

# Gradual (In)Compatibility of Fairness Criteria

Corinna Hertweck,<sup>\*1, 2</sup> Tim Rätz<sup>\*3</sup>

<sup>1</sup>Institute for Data Analysis and Process Design, Zurich University of Applied Sciences, Winterthur, Switzerland

<sup>2</sup>Department of Informatics, University of Zurich, Zurich, Switzerland

<sup>3</sup>Institute of Philosophy, University of Bern, Bern, Switzerland

corinna.hertweck@zhaw.ch, tim.raetz@posteo.de

## Abstract

Impossibility results show that important fairness measures (independence, separation, sufficiency) cannot be satisfied at the same time under reasonable assumptions. This paper explores whether we can satisfy and/or improve these fairness measures simultaneously to a certain degree. We introduce information-theoretic formulations of the fairness measures and define degrees of fairness based on these formulations. The information-theoretic formulations suggest unexplored theoretical relations between the three fairness measures. In the experimental part, we use the information-theoretic expressions as regularizers to obtain fairness-regularized predictors for three standard datasets. Our experiments show that a) fairness regularization directly increases fairness measures, in line with existing work, and b) some fairness regularizations indirectly increase other fairness measures, as suggested by our theoretical findings. This establishes that it is possible to increase the degree to which some fairness measures are satisfied at the same time – some fairness measures are gradually compatible.

## 1 Introduction

Since the publication of Angwin et al. (2016), statistical fairness measures such as independence (also known as statistical parity and demographic parity), separation and sufficiency have been much discussed. There has been interest in the relation between these fairness measures (see, e.g., Kleinberg, Mullainathan, and Raghavan (2016); Chouldechova (2017); Mitchell et al. (2021); Heidari et al. (2019); Leben (2020); Baer, Gilbert, and Wells (2019)), and also in the relation between fairness measures and accuracy (see, e.g., Corbett-Davies et al. (2017); Menon and Williamson (2018); Cooper and Abrams (2021)).

It is well known that strict satisfaction of some combinations of these fairness measures is not possible; impossibility theorems establish that satisfaction of one fairness measure may prevent satisfaction of another under certain conditions (Kleinberg, Mullainathan, and Raghavan 2016; Chouldechova 2017). This is a fundamental limitation if one wishes to satisfy different fairness measures at the same time. However, it remains an open question to what extent

fairness criteria are (in)compatible if they are *partially fulfilled*. In this paper, we take first steps towards answering this question. This is relevant in practice because we cannot expect cases to align neatly with any one of the three fairness metrics. The application of predictive models typically involves multiple stakeholders. These stakeholders might reasonably disagree about the most appropriate fairness metric. It would therefore be helpful to know whether two fairness metrics can be improved at the same time in order to trade off the interests of multiple stakeholders.

In the present paper, we will show that it is possible to simultaneously improve the degree to which some of these fairness measures are satisfied. To show this, we first define a notion of partial satisfaction of fairness measures via information theory. The fairness measures have an exact information-theoretic counterpart, but the information-theoretic formulation also generalizes the fairness measures in providing degrees of fairness. An information-theoretic decomposition of separation and sufficiency reveals quantitative relations between the three measures that have not been explored so far: The expression of separation can be viewed as interpolating between sufficiency and independence. These theoretical relations suggest that enforcing one of the measures may quantitatively improve others.

We then put these theoretical considerations to work using so-called fairness regularization. The information-theoretic formulations of independence, separation and sufficiency serve as regularization terms during training to create fairness-regularized predictors for three standard datasets. We first examine whether fairness regularization directly increases the respective fairness measure after training, showing that this works reasonably well. We then explore whether enforcing one fairness measure indirectly increases the degree to which other fairness measures hold, in keeping with our theoretical hypotheses. We show that enforcing independence indirectly increases separation, and vice versa. We also discover that enforcing sufficiency will not increase separation, and vice versa. We offer a diagnosis of why this is the case on the basis of our theoretical findings.

It is also known that increasing fairness may depreciate accuracy (Corbett-Davies et al. 2017), and that sufficiency and separation may be better aligned with accuracy than independence – this is true at least for perfect accuracy (Hardt, Price, and Srebro 2016). For predictors that are not per-

<sup>\*</sup>These authors contributed equally.

fectly accurate, it has been shown that increases of accuracy and fairness need not go hand in hand (Räz 2021). It would be desirable to better understand how accuracy and fairness are related when neither one is perfectly satisfied. The information-theoretic decompositions of separation and sufficiency suggest that accuracy is a natural ingredient of these fairness measures.

The paper is structured as follows: In Section 2, we introduce the information-theoretic formulations of statistical fairness measures and decompose them in order to examine how the fairness measures are related. We also formulate empirical hypotheses about the effect of using these information-theoretic formulations in fairness regularization. Section 3 situates our contribution in the broader recent discussion of fair-ML while Section 4 discusses the relation of our contribution to similar works. In Section 5, we experimentally test our previously stated hypotheses. For this, we provide descriptions of the experimental setup and discuss our empirical results on direct and indirect fairness regularization. In Section 6, we summarize our findings and highlight open questions as well as avenues for future research.

## 2 Theory and Hypothesis Formulation

### 2.1 Preliminaries

We first provide some standard definitions of information theory; see Cover and Thomas (2006) for background. All random variables are discrete. Two random variables  $X, Y$  are (statistically) *independent*, written as  $X \perp Y$ , if  $P(X, Y) = P(X)P(Y)$ ; two random variables  $X, Y$  are *conditionally independent* given  $Z$ , written as  $X \perp Y \mid Z$ , if  $P(X \mid Y, Z) = P(X \mid Z)$ . The *entropy* of a random variable  $X \sim p(x)$  is given by

$$H(X) = - \sum_x p(x) \log p(x);$$

the *joint entropy* of a pair of random variables  $(X, Y) \sim p(x, y)$  is given by

$$H(X, Y) = - \sum_{x, y} p(x, y) \log p(x, y);$$

the *conditional entropy* of a pair of random variables  $(X, Y) \sim p(x, y)$  is given by

$$H(Y|X) = - \sum_{x, y} p(x, y) \log p(y|x).$$

The *mutual information* of a pair of random variables  $(X, Y) \sim p(x, y)$  is given by

$$I(X; Y) = H(X) - H(X|Y);$$

the *conditional mutual information* of three random variables  $(X, Y, Z) \sim p(x, y, z)$  is given by

$$I(X; Y \mid Z) = H(X \mid Z) - H(X|Y, Z).$$

### 2.2 Information-Theoretic Fairness

We can now define statistical notions of fairness (as provided by Barocas, Hardt, and Narayanan (2019)) in probabilistic and information-theoretic terms. We work with discrete random variables  $Y, R, A$ , where  $Y$  is the “ground truth”,  $R$  is the prediction, and  $A$  is the sensitive attribute.

**Definition 1.** The *independence gap* is given by

$$I(R; A). \quad (\text{IND})$$

Independence is satisfied if  $R \perp A$  (Barocas, Hardt, and Narayanan 2019), which is the case iff.  $I(R; A) = 0$ . We define the independence gap to be of degree  $d$  if  $I(R; A) \leq d$ .

**Definition 2.** The *separation gap* is given by

$$I(R; A \mid Y). \quad (\text{SEP})$$

Separation is satisfied if  $R \perp A \mid Y$  (Barocas, Hardt, and Narayanan 2019), which is the case iff.  $I(R; A \mid Y) = 0$ . The separation gap is of degree  $d$  if  $I(R; A|Y) \leq d$ .

**Definition 3.** The *sufficiency gap* is given by

$$I(Y; A \mid R). \quad (\text{SUF})$$

Sufficiency is satisfied if  $Y \perp A \mid R$  (Barocas, Hardt, and Narayanan 2019), which is the case iff.  $I(Y; A \mid R) = 0$ . The sufficiency gap is of degree  $d$  if  $I(Y; A|R) \leq d$ .

**Definition 4.** The *accuracy* of a predictor  $R$  is given by

$$I(Y; R). \quad (\text{ACC})$$

Note that while fairness measures are enforced by minimizing the corresponding expression, accuracy is enforced by maximizing the corresponding expression. In what follows, we will simply refer to the information-theoretic expressions of fairness metric gaps as independence, separation and sufficiency to increase readability.

The following definition is motivated in the next section.

**Definition 5.** *Balance* is given by

$$I(Y; R \mid A). \quad (\text{BAL})$$

### 2.3 Decomposition of Sufficiency and Separation

In this section, we use the information-theoretic formulations of sufficiency and separation given above to gain a clearer picture of how these notions are related to each other, as well as to independence and accuracy.

**Proposition 6.** Sufficiency has the following decomposition:

$$I(Y; A \mid R) = - \underbrace{I(Y; R)}_{\text{accuracy}} + \underbrace{I(Y; R \mid A)}_{\text{balance}} + \underbrace{I(A; Y)}_{\text{legacy}}. \quad (1)$$

The proof of the proposition can be found in Appendix A.1, which is available in the extended version of this paper.<sup>1</sup> The first term on the right of equation (1) corresponds to accuracy, while we refer to the second term as balance. If accuracy increases,  $R$  contains more information about  $Y$

<sup>1</sup><https://arxiv.org/abs/2109.04399>

and the first term becomes smaller due to the negative sign. If  $Y$  and  $R$  become decorrelated given  $A$ , the second term becomes smaller. The third term is dubbed *legacy* term because it only depends on the distribution of  $(Y, A)$ , i.e., of how we find the world to be at a certain point, measured by the “true label”  $Y$  and the characteristic  $A$ , but not on the prediction  $R$  (see Kamishima, Akaho, and Sakuma (2011); R  z (2021) for more on this). Only the first and the second term depend on the prediction  $R$ ; changing  $R$  will not affect the legacy term.<sup>2</sup>

**Proposition 7.** Separation has the following decomposition:

$$I(R; A | Y) = -\underbrace{I(Y; R)}_{\text{accuracy}} + \underbrace{I(Y; R | A)}_{\text{balance}} + \underbrace{I(A; R)}_{\text{independence}}. \quad (2)$$

The proof of the proposition is analogous to the proof of the proposition above, see also Appendix A.1. In the decomposition of separation, three terms depend on  $R$ . The first two terms on the right of equation (2) are the same as in sufficiency; the third term corresponds to independence. The fact that all three terms depend on  $R$  means that if you want to see how changing  $R$  affects separation, you need to take all three terms into account. This is a major difference between separation and sufficiency.

## 2.4 Empirical Hypotheses

In this section, we collect some observations about the relations between the three fairness measures that can be gleaned from the decompositions in the previous section. This will allow us to formulate empirical hypotheses.

A first observation concerns the fact that both sufficiency and separation contain the accuracy term with a negative sign. This is related to the fact that sufficiency and separation are both what R  z (2021) refers to as *conservative* measures, i.e., they are both minimal in case of perfect prediction ( $Y = R$ ). However, as has been shown in R  z (2021), we do not necessarily improve sufficiency and separation by increasing accuracy. The decompositions above show why this is the case: If the accuracy of the predictor  $R$  is increased, this will only lower sufficiency and separation respectively if the other summands containing  $R$  do not offset the change.<sup>3</sup> Note that while accuracy, as it is defined here, is not commonly used as a loss function in machine learning, it can be expected to be closely correlated with common loss functions such as cross-entropy. This suggests that it is possible to indirectly increase the degree to which both sufficiency and separation are satisfied by making a predictor more accurate, *given* one controls for the other terms containing  $R$  in

<sup>2</sup>The first two terms on the right are also known as *interaction information*; see, e.g., Williams and Beer (2010). Interaction information measures the extent to which the interaction between two variables (here  $Y, R$ ) is mediated by a third one (here  $A$ ). Interaction information can be negative, zero, and positive. If we can only control  $R$ , but not  $Y$  and  $A$ , then changes in interaction information determine changes in sufficiency.

<sup>3</sup>A predictor  $R'$ , which is at least as accurate as predictor  $R$ , maintains the degree of sufficiency of  $R$  if  $I(Y; R' | A) \leq I(Y; R | A)$ , and the predictor  $R'$  maintains the degree of separation of  $R$  if  $I(Y; R' | A) + I(A; R') \leq I(Y; R | A) + I(A; R)$ .

the above decompositions. This would be in line with work by Liu, Simchowitz, and Hardt (2019) who found that group calibration, a form of sufficiency, increases with increased accuracy. As models are typically optimized for some variation of accuracy, they note that sufficiency is often already achieved by unconstrained learning, i.e., when no fairness constraints are imposed.

A second observation concerns the relation between all three measures. Equation 2 shows that independence is one of the summands of the decomposition of separation. On the other hand, independence does not contribute to sufficiency. In this sense, we can think of separation as being an intermediate between independence and sufficiency, or being an interpolation between independence and sufficiency. A third observation concerns the relation between separation and sufficiency. Apart from being related to accuracy, we can observe that they both also contain the balance term. This suggests that one way of improving both measures would be to control or decrease the balance term.

These observations lead to the following idea. The various relations between the three measures just pointed out mean that the three measures co-vary to a certain extent, in that if, say, independence is higher, then separation may also be higher. These possible co-variations may be exploited when fairness measures are enforced: If we increase the degree to which a predictor satisfies one of the measures, this may indirectly increase or decrease one of the other measures.

Below we will test this general idea in the form of direct and indirect fairness regularization. Fairness regularization means that the information-theoretic formulations of the fairness measures can be used as a regularization constraint in training an ML model – Kamishima, Akaho, and Sakuma (2011) call this *prejudice remover regularization*. Indirect fairness regularization means that the relations between the fairness measures may lead to an increase of a fairness measure even if it is not directly regularized, due to the relations between the measures identified above.

First, we will test for *direct regularization*:

- *Independence*  $\rightarrow$  *Independence*, *Separation*  $\rightarrow$  *Separation*, *Sufficiency*  $\rightarrow$  *Sufficiency*: If we use the three information-theoretic expressions of independence, separation and sufficiency as regularization terms during training, this should increase the degree to which they are satisfied.

Second, we will test for *indirect regularization*:

- *Independence*  $\rightarrow$  *Separation*: If we regularize independence, then this could indirectly decrease separation, due to the fact that the independence term features in the decomposition of separation in equation 2.
- *Separation*  $\rightarrow$  *Independence*: Regularizing separation could indirectly regularize independence because independence is in the decomposition of separation.
- *Separation*  $\rightarrow$  *Sufficiency*: Balance and accuracy are in the decomposition of both sufficiency and separation, so regularizing separation could indirectly regularize sufficiency.

- *Sufficiency*  $\rightarrow$  *Separation*: Balance and accuracy are in the decomposition of both sufficiency and separation, so similarly, regularizing sufficiency could indirectly regularize separation.
- *Balance*  $\rightarrow$  *Separation*, *Sufficiency*: Balance is in the decomposition of both sufficiency and separation, so regularizing balance could indirectly regularize these two measures.
- *Accuracy*  $\rightarrow$  *Separation*, *Sufficiency*: Accuracy is in the decomposition of both sufficiency and separation, so regularizing accuracy could indirectly regularize these two measures.

## 2.5 Normalizations

It is desirable to compare normalized versions of the fairness measures to get a better sense of how the fairness measures perform during evaluation. We therefore normalize independence, separation and sufficiency in the evaluation step (but not during training). The normalizations we use here were first proposed and justified in (Steinberg et al. 2020). The normalization for *independence* is

$$\hat{I}(R; A) = \frac{I(R; A)}{H(A)}. \quad (\text{N-IND})$$

The justification given by Steinberg et al. is that in case  $\hat{I}(R; A) = 1$ , which is the maximum,  $R$  encodes all information about  $A$ , which is maximally unfair. The justifications for the normalizations of separation and sufficiency are similar. The normalization for *separation* is

$$\hat{I}(R; A|Y) = \frac{I(R; A|Y)}{H(A|Y)}. \quad (\text{N-SEP})$$

The normalization for *sufficiency* is

$$\hat{I}(Y; A|R) = \frac{I(Y; A|R)}{H(A|R)}. \quad (\text{N-SUF})$$

The normalization factors are (conditional) entropies. The entropies for independence and separation do not depend on  $R$  and have to be computed only once per dataset. The entropy for sufficiency involves  $R$  and has to be computed for each (trained) model, i.e., for every evaluation.

The normalized values for independence, separation and sufficiency thus vary between 0 and 1, where 0 means that the fairness metric is perfectly fulfilled, and 1 indicates maximal unfairness. It is an open question what (philosophical) meaning can be ascribed to values between these two extremes. What is known is that lower values indicate a better fulfillment of the respective fairness metric.

## 3 Background

In 2016, ProPublica published an article entitled "Machine Bias," in which they investigated the tool COMPAS (short for "Correctional Offender Management Profiling for Alternative Sanctions") (Angwin et al. 2016). ProPublica accused the company that developed the software, Equivant (formerly known as Northpointe), of racial bias in their tool. According to their calculations, black defendants who are

not arrested within two years are deemed more likely to be arrested than white defendants who are arrested within two years. By comparing risk predictions ( $R$ ) conditioned on re-arrest ( $Y$ ) across groups, they argued that a version of separation is violated. Equivant denied the accusation of racial bias by arguing that their risk scores are equally accurate for black and white defendants (Dieterich, Mendoza, and Brennan 2016). They argued that a version of sufficiency was fulfilled. For further contributions to this discussion see Flores, Bechtel, and Lowenkamp (2016); Berk et al. (2018). Subsequently, Kleinberg, Mullainathan, and Raghavan (2016) and Chouldechova (2017) demonstrated that separation has to be violated when sufficiency is fulfilled and vice versa – except under circumstances that cannot be expected to arise in practice. This incompatibility debate has been extended in Barocas, Hardt, and Narayanan (2019): any two of the three fairness criteria, i.e., independence, separation and sufficiency, are mutually incompatible under reasonable assumptions and at least for binary predictors.

Our paper investigates whether mutually exclusive fairness metrics can still be jointly improved to a certain extent. In doing so, we use an approach that Green (2021) calls "formal": We seek to improve predictive models and do not consider possible changes in the social context of their application. Green criticizes that while the "formalism response" can sometimes increase social justice, its narrow focus on mathematical formalizations of egalitarian principles can also be harmful because the social context of the application is not considered. As pointed out by Green and Hu (2018) and Bao et al. (2021), incremental improvements in the fulfillment of mathematical definitions can distract from real reforms that target unjust systems (over predictive models that are just a small part of these systems).

We agree that the social context in which models are deployed must not be ignored when judging the fairness of models. In this paper, we merely investigate the relations between three mathematical definitions of fairness. Our results are proofs-of-concept for the partial fulfillment of multiple fairness metrics. However, the possibility of improving fairness metrics does *not* establish that fairness has been achieved, even incrementally. If direct reforms of unjust systems are not a viable option, improving the fulfillment of fairness metrics may be necessary. In such cases, understanding the relationship between the fairness metrics and possibilities for tradeoffs between them is important for improving fairness.

In our experiments, we also investigate the effects of fairness regularization on the accuracy of the predictive model. The tradeoff between fairness and accuracy has been analyzed in several works (Hardt, Price, and Srebro 2016; Corbett-Davies et al. 2017; Menon and Williamson 2018; Chen, Johansson, and Sontag 2018). However, Cooper and Abrams (2021) point out that this framing can potentially cause more unfairness because we usually work with specific operationalizations of both fairness and accuracy, which may be problematic. The goal of recidivism predictions, for example is to predict if an incarcerated person would *commit* a crime after being released. This is typically operationalized as the question of who is *arrested*

within a given time period after release. However, it is at least questionable if this operationalization of *crime* as *arrest* works as intended due to unjustified arrests and undetected crimes; see also Lum and Isaac (2016); Mehrotra et al. (2021). Moreover, accuracy does not take long-term effects of, say, imposing fairness constraints into account. So even though avoiding a loss in accuracy appears desirable from a mathematical perspective, we have to keep in mind that accuracy, as it is typically measured in ML, does not necessarily represent real-world accuracy.

## 4 Related Work

Kamishima, Akaho, and Sakuma (2011) proposed mitigating bias in predictive models by using a version of independence for regularization. For this, they formalized independence in information-theoretic terms and simplified the expression further to make it computationally tractable. They used the resulting simplified expression as a regularizer during training, and they found that this improved their definition of fairness, which is closely related to independence, but compromised on accuracy. Steinberg et al. (2020) proposed approximations of information-theoretic terms of independence, separation and sufficiency for mixed variables that they find to be faster to compute than the exact terms, but lead to similar results when used as regularizers. They considered decompositions of fairness measures similar to our equations (1) and (2) above, but they did not further investigate the relations between fairness measures. Ghasami, Khodadadian, and Kiyavash (2018) propose a method for maximizing accuracy while fulfilling separation using an information-theoretic version of it. The described approaches focus on fairness / accuracy tradeoffs; our focus, on the other hand, is wider because we consider accuracy as only one component of fairness measures as found in the fairness decompositions.

Fairness regularization of some fairness measures using non-information-theoretic regularization terms have also been investigated, e.g., in Cho, Hwang, and Suh (2020), Narasimhan (2018), Padala and Gujar (2020) and (Chuang and Mroueh 2021). These works focus on improving fairness measures through regularization, but they do not investigate how the regularization of one measure affects other measures through indirect regularization, as far as we can see.

Partial fulfillment of fairness metrics has previously been investigated by Pleiss et al. (2017) who showed that sufficiency is compatible with a relaxation of separation, i.e., parity in the false negative rates. Loi, Herlitz, and Heidari (2019) discuss conditions for the enforcement of separation and sufficiency. They point out that "when neither set of conditions obtains, fair equality of chances is satisfied by neither sufficiency nor separation, but, possibly, some other statistical constraint" (Loi, Herlitz, and Heidari 2019, p. 25). This other statistical constraint might be found in a tradeoff between the fairness metrics. Note, however, that even if it can be shown that some conditions can be partially fulfilled or improved simultaneously, it is still necessary to address the philosophical questions as to what partial fulfillment of a fairness metrics means, and whether this is desirable.

## 5 Experiments

The code for the experimental part can be found in the public repository <https://github.com/hcorinna/gradual-compatibility>.

### 5.1 Description of Datasets

Experiments were performed on three standard benchmark datasets, which are briefly described in Table 1: "Adult income", "German credit", and "ProPublica recidivism" (the latter is also known as the "COMPAS" dataset). Friedler et al.'s (2019) preprocessed versions of these datasets were used. Specifically, we used the preprocessed versions in which categorical values were one-hot encoded and in which sensitive attributes were transformed to binary variables.<sup>4</sup> In addition to Friedler et al.'s (2019) preprocessing, we preprocess the ProPublica recidivism dataset ourselves by removing the column "c\_charge\_desc" (more specifically, the columns resulting from the one-hot encoding of this column). The column describes the charges raised against the defendant, but as it can take on many different values, it creates a lot of additional columns when one-hot encoded. We opted for the removal of these columns because this reduces the number of columns from 403 to 14 and thus allows for far faster training.

Dataset	Y=1	A	Size
German	Creditworthiness	Sex	1 000
ProPublica	Rearrest within 2 years	Race	6 167
Adult	Income >50k	Race	30 162

Table 1: Datasets used in experiments.

### 5.2 Description of Experiments

In contexts where fairness is particularly salient, it is essential that the decision-making process can be fully understood. Therefore, we chose logistic regression as an easily interpretable model for our proof-of-concept. We use a logistic regression model that is optimized by `scipy`'s (Virtanen et al. 2020) `minimize` function with L-BFGS-B (Byrd et al. 1995; Zhu et al. 1997) as the solver. The loss function takes the following form:

$$L = l_{fit} + \lambda \cdot l_2 + \mu \cdot l_{fair}. \quad (3)$$

The algorithm optimizes the cross-entropy loss (dubbed  $l_{fit}$ ) with  $l_2$  regularization and fairness regularization terms (dubbed  $l_{fair}$ ). The parameter  $\lambda$  is chosen using 5-fold cross validation, for each dataset and before fairness regularization. The value of  $\lambda$  was left unchanged during fairness regularization. We used the following values for  $\lambda$ : 0.00001 ("Adult income"), 0.001 ("German credit"), and

<sup>4</sup>These are the datasets that have the ending "numerical-binsensitive" in Friedler et al.'s repository: <https://github.com/algofairness/fairness-comparison/tree/master/fairness/data/preprocessed>.

0.001 (“ProPublica recidivism”).<sup>5</sup> The fairness regularization terms  $l_{fair}$  take values in the range {IND, SEP, SUF, BAL, -ACC}. We regularize (and thus try to minimize) *negative accuracy*, which is the same as maximizing accuracy because it is desirable to obtain an accurate predictor, and because negative accuracy features in the fairness decompositions. For each fairness regularization term, we choose values for the fairness parameter  $\mu$  in the range  $[0, 100]$ .<sup>6</sup> The model is trained and evaluated separately for each  $\mu$  and  $l_{fair}$  using 5-fold cross validation. In the evaluation step, we evaluate the three normalized fairness measures independence, sufficiency, and separation, as well as the negative accuracy, and the non-normalized balance term.<sup>7</sup> Normalization was only used during evaluation. More details on the experimental setup can be found in Appendix A.2.

### 5.3 Direct Regularization

We first evaluate whether the regularization of the three fairness metrics improves the respective (normalized) fairness metric itself, see Figure 1. In Figure 1a, we use independence as the fairness regularizer, that is, we let  $l_{fair} = \text{IND}$ , train models for different values of  $\mu$ , and measure N-IND. As expected, independence improves: it decreases in the plot for increasing  $\mu$ , as more weight is put on the independence regularizer. This is in line with the findings by Kamishima, Akaho, and Sakuma (2011). Figure 1b shows that regularizing with separation (SEP) improves separation (N-SEP) globally for two datasets, and in a low parameter range for “Adult income”. The regularization of sufficiency (SUF) seems to improve sufficiency (N-SUF) marginally and, for “Adult income”, only in a low parameter range, see Figure 1c. Thus, direct regularization seems to work, but not equally well for all three metrics.

As Figure 6 in the appendix shows, the independence and separation regularization have comparable negative effects on accuracy, while sufficiency regularization does not have a pronounced effect on accuracy. We made similar observations for cross-entropy.

We can interpret these results based on the decompositions of sufficiency and separation in equations (1) and (2). Regularizing sufficiency does not yield the same improvements as regularizing independence and separation because, first, regularization using the IND term works well and affects independence as well as separation, which contains independence in its decomposition, while independence is not in the decomposition of sufficiency. Second, sufficiency is more closely associated with increasing accuracy, as noted

<sup>5</sup>Note that  $\lambda$  could also be determined by cross-validation for each fold of the fairness regularization separately in order to keep parameter tuning and testing strictly separate. We decided to use one value of  $\lambda$  per dataset for all fairness regularization experiments because our main goal is not to optimize for accuracy, but to understand relations between fairness measures.

<sup>6</sup>We experimented with  $\mu \in [0, 190]$ , but did not observe notable changes for  $\mu > 100$ , so we restrict the plots to the range  $[0, 100]$ .

<sup>7</sup>We also evaluated cross-entropy, but do not show the results for cross-entropy here because accuracy and cross-entropy are very similar in our experiments.

by Liu, Simchowitz, and Hardt (2019). However, accuracy is already enforced by the cross-entropy optimization ( $l_{fit}$ ). It is therefore plausible that sufficiency is already naturally close to optimal in unconstrained learning and does not improve as much through fairness regularization. We also note that the “German credit” dataset has the highest legacy, which might explain why the  $\text{SUF} \rightarrow \text{N-SUF}$  effect is most pronounced for this dataset: the higher legacy implies a higher initial N-SUF that can then be lowered.

### 5.4 Indirect Regularization

Now we examine whether the regularization of one fairness metric has effects on other fairness metrics.

**Independence and Separation** We expect that the regularization of independence and separation have an indirect effect on each other because they share the IND term. Figure 2a shows the effects of the independence regularizer (IND) on separation (N-SEP). As we can see, separation indeed improves with increasing independence regularization. Figure 2b shows that the regularization of separation (SEP) has a positive effects on independence (N-IND).

While the impossibility results by Barocas, Hardt, and Narayanan (2019) proved that it is impossible to perfectly fulfill independence and separation when “ $Y$  is binary,  $A$  is not independent of  $Y$ , and  $R$  is not independent of  $Y$ ,” our results show that the two metrics can both be improved simultaneously – provided they are not perfectly satisfied.

**Separation and Sufficiency** We are hoping for some indirect regularization effects between separation and sufficiency because they share the BAL and -ACC terms. However, we get mixed or negative results, depending on the dataset. In Figure 2c, we show the effect of the separation regularizer (SEP) on sufficiency (N-SUF), and in Figure 2d, we plot the effect of sufficiency (SUF) regularization on separation (N-SEP). The only positive outcome is that regularizing sufficiency seems to improve separation for the “German credit” dataset.

Indirect regularization does not achieve the desired effect of improving *both* separation and sufficiency at the same time. From this, we conclude that it is difficult to circumvent the impossibility results concerning separation and sufficiency.

**Accuracy and Balance** We examined the effects of using negative accuracy and balance as regularizers because they are part of the decomposition of both sufficiency and separation. We found that using -ACC and BAL did not yield stable improvements of either separation (N-SEP) or sufficiency (N-SUF) when used as fairness regularizers; results are in Figure 3.

We investigated why this indirect regularization does not work. For this, we examined the effect of -ACC regularization on the components of separation and sufficiency (N-IND, BAL, -ACC) and found that the additional accuracy regularization does not substantially improve any of the components (see Figure 4 in the appendix). As noted above, this may be due to the fact that optimizing the model for

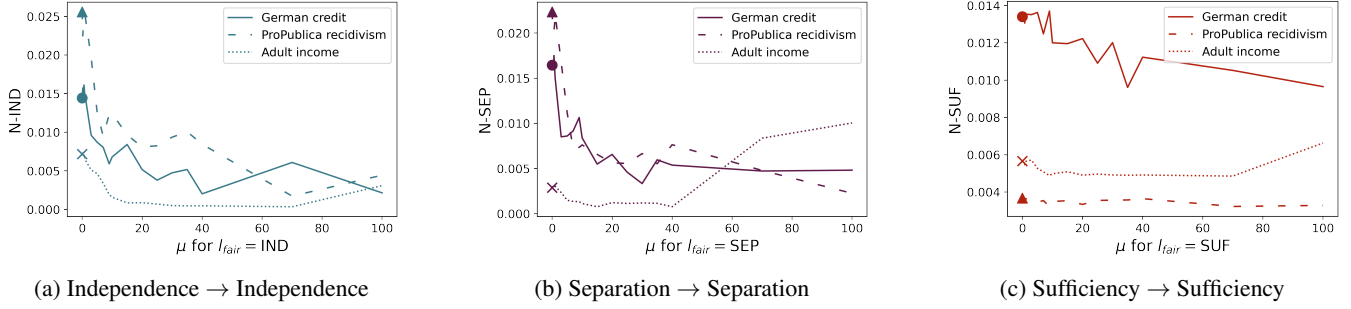


Figure 1: Direct regularization. x-axis: Value of  $\mu$  in  $l_{fair}$ . y-axis: Normalized values of regularized metric where lower values are better. The different curves represent the different datasets.

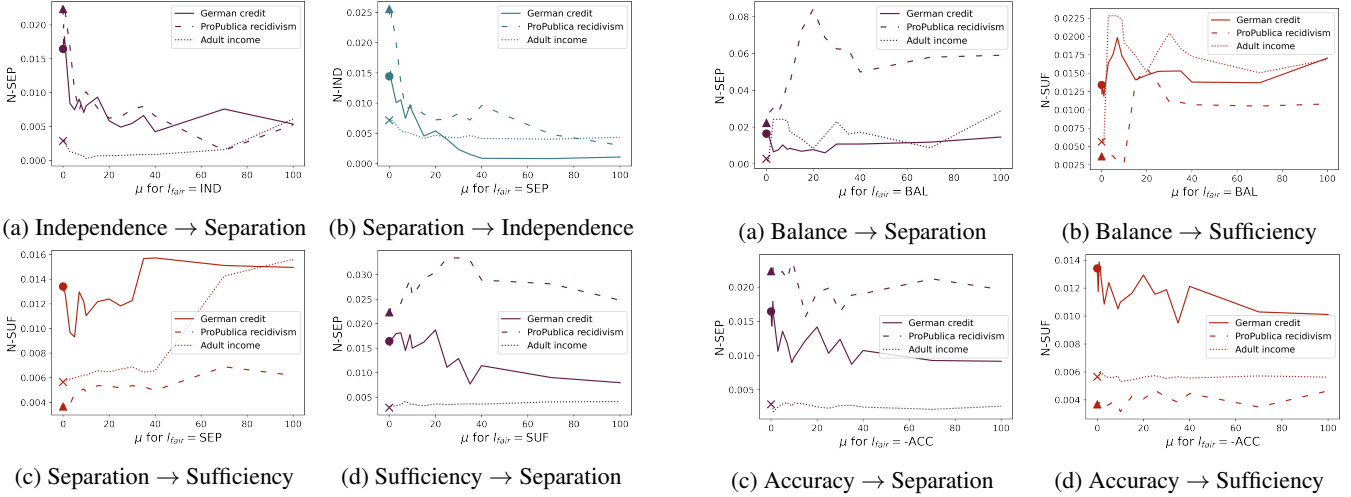


Figure 2: Indirect regularization between independence and separation and between separation and sufficiency.

cross-entropy already indirectly enforces accuracy. Examining the effect of BAL regularization on the components of separation and sufficiency (N-IND, BAL, -ACC), we found that while balance itself improves, regularizing with balance also substantially decreases accuracy (see Figure 5 in the appendix). This may be the reason why the indirect regularization between sufficiency and separation does not work: enforcing balance may decrease accuracy too much.

## 6 Conclusion

In this paper, we have used information-theoretic formulations of fairness metrics to perform fairness regularization. We have found, first, that using information-theoretic formulations of independence, sufficiency and separation in regularization works reasonably well. Second, there are indirect regularization relations between independence and separation: Enforcing independence improves separation and vice versa. This shows that it is possible to increase the degree to which fairness measures are satisfied, even though it is mathematically impossible to satisfy these measures perfectly at the same time. This finding is valuable for cases where stakeholders find both independence and separation

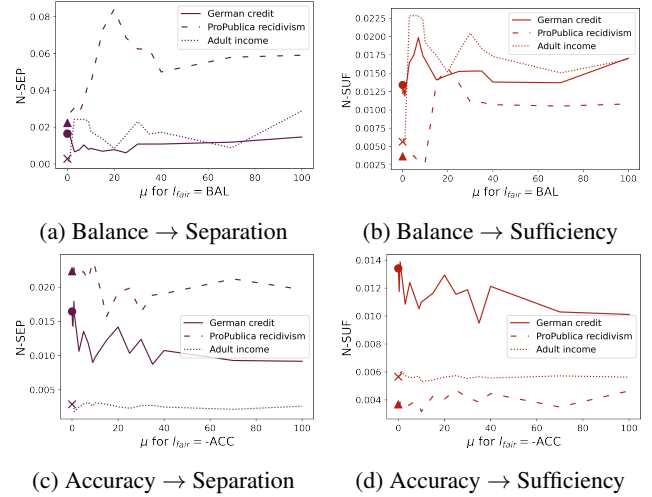


Figure 3: Indirect regularization of separation and sufficiency through negative accuracy and balance.

to be appropriate fairness metrics. Third, we have not been able to verify indirect regularization between separation and sufficiency. We have traced this back to the balance term, which seems to counteract accuracy.

Future work includes the following issues: Formulations of degrees of fairness do not need to be based on information theory; it would be valuable to compare our results to those of other formulations of “degrees of fairness”, also to better understand the advantages and drawbacks of the information-theoretic formulation. In particular, it is possible to combine non-information-theoretic regularization with information-theoretic evaluation. Then, many conceptual problems are open: How exactly should we interpret the tradeoffs we found in the present work? What meaning can we assign to degrees of fairness – outside the extreme values? Finally, we have not fully explored the parameter space, in that we have only explored linear scalings of terms in the decompositions of separation and sufficiency. For example, it is possible that separation and sufficiency can be improved at the same time by a judicious combination of weights for accuracy, balance, and independence.

## Acknowledgments

We thank the members of the NRP-77 project "Socially acceptable AI" for their useful feedback. We also appreciate the helpful feedback from the participants of the seminar "Scientific Writing II" at the University of Zurich. TR was supported by the Swiss National Science Foundation, grant number 197504. CH was supported by the National Research Programme "Digital Transformation" (NRP 77) of the Swiss National Science Foundation (SNSF), grant number 187473.

## References

- Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica.
- Baer, B. R.; Gilbert, D. E.; and Wells, M. T. 2019. Fairness criteria through the lens of directed acyclic graphical models. arXiv:1906.11333.
- Bao, M.; Zhou, A.; Zottola, S.; Brubach, B.; Desmarais, S.; Horowitz, A.; Lum, K.; and Venkatasubramanian, S. 2021. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. arXiv:2106.05498.
- Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org.
- Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; and Roth, A. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*.
- Byrd, R. H.; Lu, P.; Nocedal, J.; and Zhu, C. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5): 1190–1208.
- Chen, I. Y.; Johansson, F. D.; and Sontag, D. 2018. Why is my classifier discriminatory? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 3543–3554.
- Cho, J.; Hwang, G.; and Suh, C. 2020. A fair classifier using kernel density estimation. *Advances in Neural Information Processing Systems*, 33: 15088–15099.
- Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163.
- Chuang, C.-Y.; and Mroueh, Y. 2021. Fair Mixup: Fairness via Interpolation. arXiv:2103.06503.
- Cooper, A. F.; and Abrams, E. 2021. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 46–54. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.
- Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806.
- Cover, T. M.; and Thomas, J. A. 2006. *Elements of Information Theory*. Wiley, second edition.
- Dieterich, W.; Mendoza, C.; and Brennan, T. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Technical report, Northpointe Inc.
- Flores, A. W.; Bechtel, K.; and Lowenkamp, C. T. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Federal Probation*, 80(2): 38–46.
- Friedler, S. A.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E. P.; and Roth, D. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, 329–338.
- Ghassami, A.; Khodadadian, S.; and Kiyavash, N. 2018. Fairness in supervised learning: An information theoretic approach. In *2018 IEEE International Symposium on Information Theory (ISIT)*, 176–180. IEEE.
- Green, B. 2021. Escaping the "Impossibility of Fairness": From Formal to Substantive Algorithmic Fairness. arXiv:2107.04642.
- Green, B.; and Hu, L. 2018. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*.
- Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29: 3315–3323.
- Heidari, H.; Loi, M.; Gummadi, K. P.; and Krause, A. 2019. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the conference on fairness, accountability, and transparency*, 181–190.
- Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware Learning through Regularization Approach. 2011 IEEE 11th International Conference on Data Mining Workshops.
- Kleinberg, J. M.; Mullainathan, S.; and Raghavan, M. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:1609.05807.
- Leben, D. 2020. Normative principles for evaluating fairness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 86–92.
- Liu, L. T.; Simchowitz, M.; and Hardt, M. 2019. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, 4051–4060. PMLR.
- Loi, M.; Herlitz, A.; and Heidari, H. 2019. A Philosophical Theory of Fairness for Prediction-Based Decisions. Available at SSRN 3450300.
- Lum, K.; and Isaac, W. 2016. To predict and serve? *Significance*, 13(5): 14–19.
- Mehrotra, D.; Mattu, S.; Gilbertson, A.; and Sankin, A. 2021. How We Determined Crime Prediction Software Disproportionately Targeted Low-Income, Black, and Latino Neighborhoods. The Markup.



- Menon, A. K.; and Williamson, R. C. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, 107–118. PMLR.
- Mitchell, S.; Potash, E.; Barocas, S.; D’Amour, A.; and Lum, K. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8: 141–163.
- Narasimhan, H. 2018. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, 1646–1654. PMLR.
- Padala, M.; and Gujar, S. 2020. Fnnc: Achieving fairness through neural networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, {IJCAI-20}, International Joint Conferences on Artificial Intelligence Organization*.
- Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On Fairness and Calibration. arXiv:1709.02012.
- Räz, T. 2021. Group Fairness: Independence Revisited. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 129–137.
- Steinberg, D.; Reid, A.; O’Callaghan, S.; Lattimore, F.; McCalman, L.; and Caetano, T. 2020. Fast Fair Regression via Efficient Approximations of Mutual Information. arXiv:2002.06200.
- Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272.
- Williams, P. L.; and Beer, R. D. 2010. Nonnegative Decomposition of Multivariate Information. arXiv:1004.2515.
- Zhu, C.; Byrd, R. H.; Lu, P.; and Nocedal, J. 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, 23(4): 550–560.