Auditing and Achieving Intersectional Fairness in Classification Problems

Giulio Morina*
QuantumBlack, a McKinsey company
London, United Kingdom
fairness@quantumblack.com

Viktoriia Oliinyk
QuantumBlack, a McKinsey company
London, United Kingdom
viktoriia.oliinyk@quantumblack.com

Julian Waton
QuantumBlack, a McKinsey company
London, United Kingdom
julian.waton@quantumblack.com

Ines Marušić
QuantumBlack, a McKinsey company
London, United Kingdom
ines.marusic@quantumblack.com

Konstantinos Georgatzis
QuantumBlack, a McKinsey company
London, United Kingdom
konstantinos.georgatzis@quantumblack.com

ABSTRACT

Machine learning algorithms are extensively used to make increasingly more consequential decisions, so that achieving optimal predictive performance can no longer be the only focus. This paper explores intersectional fairness, that is fairness when intersections of multiple sensitive attributes – such as race, age, nationality, etc. – are considered. Previous research has mainly been focusing on fairness with respect to a single sensitive attribute, with intersectional fairness being comparatively less studied despite its critical importance for modern machine learning applications. We introduce intersectional fairness metrics by extending prior work, and provide different methodologies to audit discrimination in a given dataset or model outputs. Secondly, we develop novel post-processing techniques to mitigate any detected bias in a classification model. Our proposed methodology does not rely on any assumptions regarding the underlying model and aims at guaranteeing fairness while preserving good predictive performance. Finally, we give guidance on a practical implementation, showing how the proposed methods perform on a real-world dataset.

1 INTRODUCTION

Fairness is a growing topic in the field of machine learning as models are being built to determine life-changing events such as loan approvals and parole decisions. Thus, it is critical that these models do not discriminate against individuals on the basis of their race, gender or any other sensitive attribute, by learning to replicate and reinforce the biases inherent in society or indeed introduce new biases. Whilst much of the algorithmic fairness literature thus far has focused on fairness with respect to an individual sensitive attribute, in this work we consider fairness for an *intersection of sensitive attributes*. That is, we aim to ensure fairness against groups defined by multiple sensitive attributes, for example, "black women" instead of just "black people" or "women".

Ensuring *intersectional fairness* is critical for safe deployment of modern machine learning systems. A stark example of intersectional bias in deployed systems was discovered by Buolamwini and Gebru [4] who showed that several commercially available gender classification systems from facial image data had substantial intersectional accuracy disparities when considering gender and race (represented via Fitzpatrick skin type), with darker-skinned women being the most misclassified group – having an accuracy drop of over 30% compared to lighter skinned men. Buolamwini and Gebru [4] emphasize the need for investigating the *intersectional* error rates, noting that gender and skin type alone do not paint the full picture regarding the distribution of misclassifications.

One cause for bias is the data itself. A known issue in many consequential application domains is that the recorded data often does not appropriately reflect the full diversity of the population. This disproportionate representation is further exacerbated for intersectional subgroups. Indeed, Buolamwini and Gebru [4] note that common facial analysis benchmarks are overwhelmingly composed of lighter-skinned subjects (79.6% for IJB-A and 86.2% for Adience). In their study on the increased risk of maternal death among ethnic minority women in the UK, Ameh and Van Den Broek [2] noted that there was limited data specifically for black and ethnic minority women born in the UK, and emphasized the need for reliable statistics to understand the scale of the problem.

Although of crucial importance, it is only recently that greater focus in the algorithmic fairness literature has been posed on intersectional fairness. Commonly used fairness metrics were developed with a single sensitive attribute in mind, and cannot be directly applied under this setting.

Our contribution. We present a comprehensive framework for assessing and achieving intersectional fairness, based on:

- Extending well-established fairness metrics to the case of intersectionalities, allowing for a thorough analysis of discrimination in both datasets and model outputs.
- Proposing novel ways to robustly estimate such metrics, even in the case where subgroups of the population are under-represented.
- Developing post-processing methodologies that can improve the intersectional fairness of any already available classification model.

Our work builds upon the concept of ϵ-differential fairness introduced by Foulds et al. [16] and extends it to several of the most widely used fairness metrics for auditing bias in datasets and model outputs. More specifically, we extend their definition of differential fairness for statistical parity to: 1) elift and impact ratio metric for data, and 2) equal opportunity and equalized odds metrics for model outputs. Our work aims to give practitioners the ability to assess

*Also with University of Warwick, Department of Statistics. Work done as an intern at QuantumBlack.

intersectional fairness through multiple, not mutually exclusive, lenses.

We are also interested in studying real-world scenarios, where certain intersectional subgroups are often disadvantaged due to societal or data collection bias, and therefore under-represented. This presents a challenge when trying to audit for intersectional fairness from a given finite dataset, where such biases are often found. We propose to move further from the smoothed empirical estimator proposed originally by Foulds et al. [16] and provide robust estimation procedures via bootstrap and fully Bayesian techniques. Importantly, we provide theoretical guarantees and demonstrate the performance of the estimators qualitatively and experimentally on a synthetic dataset.

Furthermore, we want to provide viable solutions to mitigate any detected discriminating bias: we develop post-processing methods for binary classification models that threshold risk scores and randomize predictions separately for each intersection of sensitive attributes, combining and extending the work of Hardt et al. [17] and Corbett-Davies et al. [10]. Our methods maximize predictive performance whilst guaranteeing intersectional fairness. A key advantage of our formulation is that it allows the practitioner to focus on multiple fairness metrics at the same time, thus allowing to control for multiple facets of model bias simultaneously. We provide implementation details and demonstrate the utility of our methods experimentally on the Adult Income Prediction problem [12], achieving intersectional fairness when 2 and 3 sensitive attributes, respectively, are considered.

Paper structure. We discuss related work in Section 2. We extend common fairness metrics to accommodate for intersectionalities in Section 3, proving some of their theoretical properties in Section 3.1 and presenting methods for robustly estimating them in Section 3.2. In Section 4, we phrase post-processing as an optimization problem which aims to preserve good predictive performance while ensuring intersectional fairness; we introduce the formulations in Sections 4.2 and 4.3 for binary and score predictors, respectively. Practical implementation of post-processing methods is discussed in Section 5. We demonstrate the utility of our methods experimentally on a synthetic experiment and on the Adult dataset [12] in Section 6. In Section 7, we conclude and suggest future work. Proof of all the results stated in the paper are presented in the supplementary material.

2 RELATED WORK

There is no universally accepted "best" fairness definition, nor is there one that is considered suitable for all use cases and application domains. There exist more than 20 different fairness metrics [34, 42], and some have been shown to be mutually incompatible [29, 36]. Therefore, the appropriate fairness definition and a corresponding metric need to be selected depending on the application, context, and any regulatory or other requirements.

One can broadly divide the abundant fairness definitions into group and individual fairness. Group fairness splits the population into groups according to the sensitive attributes and aims to ensure similar treatment with respect to a fixed statistical measure; individual fairness seeks for individuals with similar features to be treated similarly regardless of their sensitive attributes. Our work focuses on group fairness metrics.

Assessing group fairness of a dataset or model output becomes much more challenging when considering potentially dozens of sensitive attributes [20]. The number of generated subgroups grows exponentially with the number of attributes considered, making it difficult to inspect every subgroup for fairness due to both computational as well as data sparsity issues. A first challenge is, therefore, to come up with fairness metrics (either by modifying the widely used metrics or developing new ones) that can accommodate a large number of intersectional subgroups [11, 19, 25]. Our work builds most directly upon the *ϵ-differential fairness* metric introduced by Foulds et al. [16], which we suggest to interpret as a generalization of statistical parity – a notion of fairness also found in certain legal requirements [14]. Such a metric satisfies important desiderata, overlooked by other multi-attribute metrics [19, 25]. It 1) considers multiple sensitive attributes, 2) protects subgroups of the population as defined by their intersectionalities (e.g., "black women") as well as by individual sensitive attributes (e.g., "women"), 3) safeguards minority groups, and 4) aims at rectifying systematic differences between groups. ϵ-differential fairness also satisfies other important properties, such as providing privacy, economical, and generalization guarantees. While in Foulds et al. [16] the focus was mainly to enable a more subtle understanding of unfairness than with a single sensitive attribute, our work presents multiple metrics that allow more nuanced analysis of intersectional discrimination.

Nevertheless, all the introduced metrics pose algorithmic challenges when auditing for intersectional bias is of interest. While Foulds et al. [16] proposes a pointwise estimate, we have found that it can be unstable in the case of sparse data; we illustrate this in Example 6.1. Several other methods have been proposed in the literature for handling intersectional fairness that either make use of ad-hoc algorithms or are based on visual analytic tools [6, 28]. For intersectional bias detection, Chung et al. [9] suggest a top-down method to find underperforming subgroups. The dataset is divided into more granular groups by considering more features until a subgroup with statistically significant loss is found. In contrast, Lakkaraju et al. [31] use approximate rule-based explanations to describe subgroup outcomes.

As well as detecting discriminatory bias, another line of research has been focusing on *achieving* "fairer" models. There are three possible points of intervention to mitigate unwanted bias in the machine learning pipeline: the training data, the learning algorithm, and the predicted outputs, which are associated with three classes of bias mitigation algorithms: pre-processing, in-processing, and post-processing.

Pre-processing methods a-priori transform the data to remove bias or extract representations that do not contain information related to sensitive attributes [7, 13, 22, 32, 35, 39]; *in-processing* methods modify the model construction mechanism to take fairness into account [8, 23, 37, 38, 40]; while *post-processing* methods transform the output of a black-box model in order to decrease discriminatory bias [10, 17, 21]. Kearns et al. [25, 26] propose and demonstrate the performance of an in-processing training algorithm which mitigates intersectional bias by imposing fairness constraints on the protected subgroups. Their work is a generalisation of the "oracle efficient" algorithm by Agarwal et al. [1] to the case of infinitely

many protected subgroups. Foulds et al. [16] also proposes an in-processing learning algorithm based on the construction of a "fair" neural network.

In contrast to this approach, we develop a novel post-processing methodology. Post-processing procedures received great attention in applications as they do not interfere with the training process and therefore are suitable for run-time environments. In addition, post-processing techniques are model agnostic and privacy preserving as they do not require access to the model or features other than sensitive attributes [22]. The work of Hardt et al. [17] aims to ensure equal opportunity for two subgroups of the population, as defined by a single binary sensitive attribute. They achieve this by carefully randomly flipping some of the predictions in order to mitigate discriminatory bias. Another approach is explored by Corbett-Davies et al. [10], where fairness is guaranteed by treating model predictions differently according to the subgroup individuals belong to. We combine and expand their approach to the case of intersectional fairness.

3 METRICS FOR INTERSECTIONAL FAIRNESS

In this section, we introduce fairness metrics that can handle intersections of multiple sensitive attributes. Such metrics can be applied to asses fairness in either the data or in model outputs. Robustly estimating them is non-trivial, especially when more sensitive attributes are considered, as some subgroups may be under-represented in the dataset. Indeed, minorities in the population may be even more severely under-represented in a dataset compared to their true representation in the general population, one cause of which is bias in the data collection practices. After defining the metrics in Section 3.1, in Section 3.2 we present three different approaches for robustly estimating intersectional fairness for impact ratio, but the same notions can be applied to other intersectional metrics.

Notation. Let p be the number of different sensitive attributes. We denote by $A_1, \dots, A_p$ disjoint sets of discrete-valued sensitive attributes; e.g., A_1 can represent gender, A_2 race, A_3 nationality and so forth. The space of the intersections is denoted as $A = A_1 \times \cdots \times A_p$. Therefore, a specific element $s_i \in A$ is a particular combination of attributes; e.g., s_i = (Woman, Black, Italian).

Suppose we have access to a finite dataset with n observations denoted as $\mathcal{D} = \{(x_i, y_i)\}_{i=1,\dots,n}$; x_i represents the individual's features – including their sensitive attributes – and $y_i \in \{0, 1\}$ a binary outcome. We interpret $y_i = 1$ as a "positive" outcome and "negative" otherwise, denoting by Y the random variable describing the true population's outcomes. Furthermore, we let S be a discrete random variable with support on A. For brevity, we denote its probability mass function by $\mu_{s_i} = \mathbb{P}(S = s_i)$; i.e., the probability that any individual has sensitive attributes $s_i \in A$. Analogously, we denote by $\mu_1 = \mathbb{P}(Y = 1)$ the probability that a given individual has positive outcome. Finally, we will also denote the probability that an individual with sensitive attributes s_i has positive outcome as $\mu_{1|s_i} = \mathbb{P}(Y = 1|S = s_i)$. We do not make explicit assumptions on the distribution of Y or S but we shall assume $\mu_{s_i} > 0, \mu_{1|s_i} > 0, \forall s_i \in A$.

Table 1: ϵ-differential fairness metrics on the data

Fairness metric	Intersectional definition
elift	$e^{-\epsilon} \leq \frac{\mathbb{P}(Y=1\|S=s_i)}{\mathbb{P}(Y=1)} \leq e^{\epsilon}, \forall s_i \in A$
impact ratio (slift)	$e^{-\epsilon} \leq \frac{\mathbb{P}(Y=1\|S=s_i)}{\mathbb{P}(Y=1\|S=s_j)} \leq e^{\epsilon}, \forall s_i, s_j \in A$

Table 2: ϵ-differential fairness metrics on the model

Fairness metric	Intersectional definition
statistical parity (demographic parity)	$e^{-\epsilon} \leq \frac{\mathbb{P}(\hat{Y}=1\|S=s_i)}{\mathbb{P}(\hat{Y}=1\|S=s_j)} \leq e^{\epsilon}, \forall s_i, s_j \in A$
TPR parity (equal opportunity)	$e^{-\epsilon} \leq \frac{\mathbb{P}(\hat{Y}=1\|Y=1,S=s_i)}{\mathbb{P}(\hat{Y}=1\|Y=1,S=s_j)} \leq e^{\epsilon}, \forall s_i, s_j \in A$
FPR parity	$e^{-\epsilon} \leq \frac{\mathbb{P}(\hat{Y}=1\|Y=0,S=s_i)}{\mathbb{P}(\hat{Y}=1\|Y=0,S=s_j)} \leq e^{\epsilon}, \forall s_i, s_j \in A$
equalized odds	If ϵ-differential fairness is satisfied for both TPR and FPR parity

When a classifier model is available, we denote by $\hat{y}_i \in \{0, 1\}$ the prediction for the i^{th} individual and by $\hat{Y}$ the corresponding random variable. Importantly, we do not make any assumptions on how the model has been constructed and regard it as a black-box.

3.1 Definitions of Metrics

We now introduce intersectional fairness metrics for data and model outputs. We build on the definition of differential fairness introduced by Foulds et al. [16]. The definitions we introduce in this paper can be seen as a relaxation of the widely-used ones, to account for the fact that the number of intersections of sensitive attributes grows exponentially. In Table 1 we define fairness metrics to asses bias in the data, while Table 2 defines metrics to asses bias in model outputs. With the exception of ϵ-differential fairness for statistical parity, intersectional fairness definitions for the other metrics (cf. Table 1 and 2) are, to the best of our knowledge, novel contributions of this paper. We prove some of their theoretical properties later in Theorem 3.1. Although we restrict our analysis to fairness metrics for binary outcomes, they can be easily extended to the categorical case by simply requiring them to hold for all possible outcomes.

We refer the reader to Foulds et al. [16] for an interpretation of ϵ-differential fairness in terms of differential privacy. We note that $\epsilon = 0$ corresponds to achieving *perfect fairness* with respect to a given metric. Moreover, ϵ-differential fairness allows us to compare bias between two different models. In particular, if we assume that two models achieve ϵ-differential fairness for ϵ_1 and ϵ_2 respectively, then the quantity $\exp(\epsilon_2 - \epsilon_1)$ can be interpreted as a multiplicative increase/decrease of one model's bias with respect to the other, a phenomenon known as bias amplification [41].

A key question is whether these intersectional fairness definitions guarantee fairness with respect to individual sensitive attributes or any arbitrary subset of them. In other words, we would like to prove that if ϵ-differential fairness is satisfied for

$A = A_1 \times \cdots \times A_p$, then it is also satisfied when only A_1 is considered, $A_1 \times A_2$ and any other possible combination. Theorem 3.1 proves that this is indeed the case and 2ϵ-differential fairness is guaranteed to hold. Notice that this is a lower bound and in practice fairness for the subgroups may be satisfied for smaller values than 2ϵ. This is certainly the case when the elift metric is considered, as we prove that ϵ-differential fairness holds for the same value of ϵ in all subgroups.

THEOREM 3.1. *Let $A' = A_{c_1} \times \cdots \times A_{c_k}$, where $c_i \in \{1, \ldots, p\}, k \leq p$. If ϵ-differential fairness is satisfied for any of the metrics in Tables 1 and 2 on the space of intersections A, then ϵ-differential fairness is also satisfied on the space A' for the same metric.*

3.2 Robust Estimation of Intersectional Fairness

We now tackle the problem of auditing discriminatory bias having only access to a finite dataset $\mathcal{D}$. In particular, we are interested in the case where some combinations of sensitive attributes may be under-represented in the data. This is often the case in real-world datasets, usually due to historical reasons or inherent bias. We first make clear what we mean by auditing for intersectional fairness. We then explore three different methodologies to achieve this: 1) *smoothed empirical estimation*, where fairness metrics are directly computed from the data, 2) *bootstrap estimation*, to measure uncertainty in the empirical estimates, and 3) *Bayesian modelling*, to provide credible intervals.

We measure discriminatory bias in the data by computing the minimum value of ϵ such that one or more of the differential fairness definitions proposed in Section 3.1 holds. For the sake of exposition, we shall focus on estimating ϵ for the impact ratio metric, but the same reasoning can be readily extended to the other metrics. We then consider the problem of computing, as per Table 1:

$$\epsilon_{IR} := \min_{\epsilon \geq 0} \left\{ e^{-\epsilon} \leq \frac{\mu_{1|s_i}}{\mu_{1|s_j}} \leq e^{\epsilon}, \forall s_i, s_j \in A \right\}. \tag{1}$$

In general, the practitioner will not only be interested in computing ϵ_{IR}, but also in checking which attributes s_i, s_j determine big values of the ratios $\frac{\mu_{1|s_i}}{\mu_{1|s_j}}$.

Computing ϵ_{IR} may appear straightforward: we could just calculate $\mu_{1|s_i}$ for all $s_i \in A$ and let $\epsilon_{IR} = \log\left(\max_{s_i, s_j \in A} \left\{ \frac{\mu_{1|s_i}}{\mu_{1|s_j}} \right\}\right)$. However, the values of $\mu_{1|s_i}$ are usually unknown and estimating them from the data for all the values of $s_i \in A$ can be challenging, as few instances of a particular combination of attributes s_i may be available in the dataset $\mathcal{D}$. Moreover, as previously mentioned, minority subgroups may be even more severely under-represented in the dataset compared to their true representation in the general population, making the problem even harder. Therefore, we now introduce three different methods to estimate $\mu_{1|s_i}$: empirically from the data, via a bootstrap procedure, and with a Bayesian approach.

3.2.1 Smoothed Empirical Estimation. A simple approach is to directly estimate $\mu_{1|s_i}$ from the data, as proposed by Foulds et al. [16]. In particular, we can set

$$\hat{\mu}_{1|s_i} = \frac{N_{1,s_i} + \alpha}{N_{s_i} + 2\beta}, \tag{2}$$

where N_{1,s_i} is the empirical count of occurrences of individuals with attributes s_i and positive outcome in the dataset $\mathcal{D}$, while N_{s_i} is the total number of individuals with attributes s_i. We introduce smoothing parameters α, β as N_{s_i} or N_{1,s_i} may be small, due to data sparsity. Note that Equation 2 represents the expected posterior value of a Beta-Binomial model with prior parameters α, β. The final estimate of ϵ can be obtained as:

$$\hat{\epsilon}_{IR} := \log\left(\max_{s_i, s_j \in A} \left\{ \frac{\hat{\mu}_{1|s_i}}{\hat{\mu}_{1|s_j}} \right\}\right) = \log\left(\frac{\max_{s_i \in A} \hat{\mu}_{1|s_i}}{\min_{s_j \in A} \hat{\mu}_{1|s_j}}\right).$$

This estimation procedure requires computing $\hat{\mu}_{1|s_i}$ for all possible combinations of attributes $s_i \in A$, leading to $O(|A|)$ computational complexity. In general, it can be hard to tune the parameters α and β properly. In particular, big values of either α or β will introduce additional bias, while small values of β will not solve the data sparsity problem. Therefore, this procedure is not robust; $\hat{\epsilon}_{IR}$ will generally be biased and no uncertainty quantification can be provided. Nevertheless we now prove in Proposition 3.2 the appealing property that, as the dataset size grows, the smoothed empirical estimator converges to the true value regardless of the chosen smoothing parameters. Although the result holds for $\alpha, \beta \in \mathbb{R}$, in practice one would choose them to be non-negative, and set them both to zero when no smoothing is desired.

PROPOSITION 3.2. *The smoothed empirical estimate of ϵ for any ϵ-differential fairness metric is consistent $\forall \alpha, \beta \in \mathbb{R}$.*

3.2.2 Bootstrap Estimation. We propose to resort to bootstrap estimation to provide confidence intervals for the estimate $\hat{\epsilon}_{IR}$. We generate B different datasets by taking with replacement n observations from the original dataset $\mathcal{D}$. For each bootstrap sample, we obtain an estimate $\hat{\epsilon}^{(b)}_{IR}, b = 1, \ldots, B$ as in Equation 2. The final estimate $\hat{\epsilon}_{IR}$ is obtained by averaging over the sample and empirical confidence intervals can be easily constructed. The computational complexity is $O(B|A|)$, but in practice we also observe a computational overhead due to the construction of the B datasets. Notice that some of the generated datasets may not contain instances of specific attributes $s_i \in A$, producing undefined values if the smoothing parameters α, β are set to zero. This observation motivates why bootstrap estimates are not consistent (cf. Proposition 3.2) unless also $B \to \infty$, as some of the combinations of sensitive attributes may not be represented in any of the bootstrapped datasets. This is usually not problematic in practice, provided that the fixed size of the bootstrapped datasets is big enough.

3.2.3 Bayesian Estimation. Motivated by the form of Equation 2, we propose a Bayesian approach by considering the likelihood $N_{1,s_i}|\mu_{1|s_i} \sim Binom(N_{s_i}, \mu_{1|s_i})$ and setting its conjugate prior $\mu_{1|s_i} \sim Beta(\alpha, \beta)$. The posterior is therefore tractable and given by

$$\mu_{1|s_i}|N_{1,s_i} \sim Beta(\alpha + N_{1,s_i}, \beta + N_{s_i} - N_{1,s_i}).$$

We resort to Monte Carlo simulation techniques to get an estimate of ϵ_{IR}. In particular, we simulate N values of $\mu_{1|s_i}$ from the posterior and use them to compute the estimate of ϵ_{IR} as in Equation 1, with a computational complexity of $O(N|A|)$. By considering the average of the so-constructed sample we can obtain the final estimate of ϵ_{IR}. Moreover, this procedure promptly provides credible intervals. Finally, we note that the simulated values of $\mu_{1|s_i}$

will always be greater than zero, so that we do not need to resort to any further smoothing. The prior parameters α, β can incorporate a practitioner's domain knowledge of the problem or can be set close to zero to suggest no prior information. It follows from Proposition 3.3 that this estimator is also consistent.

Proposition 3.3. *The Bayesian estimate of ϵ for any ϵ-differential fairness metric is consistent $\forall \alpha, \beta > 0$.*

4 POST-PROCESSING OF CLASSIFIER MODEL

Often, we have access to outputs of a classification model that has already been trained and calibrated, but we may not have any knowledge on how such predictions were made either because the model is hard to interpret or because we do not have access to the model itself. Therefore, we will always assume that we only have access to a black-box predictor. We will refer to it as a "binary predictor" if its outputs are either 0 or 1 (or any other binary labels) and as a "score predictor" if its outputs are in $[0, 1]$.

We showed in Section 3 how to asses intersectional fairness of model outputs via different metrics. A natural next question is how to mitigate any detected bias. We argue that when possible, the best way to ensure fairness is to collect more representative data and retrain the model. Nevertheless, it is commonly the case that only historical data — where conscious or unconscious bias is often present — is available, so that it is impossible to gather more information. Moreover, training a new classifier may be impractical due to cost and time constraints. This motivates the need to develop post-processing techniques that are model agnostic. Indeed, we shall make no assumptions on the model training mechanism, and only require access to its outputs and on the sensitive attributes.

We follow the approach taken by Hardt et al. [17] and aim to construct a *derived predictor* $\tilde{Y}$ that achieves better fairness with respect to one or more chosen metrics. In particular, we propose a class of derived predictors that can handle classifiers returning either binary predictions or scores. In the following, we will denote by either $\tilde{Y}$ or $Y^\star$ the post-processed predictions (the distinction between the two will be made clear in Section 4.3.1), while as usual $\hat{Y}$ will denote the given predictor outcomes. Section 4.1 discusses a general framework for the construction of derived predictors. We explore how to compute them for a binary and score predictor in Sections 4.2 and 4.3 respectively.

The main characteristic of a derived predictor is that its value depends only on the given prediction $\hat{Y}$ and on the individual's combination of sensitive attributes S. We formally define it as follows:

Definition 4.1 (Derived Predictor, [17]). A *derived predictor* $\tilde{Y}$ is a random variable whose distribution depends solely on a classifier predictions $\hat{Y}$ and a combination of sensitive attributes S.

Our aim is to construct a derived predictor that, by transforming predictions of a given classifier, achieves better fairness in terms of one or more ϵ-differential fairness metric(s). If the model only returns binary predictions $\hat{Y} \in \{0, 1\}$, we can resort to *randomization*, that is, randomly flipping some of the predictions. On the other hand, when the model returns scores, constructing a derived predictor becomes more challenging. We focus on a specific class of derived predictors:

Definition 4.2 (Randomized Thresholding Derived Predictor). Given a classifier returning predictions $\hat{Y} \in [0, 1]$, the Randomized Thresholding Derived Predictor (RTDP) $\tilde{Y}$ is a Bernoulli random variable such that

$$\mathbb{P}(\tilde{Y} = 1 | \hat{Y} = \hat{y}, S = s_i) = \tilde{p}_{1,s_i} \mathbb{I}(\hat{y} \geq \tau_{s_i}) + \tilde{p}_{0,s_i} \mathbb{I}(\hat{y} < \tau_{s_i}) \quad (3)$$

where $\mathbb{I}$ is the indicator function and $\tau_{s_i}, \tilde{p}_{1,s_i}, \tilde{p}_{0,s_i} \in [0, 1], \forall s_i \in A$, are unknown parameters.

We interpret Equation 3 as follows: given an individual with predicted score $\hat{y}$ and attributes s_i, we first construct a binary prediction by considering a threshold τ_{s_i} and then, with a specific probability, we accommodate the possibility to reverse it or keep it. In particular we can equivalently write:

$$\tilde{p}_{0,s_i} = \mathbb{P}(\tilde{Y} = 1 | \hat{Y} < \tau_{s_i}, S = s_i),$$
$$\tilde{p}_{1,s_i} = \mathbb{P}(\tilde{Y} = 1 | \hat{Y} \geq \tau_{s_i}, S = s_i),$$

so that $\tilde{p}_{0,s_i}$ is the probability of flipping what would have been a negative prediction, while $\tilde{p}_{1,s_i}$ is the probability of keeping a positive prediction.

Note that Definition 4.2 covers also the case where the model returns only binary predictions $\hat{Y} \in \{0, 1\}$ and we explore this case in more details in Section 4.2. In consequential applications, randomization may not be desired or cannot be employed, due to legal or other requirements. Definition 4.2 allows to construct a deterministic derived predictor by setting $\tilde{p}_{1,s_i} = 1$ and $\tilde{p}_{0,s_i} = 0, \forall s_i \in A$.

4.1 Formulation as an optimization problem

We construct the RTDP by solving an optimization problem. In order to asses performance of the post-processed model, we introduce a loss function $l(y, \tilde{y}) : \{0, 1\}^2 \to \mathbb{R}$ that given the true and the predicted outcomes, returns the cost of making such a prediction, following the approach of Hardt et al. [17]. Without loss of generality, we will assume $l(0, 0) = l(1, 1) = 0$, so that making correct predictions does not contribute to the loss. Indeed, if either a bonus or a penalty is desired for correct predictions, it can be incorporated by changing the values of $l(0, 1)$ and $l(1, 0)$. Therefore, by minimizing the expected loss function we preserve good predictive performance.

To control the discriminatory bias of the post-processed model, the user has to select a value of ϵ that they wish to achieve for one or more of the intersectional metrics (cf. Table 2). We consider two possible approaches to find the unknown parameters $\tau_{s_i}, \tilde{p}_{0,s_i}, \tilde{p}_{1,s_i}$: 1) minimizing the expected loss subject to the selected fairness metric(s) being satisfied for the chosen ϵ, or 2) adding a penalty term to the expected loss for values of the parameters that do not satisfy the required fairness constraint.

For instance, one established fairness guideline is the 80% rule for statistical parity [14]; corresponding to requiring ϵ-differential fairness for statistical parity to hold for $\epsilon \leq -\log(0.8)$ (cf. Theorem 3.1). Therefore, we require to have

$$\frac{\mathbb{P}(\tilde{Y} = 1 | S = s_i)}{\mathbb{P}(\tilde{Y} = 1 | S = s_j)} \leq \exp(-\log(0.8)), \forall s_i, s_j \in A.$$

We can either consider this as a constraint in the parameter space of the optimization problem or consider minimizing

$$\mathbb{E}[l(Y,\tilde{Y})] + t \cdot \mathbb{I}\left\{\exists s_i, s_j \in A : \frac{\mathbb{P}(\tilde{Y}=1|S=s_i)}{\mathbb{P}(\tilde{Y}=1|S=s_j)} > 0.8)\right\},$$

for t appropriately large. The two approaches are in principle equivalent, but their practical implementation may differ as different numerical optimization routines need to be used. Note that statistical parity is not the only fairness constraint that can be considered; for instance in Section 6 we will aim to achieve better equalized odds intersectional fairness.

We now show in Proposition 4.3 how to rewrite the expected loss as a weighted sum of the False Positive Rate $F\tilde{P}R = \mathbb{P}(\tilde{Y}=1|Y=0)$ and of the False Negative Rate $F\tilde{N}R = \mathbb{P}(\tilde{Y}=0|Y=1)$ of the post-processed model, where the weights depend on $\mu_1 := \mathbb{P}(Y=1)$.

Proposition 4.3. *Minimizing $\mathbb{E}[l(Y,\tilde{Y})]$ is equivalent to minimizing*

$$F\tilde{P}R\,(1-\mu_1)\,l(0,1) + F\tilde{N}R\,\mu_1\,l(1,0). \tag{4}$$

Another approach to construct the RTDP is by maximizing a utility function. For instance, Corbett-Davies et al. [10] consider the immediate utility function, defined as $\mathbb{E}[Y\tilde{Y} - c\tilde{Y}], c \in (0,1)$. This approach may be preferable, as it only requires tuning a constant c that can be interpreted as the cost of making a positive prediction. We prove in Proposition 4.4 that our optimization framework accommodates this approach.

Proposition 4.4. *Let the immediate utility function be*

$$\mathbb{E}[Y\tilde{Y} - c\tilde{Y}] \text{ for a constant } c \in (0,1).$$

Then, maximizing this function is equivalent to minimizing Equation 4 when setting $l(0,1) = c$ and $l(1,0) = 1-c$.

4.2 Post-processing of a Binary Predictor

In this section we consider having access to a classifier that returns binary predictions $\hat{Y} \in \{0,1\}$. In this case, as only binary predictions are available, we set $\tau_{s_i} = 1, \forall s_i \in A$ and tune the probabilities $\tilde{p}_{1,s_i}$ and $\tilde{p}_{0,s_i}$ to construct the derived predictor. To find the unknown parameters we minimize the expected loss subject to the required fairness constraint. Proposition 4.5 shows that this optimization problem can be efficiently solved via linear programming.

Proposition 4.5. *Assume setting $\tau_{s_i} = 1, \forall s_i \in A$ in Definition 4.2 and optimizing the variables $\tilde{p}_{1,s_i}, \tilde{p}_{0,s_i}$ such that the expected loss is minimized and any of the model output metrics (cf. Table 2) is below a user-defined threshold. Then, the optimization problem is a linear programming problem.*

We conclude that in the case of a binary predictor, a RTDP can be computed in polynomial time [24].

4.3 Post-processing of a Score Predictor

In this section we assume that we have access to model outputs in the form of scores $\hat{Y} \in [0,1]$, where high scores indicate high probability of a positive outcome. We emphasize that we do not need to know any further information on how these scores were computed, and can treat the underlying model as a black-box. To construct the RTDP we can optimize both the probabilities $\tilde{p}_{1,s_i}$, $\tilde{p}_{0,s_i}$ and the thresholds τ_{s_i} for all $s_i \in A$, corresponding to a total of $3|A|$ parameters to optimize. Although we don't observe overfitting in the experiments we run in Section 6, in other applications it may be necessary to use cross-validation or to add regularization terms to reduce the degrees of freedom (e.g., imposing $\tau_{s_i} = \tau_{s_j}$ for some $s_i, s_j \in A$). In consequential application it is often undesirable to have random predictions, therefore we explore in detail the "deterministic" scenario in Section 4.3.1. The case where both the thresholds and the probabilities are optimized is discussed in Section 4.3.2.

4.3.1 Deterministic post-processing. If no randomization is desired, we construct an RTDP fixing $\tilde{p}_{1,s_i} = 1$ and $\tilde{p}_{0,s_i} = 0, \forall s_i \in A$. This case is of particular interest as randomization may be undesirable in real-world applications, for instance when assessing judicial decisions [3]. Moreover, we carefully tune the thresholds τ_{s_i}, as they will drive the predictive performance of the post-processed model.

To explicitly distinguish this case, we denote the post-processed prediction as $Y^\star$ and define post-processed model performance metrics as follows:

Definition 4.6. Define the post-processed model performance metrics of the RTDP when no randomization is used as

$$FPR^\star_{s_i} = \mathbb{P}(\hat{Y} \geq \tau_{s_i}|Y=0, S=s_i), \;\; TNR^\star_{s_i} = 1 - FPR^\star_{s_i},$$
$$FNR^\star_{s_i} = \mathbb{P}(\hat{Y} < \tau_{s_i}|Y=1, S=s_i), \;\; TPR^\star_{s_i} = 1 - FNR^\star_{s_i}.$$

Note that although not explicitly stated, the metrics introduced in Definition 4.6 are functions of the thresholds τ_{s_i}.

Although intuitive, applying randomization on top of a well-performing model will in general worsen its performance. This is formally proved in Proposition 4.7 where we show that, under reasonable conditions, it is indeed counter-productive to apply randomization when achieving better fairness is not an objective.

Proposition 4.7. *Let us consider a predictor returning scores $\hat{Y} \in [0,1]$ and solving*

$$min_{\tau_{s_i}, \tilde{p}_{0,s_i}, \tilde{p}_{1,s_i}} \mathbb{E}[l(Y,\tilde{Y})],$$

where $\tilde{Y}$ is the RTDP of Definition 4.1. This is equivalent to setting $\tilde{p}_{1,s_i} = 1, \tilde{p}_{0,s_i} = 0, \forall s_i \in A$ and solving

$$\min_{\tau_{s_i}} \mathbb{E}[l(Y,Y^\star)],$$

if and only if

$$\frac{TNR^\star_{s_i}}{FNR^\star_{s_i}} > \frac{\mu_{1|s_i}}{1-\mu_{1|s_i}}\frac{l(1,0)}{l(0,1)}, \quad \frac{TPR^\star_{s_i}}{FPR^\star_{s_i}} > \frac{1-\mu_{1|s_i}}{\mu_{1|s_i}}\frac{l(0,1)}{l(1,0)}, \forall s_i \in A. \tag{5}$$

Notice that the assumption of Equation 5 requires that the given model performs sufficiently well across all the intersections of sensitive attributes $s_i \in A$. As an example, if $l(0,1) = l(1,0)$ and $\mu_{1|s_i} = \frac{1}{2}, \forall s_i \in A$, Equation 5 translates as requiring greater $TPR^\star$ and $TNR^\star$ than $FPR^\star$ and $FNR^\star$ respectively for all the subgroups.

4.3.2 Post-processing using randomization. We now focus on constructing an RTDP by finding both the optimal thresholds τ_{s_i} and probabilities $\tilde{p}_{1,s_i}, \tilde{p}_{0,s_i}$. We first consider a simple approach that we name "sequential post-processing". We first find optimal thresholds

τ_{s_i} when no fairness constraints are imposed. By applying such thresholds, we convert the scores $\hat{Y}$ to binary predictions, so that we can find optimal probabilities $\tilde{p}_{1,s_i}, \tilde{p}_{0,s_i}$ that achieve the desired fairness constraints. This procedure appears appealing as we can loosely interpret the thresholding as a way to maximize predictive performance, and the randomization as a method to achieve better fairness. However, although the final result may be acceptable for the case at hand, there is no theoretical guarantee that this procedure will return the actual optimum.

A different solution which we will refer to as "overall post-processing", is to solve the following optimization problem:

$$\min_{\tau_{s_i}} f(\tau_{s_i}), \quad \text{s.t. } \tau_{s_i} \in [0,1], \forall s_i \in A, \tag{6}$$

where $f(\tau_{s_i})$ is the optimal cost function value found by solving the optimization problem only in the variables $\tilde{p}_{1,s_i}, \tilde{p}_{0,s_i}$, for a fixed τ_{s_i} (cf. Section 4.2). Although this may seem as adding an extra layer of complexity, we note that values of $f(\tau_{s_i})$ can be efficiently computed via linear programming. In general, $f(\tau_{s_i})$ will not be a differentiable function, so that optimizing it will always pose a challenge.

5 IMPLEMENTATION

We first discuss practical implementation of post-processing techniques for a binary predictor. In this case, we showed in Proposition 4.5 that the RTDP can be obtained by solving a linear programming problem. In practice, we need to compute the unknown model metrics $\hat{FPR}, \hat{FNR}$. We propose to use the same techniques introduced in Section 3 to estimate them; i.e., directly from the data, via bootstrap estimation, or Bayesian modelling.

Practical implementation of computing an RTDP from a score predictor is more challenging. We first provide a way to evaluate the expected loss function for any arbitrary value of τ_{s_i} and of $\tilde{p}_{1,s_i}, \tilde{p}_{0,s_i}$. The expected loss function is given in Equation 4 and, by applying the law of total probability, can be rewritten as:

$$\sum_{s_i \in A} \mu_{s_i} \left[\tilde{FPR}_{s_i}(1-\mu_{1|s_i})\, l(0,1) + \tilde{FNR}_{s_i}\, \mu_{1|s_i}\, l(1,0) \right].$$

The unknown constant base rates μ_{s_i} and $\mu_{1|s_i}$ can be estimated from the data via any of the techniques introduced in Section 3. To compute the post-processed metrics $FNR^{\star}_{s_i}$ and $FPR^{\star}_{s_i}$ (cf. Definition 4.6), we first apply the thresholds to the scores available on a validation dataset. We then estimate the metrics by using either bootstrap or Bayesian techniques as in Section 3. The values of $\tilde{FNR}_{s_i}$ and $\tilde{FPR}_{s_i}$ can then be readily computed as:

$$\begin{aligned} \tilde{FPR}_{s_i} &= TNR^{\star}_{s_i}\, \tilde{p}_{0,s_i} + FPR^{\star}_{s_i}\, \tilde{p}_{1,s_i}, \\ \tilde{FNR}_{s_i} &= FNR^{\star}_{s_i}\, (1-\tilde{p}_{0,s_i}) + TPR^{\star}_{s_i}\, (1-\tilde{p}_{1,s_i}). \end{aligned}$$

Note that since $FNR^{\star}_{s_i}$ and $FPR^{\star}_{s_i}$ are estimated directly from the data, they will be piecewise constant functions of the threshold τ_{s_i}. Therefore, gradient-based optimization routines are unlikely to succeed when this approach is considered as the gradient of the objective function – if defined – will be zero at all points. Moreover, the optimum will not be unique.

To address this issue, we propose to smoothen the objective function by smoothing out the model performance metrics. In particular,

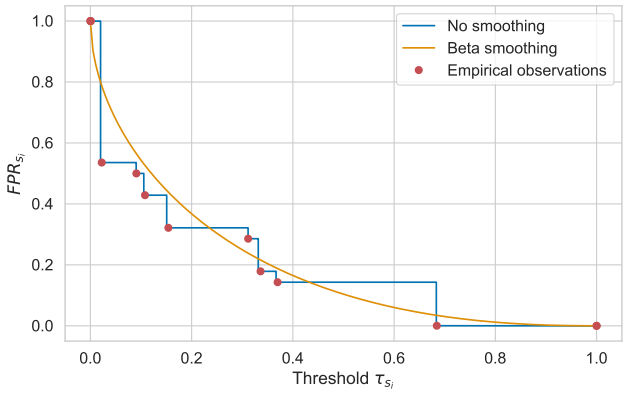

Figure 1: Illustrative example of FPR for individuals with combination of attributes s_i as a function of the threshold τ_{s_i}. Red points represent estimates of the FPR, the blue interpolating line depicts the non-differentiable step function, and the orange line is the smoothed curve obtained via a Beta modelling approach.

we propose a modelling approach by constructing the random variables $\hat{Y}_{0,s_i} = (\hat{Y}|Y=0, S=s_i)$ and $\hat{Y}_{1,s_i} = (\hat{Y}|Y=1, S=s_i)$. We model them both as Beta random variables and estimate their parameters by maximum likelihood estimation. We finally let for any arbitrary threshold τ_{s_i}, $FPR^{\star}_{s_i} = 1 - CDF_{\hat{Y}_{0,s_i}}(\tau_{s_i})$ and $FNR^{\star}_{s_i} = CDF_{\hat{Y}_{1,s_i}}(\tau_{s_i})$. This leads to a smooth function that preserves monotonicity as depicted in Figure 1.

6 EXPERIMENTS

We first propose an ad-hoc experiment using a generated dataset to compare ϵ-differential fairness estimation techniques of Section 3.2. We then apply them, together with the proposed post-processing methods, to the Adult's income prediction problem in Section 6.2.

6.1 Synthetic Experiment

We design this experiment to compare the estimation techniques of ϵ-differential fairness proposed in Section 3.1. We first discuss how we generated the synthetic datasets so that we have access to the true value of ϵ, thus allowing for a full comparison. We then discuss results and compare the different methodologies.

Dataset generation. We consider a set A_1, consisting of a binary sensitive attribute, and A_2, consisting of a different sensitive attribute with 3 possible values. Therefore, the space $A = A_1 \times A_2$ encompasses 6 different intersections. We fix true base rates as follows:

$$\begin{aligned} &\mu_{s_1} = 0.05, \quad \mu_{s_2} = 0.55, \quad \mu_{s_3} = \ldots = \mu_{s_6} = 0.1, \\ &\mu_{1|s_1} = 0.05, \quad \mu_{1|s_2} = 0.95, \quad \mu_{1|s_3} = \ldots = \mu_{1|s_6} = 0.5. \end{aligned} \tag{7}$$

As a result of this choice, the intersection of attributes s_1 is not going to be well represented in the dataset and, moreover, there will be few positive outcomes for individuals with such characteristics. Indeed, we purposely fixed base rates as above to mimic real-world scenarios where a particular subgroup can be under-represented,

either in the general population or in a particular dataset. The true value of ϵ_{IR} can be exactly computed as $\epsilon_{IR} = \log\left(\frac{0.95}{0.05}\right) \approx 2.94$.

Results. We first observe how the estimate behaves as the size of the dataset increases and we analyze the confidence intervals obtained via either bootstrap or Bayesian estimation. We fix $B = 1{,}000$ as the number of bootstrapped datasets, each of size equal to the original one, and smoothing parameters $\alpha = \beta = 0.01$ to avoid divisions by zero. When considering the Bayesian approach, we generate $N = 1{,}000$ Monte Carlo samples and consider a non-informative prior by setting $\alpha = \beta = \frac{1}{3}$, as proposed by Kerman [27].

Results are presented in Figure 2. As Propositions 3.2 and 3.3 prove, we observe that all the methods converge to the true value as the dataset size grows. Clearly, no confidence intervals can be obtained with the smoothed empirical estimator. On the other hand, we notice that for small values of the dataset size, the confidence intervals provided by the bootstrap method are generally wider than the ones obtained via a Bayesian approach. This is not surprising, as the estimate of ϵ_{IR} is particularly unstable if any instances with combination of attributes s_1 are not replicated in one of the bootstrapped datasets. Indeed, when we look at the distribution of $\hat{\epsilon}_{IR}^{(b)}$ for low values of n in these experiments, we observe one peak near the true value ϵ_{IR} and another peak determined by our choice of smoothing parameters.

To further assess properties of the proposed estimators, we approximate their Mean Squared Error (MSE). To do so, we generate 1,000 different datasets of increasing size with the same true base rates as in Equation 7. For each dataset, we obtain an estimate of ϵ_{IR} using the proposed estimation techniques. Results are presented in Figure 3. We notice that the estimate obtained via a Bayesian approach performs better for all considered dataset sizes. On the other hand, bootstrap performs slightly worse than empirical estimation for small dataset sizes. As mentioned above, this is due to the fact that when one attribute intersection, such as s_1 in our experiment, is poorly represented in the bootstrapped dataset, we obtain biased estimates of ϵ_{IR}.

We conclude that the smoothed empirical estimator requires considerably less computational effort than the other two proposed methods, however it does not provide any insight on how reliable the fairness metric estimate is. When this is desired, we suggest using either bootstrap estimation or Bayesian modelling, or possibly both. We finally observe that the Bayesian procedure is in general faster than the bootstrap one, as the posterior parameters need to be computed only once and no overhead is observed.

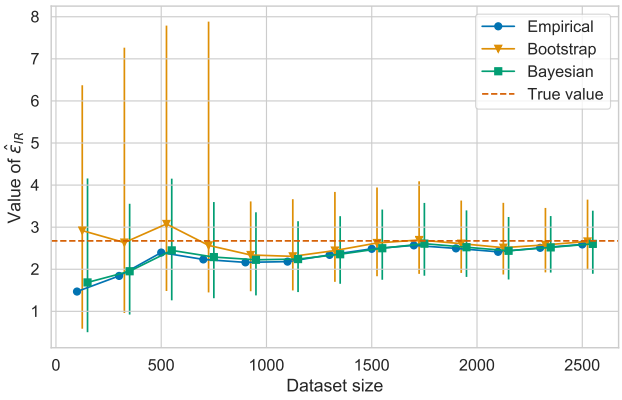

Figure 2: Comparison of different estimators of impact ratio fairness metric on synthetic datasets of increasing size. Vertical bars represent 95% confidence intervals for bootstrap and Bayesian estimation where 1,000 bootstrapped dataset and Monte Carlo sample have been drawn respectively.

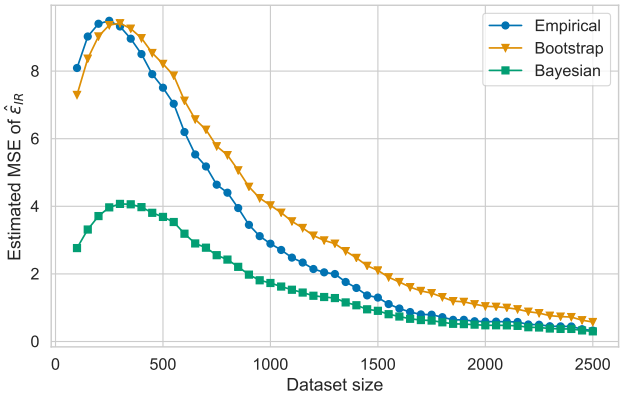

Figure 3: Comparison of Mean Squared Error of the estimator $\hat{\epsilon}_{IR}$ on synthetic datasets of increasing size. MSE of the estimators has been estimated by generating 1,000 different datasets with same base rates.

6.2 Adult Income Prediction

6.2.1 Dataset and Model. We consider a practical application of auditing and mitigating bias using the 1994 U.S. census Adult dataset from the UCI repository [12]. The aim is to predict whether an individual's income is greater than $50,000, using socio-demographic attributes. The data has already been split by the provider into a training set, consisting of 32,561 observations, and a test set, with 16,281 data points.

In the following we will focus on three sensitive attributes: age, gender, and race. We represent age as a binary categorical variable indicating which individuals are over 50. Gender is considered as a binary attribute in the Adult dataset. Race is encoded in the dataset into 5 different categories. For the purpose of this experiment, since the dataset contains few instances of categories "Eskimos and American Indians" and "Other", we encode them together under the label "Other".

We build a classifier returning scores in $[0, 1]$ and we apply a fixed threshold to obtain binary predictions, as is commonly done in practice. To allow for a simpler exposition, we first look into applying our proposed methodologies where only two binary sensitive attributes are considered. We then explore a more complex scenario with more than 2 sensitive attributes. Additional tables, figures and implementation details are reported in the supplementary material.

6.2.2 Two sensitive attributes. We treat age and gender as sensitive attributes. First, we look into auditing intersectional fairness on the

Table 3: Predictive performance of given binary predictor and post-processed models on the Adult training set with gender and age as sensitive attributes.

	No fairness constraints		*With fairness constraint* $\epsilon \leq 1.77 - \log(4) \approx 0.384$			
	Given binary predictor	Optimal score model	Randomization only	Deterministic	Sequential	Overall
TPR	0.5450	0.5481	0.5131	0.5789	0.5145	0.5699
FPR	0.0422	0.0427	0.0497	0.0591	0.0470	0.0551
Expected loss function	0.1416	0.1413	0.1550	0.1463	0.1526	0.1454

dataset and on the model outputs. We then compare performances of all the different post-processing techniques.

Auditing intersectional fairness

Figure 4 shows the minimum values of ϵ such that ϵ-differential fairness is satisfied for different intersectional metrics, both on the data and on the outputs of the binary classifier. We compare the methods introduced in Section 3, considering smoothing parameters $\alpha = \beta = 0.01$ to avoid division by zero and prior parameters for the Beta distribution both equal to $\frac{1}{3}$. We note that the three methodologies produce similar answers and that the model may be deemed unfair as the different intersections are subject to varying performance in correctly predicting negative outcomes (i.e., income less than \$50,000). It is of interest to check which subgroup of the population (as defined by the intersection of sensitive attributes) drives the value of ϵ for FPR parity. Further inspection reveals that the model is $\exp(1.77) \approx 6$ times better at correctly predicting negative outcomes for men of age > 50 than women of age ≤ 50.

Post-processing for intersectional fairness

We now aim to mitigate this detected discriminatory bias. We first consider the scenario where we only have access to binary predictions. As we assume no further knowledge of the underlying model, the only possible choice is to use randomization as a post-processing technique (cf. Section 4.2). We set a loss function that gives equal weights to false positive and false negative predictions;

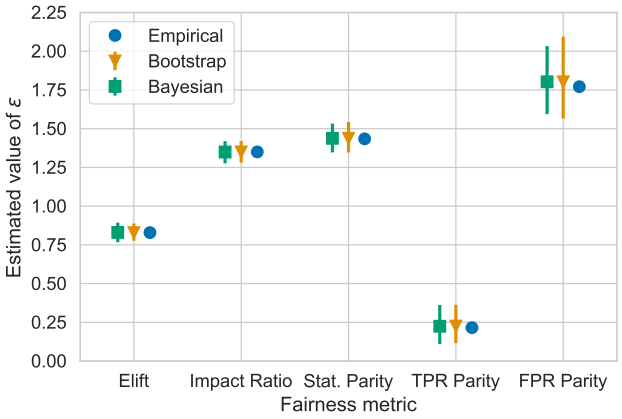

Figure 4: Estimate of ϵ-differential fairness for both data and model outputs metrics on the Adult training set when gender and age are considered as sensitive attributes. Vertical bars represent 95% confidence intervals.

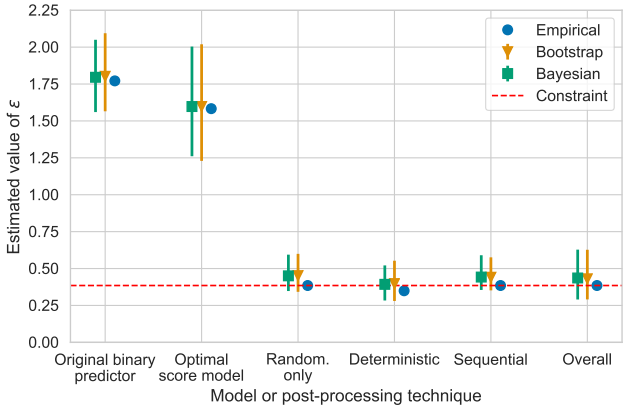

Figure 5: Estimate of ϵ-differential fairness for equalized odds across the original and the post-processed models. Results are based on the Adult training set when gender and age are considered as sensitive attributes. The constraint is set at $\epsilon \leq 1.77 - \log(4) \approx 0.384$.

i.e., $l(0,1) = l(1,0) = 1$. For instance, we decide to improve fairness by a multiplicative factor of 4; i.e., reaching an ϵ-differential fairness for FPR parity equal to $1.77 - \log(4) \approx 0.384$. As we do not want to deteriorate the performance in terms of TPR parity, we impose as a constraint to have ϵ-differential fairness for equalized odds to be less than 0.384. The calculated optimal probabilities of changing the predictions are provided in the supplementary material, but in particular we notice that we should flip positive predictions for men of age > 50 approximately 25% of the time.

We now focus on constructing a post-processed model when scores are available. The RTDP that achieves the best predictive performance can be obtained when no fairness constraints are imposed (cf. Section 4.3.1). This model represents a baseline for comparison with the other post-processed models, as it allows us to check whether imposing fairness constraints deteriorates predictive performance excessively. We refer to it as the "optimal score model".

We now aim to achieve the same value of ϵ-differential fairness for equalized odds as before, that is $\epsilon \leq 0.384$, having access to the scores as well. We construct the following three post-processed models:

- "Deterministic post-processing", where we optimize the thresholds only,
- "Sequential post-processing", where we consider the optimal score model and apply randomization on top of it,

- "Overall post-processing", where we optimize both the thresholds and probabilities simultaneously.

Figure 5 shows the value of ϵ-differential fairness for equalized odds achieved by the different post-processing techniques. We notice that, as expected, all of them reach the desired fairness constraint. Their predictive performances are compared in Table 3. When only randomization is used, the post-processed model performs significantly worse than the given binary predictor in terms of expected loss value. On the other hand, the "deterministic" and "overall" post-processed models perform almost as well as the optimal model. When comparing the "sequential" and "overall" post-processed models, we observe that the former performs significantly worse than the latter.

Whilst we chose to improve fairness by a factor of 4, in general the constraint may be set by the user according to their needs or any regulatory or other requirements.

6.2.3 More than two sensitive attributes. We repeat the same experiment by considering race as an additional sensitive attribute and using the same classifier.

Auditing intersectional fairness

Figure 6 reports the value of ϵ-differential fairness for the different metrics. Results are drastically different than the example where only age and gender were considered as sensitive attributes. We observe that the model is now even more unfair across all the different metrics, with ϵ-differential fairness for FPR parity being the worst ($\epsilon \approx 8.14$).

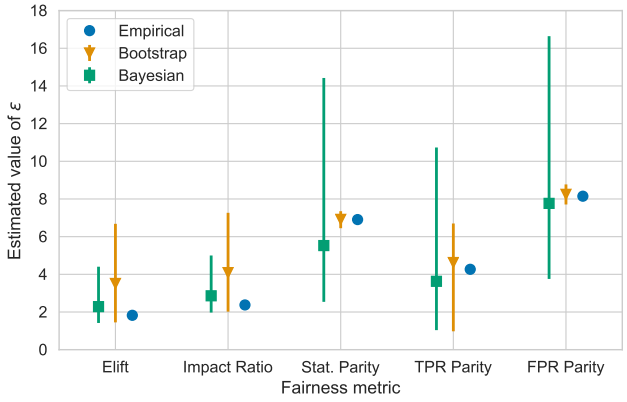

Figure 6: Estimate of ϵ-differential fairness for both data and model outputs metrics on the Adult training set when gender, age, and race are considered as sensitive attributes. Vertical bars represent 95% confidence intervals.

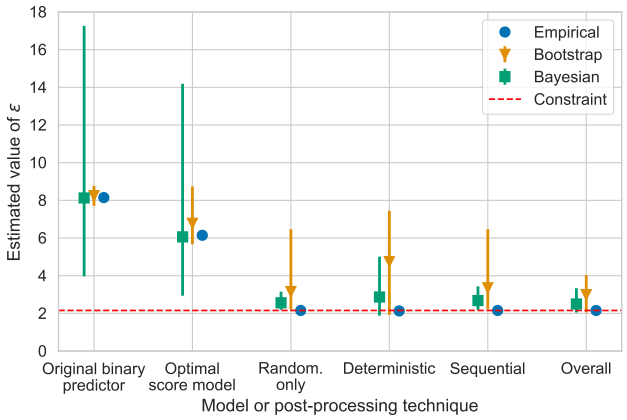

Figure 7: Estimate of ϵ-differential fairness for equalized odds across the original and the post-processed models. Results are based on the Adult training set when gender, age, and race are considered as sensitive attributes. The constraint is set at $\epsilon \leq 8.14 - \log(400) \approx 2.15$.

Post-processing for intersectional fairness

We focus on improving the equalized odds intersectional fairness metric. We proceed as above, first constructing the "optimal score model" and then building 4 different post-processing models. The first one is built on top of the binary predictor using randomization only. The other three rely on having access to the scores. We use the same loss function as before, now choosing as constraint $\epsilon \leq 8.14 - \log(400) \approx 2.15$. This constraint can be interpreted as reducing bias amplification by a multiplicative factor of 400.

The achieved fairness metrics for the different models are reported in Figure 7. We note that all the post-processed models achieve the desired fairness constraint according to the smoothed empirical estimator. The required value is also contained in the 95% confidence intervals produced by either the bootstrap or the Bayesian estimation procedure.

Table 4 reports models' predictive performances. Note that there is almost no loss in predictive performance when only randomization is used on top of the given binary predictor. However, further inspection of the post-processed probabilities of flipping the predictions reveals that one should always change positive predictions for the intersection $s_j = \{\text{Woman, Age} > 50, \text{Asian-Pac-Islander}\}$ into negative ones. Indeed, this is due to the fact that the model

Table 4: Predictive performance of given binary predictor and post-processed models on the Adult training set with gender, age, and race as sensitive attributes.

	No fairness constraints		*With fairness constraint $\epsilon \leq 8.14 - \log(400) \approx 2.15$*			
	Given binary predictor	Optimal score model	Randomization only	Deterministic	Sequential	Overall
TPR	0.5450	0.5481	0.5434	0.5995	0.5465	0.5376
FPR	0.0422	0.0427	0.0426	0.0759	0.0425	0.0400
Expected loss function	0.1416	0.1412	0.1423	0.1540	0.1415	0.1417

produces wrong predictions for this intersection more often than correct ones. This leads to a more general observation that the post-processed model represent also a valuable tool for assessing the quality of the given predictor.

Clearly, the "optimal score model" performs better in terms of predictive performance, but does not reach the desired fairness constraint. On the other hand, the deterministic post-processing model reaches the fairness constraint but the expected loss is significantly greater than for the other models. Finally, we observe that "sequential" and "overall" post-processing models perform very similarly and close to the "optimal score model".

We leave as future work a more theoretically grounded analysis of the difference in credible intervals produced by the bootstrap and Bayesian procedures.

7 CONCLUSION AND FUTURE WORK

We presented novel methods to assess and achieve intersectional fairness, where multiple sensitive attributes are considered jointly. We proposed different metrics to asses intersectional fairness of both the data and the model outputs. We outlined three different methods to robustly estimate these metrics: smoothed empirical, bootstrap, and Bayesian estimation. The last two methods allow us to asses confidence in the estimates, including rapidly evaluating which subgroups are misrepresented in the data or particularly discriminated by the model. Furthermore, we established post-processing techniques to transform the output of any given binary classifier so as to achieve better fairness with respect to the chosen intersectional fairness metric. Our methodology is particularly appealing in that it allows a practitioner to choose whether random flipping of a model prediction is desirable or not. We implemented the proposed auditing and post-processing methods on the Adult dataset.

Intersectional fairness is crucial for safe deployment of modern machine learning systems, yet most of the algorithmic fairness literature has thus far focused on fairness with respect to an individual sensitive attribute only. Our framework addresses several challenges related to auditing and achieving intersectional fairness. There are many remaining open problems that we hope future work will address, including defining other intersectional fairness metrics (e.g., for calibration) and further refining estimation procedures of fairness metrics, for instance by weighting the bootstrap sample or by differently tuning the prior parameters of the Bayesian procedure. Although we focused on post-processing, research on pre- and in-processing techniques that achieve intersectional fairness can also be carried out. Another future work avenue would be to extend our proposed post-processing methodology to regression and categorical classification problems.

ACKNOWLEDGEMENTS

We thank Imran Ahmed, Anil Choudhary, and Stavros Tsadelis for helpful comments and discussions. We would also like to thank anonymous referees for their valuable feedback, which helped us to improve the paper.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *FATMLâĂŹ17.* Association for Computing Machinery.

[2] Charles Anawo Ameh and Nynke Van Den Broek. 2008. Increased risk of maternal death among ethnic minority women in the UK. *The Obstetrician & Gynaecologist* 10, 3 (2008), 177–182.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: thereâĂŹs software used across the country to predict future criminals. And itâĂŹs biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing/. ProPublica.

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[5] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.* 16, 5 (Sept. 1995), 1190–1208. https://doi.org/10.1137/0916069

[6] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning. *arXiv preprint arXiv:1904.05419* (2019).

[7] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3992–4001. http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf

[8] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 319–328. https://doi.org/10.1145/3287560.3287586

[9] Yeounoh Chung, Tim Kraska, Neoklis Polyzotis, Kihyun Tae, and Steven Euijong Whang. 2019. Automated Data Slicing for Model Validation: A Big data - AI Integration Approach. *IEEE Transactions on Knowledge and Data Engineering* PP (05 2019), 1–1. https://doi.org/10.1109/TKDE.2019.2916074

[10] 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)*. ACM, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[11] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly Fair Representation Learning by Disentanglement. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 1436–1445. http://proceedings.mlr.press/v97/creager19a.html

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

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS âĂŹ12)*. ACM, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[14] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[15] John Forrest, Ted Ralphs, Stefan Vigerske, LouHafer, Bjarni Kristjansson, jpfasano, EdwinStraver, Miles Lubin, Haroldo Gambini Santos, rlougee, and Matthew Saltzman. 2018. coin-or/Cbc: Version 2.9.9. https://doi.org/10.5281/zenodo.1317566

[16] James Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2018. An Intersectional Definition of Fairness. arXiv:cs.LG/1807.08362

[17] Moritz Hardt, Eric Price, ecprice, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 3315–3323. http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf

[18] Tim Head, MechCoder, Gilles Louppe, Iaroslav Shcherbatyi, Fcharras, ZÃŀ VinÃŋcius, Cmmalone, Christopher Schröder, Nel215, Nuno Campos, Todd Young, Stefano Cereda, Thomas Fan, Rene-Rex, Kejia (KJ) Shi, Justus Schwabedal, Carlosdanielcsantos, Hvass-Labs, Mikhail Pak, SoManyUsernamesTaken, Fred Callaway, LoÃŕc EstÃĺve, Lilian Besson, Mehdi Cherti, Karlson Pfannschmidt, Fabian Linzberger, Christophe Cauet, Anna Gut, Andreas Mueller, and Alexander Fabisch. 2018. Scikit-Optimize/Scikit-Optimize: V0.5.2. https://doi.org/10.5281/zenodo.1207017

[19] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In

Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research), Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, StockholmsmÃďssan, Stockholm Sweden, 1939–1948. http://proceedings.mlr.press/v80/hebert-johnson18a.html

[20] Minna J. Kotkin. 2008. Diversity and Discrimination: A Look at Complex Bias. *William and Mary Law Rev.* 50 (04 2008).

[21] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi Malvajerdi, and Jonathan Ullman. 2019. Differentially Private Fair Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 3000–3008. http://proceedings.mlr.press/v97/jagielski19a.html

[22] Faisal Kamiran and Toon Calders. 2012. Data Preprocessing Techniques for Classification Without Discrimination. *Knowl. Inf. Syst.* 33, 1 (Oct. 2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[23] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*, Peter A. Flach, Tijl De Bie, and Nello Cristianini (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 35–50.

[24] Narendra Karmarkar. 1984. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*. ACM, 302–311.

[25] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, StockholmsmÃďssan, Stockholm Sweden, 2564–2572. http://proceedings.mlr.press/v80/kearns18a.html

[26] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An Empirical Study of Rich Subgroup Fairness for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 100–109. https://doi.org/10.1145/3287560.3287592

[27] Jouni Kerman. 2011. Neutral noninformative and informative conjugate beta and gamma prior distributions. *Electron. J. Statist.* 5 (2011), 1450–1470. https://doi.org/10.1214/11-EJS648

[28] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. ACM, New York, NY, USA, 247–254. https://doi.org/10.1145/3306618.3314287

[29] Jon Kleinberg. 2018. Inherent Trade-Offs in Algorithmic Fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS âĂŹ18)*. ACM, New York, NY, USA, 40–40. https://doi.org/10.1145/3219617.3219634

[30] Dieter Kraft. 1988. A Software Package for Sequential Quadratic Programming. (1988). https://books.google.co.uk/books?id=4rKaGwAACAAJ

[31] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*. AAAI Press, 2124–2132. http://dl.acm.org/citation.cfm?id=3298483.3298546

[32] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, StockholmsmÃďssan, Stockholm Sweden, 3384–3393. http://proceedings.mlr.press/v80/madras18a.html

[33] Edward B. Manoukian. 1986. *Modern Concepts and Theorems of Mathematical Statistics*. Springer New York. https://doi.org/10.1007/978-1-4612-4856-9

[34] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Conference on Fairness, Accountability, and Transparency*.

[35] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware Data Mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD âĂŹ08)*. ACM, New York, NY, USA, 560–568. https://doi.org/10.1145/1401890.1401959

[36] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., USA, 5684–5693. http://dl.acm.org/citation.cfm?id=3295222.3295319

[37] Edward Raff, Jared Sylvester, and Steven Mills. 2018. Fair Forests: Regularized Tree Induction to Minimize Model Bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. ACM, New York, NY, USA, 243–250. https://doi.org/10.1145/3278721.3278742

[38] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. In *Proceedings of the 2017 Conference on Learning Theory (Proceedings of Machine Learning Research)*, Satyen Kale and Ohad Shamir (Eds.), Vol. 65. PMLR, Amsterdam, Netherlands, 1920–1953. http://proceedings.mlr.press/v65/woodworth17a.html

[39] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28 (ICML'13)*. JMLR.org, III–325–III–333. http://dl.acm.org/citation.cfm?id=3042817.3042973

[40] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. ACM, New York, NY, USA, 335–340. https://doi.org/10.1145/3278721.3278779

[41] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *arXiv e-prints*, Article arXiv:1707.09457 (July 2017). arXiv:1707.09457

[42] Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. *arXiv e-prints*, Article arXiv:1511.00148 (Oct. 2015). arXiv:1511.00148

A PROOFS OF SECTION 3

Proof of Theorem 3.1. Theorem VIII.1 of Foulds et al. [16] proves the result in the case of ϵ-differential fairness for statistical parity. The proof is based on the following reformulation of the original definition (Lemma VIII.1, [16]):

$$\log\left(\max_{s_i\in A}\hat{\mu}_{1|s_i}\right) - \log\left(\min_{s_i\in A}\hat{\mu}_{1|s_i}\right) \leq \epsilon,$$

and on proving

$$\begin{aligned}\log\left(\max_{s_i\in A}\hat{\mu}_{1|s_i}\right) &\geq \log\left(\max_{s_i\in A'}\hat{\mu}_{1|s_i}\right), \\ \log\left(\min_{s_i\in A}\hat{\mu}_{1|s_i}\right) &\leq \log\left(\min_{s_i\in A'}\hat{\mu}_{1|s_i}\right).\end{aligned} \tag{8}$$

An analogous reformulation holds for the definitions of ϵ-differential fairness for impact ratio, TPR parity and FPR parity. Therefore, the desired result hold for these metrics by reproducing the proof of Theorem VIII.1 of Foulds et al. [16].

The definition of ϵ-differential fairness for the elift metric can be reformulated as:

$$\log\left(\max_{s_i\in A}\hat{\mu}_{1|s_i}\right) - \log(\mu_1) \leq \epsilon,$$

and so from Equation 8 it follows

$$\log\left(\max_{s_i\in A'}\hat{\mu}_{1|s_i}\right) - \log(\mu_1) \leq \log\left(\max_{s_i\in A}\hat{\mu}_{1|s_i}\right) - \log(\mu_1) \leq \epsilon,$$

as desired. □

Proof of Proposition 3.2. We prove the result for impact ratio, but similar reasoning can be applied to prove consistency for all the ϵ-differential fairness metrics introduced in Tables 1 and 2. Assume we have access to a dataset containing n observations; we make the dependency on n explicit by using superscript n. We shall prove that $\hat{\epsilon}^n_{IR}$ converges in probability to ϵ_{IR}, as defined in Equation 1.

Recall that we defined N_{1,s_i} as the number of occurrences in the dataset of individuals with attributes s_i and positive outcome, while N_{s_i} is the number of individuals with attribute s_i. Define the following estimators of $\mu_{1,s_i} := \mathbb{P}(Y=1, S=s_i)$ and $\mu_{s_i} := \mathbb{P}(S=s_i)$:

$$\hat{\mu}^n_{1,s_i} = \frac{N_{1,s_i}}{n}, \quad \hat{\mu}^n_{s_i} = \frac{N_{s_i}}{n},$$

respectively. The two estimators are consistent by the Strong Law of Large Numbers. We can now apply Slutsky's theorem [33, p. 76] and show:

$$\hat{\mu}^n_{1|s_i} = \frac{N_{1,s_i}+\alpha}{N_{s_i}+2\beta} = \frac{\hat{\mu}^n_{1,s_i}+\frac{\alpha}{n}}{\hat{\mu}^n_{s_i}+\frac{2\beta}{n}} \xrightarrow{p} \frac{\mu_{1,s_i}}{\mu_{s_i}} = \mu_{1|s_i},$$

assuming $\mu_{1|s_i} > 0, \forall s_i \in A$. Moreover, by applying again Slutsky's theorem, it follows:

$$\frac{\hat{\mu}^n_{1|s_i}}{\hat{\mu}^n_{1|s_j}} \xrightarrow{p} \frac{\mu_{1|s_j}}{\mu_{1|s_j}}.$$

Finally, by the Continuous Mapping Theorem, we conclude that $\hat{\epsilon}^n_{IR}$ is a consistent estimator of ϵ_{IR}. □

Proof of Proposition 3.3. Notice that the expected value of the posterior distribution is given by Equation 2, while the variance is $o\left(\frac{1}{n}\right)$. Therefore, as $n\to\infty$ the posterior distribution converges to a Dirac delta concentrated in $\hat{\mu}_{1|s_i}$. In the proof of Proposition 3.2 we showed that $\hat{\mu}_{1|s_i}$ converges in probability to $\mu_{1|s_i}$. It then follows by the Central Limit Theorem that the Monte Carlo procedure yields consistent estimates. □

B PROOFS OF SECTION 4

Proof of Proposition 4.3. Recall we assumed w.l.o.g. that $l(0,0) = l(1,1) = 0$. Then:

$$\begin{aligned}\mathbb{E}[l(Y,\tilde{Y})] &= \mathbb{P}(Y=0,\tilde{Y}=1)\,l(0,1) + \mathbb{P}(Y=1,\tilde{Y}=0)\,l(1,0) \\ &= \mathbb{P}(\tilde{Y}=1|Y=0)\,\mathbb{P}(Y=0)\,l(0,1) \\ &\quad + \mathbb{P}(\tilde{Y}=0|Y=1)\,\mathbb{P}(Y=1)\,l(1,0) \\ &= F\tilde{P}R\,(1-\mu_1)\,l(0,1) + F\tilde{N}R\,\mu_1 l(1,0).\end{aligned}$$

Therefore:

$$\min\mathbb{E}[l(Y,\tilde{Y})] = \min\{F\tilde{P}R\,(1-\mu_1)\,l(0,1) + F\tilde{N}R\,\mu_1\,l(1,0)\},$$

as desired. □

Proof of Proposition 4.4. Consider

$$\begin{aligned}\mathbb{E}[Y\tilde{Y} - c\tilde{Y}] &= \mathbb{P}(Y=1,\tilde{Y}=1) - c\,\mathbb{P}(\tilde{Y}=1) \\ &= \mathbb{P}(\tilde{Y}=1|Y=1)\,\mathbb{P}(Y=1) \\ &\quad - c\left(\mathbb{P}(\tilde{Y}=1|Y=0)\,\mathbb{P}(Y=0)\right. \\ &\quad \left.+\mathbb{P}(\tilde{Y}=1|Y=1)\,\mathbb{P}(Y=1)\right) \\ &= T\tilde{P}R\,\mu_1 - c\,F\tilde{P}R\,(1-\mu_1) - c\,T\tilde{P}R\,\mu_1 \\ &= (1-c)\,\mu_1\,(1-F\tilde{N}R) - c\,(1-\mu_1)\,F\tilde{P}R.\end{aligned}$$

Therefore by Proposition 4.3:

$$\begin{aligned}\max\mathbb{E}[Y\tilde{Y} - c\tilde{Y}] &= \min c\,(1-\mu_1)\,F\tilde{P}R + (c-1)\,\mu_1\,(1-F\tilde{N}R) \\ &= \min c\,(1-\mu_1)\,F\tilde{P}R + (1-c)\,\mu_1\,F\tilde{N}R \\ &= \min\mathbb{E}[l(Y,\tilde{Y})]\end{aligned}$$

where $l(0,1) = c$ and $l(1,0) = 1-c$. □

Proof of Proposition 4.5. Denote the FPR for individuals with attribute s_i of the given model as $F\hat{P}R_{s_i} := \mathbb{P}(\hat{Y}=1|Y=0,S=s_i)$ and the FNR as $F\hat{N}R_{s_i} := \mathbb{P}(\hat{Y}=0|Y=1,S=s_i)$. It follows:

$$\begin{aligned}F\tilde{P}R_{s_i} &= \tilde{p}_{0,s_i}(1-F\hat{P}R_{s_i}) + \tilde{p}_{1,s_i}F\hat{P}R_{s_i}, \\ F\tilde{N}R_{s_i} &= (1-\tilde{p}_{0,s_i})\,F\hat{N}R_{s_i} + (1-\tilde{p}_{1,s_i})\,(1-F\hat{N}R_{s_i}).\end{aligned}$$

Therefore $F\tilde{P}R(1-\mu_1)l(0,1) + F\tilde{N}R\mu_1 l(1,0)$ is a linear combination of the variables $\tilde{p}_{0,s_i}$ and $\tilde{p}_{1,s_i}$. By Proposition 4.3, minimizing Equation 4 is equivalent to minimizing $\mathbb{E}[l(Y,\tilde{Y})]$, so that the objective function is indeed linear. All that remains now is to also show that the optimization constraints are linear.

Consider for instance using statistical parity as the fairness constraint, that is $e^{-\epsilon} \leq \frac{\mathbb{P}(\tilde{Y}=1|S=s_i)}{\mathbb{P}(\tilde{Y}=1|S=s_j)} \leq e^{\epsilon}$ for all $s_i, s_j \in A$. Notice that

by the law of total probability it follows:

$$\begin{aligned}\mathbb{P}(\tilde{Y}=1|S=s_i) &= \mathbb{P}(\tilde{Y}=1|Y=0,S=s_i)\mathbb{P}(Y=0|S=s_i)\\ &\quad+\mathbb{P}(\tilde{Y}=1|Y=1,S=s_i)\mathbb{P}(Y=1|S=s_i)\\ &= F\tilde{P}R_{s_i}(1-\mu_{1|s_i})+(1-F\tilde{N}R_{s_i})\mu_{1|s_i},\end{aligned}$$

and that we have already shown that $F\tilde{P}R_{s_i}$ and $F\tilde{N}R_{s_i}$ are linear in the variables to be optimized. The same conclusion holds when equal opportunity or FPR parity are considered as constraints, and therefore also when equalized odds is considered. Indeed, we can require (as our fairness constraint) multiple ϵ-differential fairness definitions to hold simultaneously, possibly each for different values of ϵ. □

Proof of Proposition 4.7. Following the same steps as in the proof of Proposition 4.3, we first notice that the expected loss function marginalizes as:

$$\begin{aligned}\mathbb{E}[l(Y,\tilde{Y})] &= \sum_{s_i\in A}\big[\mathbb{P}(\tilde{Y}=1|Y=0,S=s_i)\,\mu_{s_i}\,(1-\mu_{1|s_i})\,l(0,1)\\ &\quad+\mathbb{P}(\tilde{Y}=0|Y=1,S=s_i)\,\mu_{s_i}\,\mu_{1|s_i}\,l(1,0)\big]\\ &= \sum_{s_i\in A}\mu_{s_i}\,\left[F\tilde{P}R_{s_i}(1-\mu_{1|s_i})\,l(0,1)+F\tilde{N}R_{s_i}\,\mu_{1|s_i}\,l(1,0)\right],\end{aligned} \tag{9}$$

so that it suffices to prove the result when solving

$$\min_{\tau_{s_i},\tilde{p}_{0,s_i},\tilde{p}_{1,s_i}}\left\{F\tilde{P}R_{s_i}\,(1-\mu_{1|s_i})\,l(0,1)+F\tilde{N}R_{s_i}\,\mu_{1|s_i}\,l(1,0)\right\},$$

for an arbitrary $s_i \in A$. From Definition 4.2 of definition of RTDP, it follows:

$$\begin{aligned}F\tilde{P}R_{s_i} &= TNR^{\star}_{s_i}\,\tilde{p}_{0,s_i}+FPR^{\star}_{s_i}\,\tilde{p}_{1,s_i},\\ F\tilde{N}R_{s_i} &= FNR^{\star}_{s_i}\,(1-\tilde{p}_{0,s_i})+TPR^{\star}_{s_i}\,(1-\tilde{p}_{1,s_i}),\end{aligned}$$

where, although not explicitly stated, $F\tilde{P}R_{s_i}$ and $F\tilde{N}R_{s_i}$ are functions of the variables $\tau_{s_i},\tilde{p}_{1,s_i},\tilde{p}_{0,s_i}$. Therefore:

$$\begin{aligned}&\min_{\tau_{s_i},\tilde{p}_{0,s_i},\tilde{p}_{1,s_i}}\left\{F\tilde{P}R_{s_i}\,(1-\mu_{1|s_i})\,l(0,1)+F\tilde{N}R_{s_i}\,\mu_{1|s_i}\,l(1,0)\right\}\\ &=\min_{\tau_{s_i},\tilde{p}_{0,s_i},\tilde{p}_{1,s_i}}\left\{[TNR^{\star}_{s_i}\,\tilde{p}_{0,s_i}+FPR^{\star}_{s_i}\,\tilde{p}_{1,s_i}](1-\mu_{1|s_i})\,l(0,1)\right.\\ &\qquad\left.+[FNR^{\star}_{s_i}\,(1-\tilde{p}_{0,s_i})+TPR^{\star}_{s_i}\,(1-\tilde{p}_{1,s_i})]\mu_{1|s_i}\,l(1,0)\right\}\\ &=\min_{\tau_{s_i},\tilde{p}_{0,s_i},\tilde{p}_{1,s_i}}\left\{\tilde{p}_{1,s_i}[FPR^{\star}_{s_i}\,(1-\mu_{1|s_i})\,l(0,1)-TPR^{\star}_{s_i}\,\mu_{1|s_i}\,l(1,0)]\right.\\ &\qquad+\tilde{p}_{0,s_i}[TNR^{\star}_{s_i}\,(1-\mu_{1|s_i})\,l(0,1)-FNR^{\star}_{s_i}\,\mu_{1|s_i}\,l(1,0)]\\ &\qquad\left.+TPR^{\star}_{s_i}\,\mu_{1|s_i}\,l(1,0)+FNR^{\star}_{s_i}\,\mu_{1|s_i}\,l(1,0)\right\}.\end{aligned}$$

Under the assumptions of Equation 5, it follows:

$$\begin{aligned}FPR^{\star}_{s_i}\,(1-\mu_{1|s_i})\,l(0,1)-TPR^{\star}_{s_i}\,\mu_{1|s_i}\,l(1,0)<0,\\ TNR^{\star}_{s_i}\,(1-\mu_{1|s_i})\,l(0,1)-FNR^{\star}_{s_i}\,\mu_{1|s_i}\,l(1,0)>0,\end{aligned}$$

so that to minimize the desired quantity, we must set $\tilde{p}_{1,s_i}=1$ and $\tilde{p}_{0,s_i}=0$ as desired. □

C EXTRA MATERIAL EXPERIMENTS

C.1 Implementation details

We now discuss the optimization algorithms we have used to run the experiments of Section 6 for the different post-processing techniques:

- Linear programming: we use the coin-or branch and cut solver [15],
- Constrained optimization: we use sequential quadratic programming [30],
- Unconstrained optimization: we consider two different approaches. The first one resorts to the L-BFGS-B algorithm [5]. This algorithm approximates gradient information and therefore we make use of the smoothing technique proposed in the previous paragraph. The second one is a Bayesian optimizer that approximates the objective function with a Gaussian process [18]. As a result, this optimizer can deal with non-differentiable functions as it does not rely on gradient information. However, its outputs are stochastic, so that assessing convergence is more complicated.

C.2 Two sensitive attributes (Section 6.2.2)

Table 5: Predictive performance of given binary predictor and post-processed models on the Adult test set with gender and age as sensitive attributes.

	No fairness constraints		*With fairness constraint* $\epsilon \leq 1.77 - \log(4) \approx 0.385$			
	Given binary predictor	Optimal score model	Randomization only	Deterministic	Sequential	Overall
TPR	0.5216	0.5257	0.4919	0.5530	0.4943	0.5438
FPR	0.0433	0.0545	0.0510	0.0622	0.0497	0.0574
Expected loss function	0.1461	0.1465	0.1590	0.1531	0.1574	0.1516

Table 6: Probabilities of flipping the original predictions for the "randomization only" post-processing method. The post-processed model has been constructed on a binary classifier trained on the Adult training set when gender and age are considered as sensitive attributes.

	Model prediction	
	Income $\leq$ 50k	Income $>$ 50k
Female, Age $\leq$ 50	0.03	0
Female, Age $>$ 50	0.02	0
Male, Age $\leq$ 50	0	0
Male, Age $>$ 50	0	0.26

(a) Log-TPR Parity Ratio **(b) Log-FPR Parity Ratio**

Figure 8: Log-value of the ratios in the definition of ϵ-differential fairness for TPR and FPR parity when gender and age are considered as sensitive attributes. The model has been built on the Adult training set and it exhibits $\epsilon = 0.22$ for TPR parity and $\epsilon = 1.77$ for FPR parity.

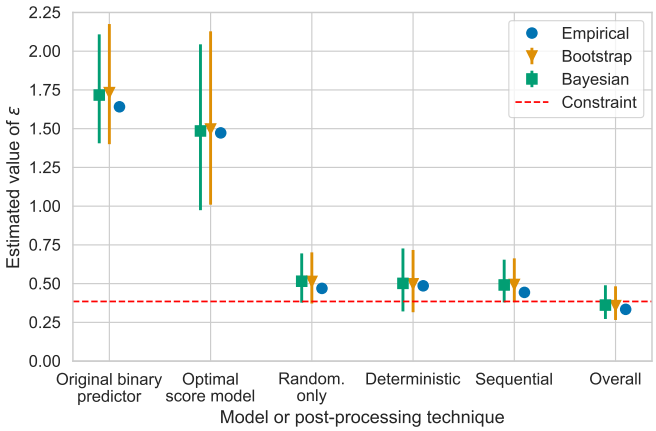

Figure 9: Estimate of ϵ-differential fairness for equalized odds across the original and the post-processed models. Results are based on the Adult test set when gender and age are considered as sensitive attributes. The constraint is set at $\epsilon \leq 1.77 - \log(4)$.

C.3 More than Two Sensitive Attributes (Section 6.2.3)

Table 7: Predictive performance of given binary predictor and post-processed models on the Adult test set with gender, age, and race as sensitive attributes.

	No fairness constraints		*With fairness constraint* $\epsilon \leq 8.14 - \log(400) \approx 0.157$			
	Given binary predictor	Optimal score model	Randomization only	Deterministic	Sequential	Overall
TPR	0.5216	0.5258	0.5197	0.5785	0.5238	0.5139
FPR	0.0433	0.0451	0.0439	0.0762	0.0452	0.0413
Expected loss function	0.1416	0.1465	0.1470	0.1578	0.1470	0.1464

Table 8: Probabilities of flipping the original predictions for the "randomization only" post-processing method. The post-processed model has been constructed on a binary classifier trained on the Adult training set when gender and age are considered as sensitive attributes. Unreported combinations of sensitive attributes have probability of flipping equal to 0.

	Model prediction	
	Income $\leq$ 50k	Income $>$ 50k
Female, Age $\leq$ 50, Asian-Pacific Islander	0.01	0
Female, Age $\leq$ 50, Black	0.01	0
Female, Age $>$ 50, Asian-Pacific Islander	0.09	1
Female, Age $\leq$ 50, Black	0.01	0
Female, Age $>$ 50, Other	0.07	0
Male, Age $\leq$ 50, Asian-Pacific Islander	0	0.01
Female, Age $>$ 50, Asian-Pacific Islander	0	0.35

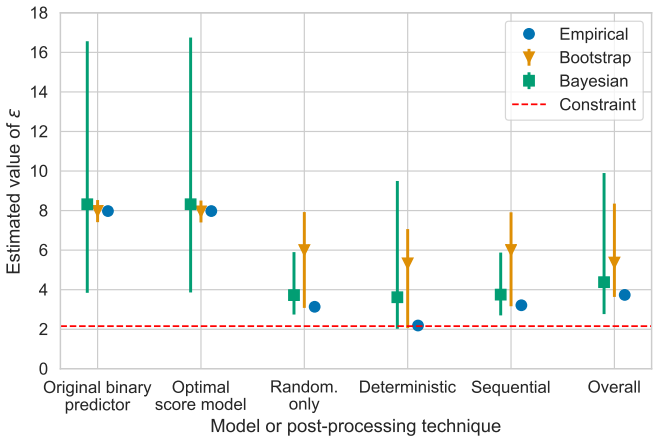

Figure 10: Estimate of ϵ-differential fairness for equalized odds across the original and the post-processed models. Results are based on the Adult test set when gender, age, and race are considered as sensitive attributes. The constraint is set at $\epsilon \leq 8.14 - \log(400) \approx 2.15$.