Debiasing Through a Causal Lens

Ashrya Agrawal
Birla Institute of Technology and Science
Pilani, India
ashryaagr@gmail.com

Florian Pfisterer
Ludwig-Maximilians-University
Münich, Germany
florian.pfisterer@stat.uni-muenchen.de

Jiahao Chen
Parity AI
New York, New York, USA
jiahao@parity.ai

Sebastian Vollmer
TU Kaiserslautern
Kaiserslautern, Germany
sebastian.vollmer@dfki.de

Abstract

Recent research has shown, that debiasing methods often do not reliably increase fairness in practical applications while simultaneously decreasing a model's accuracy. We study the effect of debiasing methods through a causal lens in order to develop a better understanding of factors determining whether a debiasing method works as intended. Simultaneously, a causal perspective on the phenomenon introduces the necessity to differentiate in debiasing methods, between discriminatory and non-discriminatory causal effects e.g. based on business necessity. Using this perspective on path specific effects, we study the effects of different debiasing methods on the underlying causal path specific effects (PSEs), observing empirically that reweighing reduces the direct effect of the protected attribute on the predicted label, while other PSEs are simultaneously increased. We provide an explanation for this phenomenon using an information theoretic approach. This perspective opens up the discussion for a need of incorporating causal perspectives into the development of debiasing methods in order to better capture the need for differentiating between discriminatory and non-discriminatory causal pathways.

1 Introduction

Machine Learning is being increasingly used in high-stake decision support systems like university admissions[Kung and Yu, 2020], criminal justice [Angwin et al., 2016, Berk et al., 2021], credit decisions [Byanjankar et al., 2015, Malekipirbazari and Aksakalli, 2015], etc. Multiple studies have established that these systems are unfair and exhibit a discriminatory behaviour against specific groups based on sensitive attributes like race and gender. In order to increase the fairness of machine learning systems, multiple debiasing algorithms[Pessach and Shmueli, 2020] have been proposed. But studies[Agrawal et al., 2020] have shown that this increase in fairness post debiasing is not statistically significant, and simultaneously decreases model's accuracy. We use a causal perspective to analyse the behaviour debiasing methods and draw deeper insights than what is provided by statistical analysis of fairness.

1.1 Causal Graphs and Notation

Causal Graphs [Pearl, 2009, Pearl et al., 2016] serve as a form for organizing assumptions about underlying data generating process and to represent decision-making process for ML models. We shall consider a directed acyclic causal graph, with nodes representing random variables corresponding to protected attribute Z, a set of non protected attributes, a predictor $\hat{Y}$ and sometimes observed

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

ground truth Y. A directed causal path is a sequence of distinct nodes $V_1, V_2, ...V_k$ for $k \geq 2$, such that $V_i \in pa(V_{i+1}) \forall i \in \{1, ..., k-1\}$, where $pa(V_i)$ are the parents of V_i i.e. they causally influence V_i. Mathematically, the causal model can be presented as a set of equations

$$V_i = f_i(pa(V_i), N_i) \forall i \in \{1, ..., k\}$$

where N_i denotes independent noise variables.

1.2 Metrics for Causal Fairness

Recently, causal notions have garnered attention in fairness as a tool for developing better insights on the sources of discrimination. Multiple causal notions based on comparing counterfactuals[Kusner et al., 2017] have been proposed. One of these is path-specific effect(PSE)[Chiappa, 2019]. PSE allows consideration of the effect of individual causal paths on predictions, thereby providing a greater analysis of decision making process and the fairness of the model. We employ this metric in our study to understand the effect of individual causal paths (or a set of paths) on the fairness and accuracy of our model. These causal paths can be categorised into discriminatory and non-discriminatory based on whether the path passed through a resolving variable(business necessity) [Kilbertus et al., 2017]. A causal path $V_1 \rightarrow ... \rightarrow V_k$ is non-discriminatory if $\exists i \in \{1..k\}$ such that V_i is a business necessity for the specific fairness problem. If no such i exists, the path is considered discriminatory. This categorisation holds in both legal and practical considerations.

1.3 Information Theoretic Analysis of Discrimination

Dutta et al. [2020] provide a information theoretic quantification of fairness by employing partial information decomposition(PID). For the causal graph in figure 1, with Z as the protected attribute and $\hat{Y}$ as the predictor, the mutual information I(Z; (A, B)) about Z in random variables A and B, available to the predictor $\hat{Y}$, can be expressed using PID framework as:

$$I(Z; (A; B)) = Uni(Z : A \backslash B) + Uni(Z : B \backslash A) + Red(Z : (A; B)) + Syn(Z : (A; B))$$

Here, $Uni(Z : A \backslash B)$ denotes the unique information component about random variable Z present only in A and not in B. Similarly, $Uni(Z : B \backslash A)$ denotes the unique information component about Z present only in B and not in A. Red(Z : (A;B)) denotes the redundant information about Z, present in both A and B, and Syn(Z : (A;B)) denotes the synergistic information not present in either of A or B individually, but present jointly in (A;B). Note that the mutual information about protected attribute through variables considered as business necessity(critical variables) is deemed as non-discriminatory.

2 Experiments

We evaluate path-specific effects(PSEs) for three models: Logistic Classifier and its combination with two different debiasing methods. Reweighing(RW)[Kamiran and Calders, 2012] and Equalized Odds(EOds)[Hardt et al., 2016] are used as the two debiasing methods. While reweighing is a preprocessing method, equalized odds is a post-processing method. Reweighing pre-processes the dataset, thereby resulting in a change in the underlying data generating processing(DGP) itself. Equalized odds operates on the predictions from the ML model, hence does not change underlying DGP.

We use a synthetic causal dataset for Law School Admissions[Kusner et al., 2017]. The causal model underlying the dataset is shown in figure 2. The dataset has 5 feature variables, of which **R**ace is considered as the protected attribute for this problem. **L**SAT and **G**PA are the business necessity variables, and thus the causal paths($R \rightarrow L \rightarrow \hat{Y}$, $R \rightarrow G \rightarrow \hat{Y}$, $R \rightarrow S \rightarrow L \rightarrow \hat{Y}$, $R \rightarrow S \rightarrow G \rightarrow \hat{Y}$) originating in Race and blocked by LSAT or GPA are deemed as non-discriminatory. Other causal paths originating in Race, $R \rightarrow \hat{Y}$ and $R \rightarrow S \rightarrow \hat{Y}$, are considered discriminatory.

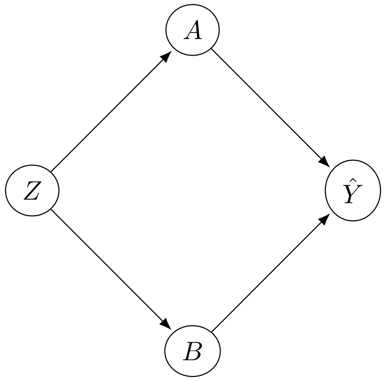


Figure 1: Causal graph for Information-decomposition illustration in 1.3

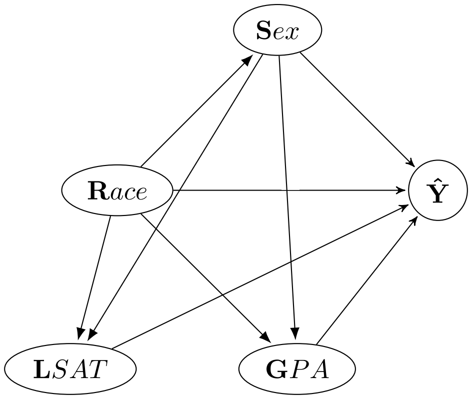


Figure 2: Causal Model for Law School Admissions used in experiments

In this study we use the notion of partial debiasing, which allows to interpolate between no debiasing and full-debiasing. Such interpolation enables a deeper analysis of the effect of debiasing on causal paths and decision making process. This interpolation is controlled by a parameter, which we refer in the text as α. The figures 3 and 4 have α on the X-axis, denoting the extent of debiasing.

We observe in figure 3 that the path specific effects(PSEs) show a different behaviour in case of reweighing and equalized odds. While the PSE decreases for the causal path $R \rightarrow \hat{Y}$ in both debiasers, the PSEs of other causal paths exhibit opposite trends. For the causal paths $R \rightarrow L \rightarrow \hat{Y}$, $R \rightarrow G \rightarrow \hat{Y}$, $R \rightarrow S \rightarrow L \rightarrow \hat{Y}$, $R \rightarrow S \rightarrow G \rightarrow \hat{Y}$ and $R \rightarrow S \rightarrow \hat{Y}$, the PSE for Equalized Odds decreases with α, while it increases for Reweighing. Thus, Reweighing increases the use of non-discriminatory paths while decreasing the use of the $R \rightarrow \hat{Y}$ path, while EOd on the other hand reduces all effects based on Race.

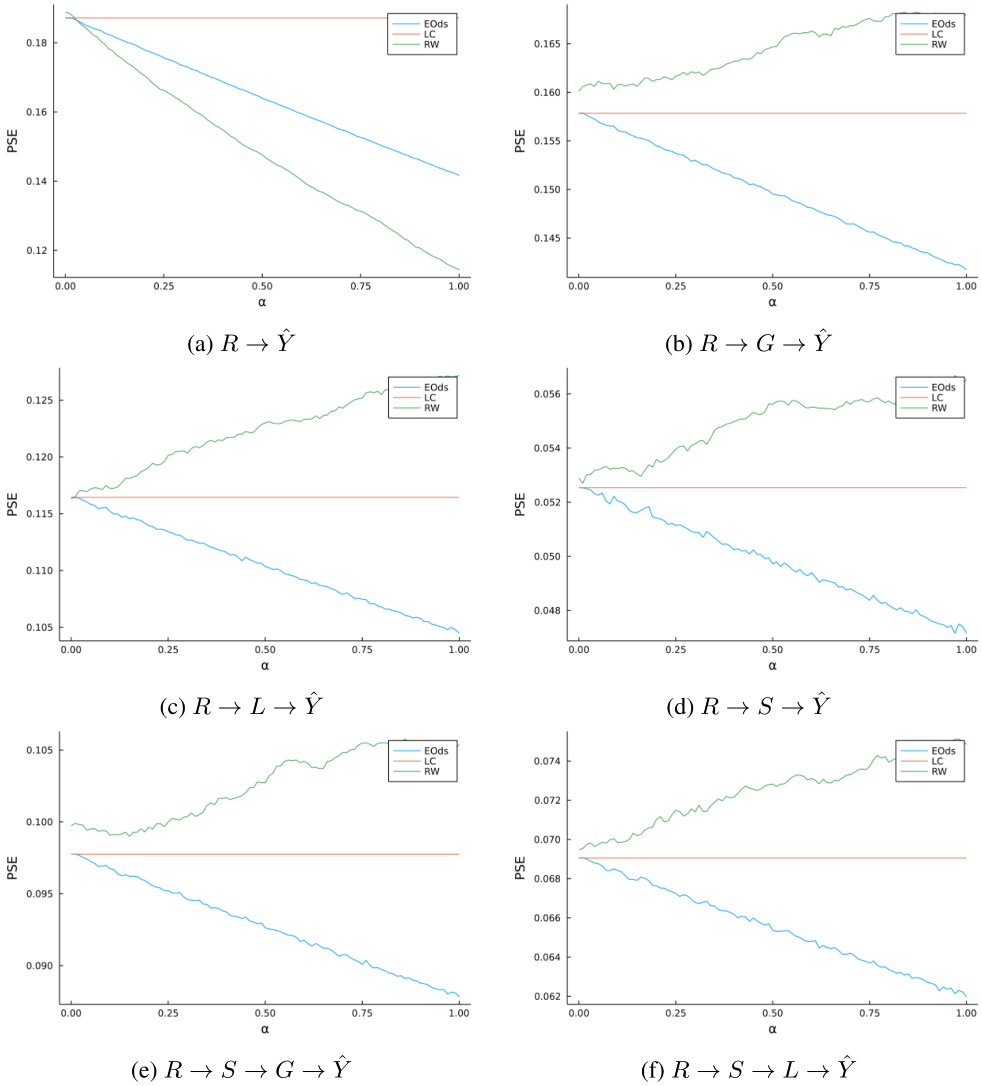
We also fit a logistic classifier on the predictions and obtain weights for the four features. These weights are the logistic regression coefficients of the fitted classifier. The weights are plotted for different values of partial debiasing parameter α in figure 4. Weights of the features reflect the pattern observed for PSEs in figure 3.

Note that this difference in trends of PSEs of both debiasers is not due to awareness or unawareness of discriminatory causal paths, but due to the choice of fairness criteria and the construction of debiaser, as explained in section 3. Equalized Odds reduces PSE of both, discriminatory paths, while reweighing reduces PSE of one discriminatory path ($R \rightarrow \hat{Y}$) and increases PSE of the other discriminatory path ($R \rightarrow S \rightarrow \hat{Y}$) and all non-discriminatory path. Even though Reweighing, unlike Equalized Odds rightly increases PSE of non-discriminatory paths, it shows different trends with the two discriminatory causal paths. This is because these trends in PSEs are due to the construction of debiasers, as stated previously. The debiasing algorithm(reweighing) is unaware of the set of discriminatory paths, thereby showing different trends for the two discriminatory paths. This highlights the need for development of debiasing methods which are aware of discriminatory and non-discriminatory paths.

Methodology We have carefully implemented the debiasing techniques and metrics in our Julia implementation, to verify that the effects observed are not a result of undiagnosed implementation issues. The classifier used in all experiments is a logistic classifier provided by the BSD-3 licensed ScikitLearn.jl [St-Jean, 2021] Julia package. We keep all the hyperparameters as default values to facilitate comparison across experiments and eliminating the variation due to different hyperparameter choices. The weights and PSEs are computed over 50 replications for $\alpha \in \{0.00, 0.01, ...1.00\}$ to eliminate the effect of noise and ascertain statistically significant treatment effect.

3 Information theoretic Perspective on Debiasing

Dutta et al. [2020] divide mutual information about race in $\hat{Y}$ into 4 mutually exclusive components:



(a) $R \rightarrow \hat{Y}$

(b) $R \rightarrow G \rightarrow \hat{Y}$

(c) $R \rightarrow L \rightarrow \hat{Y}$

(d) $R \rightarrow S \rightarrow \hat{Y}$

(e) $R \rightarrow S \rightarrow G \rightarrow \hat{Y}$

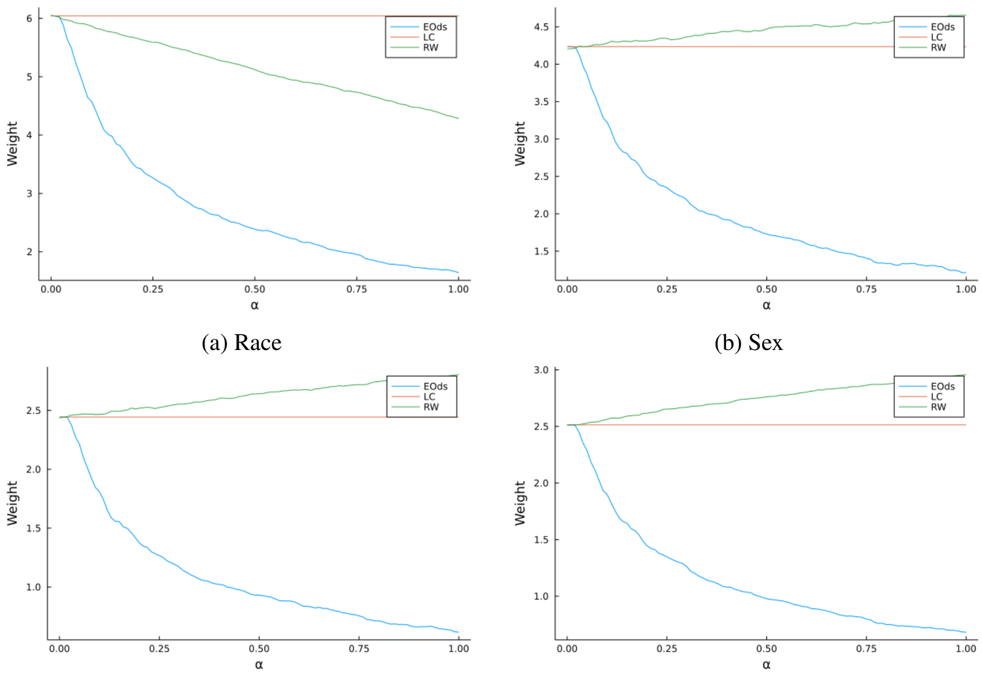
(f) $R \rightarrow S \rightarrow L \rightarrow \hat{Y}$

Figure 3: Path based Counterfactual effect for various causal paths and fairness algorithms

$$I(Race; \hat{Y}) = M_{V;NE} + M_{M;NE} + M_{V;E} + M_{M;E}$$

Here $M_{V;NE}$ is the visible non-exempt component, while $M_{M;NE}$ is the masked non-exempt component. The exempt component originates in non-discriminatory paths, while the non-exempt component originates in discriminatory paths and mainly from the direct $R \rightarrow \hat{Y}$ causal path.

Using the partial information decomposition framework stated in section 1.3, we can draw insights on the behaviour observed in the experiments. On applying reweighing, {Race, Y} stratification is done, thereby reducing correlation between Race and Y and reducing $M_{V;NE} + M_{M;NE}$. But this reduction does not affect $I(Race; \hat{Y})$ proportionately due to the presence of redundant and masked information, which was unused earlier due to direct availability of race variable. Post debiasing, redundant information Red(Race; (A, B)) provides the mutual information of race. Naturally, this redundant information comes from indirect (exempt and non-exempt) paths. This is why the path-specific effect for all the indirect paths (causal paths originating in race, excluding $R \rightarrow \hat{Y}$), as observed in figure 3. On the other hand Equalized Odds, by flipping predictions(with certain computed probabilities), results in a decrease in all mutual information from both race and indirect variables. Thus, EOds results in a decrease in exempt component along with non-exempt component, thereby resulting in a large decrease in accuracy. But reweighing increases exempt component, which partially compensates



(a) Race

(b) Sex

(c) GPA

(d) LSAT

Figure 4: Weights of different features for Logisitic Classifiers fitted on predictions across $\alpha \in \{0.00, 0.01, ..., 1.00\}$

for the decrease in non-exempt mutual information, thereby resulting in similar accuracy and higher fairness.

4 Discussion

In this work we study the effect of debiasing methods through a causal lens. We emperically show how debiasing methods, in particularly reweighing and equalized odds affect the causal pathways, which represent the decision making process of the ML model. The information theoretic perspective explains the reason for trends observed in PSEs. The causal analysis and the PID analysis also explain the behaviour of debiasing algorithms in observational settings. The extensive evaluation of debiasing algorithms by [Agrawal et al., 2020] shows that in most datasets, reweighing gives higher accuracy and a statistically significant increase in fairness compared to Equalized Odds and other post-processing methods. This behaviour is easily explained by the analysis in sections 2 and 3, which show that reweighing, in contrast to equalized odds, increases PSEs of non-discriminatory causal paths in decision making process, and also increases the exempt mutual information of race in $\hat{Y}$.

The causal perspective addresses the limitations of statistical approaches and provides deeper insights on effect of debiasing methods. We hope that these results are first steps for future research on incorporating causal perspectives into development of debiasing methods.

References

A. Agrawal, F. Pfisterer, B. Bischl, J. Chen, S. Sood, S. Shah, F. Buet-Golfouse, B. A. Mateen, and S. J. Vollmer. Debiasing classifiers: is reality at variance with expectation? *arXiv preprint arXiv:2011.02407*, 2020.

J. Angwin, J. Larson, S. Mattu, and L. Kichner. Machine bias, May 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods Research*, 50(1):3–44, 2021.

A. Byanjankar, M. Heikkila, and J. Mezei. Predicting credit risk in peer-to-peer lending: A neural network approach. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 719–725, 2015.

S. Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

S. Dutta, P. Venkatesh, P. Mardziel, A. Datta, and P. Grover. An information-theoretic quantification of discrimination with exempt features. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3825–3833, Apr. 2020. doi: 10.1609/aaai.v34i04.5794. URL `https://ojs.aaai.org/index.php/AAAI/article/view/5794`.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems*, volume 29, pages 3323–3331, 2016.

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*, volume 30, pages 656–666, 2017.

C. Kung and R. Yu. Interpretable models do not compromise accuracy or fairness in predicting college success. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, pages 413–416, 2020.

M. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, pages 4069–4079, 2017.

M. Malekipirbazari and V. Aksakalli. Risk assessment in social lending via random forests. *Expert Systems With Applications*, 42(10):4621–4631, 2015.

J. Pearl. *Causality: models, reasoning, and inference*, volume 64. 2009.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics : a primer*. 2016.

D. Pessach and E. Shmueli. Algorithmic fairness. *arXiv: Computers and Society*, 2020.

C. St-Jean. `ScikitLearn.jl`, v0.6.4, 2021. URL `https://github.com/cstjean/ScikitLearn.jl`.