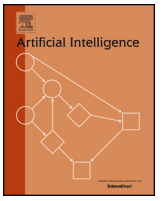
Artificial Intelligence 326 (2024) 104035

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



Gerrymandering individual fairness

Tim Räz

ARTICLE INFO

Keywords:
Fair-ML
Individual fairness
Group fairness
Gerrymandering
Ethics
Mathematics

ABSTRACT

Individual fairness requires that similar individuals are treated similarly. It is supposed to prevent the unfair treatment of individuals on the subgroup level and to overcome the problem that group fairness measures are susceptible to manipulation or gerrymandering. The goal of the present paper is to explore the extent to which individual fairness itself can be gerrymandered. It will be proved that individual fairness can be gerrymandered in the context of predicting scores. Then, it will be argued that individual fairness is a very weak notion of fairness for some choices of feature space and metric. Finally, it will be discussed which properties of (individual) fairness are desirable.

1. Introduction

The debate on fairness in machine learning distinguishes group fairness and individual fairness [17,24]. Group fairness is supposed to prevent the unequal treatment of groups, with varying notions of equal treatment. One of the problems of group fairness, pointed out by Dwork et al. [12], is that predictors can be manipulated such that they satisfy measures of group fairness, yet the predictions contradict fairness at the subgroup level. As a remedy, Dwork et al. propose individual fairness (IF), which requires that similar people are treated similarly. This is supposed to prevent the unfair treatment of individuals within groups. The goal of the present paper is to explore whether it is possible to apply *fairness gerrymandering* to IF itself. Informally, gerrymandering a fairness measure means that one can manipulate a predictive model such that the gerrymandered measure is not changed by the manipulation, but a different fairness measure is violated or changed. It will be proved that it is possible to gerrymander IF with respect to group fairness measures.

If one wants to show that gerrymandering IF is possible, one has to make choices, because the definition of IF is very general. In particular, a lot hinges on the notion of similarity. If one wants to turn IF into a substantive fairness measure, one has to choose which features of individuals are relevant to similarity, and how to measure similarity with a metric. Dwork et al. [12] argue that this is not a problem, but a feature of IF: choosing an appropriate metric is part and parcel of spelling out what we mean by fairness in a particular context. The present paper sets out to show that gerrymandering is possible for specific choices of predictors, features and metrics.

Sec. 2 provides formal definitions of IF and group fairness. Related work is discussed in Sec. 3. Gerrymandering IF predictors in the context of real-valued predictions (scores) is explored in Sec. 4. Two ways in which IF can be gerrymandered are identified. First, IF leaves room for treating people from certain groups *more* similarly than others. This can be exploited to construct unfair predictors, if different groups do not have similar distributions of scores. Second, IF leaves room for reversing the order of scores, as long as close-by individuals remain close. This can be exploited to construct unfair predictors by targeting a range of the score where certain groups have a high density. Gerrymandering IF by choosing particular feature spaces and metrics is explored in Sec. 5. The feature space determines which properties of individuals can be used to determine the similarity of individuals. It will be shown

E-mail address: tim.raez@posteo.de.

https://doi.org/10.1016/j.artint.2023.104035
Received 16 May 2022; Received in revised form 16 October 2023; Accepted 18 October 2023

Available online 24 October 2023

that large feature spaces, which can capture many properties of individuals, let us treat each individual as we please. Thus, in some contexts, IF provides a very weak notion of fairness. This result is generalized to probabilistic predictors. Finally, it will be discussed which aspects individual fairness are problematic and which aspects should be preserved (Sec. 6).

2. Formal background

2.1. Individual fairness

Individual fairness was defined by Dwork et al. [12]. Let the variable X represent (features of) individuals, and R scores (outcomes). The outcomes are given by a randomized mapping M from X to distributions Δ over R. Below, the special case where M is a (non-randomized) function will also be considered.[1] The similarity between individuals is represented by a metric d between their features, $d : X \times X \to \mathbb{R}_{\geq 0}$, and the similarity between outcomes is represented by a metric D between distributions of outcomes, $D : \Delta(R) \times \Delta(R) \to \mathbb{R}_{\geq 0}$.[2]

Definition 1. (IF): Let D, d be metrics. A randomized mapping $M : X \to \Delta(R)$ satisfies **individual fairness** if it satisfies the (D, d)-Lipschitz property, i.e., if for $p, q \in X$, we have: $D(Mp, Mq) \leq d(p, q)$.

Note that IF is usually applied to (distributions over) scores R. If a (binary) decision rule $\mathbb{1}[r \geq t]$ is constructed by applying a threshold t to a predictor R that satisfies IF, the decision rule need not satisfy IF.

2.2. Group fairness

The most important group fairness measures are *independence, sufficiency and separation*, cf. Barocas et al. [3]. Here a version of independence, statistical parity, will be considered. Let R represent scores and A group membership with values $A = a$ and $A = b$ (two groups).[3] The distributions of scores for groups a, b are given by real-valued probability densities $f_{R|a}(r) := f_{R|A=a}(r)$ and $f_{R|b}(r) := f_{R|A=b}(r)$. We define[4]:

Definition 2. A score R satisfies **statistical parity** with respect to groups $a, b \in A$ if

$$\mathbb{E}[R|a] = \mathbb{E}[R|b]. \tag{1}$$

Statistical parity requires that both groups have the same expected score, or that the expected score does not depend on the group variable. This requirement, which is quite stringent, can be weakened by requiring that the expectation of both groups are within a certain range, or by comparing degrees of statistical parity. Below we will examine how transformations of scores affect statistical parity quantitatively.

Definition 3. Let R be the score and ϕ a function of R with $\phi(R) = R'$. ϕ is **parity non-increasing** with respect to groups $A = a, b$ if applying ϕ to R does not increase the expectations with respect to these groups:

$$\mathbb{E}[R'|a] \leq \mathbb{E}[R|a]; \;\; \mathbb{E}[R'|b] \leq \mathbb{E}[R|b].$$

Such a transformation of scores can affect one group more than the other, which leads to a quantitative gap:

Definition 4. Let R be the score, $a, b \in A$ groups, and $\phi(R) = R'$ a parity non-increasing transformation of R. Group a has a (negative) **statistical parity gap** of $\epsilon \geq 0$ with respect to group b if:

$$\mathbb{E}[R'|a] - \mathbb{E}[R|a] + \epsilon = \mathbb{E}[R'|b] - \mathbb{E}[R|a]. \tag{2}$$

Group a has a negative statistical parity gap if the expected score of this group decreases more (by amount ϵ) under the transformation $\phi(R)$. Of course, it is also possible to define and examine positive statistical parity gaps. The following remark follows immediately from the definitions:

Remark 5. Let R be the score and a, b groups. If R satisfies statistical parity and we apply a parity non-increasing transformation ϕ to R such that group a has a negative statistical parity gap of $\epsilon > 0$, then $R' = \phi(R)$ violates statistical parity.

[1] Note that a predictor $M : X \to [0, 1]$ can be interpreted as a probabilistic version of $M : X \to \{0, 1\}$.

[2] See Çınlar and Vanderbei [10] for definitions of metric spaces and metrics. Strictly speaking, D and d are only pseudo-metrics, because metrics presuppose that $d(p, q) = 0$ iff. $p = q$, while Dwork et al. (p. 1) allow for the case that $d(p, q) = 0$ for $p \neq q$. Pseudo-metrics will be used throughout the paper.

[3] Note that group fairness measures do not make reference to the predictor function $M : X \to \Delta(R)$, but only to the resulting distribution of R with respect to groups.

[4] See, e.g., Corbett-Davies et al. [11] for a similar formulation of statistical parity.

2.3. Fair orderings

The order of scores provides a second example of a fairness measure that can be changed while preserving individual fairness. If we assign different, real-valued scores to different individuals, these scores induce an ordered list (ranking) of the individuals, with higher positions for individuals with higher scores.

Remark 6. Let X represent individuals, $M : X \to R$ a predictor. If n individuals are scored such that $M(x) \neq M(x')$, for $x \neq x'$, the predictor induces a strict linear order between individuals, represented by the set $[n] = \{1, ..., n\}$ with the natural order.

To define the degree to which a reordered list is unfair in comparison to a given list $[n]$, we need a notion of distance between orderings. The following is a standard measure of the distance between the list $[n]$ and a permutation σ of $[n]$, with $\sigma \in S_n$, the permutation group of n elements, cf. Kendall [20], Kemeny [19], Kumar and Vassilvitskii [21]:

Definition 7. Kendall's tau distance between $[n]$ and $\sigma \in S_n$ is given by

$$K(\sigma) = \sum_{(i,j):\, i>j} \mathbb{1}[\sigma(i) < \sigma(j)],$$

the number of pairs (i, j) with $i, j \in [n]$, such that σ inverts the order.

Remark 8. Kendall's tau can be formulated for two permutations $K(\sigma, \sigma')$, in which case it is an actual metric. Below, we will need the following properties [19,7] of K: (1.) If a permutation only affects the middle segment $\{k, k+1, ..., k+l\} \subset [n]$ with $1 \leq k \leq k+l \leq n$, then only pairs in this segment affect the value of K (Can and Storcken [7] call this *reducibility*). (2.) If the order of a list $[n]$ is completely reversed by σ, i.e., $\sigma(i) < \sigma(j)$ for all i, j with $i > j$, we get $K(\sigma) = \binom{n}{2} = \frac{n(n-1)}{2}$.

Definition 9. Let $M : X \to R$ be a predictor and $A = a, b$ groups. Abusing notation, denote the sizes of groups with $a, b \in \mathbb{N}$. Assume that M induces strict linear orders $[a]$, $[b]$ of group-specific scores. Let $R' = \phi(R)$ be a function of R that induces a strict linear reordering on $[a], [b]$, written as $\phi[a], \phi[b]$. Group a has a **strict order gap** of $\epsilon \geq 0$ with respect to group b if:

$$K(\phi[a]) - \epsilon = K(\phi[b]).$$

In this definition, strict linear orders are assumed. Also, this definition measures order gaps in absolute terms. By dividing each group-specific K by the size of the groups, one obtains a notion of order gap between densities. We will also consider the case where strictly ordered scores become equal.

Definition 10. Let $K'(\sigma)$ be a version of Kendall's tau that counts the pairs $i > j$ such that $\sigma(i) = \sigma(j)$, i.e., pairs that become equal in a non-strict linear reordering. With the assumptions of Definition 9 and ϕ a non-strict reordering, group a has an **non-strict order gap** of $\epsilon \geq 0$ with respect to group b if:

$$K'(\phi[a]) - \epsilon = K'(\phi[b]).$$

The non-strict order gap only measures unequal pairs of scores that receive the same scores after transformation ϕ. K and K' could be combined into a single measure of order gap (cf. Kemeny [19]), but we will not need this here. Finally, the following definition will be used:

Definition 11. A uniform distribution of n scores in an interval $[t, t'] \subset \mathbb{R}$ of length $l = t' - t$ means that the scores are equally spaced in the interval, e.g. at $t + i, i \in \{\frac{l}{n+1}, \frac{2l}{n+1}, ..., \frac{nl}{n+1}\}$.

3. Related work

The discussion of different kinds of group fairness picked up speed in reaction to the publication of Angwin et al. [2], which examined COMPAS, a risk assessment instrument, and found that COMPAS violates one kind of group fairness. There are now many surveys of fairness in machine learning (fair-ML), e.g., Mitchell et al. [24]. Barocas et al. [3] provide a regularly updated, book-length treatment of fair machine learning (with a focus on measures of group fairness); Kearns and Roth [18] provide a book-length introduction to ethical ML algorithms. The problem of gerrymandering statistical parity was noted in Dwork et al. [12]. In fact, the possibility of gerrymandering statistical parity was one of the main motivations of Dwork et al. to propose individual fairness. It has since been pointed out [28] that it is also possible to gerrymander other measures of group fairness, such as sufficiency (calibration) and separation (equalized odds). The issue of fair orders or rankings have been discussed, e.g., by Bower et al. [6], who consider the problem of constructing an individually fair ranking, such that similar individuals from different groups are ranked similarly. Bower et al. use Kendall's tau correlation as a way of comparing two rankings, which is related to, but distinct from Kendall's tau distance,

which is used here. Chakraborty et al. [9] examine the problem of fairly aggregating rankings from different groups using Kendall's tau distance.

Alternative variants to IF as defined by Dwork et al. [12] have been proposed. Friedler et al. [14] require that if individuals are ϵ-close in so-called Construct Space, then they should be mapped to ϵ'-close predictions in so-called Decision Space. Sharifi-Malvajerdi et al. [29] define a similar notion of *average individual fairness*. Kearns et al. [17] investigate the possibility of preventing gerrymandering group fairness by considering measures on rich subgroups, interpolating between group and individual fairness. Lahoti et al. [22,23] investigate formulations of individual fairness that are inspired by Dwork et al. [12], but use different formal expressions, such as consistency (the average distance of predictions for pairs of similar individuals), instead of the Lipschitz condition. It is noted in Dwork et al. [12] that IF is a strong requirement which may be hard to enforce. There have been several works addressing the problem of constructing IF predictors, e.g., through elicitation of the metric [15,25]. A framework for enforcing a fairness measure related to IF, which preserves order structure, is proposed in Jung et al. [16]. There are some proposals to enforce IF in-training [31,32,30], and in post-processing [26].

There have been few critical examinations of IF. Binns [4] provides a useful overview of philosophical work relevant to fair-ML, including IF. Binns [5] examines the relation between group fairness and IF, and argues that the conflict between these two kinds of fairness is only apparent. According to Binns, IF is based on statistical generalization, as opposed to a truly individualized notion of fairness, and should be regarded as a kind of group fairness because it groups together individuals with the same features. This point will be discussed in Sec. 5. Fleisher [13] provides a detailed critique of IF. Fleisher describes two ways of gerrymandering IF: Universal rejection (constant treatment), and increasing all scores by a fixed amount. Fleisher does not explore further possibilities of gerrymandering IF for given metrics, a gap filled by the present paper. Fleisher does not discuss the question whether the cases discussed by him constitute discrimination. According to a working definition [1], discrimination means that people from a socially salient group are put at a relative disadvantage due to their group membership.

4. Gerrymandering predictors

4.1. Idea and strategy

The goal of this section is to show that IF can be gerrymandered. First, some informal examples of gerrymandering are provided. Assume that we have a set X of job applicants, and a predictor $M : X \to R$ that assigns scores $r \in R$ to these individuals, capturing how suitable they are for the job. Scores are numbers in $[0, 1]$, where 0 means that an individual is maximally unsuitable, and 1 means that an individual is maximally suitable. We use the Euclidean distance on the interval $[0, 1]$ to measure the similarity of scores. We also assume that the predictor satisfies IF.

Here is a first example. Assume that a certain group of applicants $a \in A$ is overrepresented in a certain range of scores $[0.4, 0.5]$. We can now construct a new predictor such that all applicants in that interval are assigned the same score of 0.4. IF still holds, as long as we make sure that the distance between all scores does not increase. While this manipulation will affect the score of applicants of all groups, it will particularly affect and possibly hurt group a, because this group is overrepresented in the interval. To give a second example, assume that an individual x has a score $M(x) = 0.6$, while a different individual x' has a score $M(x') = 0.5$, but that we wish to assign x' a higher score than x. Again, it is possible to construct a predictor that assigns a higher score to x', by reversing the order of scores between 0.5 and 0.6, while making sure that IF still holds by not increasing distances. This manipulation will affect the scores of other individuals, but it will help x' and hurt x in particular.

Formally, the arguments in this section use the following assumptions. Predictors are functions $M : X \to R$, where R is a linearly ordered set, such as $\mathbb{R}$; note that this is a special case of $M : X \to \Delta(R)$. The set R is equipped with the Euclidean distance $D(p, q) = |p - q|$. It is assumed that a given predictor $M : X \to R$ satisfies IF. Gerrymandering is achieved by constructing a new predictor M' from M such that M' complies with IF, but violates other fairness measures. M' is constructed from M by finding a mapping $\phi : R \to R'$, with R' also linearly ordered, such that ϕ preserves the Lipschitz property. If a mapping ϕ preserves the Lipschitz property, and the predictor M satisfies IF, then it follows directly from the definition of IF that the predictor $M' : X \to R'$, defined as $M' := \phi \circ M$, also satisfies IF. Finally, if M' violates a different fairness measure, then we have shown that it is possible to gerrymander IF predictors.

4.2. Fairness gerrymandering

Gerrymandering a fairness measure F means that one can manipulate a score R with a transformation ϕ that does not affect the fairness measure F, but creates a fairness gap with respect to a different fairness measure F'. This idea is made precise in the following definition:

Definition 12. Let R be the score, F and F' fairness measures, and $a, b \in A$ groups. F can be **fairness gerrymandered** to degree $\epsilon > 0$ with respect to R, F', and a, b if there exists a (possibly randomized) function $\phi(R) = R'$ such that the following relations hold:

$$F(R', A = a, b) = F(R, A = a, b), \tag{3}$$

$$|F'(R', A = a) - F'(R, A = a)| - \epsilon = |F'(R', A = b) - F'(R, A = b)|. \tag{4}$$

Below, the case where F is IF will be explored, and the shorthand *IF can be gerrymandered with respect to F' and R* will be used.

The transformation ϕ leaves the fairness measure F invariant, but creates a fairness gap with respect to F' and groups a and b. F can be gerrymandered with respect to F' because F does not detect how the transformation ϕ affects fairness as measured by F'. The significance of the degree ϵ of gerrymandering depends on the context, and in particular on how undesirable a fairness gap of ϵ with respect to F' is. Below we will only see examples of negative gerrymandering degrees, where the transformation ϕ decreases F' for both groups. To measure binary fairness, one can choose the discrete metric, such that $\epsilon = 0$ if a fairness measure is satisfied, and $\epsilon = 1$ else. Definition 12 can be extended to accommodate fairness measures that depend on other variables, such as ground truth Y. The definition is quite general. The author is not aware of earlier attempts to provide a general definition of fairness gerrymandering. For example, Kearns et al. [17] refer to fairness gerrymandering informally as the idea that group fairness measures do not prevent the violation of fairness measures formulated at the subgroup level. Dwork et al. [12] discuss the possibility of violating IF while maintaining a version of statistical parity, that is, to gerrymander statistical parity.

In the definition of gerrymandering, nothing is said about the motivation behind ϕ. There are at least two ways in which the transformation ϕ and gerrymandering can arise. In **non-malignant fairness gerrymandering**, the transformation ϕ in Definition 12 is *not* chosen to affect fairness measure F', and gerrymandering is an unintended consequence of the use of ϕ. In **malignant fairness gerrymandering**, ϕ is chosen with the purpose of changing F' and of achieving gerrymandering intentionally.

4.3. Lipschitz maps

This section discusses maps ϕ on metric spaces that are known to be Lipschitz, and thus preserve IF. The goal is to prepare the ground for more targeted gerrymandering attacks in later sections. First, a contraction is a map that decreases the distance between any two points in a metric space, cf. Çınlar and Vanderbei [10]:

Definition 13. Let (X, d) be a metric space. A map $\phi : X \to X$ is a **contraction** if there is a constant $k \in [0, 1)$ such that for all $p, q \in X$ it holds $d\big(\phi(p), \phi(q)\big) \leq k \cdot d(p, q)$.

If a predictor $M : X \to R$ satisfies IF, then, given a contraction $\phi : R \to R$, any predictor $M' : X \to R$ with $M' := \phi \circ M$ will also satisfy IF. A first example of a contraction is $\phi : R \to R$, with $\phi(R) = r^*$ for $r^* \in R$, i.e., a constant map sending all predictions to a single value. It is a contraction because it reduces the distance between all predictions to 0. The fact that constant predictors M satisfy IF was already noted in Dwork et al. [12]. Fleisher [13] discusses constant predictors in the form of *Universal Rejection*: Assume that we want an IF predictor determining suitability of students for college, say, as a score in $[0, 1]$, we can choose a constant map that assigns the same, low score to all applicants, and thus reject them all. Fleisher notes that this may be considered unfair to those who would be suitable for college, but are rejected. However, universal rejection need not constitute discrimination in the sense that people from different groups are treated differently because of their group membership.[5] What is more, universal rejection is very costly if the predictor is supposed to have a certain accuracy and accept a certain proportion of all individuals. A second example of a contraction is a linear map $\phi : [0, 1] \to [0, 1]$, with $\phi(x) = cx$ and $c \in [0, 1)$. To see that this is a contraction, choose $k = c$ in the definition of contractions with the Euclidean metric. If M is an IF predictor, then so is $\phi \circ M$, because $\phi \circ M$ treats individuals more similarly than M.[6] The two examples, constant maps and linear contractions, contract at different rates: A constant map contracts an entire space to a point, while a linear contraction contracts more slowly, depending on the parameter c. The fact that we can choose different rates of contraction can be used to treat different groups of people differently; this is what opens up possibilities of targeted gerrymandering; cf Sec. 4.4. A second kind of Lipschitz map preserves distances:

Definition 14. Let (X, d) be a metric space. A mapping $\phi : X \to X$ is an **isometry** on X if for all $p, q \in X$ it holds $d(p, q) = d\big(\phi(p), \phi(q)\big)$.

Here we consider isometries in the setting of the metric space $\mathbb{R}$ equipped with the Euclidean metric. There are two kinds of isometries that preserve Euclidean distances on $\mathbb{R}$: translations and reflections; cf. Petrunin [27, Sec. 1.5.].

Definition 15. A **translation** is a map $\phi : \mathbb{R} \to \mathbb{R}$, with $\phi(x) = x + c, c \in \mathbb{R}$.

Translations shift all values in $\mathbb{R}$ by a fixed amount c. In the context of predicting scores, given an IF predictor $M : X \to \mathbb{R}$, we can use any translation ϕ to obtain a new predictor $M' = \phi \circ M$, in which predictions are shifted by an amount $c \in \mathbb{R}$, and satisfies IF. It has been pointed out in the literature [13, Sec. 3.1.] that translations preserve the Lipschitz property. But this does not mean that translations are discriminatory, because it does not necessarily target any group or individual. For discrimination, it is necessary that people from a socially salient group are put at a relative disadvantage due to their group membership [1]. Translations can be used in a targeted manner to discriminate under certain conditions. Assume that we have an IF predictor $M : X \to \mathbb{R}$, which is supposed to capture suitability for a job, where high values mean high suitability. Assume further that we use a decision rule $\mathbb{1}[r \geq t]$ with

[5] Constant predictions may be unfair for other reasons. For example, if we want to assign income tax rates to individuals, and use a constant predictor, i.e. a flat income tax, this will effectively yield a degressive tax scheme, because a fixed percentage of income does not have the same utility for people with high incomes as for people with low incomes.

[6] Note, incidentally, that we can apply a contraction ϕ many times and still preserve IF. In fact, in the limit of infinitely many applications of ϕ, we obtain a constant map. This is guaranteed by the Banach fixed point theorem; cf. Çınlar and Vanderbei [10].

threshold $t \in \mathbb{R}$ to make the hiring decision. Assume, finally, that many candidates from a protected group $a \in A$ have scores in the interval $[t, t']$ above the threshold t. If a gerrymanderer now wishes to hurt group a, they can apply a translation $\phi(r) \mapsto r - t'$ to the predictor M, such that people with scores in $[t, t']$ end up below the threshold. The new predictor $M' = \phi \circ M$ still satisfies IF with respect to the score, but hurts people from group a.

Gerrymandering in the above example is only possible under very specific circumstances. A gerrymanderer will only use a translation as described above if a certain amount of people from group a is concentrated in the interval $[t, t']$ just above the threshold t, and if the gerrymanderer's preference for not hiring these people is higher than not hiring people from other groups who also have scores in $[t, t']$. This means that while gerrymandering with a translation is possible, it can only be applied for this very specific distribution of scores among the different groups. An additional problem is that this instance of gerrymandering depends on the threshold remaining fixed, while scores are translated. This kind of gerrymandering is only possible if the (malicious) gerrymanderer has control over both scores and the threshold, which needs to remain fixed. However, it may be more plausible that if a translation is applied, then the threshold would also be shifted, thus offsetting the change.[7] Let us turn to the second kind of isometry on $\mathbb{R}$:

Definition 16. A **reflection** with respect to $c \in \mathbb{R}$ is a map $\phi : \mathbb{R} \to \mathbb{R}$, with $\phi(x) = -x + 2c$.

In the context of IF predictors, the fact that reflections are isometries and thus Lipschitz means that flipping all (binary) decisions preserves IF with respect to scores. Assume that we have an IF predictor $M : X \to \mathbb{R}$ and a decision rule $\mathbb{1}[r \geq t]$. If we let $\phi(r) = -r + 2t$, then $M' := \phi \circ M$ is also an IF predictor. M' will map all individuals that are in the positive class according M to the negative class, and vice versa. Again, this does not necessarily constitute discrimination. Also, the cost of applying a reflection in terms of accuracy is presumably prohibiting if the original predictor was somewhat accurate. We will see in Sec. 4.5 that a local version of reflections can be used to gerrymander in a wider set of circumstances.

4.4. Local contractions

Now we turn to a more targeted kind of gerrymandering. We have seen above that contractions can have different rates. Here this will be used to construct predictors that apply different rates of contraction depending on the group distributions. Assume that we have a predictor $M : X \to R$ with real-valued predictions. Assume further that the individuals in X belong to two different groups, $a, b \in A$, and that group a is overrepresented in the interval $[t, t'] \subset R$, meaning that individuals from a are more likely to have scores in this interval than group b. Such distributions are likely to occur if the two groups do not have equal distributions of scores. In some situations, it is possible to gerrymander the predictor M with the goal of hurting group a by applying a contraction to the interval $[t, t']$, while applying isometries outside the interval.

Definition 17. A **local contraction** of the interval $[t, t']$ to $t^* \in [t, t']$ is a function $\phi : \mathbb{R} \to \mathbb{R}$ defined as:

$$\phi(x) = \begin{cases} x - (t' - t^*) & \text{if } x > t', \\ t^* & \text{if } x \in [t, t'] \\ x + (t^* - t) & \text{if } x < t. \end{cases}$$

This map contracts scores in the interval $[t, t']$ to a point t^* in the interval, while applying an upward translation to scores below t, and a downward translation to scores above t'; it his therefore Lipschitz and preserves IF. Local contractions can be used to gerrymander IF with respect to statistical parity. The following proposition is stated for continuous distributions of scores, represented by density functions; a similar proposition holds for discrete variables.

Proposition 18. *Let $[t, t'] \subset \mathbb{R}$ be an interval such that group-specific score densities satisfy $f_{R|a}(r) \geq f_{R|b}(r)$ for $r \in [t, t']$, and let the group-specific cumulative probabilities satisfy $\mathbb{P}(R > t' | A = a) = \mathbb{P}(R > t' | A = b)$. Then, a local contraction $\phi(R)$ of $[t, t']$ to $t^* = t$ yields a negative statistical parity gap $\epsilon \geq 0$:*

$$\mathbb{E}[\phi(R)|a] - \mathbb{E}[R|a] + \epsilon = \mathbb{E}[\phi(R)|b] - \mathbb{E}[R|b] \tag{5}$$

The proof is given in appendix A. The parity gap in Proposition 18 is $\epsilon \geq 0$. By making stronger assumptions about the score densities, e.g., $f_{R|a}(r) > f_{R|b}(r)$ on an interval of positive measure, one obtains a parity gap $\epsilon > 0$. Also, the assumption that the cumulative probabilities are equal is not strictly necessary to obtain a parity gap, if the difference between the score densities is sufficiently big on an interval of sufficient size. Finally, it is also possible to gerrymander with choices other than $t^* = t$, e.g., $t^* = t'$, which may be used to create a positive statistical parity gap. From Proposition 18 and the fact that local contractions are Lipschitz we can deduce:

Corollary 19. *IF can be gerrymandered with respect to statistical parity by using a local contraction in the setting of Proposition 18.*

[7] I thank an anonymous reviewer for pointing out this issue.

Proof. Local contractions preserve IF. At the same time, local contractions create a negative statistical parity gap of $\epsilon > 0$ in the setting of Proposition 18, if $f_{R|a}(r) > f_{R|b}(r)$ on an interval of positive measure. A negative statistical parity gap means that statistical parity decreases more for one group than for the other, which means that IF can be gerrymandered. □

A local contraction can also be used to create a non-strict order gap:

Proposition 20. *Let $[t,t'] \subset \mathbb{R}$ be an interval such that more individuals from group a have scores in $[t,t']$ than individuals from group b, i.e., $a > b$. Then, a local contraction $\phi(R)$ of $[t,t']$ to $t^* \in [t,t']$ creates a non-strict order gap.*

Proof. By Remark 8, $K'(\phi[a]) = \frac{a(a-1)}{2}, = K'(\phi[b]) = \frac{b(b-1)}{2}$; with $a > b$, this implies an non-strict order gap $\epsilon = \frac{a(a-1)}{2} - \frac{b(b-1)}{2} > 0$. Note that outside $[t,t']$, ϕ is an order isomorphism. □

From Proposition 20 and the fact that local contractions preserve IF it follows:

Corollary 21. *IF can be gerrymandered with respect to non-strict order gaps by using a local contraction.*

4.5. Local reflections (folding)

In this section, we turn to gerrymandering based on local reflections. A so-called folding map allows us to put people with a lower score above people with a higher score, without reversing the order of all scores:

Definition 22. A **folding map** $\phi : \mathbb{R} \to \mathbb{R}$ of the interval $[t,t'] \subset \mathbb{R}$, $t < t'$ is defined as:

$$\phi(x) = \begin{cases} x - 2(t' - t) & \text{if } x > t', \\ -x + 2t & \text{if } x \in [t,t'] \\ x & \text{if } x < t. \end{cases}$$

A folding map adds a fold at the points t, t', such that scores between t and t' are reversed, while the scores above t' are translated down by twice the distance between t and t', and the scores below t are left unchanged. Note that the fact that scores below t are left unchanged can be corrected for by applying an upward translation after a folding map. Folding maps preserve IF:

Proposition 23. *Folding maps are Lipschitz with respect to the Euclidean metric.*

The proof is provided in appendix B. A result analogous to Proposition 18 shows that a folding map creates a statistical parity gap:

Proposition 24. *Let $[t,t'] \subset \mathbb{R}$ be an interval such that $f_{R|a}(r) \geq f_{R|b}(r)$ for $r \in [t,t']$, and let the cumulative probabilities satisfy $\mathbb{P}(R > t' | A = a) = \mathbb{P}(R > t' | A = b)$. Then, a folding map $\phi(R)$ of $[t,t']$ yields a negative statistical parity gap $\epsilon \geq 0$, cf. equation (5).*

The proof is given in appendix C; it is analogous to the proof of Proposition 18. From Propositions 23 and 24 it follows:

Corollary 25. *IF can be gerrymandered with respect to statistical parity by using folding maps.*

A folding map can also be used to gerrymander with respect to the score order.

Proposition 26. *Let $[t,t'] \subset \mathbb{R}$ be an interval of length l; let $a, b \in A$ groups, such that strictly more individuals from group a than from group b have uniformly distributed scores in $[t,t']$, and such that both groups have the same distributions of scores in the intervals $[t', t' + l]$ and $[t - l, t]$ respectively. Assume that scores from both groups can be strictly linearly ordered. Then, a folding map $\phi(R)$ of $[t,t']$ that induces a strict reorder for both groups creates a strict order gap of $\epsilon > 0$.*

A proof of Proposition 26 is given in appendix D. An order gap yields what Dwork et al. [12] call *self-fulfilling prophecy*: By putting worse candidates of a group in front of more suitable ones from the same group, people from that group end up with a rank in the order that are not reflective of their abilities, which can later be used as a justification for future discrimination. From Propositions 23 and 26, it follows:

Corollary 27. *IF can be gerrymandered with respect to strict order gaps by using folding maps.*

Table 1
Summary of results. *Changed Measure F'* is the fairness measure that changes while IF is unchanged. *Transformation* is the mapping that yields the result.

Changed Measure F'	*Transformation ϕ*	*Corollary No.*
statistical parity	local contraction	19
non-strict order gap	local contraction	21
statistical parity	folding map	25
strict order gap	folding map	27

4.6. Discussion

In this section, it was shown that gerrymandering an IF predictor is possible by using local versions of contractions and reflections. The local versions are more attractive manipulations than their global counterparts because they can be applied in more cases and preserve predictive accuracy to a certain extent. Gerrymandering IF is only possible for certain group distributions, specifically if different groups do not have the same distributions of scores, such that it is possible to target a certain interval in which one group is overrepresented. While the examples of gerrymandering depend on the context (Euclidean metric on the real numbers), the examples may generalize to other metric spaces, because they use special cases of maps on generic metric spaces (isometries and contractions). Above is Table 1 with the four main results of the section.

It could be asked how "natural" the above examples are, and in particular if the gerrymandering transformations ϕ can only arise as malignant gerrymandering. Probably the most natural of the transformations are order-preserving contractions, which may arise as some kind of normalization or in the creation of equal-sized risk bins. Lipschitz maps that change the order of scores lead to quantitatively larger fairness gaps, but preserving the order of scores may be a desideratum that may be enforced independently. Thus, order-preserving contractions may be more problematic because they cannot be detected by checking the order of scores.[8]

How does gerrymandering IF fare in comparison to gerrymandering group fairness (e.g. Dwork et al. [12], Räz [28])? On the one hand, IF gerrymandering is more fine-grained than gerrymandering group fairness. Specific assumptions about distributions of scores have to be met for IF gerrymandering. On the other hand, certain instances of gerrymandering IF would be prevented by imposing group fairness measures. In particular, the fact that IF can be gerrymandered with respect to statistical parity shows that statistical parity is a fairness requirement that is independent of IF and needs to be enforced separately.

The above examples of gerrymandering bring a general feature of individual fairness to the fore: IF is not defined in terms of the overall distribution of predictions, but only in terms of the relations (distances) between individuals and the associated predictions. One drawback of such a relative notion of fairness is that the extent to which gerrymandering is "worth it" depends on the gerrymanderer's preferences: If a gerrymanderer has a preference of not affecting a particular group b, and if the different groups have very similar distributions of scores, then IF provides a good protection against gerrymandering for all groups because gerrymandering would affect b as well. If a gerrymanderer wants to hurt group a and does not care much about adversely affecting other groups by targeting a, or if groups are very unevenly distributed, then IF may not provide much protection against gerrymandering.

5. Gerrymandering metrics and features

5.1. Idea

Individual fairness is a requirement on (randomized) predictors, which can be used to gerrymander, as we have seen in the last section. Individual fairness also presupposes the specification of a metric between individuals and a metric between predictions. Finally, it presupposes that we specify how individuals are represented in the metric space. In this section, it is explored how the choice of metric space affects individual fairness. This will reveal to what extent metric spaces can be chosen to yield unfavorable outcomes for a particular group, or for certain individuals.

An important difference between choosing metric spaces and gerrymandering with the predictor as explored in the previous section is that choosing a metric between individuals is an intended feature of the notion of individual fairness. According to Dwork et al. [12, p. 1] the metric d is supposed to capture "ground truth" with respect to what a society considers to be fair. Dwork et al. explain that the metric is supposed to be "open to discussion and continual refinement". Only by choosing an appropriate metric d does IF become a substantive notion of fairness. The upshot of this section is that if we do not place substantive restrictions on the metric space, the resulting notion of individual fairness is almost empty, which makes gerrymandering possible.

5.2. Metrics and spaces

A first possible choice of pseudo-metric d is the trivial metric. The trivial metric is identically zero, i.e., $d(p,q) = 0$ for all $p, q \in X$.[9] With the trivial metric, we are forced to choose a constant predictor $M(x) = c$ for all $x \in X$, if the metric D between predictions is a proper metric. To see this, assume that we allow different predictions $M(p) \neq M(q)$ for two individuals $p \neq q$. This implies that

[8] I thank an anonymous reviewer for raising the issue discussed in this paragraph.
[9] Note that this is not a proper metric because it violates $d(p,q) = 0$ iff. $p = q$.

$D(M(p), M(q)) \neq 0$ because D is proper. This yields $D(M(p), M(q)) > d(p, q)$, a violation of IF. Thus, the trivial metric implies that we have to choose a constant predictor. This choice may be unfair in some contexts, cf. Sec. 4.3

A second possible choice is the discrete metric:

$$d(p,q) = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$$

In order to see how the discrete metric affects individual fairness, one has to specify the space X on which the metric is defined. We can think of X as a space that represents relevant properties of individuals.[10] One important property of X is its size. If X is large, we have many properties at our disposal with which we can measure how similar the individuals represented in X are. If X is sufficiently large, it allows for unique identifiers for individuals.

Definition 28. A feature space X has **unique identifiers** if, for pairs of individuals a, b with $a \neq b$, if their representations in X are $r(a)$ and $r(b)$ respectively, we have that $r(a) \neq r(b)$.

Unique identifies can be implemented in different ways. If X is sufficiently large, it can represent individuals by representing their complete DNA sequence, pictures of their faces, fingerprints, and so on. Thus, each individual has properties in the feature space X distinguishing it from any other individual. If a feature space X has unique identifiers and is combined with the discrete metric d, the resulting instance of individual fairness is almost empty. Specifically, the following proposition holds:

Proposition 29. *Let X be a space with unique identifiers, d the discrete metric, and D a normalized metric, i.e., $D(M(p), M(q)) \leq 1$ for all $p, q \in X$. Then* any *function $M : X \to R$ satisfies individual fairness.*

Proof. Given two distinct individuals a, b with representations $r(a) = p, r(b) = q$ in X, we have $p \neq q$ because X has unique identifiers. From this, we get that $d(p, q) = 1$, because d is discrete. However, $D(M(p), M(q)) \leq 1$ for any M, p, q, because D is normalized. This implies that $D(M(p), M(q)) \leq d(p, q)$ for any M, p, q, which means that any M satisfies IF. □

Intuitively speaking, this means that for these choices of metrics and metric spaces, IF is an empty requirement, because it is compatible with any predictor. This problem has already been discussed with respect to Aristotle's notion of *consistency*, which requires that similar cases should be treated alike and is closely related to IF, cf. Binns [5, Sec. 3.1.]; Fleisher [13]. This makes gerrymandering possible because one can freely choose the predictor that creates a desired fairness gap.

A feature space X with unique identifiers, combined with the discrete metric d, corresponds to an individualized notion of fairness, what Binns [5] calls *individualized justice*: All individuals have something that distinguishes them from all other people, and these distinguishing features can be used as a justification for treating them differently. This leads to a notion of fairness that does not allow for (statistical) generalization, but requires predictions and decisions on an individualized basis, witnessed by the fact that there are no restrictions on the predictor function. This contradicts the claim by Binns [5, Sec. 4.2.] that IF is not compatible with individualized justice: If the feature space X is sufficiently rich, IF draws distinctions between any two individuals, which turns IF into a kind of individualized justice.

Note that we can also choose X to be maximally non-discriminative, e.g., by letting X be a singleton, which implies that all individuals in X will be treated similarly by definition. This corresponds to a minimally individualized notion of fairness: If we all have the same features, "We're all alike" [14] and thus should be treated similarly (or equally).

5.3. Absolute individual fairness

In previous sections, we restricted attention to point predictors $M : X \to R$. In the present section, we assume that M predicts *distributions* over R; it is a probabilistic function $M : X \to \Delta(R)$.

We will explore a special kind of IF called *absolute individual fairness* (absolute IF). The motivation for absolute IF comes from the fact that it is not clear what the correct metric d on the space of individuals X should be. To overcome this problem, we take the setting of supervised learning as an inspiration. We assume that we have pairs (x_i, y_i) of individuals $x_i \in X$, together with labels or ground truths $y_i \in Y$. To give an example, x_i could be a job applicant, while y_i is their true suitability for a certain job. The idea behind absolute IF is that we do not let the distance between individuals depend on their features, only on their ground truths. Absolute individual fairness requires that if two individuals x_i and x_j have the same distributions of ground truth Y (suitability for a job), then they should have the same distribution of predictions R (predicted suitability for a job). Thus, absolute IF is a condition of how a function $M : X \to \Delta(R)$ should approximate the function $M : X \to \Delta(Y)$. We represent distributions of ground truths as $f_Y(x) := P(Y|X = x)$, and distributions of predictions as $f_R(x) := P(R|X = x)$. With this notation, we can define absolute IF:

Definition 30. Let d' be the discrete metric on $f_Y(.)$ and D the discrete metric on $f_R(.)$ A probabilistic predictor $M : X \to \Delta(R)$ satisfies **absolute individual fairness** with respect Y if for all p, q in X, it holds: if $f_Y(p) = f_Y(q)$, then $f_R(p) = f_R(q)$.

[10] The importance of the choice of the feature space for considerations of fairness has been emphasized before in the literature, e.g. in Friedler et al. [14].

First, we have to show the following proposition:

Proposition 31. *Absolute IF is a kind of individual fairness.*

Proof. We use the discrete metric d on X which is induced by the discrete metric d' on $f_Y(.)$: given two individuals $p, q \in X$, we let $d(p,q) = 0$ if $f_Y(p) = f_Y(q)$, and $d(p,q) = 1$ otherwise. Because d and D are both discrete metrics, the only way to violate IF is if, for some $p, q \in X$, $d(p,q) = 0$ and $D(p,q) = 1$. However, $d(p,q) = 0$ implies $f_Y(p) = f_Y(q)$, while $D(p,q) = 1$ implies $f_R(p) \neq f_R(q)$, which violates absolute IF. □

To better understand what absolute individual fairness means, it is useful to have a different characterization. This characterization is based on the fact that the two discrete metrics d and D partition the space X: they group all and only those individuals in X together that have the same distribution with respect to the ground truth and the prediction, respectively. To make this formally precise, the following definitions are needed.

Definition 32. Let X, Y be random variables and S a statistic of X. S is a *sufficient statistic* of X for Y if $P(Y \mid X, S(X)) = P(Y \mid S(X))$.

This means that, for the purpose of predicting Y, we do not need the full X, we only need $S(X)$.

Definition 33. Let X, Y be random variables and T a sufficient statistic of X for Y. T is a *minimal sufficient statistic* of X for Y if it is a function of every other sufficient statistic, i.e., if, for every other sufficient statistic $S(X)$ of X for Y, there exists a function f such that $T(X) = f(S(X))$.

This means that, of all the statistics that are sufficient for Y, $T(X)$ contains the least amount of information. On this basis, one can prove that a predictor satisfies absolute individual fairness if and only if the partition of X induced by the predictor is more coarse-grained than the partition of X induced by the ground truth;

Proposition 34. *A probabilistic predictor $M : X \to \Delta(R)$ satisfies absolute individual fairness with respect to ground truth Y iff. the minimal sufficient statistic $U(X)$ of X for f_R is a function of the minimal sufficient statistic $T(X)$ of X for f_Y.*

The proof is in appendix E. A predictor satisfies absolute individual fairness if it is more coarse-grained than ground truth: Individuals with the same distributions of ground truth have to be treated equally. This implies that how strong a requirement absolute IF is depends on how fine-grained the partition of X with respect to f_Y is. If this partition is very fine-grained, then only few individuals have the same distributions of ground truth. And if this is the case, then more predictors will satisfy absolute individual fairness, because the predictor only needs to preserve a very fine-grained partition. This is a probabilistic version of Proposition 29 above. If this partition is coarse-grained, then many individuals have the same ground truth. In this case, the requirement on the predictor is more stringent.

Now, the key question is: What kind of partition of X with respect to f_Y should we expect? Will this partition usually be fine-grained or coarse-grained? There is an answer to this question in the statistical setting. The so-called Pitman-Koopman-Darmois theorem tells us that if the random variable X does not follow a distribution from the exponential family (see Casella and Berger [8] for details), then the minimal sufficient statistic $T(X)$ of X with respect to f_Y will not be much more coarse-grained than X itself. This, in turn, means that if X captures fine-grained information about individuals, as it will be the case if X has unique identifiers, then the minimal sufficient statistic of X with respect to f_Y will be very fine-grained, and it will not provide much of a restriction on predictors, which opens up the possibility of gerrymandering. This generalizes the lesson from the previous section: If many feature of individuals are taken into account, individual fairness is a very weak requirement.

5.4. Discussion

In this section, we have seen that IF strongly depends on the metrics and feature space X. In particular, if the space representing individuals X is very fine-grained, this can be used as a justification to essentially treat all people as one pleases, while still formally satisfying IF. If X is very coarse grained, then this forces us to use a predictor that takes the same value for many people. This dependence on properties of X carries over from a setting with point predictions to a probabilistic setting with distributions over predictions. Thus, as a formal requirement, IF is rather weak in that it may not provide any restrictions on predictors.

It could be wondered whether the insights in this section are due to particular features of the discrete metric. The main feature of the metric that yields the phenomena we saw here is the weight it puts on features that distinguish different individuals. For the discrete metric, these weights yield categorical differences. However, it is also possible to use metrics that draw more quantitative distinctions. The property that matters is how much weight the metric puts on individualizing features of individuals, which allows them to be treated more or less similarly.

6. General discussion and outlook

Let us recapitulate the main issues with IF identified above. IF leaves room for gerrymandering, because it is possible to satisfy IF while violating other fairness measures (to a certain degree). These manipulations are possible because IF is formulated as a *metric* requirement. The problem is that it is not just metric structure that matters for similar treatment. Similarity can be measured on the basis of many (formal) properties, not only on the basis of metric structure. A metric requirement is too weak if fairness requires that other kinds of structure, such as order, should be preserved as well. In some contexts, it is a problem that the Lipschitz condition only requires that distances do not increase, but not that they do not decrease, i.e., that the metric structure is preserved. This leaves room for (local) contractions, which also allow for gerrymandering. A further problem is that IF, as a formal requirement, does not prevent gerrymandering, because for some choices of metrics d, D and space X, IF is an almost empty requirement, because these choices are compatible with any predictor function.

It could be argued that at least some of the problems identified above are really features of IF. In particular, part of the idea of IF is that we have to make our fairness choices explicit through metrics d, D and spaces X. This is the whole point of "fairness through awareness". The idea of "fairness through awareness" has some merit: It is important that we make our choices of fairness explicit in some form and impose restrictions on possible predictors. But this does not resolve the fundamental issue that it is unclear why we have to "show our awareness" through the specification of the *metric* structure of a prediction problem, as opposed to showing awareness through other requirements. Of course, IF is useful in that it prevents some gerrymandering at the sub-group level, if only group fairness is required; this is one of the main points of Dwork et al. [12]. IF may thus complement group fairness measures. Here is a list of recommendations for formulating notions of (individual) fairness, broadly construed, which may resolve some of the issues just pointed out.

- The requirement of only preserving metric structure, as in IF, may be too weak. Depending on the context, it may be appropriate to require that other kinds of formal structure such as order are preserved.
- The requirement of having a metric on predictions in Y may be too strong, for example if some predictions are incommensurable (see Fleisher [13]). In this case, a different, weaker formal condition may be appropriate.
- To formulate a notion of individual fairness, it is not only necessary to specify what kind of structural condition (such as the Lipschitz property) we impose on the predictor, but it is also necessary to take the properties of the domain (set of individuals X and groups A) and the co-domain (predictions R) into account. In particular, the size and structural features of both domain and co-domain determine whether a fairness measure is appropriate.
- An appropriate notion of individual fairness should be combined with appropriate notions of group fairness. Individual fairness may prevent some forms of gerrymandering group fairness (see Dwork et al. [12]), while notions of group fairness may prevent some forms of gerrymandering individual fairness.
- While a notion of fairness may depend on merit or "ground truth" of individuals (labels Y), this is not necessary according to some notions of fairness. In particular, certain group fairness measures such as statistical parity, but also IF do not depend on ground truth (see Räz [28]).

7. Conclusion

The present paper argued that it is possible to gerrymander predictors that satisfy individual fairness under certain circumstances. In particular, gerrymandering is possible if different, socially salient groups have different outcome distributions. This makes it possible to implement local fairness attacks against these groups. The paper also argued that in different contexts and for certain choices of metrics and feature spaces, individual fairness is a very weak fairness requirement.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

The author thanks two reviewers for this journal for extensive and very helpful comments. This work is funded by the Swiss National Science Foundation through grant number 197504.

Appendix A. Proof of Proposition 18

Proof. To prove the result, we have to compute group-wise expectations of scores. The key fact yielding the result is that group a has score density at least as big as group b on the contracted interval; thus the contraction decreases the expectation more for group a. First we compute the terms on the l.h.s. of (5):

$$\mathbb{E}[R|a] = \int_{-\infty}^{\infty} r \cdot f_{R|a}(r)dr$$

$$= \int_{-\infty}^{t} r \cdot f_{R|a}(r)dr + \int_{t}^{t'} r \cdot f_{R|a}(r)dr + \int_{t'}^{\infty} r \cdot f_{R|a}(r)dr,$$

$$\mathbb{E}[\phi(R)|a] = \int_{-\infty}^{t} r \cdot f_{R|a}(r)dr + \int_{t}^{t'} t \cdot f_{R|a}(r)dr + \int_{t'}^{\infty} [r - (t' - t)] \cdot f_{R|a}(r)dr.$$

The same decomposition can be found for group b. From this we compute the difference on the l.h.s. of (5):

$$\mathbb{E}[\phi(R)|a] - \mathbb{E}[R|a] = \int_{t}^{t'} t \cdot f_{R|a}(r)dr - \int_{t}^{t'} r \cdot f_{R|a}(r)dr - (t' - t) \int_{t'}^{\infty} f_{R|a}(r)dr$$

$$= -\int_{t}^{t'} (r - t) \cdot f_{R|a}(r)dr - (t' - t) \int_{t'}^{\infty} f_{R|a}(r)dr$$

$$= -\int_{t}^{t'} (r - t) \cdot f_{R|a}(r)dr - (t' - t)\mathbb{P}(R > t'|A = a),$$

Note that $\mathbb{E}[\phi(R)|a] - \mathbb{E}[R|a] \leq 0$. Both the density and the distribution are non-negative, as are $(r - t)$ under the integral for $r \in [t, t']$ and $t' - t$; this yields two non-positive summands. The difference on the r.h.s. of (5) for group b is computed in the same manner, and is also non-positive. Now we subtract the r.h.s. from the l.h.s. in (5):

$$\big(\mathbb{E}[\phi(R)|a] - \mathbb{E}[R|a]\big) - \big(\mathbb{E}[\phi(R)|b] - \mathbb{E}[R|b]\big) \tag{A.1}$$

$$= -\int_{t}^{t'} (r - t) \cdot f_{R|a}(r)dr - (t' - t)\mathbb{P}(R > t'|A = a)$$

$$+ \int_{t}^{t'} (r - t) \cdot f_{R|b}(r)dr + (t' - t)\mathbb{P}(R > t'|A = b)$$

$$= \int_{t}^{t'} (r - t)[f_{R|b}(r) - f_{R|a}(r)]dr$$

The last equality follows because the cumulative probabilities above t' are equal by assumption. Now, the factor $(r - t)$ is non-negative for $r \in [t, t']$; on the other hand, $f_{R|a}(r) \geq f_{R|b}(r)$ for $r \in [t, t']$ by assumption, which means that $f_{R|b}(r) - f_{R|a}(r) \leq 0$, and thus the sum in (A.1). This imples:

$$\mathbb{E}[\phi(R)|a] - \mathbb{E}[R|a] \leq \mathbb{E}[\phi(R)|b] - \mathbb{E}[R|b],$$

which proves the proposition. Note that both sides are non-positive. Taking absolute values on both sides thus yields:

$$\Big|\mathbb{E}[\phi(R)|a] - \mathbb{E}[R|a]\Big| \geq \Big|\mathbb{E}[\phi(R)|b] - \mathbb{E}[R|b]\Big|,$$

which means that the parity gap increases more for group a than for group b. □

Appendix B. Proof of Proposition 23

Proof. We have to prove that the folding map $\phi : \mathbb{R} \to \mathbb{R}$ for an interval $[a, b] \subset \mathbb{R}$, $a < b$ is non-expansive, i.e., that for all $p, q \in \mathbb{R}$, it holds $|\phi(p) - \phi(q)| \leq |p - q|$. W.l.o.g. we assume $p > q$. There are six cases to consider. All inequalities are estimated with the triangle inequality.

Cases 1-3, $p, q > b$; $p, q \in [a, b]$; $p, q < a$: in these cases, ϕ is an isometry, and thus non-expansive.

Case 4, $p > b, q \in [a, b]$: $|\phi(p) - \phi(q)| = |p - 2(b - a) - (-q + 2a)| = |p + q - 2b|$. If $q = b$, we have $|p + q - 2b| = |p - q|$ and we are done. If $q < b$, we define $\epsilon, \delta > 0$ with $p = b + \epsilon$, and $q = b - \delta$. We continue $|p + q - 2b| = |b - \delta + b + \epsilon - 2b| = |\epsilon - \delta| \leq |\epsilon + \delta| = |\epsilon + b - b + \delta| = |p - q|$.

Case 5, $p > b, q < a$: $|\phi(p) - \phi(q)| = |p - 2(b - a) - q| = |p - q - 2(b - a)|$. Define $\epsilon, \delta > 0$ with $p = b + \epsilon$ and $q = a - \delta$. We continue $|p - q - 2(b - a)| = |b + \epsilon - a + \delta - 2(b - a)| = |\epsilon + \delta - (b - a)| \leq |\epsilon + \delta + (b - a)| = |p - b + a - q + (b - a)| = |p - q|$.

Case 6: $p \in [a,b], q < a$: This case holds in virtue of symmetry with case 4. For completeness's sake, here is the proof: $|\phi(p) - \phi(q)| = |-p+2a-q|$. If $p = a$, we have $|-p+2a-q| = |p-q|$ and we are done. If $p > a$, we define $\epsilon, \delta > 0$ with $p = a+\epsilon$, $q = a-\delta$. We continue $|-p+2a-q| = |-a-\epsilon+2a-a+\delta| = |\delta-\epsilon| \leq |\delta+\epsilon| = |p-a+a-q| = |p-q|$. □

Appendix C. Proof of Proposition 24

Proof. We compute the terms on the l.h.s. of equation (5). $\mathbb{E}[R|a]$ is as in the proof of Proposition 18. For the second term, with ϕ a folding map, we get:

$$\mathbb{E}[\phi(R)|a] = \int_{-\infty}^{t} r \cdot f_{R|a}(r)dr + \int_{t}^{t'} (-r+2t) \cdot f_{R|a}(r)dr + \int_{t'}^{\infty} [r-2(t'-t)] \cdot f_{R|a}(r)dr.$$

We then obtain the difference on the l.h.s. of (5):

$$\mathbb{E}[\phi(R)|a] - \mathbb{E}[R|a] = -2\int_{t}^{t'} (r-t) \cdot f_{R|a}(r)dr - 2(t'-t)\mathbb{P}(R > t'|A = a).$$

Expressions for $A = b$ are analogous. Finally, we calculate the following difference:

$$\begin{aligned}
&\mathbb{E}[\phi(R)|a] - \mathbb{E}[R|a] - \mathbb{E}[\phi(R)|b] + \mathbb{E}[R|b] \\
&= -2\int_{t}^{t'} (r-t) \cdot f_{R|a}(r)dr - 2(t'-t)\mathbb{P}(R > t'|A = a) \\
&\quad +2\int_{t}^{t'} (r-t) \cdot f_{R|b}(r)dr + 2(t'-t)\mathbb{P}(R > t'|A = b) \\
&= 2\int_{t}^{t'} (r-t) \cdot [f_{R|b}(r) - f_{R|a}(r)]dr
\end{aligned}$$

This expression is non-positive, for the same reasons as in the proof of Proposition 18. Note the factor of 2: In essence, the parity gap created by a folding map is twice that of a local contraction, because scores are mirrored at the axis t, instead of collapsed to t. □

Appendix D. Proof of Proposition 26

Proof. The proposition is proved in three steps, in which special cases of group distributions in $[t-l, t'+l]$ are examined. Importantly, ϕ only affects the order of scores in $[t-l, t'+l]$: everything outside this interval does not affect K and thus the order gap. In the first step, the spaces above and below $[t,t']$ are assumed to be empty, so that the folding yields a complete order inversion. In the second step, individuals are added above t'. In the third step, individuals are added below t. Individuals from group a, b in $[t-l, t'+l]$ are strictly ordered by score; orders are represented by $[a], [b]$. By assumption $a > b$.

Step 1: Assume that the distribution of groups in $[t,t']$ is as in the assumption, but the intervals $[t', t'+l]$ and $[t-l, t]$ are empty, such that the order of scores above and below $[t,t']$ are not affected by the folding ϕ. If we apply a folding ϕ, this results in a complete order inversion for lists $[a]$ and $[b]$ respectively; this yields group results of $K(\phi[a]) = \frac{a(a-1)}{2}$ and $K(\phi[b]) = \frac{b(b-1)}{2}$ respectively (see Remark 8). With $a > b$, this implies $K(\phi[a]) > K(\phi[b])$ and thus an order gap of $\epsilon > 0$.

Step 2: We assume the same distribution in $[t,t']$ as in step 1. In the interval $[t', t'+l]$, we assume that both groups have the same distributions of scores. The interval $[t-l, t]$ is assumed to be empty. The order of pairs (i,j) entirely in $[t', t'+l]$ is unchanged by ϕ and thus does not affect $K(\phi[a])$ or $K(\phi[b])$. Inverted pairs entirely in $[t,t']$ were calculated in step 1. Remaining inverted pairs (i,j) are such that $i \in [t', t'+l]$ and $j \in [t,t']$. We will now show that there are at least as many inverted pairs of this form for group a as for group b. Fix a member $i^* \in [t', t'+l]$ from group b. For this member, there is a fixed member $i' \in [t', t'+l]$ from group a with a score equal to i^* because the groups have the same distribution on $[t', t'+l]$. Consider all possible pairs (i^*, j), with $j \in [t,t']$ from group b. If an element (i^*, j) in this set is inverted, we can find a unique $j' \in [t,t']$ from group a such that (i', j') is inverted: there is a bijection $\psi(j) = j'$ from the ordered list $[b]_{[t,t']}$ of group b in $[t,t']$ to an initial segment of the ordered list $[a]_{[t,t']}$ of group a because $a > b$ in that interval. Note that the scores S of individuals j and $\psi(j)$ in $[t,t']$ before folding satisfy $S(j) > S(\psi(j))$: ψ matches all people from group b in $[t,t']$ with the equal-sized subset of people from group a in $[t,t']$ with the lowest scores; recall that both groups are uniformly distributed in $[t,t']$, such that scores in group a are closer together. Applying the folding ϕ to scores

yields $\phi[S(j)] < \phi[S(\psi(j))]$, because ϕ is order-inverting on $[t, t']$. Thus, for every j such that (i^*, j) is inverted from group b, the pair $(i', \psi(j))$ is inverted for group a. This means that at least as many new inverted pairs for group a than for group b have to be added due to people in $[t', t' + l]$, and we still have $K(\phi[a]) > K(\phi[b])$.

Step 3: Now we assume additionally that both groups have the same distribution of scores in $[t - l, t]$. The order within $[t - l, t]$ is unchanged. We now have to compare inverted pairs (i, j) of the form $i \in [t', t' + l], j \in [t - l, t]$, and of the form $i \in [t, t'], j \in [t - l, t]$. For the first kind, there is no difference between groups because they have the same distributions in both intervals. For the second kind, an similar argument as in step 2 shows that there are at least as many new inverted pairs for group a as for group b, which yields the result. □

Appendix E. Proof of Proposition 34

Proof. "⇒": Assume that a predictor $M : X \to \Delta(R)$ satisfies absolute individual fairness with respect to Y. We have to show that the minimal sufficient statistic $U(X)$ of X with respect to f_R is a function of the minimal sufficient statistic $T(X)$ of X with respect to f_Y. Assume that $T(p) = T(q)$ for some p, q. This implies that $f_Y(p) = f_Y(q)$, because $T(X)$ is the minimal sufficient statistic of X with respect to f_Y. The definition of absolute individual fairness implies that $f_R(p) = f_R(q)$. Now if $U(X)$ is the minimal sufficient statistic of X with respect to f_R, this also implies $U(p) = U(q)$. Otherwise, $U(X)$ would send elements of X with the same distribution to different values, a contradiction with it being a m.s.s. Thus, $U(X)$ is a function of $T(X)$.

"⇐": Assume that $U(X)$, the m.s.s. of X with respect to f_R, is a function of $T(X)$, the m.s.s. of X with respect to f_Y. Assume that p, q are elements of X that satisfy $f_Y(p) = f_Y(q)$. Now, because $T(X)$ is the m.s.s. of X with respect to f_Y, this implies that $T(p) = T(q)$. By assumption, there is a function g such that $U(X) = g(T(X))$. We thus get $g(T(p)) = g(T(q))$ and thus $U(p) = U(q)$. This, in turn, implies that $f_R(p) = f_R(q)$, which means that the predictor $M : X \to \Delta(R)$ satisfies absolute individual fairness with respect to Y. □

References

[1] A. Altman, Discrimination, in: E.N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, 2020, Winter 2020 ed.
[2] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks, ProPublica, 2016.
[3] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, 2022, fairmlbook.org.
[4] R. Binns, Fairness in machine learning: lessons from political philosophy, in: Conference on Fairness, Accountability and Transparency, Proc. Mach. Learn. Res. (2018) 149–159.
[5] R. Binns, On the apparent conflict between individual and group fairness, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 514–524.
[6] A. Bower, H. Eftekhari, M. Yurochkin, Y. Sun, Individually fair ranking, arXiv preprint arXiv:2103.11023, 2021.
[7] B. Can, T. Storcken, A re-characterization of the Kemeny distance, J. Math. Econ. 79 (2018) 112–116.
[8] G. Casella, R.L. Berger, Statistical Inference, second ed., Duxbury, 2002.
[9] D. Chakraborty, S. Das, A. Khan, A. Subramanian, Fair rank aggregation, Adv. Neural Inf. Process. Syst. 35 (2022) 23965–23978.
[10] E. Çınlar, R.J. Vanderbei, Real and Convex Analysis, Undergraduate Texts in Mathematics, Springer, New York, 2013.
[11] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness, in: KKD '17, 2017, pp. 797–806.
[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R.S. Zemel, Fairness through awareness, in: Proc. ACM ITCS, 2012, pp. 214–226.
[13] W. Fleisher, What's fair about individual fairness?, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, 2021, pp. 480–490.
[14] S.A. Friedler, C. Scheidegger, S. Venkatasubramanian, On the (im) possibility of fairness, arXiv:1609.07236, 2016.
[15] C. Ilvento, Metric learning for individual fairness, arXiv preprint arXiv:1906.00250, 2019.
[16] C. Jung, M. Kearns, S. Neel, A. Roth, L. Stapleton, Z.S. Wu, An algorithmic framework for fairness elicitation, arXiv preprint arXiv:1905.10660, 2019.
[17] M. Kearns, S. Neel, A. Roth, Z.S. Wu, Preventing fairness gerrymandering: auditing and learning for subgroup fairness, Proc. Mach. Learn. Res. 80 (2018) 2564–2572.
[18] M. Kearns, A. Roth, The Ethical Algorithm: The Science of Socially Aware Algorithm Design, Oxford University Press, 2019.
[19] J.G. Kemeny, Mathematics without numbers, Daedalus 88 (4) (1959) 577–591.
[20] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93.
[21] R. Kumar, S. Vassilvitskii, Generalized distances between rankings, in: Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 571–580.
[22] P. Lahoti, K.P. Gummadi, G. Weikum, iFair: learning individually fair data representations for algorithmic decision making, in: 2019 IEEE 35th International Conference on Data Engineering (ICDE), 2019, pp. 1334–1345.
[23] P. Lahoti, K.P. Gummadi, G. Weikum, Operationalizing Individual Fairness with Pairwise Fair Representations, Proc. VLDB Endow. 13 (4) (2019) 506–518.
[24] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, Algorithmic fairness: choices, assumptions, and definitions, Annu. Rev. Stat. Appl. 8 (2021) 141–163.
[25] D. Mukherjee, M. Yurochkin, M. Banerjee, Y. Sun, Two simple ways to learn individual fairness metrics from data, in: International Conference on Machine Learning, Proc. Mach. Learn. Res. (2020) 7097–7107.
[26] F. Petersen, D. Mukherjee, Y. Sun, M. Yurochkin, Post-processing for individual fairness, Adv. Neural Inf. Process. Syst. 34 (2021).
[27] A. Petrunin, Euclidean plane and its relatives, https://math.libretexts.org/@go/page/23576, 2021. (Accessed 26 January 2022).
[28] T. Räz, Group fairness: independence revisited, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 129–137.
[29] S. Sharifi-Malvajerdi, M. Kearns, A. Roth, Average individual fairness: algorithms, generalization and experiments, Adv. Neural Inf. Process. Syst. 32 (2019).
[30] A. Vargo, F. Zhang, M. Yurochkin, Y. Sun, Individually fair gradient boosting, arXiv preprint arXiv:2103.16785, 2021.
[31] M. Yurochkin, A. Bower, Y. Sun, Training individually fair ML models with sensitive subspace robustness, arXiv preprint arXiv:1907.00020, 2019.
[32] M. Yurochkin, Y. Sun, Sensei: sensitive set invariance for enforcing individual fairness, arXiv preprint arXiv:2006.14168, 2020.