# Shedding light on underrepresentation and Sampling Bias in machine learning

Sami Zhioua, Rūta Binkytė

# SHEDDING LIGHT ON UNDERREPRESENTATION AND SAMPLING BIAS IN MACHINE LEARNING

**Sami Zhioua**
INRIA, LIX, École Polytechnique
Palaiseau, Paris, France
zhioua@lix.polytechnique.fr

**Rūta Binkytė**
INRIA, LIX, École Polytechnique
Palaiseau, Paris, France
ruta.binkyte@inria.fr

**ABSTRACT**

Accurately measuring discrimination is crucial to faithfully assessing fairness of trained machine learning (ML) models. Any bias in measuring discrimination leads to either amplification or underestimation of the existing disparity. Several sources of bias exist and it is assumed that bias resulting from machine learning is born equally by different groups (e.g. females vs males, whites vs blacks, etc.). If, however, bias is born differently by different groups, it may exacerbate discrimination against specific sub-populations. Sampling bias, is inconsistently used in the literature to describe bias due to the sampling procedure. In this paper, we attempt to disambiguate this term by introducing clearly defined variants of sampling bias, namely, sample size bias (SSB) and underrepresentation bias (URB). We show also how discrimination can be decomposed into variance, bias, and noise. Finally, we challenge the commonly accepted mitigation approach that discrimination can be addressed by collecting more samples of the underrepresented group.

***K*eywords** ML Fairness · Representation Bias · Sampling Bias

## 1 Introduction

With the ubiquitous use of machine learning (ML) systems to inform decisions with critical impacts on human lifes (e.g. job hiring, college admission, security screening), fairness is emerging as an important requirement for the safe use of these technologies. A failure to guarantee fairness may create or amplify discrimination against individuals or specific sub-populations (e.g. minority groups). Such anomaly can initiate a vicious cycle that can be perpetuated and eventually resulting in severe consequences.

Discrimination in ML decisions can originate from several types of bias as described in the literature. For instance, The Centre for Evidence-Based Medicine (CEBM) at the University of Oxford is maintaining a list of 62 different sources of bias [22]. More related to ML, Mehrabi et al. [20] classify the sources of bias into three categories depending on when the bias is introduced in the automated decision loop. For instance, measurement bias [14, 24] can be introduced at the data generation step and is a result of measuring a feature using a proxy variable instead of an ideal variable (e.g. using SAT score variable as a measure for the qualification feature).

Another, more common, category of bias occurs when the ML model is trained using a limited number of samples. This produces an inaccurate model and the inaccuracy will typically be born differently by different sub-populations which leads to a discrimination. Two famous examples of ML discrimination fall into this category of bias. The first is COMPAS software [7] used by several states in US to help predict whether a defendant will recidivate in the next two years if she is released. The software is found to be discriminatory against african-americans as the false positive rate (FPR) was higher for african-americans compared to other ethnicities, but the false negative rate (FNR) was lower [1]. The second example is related to face recognition technology (FRT). Buolamwini et al. [3] found that several commercial FRT software have a significantly lower accuracy for individuals belonging to a specific sub-population, namely, dark-skinned females.

This category of bias is inconsistently given various names in the literature (e.g. sampling bias, representation bias, data imbalance bias, etc.) and, to the best of our knowledge, is not formally defined. This paper is an attempt to

disambiguate this category of bias by proposing definitions of two sources of bias, namely, sample size bias (SSB) and underrepresentation bias (URB). SSB is the bias that results from training an ML model using a training data with a limited number of samples and where all sub-populations are represented in the same proportions as the real population. URB is the bias resulting from training an ML model using a training data with a disparity in the number of samples corresponding to each sub-population.

Although the link between the limited number of samples used for training and the disparity in the accuracy of the obtained model may seem straightforward, the magnitude of such pattern has not been thoroughly studied in the ML fairness literature. Based on the proposed definitions of SSB and URB, the empirical part of the paper tries to illustrate how the magnitude of discrimination behaves as more extreme versions of bias are considered. Several metrics of discrimination are used, namely, difference in $FPR$ (false positive rate), equal opportunity [12], difference in $ZOL$ (zero-one loss), difference in $AUC$ (area under the curve), statistical disparity [10], and, for regression problems, difference in $MSE$ (mean squared error). For the latter, we use previous results in the literature [6, 9] to decompose the discrimination into noise, bias, and variance.

The very definition of sampling bias suggests that it can be mitigated by simply using more data for training, in particular for the under represented groups. Obtaining more data is possible either through data augmentation (duplicating or creating synthetic samples) or resuming data collection. Unlike data augmentation, whose effect on discrimination has been the topic of a number of papers, in particular related to computer vision (e.g. [23, 30, 31, 28, 25, 26]), the impact of of collecting more samples on discrimination has not been well studied in the literature. The last part of the paper studies the effect of collecting more samples on discrimination.

The key findings of this paper are the following:

- Discrimination defined in terms of cost/accuracy metrics that consider a trade-off between precision and recall (e.g. $AUC$ and $ZOL$) are more resilient to limited size or imbalanced training sets.
- For extremely small or imbalanced training sets, the variance component of the bias is significant and can significantly alter the fairness conclusions.
- In presence of underrepresented groups, collecting more data samples for the underrepresented group typically amplifies discrimination rather than reduces it.

## 2 Related Work

An exhaustive list of sources of bias can be found in the survey paper of Mehrabi et al. [20] where the authors categorized them into three categories: biases in the data (measurement, representation[1], etc.), biases in the algorithm (algorithmic, evaluation, etc.), and biases introduced by users (historical, self-selection, etc.). This paper focuses on biases in the first category. Suresh et al. [24], however, categorized the sources of bias into five categories depending on where in the machine learning pipeline a bias may be introduced. Although they provide a framework for the machine learning pipeline transformations, sources of bias that can impact the transformations are only described informally (Figure 2 in [24]). Similarly, Hellstrom et al. [14] propose a taxonomy of the sources of bias while categorizing them on the basis of the machine learning pipeline. The main limitation of all previous work on sources of bias is the absence of formal definitions of biases. Therefore, these papers did not include an analysis of the correlation between the magnitude of the bias and the extent of the discrimination.

Sampling bias is related to the known problems of (1) learning using a limited size training set and (2) learning using imbalanced data [13]. Chen et al. [6] studied the effect of the training set sample size on discrimination. They used two discrimination metrics, namely, false positive rate (FPR) and false negative rate (FNR). They found that discrimination according to FNR is much more sensitive to sample size than FPR. To address the problem of imbalanced data, Yan et al. [29] compared existing techniques for balancing data and found that while they achieve better prediction, they tend to exacerbate discrimination. Farrand et al. [11] focused on the impact of using imbalanced data on the accuracy and fairness of the obtained model while learning with privacy (differencial privacy (DP)) constraints. They found that data imbalance has little effect on discrimination (equal opportunity and statistical disparity) until the imbalance between sensitive groups becomes extreme (e.g. $99.9\%$ vs $0.1\%$). Interestingly, the impact is more important for DP-learned models than with non-DP models.

Machine learning loss/error has been first decomposed into bias and variance by Dietterich and Kong [8]. The decomposition did not distinguish between variance and noise and considered noise as part of variance (variance is

[1]The term representation bias in the context of computer vision denotes a different type of bias: the presence of potential "shortcuts" that a model can exploit to accurately predict the label without learning the underlying task [18]. A simple example is when the background environment of a picture can be used to recognize the object of interest.

defined as the difference between loss and bias). They illustrated the decomposition for regression and for classification problems but only for individual examples. Domingos [9] distinguished between noise and variance and extended the decomposition to the expectation over all samples. He considered three loss functions, namely, squared loss, absolute loss, and zero-one loss. For the experimental analysis, Domingos focused on decision trees and k-nearest neighbor (KNN) learning algorithms and studied the effect of some parameters on the loss, namely, pruning parameters, level of the tree, number of rounds in boosting, and the k parameter in KNN. Chen et al. [6] leveraged these previous results to decompose the discrimination between sensitive groups. They considered two types of loss functions, namely, zero-one loss for classification and squared loss for regression. They also studied the effect of increasing training data size on the discrimination. To this end, they assumed that the losses (population and group-specific) have an inverse power-low behavior asymtotically. This allows to predict the discrimination level if training data is augmented with additional samples.

Additional data collection for minority group or data augmentation is well known as a fairness remedy for imbalanced data, particularly in computer vision [23, 30, 31, 28, 25, 26, 3]. Data augmentation can be split into two approaches. First is replicating or generating additional synthetic data points for the minority group that follows the same distribution [15]. Second approach for data augmentation consists of selectively *balancing* the data with respect to the positive label and the sensitive group[4]. In this study we explore the increase in data size, or the size of a sensitive minority group by adding additional samples from the same population distribution, which is closer to the first approach or additional data collection.

## 3 Preliminaries

Let $\mathcal{A}$ be a supervised learning algorithm for learning an unknown function $f : \boldsymbol{\mathcal{X}} \mapsto \mathcal{Y}$ where $\boldsymbol{\mathcal{X}}$ is the input variables space and $\mathcal{Y}$ is the outcome space. Without loss of generality, the outcome random variable $Y$ is assumed to be binary ($\mathcal{Y} = \{0, 1\}$, e.g. accepted/rejected). Let $\mathcal{S} = \{(\mathbf{x}_i, y_i = f(\mathbf{x}_i))\}, i = 1 \dots m$, be a training sample of size $m$. Based on the data sample $\mathcal{S}$, algorithm $\mathcal{A}$ learns a function $\mathcal{A}(\mathcal{S}) = \hat{f}_{\mathcal{S}}^{\mathcal{A}}$. Let $\hat{Y}_{\mathcal{S}}^{\mathcal{A}}$ be the predicted outcome random variable such that $\hat{f}_{\mathcal{S}}^{\mathcal{A}}(\mathbf{x}_i) = \hat{y}_i$. When there is no ambiguity, we refer to $\hat{Y}_{\mathcal{S}}^{\mathcal{A}}$ and $\hat{f}_{\mathcal{S}}^{\mathcal{A}}$ simply as $\hat{Y}$ (or $\hat{Y}_{\mathcal{S}}$) and $\hat{f}$ (or $\hat{f}_{\mathcal{S}}$).

Given the true value $y$ and the prediction $\hat{y}$, $L(y, \hat{y})$ represents the loss incurred by predicting $\hat{y}$ while the true outcome is $y$. A commonly used loss function for regression problems is the squared loss defined as $L^{SL}(\hat{y}, y) = (\hat{y} - y)^2$. Other loss functions that will be considered in this paper are the absolute loss $L^{AL}(\hat{y}, y) = |\hat{y} - y|$ and the zero-one loss $L^{ZO}(\hat{y}, y) = 0$ if $\hat{y} = y$, and 1 otherwise.

Based on a loss function, we define two special predictions, namely, the main prediction for a learning algorithm $\mathcal{A}$ and the optimal prediction.

Given a learning algorithm $\mathcal{A}$ and a set of training samples $\mathfrak{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots\}$, the main prediction random variable $\bar{\hat{Y}}_{\mathfrak{S}}^{\mathcal{A}}$ ($\bar{\hat{y}} = \bar{\hat{f}}_{\mathfrak{S}}^{\mathcal{A}}(\mathbf{x})$) represents the prediction that minimizes the loss across all training sets in $\mathfrak{S}$. That is,

$$\bar{\hat{f}}_{\mathcal{A}}^{\mathfrak{S}}(\mathbf{x}) = \underset{f'}{\operatorname{argmin}}\, \mathbb{E}_{\mathcal{S} \in \mathfrak{S}}[L(\hat{f}_{\mathcal{S}}(\mathbf{x}), f'(\mathbf{x})].$$

When there is no ambiguity, we refer to $\bar{\hat{Y}}_{\mathfrak{S}}^{\mathcal{A}}$ and $\bar{\hat{f}}_{\mathfrak{S}}^{\mathcal{A}}(\mathbf{x})$ simply as $\bar{\hat{Y}}$ and $\bar{\hat{f}}(\mathbf{x})$. Typically, the main prediction corresponds to the average prediction across all training sets in $\mathfrak{S}$. That is,

$$\bar{\hat{f}}(\mathbf{x}) = \underset{\mathcal{S} \in \mathfrak{S}}{\mathbb{E}}\, \hat{f}_{\mathcal{S}}(\mathbf{x})^2. \tag{1}$$

The optimal prediction $Y^*$ ($y^* = f^*(\mathbf{x})$) is the prediction that minimizes the loss across all possible predictors. That is,

$$f^*(\mathbf{x}) = \underset{f'}{\operatorname{argmin}}\, \mathbb{E}[L(f(\mathbf{x}), f'(\mathbf{x})].$$

It is important to note that $f^*$ is independent of the learning algorithm $\mathcal{A}$.

Assume that the sensitive attribute $A$ is a binary variable with possible values $A = a_0$ and $A = a_1$, each representing a different group (e.g. male vs female, black vs white, etc.). Let $G_0$ and $G_1$ denote these groups. That is, $G_0 = \{\mathbf{x} \in \mathcal{X} | A = a_0\}$ and $G_1 = \{\mathbf{x} \in \mathcal{X} | A = a_1\}$. Discrimination between $G_0$ and $G_1$ can be defined in terms of the disparity in prediction accuracy. Let $C_a^{\bullet}(\hat{Y})$ denote the accuracy/cost of prediction $\hat{Y}$ for group $A = a$. For classification problems, we consider four metrics, namely, false positive rate ($FPR$), false negative rate ($FNR$), true positive rate ($TPR$), and zero one loss ($ZOL$). For regression problems, we consider mean square error ($MSE$). These metrics are defined as follows:

[2]We exceptionally use the expectation on a function, instead of a random variable.

- $C_a^{FPR}(\hat{Y}) = \mathbb{E}[\hat{Y}|Y=0, A=a]$
- $C_a^{FNR}(\hat{Y}) = \mathbb{E}[1-\hat{Y}|Y=1, A=a]$
- $C_a^{TPR}(\hat{Y}) = \mathbb{E}[\hat{Y}|Y=1, A=a]$
- $C_a^{ZOL}(\hat{Y}) = \mathbb{E}[\mathbb{1}[\hat{Y} \neq Y]|A=a]$
- $C_a^{MSE}(\hat{Y}) = \mathbb{E}[(\hat{Y}-Y)^2|A=a]$

Discrimination $Disc^{\bullet}$ can be defined as the difference in $C_a^{\bullet}$ between the two sensitive groups. For instance $Disc^{FPR}(\hat{Y}) = C_{a_1}^{FPR}(\hat{Y}) - C_{a_0}^{FPR}(\hat{Y})$. Notice that $Disc^{TPR}(\hat{Y})$ corresponds to discrimination according to equal opportunity [12] and that $Disc^{TPR}(\hat{Y}) = -Disc^{FNR}(\hat{Y})$ as $TPR = 1 - FNR$. In the rest of the paper, we use $Disc^{TPR}(\hat{Y})$ and $Disc^{EO}(\hat{Y})$ interchangeably. In addition, for reference, we use $Disc^{SD}(\hat{Y}) = \mathbb{E}[\hat{Y}|A=a_1] - \mathbb{E}[\hat{Y}|A=a_0]$ to denote statistical disparity [10].

## 4 Sample Size and Underrepresentation Biases

Typically, the size of the data used to train an ML model has a significant impact on the accuracy of the obtained model. However, it is generally assumed that the loss in accuracy is equally born by the different segments of the data. As it is not usually the case, we define sample size bias (SSB) as the bias resulting from training a model with a given data size.

Let $\mathfrak{S}_{\mathfrak{m}} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots\}$ be the set of samples of size $m$, and let $\hat{f}_{\mathcal{S}_1}, \hat{f}_{\mathcal{S}_2}, \ldots$ be the models produced by applying the learning algorithm $\mathcal{A}$ on each sample ($\mathcal{A}(\mathcal{S}_1) = \hat{f}_{\mathcal{S}_1}$, etc.). Let $\bar{\hat{Y}}^{\mathcal{A}}_{\mathfrak{S}_{\mathfrak{m}}}$ ($\bar{\hat{y}}_m = \bar{\hat{f}}^{\mathcal{A}}_{\mathfrak{S}_m}(\mathbf{x})$) be the main prediction obtained using the set of training sets $\mathfrak{S}_{\mathfrak{m}}$. That is,

$$\bar{\hat{f}}^{\mathcal{A}}_{\mathfrak{S}_m}(\mathbf{x}) = \operatorname*{argmin}_{f'} \mathbb{E}_{\mathcal{S}\in\mathfrak{S}_m}[L(\hat{f}_{\mathcal{S}}(\mathbf{x}), f'(\mathbf{x})]. \tag{2}$$

When there is no ambiguity, we refer to $\bar{\hat{Y}}^{\mathcal{A}}_{\mathfrak{S}_{\mathfrak{m}}}$ and $\bar{\hat{f}}^{\mathcal{A}}_{\mathfrak{S}_m}$ simply as $\bar{\hat{Y}}_m$ and $\bar{\hat{f}}_m$.

**Definition 4.1.** Given a positive number $m > 0$ representing the training set size, sample size bias is the difference in discrimination due to the training set size:

$$SSB^{\bullet}(\mathcal{A}, m) = Disc^{\bullet}(\bar{\hat{Y}}_m) - Disc^{\bullet}(\bar{\hat{Y}}_{\infty}) \tag{3}$$

where $Disc^{\bullet}(\bar{\hat{Y}}_{\infty}) = \lim_{m\to\infty} Disc^{\bullet}(\bar{\hat{Y}}_m)$ and $\bullet$ is a placeholder for the accuracy/cost metric ($FPR, FNR, EO, ZOL$, or $MSE$ for regression problems). As a metric that combines both specificity ($FPR$) and sensitivity ($TPR$), we use also $AUC$ (area under the curve)[3]. For reference, we consider also statistical disparity that we denote as $Disc^{SD}$ (See Appendix A.3).

As $SSB$ is defined in terms of an infinite size training set ($\bar{\hat{Y}}_{\infty}$), we consider an alternative definition in terms of $M$, the size of the largest training set available:

$$SSB^{\bullet}_M(\mathcal{A}, m) = Disc^{\bullet}(\bar{\hat{Y}}_m) - Disc^{\bullet}(\bar{\hat{Y}}_M) \tag{4}$$

Another variant of $SSB$ can be defined based on a specific training set $\mathcal{S}_m$ of size $m$ as follows:

$$SSB^{\bullet}_M(\mathcal{A}, \mathcal{S}_m) = Disc^{\bullet}(\hat{Y}_{\mathcal{S}_m}) - Disc^{\bullet}(\bar{\hat{Y}}_M) \tag{5}$$

When sampling a training set from a population, it is generally assumed that the generated sample is balanced. Data is balanced if all classes are proportionally represented and is imbalanced if it suffers from severe class distribution skews [13]. For instance, if one class label is overrepresented at the expense of another underrepresented class label. If data is imbalanced in the sensitive groups (e.g. male vs female, blacks vs whites, etc.), it can have significant impact on the disparity of accuracies and consequently on discrimination between sensitive groups. We define underrepresentation bias (URB) as the bias resulting from a disparity in representation between the sensitive groups.

Let $\mathfrak{S}_{\mathfrak{m}}^{\frac{\mathfrak{m}_1}{\mathfrak{m}_0}}$ be the set of samples of size $m$ with $m_0$ and $m_1$ items from $G_0$ and $G_1$ respectively. That is, for $\mathcal{S} \in \mathfrak{S}_{\mathfrak{m}}^{\frac{\mathfrak{m}_1}{\mathfrak{m}_0}}$, $|\{\mathbf{x} \in \mathcal{S}|A=a_0\}| = m_0$, $|\{\mathbf{x} \in \mathcal{S}|A=a_1\}| = m_1$, and $m_0 + m_1 = m = |\mathcal{S}|$. We use the simpler notation $\bar{\hat{Y}}_{\frac{m_1}{m_0}}$ to refer to $\bar{\hat{Y}}^{\mathcal{A}}_{\mathfrak{S}_{\mathfrak{m}}^{\frac{\mathfrak{m}_1}{\mathfrak{m}_0}}}$.

[3]Other metrics combining specificity and sensitivity include $F_1$ score and balanced accuracy ($BA$)

**Definition 4.2.** Given, $m, m_0, m_1 > 0$ such that $m_0 + m_1 = m$, underrepresentation bias is the difference in discrimination due to the disparity in sample sizes compared to the population ratio:

$$URB^{\bullet}(\mathcal{A}, m_0, m_1) = Disc^{\bullet}(\bar{\hat{Y}}_{\frac{m_1}{m_0}}) - Disc^{\bullet}(\bar{\hat{Y}}_{\frac{m_1^p}{m_0^p}}) \tag{6}$$

where $Disc^{\bullet}(\bar{\hat{Y}}_{m_1^p/m_0^p})$ is the discrimination of the prediction based on a model trained using only samples from $\mathfrak{S}_{\mathfrak{m}}^{\mathfrak{m}_1^p/\mathfrak{m}_0^p}$, and the ratio $\frac{m_1^p}{m_0^p}$ is the same as the ratio in the population ($\frac{m_1^p}{m_0^p} \approx \frac{|G_1|}{|G_0|}$).

Similar to $SSB_M^{\bullet}(\mathcal{A}, \mathcal{S}_m)$ (Equation 5), a variant of $URB$ can be defined based on a specific training set $\mathcal{S}_{\frac{m_1}{m_0}} \in \mathfrak{S}_{\mathfrak{m}}^{\frac{\mathfrak{m}_1}{\mathfrak{m}_0}}$ as follows:

$$URB^{\bullet}(\mathcal{A}, \mathcal{S}_{\frac{m_1}{m_0}}) = Disc^{\bullet}(\hat{Y}_{\mathcal{S}_{\frac{m_1}{m_0}}}) - Disc^{\bullet}(\bar{\hat{Y}}_{\frac{m_1^p}{m_0^p}}) \tag{7}$$

## 5 Loss and Discrimination Decomposition

Domingos [9] showed that if a learning algorithm $\mathcal{A}$ learns a function $\mathcal{A}(\mathcal{S}) = \hat{f}_{\mathcal{S}}$ based on a training set $\mathcal{S} \in \mathfrak{S}$, then the expected loss between the prediction $\hat{f}_{\mathcal{S}}(\mathbf{x})$ and the true value $f(\mathbf{x})$ can be decomposed into noise, bias, and variance. In particular, for squared loss,

$$L^{SL}(\hat{f}_{\mathcal{S}}(\mathbf{x}), f(\mathbf{x})) = N^{SL}(\mathbf{x}) + B^{SL}(\mathbf{x}) + V^{SL}(\mathbf{x}) \tag{8}$$

where

- $N^{SL}(\mathbf{x}) = L^{SL}(f^*(\mathbf{x}), f(\mathbf{x}))$
- $B^{SL}(\mathbf{x}) = L^{SL}(\bar{\hat{f}}(\mathbf{x}), f^*(\mathbf{x}))$
- $V^{SL}(\mathbf{x}) = \mathbb{E}_{\mathcal{S}\in\mathfrak{S}}[L^{SL}(\hat{f}_{\mathcal{S}}(\mathbf{x}), \bar{\hat{f}}(\mathbf{x}))]$

The loss decomposition can be illustrated as follows:

$$f(\mathbf{x}) \xleftarrow[Noise]{} f^*(\mathbf{x}) \xleftarrow[Bias]{} \bar{\hat{f}}(\mathbf{x}) \xleftarrow[Variance]{} \hat{f}_{\mathcal{S}}(\mathbf{x})$$

For Zero-One loss ($L^{ZO}$), Equation 8 holds also but with coefficients different than 1 for the noise and variance terms. However, it does not hold for the absolute loss ($L^{AL}$)[4] [9]

### 5.1 Decomposing Discrimination

Chen et al. [6] showed that the accuracy/cost metric $C_a^{\bullet}(\hat{Y}_{\mathcal{S}})$ as well as the discrimination $Disc^{\bullet}(\hat{Y}_{\mathcal{S}})$ can be decomposed into noise, bias, and variance components. In particular, for MSE,

$$C_a^{MSE}(\hat{Y}_{\mathcal{S}}) = \overline{N}_a^{SL}(\hat{Y}_{\mathcal{S}}) + \overline{B}_a^{SL}(\hat{Y}_{\mathcal{S}}) + \overline{V}_a^{SL}(\hat{Y}_{\mathcal{S}}) \tag{9}$$

where:

- $\overline{N}_a^{SL}(\hat{Y}_{\mathcal{S}}) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}[N^{SL}(\mathbf{x})|A = a]$
- $\overline{B}_a^{SL}(\hat{Y}_{\mathcal{S}}) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}[B^{SL}(\mathbf{x})|A = a]$
- $\overline{V}_a^{SL}(\hat{Y}_{\mathcal{S}}) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}[(1 - 2 \times B^{SL}(\mathbf{x})) \times V^{SL}(\mathbf{x})|A = a]$

The last term ($V_a(\hat{Y}_{\mathcal{S}})$) is called *net variance* [9]. Consequently,

$$Disc^{MSE}(\hat{Y}_{\mathcal{S}}) = (\overline{N}_{a_1}^{SL}(\hat{Y}_{\mathcal{S}}) - \overline{N}_{a_0}^{SL}(\hat{Y}_{\mathcal{S}})) + (\overline{B}_{a_1}^{SL}(\hat{Y}_{\mathcal{S}}) - \overline{B}_{a_0}^{SL}(\hat{Y}_{\mathcal{S}})) + (\overline{V}_{a_1}^{SL}(\hat{Y}_{\mathcal{S}}) - \overline{V}_{a_0}^{SL}(\hat{Y}_{\mathcal{S}})) \tag{10}$$

The decomposition of Equation 10 will also hold for $Disc^{FPR}(\hat{Y}_{\mathcal{S}})$, $Disc^{EO}(\hat{Y}_{\mathcal{S}})$, and $Disc^{ZOL}(\hat{Y}_{\mathcal{S}})$ but with coefficients different than 1 for the noise and variance terms [6].

[4] Alternatively, upper and lower bounds are possible.

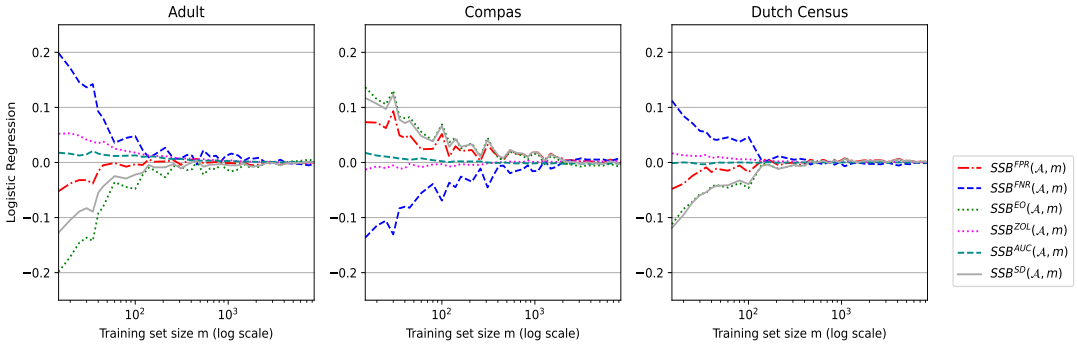


Figure 1: Magnitude of sample size bias (SSB) for increasing size of the training data.

### 5.2 Decomposing $SSB$ and $URB$

The variant $SSB^{\bullet}_{M}(\mathcal{A}, \mathcal{S}_m)$ (Eq. 5) of sample size bias has the advantage that it can be decomposed into bias and variance. The decomposition for the $MSE$ metric is as follows.

**Theorem 5.1.** *$SSB^{MSE}_{M}(\mathcal{A}, \mathcal{S}_m)$ can be decomposed into bias and variance components as follows:*

$$\begin{aligned} SSB^{MSE}_{M}(\mathcal{A}, \mathcal{S}_m) = \overline{B}^{SL}_{a_1}(\hat{Y}_{\mathcal{S}_m}) - \overline{B}^{SL}_{a_1}(\bar{\hat{Y}}_M) - (\overline{B}^{SL}_{a_0}(\hat{Y}_{\mathcal{S}_m}) - \overline{B}^{SL}_{a_0}(\bar{\hat{Y}}_M)) \\ + \overline{V}^{SL}_{a_1}(\hat{Y}_{\mathcal{S}_m}) - \overline{V}^{SL}_{a_1}(\bar{\hat{Y}}_M) - (\overline{V}^{SL}_{a_0}(\hat{Y}_{\mathcal{S}_m}) - \overline{V}^{SL}_{a_0}(\bar{\hat{Y}}_M)) \end{aligned}$$

*Proof.* The proof follows from Equation 9 and from assuming that the optimal predictor $Y^*$ coincides with the true value $Y$ and hence noise is 0 [5]. □

$URB^{\bullet}(\mathcal{A}, \mathcal{S}_{\frac{m_1}{m_0}})$ (Equation 7) can also be decomposed into bias and variance components. The decomposition for the $MSE$ metric is as follows.

**Theorem 5.2.**

$$\begin{aligned} URB^{MSE}(\mathcal{A}, \mathcal{S}_{\frac{m_1}{m_0}}) = \overline{B}^{SL}_{a_1}(\hat{Y}_{\mathcal{S}_{\frac{m_1}{m_0}}}) - \overline{B}^{SL}_{a_1}(\bar{\hat{Y}}_{\frac{m_1^p}{m_0^p}}) - (\overline{B}^{SL}_{a_0}(\hat{Y}_{\mathcal{S}_{\frac{m_1}{m_0}}}) - \overline{B}^{SL}_{a_0}(\bar{\hat{Y}}_{\frac{m_1^p}{m_0^p}})) \\ + \overline{V}^{SL}_{a_1}(\hat{Y}_{\mathcal{S}_{\frac{m_1}{m_0}}}) - \overline{V}^{SL}_{a_1}(\bar{\hat{Y}}_{\frac{m_1^p}{m_0^p}}) - (\overline{V}^{SL}_{a_0}(\hat{Y}_{\mathcal{S}_{\frac{m_1}{m_0}}}) - \overline{V}^{SL}_{a_0}(\bar{\hat{Y}}_{\frac{m_1^p}{m_0^p}})) \end{aligned}$$

*Proof.* The same as Theorem 5.1. □

## 6 Experimental Analysis

The objective of the experimental analysis is to observe the magnitude of both types of biases, namely, sample size bias $SSB$ and underrepresenation bias $URB$ as we change the parameters of data sampling. For $SSB$, we train the predictor model using training sets of increasing sizes. For $URB$, we play rather on the proportions of sensitive groups in the training set. Three benchmark datasets are used, Adult [16], Compas [1], and Dutch Census [21][6].

### 6.1 Magnitude of sample size bias ($SSB$)

To observe how sample size bias behaves as the training set size changes, we use the following process. We use a sequence of sample sizes ranging from 10 until a given portion of the full dataset size. For example, for COMPAS, we consider sample sizes ranging from 10 to 2000. For each sample size value $m$, we repeat the sampling several times (30 by default) so that we obtain 30 samples of each size $m$. Then, we train a different model using each one of the samples so that we obtain 30 models for each size $m$. We finally compute the dicrimination using each model and the returned value is the average discrimination across all models. This procedure gives a sequence of discrimination values

[5] We follow previous work (Domingos [9] and Kohavi and Wolpert [17]) in assuming a zero noise.

[6] We use the same dataset versions and learning algorithms parameters as IBM AIF360 [2]

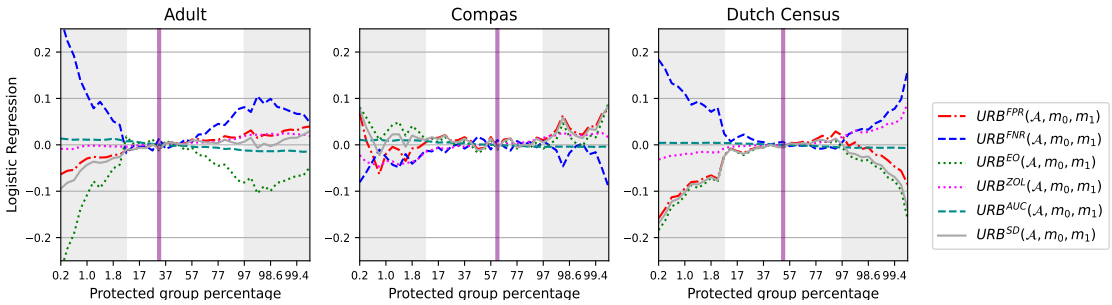


Figure 2: Underrepresentation Bias (URB) for different ratios of sensitive groups. The training set size is fixed (1000). The horizontal bar represents the same ratio as the population. The shaded sections indicate a focus on the extreme proportions (less than 2% and more than 98%).

indexed by the size. We consider five cost/accuracy metrics, namley, $FPR$ (false positive rate), $FNR$ (false negative rate), $EO$ (equal opportunity), $ZOL$ (zero one loss), and $SD$ (statistical disparity). We use five classifiers, namely, logistic regression, decision tree, random forest, nearst neighbor, and support vector machine (SVM).

Figure 1 shows the magnitude of SSB according to each metric and for each benchmark dataset and using logistic regression. Notice that $SSB^{EO}$ and $SSB^{FNR}$ are symmetric because, as mentioned above, $FNR = 1 - TPR$ and hence $SSB^{EO} = -SSB^{FNR}$. Most of the plots exhibit an expected behavior of SSB. That is, the bias is significant when the models are trained using a limited size training set. The bias disappears gradually as the training set size increases. $SSB$ behaves the same way for the other classifiers (Figure 7 in Appendix A.1). More importantly, $SSB$ results show that cost/accuracy metrics that combine specificity and sensitivity ($AUC$ and $ZOL$) are less sensitive to the training set size than the remaining metrics ($FPR$ and $EO$). A possible explanation is that for small training sets, it is more likely that a majority of the samples have the same outcome (positive or negative) which can boost precision on the expense of recall or the opposite. $AUC$ and $ZOL$ are not subject to such skewness since they consider the trade-off between precision and recall.

### 6.2 Magnitude of underrepresentation bias ($URB$)

The aim for underrepresentation bias experiment is to observe the magnitude of $URB$ while the ratio of the sensitive groups in the training set is changing. We consider different values of the splitting $\frac{m_1}{m_0}$ (see Definition 4.2) (e.g. 0.1 vs 0.9, 0.2 vs 0.8, etc.). However, as $URB$ is more significant for extreme disparities, we focus more on extreme splitting values (e.g. 0.001 vs 0.99, 0.002 vs 0.98, etc.). A similar behavior has been observed previously by Farrand et al. [11]. Assuming a fixed sample size (e.g. 1000), for each splitting value, we sample the data so that the proportions of sensitive groups (e.g. male vs female) match the splitting value. Similarly to the $SSB$ experiment, we repeat the sampling several times (30 by default) for the same splitting value. Then, we train a different model using each one of the samples so that we obtain 30 models for each splitting value $\frac{m_1}{m_0}$. The discriminations obtained using the different models are then averaged across all models. We finally obtain a sequence of discrimination values indexed by the splitting value. Figure 2 shows how URB changes as the proportion of the sensitive group increases for the same three datasets and for using logistic regression as learning algorithm. The purple vertical bar indicates the percentage of the sensitive group in the entire dataset (population). For instance, for adult dataset, the percentage of females is 31%. The shaded parts in the background of Figure 2's plots indicate that we are "zooming" on the extreme values (the plots are using different steps for the shaded and unshaded parts[7]). Almost all plots exhibit the same pattern for $URB$, that is, the further the proportions of sensitive groups are from the population proportions reference (vertical bar), the higher is the bias. The same expected behavior for $URB$ is obtained when using the other classifiers (Figure 8 in Appendix A.1). The resilience of $AUC$ and $ZOL$ metrics to extreme training set sizes holds also for imbalanced training sets. Notice that $URB^{AUC}$ and $URB^{ZOL}$ remain stable even for extremely imbalanced training sets.

### 6.3 Bias Decomposition

Section 5 shows that loss and discrimination can be decomposed into variance, bias, and noise. In particular, assuming that the optimal prediction ($Y^*$) coincides with the correct outcome ($Y$), Theorems 5.1 and 5.2 illustrate how $SSB_M^{MSE}(\mathcal{A}, \mathcal{S}_m)$ and $URB^{MSE}(\mathcal{A}, \mathcal{S}_{\frac{m_1}{m_0}})$ can be decomposed into variance and bias components. To illustrate the

[7]The step is very small below 2% and above 98%.

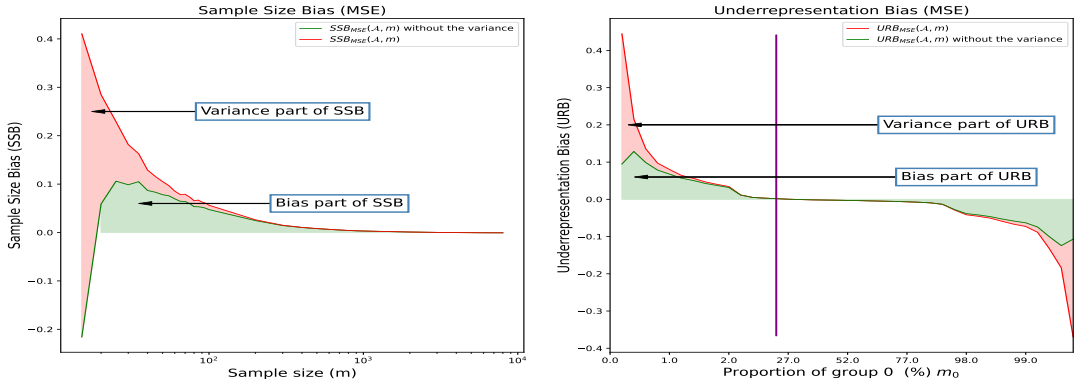


Figure 3: Decomposing $SSB^{MSE}$ (left plot) and $URB^{MSE}$ (right plot). The models are trained using linear regression. The benchmark dataset is Law School [27].

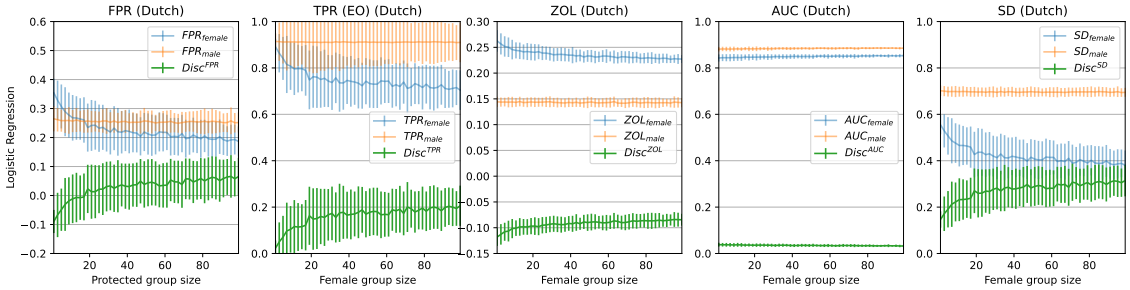


Figure 4: Discrimination while augmenting the training set with female group samples randomly. The male group size is fixed at 100. Dataset is Dutch Census and training algorithm is logistic regression.

decomposition empirically, we use the Law School benchmark dataset [27] which tracked some twenty-seven thousand law students through law school and graduation and where the sensitive attribute is gender and the outcome is the first year GPA. We use the scikit-learn linear regression algorithm to train different models using different size training sets. For $SSB$, the training size $m$ ranges from 10 to $10,000$. For $URB$, the training set size ($m$) is fixed at 1000, but the proportion of the protected group (female) is ranging from $0.1\%$ to $99.9\%$. For each training set size, the training and testing is repeated 30 times. Figure 3 shows how $SSB$ and $URB$ are decomposed into variance and bias. For $SSB$, the variance component is so significant when the training set is extremely small (less than 20) that it reverses the direction of the bias (in favor of females instead of against female). For $URB$, the variance is also significant when one of the groups is extremely underrepresented, but not to the point of reversing the direction of the bias. The main conclusion out of this emprical result is that for very small or very imbalanced training sets, $SSB$ and $URB$ variance can be so important that it can lead to unreliable conclusions about discrimination.

### 6.4 Effect of collecting more samples on discrimination

The natural approach to address sampling bias is to use more data for training, in particular for the under-represented groups. Obtaining more data is possible either through data augmentation or data collection. Data augmentation is the process of using the available data to generate more samples. In turn, this can be done in two ways: oversampling or creating fake samples. Oversampling consists in duplicating existing samples to balance the data. A simple variant is to randomly duplicate samples from the under represented group. Creating fake samples, on the other hand, is typically done using SMOTE [5]. SMOTE creates synthetic samples based on the k-nearest neighbors of every sample of the under represented group. Both techniques of data augmentation try to balance data by adding artificially generated samples. While this artificial manipulation may reduce discrimination between sensitive groups, it can lead to models which are not faithful to reality. When it is possible, collecting more data is more natural and reflects better reality. The approach is simple: if a sensitive group is under represented, collect more samples of that group. Unlike data augmentation, whose effect on discrimination has been the topic of a number of papers, in particular related to computer

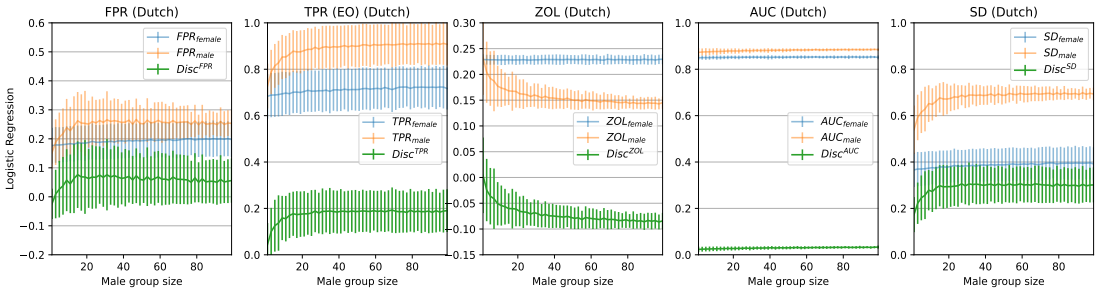


Figure 5: Discrimination while augmenting the training set with male group samples randomly. The female group size is fixed at 100. Dataset is Dutch Census and training algorithm is logistic regression.

vision (e.g. [23, 30, 31, 28, 25, 26]), the impact of of collecting more samples on discrimination has not been well studied in the literature.

In the following, we devise simple experiments to observe the effect of populating the data with more samples collected from the same population as the existing data. Using the same benchmark datasets, the aim is to train models based on an increasing number of under represented group samples while keeping the privileged group portion unchanged. For the particular case of Dutch Census dataset, we train models using a set composed of a fixed 100 privileged group (male) samples and an increasing number of protected group (female) samples starting from 2 until 100 (perfect balance between groups). Similarly to the SSB and URB experiments, it turns out that the magnitude of discrimination is manifested more with extreme values of protected groups sizes (typically less than 100) which explains the specific sample sizes considered. Figure 4 shows how the cost/accuracy metric values for each group, as well as the corresponding difference (discrimination) are changing as more protected group samples are considered for model training. We use 3-fold cross-validation and since we randomly generate 50 different samples for every size value, the plots are shown with error bars. As expected, the cost/accuracy metric value for male group maintains the same mean while for female group it is changing. Interestingly, according to all cost/accuracy metrics (except AUC), discrimination is increasing as data is more balanced. Figure 4 shows the results with logistic regression, but the pattern is similar for other classification algorithms (Figure 9 in Appendix A.2) and for other benchmark datasets (Figure 10 in Appendix A.2). This counterintuitive behavior is also observed for the reverse experiment where the protected group (female) sample size is fixed (100 samples) while the privileged group (male) is under represented and more samples are collected and considered in the training (Figure 5). It is important to mention that in all previous experiments, selecting samples to balance the training set is performed randomly to simulate, as accurately as possible, data collection in real scenarios. The fairness enhancing potential of adding more samples for the sensitive group depends on the initial fairness characteristics of the data and the goal of the classifier. Wang et al. [25] point out that adding more samples of the minority group to the data increases predictive accuracy and fairness specifically in the classification tasks, where sensitive attribute is part of the output of classification, for example face recognition [3].

If, however, training set is balanced by selecting a specific type of samples, in particular, protected group samples with positive outcome, discrimination will be decreasing as data gets balanced (Figure 6). In all three experiments (collecting more protected group samples randomly, collecting more unprotected group samples randomly, and collecting only positive outcome protected group samples), the importance of the sensitive feature (Sex) in the prediction (*shap* explanation [19]) behaves the same way (Figure 11 in Appendix A.2), that is, it contributes more to the learned model as the data is more balanced.

## 7 Conclusion

A very common source of bias in machine learning is to use a limited size or imbalanced training set. This paper defines $SSB$ and $URB$ to capture these variants. In the light of empirical analysis on benchmark datasets and using off the shelf classification algorithms, we made three important observations. First, discrimination metrics defined using $AUC$ and $ZOL$ (which consider the trade-off between precision and recall) are more resilient to sampling biases than discrimination defined using $FPR$ and $TPR$ (equal opportunity). Consequently, in presence of limited size or imbalanced training data, it is recommended to use fairness metrics based on the trade-off between precision and recall (e.g. equalized odds [12]) to reliably estimate discrimination. Second, for regression problems, discrimination defined in terms of $MSE$ is significantly affected by variance for extremely small or imbalanced training sets. Hence, it is recommended to treat discrimination values with caution in such cases. Third, collecting more samples of the

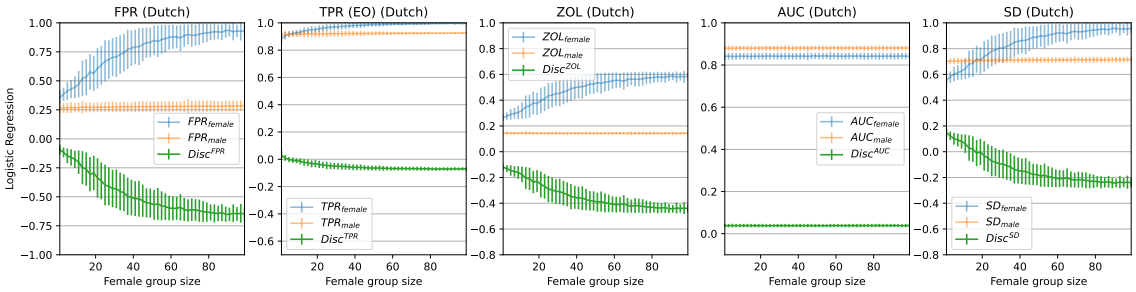


Figure 6: Discrimination while augmenting the training set with only positive outcome female group samples. The male group size is fixed at 100. Dataset is Dutch Census and training algorithm is logistic regression.

extremely underrepresented group according to the population distribution will typically amplify discrimination rather than reducing it. However, collecting more data, allows to measure discrimination more reliably.

## Acknowledgments

This work was supported by the European Research Council (ERC) project HYPATIA under the European Union's Horizon 2020 research and innovation programme. Grant agreement n. 835294.

## References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. propublica. *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*, 2016.

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

[3] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[4] K. Burns, L. A. Hendricks, K. Saenko, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. *arXiv preprint arXiv:1803.09797*, 2018.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

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

[7] COMPAS. Compas, 2020. `https://www.equivant.com/northpointe-risk-need-assessments/`.

[8] T. G. Dietterich and E. B. Kong. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Citeseer.

[9] P. Domingos. A unified bias-variance decomposition. In *Proceedings of 17th international conference on machine learning*, pages 231–238. Morgan Kaufmann Stanford, 2000.

[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[11] T. Farrand, F. Mireshghallah, S. Singh, and A. Trask. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*, pages 15–19, 2020.

[12] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, Spain, 2016.

[13] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[14] T. Hellström, V. Dignum, and S. Bensch. Bias in machine learning–what is it good for? *arXiv preprint arXiv:2004.00686*, 2020.

[15] V. Iosifidis and E. Ntoutsi. Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, 24:11, 2018.

[16] R. Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

[17] R. Kohavi, D. H. Wolpert, et al. Bias plus variance decomposition for zero-one loss functions. In *ICML*, volume 96, pages 275–83. Citeseer, 1996.

[18] Y. Li and N. Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9572–9581, 2019.

[19] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *NIPS*, pages 4765–4774, 2017.

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

[21] E. S. Nordholt, M. Hartgers, and R. Gircour. The dutch virtual census of 2001. *Analysis and Methodology*, 2004.

[22] U. of Oxford. Catalogue of bias. `https://catalogofbias.org/biases`, 2021. Accessed: 2023-03-30.

[23] I. Pastaltzidis, N. Dimitriou, K. Quezada-Tavarez, S. Aidinlis, T. Marquenie, A. Gurzawska, and D. Tzovaras. Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2302–2314, 2022.

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

[25] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5309–5318, 2018.

[26] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020.

[27] L. F. Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

[28] T. Xu, J. White, S. Kalkan, and H. Gunes. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 506–523. Springer, 2020.

[29] S. Yan, H.-t. Kao, and E. Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1715–1724, 2020.

[30] S. Yucer, S. Akçay, N. Al-Moubayed, and T. P. Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020.

[31] Y. Zhang and J. Sang. Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4346–4354, 2020.

# A Appendix

## A.1 Additional plots for the magnitude of SSB and URB (Sections 6.1 and 6.2)

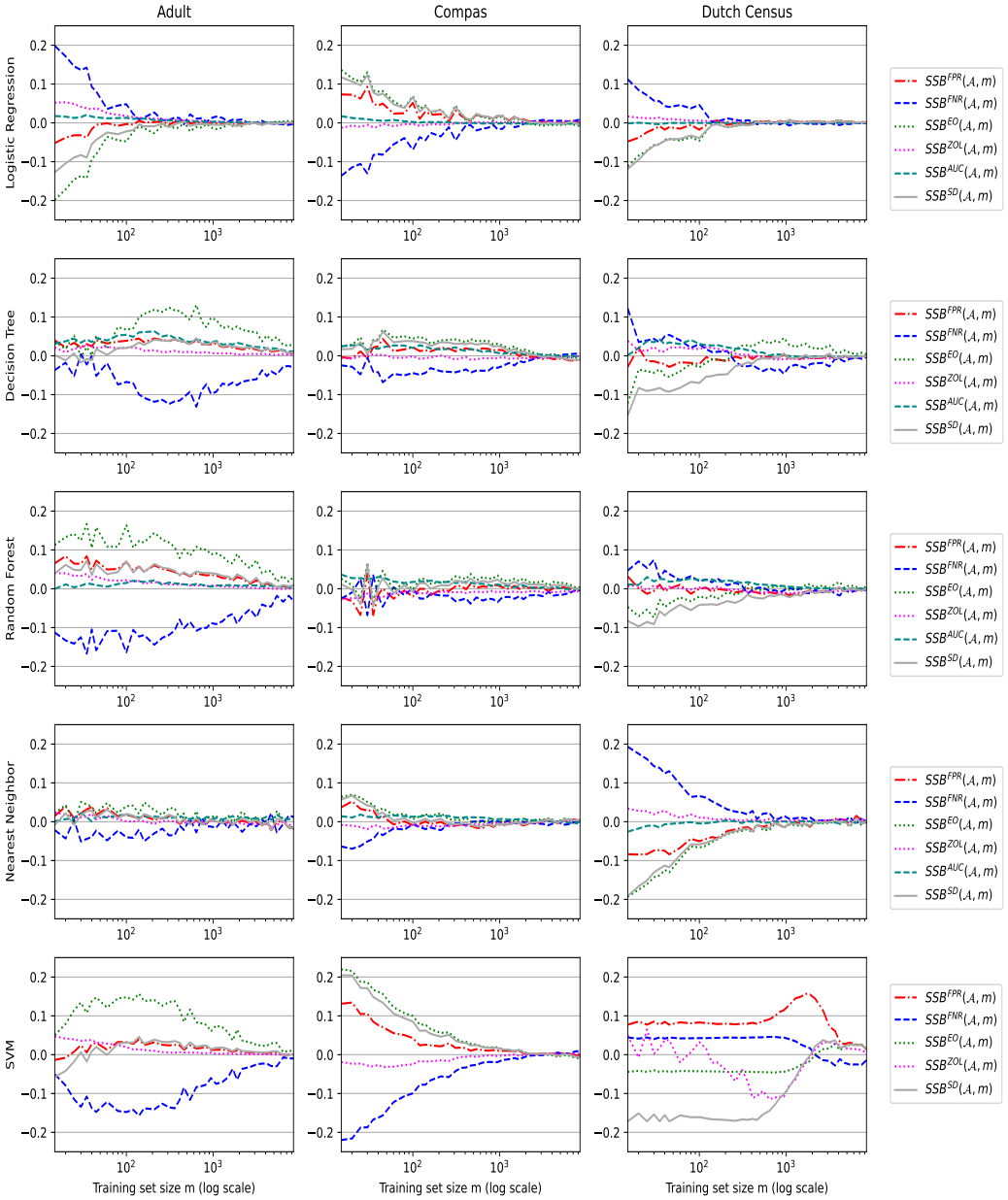


Figure 7: Magnitude of sample size bias (SSB) for increasing size of the training data.

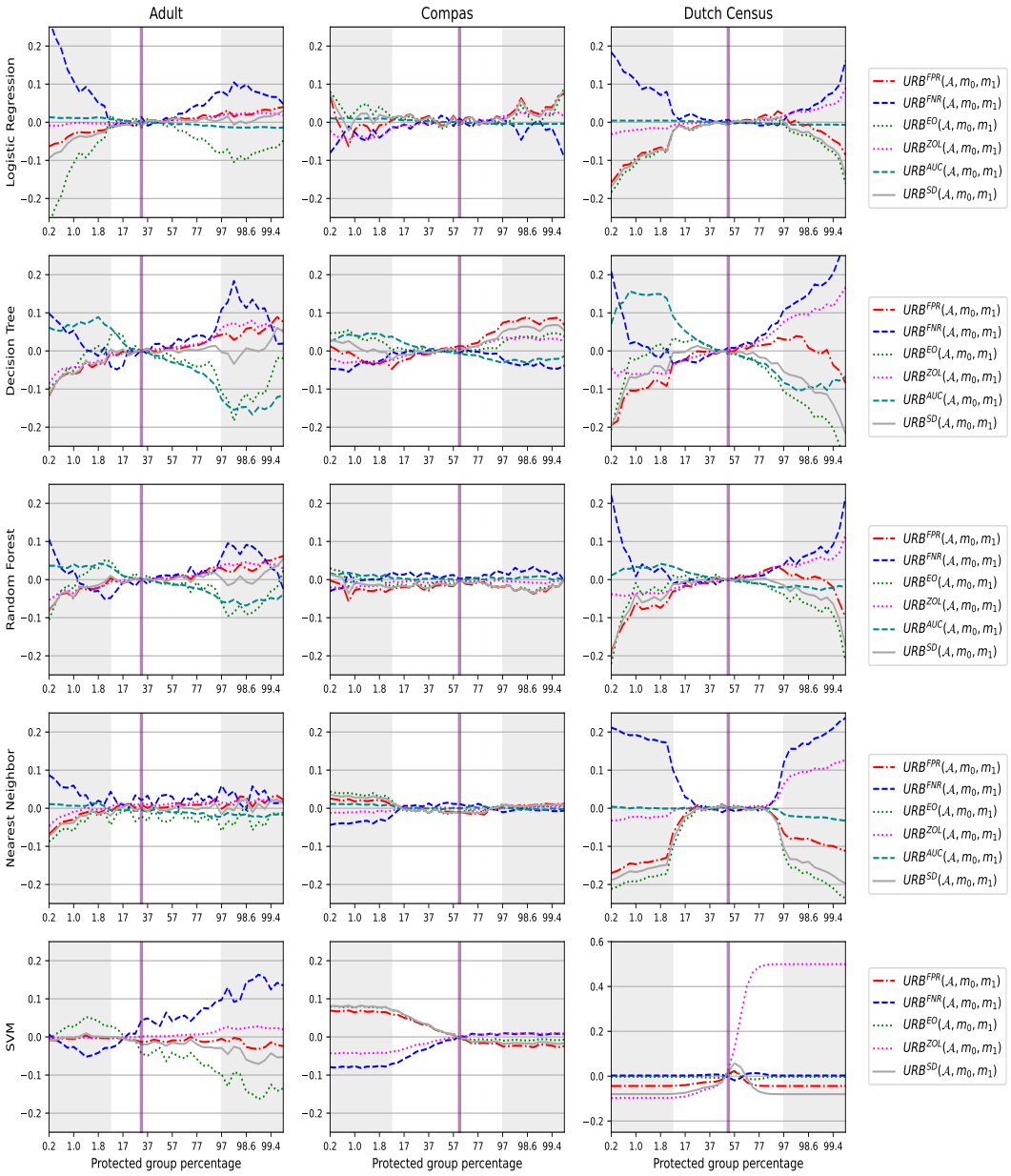


Figure 8: Underrepresentation Bias (URB) for different ratios of sensitive groups. The training set size is fixed (1000). The horizontal bar represents the same ratio as the population. The shaded sections indicate a focus on the extreme proportions (less than 2% and more than 98%).

## A.2 Additional plots for the effect of collecting more samples on discrimination (Section 6.4)

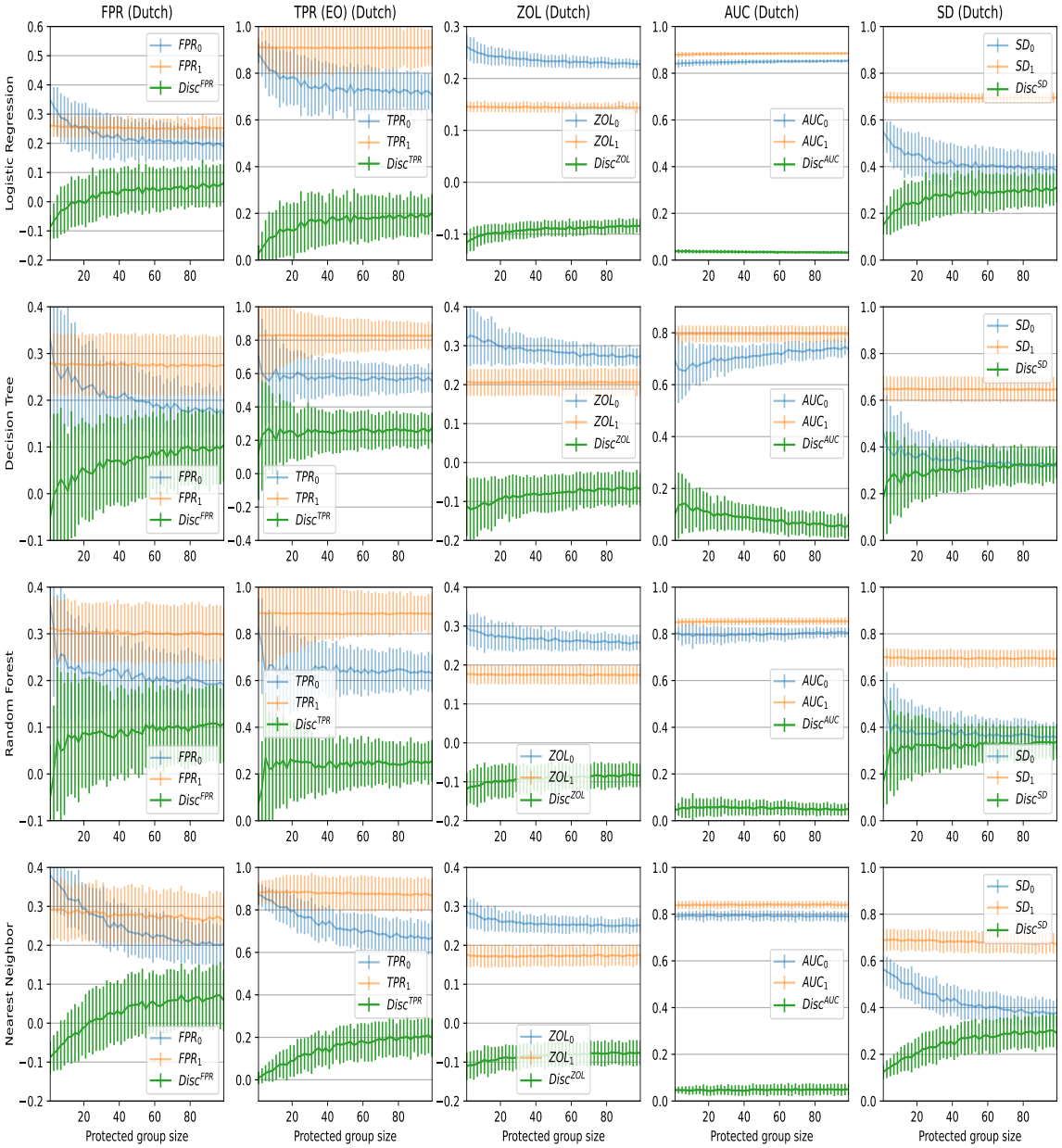


Figure 9: Discrimination values for the Dutch Census dataset while increasing the size of the protected group.

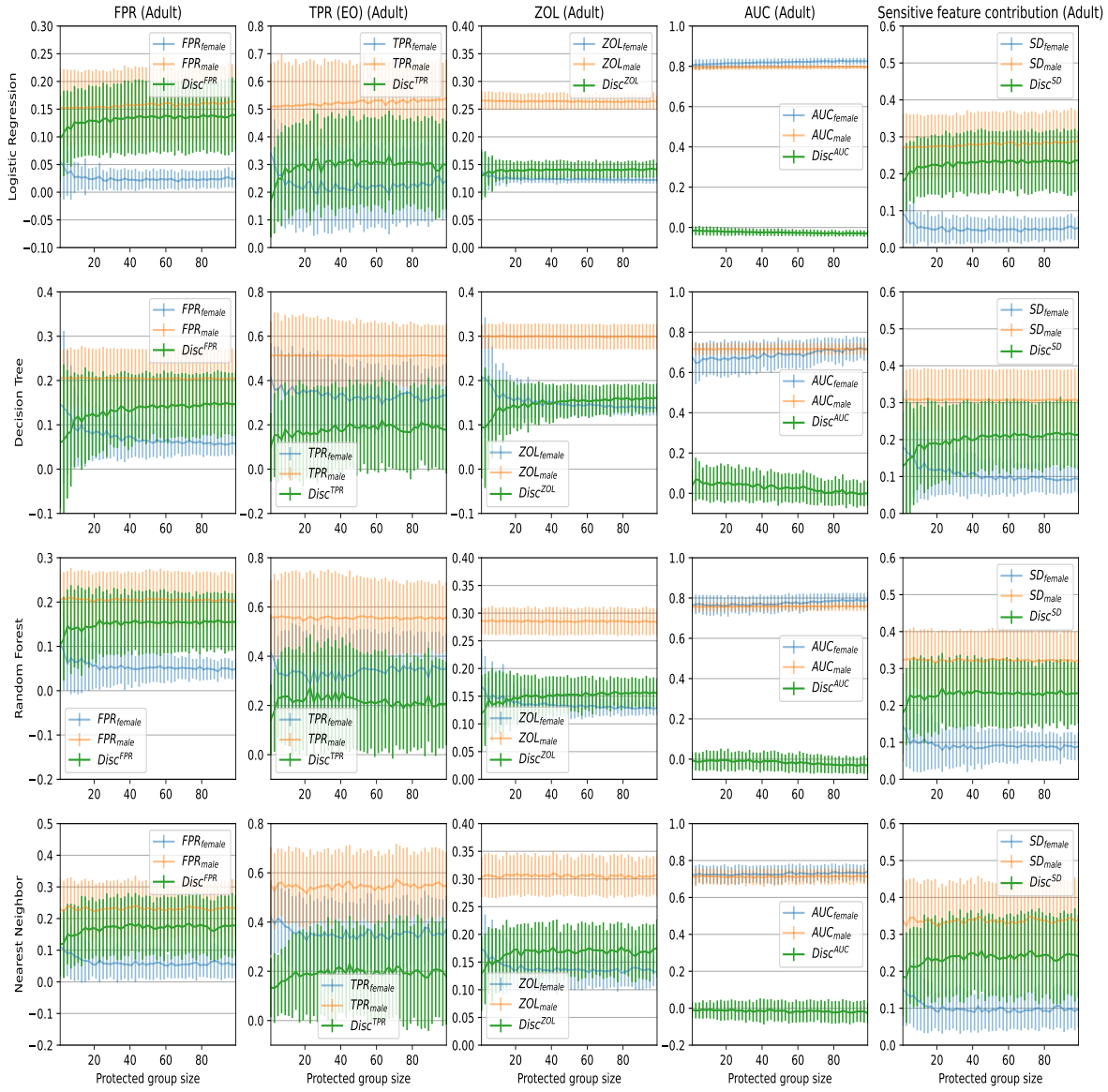


Figure 10: Discrimination value for the Adult dataset while increasing the size of the protected group.

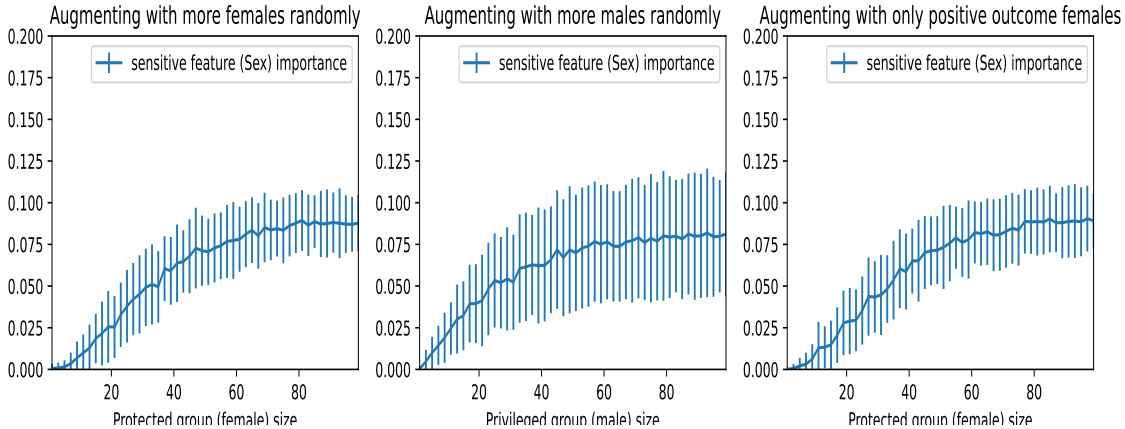


Figure 11: Sensitive feature (Sex) importance observed in the experiments of Section 6.4.

### A.3 Decomposing and bounding statistical disparity

Statistical disparity is the simplest discrimination metric and it corresponds to the difference in the expected outcomes between groups:

**Definition A.1** (Statistical Disparity)**.**

$$\begin{aligned} Disc^{SD}(\hat{Y}_{\mathcal{S}}) &= \mathbb{E}_{\mathcal{X}}[\hat{Y}_{\mathcal{S}}|A=a_1] - \mathbb{E}_{\mathcal{X}}[\hat{Y}_{\mathcal{S}}|A=a_0] \\ &= \mathop{\mathbb{E}}_{\mathbf{x}\in G_1}[\hat{f_{\mathcal{S}}}(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\in G_0}[\hat{f_{\mathcal{S}}}(\mathbf{x})] \end{aligned}$$

$Disc^{SD}(\hat{Y}_{\mathcal{S}})$ is a biased estimation of the *true* value $Disc^{SD}(Y)$. The following theorem states that the error in estimating statistical disparity can be bounded where the bounds are expressed in terms of noise, bias, and variance.

**Theorem A.2.** *The error in estimating statistical disparity is bounded as follows:*

$$\begin{aligned} |Disc^{SD}(\hat{Y}_{\mathcal{S}}) - Disc^{SD}(Y)| \leqslant\ & (\overline{N}^{AL}_{a_1}(\hat{Y}_{\mathcal{S}}) - \overline{N}^{AL}_{a_0}(\hat{Y}_{\mathcal{S}})) + (\overline{B}^{AL}_{a_1}(\hat{Y}_{\mathcal{S}}) - \overline{B}^{AL}_{a_0}(\hat{Y}_{\mathcal{S}}))+ \\ & (\overline{V}^{AL}_{a_1}(\hat{Y}_{\mathcal{S}}) - \overline{V}^{AL}_{a_0}(\hat{Y}_{\mathcal{S}})) \\ |Disc^{SD}(\hat{Y}_{\mathcal{S}}) - Disc^{SD}(Y)| \geqslant \max(\ & \\ & (\overline{N}^{AL}_{a_1}(\hat{Y}_{\mathcal{S}}) - \overline{N}^{AL}_{a_0}(\hat{Y}_{\mathcal{S}})) - (\overline{B}^{AL}_{a_1}(\hat{Y}_{\mathcal{S}}) - \overline{B}^{AL}_{a_0}(\hat{Y}_{\mathcal{S}}))- \\ & (\overline{V}^{AL}_{a_1}(\hat{Y}_{\mathcal{S}}) - \overline{V}^{AL}_{a_0}(\hat{Y}_{\mathcal{S}})), \\ & (\overline{B}^{AL}_{a_1}(\hat{Y}_{\mathcal{S}}) - \overline{B}^{AL}_{a_0}(\hat{Y}_{\mathcal{S}})) - (\overline{N}^{AL}_{a_1}(\hat{Y}_{\mathcal{S}}) - \overline{N}^{AL}_{a_0}(\hat{Y}_{\mathcal{S}}))- \\ & (\overline{V}^{AL}_{a_1}(\hat{Y}_{\mathcal{S}}) - \overline{V}^{AL}_{a_0}(\hat{Y}_{\mathcal{S}})), \\ & (\overline{V}^{AL}_{a_1}(\hat{Y}_{\mathcal{S}}) - \overline{V}^{AL}_{a_0}(\hat{Y}_{\mathcal{S}})) - (\overline{B}^{AL}_{a_1}(\hat{Y}_{\mathcal{S}}) - \overline{B}^{AL}_{a_0}(\hat{Y}_{\mathcal{S}}))- \\ & (\overline{N}^{AL}_{a_1}(\hat{Y}_{\mathcal{S}}) - \overline{N}^{AL}_{a_0}(\hat{Y}_{\mathcal{S}}))) \end{aligned}$$

*where:*

- $\overline{N}^{AL}_{a}(\hat{Y}_{\mathcal{S}}) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}[N^{AL}(\mathbf{x})|A=a]$
- $\overline{B}^{AL}_{a}(\hat{Y}_{\mathcal{S}}) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}[B^{AL}(\mathbf{x})|A=a]$
- $\overline{V}^{AL}_{a}(\hat{Y}_{\mathcal{S}}) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}[(1-2\times B^{AL}(\mathbf{x}))\times V^{AL}(\mathbf{x})|A=a]$

*Proof.* The proof is based on the triangle inequality of metrics. Recall that a metric is a function of two arguments ($dist(x,y)$) that satisfy minimality ($\forall x,y, dist(x,y) \geqslant dist(x,y)$), symmetry ($\forall x,y, dist(x,y) = dist(y,x)$), and triangle inequality ($\forall x,y,z dist(x,z) + dist(z,x) \geqslant dist(x,y)$). The full proof is very similar to the proof in [9] (Theorem 7). □ □