CIFAR-100 Accuracy	$52.87 \rightarrow 60.93 \rightarrow 68.01 \rightarrow$ **74.69**	$32.38 \rightarrow 28.59 \rightarrow 23.52 \rightarrow$ **16.76**
CIFAR-100 Benefit Ratio	$48.86 \rightarrow 54.01 \rightarrow 56.40 \rightarrow$ **66.20**	$33.39 \rightarrow 28.62 \rightarrow 22.11 \rightarrow$ **16.87**
Jigsaw (race) Accuracy	$64.88 \rightarrow 67.88 \rightarrow 72.25 \rightarrow 73.50$	$29.55 \rightarrow 17.84 \rightarrow 9.93 \rightarrow 6.86$
Jigsaw (race) Benefit Ratio	$12.09 \rightarrow 13.67 \rightarrow 14.52 \rightarrow 15.12$	$18.29 \rightarrow 15.29 \rightarrow 11.69 \rightarrow$ **7.64**
Jigsaw (gender) Accuracy	$64.50 \rightarrow 67.17 \rightarrow 73.50 \rightarrow$ **74.83**	$20.15 \rightarrow 8.68 \rightarrow 6.41 \rightarrow$ **5.99**
Jigsaw (gender) Benefit Ratio	$7.80 \rightarrow 12.54 \rightarrow 22.46 \rightarrow$ **25.81**	$41.29 \rightarrow 34.24 \rightarrow 27.30 \rightarrow 13.08$