INHERENT LIMITATIONS OF MULTI-TASK FAIR REPRESENTATIONS

Tosca Lechner[*]**, Shai Ben-David**[‡]
Cheriton School of Computer Science, University of Waterloo
Waterloo, Ontario, Canada

ABSTRACT

With the growing awareness to fairness issues in machine learning and the realization of the central role that data representation has in data processing tasks, there is an obvious interest in notions of fair data representations. The goal of such representations is that a model trained on data under the representation (e.g., a classifier) will be guaranteed to respect some fairness constraints, while still being expressive enough to model the task well. Such representations are useful when they can be fixed for training models on various different tasks and also when they serve as data filtering between the raw data (available to the representation designer) and potentially malicious agents that use the data under the representation to learn predictive models and make decisions. A long list of recent research papers strive to provide tools for achieving these goals.

However, *we prove that in most cases, such goals are inaccessible!* Roughly stated, we prove that no representation can guarantee the fairness of classifiers for different tasks trained using it (while retaining the needed expressive powers). The reasons for this impossibility depend on the notion of fairness one aims to achieve. For the basic ground-truth-independent notion of Demographic (or Statistical) Parity, the obstacle is conceptual; a representation that guarantees such fairness inevitably depends on the marginal (unlabeled) distribution of the relevant instances, and in most cases that distribution changes from one task to another. For more refined notions of fairness, that depend on some ground truth classification, like Equalized Odds (requiring equality of error rates between groups), fairness cannot be guaranteed by a representation that does not take into account the task specific labeling rule with respect to which such fairness will be evaluated (even if the marginal data distribution is known a priori). Furthermore, for tasks sharing the same marginal distribution, we prove that except for trivial cases, no representation can guarantee Equalized Odds fairness for any two different tasks while enabling accurate label predictions for both.

Fair classification, group fairness, data representation, fair data representation, Demographic Parity, Equalized Odds.

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 as stand alone, or to support human decision makers. This includes many sensitive domains which significantly impact people's livelihoods, such as granting loans, university admissions, recidivism predictions, or setting insurance rates. It was found that many such decision tools, often unintentionally, have 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 clear notions of fair decision making, and coming up with algorithms for implementing fair machine learning.

A common approach to address the important issue of fair algorithmic decision making is through *fair data representation*. The idea is that some regulator, or a responsible data curator, transforms collected data to a format (–*representation*), that can then be used for solving downstream classification tasks, while providing guarantees of fairness. This approach was put forward by the seminal paper of Zemel et al. Zemel et al. (2013). In their words: *"our intermediate representation can be used for other classification tasks (i.e., transfer learning is possible)... We further posit that such an intermediate representation is fundamental to progress in fairness in classification, since it is composable and not ad hoc; once such a representation is established, it can be used in a blackbox fashion to turn any classification algorithm into a fair classifier, by simply applying the classifier to the sanitized representation of the data"*. Many followup papers aim to realize this paradigm, solving technical and algorithmic issues Madras et al. (2018); Edwards & Storkey (2016); McNamara et al. (2019); Song et al. (2019); Creager et al. (2019) (to mention just a few). The main contribution of this paper is showing that, basically, *it is impossible to achieve that goal!*

[‡] Also Canada AI CIFAR chair and faculty at the Vector Institute, Toronto.

1.1 Our impossibility results

We prove the following impossibility results:

Demographic Parity (DP) fairness: Given any domain partitioned into two non-empty groups (say,‘privileged’ and ‘disadvantaged’), no non-trivial data representation can guarantee that every classifier expressible under that representation is DP fair for all possible probability distributions over that domain.

Equalized Odds and Predictive Rate Parity (Namely, fairness notions that take ground truth classification into account): Given any two different classification tasks over the same unlabeled data distribution (a.k.a. marginal) in which the ground truth classification has some correlation with the group membership, no data representation can simultaneously enable accurate label classifiers for both while guaranteeing that any classifier expressible over that representation is fair for both these tasks.

The “fairness of a feature" cannot be determined in isolation Several papers on fairness of representations discuss fairness as a property of each feature (e.g., Zemel et al. (2013); Creager et al. (2019); Grgic-Hlaca et al. (2018)). We show that when focussing on the outcome of a classification rule that uses a given feature, the fairness of the feature cannot be decided without considering the context of the other features used for that classification task. In particular, we show that if we consider accuracy maximizing classifiers and Equalized Odds fairness, the same feature can either increase or decrease the fairness of a representation on a fixed task, depending on the other features used in the representation.

Creager et al. Creager et al. (2019) state (in the Discussion section): *“There are two main directions of interest for future work. First is the question of fairness metrics: a wide range of fairness metrics beyond demographic parity have been proposed (Hardt et al. (2016); Pleiss et al. (2017)). Understanding how to learn flexibly fair representations with respect to other metrics is an important step in extending our approach. Secondly, robustness to distributional shift presents an important challenge in the context of both disentanglement and fairness".* Our results can be viewed as answering both questions negatively.

1.2 The source of discrepancy with previous publication

There is an apparent discrepancy between our impossibility results and the long list of papers claiming to achieve fair representations. What is the source of that discrepancy?

1. *Demographic Parity:*

   The key distinguishing component is that in most (if not all) of the papers that claim positive results about fair representations, the design of the fair representation relies (often implicitly) on having access to the data distribution with respect to which the fairness is defined.When the notion of fairness is independent of the ground truth classification (the case of Demographic Parity), the distribution in question is the marginal (unlabeled) one. Let us examine how Demographic Parity defined in the fairness literature.

   Zemel et al, Zemel et al. (2013) define in in equation (1) there as $P(Z = k|x^+ \in X^+) = P(Z = k|x^- \in X^-), \forall k$. However, they do not specify what that probability P is. Often when a probability over a finite set is not being defined, the implicit semantics is that it is the uniform distribution over that set. Here, the sets X^+ and X^- are a partition of some domain set X. The domain set is defined rather vaguely "X denotes the entire data set of individuals." Should we consider the set of all applicants for a given position, or the entire population of the city that position is in? Should we consider the local population or rather the population of the state or the world?

   Barocas et al, Barocas et al. (2019) (the “Fairness in Machine Learning” book) Definition 1 in the “Formal non-discrimination criteria" subsection of Chapter 2 reads $P[R = 1|A = a] = P[R = 1|A = b]$ without specifying what that probability distribution P is.

   Creager et al. Creager et al. (2019) express the probability their notion of fairness refers to by "$x, y, a, \sim p_{data}$" (the beginning of the Background section there) without any further elaboration of what that p_{data} may be.

   *T*hese ambiguities are in the heart of the discrepancies between the claims in those papers and the formal impossibility results we prove here.

   In situations in which the desired notion of fairness is determined with respect to the specific action or decision that the agent needs to make, the assumption that a designer of a multi-task (or "flexible") data representation has access to the relevant data distribution can be justified only in rather limited situations. For example, it is conceivable that the sought after Demographic Parity for acceptance of students to a given university program depends on the distribution of applicants *to that program,in that university* (rather than, say, the percentage of members of each group in the world). That distribution is likely to change between universities, between programs and between academic years. Therefore, based on our results, no a priory- designed data representation for accepting students to programs can be guaranteed to provide the Demographic Parity fairness it aims to establish. The situation is similar when it comes to granting loans - the distribution of applicants changes between loan granting institutions, branch locations, requested sums, dates, etc. In fact, it is hard to come up with any realistic scenarios in which a fixed data distribution remains unchanged throughout various classification tasks that may use the data representation down the road.

2. *Equalized Odds and Predictive Rate Parity:*

   When the notion of fairness of concern does involve ground truth labels (such as Equalised Odds or Predictive Rate Parity), fairness becomes harder to achieve. For a fair representation to be useful for some task (say, for concreteness, a classification task), it has to balance two opposing requirements; on one hand, the fairness requirement (in the sense of making sure that any classifier built on the representation is bound to be fair), that constrains the expressive power of the representation. On the other hand, to be useful for modeling the task at hand, the representation needs to be rather expressive - sufficiently so to allow expressing task accurate models (classifiers). In turns out that no representation can fulfil such requirements simultaneously for any two sufficiently different tasks (in a sense that we make precise in Section 4). We therefore conclude that there can be no representation that meets the desiderata stated in many of the papers aiming for “fair representations" (e.g., Zemel et al. (2013), Creager et al. (2019) and many more). In particular, for those task-dependent notions of group fairness, no data representation can meet the goal stated in those papers (namely, can "be used in a blackbox fashion to turn any classification algorithm into a fair classifier, by simply applying the classifier to the sanitized representation of the data"). Regardless of the data available to the representation designer, any representation that meets that fairness goal is bound to defeat the quest for allowing accuracy on more than a single task!

Paper road map: We begin our discussion with a concise taxonomy of the notions of fair representation that our work relates to (Subsection 1.3). Section 2 gives an overview of the related work. Section 3 introduces our basic notation and Section 4 contains our main results on the impossibility of generic fairness of a representation. In Section 5 we show similar impossibility results for Predictive Rate Equality fairness, and in Section 6 we show that the effect of a single feature on the fairness of a representation cannot be determined by looking at that feature in isolation.

1.3 WHAT IS *fair representation*?

The term ‘fair data representation’ encompasses a range of different meanings. When word embedding results in smaller distance between the vectors representing ‘woman’ and ‘nurse’ relative to the distance between the representations of ‘woman’ and ‘doctor’ and the other way around for ‘man’, is it an indication of bias in the *representation* or is it just a faithful reflection of a bias in society? Rather than delving into such issues, we discuss an arguably more concrete facet of data representation; We examine representation fairness from the perspective of its effect on the fairness of classification rules that agents using data represented that way may come up with. Such a view takes into consideration two setup characteristics:

The objective of the agent using the data We distinguish two types of classification prediction agents (formal definitions of these aspects of fairness are provided in section 3.2):

Malicious - driven by a bias against a group of subjects. To protect against such an agent, a fair representation (or feature set) should be such that *every* classifier based on data represented that way is fair. This is arguably the most common approach to fair representations in the literature e.g., Zemel et al. (2013); Madras et al. (2018).

Accuracy Driven - focusing on traditional measures of learning efficiency, ignoring fairness considerations. A representation is accuracy-driven fair if every loss minimizing classifier based on that representation is fair.

In this work we focus on representations aimed to guarantee fairness of *malicious agents*.

The notion of group fairness applied to the classification decisions The wide range of group fairness notions (for classification) can be taxonomized along several dimensions: Does the notion depend on the ground truth classification or only on the agent’s decision (like demographic parity)? Is a perfectly accurate decision (matching the ground truth classification) always considered fair (like in odds equality)? Does the fairness notion depend on unobservable features (like intention or causality)? In this work we focus on fairness notions that are ground-truth-dependent, view the ground truth classification as fair and depend only on observable features. Picking which notion of fairness one wishes to abide by depends on societal goals and may vary from one task to another. This is outside the scope of this paper. We refer the reader to Barocas et al. (2019) for further discussions of these issues.

Our running example of such a notion is Equalized Odds (EO) Hardt et al. (2016), however our results hold as well for other common notions of fairness that meet the above conditions (like Predictive Rate Parity/Calibration within groups Kleinberg et al. (2016)) We provide formal definitions of these notions in Section 3.1.

2 Related Work

Most, if not all, of the literature concerning the creation of fair data representations addresses this task in a setup where some input data (or a probability distribution over some domain of individuals) is given to the agent building the representation (e.g., Edwards & Storkey (2016); Madras et al. (2018); Zemel et al. (2013); Song et al. (2019)). Such a probability distribution is essential to any common definition of fairness. However, in many cases the probability distribution with respect to which the fairness is defined remains implicit. For example, Zemel et al. (2013) define their notion of fairness by saying: "We formulate this using the notion of statistical parity, which requires that the probability that a random element from X^+ maps to a particular prototype is equal to the probability that a random element from X^- maps to the same prototype" (where X^+ and X^- are the two groups w.r.t. which one aims to respect fairness). However, they do not specify what is the meaning of "a random element". The natural interpretation of these terms is that "random" refers to the uniform distribution over the finite set of individuals over which the algorithm selects. In that case, that information varies with each concrete tasks and is not available to the task-independent representation designer. Alternatively, one could interpret those "random" selections as picking uniformly at random from some established large training set that is fixed for all tasks. Such randomness may well be available to the representation designer, but it misses the intention of statistical parity fairness; For example, the fixed training set may have 10,000 individuals from one group and 20,000 from the other group, but when some local bank branch allocates loans it has 80 applicants from the first group and 37 applicants from the other. For the fairness of these loan allocation decisions, the relevant ratio between the groups is 80/37 rather than the 10,000/20,000 ratio available to the representation designer.

Almost all the work on fair representations focuses on the demographic parity (DP) notion of fairness Edwards & Storkey (2016); Madras et al. (2018); Zemel et al. (2013); Song et al. (2019). To achieve DP fairness, a classifier has to induce success ratio between the groups of subjects that match the ratio between these groups in the input data. However, as demonstrated above, that ratio varies from one application to another and cannot be determined a priori. We show that any fixed representation that allow expressing non-trivial classification cannot guarantee DP fairness in the face of shifting marginal (that is, unlabeled) data distribution (see section 4).

When the data marginal distribution w.r.t. which the fairness is defined is fixed and available to the designer of a representation, then, as shown by Zemel et al. (2013) and followup papers, DP fairness is indeed possible. However, we further show that even under these assumptions, no data representation can guarantee fairness with respect to notions of fairness that do rely on the correct ground truth, such as equalized odds (EO) Hardt et al. (2016), for arbitrary tasks (see Section 4).

To the best of our knowledge this fact also has not been explicitly stated (and proved) before, although it seems that some of the previous work were aware of this concern; in previous work discussing fair representation w.r.t. notions of fairness that take the ground truth classification into account, the algorithms that design the representations require access to task specific labeled data (e.g. Zhang et al. (2018); Beutel et al. (2017); Song et al. (2019); du Pin Calmon et al.). Such a requirement defies the goal of having a fixed representation that guarantees fairness for many tasks.

The effect of the motivation of the user of the representation on the fairness of the resulting decision rule has been considered by Madras et al. Madras et al. (2018) and Zhang et al. Zhang et al. (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. We address these effects as part of our taxonomy of notions of fair representations.

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 & Hansen (2016). The effect of allowing or disallowing features on fairness has been studied before, for example in Grgic-Hlaca et al. Grgic-Hlaca et al. (2018). However in previous works, the effect of a feature on fairness, has been discussed in isolation. 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.

3 Basic notation

We consider a binary classification problem with label set $\{0,1\}$ over a domain X of instances we wish to classify, e.g. individuals applying for a loan. We assume the task to be given by some distribution P over $X \times \{0,1\}$ from which instances are sampled i.i.d. We denote the ground-truth labeling rule as $t : X \to [0,1]$. 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 In a slight abuse of notation we will sometimes use $t(w)$ to indicate the label coordinate of an instance $w \in X \times \{0,1\}$.

A data representation is determined by a mapping $F : X \to Z$, for some set Z, and the learner only sees $F(x)$ for any instance x (both in the training and the test/decision stages). We denote the hypothesis class of all feature based decision rules as $\mathcal{H}_F = \{h : Z \to \{0,1\}\}$. As a loss function we consider the 0-1 loss for binary classification. We denote the true risk with respect to that loss by L_P.

3.1 Notions of group fairness

For our fairness analysis we assume the population X to be partitioned into two sub-populations A and D (namely, we restrict our discussion to the case of one binary protected attribute). We sometimes use a function notation $G : X \to \{A, D\}$ to indicate the group-membership of an instance. Of course in reality there are often many protected attributes with more than two values. However, as our goal is to show limitations and impossibility results for fair representation learning, it suffices to only consider one binary protected attribute – the same impossibilities readily follow for the more complex settings.

We now define two widely used notions of group-fairness that we will refer to throughout the paper, namely, equalized odds and demographic parity. In the following we will denote with $X_{g,l}$ the subset of X with label l and group membership g, i.e. $X_{g,l} = X \cap t^{-1}(l) \cap G^{-1}(g)$.

The notion of group-fairness we will focus on in this paper is the ground-truth-dependent notion of odds equality as introduced by Hardt et al. (2016).

Definition 1 (Group fairness; Equalized odds) *A classifier h is considered fair w.r.t. to odds equality (L^{EO}) and a distribution P if for $x \sim P$ we have the statistical independence $h(x) \perp\!\!\!\perp G(x)|t(x)$. For $g \in \{A, D\}$ let the false positive rate and the false negative rate be defined as $FPR_g(h,t,P) = \mathbb{P}_{x\sim P}[h(x) = 1|x \in X_{g,0}]$ and $FNR_g(h,t,P) = \mathbb{P}_{x\sim P}[h(x) = 0|x \in X_{g,1}]$ respectively. The EO unfairness is given then by the sum of differences in false positive rate and false negative rate between groups:*

$$L_P^{EO}(h) = \frac{1}{2}|FNR_A - FNR_D| + \frac{1}{2}|FPR_A - FPR_D|.$$

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

Definition 2 (Demographic parity) *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*

$$L_P^{DP}(h) = |\mathbb{P}_{x\sim P}[h(x) = 1|G(x) = A] - \mathbb{P}_{x\sim P}[h(x) = 1|G(x) = D]|$$

.

3.2 The role of the agent's objective

We will phrase our definitions of representation fairness in terms of a general group fairness notion L^{fair} with unfairness measure L_P^{fair}.

Most of this work considers 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 representation F is considered to be* adversarialy *fair w.r.t. the distribution P and group fairness objective L^{fair}, if every classifier $h \in \mathcal{H}_F$ is group-fair. We define the adversarial unfairness of a representation F by $U_{adv}(F) = \max_{h\in\mathcal{H}_F} L_P^{fair}(h)$.*

We also 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.

The following definition ensures that the Bayes optimal classifier for a representation is fair.

Definition 4 (Accuracy-driven fairness) *A representation F is considered to be* accuracy-driven *fair w.r.t. the fairness objective L^{fair} and distribution P, if every classifier $h \in \mathcal{H}_F$ with $L_P(h) = \min_{h \in \mathcal{H}_F} L_P(h)$ is group-fair with respect to this objective. The accuracy-driven unfairness is given by $U_{acc}(\mathcal{F}) = \max\{L_P^{fair}(h) : h \in \arg\min_{h \in \mathcal{H}_F} L_P(h)\}$.*

We note that in cases where the decision maker does not have access to the distribution P, but only to a labelled sample, this requirement might not be sufficient for guaranteeing that an accuracy-driven decision maker arrives at a fair decision.

Notions of fair representation can be defined with respect to any group-fairness notion. We will mainly focus on the equalized odds notion of fairness Hardt et al. (2016), but also have some results for demographic parity and predictive rate parity.

4 Can there be a generic fair representation?

We address the existence of a multi-task fair representation. We prove that for the adversarial agent scenario (which is the setup that most fairness representation previous work is concerned with), **it is impossible to have generic non-trivial fair representations** - no useful representation can guarantee fairness for all "downstream" classifications that are based on that representation (even if the ground truth classification remains unchanged and only the marginal may change between tasks).

We start by considering scenarios in which only the marginals shift between two tasks, e.g. two openings for different jobs, requiring similar skills, for which different pools 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. We show that we cannot guarantee fairness of a fixed data presentation for general shifts of this kind, even for the simple case of demographic parity.

Theorem 1 *No data representation can guarantee the DP fairness of any non-trivial classifier w.r.t. all possible data generating distributions (over any fixed domain set with any fixed partition into non-empty groups). That is, for every non-constant representation F, there exists a distribution P such that F is arbitrarily unfair with respect to L^{DP} and the task P (say $U_{adv}^{DP}(F) > 0.9$).*

We note that one can choose a distribution for this construction which allows for a natural interpretation. That is one can choose the marginal as a uniform distribution over finitely many points, which can be interpreted as an empirical distribution over a set of applicants. We further note that while in natural settings it might be unlikely to get a worst case selection of applicants, any shift in distribution/selection of applicants is likely to impact the fairness of the representation.

Next we prove a similar theorem for EO-fairness.

Theorem 2 *No data representation can guarantee EO fairness of any non-constant predictor based on that representation for all "downstream" classification learning tasks. Concretely,*

1. *Given any representation F that is expressive enough to allow classifiers that are not constant on each group, there is a distribution P over $X \times \{0, 1\}$ such that F is arbitrarily adversarially unfair with respect to L^{EO} and P (i.e. $U_{adv}^{EO}(F) \geq 0.5$).*
2. *Let f be a labeling rule and F any representation that can express a non-constant function that is different from f and $1 - f$. Then there exists a marginal P_X, such that F is arbitrarily adversarially unfair with respect to L^{EO} and (P_X, f) (i.e., $U_{adv}^{EO}(F) \geq 0.5$).*

The proof of this theorem can be found in the appendix. We again note that this effect occurs for fairly generic distributions.

The results above showed that there is no representation that can guarantee fairness for an arbitrary task. But what happens if we limit our discussion to a predefined selection of tasks? We will show that even in this restricted case, there can be no representation that guarantees EO fairness with respect to a general predefined selection of tasks. We say a distribution P has *equal success rates for both groups*, if both groups have the same conditional probability of label 1, i.e. $P[t(x) = 1|x \in A] = P[t(x) = 1|x \in D]$. We will now state the main result of this section.

Theorem 3 *Let P_1 and P_2 be the distributions defining two different tasks*[1] *with the same marginal $P_X = P_{1,X} = P_{2,X}$ such that both $P_X(A) \neq 0$ and $P_X(D) \neq 0$ and at least one of the tasks does not have equal success rates for both group. There can be no data representation F such that for P_1, P_2, the following criteria simultaneously hold:*

1. *F is adversarially fair w.r.t. P_1 and EO*
2. *F is adversarially fair w.r.t. P_2 and EO*
3. *F enables the expression of classifiers that have perfect accuracy w.r.t. to P_1 and P_2, i.e., there are h_1, h_2 both expressible over the representation F, such that $L_{P_1}(h_1) = L_{P_2}(h_2) = 0$.*

In order to prove this theorem we use the following lemma.

Lemma 1 *Pick any set X and a partition of X into two non-empty (disjoint) sets A and D. Let P_X be any probability distribution over X such that both $P_X(A) \neq 0$ and $P_X(D) \neq 0$. Let $f, g : X \mapsto \{0, 1\}$ such that $P_X[\{x : f(x) \neq g(x)\}] \notin \{0, 1\}$. If f is a EO fair classification w.r.t. (P_X, g) (as the labeling rule) and g is a EO fair classification w.r.t. (P_X, f) (as the labeling rule), then $P_X[f(x) = 1|A] = P_X[f(x) = 1|D]$ and $P_X[g(x) = 1|A] = P_X[g(x) = 1|D]$.*

This lemma can be deduced from the impossibility result of Kleinberg et al. (2016), namely, that a classifier cannot fulfill predictive rate parity and equalized odds in cases in which there is a difference in success rates between the two groups. For that, note that the condition of two classifiers f and g being EO fair with respect to each other as the labeling rule is equivalent to f being EO fair and having predictive rate parity with respect to some underlying task with labeling rule g.

For completeness, we provide a direct proof of the lemma in the Appendix 8.

Now we can prove our theorem.

Proof of Theorem 3: Towards a contradiction, let us assume that F was adversarially EO fair with respect to both P_1 and P_2 and that both h_1 and h_2 can be expressed by the representation. This implies that both h_1 and h_2 need to be EO fair with respect to P_1 and P_2. From Lemma 1, we know that this implies that $P_X[h_1(x) = 1|A] = P_X[h_1(x) = 1|D]$ and $P_X[h_2(x) = 1|A] = P_X[h_2(x) = 1|D]$ or that $P_X[\{x : h_1(x) \neq h_2(x)\}] = 0$. However, we have assumed that $P_X[\{x : h_1(x) \neq h_2(x)\} \neq 0$ and that P_1 or P_2 do not have equal success rates of groups. This concludes our proof.

5 Impossibility of adversarially fair representations with respect to predictive rate parity

We now show that not all common notions of group fairness always allow a adversarially fair representation, even in a single-task setting. One such notion is *predictive rate parity*.

Definition 5 *(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 $h(x)$. We denote this fairness objective with L^{Pred}.*

Theorem 4 *Adversarial fairness w.r.t. P and L^{Pred} is only possible, if P has equal success rates for both groups.*

Proof of Theorem 4: We note that in order to achieve adversarial fairness with respect to any representation, the all-one classifier needs to be fair, as any representation F admits any constant classifier. We furthermore note that the all-one classifier is fair with respect to predictive rate parity if and only if the ground truth has equal success rates. This shows our claim.

[1] Namely, their labeling rules are different from each other and are not the exact opposite of each other. Formally, both $\{x : P_1[y = 1|x] \neq P_2[y = 1|x]\}$ and $\{x : P_1[y = 1|x] \neq P_2[y = 0|x]\}$ have nonzero probability under the joint marginal distribution.

6 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. For this we assume our representation F to consist of finitely many features $f_i : X \to Y_i$ i.e. for every $x \in X : F(x) = (f_1(x), \dots, f_n(x))$ and $Z = Y_1 \times \dots \times Y_n$. We limit our discussion to cases where all Y_i are finite. While this assumption facilitates our analysis, similar results can be shown in the cases of continuous features. We will denote the set of features as $F = \{f_1, \dots, f_n\}$. Unless otherwise stated, we focus on the equalized Odds (EO) notion of group fairness. We denote by $U_{adv}(\mathcal{F})$ and $U^{\alpha}_{acc}(\mathcal{F})$ the adversarial and accuracy-driven EO fairness of the representation induced by the feature set $\mathcal{F}$ respectively. We show that it is in general not possible to determine the effect a single feature has on the fairness of a representation without considering the full representation. This is the case even if the considered feature is not correlated with the protected attribute.

6.1 Opposing effects of a feature for accuracy-driven fairness of a representation

We start our discussion with accuracy-driven fairness w.r.t. equalized odds. In this case we show that the deletion of a feature f can lead to an increase in accuracy-driven unfairness 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 unfairness for another set of other available features $\mathcal{F}'$. This implies that the fairness of the feature f cannot be evaluated without context. We show that 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-committing), and on the other hand do not reveal identity when label and group information is given (k-anonymity Samarati & Sweeney (1998))). We will start by stating the non-triviality requirements for our theorem.

Non-Triviality properties

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

1. **Non-committing** *We will call a feature* non-committing *if it leaves some ambiguity about label and group membership. That is, a feature f is non-committing 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.*

Theorem 5 *(Context-relevance for fairness of features) For every 2-anonymous non-committing 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} and P of $\mathcal{F} \cup \{f\}$ is greater than that of $\mathcal{F}$, i.e.*

$$U_{acc}(\mathcal{F} \cup \{f\}) < U_{acc}(\mathcal{F})$$

  Thus, deleting f in this context will increase unfairness.

- *The accuracy-driven fairness w.r.t L^{EO} and P of $\mathcal{F}' \cup \{f\}$ is less than that of $\mathcal{F}'$, i.e.*

$$U_{acc}(\mathcal{F}' \cup \{f\}) > U_{acc}(\mathcal{F}')$$

  Thus, deleting f in this context will decrease unfairness.

We note that this phenomenon can occur for quite general pairs (f, P) and that we mainly need to exclude pathological cases for our construction to work. In particular we want to note that this phenomenon can occur even if f is uncorrelated with the group membership and the label for ground-truth distribution P. We will give an example illustrating our last point and will refer the reader for the proof and a general discussion of the requirements on (f, P) to the appendix. Before giving our example, we need to introduce some concepts.

Feature-induced cells A set of features $\mathcal{F} = \{f_1, \dots, f_n\}$ induces an equivalence relation $\sim_{\mathcal{F}}$, by $x \sim_F$ iff $f_i(x) = f_i(y)$ for all $i = 1, \dots, n$. We call the equivalence classes with respect to $\sim_{\mathcal{F}}$ cells and denote the set of cells for a featureset $\mathcal{F}$ as $\mathcal{C}_{\mathcal{F}}$.

Ground-truth score function 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 unambiguous we will use the abbreviated notation s_t instead of s_t^P.

Bayes-optimal predictor The predictor in $\mathcal{H}_F$ that minimizes L_P is the *Bayes Optimal predictor* $t_{P,F}$ that for a cell $C \in \mathcal{C}_\mathcal{F}$ assigns the label 1 if $s_t(C) > 0.5$ and 0 otherwise.

We will now give an example in which both f and $\mathcal{F}$ are adversarially fair w.r.t. P and in which the phenomenon from Theorem 5 holds.

Example 1 *Let the domain* $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\}$ *with* $X_{A,1} = \{x_1, x_2, x_3\}, X_{D,1} = \{x_4, x_5, x_6\}, X_{A,0} = \{x_7, x_8, x_9\}$*, and* $X_{D,0} = \{x_{10}, x_{11}, x_{12}\}$*. Furthermore consider the uniform distribution* P *over* X*, i.e.* $P(\{x\}) = \frac{1}{12}$ *for every* $x \in X$*. For the construction of the feature set, we only consider binary features* $f_i : X \to \{0, 1\}$*. Now let* f *be defined by* $f^{-1}(1) = \{x_1, x_5, x_8, x_{12}\}$*. Furthermore, let* $\mathcal{F} = \{f_1, f_2, f_3\}$ *and* $\mathcal{F}' = \{f_1', f_2'\}$ *with* $f_1^{-1}(1) = \{x_1, x_2, x_3, x_5, x_8, x_{12}\}$*,* $f_2^{-1}(1) = \{x_1, x_2, x_3, x_5, x_{11}, x_{12}\}$*,* $f_3^{-1}(1) = \{x_1, x_4 x_5, x_6, x_7, x_{11}\}$*,*$f_1'^{-1}(1) = \{x_1, x_4, x_7, x_{10}\}$ *and* $f_2'^{-1}(1) = \{x_1, x_2, x_4, x_5, x_7, x_8, x_{10}, x_{11}\}$*. The resulting cells for* $\mathcal{F}$ *and* $\mathcal{F}'$ *are*

$$\mathcal{C}_\mathcal{F} = \{\{x_1, x_5\}, \{x_2, x_3, x_{12}\}, \{x_8\}, \{x_4, x_6, x_7\}, \{x_9\}, \{x_{10}, x_{11}\}\}$$

and

$$\mathcal{C}_{\mathcal{F}'} = \{\{x_1, x_4, x_7, x_{10}\}, \{x_2, x_5, x_8, x_{11}\}, \{x_3, x_6, x_9, x_{12}\}\}.$$

It is easy to see that $\mathcal{F}'$ *and* $\{f\}$ *are adversarially fair w.r.t.* P *and* L^{EO}*. Furthermore, we have:*

$$U_{acc}(\mathcal{F} \cup \{f\}) = \frac{1}{2}|\frac{3}{3} - \frac{2}{3}| + \frac{1}{2}|\frac{2}{3} - \frac{1}{3}| = \frac{1}{3} > 0 = U_{acc}(\mathcal{F})$$

and

$$U_{acc}(\mathcal{F}' \cup \{f\}) = \frac{1}{2}|\frac{3}{3} - \frac{3}{3}| + \frac{1}{2}|\frac{1}{3} - \frac{1}{3}| = 0 <$$

$$\frac{1}{6} = \frac{1}{2}|\frac{3}{3} - \frac{3}{3}| + \frac{1}{2}|\frac{1}{3} - \frac{0}{3}| = U_{acc}(\mathcal{F}').$$

Thus we see that there are indeed features f *which are adversarially fair w.r.t.* P *and equalized odds, for which there is this opposing effect of feature deletion.*

6.2 The fairness of a feature dependent on agent's objective

We will now briefly discuss the effect of a single feature on fairness for the case of an adversarial agent. In contrast to the accuracy-driven case, adding features has a monotone effect on the fairness of a malicious decision maker. We show in Theorem 6 that adding any feature in the adversarial case, will only give the decision maker more information and thus give the decision maker more chances of discrimination. However, the quantitative effect of adding a feature on the unfairness can still range from having no effect to achieving maximal unfairness. As in the accuracy-driven case, we will show (Theorem 6) that it is impossible to evaluate the quantitative effect of a feature on the fairness of a representation without considering the context of other available features.

Theorem 6

1. *For any feature* f *and any featureset* $\mathcal{F}$ *we have* $U_{adv}(\mathcal{F}) \leq U_{adv}(\mathcal{F} \cup \{f\})$*.*
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.* $U_{adv}(\mathcal{F}) = U_{adv}(\mathcal{F} \cup \{f\})$*.*
3. *There exist distributions* P*, features* f *and* $\mathcal{F}'$*, such that* $U_{adv}(\mathcal{F}') = 0$ *and* $U_{adv}(\{f\}) = 0$*, but* $U_{adv}(\mathcal{F}' \cup \{f\}) = 1$ *.*

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. In particular, an analogous statement to Theorem 6 can be made for demographic parity.

7 Conclusion

While many papers in this domain propose algorithmic solutions to fairness related issues, the main contributions of this paper are conceptual. We believe that, to a much larger extent than many other facets of machine learning, fundamental concepts of fairness in machine learning require better understanding. Some basic questions are still far from being satisfactorily elucidated; What should be considered fair decision making? (various mutually incompatible notions have been proposed, but how to pick between them for a given real life application is far from being clarified). What is a fair data representation? To what extent should accuracy or other practical utilities be compromised for achieving fairness goals? and more. 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 these as the primary issues facing the field, deserving explicit research attention (in addition to the more commonly discussed algorithmic and optimization aspects).

Our main result addressed the existence of generic fair representations. We show that even label-independent fairness notions like demographic parity are vulnerable to shifts in marginals between tasks. For fairness notions that do rely on the true classification, we show that fairness and accuracy cannot be simultaneously achieved by the same data representation for any two different tasks even if they are defined over the same marginal (unlabeled) data distributions. We conclude the impossibility of having generic data representations that guarantee (even just) DP fairness with respect to tasks whose marginal distributions are not accessible when designing the representation.

These insights stand in contrast to the impression arising from many recent papers Madras et al. (2018); Edwards & Storkey (2016); McNamara et al. (2019); Song et al. (2019); Creager et al. (2019); Madras et al. (2018) that claim to learn transferable fairness-ensuring representations.

Furthermore, we showed that some fairness notions, like predictive rate parity, do not always allow an adversarially fair representation, even if it is just for a single task.

Lastly, we also considered the question of "fairness of a feature", which has been used in legal scenarios. We showed that 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[2]

One obvious direction for further research is extending our impossibility results to quantitative accuracy-fairness trade-offs and bounds on what a data representation can guarantee over multiple tasks as a function of appropriate measures of task similarities.

Our impossibility results imply that claims about representations aimed to guarantee fairness should come with specifications of the scope of tasks/ distributions for which those guarantees apply. Ideally, the property of whether the representation is (approximately) fair with respect to a given task could be tested based on samples of a bounded size.

8 Acknowledgements

We would like to thank Sushant Agarwal and Nivasini Anathakrishnan for helpful discussions and comments as well as help with the related work section. We would like to thank the Vector Institute for their generous support.

References

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1–2), 2010.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, abs/1707.00075, 2017.

Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *ICML*, 2019.

[2] While we focused on the equalized odds notion of fairness, similar results can be shown for demographic parity (i.e. a feature that has demographic parity by itself can still make a representation demographic-parity unfair (in the adversarial sense) and for other common notions of group fairness. This is simply due to the fact that *pairwise* statistical independence for a set of random variables does not imply statistical independence of the set of random variables.

Jennifer L Doleac and Benjamin Hansen. 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, 2016.

Flávio du Pin Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems 30*.

Harrison Edwards and Amos J. Storkey. Censoring representations with an adversary. In *ICLR*, 2016.

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *AAAI*, 2018.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *ICML*, 2018.

Daniel McNamara, Cheng Soon Ong, and Robert C Williamson. Costs and benefits of fair representation learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 263–270, 2019.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5680–5689, 2017.

Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, 1998.

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

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional learning of fair representations. *CoRR*, abs/1910.07162, 2019.

Appendix

Additional remarks on Theorem 5

Theorem 5 stated that every feature f fulfilling some non-triviality requirements, there exists a distribution P and feature sets $\mathcal{F}$ and $\mathcal{F}'$ such that adding f to either of the feature sets has opposing effects on the accuracy-driven fairness of the respective representations. We will now state a condition on f and P for this phenomenon to occur. It will be easy to see that this condition is fulfilled for a very general class of distributions and features, only excluding pathological examples.

Definition 7 *In the following let $l_1 \in \{0, 1\}$ denote a label and $G_1 \in \{A, D\}$ a group. The opposing label and group will be denoted by l_2 and G_2 respectively. A pair (f, P) of a feature f and a distribution P is called* generic *if there exist sets $C_1, C_2, C_3 \subset X$ with the following properties.*

1. *$P(C_1) > P(C_2)$*
2. *C_1 and C_2 are separated by the feature f, i.e. there are $y_1 \neq y_2$ such that $C_1 \subset f^{-1}(y_1)$ and $C_2 \subset f^{-1}(y_2)$*
3. *C_1 and C_2 are label-homogeneous for different labels and C_2 is group homogeneous, i.e. $C_1 \subset t^{-1}(l_1)$ and $C_2 \subset X_{G_1,l_2}$.*
4. *C_3 is not split by the feature, i.e. there is y_3 such that $C_3 \subset f^{-1}(y_3)$*
5. *C_3 has the same majority label as C_1, i.e. $P(t^{-1}(l_1) \cap C_3) \geq P(t^{-1}(l_2) \cap C_3)$*
6. *The fraction of elements of group G_2 and label l_2 in C_3 is sufficiently big in comparison to C_2, i.e. $\frac{P(C_3 \cap X_{G_2,l_2})}{P(X_{G_2,l_2})} \geq \frac{P((C_2 \cup C_3) \cap X_{G_1,l_2})}{P(X_{G_1,l_2})}$.*

Lemma 2 *For every pair generic feature-distribution pair (f, P), there are two feature sets $\mathcal{F}$ and $\mathcal{F}'$*

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

  $$U_{acc}(\mathcal{F} \cup \{f\}) < U_{acc}(\mathcal{F})$$

  Thus, deleting f in this context will increase unfairness.

- *The accuracy-driven fairness w.r.t L^{EO} and P of $\mathcal{F}' \cup \{f\}$ is less than that of $\mathcal{F}'$, i.e.*

  $$U_{acc}(\mathcal{F}' \cup \{f\}) > U_{acc}(\mathcal{F}')$$

  Thus, deleting f in this context will decrease unfairness.

Proof: We define $\mathcal{F}$ as a representation which separates everything but a cell $C' = C_1 \cup C_2$ by labels. For such a representation $\mathcal{F} \cup \{f\}$ enables perfect accuracy and therefore perfect fairness. However $\mathcal{F}$ is constructed that the optimal classifier with respect to 0-1 loss is unfair, as only elements of X_{G_1,l_2} are misclassified. Furthermore we can define $\mathcal{F}'$ as a representation that separates all but two cells $C' = C_1 \cup C_2$ and $C'' = C_3$ perfectly by labels. As the only misclassification of Bayes classifier $t_{P,\mathcal{F}' \cup \{f\}}$ occurs on C_3 and it labels $t_{P,\mathcal{F}' \cup \{f\}} = l$ it has unfairness $L_P^{EO}(t_{P,\mathcal{F}' \cup \{f\}}) = \frac{1}{2} | \frac{P(C_3 \cap X_{G_2,l_2})}{P(X_{G_2,l_2})} - \frac{P(C_3 \cap X_{G_1,l_2})}{P(X_{G_1,l_2})} |$. Furthermore the only misclassification for the Bayes classifier $t_{P,\mathcal{F}'}$ occurs on C_2 and C_3 which are both labeled as l, yielding the unfairness $t_{P,\mathcal{F}'} = l$ it has unfairness $L_P^{EO}(t_{P,\mathcal{F}'}) = \frac{1}{2} | \frac{P(C_3 \cap X_{G_2,l_2})}{P(X_{G_2,l_2})} - \frac{P((C_2 \cup C_3) \cap X_{G_1,l_2})}{P(X_{G_1,l_2})} |$. As $\frac{P((C_2 \cup C_3) \cap X_{G_1,l_2})}{P(X_{G_1,l_2})} > \frac{P(C_3 \cap X_{G_1,l_2})}{P(X_{G_1,l_2})}$, by property (6.) of Definition 7, we thus get $L_P^{EO}(t_{P,\mathcal{F}' \cup \{f\}}) > L_P^{EO}(t_{P,\mathcal{F}'})$, concluding our proof.

Lemma 3 *For every non-committing, 2-anonymous feature f, there exists a distribution P, such that the pair (f, P) is generic.*

Proof: We need to show that it is possible to define three sets $C_1, C_2, C_3, C_4 \subset X$ and a distribution P such that the requirements of Definition 7 are fulfilled. From the fact that f is non-committing we know that there are y_1, y_2 such that none of the subsets $f^{-1}(y_1) \cap X_i$ and $f^{-1}(y_2) \cap X_i$ is empty for any $X_i \in \{X_{A,0}, X_{A,1}, X_{D,1}, X_{D,0}\}$. We can thus define the non-empty set $B = f^{-1}(y_2) \cap X_{A,0}$. Furthermore, we know that f is also 2-anonymous and thus we can split B further into two non-empty subsets C_2 and C_4. Furthermore, we can define C_1 and C_3 as disjoint non-empty subsets of $f^{-1}(y_1)$, such that $C_1 \subset f^{-1}(y_1) \cap t^{-1}(1)$ and such that $C_3 \cap X_i \neq \emptyset$ for any $X_i \in \{X_{A,0}, X_{A,1}, X_{D,1}, X_{D,0}\}$. Thus the properties (2.), (3.) and (4.) of the non-generic definition are fulfilled for the sets C_1, C_2, C_3.

We can now choose P to pick probability weights as follows:

- $P(C_1) = 0.2$
- $P(C_2) = 0.1$
- $P(C_3 \cap t^{-1}(1)) = 0.3$
- $P(C_3 \cap X_{D,0}) = 0.2$
- $P(C_4) = 0.2$

Clearly (1.) is fulfilled as $P(C_1) = 0.2 > 0.1 = P(C_2)$. Furthermore (5.) is fulfilled as, $P(C_3 \cap t^{-1}(1)) = 0.3 > 0.2 = P(C_3 \cap X_{D,0}) = P(C_3 \cap t^{-1}(0))$. Lastly, (6.) is fulfilled as:

$$\frac{P(C_3 \cap X_{D,0})}{P(X_{D,0})} = 1 < \frac{1}{3} = \frac{P(C_2 \cap X_{A,0})}{P(X_{A,0})}$$

PROOFS

Proof of Theorem 1:

Pick any domain set X and any partition of X into non-empty subsets A, D. We first show that for every non-constant function $f : X \to \{0,1\}$ there exists a probability distribution P over X such that f is arbitrarily DP-unfair w.r.t. P (i.e., $L_P^{\text{DP}}(h) > 0.9$).

If f is constant on any of the groups A or D then, since f is not a constant over X there are points in the other group on which f has the opposite value. Thus, from f not being constant, we can conclude that there are two labels $y_1 \neq y_2$, such that the sets $\{x \in A : f(x) = y_1\}$ and $\{x \in D : f(x) = y_2\}$ are both non-empty. Now we choose the marginal P_X to assign probability 0.5 to $\{x \in A : f(x) = y_1\}$ and probability 0.5 to $\{x \in D : f(x) = y_2\}$. Clearly f fails DP w.r.t. this P.

Now if a representation F is non-constant, it allows some non-constant function f using that representation. Thus no non-constant representation can fulfill adversarial demographic parity with respect to every distribution P.

Proof of Theorem 2:

We note that 1.) follows directly from 2.) and will now show 2.). We know that F can express a non-constant classifier $h : X \to \{0,1\}$ with $h \neq f$ and $h \neq 1 - f$ (that is, for some $x \in X$, $f(x) = h(x)$). We will now show that there exists a marginal P_X, such that h has high unfairness with respect to L^{EO} and $P = (P_X, f)$, (i.e. $L_P^{\text{EO}}(h) \geq 0.5$).

Let $f : X \to \{0,1\}$ be any non-constant function and $h : X \to \{0,1\}$ be any non-constant classifier with $h \neq f, 1 - f$. Then we know that at least three of the four sets $\{x \in X : f(x) = 1, h(x) = 0\}$, $\{x \in X : f(x) = 0, h(x) = 1\}$, $\{x \in X : f(x) = 1, h(x) = 1\}$ and $\{x \in X : f(x) = 0, h(x) = 0\}$ are non-empty. Thus, there exist two sets B_1 and B_2 among these sets, on which the ground truth function f assigns the same label. That is, for every $x_1 \in B_1$ and every $x_2 \in B_2$ we have $f(x_1) = f(x_2)$. W.l.o.g. $B_1 = \{x \in X : f(x) = 1, h(x) = 0\}$, $B_2 = \{x \in X : f(x) = 1, h(x) = 1\}$. Let B_3 be the remaining of the three sets that are guaranteed to be non-empty. We note, that for any set B, we have $B = (B \cap A) \cup (B \cap D)$. Thus for a non-empty set B, $B \cap A = \emptyset$ implies $B \cap D \neq \emptyset$ and $B \cap D = \emptyset$ implies $B \cap A \neq \emptyset$. We thus get a distinction into the following cases:

- Case 1: $B_1 \cap A \neq \emptyset$ and $B_2 \cap D \neq \emptyset$. Then we can choose the marginal P_X as $P_X(B_1 \cap A) = 0.5$ and $P_X(B_2 \cap D) = 0.5$. Yielding, $L_P^{\text{EO}}(h) = 0.5$
- Case 2: $B_2 \cap A \neq \emptyset$ and $B_1 \cap D \neq \emptyset$: Analogous to Case 1
- Case 3: there is $G \in \{A, D\}$, such that $B_1 \cap G = B_2 \cap G = \emptyset$. W.l.o.g. $G = A$. Then $B_3 \cap A \neq \emptyset$ and $B_1 \cap D \neq \emptyset$ and $B_2 \cap D \neq \emptyset$. In this case we can choose the marginal as $P_X(A \cap B_3) = 0.5$ and $P_X(D \cap B_1) = 0.5$. Then all elements of D will be misclassified and all elements of A will either be classified correctly or be misclassified in the opposite direction, yielding to high EO unfairness. (In the case where the ground truth labeling is constant on one group, we define the misclassification rate with respect to the label it will not achieve to be zero. Then we get $L_P^{\text{EO}}(h) \geq 0.5$.)

Lastly, $L_P^{\text{EO}}(h) \geq 0.5$ implies $U_{adv}^{EO}(F) \geq 0.5$, concluding our proof.

Proof of Lemma 1: Consider the following four sets: $S = \{x : f(x) = 1,\ g(x) = 0\}$, $T = \{x : f(x) = 1,\ g(x) = 1\}$, $U = \{x : f(x) = 0,\ g(x) = 1\}$, $V = \{x : f(x) = 0,\ g(x) = 0\}$.

Let S_A, T_A, U_A, V_A, denote the intersections of these sets with the set A, (e.g., $S_A = S \cap A$), and similarly, S_D, T_D, U_D, V_D, denote the intersections of these sets with the set D. Notice that

- $P[f(x) = 1|A] = \frac{P(S_A)+P(T_A)}{P(A)}$.
- $P[f(x) = 1|D] = \frac{P(S_D)+P(T_D)}{P(D)}$.
- $P[g(x) = 1|A] = \frac{P(T_A)+P(U_A)}{P(A)}$.
- $P[g(x) = 1|D] = \frac{P(T_D)+P(U_D)}{P(D)}$.

It follows that once one shows that each of these quantities can be expressed in terms of the false positive and false negative rates when each of f or g is considered the true classification and the other as the predicted labeling, then the conclusion of the lemma is implied by its EO assumptions.

Using the above notation, when f is the true classification,

$\text{FPR}_A(g, f, P) = \frac{P(U_A)}{P(V_A)+P(U_A)}$ and $\text{FNR}_A(g, f, P) = \frac{P(S_A)}{P(S_A)+P(T_A)}$ (and similarly for D).

And when the true classification is g,

$\text{FPR}_A(f, g, P) = \frac{P(S_A)}{P(V_A)+P(S_A)}$ and $\text{FNR}_A(f, g, P) = \frac{P(U_A)}{P(U_A)+P(T_A)}$ (and similarly for D).

We will start with the case where all eight sets U_A, V_A, S_A, T_A and U_D, V_D, S_D, T_D are non-empty. We note, that in this case equalized false positive rates and false negative rates of f with respect to g gives us the following two equations,

$$\frac{P(U_A)}{P(V_A) + P(U_A)} = \frac{P(U_D)}{P(V_D) + P(U_D)},$$

and

$$\frac{P(S_A)}{P(S_A) + P(T_A)} = \frac{P(S_D)}{P(S_D) + P(T_D)}.$$

This implies that there are two constants β_1, β_2 with $P(U_A) = \beta_1 P(V_A)$ and $P(U_D) = \beta_1 P(V_D)$ and $P(S_A) = \beta_2 P(T_A)$ and $P(S_D) = \beta_2 P(T_D)$.

Furthermore, g being EO fair with respect to f gives us

$$\frac{P(S_A)}{P(V_A) + P(S_A)} = \frac{P(S_D)}{P(V_D) + P(S_D)},$$

and

$$\frac{P(U_A)}{P(U_A) + P(T_A)} = \frac{P(U_D)}{P(U_D) + P(T_D)}.$$

This implies that there is a constant β_3 such that $P(V_A) = \beta_3 P(S_A)$ and $P(V_D) = \beta_3 P(S_D)$.

Thus,

$$P[f(x) = 1|A] = \frac{\beta_2 + 1}{\beta_2 + 1 + \beta_3\beta_2(1 + \beta_1)} = P[f(x) = 1|D],$$

and

$$P[g(x)=1|A] = \frac{1+\beta_1\beta_2\beta_3}{\beta_2+1+\beta_3\beta_2(1+\beta_1)} = P[g(x)=1|D].$$

The cases in which one or several of these sets are empty can be shown in an analogous way. This proves our claim.

Proof of Theorem 5: The result follows directly from Lemma 3 and Lemma 2.

Proof of Theorem 6:

1. We note that $\mathcal{H}_\mathcal{F} \subset \mathcal{H}_{\mathcal{F}\cup\{f\}}$. Thus any $\arg\min_{h\in\mathcal{H}_\mathcal{F}} L_P^{\text{EO}}(h) \leq \arg\min_{h\in\mathcal{H}_{\mathcal{F}\cup\{f\}}} L_P^{\text{EO}}(h)$, proving the inequality for adversarial fairness.
2. For any distribution P and feature f we can choose a representation $\mathcal{F}$ such that $\mathcal{C}_\mathcal{F} = \mathcal{C}_{\mathcal{F}\cup\{f\}}$. It is obvious that the fairness will not change between those representations.
3. The following example establishes the second claim: Consider the domain $X = \{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$ with $X_{A,1} = \{x_1,x_2\}$,$X_{D,1} = \{x_3,x_4\}$, $X_{A,0} = \{x_5,x_6\}$ and $X_{D,0} = \{x_7,x_8\}$,. Furthermore let $\mathcal{F} = \{f_1,f_2\}$ with $f_1^{-1}(1) = \{x_1,x_3,x_5,x_7\}$ and $f_2^{-1}(1) = \{x_1,x_4,x_5,x_8\}$. Furthermore let P be uniform over X,i.e. $P(\{x_1\}) = P(\{x_2\}) = P(\{x_3\}) = P(\{x_4\}) = P(\{x_5\}) = P(\{x_6\}) = P(\{x_7\}) = P(\{x_8\}) = 0.125$. Thus, we have adversarial fairness w.r.t. EO for both features, i.e.

$$\frac{P(X_{A,1}\cap f_1^{-1}(1))}{P(t^{-1}(1)\cap f_1^{-1}(1))} = \frac{P(\{x_1\})}{P(\{x_1,x_2\})} = 0.5 =$$

$$\frac{P(\{x_3\})}{P(\{x_3,x_4\})} = \frac{P(X_{D,1}\cap f_1^{-1}(1))}{P(t^{-1}(1)\cap f_1^{-1}(1))}.$$

$$\frac{P(X_{A,0}\cap f_1^{-1}(1))}{P(t^{-1}(0)\cap f_1^{-1}(1))} = \frac{P(\{x_5\})}{P(\{x_5,x_6\})} = 0.5 =$$

$$= \frac{P(\{x_7\})}{P(\{x_7,x_8\})} = \frac{P(X_{D,0}\cap f_1^{-1}(1))}{P(t^{-1}(0)\cap f_1^{-1}(1))}.$$

$$\frac{P(X_{A,1}\cap f_2^{-1}(1))}{P(t^{-1}(1)\cap f_2^{-1}(1))}\frac{P(\{x_1\})}{P(\{x_1,x_2\})} = 0.5 =$$

$$= \frac{P(\{x_4\})}{P(\{x_3,x_4\})} = \frac{P(X_{D,1}\cap f_2^{-1}(1))}{P(t^{-1}(1)\cap f_2^{-1}(1))}.$$

$$\frac{P(X_{A,0}\cap f_2^{-1}(1))}{P(t^{-1}(0)\cap f_2^{-1}(1))}\frac{P(\{x_5\})}{P(\{x_5,x_6\})} = 0.5$$

$$= \frac{P(\{x_8\})}{P(\{x_7,x_8\})} = \frac{P(X_{D,0}\cap f_2^{-1}(1))}{P(t^{-1}(0)\cap f_2^{-1}(1))}.$$

However, the featureset $\mathcal{F}$ does not have adversarial fairness w.r.t. EO: $\mathcal{C}_\mathcal{F} = \{C_1,C_2,C_3,C_4\}$ with $C_1 = \{x_1,x_5\}$, $C_2 = \{x_2,x_6\}$, $C_3 = \{x_3,x_7\}$, and $C_4 = \{x_4,x_8\}$. Consider the classifier $h \in \mathcal{H}_\mathcal{F}$ with $h^{-1}(1) = \{C_1,C_2\}$. Then $L_P^{\text{EO}}(h) = \frac{1}{2}\sum_{l\in\{0,1\}}\left|\frac{P(h^{-1}(|1-l|)\cap X_{A,l})}{P(X_{A,l})} - \frac{P(h^{-1}(|1-l|)\cap X_{D,l})}{P(X_{D,l})}\right| = \frac{1}{2}(|1-0|+|0-1|) = 1$. Thus $U_{\text{adv}}^{EO}(\mathcal{F}) = 1$.

Characterizations of different notions of fair representations

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 8 (Zero-group-knowledge) *A representation 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 7 *A representation 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 et al. (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 over the cells $\mathcal{C}_\mathcal{F}$ of a finite feature set $\mathcal{F}$. In the following we will denote the conditional probabilities given label l and group G as $P_{G,l}$. We will see that a representation is adversarially 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 8 *A feature set $\mathcal{F}$ is adversarially 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 9 (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 9 *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 10 ($\mathcal{H}$-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.

In the following let $\mathcal{H}^{\text{thres}}_{\mathcal{C}_\mathcal{F},s_t} = \{h : \mathcal{C}_\mathcal{F} \to \{0,1\} : \text{for some } \alpha,\ h(C) = 0 \text{ iff } s_t(C) < \alpha\}$ be the class of all classifiers that are a threshold in the ground-truth scoring. We can now state a quantitative theorem about the relation between the conditional alignment and the fairness of a representation:

Theorem 10 *We can bound adversarial fairness and accuracy enforced fairness of a feature set $\mathcal{F}$ w.r.t. P and L^{EO} by the $d_{\mathcal{H}_\mathcal{F}}$-difference and $d_{\mathcal{H}^{thres}_{\mathcal{C}_\mathcal{F},s_t}}$-difference of conditional distributions respectively:*

$$U_{adv}(\mathcal{F}) \leq \frac{1}{2} d_{\mathcal{H}_\mathcal{F}}(P_{A,1}, P_{D,1}) + \frac{1}{2} d_{\mathcal{H}_\mathcal{F}}(P_{A,0}, P_{D,0})$$

$$U_{acc}(\mathcal{F}) \leq \frac{1}{2} d_{\mathcal{H}^{thres}_{\mathcal{C}_\mathcal{F},s_t}}(P_{A,1}, P_{D,1}) + \frac{1}{2} d_{\mathcal{H}^{thres}_{\mathcal{C}_\mathcal{F},s_t}}(P_{A,0}, P_{D,0})$$

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

$$\frac{1}{2} d_{\mathcal{H}_\mathcal{F}}(P_{A,l}, P_{D,l}) \leq U_{adv}(\mathcal{F})$$

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.

From Theorem 4 we know 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, as we will see in the following theorem.

Definition 11 *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. $G(x) \perp\!\!\!\perp t(x) | F(x)$.*

Theorem 11 *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*

Theorem 12 *A feature set $\mathcal{F}$ has demographic parity w.r.t. P if and only if it has adversarially fair w.r.t P and the group-fairness objective L^{DP}.*

Impact of a feature on fairness for other group fairness notions

We can make another observation about the impact of feature deletion on unfairness for other notions of group fairness.

Observation 1

- *There exists a distribution P and a feature set $\mathcal{F}$ such that each $f \in \mathcal{F}$ the feature set $\{f\}$ has zero-group-knowledge w.r.t. P, but $\mathcal{F}$ is not and $U_{adv}(\mathcal{F}) = 1$*
- *There exists a distribution P and a feature set $\mathcal{F}$ such that each $f \in \mathcal{F}$, the feature set $\{f\}$ has demographic parity w.r.t. P, but $\mathcal{F}$ has not. Furthermore the group-membership can be perfectly determined by $\mathcal{F}$, i.e. for every cell $C \in \mathcal{C}_\mathcal{F}$ we have*
$$\mathbb{P}_{x \sim P}[x \in A | x \in C] \in \{0, 1\}$$
- *There exists a distribution P and a feature set $\mathcal{F}$ such that each $f \in \mathcal{F}$, the feature set $\{f\}$ has calibration parity w.r.t. P, but $\mathcal{F}$ has not. Furthermore the scores for the different groups in each cell are perfectly opposed, i.e. $C \subseteq A$ or $C \subseteq D$.*

Proofs

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}. **Proof of Theorem 7:**

$$\frac{P(h^{-1}(1) \cap X_{G,l})}{P(X_{G,l})} = \frac{\sum_{C \in \mathcal{C}_\mathcal{F} : C \in h^{-1}(1)} P(h^{-1}(1) \cap X_{G,l})}{P(X_{G,l})} = \sum_{C \in \mathcal{C}_\mathcal{F} : C \in h^{-1}(1)} \frac{P(C \cap X_{G,l})}{P(X_{G,l})}$$

$$= \sum_{C \in \mathcal{C}_\mathcal{F} : P(C \in h^{-1}(1)} \frac{P(C \cap t^{-1}(l))}{P(t^{-1}(l))} = \frac{P((h^{-1}(1) \cap t^{-1}(l))}{P(t^{-1}(l))}$$

Thus any hypothesis $h \in \mathcal{H}_{\mathcal{F}}$ is fair w.r.t. to the odds equality notion of fairness.
Assume $\mathcal{F}$ does not have zero-group-knowledge. Thus $F(x)$ and $G(x)$ are dependent given the ground truth $t(x)$. Thus there exists label $l \in \{0,1\}$, group $G \in \{A, D\}$ and a cell $C \in \mathcal{C}_{\mathcal{F}}$ with $\frac{P(C \cap X_{G,l})}{P(X_{G,l})} \neq \frac{P(C \cap t^{-1}(l))}{P(t^{-1}(l))}$. Now consider the hypothesis class h defined by $h^{-1}(1) = C$. For this hypothesis we have $\frac{P(h^{-1}(1) \cap X_{G,l})}{P(X_{G,l})} \neq \frac{P(h^{-1}(1) \cap t^{-1}(l))}{P(t^{-1}(l))}$. Thus, not every hypothesis $h \in \mathcal{H}_{\mathcal{F}}$ fulfills equalized odds.

A feature set $\mathcal{F}$ is adversarially 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 $P_{A,l}(C) = P_{D,l}(C)$. **Proof of Theorem 8:**

Assume $\mathcal{F}$ is adversarially fair w.r.t. to P and L^{EO}. This means that every $h \in \mathcal{H}_{\mathcal{F}}$ is fair w.r.t to L^{EO}. Now take any cell $C \in \mathcal{C}_{\mathcal{F}}$ and let h be defined by $h^{-1}(1) = C$. Then we know that $\frac{P(X_{A,1} \cap C)}{P(X_{A,1})} = \frac{P(X_{D,1} \cap C)}{P(X_{D,1})}$ and $\frac{P(X_{A,0} \cap C)}{P(X_{A,0})} = \frac{P(X_{D,0} \cap C)}{P(X_{D,0})}$. Thus, for each $l \in \{0,1\}$ we have $P_{A,l}(C) = P_{D,l}(C)$.

Now assume, we have for each $l \in \{0,1\}$ we have $P_{A,l}(C) = P_{D,l}(C)$. Then for any $h \in \mathcal{H}_{\mathcal{F}}$, we get

$$\frac{P(X_{A,l} \cap h^{-1}(1))}{P(X_{A,l})} = \sum_{C \in h^{-1}(1)} \frac{P(X_{A,l} \cap C)}{P(X_{A,l})} =$$

$$= \sum_{C \in h^{-1}(1)} \frac{P(X_{D,l} \cap C)}{P(X_{D,l})} = \frac{P(X_{D,l} \cap h^{-1}(1))}{P(X_{D,l})}.$$

Thus $L_P^{\text{EO}}(h) = 0$.

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 $P_{A,l}(C) = P_{D,l}(C)$. **Proof of Theorem 9:** "Conditional probabilities over score-cells align" implies "representation is accuracy-driven fair": We know for every cell $C \in \mathcal{C}_{\mathcal{F}s_t}$

$$\mathbb{P}_{x \sim P}[x \in C | x \in A, t(x) = 0] = \mathbb{P}_{x \sim P}[x \in C | x \in D, t(x) = 0]$$

and

$$\mathbb{P}_{x \sim P}[x \in C | x \in A, t(x) = 1] = \mathbb{P}_{x \sim P}[x \in C | x \in D, t(x) = 1].$$

Thus for every threshold $\alpha \in [0,1]$, we have $\mathbb{P}_{x \sim P}[t_{P,F}(x) | x \in A, t(x) = 1] = \mathbb{P}_{x \sim P}[t_{P,F}(x) | x \in D, t(x) = 1]$ and $\mathbb{P}_{x \sim P}[t_{P,F}(x) | x \in A, t(x) = 0] = \mathbb{P}_{x \sim P}[x \in C | x \in D, t(x) = 0]$. This implies equal false-positive and false-negative rates and therefore group fairness.

" conditional probabilities over score-cells do not align" implies "representation is not accuracy-driven fair": We assume that the conditional probabilities over score induced cells are not aligned. Let $\mathcal{C}_{\mathcal{F}score} = \{C_1, \dots C_{k'}\}$ such that $s_t(C_i) < s_t(C_j)$ for every $i < j$. Thus, $C_i \in \mathcal{C}_{\mathcal{F}s_t}$ with $\mathbb{P}_{x \sim P}[x \in C | \in A, t(x) = 0] \neq \mathbb{P}_{x \sim P}[x \in C_i | x \in D, t(x) = 0]$ or $\mathbb{P}_{x \sim P}[x \in C_i | \in A, t(x) = 1] \neq \mathbb{P}_{x \sim P}[x \in C_i | x \in D, t(x) = 1]$. Now consider the threshold classifier with threshold $s_t(C_i)$. We can consider two cases:

- Case 1: $\frac{P(t_{P,F}^{s(C_i)-1}(0) \cap X_{A,1})}{P(X_{A,1})} \neq \frac{P(t_{P,F}^{s(C_i)-1}(0) \cap X_{D,1})}{P(X_{D,1})}$ or $\frac{P(t_{P,F}^{s(C_i)-1}(1) \cap X_{A,0})}{P(X_{A,0})} \neq \frac{P(t_{P,F}^{s(C_i)-1}(1) \cap X_{D,0})}{P(X_{D,0})}$. This implies $L_{F,P}^{\text{fair}}(t_{P,F}^{s(C_i)}) > 0$. In this case, the Bayes classifier that cuts at $s_t(C_i)$ is unfair. Thus there exist a threshold classifier that is unfair.

- Case 2: $\frac{P(t_{P,F}^{s(C_i)-1}(0) \cap X_{A,1})}{P(X_{A,1})} = \frac{P(t_{P,F}^{s(C_i)-1}(0) \cap X_{D,1})}{P(X_{D,1})}$ and $\frac{P(t_{P,F}^{s(C_i)-1}(1) \cap X_{A,0})}{P(X_{A,0})} = \frac{P(t_{P,F}^{s(C_i)-1}(1) \cap X_{D,0})}{P(X_{D,0})}$.

  However since $\mathbb{P}_{x \sim P}[x \in C_i | \in A, t(x) = 0] \neq \mathbb{P}_{x \sim P}[x \in C_i | x \in D, t(x) = 0]$ or $\mathbb{P}_{x \sim P}[x \in C_i | \in A, t(x) = 1] \neq \mathbb{P}_{x \sim P}[x \in C_i | x \in D, t(x) = 1]$, this implies that $i > 1$. Now consider the threshold classifier with threshold $s_t(C_{i-1})$:

$$\begin{aligned}
&\mathbb{P}_{x\sim P}[t^{s(C_{i-1})}(x)=0|\in A, t(x)=1]\\
=&\mathbb{P}_{x\sim P}[t^{s(C_i)}(x)=0|\in A, t(x)=1]\\
&+\mathbb{P}_{x\sim P}[x\in C|\in A, t(x)=1]\\
\neq&\mathbb{P}_{x\sim P}[t^{s(C_i)}(x)=0|\in D, t(x)=1]\\
&+\mathbb{P}_{x\sim P}[x\in C|\in D, t(x)=1]\\
&=\mathbb{P}_{x\sim P}[t^{s(C_{i-1})}(x)=0|\in D, t(x)=1]
\end{aligned}$$

or

$$\begin{aligned}
&\mathbb{P}_{x\sim P}[t^{s(C_{i-1})}(x)=1|\in A, t(x)=0]\\
=&\mathbb{P}_{x\sim P}[t^{s(C_i)}(x)=1|\in A, t(x)=0]\\
&-\mathbb{P}_{x\sim P}[x\in C|\in A, t(x)=0]\\
\neq&\mathbb{P}_{x\sim P}[t^{s(C_i)}(x)=0|\in D, t(x)=0]\\
&-\mathbb{P}_{x\sim P}[x\in C|\in D, t(x)=1]\\
=&\mathbb{P}_{x\sim P}[t^{s(C_{i-1})}(x)=1|\in D, t(x)=0]
\end{aligned}$$

Which implies $L_P^{\text{EO}}(t^{s(C_{i-1})})>0$. Thus there exist a threshold classifier that is unfair.

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:

$$U_{\text{adv}}(\mathcal{F})\leq\frac{1}{2}d_{\mathcal{H}_\mathcal{F}}(P_{A,1},P_{D,1})+\frac{1}{2}d_{\mathcal{H}_\mathcal{F}}(P_{A,0},P_{D,0})$$

$$U_{\text{acc}}(\mathcal{F})\leq\frac{1}{2}d_{\mathcal{H}^{\text{thres}}_{\mathcal{C}_\mathcal{F},s_t}}(P_{A,1},P_{D,1})+\frac{1}{2}d_{\mathcal{H}^{\text{thres}}_{\mathcal{C}_\mathcal{F},s_t}}(P_{A,0},P_{D,0})$$

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

$$\frac{1}{2}d_{\mathcal{H}_F}(P_{A,l},P_{D,l})\leq U_{\text{adv}}(\mathcal{F})$$

for every $l\in\{0,1\}$ **Proof of Theorem 10:**

$$\begin{aligned}
&U_{\text{adv}}(\mathcal{F})=\max_{h\in\mathcal{H}_{\mathcal{C}_\mathcal{F}}}L_P^{\text{EO}}(h)\\
&=\max_{h\in\mathcal{H}_F}\sum_{l\in\{0,1\}}\frac{1}{2}|\frac{P(h^{-1}(1-l)\cap X_{A,l})}{P(X_{A,l})}-\frac{P(h^{-1}(1-l)\cap X_{D,l})}{P(X_{D,l})}|\\
&\leq\frac{1}{2}\sup_{1_B\in\mathcal{H}_F}|P_{A,1}(B)-P_{D,1}(B)|+\frac{1}{2}\sup_{1_B\in\mathcal{H}_F}|P_{A,0}(B)-P_{D,1}(B)|\\
&=\frac{1}{2}d_{\mathcal{H}_F}(P_{A,1},P_{D,1})+\frac{1}{2}d_{\mathcal{H}_F}(P_{A,0},P_{D,0})
\end{aligned}$$

$$\begin{aligned}
&U_{\text{acc}}(\mathcal{F})\leq\max_{h\in\mathcal{H}_F}L_P^{\text{EO}}(h)\\
&=\max_{h\in\mathcal{H}_F}\sum_{l\in\{0,1\}}\frac{1}{2}|\frac{P(h^{-1}(1-l)\cap X_{A,l}}{P(X_{A,l})}-\frac{P(h^{-1}(1-l)\cap X_{D,l})}{P(X_{D,l})}|\\
&\leq\frac{1}{2}\sup_{1_B\in\mathcal{H}^{\text{thres}}_{\mathcal{C}_\mathcal{F},s_t}}|P_{A,1}(B)-P_{D,1}(B)|\\
&+\frac{1}{2}\sup_{1_B\in\mathcal{H}^{\text{thres}}_{\mathcal{C}_\mathcal{F},s_t}}|P_{A,0}(B)-P_{D,1}(B)|\\
&=\frac{1}{2}d_{\mathcal{H}^{\text{thres}}_{\mathcal{C}_\mathcal{F},s_t}}(P_{A,1},P_{D,1})+\frac{1}{2}d_{\mathcal{H}^{\text{thres}}_{\mathcal{C}_\mathcal{F},s_t}}(P_{A,0},P_{D,0})
\end{aligned}$$

Furthermore, for any label $l' \in \{0,1\}$, we get

$$\begin{aligned}
U_{\text{adv}}(\mathcal{F}) &= \max_{h \in \mathcal{H}_F} L_P^{\text{EO}}(h) \\
&= \max_{h \in \mathcal{H}_F} \sum_{l \in \{0,1\}} \frac{1}{2} | \frac{P(h^{-1}(1-l) \cap X_{A,l}}{P(X_{A,l})} - \frac{P(h^{-1}(1-l) \cap X_{D,l})}{P(X_{D,l})} | \\
&\geq \frac{1}{2} \max_{h \in \mathcal{H}_F} | \frac{P(h^{-1}(1-l') \cap X_{A,l}}{P(X_{A,l'})} - \frac{P(h^{-1}(1-l') \cap X_{D,l})}{P(X_{D,l'})} | \\
&= \frac{1}{2} \sup_{1_B \in \mathcal{H}_F} |P_{A,l'}(B) - P_{D,l'}(B)| \\
&= \frac{1}{2} d_{\mathcal{H}_F}(P_{A,l}, P_{D,l})
\end{aligned}$$

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

Proof of Theorem 11: Assume $\mathcal{F}$ does not have calibration parity. Thus $t(x)$ and $G(x)$ are dependent given a feature vector. Thus there exists label $l \in \{0,1\}$, group $G \in \{A, D\}$ and a cell $C \in \mathcal{C}_{\mathcal{F}}$ with $\frac{P(C \cap X_{G,l})}{P(C \cap G)} \neq \frac{P(C \cap t^{-1}(l))}{P(C)}$. Now consider the hypothesis class h defined by $h^{-1}(1) = C$. For this hypothesis we have $\frac{P(h^{-1}(1) \cap X_{G,l})}{P(h^{-1}(l) \cap G)} \neq \frac{P(h^{-1}(1) \cap t^{-1}(l))}{P(h^{-1}(l))}$. Thus, not every hypothesis $h \in \mathcal{H}_{\mathcal{F}}$ fulfills predictive rate parity.

The reverse statement is not true. Let $\mathcal{C}_{\mathcal{F}} = \{C_1, C_2\}$ be such that $P(C_1 \cap X_{A,1}) = 0.5$, $P(C_1 \cap X_{A,0}) = 0.1$, $P(C_1 \cap X_{D,1}) = 0.2$, $P(C_1 \cap X_{A,1}) = 0.04$ and $P(C_2 \cap X_{A,1}) = P(C_2 \cap X_{A,0}) = P(C_1 \cap X_{D,1}) = P(C_1 \cap X_{A,1}) = 0.04$. The classifier h defined by $h(C) = 1$ for every $C \in \mathcal{C}_{\mathcal{F}}$ does not have predictive rate parity, since $\frac{P(h^{-1}(1) \cap X_{A,0})}{P(h^{-1} \cap A)} = \frac{P(X_{A,0})}{P(A)} = \frac{14}{68} \neq \frac{22}{100} = P(t^{-1}(1)) = \frac{P(h^{-1}(1) \cap X_{A,0})}{P(h^{-1} \cap A)} = \frac{P(X_{A,0})}{P(A)}$. Moreover, adversarial predictive parity is only possible in cases where success rates are equal, since unequal success rates always implies that the classifier h defined by $h(C) = 1$ for every $C \in \mathcal{C}_{\mathcal{F}}$ does not fulfill predictive rate parity.

A feature set $\mathcal{F}$ has demographic parity w.r.t. P if and only if it has adversarially fair w.r.t P and the group-fairness objective L^{DP}. **Proof of Theorem 12:**

- **Demographic Parity:** Assume $\mathcal{F}$ has demographic parity, then we have for every cell $C \in \mathcal{C}_{\mathcal{F}}$: $\frac{P(A \cap C)}{P(C)} = P(A)$. Thus, we have for any $h \in \mathcal{H}_{\mathcal{F}}$:
$$\frac{P(h^{-1}(1) \cap A)}{P(h^{-1}(1))} = \frac{\sum_{C \in \mathcal{C}_{\mathcal{F}}: C \in h^{-1}(1)} P(C \cap A)}{P(h^{-1}(1))}$$
$$= \frac{\sum_{C \in \mathcal{C}_{\mathcal{F}}: C \in h^{-1}(1)} P(C) P(A)}{P(h^{-1}(1))} = \frac{P(A) \sum_{C \in \mathcal{C}_{\mathcal{F}}: C \in h^{-1}(1)} P(C)}{P(h^{-1}(1))} = \frac{P(A) P(h^{-1}(1))}{P(h^{-1}(1))} = P(A)$$
Thus any $h \in \mathcal{H}_{\mathcal{F}}$ also has demographic parity.

- Assume $\mathcal{F}$ does not have demographic parity. Thus, there exists at least one cell $C \in \mathcal{C}_{\mathcal{F}}$ with $\frac{P(A \cap C)}{P(C)} \neq P(C)$. Now consider the hypothesis class h defined by $h^{-1}(1) = C$. For this hypothesis we have $\frac{P(h^{-1}(1) \cap A)}{P(h^{-1}(1))} \neq P(A)$. Thus, not every hypothesis $h \in \mathcal{H}_{\mathcal{F}}$ has demographic parity.

- There exists a distribution P and a feature set $\mathcal{F}$ such that each $f \in \mathcal{F}$ the feature set $\{f\}$ has zero-group-knowledge w.r.t. P, but $\mathcal{F}$ is not and $U_{\text{adv}}(\mathcal{F}) = 1$
- There exists a distribution P and a feature set $\mathcal{F}$ such that each $f \in \mathcal{F}$, the feature set $\{f\}$ has demographic parity w.r.t. P, but $\mathcal{F}$ has not. Furthermore the group-membership can be perfectly determined by $\mathcal{F}$, i.e. for every cell $C \in \mathcal{C}_{\mathcal{F}}$ we have
$$\mathbb{P}_{x \sim P}[x \in A | x \in C] \in \{0,1\}$$

- There exists a distribution P and a feature set $\mathcal{F}$ such that each $f \in \mathcal{F}$, the feature set $\{f\}$ has calibration parity w.r.t. P, but $\mathcal{F}$ has not. Furthermore the scores for the different groups in each cell are perfectly opposed, i.e. $C \subseteq A$ or $C \subseteq D$.

Proof of Observation 1:

- **(zero-group-knowledge)** Consider the domain $X = x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$ with $X_{A,1} = x_1, x_2$,$X_{D,1} = x_3, x_4$, $X_{A,0} = x_5, x_6$ and $X_{D,0} = x_7, x_8,$. Furthermore let $\mathcal{F} = \{f_1, f_2\}$ with $f_1^{-1}(1) = \{x_1, x_3, x_5, x_7\}$ and $f_2^{-1}(1) = \{x_1, x_4, x_5, x_8\}$. Furthermore let P be uniform over X,i.e. $P(\{x_1\}) = P(\{x_2\}) = P(\{x_3\}) = P(\{x_4\}) = P(\{x_5\}) = P(\{x_6\}) = P(\{x_7\}) = P(\{x_8\}) = 0.125$. Thus, we have zero-group-knowledge for both features, i.e.

$$\frac{P(X_{A,1} \cap f_1^{-1}(1))}{P(t^{-1}(1) \cap f_1^{-1}(1))} = \frac{P(\{x_1\})}{P(\{x_1, x_2\})} = 0.5$$

$$= \frac{P(\{x_3\})}{P(\{x_3, x_4\})} = \frac{P(X_{D,1} \cap f_1^{-1}(1))}{P(t^{-1}(1) \cap f_1^{-1}(1))}.$$

$$\frac{P(X_{A,0} \cap f_1^{-1}(1))}{P(t^{-1}(0) \cap f_1^{-1}(1))} = \frac{P(\{x_5\})}{P(\{x_5, x_6\})} = 0.5$$

$$= \frac{P(\{x_7\})}{P(\{x_7, x_8\})} = \frac{P(X_{D,0} \cap f_1^{-1}(1))}{P(t^{-1}(0) \cap f_1^{-1}(1))}.$$

$$\frac{P(X_{A,1} \cap f_2^{-1}(1))}{P(t^{-1}(1) \cap f_2^{-1}(1))} \frac{P(\{x_1\})}{P(\{x_1, x_2\})} = 0.5$$

$$= \frac{P(\{x_4\})}{P(\{x_3, x_4\})} = \frac{P(X_{D,1} \cap f_2^{-1}(1))}{P(t^{-1}(1) \cap f_2^{-1}(1))}.$$

$$\frac{P(X_{A,0} \cap f_2^{-1}(1))}{P(t^{-1}(0) \cap f_2^{-1}(1))} \frac{P(\{x_5\})}{P(\{x_5, x_6\})} = 0.5$$

$$= \frac{P(\{x_8\})}{P(\{x_7, x_8\})} = \frac{P(X_{D,0} \cap f_2^{-1}(1))}{P(t^{-1}(0) \cap f_2^{-1}(1))}.$$

  However, the featureset $\mathcal{F}$ does not have zero-group-knowledge (using Theorem 7) : $\mathcal{C}_{\mathcal{F}} = \{C_1, C_2, C_3, C_4\}$ with $C_1 = \{x_1, x_5\}$, $C_2 = \{x_2, x_6\}$, $C_3 = \{x_3, x_7\}$, and $C_4 = \{x_4, x_8\}$. Consider the classifier $h \in \mathcal{H}_{\mathcal{F}}$ with $h^{-1}(1) = \{C_1, C_3\}$. Then $L_P^{\text{EO}}(h) = \sum_{l \in \{0,1\}} \left| \frac{P(h^{-1}(|1-l|) \cap X_{A,l})}{P(X_{A,l})} - \frac{P(h^{-1}(|1-l|) \cap X_{D,l})}{P(X_{D,l})} \right| = |\frac{1}{2} - 0| + |0 - \frac{1}{2}| = 1$. Thus $U_{\text{adv}}^{EO}(\mathcal{F}) = 1$.

- **(demographic parity)** Consider the domain $X = x_1, x_2, x_3, x_4$ with $A = x_1, x_2$ and $D = x_3, x_4$. Furthermore let $\mathcal{F} = \{f_1, f_2\}$ with $f_1^{-1}(1) = \{x_1, x_3\}$ and $f_2^{-1}(1) = \{x_1, x_4\}$. Thus, $\mathcal{C}_{\mathcal{F}} = X$. Furthermore let P be uniform over X,i.e. $P(\{x_1\}) = P(\{x_2\}) = P(\{x_3\}) = P(\{x_4\}) = 0.25$. We have demographic parity for both features. However, since $\mathcal{C}_{\mathcal{F}} = X$, the featureset does not have demographic parity. Furthermore, the information from the cells suffice to perfectly predict the group-membership.

- **(calibration parity)** Consider the same domain X, the same feature set and the same probability distribution P as in the case of zero-group-knowledge. Furthermore consider the featureset $\mathcal{F} = \{f_1, f_2\}$ with $f_1^{-1}(1) = \{x_1, x_3, x_5, x_7\}$ and $f_2^{-1}(1) = \{x_1, x_4, x_6, x_7\}$. Both features of $\mathcal{F}$ have calibration parity, since both sides of each split have success-rate 0.5 for each group. Furthermore the $\mathcal{F}$ itself does not have calibration parity: We have $\mathcal{C}_{\mathcal{F}} = \{C_1, C_2, C_3, C_4\}$ with $C_1 = \{x_1, x_7\}, C_2 = \{x_2, x_8\}, C_3 = \{x_3, x_5\}$, and $C_4 = \{x_4, x_6\}$. Both cells C_1 and C_2 have one element from $X_{A,1}$ and one from $X_{D,0}$. Thus the success rate of elements of group A is 1 in these cells and the success rate of elements of group D is 0. Accordingly, both cells C_3 and C_4 have one element from $X_{D,1}$ and one from $X_{A,0}$. Thus the success rate of elements of group A is 0 in these cells and the success rate of elements of group D is 1 Thus when splitting these cells by group-membership both cells the resulting scores don't remain the same.