Debiasing classifiers: is reality at variance with expectation?

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

Florian Pfisterer
Bernd Bischl
florian.pfisterer@stat.uni-muenchen.de
bernd.bischl@stat.uni-muenchen.de
Ludwig-Maximilians-University
Munich, Bavaria, Germany

Jiahao Chen
Srijan Sood
Sameena Shah
jiahao.chen@jpmorgan.com
srijan.sood@jpmorgan.com
sameena.shah@jpmorgan.com
J. P. Morgan AI Research
New York, New York, United States

Francois Buet-Golfouse
francois.buet-golfouse@jpmorgan.com
University College London
London, United Kingdom
J. P. Morgan
London, United Kingdom

Bilal A Mateen
Alan Turing Institute
London, United Kingdom
Wellcome Trust
London, United Kingdom
b.mateen@wellcome.org

Sebastian Vollmer
University of Warwick
Warwick, United Kingdom
Alan Turing Institute
London, United Kingdom
s.vollmer.4@warwick.ac.uk

ABSTRACT

Many methods for debiasing classifiers have been proposed, but their effectiveness in practice remains unclear. We evaluate the performance of pre-processing and post-processing debiasers for improving fairness in random forest classifiers trained on a suite of data sets. Specifically, we study how these debiasers generalize with respect to both out-of-sample test error for computing fairness–performance and fairness–fairness trade-offs, and on the change in other fairness metrics that were not explicitly optimised. Our results demonstrate that out-of-sample performance on fairness and performance can vary substantially and unexpectedly. Moreover, the variance in estimation arises from class imbalances with respect to both the outcome and the protected classes. Our results highlight the importance of evaluating out-of-sample performance in practical usage.

CCS CONCEPTS

• **General and reference** → **Empirical studies**; • **Computing methodologies** → *Supervised learning by classification*; *Cross-validation.*

KEYWORDS

bias mitigation,fairness,out-of-sample,test error,cross-validation

1 INTRODUCTION

Machine learning is increasingly used in important decision-making processes at scale[31, 37], ranging from: automating decisions on extending credit [13, 40]; to medical diagnostics [39]; and even informing bail judgements based on the risk of criminal recidivism [1, 7, 16]. As a result, unfairness and bias in these decision-making tools raise many legal, regulatory and ethical issues [4, 23, 43]. However, the causes of discriminatory bias can be subtle and complex, ranging from exclusionary biases in data sets [4, 11, 35] to classification criteria that reinforce historical and systemic biases [32]. The successful mitigation of these biases must necessarily take into account the origins of these biases [3, 30, 32].

The standard workflow for mitigating biases in a machine learning model, comprises two parts: first, the identification of a relevant fairness metrics [33, 41] and; second, the selection of an appropriate method by which to debias the model. However, both aspects are fraught with practical difficulties. Principally, the issue is that it is not always obvious which fairness definitions are relevant for a particular application [1, 7], despite the recommendations that are based on the resulting interventions [36], and that debiasing is not without cost. For example, a credit decisioning model has to be accurate in order to be profitable, which naturally leads to considerations of equality of opportunity [22]. At the same time, there are reputational and regulatory risks associated with bias in incorrect decisions, leading to considerations of equalized false negative rate and equalized false positive rate [22]. Simultaneous considerations of these fairnesses is unfortunately impossible without degradation in model performance due to well-known impossibility theorems [15, 27, 28, 34]. Whilst it might sound obvious to some, poorly performing models tends not to be used in practice, and therefore the accuracy penalty incurred by debiasing is an important consideration. Therefore, we have to consider not only the fairness–fairness trade-offs, but also the fairness–performance trade-offs to determine the optimally debiased model [15, 27]. These concerns have already been acknowledged explicitly in national regulatory guidance and thus there is a clear need to address this methodological gap in the literature [23].

In this paper, we consider the situation of supervised learning where membership of protected classes is fully known, and all relevant sociotechnical concerns have been identified from domain knowledge, resulting in the need for simultaneous consideration of multiple fairness metrics and performance metrics - a simplified abstraction of the aforementioned real world use cases. Specifically, we study the technicalities of remediation at a single point in time, ignoring time-evolving concerns [30]. In addition, we ignore other practical concerns such as the limited observability of protected class membership [14, 24]. Despite this highly restrictive setting, we find that the practicalities of debiasing are already sufficiently rich for in-depth study. Previous studies have found that debiasing methods are prone to overfit on the training set [21], focusing primarily on the metrics of disparate impact and Calders-Verwer / demographic parity [12]. They observe that the debiasing outcomes vary depending on the details of the train/test split, but did not explain why. Other studies like [10] have identified technical challenges resulting from the lack of native support for fairness or debiasing concerns in major machine learning software libraries, but do not consider variance or sensitivity issues. We build upon these observations, focusing on answering this question: *When using debiasing methods, how does generalization error affect how we consider fairness–performance and fairness–fairness trade-offs?*

Our contributions. First, we present in Section 3 some generalization of existing debiasing methods, introducing the notion of partial debiasing as well as our new NLinProg debiaser which generalizes the equalized offs debiaser. Second, we present in Section 4 an empirical study of how well debiasing algorithms perform in practice across multiple data sets and fairness measures, demonstrating that debiasing with respect to one measure of fairness often comes at the cost of overfitting, performance degradation, or worsened unfairness by other measures. Third, we present in Section 4.1 simple methods for partial debiasing and study performance–fairness trade-offs as well as fairness–fairness trade-offs. Finally, we state in Section 5.2 a result from learning theory that explains how the loss of generalizability is a consequence of the well-known bias–variance trade-off in machine learning, which gives us some insight into how to debias models more robustly. We have introduced related work throughout the exposition of this paper, in lieu of a dedicated section.

Notation. Define $S \in \mathcal{S} = \{0, 1\}$ to be a binary protected class, $X \in \mathcal{X}$ to be some set of features that explicitly excludes $\mathcal{S}$, so that $\mathcal{X} \cap \mathcal{S} = \emptyset$, $Y \in \mathcal{Y} = \{0, 1\}$ to be a binary outcome variable, $Z = (X, Y)$, $\hat{Y} \in \mathcal{Y}$ to be an estimator for Y, and $\hat{\mathcal{Y}}_f \subseteq \mathcal{Y}$ be the range of prediction for some classifier f, with the subscript f dropped when obvious from context. $S = s$ means that the variable S takes the value s, and similarly for X and Y. We specialize to the cases of binary $\mathcal{Y}$ and $\mathcal{S}$ for the ease of presentation; nevertheless, our results generalize to the cases with multiple classes. Additionally, define $\mathcal{D}$ to be in-sample (training) data, and $\mathcal{D}^\star$ to be out-of-sample (testing) data. Furthermore, let $\mathrm{id}_{\mathcal{A}} : \mathcal{A} \to \mathcal{A}$ be the identity function over the set $\mathcal{A}$, $f : \mathcal{X} \to \mathcal{Y}$ be a classification function, which we call an $\mathcal{S}$-*oblivious* classifier or oblivious classifier, $f' : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$ to be a $\mathcal{S}$-*aware* classification function, $g_{\text{pre}} : \mathcal{X} \times \mathcal{S} \to \mathcal{X}$ be a pre-processing debiasing function, $g_{\text{post}} : \mathcal{Y} \times \mathcal{S} \to \mathcal{Y}$ be a post-processing debiasing function, $G : (\mathcal{X} \times \mathcal{S} \to \mathcal{Y}) \to (\mathcal{X} \times \mathcal{S} \to \mathcal{Y})$ be an endomorphism over the space of $\mathcal{S}$-aware classification functions, and $\mathcal{H} \ni f$ be a family of classifiers . $\gamma_{f,\mathcal{D}}$ is the accuracy of the classifier f on the data set $\mathcal{D}$, and $\tau_{h,f,\mathcal{D}}$ is the fairness metric as defined in Equation (1) corresponding to the fairness definition h. When clear from context, the subscripts f and $\mathcal{D}$ will be dropped for brevity.

1.1 Fairness definitions

Many technical definitions of fairness exist and they have been reviewed elsewhere [7, 27, 33, 41]. We present only the definitions of fairness that we will study in this paper in Table 1.

1.2 Fairness metrics

In addition to choosing a suitable fairness definition, we also have to choose some loss function associated with that definition. For example, the Calders-Verwer gap [12] $\Delta_{\text{DP}} = \Pr(\hat{Y} = 1|S = 1) - \Pr(\hat{Y} = 1|S = 0)$ is simply the difference of the two sides of the equation that define demographic parity, and vanishes when perfect fairness exists. In addition to absolute differences, other metrics based on ratios, relative differences, or other more complicated losses have have been proposed. In this paper, we focus on symmetrized ratio-based metrics of the form

$$\tau_{\text{DP}} = \min\left(\frac{\Pr(\hat{Y} = 1|S = 1)}{\Pr(\hat{Y} = 1|S = 0)}, \frac{\Pr(\hat{Y} = 1|S = 0)}{\Pr(\hat{Y} = 1|S = 1)}\right), \quad (1)$$

and similarly for other group fairness measures. These τs by definition are constrained to lie within the unit interval $0 \le \tau \le 1$. The symmetry with respect to interchanging $S = 0$ and $S = 1$ removes the need to assume that there exists a protected class $S = 0$ that is generally privileged. Considering these τ ratios allows us to compare different metrics across different datasets.

1.3 Debiasing methods

Pre-, in- and post-processing methods for debiasing. Various debiasing algorithms are available in toolkits such as Aequitas [38] or IBM AI Fairness 360 [6]. These algorithms are traditionally classified as pre-processing, in-processing and post-processing methods, which are depicted at the functional level in Figure 1. In Figure 1, all primed quantities denote that they have been debiased. A *pre-processing debiaser* first transforms the input features $\mathcal{X}$ using some function g_{pre}, then feeds the transformed features as input to an oblivious classifier f. The debiased classifier is then the composition $f' = f \circ g_{\text{pre}}$. A *post-processing debiaser* takes the output of some oblivious classifier, $\hat{Y} \in \mathcal{Y}$, then transforms this output using some function g_{post}, The debiased classifier is then the composition $f' = g_{\text{post}} \circ (f \times \mathrm{id}_{\mathcal{S}})$, defined elementwise by $\hat{Y} = f(X)$, $f'(X, S) = g_{\text{post}}(\hat{Y}, S)$. Finally, an *in-processing debiaser* transforms some oblivious classifier f into an $\mathcal{S}$-aware but debiased classifier $f' = G(f)$. In general, the resulting classifier cannot be written as a function composition involving the original oblivious classifier f. Some in-processing debiasers like prejudice removal [26] further require that the debiased classifier f' be $\mathcal{S}$-oblivious, which is equivalent to the defining the debiased classifier $f'(X, S) = f(X)$ to be independent of S always. In other words, pre-processing debiasers transform the features $\mathcal{X}$, post-processing debiasers transform the predictions $\hat{\mathcal{Y}}$, and in-processing debiasers transform the classifiers f.

In this paper, we focus on pre- and post-processing debiasers, due to difficulties encountered in the execution of existing in-processing methods. Nevertheless, studying just these two types of debiasers

Fairness metric	Equality statement
Equalized false omission rate (EFOR) [7]	$\Pr(Y=1\|\hat{Y}=0,S=s)=\Pr(Y=1\|\hat{Y}=0)$
Predictive parity (PP) [15]	$\Pr(Y=1\|\hat{Y}=1,S=s)=\Pr(Y=1\|\hat{Y}=1)$
Demographic parity (DP) [12]	$\Pr(\hat{Y}=1\|S=s)=\Pr(\hat{Y}=1)$
Equalized false negative rate (EFNR) [15]	$\Pr(\hat{Y}=0\|Y=1,S=s)=\Pr(\hat{Y}=0\|Y=1)$
Predictive equality (PE) [15]	$\Pr(\hat{Y}=1\|Y=0,S=s)=\Pr(\hat{Y}=1\|Y=0)$
Equality of opportunity (EOp) [22]	$\Pr(\hat{Y}=1\|Y=1,S=s)=\Pr(\hat{Y}=1\|Y=1)$
Equalized odds (EOd) [22]	EOp and PE

Table 1: Group fairness definitions used in this paper.

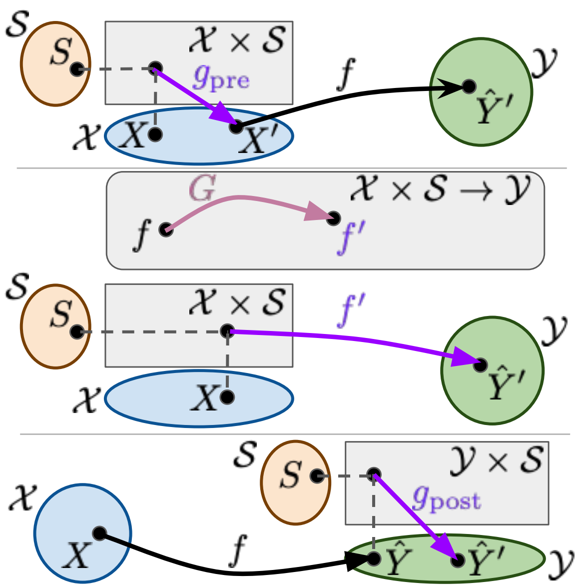

Figure 1: Overview of debiasing methods as described in Section 1.3. Top: pre-processing methods. Middle: in-processing methods. Bottom: post-processing methods.

suffices to illustrate the practical challenges of debiasing. We conclude our introduction with a description of a pre-processing debiaser and post-processing debiaser, using these as illustrations of the general class of debiasing methods.

1.4 Reweighing / resampling pre-processing

Reweighting / resampling as a preprocessing debiaser was introduced in Kamiran and Calders [25] in the context of demographic parity. They note that demographic parity is satisfied when Y and S are statistically independent, since

$$\Pr(\hat{Y}=y|S=s)=\frac{\Pr(\hat{Y}=y,S=s)}{\Pr(S=s)}=\Pr(\hat{Y}=y),$$

where the first equality is by definition of the conditional probability, and the second equality follows from the statistical independence assumption. The reweighing method simply proposes to assign each data point i a corresponding weight

$$w_{\mathrm{DP},i}=\frac{\Pr(\hat{Y}=y_i)\Pr(S=s_i)}{\Pr(\hat{Y}=y_i,S=s_i)}=\frac{\Pr(\hat{Y}=y_i)}{\Pr(\hat{Y}=y_i|S=s_i)}, \quad (2)$$

which effectively alters the measure associated with the sampled joint distribution of (Y,S) to match what would be expected from statistical independence. The resampling method simply resamples a new data distribution according to these weights w, to be used in classifiers that do not support arbitrary weights on data points.

1.5 Equalized odds post-processing

Equalized odds as a postprocessing debiaser [22, 34] consists of solving the following linear program involving equalized odds fairness (EOd) as a constraint, but making the substitution of variables $(Y,\hat{Y})\leftarrow(\hat{Y},\hat{Y}')$:

$$\begin{aligned}
&\min_{\hat{Y}'}\ \mathbb{E}\ell(\hat{Y}',\hat{Y})\\
&\text{s.t. } \Pr(\hat{Y}'=1|S=0,\hat{Y}=y)=\Pr(\hat{Y}'=1|S=0,\hat{Y}=y)\ \forall y\in\mathcal{Y}\\
&\qquad\sum_{s\in\mathcal{S}}\Pr(\hat{Y}'=1|S=s,\hat{Y}=y)=1\ \forall y\in\mathcal{Y}\\
&\text{and } \Pr(\hat{Y}'=1|S=s,\hat{Y}=y)\ \forall y\in\mathcal{Y},s\in\mathcal{S},
\end{aligned} \quad (3)$$

where the last two conditions enforce that $\hat{Y}'$ follows a valid probability distribution, and $\ell:\mathcal{Y}^2\to\mathbb{R}$ is an unspecified loss function in Hardt et al. [22] but is assumed to be effectively $1-\gamma$ in the code published with Pleiss et al. [34], where $\hat{Y}$ takes the place of the ground truth Y in the definition of accuracy. The solution to this optimization problem is the set of probabilities $\Pr(\hat{Y}'|\hat{Y}=y,S=s)$ for each $y\in\mathcal{Y}$ and $s\in\mathcal{S}$ that the labels $\hat{Y}$ should be flipped to yield the debiased labels $\hat{Y}'$.

2 GENERALIZED DEBIASERS

In this section, we present some generalizations of existing pre-processing and post-processing debiasers which we use in our studies.

2.1 Partially debiased reweighting

The reweighing pre-processing debiaser of Section 1.4 can be easily extended to some other definitions of group fairness, but not all. For example, considering $\hat{Y}\perp\!\!\!\perp S|Y=1$ instead of $\hat{Y}\perp\!\!\!\perp S$ gives an immediate generalization of reweighing for equality of opportunity instead of demographic parity. For equality of opportunity fairness (EOp),

$$\Pr(\hat{Y}=1|S=0,Y=1)=\Pr(\hat{Y}=1|S=0,Y=1),$$

the corresponding reweighting scheme would simply be

$$w_{\mathrm{EOp},i} = \frac{\Pr(\hat{Y} = y_i)}{\Pr(\hat{Y} = y_i | S = s_i, Y = 1)}. \quad (4)$$

However, there is no such reweighing scheme for equalized odds (EOd), which requires that both EOp and PE hold. Each equation demands its own reweighing scheme, with the first as in Equation (4) and the second as in

$$w_{\mathrm{ENOp},i} = \frac{\Pr(\hat{Y} = y_i)}{\Pr(\hat{Y} = y_i | S = s_i, Y = 0)}, \quad (5)$$

which will in general differ from the weights defined in Equations (2) and (4). Thus, reweighing as a method for exact debiasing works for neither composite fairness definitions that require multiple equality constraints, nor situations requiring multiple fairnesses to be satisfied simultaneously.

Given the impossibility of exact reweighing to work in general, we are led to consider the notion of *partial debiasing*. Noting that the simple weighting scheme $w_i = 1$ corresponds to no debiasing, we can consider reweighing schemes that interpolate between multiple reweighting strategies, such as the simple linear interpolation scheme:

$$w_i = \alpha_0 1 + \alpha_{\mathrm{EOp}} w_{\mathrm{EOp},i} + \alpha_{\mathrm{ENOp}} w_{\mathrm{ENOp},i}, \quad (6)$$

which under the constraints $\alpha_0 = 0$, $\alpha_{\mathrm{EOp}} + \alpha_{\mathrm{ENOp}} = 1$ defines a 2-simplex of reweighing strategies interpolating between pure EOp and pure ENOp, or alternatively under the constraints $\alpha_0 + \alpha_{\mathrm{EOp}} + \alpha_{\mathrm{ENOp}} = 1$ defines a 3-simplex of reweighing strategies interpolating between pure EOp, pure ENOp, and no debiasing.

In this paper, we consider only one of the simplest partial reweighing schemes,

$$w_i = (1 - \alpha)1 + \alpha w_{\mathrm{DP},i}, \quad (7)$$

as an exemplar of more general partial reweighing schemes that may interpolate among multiple constraints or in other spaces such as Wasserstein or spheres. As we show below, we find some surprising and nontrivial behaviors of this simple partial reweighing scheme, including empirical evidence that partial debiasing is often preferable to full debiasing ($\alpha = 1$) to maintain good generalizability of the debiased model.

For Equalized Odds and NLinProg described in 2.2, partial debiasing is done by interpolating between 0 and the flipping ratios $n2p_{S=0}, p2n_{S=0}, n2p_{S=1}, p2n_{S=1}$ obtained as a solution to the Optimisation Problem.

$$p2n_{\alpha,S=0} = \alpha * p2n_{S=0} \quad (8)$$

2.2 Nonlinear programs for post-processing (NLinProg)

In this subsection, we present a generalization of the equalized odds post-processing debiaser of Section 1.5 which we name NLinProg. NLinProg can handle arbitrary fairness definitions or combinations of fairnesses, at the cost of formulating a nonlinear constrained optimization problem as shown by Kim et al. [27]. We implement Algorithm 1 in the JuMP [19] framework for the Julia programming language [8], which uses Ipopt [42] for interior point optimization. The code is freely available on GitHub [1] .

input : predictions $\hat{Y}$, ground truth Y, protected classs S and debiasing measures F
output: modified predictions $\hat{Y}'$

$\hat{Y}'$ is constructed based on $\hat{Y}$
$n2p_{S=0} = \Pr(\hat{Y}' = 1 | S = 0, \hat{Y} = 0)$ is the probability to flip negative label for group $S = 0$ to positive and similarly $p2n_{S=0}, n2p_{S=1}, p2n_{S=1}$. All flips are independently distributed. The probabilities are chosen to solve optimizaton problem below.
P = OptimisationProblem($n2p_{S=0}, p2n_{S=0}, n2p_{S=1}, p2n_{S=1}$)
maximize($\mathbb{E}$ accuracy($\hat{Y}, \hat{Y}'$))
subject to fairness constraint that can also be expressed as quadratic constraint [27]

Algorithm 1: The NLinProg post-processing debiaser

3 GENERALIZABILITY OF DEBIASING METHODS

The preceding discussion of partially debiased reweighing in Section 4.1 already hints at a potential issue, namely that debiasing with respect to some fairness measures does not imply that the resulting classifier is also debiased with respect to other fairness measures. Such fairness–fairness trade-offs are implied in situations where impossibility theorems apply [15, 27, 28], and are further compounded by considerations of out-of-sample generalizability that is crucial to the practical utility of learned classifiers. In this section, we consider both notions of generalizability, in the senses of out-of-sample performance and also the transferability of debiasing across multiple fairness metrics.

3.1 Generalization error of fairness measures

Let $\mathcal{D} \sim \mathcal{P}(X, Y)$ be some data set consisting of $N = |\mathcal{D}|$ observations of (X, Y), and let $f : \mathcal{X} \to \mathcal{Y}$ be some learned classifier. Traditionally, the performance of f is measured using accuracy or similar measures from confusion matrices or receiver operating characteristic (ROC) curves [20]; however, we are usually more interested in the classifier's performance on unseen, out-of-sample data $\mathcal{D}^\star$, which requires us to estimate the generalization error using train-test splits, cross-validation [2], Rademacher complexity [5] or other methods. The variance of these methods usually depends on the evaluation procedure and the number of data points $|\mathcal{D}|$ and $|\mathcal{D}^\star|$.

Measuring fairness complicates the analysis of generalization error. The sample group has to be further subdivided by protected class membership S, which means that confusion matrices are no longer sufficient statistics for generalization error; instead, we must consider fairness-confusion tensors (FAcTs) [27] which consist of the stack of confusion matrices for each $S = s$ subgroup. This means that number of entries to be estimated grows by a factor of $|\mathcal{S}|$, which increases the estimation variance due to small sample counts in any one element of the FAcT. Taking once again the

[1] URL redacted for double-blind peer review.

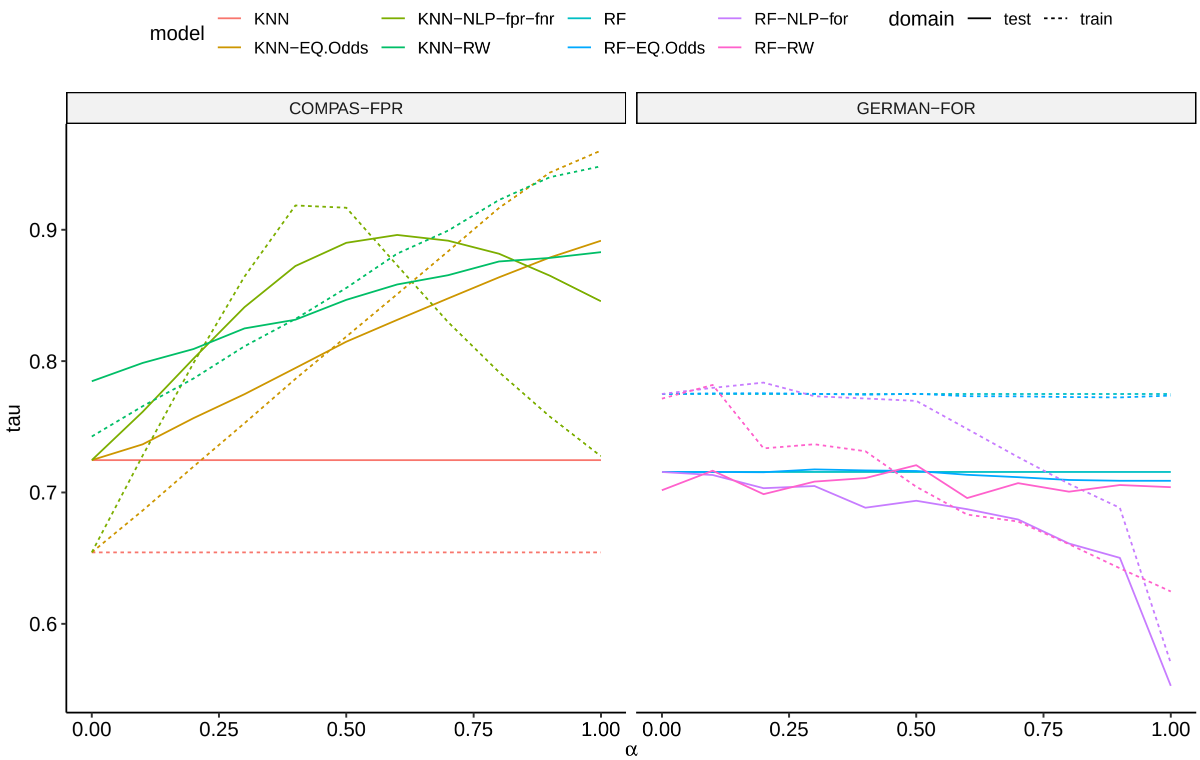

Figure 2: Examples where debiasing goes wrong - as α goes to 1 the fairness should improve (τ should get closer to 1) this is not the case

reweighing scheme in Section 1.4 as our prototypical debiasing method, we see that the joint probabilities $\Pr(\hat{Y} = y, S = s)$ in the denominator are sensitive to class imbalances not just in Y, but also in S, which further sensitizes the weights calculated and increases the uncertainty in their estimation. We show below that that similar considerations hold in general regardless of the debiaser used, which poses practical problems for the evaluation of debiasers.

3.1.1 Empirical evaluation. In this section, we specialize to classifiers f that are *random forest* models, which are either used without debiasing, or further processed using one of three debiasing strategies described in Sections 1.3 and 2.2 reweighing (RW), equalized odds (EO), or NLinProg (LP).

As a motivating example, we replicate the empirical study from Zafar et al. [44] on the COMPAS dataset [1] 100 times with different initial random seeds, reporting only results of a linear model trained with fairness constraints. Figure 3 shows the variance incurred from differing train-test splits and stochasticity in the optimization procedures. While this variance is comparably low for accuracy, the absolute difference $|\Delta FPR|$ and the relative difference τ in false positive rates between groups suffer from large variance in the estimation. In this study, $|\Delta FPR|$ ranges from 0.001 to 0.1, i.e. from almost no bias to a considerable bias.

This variance is problematic, as estimates of generalization error for fairness metrics become unreliable in practice, which e.g.

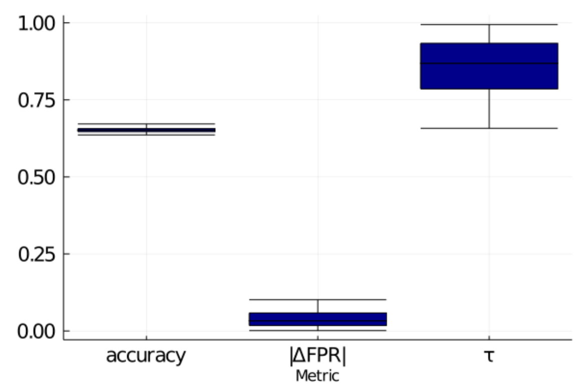

Figure 3: Results reporting accuracy and absolute as well as relative differences (τ) in FPR from 100 replications of bias mitigation on COMPAS with differing random seeds. The latter shows a large variation.

hinders the comparison of multiple de-biasing methods, as we will show below.

We conduct further empirical studies on the simulated datasets used in Zafar et al. [45] and the real-world Adult dataset [18] in order to show, that this holds for moderate dataset sizes, the additional variance does not depend on stochasticity during training,

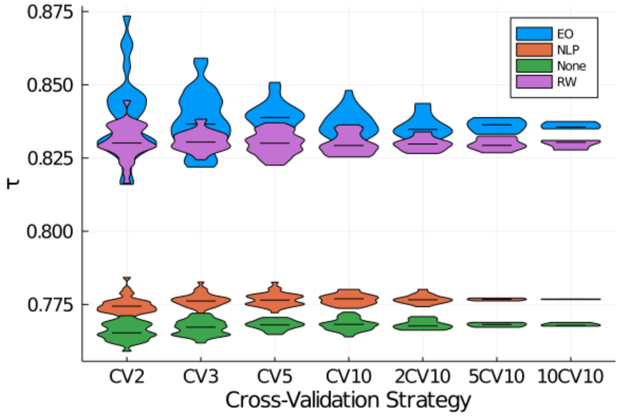

Figure 4: Estimated bias τ_{PP} across several de-biasing methods and cross-validation strategies on the Adult dataset. This demonstrates that a larger number of folds and replications are required to distinguish methods.

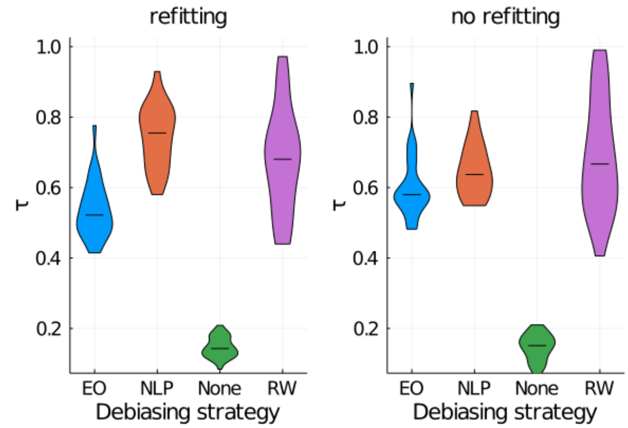

Figure 5: Estimated bias τ_{EFPR} estimated on multiple test-sets of size $n = 1000$ and several de-biasing strategies for data simulated as described in [44].

and that this variance partially stems from the de-biasing strategy. Figure 4 shows estimated τ for different re-sampling and de-biasing strategies measuring the rate of positive predictions (PPR) and *sex* as a protected class on the Adult dataset. Results depict replications of the full strategy averaged over the respective folds/replications. For fewer folds and non-replicated cross-validation, performance estimates of the different de-biasing strategies clearly overlap despite an actual difference (as evidenced by the medians). A clear separation is only possible for repeated 10-fold cross-validation. This effect can be observed despite the adult dataset containing 50.000 observations. This poses a problem in practice, when considering that results from e.g. tuning over different debiasing techniques using non-repeated (nested) CV [9] with a low amount of folds might be unreliable.

In an additional experiment shown in Figure 5 on synthetic data, we aim to separate the effect of variance incurred from different samples in the training data and differing test sets as well as the effect of stochasticity during the model fitting process by considering a synthetic data scenario. We first fit a model for each debiasing method on 10.000 datapoints simulated using the approach described in Zafar et al. [44]. Afterwards, we sample a test-set of size $n = 1000$ from the same joint distribution in order to estimate the generalization error of our fairness metric τ (using False Positive Rates). We repeat the same procedure 30 times with and without re-fitting the model in each replication in order to assess the additional variance incurred from the stochasticity of the fitting process. In summary, we can observe that: *i*) the use of debiasing strategies leads to increased variance in the estimate of τ, *ii*) the variance incurred from the fitting process is negligible in comparison to other factors and *iii*) that for debasing methods, variance again is larger than differences in medians, underlining the resulting problems for choosing an appropriate de-biasing strategy.

3.2 Optimizing for multiple fairness metrics

In various fairness problems, picking the debiasing metric can be hard and there might be no clear consensus on it. Generally, we use multiple fairness definitions for a dataset. For example, Corbett-Davies et al. [16] analysis of COMPAS recidivism algorithm uses 3 definitions while Berk et al. [7] discussion of fairness in COMPAS uses 6 definitions. While optimising a fairness definition, we might end up making other possible definitions of fairness worse.

Using the tuneable hyperparameter α described in Section 4.1, we plot the trade-off between fairness definitions in Figures 6 and 7. In each figure, we measure two different fairness definitions for increasing values of the partial debiasing parameter α across 3 ML methods: random forest (RF), k-nearest neighbours (KNN) and logistic regression (LR).

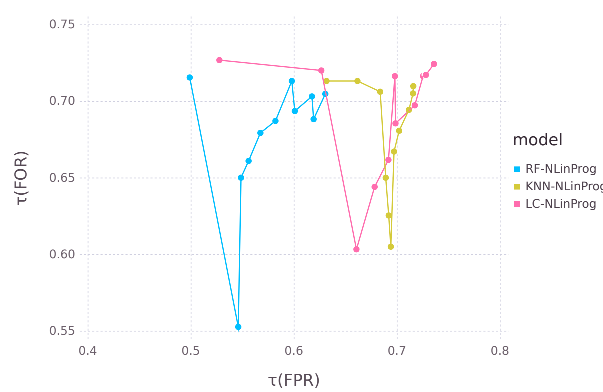

Figure 6: Fairness–fairness trade-off on German Dataset when NLinProg algorithm is used to debias with respect to EFOR fairness - across different ranges of α, see Equation (7).

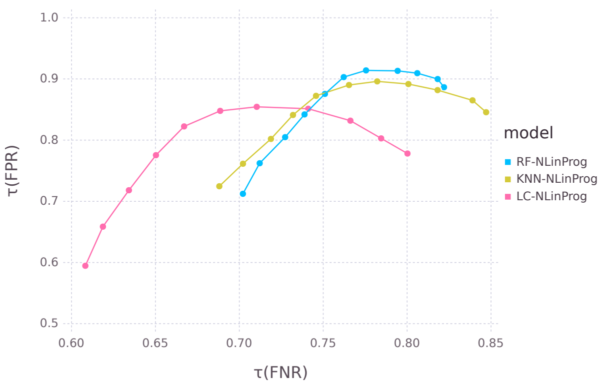

Figure 7: Fairness–fairness trade-off on COMPAS Dataset when NLinProg algorithm is used to debias both EFPR and EFNR fairness - across different ranges of α, see Equation (7).

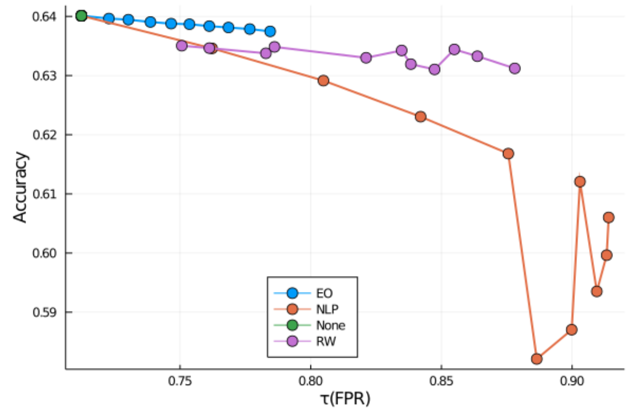

Figure 8: Fairness-Accuracy trade-off on the COMPAS dataset across several partial debiasing parameters α (see Equation (7)) for a random forest classifier.

This allows us to make several interesting observations: Some measures e.g. FOR and FPR in Figure 6, at least for larger values of τ need to be traded off against each other. Further improvement in one direction leads to decreasing performance for the other measure. The same results hold, when two fairness metrics are jointly optimized (c.f Figure Figure 7). Furthermore, for some settings (RF and KNN in Figure Figure 7), jointly optimizing two measures (FPR and FNR) without sacrificing one seems possible.

4 BENCHMARK STUDY

In order to provide additional insights into whether observations made in previous chapters hold across a broader variety of datasets, we conduct a large benchmark study across 9 datasets and 3 bias mitigation strategies. We evaluate accuracy and fairness metrics using 10 times repeated 10-fold Cross-Validation in order to reduce variance of our generalization error estimates and report τ as well as accuracy for each method and dataset. A list of data sets used throughout the experiment can be obtained from Table 2. We again resort to a *random forest* classifier as our base model combined with debiasing strategies. While we investigate partial de-biasing strategies (c.f. section section 2.1), we do not report results for partially debiased methods in more detail for brevity, and as it is unclear how exactly this hyperparameter would be tuned and selected in a multi-objective fashion. We instead report results for fixed $\alpha = 1$, i.e. full debiasing as defined by the respective methods..

4.1 Partial Debiasing

Methods considered in our benchmark exhibit a tunable hyperparameter α, allowing to trade-off emphasis on de-biasing against emphasis on accuracy. As the implication of varying this hyperparameter is rarely investigated in literature, we aim to provide additional insights into effects of this trade-off. Figure 8 shows the trade-offs available for the investigated methods by varying α between 0 and 1. As a control, a model without debiasing (None) is added, where the α has no effect and variations in accuracy and τ therefore are simply a result of stochasticity in the fitting process. Partial debiasing trades off emphasis on accuracy against emphasis on the desired fairness metric, allowing us to obtain a set of various trade-offs.

4.2 Results

The average accuracy γ as well as fairness τ for each datasets' associated fairness metric is reported in Table 3. Reported standard deviations (in brackets) correspond to the standard deviation of a 10-fold CV estimate as estimated via replications. It is important to consider the standard deviation of τ in comparison to the differences in mean, even considering computationally demanding CV10 resampling. In cases, where τ is already comparably large, the cost of increased fairness is relatively few points in accuracy. Additionally, we can observe some instances where debiasing actually leads to more unfair outcomes in comparison to a simple not-debiased model.

5 ANALYSIS OF ESTIMATOR VARIANCE FROM LEARNING THEORY

In this section, we provide some mathematical intuition behind some of the results that we have obtained. To simplify our approach, we consider the penalized (or dual) version of machine learning problems involving fairness constraints.

5.1 A simple trade-off between statistical performance and fairness

We write the optimization program: $\lambda \cdot$ statistical loss $+ (1 - \lambda) \cdot$ unfairness penalty as a trade-off between a loss function and a function accounting for unfairness. The trade-off is parameterized by λ, with $\lambda = 1$ being the unconstrained, performance-only case, and $\lambda = 0$ that accounts only for fairness.

We consider here that a model f from a family $\mathcal{H}$ has already been chosen (for instance by being fitted on a training set) and wish to study the speed of convergence of the *empirical* trade-off to the *true* trade-off. This is akin to measuring the convergence

Data set	Protected class	Fairness metric	Source
COMPAS	race	EFPR	[1]
Adult	sex	PP	[18, 29]
German Credit	marital_status	EFOR	[18]
Portuguese Bank Marketing	gender	EFOR	[18]
Communities and Crime	racepctblack	EFPR	[18]
Student Performance	sex	EFNR	[18]
Framingham Heart Study	male	EFOR	[17]
Loan Defaults	sex	EFOR	[18]
Medical Expenditure	race	EFOR	[6]

Table 2: List of data sets and associated fairness metrics used in our benchmarking study of Section 4.

Data set	γ_{None}	γ_{EO}	γ_{NLP}	γ_{RW}	τ_{None}	τ_{EO}	τ_{LP}	τ_{RW}
Adult	0.85(0.001)	0.83(0.002)	0.84(0.001)	0.84(9e-04)	0.77(0.003)	0.84(0.006)	0.78(0.003)	0.83(5e-03)
Communities & Crime	0.95(0.003)	0.74(0.051)	0.72(0.121)	0.94(2e-03)	0.53(0.050)	0.40(0.066)	0.48(0.131)	0.66(9e-02)
COMPAS	0.64(0.003)	0.64(0.004)	0.57(0.014)	0.63(3e-03)	0.82(0.013)	0.84(0.018)	0.88(0.033)	0.91(2e-02)
Framingham	0.84(0.003)	0.81(0.009)	0.57(0.041)	0.83(2e-03)	0.83(0.007)	0.82(0.008)	0.81(0.054)	0.81(1e-02)
German	0.73(0.010)	0.72(0.013)	0.50(0.022)	0.73(1e-02)	0.84(0.018)	0.83(0.025)	0.74(0.067)	0.83(3e-02)
Loan default	0.81(0.001)	0.80(0.002)	0.51(0.024)	0.81(1e-03)	0.94(0.006)	0.94(0.007)	0.79(0.067)	0.93(3e-03)
Medical expenditure	0.85(0.001)	0.80(0.005)	0.49(0.029)	0.84(1e-03)	0.70(0.006)	0.69(0.007)	0.66(0.025)	0.65(9e-03)
Portuguese	0.91(0.001)	0.91(0.001)	0.91(0.001)	0.91(9e-04)	0.99(0.001)	0.99(0.001)	0.99(0.001)	0.99(9e-04)
Student	0.62(0.017)	0.61(0.021)	0.55(0.026)	0.60(2e-02)	0.84(0.031)	0.84(0.031)	0.79(0.058)	0.84(4e-02)

Table 3: Summary of experimental results on random forest classifiers as described in Section 4. γ_h is the accuracy of the classifier after debiasing with method h, while τ_h is the fairness measure defined in Equation (1) of the classifier after debiasing with method h. Means and standard deviations (in parentheses) are reported for 10-fold cross-validation estimates via replications.

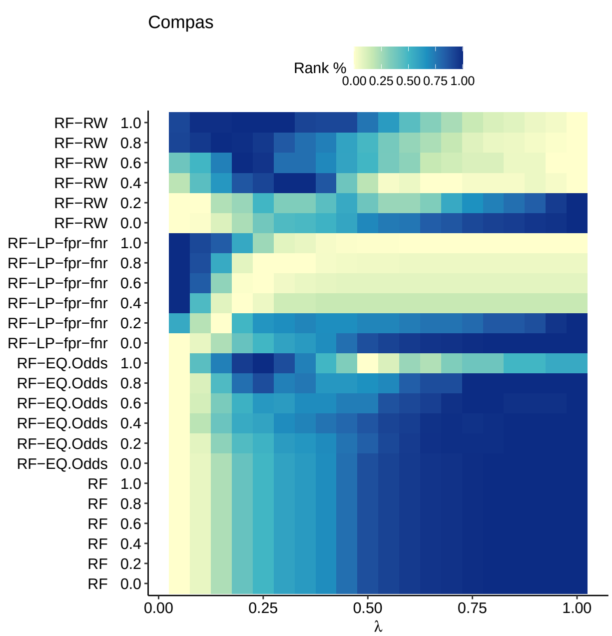

Figure 9: Rank of algorithms on generalisation error of a convex combination of accuracy and fairness: $\lambda \cdot$ accuracy + $(1-\lambda)\cdot\tau$. 0 corresponds to the best performance per λ while 1 corresponds to the worst.

of performance on the test set $\mathcal{D}^\star$ to its true value. By true trade-off, we mean the value of the penalized loss function under the true underlying distribution $(Z, S) \sim \mathcal{P}$ where Z corresponds to the vector of features and outcome (i.e., $Z = (\mathrm{x}, y)$) and S, as mentioned, is the protected class:

$$L_{\mathcal{P}}(f) = \lambda \mathbb{E}_{(Z,S)\sim\mathcal{P}}[\ell(f, Z, S)] + \\ (1-\lambda)\phi\left(\mathbb{E}_{(Z|S=0)}[\mu(f, Z, 0)], \mathbb{E}_{(Z|S=1)}[\mu(f, Z, 1)]\right).$$

Here, ℓ is a statistical loss function and ϕ is a non-negative fairness loss function such that $\phi(x, x) = 0$, i.e., it vanishes when perfect fairness is attained. μ is a function we can choose, such as the mis-classification error, $\mu(f, Z) = 1_{\{\widehat{y}\neq y\}}$. This can be succinctly rewritten as $L_{\mathcal{P}}(f) = \lambda\mathcal{L}_{\mathcal{P}}(f) + (1-\lambda)\phi\left(M_{\mathcal{P},0}(f), M_{\mathcal{P},1}(f)\right)$ with obvious notations. It is worth pointing out that the statistical loss is the average of both groups' statistical losses:

$$\mathcal{L}_{\mathcal{P}}(f) = \Pr[S=0]\mathbb{E}_{(Z|S=0)}[\ell(f, Z, 0)] \\ + (1-\Pr[S=0])\mathbb{E}_{(Z|S=1)}[\ell(f, Z, 1)].$$

5.2 Asymptotic convergence of the test trade-off to the true trade-off

As is customary, we -in general- only have access to empirical risk and thus to the sample version of the above, namely

$$
\begin{aligned}
L_{\mathcal{D}^\star}(f) &= \frac{\lambda}{m} \sum_{j \in S_0 \cup S_1} \ell(f, z_j, s_j) \\
&\quad + (1-\lambda)\phi\left(\frac{1}{m_0}\sum_{j\in S_0}\mu(f, z_j, 0), \frac{1}{m_1}\sum_{j \in S_1}\mu(f, z_j, 1)\right) \\
&= \lambda \mathcal{L}_{\mathcal{D}^\star}(f) + (1-\lambda)\phi\left(M_{\mathcal{D}^\star,0}(f), M_{\mathcal{D}^\star,1}(f)\right),
\end{aligned}
$$

where m_0 and m_1 are the sample sizes of group 0 and 1 respectively, and $m = m_0 + m_1$. Here, and in what follows, we consider that

Our objective is to derive the limiting distribution of $L_{\mathcal{D}^\star}(f)$ for a particular $f \in \mathcal{H}$. We suppose that we have drawn m samples from the overall population, each being independently and identically distributed according to the underlying distribution $\mathcal{P}$.

Proposition 1. *Assume that all observations are independent and identically distributed, $(z_j, s_j) \sim \mathcal{P}, \forall j \in \{1, ..., m\}$, the variance of $\ell(f, Z, S)$ is finite, the fairness penalty function ϕ is at least once-differentiable, and the variance of $\mu(f, Z, S)$ is finite. Then, the following asymptotic result holds:*

$$
\sqrt{m}\left[L_{\mathcal{D}^\star}(f) - L_{\mathcal{P}}(f)\right] \xrightarrow[m \to +\infty]{} N\left(0, \mathbb{V}_{\lim}(f)\right), \tag{9}
$$

with

$$
\begin{aligned}
\mathbb{V}_{\lim}(f) &= \pi_0 \lambda^2 [\sigma_0^\ell]^2 + (1-\pi_0)\lambda^2[\sigma_1^\ell]^2 \\
&\quad + \pi_0(1-\pi_0)\lambda^2\left[L_{\mathcal{P},0}(f) - L_{\mathcal{P},1}(f)\right]^2 \\
&\quad + (1-\lambda)^2 k_0^2 \frac{[\sigma_0^\mu]^2}{\pi_0} + (1-\lambda)^2 k_1^2 \frac{[\sigma_1^\mu]^2}{1-\pi_0} \\
&\quad + 2\lambda(1-\lambda)k_0\, cov\left[\ell(f, z, 0), \mu(f, z, 0)\right] \\
&\quad + 2\lambda(1-\lambda)k_1\, cov\left[\ell(f, z, 1), \mu(f, z, 1)\right],
\end{aligned}
$$

where we have denoted $\pi_0 = Pr[S = 0]$, $[\sigma_i^\ell]^2 = \mathbb{V}_{Z|S=i}[\ell(f, Z, i)]$ for $i = 0, 1$, $[\sigma_i^\mu]^2 = \mathbb{V}_{Z|S=i}[\mu(f, Z, i)]$ for $i = 0, 1$, and $(k_0, k_1)^T = \nabla\phi\left(L^1_{\mathcal{P},0}(f), L^1_{\mathcal{P},1}(f)\right)$.

This can be shown by repeated use of the central limit theorem, the delta method and Slutsky's lemma.

It is worth explaining the constitutive blocks of the limiting variance $\mathbb{V}_{\lim}(f)$ as they illustrate some of the challenges of measuring accuracy-fairness trade-offs. The first two lines illustrate the variance decomposition of the statistical loss in terms of a mixture model containing the two groups 0 and 1. In particular, the larger variance weight comes from the larger class of the two. The third line comes from the fairness requirement and shows that the limiting variance grows with $1-\lambda$ and $|k_{0,1}|$, which is intuitive: more sensitivity to fairness constraints leads to a larger proportion of variance. The most interesting feature however is the fact that these terms are *inversely* proportional to class probabilities. In other words, large class imbalance may lead to increased variance. The last two lines simply reflect the fact that both statistical and fairness components of the trade-off are based on the same sample.

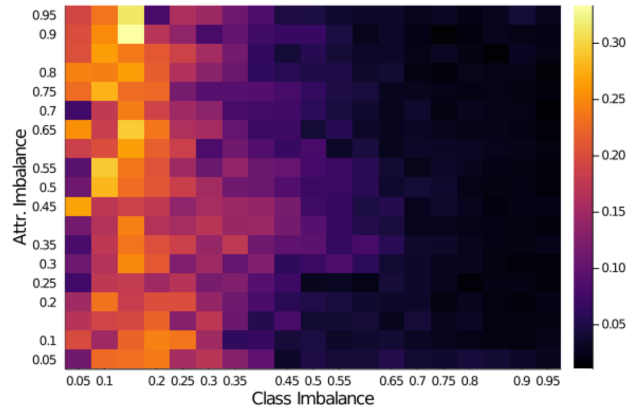

Figure 10: Standard deviations obtained for different fractions of target and protected class groups on simulated data from [44]. More class-imbalance corresponds to higher variance in the estimation of the fairness metric.

5.3 Empirical Considerations

We consider a scenario based on the synthetic data generating process described in [44] with varying fractions of samples in the respective classes (class imbalance) and in the respective protected class groups (attribute imbalance) in order to illustrate the conclusions made in section 5.2. We repeatedly sample 10.000 datapoints of the specified fraction for *train* and *test* sets and measure the standard deviation of the divergence in FPR across the different replications. Results can be obtained from Figure 10.

In this study, as a metric for the (un-)fairness of a classifier, we consider the disparity in false positive rates, again transformed through τ. Our results support the theoretical analysis above that the variance in debiasing is strongly affected by class imbalances in the data, both with respect to the class imbalance and the fraction of data points in the two protected attribute groups.

6 DISCUSSION

Group definitions. Fairness frameworks aim at learning rather complicated relationships, possibly historically inherited or representative of current socio-economic biases. It is worth pointing out that there is some subjectivity in choosing which categories are not to be discriminated against, as is exemplified by variations across different countries' legal frameworks.

Sample size effects. As shown in the expression of the limiting variance, one now needs to learn not just the global behaviour of a particular model, but its behaviour by subcategory. These subcategories have, by definition, lower sample sizes, implying that, all else being equal, we need either more data or simpler models.

Generalization. While we have tackled algorithmic fairness in this paper, it is worth emphasizing the fact that generalization is not purely statistical but also corresponds to the assumption that a distribution is the same. Debiasing algorithms may just have difficulties generalizing to different social contexts and/or definitions of protected classes.

Statistical loss function and parameters. In addition to considerations around sample size and the difficulty of learning fairness, we would not necessarily expect all algorithms to perform similarly, e.g., logistic regressions or support vector machines may be more amenable to fairness than other choices. This is still to be investigated and fully understood.

Fairness metric(s). One of the most important choices is the set of fairness metrics that is scrutinized when building or testing a model. This is not purely a statistical choice, but a more fundamental one. Certain fairness criteria may be more difficult to attain given the dataset or the statistical loss function.

Bias-mitigating algorithms. As exemplified in our analysis, bias-mitigating algorithms may have a fairly wide range of performances depending on the chosen methodology, fairness metric, sample size, etc. As shown above, this choice may lead to substantially different performance results, i.e. not all debiasing algorithms should be expected to perform well under all circumstances.

Trade-offs. Most of the time, trade-off values for different λs are computed and then "best" model is picked. This *peeking* in effect makes learning and generalization more difficult as it involves implicitly performing many model comparisons.

7 CONCLUSIONS

Our results above show that debiasing algorithms are slow to generalize and their performance is very context-dependent. Nevertheless, we have only studied the problem of *outcome fairness* in this paper, which has to be augmented with other aspects of fairness such as data fairness, design fairness and implementation fairness. Since our technical study indicates that currently available debiasing tools generalise poorly, these results highlight the need to work at a more holistic level with subject matter experts to evaluate carefully the appropriateness of classification models for their intended purposes.

ACKNOWLEDGMENTS

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content. We also thank our generous funding agencies IQVIA, UNIVERSITY of Auckland, Turing and tools practices, and Microsoft.

Disclaimer. This paper was prepared for informational purposes in part by the Artificial Intelligence Research group of JPMorgan Chase & Co and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kichner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Sylvain Arlot and Alain Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* 4 (2010), 40–79. https://doi.org/10.1214/09-SS054 arXiv:0907.4728

[3] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: Reorienting the Study of Algorithmic Fairness around Issues of Power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 167–176. https://doi.org/10.1145/3351095.3372859

[4] Solon Barocas and Andrew Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 1 (2016), 671–729. https://doi.org/10.15779/Z38BG31

[5] Peter L Bartlett and Shahar Mendelson. [n.d.]. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research* ([n. d.]), 463–482.

[6] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4:1–15.

[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* (Aug. 2018), 42. https://doi.org/10.1177/0049124118782533 arXiv:1703.09207

[8] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. 2017. Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.* 59, 1 (2017), 65–98. https://doi.org/10.1137/141000671

[9] Bernd Bischl, Olaf Mersmann, Heike Trautmann, and Claus Weihs. 2012. Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation* 20, 2 (2012), 249–275.

[10] Sumon Biswas and Hridesh Rajan. 2020. Do the Machine Learning Models on a Crowd Sourced Platform Exhibit Bias? An Empirical Study on Model Fairness. (May 2020). arXiv:cs.LG/2005.12379

[11] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research* 81 (2018), 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[12] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[13] Jiahao Chen. 2018. Fair lending needs explainable models for responsible recommendation. In *Proceedings of the 2nd FATREC Workshop on Responsible Recommendation*. 4. arXiv:1809.04684

[14] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geofry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. 339–348. https://doi.org/10.1145/3287560.3287594 arXiv:1811.11154

[15] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (June 2017), 153–163. https://doi.org/10.1089/big.2016.0047 arXiv:1703.00056

[16] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[17] L Cupples, Qiong Yang, Serkalem Demissie, Donna L Copenhafer, and Daniel Levy. 2003. Description of the Framingham Heart Study Data for Genetic Analysis Workshop 13. *BMC genetics* 4 Suppl 1 (02 2003), S2. https://doi.org/10.1186/1471-2156-4-S1-S2

[18] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[19] Iain Dunning, Joey Huchette, and Miles Lubin. 2017. JuMP: A Modeling Language for Mathematical Optimization. *SIAM Rev.* 59, 2 (2017), 295–320. https://doi.org/10.1137/15M1020575

[20] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 8 (2006), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[21] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. ACM Press, New York, New York, USA, 329–338. https://doi.org/10.1145/3287560.3287589 arXiv:1802.04422

[22] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29 (Dec. 2016), 3323–3331. https://doi.org/10.5555/3157382.3157469 arXiv:1610.02413

[23] Information Commissioner's Office. 2020. *Guidance on the AI auditing framework: Draft guidance for consultation (v1.0)*. Technical Report. Information

Commissioner's Office, Wilmslow, UK. https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

[24] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2020. Assessing algorithmic fairness with unobserved protected class using data combination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 110. https://doi.org/10.1145/3351095.3373154 arXiv:1906.00285

[25] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[26] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. *Lecture Notes in Artificial Intelligence* 7524, PART 2 (2012), 35–50. https://doi.org/10.1007/978-3-642-33486-3_3

[27] Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. 2020. Model-Agnostic Characterization of Fairness Trade-offs. In *Proceedings of the International Conference on Machine Learning*, Vol. 37. 9339–9349.

[28] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (Leibniz International Proceedings in Informatics (LIPIcs))*, Christos H. Papadimitriou (Ed.), Vol. 67. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, Article 43, 23 pages. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[29] Ron Kohavi. 1996. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. In *Proceedings of the 2nd. International Conference on Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, Portland, Oregon, 202–207.

[30] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2019. Delayed impact of fair machine learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 6196–6200.

[31] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635

[32] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology* 405 (2020), 26. https://doi.org/10.1007/s13347-020-00405-8 arXiv:2007.04068

[33] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT* 18)*. New York, USA.

[34] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Vol. 30. Curran Associates, Red Hook, NY, 5680–5689. arXiv:1709.02012 https://papers.nips.cc/paper/7151-on-fairness-and-calibration

[35] Rashida Richardson, Jason M Schultz, and Kate Crawford. 2019. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing systems, and Justice. *New York University Law Review* 94, 2 (2019), 192–233. https://www.nyulawreview.org/online-features/dirty-data-bad-predictions-how-civil-rights-violations-impact-police-data-predictive-policing-systems-and-justice/

[36] Kit T Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. 2020. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 142–153.

[37] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x arXiv:1811.10154

[38] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit. arXiv:1811.05577

[39] Eric J. Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* 25, 1 (2019), 44–56. https://doi.org/10.1038/s41591-018-0300-7

[40] Matthew Turner and Michael McBurnett. 2019. Predictive models with explanatory concepts: a general framework for explaining machine learning credit risk models that simultaneously increases predictive power. In *Proceedings of the 15th Credit Scoring and Credit Control Conference*. 11. https://crc.business-school.ed.ac.uk/wp-content/uploads/sites/55/2019/07/C12-Predictive-Models-with-Explanatory-Concepts-McBurnett.pdf

[41] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Conference on Software Engineering*. ACM, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776

[42] Andreas Wächter and Lorenz T. Biegler. 2006. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming* 106, 1 (2006), 25–57. https://doi.org/10.1007/s10107-004-0559-y

[43] Alice Xiang and Inioluwa Deborah Raji. 2019. On the Legal Compatibility of Fairness Definitions. In *Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems*. 6. arXiv:1912.00761

[44] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Geneva, Switzerland, 1171–1180. https://doi.org/10.1145/3038912.3052660

[45] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. In *Advances in Neural Information Processing Systems*, I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (Eds.), Vol. 30. Curran Associates, Inc., Red Hook, NY, 229–239. arXiv:1707.00010 https://papers.nips.cc/paper/6627-from-parity-to-preference-based-notions-of-fairness-in-classification