A Different View of Fair Data Representation

Abstract

We investigate notions of fair data representation as reflected in the fairness of classification predictions of an agent accessing training data represented that way . The resulting notions rely of course on the notion of fair classification being applied, but also on the agent's objective driving that classification. We distinguish three major types of such objectives; malicious (driven by a bias against a group of subjects), accuracy driven (ignoring fairness considerations), and fairness driven.

The main questions we analyze under this taxonomy of representation fairness notions are: 1) Can there be a generic fair representation (independent of the classification task)? 2) Can the effect of a given feature in a representation on fairness be determined independently of the other features involved? 3) What is the effect of training sample sizes on the fairness based on such representations?

We show that, for central notions of classification fairness (e.g., odds equality), the answer to the first question is negative for all the three different objectives. This is in contrast to previous results suggesting a positive answer when the underlying notion of fairness is demographic parity. We furthermore show that there are tasks which are fairness incompatible, i.e. there is no representation that prevents unfairness and allows accurate representation for both tasks simultaneously.

For the second question we also give negative answers for all three classification objectives under various fairness notions. In particular, we show that in the accuracy driven case a feature can have opposing fairness effects, dependent on the other available features.

Finally, we address the effect of sampling size on unfairness, as a property of the representation, in the accuracy driven case. We show that there are situations where the the Bayes optimal classifier is fair, but sample sizes that guarantee high accuracy do not suffice to guarantee fairness.

1 Introduction

Automated decision making has become more and more successful over the last few decades and has therefore been used in an increasing number of domains, either to fully automate decision making, or to support human decision makers. This includes many sensitive domains which significantly impact people's livelihoods, such as loan applications, university admissions, recidivism predictions, or insurance rate settings. It has been found that many such decision tools have, often unintentionally, biases against minority groups, and therefore lead to discrimination. In response to these concerns, the machine learning research community has been devoting effort to developing precise notions of fair decision making, and coming up with algorithms for implementing fair machine learning.

While many papers in this domain propose algorithmic solutions to fairness related issues, the main contributions of this paper are conceptual. We feel that unlike many other facets of machine learning, the field of fairness in machine learning is still in need of investigating its fundamental concepts and setup. Some basic questions are still far from being satisfactorily elucidated; What should be considered fair decision making? (various mutually inconsistent notions have been proposed, but the real life application of picking between them still lacks clarifications). What is a fair data representation? To what extent should accuracy or other practical utilities and costs be compromised for achieving such goals? and so on. The answers to these questions are not generic. They vary with the principles and the goals guiding the agents involved (decision makers, subjects of such a decision, policy regulators, etc.), as well as with what can be assumed regarding the underlying learning setup. We view those as the primary issues facing this area, deserving explicit research attention (in addition to the more common algorithmic and optimization aspects). In this paper we address one of these questions - the nature of fair data representation.

Machine learning tools base their classification output on the data they access for training. The features used for representing that data play a major role in the outcome of an algorithm trained on such inputs. This realization has spurred research on notions of fair data representation, which is the subject of this paper.

What is a fair data representation? Most previous works on fair representation aim to find representations that can be considered fair across many classification prediction tasks (often termed 'transferable' or 'flexibly fair'). Consequently, they impose some level of *demographic parity* or independence of the feature values from protected group membership. Such properties have the advantage that they can be verified for a feature regardless of the way it may be used by an agent (Zhao and Gordon 2019; Creager et al. 2019; Madras et al. 2018; Oneto et al. 2019). A downside of this approach is that, once there is significant correlation between group membership and the target classification, it is in conflict with

accuracy. Any classifier that is based on such fair features will fail to accurately match a classification that is highly correlated with the groups it requires parity for.

We address this question from the perspective of its effect on the classification rules that an agent using data represented that way may come up with. Such a view takes into consideration (and therefore is a function of and adapted to) two setup characteristics:

1. The type of agent using the data, depending on the objective of the agent; malicious (driven by a bias against a group of subjects), accuracy driven (ignoring fairness considerations), or fairness driven.
2. The applied notion of group fairness of classification decisions. In this paper we primarily focus on the notion of odds equality.

Applying our notion of fairness of representations to ground-truth-related notions of group fairness in the spirit of odds equality (rather than demographic parity), allows for fair representations that are aligned with accurate classification prediction (including when group membership is highly correlated with the target class).

Dependent on those two considerations we examine three questions concerning the choice of features for a fair data representation:

1. Can there be a generic fair representation (independent of the classification task the representation will be used for)?
2. Can the effect of a given feature in a representation on the fairness of the representation be determined independently of the other features involved?
3. What is the effect of training sample sizes on the fairness of a classifier trained on data under the representation?

We show that, for central notions of classification fairness (e.g., odds equality), the answer to the first question is negative for all three types of agent objectives, which are malicious, fairness driven, and accuracy driven (Observation 4, Observation 5, Theorem 2). This is in contrast with previous works cited above suggesting a positive answer when the underlying notion of fairness is demographic parity. We furthermore show that there are tasks which are fairness incompatible (Theorem 3). Namely, pair of tasks for which there is no representation that guarantees fair classification and allows accurate representation for both tasks simultaneously.

For the second question we also give negative answers for all three classification objectives under various fairness notions. In particular, we consider the fairness implications of feature deletion. A well known real world example of this is the "ban the box" policy which disallowed employers using criminal history in hiring decisions (Doleac and Hansen 2016). It was observed that such a policy could lead to less fair decision making even though it was motivated by fairness considerations. Evaluating the fairness of a representation through the classification decisions they imply provides a framework for analysing such issues. We show that it is not possible to determine the impact a single feature has on the fairness of a representation if other available features are not considered. This can be shown for many fairness notions for representation that consider a decision maker who actively tries to discriminate (Observation 2, Observation **??** and Observation **??**) and for decision makers who try to actively pursue fair classification (Observation 3). Furthermore, we show that when decision makers optimize accuracy and are indifferent to fairness, feature deletion of the same feature can have opposing effects depending on the context of other available features: When considering the fairness of the most accurate decision-rule for a given feature set, we show that the unfairness can increase when deleting a feature for a given set of available features, while deleting the same feature can decrease unfairness, if the set of other available features is different (Theorem 1).

One should note that while the data representation and task determines the Bayes classifier, in most cases the decision maker does not have access to the Bayes classifier. Instead, it usually bases its decision on some training sample. The last question we investigate is the effect of training sample sizes on unfairness, as a property of the representation. We show that there are situations where the the Bayes optimal classifier is fair, but sample sizes that guarantee good approximation of that optimal classifier in terms of accuracy do not suffice to guarantee a similar level of fairness (Theorem 4). It follows that in such scenarios, an accuracy driven agent may end up with an unfair classifier if it settles for training samples of sizes sufficient to guarantee accuracy. We give a bound of this effect in terms of the probability mass of the least likely group-label combination (Theorem 5) and suggest importance sampling as a countermeasure for unfairness due to under-sampling.

Finally, we look into the details of accuracy-driven and adversarial fairness. We give necessary and sufficient conditions on the alignment of conditional distributions given labels and groups for accuracy-driven fairness (Theorem 8) and adversarial fairness (Theorem 7) for the equalized odds notion of group fairness. For both of those cases we bound the unfairness in terms of distribution distances of conditionals (Theorem 9). We furthermore connect adversarial fairness for various group-fairness notions to independence considerations between the feature-vector, ground-truth-labeling and group-membership of random instances (Theorem 6, Theorem 10). We end this discussion by showing that there are also group-fairness notions for which adversarial fairness is not always achievable (Theorem 10), i.e., we show that there can be no adversarially fair representation with respect to predictive rate parity, if the success rates between groups is not equal.

Our paper is organized as follows: Section 2 introduces the formal setup. Section 3 introduces our taxonomy of fair representations. Section 4 discusses the impact on a single feature on the fairness of a representation. Section 5 addresses the question of transferable fair representation. Section 6 discusses unfairness due to under-sampling as a separate source of unfairness. Section 7 gives some characterization of accuracy-driven and adversarially fair representations. In Section 8 we present related work. Section 9 concludes this paper with some final remarks.

2 Formal Setup

Domain set X is the set of instances that we wish to classify (say people that may be applying for a loan)

Data generating probability distribution We assume that there is some fixed probability distribution P over $X \times \{0,1\}$. As common in PAC learning theory, we assume that the training data is generated i.i.d. by that distribution.

Ground truth labeling rule $t : X \to [0,1]$. We will think of the label 1 as denoting 'qualified' and the label 0 as 'unqualified' and $t(x) = P[y = 1|x]$. For concreteness, we focus here on the case of deterministic labeling (that is $t : X \to \{0,1\}$. Most of our discussion can readily be extended to the probabilistic labeling case as well.

Group membership : We assume that X is partitioned by sets A and D We further use the function $G : X \to \{A, D\}$ to indicate the group-membership of an instance.

Feature Set $\mathcal{F} = \{f_1, \ldots, f_n\}$ where each $f_i : X \times \{0,1\} \to Y_i$ is a feature (we think of Y_i as being finite, maybe a discretization of an infinite set of values).

Feature-induced cells $\mathcal{F}$ includes an equivalence relation $\sim_{\mathcal{F}}$ over $X \times \{0,1\}$. Namely, $x \sim_{\mathcal{F}} y$ if for every $f \in F$, $f(x) = f(y)$. Let $\mathcal{C}_{\mathcal{F}}$ be the set of the $\sim_{\mathcal{F}}$ equivalence classes. We call each such equivalence class a '*cell*'. That is, cells are formed by grouping together individuals with the same representation.

Ground truth scoring We define the *ground truth score function* $s_t : \mathcal{C}_{\mathcal{F}} \to [0,1]$. $s_t^P(C)$ is the probability, w.r.t. P, of $x \in C$ having the true-label 1. i.e.,

$$s_t^P(C) = \mathbb{E}_{x \sim P}[t(x)|x \in C]$$

In cases where the distribution is non-ambivalent we will use the abbreviated notation s_t instead of s_t^P.

Decision rule A feature based decision rule is a function $h : \mathcal{C}_{\mathcal{F}} \to \{0,1\}$. We denote the hypothesis class of all feature based decision rules as $H_{\mathcal{C}_{\mathcal{F}}}$.

Learner We consider learners that get as input a labeled sample and output a decision rule (a.k.a. label predictor). The learner (or decision maker) does not know what P or t are. Furthermore, rather than seeing the identity of each instance in the training sample (or the test data), the decision maker only sees the feature based representation $F(x) = (f_1(x), \ldots f_n(x))$.

2.1 Loss minimizing decision maker

A decision maker uses a training sample to pick predictor h aiming to minimize some loss function. In this work we consider weighted $0/1$ losses. Namely, for some given parameter $0 \leq \alpha \leq 1$, false negative instances incur a loss α and false positives $(1-\alpha)$:

$$\begin{aligned} L_P^{\alpha}(h) =& \alpha P(\{x : t(x) = 1, h(x) = 0\}) \\ &+ (1-\alpha)P(\{x : t(x) = 0, h(x) = 1\}) \end{aligned}$$

For a threshold $\alpha = 0.5$ we will use the notation $L_P = L_P^{0.5}$. We denote the set of threshold classifiers w.r.t. the ground-truth scoring function s_t by

$\mathcal{H}^{\text{thres}}_{\mathcal{C}_{\mathcal{F}}, s_t} = \{h : \mathcal{C}_{\mathcal{F}} \to \{0,1\} : \text{ for some } \alpha,\ h(C) = 0 \text{ iff } s_t(C) < \alpha\}$

Lemma 1. *(Corbett-Davies and Goel 2019) The predictor in $\mathcal{H}_{cells}$ that minimizes L_P^{α} is the Bayes Optimal predictor $t_{P,F}^{\alpha}$ that for a cell $C \in \mathcal{C}_{\mathcal{F}}$ assigns the label 1 if $s_t(C) > \alpha$ and 0 otherwise.*

3 Taxonomy of fair representation notions

In this section we propose several notions for the fairness of a representation. Since we focus on data representation for classification/decision-making purposes, our notions of fair representation are induced by notions of fair decision making. To define a notion of fair representation, we therefore need to answer two questions 1) When is a decision rule considered fair? and 2) Which decision rule will a decision maker arrive at?. The first question has been addressed in the fairness literature in various ways by the introduction of group fairness notions. Our discussion will mainly focus on the equalized odds notion of fairness (Hardt, Price, and Srebro 2016). For the second question we distinguish between three different motivations of decision makers: accuracy-driven, fairness-driven and malicious. We will start now by giving a definition of some common notions of group fairness.

Group fairness notions (for decision making)

Definition 1 (Group fairness; Equalized odds)**.** *Group fairness is a property of a classifier that takes into account the classifiers predictions and group membership. We say a group fairness notion is* ground-truth-related*, if it also considers the ground truth label. The notion of group-fairness we will focus on in this paper is the ground-truth-related notion of odds equality as introduced by (Hardt, Price, and Srebro 2016).*

A classifier h is considered fair w.r.t. to odds equality (L^{EO}) and a distribution P if $h(x) \perp\!\!\!\perp G(x)|t(x)$. The respective unfairness is given by the sum of differences in false positive rate and false negative rate: For $g \in \{A, D\}$

$$\begin{aligned} \mathit{FNR}_g(h,t,P) &= \frac{P(\{x \in X_{g,1} : h(x) = 0\})}{P(X_{g,1})} \\ \mathit{FPR}_g(h,t,P) &= \frac{P(\{x \in X_{g,0} : h(x) = 1\})}{P(X_{g,0})} \\ L_P^{EO}(h) &= |\mathit{FNR}_A - \mathit{FNR}_D| + |\mathit{FPR}_A - \mathit{FPR}_D| \end{aligned}$$

If we say a classifier is fair, without referring to any particular group-fairness notion, we mean fairness w.r.t. equalized odds.

Another widely used notion of group-fairness that we will refer to throughout the paper is demographic parity.

Definition 2. *A classifier h is considered fair w.r.t. to demographic parity (L^{DP}) and a distribution P if $h(x) \perp\!\!\!\perp G(x)$. The respective unfairness is given by difference in positive classification rates between groups*

$$L_P^{DP}(h) = |\frac{P(h^{-1}(1) \cap A)}{P(A)} - \frac{P(h^{-1}(1) \cap D)}{P(D)}|$$

.

Fairness of a representation We will now give our definitions of representation fairness in terms of a general group fairness notion L^{fair} with unfairness measure L_P^{fair}. We wish to provide fairness guarantees with respect to this notion to the decision rules a decision maker arrives at. We therefore introduce different fairness notions for representations for different classification objectives that would inform the decision rule that would be used.

We start by considering a *malicious decision maker* who tries to actively discriminate against one group. To protect against this kind of decision maker, we need to give a guarantee such that based on the feature set it is not possible to discriminate against one group. This corresponds to the notion of adversarial fairness.

Definition 3 (Adversarial fairness)**.** *A feature set $\mathcal{F}$ is considered to be* adversarial *fair w.r.t. the distribution P and group fairness objective L^{fair} , if every classifier $h \in H_{\mathcal{C}_\mathcal{F}}$ is group-fair. We define the worst case unfairness of a feature set $\mathcal{F}$ by $U_{adv}(\mathcal{F}) = \max_{h \in H_{\mathcal{C}_\mathcal{F}}} L_P^{fair}(h)$.*

Furthermore, we consider an *accuracy-driven decision maker*, who aims to label instances correctly and is agnostic about fairness. For this kind of decision maker, we only need to make sure that optimizing for correct classification results in a fair classifier. In order to achieve this, we need to provide a guarantee for the fairness of the Bayes optimal classifier. We define this to be the accuracy-driven fairness of a representation.

Definition 4 (Accuracy-driven fairness)**.** *We define accuracy-driven fairness in two steps:*

1. *A feature set $\mathcal{F}$ is considered to be* accuracy-driven *fair w.r.t. the fairness objective L^{fair}, the distribution P and weight α if every classifier $h \in H_{\mathcal{C}_\mathcal{F}}$ with $L_P^\alpha(h) = \min_{h \in H_{\mathcal{C}_\mathcal{F}}} L_P^\alpha(h)$ is group-fair. The accuracy-driven (un)fairness w.r.t. L^{fair}, P and α is defined by $U_{acc}^\alpha = L_P^{fair}(t_{P,F}^\alpha)$.*
2. *We say a feature set $\mathcal{F}$ is accuracy-driven fair w.r.t. L^{fair} and P if for every weight $\alpha \in [0,1]$ the feature set $\mathcal{F}$ is accuracy-driven fair w.r.t. to P and α. Namely,*

$$U_{acc}(\mathcal{F}) = \max_{h \in H_{\mathcal{C}_\mathcal{F}}, \alpha \in [0,1] : L_P^\alpha(h) = \min_{h \in H_{\mathcal{C}_\mathcal{F}}} L_P^\alpha(h)} L_P^{fair}(h).$$

We will see in Section 6, that in cases where the decision maker does not have access to the distribution P, but only to a labelled sample, this requirement is actually not sufficient for guaranteeing that an accuracy-driven decision maker arrives at a fair decision. In this case, we would need to further require that any classifier who is close to the Bayes optimal classifier in terms of accuracy is also close to the Bayes optimal classifier in terms of fairness. This can be phrased as another fairness criterion for a representation (*λ-robustness*, see Appendix).

Lastly, we also consider a *fairness-driven decision maker* who actively tries to find a fair and accurate decision rule, while maintaining some accuracy guarantees. For such a decision maker a representation should allow for fair and accurate decision rules. If a representation fulfills this requirement, we call it fairness-enabling.

Definition 5 ((ϵ, η)-fairness-enabling representation)**.** *A feature set $\mathcal{F}$ is considered to be* (ϵ, η)-fairness-enabling *w.r.t. a fairness objective L^{fair}, if there exists a classifier $h \in H_{\mathcal{C}_\mathcal{F}}$ that such that $L_P^\alpha(h) \leq \epsilon$ and $L_P^{fair}(h) \leq \eta$.*

Our discussion focuses primarily on the case of malicious and indifferent decision makers. These notions of fair representation can be defined with respect to any group-fairness notion. In our paper we will mainly focus on the equalized odds notion of fairness (Hardt, Price, and Srebro 2016).

4 Fairness of a feature set vs. fairness of a feature

In this section we discuss feature deletion and its impact on the fairness of a representation. We mainly focus on accuracy-driven fairness w.r.t. equalized odds. In particular, we show that the deletion of a feature f can lead to an increase in accuracy-driven fairness for some set of other given features $\mathcal{F}$ and that the deletion of the *same* feature f can lead to a decrease in accuracy-driven fairness for another set of other available features $\mathcal{F}'$. We start by giving an example of the above phenomena in a real-life scenario. Quite surprisingly, we show this phenomena holds for a general class of features that satisfy some non-triviality properties (That on the one hand do not reveal too much information about group membership and labels (non-disclosing), and on the other hand does not reveal identity when label and group information is given (k-anonymity (Samarati and Sweeney 1998))). Lastly, we will discuss the impact of a single feature on the fairness of a representation for adversarial fairness and fairness-enabling representations.

4.1 University Admissions Example

Let us assume that a university has to deploy an automated classifier to admit students to their program. They have access to some features for each applicant. They would typically have access to tens or maybe even hundreds of features, but for illustration purposes we consider a simplified version, with very few features.

1. One feature is group membership:

$$\text{Advantaged} = A.$$
$$\text{Disadvantaged} = D$$

2. Another is the feature which indicates if an applicant has volunteered before or not.

$$\text{Volunteered previously} = V.$$
$$\text{Not volunteered previously} = N$$

3. The third is the feature which indicates if an applicant is old or young.

$$\text{Old} = O.$$
$$\text{Young} = Y$$

One might be tempted to think that using feature 1 in the classifier is unfair. We show that this is not always the case. More specifically, we give two cases: In Situation 1, using the

group membership feature actually harms the fairness of the classifier. In Situation 2 however, we see the opposite effect! Contrary to popular intuition, using group membership in this situation, actually helps the fairness of the classifier. Let us see how this happens.

In both situations, an applicant needs to have a score of greater than 0.5 to be admitted.

Situation 1 The classifier has access to features 1 and 3. The following are the cells with weights and scores in the following format:

Cell(Feature 1, Feature 3) = (Weight, Score)

$$C(A, O) = (\frac{1}{4}, 0.65)$$
$$C(D, O) = (\frac{1}{4}, 0.5)$$
$$C(A, Y) = (\frac{1}{4}, 0.65)$$
$$C(D, Y) = (\frac{1}{4}, 0.5)$$

One sees that feature 3 (age) has no effect on an individuals ability to succeed in university. It is not hard to observe that using only feature 3 implies that each applicant is admitted to university. However, using both feature 1 and 3 results in every member of group A being accepted, while each member of group D is rejected. Easy calculations show us that the former case (Unfairness = 0.46) is more fair than the latter (Unfairness = 2), lending support to our intuition that using feature 1 (group membership) is unfair.

Situation 2 The classifier has access to features 1 and 2. The following are the cells with weights and scores in the following format:

Cell(Feature 1, Feature 2) = (Weight, Score)

$$C(A, V) = (\frac{3}{8}, 0.8)$$
$$C(D, V) = (\frac{1}{8}, 0.47)$$
$$C(A, N) = (\frac{1}{8}, 0.2)$$
$$C(D, N) = (\frac{3}{8}, 0.51)$$

We see that people in the disadvantaged group have a much smaller fraction of people who have volunteered. This is also something observed in real life: students in underprivileged schools do not have as much access to volunteering and other opportunities. Hence, we observe a large difference between the scores of people in the disadvantaged group who have not volunteered, and advantaged group who have not volunteered. This is not surprising: an advantaged person who has not made use of the ample volunteering opportunities available to them is less likely to succeed in university, than a disadvantaged person who has not volunteered, because the disadvantaged person had little to no opportunities to volunteer.

Using only feature 2 (volunteering) groups disadvantaged applicants with high scores ($C(D, N)$), and advantaged applicants with lower scores ($C(A, N)$). Both these cells are rejected as a consequence. Adding feature 1 (group membership) into the mix allows the classifier to separate disadvantaged applicants with high scores($C(D, N)$), and advantaged applicants with lower scores ($C(A, N)$). We can therefore now accept applicants in ($C(D, N)$) and reject applicants from ($C(A, N)$). Easy calculations show us that the former case (Unfairness = 0.96) is less fair than the latter (Unfairness = 0.43), contradicting our intuition that using feature 1 (group membership) is unfair.

Two "unfair" features together? We see that using either the feature of group membership (A and D), or of volunteering decreases the fairness w.r.t. using no feature. This might make someone think that both the group membership feature and volunteering features are unfair. Therefore, one would think that if one is using one of these two features, one should not use the other. Using one of them is unfair enough, and using both of them together would be worse. However, this intuition is proved false. Using both features results in a more fair classifier compared to using just one of the features.

Overall, we see that the effect of adding features on the fairness of the classifier depends on the set of other features available to the classifier. More specifically, we see that, in one situation, adding the group membership feature decreased fairness, and in another situation it increased fairness.

4.2 Result

As mentioned previously, this section shows that the deletion of a feature can lead to either an increase or a decrease in accuracy-enforced fairness. Moreover, the deletion of the *same* feature can have opposing effects on the accuracy-enforced fairness of different representations. These results hold for a general class of features that satisfy some non-triviality properties.

We start by detailing the non-triviality feature properties that suffice for this phenomenon.

Non-Triviality properties

Definition 6. *We define the following two non-triviality requirements for a feature:*

1. **Non-disclosing** *We will call a feature* non-disclosing *if this feature does not reveal too much information about group-membership and label. That is, a feature f is non-disclosing if there are two distinct values y_1 and y_2, such that f assigns each of these values to at least one instance of each $X_{A,0}, X_{A,1}, X_{D,1}, X_{D,0}$. i.e. $f^{-1}(y_1) \cap X_i \neq \emptyset$ and $f^{-1}(y_2) \cap X_i \neq \emptyset$ for every $X_i \in \{X_{A,0}, X_{A,1}, X_{D,1}, X_{D,0}\}$*

2. **k-anonymity** *A feature f is* k-anonymous *if knowing this feature, group-membership and label, will only reveal identity of an individuals up to a set of at least k individuals. Namely, for every combination of value of this feature, group membership and class label, there are either no instances satisfying this combination or there are at least k many such instances.*

Overview of Result We show that for every feature that is 6-anonymous and non-disclosing there is a context (i.e., a set of features) for which deleting this feature increases accuracy-driven fairness, and a context for which deleting this feature decreases accuracy-driven fairness. This shows that in most situations we cannot answer the question of how fair a feature is without regarding the context of the other features available to the classifier.

Theorem 1. *(Context-relevance for fairness of features) For every 6-anonymous non-disclosing feature f, there exists a probability function P over X and feature sets $\mathcal{F}$ and $\mathcal{F}'$ such that:*

- *The accuracy-driven fairness w.r.t L^{EO}, P and $\alpha = 0.5$ of $\mathcal{F} \cup \{f\}$ is greater than that of $\mathcal{F}$, i.e.*

$$U^{\alpha}_{acc}(\mathcal{F} \cup \{f\}) < U^{\alpha}_{acc}(\mathcal{F})$$

  Thus, deleting f in this context will increase unfairness.
- *The accuracy-driven fairness w.r.t L^{EO}, P and $\alpha = 0.5$ of $\mathcal{F}' \cup \{f\}$ is less than that of $\mathcal{F}'$, i.e.*

$$U^{\alpha}_{acc}(\mathcal{F}' \cup \{f\}) > U^{\alpha}_{acc}(\mathcal{F}')$$

  Thus, deleting f in this context will decrease unfairness.

We can also show that this is the case when natural conditions between f and P hold, such as $\{f\}$ being adversarial fair w.r.t. to P and equalized odds. For a more general theorem, details and proofs of this theorem we refer the reader to the appendix.

4.3 The fairness of a feature for different notions of fairness

We will now briefly discuss the fairness of a single feature for the case of a malicious or a fairness-driven decision maker. In contrast to the accuracy-driven case, adding features has a monotone effect on the fairness in those cases. That is, adding any feature in the malicious case, will only give the decision maker more information and thus give the decision maker more chances of discrimination. Similarly, in the fairness driven case, any feature will only give the decision maker another option for fair decision making.

Observation 1. *For any feature f and any featureset $\mathcal{F}$ we have $U_{adv}(\mathcal{F}) \leq U_{adv}(\mathcal{F} \cup \{f\})$. Similarly, if the representation $\mathcal{F}$ is (ϵ, η)-fairness-enabling, the representation $\mathcal{F} \cup \{f\}$ is also (ϵ, η)-fairness-enabling.*

Proof. $\mathcal{H}_{\mathcal{C}_{\mathcal{F}}} \subset \mathcal{H}_{\mathcal{C}_{\mathcal{F} \cup \{f\}}}$. □

The above observation of course also implies that the fairness of $\{f\}$ as a feature set is a bound on the fairness of any representation that includes that feature $\mathcal{F} \cup \{f\}$, i.e. $U_{\text{adv}}(\{f\})$ lower bounds $U_{\text{adv}}(\mathcal{F} \cup \{f\})$. However, we find that in both the malicious and the fairness-driven case, the impact of fairness of adding a single feature cannot be determined without considering the context of other available features. To illustrate this we make the following two observations.

Observation 2. *For every distribution P and feature f, there exists a feature set $\mathcal{F}$, such that adding f will not impact the fairness of the distribution, e.g. $U_{adv}(\mathcal{F}) = U_{adv}(\mathcal{F} \cap \{f\})$. Furthermore, there exist distributions P, features f and $\mathcal{F}'$, such that $U_{adv}(\mathcal{F}') = 0$ and $U_{adv}(\{f\}) = 0$, but $U_{adv}(\mathcal{F}' \cup \{f\}) = 1$.*

Observation 3. *For every distribution P and every feature f, there exists a feature set $\mathcal{F}$, such that $\mathcal{F} \cup \{f\}$ is (η, ϵ)-fairness-enabling, if and only if $\mathcal{F}$ is (η, ϵ)-fairness-enabling. Furthermore, there exists a distribution P, a feature f and a feature set $\mathcal{F}'$, such that both $\mathcal{F}'$ and $\{f\}$ are not (ϵ, η)-fairness-enabling for any $\epsilon, \eta < \frac{1}{2}$, but such that $\mathcal{F}' \cup \{f\}$ is $(0, 0)$-fairness-enabling.*

While this section focused on fairness with respect to equalized odds, we note that many of these results can be replicated for other notions of fairness. For a more general version of Observation 2, which takes into account other fairness notions, like equalized odds, we will refer the reader to the Appendix.

Thus, in general for any of motivation of a decision maker, we cannot determine the impact of a single feature without considering the context.

5 Can there be a generic fair representation?

In this part we address the question of whether there can be a representation, that is fair independent of task. A task is defined by a distribution P and consists of two components, the marginal P_X and the ground truth labeling t.
We start by considering scenarios in which the marginals shift between two tasks, e.g. two openings for different jobs, for which a different pool of people would apply. Such a distribution shift can likely affect one group more than another and would thus affect the classification rates of both groups differently.

Observation 4. *For every representation $\mathcal{F}$ and every ground truth labeling function t, that is not constant on either group, there exists a distribution P with ground truth labeling t, such that $\mathcal{F}$ is adversarial unfair w.r.t. to P and L^{EO}. The same holds for any non-constant labeling function t for the (seemingly task-independent) demographic parity notion of fairness.*

Thus, when a shift in marginal occurs between tasks, any fairness guarantee for previous tasks does not necessarily imply a fairness guarantee for a new task. Now let us consider situations, in which tasks share a marginal distribution and only the ground-truth labelling differ between them. For fairness-notions that are independent from the ground-truth, like demographic parity, the question of the existence of general fair representation under this additional restriction can be answered to the positive; we only need demographic parity in every cell to achieve that goal (see Appendix). However,

in the case of a ground-truth-dependent notions of fairness like equalized odds we show that there is no representation that guarantees fairness with respect to all ground-truth labelings. This is true for all three decision objectives. We start by making the following observation for the fairness-driven case.

Observation 5. *For every marginal distribution P_X and every representation $\mathcal{F}$, for which there exists a classifier $h_{DP} \in \mathcal{H}_{\mathcal{C}_\mathcal{F}}$ with demographic parity that labels exactly half the mass as 1, i.e. $P_X(h_{DP}^{-1}(1)) = \frac{1}{2}$, there is a task given by distribution P with marginal P_X, such that $\min_{h' \in \mathcal{H}} L_P(h') \leq \frac{1}{4}$ such that any $h \in \mathcal{H}_{\mathcal{C}_\mathcal{F}}$ based on the representation is either unfair with respect to P and equalized odds or has loss $L_P(h) \geq \frac{1}{2}$.*

We get a similar result for the other two classification objectives. In the next theorem, we will consider a feature f to be *non-trivial* with respect to a marginal P_X, if it splits the support of P_X according to something other than group-membership, i.e. $P_X(f^{-1}(1) \cap A) \notin \{0, P(A)\}$ or $P_X(f^{-1}(1) \cap D) \notin \{0, P_X(D)\}$.

Theorem 2. *For every marginal distribution P_X and every representation $\mathcal{F}$ containing at least one feature that is non-trivial w.r.t. P_X, there exists a task given by a distribution P with marginal P_X, such that $\mathcal{F}$ is neither accuracy-driven nor adversarial fair w.r.t. P.*

We will now look at a slightly more restricted setting and analyse the case of multi-task learning, where instead of asking for a representation that is fair for every task, we only consider fairness with respect to a fixed (finite) set of tasks that we want to learn. We again make the distinction of the three motivations for decision makers. For a decision maker who is accuracy-driven or fairness-driven, it is always possible to have such a fair multi-task representation with respect to equalized odds. To see this, we consider a representation, that allows perfect classification for all considered (finitely many) tasks. Such a representation can be achieved by including the ground-truth labeling for each task as a feature. Since every classifier with perfect accuracy is also fair w.r.t. equalized odds, this representation would be considered fair in the accuracy-driven and the fairness-driven case.However, when the goal is to prevent unfair classification, i.e. if we want to address the malicious case, it is not always possible to find a representation that fulfills both fairness and accuracy requirements. In the next theorem, we show that there are *fairness-incompatible tasks*, i.e., tasks for which there is no joint representation which prevents unfair classification while enabling accurate prediction for both tasks.

Definition 7 (Success-rate)**.** *For a task given by distribution P, the success-rate of a group G is defined by $SR_G(P) = \frac{P(X_{G,1})}{P(G)}$. We say a distribution P has equal success rates if $SR_A(P) = SR_D(P)$.*

Theorem 3. *Given two tasks with distributions P_1 and P_2 with the same marginal $P_X = P_{1,X} = P_{2,X}$, such that their success rates $SR_A(P_1), SR_D(P_1), SR_A(P_2).SR_D(P_2) \in (0,1)$ are not equal to 0 or 1 and have different ratios between groups, i.e. $\frac{SR_A(P_1)}{SR_D(P_1)} \geq \max\{\frac{SR_A(P_2)}{SR_D(P_2)}, \frac{1-SR_A(P_2)}{1-SR_D(P_2)}\}$, then there is no representation $\mathcal{F}$, such that*

- *$\mathcal{F}$ is worst case fair w.r.t. P_1 and P_2 and L^{EO}*
- *$\mathcal{F}$ allows for perfect accuracy w.r.t. to P_1, i.e. there is a $h \in \mathcal{H}_{\mathcal{C}_\mathcal{F}}$, such that $L_{P_1}(h) = 0$.*

Therefore, if the goal is to prevent discrimination from a possibly adversarial decision maker, while also enabling accurate prediction, each task requires its task-specific feature representation.

6 Effect of sampling on group fairness

In previous sections, we have examined the group-unfairness of feature based loss-minimizing classifiers, $t^\alpha_{P,F}$. However, in reality, a classifier is derived based on finite samples and can only approximate such loss minimizers. In this section, we turn our attention to the effect of sampling errors on the fairness of feature based learnt classifiers.

Our main finding on this issue is that while under some conditions the goals of maximizing accuracy and minimizing unfairness go hand in hand, when these conditions fail, sample sizes that suffice for achieving high accuracy may still result in learnt classifiers whose unfairness is significantly high.

We consider learning through empirical risk minimization. The sampling method we consider is i.i.d. sampling of a selected sample size. All of our results are stated with respect to the following sample-based classifier.

Definition 8 (empirical α-threshold predictor)**.** *We define the* empirical α-threshold predictor $t^\alpha_F(S) : \mathcal{C}_\mathcal{F} \to \{0,1\}$ *as a function that given access to a labeled sample $S \subset (\mathcal{C}_\mathcal{F} \times \{0,1\})^m$ of some sample size m, assigns label 1 to a cell if and only if more than α fraction of samples in the cell have label 1,i.e.,*

$$t^\alpha_F(S)(C) = \begin{cases} 1 & \text{if } |S \cap (C \times \{0,1\})| > 0 \\ & \text{and } \frac{|S \cap C \times \{1\}|}{|S \cap C \times \{0,1\}|} > \alpha \\ 0 & \text{otherwise} \end{cases}$$

Note that, if there is no sample point in a cell C the cell is assigned label 0.

We show that the empirical α-threshold predictor is indeed an Empirical Risk Minimizer.

Lemma 2. *For a sample S, $t^\alpha_F(S)$ is an Empirical Risk Minimizer of the class $\mathcal{H}_{cells}$.*

We can now state the main theorem of this section.

Theorem 4. *There is a feature set $\mathcal{F}$ such that for every $\delta > 0$ and every $\epsilon > 0$, there is a probability distribution P over the domain X and a sample size $M \in \mathbb{N}$ such that for every sample $S \sim P^M$ the empirical risk minimizing classifier $t^\alpha_F(S)$*

- *$\mathcal{F}$ is accuracy-driven fair w.r.t. P*
- *$t^\alpha_F(S)$ has estimation error $L^\alpha_P(t^\alpha_F(S)) - L^\alpha_P(t^\alpha_{P,F}) < \epsilon$ and*
- *$t^\alpha_F(S)$ has group unfairness $L^{fair}_P(t^\alpha_F(S)) = 1$*

with probability at least $1 - \delta$ over the sample generation.

Just as how the learnt classifier can have worse fairness than the Bayes classifier, the learnt classifier could also have better fairness than the Bayes classifier.

Observation 6. *There are situations in which a sample-based classifier is more fair than the respective Bayes classifier, with significant probability over the sampling.*

When the feature set and distribution satisfy some properties, any classifier with loss close to the loss of $t^{\alpha}_{P,F}$ also has unfairness close to the unfairness of $t^{\alpha}_{P,F}$. When these properties are satisfied, any accuracy guarantee we get for learning from a sample implies a fairness guarantee.

Theorem 5. *For every distribution P, every threshold α and every feature set $\mathcal{F}$, for every feature based classifier h*

$$L^{fair}_P(h) - L^{fair}_P(t^{\alpha}_{P,F}) \leq \lambda(L^{\alpha}_P(h) - L^{\alpha}_P(t^{\alpha}_{P,F}))$$

where $\lambda = \max\left\{\frac{1}{\alpha}, \frac{1}{1-\alpha}\right\} \cdot \max_{(g,l)\in\{A,D\}\times\{0,1\}} \frac{1}{P(X_{g,l})}$

The property that allows for accuracy guarantee to imply a fairness guarantee is that all groups and labels are balanced. That is, no group-label set $X_{g,l}$ has low measure. Based on an i.i.d. sample, we can detect if a there is a set $X_{g,l}$ with low measure. We can use the fraction of the sample belonging to the set $X_{g,l}$ as an estimate of the probability measure of $X_{g,l}$. With high probability over i.i.d. samples of size m, this estimate is accurate up to order of $1/\sqrt{m}$.

If we find that in an i.i.d. sample, a set $X_{g,l}$ occurs with low frequency, we might be in a situation where an ERM classifier based on i.i.d. sampling would have high unfairness despite having good accuracy. We could prevent this by changing our sampling strategy. One way is by increasing the sample size of i.i.d. sampling. Another way is to shift away from i.i.d. sampling into a more targeted sampling and collect more data points from the low probability set.

7 Analysing Worst-case and Accuracy-driven fairness

In this section we characterize accuracy-driven and adversarial representation fairness w.r.t. the odds equality notion of classification fairness. We will start by introducing a property we call zero-group knowledge. it is aimed to prevent an adversary from inferring the group membership from the representation, when given access to the ground-truth labels. To ensure that an adversarial agent won't be able to infer group-membership, one would of course require the representation to have demographic parity. However, in situations where label information is correlated with group membership, demographic parity of all features will hurt classification accuracy. In such cases, zero-group-knowledge might be a better tool for concealing group-information.

We will then see that this property is closely related to adversarial fairness.

Definition 9 (Zero-group-knowledge)**.** *A feature set $\mathcal{F}$ has* zero-group-knowledge *w.r.t. a distribution P, if for $x \sim P$, knowing the feature vector $F(x)$ will not reveal more information about the group membership $G(x)$ than knowing just the ground truth, $t(x)$. Namely, $G(x) \perp\!\!\!\perp F(x)|t(x)$.*

It turns out that this property is equivalent to adversarial fairness with respect to equalized odds.

Theorem 6. *A feature set $\mathcal{F}$ has zero-group knowledge w.r.t. P if it has adversarial fairness w.r.t to P and the group-fairness measure L^{EO}.*

A similar observation has been made and shown by Zhang et al (Zhang, Lemoine, and Mitchell 2018), relating the optimization criteria for the goal of concealing group-membership and preventing unfair classification with respect to equalized odds in a representation learning setting with GANs.

We will now give a characterization of accuracy-driven and worse-case fairness in terms of the conditional distributions given label and group-membership. In the following we will denote the conditional probabilities given label l and group G as $P_{G,l}$. We will see that a feature set is adversarial fair, if and only if the conditional probabilities are aligned. It has already been shown in (Zhao et al. 2019) that if conditional probabilities are aligned over a representation, every classifier based on that representation is fair. We go a step further here, by noting, that this is indeed a necessary condition for adversarial fairness.

Theorem 7. *A feature set $\mathcal{F}$ is adversarial fair w.r.t. distribution P if and only if for each cell $C \in \mathcal{C}_{\mathcal{F}}$ and for each $l \in \{0,1\}$ we have $P_{A,l}(C) = P_{D,l}(C)$.*

We now give a similar statement for accuracy enforced fairness. Here, the same statement holds, if instead of considering the probability distributions over the set of cells $\mathcal{C}_{\mathcal{F}}$, we consider the set of cells that results from merging all cells of the same score:

Definition 10 (Score-induced cells)**.** *For a set of cells $\mathcal{C}_{\mathcal{F}}$, the corresponding set of* score-induced cells *$\mathcal{C}_{\mathcal{F}s_t}$ is the set of cells that is obtained by merging all cells with the same score together. More formally, each feature set and scoring function, induce an equivalence relation $\sim_{\mathcal{F},s_t}$, such that $x \sim_{\mathcal{F},s_t} y$ if and only if there are cells $C_x, C_y \in \mathcal{C}_{\mathcal{F}}$ such that $x \in C_x, y \in C_y$ and $s_t(C_x) = s_t(C_y)$. The set $\mathcal{C}_{\mathcal{F}s_t}$ is then defined as the set of $\sim_{\mathcal{F},s_t}$ equivalence classes.*

Theorem 8. *A feature set $\mathcal{F}$ is accuracy-driven fair w.r.t. distribution P if and only if for each cell in the score-induced $C \in \mathcal{C}_{\mathcal{F}s_t}$ and for each $l \in \{0,1\}$ we have $P_{A,l}(C) = P_{D,l}(C)$.*

We can now bound the unfairness in terms of accuracy-driven and adversarial fairness of a representation by the distribution distance of conditional probabilities. For this we take the $\mathcal{H}$-distance as introduced by (Ben-David et al. 2010).

Definition 11 ($\mathcal{H}$-distance)**.** *Given two distributions P and Q over X, we define their $\mathcal{H}$-distance by*

$$d_{\mathcal{H}}(P,Q) = \sup_{1_B \in \mathcal{H}} |P(B) - Q(B)|,$$

where 1_B denotes the indicator function of set B.

Theorem 9. *We can bound adversarial fairness and accuracy enforced fairness of a feature set $\mathcal{F}$ w.r.t. P and L^{EO} by the $d_{\mathcal{C}_{\mathcal{F}}}$-difference and $d_{\mathcal{C}_{\mathcal{F}s_t}}$-difference of conditional distributions respectively:*

$$U_{adv}(\mathcal{F}) \leq d_{\mathcal{H}_{cells}}(P_{A,1}, P_{D,1}) + d_{\mathcal{H}_{\mathcal{C}_{\mathcal{F}}}}(P_{A,0}, P_{D,0})$$

$$U_{acc}(\mathcal{F}) \leq d_{\mathcal{H}^{thres}_{\mathcal{C}_\mathcal{F},s_t}}(P_{A,1}, P_{D,1}) + d_{\mathcal{H}^{thres}_{\mathcal{C}_\mathcal{F},s_t}}(P_{A,0}, P_{D,0})$$

Furthermore, we can lower bound the adversarial fairness of a representation by

$$d_{\mathcal{H}_{cells}}(P_{A,l}, P_{D,l}) \leq U_{adv}(\mathcal{F}) \text{ for every } l \in \{0,1\}$$

Note that for both bounds there exist probability distributions P such that equality holds in all cases. Furthermore we note that since the $\mathcal{H}$-distance between two distributions can be estimated, if $\mathcal{H}$ has a finite VC-dimension (Ben-David et al. 2010), we can estimate both the upper and the lower bound with a sample size dependent on $|\mathcal{C}_\mathcal{F}|$, when given access to i.i.d. samples from $P_{A,1}, P_{D,1}, P_{D,0}$ and $P_{A,0}$ each.

7.1 Other fairness notions

So far we have demonstrated our approach mainly with respect to the odds equality notion of fair decision making. In this subsection we briefly discuss the implications on fair representations of using other notions of decision making fairness.

For some of the common such notions, it is easy to see that adversarial representation fairness is achievable and easy to characterize: For *demographic parity*, requiring cells representation induced cell have the same proportions for all groups is both necessary and sufficient for achieving adversarial fairness of the representation. This is probably an already known insight and we provide a formal statement and proof in the appendix.

However, some other acceptable notions of fairness do not always allow a adversarial fair representation. One such notion is *predictive rate parity*.

Definition 12. *(Predictive rate parity (PRP)) A classifier h is considered PRP fair w.r.t. to a marginal data distribution P and true classification t if the random variable $t(x)$ is independent of the group membership, $G(x)$ given the classification $h(x)$. The respective measure of unfairness is given by the sum of differences in positive predictive rate and negative predictive rate, where the positive predictive rate for a hypothesis h and group G is defined as $PPR_G(h) = \frac{P(h^{-1}(1)\cap X_{G,1})}{P(h^{-1}(1)\cap G)}$ and the negative predictive rate is defined as $NPR_G(h) = \frac{P(h^{-1}(0)\cap X_{G,0})}{P(h^{-1}(0)\cap G)}$, i.e.*

$$L_P^{Pred}(h) = |PPR_A(h) - PPR_D(h)| + |NPR_A(h) - NPR_D(h)|$$

Theorem 10 below shows that there are distributions for which there is no representation that has adversarial fairness with respect to predictive rate parity. In cases, where such a adversarial representation is achievable, however, we can characterize it by the following natural requirement on the representation.

Definition 13. *A feature set $\mathcal{F}$ has* calibration parity *w.r.t. a distribution P if for every cell $C \in \mathcal{C}_\mathcal{F}$ both groups have equal success probability. Equivalently, one can say that for a random instance $x \in P$ the ground truth labeling $t(x)$ and the group membership $G(x)$ are statistically independent, when the feature vector $F(x)$ of x is known, i.e. $G(x) \perp\!\!\!\perp t(x)|F(x)$.*

We can now state the main theorem of this subsection.

Theorem 10. *A feature set $\mathcal{F}$ has calibration parity w.r.t. P if it has adversarial fairness w.r.t P and the group-fairness measure L^{Pred}. The other direction does not hold. In particular, adversarial fairness w.r.t. P and L^{Pred} is only possible, if P has equal success rates for both groups.*

8 Related Work

Most recent works on fair representation learning focus on learning fair intermediate representations that is then used by a learning algorithm to build a decision rule (Zemel et al. 2013; Madras et al. 2018; Zhao et al. 2019; Adel et al. 2019; Zhang, Lemoine, and Mitchell 2018). Most of the papers in this area use the demographic parity notion of fairness (Edwards and Storkey 2016; Madras et al. 2018; Zemel et al. 2013). Our work uses the equalized odds notion of fairness (Hardt, Price, and Srebro 2016). Some other papers on fair representation also use equalized odds fairness (Zhang, Lemoine, and Mitchell 2018; Beutel et al. 2017; Song et al. 2019).

The connection between the motivation of the decision maker using the representation and the study of its fairness has been made before (Madras et al. 2018; Zhang, Lemoine, and Mitchell 2018). These papers identify two motivations. The first is malicious which is the intent to discriminate without regard for accuracy. The second is accuracy-driven which is the intent to maximize accuracy. Our results are expressed according to these motivations. Additionally, we provide results for the motivation of fairness-driven which is the intent to provide the most fair outcome while maintaining some level of accuracy.

Many algorithmic approaches to learning fair representations have been proposed (Zhang, Lemoine, and Mitchell 2018; Madras et al. 2018; Edwards and Storkey 2016; Song et al. 2019). Providing algorithmic approaches for fair representation learning is not our goal in this paper. We instead aim to answer questions about what we can and cannot achieve for a fair representation. One question in this spirit is the fairness-accuracy trade-off due to fair representations . This question has been studied for demographic parity (Zhao and Gordon 2019; McNamara, Ong, and Williamson 2019) . A lower bound on the error of any classifier based on a representation that guarantees demographic parity, when the success rates are different between groups, is shown in Zhao et al. (Zhao and Gordon 2019). The existence of situations when we can build a more accurate and more fair classifier based on the original representation than any classifier built using a learnt demographic parity satisfying representation is shown in Mcnamara et al. (McNamara, Ong, and Williamson 2019).

With demographic parity, there are works aiming to construct fair representations that can be used for different tasks (Creager et al. 2019; Madras et al. 2018; Oneto et al. 2019). Representations for different ground truths are considered in Madras et al. (Madras et al. 2018), and Oneto et al. (Oneto et al. 2019). Representations for different fairness notions - demographic parity w.r.t. different subgroups are considered in Creager et al. (Creager et al. 2019). We prove that for a ground truth dependent fairness notion like equalized

odds, we cannot construct fair representations that transfer to different ground truths, other than in trivial situations. We also note that representations that are fair with respect to seemingly task-independent notions like demographic parity, are also not transferable to other tasks, if there is a change in marginals between tasks. Such a scenario is likely to occur in a job-application setting where the pool of applicants might be different between jobs.

The question of feature deletion has also been considered in real world examples, such as in the "ban the box" policy which disallowed employers using criminal history in hiring decisions (Doleac and Hansen 2016). The effect of allowing or disallowing features on fairness has been studied before, for example in Grgić-Hlača et al. (Grgic-Hlaca et al. 2018). However, the effect of a feature on fairness, has been discussed in isolation (as in (Grgic-Hlaca et al. 2018)). In contrast, we show that fairness of a feature should not be considered in isolation, but should also take into account the remaining features available.

Under-sampling has been identified as a source of unfairness (Chen, Johansson, and Sontag 2018; Balashankar and Lees 2019). Both these papers note the possibility of learning accurate classifiers that are unfair, even though there is a classifier that is both fair and accurate. This is done using an assumption on how generalization error behaves as a function of the sample size in Chen et al. (Chen, Johansson, and Sontag 2018). In Balashankar et al. (Balashankar and Lees 2019), this is done by noting that the lower bound on sample complexity for good generalization across multiple groups is strictly higher than the upper bound on sample complexity for good overall generalization. We prove that this situation of learning accurate but unfair classifiers due to sampling can occur, even with significant probability over the sampling. Increasing sample sizes and performing targeted sampling have been suggested as ways to prevent unfairness from sampling (Chen, Johansson, and Sontag 2018; Balashankar and Lees 2019; Zelaya, Missier, and Prangle 2019).

9 Conclusion

In this paper we introduced a general taxonomy of notions of fair representation, taking into consideration both different objectives of decision makers using the representation, and different group fairness notions. Within this taxonomy we addressed several questions about fair representations.

We first considered the question of "fairness of a feature", which has been used in legal scenarios. We showed that for notions of decision-making fairness other than demographic parity, the fairness of a single feature is an ill defined notion. Namely, the impact of a feature on the fairness of a decision cannot be determined without considering the other features of the representation.

Next, we addressed the existence of generic fair representations and of fair transfer learning. We show that even seemingly task-independent fairness notions like demographic parity are vulnerable to shifts in marginals between tasks. We have also proved that for the odds equality notion of fairness, there is no fixed representation that is fair with respect to all ground-truth labelings, even when the marginal distribution is fixed. Furthermore, when representations are aimed to prevent an adversary from unfair decision making, we show that there exist fairness incompatible tasks, i.e. tasks that do not allow an adversarial fair representation which also enables accurate predictions of class labels for both tasks. These insights stand in contrast to the impression arising from recent papers (Madras et al. 2018; Oneto et al. 2019) that claim to learned transferable fair decisions. Such results are inevitably limited to ground-truth oblivious notions of fairness (such as demographic parity).

Following this, we address the issue of unfairness due to under-sampling. We analyze this issue as a property of the distribution and show that it differs from the unfairness of a representation due to an unfair Bayes classifier. We show that there are distributions and representations such that an empirical risk minimizing algorithm is unfair with high probability over the sample generation, while also being accurate with high probability. We show that this can occur even if the Bayes classifier based on the representation is fair. This suggests that further stability criteria need to be taken into account when considering fair representations for accuracy-driven decision makers who base their decision on a finite sample. As a solution for ensuring fairness we suggest importance sampling.

Lastly, we provide a characterization of accuracy-driven and adversarial fairness of a representation with respect to the equalized odds notion of fairness in terms of differences between conditional distribution given labels and groups. Furthermore we show that some fairness notions, like predictive rate parity, do not always allow an adversarially fair representation.

unsrt

References

Adel, T.; Valera, I.; Ghahramani, Z.; and Weller, A. 2019. One-network adversarial fairness. In *AAAI*.

Balashankar, A.; and Lees, A. 2019. Fairness Sample Complexity and the Case for Human Intervention. *arXiv preprint arXiv:1910.11452* .

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A Theory of Learning from Different Domains. *Mach. Learn.* 79(1–2).

Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *CoRR* abs/1707.00075.

Chen, I.; Johansson, F. D.; and Sontag, D. 2018. Why is my classifier discriminatory? In *NIPS*.

Corbett-Davies, S.; and Goel, S. 2019. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR* abs/1808.00023.

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly Fair Representation Learning by Disentanglement. In *ICML*.

Doleac, J. L.; and Hansen, B. 2016. Does “ban the box” help or hurt low-skilled workers? Statistical discrimination and employment outcomes when criminal histories are hidden. Technical report, National Bureau of Economic Research.

Edwards, H.; and Storkey, A. J. 2016. Censoring Representations with an Adversary. In *ICLR*.

Grgic-Hlaca, N.; Bilal Zafar, M.; P. Gummadi, K.; and Weller, A. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *AAAI*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *NIPS*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning Adversarially Fair and Transferable Representations. In *ICML*.

McNamara, D.; Ong, C. S.; and Williamson, R. C. 2019. Costs and benefits of fair representation learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 263–270.

Oneto, L.; Donini, M.; Maurer, A.; and Pontil, M. 2019. Learning fair and transferable representations. *arXiv preprint arXiv:1906.10673* .

Samarati, P.; and Sweeney, L. 1998. Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. Technical report.

Song, J.; Kalluri, P.; Grover, A.; Zhao, S.; and Ermon, S. 2019. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2164–2173.

Zelaya, V.; Missier, P.; and Prangle, D. 2019. Parametrised data sampling for fairness optimisation. *KDD XAI* .

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *ICML*.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *AAAI/ACM Conference on AI, Ethics, and Society*.

Zhao, H.; Coston, A.; Adel, T.; and Gordon, G. J. 2019. Conditional Learning of Fair Representations. *CoRR* abs/1910.07162.

Zhao, H.; and Gordon, G. J. 2019. Inherent Tradeoffs in Learning Fair Representations. *CoRR* abs/1906.08386.
otherwise