Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews

Samuel Yeom
Carnegie Mellon University
syeom@cs.cmu.edu

Michael Carl Tschantz
International Computer Science Institute
mct@icsi.berkeley.edu

Abstract

We mathematically compare three competing definitions of group-level nondiscrimination: *demographic parity*, *equalized odds*, and *calibration*. Using the theoretical framework of Friedler et al., we study the properties of each definition under various *worldviews*, which are assumptions about how, if at all, the observed data is biased. We prove that different worldviews call for different definitions of fairness, and we specify when it is appropriate to use demographic parity and equalized odds. In addition, we show that calibration is insufficient because it allows an arbitrarily large inter-group disparity. Finally, we define a worldview that is more realistic than the previously considered ones, and we introduce a new notion of fairness that is suitable for this worldview.

1. Introduction

As the field of machine learning has grown to influence many aspects of our lives, so have concerns about the effects of its potential biases. These biases are especially problematic if they disproportionately favor one demographic group over another. Legally, many jurisdictions around the world prohibit discrimination on the basis of a *protected attribute*, such as race or gender. One category of prohibited discrimination is *disparate impact*, or *indirect discrimination*, which means that the use of the model must not result in discriminatory effect even if the discrimination is not intentional. As a result, it is not sufficient for the model to simply avoid using the protected attribute since it may arrive at the same result by using another attribute that is correlated with the protected one.

Many different notions of fairness have been proposed in efforts to better understand the issue of discrimination in machine learning. A common notion is *demographic parity*, which requires that the model give the favorable outcome to both groups of people at equal rates. However, sometimes there are reasons to believe that the model should not give equal outcomes, such as when predicting physical strength for different genders. Because of this, jurisdictions that recognize disparate impact also make exceptions for cases where there is sufficient justification for the discriminatory effect, such as a *business necessity* (e.g., Grover, 1995; Barocas and Selbst, 2016). This motivates moving away from demographic parity to definitions that take the ground truth into account. One such definition, called *equalized odds* by Hardt et al. (2016), requires equal false positive and false negative rates for each protected group. Another commonly used notion of fairness is *calibration* (e.g., Chouldechova, 2017), which roughly corresponds to the requirement of equal positive predictive values. However, one drawback of equalized odds and calibration is that the "ground truth" may be tainted by past discrimination, in which case consulting the ground truth will help perpetuate the discrimination.

In this work, we handle the issue of biased ground truth by adopting the framework of Friedler et al. (2016), who make the distinction between the observed ground truth and the *construct*, which is the attribute that is truly relevant for prediction. Using this framework, we define what it means for a model to be *discriminative*, that is, able to accurately predict the construct, and *discriminatory*, that is, disproportionately favoring one protected group over another in an unjustified way. Our definition of discrimination deals with the disparity in positive classification rates, which is a widely accepted measure of discriminatory effect in both law (Equal Employment Opportunities Commission, 1978) and computer science (Calders et al., 2009; Calders and Verwer, 2010; Kamishima et al., 2012; Zemel et al., 2013; Feldman et al., 2015; Zafar et al., 2017b). Unlike most prior definitions, our definition also considers the benefit of having a discriminative model, stipulating that a disparity in the output of the model is justified by a commensurate disparity in the construct. We note that we do not address other aspects of discrimination, such as intentional discrimination (Barocas and Selbst, 2016, §II-A).

Because the construct is usually unmeasurable, Friedler

et al. introduce and analyze two assumptions, or *worldviews*, about the construct. More specifically, they define the We're All Equal (WAE) worldview, under which there is no association between the construct and the protected attribute, and the WYSIWYG worldview, under which the observations accurately reflect the construct. We tie in the previously defined notions of fairness with these worldviews, arguing that demographic parity is appropriate for the WAE worldview and that equalized odds is appropriate for the WYSIWYG worldview. We also show that calibration does not impose any restrictions on the extent to which a model discriminates. Since equalized odds and calibration are incompatible (Darlington, 1971; Chouldechova, 2017; Kleinberg et al., 2017), our result is an argument for the use of equalized odds instead of calibration. Furthermore, we compare our approach to that of Zafar et al. (2017a) in their work on *disparate mistreatment*, or disparate misclassification rates, showing that our definition of nondiscrimination can be modified to apply in their setting.

Although the WAE and WYSIWYG worldviews are useful for theoretical analysis, they are unlikely to be true in practice. To remedy this issue, we introduce a family of hybrid worldviews that is parametrized by a measure of how biased the observed data is against a protected group of people. This allows us to model many real-world situations by simply adjusting the parameter. We then create a parametrized fairness definition that is suitable for the new family of worldviews, showing how one can apply the analysis in our paper to real-world scenarios.

Our most fundamental contribution is the introduction of a framework in which to evaluate proposed definitions of fairness. We do not claim that the definition of nondiscrimination that we use in this paper captures the only relevant notion of nondiscrimination. Rather, we view our definition as more of a diagnostic than a goal in and of itself. Indeed, we do not provide an algorithm for ensuring that a model complies with our definition of nondiscrimination since, in our view, doing so would be treating the symptom rather than the cause. Such algorithms can eliminate one aspect of discrimination, but may in the process create a model that is obviously discriminatory from another angle. When a model does not satisfy a definition of nondiscrimination, it should be a starting point for investigation as to why. While it could be that the model itself is corrupt, it could also be due to a mismatch between the construct and the observed data, or a need for better features. We remind the reader that no algorithm can fix all ills.

2. Related Work

Barocas and Selbst (2016) discuss in detail the potential legal issues with discrimination in machine learning. One widely consulted legal standard for detecting disparate impact is the *four-fifths rule* (Equal Employment Opportunities Commission, 1978). The four-fifths rule is a guideline that checks whether the ratio of the rates of favorable outcomes for different demographic groups is at least four-fifths. This guideline can be considered a relaxation of demographic parity, which would instead require that the ratio of the positive classification rates be exactly one.

The four-fifths rule has inspired the work of Feldman et al. (2015) and Zafar et al. (2017b), who deal with a generalization of the four-fifths rule, called the $p\%$ *rule*, in their efforts to remove disparate impact. On the other hand, many others (Calders et al., 2009; Calders and Verwer, 2010; Kamishima et al., 2012; Zemel et al., 2013) consider the difference, rather than the ratio, of the positive classification rates. Our definition of nondiscrimination is a generalization of this difference-based measure, but it differs from the others in that it uses the construct rather than the observed data.

Friedler et al. (2016) introduced the concept of the construct in fair machine learning. Although they also use the construct in their definition of nondiscrimination, their definition uses the Gromov–Wasserstein distance and as a result is more difficult to compute and reason about. One benefit of their approach is that it enables their treatment of fairness at both an individual level and a group level. By contrast, we adopt a definition that is specialized for the consideration of group nondiscrimination only.

Other works in the field of fair machine learning deal with aspects of discrimination that are not well described by positive classification rates. Datta et al. (2017) tackle the issue that some parts of a model could be discriminatory even if the model, when taken as a whole, does not appear to have discriminatory effect. Zafar et al. (2017a) point out a that a model can have a higher misclassification rate for one protected group than another, and they propose a method for mitigating this form of discrimination. Hardt et al. (2016) characterize nondiscrimination through *equalized odds*, which requires that two measures of misclassification, false positive and false negative rates, be equal for all protected groups. Finally, *calibration*, Chouldechova (2017) points out, is widely accepted in the "educational and psychological testing and assessment literature". In this paper, we prove that equalized odds, but not calibration, is sometimes a useful way to satisfy our definition of nondiscrimination. We refer the reader to a survey by Romei and Ruggieri (2014) for a discussion of other measures of discrimination.

As mentioned previously, discriminatory effect can be justified if there is a sufficient reason. For prediction tasks, it is natural to think of accuracy as a sufficient justification. Zafar et al. (2017b) handle this by solving an optimization problem to maximize fairness subject to some ac-

curacy constraints. This reflects the idea that a classifier is justified in sacrificing fairness for accuracy. To a lesser extent, equalized odds and calibration can also be thought of as motivated by the dual desires for accuracy and fairness. Our approach to justification is also motivated by these desires, but we use the construct and say that a classifier is justified in predicting the construct correctly.

3. Notation

In the framework introduced by Friedler et al. (2016), there are three spaces that describe the target attribute of a prediction model. The *construct space* represents the value of the attribute that is truly relevant for the prediction task. This value is usually unmeasurable, so prediction models in a supervised learning problem are instead trained with a related measurable label, whose values reside in the *observed space*. Finally, the *prediction space* (called *decision space* by Friedler et al.) describes the output of the model. We will use Y', Y, and $\hat{Y}$ as the random variables representing values from the construct, observed, and prediction spaces, respectively. (See Figure 1.)

In addition, we will use Z to denote the protected attribute at hand, and we will assume that $Z \in \{0, 1\}$. For example, if Z is gender, the values 0 and 1 could represent male and female, respectively. Although the input features $X = (X_1, \ldots, X_n)$ are also critical for both the training and the prediction of the model, they are rarely used in this paper.

Example 1. Some jurisdictions have started to use machine learning models to predict how much risk a criminal defendant poses (Liptak, 2017). Judges are then allowed to consider the risk score as one of many factors when making bail or sentencing decisions (Supreme Court of Wisconsin, 2016). Using the three-space framework of Friedler et al. (2016), we can represent the risk score output by the model as $\hat{Y}$. The model would be trained with the observation Y, which in this case may be recorded data about past criminal defendants and their failures to appear in court (bail) or recidivism (sentencing). These models would also be trained with features X from the input space, such as age and criminal history.

For sentencing decisions, presumably we want to know whether the defendant will commit another crime in the future, regardless of whether the defendant will be caught committing the crime. Therefore, we argue that the recorded recidivism rate Y is merely a proxy for the actual reoffense rate Y', which is the relevant attribute for the prediction task. There is evidence that black Americans are arrested at a higher rate than white Americans for the same crime (Mueller, 2018), so it is reasonable to suspect that Y is a racially biased proxy for Y'.

Example 2. Universities want the students that they admit to the university to be successful in the university (Y'). Because success is a vague term that encompasses many factors, a model that predicts success in university would instead be trained with a more concrete measure, such as graduating within six years (Y). This model may take inputs such as a student's high-school grades and standardized test scores (X), and will output a prediction of how likely the student is to graduate within six years ($\hat{Y}$). Admissions officers can then use this prediction to guide their decision about whether to admit the student.

It is important to note that the models in the above examples do not make the final decision and that human judgments are a major part of the decision process. However, we are concerned about the fairness of the model rather than that of the entire decision process. Thus, we focus on $\hat{Y}$, the output of the model, rather than the final decision made using it. In addition, although the use of the protected attribute Z is generally prohibited, it is sometimes allowed for *affirmative action*, or *positive discrimination*, that aims to correct a demographic disparity. For example, many U.S. universities take the race and gender of the students into account with the goal of admitting a diverse group of students. We limit the scope of our work to the case where the model is not trained with affirmative action or diversity as one of its objectives. The human-led decision process that uses $\hat{Y}$ may still consider these objectives.

4. Preliminary Definitions

In this work, we use the total variation distance to measure the extent to which two categorical random variables differ.

Definition 1 (Total Variation Distance)**.** *Let Y_0 and Y_1 be categorical random variables with finite support $\mathcal{Y}$. Then, the* total variation distance *between Y_0 and Y_1 is*

$$d_{\mathrm{tv}}(Y_0, Y_1) = \frac{1}{2} \sum_{y \in \mathcal{Y}} \Big| \Pr[Y_0{=}y] - \Pr[Y_1{=}y] \Big|.$$

In the special case where $Y_0, Y_1 \in \{0, 1\}$, the total variation distance can also be expressed as $|\Pr[Y_0{=}1] - \Pr[Y_1{=}1]|$.

4.1. Fairness Definitions as Empirical Tests

Many definitions of fairness for prediction models have been proposed previously, and here we restate three of them. In all three definitions, the probabilities are taken over random draws of data points from the data distribution, as well as any randomness used by the model.

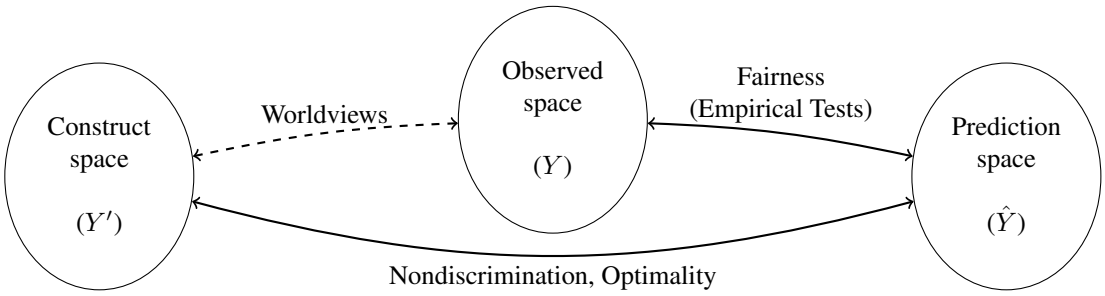

Figure 1. Three relevant spaces for prediction models. The space of input features $X = (X_1, \ldots, X_n)$ is not depicted here. The observed space and the prediction space are measurable, and the fairness definitions, which are interpreted as empirical tests (Definitions 2, 3, 4), impose constraints on the relationship between the two spaces. On the other hand, the construct space is usually unmeasurable, so we must assume a particular worldview (e.g., Worldview 1 or 2) about how the construct space relates to the observed space, if at all. Then, we can characterize nondiscrimination and optimality, which relate the construct space to the prediction space.

Definition 2 (Demographic Parity Test)**.** *A model passes the* demographic parity test *if, for all $\hat{y}$,*

$$\Pr[\hat{Y}{=}\hat{y} \mid Z{=}0] = \Pr[\hat{Y}{=}\hat{y} \mid Z{=}1].$$

Definition 3 (Equalized Odds Test (Hardt et al., 2016))**.** *A model passes the* equalized odds test *if, for all y and $\hat{y}$,*

$$\Pr[\hat{Y}{=}\hat{y} \mid Y{=}y, Z{=}0] = \Pr[\hat{Y}{=}\hat{y} \mid Y{=}y, Z{=}1].$$

Definition 4 (Calibration Test (Chouldechova, 2017))**.** *A model passes the* calibration test *if, for all y and $\hat{y}$,*

$$\Pr[Y{=}y \mid \hat{Y}{=}\hat{y}, Z{=}0] = \Pr[Y{=}y \mid \hat{Y}{=}\hat{y}, Z{=}1].$$

Because much of the previous work does not make the distinction between the construct space and the observed space, there is some ambiguity about whether Y' or Y is the appropriate variable to use these definitions. In this paper, we interpret these definitions to be empirical tests that can verify whether a model is fair. As a result, none of the above definitions include the construct Y'.

Although all of the above definitions were created for the purpose of characterizing nondiscrimination, we will refer to them as *fairness definitions* or *empirical tests* in order to clearly contrast them with the definition of nondiscrimination that we later introduce.

4.2. Worldviews

Our intuitive notion of nondiscrimination involves the relationship between the construct space and the prediction space. For example, consider the context of recidivism prediction described in Example 1. Suppose that one group of people is much more likely to be arrested for the same crime than another group. Then, the disparity in arrest rates can cause the recorded recidivism rate Y to be biased, and a model trained using such Y would likely learn to discriminate as a result. If in fact the two groups have equal reoffense rates Y', it would hardly be considered justified that one group tends to be given longer sentences as a result of the bias in Y.

However, because Y' is typically unmeasurable, in practice we do not know whether Y' is the same for both groups. Therefore, to reason about nondiscrimination using the construct space, we must make assumptions about the construct space. Two such assumptions, or *worldviews*, have previously been introduced by Friedler et al. (2016) and are described below. Our versions of these worldviews are simpler than the original because they are exact, whereas the original versions allow deviations by some parameter ϵ.

Worldview 1 (We're All Equal)**.** *Under the* We're All Equal (WAE) *worldview, every group is identical with respect to the construct space. More formally, Y' is independent of Z, i.e., $Y' \perp Z$.*

Worldview 2 (WYSIWYG)**.** *Under the* What You See Is What You Get (WYSIWYG) *worldview, the observed space accurately reflects the construct space. More formally, $Y' = Y$.*

5. Discriminatory Association

In this section, we introduce our definition of nondiscrimination and use it to analyze the suitability of existing fairness definitions under different worldviews. We first begin with case where Y' and $\hat{Y}$ are categorical, and in the appendix we show how to generalize the definition to numerical Y'.

Our definition of nondiscrimination is motivated by the following two concerns: First, we want the model to be *discriminative*, i.e., the output of the model should accurately reflect the value of Y'. We formalize this goal with the following definition of optimality.

Definition 5 (Optimality)**.** *We say that a model is* optimal

if its output $\hat{Y}$ and the construct Y' are always equal.

It is important to note that enforcing a fairness definition does not necessarily result in an optimal model. In fact, because the construct Y' is usually unmeasurable, it is likely impossible in practice to train a model that is optimal under our definition. However, this definition is sufficient for our purposes because we simply use it to argue that a fairness definition should not preclude a perfectly predictive model.

Second, we do not want the model to have a *discriminatory* effect. When $\hat{Y}$ is binary, the size of the discriminatory effect is commonly measured by $|\Pr[\hat{Y}{=}1|Z{=}0] - \Pr[\hat{Y}{=}1|Z{=}1]|$, or the difference in positive classification rates. Definition 6 is a generalization of this measure for the case of non-binary categorical $\hat{Y}$.

Definition 6 (Output Disparity)**.** *Let the output $\hat{Y}$ of a model be categorical. The* output disparity *of the model is the quantity $d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1)$.*

However, not all output disparities are bad. In particular, because we want the model to be discriminative, we allow an output disparity insofar as it can be explained by the inter-group disparity in Y'. This happens when

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) \leq d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1). \quad (1)$$

Since a model can have issues with discrimination that are not well characterized by output disparity (discussed below), Equation 1 is not the conclusive definition of nondiscrimination. Therefore, we use the logical negation of Equation 1 as the definition of one particular form of discrimination, which occurs when an output disparity is *not* explained by Y'.

Definition 7 (Discriminatory Association)**.** *Let Y' and $\hat{Y}$ be categorical. Then, a model exhibits discriminatory association that is unexplainable by Y' (DAU-Y') if*

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) > d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1). \quad (2)$$

Note that the equality in Equation 1 holds for every optimal model. In other words, an optimal model displays the maximum amount of output disparity allowed by Definition 7. On the other hand, if the output disparity is greater than the disparity in Y', the model must be discriminating in a way that cannot be explained by Y'.

Of course, there are forms of discrimination that are not well described by output disparity alone. For example, a model could have a higher misclassification rate for one group of people (Zafar et al., 2017a), and Definition 7 is not well suited for detecting such errors. In addition, even if Definition 7 does not show a violation (i.e., Equation 2 does not hold) for the entire model, it is possible that some part of the model is a proxy for the protected attribute and that it causes a discriminatory effect. In their work on proxy use, Datta et al. (2017) show that the input/output behavior of the model does not give enough information to decide whether a model uses a proxy of the protected attribute. As a result, we would have to look at the internal details of the model to determine whether any part of the model is discriminatory. On the other hand, Definition 7 is intended to be a general, model-agnostic way to incriminate, but not necessary absolve, a model.

Table 1. Summary of the results in Section 5. We say that a fairness definition is appropriate for a worldview if it precludes DAU-Y' (Definition 7) but does not preclude a perfectly predictive model. The demographic parity test is appropriate when the We're All Equal (WAE) worldview holds, but otherwise the resulting models are necessarily suboptimal. The equalized odds test is appropriate when the WYSIWYG worldview holds, but otherwise it does not effectively prevent discrimination. Finally, regardless of the worldview, the calibration test does not effectively prevent discrimination. Here, we assume that WAE and WYSIWYG do not hold simultaneously.

	We're All Equal (Worldview 1)	WYSIWYG (Worldview 2)
Dem. Parity (Definition 2)	✔ Theorem 1	Suboptimal Theorem 2
Eq. Odds (Definition 3)	Possibly discrim. Theorem 4	✔ Theorem 3
Calibration (Definition 4)	Possibly discriminatory Theorem 5	

We are now ready to argue about the suitability of an empirical test for a particular worldview. If an empirical test disallows an optimal model, we can conclude that the test is too strict in a way that lowers the utility of the model. On the other hand, if a test does not guarantee the lack of DAU-Y', it is insufficient as an anti-discrimination measure. Therefore, to argue that an empirical test is appropriate, we will prove the following two statements: (a) Every model that passes the empirical test does not have DAU-Y', and (b) every optimal model passes the empirical test.

We apply this reasoning to demographic parity (Definition 2) and equalized odds (Definition 3), showing that these empirical tests are appropriate for the WAE and WYSIWYG worldviews, respectively. More formally, we will prove that the above statements are true for every joint distribution of Y', Y, $\hat{Y}$, and Z that is consistent with the worldview. Table 1 summarizes these results.

5.1. Demographic Parity and WAE

Theorem 1. *A model that passes the demographic parity test does not have DAU-Y' under Definition 7. Moreover, if the WAE worldview holds, every optimal model satisfies*

demographic parity.

Proof. By the definition of demographic parity, the left-hand side of Equation 2 is $d_{\mathrm{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) = 0$. Since the total variation distance is always nonnegative, demographic parity ensures the lack of DAU-Y'.

If the WAE worldview holds, we have $Y' \perp Z$, so every optimal model satisfies $\hat{Y} \perp Z$. This implies demographic parity by Definition 2. □

The first part of Theorem 1 shows that we can guarantee that a model will not have DAU-Y' by training it to pass the demographic parity test. However, this does not mean that demographic parity is appropriate for every situation. First, we remind the reader that the lack of DAU-Y' does not mean that the model will be free of all issues related to discrimination. In particular, DAU-Y' is only designed to catch the type of discrimination akin to *disparate impact*. If the WAE worldview holds, demographic parity is the only way to avoid DAU-Y', so it makes sense to enforce demographic parity. On the other hand, if the WAE worldview does not hold, enforcing demographic parity may introduce other forms of discrimination. For example, the U.S. Supreme Court held in *Ricci v. DeStefano* (2009) that the prohibition against intentional discrimination can sometimes override the consideration of disparate impact, ruling that an employer unlawfully discriminated by discarding the results of a bona fide job-related test simply because of a racial performance gap.

Second, demographic parity can unnecessarily lower the utility of a model. In fact, Theorem 2 shows that, if the joint distribution of Y', Y, $\hat{Y}$, and Z is not consistent with the WAE worldview, any model that satisfies demographic parity must be suboptimal.

Theorem 2. *If the WAE worldview does not hold, no optimal model satisfies demographic parity.*

Proof. If the WAE worldview does not hold, $d_{\mathrm{tv}}(Y'|Z{=}0, Y'|Z{=}1)$, the right-hand side of Equation 2, is positive. Therefore, the left-hand side $d_{\mathrm{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1)$ must be positive for an optimal model. On the other hand, if a model satisfies demographic parity, the left-hand side must be zero. Therefore, no optimal model can satisfy demographic parity. □

Theorem 1 and 2 demonstrate that the demographic parity test is best suited for a setting where the WAE worldview holds.

5.2. Equalized Odds and WYSIWYG

We now argue that a similar relationship exists between the equalized odds test and the WYSIWYG worldview.

Theorem 3. *If the WYSIWYG worldview holds, a model that passes the equalized odds test does not have DAU-Y' under Definition 7. Moreover, if the WYSIWYG worldview holds, every optimal model satisfies equalized odds.*

Proof. Let $\mathcal{Y}'$ and $\hat{\mathcal{Y}}$ be the supports of Y' and $\hat{Y}$, respectively. Applying the WYSIWYG worldview to the definition of equalized odds, we get $\Pr[\hat{Y}{=}\hat{y} \mid Y'{=}y', Z{=}0] = \Pr[\hat{Y}{=}\hat{y} \mid Y'{=}y', Z{=}1] = \Pr[\hat{Y}{=}\hat{y} \mid Y'{=}y']$ for all $y' \in \mathcal{Y}'$ and $\hat{y} \in \hat{\mathcal{Y}}$. Therefore, we have

$$
\begin{aligned}
&d_{\mathrm{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) \\
&= \tfrac{1}{2} \textstyle\sum_{\hat{y}\in\hat{\mathcal{Y}}} \big|\Pr[\hat{Y}{=}\hat{y} \mid Z{=}0] - \Pr[\hat{Y}{=}\hat{y} \mid Z{=}1]\big| \\
&= \tfrac{1}{2} \textstyle\sum_{\hat{y}\in\hat{\mathcal{Y}}} \Big|\sum_{y'\in\mathcal{Y}'} \Pr[\hat{Y}{=}\hat{y} \mid Y'{=}y'] \\
&\qquad\qquad \cdot \big(\Pr[Y'{=}y' \mid Z{=}0] - \Pr[Y'{=}y' \mid Z{=}1]\big)\Big| \\
&\le \tfrac{1}{2} \textstyle\sum_{\hat{y}\in\hat{\mathcal{Y}}} \sum_{y'\in\mathcal{Y}'} \Pr[\hat{Y}{=}\hat{y} \mid Y'{=}y'] \\
&\qquad\qquad \cdot \big|\Pr[Y'{=}y' \mid Z{=}0] - \Pr[Y'{=}y' \mid Z{=}1]\big| \\
&= \tfrac{1}{2} \textstyle\sum_{y'\in\mathcal{Y}'} \Big(\big|\Pr[Y'{=}y' \mid Z{=}0] - \Pr[Y'{=}y' \mid Z{=}1]\big| \\
&\qquad\qquad \cdot \textstyle\sum_{\hat{y}\in\hat{\mathcal{Y}}} \Pr[\hat{Y}{=}\hat{y} \mid Y'{=}y']\Big) \\
&= \tfrac{1}{2} \textstyle\sum_{y'\in\mathcal{Y}'} \big|\Pr[Y'{=}y' \mid Z{=}0] - \Pr[Y'{=}y' \mid Z{=}1]\big| \\
&= d_{\mathrm{tv}}(Y'|Z{=}0, Y'|Z{=}1).
\end{aligned}
$$

This concludes the proof of the first statement.

For an optimal model, we have $\hat{Y} = Y' = Y$ by the WYSIWYG worldview. Because Y fully determines the value of $\hat{Y}$, it follows from Definition 3 that every optimal model satisfies equalized odds. □

We conclude this section with the following theorem, which states that equalized odds does not guarantee nondiscrimination unless the WYSIWYG worldview holds. We omit the proof due to its simplicity, but this result is consistent with our intuition that an output disparity is unjustified if it is due to a bias in the observation process.

Theorem 4. *If the WYSIWYG worldview does not hold, a model with equalized odds can still have DAU-Y'.*

5.3. Insufficiency of Calibration

Switching Y' and $\hat{Y}$ in the proof of Theorem 3 shows that the calibration test, when combined with the WYSIWYG worldview, ensures that $d_{\mathrm{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) \ge d_{\mathrm{tv}}(Y'|Z{=}0, Y'|Z{=}1)$. The inequality here is in the opposite direction of that in Equation 1, so the calibration test does not place any upper bound on the output disparity and guarantees that it is at least as large as can be explained by Y'. In fact, the following theorem shows that, regardless of the worldview, even a model with almost the maximum output disparity can still pass the calibration test.

Theorem 5. *Let Y be a categorical random variable with finite support such that $\Pr[Y{=}y \mid Z{=}z]$ is positive for all y and z. Then, for any sufficiently small $\epsilon > 0$, there exists a model that passes the calibration test such that $d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) = 1 - \epsilon$.*

Proof. The main idea behind the proof is that the model simply outputs the value of Z. However, because calibration is not well-defined if $\Pr[\hat{Y}{=}\hat{y}, Z{=}z] = 0$ for any $\hat{y}$ and z, we must allow the model to output the other value with some very small probability. More specifically, we construct a model such that

$$\Pr[\hat{Y}{=}\hat{y} \mid Z{=}z] = \begin{cases} 1 - \frac{\epsilon}{2}, & \text{if } \hat{y} = z \\ \frac{\epsilon}{2}, & \text{if } \hat{y} \neq z. \end{cases}$$

We can choose which values our constructed model outputs, so assume without loss of generality that $\hat{Y} \in \{0, 1\}$.

Let $\mathcal{Y}$ be the support of Y. By the calibration test, we have $\Pr[Y{=}y \mid \hat{Y}{=}\hat{y}, Z{=}0] = \Pr[Y{=}y \mid \hat{Y}{=}\hat{y}, Z{=}1] = \Pr[Y{=}y \mid \hat{Y}{=}\hat{y}]$ for all $y \in \mathcal{Y}$ and $\hat{y} \in \{0, 1\}$. Let $p_{y\hat{y}} = \Pr[Y{=}y \mid \hat{Y}{=}\hat{y}]$. Our goal is to find the values of p_{y0} and p_{y1} that are consistent with the fixed observed probabilities $\Pr[Y{=}y \mid Z{=}0]$ and $\Pr[Y{=}1 \mid Z{=}1]$.

By the law of total probability, our model must satisfy

$$\begin{pmatrix} \Pr[Y{=}y \mid Z{=}0] \\ \Pr[Y{=}y \mid Z{=}1] \end{pmatrix} = \begin{pmatrix} 1 - \frac{\epsilon}{2} & \frac{\epsilon}{2} \\ \frac{\epsilon}{2} & 1 - \frac{\epsilon}{2} \end{pmatrix} \begin{pmatrix} p_{y0} \\ p_{y1} \end{pmatrix}.$$

Solving for p_{y0} and p_{y1}, we see that they converge to $\Pr[Y{=}y \mid Z{=}0]$ and $\Pr[Y{=}y \mid Z{=}1]$, respectively, as ϵ approaches zero. By assumption, these probabilities are positive. Since $\mathcal{Y}$ is finite, this means that there exists a small enough $\epsilon > 0$ such that $p_{y0}, p_{y1} > 0$ for all $y \in \mathcal{Y}$. Moreover, it is easy to verify that $\sum_{y \in \mathcal{Y}} p_{y0} = \sum_{y \in \mathcal{Y}} p_{y1} = 1$, making them valid probability distributions.

Now, when given $Y{=}y$ and $Z{=}z$, our model can output $\hat{Y}{=}\hat{y}$ with probability

$$\Pr[\hat{Y}{=}\hat{y} \mid Y{=}y, Z{=}z] = \frac{p_{y\hat{y}} \cdot \Pr[\hat{Y}{=}\hat{y} \mid Z{=}z]}{\Pr[Y{=}y \mid Z{=}z]},$$

where $\Pr[\hat{Y}{=}\hat{y} \mid Z{=}z]$ is either $\frac{\epsilon}{2}$ or $1 - \frac{\epsilon}{2}$ depending on whether $\hat{y} = z$. □

Because the calibration test allows models, such as the one we constructed in the above proof, that are clearly discriminatory, it is unsuitable for ensuring nondiscrimination as characterized by output disparity. As a result, in the rest of the paper we focus on the equalized odds test rather than the calibration test.

5.4. Connection to Misclassification

Here, we show that our definition of nondiscrimination is closely related to the that given by Zafar et al. (2017a) in their treatment of disparate misclassification rates. First, we motivate the issue of disparate misclassification rates with an example. Let Y' and Z be independent and uniformly random binary variables. If $\hat{Y} = Y' \oplus Z$, where $\oplus$ is the bitwise XOR, both protected groups are given the positive label exactly half of the time, so there is no output disparity. However, one group always receives the correct classification and the other always receives the incorrect classification, so the disparity in the misclassification rates is as large as it can be. This shows that a lack of DAU-Y' does not imply a lack of disparity in misclassification rates.

Conversely, a lack of disparity in misclassification rates does not imply a lack of DAU-Y'. To see this, modify the above example so that $\hat{Y} = Z$ instead. In this case, both groups have half of its members misclassified since Z is independent of Y', so they have the same overall misclassification rate. On the other hand, we have $d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1) = d_{\text{tv}}(Y', Y') = 0$ and $d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) = d_{\text{tv}}(Z|Z{=}0, Z|Z{=}1) = 1$. Thus, $\hat{Y}$ has DAU-Y'.

However, we can still find a connection between misclassification parity and DAU-Y'. Let $C = \mathbb{1}(Y' = \hat{Y})$, and replace $\hat{Y}$ with C in the definition of output disparity (Definition 6). Since C is binary, the resulting expression $d_{\text{tv}}(C|Z{=}0, C|Z{=}1)$ is simply the difference in the misclassification rates. We would like to compare this value to some measure of disparity in the construct space. Since our standard measure of $d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1)$ does not necessarily justify inter-group differences in C, it may not be a correct measure to use. Exploring what measures provide justification for disparate misclassification rates is interesting future work.

6. Hybrid Worldviews

So far, we have assumed either the WAE or the WYSIWYG worldview. While these worldviews are interesting from a theoretical perspective, in practice it is unlikely that these worldviews hold. For example, for recidivism prediction as described in Example 1, there are reasons to believe that Y is tainted by past discrimination, rendering the WYSIWYG worldview unsuitable. On the other hand, the recorded recidivism rate Y is still strongly related to the actual reoffense rate Y', so it is plausible that at least some of the racial disparity in Y is also present in the actual reoffense rate Y'. As a result, the WAE worldview is also unsuitable.

In this section, we propose a family of more realistic worldviews for the case where Y' and Y are categorical. As we have depicted in Figure 1, worldviews describe the relation-

ship between the construct and observed spaces. Because our definition of DAU-Y' has to do with inter-group disparities, here we focus specifically on the inter-group disparities in Y' and Y. Note that the WAE worldview has the effect of assuming that none of the disparity in Y is explained by Y'. By contrast, under the WYSIWYG worldview, all of the disparity in Y is explained by Y'. Described below is the α-Hybrid worldview, which is a family of worldviews that occupy the space between the two extremes of WAE and WYSIWYG.

Worldview 3 (α-Hybrid)**.** *Let $\alpha \in [0, 1]$. Under the α-*Hybrid *worldview, exactly α fraction of the disparity in Y is explained by Y'. More formally,*

$$d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1) = \alpha \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1) \quad (3)$$

It is easy to see that the WAE worldview is equivalent to the 0-Hybrid worldview. On the other hand, the relationship between the WYSIWYG and 1-Hybrid worldviews is only unidirectional. Although the WYSIWYG worldview implies the 1-Hybrid worldview, there are plenty of ways to satisfy $d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1) = d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1)$ even when the equality $Y' = Y$ does not hold. If we wanted to make the relationship bidirectional, we could instead have assumed that Y' can be broken down into two components, one of which satisfies WAE and the other WYSIWYG. However, this would mean that every component of Y' is either equal with respect to Z (WAE) or measurable (WYSIWYG), whereas in practice many inter-group disparities in the construct space are not easily measurable. Therefore, to make the α-Hybrid worldview more realistic, we sacrifice one direction of the relationship between the WYSIWYG and 1-Hybrid worldviews.

Now we introduce the α-disparity test and prove that it is suitable for use with the α-Hybrid worldview. Unlike the demographic parity and equalized odds tests, the α-disparity test is parametrized and therefore can be applied to various real-world situations.

Definition 8 (α-Disparity Test)**.** *A model passes the α-*disparity test *if*

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) \leq \alpha \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1). \quad (4)$$

Theorem 6. *If the α-Hybrid worldview holds, a model that passes the α-disparity test does not have DAU-Y' under Definition 7. Moreover, if the α-Hybrid worldview holds, every optimal model satisfies the α-disparity test.*

Proof. To prove the first part of the theorem, we simply combine the inequality guaranteed by the α-disparity test (Equation 4) with the equation that defines the α-Hybrid worldview (Equation 3). Then, we get

$$\begin{aligned} d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) &\leq \alpha \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1) \\ &= d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1), \end{aligned}$$

which is what we want.

For the second part of the theorem, an optimal model has $Y' = \hat{Y}$, so we can substitute the Y' in Equation 3 with $\hat{Y}$ to get

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) = \alpha \cdot d_{\text{tv}}(Y|Z{=}0, Y|Z{=}1).$$

This is simply the equality in Equation 4, so we are done. □

The α-disparity test is closely related to demographic parity and equalized odds. 0-disparity is satisfied if and only if the output disparity is zero, so it is equivalent to demographic parity. In addition, we can easily adapt the proof of Theorem 3 to show that equalized odds implies 1-disparity. However, because equalized odds imposes a condition for each possible value of Y, 1-disparity does not imply equalized odds. Although it may thus seem that equalized odds is stronger and better than 1-disparity, recent results by Corbett-Davies and Goel (2018) show that the threshold rule, which they argue is optimal, does not lead to equalized odds in general. Therefore, there is a trade-off between the stronger fairness guarantee provided by equalized odds and the higher utility that is attainable under 1-disparity. Of course, the 1-disparity test has the additional benefit that it can be generalized to other values of α.

7. Conclusion

We showed that demographic parity and equalized odds are related and that the difference between them boils down to one's worldview. In addition, we proved that calibration allows a model with an arbitrarily large output disparity regardless of the worldview, arguing that calibration should not be used to enforce nondiscrimination as characterized by output disparity.

Our work differs from much of the prior work in that we consider the construct as separate from the observed data. In particular, we interpreted the existing fairness definitions as acting on the observed data, whereas discrimination was viewed as a property of the construct. This bifurcation allowed us to handle the following issues simultaneously: (a) Disparate impact can be justified by a sufficiently good reason such as a business necessity, but (b) due to past discrimination, the observed data can be biased in an unjustified way. It is the second of these points that motivates our use of worldviews to characterize how biased the observed data is.

To illustrate how this might work in practice, let us revisit the examples in Section 3. In Example 1, there are reasons to believe that the observed recidivism rate is a racially bi-

ased measurement of the actual reoffense rate. In Example 2, for various socioeconomic reasons, some protected groups may have disproportionately many people who take longer than six years to graduate but are eventually considered successful in the university. Therefore, the α-Hybrid worldview best characterizes these real-world scenarios, and the value of α reflects one's beliefs about how much more biased the observed data is than the construct. The creator of a model could then use the α-disparity test to make sure that the model does not have a discriminatory effect that is unexplainable by the construct.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1704985. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104:671–732, 2016.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

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

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

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.

Richard B Darlington. Another look at "cultural fairness". *Journal of Educational Measurement*, 8(2):71–82, 1971.

Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Proxy discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120*, 2017.

Equal Employment Opportunities Commission. Uniform guidelines on employee selection procedures. 29 CFR Part 1607, 1978.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

Susan S Grover. The business necessity defense in disparate impact discrimination cases. *Georgia Law Review*, 30:387–430, 1995.

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

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science*, pages 43:1–43:23, 2017.

Adam Liptak. Sent to prison by a software program's secret algorithms. *The New York Times*, 2017.

Benjamin Mueller. Using data to make sense of a racial disparity in NYC marijuana arrests. *The New York Times*, 2018.

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

Supreme Court of the United States. *Ricci v. DeStefano*. 557 U.S. 557, 2009.

Supreme Court of Wisconsin. *State v. Loomis*. 881 N.W.2d 749, 2016.

Cédric Villani. *Optimal transport: old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften: Comprehensive Studies in Mathematics*. Springer-Verlag Berlin Heidelberg, 2008.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, pages 1171–1180, 2017a.

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

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

A. Stronger Definition of Nondiscrimination

In this section, we strengthen our definition of nondiscrimination in a way that is applicable even in the case where Y' is not categorical.

A.1. Additional Preliminary Definitions

When dealing with numerical random variables, we want our notion of distance to take into account the magnitude of the difference in the numerical values. We will thus use the earthmover distance throughout this section. The following definition assumes that the random variables are continuous, but a similar definition is applicable when they are discrete.

Definition 9 (Earthmover Distance)**.** *Let Y_0 and Y_1 be continuous numerical random variables with probability density functions p_0 and p_1 defined over support $\mathcal{Y}$. Furthermore, let Γ be the set of joint probability density functions $\gamma(u,v)$ such that $\int_{\mathcal{Y}} \gamma(u,v)\,dv = p_0(u)$ for all $u \in \mathcal{Y}$ and $\int_{\mathcal{Y}} \gamma(u,v)\,du = p_1(v)$ for all $v \in \mathcal{Y}$. Then, the* earthmover distance *between Y_0 and Y_1 is*

$$d_{\mathrm{em}}(Y_0, Y_1) = \inf_{\gamma\in\Gamma} \int_{\mathcal{Y}} \int_{\mathcal{Y}} \gamma(u,v)\, d(u,v)\, du\, dv,$$

where d is a distance metric defined over $\mathcal{Y}$.

The joint probability density function γ has marginal distributions that correspond to Y_0 and Y_1. Intuitively, if we use the graphs of the probability density functions p_0 and p_1 to represent mounds of sand, γ corresponds to a transportation plan that dictates how much sand to transport in order to reshape the p_0 mound into the p_1 mound. In particular, the value of $\gamma(u,v)$ is the amount of sand to be transported from u to v. The distance $d(u,v)$ can then be interpreted as the cost of transporting one unit of sand from u to v, and the earthmover distance is simply the cost of the transportation plan γ that incurs the least cost.

Now we define Lipschitz continuity.

Definition 10. *Let $f : \mathcal{Y} \to \mathbb{R}$ be a function, and let d be a distance metric defined over $\mathcal{Y}$. f is* ρ-Lipschitz continuous *if, for all $u, v \in \mathcal{Y}$,*

$$|f(u) - f(v)| \le \rho \cdot d(u,v). \tag{5}$$

A.2. Main Result

Definition 7 allows an output disparity if there *exists* an equally large disparity in Y', but it does not explicitly reflect the fact that we care about *how* the model came to exhibit the disparity. The only reason why we allow the disparity is that Y' is the right attribute to use. Thus, if the model does not use Y' at all, then there should be no output disparity. More formally, we want that if $Y' \perp \hat{Y}$, then $\hat{Y} \perp Z$.

Definition 11 generalizes this requirement and, unlike Definition 7, is applicable for both categorical and numerical Y' at the expense of limiting $\hat{Y}$ to be binary. The generalization deals with cases where $\hat{Y}$ is not completely independent of Y' by measuring of how much $\hat{Y}$ depends upon Y'. For binary $\hat{Y}$, this dependence is captured by the likelihood function $\ell(y') = \Pr[\hat{Y}{=}1 \mid Y'{=}y']$, and we use the Lipschitz continuity of this function to measure the dependence.

Definition 11 (Discriminatory Association, Stronger)**.** *For $\hat{Y} \in \{0,1\}$ and $\ell(y') = \Pr[\hat{Y}{=}1 \mid Y'{=}y']$, let ρ_ℓ^* be the smallest nonnegative ρ such that ℓ is ρ-Lipschitz continuous.*[1] *Then, a model exhibits discriminatory association that is unexplainable by Y' (DAU-Y') if*

$$d_{\mathrm{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) > \rho_\ell^* \cdot d_{\mathrm{em}}(Y'|Z{=}0, Y'|Z{=}1). \tag{6}$$

ρ_ℓ^* characterizes how much impact Y' can have on the output of the model. If the impact is small, we can conclude that the model is not using Y' much, so not much output disparity can be explained by Y'. On the other hand, if a small change in Y' can cause a large change in the probability distribution of $\hat{Y}$, then even a large output disparity can possibly be due to a small inter-group difference in Y'. In fact, the use of ρ_ℓ^* makes Definition 11 invariant to scaling in Y'. If a numerical Y' is increased by some factor, ρ_ℓ^* will decrease by the same factor, so the quantity on the right-hand side of Equation 6 will not change.

We now give two arguments that Definition 11 is the correct refinement of the previous definition (Definition 7). First, we show that the new definition is a broader definition of unexplainable discrimination than the previous one. The previous definition assumes that Y' is categorical, and in this case a natural distance metric for its support $\mathcal{Y}'$ is the indicator $d(u,v) = \mathbb{1}(u \neq v)$. With this distance metric, we can relate the total variation distance used in the right-hand side of Equation 2 with the earthmover distance used in Equation 6.

Theorem 7. *Let the construct Y' be categorical with support $\mathcal{Y}'$, which has distance metric $d(u,v) = \mathbb{1}(u \neq v)$.*

[1] Technically, ρ_ℓ^* should be the *infimum* of all ρ such that ℓ is ρ-Lipschitz continuous, but it is not difficult to show then that ℓ is in fact ρ_ℓ^*-Lipschitz continuous.

If a model has DAU-Y' under Definition 7, the model has DAU-Y' under Definition 11 as well.

Proof. We proceed by showing that $\rho_\ell^* \cdot d_{\text{em}}(Y'|Z{=}0, Y'|Z{=}1) \le d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1)$.

Since the likelihood function ℓ in Definition 11 is always between 0 and 1, we have $|\ell(u){-}\ell(v)| \le 1 = d(u,v)$ when $u \ne v$, so ℓ is 1-Lipschitz continuous. Therefore $\rho_\ell^* \le 1$, and it suffices to show that $d_{\text{em}}(Y'|Z{=}0, Y'|Z{=}1) \le d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1)$.

By (Gibbs and Su, 2002, Theorem 4), we get

$$\begin{aligned} &d_{\text{em}}(Y'|Z{=}0, Y'|Z{=}1) \\ &\le \Big(\max_{u,v\in\mathcal{Y}'} d(u,v)\Big) \cdot d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1) \\ &= d_{\text{tv}}(Y'|Z{=}0, Y'|Z{=}1), \end{aligned}$$

so we are done. □

Second, we show that Theorems 1 and 3 still hold under the refined definition of DAU-Y'. Since the definitions of optimality and the empirical tests have not changed, we focus strictly on the nondiscrimination portions of the theorems.

Theorem 8. *A model that passes the demographic parity test does not have DAU-Y' under Definition 11.*

The proof of Theorem 8 is very similar to that of Theorem 1 and will thus be omitted.

Theorem 9. *If the WYSIWYG worldview holds, then a model that passes the equalized odds test does not have DAU-Y' under Definition 11.*

Proof. We present the proof for the case where Y' is continuous, but the proof for the discrete case is very similar. Let p_0 and p_1 be the probability density functions of $Y'|Z{=}0$ and $Y'|Z{=}1$, respectively. By Kantorovich duality (Villani, 2008, Equation 5.4), we have

$$\begin{aligned} &d_{\text{em}}(Y'|Z{=}0, Y'|Z{=}1) \\ &\qquad \ge \int_{\mathcal{Y}'} \phi(v)\, p_1(v)\, dv - \int_{\mathcal{Y}'} \psi(u)\, p_0(u)\, du \end{aligned} \tag{7}$$

for all ϕ and ψ such that $\phi(v) - \psi(u) \le d(u,v)$ for all $u, v \in \mathcal{Y}'$. We set $\phi(v) = \psi(v) = \ell(v)/\rho_\ell^*$, where ℓ and ρ_ℓ^* are defined as in Definition 11. Then, $\phi(v) - \psi(u) = (\ell(v) - \ell(u))/\rho_\ell^* \le d(u,v)$ by Lipschitz continuity. Thus, Equation 7 applies and implies that

$$\begin{aligned} &\rho_\ell^* \cdot d_{\text{em}}(Y'|Z{=}0, Y'|Z{=}1) \\ &\qquad \ge \int_{\mathcal{Y}'} \ell(v)\, p_1(v)\, dv - \int_{\mathcal{Y}'} \ell(u)\, p_0(u)\, du. \end{aligned} \tag{8}$$

By the WYSIWYG worldview and the definition of equalized odds, we have $\ell(y) = \Pr[\hat{Y}{=}1 \mid Y'{=}y] = \Pr[\hat{Y}{=}1 \mid Y'{=}y, Z{=}0] = \Pr[\hat{Y}{=}1 \mid Y'{=}y, Z{=}1]$. Therefore, we can use the law of total probability to rewrite the first term on the right-hand side of Equation 8 as $\Pr[\hat{Y}{=}1 \mid Z{=}1]$, and similarly the second term becomes $\Pr[\hat{Y}{=}1 \mid Z{=}0]$.

If we let $\phi(v) = \psi(v) = -\ell(v)/\rho_\ell^*$ in Equation 7 instead, we get $\rho_\ell^* \cdot d_{\text{em}}(Y'|Z{=}0, Y'|Z{=}1) \ge \Pr[\hat{Y}{=}1 \mid Z{=}0] - \Pr[\hat{Y}{=}1 \mid Z{=}1]$. Finally, combining this inequality with the previous one gives us

$$\begin{aligned} &\rho_\ell^* \cdot d_{\text{em}}(Y'|Z{=}0, Y'|Z{=}1) \\ &\ge \Big|\Pr[\hat{Y}{=}1 \mid Z{=}0] - \Pr[\hat{Y}{=}1 \mid Z{=}1]\Big| \\ &= d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1), \end{aligned}$$

which is what we want. □

We now briefly discuss the tightness of the above result. In the extreme example where ℓ is a step function over real-valued y', ρ_ℓ^* is infinite, so we trivially have nondiscrimination under Definition 11. Therefore, in order to receive meaningful fairness guarantees from Theorem 9, we must make sure that ρ_ℓ^* is not too large. One way to achieve this is to apply the function ℓ to the construct space and reason about the transformed construct space. If any transformation of the construct space results in a finding of discrimination under Definition 11, then it is evidence that there could be a problem with the model with respect to discrimination. Let $\tilde{y}' = \ell(y')$ be a value in the transformed construct space, and $\tilde{\ell}$ denote the likelihood function on this space. Then,

$$\begin{aligned} \tilde{\ell}(\tilde{y}') &= \Pr[\hat{Y}{=}1 \mid \tilde{Y}'{=}\tilde{y}'] \\ &= \Pr[\hat{Y}{=}1 \mid Y'{=}y'] = \ell(y') = \tilde{y}', \end{aligned}$$

so the transformation ensures that $\rho_{\tilde{\ell}}^* = 1$.

A.3. Connection to the α-Disparity Test

When Y' and Y are numerical, a natural extension of the α-disparity test (Definition 8) is

$$d_{\text{tv}}(\hat{Y}|Z{=}0, \hat{Y}|Z{=}1) \le \rho_\ell^* \cdot \alpha \cdot d_{\text{em}}(Y|Z{=}0, Y|Z{=}1). \tag{9}$$

For this to work, Worldview 3 would have to change to use the earthmover distance rather than the total variation distance. Since the earthmover distance is defined over a distance metric, the parameter α is not very meaningful unless Y' and Y have the same scale. As a result, here we consider the case where Y' and Y are defined over the same metric space $(\mathcal{Y}, d)$.

Unfortunately, Equation 9 is still not an empirical test because ρ_ℓ^* is defined in terms of Y'. Although it is tempting

to redefine ρ_ℓ^* in terms of Y, it is possible for Y' and Y to have vastly different likelihood functions while having the same disparity, so this new empirical test will not guarantee the lack of DAU-Y' under Definition 11. We leave as future work the discovery of the appropriate empirical test for numerical Y' and Y under the α-Hybrid worldview.