Carnegie Mellon University

DIETRICH COLLEGE OF HUMANITIES AND SOCIAL SCIENCES

Master Thesis

Submitted in Fulfillment of the Requirements for the Degree

Master of Science in Logic Computation and Methodology

Title: Think Twice Before Enforcing A Notion: A Reflection and a Case Study on Fairness in Machine Learning
Presented by: Zeyu Tang
Accepted by: Department of Philosophy

Name: [redacted] Zh[redacted] (Sep 8, 2021 16:59 EDT) Date: Sep 8, 2021

Name: [redacted] Alex London (Sep 9, 2021 09:37 EDT) Date: Sep 9, 2021

Name: [redacted] Hod aHo[redacted] (Sep 11, 2021 2:03 EDT) Date: Sep 11, 2021

Name: [redacted] Date: Sep 10, 2021

Approved by the Committee on Graduate Degrees

[redacted] 10/25/21

(Dean) Date

Think Twice Before Enforcing A Notion: A Reflection and a Case Study on Fairness in Machine Learning

Zeyu Tang

August, 2021

Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Prof. Kun Zhang (Chair)
Prof. Hoda Heidari
Prof. Yang Liu
Prof. Alex London

Submitted in fulfillment for the requirement of
Master of Science in Logic, Computation, and Methodology.

Abstract

The thesis presents a reflection and a case study of fairness notions in machine learning. I review commonly used fairness notions and reflect on the subtleties with respect to the role played by causality in fairness analysis. Then focusing on the Equalized Odds notion of fairness, I consider the theoretical attainability of Equalized Odds and, furthermore, if it is attainable, the optimality of the prediction performance under various settings. In particular, for prediction performed by a deterministic function of input features, I give the conditions under which Equalized Odds can hold true; if the stochastic prediction is acceptable, I show that under mild assumptions, fair predictors can always be derived. For classification tasks, I further prove that compared to enforcing fairness by post-processing, one can further benefit from exploiting all available features during training and get potentially better prediction performance while remaining fair.

Contents

Chapter 1

Introduction

With the widespread of the utilization of machine learning models in our daily life, researchers have been thinking about the potential social consequences of the prediction/decision made by algorithms. 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 [75]; the gender discrimination in job-related ads push [22]; stereotypes associated with gender in word embeddings [9]; the bias against certain ethnic groups in the assessment of recidivism risk [3, 8]; the violation of antidiscriminatin law (e.g., Title VII of the 1964 Civil Rights Act) emerged in data mining [6].

In the effort of enforcing fairness in machine learning, various notions as well as techniques to regulate discrimination under different scenarios have been proposed in the literature. There are multiple different perspectives of fairness analysis: in terms of the type of relation between variables that is encoded in the fairness criterion, there are associative notions of fairness that characterize correlation or dependence between variables (e.g., *Demographic Parity* [25], *Equalized Odds* [32], *Predictive Rate Parity* [23, 14, 84]) and causal notions of fairness that characterize causal relation between variables (e.g., *Counterfactual Fairness* [48], *No Unresolved Discrimination* [42], *Path-specific Counterfactual Fairness* [12, 83]); in terms of the scope of application, there are group-level fairness notions (e.g., *Equalized Odds* [32], *Fairness on Average Causal Effect* [41], *Equality of Effort* [36]) and individual-level fairness notions (e.g., *Individual Fairness* [25], *Counterfactual Fairness* [48], *Individual Fairness on Hindsight* [31]); in terms of the technical detail with respect to how discrimination is eliminated or suppressed, there are *pre-processing* approaches [10, 25, 86, 87, 51, 17, 99], *in-processing* approaches [40, 61, 84, 85, 24, 72, 53, 5, 63], and *post-processing* approaches [32, 28, 26]; in terms of the time span within which fairness is considered, the literature also includes fairness analysis in the dynamical setting [49, 33, 35, 96, 18, 98, 80, 34] other than considering fairness merely with respect to a snapshot of reality.

While there are explications with respect to available choices to quantify discrimination and enforce fairness in recent survey papers [64, 20, 50, 15, 57, 11, 52, 97, 54] as well as an investigation into public attitudes towards different notions [67], it is not well addressed whether or not a given notion of fairness can always be attained, even if with unlimited amount of data. The attainability of a fairness notion, 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 that has important practical implications. It characterizes a fundemantally different kind of violation of fairness compared to the empirical error bound of discrimination in finite-sample cases. If we deploy a "fair" prediction system in practice whose output is actually biased (since fairness is in fact not attainable with the chosen prediction scheme), the

discrimination would be easily neglected and hard to eliminate.

My thesis proceeds as follows: in Chapter 2 I introduce the notation conventions and present a brief introduction to causal modeling; in Chapter 3 I review commonly used fairness notions and reflect on the role played by causality in fairness analysis; in Chapter 4 I extensively discuss the *Equalized Odds* notion in terms of the theoretical attainability and, if attainable, the optimality of prediction in various settings; I summarize with conclusion remarks and future works in Chapter 5.

Chapter 2

Preliminaries

In this chapter, I first present the notation conventions used throughout this thesis in Section 2.1; and then, I present a brief introduction to causal reasoning in Section 2.2.

2.1 Setup and Notation

I use uppercase letters to refer to variables, lowercase letters to refer to specific values that variables can take, and calligraphic letters to refer to domains of value. For instance, I denote the protected feature by A (which may take values a or a') with domain of value $\mathcal{A}$, additional (observable) features(s) by X, with domain of value $\mathcal{X}$, ground truth (label) variable by Y and its predictor by $\widehat{Y}$, with domain of value $\mathcal{Y}$. Throughout the paper, without loss of generality I assume that there is only one protected feature and one ground truth variable for the purpose of simplifying notation. However, since the protected feature (e.g., race, sex, ratio of ethnic groups within community) and the ground truth variable (e.g., recidivism, annual income) can be discrete or continuous, I do not explicitly assume discreteness of the corresponding variables.[1] Fairness is also considered in the dynamical setting [49, 33, 35, 96, 18, 98, 80, 34] where there is a temporal axis for (some) variables. Considering the fact that the technical nuances involved in the quantification of discrimination (for each time step or in average) does not change the fairness we would like to enforce in terms of the notions themselves, without further clarification I do not include the temporal axis into discussion.

2.2 Causal Modeling

Since I will review commonly used causal notions of fairness and discuss subtleties regarding the role played by causality in fairness analysis, I give a brief introduction to causal modeling and inference in this section. Another field in the causality study is causal discovery where the primary goal is to recover the causal relations among variables from the data [73, 74, 13, 68, 90, 91, 89]. Causal discovery is not directly related to characterization of fairness in machine learning and therefore is not reviewed in this thesis. Readers that are already familiar with the related topics can feel free to skip the content.

[1] There might be some technical difficulties for certain fairness notions to be able to apply to different types of variables. However, this will not impede us from discussing and reflecting on the intuitions and insights behind fairness notions.

2.2.1 Definition and Representation of Causality

For two random variables X and Y, we say that X is a *cause* of Y if there is a change of distribution for Y when we apply different *interventions* on X [73, 60]. We can represent a causal model with a tuple $(U, V, \mathbf{F})$ such that:

(1) V is a set of observed variables involved in the system of interest;
(2) U is a set of exogenous variables that we cannot directly observe but contains the background information representing all causes of V and jointly follows a distribution $P(U)$;
(3) $\mathbf{F}$ is a set of functions (also known as structural equations) $\{f_1, f_2, \ldots, f_n\}$ where each f_i corresponds to one variable $V_i \in V$ and is a mapping $U \cup V \to V \setminus \{V_i\}$.

The triplet $(U, V, \mathbf{F})$ is known as the structural causal model (SCM). We can also capture causal relations among variables via a directed acyclic graph (DAG) $\mathcal{G}$, where nodes (vertices) represent variables and edges represent functional relations among variables.[2]

2.2.2 Interventions and Counterfactuals

Following Pearl [60], I use the $do(\cdot)$ operator to denote an intervention, which is a manipulation of the model such that the value of a variable (or a set of variables) is set to one specific value regardless of the corresponding structural equation(s). For example, the distribution of Y under the intervention $do(X = x)$ where $X \subseteq V$, is denoted by $P\big(Y \mid do(X = x)\big)$, which reads "what is the distribution of Y if we were to force $X = x$ in the population (regardless of the value X takes originally)". The full knowledge about the structural equations $\mathbf{F}$ is a rather strong assumption, but it also allows us to infer counterfactual quantities. For example, let $O, X \subseteq V$, with an observation $O = o$ the counterfactual distribution of Y if X had taken value x is denoted by $P\big(Y_{X \leftarrow x}(U) \mid O = o\big)$, which reads "what is the distribution of Y had X been set to x given that we actually observe $O = o$". The inference of the counterfactual quantity $P\big(Y_{X \leftarrow x}(U) \mid O = o\big)$ involves a three-step procedure (as explained in more detail in Pearl [60]):

(1) Abduction: for a given prior on U, compute the posterior distribution of U given the observation $O = o$;
(2) Action: substitute the structural equation that determines the value of X with the intervention $X = x$ and get modified set of structural equations $\mathbf{F}_{\text{modify}}$;
(3) Prediction: compute the distribution of Y using $\mathbf{F}_{\text{modify}}$ and the posterior $P(U \mid O = o)$.

The identifiability of various causal quantities has been extensively studied [77, 4, 37, 69, 70, 71].

[2]The causal graphs discussed in this thesis are limited to DAGs, and causal models represented by cyclic graphs are beyond the scope of the discussion.

Chapter 3

A Review and Reflection on Fairness Notions

In this chapter, when introducing commonly used fairness notions in Section 3.1, I unify the notations for consistency while keeping the semantics identical to original works when they are proposed in the literature; and then, I present a reflection on the role played by causality in fairness analysis in Section 3.2.

3.1 Notions of Fairness

In this section, I introduce commonly used fairness notions proposed in the literature. When presenting the definition of various fairness notions, I unify the notations while keeping the semantics identical to the original works.

3.1.1 Demographic Parity

Demographic Parity, also known as *Statistical Parity*, is one of the earliest fairness notions proposed in the literature [10, 25, 86, 27]. In the context of binary classification ($\mathcal{Y} = \{0, 1\}$), *Demographic Parity* requires that the ratio of positive decisions among different groups equals to each other:

$$\forall a, a' \in \mathcal{A} : \; P(\widehat{Y} = 1 \mid A = a) = P(\widehat{Y} = 1 \mid A = a'). \tag{3.1}$$

In general contexts, *Demographic Parity* is characterized via the independence between the prediction $\widehat{Y}$ and the protected feature A:

Definition 3.1. Demographic Parity We say that a predictor $\widehat{Y}$ is fair in terms of *Demographic Parity* with respect to the protected feature A, if $\widehat{Y}$ is independent from A, i.e., $\widehat{Y} \perp\!\!\!\perp A$.

While it is intuitive to characterize fairness through the aforementioned independence, the notion has significant drawbacks [25]. For instance, whenever the ground truth Y and the protected feature A are dependent, *Demographic Parity* would rule out the perfect prediction $\widehat{Y} = Y$, which would inevitably introduce substantial loss of utility.

3.1.2 Equalized Odds

In light of the limitation of *Demographic Parity*, Hardt et al. (2016) [32] propose *Equalized Odds* notion of fairness. In the context of binary classification, *Equalized Odds* requires that the True Positive Rate

(TPR) and False Positive Rate (FPR) of each group match the population TPR and FPR respectively:

$$\forall a \in \mathcal{A}, y \in \{0,1\}: \ P(\widehat{Y}=1 \mid A=a, Y=y) = P(\widehat{Y}=1 \mid Y=y). \tag{3.2}$$

In general contexts, this notion is characterized by stating the conditional independence between the prediction $\widehat{Y}$ and the protected feature A given the ground truth of the target Y:

Definition 3.2. Equalized Odds We say that a predictor $\widehat{Y}$ is fair in terms of *Equalized Odds* with respect to the protected feature A and the outcome Y, if $\widehat{Y}$ is conditionally independent from A given Y, i.e., $\widehat{Y} \perp\!\!\!\perp A \mid Y$.

The intuition behind this group-level fairness notion is that, once we know the true value of the target (in the hypothetical ideal world), the additional information of the value of the protected feature should not further alter our prediction results.

3.1.3 Predictive Rate Parity

First proposed by Dieterich et al. (2016) [23], *Predictive Rate Parity* is another group-level fairness notion, which is also referred to as calibration, *Test Fairness* [14] and *No Disparate Mistreatment* [84]. In the context of binary classification, *Predictive Rate Parity* requires that among those whose predicted value is positive (negative), their probability of actually having a positive label should be the same regardless of the value of the protected feature:

$$\forall a \in \mathcal{A}, \hat{y} \in \{0,1\}: \ P(Y=1 \mid A=a, \widehat{Y}=\hat{y}) = P(Y=1 \mid \widehat{Y}=\hat{y}). \tag{3.3}$$

Similar to *Equalized Odds*, *Predictive Rate Parity* can also be characterized through the conditional independence relation among $(A, Y, \widehat{Y})$:

Definition 3.3. Predictive Rate Parity We say that a predictor $\widehat{Y}$ is fair in terms of *Predictive Rate Parity* with respect to the protected feature A and the outcome Y, if Y is conditionally independent from A given $\widehat{Y}$, i.e., $Y \perp\!\!\!\perp A \mid \widehat{Y}$.

Although looks similar to *Equalized Odds*, *Predictive Rate Parity* and *Equalized Odds* are actually incompatible. It is shown independently by Kleinberg et al. [45] and Chouldechova [14] that both conditions can not be attained at the same time except in very special cases, e.g., Y and A are independent or $\widehat{Y}$ perfectly predict Y with no prediction error.

3.1.4 No Direct/Indirect Discrimination

Previous fairness notions (*Demographic Parity*, *Equalized Odds*, and *Predictive Rate Parity*) are based on associative relations among variables. Going beyond these observational criteria, it is desirable if we can further capture the structure of the data generating process by making use of causal modeling.

In the legislation literature, the discrimination is commonly divided into two categories: direct discrimination (e.g., rejecting a well-qualified loan applicator only because of his/her race) and indirect discrimination (e.g., refusing service to areas with certain Zip code). The motivation behind detecting indirect discrimination is that: among the non-protected attributes X, there is a set of attrites whose usage may still remain (potentially) unjustified although they are not the protected feature itself, i.e., redlining attributes R. In the language of causal reasoning, given a causal graph, we can start from the node for the protected feature and trace along the paths all the way to the node of interest by following the arrowheads in the graph. Therefore, we can characterize direct and indirect discrimination as different path-specific causal effects with respect to the protected feature [60, 93, 94, 92, 95, 59, 88]:

Definition 3.4. No Direct Discrimination Let us denote π_d as the path set that contains only the direct path from the protected feature A to the predictor $\widehat{Y}$, i.e., $A \to \widehat{Y}$. We say that a predictor $\widehat{Y}$ is fair in terms of *No Direct Discrimination* with respect to the protected feature A and the path set π_d, if for any $a, a' \in \mathcal{A}$ and $\hat{y} \in \mathcal{Y}$ the π_d-specific causal effect of the change in A from a to a' on $\widehat{Y} = \hat{y}$ satisfies:

$$P\big(\widehat{Y} = \hat{y} \mid do(A = a'|_{\pi_d})\big) - P\big(\widehat{Y} = \hat{y} \mid do(A = a)\big) = 0. \tag{3.4}$$

Definition 3.5. No Indirect Discrimination Let us denote π_i as the path set that contains all causal paths from the protected feature A to the predictor $\widehat{Y}$ which go though redlining attributes R, i.e., each path within the set π_i includes at least one node from R. We say that a predictor $\widehat{Y}$ is fair in terms of *No Indirect Discrimination* with respect to the protected feature A and the path set π_i, if for any $a, a' \in \mathcal{A}$ and $\hat{y} \in \mathcal{Y}$ the π_i-specific causal effect of the change in A from a to a' on $\widehat{Y} = \hat{y}$ satisfies:

$$P\big(\widehat{Y} = \hat{y} \mid do(A = a'|_{\pi_i})\big) - P\big(\widehat{Y} = \hat{y} \mid do(A = a)\big) = 0. \tag{3.5}$$

Motivated by the idea of capturing discrimination through different types of causal effects of the protected feature on the predictor, similar notions are also proposed by Kilbertus et al. (2017) [42] to further distinguish different types of attributes that are decendants of the protected feature. In particular, for attributes that are influenced by the protected feature A in a manner that we deem as non-discriminatory, i.e., resolving variables, the path-specific causal effects of A on $\widehat{Y}$ through these attributes are "resolved": for attributes that are influence by A in a unjustifiable way, i.e., proxy variables, the path-specific causal effects of A on $\widehat{Y}$ through these attributes are "unresolved":

Definition 3.6. No Unresolved Discrimination We say that a predictor $\widehat{Y}$ is fair in terms of *No Unresolved Discrimination*, if each path from A to $\widehat{Y}$ is blocked by a resolving variable in the corresponding causal graph.

Definition 3.7. No Proxy Discrimination We say that a predictor $\widehat{Y}$ is fair in terms of *No Proxy Discrimination* with respect to a proxy R, if for any $r, r' \in \mathcal{R}$ and $\hat{y} \in \mathcal{Y}$:

$$P\big(\widehat{Y} = \hat{y} \mid do(R = r)\big) = P\big(\widehat{Y} = \hat{y} \mid do(R = r')\big). \tag{3.6}$$

Similar to related notions like "explanatory feature" [39], "readlining attribute" [95], and "admissible variables" [66], the notion of "resolving variable" and "proxy variable" are just decendants of A with different user-specified characteristics. Compared to *No Indirect Discrimination*, although *No Proxy Discrimination* is also capturing indirect discrimination through proxy variables, i.e., redlining attributes, the intervention based on the proxy variable is conceptually easier to parse compared to the intervention on the protected feature itself – especially considering the fact that the protected feature, e.g., gender or race, is a deeply rooted personal property and it is impossible to perform a randomized trial [78].

3.1.5 Counterfactual Fairness

So far, the causal notions of fairness (*No Direct/Indirect Discrimination*, *No Unresolved Discrimination*, *No Proxy Discrimination*) are quantifying the discrimination on the group level. *Counterfactual Fairness* proposed by Kusner et al. (2017), compared to previous ones, is more fine-grained since it captures individual-level notion of fairness.

The canonical individual-level fairness notion is *Individual Fairness* proposed by Dwork et al. (2012). The intuition behind *Individual Fairness* is that we want similar predicted outcome for similar individuals

(in terms of the user-specified similarity metric). While *Individual Fairness* is general enough to be applicable in various practical scenarios, the specification of the similarity metric is not often straightforward. *Counterfactual Fairness*, on the other hand, approaches the individual-level fairness problem from a different angle. In particular, the intuition behind *Counterfactual Fairness* is that a decision is fair towards an individual if the decision remains the same in the actual world (the current reality) and a counterfactual world (the hypothetical world where this individual had a different demographic property):

Definition 3.8. Counterfactual Fairness Given a causal model $(U, V, \mathbf{F})$ where V consists of all features $V \coloneqq \{A, X\}$, we say that a predictor $\widehat{Y}$ is fair in terms of *Counterfactual Fairness* with respect to the protected feature A, if for any $a, a' \in \mathcal{A}, x \in \mathcal{X}, \hat{y} \in \mathcal{Y}$ the following holds true:

$$P\big(\widehat{Y}_{A\leftarrow a}(U) = \hat{y} \mid A = a, X = x\big) = P\big(\widehat{Y}_{A\leftarrow a'}(U) = \hat{y} \mid A = a, X = x\big). \tag{3.7}$$

3.1.6 Path-specific Counterfactual Fairness

First introduced by Chiappa [12], the *Path-specific Counterfactual Fairness* notion, as indicated by the name, shares the similar intuition with *Counterfactual Fairness* and captures the difference in decision between the actual world and a counterfactual world. Different from *Counterfactual Fairness*, more fine-grained causal effects are utilized by *Path-specific Counterfactual Fairness* – path-specific counterfactual effects, i.e., the counterfactual causal effects (that compare the factual world with the counterfactual world) are characterized only through unfair paths. Wu et al. (2019) [83] uses the abbreviated term *PC Fairness* to denote a unified formula for various causal notions of fairness:

Definition 3.9. Path-specific Counterfactual Fairness (PC Fairness) Given a causal model $(U, V, \mathbf{F})$ and a factual observation $O = o$, where V consists of all features $V \coloneqq \{A, X\}$ and $O \subseteq \{A, X, Y\}$, we say that a predictor $\widehat{Y}$ is fair in terms of *Path-specific Counterfactual Fairness (PC Fairness)* with respect to the protected feature A and the path set π, if for any $a, a' \in \mathcal{A}, \hat{y} \in \mathcal{Y}$ the π-specific counterfactual causal effect of the change in A from a to a' on $\widehat{Y} = \hat{y}$ satisfies (let $\bar{\pi}$ denote the set containing all other paths in the graph that are not elements of π):

$$P\big(\widehat{Y}_{A\leftarrow a'|\pi, A\leftarrow a|\bar{\pi}}(U) = \hat{y} \mid O = o\big) - P\big(\widehat{Y}_{A\leftarrow a}(U) = \hat{y} \mid O = o\big) = 0. \tag{3.8}$$

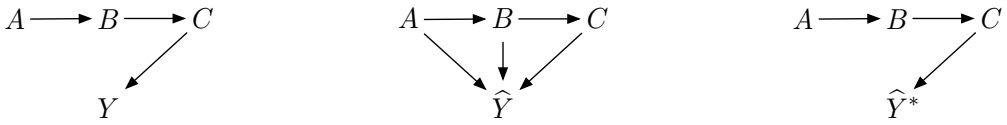
For different configurations of the observation $O = o$ and the path set of interest π, *PC Fairness* can capture different types of causal effect, which results in various flavors of fairness notions. For example, if π consists of all paths in the graph and $O = \{A, X\}$, this configuration of *PC Fairness* reduces to *Counterfactual Fairness*.

3.2 Subtlety: The Role of Causality in Fairness Analysis

In Section 3.1 I presented multiple commonly used fairness notions in the literature, many of which leverage the power of causal reasoning. In light of this, it is necessary and important to reflect on the subtleties regarding the role of causality in fairness analysis. In particular, I argue that we should always perform sanity checks to make sure that we are quantifying the discrimination in the way that best fulfills the intuition behind the fairness notion.

3.2.1 Causal Modeling on the Object of Interest

It is widely recognized in the fairness literature that we can leverage the power of causal reasoning to help us better understand how discrimination propagates through the data generating process [42, 48,



(a) The data generating process for the ground truth Y

(b) The data generating process for the classifier/regressor $\widehat{Y}$

(c) The data generating process for the inference output $\widehat{Y}^*$

Figure 3.1: The comparison between the causal graphs that represent different data generating processes for the ground truth Y, the prediction result via regression $\widehat{Y}$, and the prediction result via inference $\widehat{Y}^*$.

65, 95, 59, 88, 50, 12, 83]. While the assumption of the availability of additional information about the data generating process, e.g., a causal graph, is in general acceptable, we should think twice before directly assuming that the prediction variable $\widehat{Y}$ shares the exactly same causal graph with the ground truth variable Y.[1]

Let us consider a simple example of the performance of basketball players where there are four variables: the gender of the player (A), the height of the player (B), the player's position (C), the total points scored by the player in this season (Y). Suppose that the data generating process with respect to the ground truth, i.e., the relation among the measured variables A, B, C, and Y, can be described by Figure 3.1a: the gender is a cause of the height; the height determines the position of the player on court; the position determines the total points that the player scores in the season.[2] The task is to come up with the prediction ($\widehat{Y}$) for the total points of this season (Y) based on the information available (the gender A, the height B, and the position C): $\widehat{Y} = f(A, B, C)$ where $f : \mathcal{A} \times \mathcal{B} \times \mathcal{C} \to \mathcal{Y}$ is a classification/regression algorithm. In this case, the prediction result itself can be viewed as a random variable. If we were to draw a graph that represents the how $\widehat{Y}$ is generated from (A, B, C), we will have a data generating process as shown in Figure 3.1b. The reason of the extra arrows in Figure 3.1b compared to Figure 3.1a is that the classification/regression algorithm, regardless of the loss function and the optimization techniques, treats available variables merely as input features, which does not really respect the original data generating process in Figure 3.1a. However, if for example we use a generative model and perform an probabilistic inference task on the outcome where we follow the underlying data generating process, the inference result $\widehat{Y}^*$ can share the causal graph with the ground truth variable Y.[3] The data generating process for the prediction result via inference $\widehat{Y}^*$ (Figure 3.1c) is only different (in terms of the causal graph) from its counterpart for the ground truth variable Y (Figure 3.1a) up to a substitution of the outcome variable.

As we can see in the previous example, when performing causal reasoning in fairness analysis we should always be aware of the object of interest, i.e., the variable whose data generating process is subject to fairness consideration. When we directly assume that the causal graph can be shared by the ground truth and the prediction, there could be a mismatch between the causal model (based on which the discrimination is quantified) and the object (whose data generating process is in fact *not* described by this model). If there is a mismatch between the causal model and the object of interest, the result of discrimination quantification could be unpredictable and therefore is hardly justifiable.

[1]For the tasks like prediction, the output $\widehat{Y}$ is usually generated by a classification or regression algorithm in the literature.

[2]This is a simplified model with a very limited number of variables involved for the purpose of illustration.

[3]Usually, we need stronger assumptions regarding the underlying data generating process in order to perform the inference tasks, e.g., the availability of a structural equation model (SCM) instead of only the causal graph.

3.2.2 Causal Modeling with the Intended Semantics

When we mention causal modeling in the fairness analysis, there are several potential candidates in terms of which causal model we are referring to. For instance, a causal model that describes "the data generating process for the ground truth" could be interpreted with multiple different semantics: the causal model recovered from the data at hand through causal discovery (i.e., the model that corresponds to the current reality), the external knowledge from experts or just an assumption on the data generating process (i.e., the model that we assume with professional knowledge to be able to describe the current reality), the relations among variables in the hypethetical ideal world where there is no discrimination (i.e., the model that corresponds to the ideal reality which currently is not the case), and so on. Among these various possible interpretations, it is not always self-explainable from the fairness notions themselves regarding which interpretation really corresponds to the causal model presented to us, if without further clarifications. Therefore in practice, we should not only keep in mind the intuition behind the fairness notions, but also make sure that the semantics of the causal model we are using truly matches the type of the intended task.

In the basketball player example, if we were to quantify the discrimination hidden within the current data, we need to refer to Figure 3.1a and at the same time make sure that it reflects the relation among variables within the current data at hand. If we decide that there is historical discrimination in the data and we want to see how the distribution of Y would have been like were there no discrimination, we need to refer to a graph that corresponds to a hypothetical ideal world (which may not necessarily be the same as Figure 3.1a). A mismatch between the intended task and the causal model with correct scope can easily introduce unforeseeenable new discriminations into the system/data.

3.2.3 Potential Modifications to Previous Causal Notions of Fairness

Multiple causal notions of fairness have been proposed in the literature [93, 42, 48, 95, 59, 88, 41, 12, 83, 66]. However, in light of the frequently neglected subtleties that I discussed in Section 3.2.1 and Section 3.2.2, we might need to modify causal fairness notions to remedy the mismatch between the intended task (which type of discrimination we would like to quantify) and the object or semantic of interest, so that the intuition behind the notion can be properly expressed. Here for the purpose of illustration, I present the modified versions of *No Direct/Indirect Discrimination* (Definition 3.4 and Definition 3.5), *Counterfactual Fairness* (Definition 3.8), and *Path-specific Counterfactual Fairness* (Definition 3.9) that I reviewed in Section 3.1:

Definition 3.10. No Direct Discrimination (modified) Given the causal graph that describes the data generating process of the current reality, let us denote π_d as the path set that contains only the direct path from the protected feature A to the outcome Y, i.e., $A \to Y$. We say that the outcome Y is fair in terms of *No Direct Discrimination* with respect to the protected feature A and the path set π_d, if for any $a, a' \in \mathcal{A}$ and $y \in \mathcal{Y}$ the π_d-specific causal effect of the change in A from a to a' on $Y = y$ satisfies:

$$P\big(Y = y \mid do(A = a'|_{\pi_d})\big) - P\big(Y = y \mid do(A = a)\big) = 0. \tag{3.9}$$

Definition 3.11. No Indirect Discrimination (modified) Given the causal graph that describes the data generating process of the current reality, let us denote π_i as the path set that contains all causal paths from the protected feature A to the outcome Y which go though redlining attributes R, i.e., each path within the set π_i includes at least one node from R. We say that the outcome Y is fair in terms of *No Indirect Discrimination* with respect to the protected feature A and the path set π_i, if for any $a, a' \in \mathcal{A}$ and $y \in \mathcal{Y}$

the π_i-specific causal effect of the change in A from a to a' on $Y = y$ satisfies:

$$P\big(Y = y \mid do(A = a'|_{\pi_i})\big) - P\big(Y = y \mid do(A = a)\big) = 0. \tag{3.10}$$

Definition 3.12. Counterfactual Fairness (modified) Given a causal model $(U, V, \mathbf{F})$ that describes the data generating process of the current reality, where V consists of all features $V \coloneqq \{A, X\}$, we say that the outcome Y is fair in terms of *Counterfactual Fairness* with respect to the protected feature A, if for any $a, a' \in \mathcal{A}, x \in \mathcal{X}, y \in \mathcal{Y}$ the following holds true:

$$P\big(Y_{A\leftarrow a}(U) = y \mid A = a, X = x\big) = P\big(Y_{A\leftarrow a'}(U) = y \mid A = a, X = x\big). \tag{3.11}$$

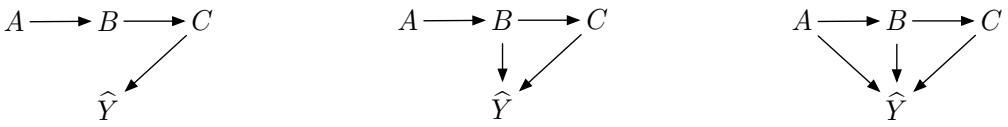
Definition 3.13. Path-specific Counterfactual Fairness (modified) Given a causal model $(U, V, \mathbf{F})$ that describes the data generating process of the current reality and a factual observation $O = o$, where V consists of all features $V \coloneqq \{A, X\}$ and $O \subseteq \{A, X, Y\}$, we say that the outcome Y is fair in terms of *Path-specific Counterfactual Fairness (PC Fairness)* with respect to the protected feature A and the path set π, if for any $a, a' \in \mathcal{A}, y \in \mathcal{Y}$ the π-specific counterfactual causal effect of the change in A from a to a' on $Y = y$ satisfies (let $\bar{\pi}$ denote the set containing all other paths in the graph that are not elements of π):

$$P\big(Y_{A\leftarrow a'|\pi, A\leftarrow a|\bar{\pi}}(U) = y \mid O = o\big) - P\big(Y_{A\leftarrow a}(U) = y \mid O = o\big) = 0. \tag{3.12}$$

Compared to the original notions (Definition 3.4, 3.5, 3.8, 3.9), the modified causal notions (Defintion 3.10, 3.11, 3.12, 3.13) are quantifying discrimination with respect to the outcome variable Y instead of the prediction $\widehat{Y}$, using the data generating process behind Y with respect to the current reality. This seemingly trivial modification is more than just exchanges of variables. In practical applications, when we assume the availability (either via a educated guess or from the expertise knowledge) of a causal graph that characterizes underlying properties of the data, we are referring to the data generating process with respect to the outcome variable Y, instead of the predictor $\widehat{Y}$. Furthermore, even if we can draw the causal graph for predictions as illustrated in Figure 3.1b (for prediction via classification/regression) and Figure 3.1c (for prediction via inference), we will still need to make sure that we pair up the object of interest and the technical detail of the corresponding analyzing scheme (e.g., path-based criterion, or causal effect estimation that involves additional information/assumption on the functional class).

Let us revisit the basketball player performance example in Section 3.2.1. Suppose that a practitioner would like to audit fairness with respect to the prediction and at the same time understand the source of discrimination, and that the practitioner thinks that a causal notion of fairness could be very handy. Let's say, for example, the practitioner picks *Counterfactual Fairness* (Definition 3.8, which is the original notion proposed by Kusner et al., 2017 [48]) since this causal notion is with respect to $\widehat{Y}$. There are multiple strategies a practitioner might choose to audit fairness, and for each one of them it is possible to have a mismatch between the mission (which kind of fairness we really would like to capture) and the means (how exactly fairness audit is carried out):

Strategy (1) The practitioner makes an educated guess regarding how attributes could relate to each other in the data set and draws the causal graph Figure 3.1a. Considering the task is to audit fairness on $\widehat{Y}$, the practitioner directly exchanges the variable Y in the graph to $\widehat{Y}$ and draws the graph shown in Figure 3.2a. The practitioner then proceeds to the fairness audit via *Counterfactual Fairness* (Definition 3.8) without knowing the detail regarding how $\widehat{Y}$ is computed (which is the output of a regressor).



(a) Strategy 1 & 2 (b) Strategy 3 (c) Strategy 4

Figure 3.2: The comparison between graphs that the practitioner draws in different strategies.

Strategy (2) The practitioner utilizes the exactly same strategy to audit fairness as in Strategy (1), without knowing the detail regarding how $\widehat{Y}$ is computed (which is in fact output of a inference model shown as in Figure 3.1c).

Strategy (3) The practitioner first pictures an idealized fair world where both the height B and the position C are causes of the total points scored by the player Y. Then the practitioner realizes that the task is to audit fairness on $\widehat{Y}$ and draws the graph shown in Figure 3.2b. The practitioner proceeds to the fairness audit via *Counterfactual Fairness* (Definition 3.8) without knowing the detail regarding how $\widehat{Y}$ is computed (which is the output of a regressor).

Strategy (4) The practitioner notices that $\widehat{Y}$ is the output of a regression algorithm and draws the causal graph that corresponds the data generating process of $\widehat{Y}$ as shown in Figure 3.2c. The practitioner then proceeds to the fairness audit via *Counterfactual Fairness* (Definition 3.8) with respect to $\widehat{Y}$ (which is the output of a regressor).

For Strategy 1, there is a mismatch between the object of interest ($\widehat{Y}$) and the corresponding data generating process (it should be the graph shown in Figure 3.1b, instead of the one shown in Figure 3.2a. For Strategy 2, there seems to be no mismatch between the object of interest ($\widehat{Y}$) and the corresponding data generating process in terms of the causal graph, since Figure 3.2a happens to be identical to Figure 3.1c (except for the asterisk symbol in Figure 3.1c). Although the causal graphs agree with each other, the details of causal modeling (e.g., functional classes in the SCM) may differ across the algorithm builder (who generates $\widehat{Y}$) and the practitioner (who audits fairness on $\widehat{Y}$), which may still incur a mismatch between the object of interest and the corresponding data generating process. For Strategy 3, there is a mismatch of the causal modeling both in terms of the intended semantics (using the graph which reflects the hypothetical ideal world) and the object of interest (substituting Y with $\widehat{Y}$ without justification). For Strategy 4, there seems to be no mismatch since Figure 3.2c is identical to Figure 3.1b. However, while there is no significant difference in terms of technical treatments when estimating causal effects on Y and $\widehat{Y}$ (if we were to draw a causal graph for the regression output), only the data generating process behind Y reflects what happens in the real world. After all, one of the strongest motivations behind the usage of a causal notion is the insight into the data generating process behind the outcome Y in the current reality, but this purpose does not seem to be well-served if we consider the data generating process behind the prediction $\widehat{Y}$.

As we can see from different possible strategies in this example, there are subtleties involved in enforcing/auditing causal notions of fairness. Neglecting these subtleties may result in mismatches between the mission and the means. Unfortunately, the precautions against these negligence are often not well packed into the causal notions of fairness themselves in current literature. To some extent, the causal notions of fairness with respect to $\widehat{Y}$ (unintentionally) invites the negligence of subtleties discussed in Section 3.2.1 and Section 3.2.2. In fact, it is not uncommon to see (variants of) the aforementioned Strategy (1) utilized in current literature [42, 48, 95, 12, 83]. Therefore, thinking twice before enforcing a

fairness notion and making sure that the intuition behind the notion is fulfilled in a sensible and justifiable way, are always desirable.

Chapter 4

Attainability and Optimality: An Extensive Discussion on Equalized Odds

In this chapter, I focus on one particular fairness notion, namely, *Equalized Odds* proposed by Hardt et al. (2016) [32], discussing the attainability of *Equalized Odds* and, furthermore, if attainable, the optimality of performance among fair predictors. The analysis on the theoretical attainability is a fundementally different type of fairness audit compared to error bounds in the finite sample cases. Actually, as we can see in this chapter, *Equalized Odds* is not always attainable for regression and even for classification tasks, if we use deterministic prediction functions. Throughout the chapter, I focus on the implication and illustration of the results and defer all the proofs to Appendix A. The research presented in this chapter is based on a joint work with Kun Zhang and I am the primary contributor.

4.1 Motivation

In the algorithmic fairness literature, the phenomenon of the "tradeoff between fairness and accuracy" for the prediction has been widely observed and discussed [38, 64, 27, 14, 7, 16, 45, 56, 1, 53, 81, 5, 79]. However, only when we assume/know that the data does not contain discrimination can we really justify the practice of enforcing fairness and accuracy at the same time for the prediction result. After all, if Y contains discrimination, enforcing the prediction $\widehat{Y}$ to be close to Y (even if with fairness regularization) is not desirable. Actually as we shall see in this chapter, even if the data does not contain any discrimination, the utilization of the data for prediction can still introduce new discriminations, for instance, the theoretical unattainability of fairness with a particular prediction scheme.

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 that has important practical implications. It characterizes a completely different kind of violation of fairness compared to the empirical error bound of discrimination in finite-sample cases. In practice, although one can always audit violation of *Equalized Odds* via the empirical quantification, because of the finite sample size one cannot expect the empirical fairness violation to be exactly zero. The absolute magnitude of the empirical fairness violation is often not informative enough since it is not clear how small an empirical fairness violation is small enough such that the prediction result can be deemed as "fair", i.e., the fairness notion of interest will be attained with zero violation in the large sample limit. Therefore it is desirable to develop prediction schemes that come with theoretical guarantees with respect to the method itself so that the

fairness notion is proved to be attainable in the large sample limit.

4.2 Equalized Odds in Regression

Various regularization terms have also been proposed to suppress discrimination when predicting a continuous target [7, 87, 53, 63]. However, whether or not one can always achieve 0-discrimination for regression, even if with an unlimited amount of data, is not clear. In this section I first consider a simple setup with linearly correlated continuous data as a witnessing observation that Equalized Odds is not always attainable. Then I consider more general cases where regression is performed by a deterministic prediction function and derive the condition under which Equalized Odds can hold true. Finally, when stochastic prediction is utilized, I show that under mild assumptions one can always find a non-trivial fair predictor, i.e., Equalized Odds is guaranteed to be (non-trivially) attainable.[1]

4.2.1 Unattainability of Equalized Odds in Linear Non-Gaussian Regression

Recall that in this section I consider the bias that results from data utilization of the prediction scheme, any possible bias introduced by the data generating procedure itself is beyond the scope of the discussion. Let us start with the specific linear, non-Gaussian situation where the data is generated as follows (H is not measured in the dataset):

$$X = qA + E_X, \;\; H = bA + E_H, \;\; Y = cX + dH + E_Y, \tag{4.1}$$

where (A, E_X, E_H, E_Y) are mutually independent and q, b, c, d are constants.

Let us denote $E := E_Y + dE_H$ and let $\widehat{Y}$ be a linear combination of A and X, i.e., $\widehat{Y} = \alpha A + \beta X = (\alpha + q\beta)A + \beta E_X$, with linear coefficients α and β, where $\beta \neq 0$. In Theorem 4.1, I present the general result in linear non-Gaussian cases, where one cannot achieve the conditional independence between $\widehat{Y}$ and A given Y.

Theorem 4.1. ***(Unattainability of Equalized Odds in the Linear Non-Gaussian Case)*** *Assume that X has a causal influence on Y, i.e., $c \neq 0$ in Equation 4.1, and that 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, $\widehat{Y}$ is always conditionally dependent on A given Y.*

From Theorem 4.1, we can see that if at most one of E_X and E 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 impossible to (non-trivially) achieve Equalized Odds with a linear model in the linear non-Gaussian case. If both E_X and E are Gaussian (A is not necessarily Gaussian), then one can easily find the constraints on α and β such that $\widehat{Y}$ is conditionally independent from A given Y, as stated in the following corollary.

Corollary 4.2. *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 $\widehat{Y} \perp\!\!\!\perp A \mid Y$ if and only if*

$$\frac{\alpha}{\beta} = \frac{bdc \cdot \sigma^2_{E_X} - q \cdot \sigma^2_E}{c^2 \cdot \sigma^2_{E_X} + \sigma^2_E}. \tag{4.2}$$

[1] Here by "non-trivial" predictors I am referring to predictors that perform better than a random guess or a constant output.

4.2.2 Regression with Deterministic Prediction

We have seen that in the linear non-Gaussian case, any non-zero linear combination of the feature (which is a deterministic prediction function of the input feature) will not satisfy Equalized Odds. One may naturally wonder, instead of only considering linear models, whether or not Equalized Odds can be achieved in general regression cases. It is therefore desirable to derive the condition under which Equalized Odds can hold true for regression with deterministic prediction functions.

Theorem 4.3. ***(Condition to Achieve Equalized Odds for Regression with Deterministic Prediction)*** *Assume that the protected feature A and the continuous target variable Y are dependent and that their joint probability density $p(A, Y)$ 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 prediction $\widehat{Y}$ be characterized by a deterministic function $f : \mathcal{A} \times \mathcal{X} \to \mathcal{Y}$. Equalized Odds holds true if and only if the following condition is satisfied ($\delta(\cdot)$ is the delta function):*

$$\forall y, \hat{y} \in \mathcal{Y},\ \forall a, a' \in \mathcal{A}, a \neq a' :\ Q(a, y, \hat{y}) = Q(a', y, \hat{y}),$$
$$\textit{where}\ \ Q(a, y, \hat{y}) \triangleq \int_{\mathcal{X}} \delta\big(\hat{y} - f(a, x)\big) p_{X|AY}(x|a, y) dx.$$

In special cases when $X \perp\!\!\!\perp A \mid Y$ and f is a function of only X (e.g., a linear function with 0 coefficient for A), this condition is always satisfied. In general cases, however, this condition specifies a rather strong constraint on the relation between the conditional probability density $p_{X|AY}(x \mid a, y)$ and the function f. Because of the integration over the product of $\delta(\cdot)$ function and $p_{X|AY}(x \mid a, y)$, the aforementioned condition requires f to "pick out" certain x's (for any given $\hat{y} \in \mathcal{Y}$ and $a \in \mathcal{A}$) such that the sum of the corresponding conditional probability density values stays the same. Generally speaking, if the way f "picks out" x is not strictly coupled with the conditional probability density $p_{X|AY}(x \mid a, y)$, the condition specified in Theorem 4.3 would be violated, i.e., Equalized Odds cannot hold true. In order to further illustrate this constraint, let us consider a special case where variables are linear correlated with Gaussian exogenous terms (i.e., E_X and $E = E_Y + dE_H$ in Equation 4.1 are Gaussian). Here, the conditional distribution of X given A and that of Y given A are both Gaussian. Let $\widehat{Y} = f(A, X)$ be a linear function of (A, X) with linear coefficients α and non-zero β. Then for any given value of $(\widehat{Y}, A)$ we have $X = (\widehat{Y} - \alpha A)/\beta$. Therefore $\delta\big(\hat{y} - f(a, x)\big)$ does not equal to 0 only on the singleton set $\{x \mid f(a, x) = \hat{y}\} = \{(\hat{y} - \alpha a)/\beta\}$. We have

$$Q(a, y, \hat{y}) = p_{X|AY}(\frac{\hat{y} - \alpha a}{\beta} \mid a, y).$$

Then the aforementioned constraint requires that

$$p_{X|AY}(\frac{\hat{y} - \alpha a}{\beta} \mid a, y) = p_{X|AY}(\frac{\hat{y} - \alpha a'}{\beta} \mid a', y). \tag{4.3}$$

Since the conditional distribution of X given $A = a$, i.e., $\mathcal{N}(qa, \sigma_X^2)$, and that of Y given $A = a$, i.e., $\mathcal{N}\big((qc + bd)a, c^2\sigma_X^2 + \sigma_E^2\big)$, are both Gaussian, we have $p_{X|AY}(x \mid a, y) = \frac{1}{\sqrt{2\pi\sigma_{ay}^2}} \exp\{\frac{(x - \mu_{ay})^2}{2\sigma_{ay}^2}\}$, where $\mu_{ay} = qa + \frac{c\sigma_X^2}{c^2\sigma_X^2 + \sigma_E^2}\big(y - (qc + bd)a\big),\ \sigma_{ay}^2 = (1 - \frac{c^2\sigma_X^2}{c^2\sigma_X^2 + \sigma_E^2})\sigma_X^2$.

Then the Equation 4.3 hold true if and only if the absolute distance between $\frac{\hat{y} - \alpha a}{\beta}$ and μ_{ay} is not a function of a (for any fixed y and $\hat{y}$), which yields the following relation:

$$\frac{\alpha}{\beta} = \frac{bdc \cdot \sigma_{E_X}^2 - q \cdot \sigma_E^2}{c^2 \cdot \sigma_{E_X}^2 + \sigma_E^2},$$

which is the exact relation between α and β derived from a different perspective in Corollary 4.2.

As we can see in this example, in order to achieve Equalized Odds, the coefficients of linear function $f(A, X) = \alpha A + \beta X$ have to satisfy a very specific relation. In more general cases, for example with linear non-Gaussian data, as we have seen in Theorem 4.1, Equalized Odds is not attainable with deterministic linear regression. This indicates the general restrictiveness of the aforementioned constraint on the relation between $p_{X|AY}(x \mid a, y)$ and the deterministic function f, i.e., Equalized Odds cannot hold true. Actually as we will see in Theorem 4.5, similar phenomena also occur for classification with deterministic predictors.

4.2.3 Regression with Stochastic Prediction

In light of the unattainability of Equalized Odds for prediction with deterministic functions of A and X, it is important to ask whether Equalized Odds can be attainable with stochastic prediction, where the output given input features is no longer a single value. As we shall see in Theorem 4.4, with stochastic prediction functions, under mild assumptions one can always find a non-trivial predictor $\widetilde{Y}$ that satisfies Equalized Odds, even if both the target Y and the protected feature A are continuous.[2]

Theorem 4.4. ***(Attainability of Equalized Odds for Regression with Stochastic Prediction)*** *Let A, X, and Y be continuous variables with domain of value $\mathcal{A}$, $\mathcal{X}$, and $\mathcal{Y}$, respectively. Assume that their joint distribution is fixed and known. Further assume $Y \not\perp\!\!\!\perp A$, $Y \not\perp\!\!\!\perp X$, and $Y \not\perp\!\!\!\perp X \mid A$. Without loss of generality let the conditional probability density $p_{X|AY}(x \mid a, y)$ be non-negative and finite. Then there exists $\widetilde{Y}$ with domain of value $\mathcal{Y}$ whose distribution is fully determined by $p_{\widetilde{Y}|AX}(\tilde{y} \mid a, x)$, such that $\widetilde{Y}$ is not independent from (A, X) but $\widetilde{Y} \perp\!\!\!\perp A \mid Y$, i.e., the Equalized Odds is non-trivially attainable.*

From Theorem 4.4 we can see that with stochastic prediction, one can guarantee the attainability of Equalized Odds for regression under rather mild assumptions. It would then be desirable to construct general, nonlinear prediction models to produce a stochastic prediction (i.e., with a certain type of randomness in the prediction). One possible way follows the framework of Generative Adversarial Networks (GANs) [30]: I use random standard Gaussian noise E, 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, as a function of A, X, and E, at the same time. Given the loss function $\mathcal{L}$, a class of function $\mathcal{F}$, and a fairness penalty $\mathcal{G}$, I propose the following objective function for fitting an Equalized Odds model with stochastic prediction:

$$\min_{f \in \mathcal{F}} \mathbb{E}\big[\mathcal{L}\big(f(A, X, E), Y\big) + \lambda \mathcal{G}(f(A, X, E), A, Y)\big]. \tag{4.4}$$

The first term measures the prediction error, for instance the mean squared error for regression. The second term is the fairness penalty that imposes Equalized Odds. I use the kernel measure of conditional dependence [29] between $\widetilde{Y}$ and A given Y as the fairness penalty term. The hyperparameter λ reflects the trade-off between accuracy and fairness.

4.3 Equalized Odds in Classification

In this section, I consider the attainability of Equalized Odds for binary classifiers (with a deterministic or stochastic prediction function), and furthermore, if attainable, the optimality of performance of various computational procedures under the fairness criterion.

[2] In this chapter, I use $\widetilde{Y}$ to denote Equalized Odds predictors to distinguish from (not necessarily fair) predictors $\widehat{Y}$.

4.3.1 Classification with Deterministic Prediction

I begin with considering cases when classification is performed by a deterministic function of the input. Similar to the previous analysis for regression tasks with deterministic prediction functions in Section 4.2.2, I derive the conditions under which Equalized Odds can possibly hold true for classification with deterministic prediction functions.

Theorem 4.5. ***(Condition to Achieve Equalized Odds for Classification with Deterministic Prediction)*** *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 $\widehat{Y}$ be a deterministic function $f : \mathcal{A} \times \mathcal{X} \to \mathcal{Y}$. Let $S_A^{(\hat{y})} := \{a \mid \exists x \in \mathcal{X} \text{ s.t. } f(a, x) = \hat{y}\}$, and $S_{X|a}^{(\hat{y})} := \{x \mid f(a, x) = \hat{y}\}$. Equalized Odds is attained if and only if the following conditions hold true (for continuous X, replace summation with integration accordingly):*

(i) $\forall \hat{y} \in \mathcal{Y} : \; S_A^{(\hat{y})} = \mathcal{A}$,
(ii) $\forall \hat{y} \in \mathcal{Y}, \; \forall a, a' \in \mathcal{A} : \sum_{x \in S_{X|a}^{(\hat{y})}} P_{X|AY}(x \mid a, y) = \sum_{x \in S_{X|a'}^{(\hat{y})}} P_{X|AY}(x \mid a', y)$

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 rather restrictive, condition (ii) specifies an even stronger constraint on the relation between $P_{X|AY}(x|a, y)$ (or $p_{X|AY}(x|a, y)$ for continuous X) and the set $S_{X|a}^{(\hat{y})}$ (which is determined by the function f). Generally speaking, in order to score a better classification accuracy, one would like to make $P_{\widehat{Y}|A,X}(\hat{y}|a, x)$ as close as possible to $P_{Y|A,X}(y|a, x)$, and if the set $S_{X|a}^{(\hat{y})}$ and $P_{X|AY}(x|a, y)$ are not strictly coupled, condition (ii) would be violated, i.e., Equalized Odds cannot be attained. For better illustration I present concrete examples where those conditions can or cannot be satisfied. For the purpose of simplifying the notation, let us consider cases when A, X, and Y are binary, and the joint probability of (A, Y) is specified as following:

$$\begin{aligned} P(A = 0, Y = 0) &= 0.2, \\ P(A = 0, Y = 1) &= 0.4, \\ P(A = 1, Y = 0) &= 0.3, \\ P(A = 1, Y = 1) &= 0.1. \end{aligned} \tag{4.5}$$

Then in the special case when $X \perp\!\!\!\perp A \mid Y$, Equalized Odds can hold true if f is a function of only X. For example, if $f = \mathbb{1}\{X = 1\}$, one can quickly verify that $P(\widetilde{Y} = 1 \mid A = a, Y = 0) = P(X = 1 \mid Y = 0)$ and that $P(\widetilde{Y} = 1 \mid A = a, Y = 1) = P(X = 1 \mid Y = 1)$ (for $a \in \{0, 1\}$), i.e., Equalized Odds holds true. If $X \not\perp\!\!\!\perp A \mid Y$, let us compare the following two cases:

$$\begin{aligned} P(X = 1 \mid A = 0, Y = 0) &= 0.3, \\ P(X = 1 \mid A = 1, Y = 0) &= 0.7, \\ P(X = 1 \mid A = 0, Y = 1) &= 0.8, \\ P(X = 1 \mid A = 1, Y = 1) &= 0.2; \end{aligned} \tag{4.6}$$

$$
\begin{aligned}
&P(X=1 \mid A=0, Y=0)=0.4,\\
&P(X=1 \mid A=1, Y=0)=0.7,\\
&P(X=1 \mid A=0, Y=1)=0.6,\\
&P(X=1 \mid A=1, Y=1)=0.2.
\end{aligned}
\tag{4.7}
$$

Here, the joint distribution of (A, Y) is specified the same as in Equation 4.5. In Equation 4.6, the conditional probability of X is crafted so that $P(X=0 \mid A=0, Y=0)=P(X=1 \mid A=1, Y=0)$ and $P(X=0 \mid A=0, Y=1)=P(X=1 \mid A=1, Y=1)$. Then if we choose $f=\mathbb{1}\{A=X\}$ (or flip the prediction, let $f=1-\mathbb{1}\{A=X\}$) to exploit this property, we can have a predictor that satisfies Equalized Odds. One can quickly verify this: $P(\widetilde{Y}=1 \mid A=0, Y=0)=P(X=0 \mid A=0, Y=0)=0.6=P(X=1 \mid A=1, Y=0)=P(\widetilde{Y}=1 \mid A=1, Y=0)$, and $P(\widetilde{Y}=1 \mid A=0, Y=1)=P(X=0 \mid A=0, Y=1)=0.2=P(X=1 \mid A=1, Y=1)=P(\widetilde{Y}=1 \mid A=1, Y=1)$. However, there is no obvious reason to believe that the conditional probability of X given A and Y should satisfy such "crafted" property in general cases. In the case shown in Equation 4.7, one cannot find an f that satisfies conditions (i) and (ii) in Theorem 4.5, and therefore in this case there is no deterministic prediction function that satisfies Equalized Odds.

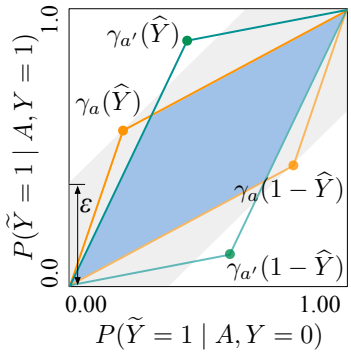
4.3.2 Classification with Stochastic Prediction

In this subsection, I consider cases when stochastic prediction is acceptable, namely, the classifier would output class labels with certain probabilities. Among different categories of approaches to derive a fair classifier, recent efforts to impose Equalized Odds in the pre-processing manner [51, 99, 76] 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. Therefore, I omit pre-processing approaches from the discussion and focus on post-processing and in-processing fair classifiers.

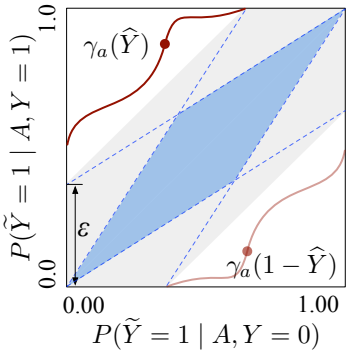
I first derive the relation between positive rates (TPR and FPR) of binary classifiers before and after the post-processing step, i.e., $\widehat{Y}_{\text{opt}}$ (the unconstrainedly optimized classifier) and $\widetilde{Y}_{\text{post}}$ (the fair classifier derived by post-processing $\widehat{Y}_{\text{opt}}$), and show that under mild assumptions, one can always derive a non-trivial Equalized-Odds $\widetilde{Y}_{\text{post}}$ via a post-processing step. Then, from the ROC feasible area perspective, I 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.

The Post-processing Step

The post-processing step of a predictor $\widehat{Y}$ (here I drop the subscript when there is no 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, with the *shifted decision boundary* [28], the *derived predictor* [32], or the *(monotonic) joint loss optimization* over *decoupled classifiers* [26], is then fully specified by a (possibly randomized) function of $(A, \widehat{Y})$. This implies the conditional independence $\widetilde{Y}_{\text{post}} \perp\!\!\!\perp Y \mid A, \widehat{Y}$. Then if we denote the positive rates of $\widehat{Y}$ as $P_{\widehat{Y}|AY}(1|a, y)$, positive rates of $\widetilde{Y}$ as $P_{\widetilde{Y}|AY}(1|a, y)$, we can factorize the positive rates under this conditional independence: , where $\beta_a^{(\hat{y})} := P(\widetilde{Y}=1 \mid A=a, \widehat{Y}=\hat{y})$.



(a) Discrete protected feature A



(b) Continuous protected feature A

Figure 4.1: ROC feasible area illustrations for binary classifiers with stochastic prediction functions.

Therefore, post-processing an existing predictor boils down to optimizing parameters (for discrete A) or functions (for continuous A) $\beta_a^{(\hat{y})}$.

The 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 necessarily a fair one) with a 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 $\gamma_{ay}(\widehat{Y})$:

$$\gamma_a(\widehat{Y}) = \big(\gamma_{a0}(\widehat{Y}), \gamma_{a1}(\widehat{Y})\big) := \big(P_{\widehat{Y}|AY}(1|a,0), P_{\widehat{Y}|AY}(1|a,1)\big). \tag{4.8}$$

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

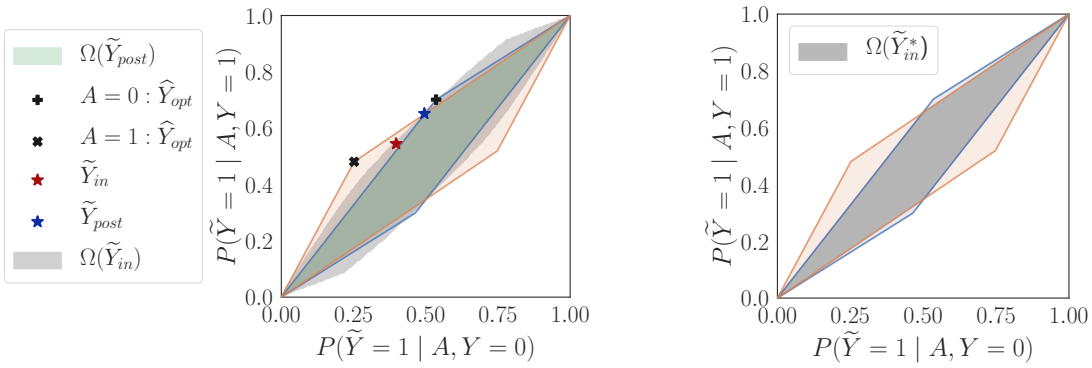
$$\mathcal{C}_a(\widehat{Y}) := \mathrm{Conv}\big\{(0,0), \gamma_a(\widehat{Y}), \gamma_a(1-\widehat{Y}), (1,1)\big\}, \tag{4.9}$$

and then, as already stated in Hardt et al. [32], the (FPR, TPR) pair corresponding to a post-processing predictor falls within (including the boundary of) $\mathcal{C}_a(\widehat{Y})$.

Definition 4.6. 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.

Hardt et al. [32] proposed that the post-processing fair predictor can be derived by solving a linear programming problem. However, it is not given whether this problem always has a non-trivial solution. Following Hardt al el. [32], I 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. As I show in Theorem 4.7, under mild assumptions an Equalized-Odds predictor $\widetilde{Y}_{\text{post}}$ derived via post-processing $\widehat{Y}$ always has non-empty ROC feasible area.

Theorem 4.7. ***(Attainability of Equalized Odds for Classification with Stochastic Prediction)*** *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., $\Omega(\widetilde{Y}_{post}) \neq \emptyset$.*



(a) Feasible areas for $\widetilde{Y}_{\text{in}}$, $\widetilde{Y}_{\text{post}}$, and $\widehat{Y}_{\text{opt}}$

(b) $\Omega(\widetilde{Y}^*_{\text{in}})$ with "pseudo" constraints

Figure 4.2: The comparison between ROC feasible areas: $\Omega(\widetilde{Y}_{\text{in}})$ for the in-processing fair classifier, $\Omega(\widetilde{Y}_{\text{post}})$ for the post-processing fair classifier, $\Omega(\widehat{Y}_{\text{opt}})$ for the unconstrainedly optimized classifier, and $\Omega(\widetilde{Y}^*_{\text{in}})$ for the in-processing fair classifier with "pseudo" constraints.

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 Figure 4.1a and Figure 4.1b 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 no better and often worse than the weakest performance across different groups, which is obviously suboptimal.

4.3.3 Optimality of Performance among Classifiers that Satisfy Equalized Odds

In this subsection I discuss the optimality of performance of fair classifiers derived via in-processing and post-processing approaches. Let us consider a more general setting for the post-processing approach, where the optimal predictor to be post-processed is stochastic (while most previous approaches use deterministic predictors). The derivation of the in-processing fair predictor $\widetilde{Y}_{\text{in}}$ and the unconstrained statistical optimal predictor $\widehat{Y}_{\text{opt}}$ take the following forms, respectively:

$$\begin{aligned} \min_{f\in\mathcal{F}} \quad & \mathbb{E}[\mathcal{L}(\widetilde{Y}_{\text{in}}, Y)] \\ \text{s.t.} \quad & P_{\widetilde{Y}_{\text{in}}|AY}(\tilde{y} \mid a, y) = P_{\widetilde{Y}_{\text{in}}|Y}(\tilde{y} \mid y) \\ \text{where} \quad & \widetilde{Y}_{\text{in}} \sim \text{Bernoulli}\big(f(A, X)\big), \end{aligned} \tag{4.10}$$

$$\begin{aligned} \min_{f\in\mathcal{F}} \quad & \mathbb{E}[\mathcal{L}(\widehat{Y}_{\text{opt}}, Y)] \\ \text{where} \quad & \widehat{Y}_{\text{opt}} \sim \text{Bernoulli}\big(f(A, X)\big). \end{aligned} \tag{4.11}$$

It is natural to wonder, now that one can always directly solve for $\widetilde{Y}_{\text{in}}$ from Equation 4.10, how it is related to $\widetilde{Y}_{\text{post}}$, which is derived by post-processing the $\widehat{Y}_{\text{opt}}$ solved from Equation 4.11? 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 I summarize in the following theorem.

Theorem 4.8. ***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 $\beta^{(\hat{y})}_{ay} = \sum_{x\in\mathcal{X}} P(\widetilde{Y}_{in} = 1 \mid A = a, X = x)P(X = x \mid A = a, Y = y, \widehat{Y}_{opt} = \hat{y})$.

As we can see in Figure 4.2a and Figure 4.2b, if the additional "pseudo" constraints are introduced when optimizing $\widetilde{Y}^*_{\text{in}}$, we have $\Omega(\widetilde{Y}_{\text{in}}) \supseteq \Omega(\widetilde{Y}_{\text{post}}) = \Omega(\widetilde{Y}^*_{\text{in}})$. Therefore, with the same objective function and fairness constraint, the fair classifier derived from an in-processing approach always outperforms (or performs equally well with) the one derived from a post-processing approach.

4.3.4 Equalized Odds Experiments

In this section, I provide numerical results of enforcing *Equalized Odds* in various settings. I first present illustrations of unattained Equalized Odds with deterministic regression. Then, to demonstrate the benefit of stochastic prediction with respect to imposing fairness, I present the results for regression with stochastic prediction on both simulated data and the real-world *Communities and Crime* data set. Finally for classification tasks I compare the performance of several existing methods in the literature on multiple real-world data sets. The description of the data set as well as the data processing procedure can be found at the end of this subsection.

Regression Tasks

The regression experiments consist of two base models, namely, a linear regression model and a neural network model, each of which appears in the form of both deterministic and stochastic prediction. For the neural network model, following Mary et al. [53] I use a simple regressor for the experiments: two hidden layers (with the same number of neurons ranging from 30 to 100 for each layer, depending on the size of the data set) with SELUs nonlinearity [44]. The network is optimized by minimizing the MSE loss combined with the λ-weighted fairness penalty term, using the Adam optimizer [43]. I use the kernel measure of conditional dependence (KMCD) [29] as the fairness penalty term (for stochastic prediction), and the weight λ ranges from $\{2^{-2}, 2^{-1}, \ldots, 2^{14}\}$. The range of the absolute value of λ may differ for different experiment setups since the relative value difference between the MSE loss and the KMCD measure. The learning rate is chosen from $\{10^{-4}, 10^{-5}, 10^{-6}\}$ (smaller learning rate is preferable especially for larger values of λ, i.e., more emphasis on the fairness penalty), and the batch size is chosen from $\{64, 128, 256\}$. As I have stated in Section 4.2.3, for stochastic prediction, a random sample from the standard Gaussian distribution is concatenated to each batch of data, and the input dimension of the model changes accordingly. The random sample to concatenate is redrawn each time the data point is seen by the model.

In the experiments on the simulated data, for linear cases, the data is generated as stated in Equation 4.1, with non-Gaussian distributed exogenous terms (E_X, E_H, and E_Y). Figure 4.3a and Figure 4.3b correspond to different distributions for the noise terms, specifically, the Laplace distribution and uniform distribution, respectively. I use linear regression with the *Equalized Correlations* constraint [82], 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 involved and Gaussian distributed exogenous terms. I use a neural network regressor with an Equalized Odds regularization term [53] to

perform nonlinear fair regression. As we can see in Figure 4.3c and Figure 4.3d, for nonlinear regression tasks, Equalized Odds may not be attained even if every exogenous term is Gaussian distributed. In fact, according to the kernel-based conditional independence (KCI) test [91] (with the null hypothesis $\widehat{Y} \perp\!\!\!\perp A \mid Y$), the p-value is smaller than 0.05 in all four examples, which indicates that Equalized Odds does not hold true.

In Figure 4.4 I compare deterministic and stochastic regression on simulated linear non-Gaussian data. Panel (a) illustrates the trade-off between fairness in terms of the kernel measure of conditional dependence (KMCD) and prediction error in terms of the mean squared error (MSE) for different values of the hyperparameter λ. Panel (b) summarizes the p-value outputs of the KCI test. Note that in (b), the green boxes are not visible since p-values in the deterministic prediction case are always close to 0. The distribution of the p-value outputs clearly indicate that stochastic prediction often achieve Equalized Odds (the test fails to reject the null quite often), while deterministic prediction may not (the test rejects the null almost all the time). For panel (c) and (d), I fix the training sample size at 200 while the sample size for testing ranges from 100 to 3000.[3] The purpose is to see whether the prediction can remain fair with more and more new testing data. A practical example of the scenario would be an automated hiring system to aid recruitment (each future applicant is a new data point for the system as long as the system is deployed and keeps running). As we can see in panel (d), while the p-value for deterministic prediction is always close to 0 (invisible green box, which indicates violated Equalized Odds), the p-value for stochastic prediction spreads over $[0, 1]$ if the test set sample size is not too big compared to the training data.

In the experiments on the real-world data, I perform 20 random splits of *Communities and Crime* data into training (80%) and testing (20%) sets, and present the testing performance for deterministic and stochastic prediction models (for the linear regressor as well as the neural network regressor) in Figure 4.5. Compare panel (b) and panel (d), we can see that unlike the nonlinear stochastic prediction (the neural network predictor in panel (d)), the distribution of p-value outputs for stochastic prediction with a linear base model does not dominate its deterministic counterpart. This is not surprising since for stochastic prediction with a linear model, the prediction is in essence produced by a deterministic linear model and an additive noise. Therefore, if the deterministic part of the linear model does not satisfy Equalized Odds, adding an independent noise will not help impose fairness. In situations where the protected feature is not available when testing, it is desirable to exclude the protected feature from the attributes when performing prediction. To this end, during the training process, the input to the regression model does not include the protected feature, and the protected feature is only used to compute the fairness penalty. The related result on the *Communities and Crime* data set is summarized in Figure 4.6.

Classification Tasks

In Figure 4.7, I compare the performance under Equalized Odds of multiple methods proposed in the literature. Hardt et al. [32] propose a post-processing approach where the prediction is randomized to minimize violation of fairness; Zafar et al. [84] use a covariance proxy measure as the regularization term when optimizing classification accuracy; Agarwal et al. [1] take the reductions approach and reduce fair classification into solving a sequence of cost-sensitive classification problems; Rezaei et al. [62] minimize the worst-case log loss using an approximated regularization term; Baharlouei et al. [5] propose to use Rényi correlation as the regularization term to account for nonlinear dependence between variables.

[3] Considering the fact that the computational cost for KCI test p-value increases dramatically with large sample size, I use linearly correlated data with relatively small sample size for training and consider increasingly large test sets in the repetitive experiments (with fixed hyperparameter λ).

To measure the violation of the fairness criterion, I 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. [1], I 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 I plot the testing accuracy versus the violation of Equalized Odds.

In Figure 4.7, I compare the performance under Equalized Odds of multiple methods proposed in the literature. Although a probabilistic classification model is used for across each method (logistic regression here), if an algorithm outputs the class label where the prediction likelihood is maximized, the prediction is in essence deterministic. Therefore, although here we are considering finite data cases, we can still anticipate a lower level of fairness violation with stochastic prediction. This is validated by the numerical experiment: while the post-processing approach [32] does not score the lowest test error, the violation of Equalized Odds is the lowest compared to other approaches.

Description of Data Sets

(1) **Communities and Crime**: The UCI Communities and Crime data set contains 122 features for 1994 records.[4] The regression task is to predict the number of violent crimes per population for US cities given various census information of the corresponding communities. The data is preprocessed following Mary et al. [53], and the (continuous) protected feature is the ratio of African-American people in the population.

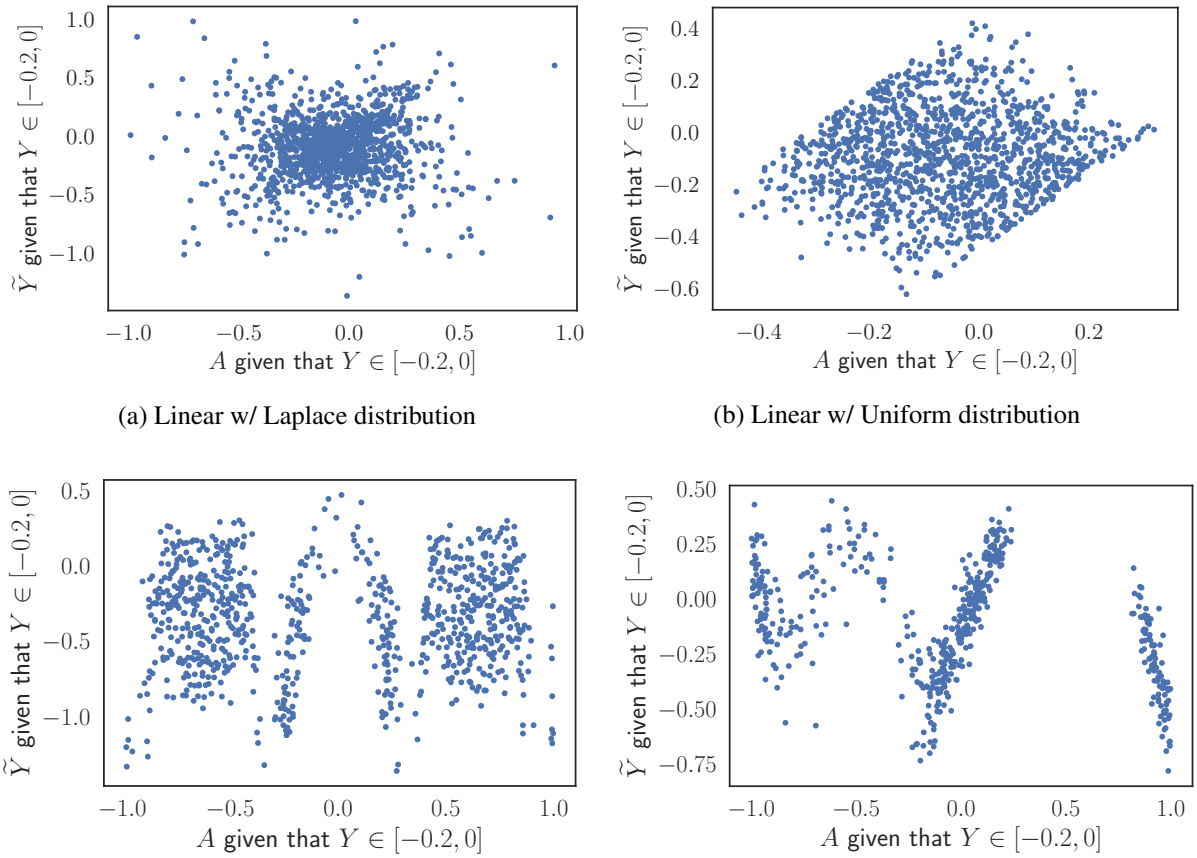
(2) **Adult** [46]: The UCI Adult data set contains 14 features for 45,222 individuals (32,561 samples for training and 12,661 samples for testing).[5] The census information includes gender, marital status, education, capital gain, etc. The classification task is to predict whether a person's annual income exceeds 50,000 USD. I 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.

(3) **Bank** [58]: The UCI Bank Marketing data set is related with marketing campaigns of a banking institution, containing 16 features of 45,211 individuals.[6] 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, I 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.

(4) **COMPAS** [3]: 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. [84] and Nabi et al. [59], I consider the subset consisting of African-Americans and Caucasians defendants. The features I use include age, gender, race, number of priors, and degree of charges. The task is to predict the recidivism of the defendant and I choose race as the protected feature.

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

[4] `https://archive.ics.uci.edu/ml/datasets/communities+and+crime`
[5] `http://archive.ics.uci.edu/ml/datasets/Adult`
[6] `https://archive.ics.uci.edu/ml/datasets/bank+marketing`
[7] `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`



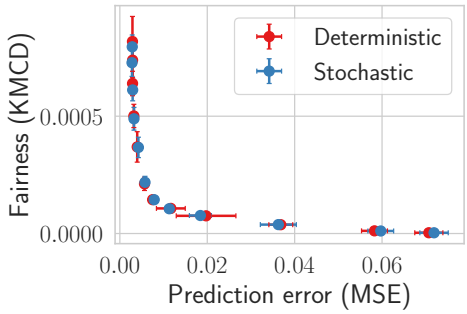
(a) Linear w/ Laplace distribution

(b) Linear w/ Uniform distribution

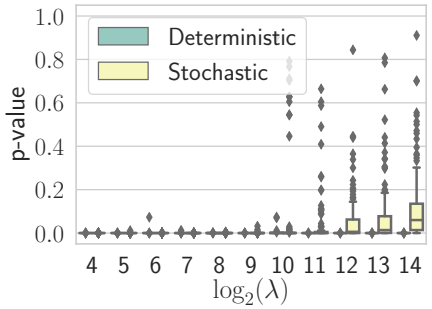
(c) Nonlinear w/ Gaussian distribution

(d) Nonlinear w/ Gaussian distribution

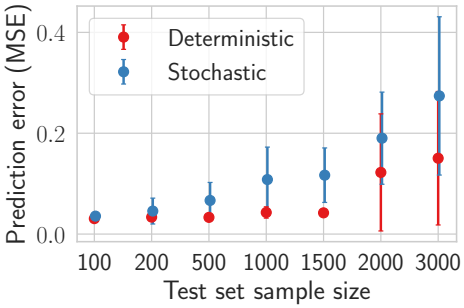
Figure 4.3: Illustration of unattainable Equalized Odds for regression with deterministic prediction. Panel (a)-(b): Linear regression on the data generated with linear transformations and non-Gaussian distributed exogenous terms (following Laplace, Uniform distribution respectively). Panel (c)-(d): Nonlinear regression with a neural net regressor [53] 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.



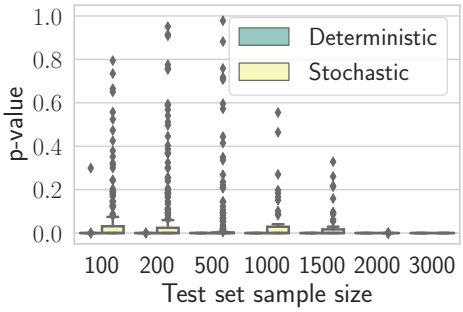
(a) Tradeoff between fairness and accuracy



(b) KCI test outcome with different λ

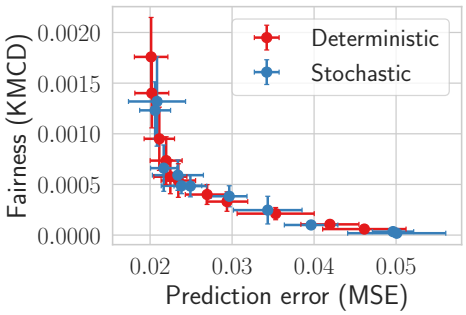


(c) MSE on test set of different sizes

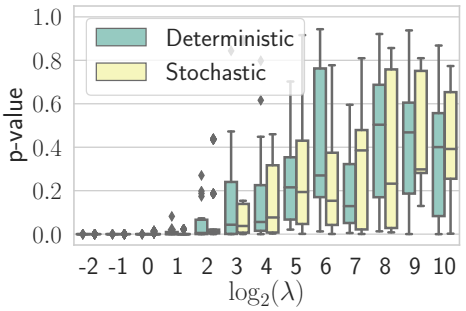


(d) p-values on test set of different sizes

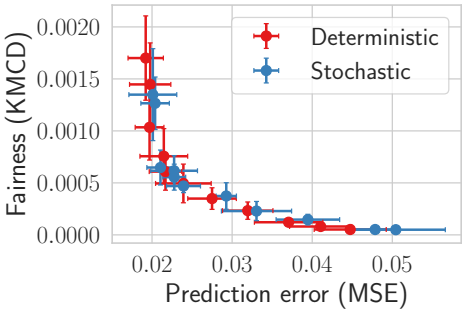
Figure 4.4: Illustration of the benefit of the stochastic prediction for the purpose of imposing Equalized Odds. The data is linear non-Gaussian, and the predictor is a neural network (nonlinear) regressor with deterministic or stochastic output. The green boxes in panel (b) and (d) are barely visible, meaning that p-values for deterministic prediction are always close to 0.



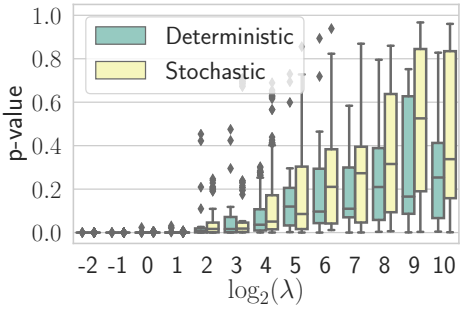
(a) KMCD vs. MSE (linear model)



(b) KCI test outcome (linear model)

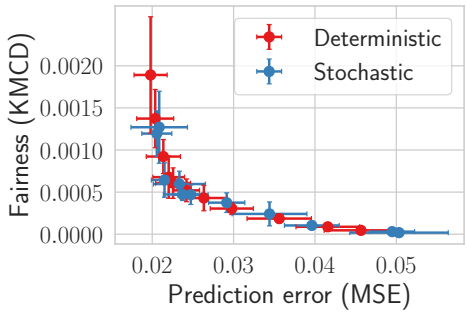


(c) KMCD vs. MSE (neural network)

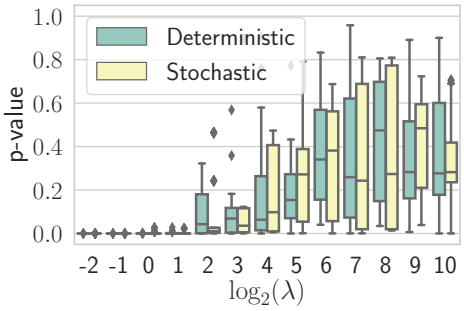


(d) KCI test outcome (neural network)

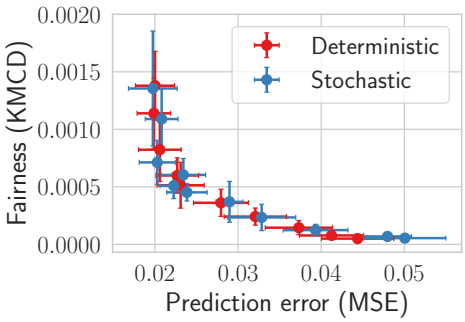
Figure 4.5: Results for regression with deterministic and stochastic prediction functions on the *Communities and Crime* data set. The base model could be a linear regressor or a neural network regressor, and the stochastic prediction is achieved by introducing an independent noise input sampled from a standard Gaussian distribution.



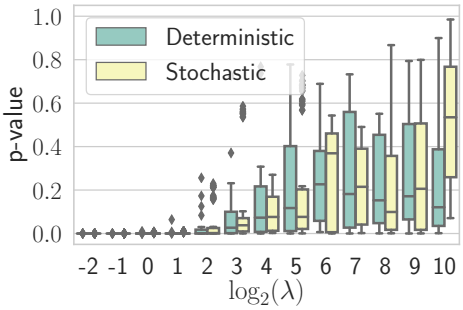
(a) KMCD vs. MSE (linear model)



(b) KCI test outcome (linear model)



(c) KMCD vs. MSE (neural network)



(d) KCI test outcome (neural network)

Figure 4.6: Results for regression with deterministic and stochastic prediction functions on the *Communities and Crime* data set (the protected feature is *not* included as an attribute). The base model could be a linear regressor or a neural network regressor, and the stochastic prediction is achieved by introducing an independent noise input sampled from a standard Gaussian distribution.

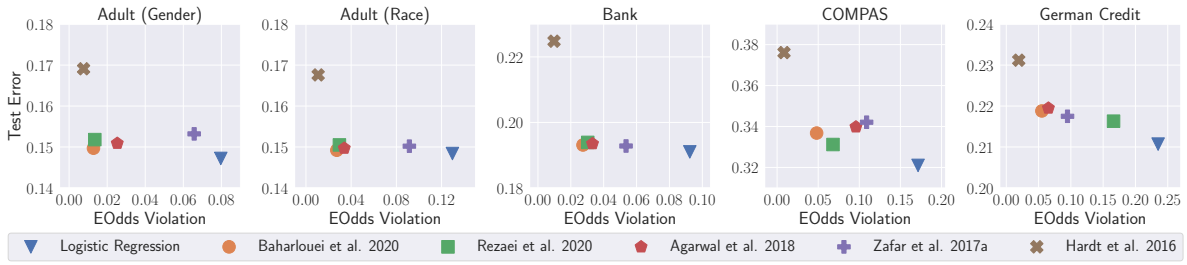


Figure 4.7: Results for classification with Equalized Odds criterion.

gender as the protected feature.

4.4 Summary

In this chapter I focus on *Equalized Odds* and consider the attainability of fairness, and furthermore, if it is attainable, the optimality of the prediction performance under various settings. I present theoretical guarantee that under mild assumptions, one can always find a non-trivial stochastic predictor that satisfies *Equalized Odds* in the large sample limit, while in general fairness is not attainable with a deterministic predictor. Therefore I propose a stochastic prediction scheme that has the distribution-free theoretical guarantee of fairness attainability.

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. [32], 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 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. The system in fact only has one fixed threshold (i.e., the original lower threshold), which is in essence a deterministic predictor (therefore in general, does not satisfy *Equalized Odds*).

As we can see from the extensive discussion with respect to enforcing the *Equalized Odds* notion, it is always preferrable to develop prediction schemes that come with theoretical guarantees on the fairness attainability, so that one can better regulate the way that information is utilized to perform prediction (under the assumption that the data at hand is "clean"). That being said, before picking a fairness notion we should also be aware of the potential negative social impact of enforcing such notion, since the choice of fairness notion largely depends on the question at hand and there is no one-size-fits-all solution.

Chapter 5

Conclusion and Future Work

In this thesis, I present a reflection and a case study on fairness notions in machine learning. Considering the often neglected subtleties regarding the role played by causality in fairness analysis, I propose necessary modifications to previous causal notions of fairness. I extensively discuss the *Equalized Odds* notion of fairness to illustrate the importance of theoretical attainability of the fairness notion. In particular, I consider the attainability of *Equalized Odds* and, furthermore, if it is attainable, the optimality of the prediction performance under various settings. I present theoretical guarantee that under mild assumptions, one can always find a non-trivial stochastic predictor that satisfies *Equalized Odds* in the large sample limit, while in general fairness is not attainable with a deterministic predictor.

The future works naturally span across different types of fairness considerations. If we are interested in quantifying discrimination within the data at hand, it is desirable to develop methods to evaluate and guarantee the effectiveness of fairness pursuit with respect to the underlying data generating process, especially for the potential correction (going beyond detection) of the data to eliminate discriminations within the data. If we are willing to assume that the data at hand is not biased, a thorough understanding of the fairness notion of interest (e.g., the one that is, or will be, deployed in real world) calls for analysis with respect to attainability and optimality, which, if carefully characterized, is very informative and helpful both in terms of theoretical rigorousness and practical significance (e.g., the development of better learning strategies that come with theoretical guarantees). If we consider fairness from a broader scope, the potential unification of the findings from fairness audits conducted in separated but highly-related scenarios (e.g., school admission, loan application, occupational outlook, etc.) would be very helpful to identify potential ways to systematically promote fairness from a wider scope.

Bibliography

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

[2] Edward J Anderson and Peter Nash. *Linear programming in infinite-dimensional spaces: theory and applications*. John Wiley & Sons, 1987.

[3] 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.

[4] Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 2005.

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

[6] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *CALIFORNIA LAW REVIEW*, pages 671–732, 2016.

[7] 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.

[8] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[9] 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*, pages 4349–4357, 2016.

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

[11] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

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

[13] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[14] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism

prediction instruments. *Big data*, 5(2):153–163, 2017.

[15] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[16] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[17] 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*, pages 1436–1445, 2019.

[18] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

[19] Daniel Daners. Introduction to functional analysis. *The University of Sydney*, 2008.

[20] David Danks and Alex John London. Algorithmic bias in autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4691–4697, 2017.

[21] George Bernard Dantzig. *Linear programming and extensions*, volume 48. Princeton university press, 1965.

[22] 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.

[23] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 2016.

[24] 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*, pages 2791–2801, 2018.

[25] 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*, pages 214–226, 2012.

[26] 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*, pages 119–133, 2018.

[27] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[28] 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*, pages 144–152. SIAM, 2016.

[29] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007.

[30] 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*, pages 2672–2680, 2014.

[31] Swati Gupta and Vijay Kamble. Individual fairness in hindsight. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 805–806, 2019.

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

[33] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[34] Hoda Heidari and Jon Kleinberg. Allocating opportunities in a dynamic model of intergenerational mobility. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 15–25, 2021.

[35] Hoda Heidari, Vedant Nanda, and Krishna Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In *36th International Conference on Machine Learning*, pages 2692–2701, 2019.

[36] Wen Huang, Yongkai Wu, Lu Zhang, and Xintao Wu. Fairness through equality of effort. In *Companion Proceedings of the Web Conference 2020*, pages 743–751, 2020.

[37] Yimin Huang and Marco Valtorta. Identifiability in causal bayesian networks: A sound and complete algorithm. In *AAAI*, pages 1149–1154, 2006.

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

[39] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems*, 35(3):613–644, 2013.

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

[41] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pages 2907–2914, 2019.

[42] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 656–666, 2017.

[43] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[44] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pages 971–980, 2017.

[45] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*.

Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[46] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

[47] Mark Krein and David Milman. On extreme points of regular convex sets. *Studia Mathematica*, 9:133–138, 1940.

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

[49] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

[50] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

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

[52] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

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

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

[55] Reinhold Meise and Dietmar Vogt. *Introduction to functional analysis*. Clarendon Press, 1997.

[56] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118, 2018.

[57] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

[58] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

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

[60] Judea Pearl. *Causality*. Cambridge university press, 2009.

[61] 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*, pages 339–355. Springer, 2017.

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

[63] Yaniv Romano, Stephen Bates, and Emmanuel J Candès. Achieving equalized odds by resampling sensitive attributes. *arXiv preprint arXiv:2006.04292*, 2020.

[64] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.

[65] 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*, pages 6414–6423, 2017.

[66] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810, 2019.

[67] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106, 2019.

[68] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[69] Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444, 2006.

[70] Ilya Shpitser and Judea Pearl. What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 2007.

[71] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

[72] 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*, pages 2164–2173, 2019.

[73] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. Springer New York, 1993.

[74] Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 499–506, 1995.

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

[76] Zilong Tan, Samuel Yeom, Matt Fredrikson, and Ameet Talwalkar. Learning fair representations for kernel models. In *International Conference on Artificial Intelligence and Statistics*, pages 155–166. PMLR, 2020.

[77] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, 2002.

[78] Tyler J VanderWeele and Whitney R Robinson. On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology (Cambridge, Mass.)*, 25(4):473, 2014.

[79] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In

Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 526–536, 2021.

[80] Min Wen, Osbert Bastani, and Ufuk Topcu. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*, pages 1144–1152. PMLR, 2021.

[81] Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. In *Advances in Neural Information Processing Systems*, pages 8780–8789, 2019.

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

[83] 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*, pages 3399–3409, 2019.

[84] 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*, pages 1171–1180, 2017.

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

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

[87] 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*, pages 335–340, 2018.

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

[89] Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1347–1353, 2017.

[90] Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 647–655. AUAI Press, 2009.

[91] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press, 2011.

[92] Lu Zhang and Xintao Wu. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics*, 4(1):1–16, 2017.

[93] Lu Zhang, Yongkai Wu, and Xintao Wu. Situation testing-based discrimination discovery: A causal inference approach. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, volume 16, pages 2718–2724, 2016.

[94] Lu Zhang, Yongkai Wu, and Xintao Wu. Achieving non-discrimination in data release. In *Pro-*

ceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1335–1344, 2017.

[95] 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*, pages 3929–3935, 2017.

[96] Xueru Zhang, Mohammadmahdi Khaliligarekani, Cem Tekin, et al. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. *Advances in Neural Information Processing Systems*, 32:15269–15278, 2019.

[97] Xueru Zhang and Mingyan Liu. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*, pages 525–555. Springer, 2021.

[98] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33:18457–18469, 2020.

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

Appendix A

Proof for Theorems

A.1 Proof for Theorem 4.1

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*

$$p_{V_1,V_2|V_3}(v_1, v_2 \mid v_3) = h(v_1, v_3) \cdot g(v_2, v_3). \tag{A.1}$$

Proof. First, if V_1 and V_2 are conditionally independent given variable V_3, then Equation A.1 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 A.1 w.r.t. v_1 and v_2, we have:

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

respectively. Bearing in mind Equation A.1, 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. Recall that the data is generated as follows (H is not measured in the dataset):

$$\begin{aligned} X &= qA + E_X, \\ H &= bA + E_H, \\ Y &= cX + dH + E_Y, \end{aligned} \tag{A.2}$$

where (A, E_X, E_H, E_Y) are mutually independent and q, b, c, d are constants.

Theorem. ***(Unattainability of Equalized Odds in the Linear Non-Gaussian Case)*** *Assume that X has a causal influence on Y, i.e., $c \neq 0$ in Equation A.2, and that 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, $\widehat{Y}$ is always conditionally dependent on A given Y.*

Proof. According to Equation A.2, we have

$$\begin{bmatrix} A \\ \widehat{Y} \\ Y \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \alpha + q\beta & \beta & 0 \\ qc + bd & c & 1 \end{bmatrix} \cdot \begin{bmatrix} A \\ E_X \\ E \end{bmatrix}. \tag{A.3}$$

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 A.3, 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

$$\tilde{\alpha} := \frac{\alpha + q\beta}{\beta}, \;\; \tilde{r} := bd - \frac{c\alpha}{\beta}, \;\; \text{and} \;\; \tilde{c} := \frac{c}{\beta}. \tag{A.4}$$

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

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

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

$$\begin{aligned} J &:= \log p_{A,\widehat{Y},Y}(a, \hat{y}, y) \\ &= \log p_A(a) + \log p_{E_X}(\frac{1}{\beta}\hat{y} - \tilde{\alpha}a) \\ &\quad + \log p_E(y - \tilde{r}a - \tilde{c}\hat{y}) - log|\beta| \\ &= f_1(a) + f_2(\frac{1}{\beta}\hat{y} - \tilde{\alpha}a) \\ &\quad + f_3(y - \tilde{r}a - \tilde{c}\hat{y}) - \log|\beta|. \end{aligned} \tag{A.5}$$

According to Lemma A.1, $A \perp\!\!\!\perp \widehat{Y} \mid Y$ if and only if $p_{A,\widehat{Y}|Y}(a, \hat{y} \mid y)$ is a product of a function of a and y and a function of t and y. $p_{A,\widehat{Y},Y}(a, \hat{y}, y)$ is further a product of the above function and a function of only y. This property, under the conditions in Theorem 4.1, is equivalent to the constraint

$$\frac{\partial^2 J}{\partial A \partial \widehat{Y}} \equiv 0. \tag{A.6}$$

According to Equation A.5, we have

$$\frac{\partial J}{\partial \hat{y}} = \frac{1}{\beta} \cdot f_2'(\frac{1}{\beta}\hat{y} - \tilde{\alpha}a) - \tilde{c} \cdot f_3'(y - \tilde{r}a - \tilde{c}\hat{y}),$$

and therefore

$$\frac{\partial^2 J}{\partial a \partial \hat{y}} = -\frac{\tilde{\alpha}}{\beta} \cdot f_2''(\frac{1}{\beta}\hat{y} - \tilde{\alpha}a) + \tilde{r}\tilde{c} \cdot f_3''(y - \tilde{r}a - \tilde{c}\hat{y}). \tag{A.7}$$

Combining Equations A.6 and A.7 gives

$$\tilde{r}\tilde{c} \cdot f_3''(y - \tilde{r}a - \tilde{c}\hat{y}) = \frac{\tilde{\alpha}}{\beta} \cdot f_2''(\frac{1}{\beta}\hat{y} - \tilde{\alpha}a). \tag{A.8}$$

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

$$\tilde{r}\tilde{c} \cdot f_3'''(y - \tilde{r}a - \tilde{c}\hat{y}) \equiv 0. \tag{A.9}$$

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 A.8, we have $\frac{\tilde{\alpha}}{\beta} \cdot f_2''(\frac{1}{\beta}\hat{y} - \tilde{\alpha}a) \equiv 0$, implies either $\tilde{\alpha} = 0$ or $f_2''(\frac{1}{\beta}\hat{y} - \tilde{\alpha}a) \equiv 0$. 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 A.4, 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) $f_3'''(y - \tilde{r}a - \tilde{c}\hat{y}) \equiv 0$. 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 A.8 is a nonzero constant. Hence f_2 is a quadratic function and E_X also follows a Gaussian distribution. Therefore if $A \perp\!\!\!\perp \widehat{Y} \mid Y$ were to be true, then E_X and E are both Gaussian. Its contrapositive gives the conclusion of this theorem. □

Corollary A.2. *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 $\widehat{Y} \perp\!\!\!\perp A \mid Y$ if and only if*

$$\frac{\alpha}{\beta} = \frac{bdc \cdot \sigma^2_{E_X} - q \cdot \sigma^2_E}{c^2 \cdot \sigma^2_{E_X} + \sigma^2_E}. \tag{A.10}$$

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 4.1, the Equalized Odds condition $A \perp\!\!\!\perp \widehat{Y} \mid Y$ is equivalent to Equation A.8, which, together with Equation A.4 as well as the fact that $f_2'' = -\frac{1}{\sigma^2_{E_X}}$ and $f_3'' = -\frac{1}{\sigma^2_E}$, yields Equation A.10. □

A.2 Proof for Theorem 4.3

Theorem. ***(Condition to Achieve Equalized Odds for Regression with Deterministic Prediction)*** *Assume that the protected feature A and the continuous target variable Y are dependent and that their joint probability density $p(A, Y)$ 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 prediction $\widehat{Y}$ be characterized by a deterministic function $f : \mathcal{A} \times \mathcal{X} \to \mathcal{Y}$. Equalized Odds holds true if and only if the following condition is satisfied ($\delta(\cdot)$ is the delta function):*

$$\forall y, \hat{y} \in \mathcal{Y},\ \forall a, a' \in \mathcal{A}, a \neq a' :\ Q(a, y, \hat{y}) = Q(a', y, \hat{y}),$$
$$\textit{where}\ \ Q(a, y, \hat{y}) \stackrel{\triangle}{=} \int_{\mathcal{X}} \delta\big(\hat{y} - f(a, x)\big) p_{X|AY}(x|a, y) dx.$$

Proof. Without loss of generality, let us assume that A and X are continuous. By Lemma A.1, the Equalized Odds criterion can be written into terms of the conditional probability density functions:

$$\forall y, \hat{y} \in \mathcal{Y},\ a \in \mathcal{A} :\ p_{\widehat{Y}|AY}(\hat{y}|a, y) = p_{\widehat{Y}|Y}(\hat{y}|y) \tag{A.11}$$

Since $\widehat{Y} = f(A, X)$ where f is a deterministic function of (A, X), f is an injective mapping from $\mathcal{A} \times \mathcal{X}$ to $\mathcal{Y}$. One can derive the joint probability density of $(\widehat{Y}, A, X, Y)$ by making use of the change of variables. Define the mapping $\mathcal{H}_f$ as following:

$$\mathcal{H}_f \begin{pmatrix} V \\ A \\ X \\ Y \end{pmatrix} = \begin{pmatrix} V + f(A, X) \\ A \\ X \\ Y \end{pmatrix},$$

where V is a constant 0 whose probability density function is $\delta(v)$. Notice that $\mathcal{H}$ is bijective:

$$\mathcal{H}_f^{-1} \begin{pmatrix} \widehat{Y} \\ A \\ X \\ Y \end{pmatrix} = \begin{pmatrix} \widehat{Y} - f(A, X) \\ A \\ X \\ Y \end{pmatrix},$$

with the Jacobian matrix

$$J(\mathcal{H}_f^{-1}) = \begin{bmatrix} 1 & -\frac{\partial f}{\partial A} & -\frac{\partial f}{\partial X} & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Therefore by making use of the change of variable technique, we have:

$$\begin{aligned} p_{\widehat{Y}AXY}(\hat{y}, a, x, y) &= p_{VAXY}(v, a, x, y)|J(\mathcal{H}_f^{-1})| \\ &= \delta\big(\hat{y} - f(a, x)\big) p_{AXY}(a, x, y). \end{aligned}$$

Expand the LHS of Equation A.11, we have:

$$\begin{aligned} p_{\widehat{Y}|AY}(\hat{y}|a, y) &= \int_{\mathcal{X}} p_{\widehat{Y}X|AY}(\hat{y}, x|a, y) dx \\ &= \int_{\mathcal{X}} \delta\big(\hat{y} - f(a, x)\big) p_{X|AY}(x|a, y) dx \\ &:= Q(a, y, \hat{y}). \end{aligned}$$

Since Equalized Odds holds true if and only if Equation A.11 holds true, the LHS of the equation does not involve a (as in RHS of the equation), i.e., $Q(a, y, \hat{y})$ does not change with a. Then, rewrite the RHS of Equation A.11, we have:

$$\begin{aligned} p_{\widehat{Y}|Y}(\hat{y}|y) &= \int_{\mathcal{A}} p_{\widehat{Y}|AY}(\hat{y}|a, y) p_{A|Y}(a|y) da \\ &= \int_{\mathcal{A}} Q(a, y, \hat{y}) p_{A|Y}(a|y) da \\ &= Q(a, y, \hat{y}) \int_{\mathcal{A}} p_{A|Y}(a|y) da \\ &= Q(a, y, \hat{y}). \end{aligned}$$

Therefore, Equalized Odds implies the condition that $Q(a, y, \hat{y}) = Q(a', y, \hat{y})$. On the other hand, it is easy to see that when the aforementioned condition holds true, Equation A.11 holds true, i.e., Equalized Odds holds true. □

A.3 Proof for Theorem 4.4

In order to prove Theorem 4.4, we need the following results from functional analysis (a fuller treatment can be found, for example, in [55, 19]):

Lemma A.3. *Let $\mathcal{S}$ be a set and $\mathcal{E} = (\mathcal{E}, \|\cdot\|)$ a Banach space. For a function $f : \mathcal{S} \to \mathcal{E}$, we define the supremum norm by $\|f\|_\infty := \sup_{s \in \mathcal{S}} \|f(s)\|$, and the space of bounded functions by $B(\mathcal{S}, \mathcal{E}) := \{f : \mathcal{S} \to \mathcal{E} \mid \|f\|_\infty < \infty\}$. Then $B(\mathcal{S}, \mathcal{E})$ is a Banach space with the supremum norm.*

Definition A.4 (Extreme Point)**.** If $\mathcal{K}$ is a non-empty convex subset of a vector space, a point $x \in \mathcal{K}$ is called an extreme point of $\mathcal{K}$ if whenever $x = \lambda x_1 + (1 - \lambda) x_2$ with $0 < \lambda < 1$ and $x_1, x_2 \in \mathcal{K}$, then $x = x_1 = x_2$.

Theorem A.5 (Krein-Milman Theorem [47])**.** *Let $\mathcal{K}$ be a non-empty compact subset of a locally convex Hausdorff topological vector space. Then the set of extreme points of $\mathcal{K}$ is not empty. If $\mathcal{K}$ is also convex, then $\mathcal{K}$ is the closed convex hull of its extreme points.*

It is a known result in finite-dimensional linear programming (see, for example, [21]) that if the specified feasible region $\mathcal{K}$ (which is the intersection of a finite number of half spaces) is non-empty, then the set of extreme points is not empty and finite, and the set of extreme directions is empty if and only if $\mathcal{K}$ is bounded. If $\mathcal{K}$ is unbounded, the set of extreme directions is not empty and finite. Furthermore, a point is in $\mathcal{K}$ if and only if it can be represented as a convex combination of the extreme points plus a non-negative linear combination of extreme directions (if there is any). Theorem A.5 extends this result to infinite dimensional vector spaces, allowing us to establish the existence of extreme points, and therefore a non-empty feasible region for certain infinite-dimensional linear programming problems. Note that in order to prove attainability of Equalized Odds, it is sufficient to show that the corresponding infinite-dimensional linear programming problem yields a non-empty feasible region. The establishment of duality is relatively complicated for infinite-dimensional linear programming [2] and is beyond the scope of the discussion.

Theorem. ***(Attainability of Equalized Odds for Regression with Stochastic Prediction)***
Let A, X, and Y be continuous variables with domain of value $\mathcal{A}$, $\mathcal{X}$, and $\mathcal{Y}$, respectively. Assume that their joint distribution is fixed and known. Further assume $Y \not\perp\!\!\!\perp A$, $Y \not\perp\!\!\!\perp X$, and $Y \not\perp\!\!\!\perp X \mid A$. Without loss

of generality let the conditional probability density $p_{X|AY}(x \mid a, y)$ be non-negative and finite. Then there exists $\widetilde{Y}$ with domain of value $\mathcal{Y}$ whose distribution is fully determined by $p_{\widetilde{Y}|AX}(\tilde{y} \mid a, x)$, such that $\widetilde{Y}$ is not independent from (A, X) but $\widetilde{Y} \perp\!\!\!\perp A \mid Y$, i.e., the Equalized Odds is non-trivially attainable.

Proof. Let us denote $h : \mathcal{A} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ as the fixed function satisfying $h(a, x, y) = p_{X|AY}(x \mid a, y)$. We want to show that there exists a function $f : \mathcal{A} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ characterizing the conditional probability density $f(a, x, \tilde{y}) = p_{\widetilde{Y}|AX}(\tilde{y} \mid a, x)$ that satisfies the following conditions:

(i) $\forall a \in \mathcal{A}, x \in \mathcal{X}, \tilde{y} \in \mathcal{Y}, f(a, x, \tilde{y})$ is non-negative and finite;
(ii) $\forall a \in \mathcal{A}, x \in \mathcal{X}, \int_{\mathcal{Y}} f(a, x, \tilde{y}) dt = 1$;
(iii) let $Q(a, y, \tilde{y}) := \int_{\mathcal{X}} f(a, x, \tilde{y}) h(a, x, y) dx$,
and then $\forall y, \tilde{y} \in \mathcal{Y}, a, a' \in \mathcal{A}, a \neq a', Q(a, y, \tilde{y}) = Q(a', y, \tilde{y})$;
(iv) the function $f(a, x, \tilde{y})$ cannot be written into a function of only $\tilde{y}$.

Conditions (i) and (ii) guarantee that f corresponds to a valid conditional probability density for some $\widetilde{Y}$ conditioned on A and X. Condition (iii) is the necessary and sufficient condition for Equalized Odds to hold true for regression tasks (as we have seen in Theorem 4.3). Condition (iv) guarantees that the corresponding $\widetilde{Y}$ is not independent from (A, X) – otherwise $\widetilde{Y}$ will be a trivial prediction.

Through the lens of functional analysis, we view f as a point in the infinite-dimensional vector space. In order to prove the existence of such a function f, we begin by showing that condition (i) and (iii) form a set of constraints of an infinite-dimensional linear programming problem, and that the corresponding feasible region is non-empty. Furthermore, we can always find a point (which is a function) in the aforementioned feasible region that satisfies condition (iv). Finally, condition (ii) is satisfied by "normalizing" over $\tilde{y} \in \mathcal{Y}$.

To begin with, any function that satisfies condition (i) is a bounded function defined on the set $\mathcal{A} \times \mathcal{X} \times \mathcal{Y}$. $\mathbb{R}$ with absolute value norm forms a Banach space. By Lemma A.3, the functions that satisfy condition (i) form a Banach space with the supremum norm $\mathcal{F} = (\mathcal{F}, \|\cdot\|_\infty)$. Recall that we assume $h : \mathcal{A} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ satisfying $h(a, x, y) = p_{X|AY}(x \mid a, y)$ (which is fixed and known) is non-negative and finite. Condition (iii) specifies a set of equality constraints, each of which is characterized with a linear combination of points f in the space $\mathcal{F}$. Therefore conditions (i) and (iii) form a set of constraints of an infinite-dimensional linear programming problem (since we focus on the feasible region of such problem, the exact form of the objective function is omitted). Since for any function $f \in \mathcal{F}$ that can be written into a function of the form $g : \mathcal{Y} \to \mathbb{R}$, f trivially satisfies condition (iii), one can easily construct a non-empty compact subset $\mathcal{K} \subset \mathcal{F}$ such that each point $f \in \mathcal{K}$ satisfies conditions (i) and (iii). In other words, the infinite-dimensional linear programming problem has a non-empty convex feasible region, which contains $\mathcal{K}$.

We now prove that we can find a point in this feasible region that satisfies condition (iv) by showing that any point that violates condition (iv) cannot be an extreme point (as defined in Definition A.4) of the feasible region. We have shown that there is a non-empty compact subset $\mathcal{K} \subset \mathcal{F}$, and that $\mathcal{F}$ is a Banach space, and therefore also a locally convex Hausdorff topological vector space. This also implies that $\mathcal{K}$ is closed (since it is a compact subset of a Hausdorff space). By the Krein-Milman Theorem (Theorem A.5), the set of extreme points for $\mathcal{K}$ is non-empty. Notice that for any $f_0 \in \mathcal{K}$ that violates condition (iv), f_0 can be written into a convex combination of f_1 and f_2 ($f_1, f_2 \in \mathcal{K}$, $f_1 \neq f_2$) that also violate condition (iv). For example let $f_i(a, x, \tilde{y}) = g_i(\tilde{y}), i = 0, 1, 2$ (since each f_i violates condition (iv)). Let $g_1(\tilde{y}) = 2 \cdot \mathbb{1}_{(-\infty, 0]}(\tilde{y}) \cdot g_0(\tilde{y})$ and $g_2(\tilde{y}) = 2 \cdot \mathbb{1}_{(0, +\infty)}(\tilde{y}) \cdot g_0(\tilde{y})$ where $\mathbb{1}_\cdot(\cdot)$ is the indicator function (without loss of generality assume that the corresponding $f_1, f_2 \in \mathcal{K}$). It is easy to verify that $f_1 \neq f_2$,

and $\frac{1}{2}f_1 + \frac{1}{2}f_2 = f_0$. Therefore any f_0 that violates condition (iv) cannot be an extreme point of $\mathcal{K}$, which is closed with a non-empty set of extreme points. In other words, there exist a point $f^* \in \mathcal{F}$ such that f^* is an extreme point of $\mathcal{K}$ (therefore f^* is within the feasible region) and that f^* satisfies condition (iv).

Finally, for such f^* within the feasible region, one can find a unique function $\tilde{f} : \mathcal{A} \times \mathcal{X} \times \mathcal{Y}$ such that condition (ii) is satisfied by normalizing over $\tilde{y} \in \mathcal{Y}$ for each possible combination of $(a, x) \in \mathcal{A} \times \mathcal{X}$:

$$\forall a \in \mathcal{A}, x \in \mathcal{X}, \tilde{f}(a, x, \tilde{y}) := \frac{f^*(a, x, \tilde{y})}{\int_{\mathcal{Y}} f^*(a, x, \xi) d\xi}.$$

Therefore, there exists a function $\tilde{f}$ that satisfies conditions (i) to (iv), i.e., Equalized Odds can be non-trivially attained. □

A.4 Proof for Theorem 4.5

Theorem. ***(Condition to Achieve Equalized Odds for Classification with Deterministic Prediction)*** *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 $\widehat{Y}$ be a deterministic function $f : \mathcal{A} \times \mathcal{X} \to \mathcal{Y}$. Let $S_A^{(\hat{y})} := \{a \mid \exists x \in \mathcal{X} \text{ s.t. } f(a, x) = \hat{y}\}$, and $S_{X|a}^{(\hat{y})} := \{x \mid f(a, x) = \hat{y}\}$. Equalized Odds is attained if and only if the following conditions hold true (for continuous X, replace summation with integration accordingly):*

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

(ii) $\forall \hat{y} \in \mathcal{Y}, \; \forall a, a' \in \mathcal{A} :$ $\sum_{x \in S_{X|a}^{(\hat{y})}} P_{X|AY}(x \mid a, y) = \sum_{x \in S_{X|a'}^{(\hat{y})}} P_{X|AY}(x \mid a', y)$

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:

$$\begin{aligned} &\forall a \in \mathcal{A}, y, \hat{y} \in \mathcal{Y} : \\ &P_{\widehat{Y}|AY}(\hat{y} \mid a, y) = P_{\widehat{Y}|Y}(\hat{y} \mid y). \end{aligned} \tag{A.12}$$

Expand the LHS of Equation A.12:

$$\begin{aligned} &P_{\widehat{Y}|AY}(\hat{y} \mid a, y) \\ &= \sum_{x \in \mathcal{X}} P_{\widehat{Y}|AXY}(\hat{y} \mid a, x, y) P_{X|AY}(x \mid a, y), \end{aligned}$$

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

$$\begin{aligned} &P_{\widehat{Y}|AXY}(\hat{y} \mid a, x, y) \\ &= P\big(f(A, X) = \hat{y} \mid A = a, X = x, Y = y\big) \\ &= P\big(f(A, X) = \hat{y} \mid A = a, X = x\big) \in \{0, 1\}. \end{aligned} \tag{A.13}$$

From Equation A.13 we can see that the conditional probability $P_{X|AY}(x \mid a, y)$ can contribute to the summation only when $f(a, x) = \hat{y}$. We can rewrite the LHS of Equation A.12:

$$P_{\widehat{Y}|AY}(\hat{y} \mid a, y) = \sum_{x \in S_{X|a}^{(\hat{y})}} P_{X|AY}(x \mid a, y) := Q^{(\hat{y})}(a, y).$$

Similarly, for the RHS of Equation A.12, we have:

$$\begin{aligned}
&P_{\widehat{Y}|Y}(\hat{y} \mid y) \\
=&\sum_{a\in\mathcal{A}}\sum_{x\in\mathcal{X}} P_{\widehat{Y}|AXY}(\hat{y} \mid a, x, y) P_{A,X|Y}(a, x \mid y) \\
=&\sum_{a\in S_A^{(\hat{y})}}\sum_{x\in S_{X|a}^{(\hat{y})}} P_{X|AY}(x \mid a, y) P_{A|Y}(a \mid y) \\
=&\sum_{a\in S_A^{(\hat{y})}} Q^{(\hat{y})}(a, y) P_{A|Y}(a \mid y).
\end{aligned}$$

Since Equalized Odds holds true if and only if Equation A.12 holds true, then the LHS of the equation does not involve a (as is the case for the RHS), i.e., $Q^{(\hat{y})}(a, y)$ does not change with a. Then Equation A.12 becomes:

$$\begin{aligned}
Q^{(\hat{y})}(a, y) =& \sum_{a\in S_A^{(\hat{y})}} Q^{(\hat{y})}(a, y) P_{A|Y}(a \mid y) \\
=&\ Q^{(\hat{y})}(a, y) \sum_{a\in S_A^{(\hat{y})}} P_{A|Y}(a \mid y),
\end{aligned}$$

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

$$\begin{aligned}
&\forall a, a' \in \mathcal{A}, a \neq a' : \\
&\sum_{x\in S_{X|a}^{(\hat{y})}} P_{X|AY}(x \mid a, y) = \sum_{x\in S_{X|a'}^{(\hat{y})}} P_{X|AY}(x \mid a', y),
\end{aligned}$$

which gives 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 A.12 holds true, i.e., Equalized Odds holds true.

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

A.5 Proof for Theorem 4.7

Theorem. ***(Attainability of Equalized Odds for Classification with Stochastic Prediction)*** *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:*

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

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 ($\forall a \in \mathcal{A}$):

$$\left|P_{\widehat{Y}|AY}(1 \mid a, 1) - P_{\widehat{Y}|AY}(1 \mid a, 0)\right| \geq \epsilon. \tag{A.14}$$

Equation A.14 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[1]. As illustrated in Figure 4.1a and Figure 4.1b, 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.6 Proof for Theorem 4.8

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 $\beta^{(\hat{y})}_{ay} = \sum_{x\in\mathcal{X}} P(\widetilde{Y}_{in} = 1 \mid A = a, X = x)P(X = x \mid A = a, Y = y, \widehat{Y}_{opt} = \hat{y})$.

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 $\gamma_{ay}(\widehat{Y}_{\text{opt}})$:

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

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 $\widehat{Y}_{\text{opt}}$:

$$P_{\widetilde{Y}_{\text{in}}|AY}(1|a, y) = \sum_{\hat{y}\in\mathcal{Y}} P_{\widehat{Y}_{\text{opt}}|AY}(\hat{y}|a, y)\Big[\sum_{x\in\mathcal{X}} P_{\widetilde{Y}_{\text{in}}|AX}(1|a, x) \cdot P_{X|AY\widehat{Y}_{\text{opt}}}(x|a, y, \hat{y})\Big]. \tag{A.15}$$

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

$$\begin{aligned}
\gamma_{ay}(\widetilde{Y}_{\text{in}}) &= \beta^{(0)}_{ay}\ \gamma_{ay}(1 - \widehat{Y}_{\text{opt}}) + \beta^{(1)}_{ay}\ \gamma_{ay}(\widehat{Y}_{\text{opt}}),\\
\beta^{(0)}_{ay} &= \sum_{x\in\mathcal{X}} P_{\widetilde{Y}_{\text{in}}|AX}(1 \mid a, x)P_{X|AY\widehat{Y}_{\text{opt}}}(x \mid a, y, 0),\\
\beta^{(1)}_{ay} &= \sum_{x\in\mathcal{X}} P_{\widetilde{Y}_{\text{in}}|AX}(1 \mid a, x)P_{X|AY\widehat{Y}_{\text{opt}}}(x \mid a, y, 1).
\end{aligned} \tag{A.16}$$

If there is more than one variable in X in Equation A.16, one can expand the summation if needed; if some variables are continuous, one may also substitute the summation with integration accordingly.

[1] 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 A.14 holds true.

From Equation A.16, $\beta_{ay}^{(0)}$ and $\beta_{ay}^{(1)}$ depend on the value of Y:

$$\begin{aligned}\beta_{ay}^{(0)} &= P(\widetilde{Y}_{\text{in}} = 1 \mid A = a, Y = y, \widehat{Y}_{\text{opt}} = 1),\\ \beta_{ay}^{(0)} &= P(\widetilde{Y}_{\text{in}} = 1 \mid A = a, Y = y, \widehat{Y}_{\text{opt}} = 0).\end{aligned} \tag{A.17}$$

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. □