Dissecting Causal Biases

Rūta Binkytė, Sami Zhioua, Yassine Turki

DISSECTING CAUSAL BIASES

Rūta Binkytė* **& Sami Zhioua**† **& Yassine Turki**
Ecole Polytechnique, INRIA
1 Rue Honoré d'Estienne d'Orves,
Bâtiment Alan Turing,
Palaiseau, 91120, France
{ruta.binkyte-sadauskiene, zhioua}@lix.polytechnique.fr

ABSTRACT

Accurately measuring discrimination in machine learning-based automated decision systems is required to address the vital issue of fairness between subpopulations and/or individuals. Any bias in measuring discrimination can lead to either amplification or underestimation of the true value of discrimination. This paper focuses on a class of bias originating in the way training data is generated and/or collected. We call such class causal biases and use tools from the field of causality to formally define and analyze such biases. Four sources of bias are considered, namely, confounding, selection, measurement, and interaction. The main contribution of this paper is to provide, for each source of bias, a closed-form expression in terms of the model parameters. This makes it possible to analyze the behavior of each source of bias, in particular, in which cases they are absent and in which other cases they are maximized. We hope that the provided characterizations help the community better understand the sources of bias in machine learning applications.

1 INTRODUCTION

Machine learning (ML) is being used to inform decisions with critical consequences on human lives such as job hiring, college admission, loan granting, and criminal risk assessment. Unfortunately, these automated decision systems have been found to consistently discriminate against certain individuals or sub-populations, typically minorities Angwin et al. (2016); Buolamwini & Gebru (2018); O'Neill (2016); Quick (2015); Obermeyer et al. (2019). Addressing the problem of discrimination involves two main tasks. First, measuring discrimination as accurately and reliably as possible. Second, mitigating discrimination. The first task is clearly a prerequisite for the appropriate implementation of the second task. Proposing mitigation policies on the ground that discrimination is significant while it is in fact less significant (let alone not existing) may lead to undesirable consequences.

In this paper, we make a distinction between discrimination and bias. We use the term discrimination to refer to *the unjust or prejudicial treatment of different categories of people, on the ground of race, age, gender, disability, religion, political belief, etc.*. Whereas the term bias is used to refer to *the deviation of the expected value from the quantity it estimates.*

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 of Oxford (2021). More related to ML, Mehrabi et al. Mehrabi et al. (2021) classify the sources of bias into three categories depending on when the bias is introduced in the automated decision loop. In this paper we focus on a class of biases, we call causal biases, which arise from the way data is generated and/or collected. We use tools from the field of causality Pearl (2009); Imbens & Rubin (2015) as the latter emerged as a way to reliably estimate the effects between variables in presence of data imbalance leading to a deviation between the population distribution and the training data distribution.

The main contribution of the paper is to use tools and existing results from the field of causality to generate closed-form expressions of four sources of bias in the binary and the linear cases. These

*http://www.lix.polytechnique.fr/Labo/Ruta.BINKYTE-SADAUSKIENE/
†https://www.lix.polytechnique.fr/ zhioua/

sources of biases correspond to four different causal structures, namely, confounding, colliding (selection), measurement, and interaction. This has at least two advantages. First, understand how bias is expressed in terms of model parameters. Second, analyze the magnitude of each type of bias, in particular, when it is absent and when it is optimal. Finally, we empirically show the extent of causal biases in ML fairness benchmark datasets. All proofs of the closed-form expressions can be found in the supplementary material.

2 PRELIMINARIES AND PREVIOUS RESULTS USED IN THE PROOFS

Variables are denoted by capital letters. In particular, A is used for the sensitive variable (e.g., gender, race, age) and Y is used for the outcome of the automated decision system (e.g., hiring, admission, releasing on parole). Small letters denote specific values of variables (e.g., $A = a'$, $W = w$). Bold capital and small letters denote a set of variables and a set of values, respectively.

Consider a pair of variables X and Y. The variance of a variable X, ${\sigma_x}^2$, is a measure of dispersion which quantifies how far a set of values deviate from their mean and is defined as: ${\sigma_x}^2 = \mathbb{E}[X - \mathbb{E}[X]]^2$. Covariance of X and Y, σ_{xy}, is a measure of the joint variability of two random variables and is defined as: $\sigma_{xy} = \mathbb{E}[[X - \mathbb{E}[X][Y - \mathbb{E}[Y]]]$. Assuming a linear relationship between X and Y (X is the predictor variable, while Y is the response variable), the regression coefficient of Y given X, β_{yx}, represents the slope of the regression line in the prediction of Y given X ($\frac{\partial}{\partial x}\mathbb{E}[Y|X = x]$) and is equal to $\beta_{yx} = \frac{\sigma_{xy}}{{\sigma_x}^2}$. Correlation coefficient ρ_{yx}, however, represents the slope of the least square error line in the prediction of Y given X. The relationships between σ_{yx}, β_{yx}, and ρ_{yx} are as follows:

$$\beta_{yx} = \frac{\sigma_{yx}}{\sigma_x^2} = \rho_{yx}\frac{\sigma_y}{\sigma_x}$$
$$\rho_{yx} = \rho_{xy} = \frac{\sigma_{yx}}{\sigma_x\sigma_y} = \beta_{yx}\frac{\sigma_x}{\sigma_y} = \beta_{xy}\frac{\sigma_y}{\sigma_x}$$

Partial regression coefficient, $\beta_{yx.z}$, represents the slope of the regression line of Y on X when we hold variable Z constant ($\frac{\partial}{\partial x}\mathbb{E}[Y|X = x, Z = z]$). A well known result by Cramer Cramér (1999) allows to express $\beta_{yx.z}$ in terms of covariance between pairs of variables Pearl (2013):

$$\beta_{yx.z} = \frac{{\sigma_z}^2\sigma_{xy} - \sigma_{yz}\sigma_{zx}}{{\sigma_x}^2{\sigma_z}^2 - {\sigma_{xz}}^2} \tag{1}$$

For standardized variables (all variables are normalized to have a zero mean and a unit variance), the partial regression coefficient has a simpler expression since $\beta_{yx} = \sigma_{yx}$:

$$\beta_{yx.z} = \frac{\sigma_{xy} - \sigma_{yz}\sigma_{zx}}{1 - {\sigma_{xz}}^2} \tag{2}$$

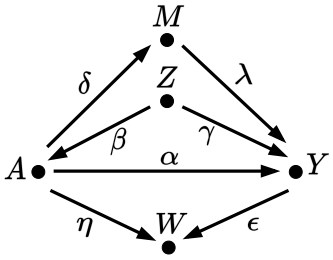

Figure 1: Causal graph with linearly related variables. Arrow labels represent linear regression coefficients.

Another known result by Wright Wright (1921); Pearl (2013) allows to represent the covariance of two variables in terms of the regression coefficients of the different paths (causal and non-causal, but not passing through any collider variable) between those two variables. More precisely, σ_{yx}

is equal to the sum of the regression coefficients of every path between x and y, weighted by the variance of the root variable of each path. For instance, in Figure 1, $\sigma_{ya} = {\sigma_a}^2\alpha + {\sigma_z}^2\beta\gamma + {\sigma_a}^2\delta\lambda$. Notice that the coefficients η and ϵ are not included as the path $A \rightarrow W \leftarrow Y$ is not $d-$connected (W is a collider variable). For standardized variables, the expression is simpler as all variables are normalized to have a unit variance. For the same example (Figure 1), $\sigma_{ya} = \alpha + \beta\gamma + \delta\lambda$. For linear models, regression coefficients can be interpreted causally. For instance, using the same example of Figure 1, α repesents the direct causal effect of A on Y. In more general models, the causal effect between two variables is typically expressed in terms of intervention probabilities. Intervention, noted $do(V = v)$ Pearl (2009), is a manipulation of the model that consists in fixing the value of a variable (or a set of variables) to a specific value regardless of the causes of that variable. The intervention $do(V = v)$ induces a different distribution on the other variables. Intuitively, while $\mathbb{P}(Y|A = a)$ reflects the population distribution of Y among individuals whose A value is a, $\mathbb{P}(Y|do(A = a)$[1] reflects the population distribution of Y if *everyone in the population* had their A value fixed at a. The obtained distribution $\mathbb{P}(Y|do(A = a)$ can be considered as a *counterfactual* distribution since the intervention forces a to take a value different from the one it would take in the actual world. $\mathbb{P}(Y|do(A = a)$ is not always computable from the data, a problem known as identifiability. For instance, if all counfounder variables are observable, the intervention probability, $\mathbb{P}(Y|do(A = a))$, can be computed by adjusting on the counfounder(s). For instance, assuming Z is the only confounder of A and Y,

$$\mathbb{P}(Y|do(A = a) = \sum_{z \in Z} \mathbb{P}(Y|A = a, Z = z).\mathbb{P}(Z = z) \tag{3}$$

Equation 3 is called the backdoor formula.

2.1 Statistical Disparity

Statistical disparity Rawls (2020) between groups $A = 0$ and $A = 1$, denoted as *StatDisp*(Y, A), is the difference between the conditional probabilities: $\mathbb{P}(y_1|a_1) - \mathbb{P}(y_1|a_0)$:

Definition 2.1.

$$\textit{StatDisp}(Y, A) = \mathbb{P}(y_1|a_1) - \mathbb{P}(y_1|a_0). \tag{4}$$

In presence of a confounder variable, Z, between A and Y, statistical disparity is a biased estimation of the discrimination as it does not filter out the spurious effect due to the confounding. For the sake of the proofs, we define the following variant of statistical disparity:

Definition 2.2.

$$\textit{StatDisp}(Y, A)_Z = \sum_{z \in Z} \big(\mathbb{P}(y_1|a_1, z) - \mathbb{P}(y_1|a_0, z)\big).\mathbb{P}(z). \tag{5}$$

Notice that if Z d-separates[2] A and Y, *StatDisp*$(Y, A)_Z$ coincides with the average causal effect ACE which defined using the do-operator (Equation 3):

Definition 2.3.

$$ACE(Y, A) = \mathbb{P}(y_1|do(a_1)) - \mathbb{P}(y_1|do(a_0)). \tag{6}$$

3 Types of Bias

Measuring discrimination without taking into consideration the causal structure underlying the relationships between variables may lead to misleading conclusions. That is, a biased estimation of discrimination. In extreme cases, such as Simpson's paradox, the bias may lead to reversing the conclusions (e.g. the biased estimation indicates a positive discrimination, while the unbiased estimation is actually a negative discrimination).

[1]The notations $Y_{A \leftarrow a}$ and $Y(a)$ are used in the literature as well. $\mathbb{P}(Y = y|do(A = a)) = \mathbb{P}(Y_{A=a} = y) = \mathbb{P}(Y_a = y) = \mathbb{P}(y_a)$ is used to define the causal effect of A on Y.

[2]For the definition of d-separation, we refer the reader to Definition 1.2.3 in Pearl (2009).

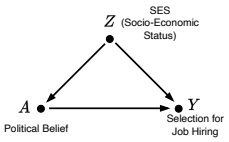

Figure 2: Confounding bias example.

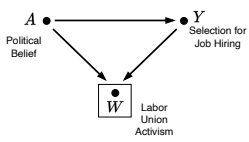

Figure 3: Collider bias example.

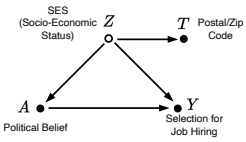

Figure 4: Measurement bias example.

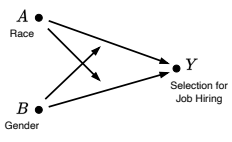

Figure 5: Interaction Bias example.

3.1 CONFOUNDING BIAS

The first type of bias, confounding bias, is due to a failure to consider a confounder variable. Consider the hypothetical example in Figure 2 of an automated system to select candidates for job positions. Assume that the system takes as input two features, namely, the socio-economic status (SES) denoted as Z and the political belief of the candidate A. The outcome Y is whether the candidate is selected for the next stage of hiring (or the probability the candidate is selected). The outcome Y is influenced by the SES (A better SES makes it possible for candidates to attend more reputable academic institutions and to be enrolled in costly trainings). Both variables can be either binary (Z might be either rich or poor while A might be either liberal or conservative) or continuous (how rich/poor a candidate is for Z and the degree of conservativeness of the candidate for A). The political belief A of a candidate can be influenced by several variables, but in this example, assume that it is only influenced by the SES of the candidate. Finally, assume that the automated decision system is suspected to be biased by the political belief of candidates. That is, it is claimed that the system will more likely select candidates with a particular political belief.

A simple approach to check the fairness of the automated selection Y with respect to the sensitive attribute A is to contrast the conditional probabilities: $\mathbb{P}(Y = 1 \mid A = 0)$ and $\mathbb{P}(Y = 1 \mid A = 1)$[3], corresponding to statistical disparity, which quantifies the disparity in the selection rates between both types of candidates (conservatives and liberals). However such estimation of discrimination is biased due to the confounding path through Z. As Z variable causes both the sensitive variable A and the outcome Y, it creates a correlation between A and Y which is not causal. In other words, high SES (rich) candidates tend to have a more conservative political belief and at the same time more chances to be selected for the job (better academic institutions and training) which creates the following correlation in the data: employers will have more candidates with convervative political beliefs, and hence less candidates with liberal political beliefs. Such correlation is due to the confounder Z and should not count as discrimination. We call such bias in estimating discrimination, confounding bias.

3.2 SELECTION BIAS

The second type of bias, selection bias, is due to the presence of common effect (collider) variable and a data generation process implicitly conditioning on that variable. Using the same hypothetical example of job selection, consider the causal graph in Figure 3. A and Y are the same as in the previous example. Assume that data for training the automated decision system is collected from different sources, but mainly from labor union records. Assume also that variable W representing the labor union activism of the candidate is caused by both A and Y. On one hand, the political belief A influences whether a candidate is an active member of labor union (individuals with liberal political beliefs are more likely to enroll in labor unions). On the other hand, if a candidate is selected/hired, then there are higher chances that she becomes a member of labor union and consequently that her case is recorded in the labor union records. Consistent with previous work, a box around a variable (W) indicates that data is generated by implicitly conditioning on that variable.

Again the simple approach of constrasting the selection rates between both types of candidates (conservatives and liberals) leads to a biased estimation of discrimination due to the colliding path through W. Intuitively, an individual has a record in the collected data either because she has liberal political beliefs or because she is selected for the job. Individuals who happen to have liberal political beliefs and at the same time selected for the job are still present in the data, however conditioning on labor union activism creates a correlation between A and Y which is not causal: data coming from labor union records includes fewer liberal candidates which are selected for the job than conservative

[3]

candidates. Again, this is a discrimination against candidates with liberal political beliefs. Such correlation is due to the colliding structure and should not count as discrimination. We call such bias in estimating discrimination, selection bias.

3.3 Measurement Bias

The third type of bias, measurement bias, is due to the use of a proxy variable to estimate discrimination instead of an ideal but unmeasurable variable. Consider a third variant of the same job selection example having the causal graph of Figure 4. Unlike in the causal graph of confounding bias (Figure 2), the confounder variable Z is unmeasurable (empty bullet instead of a filled one). In practice, it is difficult to find a variable that represents accurately the socio-economic status (salary, possessions, etc.). Being unmeasurable, Z cannot be used to estimate discrimination while blocking the confounding path through Z. For practical reasons, the (measurable) variable T representing the postal/zip code of the candidate's address can be used instead. T is considered a proxy of Z as it is highly correlated with (but not identical to) Z[4]. Using variable T as a proxy to measure Z may lead to an additional bias, we call measurement bias.

3.4 Interaction Bias

The fourth type of bias, interaction bias, is observed when two causes of the outcome interact with each other, making the joint effect smaller or greater than the sum of individual effects. Consider the same job hiring example but where two sensitive attributes, political belief (liberals vs conservatives) and gender have an effect on the hiring decision. In the presence of interaction between political belief and gender, statistical disparity will not accurately measure the individual effects of Political Belief and Gender even if no confounding condition is satisfied. For example, it is possible to observe a situation where statistical parity is almost satisfied for both individual sensitive variables, but the intersectional sensitive group is discriminated Buolamwini & Gebru (2018). Following our previous example, we would define liberal females as an unprivileged intersectional group and conservative males as a privileged intersectional group. In the presence of interaction, the discrimination against liberal females is not equal to the sum of discrimination against conservative and females individually. In addition, the average discrimination value for liberals or females, as measured by statistical disparity, will also be biased, as it does not take into account the interaction between the two sensitive variables.

4 Confounding Bias

Confounding bias occurs when both the sensitive variable and the outcome have a common cause, the counfounder variable (Figure 6). Consequently, the mechanism of selecting samples from the two groups (protected and privileged) is not independent of the outcome. This creates a bias when measuring the causal effect of the sensitive attribute on the outcome.

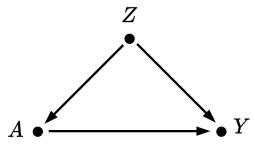

Figure 6: Simple confounding structure

4.1 Binary Model Case

For a concise notation, let y_1 and y_0 denote the propositions $Y = 1$ and $Y = 0$, respectively, and the same for the variables A and Z. For instance, $\mathbb{P}(Y = 1|A = 0)$ is written simply as $\mathbb{P}(y_1|a_0)$.

Statistical disparity Rawls (2020) between groups $A = 0$ and $A = 1$, denoted as *StatDisp*(Y, A), is the difference between the conditional probabilities: $\mathbb{P}(y_1|a_1) - \mathbb{P}(y_1|a_0)$. In presence of a confounder

[4]The candidate's address gives a strong indicator of the socio-economic status.

variable, Z, between A and Y, statistical disparity is a biased estimation of the discrimination as it does not filter out the spurious effect due to the confounding.

Definition 4.1. *Confounding bias is defined as*[5]*:*

$$\textit{ConfBias}(Y, A) = \textit{StatDisp}(Y, A) - ACE(Y, A) \tag{7}$$

where $ACE(Y, A)$ is the causal effect of A on Y (Definition 2.3). For the simple confounding structure of Figure 6, ACE coincides with $\textit{StatDisp}_Z(Y, A)$ (Definition 2.2).

Theorem 4.2. *The difference in discrimination due to confounding bias is equal to:*

$$\begin{aligned}\textit{ConfBias}(Y, A) = &(1 - \mathbb{P}(z_0|a_0) - \mathbb{P}(z_1)) \\ &\times (\alpha - \beta - \gamma + \delta + \frac{\gamma}{\mathbb{P}(a_1)} - \frac{\delta}{\mathbb{P}(a_1)})\end{aligned} \tag{8}$$

where α, β, γ, and δ denote, respectively, $\mathbb{P}(y_1|a_0, z_0), \mathbb{P}(y_1|a_0, z_1), \mathbb{P}(y_1|a_1, z_0)$, and $\mathbb{P}(y_1|a_1, z_1)$.

Proof. Let $\mathbb{P}(z_1) = \epsilon$ ($\epsilon \in]0, 1[$) and hence $\mathbb{P}(z_0) = 1 - \epsilon$. Similarly, let $\mathbb{P}(a_1) = \lambda$ and hence $\mathbb{P}(a_0) = 1 - \lambda$. Let $\mathbb{P}(y_1|a_0, z_0) = \alpha$, $\mathbb{P}(y_1|a_0, z_1) = \beta$, $\mathbb{P}(y_1|a_1, z_0) = \gamma$, and $\mathbb{P}(y_1|a_1, z_1) = \delta$. Finally, let $\mathbb{P}(z_0|a_0) = \tau$. The remaining conditional probabilities of Z given A are equal to the following:

$$\mathbb{P}(z_1|a_0) = 1 - \mathbb{P}(z_0|a_0) = 1 - \tau \tag{9}$$

$$\begin{aligned}\mathbb{P}(z_1|a_1) &= \frac{\mathbb{P}(z_1) - \mathbb{P}(z_1|a_0)\mathbb{P}(a_0)}{\mathbb{P}(a_1)} \\ &= \frac{\epsilon - (1 - \tau)(1 - \lambda)}{\lambda} \\ &= \frac{\epsilon - 1 + \tau + \lambda - \tau\lambda)}{\lambda}\end{aligned} \tag{10}$$

$$\begin{aligned}\mathbb{P}(z_0|a_1) &= \frac{\mathbb{P}(z_0) - \mathbb{P}(z_0|a_0)\mathbb{P}(a_0)}{\mathbb{P}(a_1)} \\ &= \frac{(1 - \epsilon) - \tau(1 - \lambda)}{\lambda} \\ &= \frac{1 - \epsilon - \tau + \tau\lambda}{\lambda}\end{aligned} \tag{11}$$

Equation (9) follow from the fact that, given u_i events are exhaustive and mutually exclusive, $\sum_i \mathbb{P}(a_i|X) = 1$. Equations (10) and (11) follow from the fact that, given u_i events are exhaustive and mutually exclusive, $\sum_i \mathbb{P}(X|u_i)\mathbb{P}(u_i) = \mathbb{P}(X)$. $\textit{StatDisp}(Y, A)$ can then be expressed in terms of the above parameters:

$$\begin{aligned}\mathbb{P}(y_1|a_1) - \mathbb{P}(y_1|a_0) &= \sum_{z \in Z} (\mathbb{P}(y_1|a_1, z)\mathbb{P}(z|a_1) - \mathbb{P}(y_1|a_0, z)\mathbb{P}(z|a_0) \\ &= \mathbb{P}(y_1|a_1, z_0)\mathbb{P}(z_0|a_1) - \mathbb{P}(y_1|a_0, z_0)\mathbb{P}(z_0|a_0) \\ &\quad + \mathbb{P}(y_1|a_1, z_1)\mathbb{P}(z_1|a_1) - \mathbb{P}(y_1|a_0, z_1)\mathbb{P}(z_1|a_0) \\ &= \gamma\big(\frac{1 - \epsilon - \tau\lambda}{1 - \lambda}\big) - \alpha\tau + \delta\big(\frac{\epsilon - \lambda + \tau\lambda}{1 - \lambda}\big) - \beta(1 - \tau)\end{aligned}$$

$ACE(Y, A)$, on the other hand can be expressed as follows:

$$\begin{aligned}\mathbb{P}(y_1|do(a_1)) - \mathbb{P}(y_1|do(a_0)) &= \sum_{z \in Z} (\mathbb{P}(y_1|a_1, z) - \mathbb{P}(y_1|a_0, z))\mathbb{P}(z) \\ &= \mathbb{P}(y_1|a_1, z_0) - \mathbb{P}(y_1|a_0, z_0))\mathbb{P}(z_0) \\ &\quad + \mathbb{P}(y_1|a_1, z_1) - \mathbb{P}(y_1|a_0, z_1))\mathbb{P}(z_1) \\ &= (\gamma - \alpha)(1 - \epsilon) + (\delta - \beta)\epsilon\end{aligned}$$

[5]In this paper, bias is defined by substracting the correct value of discrimination from the biased estimation.

Confounding bias is then equal to:

$$\begin{aligned}
\textit{StatDisp}(Y,A) - ACE(Y,A) &= \mathbb{P}(y_1|a_1) - \mathbb{P}(y_1|a_0) - (\mathbb{P}(y_1|do(a_1)) - \mathbb{P}(y_1|do(a_0))) \\
&= \gamma\big(\frac{1-\epsilon-\tau\lambda}{1-\lambda}\big) - \alpha\tau + \delta\big(\frac{\epsilon-\lambda+\tau\lambda}{1-\lambda}\big) - \beta(1-\tau) \\
&\quad - \big((\gamma-\alpha)(1-\epsilon) + (\delta-\beta)\epsilon\big) \\
&= \frac{\gamma}{\lambda} - \frac{\gamma\epsilon}{\lambda} - \frac{\gamma\tau}{\lambda} + \gamma\tau - \alpha\tau + \frac{\delta\epsilon}{\lambda} - \frac{\delta}{\lambda} + \frac{\delta\tau}{\lambda} \\
&\quad + \delta - \delta\tau - \beta + \beta\tau - \gamma + \alpha + \epsilon\gamma - \alpha\epsilon - \delta\epsilon + \beta\epsilon \\
&= \big(\alpha + \delta - \beta - \gamma + \frac{\gamma}{\lambda} - \frac{\delta}{\lambda}\big) - \tau\big(\alpha + \delta - \beta - \gamma + \frac{\gamma}{\lambda} - \frac{\delta}{\lambda}\big) \\
&\quad - \epsilon\big(\alpha + \delta - \beta - \gamma + \frac{\gamma}{\lambda} - \frac{\delta}{\lambda}\big) \\
&= \big(1 - \tau - \epsilon\big)\big(\alpha + \delta - \beta - \gamma + \frac{\gamma}{\lambda} - \frac{\delta}{\lambda}\big)
\end{aligned}$$

□

For the specific case of equal proportions between sensitive groups (e.g. no under or over representation of a certain sensitive group), confounding bias can be characterized by a simpler closed-form expression.

Theorem 4.3. *Assuming that $\mathbb{P}(a_0) = \mathbb{P}(a_1) = \frac{1}{2}$, the difference in discrimination due to confounding bias is equal to:*

$$\textit{ConfBias}(Y,A) = (1 - \mathbb{P}(z_0|a_0) - \mathbb{P}(z_1))(\alpha - \beta + \gamma - \delta) \tag{12}$$

where α, β, γ, and δ are defined similarly to Theorem 4.2.

Proof. Let $\mathbb{P}(z_1) = \epsilon\,(\epsilon \in]0,1[)$ and hence $\mathbb{P}(z_0) = 1-\epsilon$. And let $\mathbb{P}(y_1|a_0,z_0) = \alpha$, $\mathbb{P}(y_1|a_0,z_1) = \beta$, $\mathbb{P}(y_1|a_1,z_0) = \gamma$, and $\mathbb{P}(y_1|a_1,z_1) = \delta$. Finally, let $\mathbb{P}(z_0|a_0) = \tau$. The remaining conditional probabilities of Z given A are equal to the following:

$$\mathbb{P}(z_1|a_0) = 1 - \mathbb{P}(z_0|a_0) = 1 - \tau \tag{13}$$

$$\begin{aligned}
\mathbb{P}(z_1|a_1) &= \frac{\mathbb{P}(z_1) - \mathbb{P}(z_1|a_0)\mathbb{P}(a_0)}{\mathbb{P}(a_1)} \\
&= 2\epsilon + \tau - 1
\end{aligned} \tag{14}$$

$$\begin{aligned}
\mathbb{P}(z_0|a_1) &= 1 - \mathbb{P}(z_1|a_1) \\
&= 2 - 2\epsilon - \tau
\end{aligned} \tag{15}$$

Equations (13) and (15) follow from the fact that, given u_i events are exhaustive and mutually exclusive, $\sum_i \mathbb{P}(a_i|X) = 1$. Equation (14) follows from the fact that, given u_i events are exhaustive and mutually exclusive, $\sum_i \mathbb{P}(X|u_i)\mathbb{P}(u_i) = \mathbb{P}(X)$. *StatDisp*$(Y,A)$ can then be expressed in terms of the above parameters:

$$\begin{aligned}
\mathbb{P}(y_1|a_1) - \mathbb{P}(y_1|a_0) &= \sum_{z\in Z} (\mathbb{P}(y_1|a_1,z)\mathbb{P}(z|a_1) - \mathbb{P}(y_1|a_0,z)\mathbb{P}(z|a_0) \\
&= \mathbb{P}(y_1|a_1,z_0)\mathbb{P}(z_0|a_1) - \mathbb{P}(y_1|a_0,z_0)\mathbb{P}(z_0|a_0) \\
&\quad + \mathbb{P}(y_1|a_1,z_1)\mathbb{P}(z_1|a_1) - \mathbb{P}(y_1|a_0,z_1)\mathbb{P}(z_1|a_0) \\
&= \gamma(2 - 2\epsilon - \tau) - \alpha\tau + \delta(2\epsilon + \tau - 1) - \beta(1-\tau) \\
&= \tau(-\alpha + \beta - \gamma + \delta) + 2\epsilon(\delta - \gamma) + 2\gamma - \delta - \beta
\end{aligned}$$

$ACE(Y,A)$, on the other hand can be expressed as follows:

$$\begin{aligned}
\mathbb{P}(y_1|do(a_1)) - \mathbb{P}(y_1|do(a_0)) &= \sum_{z\in Z} (\mathbb{P}(y_1|a_1,z) - \mathbb{P}(y_1|a_0,z))\mathbb{P}(z) \\
&= \mathbb{P}(y_1|a_1,z_0) - \mathbb{P}(y_1|a_0,z_0))\mathbb{P}(z_0) \\
&\quad + \mathbb{P}(y_1|a_1,z_1) - \mathbb{P}(y_1|a_0,z_1))\mathbb{P}(z_1) \\
&= (\gamma - \alpha)(1-\epsilon) + (\delta - \beta)\epsilon
\end{aligned}$$

Confounding bias is then equal to:

$$\begin{aligned}
\textit{StatDisp}(Y,A) - ACE(Y,A) &= \mathbb{P}(y_1|a_1) - \mathbb{P}(y_1|a_0) - (\mathbb{P}(y_1|do(a_1)) - \mathbb{P}(y_1|do(a_0))) \\
&= \tau(-\alpha+\beta-\gamma+\delta) + 2\epsilon(\delta-\gamma) + 2\gamma - \delta - \beta \\
&\quad - ((\gamma-\alpha)(1-\epsilon) + (\delta-\beta)\epsilon) \\
&= \tau(-\alpha+\beta-\gamma+\delta) + 2\epsilon\delta - 2\epsilon\gamma + 2\gamma - \delta - \beta \\
&\quad - \gamma + \gamma\epsilon + \alpha - \alpha\epsilon - \delta\epsilon + \beta\epsilon \\
&= \tau(-\alpha+\beta-\gamma+\delta) + \epsilon(2\delta - 2\gamma + \gamma - \alpha - \delta + \beta) \\
&\quad + 2\gamma - \delta - \beta - \gamma + \alpha \\
&= \tau(-\alpha+\beta-\gamma+\delta) + \epsilon(-\alpha+\beta-\gamma+\delta) + \alpha - \beta + \gamma - \delta \\
&= (1-\tau-\epsilon)(\alpha-\beta+\gamma-\delta)
\end{aligned}$$

□

4.2 Linear Model Case

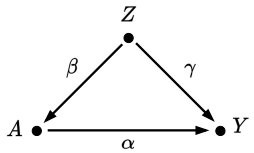

Figure 7: Confounding structure in linear model

Theorem 4.4. *Let A, Y, and Z variables with linear regressions coefficients as in Figure 7 which represents the basic confounding structure. The confounding bias can be expressed in terms of covariances of pairs of variables as follows:*

$$\textit{ConfBias}(Y,A) = \frac{\sigma_{za}\sigma_{yz} - \frac{\sigma_{ya}}{\sigma_a^2}\sigma_{za}^2}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2} \tag{16}$$

Confounding bias can also be expressed in terms of the linear regression coefficients as follows:

$$\textit{ConfBias}(Y,A) = \frac{{\sigma_z}^2}{{\sigma_a}^2}\beta\gamma \tag{17}$$

Proof. For Equation (16),

$$\begin{aligned}
\textit{ConfBias}(Y,A) &= \beta_{ya} - \beta_{ya.z} \\
&= \frac{\sigma_{ya}}{\sigma_a^2} - \frac{\sigma_z^2\sigma_{ya} - \sigma_{yz}\sigma_{za}}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2} \\
&= \frac{\frac{\sigma_{ya}}{\sigma_a^2}(\sigma_a^2\sigma_z^2 - \sigma_{za}^2) - (\sigma_z^2\sigma_{ya} - \sigma_{yz}\sigma_{za})}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2} \\
&= \frac{\cancel{\frac{\sigma_{ya}}{\sigma_a^2}\sigma_a^2\sigma_z^2} - \frac{\sigma_{ya}}{\sigma_a^2}\sigma_{za}^2 - \cancel{\sigma_z^2\sigma_{ya}} + \sigma_{yz}\sigma_{za}}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2} \\
&= \frac{\sigma_{za}\sigma_{yz} - \frac{\sigma_{ya}}{\sigma_a^2}\sigma_{za}^2}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2}
\end{aligned}$$

For Equation (17),

$$\begin{aligned}
\textit{ConfBias}(Y, A) &= \beta_{ya} - \beta_{ya.z} \\
&= \frac{\sigma_{ya}}{\sigma_a^2} - \frac{\sigma_z^2\sigma_{ya} - \sigma_{yz}\sigma_{za}}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2} \\
&= \frac{\sigma_a^2\alpha + \sigma_z^2\beta\gamma}{\sigma_a^2} - \frac{\sigma_z^2(\sigma_a^2\alpha + \sigma_z^2\beta\gamma) - (\sigma_z^2\gamma + \sigma_z^2\beta\alpha)(\sigma_z^2\beta)}{\sigma_a^2\sigma_z^2 - (\sigma_z^2\beta)^2} \\
&= \frac{\cancel{\sigma_a^2}\alpha}{\cancel{\sigma_a^2}} + \frac{\sigma_z^2\beta\gamma}{\sigma_a^2} - \frac{\sigma_z^2\sigma_a^2\alpha + \cancel{\sigma_z^4\beta\gamma} - \cancel{\sigma_z^4\beta\gamma} - \sigma_z^4\beta^2\alpha}{\sigma_a^2\sigma_z^2 - \sigma_z^4\beta^2} \\
&= \cancel{\alpha} + \frac{\sigma_z^2\beta\gamma}{\sigma_a^2} - \frac{\cancel{\alpha(\sigma_z^2\sigma_a^2 - \sigma_z^4\beta^2)}}{\cancel{\sigma_a^2\sigma_z^2 - \sigma_z^4\beta^2}} \\
&= \frac{\sigma_z^2}{\sigma_a^2}\beta\gamma
\end{aligned}$$

□

Corollary 4.5. *For standardized variables A, Y, and Z, confounding bias can be expressed in terms of covariances as:*

$$\textit{ConfBias}(Y, A) = \frac{\sigma_{za}\sigma_{yz} - \sigma_{ya}\sigma_{za}^2}{1 - \sigma_{za}^2} \tag{18}$$

And in terms of regression coefficient, simply as (Pearl (2013)):

$$\textit{ConfBias}(Y, A) = \beta\gamma \tag{19}$$

Equations (18) and (19) can be obtained from Equations (16) and (17) as $\sigma_z = \sigma_a = 1$.

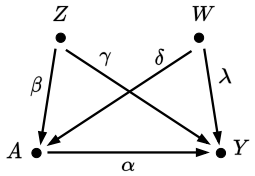

Figure 8: Confounding structure with two confounders

Theorem 4.6. *Let A, Y, Z, W variables as in Figure 8. Assuming that all variables are standardized and that W and Z are independent, the regression coefficent of Y on A conditioning on Z and W, the confounding bias is equal:*

$$\textit{ConfBias}(Y, A) = \frac{\sigma_{za}\sigma_{yz} + \sigma_{wa}\sigma_{yw} - \sigma_{ya}(\sigma_{za}^2 + \sigma_{wa}^2)}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \tag{20}$$

And in terms of the regression coefficients:

$$\textit{ConfBias}(Y, A) = \beta\gamma + \delta\lambda \tag{21}$$

Proof. The proof is based on proving that:

$$\beta_{ya.zw} = \frac{\sigma_{ya} - \sigma_{za}\sigma_{yz} - \sigma_{wy}\sigma_{wa}}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \tag{22}$$

From Cramér Cramér (1999) (Page 307), we know that the partial regression coefficient can be expressed as:

$$\beta_{ya.zw} = \rho_{ya.zw}\frac{\sigma_{y.zw}}{\sigma_{a.zw}} \tag{23}$$

Where $\rho_{ya.zw}$ denotes the partial correlation and $\sigma_{a.zw}, \sigma_{y.zw}$ denote the residual variances.

Based on the correlation matrix:

$$\begin{bmatrix} 1 & \rho_{ya} & \rho_{yz} & \rho_{yw} \\ \rho_{ay} & 1 & \rho_{az} & \rho_{aw} \\ \rho_{zy} & \rho_{za} & 1 & \rho_{zw} \\ \rho_{wy} & \rho_{wa} & \rho_{wz} & 1 \end{bmatrix}$$

the partial correlation $\rho_{ya.zw}$ can be expressed in terms of cofactors as follows[6]:

$$\rho_{ya.zw} = -\frac{C_{ya}}{\sqrt{C_{yy}C_{aa}}} \tag{24}$$

where C_{ij} denotes the cofactor of the element ρ_{ij} in the determinant of the correlation matrix and are equal to the following:

$$\begin{aligned} C_{ya} = -(&\rho_{ya} - \rho_{ya}\rho_{zw}^2 - \rho_{za}\rho_{yz} - \rho_{wa}\rho_{yw} \\ &+ \rho_{za}\rho_{yw}\rho_{wz} + \rho_{wa}\rho_{yz}\rho_{zw}) \end{aligned} \tag{25}$$

$$C_{yy} = 1 - \rho_{zw}^2 - \rho_{za}^2 - \rho_{wa}^2 + 2\rho_{za}\rho_{aw}\rho_{wz} \tag{26}$$

$$C_{aa} = 1 - \rho_{zw}^2 - \rho_{zy}^2 - \rho_{wy}^2 + 2\rho_{yz}\rho_{yw}\rho_{wz} \tag{27}$$

Residual variances in Equation 23 can be expressed in terms of total and partial correlation coefficients as follows Cramér (1999)(Equation 23.4.5 in page 307):

$$\sigma_{y.zw}^2 = \sigma_y^2(1-\rho_{yz}^2)(1-\rho_{yw.z}^2)(1-\rho_{ya.zw}^2) \tag{28}$$

$$\sigma_{a.zw}^2 = \sigma_a^2(1-\rho_{az}^2)(1-\rho_{aw.z}^2)(1-\rho_{ay.zw}^2) \tag{29}$$

As the last term is the same, we have:

$$\frac{\sigma_{y.zw}}{\sigma_{a.zw}} = \frac{\sigma_y\sqrt{(1-\rho_{yz}^2)(1-\rho_{yw.z}^2)}}{\sigma_a\sqrt{(1-\rho_{az}^2)(1-\rho_{aw.z}^2)}} \tag{30}$$

The partial correlation coefficients in Equation 30 can be expressed in terms of total correlation coefficients as follows Cramér (1999) (Equation 23.4.3 in page 306):

$$\rho_{yw.z} = \frac{\rho_{yw} - \rho_{yz}\rho_{wz}}{\sqrt{(1-\rho_{yz}^2)(1-\rho_{wz}^2)}} \tag{31}$$

After simple algebraic steps, we obtain:

$$\frac{\sigma_{y.zw}}{\sigma_{a.zw}} = \frac{\sigma_y}{\sigma_a}\frac{\sqrt{1-\rho_{zw}^2-\rho_{yz}^2-\rho_{yw}^2+2\rho_{zy}\rho_{yw}\rho_{wz}}}{\sqrt{1-\rho_{zw}^2-\rho_{az}^2-\rho_{aw}^2+2\rho_{za}\rho_{aw}\rho_{wz}}} \tag{32}$$

Finally, $\beta_{ya.zw}$ in Equation 23 can be expressed in terms of total correlation coefficients as follows:

$$\beta_{ya.zw} = \frac{\sigma_y}{\sigma_a}\frac{Q}{1-\rho_{zw}^2-\rho_{za}^2-\rho_{wa}^2+2\rho_{za}\rho_{aw}\rho_{wz}} \tag{33}$$

where

$$\begin{aligned} Q = &\rho_{ya} - \rho_{ya}\rho_{zw}^2 - \rho_{za}\rho_{yz} - \rho_{wy}\rho_{wa} \\ &+ \rho_{za}\rho_{yw}\rho_{zw} + \rho_{wa}\rho_{yz}\rho_{zw} \end{aligned}$$

Recall that $\rho_{ya} = \frac{\sigma_{ya}}{\sigma_y\sigma a}$. The formula becomes:

[6]The proof is sketched in https://en.wikipedia.org/wiki/Partial_correlation.

$$\beta_{ya.zw} = \frac{Q}{R} \tag{34}$$

Where

$$\begin{aligned} Q = \sigma_{ya}(\sigma_z^2\sigma_w^2 - \sigma_{zw}^2) + \sigma_{yz}(\sigma_{wa}\sigma_{zw} - \sigma_{za}\sigma_w^2) \\ + \sigma_{wy}(\sigma_{za}\sigma_{zw} - \sigma_{wa}\sigma_z^2) \end{aligned}$$

And

$$\begin{aligned} R = \sigma_a^2\sigma_z^2\sigma_w^2 - \sigma_a^2\sigma_{zw}^2 - \sigma_a\sigma_z\sigma_w^2\sigma_{za}^2 \\ - \sigma_z^2\sigma_{aw}^2 + 2\sigma_a\sigma_z\sigma_{az}\sigma_{aw}\sigma_{zw} \end{aligned}$$

For standardized variables, $\forall v, \sigma_v = 1$, and hence $\forall u, v \sigma_{uv} = \rho_{uv}$. Equation 33 becomes:

$$\beta_{ya.zw} = \frac{Q}{1 - \sigma_{zw}^2 - \sigma_{za}^2 - \sigma_{wa}^2 + 2\sigma_{za}\sigma_{aw}\sigma_{wz}} \tag{35}$$

Where

$$\begin{aligned} Q = \sigma_{ya}(1 - \sigma_{zw}^2) + \sigma_{yz}(\sigma_{wa}\sigma_{zw} - \sigma_{za}) \\ + \sigma_{yw}(\sigma_{za}\sigma_{zw} - \sigma_{wa}) \end{aligned}$$

If we further assume that confounders are uncorrelated, that is, $\sigma_{zw} = 0$, then we have the simpler expression:

$$\beta_{ya.zw} = \frac{\sigma_{ya} - \sigma_{za}\sigma_{yz} - \sigma_{wy}\sigma_{wa}}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \tag{36}$$

For Equation (20):

$$\begin{aligned} \textit{ConfBias}(Y, A) &= \beta_{ya} - \beta_{ya.zw} \\ &= \sigma_{ya} - \frac{\sigma_{ya} - \sigma_{za}\sigma_{yz} - \sigma_{wa}\sigma_{yw}}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \\ &= \frac{\sigma_{ya}(1 - \sigma_{za}^2 - \sigma_{wa}^2) - \sigma_{ya} + \sigma_{za}\sigma_{yz} + \sigma_{wa}\sigma_{yw}}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \\ &= \frac{\cancel{\sigma_{ya}} - \sigma_{ya}\sigma_{za}^2 - \sigma_{ya}\sigma_{wa}^2 - \cancel{\sigma_{ya}} + \sigma_{za}\sigma_{yz} + \sigma_{wa}\sigma_{yw}}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \\ &= \frac{\sigma_{za}\sigma_{yz} + \sigma_{wa}\sigma_{yw} - \sigma_{ya}(\sigma_{za}^2 + \sigma_{wa}^2)}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \end{aligned} \tag{37}$$

For Equation (21):

$$\begin{aligned} \textit{ConfBias}(Y, A) &= \beta_{ya} - \beta_{ya.zw} \\ &= \alpha + \beta\gamma + \lambda\delta - \frac{\alpha + \beta\gamma + \lambda\delta - \beta(\gamma + \beta\alpha) - \delta(\lambda + \delta\alpha)}{1 - \beta^2 - \delta^2} \\ &= \alpha + \beta\gamma + \lambda\delta - \frac{\alpha + \cancel{\beta\gamma} + \cancel{\lambda\delta} - \cancel{\beta\gamma} - \beta^2\alpha - \cancel{\delta\lambda} - \delta^2\alpha}{1 - \beta^2 - \delta^2} \\ &= \cancel{\alpha} + \beta\gamma + \lambda\delta - \frac{\cancel{\alpha(1 - \beta^2 - \delta^2)}}{\cancel{1 - \beta^2 - \delta^2}} \\ &= \beta\gamma + \lambda\delta \end{aligned} \tag{38}$$

□

It is important to mention that although Theorem 4.6 assumes that the variables are standardized, the equations can be easily generalized to the non-standardized variables case. Moreover, the proof is general and can be extended to the case where the two confounders are not independent.

5 SELECTION BIAS

Selection bias occurs when there is collider variable caused by both the sensitive attribute A and the outcome variable Y and the data generation process implicitly conditions on that collider variable. The simplest case is illustrated in Figure 9. Consistent with previous work, a box around a variable indicates that data is generated by conditioning on that variable.

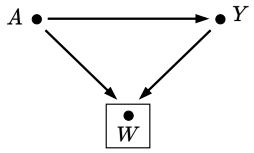

Figure 9: Simple collider structure

5.1 BINARY MODEL

Definition 5.1. *Given the basic collider structure (Figure 9), selection bias is defined as:*

$$\textit{SelBias}(Y, A, W) = StatDisp(Y, A)_W - StatDisp(Y, A) \tag{39}$$

Theorem 5.2. *Assuming that $\mathbb{P}(a_0) = \mathbb{P}(a_1) = \frac{1}{2}$, the difference in discrimination due to selection bias is equal to:*

$$\textit{SelBias}(Y, A) = (1 - \mathbb{P}(w_0|a_0) - \mathbb{P}(w_1))(-\alpha + \beta - \gamma + \delta) \tag{40}$$

where α, β, γ, and δ denote, respectively, $\mathbb{P}(y_1|a_0, z_0), \mathbb{P}(y_1|a_0, z_1), \mathbb{P}(y_1|a_1, z_0)$, and $\mathbb{P}(y_1|a_1, z_1)$.

Proof. The proof is based on the proof of Theorem 4.3. Notice that, conditioning on variable Z in $ACE(Y, A)$ has the same formulation as conditioning on W in $StatDisp(Y, A)_W$. The difference is that the conditioning is on W instead of Z. The other important difference is that in Theorem 4.3, the unconditional expression $StatDisp(A, Y)$ is the biased estimation of the discrimination and the conditional expression $ACE(Y, A)$ is the unbiased estimation. Whereas in Theorem 5.2, it is the opposite: the unconditional expression $StatDisp(A, Y)$ is the unbiased estimation of discrimination and the conditional expression $StatDisp(Y, A)_W$ is the biased estimation. Hence, selection bias is just the opposite of Equation (12) while replacing the variable Z by the variable W. □

5.2 LINEAR MODEL CASE

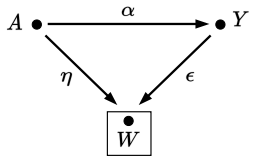

Figure 10: Simple collider structure with linear coefficients.

Theorem 5.3. *Let A, Y, and Z variables with linear regressions coefficients as in Figure 10. Bias due to selection is equal to:*

$$\textit{SelBias}(Y, A) = \frac{\frac{\sigma_{ya}}{\sigma_a^2}\sigma_{wa}^2 - \sigma_{wa}\sigma_{yw}}{\sigma_a^2\sigma_w^2 - \sigma_{wa}^2} \tag{41}$$

Selection bias can also be expressed in terms of the linear regression coefficients as follows:

$$\textit{SelBias}(Y, A) = \epsilon\ \frac{\sigma_a^4\alpha^2\eta + \sigma_a^4\alpha^3\epsilon - \sigma_y^2\sigma_a^2\eta - \sigma_y^2\sigma_a^2\alpha\epsilon}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2} \tag{42}$$

Corollary 5.4. *For standardized variables A, Y, and W, selection bias can be expressed in terms of convariances as:*

$$\textit{SelBias}(Y, A) = \frac{\sigma_{ya}\sigma_{wa}^2 - \sigma_{wa}\sigma_{yw}}{1 - \sigma_{wa}^2} \tag{43}$$

And in terms of regression coefficient:

$$\textit{SelBias}(Y, A) = \epsilon \ \frac{\alpha^2\eta + \alpha^3\epsilon - \eta - \alpha\epsilon}{1 - (\eta + \alpha\epsilon)^2} \tag{44}$$

Equations (43) and (44) can be obtained from Equations (41) and (42) as $\sigma_a = \sigma_w = \sigma_y = 1$.

Proof. For Equation (41),

$$\begin{aligned}
\textit{SelBias}(Y, A) &= \beta_{ya.w} - \beta_{ya} \\
&= \frac{\sigma_w^2\sigma_{ya} - \sigma_{yw}\sigma_{wa}}{\sigma_a^2\sigma_w^2 - \sigma_{wa}^2} - \frac{\sigma_{ya}}{\sigma_a^2} \\
&= \frac{(\sigma_w^2\sigma_{ya} - \sigma_{yw}\sigma_{wa}) - \frac{\sigma_{ya}}{\sigma_a^2}(\sigma_a^2\sigma_w^2 - \sigma_{wa}^2)}{\sigma_a^2\sigma_w^2 - \sigma_{wa}^2} \\
&= \frac{\cancel{\sigma_w^2\sigma_{ya}} - \sigma_{yw}\sigma_{wa} - \frac{\sigma_{ya}}{\cancel{\sigma_a^2}}\cancel{\sigma_a^2}\sigma_w^2 + \frac{\sigma_{ya}}{\sigma_a^2}\sigma_{wa}^2}{\sigma_a^2\sigma_w^2 - \sigma_{wa}^2} \\
&= \frac{\frac{\sigma_{ya}}{\sigma_a^2}\sigma_{wa}^2 - \sigma_{wa}\sigma_{yw}}{\sigma_a^2\sigma_w^2 - \sigma_{wa}^2}
\end{aligned}$$

For Equation (42),

$$\begin{aligned}
\textit{SelBias}(Y, A) &= \beta_{ya.w} - \beta_{ya} \\
&= \frac{\sigma_w^2\sigma_{ya} - \sigma_{yw}\sigma_{wa}}{\sigma_a^2\sigma_w^2 - \sigma_{wa}^2} - \frac{\sigma_{ya}}{\sigma_a^2} \\
&= \frac{\sigma_w^2\sigma_a^2\alpha - (\sigma_y^2\epsilon + \sigma_a^2\alpha\eta)(\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2} - \frac{\cancel{\sigma_a^2}\alpha}{\cancel{\sigma_a^2}} \\
&= \frac{\sigma_w^2\sigma_a^2\alpha - \sigma_y^2\sigma_a^2\epsilon\eta - \sigma_y^2\sigma_a^2\alpha\epsilon^2 - \sigma_a^4\alpha\eta^2 - \sigma_a^4\alpha^2\eta\epsilon}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2} \\
&\quad - \frac{\alpha(\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2)}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2} \\
&= \frac{\cancel{\sigma_w^2\sigma_a^2\alpha} - \sigma_y^2\sigma_a^2\epsilon\eta - \sigma_y^2\sigma_a^2\alpha\epsilon^2 - \cancel{\sigma_a^4\alpha\eta^2} - \cancel{\sigma_a^4\alpha^2\eta\epsilon}}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2} \\
&\quad + \frac{\cancel{-\sigma_a^2\sigma_w^2\alpha} + \cancel{\sigma_a^4\alpha\eta^2} + \cancel{2}\sigma_a^4\alpha^2\eta\epsilon + \sigma_a^4\alpha^3\epsilon^2}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2} \\
&= \epsilon \ \frac{\sigma_a^4\alpha^2\eta + \sigma_a^4\alpha^3\epsilon - \sigma_y^2\sigma_a^2\eta - \sigma_y^2\sigma_a^2\alpha\epsilon}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2}
\end{aligned}$$

□

6 MEASUREMENT BIAS

Measurement bias arises from how particular variable(s) are measured. A common example is when the ideal variable for a model is not measurable/observable and instead we rely on a proxy variable which behaves differently in different groups. Figure 11 shows a simple scenario when measuring accurately the discrimination based on A requires adjusting on variable Z. However, if Z is not measurable but a proxy variable T is measurable, measurement bias occurs when we adjust on T instead of Z.

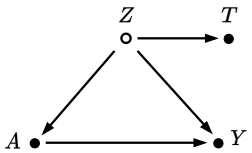

Figure 11: Simple measurement bias structure

6.0.1 BINARY MODEL

Definition 6.1. *Given variables A, Y, Z, and T with causal relations as in Figure 11, measurement bias can be defined as:*

$$\textit{MeasBias}(Y, A) = \textit{StatDisp}(Y, A)_T - \textit{StatDisp}_Z(Y, A) \tag{45}$$

Theorem 6.2. *Assuming that Z is not measurable, but only the error mechanism ($\mathbb{P}(T|Z)$) is available, and that $\mathbb{P}(a_0) = \mathbb{P}(a_1) = \frac{1}{2}$, the difference in discrimination due to measurement bias, $\textit{MeasBias}(Y, A)$ can be expressed in terms of $\mathbb{P}(T|Z)$ as follows:*

$$\begin{aligned}
&\epsilon(\delta - \beta) + (1 - \epsilon)(\gamma - \alpha) \\
&\quad - \epsilon\big(\delta - \beta + 4\mathbb{P}(t_1|z_0)(\beta - \delta + \gamma\Theta + \gamma\Psi)\big)\, Q \\
&\quad - (1 - \epsilon)\big(\gamma - \alpha + 4\mathbb{P}(t_0|z_1)(\alpha - \gamma + \delta + \delta\Psi^{-1} + \beta\Theta^{-1})\big)\, R
\end{aligned} \tag{46}$$

where:

$$\alpha = \mathbb{P}(y_1|a_0, t_0) \quad \gamma = \mathbb{P}(y_1|a_1, t_0) \quad Q = \frac{1 - \frac{\mathbb{P}(t_0|z_1)}{\epsilon}}{1 - \frac{\mathbb{P}(t_0|z_1)}{2\epsilon}} \quad \Phi = \frac{\epsilon + \frac{\tau}{2} - 1}{\epsilon + \frac{\tau}{2} - \frac{1}{2}} \quad \epsilon = \mathbb{P}(t_1)$$

$$\beta = \mathbb{P}(y_1|a_0, t_1) \quad \delta = \mathbb{P}(y_1|a_1, t_1) \quad R = \frac{1 - \frac{\mathbb{P}(t_1|z_0)}{1-\epsilon}}{1 - \frac{\mathbb{P}(t_1|z_0)}{2-2\epsilon}} \quad \Psi = \frac{1 - \tau}{\tau} \qquad \tau = \mathbb{P}(t_0|a_0)$$

Proof. Let $\mathbb{P}(t_1) = \epsilon$ ($\epsilon \in]0, 1[$) and hence $\mathbb{P}(t_0) = 1 - \epsilon$. And let $\mathbb{P}(y_1|a_0, t_0) = \alpha$, $\mathbb{P}(y_1|a_0, t_1) = \beta$, $\mathbb{P}(y_1|a_1, t_0) = \gamma$, and $\mathbb{P}(y_1|a_1, t_1) = \delta$. Finally, let $\mathbb{P}(t_0|a_0) = \tau$. The remaining conditional probabilities of T given A are equal to the following:

$$\begin{aligned}
\mathbb{P}(t_1|a_0) &= 1 - \mathbb{P}(t_0|a_0) = 1 - \tau \\
\mathbb{P}(t_1|a_1) &= \frac{\mathbb{P}(t_1) - \mathbb{P}(t_1|a_0)\mathbb{P}(a_0)}{\mathbb{P}(a_1)} \\
&= 2\epsilon + \tau - 1 \qquad (47) \\
\mathbb{P}(z_0|a_1) &= 1 - \mathbb{P}(z_1|a_1) \qquad (48) \\
&= 2 - 2\epsilon - \tau
\end{aligned}$$

According to Definition 6.1:

$$\textit{MeasBias}(Y, A) = \textit{StatDisp}_T(Y, A) - \textit{StatDisp}_Z(Y, A)$$

By the proof of Theorem 4.3, the first term:

$$\textit{StatDisp}_T(Y, A) = \epsilon(\delta - \beta) + (1 - \epsilon)(\gamma - \alpha)$$

The rest of the proof consists in expressing $\textit{StatDisp}_Z(Y, A)$ in terms of the error term $\mathbb{P}(T|Z)$.

$$\textit{StatDisp}_Z(Y, A) = \mathbb{P}(y_1|do(a_1)) - \mathbb{P}(y_1|do(a_0)) \tag{49}$$

where:

$$\begin{aligned}
&\mathbb{P}(y_1|do(a)) = \\
&\frac{\mathbb{P}(y_1, a, t_1)}{\mathbb{P}(a|t_1)} \frac{\left(1 - \frac{\mathbb{P}(t_1|z_0)}{\mathbb{P}(t_1|a,y_1)}\right)\left(1 - \frac{\mathbb{P}(t_1|z_0)}{\mathbb{P}(t_1)}\right)}{1 - \mathbb{P}(t_1|z_0)\frac{\mathbb{P}(a)}{\mathbb{P}(t_1)}} \\
&+ \frac{\mathbb{P}(y_1, a, t_0)}{\mathbb{P}(a|t_0)} \frac{\left(1 - \frac{\mathbb{P}(t_0|z_1)}{\mathbb{P}(t_0|a,y_1)}\right)\left(1 - \frac{\mathbb{P}(t_0|z_1)}{\mathbb{P}(t_0)}\right)}{1 - \mathbb{P}(t_0|z_1)\frac{\mathbb{P}(a)}{\mathbb{P}(t_0)}}
\end{aligned} \tag{50}$$

The proof can be found in Pearl (2010) (Section 3). Using Bayes rule, we can easily show that

$$\begin{aligned}
\mathbb{P}(y_1, a_1, t_1) &= \epsilon\delta + \frac{\delta\tau}{2} - \frac{\delta}{2} \\
\mathbb{P}(y_1, a_1, t_0) &= \gamma - \epsilon\gamma + \frac{\delta\tau}{2} - \frac{\tau\gamma}{2} \\
\mathbb{P}(y_1, a_0, t_1) &= \frac{\beta}{2} - \frac{\beta\tau}{2} \\
\mathbb{P}(y_1, a_0, t_0) &= \frac{\gamma\tau}{2}
\end{aligned}$$

Using Bayes rule and the marginal conditional probability rule: $\mathbb{P}(A|B) = \sum_{z\in Z} \mathbb{P}(A|B, z)\mathbb{P}(z|B)$, we can easily show that:

$$\begin{aligned}
\mathbb{P}(t_1|a_0, y_1) &= \frac{1}{4}\frac{\beta - \beta\tau}{\alpha\tau + \beta - \beta\tau} \\
\mathbb{P}(t_0|a_0, y_1) &= \frac{1}{4}\frac{\gamma\tau}{\gamma\tau + \beta - \beta\tau} \\
\mathbb{P}(t_1|a_1, y_1) &= \frac{\epsilon\delta + \frac{\delta\tau}{2} - \frac{\delta}{2}}{4\gamma - 4\epsilon\gamma - 2\tau\gamma + 4\epsilon\delta + 2\delta\tau - 2\delta} \\
\mathbb{P}(t_0|a_1, y_1) &= \frac{\gamma - \epsilon\gamma - \frac{\tau\gamma}{2}}{4\gamma - 4\epsilon\gamma - 2\tau\gamma + 4\epsilon\delta + 2\delta\tau - 2\delta}
\end{aligned}$$

Finally, using Bayes rule, we can show that:

$$\begin{aligned}
\mathbb{P}(a_0|t_1) &= \frac{\mathbb{P}(t_1|a_0)\mathbb{P}(a_0)}{\mathbb{P}(t_1)} = \frac{(1-\tau)}{2\epsilon} \\
\mathbb{P}(a_0|t_0) &= \frac{\mathbb{P}(t_0|a_0)\mathbb{P}(a_0)}{\mathbb{P}(t_0)} = \frac{\tau}{2-2\epsilon} \\
\mathbb{P}(a_1|t_1) &= \frac{\mathbb{P}(t_1|a_1)\mathbb{P}(a_1)}{\mathbb{P}(t_1)} = \frac{\epsilon + \frac{\tau}{2} - \frac{1}{2}}{\epsilon} \\
\mathbb{P}(a_1|t_0) &= \frac{\mathbb{P}(t_0|a_1)\mathbb{P}(a_1)}{\mathbb{P}(t_0)} = \frac{(2 - 2\epsilon - \tau}{2-2\epsilon}
\end{aligned}$$

After some algebra, we have:

$$\begin{aligned}
ACE(Y, A) = \epsilon\big(\delta - \beta + 4\mathbb{P}(t_1|z_0)(\beta - \delta + \gamma\Phi + \gamma\Psi)\big)\ Q \\
+ (1-\epsilon)\big(\gamma - \alpha + 4\mathbb{P}(t_0|z_1)(\alpha - \gamma + \delta + \delta\Psi^{-1} + \beta\Phi^{-1})\big)\ R
\end{aligned} \tag{51}$$

□

6.0.2 Linear Model Case

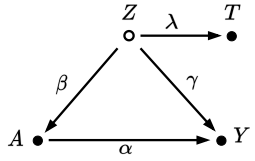

Figure 12: Simple measurement bias structure with linear coefficients.

Theorem 6.3. *Let A, Y, Z, and T variables with linear regressions coefficients as in Figure 12 which represents the basic measurement bias structure. Bias due to measurement error is equal to:*

$$\textit{MeasBias}(Y, A) = \frac{{\sigma_z}^2\beta\gamma({\sigma_t}^2 - {\sigma_z}^2\lambda^2)}{{\sigma_a}^2{\sigma_t}^2 - {\sigma_z}^4\lambda^2\beta^2} \tag{52}$$

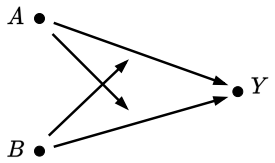

Figure 13: Interaction Bias, where A and B are sensitive variables and Y is an outcome.

Corollary 6.4. *For standardized variables A, Y, Z, and T, measurement bias is equal to:*

$$\textit{MeasBias}(Y, A) = \frac{\beta\gamma(1-\lambda^2)}{1-\lambda^2\beta^2} \tag{53}$$

Proof.

$$\begin{aligned}
\textit{MeasBias}(Y, A) &= \beta_{ya.t} - \beta_{ya.z} \\
&= \frac{\sigma_t^2\sigma_{ya} - \sigma_{yt}\sigma_{ta}}{\sigma_a^2\sigma_t^2 - \sigma_{ta}^2} - \frac{\sigma_z^2\sigma_{ya} - \sigma_{yz}\sigma_{za}}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2} \\
&= \frac{\sigma_t^2(\sigma_a^2\alpha + \sigma_z^2\beta\gamma) - (\sigma_z^2\gamma\lambda + \sigma_z^2\alpha\beta\lambda)(\sigma_z^2\lambda\beta)}{\sigma_a^2\sigma_t^2 - \sigma_z^4\lambda^2\beta^2} - \alpha \qquad (54) \\
&= \frac{\sigma_t^2\sigma_a^2\alpha + \sigma_t^2\sigma_z^2\beta\gamma - \sigma_z^4\gamma\lambda^2\beta - \sigma_z^4\lambda^2\beta^2\alpha}{\sigma_a^2\sigma_t^2 - \sigma_z^4\lambda^2\beta^2} \\
&= \frac{\cancel{\alpha}\,(\cancel{\sigma_t^2\sigma_a^2 - \sigma_z^2\lambda^2\beta^2})}{\cancel{\sigma_a^2\sigma_t^2 - \sigma_z^4\lambda^2\beta^2}} + \frac{\sigma_t^2\sigma_z^2\beta\gamma - \sigma_z^4\gamma\lambda^2\beta}{\sigma_a^2\sigma_t^2 - \sigma_z^4\lambda^2\beta^2} - \cancel{\alpha} \\
&= \frac{{\sigma_z}^2\beta\gamma({\sigma_t}^2 - {\sigma_z}^2\lambda^2)}{{\sigma_a}^2{\sigma_t}^2 - {\sigma_z}^4\lambda^2\beta^2}
\end{aligned}$$

In step (54), $\beta_{ya.z}$ is replaced by α (see proof of Theorem 4.4). □

7 INTERACTION BIAS

Interaction bias takes place in the presence of two sensitive attributes when the value of one sensitive attribute influences the effect of the other sensitive attribute on the outcome. Interaction bias is graphically illustrated in Figure 13. Note that regular DAGs are not able to express interaction. For this reason, we are employing the graphical representation proposed by Weinberg (2007). The arrows pointing to arrows, instead of nodes account for the interaction term. In a binary model, interaction bias coincides with interaction term ($Interaction$) in the case of an intersectional sensitive attribute. Interaction bias also affects the individual measurement of the effect of sensitive attribute A or B.

7.1 BINARY MODEL, INTERSECTIONAL SENSITIVE VARIABLE

Given binary sensitive variables A, B and a binary outcome Y, the joint discrimination of $A = 0$ and $B = 0$ with respect to Y can be defined as follows:

Definition 7.1.

$$\textit{StatDisp}(Y, A, B) = P(Y_1|a_1, b_1) - P(Y_1|a_0, b_0) \tag{55}$$

Here $Y = 1$ is a positive outcome, $A = 1$ and $B = 1$ (a_1, b_1) represent the disadvantaged group.

Theorem 7.2. *Under the assumption of no common parent for A and Y and B and Y [7] we can express StatDisp(Y, A, B) in terms of causal effects of A and B and interaction between A and B on the additive scale:*

[7]This assumption is relatively easy to satisfy in case of immutable sensitive attributes such as gender or race because they are unlikely to have external causes. It is important to control for possible confounders when sensitive attributes can have external causes, for example, political beliefs can be influenced by education.

$$\begin{aligned}\textit{StatDisp}(Y,A,B) &= \big[P(Y_1|a_1,b_0) - P(Y_1|a_0,b_0)\big] + \big[P(Y_1|a_0,b_1) - P(Y_1|a_0,b_0)\big] \\ &\quad + \textit{Interaction}(A,B) \\ \textit{where Interaction}(A,B) &= P(Y_1|a_1,b_1) - P(Y_1|a_0,b_1) - P(Y_1|a_1,b_0) + P(Y_1|a_0,b_0)\end{aligned}$$

Proof. By Definition 7.1:

$$\begin{aligned}\textit{StatDisp}(Y,A,B) &= P(Y_1|a_1,b_1) - P(Y_1|a_0,b_0) \\ &= P(Y_1|a_1,b_1) - P(Y_1|a_0,b_0) + P(Y_1|a_0,b_0) - P(Y_1|a_0,b_0) \\ &\quad + P(Y_1|a_1,b_0) - P(Y_1|a_1,b_0) + P(Y_1|a_0,b_1) - P(Y_1|a_0,b_1) \\ &= P(Y_1|a_1,b_0) - P(Y_1|a_0,b_0) + P(Y_1|a_0,b_1) - P(Y_1|a_0,b_0) \\ &\quad + P(Y_1|a_1,b_1) - P(Y_1|a_0,b_1) - P(Y_1|a_1,b_0) + P(Y_1|a_0,b_0) \\ &= \big[P(Y_1|a_1,b_0) - P(Y_1|a_0,b_0)\big] + \big[P(Y_1|a_0,b_1) - P(Y_1|a_0,b_0)\big] \\ &\quad + \textit{Interaction}(A,B)\end{aligned}$$

□

Notice that: $P(Y_1|a_1,b_0) - P(Y_1|a_0,b_0)$ is the effect of A on Y in case there is no interaction, and similarly for B: $P(Y_1|a_0,b_1) - P(Y_1|a_0,b_0)$ is the effect of B on Y in case there is no interaction. To avoid confusion, we denote such expressions as $SD_{\not{Int}}(Y,A)$ and $SD_{\not{Int}}(Y,B)$ respectively.

Definition 7.3. *Under the assumption of no confounders between A and Y on one hand, and between B and Y on the other hand, adding up the single effects of A and B on Y to estimate the discrimination due to both sensitive variables StatDisp(Y,A,B) leads to a biased estimation. The amount of the bias (StatDisp) coincides with the interaction term as follows:*

$$\begin{aligned}\textit{IntBias}(Y,A,B) &= \textit{StatDisp}(Y,A,B) \\ &\quad - \big[SD_{\not{Int}}(Y,A) + SD_{\not{Int}}(Y,B)\big] \\ &= \textit{Interaction}(A,B)\end{aligned}$$

The presence of the *Interaction*(A,B) in the case of two sensitive variables is very common. Rothman et al. (2008) distinguish 16 combinations of the effects of binary A and B on the binary outcome Y. Only six of those cases correspond to *Interaction*$(A,B) = 0$, which means that there is no interaction. Indeed, the interaction is absent only when at least one of the terms A and B has no effect on Y Rothman et al. (2008). Unfortunately, most of the time the numeric value of interaction does not indicate which particular case (out of 16 combinations of the effects of A and B) is dominant in the data. However, VanderWeele and Robins show that under sufficient-component-cause framework and assumption of monotonic effect of A and B [8] if $P(Y_1|a_1,b_1) - P(Y_1|a_0,b_1) - P(Y_1|a_1,b_0) > 0$, then the synergism between $A = 1$ and $B = 1$ must be present Rothman et al. (2008); VanderWeele & Robins (2007). In the fairness scenario, this means that two privileged groups have a synergetic effect on the positive outcome. In terms of the previous example, it is a situation where only conservative men are hired.

7.2 Binary model, Individual Sensitive Variable

Given binary sensitive variables A, B and a binary outcome Y, the discrimination with respect to only A (and similarly for B) can be expressed as follows:

$$\textit{StatDisp}(Y,A) = P(Y_1|a_1) - P(Y_1|a_0) \tag{56}$$

Theorem 7.4. *Under previously introduced assumption of no confounding, the discrimination with respect to A can be decomposed into an interaction free discrimination and the interaction between A and B:*

$$\textit{StatDisp}(Y,A) = SD_{\not{Int}}(Y,A) + P(b_1)\textit{Interaction}(A,B)$$

[8] monotonic effect means, that an intervention either increases or decreases outcome Y for every individual.

Proof.

$$\begin{aligned}
\textit{StatDisp}(Y,A) &= \mathbb{P}(Y_1|a_1) - \mathbb{P}(Y_1|a_0) \\
&= \sum_b \mathbb{P}(Y_1|a_1,b)\mathbb{P}(b|a_1) - \sum_b \mathbb{P}(Y_1|a_0,b)\mathbb{P}(b|a_0) \\
&= \mathbb{P}(Y_1|a_1,b_1)\mathbb{P}(b_1|a_1) + \mathbb{P}(Y_1|a_1,b_0)\mathbb{P}(b_0|a_1) \\
&\quad - \mathbb{P}(Y_1|a_0,b_1)\mathbb{P}(b_1|a_0) - \mathbb{P}(Y_1|a_0,b_0)\mathbb{P}(b_0|a_0) \\
&= \mathbb{P}(Y_1|a_1,b_1)\mathbb{P}(b_1|a_1) + \mathbb{P}(Y_1|a_1,b_0)\mathbb{P}(1-\mathbb{P}(b_1|a_1) \\
&\quad - \mathbb{P}(Y_1|a_0,b_1)\mathbb{P}(b_1|a_0) - \mathbb{P}(Y_1|a_0,b_0)\mathbb{P}(1-\mathbb{P}(b_1|a_0) \\
&= \mathbb{P}(b_1|a_1)\big(\mathbb{P}(Y_1|a_1,b_1) - \mathbb{P}(Y_1|a_1,b_0)\big) + \mathbb{P}(Y_1|a_1,b_0) \\
&\quad + \mathbb{P}(b_1|a_0)\big(\mathbb{P}(Y_1|a_0,b_0) - \mathbb{P}(Y_1|a_0,b_1)\big) - \mathbb{P}(Y_1|a_0,b_0)
\end{aligned}$$

Since A and B are independent $\mathbb{P}(b_1|a_1) = \mathbb{P}(b_1|a_0) = \mathbb{P}(b_1)$. It follows that:

$$\begin{aligned}
\textit{StatDisp}(Y,A) &= \mathbb{P}(b_1)Int(A,B) + \mathbb{P}(Y_1|a_1,b_0) - \mathbb{P}(Y_1|a_0,b_0) \\
&= \mathbb{P}(b_1)Int(A,B) + SD_{\cancel{Int}}(Y,A)
\end{aligned}$$

□

StatDisp(Y,B) can be decomposed in a similar way. Interaction bias *IntBias*(Y,A) can then be defined as:

Definition 7.5.

$$\begin{aligned}
\textit{IntBias}(Y,A) &= \textit{StatDisp}(Y,A) - SD_{\cancel{Int}}(Y,A) \\
&= P(b_1)\textit{Interaction}(A,B)
\end{aligned}$$

IntBias(Y,B) can be defined similarly:

$$\begin{aligned}
\textit{IntBias}(Y,B) &= \textit{StatDisp}(Y,B) - SD_{\cancel{Int}}(Y,B) \\
&= P(a_1)\textit{Interaction}(A,B)
\end{aligned}$$

7.3 Linear Model Case

Given the true model:

$$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 AB + \beta_4 C \tag{57}$$

And biased model, that does not include interaction term β_3:

$$Y = \beta_0^{'} + \beta_1^{'} A + \beta_2^{'} B + \beta_4^{'} C \tag{58}$$

Where A and B are binary sensitive attributes, C is a set of covariates and Y is a continuous outcome (for example a credit score).

The change in Y due to A is $\beta_1 + \beta_3 B$ and, similarly the change in Y due to B is $\beta_2 + \beta_3 A$ Keele & Stevenson (2021). In this case, a measure of effect of A ($\beta_1^{'}$) or B ($\beta_2^{'}$) without an interaction term would be inaccurate. Next we define the bias introduced by not accounting for the interaction between two sensitive attributes.

7.3.1 Linear model, Intersectional Sensitive Variable

Theorem 7.6. *Let A,B and Y be variables with linear regression coefficients as in Equation 58. In a linear model with binary A and B the bias due to interaction, when measuring the effect of intersectional sensitive variable A and B on Y (StatDisp(Y,A,B)) is equal to:*

$$\begin{aligned}
\textit{IntBias}(Y,A,B) &= (\beta_1^{'} + \beta_2^{'}) - (\beta_1 + \beta_2) \\
&= \beta_3
\end{aligned}$$

Proof. $\beta_1^{'} + \beta_2^{'}$ represents the causal effect of A and B on Y including interaction, whereas $\beta_1 + \beta_2$ represents the same causal effect but without interaction. The difference coincides with the interaction coefficient β_3. □

Intuitively, β_3 is part of an effect of the intersectional sensitive variable $A = 1, B = 1$ on Y that is left out of the estimation when fitting linear regression without the interaction term.

7.3.2 Linear model, Individual Sensitive Variable

The difference of measurement of effect of A on Y with interaction term ($\beta_1^{'}$) and without interaction term (β_1) depends on the value of B.

Theorem 7.7. *Let A, B and Y be variables with linear regression coefficients as in Equation 58. In a linear model with binary A and B the bias due to interaction, when measuring the effect of A on Y (StatDisp(Y, A)) is equal to:*

$$\begin{aligned} \textit{IntBias}(Y, A) &= \beta_1^{'} - \beta_1 \\ &= \beta_3 \mathbb{P}(B_1) \end{aligned}$$

Proof. *StatDisp*(Y, A) measures how wrong is the evaluation of effect of $A = 1$ on average, for cases where $B = 1$ or $B = 0$, which are as follows:

$$\beta_1^{'} = \begin{cases} \beta_1 + \beta_3 & \text{when} \quad B = 1 \\ \beta_1 & \text{when} \quad B = 0 \end{cases} \tag{59}$$

Note that the *StatDisp*(Y, A) is dependent on β_3 and the probability of $B = 1$ (and, similarly, *StatDisp*(B, A) is dependent on β_3 and the probability of $A = 1$). □

8 Bias Analysis

Expressing different types of bias in terms of the model parameters (conditional probabilities and regression coefficients) allows to study the behavior of bias and how it is impacted by the different parameters. In particular, at which parameters value it is peaked and at which other values it is absent. The aim is to identify the cases where a given estimation of discrimination is biased and at which extent.

8.1 Binary Case

Confounder Bias is absent when at least one of the two terms of Equation 12 is equal to 0. For the first term ($1 - \mathbb{P}(z_0|a_0) - \mathbb{P}(z_1) = 0$), it is easy to show that it is equivalent to $\mathbb{P}(z_0|a_1) = \mathbb{P}(z_0)$ which is in turn means that Z and A are independent ($A \perp\!\!\!\perp Z$).

The second term is equal to 0 when :

$$\mathbb{P}(y_1|a_0, z_0) - \mathbb{P}(y_1|a_0, z_1) = -(\mathbb{P}(y_1|a_1, z_0) - \mathbb{P}(y_1|a_1, z_1)) \tag{60}$$

Thre right-hand side can be interpreted as the Contolled Direct Effect (CDE) VanderWeele (2011) of Z on Y when $A = 0$ whereas the left-hand side is the opposite of $\mathbb{P}(y_1|a_1, z_0) - \mathbb{P}(y_1|a_1, z_1)$ which is the CDE of Z on Y when $A = 1$. Confounding bias is equal zero, when the CDE of Z on Y when $A = 1$ is the exact opposite of to that when $A = 0$. In the job hiring example of Figure 2, it means that we privilege poor liberals as much as we privilege rich conservatives, therefore the effect $Z- > Y$ is canceled out. Equation 60 can also hold when both sides are equal to 0. This means that Z has no direct effect on Y (no edge between Z and Y). Z can still have effect on Y which is mediated through A, but it does not have a role as a confounder. To summarize, confounding bias is absent in three cases: either $A \perp\!\!\!\perp Z$ (A and Z are independent) or the edge $Z \rightarrow Y$ is absent, or the CDE of Z on Y when $A = 0$ and $A = 1$ are opposite and hence cancel each others.

Confounding bias is peaked when the first term ($1 - \mathbb{P}(z_0|a_0) - \mathbb{P}(z_1)$) is equal to 1 or -1 *and* the second term ($-\alpha + \beta - \gamma + \delta$) is equal to 2 or -2. The first term is equal to 1 when $\mathbb{P}(z_1) = 0$ and

$\mathbb{P}(z_0|a_0) = 0$. This is an extreme situation when all data instances have the same values of A and Z variables, that is, a_1 and z_0. The same term is equal to -1 when $\mathbb{P}(z_1) = 1$ and $\mathbb{P}(z_0|a_0) = 1$ which corresponds to the other extreme situation of all data instances have a_0 and z_0. In the job hiring example, both cases correspond to a situation when all candidates are of the same type: poor liberals or rich liberals. The second term reaches a peak value (2.0 or -2.0) when the CDE of Z on Y is maximum (1 or -1) for both a_0 and a_1. To summarize, confounding bias is optimal when the effect through the edge $Z \rightarrow A$ is very strong (first term) *and* the effect through the edge $Z \rightarrow Y$ is very strong (second term). This optimal situation can be seen as an extreme case of Simpson's paradox Simpson (1951).

Collider Bias Collider bias can be viewed as an inverse case of a confounder bias. While confounder bias compromises internal validity, selection bias is a threat to external validity Haneuse (2016). Similarly as confounder bias, collider bias does not manifest if the direct link between A and W or Y and W is absent, or the link between W and Y is the opposite for the values $A = 1$ and $A = 0$. The bias is maximized when the group corresponding to $A = 1$ and $W = 0$ is very large (the negative bias case would occur if the group $A = 1$ and $W = 1$ is dominant). Maximization of bias also requires that the link from Y to W is deterministic and has the same direction for both values of A.

Measurement Bias depends heavily on $\mathbb{P}(T|Z)$. For instance, from Theorem 6.2, it is easy to show that if $\mathbb{P}(t_0|z_1) = \mathbb{P}(t_1|z_0) = 0$ (T and Z are fully dependent), then $Q = R = 1$, and consequently measurement bias disappears. Conversely, if $\mathbb{P}(t_0|z_1) = \mathbb{P}(t_1) = \epsilon$ and $\mathbb{P}(t_1|z_0) = \mathbb{P}(t_0) = 1 - \epsilon$ (T and Z are independent), then $Q = R = 0$, and consequently, measurement bias is maximized as the two negative terms of Equation equation 46 disappear. The maximum value of measurement bias in that case is $\epsilon(\delta - \beta) + (1 - \epsilon)(\gamma - \alpha)$.

Interaction Bias Interaction bias for the intersectional case coincides with the interaction term. More precisely, it is maximized when the interaction is maximized and diminishes when the interaction is small. Note that the interaction is equal 0 when one of the sensitive attributes does not have an effect on Y Rothman et al. (2008). The interaction bias when measuring the effect of one sensitive attribute A or B on Y depends on the interaction term and the probability of $B = 1$ and $A = 1$, respectively. The bias increases with the probability of $A = 1$ or $B = 1$ and the interaction term. Interaction bias is equal to zero when either interaction, to the probability of $B = 1$ or $A = 1$, respectively, is equal to 0.

8.2 Linear Case

To analyze the different types of bias in the linear case, we generate synthetic data according to the following models. Without loss of generality, the range of possible values of all coefficients ($\alpha, \beta, \gamma, \eta, \epsilon$, and δ) is $[-1.0, 1.0]$:

Confounding Structure:	*Colliding Structure:*
$Z = \mathcal{U}_z,$	$A = \mathcal{U}_a,$
$A = \beta Z + \mathcal{U}_a,$	$Y = \alpha A + \mathcal{U}_y,$
$Y = \alpha A + \gamma Z + \mathcal{U}_y$	$W = \eta A + \epsilon Y + \mathcal{U}_w$

Measurement Structure:	
$Z = \mathcal{U}_z,$	$\mathcal{U}_z \sim \mathcal{N}(0, 1),$
$A = \beta Z + \mathcal{U}_a,$	$\mathcal{U}_a \sim \mathcal{N}(0, 1),$
$Y = \alpha A + \gamma Z + \mathcal{U}_y,$	$\mathcal{U}_y \sim \mathcal{N}(0, 1),$
$T = \delta Z + \mathcal{U}_t$	$\mathcal{U}_w \sim \mathcal{N}(0, 1),$
	$\mathcal{U}_t \sim \mathcal{N}(0, 1).$

Figure 16 shows the magnitude of each type of bias based on the expressions obtained in Sections 4, 5, and 6. In particular, Equations 17 for confounding bias, 42 for selection bias, and 52 for measurement bias. Three dimensions plot is used for confounding bias (Figure 16(a)) as bias is expressed in terms of two variables (β and γ) whereas four dimensions plots are used for selection and measurement biases (three variables). Confounding bias is maximized when both β and γ have extreme values ($+1.0$ or -1.0): positive bias when β and γ are of the same sign, and negative otherwise. Bias is absent when at least one of the coefficients is zero. In between these extreme cases, confounding bias has strictly linear relation with β whereas a non-linear relation with γ. More importantly, confounding bias is more sensitive to β than to γ particularly for extreme values (when coefficients are close to $+1.0$ or -1.0). That is, modifying the effect of the confounder (e.g. Z) on the sensitive variable (e.g. A) has more impact on the confounding bias than modifying the effect of the confounder on the outcome variable (e.g. Y) with the same amount. In the job hiring example (Section 3.1) this means that the effect of Socio-Economic status on politicial belief has more impact on the counfounding

bias than the effect of socio-economic status on job hiring. However, if the variables are standardized, both effects contribute equally to confounding bias (Corollary 4.5).

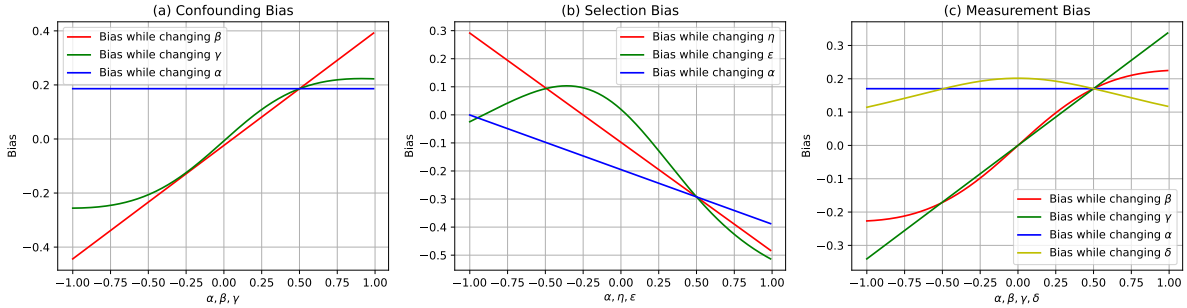

Figure 14: Bias Magnitude while changing one variable and holding the other variables at 0.5.

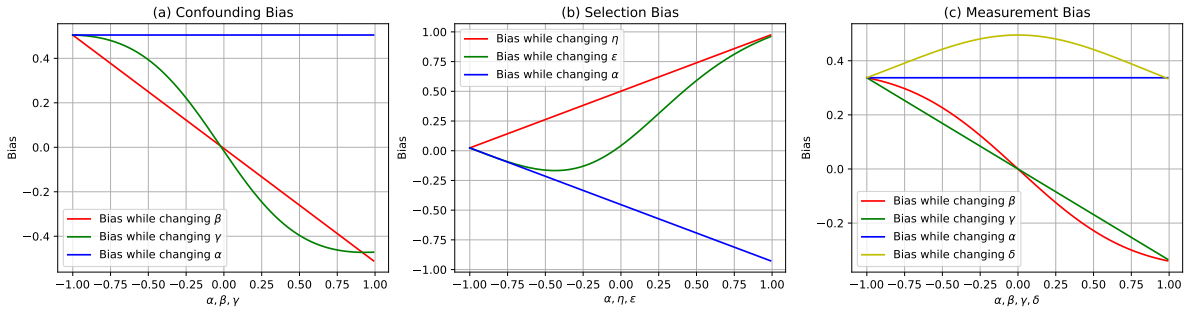

Figure 15: Bias Magnitude while changing one variable and holding the other variables at -1.0.

Unlike confounding bias, the magnitude of selection bias (Figure 16(b)) depends also on the regression coefficient of Y on A (α). Selection bias is peaked in two cases depending on the value of α. First, when η and ϵ have the same extreme values (1 or -1) and $\alpha = 1$. This leads to maximal negative bias. Second, when η and ϵ have extreme but different sign values (1 or -1) and $\alpha = -1$. This corresponds to maximal positive bias. Intuitively, conditioning on the collider variable W introduces a spurious effect between the two causes A on Y: any information "explaining away" one cause will make the other cause more plausible. Using the job hiring example (Figure 3), if there is maximum negative discrimination based on the political beliefs of the candidates ($\alpha = -1$) and we measure discrimination using only labor union records, while political belief and job hiring have strong but opposite effects on labor union membership, the selection bias will be maximum to the point it cancels out all positive discrimination and leads to a conclusion of no discrimination. Figure 16(b) shows also that selection bias disappears when ϵ is zero, but not when η is zero. When $\epsilon \neq 0$, selection bias can be zero depending on the value of ϵ as follows: $\epsilon = 1$ and $\alpha = -\eta$ or $\epsilon = -1$ and $\alpha = \eta$. Overall, selection bias has linear relation with both α and η, whereas non-linear relation with ϵ[9].

Similarly to confounding and selection, measurement bias (Figure 16(c)) is peaked when β and γ have extreme values (1.0 or -1.0) but when $\delta = 0$. This is expected as, by definition, the more Z and T are independent, the higher measurement bias is. Conversely, the plot shows that measurement bias fades away as δ departs from 0[10].

[9]Such relations can be observed more clearly using 2D plots (Figures 14 and Figure 15).

[10]The 2D plots in the appendix show clearly these observations.

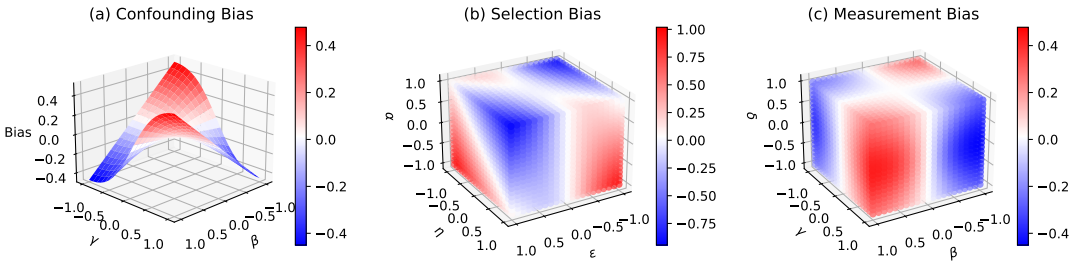

Figure 16: Bias magnitude in the linear case

9 Bias Magnitude: An empirical analysis

We use well-known fairness benchmark data sets Le Quy et al. (2022) for the experiments on real data: Adult [11], Boston housing [12], Compas Angwin et al. (2022), Communities and crimes [13] and Dutch census [14] data. The causal experiments on the real data are limited by the availability of true causal graphs for the benchmark fairness datasets. Furthemore, Binkytė-Sadauskienė et al. (2022) shows, that obtaining reliable causal graphs with causal discovery algorithms is a complicated task. However, we assume that the graphs in the literature are true for a given real dataset. We use the graphs by Zhang et al. (2018); Huan et al. (2020) for Adult and Dutch data sets to measure the interaction bias. For measuring confounder and collider biases we rely on graphs obtained by Binkytė-Sadauskienė et al. (2022) for Communities and Crimes, Boston Housing, Compas, and Dutch datasets (Appendix A). For measurement bias, we use synthetic data because the required structure is not present in the available graphs for the benchmark data sets. Synthetic data is generated according to the following models:

The variables Z, A, T, and Y are binary Bernoulli variables controlled by the parameter p_1. Conditional dependencies of the measurement bias structure define how the parameter p_1 depends on the value of the parent variables.

$$Z \sim (p) = \begin{cases} p_1, \\ p_0 = 1 - p_1 \end{cases} \qquad A \sim (Z; p) = \begin{cases} p_1, & \textit{if } Z = 1, \\ p_0 = 1 - p_1 & \\ p_1^{'}, & \textit{if } Z = 0. \\ p_0^{'} = 1 - p_1^{'} & \end{cases}$$

$$T \sim (Z; p) = \begin{cases} p_1, & \textit{if } Z = 1, \\ p_0 = 1 - p_1 & \\ p_1^{'}, & \textit{if } Z = 0. \\ p_0^{'} = 1 - p_1^{'} & \end{cases} \qquad Y \sim (p; Z, A) = \begin{cases} p_1 = 0.5 * z + 0.5 * a, \\ p_0 = 1 - p_1 \end{cases}$$

The parameters p_1, p_0, $p_1^{'}$ and $p_0^{'}$ are generated randomly and take value between 0 and 1.

Although we cannot claim that the causal structure that we use for the experiments is the ground truth, it is useful for experimentally demonstrating the behavior of causal biases. In addition, the considered causal structures most often show the presence of multiple causal biases at once. However, for the purposes of illustration, we control for a single type of bias separately. More precisely, we consider the difference in measured discrimination with the presence of the absence of a certain type of bias.

The experimental results for confounder bias show that the biases for each individual confounding variable are not significant (Figure 17). However, its magnitude increases and can cancel out the value for statistical disparity (Dutch data set), when multiple confounders are considered simultaneously

[11] https://archive.ics.uci.edu/dataset/2/adult
[12] http://lib.stat.cmu.edu/datasets/boston
[13] https://archive.ics.uci.edu/dataset/183/communities+and+crime
[14] https://microdata.worldbank.org/index.php/catalog/2102/data-dictionary

(Figure 18). Measurement bias takes the highest value for *Synthetic2* dataset (Figure 20). The effect of A on Y when controlling for T appears smaller than when controlling for Z. Here, the value of T is highly dependent on Z if $Z = 0$, but only loosely dependent on Z if $Z = 1$. The prior probability of Z conditions it to take value $Z = 1$ with probability 0.95. Therefore, the link between Z and T is weak. The weak link between the variables makes T a bad predictor for Z and introduces a high measurement bias. Collider bias (Figure 19) is significant if it was introduced by conditioning on income (adult data), age (Compas data), economic status (Dutch data), poverty, unemployment, or divorce (Communities and crime data). Collider bias would reverse the value of statistical disparity, showing discrimination against the privileged group instead of discrimination against the disadvantaged group. We observe a portion of the interaction in all cases of the intersectional sensitive attribute (Figure 21). However, the value of synergism is negative, which means that it is not present in the data. Measurement of interaction bias for A and B individually can yield different values of interaction bias (Figure 22). Although the interaction term is symmetric for A and B, the interaction bias value is also dependent on the probability $B = 1$ (when measuring *IntBias*(Y, A)) or $A = 1$ (when measuring *IntBias*(Y, B)). Therefore, for example, the interaction bias for sex is higher than for age in the Adult data set, because the probability of value 1 for age is higher than the probability of the sex variable taking value 1. Furthermore, we observe that the statistical disparity does not always correspond to the sum of interaction bias and statistical disparity without interaction ($StatDisp(Y, A) \neq SD_{\not{Int}}(Y, A) + P(b_1)Interaction(A, B)$), as required in Theorem 7.4. This observation suggests that the two sensitive variables A and B are not independent as suggested by the graphs provided by Zhang et al. (2018); Huan et al. (2020). Indeed, the graphs discovered by Binkytė-Sadauskienė et al. (2022) show the dependency between age and sex variables in the Dutch data set (Appendix A, Figure 29).

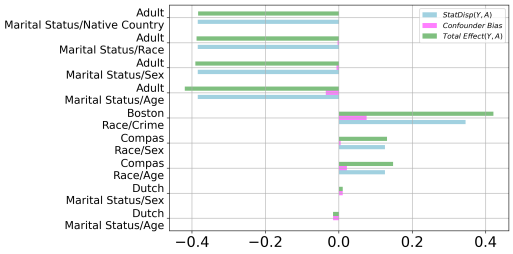

Figure 17: Confounder bias, when treating each confounder separately.

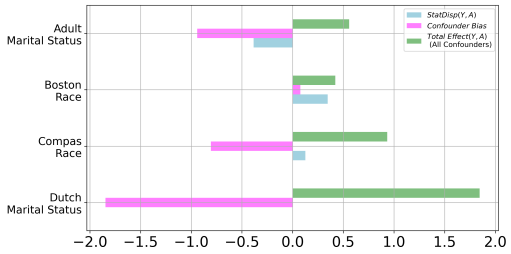

Figure 18: Confounder bias when treating all confounders together.

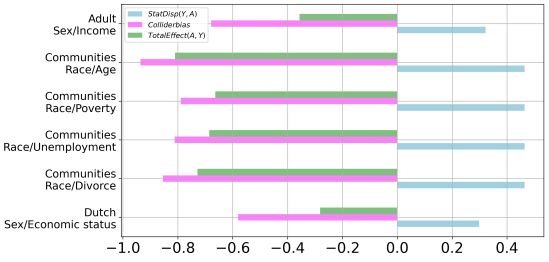

Figure 19: Collider bias, when treating each confounder separately.

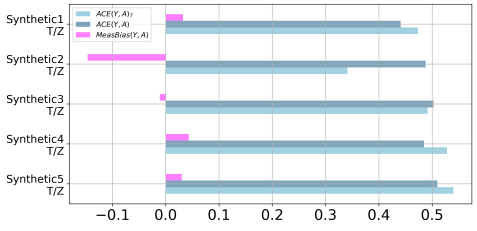

Figure 20: Measurement bias. Synthetic data.

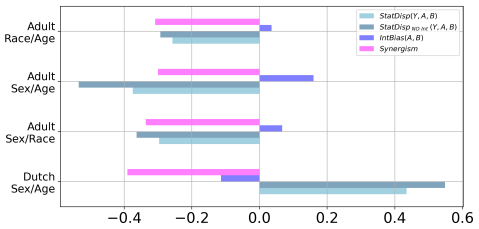

Figure 21: Interaction bias, intersectional sensitive variable.

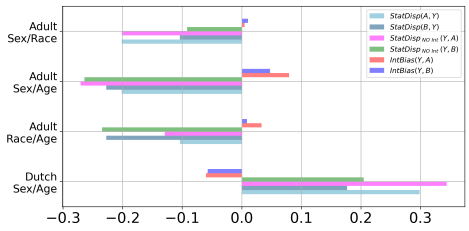

Figure 22: Interaction bias for individual sensitive attributes.

10 CONCURRENT BIASES

Confounding and selection biases. In presence of one or several confounder and collider variables, the estimation of discrimination can suffer from both confounding and selection biases simultaneously. Figure 23 shows the simplest case. According to Definitions 4.1 and 5.1, confounding bias can be isolated by adjusting on the confounder variable $\mathit{ConfBias}(Y,A) = \mathit{StatDisp}(Y,A) - \mathit{StatDisp}_Z(Y,A)$[15] ($\beta_{ya} - \beta_{ya.z}$ in the linear case), whereas selection bias can be isolated by cancelling the adjustment on the collider variable $\mathit{SelBias}(Y,A) = \mathit{StatDisp}_W(Y,A) - \mathit{StatDisp}(Y,A)$ ($\beta_{ya.w} - \beta_{ya}$ in the linear case). The total bias in presence of both types of bias can then be estimated as $\mathit{StatDisp}_W(Y,A) - \mathit{StatDisp}_Z(Y,A)$ in the binary case and $\beta_{ya.w} - \beta_{ya.z}$ in the linear case.

Confounding and measurement biases. Measurement bias (Figure 11) is defined as the difference in estimating *StatDisp* when adjusting on the proxy variable (T) instead of the unobservable/unmeasurable confounder variable (Z). For the binary case, it corresponds to the difference $\mathit{StatDisp}_T(Y,A) - \mathit{StatDisp}_Z(Y,A)$. For the linear case, it corresponds to the difference between the partial regression coefficients $\beta_{ya.t} - \beta_{ya.z}$. The difference between the adjustment free estimation of $\mathit{StatDisp}(Y,A)$ (the regression coefficient β_{ya} in the linear case) and $\mathit{StatDisp}_T(Y,A)$ ($\beta_{ya.t}$) corresponds to the total of both confounder and measurement biases.

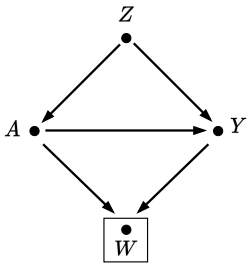

Figure 23: Confounding and colliding bias.

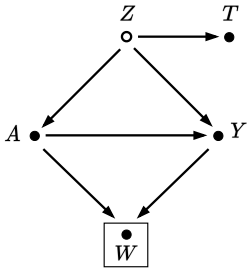

Figure 24: Confounding, colliding, and measurement bias

Selection and measurement biases. Figure 24 shows the simplest case where measurement and selection biases occur simultaneously. Adjusting on both the proxy (T) and the collider (W) variables ($\mathit{StatDisp}_{TW}(Y,A)$ and $\beta_{ya.tw}$) leads to both types of biases occurring simultaneously. Substracting $\mathit{StatDisp}_Z(Y,A)$ (respectively β_{ya}) from $\mathit{StatDisp}_{TW}(Y,A)$ (respectively $\beta_{ya.tw}$) coincides with the sum of selection and measurement biases in the binary and linear cases respectively.

Confounding, selection, and measurement biases. In the same simple case of Figure 24, the difference between adjusting on variables T and W on one hand and adjusting on Z on the other hand ($\mathit{StatDisp}_{TW}(Y,A) - \mathit{StatDisp}_Z(Y,A)$ in the binary case and $\beta_{ya.tw} - \beta_{ya.z}$ in the linear case) encompasses the three types of bias.

Confounding and interaction biases. In presence of interaction between two sensitive variables, confounding bias can be decomposed into interaction free portion and an interaction term. Figure 25

[15]Notice that, by the backdoor formula, $\mathit{StatDisp}_Z(Y,A)$ coincides with $ACE(Y,A)$.

shows a simple confounding structure between A and Y and a second sensitive variable B which is interacting with the effect of A on Y. In the binary case, the confounding bias *ConfBias*(Y, A) (Definition 4.1) can be decomposed as follows:

Proposition 10.1.

$$\begin{aligned} \textit{ConfBias}(Y,A) &= \textit{StatDisp}(Y,A) - \textit{StatDisp}_Z(Y,A) \\ &= SD_{\cancel{Int}}(Y,A) - SD_{\cancel{Int}_Z}(Y,A) \\ &\quad + P(b_1)(Interaction(A,B) - Interaction_Z(A,B)) \end{aligned} \tag{61}$$

$$\tag{62}$$

where

$$SD_{\cancel{Int}_Z}(Y,A) = \sum_Z (P(y_1|a_1,b_0,z) - P(y_1|a_0,b_0,z))P(z)$$

$$\begin{aligned} Interaction_Z(A,B) = \sum_Z \big(&P(y_1|a_1,b_1,z) - P(y_1|a_0,b_1,z) \\ &- P(y_1|a_1,b_0,z) + P(y_1|a_0,b_0,z)\big)P(z) \end{aligned}$$

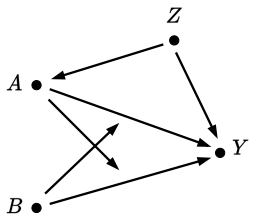

Figure 25: Interaction and confounding bias

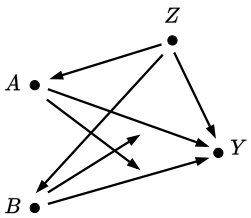

Figure 26: Interaction and confounding bias

In the same example of Figure 25, the confounding bias in case of intersectionality (two interacting sensitive variables) can be decomposed as follows:

Proposition 10.2.

$$\begin{aligned} \textit{ConfBias}(Y,A,B) &= \textit{StatDisp}(Y,A,B) - \textit{StatDisp}_Z(Y,A,B) \\ &= SD_{\cancel{Int}}(Y,A) - SD_{\cancel{Int}_Z}(Y,A) \\ &\quad + Interaction(A,B) - Interaction_Z(A,B) \end{aligned} \tag{63}$$

$$\tag{64}$$

In the slightly different structure where Z is also a confounder between B and Y (Figure 26), the term $SD_{\cancel{Int}}(Y,B) - SD_{\cancel{Int}_Z}(Y,B)$ needs to be added to the *ConfBias*(Y, A, B) expression above.

11 Conclusion

Several sources of bias have been described in the literature of Oxford (2021); Mehrabi et al. (2021). However, unlike existing work which typically do not define sources of bias formally, we provide closed-form expressions of a specific class of biases, namely causal biases. By analyzing the magnitude of bias in terms of the model parameters, we could establish an intuitive interpretation of bias based on the causal graph structure underlying each type of bias. Additionally, we provide in Appendix 11 an analysis of cases where two or more types of biases are present simultaneously. We strongly believe that a better understanding of the magnitude of causal biases, and more generally all sources of bias, will help ML fairness practitioners accurately predict the impact of proposed policies (e.g. training programs, awarness campaigns, establishing quotas, etc.) on existing discrimination.

REFERENCES

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. propublica. *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*, 2016.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications, 2022.

Rūta Binkytė-Sadauskienė, Karima Makhlouf, Carlos Pinzón, Sami Zhioua, and Catuscia Palamidessi. Causal discovery for fairness. *arXiv preprint arXiv:2206.06685*, 2022.

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

Harald Cramér. *Mathematical methods of statistics*, volume 43. Princeton university press, 1999.

Sebastien Haneuse. Distinguishing selection bias and confounding bias in comparative effectiveness research. *Medical care*, 54(4):e23, 2016.

Wen Huan, Yongkai Wu, Lu Zhang, and Xintao Wu. Fairness through equality of effort. In *Companion Proceedings of the Web Conference 2020*, pp. 743–751, 2020.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Luke Keele and Randolph T. Stevenson. Causal interaction and effect modification: same model, different concepts. *Political Science Research and Methods*, 9(3):641–649, 2021. doi: 10.1017/psrm.2020.12.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

University of Oxford. Catalogue of bias. https://catalogofbias.org/biases, 2021. Accessed: 2023-03-30.

Catherine O'Neill. *Weapons of math destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishers, 2016. ISBN 0553418815.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pp. 425–432, 2010.

Judea Pearl. Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference*, 1(1):155–170, 2013.

Kimberly Quick. The unfair effects of impact on teachers with the toughest jobs. *The Century Foundation*, 2015. https://tcf.org/content/commentary/the-unfair-effects-of-impact-on-teachers-with-the-toughest-jobs/?agreed=1.

John Rawls. *A theory of justice: Revised edition*. Harvard university press, 2020.

Kenneth J Rothman, Sander Greenland, Timothy L Lash, et al. *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008.

Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

Tyler J VanderWeele. Controlled direct and mediated effects: definition, identification and bounds. *Scandinavian Journal of Statistics*, 38(3):551–563, 2011.

Tyler J VanderWeele and James M Robins. The identification of synergism in the sufficient-component-cause framework. *Epidemiology*, 18(3):329–339, 2007.

Clarice R Weinberg. Can dags clarify effect modification? *Epidemiology (Cambridge, Mass.)*, 18(5): 569, 2007.

Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

Lu Zhang, Yongkai Wu, and Xintao Wu. Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 31(11):2035–2050, 2018.

A CAUSAL GRAPHS

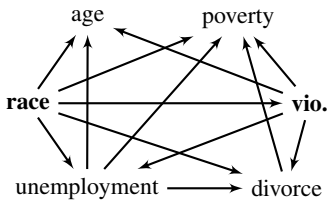

Figure 27: The graph for the communities and crime dataset. 'divorce', 'age', 'poverty' and 'unemployement' are the colliders between 'race' and 'violence' (vio.). The graph is produced using LiNGAM algorithm.

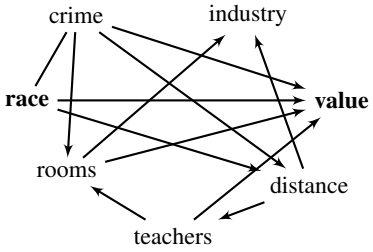

Figure 28: The graph for the Boston housing data set. 'Crime' is a possible confounder between 'race' and 'value'.The graph is produced using GES algorithm.

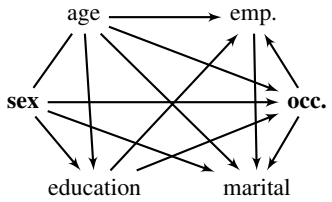

Figure 29: The graph for the Dutch data set. 'Marital Status' is a collider between 'sex' and 'occupation' (occ.). The graph is produced using GES algorithm.

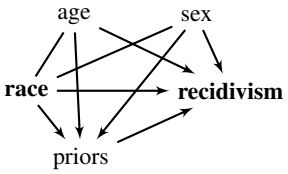

Figure 30: The graph for the Compas dataset. 'Age' and 'sex' are possible confounders between 'race' and 'recidivism'. The graph is produced using PC algorithm.