ATTAINABILITY AND OPTIMALITY: THE EQUALIZED-ODDS FAIRNESS REVISITED

Anonymous authors
Paper under double-blind review

ABSTRACT

Fairness of machine learning algorithms has been of increasing interest. In order to suppress or eliminate discrimination in prediction, various notions as well as approaches to impose fairness have been proposed. However, in different scenarios, whether or not the chosen notion of fairness can always be attained, even if with unlimited amount of data, is not well addressed. In this paper, focusing on the Equalized Odds notion of fairness, we consider the attainability of this criterion, and furthermore, if attainable, the optimality of the prediction performance under various settings. In particular, for classification with a deterministic prediction function of the input, we give the condition under which Equalized Odds can hold true; if randomized prediction is acceptable, we show that under mild assumptions, fair classifiers can always be derived. Moreover, we prove that compared to enforcing fairness by post-processing, one can always benefit from exploiting all available features during training and get better prediction performance while remaining fair. However, for regression tasks, Equalized Odds is not always attainable if certain conditions on the joint distribution of the features and the target variable are not met. This indicates the inherent difficulty in achieving fairness in certain cases and suggests a broader class of prediction methods might be needed for fairness.

1 INTRODUCTION

As machine learning models become widespread in automated decision making systems, apart from the efficiency and accuracy of the prediction, their potential social consequence also gains increasing attention. To date, there is ample evidence that machine learning models have resulted in discrimination against certain groups of individuals under many circumstances, for instance, the discrimination in ad delivery when searching for names that can be predictive of the race of individual (Sweeney, 2013); the gender discrimination in job-related ads push (Datta et al., 2015); stereotypes associated with gender in word embeddings (Bolukbasi et al., 2016); the bias against certain ethnicities in the assessment of recidivism risk (Angwin et al., 2016).

The call for accountability and fairness in machine learning has motivated various (statistical) notions of fairness. The *Demographic Parity* criterion (Calders et al., 2009) requires the independence between prediction (e.g., of a classifier) and the protected feature (sensitive attributes of an individual, e.g., gender, race). *Equalized Odds* (Hardt et al., 2016), also known as *Error-rate Balance* (Chouldechova, 2017), requires the output of a model be conditionally independent of protected feature(s) given the ground truth of the target. *Predictive Rate Parity* (Zafar et al., 2017a), on the other hand, requires the actually proportion of positives (negatives) in the original data for positive (negative) predictions should match across groups (well-calibrated).

On the theoretical side, results have been reported regarding relationships among fairness notions. It has been independently shown that if base rates of true positives differ among groups, then Equalized Odds and Predictive Rate Parity cannot be achieved simultaneously for non-perfect predictors (Kleinberg et al., 2016; Chouldechova, 2017). Any two out of three among Demographic Parity, Equalized Odds, and Predictive Rate Parity are incompatible with each other (Barocas et al., 2017). At the interface of privacy and fairness, the impossibility of achieving both *Differential Privacy* (Dwork et al., 2006) and *Equal Opportunity* (Hardt et al., 2016) while maintaining non-trivial accuracy is also established (Cummings et al., 2019).

In practice, one can broadly categorize computational procedures to derive a fair predictor into three types: pre-processing approaches (Calders et al., 2009; Dwork et al., 2012; Zemel et al., 2013; Zhang et al., 2018; Madras et al., 2018; Creager et al., 2019; Zhao et al., 2020), in-processing approaches (Kamishima et al., 2011; Pérez-Suay et al., 2017; Zafar et al., 2017a;b; Donini et al., 2018; Song et al., 2019; Mary et al., 2019; Baharlouei et al., 2020), and post-processing approaches (Hardt et al., 2016; Fish et al., 2016; Dwork et al., 2018). In accord with the fairness notion of interest, a pre-processing approach first maps the training data to a transformed space to remove discriminatory information between protected feature and target, and then pass on the data to make prediction. In direct contrast, a post-processing approach treats the off-the-shelf predictor(s) as uninterpretable black-box(es), and imposes fairness by outputting a function of the original prediction. For in-processing approaches, various kinds of regularization terms are proposed so that one can optimize the utility function while suppressing the discrimination at the same time. Approaches based on estimating/bounding causal effect between the protected feature and final target have also been proposed (Kusner et al., 2017; Russell et al., 2017; Zhang et al., 2017; Nabi & Shpitser, 2018; Zhang & Bareinboim, 2018; Chiappa, 2019; Wu et al., 2019).

Focusing on the Equalized-Odds criterion, although various approaches have been proposed to impose the fairness requirement, whether or not it is always attainable is not well addressed. The attainability of Equalized Odds, namely, the existence of the predictor that can score zero violation of fairness in the large sample limit, is an asymptotic property of the fairness criterion. This characterizes a completely different kind of violation of fairness compared to the empirical error bound of discrimination in finite-sample cases. If utilizing a "fair" predictor which is actually biased, the discrimination would become a snake in the grass, making it hard to detect and eliminate. Actually, as we illustrate in this paper, Equalized Odds is not always attainable for regression and even classification tasks, if we use deterministic prediction functions. This calls for alternative definitions in the same spirit as Equalized Odds that can always be achieved under various circumstances. Our contributions are mainly:

- For regression and classification tasks with deterministic prediction functions, we show that Equalized Odds is not always attainable if certain (rather restrictive) conditions on the joint distribution of the features and the target variable are not met.
- Under mild assumptions, for binary classification we show that if randomized prediction is taken into consideration, one can always derive a non-trivial Equalized Odds classifier.
- Considering the optimality of performance under fairness constraint(s), when exploiting all available features, we show that the predictor derived via an in-processing approach would always outperform the one derived via a post-processing approach (unconstrained optimization followed by a post-processing step).

2 PRELIMINARIES

In this section, we first illustrate the difference between *prediction fairness* and *procedure fairness*, and then, we present the formal definition of Equalized Odds (Hardt et al., 2016).

2.1 HIERARCHY OF FAIRNESS

Before presenting the formulation of fairness, it is important to see the distinction between different levels of fairness when discussing fair predictors. When evaluating the performance of the proposed fair predictor, it is a common practice to compare the loss (with respect to the utility function of choice, e.g., accuracy for binary classification) computed on target variable and the predicted value. There is an implicit assumption lying beneath this practice: the generating process of the data, which is just describing a real-world procedure, is *not* biased in any sense (Danks & London, 2017). Only when we treat the target variable (recorded in the dataset) as unbiased can we justify the practice of loss evaluation and the conditioning on target variable when imposing fairness (as we shall see in the definition of Equalized Odds in Equation 1).

One may consider a music school admission example. The music school committee would decide if they admit a student to the violin performance program based on the applicant's personal information, educational background, instrumental performance, and so on. When evaluating whether or not

the admission is "fair", there are actually two levels of fairness. First, based on the information at hand, did the committee evaluate the qualification of applicants without bias (How committee evaluate the applicants)? And second, is committee's procedure of evaluating applicants' qualification reasonable (How other people view the evaluation procedure used by the committee)?

In this paper, we consider *prediction fairness*, namely, assuming the data recorded is unbiased, the prediction (made with respect to current reality) itself should not include any biased utilization of information. The fairness with respect to the data generating procedure as well as the potential future influence of the prediction are beyond the scope of this paper.

2.2 Equalized-Odds Fairness

Hardt et al. (2016) proposed *Equalized Odds* which requires conditional independence between prediction and protected feature(s) given ground truth of the target. Let us denote the protected feature by A, with domain of value $\mathcal{A}$, additional (observable) feature(s) by X, with domain of value $\mathcal{X}$, target variable by Y, with domain $\mathcal{Y}$, (not necessarily fair) predictors by $\widehat{Y}$, and fair predictors by $\widetilde{Y}$. Equalized-Odds fairness requires

$$\widetilde{Y} \perp\!\!\!\perp A \mid Y. \tag{1}$$

For classification tasks, one can conveniently use the probability distribution form:

$$\forall a \in \mathcal{A}, t, y \in \mathcal{Y} : P(\widetilde{Y} = t \mid A = a, Y = y) = P(\widetilde{Y} = t \mid Y = y), \tag{2}$$

$$\text{or more concisely, } P_{\widetilde{Y}|AY}(t \mid a, y) = P_{\widetilde{Y}|Y}(t \mid y). \tag{3}$$

For better readability, we also use the formulation in Equation 3 in cases without ambiguity. In the context of binary classification ($\mathcal{Y} = \{0, 1\}$), Equalized Odds requires that the True Positive Rate (TPR) and False Positive Rate (FPR) of each certain group match population positive rates. Throughout the paper, without loss of generality we assume there is only one protected feature for the purpose of simplifying notation. However, considering the fact that the protected feature can be discrete (e.g., race, gender) or continuous (e.g., the ratio of ethnic group in the population for certain district of a city), we do not assume discreteness of the protected feature. Due to the space limit, we will focus on the illustration and implication of our results and defer all the proofs to the appendix.

3 Fairness in Regression May Not be Attained

In this section we consider the attainability of Equalized Odds for regression tasks, namely, whether or not it is possible to find a predictor that is conditionally independent from the protected feature given true value of the target. For linearly Gaussian cases, one can attain Equalized Odds by constraining zero partial correlation between the prediction and the protected feature given target variable (Woodworth et al., 2017). Various regularization terms have also been proposed to suppress discrimination when predicting a continuous target (Berk et al., 2017; Mary et al., 2019). However, whether or not one can always achieve 0-discrimination for regression, even if with an unlimited amount of data, is not clear yet.

If "fair" predictors are deployed without carefully checking the attainability of fairness, the discrimination would become a hidden hazard, making it hard to detect and eliminate. Actually as we will show in this section, even in the simple setup of linearly correlated continuous data, Equalized Odds is not always attainable.

3.1 Unattainability of Equalized Odds in Linear non-Gaussian Regression

As stated in Section 2.1, in this paper we consider *prediction fairness*, and therefore any possible bias introduced by the data generating procedure itself is beyond the scope of the discussion. Consider the situation where the data is generated as following (H is not measured in the dataset):

$$\begin{aligned} X &= qA + E_X, \\ H &= bA + E_H, \\ Y &= cX + dH + E_Y, \end{aligned} \tag{4}$$

where (A, E_X, E_H, E_Y) are mutually independent. In fact, if at most one of E_X and $E := E_Y + dE_H$ is Gaussian, then any linear combination of A and X with non-zero coefficients will not be conditionally independent from A given Y, meaning that it is not possible to achieve Equalized-Odds fairness. Let Z be a linear combination of A and X, i.e., $Z = \alpha A + \beta X = (\alpha + q\beta)A + \beta E_X$, with linear coefficients α and β, where $\beta \neq 0$. In Theorem 3.1, we present the general result in linear non-Gaussian cases, where one cannot achieve the conditional independence between Z and A given Y.

Theorem 3.1. ***(Unattainability of Equalized Odds in the Linear Non-Gaussian Case)*** *Assume that feature X has a causal influence on Y, i.e., $c \neq 0$ in Equation 4, and that the protected feature A and Y are not independent, i.e., $qc + bd \neq 0$. Assume p_{E_X} and p_E are positive on $\mathbb{R}$. Let $f_1 := \log p_A$, $f_2 := \log p_{E_X}$, and $f_3 := \log p_E$. Further assume that f_2 and f_3 are third-order differentiable. Then if at most one of E_X and E is Gaussian, Z is always conditionally dependent on A given Y.*

From Theorem 3.1, we see that in linear non-Gaussian cases, any non-zero linear combination of the feature (which is a deterministic function of the input) will not satisfy Equalized Odds. One may wonder whether Equalized Odds can be achieved by nonlinear regression, instead of a linear model. Although a proof with general nonlinear models is rather complicated, our simulation results in Section 5.1 strongly suggest that the unattainability of Equalized Odds persists in nonlinear regression cases.

In light of the unattainability of Equalized Odds for prediction with deterministic functions of A and X, it is desirable to develop general, nonlinear prediction algorithms to produce a probabilistic prediction (i.e., with a certain type of randomness in the prediction). One possible way follows the framework of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014): we use random standard Gaussian noise, in addition to A and X, as input, such that the output will have a specific type of randomness. The parameters involved are learned by minimizing prediction error and enforcing Equalized Odds on the "randomized" output at the same time. Given that this approach is not essential to illustrate the claims made in this paper and that theoretical properties of such nonlinear regression algorithms with randomized output are not straightforward to establish, this is left as future work.

4 Fairness in Classification

In this section, we consider the attainability of Equalized Odds for binary classifiers (with a deterministic or randomized prediction function), and furthermore, if attainable, the optimality of performance under the fairness criterion. Admittedly, as is already pointed out by Woodworth et al. (2017), we generally cannot have 0-discriminatory predictors with a finite number of samples; instead, one should consider imposing δ-discrimination in practice (δ is the violation of Equalized Odds). However, this does not guarantee nor rule out the possibility of attaining 0-discrimination on population when the sample size goes to infinity.

4.1 Classification with Deterministic Prediction

We begin with considering cases when the classification is performed by a deterministic function of the input. In particular, we derive the condition under which Equalized Odds can possibly hold true.

Theorem 4.1. *Assume that the protected feature A and Y are dependent and that their joint probability $P(A, Y)$ (for discrete A) or joint probability density $p(A, Y)$ (for continuous A) is positive for every combination of possible values of A and Y. Further assume that Y is not fully determined by A, and that there are additional features X that are not independent of Y. Let the output of the classifier $\widetilde{Y}$ be a deterministic function $f : \mathcal{A} \times \mathcal{X} \to \mathcal{Y}$. Let $S_A^{(t)} := \{a \mid \exists x \in \mathcal{X} \text{ s.t. } f(a, x) = t\}$, and $S_{X|a}^{(t)} := \{x \mid f(a, x) = t\}$. Equalized Odds holds true if and only if the following two conditions are satisfied:*

(i) $\forall t \in \mathcal{Y} :\ S_A^{(t)} = \mathcal{A}$,

(ii) $\forall t \in \mathcal{Y},\ \forall a, a' \in \mathcal{A}$ *(for continuous X, replace summation with integration accordingly):* $\Sigma_{x \in S_{X|a}^{(t)}} P_{X|AY}(x \mid a, y) = \Sigma_{x \in S_{X|a'}^{(t)}} P_{X|AY}(x \mid a', y)$.

Let us take a look at the two conditions. Condition (i) says that within each class determined by the classification function f, A should be able to take all possible values in $\mathcal{A}$. While condition (i) is already pretty restrictive, condition (ii) specifies an even stronger constraint on the relation between the conditional probability $P_{X|AY}(x|a,y)$ (or the conditional probability density $p_{X|AY}(x|a,y)$ for continuous X) and the set $S^{(t)}_{X|a}$ (which is determined by the function f). By definition $S^{(t)}_{X|a}$ has following properties: (1) for any fixed value of $a \in \mathcal{A}$, if $t \neq t'$, then $S^{(t)}_{X|a} \bigcap S^{(t')}_{X|a} = \emptyset$; (2) for any fixed value of $a \in \mathcal{A}$, $\bigcup_{t\in\mathcal{Y}} S^{(t)}_{X|a} = \mathcal{X}$. Condition (ii) says that the set $S^{(t)}_{X|a}$ and the conditional distribution $P_{X|AY}(x|a,y)$ are coupled in some specific way so that they happen to satisfy the specified equality.

In special cases when $X \perp\!\!\!\perp A \mid Y$, if f is a function of only X, condition (ii) would always hold true. In general situations, if there does not exist any subsets $K_a, K_{a'} \subseteq \mathcal{X}$ for different values of $a, a' \in \mathcal{A}$ such that $\Sigma_{x\in K_a} P_{X|AY}(x|a,y) = \Sigma_{x\in K_{a'}} P_{X|AY}(x|a',y)$, then condition (ii) can never hold true (i.e., we cannot find a deterministic function $f(A,X)$ that satisfies Equalized Odds). Generally speaking, in order to score a better classification accuracy, one would like to make $P_{\widetilde{Y}|A,X}(t|a,x)$ as close as possible to $P_{Y|A,X}(y|a,x)$, and if the set $S^{(t)}_{X|a}$ and $P_{X|AY}(x|a,y)$ are not strictly coupled, condition (ii) would be violated.

4.2 Classification with Randomized Prediction

In this section, we consider cases when randomized prediction is acceptable, namely, the classifier would output class labels with certain probabilities. We first derive the relation between positive rates (TPR and FPR) of binary classifiers before and after the post-processing step, i.e., $\widehat{Y}_{\text{opt}}$ (the unconstrainedly optimized classifier) and $\widetilde{Y}_{\text{post}}$ (the fair classifier derived by post-processing $\widehat{Y}_{\text{opt}}$), and show that under mild assumptions, one can always derive a non-trivial Equalized-Odds (on population level) $\widetilde{Y}_{\text{post}}$ via a post-processing step. Then, from the ROC feasible area perspective, we prove that post-processing approaches are actually equivalent to in-processing approaches but with additional "pseudo" constraints enforced. Therefore, using the same loss function, post-processing approaches can perform no better than in-processing approaches.

4.2.1 The Post-processing Step

The post-processing step of a predictor $\widehat{Y}$ (here we drop the subscript if without ambiguity) only utilizes the information in the joint distribution $(A, Y, \widehat{Y})$. A fair predictor $\widetilde{Y}_{\text{post}}$ derived via a post-processing step, for instance, the *shifted decision boundary* (Fish et al., 2016), the *derived predictor* (Hardt et al., 2016), or the *(monotonic) joint loss optimization* over *decoupled classifiers* (Dwork et al., 2018), is then fully specified by a (possibly randomized) function of $(A, \widehat{Y})$. This implies the conditional independence $\widetilde{Y}_{\text{post}} \perp\!\!\!\perp Y \mid A, \widehat{Y}$. Since we can denote the positive rates of $\widehat{Y}$ as $P_{\widehat{Y}|AY}(1|a,y)$[1], positive rates of $\widetilde{Y}$ (here we drop the subscript for readability) as $P_{\widetilde{Y}|AY}(1|a,y)$, the relation between positive rates of binary classifiers before and after a post-processing step would satisfy (for every $a \in \mathcal{A}, u, y \in \mathcal{Y}$):

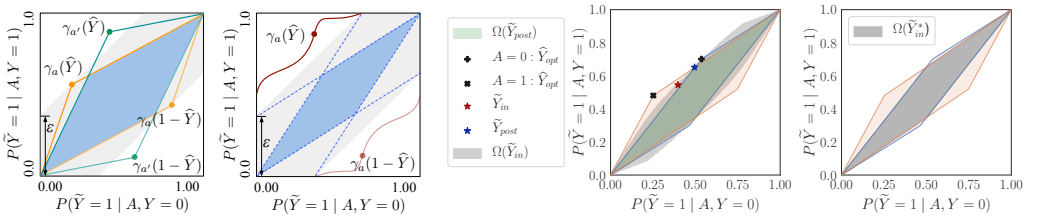
$$P_{\widetilde{Y}|AY}(1|a,y) = \Sigma_{u\in\mathcal{Y}}\, \beta^{(u)}_a P_{\widehat{Y}|AY}(u|a,y), \text{ where } \beta^{(u)}_a := P(\widetilde{Y} = 1 \mid A = a, \widehat{Y} = u). \tag{5}$$

Notice that Equation 5 is just a factorization of probability under the conditional independence (between $\widetilde{Y}_{\text{post}}$ and Y given A and $\widehat{Y}$). Therefore, post-processing an existing predictor boils down to optimizing parameters (for discrete A) or functions (for continuous A) $\beta^{(u)}_a$.

4.2.2 ROC Feasible Area

On the Receiver Operator Characteristic (ROC) plane, a two-dimensional plane with horizontal axis denoting FPR and vertical axis denoting TPR, the performance of any binary predictor $\widehat{Y}$ (not

[1]Recall that $P_{\widehat{Y}|AY}(u|a,y) = P(\widehat{Y} = u \mid A = a, Y = y)$. When $u = 1$, $P_{\widehat{Y}|AY}(1|a,y) = P(\widehat{Y} = 1 \mid A = a, Y = y)$ represents positive rates of $\widehat{Y}$; When $u = 0$, $P_{\widehat{Y}|AY}(0|a,y) = P(\widehat{Y} = 0 \mid A = a, Y = y)$ represents positive rates of $1 - \widehat{Y}$ (the classifier that flips the prediction of $\widehat{Y}$).



(a) Discrete protected feature (b) Continuous protected feature (c) Feasible areas for $\widetilde{Y}_{\text{in}}$, $\widetilde{Y}_{\text{post}}$ and $\widehat{Y}_{\text{opt}}$ (d) $\Omega(\widetilde{Y}^*_{\text{in}})$ with "pseudo" constraints

Figure 1: ROC feasible area illustrations. Panels (a)-(b): Attainability of Equalized Odds for binary classifiers with discrete or continuous protected feature. Panels (c)-(d): ROC feasible areas comparison between $\Omega(\widetilde{Y}_{\text{in}})$, $\Omega(\widetilde{Y}_{\text{post}})$, $\Omega(\widehat{Y}_{\text{opt}})$, and $\Omega(\widetilde{Y}^*_{\text{in}})$.

necessarily a fair one) with certain value of protected feature $A = a$ corresponds to a point $\gamma_a(\widehat{Y}) =$ (FPR, TPR) on the plane. Denote each coordinate according to the value of Y as $\gamma_{ay}(\widehat{Y})$:

$$\gamma_a(\widehat{Y}) = \big(\gamma_{a0}(\widehat{Y}), \gamma_{a1}(\widehat{Y})\big) := \big(P_{\widehat{Y}|AY}(1|a,0), P_{\widehat{Y}|AY}(1|a,1)\big). \tag{6}$$

Further denote the corresponding convex hull of $\widehat{Y}$ on the ROC plane as $\mathcal{C}_a(\widehat{Y})$ using vertices:

$$\mathcal{C}_a(\widehat{Y}) := \text{convhull}\big\{(0,0), \gamma_a(\widehat{Y}), \gamma_a(1-\widehat{Y}), (1,1)\big\}, \tag{7}$$

and then, as already stated in Hardt et al. (2016), the (FPR, TPR) pair corresponding to a post-processing predictor falls within (including the boundary of) $\mathcal{C}_a(\widehat{Y})$.

Definition 4.1. (**ROC feasible area**) The feasible area of a predictor $\Omega(\widehat{Y})$, specified by the hypothesis space of available predictors $\widehat{Y}$, is the set containing all attainable (FPR, TPR) pairs by the predictor on the ROC plane satisfying Equalized Odds.

In Hardt et al. (2016) it is proposed that the post-processing fair predictor can be derived by solving a linear programming problem on the ROC plane. However, it is not clearly stated whether or not such problem always has a non-trivial solution. Following Hardt et al. (2016), we analyze the relation between the (FPR, TPR) pair of predictors on the ROC plane and formally establish the existence of the non-trivial Equalized-Odds predictor. Actually as we shall see in Theorem 4.2, under mild assumptions, an Equalized-Odds predictor $\widetilde{Y}_{\text{post}}$ derived via post-processing $\widehat{Y}$ (a predictor optimized without fairness concern) always has non-empty ROC feasible area.

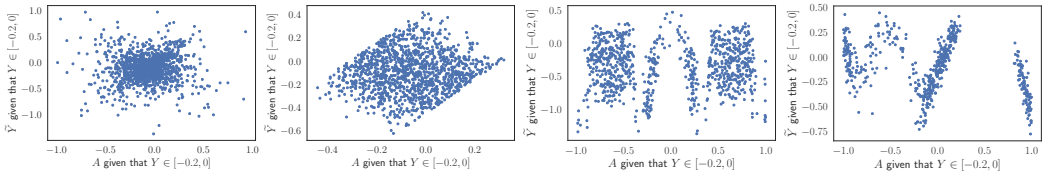
Theorem 4.2. ***(Attainability of Equalized Odds)***
Assume that the feature X is not independent from Y, and that $\widehat{Y}$ is a function of A and X. Then for binary classification, if $\widehat{Y}$ is a non-trivial predictor for Y, there is always at least one non-trivial (possibly randomized) predictor $\widetilde{Y}_{post}$ derived by post-processing $\widehat{Y}$ that can attain Equalized Odds:

$$\Omega(\widetilde{Y}_{post}) \neq \emptyset.$$

Here $\widetilde{Y}_{\text{post}}$ is a possibly randomized function of only A and $\widehat{Y}$, trading off TPR with FPR across groups with different value of protected feature. From the panels (a) and (b) of Figure 1 we can also see that $\Omega(\widetilde{Y}_{\text{post}})$, the ROC feasible area of $\widetilde{Y}_{\text{post}}$, is the intersection of $\Omega_a(\widehat{Y})$, indicating that although Equalized Odds is attained, the performance of $\widetilde{Y}_{\text{post}}$ is always worse than the weakest performance across different groups, which is obviously suboptimal.

4.2.3 Optimality of Performance among Fair Classifiers

In this subsection we discuss the optimality of performance of fair classifiers derived via different approaches. Considering the fact that recent efforts to impose Equalized Odds in the pre-processing manner (Madras et al., 2018; Zhao et al., 2020) approach the problem from a representation learning perspective, where the main focus is to learn fair representations that at the same time preserve sufficient information from the original data, we omit pre-processing approaches from the discussion and compare the performance of post-processing and in-processing fair classifiers.



(a) Linear w/ Laplace distribution (b) Linear w/ Uniform distribution (c) Nonlinear w/ Gaussian distribution (d) Nonlinear w/ Gaussian distribution

Figure 2: Illustration of unattainable Equalized Odds for regression tasks. Panel (a)-(b): Linear regression on the data generated with linear transformations and non-Gaussian distributed exogenous terms (following Laplace, Uniform distribution respectively). Panel (c)-(d): Nonlinear regression with a neural net regressor (Mary et al., 2019) on the data generated with nonlinear transformations and Gaussian exogenous terms. We can observe obvious dependencies between $\widetilde{Y}$ and A on a small interval of Y. This indicates the conditional dependency between $\widetilde{Y}$ and A given Y, i.e., the Equalized Odds is not achieved.

Admittedly, when only the information about joint distribution of $(A, Y, \widehat{Y})$ is available, post-processing is the best we can do. However, this is not the case when we have access to additional available features during training. For any predictor specified by parameters $\theta \in \Theta$, the derivation of the in-processing fair predictor $\widetilde{Y}_{\text{in}}$ and the unconstrained statistical optimal predictor $\widehat{Y}_{\text{opt}}$ take following forms respectively:

$$\begin{aligned} \min_{\theta\in\Theta} \quad & \mathbb{E}[l\left(f(A, X; \theta), Y\right)] \\ \text{s.t.} \quad & P_{\widetilde{Y}_{\text{in}}|AY}(t \mid a, y) = P_{\widetilde{Y}_{\text{in}}|Y}(t \mid y) \quad (8) \\ \text{where} \quad & \widetilde{Y}_{\text{in}} = f(A, X; \theta), \end{aligned} \qquad \begin{aligned} \min_{\theta\in\Theta} \quad & \mathbb{E}[l\left(f(A, X; \theta), Y\right)] \\ \text{where} \quad & \widehat{Y}_{\text{opt}} = f(A, X; \theta). \end{aligned} \quad (9)$$

It is natural to wonder, now that one can always directly solve for $\widetilde{Y}_{\text{in}}$ from Equation 8, how is it related to $\widetilde{Y}_{\text{post}}$, which is derived by post-processing the $\widehat{Y}_{\text{opt}}$ solved from Equation 9? Interestingly, although $\widetilde{Y}_{\text{in}}$ and $\widetilde{Y}_{\text{post}}$ are solved separately using different constrained optimization schemes, one can draw a connection between them by utilizing $\widehat{Y}_{\text{opt}}$ as a bridge and reason about the relation between their ROC feasible areas $\Omega(\widetilde{Y}_{\text{in}})$ and $\Omega(\widetilde{Y}_{\text{post}})$, as we summarize in the following theorem.

Theorem 4.3. ***(Equivalence between ROC feasible areas)***
Let $\Omega(\widetilde{Y}_{post})$ denote the ROC feasible area specified by the constraints enforced on $\widetilde{Y}_{post}$. Then $\Omega(\widetilde{Y}_{post})$ is identical to the ROC feasible area $\Omega(\widetilde{Y}^*_{in})$ that is specified by the following set of constraints:

(i) constraints enforced on $\widetilde{Y}_{in}$;

(ii) additional "pseudo" constraints: $\forall a \in \mathcal{A}$, $\beta^{(0)}_{a0} = \beta^{(0)}_{a1}$, $\beta^{(1)}_{a0} = \beta^{(1)}_{a1}$, where $\beta^{(u)}_{ay} = \Sigma_{x\in\mathcal{X}} P(\widetilde{Y}_{in} = 1 \mid A = a, X = x) P(X = x \mid A = a, Y = y, \widehat{Y}_{opt} = u)$.

As we can see from panels (c) and (d) of Figure 1, if the additional "pseudo" constraints are introduced when optimizing $\widetilde{Y}^*_{\text{in}}$, we have $\Omega(\widetilde{Y}_{\text{in}}) \supseteq \Omega(\widetilde{Y}_{\text{post}}) = \Omega(\widetilde{Y}^*_{\text{in}})$. The ROC feasible area is fully specified by the hypothesis class and the fairness constraint. Therefore, with the same objective function and fairness constraint, the fair classifier derived from an in-processing approach always outperforms the one derived from a post-processing approach. We can see that when we have access to additional features and choose a post-processing approach, we lose performance (compared to $\widetilde{Y}_{\text{in}}$) by unintentionally introducing "pseudo" constraints during optimization. These "pseudo" constraints actually offset the benefit of utilizing additional features (in the hope to score a better performance while remaining fair).

5 Experiments

In order to intuitively illustrate the claims, we provide numerical results for various settings. We first present the result for (linear non-Gaussian and nonlinear) regression tasks when Equalized Odds is not attained. We demonstrate the dependence between the prediction and the protected feature given true value of the target variable. Then for classification tasks we compare the performance of several

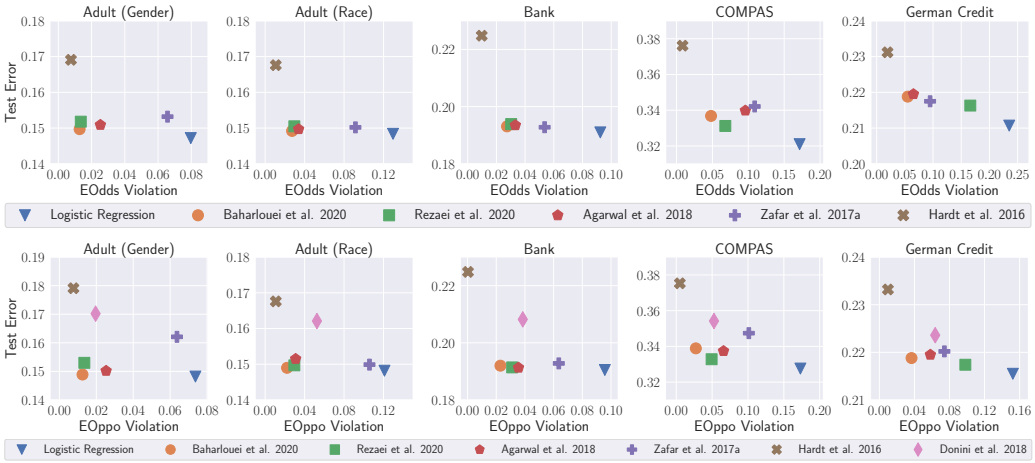


Figure 3: Results for classification with Equalized Odds/Equal Opportunity criterion.

existing methods in the literature on the *Adult*, *Bank*, *COMPAS*, and *German Credit* data sets. The detailed description of the data sets is available in the appendix.

5.1 Regression with Linear Non-Gaussian and Nonlinear Data

In Section 3.1 we showed the unattainability of Equalized Odds for regression with linear non-Gaussian data. Although a proof for similar results in nonlinear cases does not seem straightforward, as strongly suggested by our numerical illustrations, the unattainability of Equalized Odds persists in nonlinear regression cases. In Figure 2 we present scatter plots of $\widetilde{Y}$ versus A for Y in a small (compared to its support) interval, for linear non-Gaussian as well as nonlinear regression cases.

For linear cases, the data is generated as stated in Equation 4, with non-Gaussian distributed exogenous terms (E_X, E_H, and E_Y). We use linear regression with the *Equalized Correlations* constraint (Woodworth et al., 2017), a weaker notion of Equalized Odds for linearly correlated data, as the predictor. For nonlinear cases, the data is generated using a similar scheme but with nonlinear transformations (e.g., combinations of $\sin(\cdot), \log(\cdot)$, and polynomials) and Gaussian distributed exogenous terms. We use a neural net regressor with an Equalized Odds regularization term (Mary et al., 2019) to perform nonlinear fair regression. As we can see in Figure 2, for nonlinear regression tasks, Equalized Odds may not be attained even if every exogenous term is Gaussian distributed.

5.2 Fair Classification

In Figure 3, we compare the performance under Equalized Odds of multiple methods proposed in the literature. Hardt et al. (2016) propose a post-processing approach where the prediction is randomized to minimize violation of fairness; Zafar et al. (2017a) use a covariance proxy measure as the regularization term when optimizing classification accuracy; Agarwal et al. (2018) take the reductions approach and reduce fair classification into solving a sequence of cost-sensitive classification problems; Rezaei et al. (2020) minimize the worst-case log loss using an approximated regularization term; Baharlouei et al. (2020) propose to use Rényi correlation as the regularization term to account for nonlinear dependence between variables. To measure the violation of the fairness criterion, we use Equalized Odds (EOdds) violation, defined as $\max_{y\in\mathcal{Y}\ a,a'\in\mathcal{A}} \left|P_{\widetilde{Y}|AY}(1|a,y) - P_{\widetilde{Y}|AY}(1|a',y)\right|$. Following Agarwal et al. (2018), we pick 0.01 as the default violation bound that the EOdds violation does not exceed (if practically achievable for the method) during training. For each method we plot the testing accuracy versus the violation of Equalized Odds. Although a probabilistic classification model is used across each method (here is logistic regression), if an algorithm output the class label where the prediction likelihood is maximized, the prediction is in essence performed by a deterministic function of input features (e.g., Rezaei et al. (2020); Baharlouei et al. (2020)). As we have shown in Section 4.1, for classification with a deterministic function, in general cases the conditions specified in Theorem 4.1 are easily violated, i.e., Equalized Odds may not be

attained even if there is an unlimited amount of data. Therefore, although here we are considering finite data cases, we can still anticipate a lower level of fairness violation with a randomized prediction. This is validated by the numerical experiment: while the approach by Hardt et al. (2016) does not score the lowest test error, the violation of Equalized Odds is the lowest compared to other approaches. The benefit of introducing randomization can also be witnessed by the Pareto frontier presented in Agarwal et al. (2018), where the approach can potentially achieve any desired fairness-accuracy trade-off between that of the post-processing approach and that of the unconstrainedly optimized classifier[2]. In some scenarios people tend to only care about equal TPR (e.g., the rate of acceptance/admission) across groups, i.e., the Equal Opportunity (Hardt et al., 2016) notion of fairness. The related numerical result on real-world data sets is also presented.

6 Conclusion and Future Work

In this paper, we focus on the Equalized-Odds criterion and consider the attainability of fairness, and furthermore, if attainable, the optimality of the prediction performance under various settings. We first show that, for fair regression, one can only achieve Equalized Odds when certain conditions on the joint distribution of the features and the target variable are met. Then for classification tasks with deterministic classifiers, we give the condition under which Equalized Odds can hold true; we also show that under mild assumptions, one can always find a non-trivial Equalized-Odds (randomized) predictor, even with a continuous protected feature; in terms of the optimality of performance, one can always (if conditions permit) benefit from exploiting all available features during training. Future work would naturally consider nonlinear regression algorithms with randomized output and fairness guarantees, and the attainability of more fine-grained (compared to group fairness) criteria of fairness (e.g., individual fairness) as well as the *procedure fairness* in the fairness hierarchy.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69, 2018.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks. *ProPublica*, 2016.

Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *International Conference on Learning Representations*, 2020.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pp. 4349–4357, 2016.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

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

[2]In the approach proposed by Agarwal et al. (2018), the randomization can come in two folds: the first kind of randomization comes from picking a classifier from the distribution of multiple available classifiers; the second kind of randomization comes from the probabilistic prediction (if the hypothesis class contains probabilistic prediction models).

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, pp. 1436–1445, 2019.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pp. 309–315, 2019.

David Danks and Alex John London. Algorithmic bias in autonomous systems. In *IJCAI*, pp. 4691–4697, 2017.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pp. 119–133, 2018.

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 144–152. SIAM, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650. IEEE, 2011.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pp. 4382–4391, 2019.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 339–355. Springer, 2017.

Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian Ziebart. Fairness for robust log loss classification. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pp. 6414–6423, 2017.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2164–2173, 2019.

Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10–29, 2013.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pp. 1920–1953, 2017.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, pp. 3399–3409, 2019.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3929–3935, 2017.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional learning of fair representations. In *International Conference on Learning Representations*, 2020.

A APPENDIX

A.1 PROOF FOR THEOREM 3.1

To prove the unattainability of Equalized Odds in regression, we will need the following lemma, which provides a way to characterize conditional independence/dependence with conditional or joint distributions.

Lemma A.1. *Variables V_1 and V_2 are conditionally independent given variable V_3 if and only if there exist functions $h(v_1, v_3)$ and $g(v_2, v_3)$ such that*

$$p_{V_1,V_2|V_3}(v_1, v_2 \mid v_3) = h(v_1, v_3) \cdot g(v_2, v_3). \tag{10}$$

Proof. First, if V_1 and V_2 are conditionally independent given variable V_3, then Equation 10 holds:

$$p_{V_1,V_2|V_3}(v_1, v_2 \mid v_3) = p_{V_1|V_3}(v_1 \mid v_3) \cdot p_{V_2|V_3}(v_2 \mid v_3).$$

We then let $\tilde{h}(v_3) := \int h(v_1, v_3)dv_1$ and $\tilde{g}(v_3) := \int g(v_2, v_3)dv_2$. Take the integral of Equation 10 w.r.t. v_1 and v_2, we have:

$$p_{V_2|V_3}(v_2 \mid v_3) = \tilde{h}(v_3) \cdot g(v_2, v_3),$$
$$p_{V_1|V_3}(v_1 \mid v_3) = \tilde{g}(v_3) \cdot h(v_1, v_3),$$

respectively. Bearing in mind Equation 10, one can see that the product of the two equations above is

$$\begin{aligned} p_{V_2|V_3}(v_2 \mid v_3) \cdot p_{V_1|V_3}(v_1 \mid v_3) &= \tilde{h}(v_3) \cdot g(v_2, v_3) \cdot \tilde{g}(v_3) \cdot h(v_2, v_3) \\ &= \tilde{h}(v_3) \cdot \tilde{g}(v_3) \cdot p_{V_1,V_2|V_3}(v_1, v_2 \mid v_3). \end{aligned}$$

Take the integral of the equation above w.r.t. v_1 and v_2 gives $\tilde{h}(v_3) \cdot \tilde{g}(v_3) \equiv 1$. The above equation then reduces to

$$p_{V_2|V_3}(v_2 \mid v_3) \cdot p_{V_1|V_3}(v_1 \mid v_3) = p_{V_1,V_2|V_3}(v_1, v_2 \mid v_3).$$

That is, V_1 and V_2 are conditionally independent given V_3. □

Now we are ready to prove the unattainability of Equalized Odds in linear non-Gaussian regression:

Theorem. ***(Unattainability of Equalized Odds in the Linear Non-Gaussian Case)*** *Assume that feature X has a causal influence on Y, i.e., $c \neq 0$ in Equation 4, and that the protected feature A and Y are not independent, i.e., $qc + bd \neq 0$. Assume p_{E_X} and p_E are positive on $\mathbb{R}$. Let $f_1 := \log p_A$, $f_2 := \log p_{E_X}$, and $f_3 := \log p_E$. Further assume that f_2 and f_3 are third-order differentiable. Then if at most one of E_X and E is Gaussian, Z is always conditionally dependent on A given Y.*

Proof. According to Equation 4, we have

$$\begin{bmatrix} A \\ Z \\ Y \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \alpha + q\beta & \beta & 0 \\ qc + bd & c & 1 \end{bmatrix} \cdot \begin{bmatrix} A \\ E_X \\ E \end{bmatrix}. \tag{11}$$

The determinant of the above linear transformation is β, which relates the probability density function of the variables on the LHS and that of the variables on the RHS of the equation. Therefore, according to Equation 11, we can rewrite the joint probability density function by making use of the Jacobian determinant and factor the joint density into marginal density functions (A, E_X, E are mutually independent according to the data generating process). Further let

$$\tilde{\alpha} := \frac{\alpha + q\beta}{\beta}, \;\; \tilde{r} := bd - \frac{c\alpha}{\beta}, \;\; \text{and} \;\; \tilde{c} := \frac{c}{\beta}. \tag{12}$$

Then we have $E_X = \frac{1}{\beta}Z - \tilde{\alpha}A$, $E = Y - \tilde{r}A - \tilde{c}Z$, and

$$\begin{aligned} p_{A,Z,Y}(a, z, y) &= p_{A,E_X,E}(a, e_x, e)/|\beta| \\ &= \frac{1}{|\beta|} p_A(a) p_{E_X}(e_x) p_E(e) \\ &= \frac{1}{|\beta|} p_A(a) p_{E_X}(\frac{1}{\beta}z - \tilde{\alpha}a) p_E(y - \tilde{r}a - \tilde{c}z). \end{aligned}$$

On its support, the log-density can be written as

$$\begin{aligned} J &:= \log p_{A,Z,Y}(a,z,y) \\ &= \log p_A(a) + \log p_{E_X}(\frac{1}{\beta}z - \tilde{\alpha}a) + \log p_E(y - \tilde{r}a - \tilde{c}z) - log|\beta| \\ &= f_1(a) + f_2(\frac{1}{\beta}z - \tilde{\alpha}a) + f_3(y - \tilde{r}a - \tilde{c}z) - \log|\beta|. \end{aligned} \tag{13}$$

According to Lemma A.1, $A \perp\!\!\!\perp Z \mid Y$ if and only if $p_{A,Z|Y}(a, z \mid y)$ is a product of a function of a and y and a function of z and y. $p_{A,Z,Y}(a,z,y)$ is further a product of the above function and a function of only y. This property, under the conditions in Theorem 3.1, is equivalent to the constraint

$$\frac{\partial^2 J}{\partial A \partial Z} \equiv 0. \tag{14}$$

According to Equation 13, we have

$$\begin{aligned} \frac{\partial J}{\partial z} &= \frac{1}{\beta} \cdot f_2'(\frac{1}{\beta}z - \tilde{\alpha}a) - \tilde{c} \cdot f_3'(y - \tilde{r}a - \tilde{c}z) \\ \Rightarrow \frac{\partial^2 J}{\partial a \partial z} &= -\frac{\tilde{\alpha}}{\beta} \cdot f_2''(\frac{1}{\beta}z - \tilde{\alpha}a) + \tilde{r}\tilde{c} \cdot f_3''(y - \tilde{r}a - \tilde{c}z). \end{aligned} \tag{15}$$

Combining Equations 14 and 15 gives

$$\tilde{r}\tilde{c} \cdot f_3''(y - \tilde{r}a - \tilde{c}z) = \frac{\tilde{\alpha}}{\beta} \cdot f_2''(\frac{1}{\beta}z - \tilde{\alpha}a). \tag{16}$$

Further taking the partial derivative of both sides of the above equation w.r.t. y yields

$$\tilde{r}\tilde{c} \cdot f_3'''(y - \tilde{r}a - \tilde{c}z) \equiv 0. \tag{17}$$

There are three possible situations where the above equation holds:

(i) $\tilde{c} = 0$, which is equivalent to $c = 0$ and contradicts with the theorem assumption.

(ii) $\tilde{r} = 0$. Then according to Equation 16, we have $\frac{\tilde{\alpha}}{\beta} \cdot f_2''(\frac{1}{\beta}z - \tilde{\alpha}a) \equiv 0$, implies either $\tilde{\alpha} = 0$ or $f_2''(\frac{1}{\beta}z - \tilde{\alpha}a) \equiv 0$. If the latter is the case, then f_2 is a linear function and, accordingly, $\exp(f_2)$ is not integrable and does not correspond to any valid density function. If the former is true, i.e., $\tilde{\alpha} = 0$, then according to Equation 12, we have $\alpha = -q\beta$, which further implies $\tilde{r} = bd - \frac{c\alpha}{\beta} = bd + qc$. Therefore, in this situation, $bd + qc = 0$, which again contradicts with the theorem assumption.

(iii) $f_3'''(y - \tilde{r}a - \tilde{c}z) \equiv 0$. That is, f_3 is a quadratic function with a nonzero coefficient for the quadratic term (otherwise f_3 does not correspond to the logarithm of any valid density function). Thus E follows a Gaussian distribution.

Only situation (iii) is possible, i.e., $\tilde{r}\tilde{c} \neq 0$ and E follows a Gaussian distribution. This further tells us that the RHS of Equation 16 is a nonzero constant. Hence f_2 is a quadratic function and E_X also follows a Gaussian distribution. Therefore if $A \perp\!\!\!\perp Z \mid Y$ were to be true, then E_X and E are both Gaussian. Its contrapositive gives the conclusion of this theorem. □

Corollary. *Suppose that both E_X and E are Gaussian, with variances $\sigma^2_{E_X}$ and σ^2_E, respectively. (The protected feature A is not necessarily Gaussian.) Then $Z \perp\!\!\!\perp A \mid Y$ if and only if*

$$\frac{\alpha}{\beta} = \frac{bdc \cdot \sigma^2_{E_X} - q \cdot \sigma^2_E}{c^2 \cdot \sigma^2_{E_X} + \sigma^2_E}. \tag{18}$$

Proof. Under the condition that E_X and E are Gaussian, their log-density functions are third-order differentiable. Then according to the proof of Theorem 3.1, the Equalized Odds condition $A \perp\!\!\!\perp Z \mid Y$ is equivalent to Equation 16, which, together with Equation 12 as well as the fact that $f_2'' = \frac{1}{\sigma^2_{E_X}}$ and $f_3'' = \frac{1}{\sigma^2_E}$, yields Equation 18. □

A.2 PROOF FOR THEOREM 4.1

Theorem. *Assume that the protected feature A and Y are dependent and that their joint probability $P(A, Y)$ (for discrete A) or joint probability density $p(A, Y)$ (for continuous A) is positive for every combination of possible values of A and Y. Further assume that Y is not fully determined by A, and that there are additional features X that are not independent of Y. Let the output of the classifier $\widetilde{Y}$ be a deterministic function $f : \mathcal{A} \times \mathcal{X} \to \mathcal{Y}$. Let $S_A^{(t)} := \{a \mid \exists x \in \mathcal{X} \text{ s.t. } f(a, x) = t\}$, and $S_{X|a}^{(t)} := \{x \mid f(a, x) = t\}$. Equalized Odds holds true if and only if the following two conditions are satisfied:*

(i) $\forall t \in \mathcal{Y} :\ S_A^{(t)} = \mathcal{A}$,

(ii) $\forall t \in \mathcal{Y},\ \forall a, a' \in \mathcal{A}, a \neq a' :$
$$\sum_{x \in S_{X|a}^{(t)}} P(X = x \mid A = a, Y = y) = \sum_{x \in S_{X|a'}^{(t)}} P(X = x \mid A = a', Y = y).$$

Proof. We begin by considering the case when A and X are discrete (for the purpose of readability). The Equalized Odds criterion can be written in terms of the conditional probabilities:

$$\forall a \in \mathcal{A},\ t, y \in \mathcal{Y} :\ P(\widetilde{Y} = t \mid A = a, Y = y) = P(\widetilde{Y} = t \mid Y = y). \tag{19}$$

Expand the LHS of Equation 19:

$$P(\widetilde{Y} = t \mid A = a, Y = y) = \sum_{x \in \mathcal{X}} P(\widetilde{Y} = t \mid A = a, X = x, Y = y) P(X = x \mid A = a, Y = y),$$

and bear in mind that $\widetilde{Y} := f(A, X)$ is a deterministic function of (A, X), we have:

$$P\big(f(A, X) = t \mid A = a, X = x, Y = y\big) = P\big(f(A, X) = t \mid A = a, X = x\big) \in \{0, 1\}. \tag{20}$$

From Equation 20 we can see that the conditional probability $P(X = x \mid A = a, Y = y)$ can contribute to the summation only when $f(a, x) = t$. We can rewrite the LHS of Equation 19:

$$P(\widetilde{Y} = t \mid A = a, Y = y) = \sum_{x \in S_{X|a}^{(t)}} P(X = x \mid A = a, Y = y) := Q^{(t)}(a, y).$$

Similarly, for the RHS of Equation 19, we have:

$$\begin{aligned} P(\widetilde{Y} = t \mid Y = y) &= \sum_{a \in \mathcal{A}} \sum_{x \in \mathcal{X}} P(\widetilde{Y} = t \mid A = a, X = x, Y = y) P(X = x, A = a \mid Y = y) \\ &= \sum_{a \in S_A^{(t)}} \sum_{x \in S_{X|a}^{(t)}} P(X = x \mid A = a, Y = y) P(A = a \mid Y = y) \\ &= \sum_{a \in S_A^{(t)}} Q^{(t)}(a, y) P(A = a \mid Y = y). \end{aligned}$$

Since Equalized Odds holds true if and only if Equation 19 holds true, then the LHS of the equation does not involve a (as is the case for the RHS), i.e., $Q^{(t)}(a, y)$ does not change with a. Then Equation 19 becomes:

$$Q^{(t)}(a, y) = \sum_{a \in S_A^{(t)}} Q^{(t)}(a, y) P(A = a \mid Y = y) = Q^{(t)}(a, y) \sum_{a \in S_A^{(t)}} P(A = a \mid Y = y),$$

which gives the condition (i) that $S_A^{(t)}$ contains all possible values of A, i.e., $\mathcal{A} = S_A^{(t)}$ (otherwise $\Sigma_{a \in S_A^{(t)}} P(A = a \mid Y = y) < 1$). Since $Q^{(t)}(a, y)$ does not change with a, we have:

$$\forall a, a' \in \mathcal{A}, a \neq a' : \sum_{x \in S_{X|a}^{(t)}} P(X = x \mid A = a, Y = y) = \sum_{x \in S_{X|a'}^{(t)}} P(X = x \mid A = a', Y = y),$$

which gives the condition (ii). Therefore, Equalized Odds implies conditions (i) and (ii). On the other hand, it is easy to see that when conditions (i) and (ii) are satisfied, Equation 19 holds true, i.e., Equalized Odds holds true.

When A and X are continuous, one can replace the summation with integration accordingly. □

A.3 PROOF FOR THEOREM 4.2

Theorem. ***(Attainability of Equalized Odds)***
Assume that the feature X is not independent from Y, and that $\widehat{Y}$ is a function of A and X. Then for binary classification, if $\widehat{Y}$ is a non-trivial predictor for Y, there is always at least one non-trivial predictor $\widetilde{Y}_{post}$ derived by post-processing $\widehat{Y}$ that can attain Equalized Odds, i.e.,

$$\Omega(\widetilde{Y}_{post}) \neq \emptyset.$$

Proof. Since $\widehat{Y}$ is a function of (A, X) and $X \not\perp\!\!\!\perp Y$, $\widehat{Y}$ is not conditionally independent from Y given protected feature A. Furthermore, since $\widehat{Y}$ is a non-trivial estimator of the binary target Y, there exists a positive constant $\epsilon > 0$, such that:

$$\big|P(\widehat{Y} = 1 \mid A = a, Y = 1) - P(\widehat{Y} = 1 \mid A = a, Y = 0)\big| \geq \epsilon, \ \forall a \in \mathcal{A}. \tag{21}$$

Equation 21 implies that for each value of A, the corresponding true positive rate of the non-trivial predictor is always strictly larger than its false positive rate[3]. As illustrated in panels (a) and (b) of Figure 1, the (FPR, TPR) pair of the predictor $\widehat{Y}$ when $A = a$, i.e., the point $\gamma_a(\widehat{Y})$ on ROC plane, will never fall in the gray shaded area, and its coordinates are bounded away from the diagonal by at least ϵ. Therefore, the intersection of all $\mathcal{C}_a(\widehat{Y})$ would always form a parallelogram with non-empty area, which corresponds to attainable non-trivial post-processing fair predictors $\widetilde{Y}_{\text{post}}$. □

A.4 PROOF FOR THEOREM 4.3

Theorem. ***(Equivalence between ROC feasible areas)***
Let $\Omega(\widetilde{Y}_{post})$ denote the ROC feasible area specified by the constraints enforced on $\widetilde{Y}_{post}$. Then $\Omega(\widetilde{Y}_{post})$ is identical to the ROC feasible area $\Omega(\widetilde{Y}^*_{in})$ that is specified by the following set of constraints:

(i) constraints enforced on $\widetilde{Y}_{in}$;

(ii) additional "pseudo" constraints: $\forall a \in \mathcal{A}$, $\beta^{(0)}_{a0} = \beta^{(0)}_{a1}$, $\beta^{(1)}_{a0} = \beta^{(1)}_{a1}$, where $\beta^{(u)}_{ay} = \Sigma_{x\in\mathcal{X}} P(\widetilde{Y}_{in} = 1 \mid A = a, X = x)P(X = x \mid A = a, Y = y, \widehat{Y}_{opt} = u)$.

Proof. Since the post-processing predictor $\widetilde{Y}_{\text{post}}$ is derived by optimizing over parameters or functions (of A) $\beta^{(u)}_a$. Therefore, considering the fact that $P_{\widetilde{Y}_{\text{post}}|AY}(1|a, y) = \gamma_{ay}(\widetilde{Y}_{\text{post}})$, $P_{\widehat{Y}_{\text{opt}}|AY}(1|a, y) = \gamma_{ay}(\widehat{Y}_{\text{opt}})$, we have the relation between $\gamma_{ay}(\widetilde{Y}_{\text{post}})$ and $\gamma_{ay}(\widehat{Y}_{\text{opt}})$:

$$\begin{aligned}
\gamma_{ay}(\widetilde{Y}_{\text{post}}) &= \beta^{(0)}_a \, \gamma_{ay}(1 - \widehat{Y}_{\text{opt}}) + \beta^{(1)}_a \, \gamma_{ay}(\widehat{Y}_{\text{opt}}), \\
\beta^{(0)}_a &= P(\widetilde{Y}_{\text{post}} = 1 \mid A = a, \widehat{Y}_{\text{opt}} = 0), \\
\beta^{(1)}_a &= P(\widetilde{Y}_{\text{post}} = 1 \mid A = a, \widehat{Y}_{\text{opt}} = 1).
\end{aligned} \tag{22}$$

Similarly, consider the relation between positive rates of $\widetilde{Y}_{\text{in}}$ and those of $\widehat{Y}_{\text{opt}}$, i.e., $P_{\widetilde{Y}_{\text{in}}|AY}(1|a, y)$ and $P_{\widehat{Y}_{\text{opt}}|AY}(1|a, y)$, by factorizing $P_{\widetilde{Y}_{\text{in}}|AY}(1|a, y)$ over X and $\widehat{Y}_{\text{opt}}$:

$$P_{\widetilde{Y}_{\text{in}}|AY}(1|a, y) = \sum_{u\in\mathcal{Y}} \big[\sum_{x\in\mathcal{X}} P_{\widetilde{Y}_{\text{in}}|AX}(1|a, x) P_{X|AY\widehat{Y}_{\text{opt}}}(x|a, y, u) \big] P_{\widehat{Y}_{\text{opt}}|AY}(u|a, y). \tag{23}$$

Therefore, we have the relation between $\gamma_{ay}(\widetilde{Y}_{\text{in}})$ and $\gamma_{ay}(\widehat{Y}_{\text{opt}})$:

$$\begin{aligned}
\gamma_{ay}(\widetilde{Y}_{\text{in}}) &= \beta^{(0)}_{ay} \, \gamma_{ay}(1 - \widehat{Y}_{\text{opt}}) + \beta^{(1)}_{ay} \, \gamma_{ay}(\widehat{Y}_{\text{opt}}), \\
\beta^{(0)}_{ay} &= \Sigma_{x\in\mathcal{X}} P(\widetilde{Y}_{\text{in}} = 1 \mid A = a, X = x)P(X = x \mid A = a, Y = y, \widehat{Y}_{\text{opt}} = 0), \\
\beta^{(1)}_{ay} &= \Sigma_{x\in\mathcal{X}} P(\widetilde{Y}_{\text{in}} = 1 \mid A = a, X = x)P(X = x \mid A = a, Y = y, \widehat{Y}_{\text{opt}} = 1).
\end{aligned} \tag{24}$$

[3]If the TPR of the predictor is always smaller than its FPR, one can simply flip the prediction (since the target is binary) and then Equation 21 holds true.

If there is more than one variable in X in Equation 24, one can expand the summation if needed; if some variables are continuous, one may also substitute the summation with integration accordingly.

From Equation 24, $\beta_{ay}^{(0)}$ and $\beta_{ay}^{(1)}$ depend on the value of Y:

$$\begin{aligned}\beta_{ay}^{(0)} &= P(\widetilde{Y}_{\text{in}} = 1 \mid A = a, Y = y, \widehat{Y}_{\text{opt}} = 1),\\ \beta_{ay}^{(0)} &= P(\widetilde{Y}_{\text{in}} = 1 \mid A = a, Y = y, \widehat{Y}_{\text{opt}} = 0).\end{aligned} \tag{25}$$

Apart from Equalized Odds constraints (which are shared by $\widetilde{Y}_{\text{in}}$ and $\widetilde{Y}_{\text{post}}$), when enforcing additional "pseudo" constraints $\beta_{a0}^{(0)} = \beta_{a1}^{(0)}$ and $\beta_{a0}^{(1)} = \beta_{a1}^{(1)}$, conditional independence $\widetilde{Y}_{\text{in}} \perp\!\!\!\perp Y \mid A, \widehat{Y}_{\text{opt}}$ is enforced, making $\beta_{ay}^{(0)}$ and $\beta_{ay}^{(0)}$ no longer depend on Y. This is exactly the inherent constraint $\widetilde{Y}_{\text{post}}$ satisfies. Therefore the stated equivalence between ROC feasible areas $\Omega(\widetilde{Y}_{\text{post}})$ (specified by the constraints enforced on $\widetilde{Y}_{\text{post}}$) and $\Omega(\widetilde{Y}_{\text{in}}^*)$ (specified by the constraints enforced on $\widetilde{Y}_{\text{in}}$ together with the additional "pseudo" constraints) hold true. □

A.5 Description of the Data Sets

(1) **Adult**[4]: The UCI Adult data set contains 14 features for 45,222 individuals (32,561 samples for training and 12,661 samples for testing). The census information includes gender, marital status, education, capital gain, etc. The classification task is to predict whether a person's annual income exceeds 50,000 USD. We use the provided testing set for evaluations and present the result with gender and race (consider white and black people only) set as the protected feature respectively.

(2) **Bank**[5]: The UCI Bank Marketing data set is related with marketing campaigns of a banking institution, containing 16 features of 45,211 individuals. The assigned classification task is to predict if a client will subscribe (yes/no) to a term deposit. The original data set is very unbalanced with only 4,667 positives out of 45,211 samples. Therefore, we combine "yes" points with randomly subsampled "no" points and perform experiments on the down-sampled data set with 10,000 data points. The protected feature is the marital status of the client.

(3) **COMPAS** (Angwin et al., 2016): The COMPAS data set contains records of over 11,000 defendants from Broward County, Florida, whose risk (of recidivism) was assessed using the COMPAS tool. Each record contains multiple features of the defendant, including demographic information, prior convictions, degree of charge, and the ground truth for recidivism within two years. Following Zafar et al. (2017a); Nabi & Shpitser (2018), we limit our attention to the subset consisting of African-Americans and Caucasians defendants. The features we use include age, gender, race, number of priors, and degree of charges. The task is to predict the recidivism of the defendant and we choose race as the protected feature.

(4) **German Credit**[6]: The UCI German Credit data contains 20 features (7 numerical, 13 categorical) describing the social and economical status of 1,000 customers. The prediction task is to classify people as good or bad credit risks. We use the provided numerical version of the data and choose gender as the protected feature.

A.6 Additional discussion

For classification, while randomization can ensure group level of fairness, there is still some inherent shortcoming of the criterion that we should pay attention to. For example, in the FICO case study in Hardt et al. (2016), for a specific client from certain demographic group, the decision of approve/deny the loan actually comes in two folds: if his/her credit score is above (below) the upper (lower) threshold, the bank approve (deny) the application for sure; if the score falls in the interval between two thresholds, the bank would flip a coin to make a decision. Then we can imagine the following situation when a client whose credit score falls within the interval between the upper and

[4] http://archive.ics.uci.edu/ml/datasets/Adult

[5] https://archive.ics.uci.edu/ml/datasets/bank+marketing

[6] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

lower thresholds goes to a bank to apply a loan. He/she can ask (if conditions permit) the bank to run the model multiple times until the decision is approval. This would make the randomization that was built into the system for the sake of fairness no longer effective, and the system in essence only has one fixed threshold (i.e., the original lower threshold).