Mapping the Tradeoffs and Limitations of Algorithmic Fairness

Etam Benger
The Hebrew University of Jerusalem, Israel

Katrina Ligett
The Hebrew University of Jerusalem, Israel

Abstract

Sufficiency and separation are two fundamental criteria in classification fairness. For binary classifiers, these concepts correspond to subgroup calibration and equalized odds, respectively, and are known to be incompatible except in trivial cases. In this work, we explore a relaxation of these criteria based on f-divergences between distributions – essentially the same relaxation studied in the literature on approximate multicalibration – analyze their relationships, and derive implications for fair representations and downstream uses (post-processing) of representations. We show that when a protected attribute is determinable from features present in the data, the (relaxed) criteria of sufficiency and separation exhibit a tradeoff, forming a convex Pareto frontier. Moreover, we prove that when a protected attribute is not fully encoded in the data, achieving full sufficiency may be impossible. This finding not only strengthens the case against "fairness through unawareness" but also highlights an important caveat for work on (multi-)calibration.

2012 ACM Subject Classification Theory of computation → Machine learning theory; Mathematics of computing → Information theory; Social and professional topics → Computing / technology policy

Keywords and phrases Algorithmic fairness, information theory, sufficiency-separation tradeoff

Digital Object Identifier 10.4230/LIPIcs.FORC.2025.19

Funding This work was supported in part by a gift to the McCourt School of Public Policy and Georgetown University, Simons Foundation Collaboration 733792, Israel Science Foundation (ISF) grant 2861/20, ERC grant 101125913, and a grant from the Israeli Council for Higher Education. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Acknowledgements We thank Flavio Calmon for an insightful suggestion that inspired the formulation using f-information.

1 Introduction

Machine learning algorithms increasingly influence many aspects of our lives – ranging from seemingly minor decisions, such as which ads appear on a website, to critical financial determinations like loan approvals, and even life-altering outcomes in healthcare and the judicial system. Determining whether and how fairness should be taken into consideration in a specific context; examining whether a particular fairness criterion is suitable, aligns with legal standards, or reflects ethical norms; deciding what trade-offs to strike between the many desiderata one might hold – these quandaries cannot be addressed by math alone. What theoretical computer science can do and has been doing for fairness is to reveal what is possible and what is impossible – for example, what notions can be enforced computationally efficiently and when? What tradeoffs between desiderata are fundamental?

A substantial literature has focused on statistical criteria of group fairness, formalizing various notions of equality across groups or attributes. In this vein, a landmark early result of Chouldechova [3] and of Kleinberg, Mullainathan, and Raghavan [11], revealed that the multiple natural statistical fairness criteria in classification are often inherently incompatible, except in trivial cases.

In this setting, a classification algorithm takes data of an individual as input (for example, her financial history) and outputs some predicted label (for example, whether she is expected to repay a loan). The utility of a classifier or predictor is measured with respect to the true label of the individual (would she actually eventually repay the loan?). Group fairness notions consider some protected attribute (for example, the individual's ethnicity); each set of people with a particular value of the protected attribute is sometimes referred to as a group or subgroup. The main statistical criteria of fairness in classification are concerned with the statistical relationships that the algorithm induces between the true label, the algorithm's output, and the protected attribute.

Independence, also known as statistical parity or demographic parity, requires that the predictor's outcome be independent of the protected attribute. This notion is very restrictive and has been criticized for its seeming unfairness in settings where there exists a correlation between the true label and the protected attribute (see, for example, [5]). It is not a major focus of our study.

Separation, also known as equalized odds [8], requires that the predictor's outcome be conditionally independent of the protected attribute, given the true label. This means that the distribution of outcomes should be the same across different groups, as long as individuals share the same true label. When the label and the prediction are both binary, this criterion is equivalent to asking for the same true positive rate and true negative rate across the different protected groups. In other words, it guarantees that an imperfect classifier would not err in either direction more on one subgroup than on others.

Sufficiency requires that the protected attribute and the true label be conditionally independent, given the classifier's prediction. This criterion requires that, among individuals that are given the same prediction, there would be no correlation between the true label and membership in a protected group. In other words, given the outcome of the classifier for an individual, the probability that she has a certain true label should be the same, regardless of her protected attribute. When the true label is binary, sufficiency is equivalent to calibration within subgroups [1]; roughly speaking, a predictor is calibrated if its prediction can be interpreted as the probability of the true label being positive.

1.1 Our contribution

As mentioned above, it is known that separation and sufficiency are often incompatible (although, as we show in Appendix A, they are sometimes compatible in non-trivial scenarios). A common consequence of this impossibility is that researchers and practitioners have been told that they must choose one or the other – separation or sufficiency – but not both. In this work, we propose a relaxation of these statistical criteria of fairness (Section 3.2), study the combinations of these relaxed criteria that are achievable (Lemma 10 and Theorem 12), and show that, in certain scenarios, there exists a continuous, meaningful tradeoff between the relaxed notions, forming a Pareto frontier (Theorem 14). We explore the relationships between the relaxed criteria, see that full sufficiency is not always achievable (Theorem 15), and draw conclusions regarding the post-processing of fair representations – post-processing cannot degrade our notion of approximate separation, but there are conditions on representations under which full sufficiency cannot be recovered by any post-processing (Theorem 17). These results enrich our understanding of the possible fairness tradeoffs that can be achieved, giving policymakers and domain experts a richer design space from which to select.

1.2 Related Work

There has been substantial work on relaxations of the statistical criteria of group fairness. Some approaches weaken the criteria in some non-continuous fashion. This includes, for example, relaxed equalized odds [12] (weakening separation) and equal precision [3] (weakening sufficiency), both of which are motivated by the incompatibility results, in an attempt to find weaker fairness notions that are compatible with each other. A major motivation for equal opportunity [8] (a relaxation of separation) in contrast, is compatibility with higher prediction accuracy than the stronger notion of equalized odds.

More continuous relaxations have primarily appeared in the contexts of learning and optimization. They typically involve the definition of some approximation or error term that quantifies how much a predictor violates the desired criterion, and then use this term either as an objective or as a penalty or regularization term in the learning objective of the predictor; such work has not focused on the *compatibilities* that such relaxations enable. Examples include Zafar et al. [16] and Woodworth et al. [15], who propose moment-based approximations in order to solve the intractability of the optimization problem, as well as Zemel et al. [17] and Bechavod and Ligett [2], who define a regularization term in order to achieve independence and separation, respectively. Of particular relevance to our work are Kashima et al. [10], who use mutual information as a regularization term for learning classifiers that satisfy independence, and Gopalan et al. [6], who define the objective of multicalibration [9] in terms of average conditional covariance (as discussed in Section 3.3).

Perhaps most closely related to our work is the analysis of Hamman and Dutta [7], who adopt an information-theoretic perspective with very similar definitions to ours. However, their analysis – based on partial information decomposition – focuses only on scenarios where at least one of the fairness criteria is fully met.

2 Preliminaries

Statistical notions of fairness focus on the aggregate properties of a population rather than of any particular individual. The literature on fair classification typically considers the relationships between three variables when defining statistical criteria for fairness: the true label (or target variable) Y, the protected attribute (or subgroup) A, and the predictor's outcome R. In principle, all relationships between these variables are possible (see Figure 1(a)), and various fairness notions can be seen as restrictions on the relationships between these variables.

Interestingly, the data upon which the predictor makes the prediction (or is learned), X, is usually not explicitly modeled. In the multicalibration literature [9], however, the data does play a central role, as all the variables Y, A and R are defined as (possibly stochastic) functions of X (see Figure 1(b)). If this is the case, not all relationships between Y, A, and R are possible. In particular, under this model, $A \perp Y \mid X$, meaning that all the information about the true label that is encoded in the protected attribute is also encoded in the data X.

In this work we consider a model that generalizes these two approaches: we explicitly include the data X in our model, but we do not assume that A and Y are functions only of X, thus allowing for more complex relationships between them (see Figure 1(c)). This model can capture any joint distribution $P(X, Y, A)$ without any further assumptions, with R determined exclusively by the conditional distribution $P(R|X)$.

Moreover, we prefer to view R as a representation of the data X – either an explicit representation that can be passed to some end-user, or an inner state of a model – and not necessarily as the final prediction or classification, which can be made upon this representation.

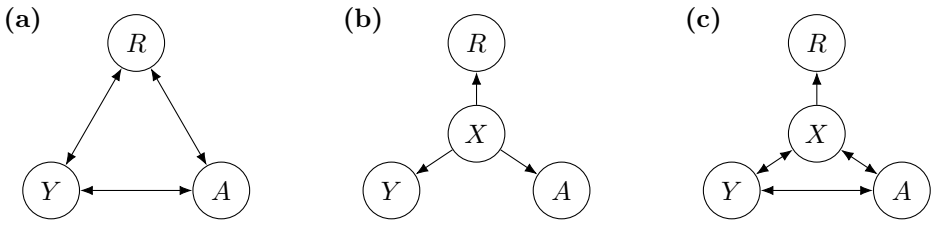

Figure 1 Diagrams of the relationships between the variables in models of fair classification: **(a)** a model where the data X is implicit and all relationships are permissible; **(b)** all variables are functions of X; **(c)** our model, where R is a function of X, but otherwise, all the relationships between X, Y and A are possible.

This view allows for greater flexibility and raises interesting points with respect to post-processing.

2.1 Statistical Criteria for Fair Classification

As mentioned above, the main statistical criteria for fair classification are defined in terms of statistical independence of the three variables Y, A and R:

▶ **Definition 1** (Statistical criteria for fair classification)**.** *Let Y, A and R be jointly distributed random variables, representing the true label, a protected attribute and the predictor's outcome, accordingly. Then,*

Independence *requires the predictor's outcome to be independent of the protected attribute: $R \perp A$. Specifically, for all r and a, $P(r|a) = P(r)$.*

Separation *requires the predictor's outcome to be conditionally independent of the protected attribute, given the value of the true label: $R \perp A \mid Y$. Specifically, for all r, a and y, $P(r|a, y) = P(r|y)$.*

Sufficiency *requires the true label and the protected attribute to be conditionally independent of each other, given the predictor's outcome: $Y \perp A \mid R$. Specifically, for all y, a and r, $P(y|a, r) = P(y|r)$.*

In this work we focus on separation and sufficiency, as independence is usually too restrictive.

2.2 Notation

Throughout this paper, we use capital letters for random variables, their respective lower case letters for their realizations, and script for the alphabet, as in $X = x \in \mathcal{X}$. The notation $P(X|y)$ is short for $P(X|Y = y)$. For simplicity, we assume that all variables are finite; if, for example, R naturally takes values in $[0, 1]$, we consider a suitable quantization.

3 Relaxing the Statistical Criteria of Fairness

A notable result in algorithmic fairness due to Chouldechova [3] and Kleinberg, Mullainathan, and Raghavan [11] states that for binary classification, separation and sufficiency cannot simultaneously hold except in trivial cases – either when Y and A are independent (meaning that different groups share the same base rate for the true label, eliminating any inherent fairness concerns), or when the predictor R perfectly predicts Y. More generally, they show that if Y and A are not independent and the joint distribution of Y, A, and R has full support (meaning that every combination of these variables has nonzero probability), then sufficiency and separation are fundamentally incompatible. (In Appendix A we show that there are non-trivial (though limited) settings where separation and sufficiency are compatible.)

Considering this incompatibility, we may look for suitable relaxations of the notions of separation and sufficiency, and analyze the relationships between the relaxed criteria. Since both criteria are defined in terms of statistical independence, one natural relaxation is to quantify the deviation from independence using (conditional) mutual information. We adopt a slightly more general approach, using f-information – a measure of statistical association based on f-divergences, which generalizes the well-established concept of mutual information.

3.1 f-Divergences and f-Information

The next section provides the essential background on f-divergences and f-information, including definitions and key properties.

▶ **Definition 2** (f-Divergence). *Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, and let P and Q be two probability distributions over the same space. The* f-divergence of P from Q *is defined as*

$$D_f[P\|Q] := \mathbb{E}_Q\, f\left(\frac{dP}{dQ}\right) = \sum_u Q(u)\, f\left(\frac{P(u)}{Q(u)}\right),$$

where the second equality holds for discrete distributions with the conventions that $f(0) = \lim_{t\downarrow 0} f(t)$, $0f(\frac{0}{0}) = 0$ and $0f(\frac{c}{0}) = \lim_{t\downarrow 0} t\, f(\frac{c}{t})$ for $c > 0$. For a complete definition and analysis, see [13, Chapter 7].

Some important examples of f-divergences are the Kullback-Leibler (KL) divergence, taking $f(t) = t \log t$ to obtain $D_{\mathrm{KL}}[P\|Q] = \mathbb{E}_P \log \frac{P}{Q}$ (note that the expectation is over P); the total variation (TV), with $f(t) = \frac{1}{2}|t - 1|$ yielding $\mathrm{TV}(P, Q) = \frac{1}{2}\|P - Q\|_1$; and the χ^2-divergence, with $f(t) = (x - 1)^2$. The Rényi divergences, although not f-divergences, are closely related and share many of their properties.

The following is a useful property of f-divergences.

▶ **Proposition 3.** *[13, Theorem 7.5] For any f-divergence D_f, $D_f[P\|Q] \geq 0$. If f is strictly convex at 1, then $D_f[P\|Q] = 0$ iff $P = Q$.*

In all the examples above, f is indeed strictly convex at 1.

Inspired by the definition of mutual information, a similar measure of association between random variables can be defined using f-divergences:

▶ **Definition 4** (f-Information). *Let U and V be random variables with joint distribution $P(U, V)$ and respective marginals $P(U)$ and $P(V)$. The* f-information between U and V *is defined as*

$$I_f(U; V) := D_f[P(U, V)\|P(U)P(V)],$$

that is, the f-divergence between their joint distribution and the product distribution of their marginals.

Let W be another jointly distributed random variable. Then, the conditional f-information between U and V given W *is defined as*

$$\begin{aligned} I_f(U; V|W) &:= \mathbb{E}_W\, D_f[P(U, V|W)\|P(U|W)P(V|W)] \\ &= \sum_w P(w)\, D_f[P(U, V|w)\|P(U|w)P(V|w)]. \end{aligned}$$

For KL divergence, f-information is exactly Shannon's mutual information. For TV, I_{TV} is closely related to the covariance (see Proposition 8).

The following property of f-information is an immediate consequence of Proposition 3 and the definition of statistical independence.

▶ **Proposition 5** (f-Information and independence). *If f is strictly convex at 1, then*

$$I_f(U;W) = 0 \iff U \perp W,$$
$$I_f(U;W|V) = 0 \iff U \perp W \mid V.$$

(If f is not strictly convex at 1, then independence still implies zero information, but not vice versa.)

We highlight another important property of f-information – the data processing inequality (DPI):

▶ **Proposition 6** (Data processing inequality). *[13, Theorem 7.16] Let random variables $U \to V \to W$ form a Markov chain in that order, meaning that $P(U|V,W) = P(U|V)$. Then $I_f(U;W) \leq I_f(U;V)$.*

3.2 Relaxing the Fairness Criteria

In light of Proposition 5, the statistical criteria for fair classification, which are defined in terms of conditional independence, can be relaxed using the notion of f-information.

▶ **Definition 7** (Deviations from fairness criteria). *Let Y, A and R be jointly distributed random variables, and let I_f be an f-information. We define*

Deviation from Independence *as* $\Delta_{\mathrm{ind}}(R) := I_f(A;R)$,

Deviation from Separation *as* $\Delta_{\mathrm{sep}}(R) := I_f(A;R|Y)$, *and*

Deviation from Sufficiency *as* $\Delta_{\mathrm{suff}}(R) := I_f(A;Y|R)$.

Each of these quantities measures to what extent the joint distribution of Y, A and R diverges from the statistical independence that defines the respective criterion, where a value of zero means that the criterion is fully satisfied. When the specific f-information is of relevance, we may write Δ^f_{sep} for clarity. We sometimes refer to the original statistical criteria as, e.g., "full separation" and "full sufficiency" to distinguish them from these relaxed notions.

An additional relevant quantity that is closely related to the deviations above is $I_f(R;Y)$, which expresses the amount of "information" that the predictor R conveys about the true label Y. This can be seen as a measure of the predictor's overall performance, similar to its accuracy. It can be shown that maximizing $I(R;Y)$ (using KL divergence) is equivalent to minimizing the expected cross-entropy loss, $-\mathbb{E}_{Y,R} \log P(Y|R)$, which is a common practice in the optimization of classifiers.[1]

[1] Note that $I(R;Y) = H(Y) - H(Y|R) = -\mathbb{E}_Y \log P(Y) + \mathbb{E}_R \mathbb{E}_{Y|R} \log P(Y|R)$, where the first term, $H(Y)$, is the entropy of Y and does not depend on R, and the second term is the negative of the expected cross-entropy loss (see, e.g., [4]).

3.3 Interpreting the Relaxed Criteria

The definition of the relaxed criteria in terms of f-information (Definition 7) may not seem intuitive at first. However, applying Bayes' rule and some algebra transforms these expressions into a form that more closely resembles the original definitions of the fairness criteria. Consider, for example, the deviation from separation:

$$\begin{aligned}\Delta_{\text{sep}}(R) = I_f(A;R|Y) &= \mathbb{E}_Y\, D_f[P(A,R|Y)\|P(A|Y)P(R|Y)] \\ &= \sum_y P(y) \sum_{a,r} P(a|y)P(r|y)\, f\left(\frac{P(a,r|y)}{P(a|y)P(r|y)}\right) \\ &= \sum_y P(y) \sum_a P(a|y) \sum_r P(r|y)\, f\left(\frac{P(r|a,y)}{P(r|y)}\right) \\ &= \sum_{a,y} P(a,y) \sum_r P(r|y)\, f\left(\frac{P(r|a,y)}{P(r|y)}\right) \\ &= \mathbb{E}_{A,Y}\, D_f[P(R|A,Y)\|P(R|Y)].\end{aligned}$$

This expression represents the expected divergence between the conditional distribution of R given both A and Y, and its conditional distribution given only Y. In other words, it quantifies how much, on average (over the values of A and Y), the predictor's outcome distribution within the subgroups differs from the average in the population, among individuals that share the same true label.

Similarly, for the deviation from sufficiency we have

$$\Delta_{\text{suff}}(R) = I_f(A;Y|R) = \mathbb{E}_{A,R}\, D_f[P(Y|A,R)\|P(Y|R)], \tag{1}$$

that is, the average difference between the the true label distribution within the subgroups and its average in the population, among individuals that share the same predicted value.

One potential drawback of using f-information to define relaxed fairness criteria is its average-case nature. In particular, smaller subgroups – especially those where certain labels or prediction values are rare – contribute less to the deviation. One could argue that this is an unfair approach to fairness. In any case, some of our results treat scenarios where one or more of the deviations is zero, in which case taking the average or the maximum yields the same result.

To address this drawback, some works instead use a maximum-based approach to defining relaxed group fairness criteria (see, for example, [15]). However, placing excessive weight on rare examples can introduce statistical biases and negatively impact the learning process of fair predictors and representations. This very concern has led the multicalibration literature [9, 6] to adopt a relaxed objective that is essentially equivalent to TV-information, as we demonstrate next.

Following the definition in [6], a predictor R is said to be α-approximate subgroup calibrated with respect to A if it satisfies $\mathbb{E}_R\,|\operatorname{Cov}[A,Y|R]| \le \alpha$.

▶ **Proposition 8.** *Let A, Y and R be jointly distributed random variables, and assume that A and Y take values in $\{0,1\}$. Then, R is α-approximate subgroup calibrated with respect to A iff $\Delta^{\text{TV}}_{\text{suff}}(R) \le 2\alpha$.*

The proof, provided in Appendix B, relies on the fact that in the binary case, $\mathbb{E}_R\,|\operatorname{Cov}[A,Y|R]| = \frac{1}{2}\Delta^{\text{TV}}_{\text{suff}}(R)$. More generally, for A and Y that take on numeric values, denote the ranges $\mathcal{R}_Y = \max Y - \min Y$ and $\mathcal{R}_A = \max A - \min A$. Then we have $\mathbb{E}_R\,|\operatorname{Cov}[A,Y|R]| \le \frac{1}{2}\mathcal{R}_A\mathcal{R}_Y\Delta^{\text{TV}}_{\text{suff}}(R)$.

Finally, Pinsker's inequality [13, Theorem 7.10] states that $\mathrm{TV}(P,Q) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}[P\|Q]}$ (when KL divergence is taken with the natural logarithm), meaning that using KL divergence for the deviation dominates the use of total variation. In particular, this gives $\mathbb{E}_R\,|\,\mathrm{Cov}[A,Y|R]| \leq \mathcal{R}_A \mathcal{R}_Y \sqrt{\frac{1}{8}\Delta^{\mathrm{KL}}_{\mathrm{suff}}(R)}$.

4 Main Results

Given that the criteria of full separation and full sufficiency cannot always be satisfied simultaneously, our primary goal is to address the following question:

For any given setting (defined by X, Y, and A), what are the best combinations of deviations from sufficiency and separation that are achievable?

This question leads us to formalize the following definition:

▶ **Definition 9** (The separation-sufficiency function)**.** *Let X, Y and A be jointly distributed finite random variables. We define* the separation-sufficiency function *as*

$$\Phi(\xi) := \inf\left\{\Delta_{\mathrm{suff}}(R) \,\middle|\, R \text{ is finite, } (A,Y) \to X \to R \text{ is Markov, and } \Delta_{\mathrm{sep}}(R) = \xi\right\},$$

for all ξ, such that the set above is not empty.

Since $\Delta_{\mathrm{suff}}(R) = I_f(A;Y|R)$, we know from Proposition 3 that the separation-sufficiency tradeoff function Φ is nonnegative. In what follows, we prove that Φ is attained as a minimum and it is continuous and convex. Moreover, we show that in certain settings it is monotonically nonincreasing, and in others it is bounded away from zero – two properties with important implications.

4.1 The Achievable Region and the Separation-Sufficiency Curve

Consider any joint distribution $P(X,Y,A)$. Any choice of a predictor R – that is, a choice of a conditional distribution $P(R|X)$ – corresponds to a point in the nonnegative octant representing the triplet $\big(\Delta_{\mathrm{sep}}(R), \Delta_{\mathrm{suff}}(R), I_f(R;Y)\big)$. We refer to the set of all such possible points as the *achievable region*. We first show that the achievable region (as illustrated in Figure 2) is convex and compact.

▶ **Lemma 10** (Convexity of the achievable region)**.** *Let X, Y and A be jointly distributed finite random variables. The associated* achievable region*, that is, the set*

$$\mathcal{S} := \left\{\big(\Delta_{\mathrm{sep}}(R), \Delta_{\mathrm{suff}}(R), I_f(R;Y)\big) \,\middle|\, R \text{ is finite and } (A,Y) \to X \to R \text{ is Markov}\right\} \subseteq \mathbb{R}^3_{\geq 0}$$

is convex and compact.

Proof. Denote by $\mathcal{P}(\mathcal{X})$ the set of all probability distributions over the alphabet of X. For all $Q \in \mathcal{P}(\mathcal{X})$, define the following functions:

$$\boldsymbol{\xi}(Q) := \sum_{a,y} P(a|y) \left(\sum_x P(y|x)Q(x)\right) f\left(\frac{\sum_x P(a,y|x)Q(x)}{P(a|y)\left(\sum_x P(y|x)Q(x)\right)}\right),$$

$$\boldsymbol{\eta}(Q) := \sum_{a,y} \left(\sum_x P(a|x)Q(x)\right)\left(\sum_x P(y|x)Q(x)\right) f\left(\frac{\sum_x P(a,y|x)Q(x)}{\left(\sum_x P(a|x)Q(x)\right)\left(\sum_x P(y|x)Q(x)\right)}\right),$$

$$\boldsymbol{\zeta}(Q) := \sum_{y} P(y)\, f\left(\frac{\sum_x P(y|x)Q(x)}{P(y)}\right).$$

Define the function $F : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X}) \times \mathbb{R}^3$ as $F(Q) := \big(Q, \boldsymbol{\xi}(Q), \boldsymbol{\eta}(Q), \boldsymbol{\zeta}(Q)\big)$, and let $\mathcal{C} := \operatorname{co} F(\mathcal{P}(\mathcal{X}))$ be the convex hull of the image of F in $\mathcal{P}(\mathcal{X}) \times \mathbb{R}^3$. Finally, define $\mathcal{C}_{P(X)} := \{(\xi, \eta, \zeta) \mid \big(P(X), \xi, \eta, \zeta\big) \in \mathcal{C}\} \subseteq \mathbb{R}^3$, that is, the "slice" of $\mathcal{C}$ that corresponds to the marginal of X in the first coordinate.

Since X is finite, $\mathcal{P}(\mathcal{X})$ is finite dimensional and compact, and since F is continuous, the set $\mathcal{C}$ is both convex and compact. Consequently, so is $\mathcal{C}_{P(X)}$, as an intersection with a hyperplane.

We will now show that $\mathcal{C}_{P(X)} = \mathcal{S}$, that is, the achievable region of $P(X, Y, A)$. Indeed, if $(\xi, \eta, \zeta) \in \mathcal{C}_{P(X)}$, then by its definition and the definition of a convex hull, there exist $k \in \mathbb{N}$, $Q_1, \ldots, Q_k \in \mathcal{P}(\mathcal{X})$ and $\alpha_1, \ldots, \alpha_k > 0$ with $\sum_{r=1}^{k} \alpha_r = 1$, such that

$$P(X) = \sum_{r=1}^{k} \alpha_r Q_r(X), \quad \xi = \sum_{r=1}^{k} \alpha_r \boldsymbol{\xi}(Q_r), \quad \eta = \sum_{r=1}^{k} \alpha_r \boldsymbol{\eta}(Q_r) \quad \zeta = \sum_{r=1}^{k} \alpha_r \boldsymbol{\zeta}(Q_r). \tag{2}$$

Define a predictor $R \in [k]$ according to the conditional distribution $P(r|x) = P(R = r|X = x) := \alpha_r \frac{Q_r(x)}{P(x)}$, for all $x \in \mathcal{X}$ and $r \in [k]$. This distribution is well-defined, since by (2), $\sum_{r=1}^{k} P(r|x) = \frac{\sum_{r=1}^{k} \alpha_r Q_r(x)}{P(x)} = \frac{P(x)}{P(x)} = 1$. In addition, we have $P(r) = \sum_x P(x)P(r|x) = \alpha_r \sum_x Q_r(x)\frac{P(x)}{P(x)} = \alpha_r$, and therefore by Bayes' rule, $P(x|r) = Q_r(x)$. Moreover, $(A, Y) \to X \to R$ clearly form a Markov chain, because R is defined in terms of X alone. Now, substituting α_r and Q_r in (2), it can be shown that $\xi = I_f(A; R|Y) = \Delta_{\text{sep}}(R)$, $\eta = I_f(A; Y|R) = \Delta_{\text{suff}}(R)$ and $\zeta = I_f(R; Y)$ (see Appendix C for a detailed derivation).

Conversely, if there exists a finite random variable R, such that $(A, Y) \to X \to R$ form a Markov chain, then following the approach in Appendix C, we get that $\Delta_{\text{sep}}(R) = I_f(A; R|Y) = \sum_r P(r)\boldsymbol{\xi}(P(X|r))$, $\Delta_{\text{suff}}(R) = I_f(A; Y|R) = \sum_r P(r)\boldsymbol{\eta}(P(X|r))$ and $I_f(R; Y) = \sum_r P(r)\boldsymbol{\zeta}(P(X|r))$. In other words, $\big(P(X), \Delta_{\text{sep}}(R), \Delta_{\text{suff}}(R), I_f(R; Y)\big)$ is a convex combination of the points $\big(P(X|r), \boldsymbol{\xi}(P(X|r)), \boldsymbol{\eta}(P(X|r)), \boldsymbol{\zeta}(P(X|r))\big) \in F(\mathcal{P}(\mathcal{X}))$, with the weights given by $P(R)$, and thus by its definition, $\big(P(X), \Delta_{\text{sep}}(R), \Delta_{\text{suff}}(R), I_f(R; Y)\big) \in \mathcal{C}$. Consequently, $\big(\Delta_{\text{sep}}(R), \Delta_{\text{suff}}(R), I_f(R; Y)\big) \in \mathcal{C}_{P(X)}$. ◀

It is worth noting that the definition of the achievable region can be extended to include $\Delta_{\text{ind}}(R) = I_f(A; R)$, while preserving its convexity and compactness. The proof is a straightforward extension of the arguments above.

The connection between the achievable region and the separation-sufficiency function is emphasized in the following corollary:

▶ **Corollary 11.** *The separation-sufficiency function Φ is a continuous convex curve, and it is attained as a minimum.*

Proof. Let $\bar{\mathcal{S}} = \{(\xi, \eta) \,|\, (\xi, \eta, \zeta) \in \mathcal{S}\}$, that is, the projection of the achievable region $\mathcal{S}$ onto the first two coordinates, which correspond to $\Delta_{\text{sep}}(R)$ and $\Delta_{\text{suff}}(R)$. By Lemma 10, $\mathcal{S}$ is convex and compact, implying the same for the projection $\bar{\mathcal{S}}$. The claim now follows from the observation that the curve $\Phi(\xi)$ is the lower boundary of $\bar{\mathcal{S}}$. ◀

The next result further characterizes the function Φ and establishes its domain.

▶ **Theorem 12.** *Let X, Y and A be jointly distributed finite random variables and denote by*

$$\mathcal{D} = \{\Delta_{\text{sep}}(R) \,|\, R \text{ is finite and } (A, Y) \to X \to R \text{ is Markov}\}$$

the set of all achievable values of deviation from separation, corresponding to all possible predictors. Then

$$\mathcal{D} = [0, I_f(A; X|Y)].$$

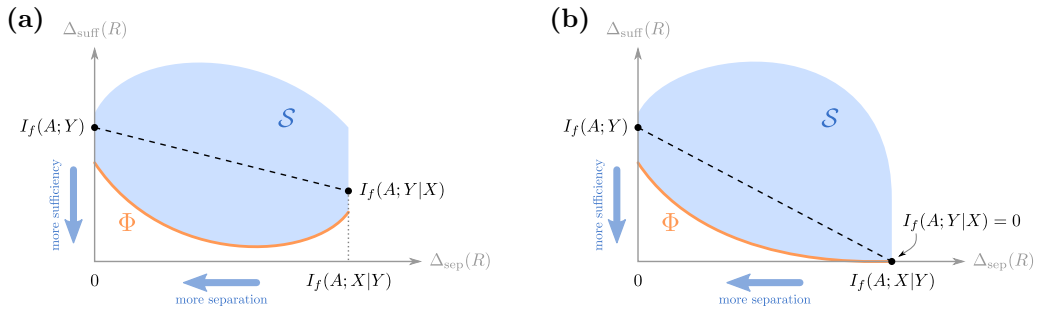

Figure 2 The achievable region: **(a)** A schematic illustration of the achievable region $\mathcal{S}$ (blue; in fact, what is depicted is the projection of $\mathcal{S}$ onto the first two coordinates, corresponding to $\Delta_{\text{sep}}(R)$ and $\Delta_{\text{suff}}(R)$), as well as the separation-sufficiency curve Φ (orange). Note that, in general, $I_f(A;Y)$ is not necessarily larger than $I_f(A;Y|X)$. **(b)** When A and Y are conditionally independent given X, i.e., $I_f(A;Y|X)=0$, the curve Φ is monotonically nonincreasing, attains the value zero, and represents a Pareto frontier of fairness, reflecting the tradeoff between separation and sufficiency.

For the proof, we use the following bound:

▶ **Proposition 13.** *Let X, Y, A and R be jointly distributed random variables, such that $(A,Y) \to X \to R$ form a Markov chain. Then $\Delta_{\text{sep}}(R) = I_f(A;R|Y) \le I_f(A;X|Y)$.*

Proof. This follows from the data processing inequality for the Markov chain $A \to X \to R$, conditioned on each value of Y, so that $I_f(A;R|y) \le I_f(A;X|y)$. ◀

Proof of Theorem 12. Since $\Delta_{\text{sep}}(R) = I_f(A;R|Y)$, Proposition 3 implies that all $\xi \in \mathcal{D}$ satisfy $\xi \ge 0$. Let $R_0 \equiv 0$ be a constant predictor, that is, $P(R_0 = 0) = 1$. Then R_0 meets the Markov condition and $\Delta_{\text{sep}}(R_0) = I_f(A;R_0|Y) = 0$. Therefore, the bound $\mathcal{D} \ge 0$ is tight.

On the other hand, by Proposition 13, all $\xi \in \mathcal{D}$ satisfy $\xi \le I_f(A;X|Y)$. Let $R_{id} = X$ be the identity predictor. Then R_{id} satisfies the Markov condition and $\Delta_{\text{sep}}(R_{id}) = I_f(A;R_{id}|Y) = I_f(A;X|Y)$. Therefore, the bound $\mathcal{D} \le I_f(A;X|Y)$ is also tight.

Finally, since $\mathcal{D}$ is the projection of the achievable set S onto the first coordinate, and since S is convex and compact by Lemma 10, $\mathcal{D}$ must be a closed interval in $\mathbb{R}$. Consequently, following the bounds above, $\mathcal{D} = [0, I_f(A;X|Y)]$. ◀

The proof shows that the predictors R_0 and R_{id} attain the left and right bounds of $\mathcal{D}$, respectively. Since we have $\Delta_{\text{suff}}(R_0) = I_f(A;Y|R_0) = I_f(A;Y)$ and $\Delta_{\text{suff}}(R_{id}) = I_f(A;Y|R_{id}) = I_f(A;Y|X)$, and the achievable region $\mathcal{S}$ is convex, we conclude that the line segment between the points $\big(0, I_f(A;Y)\big)$ and $\big(I_f(A;X|Y), I_f(A;Y|X)\big)$ is contained in $\bar{\mathcal{S}}$, the projection of S onto the first two coordinates (see Figure 2(a)). In particular, this implies the following bounds on the right endpoint of Φ:

$$0 \le \Phi\big(I_f(A;X|Y)\big) \le I_f(A;Y|X). \tag{3}$$

4.2 The Separation-Sufficiency Tradeoff

If both A and Y are determined by (possibly stochastic) functions of X, then all the information about the true label that is encoded in the protected attribute is also encoded in the data X. This assumption is common in the multicalibration literature, where A denotes

either a subset of $\mathcal{X}$ [9], or more generally a real valued function of X [6]. It is equivalent to saying that $Y \to X \to A$ form a Markov chain, that $A \perp Y|X$, or – most importantly for our analysis – that $I_f(A;Y|X) = 0$. In this case, we say that the data itself satisfies sufficiency.

In this specific setting, as we prove in the following theorem, the separation-sufficiency curve is monotonically nonincreasing, and thus represents a Pareto frontier of fairness, reflecting the fundamental tradeoff between separation and sufficiency (see Figure 2(b) for a schematic, and the left panel of Figure 3 for a simulation).

▶ **Theorem 14** (The separation-sufficiency tradeoff)**.** *Let X, Y and A be jointly distributed finite random variables, such that $A \perp Y \mid X$. Then the separation-sufficiency curve $\Phi(\xi)$ is convex, nonincreasing in ξ, and attains the value zero. In particular, there exist predictors that satisfy sufficiency.*

Proof. From Corollary 11 and Theorem 12, we know that $\Phi(\xi)$ is convex for all $\xi \in [0, I_f(A;X|Y)]$. In addition, the assumption that $A \perp Y \mid X$, meaning that $I_f(A;Y|X) = 0$, together with (3), implies that $\Phi\big(I_f(A;X|Y)\big) = 0$. Since Φ is attained as a minimum, it follows that there exist predictors that satisfy sufficiency (in particular, the identity predictor $R_{id} = X$).

Let $0 \le \xi_1 < \xi_2 < I_f(A;X|Y)$. From convexity, we have

$$\frac{\Phi(\xi_2) - \Phi(\xi_1)}{\xi_2 - \xi_1} \le \frac{\Phi\big(I_f(A;X|Y)\big) - \Phi(\xi_2)}{I_f(A;X|Y) - \xi_2} = \frac{0 - \Phi(\xi_2)}{I_f(A;X|Y) - \xi_2}.$$

Since $\Phi(\xi_2) \ge 0$, $I_f(A;X|Y) - \xi_2 > 0$ and $\xi_2 - \xi_1 > 0$, we conclude that $\Phi(\xi_2) - \Phi(\xi_1) \le 0$, that is, $\Phi(\xi_2) \le \Phi(\xi_1)$. This means that $\Phi(\xi)$ is monotonically nonincreasing in ξ. ◀

This result shows that, rather than needing to abandon either separation or sufficiency due to their incompatibility, a parametric approach reveals a continuous tradeoff between the measures of deviation from those criteria, exposing a range of intermediate combinations that can be achieved.

For example, suppose that a predictor was trained to achieve no worse than a certain subgroup calibration error. This objective corresponds to a level of deviation from sufficiency, and may be compatible with a quite good level of separation. The convexity of the curve suggests that a small relaxation of the subgroup calibration objective can result in a relatively large improvement in the separation measure.

4.3 Full Sufficiency Is Unachievable When the Data Is Not Rich Enough

The previous result relies heavily on the assumption that the data itself satisfies sufficiency. When this is not the case, that is, when $I_f(A;Y|X) > 0$, it means that the data X is not expressive enough to convey all the information about the relationship between the protected attribute A and the true label Y. This can occur if the protected attribute is not an inherent part of the data, yet it bears information about the true label. In particular, this can happen when the protected attribute is deliberately stripped from the data, as in the practice of "fairness through unawareness."

In this case, there is no guarantee that the separation-sufficiency curve is monotonic (see an example of such a situation in the middle panel of Figure 3), nor that it attains the value zero (see Figure 3, right panel). In other words, there is no guarantee that there exist predictors that satisfy perfect sufficiency. In fact, we show that in certain scenarios such predictors do not exist.

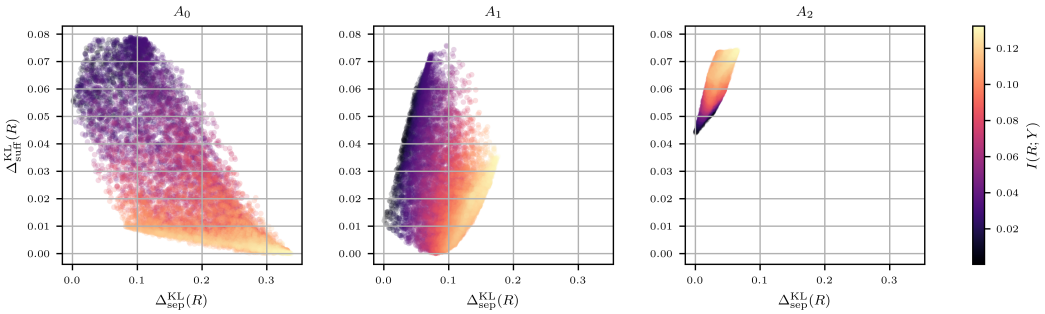

Figure 3 Simulations of the achievable region (using KL divergence) for a fixed joint distribution $P(X,Y)$ and three different scenarios of $P(A|X,Y)$. Left: $A_0 \perp Y \mid X$; it can be seen that the lower boundary of the achievable region, $\Phi(\xi)$, is indeed convex, monotonically nonincreasing, and attains the value zero. Middle: $A_1 \not\perp Y \mid X$, but $I(A_1;Y) < I(A_1;X)$; the lower boundary in not monotonic, but there exist predictors that achieve $\Delta^{\mathrm{KL}}_{\mathrm{suff}} = 0$. Right: $A_2 \not\perp Y \mid X$ and $I(A_2;Y) > I(A_2;X)$; the entire achievable region is bounded away from zero. See Appendix E for a detailed description of the simulation.

In the following analysis, we rely exclusively on Shannon's mutual information, as it allows us to apply its chain rule (see, for example, [4]), which does not necessarily have a counterpart in other forms of f-information.

▶ **Theorem 15.** *Let X, Y and A be jointly distributed finite random variables. Then for all ξ in its domain, the separation-sufficiency function using mutual information satisfies $\Phi(\xi) \geq \xi + I(A;Y) - I(A;X)$.*

Proof. Let R be a predictor defined by the conditional distribution $P(R|X)$, meaning that $(A,Y) \to X \to R$ form a Markov chain. In particular, $A \to X \to R$ form a Markov chain too, and thus by the data processing inequality we have

$$I(A;R) \leq I(A;X). \tag{4}$$

Now, using the chain rule of mutual information in two different ways, we get

$$\begin{aligned} I(A;R,Y) &= I(A;R) + I(A;Y|R) = I(A;R) + \Delta^{\mathrm{KL}}_{\mathrm{suff}}(R) \\ &= I(A;Y) + I(A;R|Y) = I(A;Y) + \Delta^{\mathrm{KL}}_{\mathrm{sep}}(R). \end{aligned}$$

Therefore, $\Delta^{\mathrm{KL}}_{\mathrm{suff}}(R) = \Delta^{\mathrm{KL}}_{\mathrm{sep}}(R) + I(A;Y) - I(A;R)$, and from (4) it follows that

$$\Delta^{\mathrm{KL}}_{\mathrm{suff}}(R) \geq \Delta^{\mathrm{KL}}_{\mathrm{sep}}(R) + I(A;Y) - I(A;X). \tag{5}$$

Finally, let $\xi \in [0, I(A;X|Y)]$. Then by Corollary 11, $\Phi(\xi)$ is attained by some predictor R, such that $\xi = \Delta^{\mathrm{KL}}_{\mathrm{sep}}(R)$ and $\Phi(\xi) = \Delta^{\mathrm{KL}}_{\mathrm{suff}}(R)$. Substituting these equalities in (5) yields $\Phi(\xi) \geq \xi + I(A;Y) - I(A;X)$. ◀

▶ **Corollary 16.** *If $I(A;X) < I(A;Y)$ then there exists no predictor that satisfies sufficiency.*

Proof. By Theorem 15 we have $\Phi(\xi) \geq \xi + I(A;Y) - I(A;X) > \xi \geq 0$. Since Φ is the lower boundary of the achievable region, this means that no predictor attains $\Delta^{\mathrm{KL}}_{\mathrm{suff}}(R) = I(A;Y|R) = 0$. ◀

(Note that, although the statement of the corollary is in terms of Shannon's mutual information, the result that $\Delta_{\text{suff}}(R) > 0$ holds for all f-information with strictly convex f at 1, since in that case $\Delta_{\text{suff}}(R) = I_f(A; Y|R) = 0$ implies sufficiency.)

It is worth noting that when $Y \to X \to A$ form a Markov chain, the data processing inequality guarantees that $I(A; Y) \leq I(A; X)$, and thus this corollary does not contradict Theorem 14.

This result has important implications when considering downstream uses of representations (post-processing), as we explore in the next section.

4.4 Post-Processing

Our analysis of the achievable region, along with the tradeoffs and limitations of separation and sufficiency, has some straightforward implications for the downstream use of fair representations. Specifically, note that when R is regarded as a representation, upon which a further prediction R' will be made, then R takes the place of X in our model (Figure 1(c)). This is formalized in the next result.

▶ **Theorem 17** (Post-processing). *Let X, Y and A be jointly distributed finite random variables. If R and R' are two finite random variables, such that $(A, Y) \to X \to R \to R'$ form a Markov chain in that order, we say that the predictor R' is a* post-processing *of R, or a* downstream predictor *based on R. We have the following results:*

1. *If R' is a post-processing of R, then $\Delta_{\text{sep}}(R') \leq \Delta_{\text{sep}}(R)$, meaning that the deviation from separation cannot increase with post-processing.*
2. *If R satisfies sufficiency, then the separation-sufficiency curve restricted to downstream predictors is convex, nonincreasing, and attains the value zero.*
3. *If R satisfies $I(A; R) < I(A; Y)$, then for any post-processing R' we have $\Delta_{\text{suff}}(R') > 0$, meaning that no downstream predictor can satisfy sufficiency.*

Proof. Note that if R' is a post-processing of R then, in particular, $(A, Y) \to R \to R'$ form a Markov chain. Therefore,

1. From Theorem 12 we have $\Delta_{\text{sep}}(R') \leq I_f(A; R|Y) = \Delta_{\text{sep}}(R)$.
2. If R satisfies sufficiency, then $A \perp Y \mid R$, and the claim follows from Theorem 14.
3. This is an immediate consequence of Corollary 16. ◀

Note that, as a consequence of the first result above, if R satisfies full separation then any downstream predictor will necessarily satisfy full separation too – essentially collapsing the space of possible tradeoffs with sufficiency. This is especially restrictive when taking into account that separation is fundamentally at odds with $I_f(R; Y)$, the amount of relevant information that the predictor conveys about the true label (see Appendix D).

Regarding statement 3, it should be noted that by the data processing inequality, $I(A; X) \geq I(A; R) \geq I(A; R')$, meaning that the quantity $I(A; R)$ can only decrease with post-processing, running the risk that repeated or aggressive post-processing of a representation can result in a state where full sufficiency is no longer achievable.

5 Discussion

We hope that this work contributes to elucidating the possible tradeoffs between various desiderata that are achievable in the design of classification algorithms, providing a richer space of options to domain experts and decision-makers.

Although we have primarily focused on separation-sufficiency tradeoffs, the tradeoff is, in fact, triple – between separation, sufficiency, and performance, as measured by $I_f(R;Y)$. In particular, there may exist several predictors that attain a specific point on the separation-sufficiency curve, with different levels of overall performance. In Appendix D we make a first step in this direction, and provide an analysis of the tension between separation and the predictor's performance. This is a promising direction for future research.

Furthermore, our focus in this work has been information theoretic – exploring the *possible* tradeoffs that predictors can achieve; developing algorithms to efficiently compute such predictors is an interesting open question.

References

1 Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

2 Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

3 Alex Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, October 2016. doi:10.1089/BIG.2016.0047.

4 Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd ed. edition, 2006.

5 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *3rd Innovations in Theoretical Computer Science Conference (ITCS)*, pages 214–226, 2012. doi:10.1145/2090236.2090255.

6 Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS)*, volume 215, pages 79:1–79:21, 2022. doi:10.4230/LIPICS.ITCS.2022.79.

7 Faisal Hamman and Sanghamitra Dutta. A unified view of group fairness tradeoffs using partial information decomposition. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 214–219, 2024. doi:10.1109/ISIT57864.2024.10619698.

8 Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

9 Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

10 Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650, 2011. doi:10.1109/ICDMW.2011.83.

11 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS)*, pages 43:1–43:23, 2017. doi:10.4230/LIPICS.ITCS.2017.43.

12 Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in Neural Information Processing Systems*, 30, 2017.

13 Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

14 Hans S. Witsenhausen and Aaron D. Wyner. A conditional entropy bound for a pair of discrete random variables. *IEEE Transactions on Information Theory*, 21(5):493–501, 1975.

15 Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017. URL: http://proceedings.mlr.press/v65/woodworth17a.html.

16 Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180, 2017. doi:10.1145/3038912.3052660.

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

A Achieving Separation and Sufficiency Together

If Y and A are not independent, then separation and sufficiency cannot simultaneously hold if the joint distribution of A, Y and R has full support [3, 11]. In the binary case, this means that separation and sufficiency can be achieved together only if R perfectly predicts Y. However, in the general case, this does not always result in trivial or degenerate conditions.

Indeed, assume that R satisfies both $A \perp R \mid Y$ and $A \perp Y \mid R$. The first condition means that $A \to Y \to R$ form a Markov chain in that order. From the data processing inequality for mutual information we have $I(R;A) \le I(Y;A)$. The second condition means that $A \to R \to Y$ form a Markov chain, and thus $I(Y;A) \le I(R;A)$. Together, we get that $I(Y;A) = I(R;A)$, meaning that R and Y convey the same information about A. In other words, this implies that R is a sufficient statistic of Y for A (see, for example, [4]).

The following is an example of such a case, where full separation and full sufficiency are not incompatible:

▶ **Example 18.** Let $A \sim \mathrm{Ber}(q)$ for some $q \in (0,1)$. For $i = 1,2,3$, let $X_i|A \sim \mathrm{Ber}(p_A)$, where $p_0 = \frac{1}{2}$ and $p_1 = \frac{1}{4}$. Define $Y = \sum_{i=1}^{3} X_i 2^{i-1}$, that is, the value of the binary word $X_1X_2X_3$, and $R = \sum_{i=1}^{3} X_i$. Then

$$P(Y|A) = \begin{array}{c} \\ A \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{bmatrix} \frac{1}{8} & \frac{1}{8} & & & \cdots & & & \frac{1}{8} \\ \frac{27}{64} & \frac{9}{64} & \frac{9}{64} & \frac{3}{64} & \frac{9}{64} & \frac{3}{64} & \frac{3}{64} & \frac{1}{64} \end{bmatrix} \end{array} \begin{array}{c} Y \\ \\ 0 \\ 1 \end{array}, \quad \text{and} \quad R = R(Y) = \begin{cases} 0, & Y = 0 \\ 1, & Y \in \{1,2,4\} \\ 2, & Y \in \{3,5,6\} \\ 3, & Y = 7 \end{cases}$$

From this definition of R, we see that the sequence $A \to Y \to R$ forms a Markov chain. In other words, R is conditionally independent of A given Y, meaning that it satisfies full separation.

A simple calculation shows that R also satisfies full sufficiency. For example,

$$P(Y=3|R=2,A=0) = \frac{P(Y=3,R=2|A=0)}{P(R=2|A=0)} = \frac{P(Y=3|A=0)}{P(Y\in\{3,5,6\}|A=0)} = \frac{\frac{1}{8}}{3\cdot\frac{1}{8}} = \frac{1}{3},$$

$$P(Y=3|R=2,A=1) = \frac{P(Y=3,R=2|A=1)}{P(R=2|A=1)} = \frac{P(Y=3|A=1)}{P(Y\in\{3,5,6\}|A=1)} = \frac{\frac{3}{64}}{3\cdot\frac{3}{64}} = \frac{1}{3}.$$

Consequently, even though Y is not independent of A, we have an example of a predictor that satisfies both separation and sufficiency, and yet does not allow for a perfect estimation of Y (for example, if $R = 1$, then Y can be either 1, 2 or 4). The predictor R is indeed a sufficient statistic of Y for A—it can be seen in the matrix above that for all A, the probability $P(Y|A)$ does not distinguish between values of Y that correspond to binary words $X_1X_2X_3$ with the same number of ones, which is exactly the value of R.

B Deviation from Sufficiency and Multicalibration

Following Definition 5.3 in [6], a predictor R is said to be α-approximate subgroup calibrated with respect to A if it satisfies $\mathbb{E}_R\,|\operatorname{Cov}[A, Y|R]| \le \alpha$.

▶ **Proposition 8.** *Let A, Y and R be jointly distributed random variables, and assume that A and Y take values in $\{0,1\}$. Then R is α-approximate subgroup calibrated with respect to A iff $\Delta^{\mathrm{TV}}_{\mathrm{suff}}(R) \le 2\alpha$.*

Proof. We show that when A and Y are binary, $\Delta^{\mathrm{TV}}_{\mathrm{suff}}(R) = 2\,\mathbb{E}_R\,|\operatorname{Cov}[A, Y|R]|$. Indeed,

$$\begin{aligned}
\operatorname{Cov}[A, Y|R] &= \mathbb{E}_{A,Y|R}\,AY - \mathbb{E}_{A|R}\,A\;\mathbb{E}_{Y|R}\,Y \\
&= \sum_{a,y\in\{0,1\}^2} P(a, y|R)\,ay - \sum_{a\in\{0,1\}} P(a|R)\,a \sum_{y\in\{0,1\}} P(y|R)\,y \\
&= \sum_{a,y\in\{0,1\}^2} P(a|R)P(y|a, R)\,ay - \sum_{a\in\{0,1\}} P(a|R)\,a \sum_{y\in\{0,1\}} P(y|R)\,y \\
&= \sum_{a\in\{0,1\}} P(a|R)\,a \left(\sum_{y\in\{0,1\}} P(y|a, R)\,y - \sum_{y\in\{0,1\}} P(y|R)\,y \right) \\
&= P(A = 1|R)\Big(P(Y = 1|A = 1, R) - P(Y = 1|R)\Big),
\end{aligned}$$

so we have $|\operatorname{Cov}[A, Y|R]| = P(A = 1|R)\big|P(Y = 1|A = 1, R) - P(Y = 1|R)\big|$.

In a similar way, it can be also shown that

$$\begin{aligned}
|\operatorname{Cov}[A, Y|R]| &= P(A = 0|R)\big|P(Y = 1|A = 0, R) - P(Y = 1|R)\big| \\
&= P(A = 1|R)\big|P(Y = 0|A = 1, R) - P(Y = 0|R)\big| \\
&= P(A = 0|R)\big|P(Y = 0|A = 0, R) - P(Y = 0|R)\big|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
|\operatorname{Cov}[A, Y|R]| &= \frac{1}{4} \sum_{a\in\{0,1\}} P(a|R) \sum_{y\in\{0,1\}} \big|P(y|a, R) - P(y|R)\big| \\
&= \frac{1}{4}\,\mathbb{E}_{A|R} \sum_{y\in\{0,1\}} \big|P(y|A, R) - P(y|R)\big|.
\end{aligned}$$

Now, using total variation to measure the deviation from sufficiency, we get from (1) that

$$\begin{aligned}
\Delta^{\mathrm{TV}}_{\mathrm{suff}}(R) = I_{\mathrm{TV}}(A; Y|R) &= \mathbb{E}_{A,R}\,\mathrm{TV}\left(P(Y|A, R), P(Y|R)\right) \\
&= \frac{1}{2}\,\mathbb{E}_{A,R} \sum_{y\in\{0,1\}} \big|P(y|A, R) - P(y|R)\big| \\
&= \frac{1}{2}\,\mathbb{E}_R\,\mathbb{E}_{A|R} \sum_{y\in\{0,1\}} \big|P(y|A, R) - P(y|R)\big| = 2\,\mathbb{E}_R\,|\operatorname{Cov}[A, Y|R]|. \;\blacktriangleleft
\end{aligned}$$

C Details for the Proof of Lemma 10

From (2) we have $\xi = \sum_{r=1}^{k} \alpha_r \boldsymbol{\xi}(Q_r)$. Substituting $P(r)$ for α_r and $P(X|r)$ for Q_r we get

$$\begin{aligned}
\xi &= \sum_r P(r)\boldsymbol{\xi}(P(X|r)) \\
&= \sum_r P(r) \sum_{a,y} P(a|y) \left(\sum_x P(y|x)P(x|r) \right) f\left(\frac{\sum_x P(a,y|x)P(x|r)}{P(a|y)\left(\sum_x P(y|x)P(x|r)\right)} \right) && (6) \\
&= \sum_r P(r) \sum_{a,y} P(a|y)P(y|r) f\left(\frac{P(a,y|r)}{P(a|y)P(y|r)} \right) && (7) \\
&= \sum_r P(r) \sum_{a,y} P(a|y)P(y|r) f\left(\frac{P(a,r|y)}{P(a|y)P(r|y)} \right) && (8) \\
&= \sum_y P(y) \sum_{a,r} P(a|y)P(r|y) f\left(\frac{P(a,r|y)}{P(a|y)P(r|y)} \right) = I_f(A;R|Y) = \Delta_{\text{sep}}(R), && (9)
\end{aligned}$$

where (6) is simply the definition of $\boldsymbol{\xi}$; (7) follows from the Markov assumption, $(A,Y) \to X \to R$, so $P(y|r) = \sum_x P(y|x)P(x|r)$ and $P(a,y|r) = \sum_x P(a,y|x)P(x|r)$; and (8) and (9) are due to Bayes' rule, as $\frac{P(a,y|r)}{P(y|r)} = \frac{P(a,r|y)}{P(r|y)}$ and $P(r)P(y|r) = P(y)P(r|y)$.

Similar steps show that $\sum_r P(r)\boldsymbol{\eta}(P(X|r)) = I_f(A;Y|R) = \Delta_{\text{suff}}(R)$ and $\sum_r P(r)\boldsymbol{\zeta}(P(X|r)) = I_f(R;Y)$.

D Separation Is at Odds with the Predictor's Overall Performance

The criterion of separation has sometimes been criticized (see, for example, [1, Chapter 3]) because imposing equal error rates across subgroups often leads to suboptimal performance for some of them. In this appendix we formalize this argument—that separation may conflict with the overall performance quality of the predictor.

Indeed, in addition to the tradeoff between separation and sufficiency, the convexity of the achievable region (Lemma 10) also implies a similar tradeoff between separation and $I_f(R;Y)$. As noted earlier, this quantity can be interpreted as a measure of the predictor's overall performance. The analysis below builds on the same reasoning as that of the separation-sufficiency tradeoff.

▶ **Definition 20** (The separation-performance function)**.** *Let X, Y and A be jointly distributed finite random variables. We define* the separation-performance function *as*

$$\Psi(\xi) := \sup\left\{ I_f(R;Y) \,\middle|\, R \text{ is finite, } (A,Y) \to X \to R \text{ is Markov, and } \Delta_{\text{sep}}(R) = \xi \right\},$$

for all $\xi \in [0, I_f(A;X|Y)]$.

Note that the domain of Ψ follows from Theorem 12.

▶ **Theorem 21** (The separation-performance tradeoff)**.** *The separation-performance function $\Psi(\xi)$ is continuous, concave, monotonically nondecreasing in ξ, and attained as a maximum.*

Proof. Recall the definition of the achievable region,

$$\mathcal{S} := \left\{ \left(\Delta_{\text{sep}}(R), \Delta_{\text{suff}}(R), I_f(R;Y)\right) \,\middle|\, R \text{ is finite and } (A,Y) \to X \to R \text{ is Markov} \right\} \subseteq \mathbb{R}^3_{\geq 0},$$

and denote by $\tilde{\mathcal{S}} = \{(\xi,\zeta) \,|\, (\xi,\eta,\zeta) \in \mathcal{S}\}$ the projection of $\mathcal{S}$ onto the first and third coordinates, corresponding to $\Delta_{\text{sep}}(R)$ and $I_f(R;Y)$. By Lemma 10, the achievable region

$\mathcal{S}$ is convex and compact, implying the same for its projection $\tilde{\mathcal{S}}$. Since $\Psi(\xi)$ is the upper boundary of $\tilde{\mathcal{S}}$, it follows that it is a continuous concave curve. In addition, the compactness of $\tilde{\mathcal{S}}$ implies that $\Psi(\xi)$ is attained as a maximum.

Let $\xi \in [0, I_f(A;X|Y)]$. Since $\Psi(\xi)$ is attained as a maximum, there exists a predictor R forming a Markov chain $(A,Y) \to X \to R$, such that $\Delta_{\text{sep}}(R) = \xi$ and $I_f(R;Y) = \Psi(\xi)$. By the data processing inequality we have $I_f(R;Y) \leq I_f(X;Y)$, and thus

$$0 \leq \Psi(\xi) \leq I_f(X;Y). \tag{10}$$

Now, denote by $R_{id} = X$ the identity predictor. Then $\Delta_{\text{sep}}(R_{id}) = I_f(A;R_{id}|Y) = I_f(A;X|Y)$ and $I_f(R_{id};Y) = I_f(X;Y)$. Therefore, by its definition, $\Psi\big(I_f(A;X|Y)\big) \geq I_f(X;Y)$, and together with (10) we get that $\Psi\big(I_f(A;X|Y)\big) = I_f(X;Y)$.

Finally, let $0 \leq \xi_1 < \xi_2 < I_f(A;X|Y)$. From concavity, we have

$$\frac{\Psi(\xi_2) - \Psi(\xi_1)}{\xi_2 - \xi_1} \geq \frac{\Psi\big(I_f(A;X|Y)\big) - \Psi(\xi_2)}{I_f(A;X|Y) - \xi_2} = \frac{I_f(X;Y) - \Psi(\xi_2)}{I_f(A;X|Y) - \xi_2}.$$

Since by (10), $I_f(X;Y) - \Psi(\xi_2) \geq 0$, $I_f(A;X|Y) - \xi_2 > 0$ and $\xi_2 - \xi_1 > 0$, we conclude that $\Psi(\xi_2) - \Psi(\xi_1) \geq 0$; that is, $\Psi(\xi_1) \leq \Psi(\xi_2)$. This means that $\Psi(\xi)$ is monotonically nondecreasing in ξ. ◀

This result is of particular concern if R serves as a representation, upon which further predictions will be made. The reason is that imposing a low deviation from separation already at an upstream stage may decrease the information that R bears on the true label Y; by the data processing inequality, this in turn will constrain all downstream predictors.

E Simulation of the Achievable Region

We provide the details and full results of the simulation that we ran in order to visualize the achievable region in concrete examples, corresponding to different settings of X, Y and A.

E.1 Method

First, we fix a joint distribution $P(X,Y)$. We then consider three different cases of $P(A|X,Y)$, corresponding to three distinct scenarios:

1. $A_0 \perp Y \mid X$, so we only need to define $P(A_0|X)$; this corresponds to the conditions of Theorem 14.
2. $A_1 \not\perp Y \mid X$, with $I(A_1;X) > I(A_1;Y)$.
3. $A_2 \not\perp Y \mid X$, with $I(A_2;X) < I(A_2;Y)$; corresponding to the conditions of Corollary 16.

For each of these scenarios we draw at random 10,000 conditional distributions $P(R|X)$ in the following manner: denote by $n = |\mathcal{X}|$ the size of the alphabet of X, then by Carathéodory-Fenchel-Eggleston theorem (see, e.g., [14, Lemma 2.2]), every point in the achievable region can be attained by a predictor R taking as most $n+1$ values. Hence, for each sampling of a distribution $P(R|X)$, it suffices to draw n distributions from $\mathcal{P}([n+1])$, the probability simplex over $n+1$ points. We do that using an $(n+1)$-dimensional Dirichlet distribution.

For each choice of $P(R|X)$, we calculate $\Delta_{\text{sep}}(R)$, $\Delta_{\text{suff}}(R)$ and $I_f(R;Y)$, using different f-divergences. Finally, we plot the points corresponding to the 10,000 sampled predictors in the $\Delta_{\text{sep}} - \Delta_{\text{suff}}$ plane, colored according to their respective value of $I_f(R;Y)$.

Table 1 Main information values of the simulated distributions for A_0, A_1 and A_2, using KL divergence, total variation and χ^2-divergence.

	KL divergence			Total variation			χ^2-divergence		
	A_0	A_1	A_2	A_0	A_1	A_2	A_0	A_1	A_2
$I_f(A;Y\|X)$	0.000	0.036	0.074	0.000	0.045	0.098	0.000	0.041	0.107
$I_f(A;X\|Y)$	0.338	0.177	0.066	0.200	0.162	0.096	0.553	0.267	0.090
$I_f(A;X)$	0.393	0.152	0.035	0.224	0.148	0.070	0.583	0.251	0.049
$I_f(A;Y)$	0.055	0.011	0.044	0.098	0.048	0.099	0.081	0.016	0.055

E.2 Details

We considered X taking 4 values, and binary Y and A. Specifically, we fixed $P(X) = [0.1, 0.1, 0.1, 0.7]$ and $P(Y = 1|X) = [0.8, 0.6, 0.4, 0.2]$, as well as the following conditional distributions for A:

1. $P(A_0 = 1|X) = [1.0, 0.4, 0.8, 0.2]$ with $P(A_0|X, Y) = P(A_0|X)$, so that $A_0 \perp Y \mid X$.
2. $P(A_1 = 1|X, Y = 0) = [0.9, 1.0, 0.8, 0.2]$ and $P(A_1 = 1|X, Y = 1) = [0.9, 0.6, 0.4, 0.4]$.
3. $P(A_2 = 1|X, Y = 0) = [0.9, 0.6, 0.8, 1.0]$ and $P(A_2 = 1|X, Y = 1) = [0.7, 0.4, 0.3, 0.4]$.

See Table 1 for the resulting values of $I_f(A;X|Y)$, $I_f(A;Y|X)$, $I_f(A;X)$ and $I_f(A;Y)$.

For each $P(R|X)$, as explained above, we sampled 4 distribution vectors (corresponding to the values of X) from a 5-dimensional symmetric Dirichlet distribution with parameter $\alpha = 0.1$. The resulting plots are shown in Figure 4.

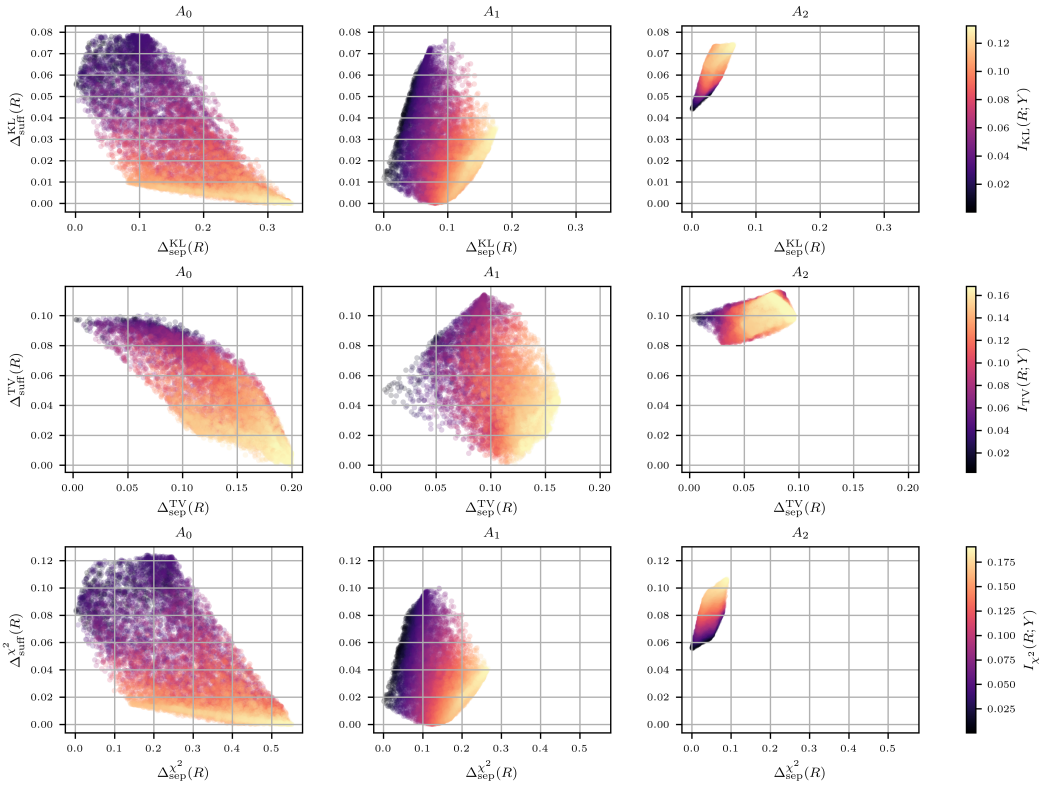

Figure 4 Simulations of the achievable region using KL divergence (top row; shown also in Figure 3), total variation (middle row) and χ^2-divergence (bottom row). The left plots correspond to the setting where $A \perp Y \mid X$, the middle and right plots correspond to the setting where $A \not\perp Y \mid X$, with $I(A;Y) < I(A;X)$ and $I(A;Y) > I(A;X)$, respectively.