Obstructing Classification via Projection

Pantea Haghighatkhah
TU Eindhoven, The Netherlands

Wouter Meulemans
TU Eindhoven, The Netherlands

Bettina Speckmann
TU Eindhoven, The Netherlands

Jérôme Urhausen
Utrecht University, The Netherlands

Kevin Verbeek
TU Eindhoven, The Netherlands

Abstract

Machine learning and data mining techniques are effective tools to classify large amounts of data. But they tend to preserve any inherent bias in the data, for example, with regards to gender or race. Removing such bias from data or the learned representations is quite challenging. In this paper we study a geometric problem which models a possible approach for bias removal. Our input is a set of points P in Euclidean space $\mathbb{R}^d$ and each point is labeled with k binary-valued properties. A priori we assume that it is "easy" to classify the data according to each property. Our goal is to obstruct the classification according to one property by a suitable projection to a lower-dimensional Euclidean space $\mathbb{R}^m$ $(m < d)$, while classification according to all other properties remains easy.

What it means for classification to be easy depends on the classification model used. We first consider classification by linear separability as employed by support vector machines. We use Kirchberger's Theorem to show that, under certain conditions, a simple projection to $\mathbb{R}^{d-1}$ suffices to eliminate the linear separability of one of the properties whilst maintaining the linear separability of the other properties. We also study the problem of maximizing the linear "inseparability" of the chosen property. Second, we consider more complex forms of separability and prove a connection between the number of projections required to obstruct classification and the Helly-type properties of such separabilities.

2012 ACM Subject Classification Theory of computation → Computational geometry; Theory of computation → Models of learning

Keywords and phrases Projection, classification, models of learning

Digital Object Identifier 10.4230/LIPIcs.MFCS.2021.56

Funding *Jérôme Urhausen*: Supported by the Dutch Research Council (NWO); 612.001.651.

Acknowledgements Research on the topic of this paper was initiated at the 5th Workshop on Applied Geometric Algorithms (AGA 2020) in Langbroek, NL. The authors thank Jordi Vermeulen for initial discussions on the topic of this paper.

1 Introduction

Classification is one of the most basic data analysis operators: given a (very) large set of high-dimensional input data with a possibly large set of heterogeneous properties, we would like to classify the data according to one or more of these properties to facilitate further analysis and decision making. Machine learning and data mining techniques are frequently employed in this setting, since they are effective tools to classify large datasets. However, just as any data-driven techniques, they tend to preserve any bias inherent in the data,

for example, with regards to gender or race. Such bias arises from under-representation of minority groups in the data or is caused by historical data, which reflect outdated societal norms. Bias in the data might be inconsequential, for example in music recommendations, but it can be harmful when classification algorithms are used to make life-changing decisions on, for example, loans, recruitment, or parole [23].

Naturally, the identification and removal of bias receives a significant amount of attention, although the problem is still far from solved. For example, Mehrabi et al. [19] provide a taxonomy of fairness definitions and bias types. They list the biases caused by data and the types of discrimination caused by machine learning techniques. Many approaches have been considered to eliminate or reduce bias in machine learning models. Some researchers have used a statistical approach to address this problem (e.g., [13]), while others focus on data preprocessing or controlling the sampling to compensate for bias or under-representation in the data (e.g., [2, 15]). Another approach is to use an additional (adversarial) machine learning model to eliminate bias in the first model (e.g., [11, 18, 27]). One major problem of attempting to eliminate bias (or increasing fairness) in machine learning is that it may negatively affect the accuracy of the learned model. This trade-off has also been studied extensively (e.g., [3, 25]).

We are particularly interested in data that is represented by vectors in high-dimensional Euclidean space. Such data arises, for example, from word embeddings for textual data. Several studies show that the bias present in the training corpora is also present in the learned representation (e.g. [7, 8]). Abbasi et al. [1] recently introduced a geometric notion of stereotyping. In this paper we follow the same premise that bias is in some form encoded in the geometric or topological features of the high-dimensional vector representation and that manipulating this geometry can remove the bias. This premise has been the basis for many papers on algorithmic fairness (e.g., [11, 12, 26]).

Several papers investigate the theory that gender is captured in certain dimensions of the data. Bolukbasi et al. [5] postulate that the bias manifests itself in specific "particularly gendered" words and that equalizing distances to these special words removes bias. Zhao et al. [28] devise a model which attempts to represent gender in one dimension which can be removed after training to arrive at a (more) gender-neutral word representation. Bordia and Bowman [6] remove bias by minimizing the projection of the embeddings on the gender subspace (using a regularization term in the training process). Very recently, various papers [9, 10, 14, 22] explored the direct use of projection to remove sensitive properties of the data. In some cases the data is not projected completely, as removing sensitive properties completely may negatively affect the quality of the model.

In this paper we take a slightly more general point of view. We say that a property is present in the data representation if it is "easy" to classify the data according to that property. That is, a property (such as gender) can be described by more complicated geometric relations than a subspace. Given the premise that the geometry of word embeddings encodes important relations between the data, then any bias removal technique needs to preserve as much as possible of these relations. Hence we investigate the use of projection to eliminate bias while maintaining as many other relations as possible. We say that the relation of data points with respect to specific properties is maintained by a projection, if it is still easy to classify according to these properties after projection. Our paper explores how well projection can obstruct classification according to a specific property (such as gender) for certain classification models.

Problem statement. Our input is a set of n points $P = \{p_1, \ldots, p_n\}$ in general position in $\mathbb{R}^d$. For convenience we identify the points with their corresponding vector. We model the various properties of the data (such as gender) as binary labels.[1] Hence, for all points in P we are also given k binary-valued *properties*, represented as functions $a_i \colon P \to \{-1, 1\}$ for $1 \leq i \leq k$. We denote the subset of points $p \in P$ with $a_i(p) = 1$ as P^i_+, and the subset of points $p \in P$ with $a_i(p) = -1$ as P^i_- for $1 \leq i \leq k$. For a point $p \in P$, we refer to the tuple $(a_1(p), \ldots, a_k(p))$ as the *label* of p. Note that there are 2^k different possible labels. Generally speaking, we do not know which specific properties a dataset has. However, to study the influence of projection on all relevant properties of a dataset, we assume that these properties are given.

We assume that it is "easy" to classify the points in P according to the properties by using the point coordinates. Throughout the paper, we consider different definitions for what is considered easy or difficult to classify. Our goal is to compute a projection P' of P to lower dimensions such that the first property a_1 becomes difficult to classify in P', and the other properties $a_2, \ldots, a_k$ remain easy to classify in P'. As a shorthand we use the notation $P_- = P^1_-$ and $P_+ = P^1_+$ for the point sets in which the special property a_1 is set to -1 and $+1$, respectively. Similarly, we use the notation P'_- and P'_+ for the point sets P_- and P_+ after projection. In most cases we will consider a projection along a single unit vector w ($\|w\| = 1$), mapping points in $\mathbb{R}^d$ to points in $\mathbb{R}^{d-1}$. For a point $p_i \in P$, we denote its projection as $p'_i = p_i - (p_i \cdot w)w$, where $(p_i \cdot w)$ denotes the dot product between the vectors $p_i, w \in \mathbb{R}^d$. To assign coordinates to p'_i in $\mathbb{R}^{d-1}$, we need to establish a basis for the projected space. We therefore often consider p'_i to lie in the original space $\mathbb{R}^d$, where the coordinates of p'_i are restricted to the hyperplane that is orthogonal to w and passes through the origin. Sometimes we will consider projections along multiple vectors $w_1, \ldots, w_r$. In that case we assume that $\{w_j\}_{j=1}^r$ form an orthonormal system, such that we can write the projection as $p'_i = p_i - \sum_{j=1}^r (p_i \cdot w_j)w_j$. Again, we assume that p'_i still lies in $\mathbb{R}^d$, but is restricted to the $(d-r)$-dimensional flat that is orthogonal to $w_1, \ldots, w_r$ and passes through the origin.

We consider different models for defining what is easy or difficult to classify, resulting in different computational problems. These models typically rely on a form of "separability" between two point sets. For a specific definition of separability, using a slight abuse of notation, we will often state that a property a_i is separated in a point set P when we actually mean that P^i_- and P^i_+ are separated (see Figure 1 for a simple example in $\mathbb{R}^2$). The specific models, along with the relevant definitions, are described in detail in the respective sections.

[1] Neither gender nor many other societally relevant properties are binary, however, we restrict ourselves to binary properties to simplify our mathematical model.

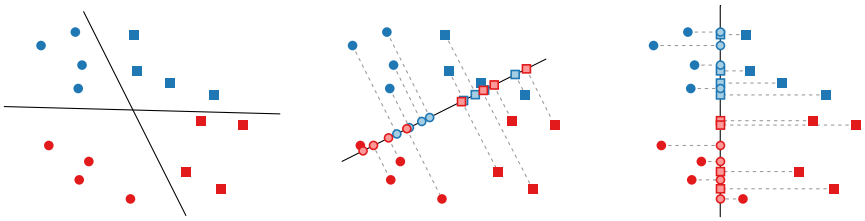

Figure 1 Left: data points with two linearly-separable properties: shape and color. Middle: a projection which keeps shape separated, but not color. Right: a projection with the opposite effect.

Contributions and organization. In Section 2 we consider linear separability as the classification model. We first show that, if even one possible label is missing from P, then there may be no projection that eliminates the linear separability of a_1 whilst keeping the linear separability of the other properties. On the other hand, if all possible labels are present in the point set, then we show that it is always possible to achieve this goal. In Appendix A we discuss a related question: given a measure to quantify how far removed a labeled point set is from linear separability, how can we optimize this measure for a_1 after projection? We show that the optimal projection can be computed efficiently under certain specific conditions, but may be hard to compute efficiently in general. In Section 3 we introduce (b, c)-separability, which is a generalization of linear separability. Although a single projection is no longer sufficient to avoid (b, c)-separability of a_1 after projection, we show that, in general, the number of projections needed to achieve this is linked to the Helly number of the respective separability predicate. We then establish bounds on the Helly numbers of (b, c)-separability for specific values of b and c. Omitted proofs can be found in Appendix B.

2 Linear separability

In this section we consider linear separability for classification. For a point set P and property $a_i \colon P \to \{-1, 1\}$, we say that a_i is easy to classify on P if P^i_- and P^i_+ are (strictly) linearly separable; we say that a_i is difficult to classify otherwise. Two point sets P and Q ($P, Q \subset \mathbb{R}^d$) are *linearly separable* if there exists a hyperplane H separating P from Q. The point sets are *strictly linearly separable* if we can additionally require that none of the points lie on H. Equivalently, the point sets P and Q are linearly separable if there exists a unit vector $v \in \mathbb{R}^d$ and constant $c \in \mathbb{R}$ such that $(v \cdot p) \leq c$ for all $p \in P$ and $(v \cdot q) \geq c$ for all $q \in Q$ (v is the normal vector of the hyperplane H). We say that P and Q are linearly separable *along* v. If the inequalities can be strict, then the point sets are strictly linearly separable.

One of the machine learning techniques that use linear separability for classification are *support vector machines* (SVMs). SVMs compute the (optimal) hyperplane that separates two classes in the training data (if linearly separable), and use that hyperplane for further classifications. Linear separability is therefore a good first model to consider for classification.

Let $CH(P)$ denote the convex hull of a point set P. By definition, we have that $x \in CH(P)$ if and only if there exist coefficients $\lambda_i \geq 0$ such that $x = \sum_{i=1}^n \lambda_i p_i$ and $\sum_{i=1}^n \lambda_i = 1$. We use the following basic results on convex geometry and linear algebra.

▶ **Fact 1.** *Two point sets P and Q are linearly separable iff $CH(P)$ and $CH(Q)$ are interior disjoint. P and Q are strictly linearly separable iff $CH(P) \cap CH(Q) = \emptyset$.*

▶ **Observation 2.** *Let $P' = \{p'_1, \ldots, p'_n\}$ be the point set obtained from $P = \{p_1, \ldots, p_n\}$ by projecting along a unit vector w. If $x = \sum_{i=1}^n \lambda_i p_i$ (for $\lambda_i \in \mathbb{R}$), then $x' = x - (w \cdot x)w = \sum_{i=1}^n \lambda_i p'_i$. Specifically, if $x \in CH(P)$, then $x' \in CH(P')$.*

▶ **Lemma 3.** *Let P and Q be two point sets. If we project both P and Q along a unit vector w to obtain P' and Q', then P' and Q' are not strictly linearly separable iff there exists a line ℓ parallel to w that intersects both $CH(P)$ and $CH(Q)$. If ℓ intersects the interior of $CH(P)$ or $CH(Q)$, then P' and Q' are not linearly separable.*

Proof. Assume that the line ℓ exists, and it contains $x_P \in CH(P)$ and $x_Q \in CH(Q)$ (see Figure 2). By construction, $x' = x_P - (w \cdot x_P)w = x_Q - (w \cdot x_Q)w$. Hence, by Observation 2, $x' \in CH(P') \cap CH(Q')$. Thus, by Fact 1, P' and Q' are not strictly linearly separable.

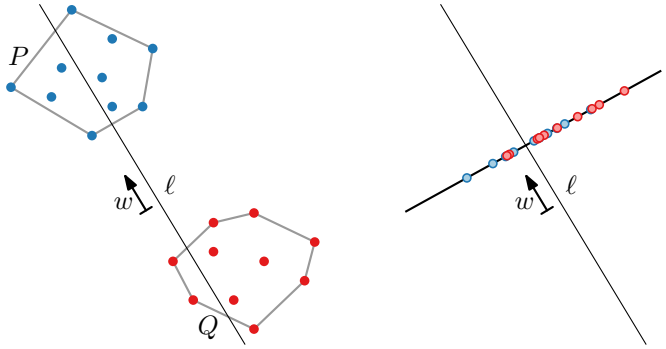

Figure 2 Line ℓ intersects $CH(P)$ and $CH(Q)$; after projection the convex hulls intersect.

For the other direction, choose $x' \in CH(P') \cap CH(Q')$. The line parallel to w and passing through x' must clearly intersect both $CH(P)$ and $CH(Q)$. The extension to (non-strict) linear separability is straightforward. ◀

Assume now that the properties $a_1, \ldots, a_k$ are strictly linearly separable in P. Can we project P along a unit vector w so that $a_2, \ldots, a_k$ are still strictly linearly separable in P', but a_1 is not? We consider two variants: (1) *separation preserving* and (2) *separability preserving* projections. The former preserves a fixed set of separating hyperplanes $H_2, \ldots, H_k$ for properties $a_2, \ldots, a_k$, the latter preserves only linear separability of $a_2, \ldots, a_k$.

Lemma 4 proves there exist point sets using only $2^k - 1$ possible labels for which every separability preserving projection also keeps a_1 strictly linearly separable after projection. The idea is to use the properties $a_2, \ldots, a_k$ to sufficiently restrict the direction of a separability preserving projection to make it impossible for this projection to eliminate the linear separability of a_1. A simple example for $d = k = 2$ is shown in Figure 3.

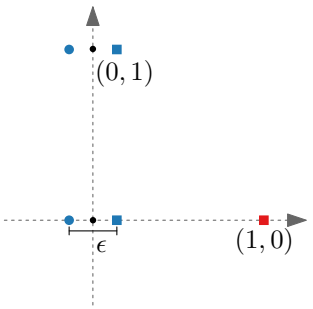

Figure 3 A point set with 5 points and 2 properties: a_1 (color) and a_2 (shape). To keep a_2 linearly separable after projection, the projection vector w should be nearly vertical, but then a_1 will also remain linearly separable.

▶ **Lemma 4.** *For all $k > 1$ and $d \geq k$, there exist point sets P in $\mathbb{R}^d$ with properties $a_1, \ldots, a_k$ using $2^k - 1$ labels such that any separability preserving projection along a unit vector w also keeps a_1 strictly linearly separable after projection.*

We now assume that all 2^k labels are used in P. Note that this assumption directly implies that $d \geq k$: take any set of k separating hyperplanes $H_1, \ldots, H_k$ for the k properties and consider the arrangement formed by the hyperplanes in $\mathbb{R}^d$. Clearly, all points in the same cell of the arrangement must have the same label. However, it is well-known that it is

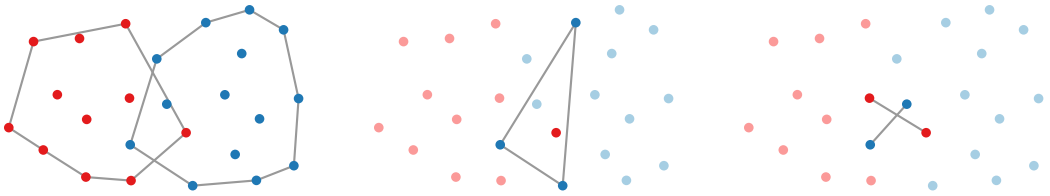

Figure 4 Theorem 5 in 2D: 4 points are needed to construct two intersecting convex hulls.

not possible to create 2^k cells in $\mathbb{R}^d$ with only k hyperplanes if $d < k$. This has also interesting implications for the case when $d = k$: if we apply a separation preserving projection to P, then a_1 cannot be linearly separable in P', since P' is embedded in $\mathbb{R}^{k-1}$.

We now show that, if $d \geq k$, then there always exists a separation preserving projection that eliminates the strict linear separability of a_1 (see Figure 4). Our proof uses Kirchberger's theorem [16]. Below we restate this theorem in our own notation. We also include our own proof, since the construction in the proof is necessary for efficient computation of our result.

▶ **Theorem 5** ([16]). *Let P and Q be two points sets in $\mathbb{R}^d$ such that $CH(P) \cap CH(Q) \neq \emptyset$. Then there exist subsets $P^* \subseteq P$ and $Q^* \subseteq Q$ such that $CH(P^*) \cap CH(Q^*) \neq \emptyset$ and $|P^*| + |Q^*| = d + 2$.*

Proof. Let $|P| = n$ and $|Q| = m$. We show that, if $n + m \geq d + 3$, then we can remove one of the points from either P or Q. Pick a point $x \in CH(P) \cap CH(Q)$. By definition, we can find coefficients $\lambda_1, \ldots, \lambda_n \geq 0$ and $\mu_1, \ldots, \mu_m \geq 0$ such that $\sum_{i=1}^{n} \lambda_i p_i = x = \sum_{j=1}^{m} \mu_j q_j$, $\sum_{i=1}^{n} \lambda_i = 1$, and $\sum_{j=1}^{m} \mu_j = 1$. If any of these coefficients is zero, then we can remove the corresponding point whilst keeping x in the intersection of the two convex hulls. Otherwise, we find nonzero coefficients $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$ such that $\sum_{i=1}^{n} a_i p_i = \sum_{j=1}^{m} b_j q_j$, $\sum_{i=1}^{n} a_i = 0$, and $\sum_{j=1}^{m} b_j = 0$. As this is a linear system with $d + 2$ constraints and $n + m \geq d + 3$ variables, there must exist a set of nonzero coefficients that satisfy these constraints. Let $\rho_\lambda = \min\{\lambda_i / a_i \mid a_i > 0\}$, $\rho_\mu = \min\{\mu_j / b_j \mid b_j > 0\}$, and $\rho = \min(\rho_\lambda, \rho_\mu)$. Now consider the new coefficients $\lambda'_i = \lambda_i - \rho a_i$ and $\mu'_j = \mu_j - \rho b_j$. By construction we have that $\lambda'_i \geq 0$ for $1 \leq i \leq n$, $\mu'_j \geq 0$ for $1 \leq j \leq m$, $\sum_i \lambda'_i = \sum_j \mu'_j = 1$, and $\sum_i \lambda'_i p_i = \sum_j \mu'_j q_j = x'$. Additionally, one of the new coefficients is zero, and we can remove the corresponding point. We can repeat this process until $n + m = d + 2$. ◀

The following proof constructs a suitable projection vector using four main steps:

1. We project the points orthogonally onto the linear subspace A spanned by the normals of the separating hyperplanes $H_2, \ldots, H_k$.
2. We argue that, since P uses all 2^k labels, a_1 is not linearly separable in A.
3. We find a small subset of points P^* for which a_1 is not linearly separable in A.
4. We construct a separation preserving projection that maps all points in P^* to an affine transformation of A. As a result, a_1 is not strictly linearly separable after projection.

We assume that the points in P, along with the chosen separating hyperplanes, are in *general position*. Specifically, we assume that any set of d vectors, where each vector is either a distinct difference vector of two points in P or the normal vector of one of the separating hyperplanes, is linearly independent. Note that, since all properties are initially strictly linearly separable, it is always possible to perturb the separating hyperplanes to ensure general position, assuming that P is also in general position.

▶ **Theorem 6.** *If P is a point set in $\mathbb{R}^d$ in general position with $k \leq d$ strictly linearly separable properties $a_1, \ldots, a_k$ using all 2^k labels, then there exists a separation preserving projection along a unit vector w that eliminates the strict linear separability of a_1.*

Proof. We provide an explicit construction of the vector w. Let $H_2, \ldots, H_k$ be any separating hyperplanes (in general position with P) for each of the properties $a_2, \ldots, a_k$ in P, respectively. Let v_i be the normal of hyperplane H_i for $2 \leq i \leq k$, and let $A \subset \mathbb{R}^d$ be the $(k-1)$-dimensional linear subspace spanned by $v_2, \ldots, v_k$. Furthermore, let $H^* = \bigcap_{i=2}^{k} H_i$ be the $(d-k+1)$-dimensional flat that is the intersection of the separating hyperplanes. Note that a projection along a vector w is separation preserving if and only if w is parallel to H^*. Let $T(p)$ be the result of an orthogonal projection of a point $p \in P$ onto A. For ease of argument, we also directly apply an affine transformation that maps H^* (which intersects A in one point by construction) to the origin, and maps $v_2, \ldots, v_k$ to the standard basis vectors of $\mathbb{R}^{k-1}$.

Now define $Q_- = \{T(p) \mid p \in P_-\}$ and $Q_+ = \{T(p) \mid p \in P_+\}$. By construction, since all labels are used by P, both Q_- and Q_+ must have a point in each orthant of $\mathbb{R}^{k-1}$. If a point set Q has a point in each orthant, then $CH(Q)$ must contain the origin; because if it does not, then there exists a vector v such that $(v \cdot q) > 0$ for all $q \in Q$. But there must exist a point $q^* \in Q$ whose sign for each coordinate is opposite from that of v (or zero), which means that $(v \cdot q^*) \leq 0$, a contradiction. Thus, both $CH(Q_-)$ and $CH(Q_+)$ contain the origin, and $CH(Q_-) \cap CH(Q_+) \neq \emptyset$. We now apply Theorem 5 to Q_- and Q_+ to obtain Q_-^* and Q_+^* consisting of $k+1$ points in total. Let $P^* \subseteq P$ be the corresponding set of original points that map to $Q_-^* \cup Q_+^*$. We can now construct w as follows. Pick a point $p^* \in P^*$, and let F_1 be the unique $(k-1)$-dimensional flat that contains the remaining points in P^*. Let F_2 be the flat obtained by translating H^* to contain p^*. Since F_1 is $(k-1)$-dimensional and F_2 is $(d-k+1)$-dimensional, $F_1 \cap F_2$ consists of a single point $r \in \mathbb{R}^d$ (assuming general position). The desired projection vector is now simply $w = r - p^*$ (normalized if necessary).

We finally show that the constructed vector w has the correct properties. First of all, w is parallel to H^* by construction, and hence the projection along w is separation preserving. Second, since $r \in F_1$ and p^* is projected to coincide with r (as $w = r - p^*$), all points in P^* will lie on the same $(k-1)$-dimensional flat F_1' after projection. Also, since w is orthogonal to A, there exists an affine map from $Q_-^* \cup Q_+^*$ to P^* (after projection). Thus, we obtain that $CH(P_-') \cap CH(P_+') \neq \emptyset$; in particular, the convex hulls must intersect on F_1'. By Fact 1 this implies that a_1 is not strictly linearly separable after projection. ◀

The result of Theorem 6 has one shortcoming: the resulting projected point set P' is degenerate by construction and property a_1 may still be (non-strictly) linearly separable after projection. This is simply an artifact of the proof and can be avoided by slightly perturbing the projection vector w. The following lemma can be used to remedy this shortcoming. Here we again assume that, before projection, the point set P and the separating hyperplanes are in general position, and hence the only degeneracy in P' is the one introduced by construction.

▶ **Lemma 7.** *Let P and Q be two point sets in $\mathbb{R}^d$ in general position and let P' and Q' be the point sets obtained by projecting P and Q along a vector w, respectively. If $CH(P') \cap CH(Q') \neq \emptyset$, then we can perturb w to obtain projections P'' and Q'' such that P'' and Q'' are not linearly separable and $P'' \cup Q''$ is in general position.*

Proof. Let $P = \{p_1, \ldots, p_n\}$ and $Q = \{q_1, \ldots, q_m\}$, and similarly $P' = \{p_1', \ldots, p_n'\}$ and $Q' = \{q_1', \ldots, q_m'\}$. We may assume that $m + n \geq d + 1$, for otherwise P and Q do not really span $\mathbb{R}^d$. Since $CH(P') \cap CH(Q') \neq \emptyset$, there exist coefficients $\lambda_i \geq 0$ $(1 \leq i \leq n)$ and $\mu_j \geq 0$ $(1 \leq j \leq m)$ such that $\sum_i \lambda_i = 1$, $\sum_j \mu_j = 1$, and $\sum_i \lambda_i p_i' = \sum_j \mu_j q_j'$. We can ignore some

points with a zero coefficient so that we have exactly $d+1$ points left, and we assume in the remainder of this proof that $m+n=d+1$. Now assume w.l.o.g. that $\lambda_1 > 0$. We use the remaining points $(P' \cup Q') \setminus \{p'_1\}$ to set up a barycentric coordinate system for the points in $P' \cup Q'$. This has the advantage that only the coordinates of p_1 are affected when changing the projection vector w. Next, we slightly perturb the coefficients to obtain $\lambda'_i > 0$ $(1 \leq i \leq n)$, $\mu'_j > 0$ $(1 \leq j \leq m)$ with $\lambda'_1 = \lambda_1$, $\sum_i \lambda'_i = 1$ and $\sum_j \mu'_j = 1$ (this is clearly possible). There then exist a vector v (in barycentric coordinates) and $\epsilon > 0$ (ϵ can be arbitrarily small by scaling the perturbation of the coefficients) such that $\epsilon v + \sum_i \lambda'_i p'_i = \sum_j \mu'_j q'_j$. Now consider the point $p_1^{\perp}$ which has the same barycentric coordinates as p'_1, but then with the barycentric coordinate system defined by $(P \cup Q) \setminus \{p_1\}$. Then, by Observation 2, we must have that $p_1 - p_1^{\perp} = \alpha w$ for some constant $\alpha \neq 0$. Now we perturb $p_1^{\perp}$ to p^* such that p^* has the same barycentric coordinates as $p'_1 + (\epsilon/\lambda_1)v$, but then again with the barycentric coordinate system defined by $(P \cup Q) \setminus \{p_1\}$. Additionally, we perturb w to $w' = p_1 - p^*$. Let $P'' = \{p''_1, \ldots, p''_n\}$ and $Q'' = \{q''_1, \ldots, q''_m\}$ be the point sets obtained by projecting P and Q along w'. We then have by construction that $\sum_i \lambda'_i p''_i = \sum_j \mu'_j q''_j$. Now assume for the sake of contradiction that P'' and Q'' are linearly separable by a hyperplane H. Then H must contain $CH(P'') \cap CH(Q'')$ and, consequently, all points that have a nonzero coefficient for the convex combination of a point $x \in CH(P'') \cap CH(Q'')$ (since all points of either P'' or Q'' lie on the same side of H). By construction there are $d+1$ of these points in $P'' \cup Q''$. Since H is $(d-2)$-dimensional and we performed only a single projection, this also implies that there were $d+1$ points on a $(d-1)$-dimensional hyperplane in $P \cup Q$. This contradicts the assumption that $P \cup Q$ is in general position. Finally, since $CH(P'')$ and $CH(Q'')$ are not interior disjoint by Fact 1, this property cannot be broken by slightly perturbing the projection vector w'. Thus, we can also ensure that $P'' \cup Q''$ is in general position. ◀

Computation. The proof of Theorem 6 is constructive and implies an efficient algorithm to compute the desired projection. Most steps in the construction involve simple linear algebra operations, like projections and intersecting flats (Gaussian elimination), which can easily be computed in polynomial time. The only nontrivial computational step is the application of Theorem 5, for which the proof is also constructive. If a point $x \in CH(P) \cap CH(Q)$ is given along with the coefficients for the convex combination, then we can simply obtain P^* and Q^* by repeatedly solving a linear system of equations and eliminating a point. Note that the linear system needs to involve only $d+3$ points (arbitrarily chosen), so the linear system of equations can be solved in $O(d^3)$ time, and we can eliminate a point and update the coefficients in the same amount of time. Thus, we can compute P^* and Q^* in $O(nd^3)$ time, where $n = |P| + |Q|$ (similar arguments were used in [20]). If we are not given a point in $x \in CH(P) \cap CH(Q)$ along with the coefficients for the convex combination, then this must be computed first. This can be computed efficiently using linear programming.

The proof of Theorem 6 suggests how to check, if P does not use all 2^k labels, if there exists a separability preserving projection that eliminates the linear separability of a_1: If we can find a set of separating hyperplanes $H_2, \ldots, H_k$ such that $CH(P_-)$ and $CH(P_+)$ intersect *after* projecting them orthogonally onto the space spanned by the normals of $H_2, \ldots, H_k$, then the remainder of the proof holds. However, finding such suitable separating hyperplanes might be computationally hard in general.

3 Generalized separability

In this section we consider a generalization of linear separability for classification. One approach to achieve more complicated classification boundaries is to use clustering: the label of a point is determined by the label of the "nearest" cluster. If we use more than one

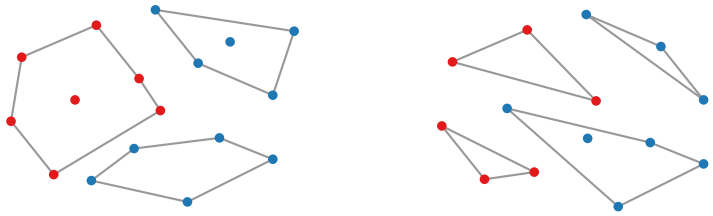

Figure 5 Left: two point sets P (red) and Q (blue) that are $(1,2)$-separable, but not linearly separable. Right: two point sets that are $(2,2)$-separable, but not $(1,x)$-separable for any value of x.

cluster per class, then the resulting classification is more expressive than classification by linear separation. This approach is also strongly related to nearest-neighbor classification, another common machine learning technique: the points decompose the space into convex subsets, each of which is associated with exactly one point; given enough clusters, we can thus exactly capture this behavior. But even with few clusters (convex sets), it may be possible to reasonably approximate the decomposition by using a single cluster to capture the same of many points with the same label. Hence, Our generalized definition of separability is inspired by such clustering-based classifications, with convex sets modeling the clusters.

Let P and Q be two point sets in $\mathbb{R}^d$. We say that P and Q are (b,c)-*separable* if there exist b convex sets $S_1,\ldots,S_b$ and c convex sets $T_1,\ldots,T_c$ such that for every point $p \in P$ we have that $p \in S = \bigcup_i S_i$, for every point $q \in Q$ we have that $q \in T = \bigcup_j T_j$, and that $S \cap T = \emptyset$ (see Figure 5). We can assume that $b \leq c$. Furthermore, we generally assume w.l.o.g. that any convex set S_i is the convex hull of its contained points. It is easy to see that linear separability and $(1,1)$-separability are equivalent.

Given a point set P along with k properties $a_1,\ldots,a_k$, the goal is now to compute a separation preserving projection to a point set P' such that a_1 is not (b,c)-separable in P'. We again assume that all k properties are strictly linearly separable in P. To achieve this goal, we may need to project along multiple vectors $w_1,\ldots,w_r$. As mentioned in Section 1, we assume that $\{w_j\}_{j=1}^r$ form an orthonormal system and that we can compute the projected points as $p_i' = p_i - \sum_{j=1}^r (w_j \cdot p_i) w_j$.

To extend Theorem 6 to (b,c)-separability, recall the four main steps of the proof described before Theorem 6. Step 3 is the most important. If a_1 was not linearly separable in A, then not even multiple separation preserving projections can eliminate the linear separability of a_1. In that sense, A is the "worst we can do" with separation preserving projections. Step 3 is actually exploiting a Helly-type property [24] for linear separability: If two sets of points P and Q are not linearly separable, then there exist small subsets $P^* \subseteq P$ and $Q^* \subseteq Q$ such that P^* and Q^* are not linearly separable (Theorem 5). Hence, if we use a different type of separability that also has a Helly-type property, then we may be able to use the same approach as for linear separability. Generally speaking, let $F(P,Q)$ be a predicate that determines if point sets $P,Q \subset \mathbb{R}^d$ are "separable" (for some arbitrary definition of separable)[2]. If, in the case that $F(P,Q)$ does not hold, there exist small (bounded by a constant) subsets $P^* \subseteq P$ and $Q^* \subseteq Q$ such that $F(P^*,Q^*)$ also does not hold, then F has the *Helly-type property*. The worst-case size of $|P^*| + |Q^*|$ often depends on the number of dimensions d of P and Q, and is referred to as the *Helly number* $m_F(d)$ of F. For technical reasons, we will require the following three natural conditions on F:

[2] We assume that F is defined independently from the dimensionality of P and Q (like (b,c)-separability). We do require that P and Q are embedded in the same space.

1. If $F(P,Q)$ does not hold, then $F(P',Q')$ does not hold, where P' and Q' are obtained by projecting P and Q along a single unit vector, respectively.
2. If $P' \subseteq P$ and $Q' \subseteq Q$, then $F(P,Q)$ implies $F(P',Q')$.
3. If $\mathcal{A}$ is an affine map, then $F(P,Q)$ holds if and only if $F(\mathcal{A}(P),\mathcal{A}(Q))$ holds.

We call a separation predicate F *well-behaved* if it satisfies these conditions. It is easy to see that (b,c)-separability is well-behaved. For Condition 1, note that any collection of convex sets for P' and Q' can easily be extended along the projection vector for P and Q without introducing an overlap between S and T. Condition 2 also holds, since we can simply use the same covering sets. Finally, Condition 3 holds since affine transformations preserve convexity. We summarize this generalization in the following generic theorem.

▶ **Theorem 8.** *Let P be a point set in $\mathbb{R}^d$ with k ($d \geq k$) properties $a_1,\ldots,a_k$ and let F be a well-behaved separation predicate in $\mathbb{R}^d$. Either we can use at most $\min(m_F(k-1)-k, d-k+1)$ separation preserving projections to eliminate $F(P_-,P_+)$, or this cannot be achieved with any number of separation preserving projections.*

Proof. Following the proof of Theorem 6, we first orthogonally project the points in P onto the $(k-1)$-dimensional linear subspace A that is spanned by the normals $v_2,\ldots,v_k$ of the separating hyperplanes $H_2,\ldots,H_k$ of the properties $a_2,\ldots,a_k$. Let $T(p)$ be the resulting projected point for a point $p \in P$. Now define $Q_- = \{T(p) \mid p \in P_-\}$ and $Q_+ = \{T(p) \mid p \in P_+\}$. If $F(Q_-,Q_+)$ holds, then no sequence of separation preserving projections can eliminate the separability (as defined by F) of a_1, due to Condition 1 of a well-behaved separation predicate. Otherwise, we can find $Q_-^* \subseteq Q_-$ and $Q_+^* \subseteq Q_+$ such that $F(Q_-^*,Q_+^*)$ does not hold, and $|Q_-^*|+|Q_+^*| \leq m_F(k-1)$. Let $P^* \subseteq P$ be the set of original points that map to $Q_-^* \cup Q_+^*$. The points in P^* span a linear subspace B. Next, we construct an orthonormal basis $\{w_j\}_{j=1}^r$ for the set of vectors in B that are orthogonal to A (orthogonal to $v_2,\ldots,v_k$). Since B has at most $m_F(k-1)-1$ dimensions, and A has $k-1$ dimensions, we conclude that the orthonormal basis contains $r \leq m_F(k-1)-1-(k-1) = m_F(k-1)-k$ vectors. We then choose to project P along the vectors $w_1,\ldots,w_r$. Since every w_j for $1 \leq j \leq r$ is orthogonal to A, these projections are all separation preserving. Furthermore, since we eliminate all vectors orthogonal to A from B, there exists an affine map from $Q_-^* \cup Q_+^*$ to P^* after projection. By using Condition 2 and Condition 3 of a well-behaved separation predicate, we can then conclude that $F(P'_-,P'_+)$ does not hold. Alternatively, we can simply project P to A, which requires $d-k+1$ separation preserving projections. Hence, we need at most $\min(m_F(k-1)-k, d-k+1)$ projections. ◀

We now focus on (b,c)-separability for different values of b and c. Unfortunately, not every form of (b,c)-separability has the Helly-type property.

▶ **Lemma 9.** *In $d \geq 2$ dimensions, $(1,2)$-separability does not have the Helly-type property.*

Proof. We prove the statement for $d=2$, which automatically implies it for $d>2$. Consider a set of n points $P = \{p_1,\ldots,p_n\}$ equally spaced on the unit circle, where n is odd. For every point p_i we can define a wedge W_i formed between the rays from p_i to the two opposite points on the circle (which are well defined, since n is odd). By the Central Angle Theorem, the angle of this wedge is $\frac{\pi}{n}$. Furthermore, the distance of the rays to the origin is exactly $\sin\left(\frac{\pi}{n}\right)$. Now, for some $\epsilon > 0$ and for each point p_i, we add a point q_i on the circle centered at the origin with radius $\sin\left(\frac{\pi}{n}\right)+\epsilon$, such that q_i lies outside of W_i to the left (counterclockwise). By construction there will also be a point q_j to the right of W_i, added by the point p_j that is the opposite point of p_i on the right (clockwise) side. We choose ϵ small enough such that any wedge W_i contains exactly $n-2$ points from $Q = \{q_1,\ldots,q_n\}$, having one point of Q outside of W_i on each side (see Figure 6).

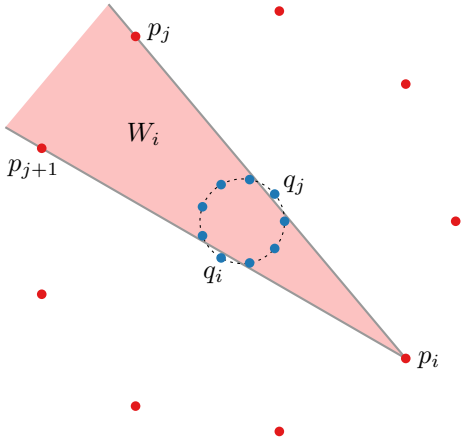

Figure 6 The construction for Lemma 9 with P in red and Q in blue.

Assume for the sake of contradiction that P and Q are $(1,2)$-separable. Since $Q \subset CH(P)$, we must cover Q with one set, and hence $S_1 = CH(Q)$. Now consider $P_1 = T_1 \cap P$ and $P_2 = T_2 \cap P$. Since the line segments between a point $p_i \in P_1$ and its opposite points p_j and p_{j+1} intersect $CH(Q)$, we get that p_j and p_{j+1} must both be in P_2. We can repeat this argument for all points p_i to conclude that all pairs of consecutive points of P must be in the same set (P_1 or P_2). Since not all points in P can belong to the same set ($Q \subset CH(P)$), we obtain a contradiction. Thus, P and Q are not $(1,2)$-separable.

Now consider removing a single point p_i from P, and consider the line ℓ through the origin and p_i. The line ℓ splits $P \setminus \{p_i\}$ into two sets P_1 and P_2. It is easy to see that, if we pick ϵ small enough, $CH(P_1)$ and $CH(P_2)$ do not intersect $CH(Q)$. Hence, $P \setminus \{p_i\}$ and Q are $(1,2)$-separable. If we remove a single point q_i from Q, then the line segment between p_i and one of its opposite points p_j does not intersect $CH(Q \setminus \{q_i\})$. We can again split P into P_1 and P_2 using the line ℓ through p_i and p_j (and shifted slightly towards the origin). Then it is again easy to see that, if we pick ϵ small enough, $CH(P_1)$ and $CH(P_2)$ do not intersect $CH(Q \setminus \{q_i\})$. Hence, P and $Q \setminus \{q_i\}$ are $(1,2)$-separable.

As a result, there exist no subsets of P and Q that are not $(1,2)$-separable. Thus, we get that the Helly number for $(1,2)$-separability is at least $|P| + |Q| = 2n$, and hence $(1,2)$-separability does not have the Helly-type property. ◀

Hence we cannot apply Theorem 8 to eliminate $(1,2)$-separability of a_1 in few separation preserving projections, if possible at all. However, this does not mean that it is not possible to provide this guarantee using different arguments. Nonetheless, we can use a similar construction as in the proof of Lemma 9 (using many more dimensions) to show that many separation preserving projections are needed to eliminate $(1,2)$-separability for a_1 (as many projections as needed to reach the 2-dimensional construction in the proof of Lemma 9).

Next, we consider $(1,\infty)$-separability. This means that one of the point sets, say P, must be covered with one convex set, but we can use arbitrarily many convex sets to cover Q. Equivalently, P and Q are $(1,\infty)$-separable if $CH(P) \cap Q = \emptyset$ or $P \cap CH(Q) = \emptyset$.

▶ **Lemma 10.** *In $d \geq 1$ dimensions, $(1,\infty)$-separability has the Helly-type property with Helly number $2d+2$.*

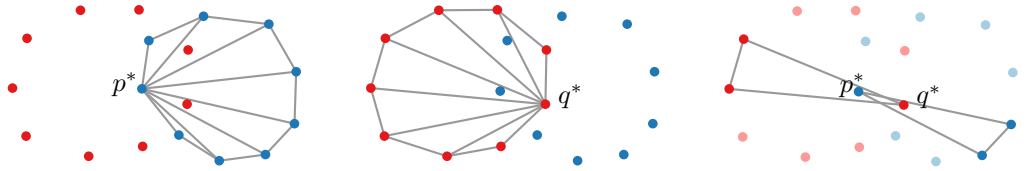

Figure 7 Lemma 10: constructing a small point set that is not $(1, \infty)$-separable.

Proof. Let P and Q be point sets in $\mathbb{R}^d$ such that P and Q are not $(1, \infty)$-separable. Then there must be a point $p^* \in CH(Q)$ and a point $q^* \in CH(P)$. We can construct a star triangulation $\mathcal{T}(P)$ of $CH(P)$ with p^* as center (that is, all d-dimensional simplices have p^* as a vertex) and a star triangulation $\mathcal{T}(Q)$ of $CH(Q)$ with q^* as center (see Figure 7). We identify the unique simplex $\sigma_P \in \mathcal{T}(P)$ that contains q^*, and similarly the unique simplex $\sigma_Q \in \mathcal{T}(Q)$ that contains p^*. Now let $P^* \subseteq P$ be the vertices of σ_P and let $Q^* \subseteq Q$ be the vertices of σ_Q. Note that $p^* \in P^*$ and $q^* \in Q^*$. Then P^* and Q^* are not $(1, \infty)$-separable, since $q^* \in CH(P^*) \cap Q^*$ and $p^* \in CH(Q^*) \cap P^*$. Finally, since a d-dimensional simplex contains $d + 1$ vertices, we obtain Helly number $2d + 2$. ◀

▶ **Corollary 11.** *Let P be a point set in $\mathbb{R}^d$ with k $(d \geq k)$ properties $a_1, \ldots, a_k$. Either we can use at most $\min(k, d-k+1)$ separation preserving projections to eliminate $(1, \infty)$-separability of a_1, or this cannot be achieved with any number of separation preserving projections.*

It may initially seem counter-intuitive that $(1, \infty)$-separability has the Helly-type property (requiring only few projections to eliminate $(1, \infty)$-separability), while the strictly stronger $(1, 2)$-separability does not have the Helly-type property (and may require many projections to eliminate $(1, 2)$-separability). Note however that Theorem 8 includes the clause that it simply may not be possible to eliminate separability of a_1 via any number of separation preserving projections. This case occurs more often with $(1, \infty)$-separability than with $(1, 2)$-separability, which explains why we can provide better guarantees on the number of projections for a strictly weaker separability condition.

We finally briefly consider $(2, \infty)$-separability in $\mathbb{R}^2$. Two point sets P and Q are not $(2, \infty)$-separable in $\mathbb{R}^2$ if we need at least three convex sets disjoint from Q to cover P (and vice versa). This implies that $CH(P)$ must contain at least 3 points of Q; if not, then we can draw a single line through all points in $Q \cap CH(P)$ to separate P into P_1 and P_2, and $CH(P_1)$ and $CH(P_2)$ both cover P and are disjoint from Q. More generally, assume that we can cover P with two sets $CH(P_1)$ and $CH(P_2)$ that are disjoint from Q, and let ℓ be a line that separates $CH(P_1)$ and $CH(P_2)$ (Fact 1). Now consider the set of all triangles $\mathcal{T}_P$ that are formed by three points of P such that a point of Q is contained in the triangle. We must have that ℓ transverses (intersects) all triangles in $\mathcal{T}_P$, otherwise the triangle is contained in P_1 or P_2, and hence there is a point of Q in either $CH(P_1)$ or $CH(P_2)$. Furthermore, if there is point $q \in Q$ contained in, say, $CH(P_1)$, then there is also a triangle $\Delta \in \mathcal{T}_P$ in P_1 (Carathéodory's theorem), and hence ℓ does not intersect all triangles in $\mathcal{T}_P$. Thus, P and Q are $(2, \infty)$-separable (assuming we cover P with 2 convex sets) if and only if there exists a line ℓ that transverses $\mathcal{T}_P$. As a result, if we can show a Helly-type property for line transversals of triangles, then we also obtain a Helly-type property for $(2, \infty)$-separability. Unfortunately, there is no Helly-type property for line transversals of general sets of triangles [17]. We leave it as an open question to determine if there exists a Helly-type property for line transversals of these special sets of triangles $\mathcal{T}_P$.

4 Conclusion

We studied the use of projections for obstructing classification of high-dimensional Euclidean point data. Our results show that, if not all possible labels are present in the data, then it may not be possible to eliminate the linear separability of one property while preserving it for the other properties. This is not surprising if a property that we aim to keep is strongly correlated with the property we aim to hide. Nonetheless, one should be aware of this effect when employing projections in practice. When going beyond linear separability, we see that the number of projections required to hide a property increases significantly in theory, and we expect a similar effect when using, for example, neural networks for classification in practice. In other words, projecting a dataset once (or few times) may not be sufficient to hide a property from a smart classifier. Projection, as a linear transformation, can however be effective in eliminating certain linear relations in the data.

One potential direction of future work is to consider other separability predicates for labeled point sets, beyond linear separability and (b, c)-separability. Are there other types of separability that also have the Helly-type property used in Theorem 8? Or is there another way to show that few projections suffice to eliminate the separability of one of the properties? There are many other types of separability (for example, via boxes or spheres) for which this can be evaluated.

In this paper we focused on eliminating bias based on one property (such as gender). Intersectionality posits that discrimination due to multiple properties should be considered in a holistic manner, instead of one property at a time. In fact, any one property might not be a cause for discrimination, but their combination is. The following challenge arises: say we used projection successfully to eliminate the linear separability of gender. However, if we now restrict the data to one particular sub-class, for example black people, then the linear separability of gender might still be preserved within this subclass and hence discrimination against black women can still be possible. Under which conditions is it possible to eliminate the linear separability of one property not only in the full data, but also in specific (or all) subclasses? We leave this question as an open problem.

References

1 Mohsen Abbasi, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. Fairness in representation: quantifying stereotyping as a representational harm. In *Proc. SIAM International Conference on Data Mining*, pages 801–809, 2019. doi:10.1137/1.9781611975673.90.

2 Alexander Amini, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019. doi:10.1145/3306618.3314243.

3 Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv:1706.02409*, 2017.

4 Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

5 Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, pages 4349–4357, 2016.

6 Shikha Bordia and Samuel R. Bowman. Identifying and reducing gender bias in word-level language models. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 7–15, 2019. doi:10.18653/v1/n19-3002.

7 Marc E. Brunet, Colleen A. Houlihan, Ashton Anderson, and Richard S. Zemel. Understanding the origins of bias in word embeddings. In *Proc. 36th International Conference on Machine Learning*, volume 97, pages 803–811, 2019.

8 Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. doi:10.1126/science.aal4230.

9 Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. On measuring and mitigating biased inferences of word embeddings. In *Proc. AAAI Conference on Artificial Intelligence*, pages 7659–7666, 2020.

10 Sunipa Dev and Jeff Phillips. Attenuating bias in word vectors. In *Proc. Machine Learning Research*, pages 879–887, 2019.

11 Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv*, 2016. arXiv:1511.05897.

12 Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proc. 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 259–268, 2015.

13 Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, pages 3315–3323, 2016.

14 Yuzi He, Keith Burghardt, and Kristina Lerman. A geometric solution to fair representations. In *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, page 279–285, 2020. doi:10.1145/3375627.3375864.

15 Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1–33, 2011. doi:10.1007/s10115-011-0463-8.

16 Paul Kirchberger. Über Tchebychefsche Annäherungsmethoden. *Mathematische Annalen*, 57:509–540, 1903. doi:10.1007/BF01445182.

17 Ted Lewis. Two counterexamples concerning transversals for convex subsets of the plane. *Geometriae Dedicata*, 9:461–465, 1980. doi:10.1007/BF00181561.

18 David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proc. 35th International Conference on Machine Learning*, pages 3384–3393, 2018.

19 Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv*, 2019. arXiv:1908.09635.

20 Frédéric Meunier, Wolfgang Mulzer, Pauline Sarrabezolles, and Yannik Stein. The rainbow at the end of the line - A PPAD formulation of the colorful Carathéodory theorem with applications. In *Proc. 28th ACM-SIAM Symposium on Discrete Algorithms*, pages 1342–1351, 2017. doi:10.1137/1.9781611974782.87.

21 Yingjie Tian Naiyang Deng and Chunhua Zhang. *Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions*. CRC Press, 2012.

22 Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, 2020.

23 Jennifer L. Skeem and Christopher T. Lowenkamp. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54(4):680–712, 2016. doi:10.1111/1745-9125.12123.

24 Rephael Wenger. Helly-type theorems and geometric transversals. In Jacob E. Goodman and Joseph O'Rourke, editors, *Handbook of Discrete and Computational Geometry, second edition*, pages 73–96. Chapman and Hall/CRC, 2004. doi:10.1201/9781420035315.ch4.

25 Muhammad Bilal Zafar, Isabel Valera, Manuel G. Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proc. 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 962–970, 2017.

26 Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proc. 30th International Conference on Machine Learning*, pages 325–333, 2013.

27 Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. doi:10.1145/3278721.3278779.

28 Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai Wei Chang. Learning gender-neutral word embeddings. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, 2018.

A Maximizing inseparability

In this section we consider the problem of not only eliminating the linear separability of a_1, but additionally to maximize the "linear inseparability" (or overlap) of a_1 after projection. For that we need to define the overlap between two point sets P and Q. For a unit vector v, consider the intervals $I_P(v) = CH(\{v \cdot p_i \mid p_i \in P\}$ and $I_Q(v) = CH(\{v \cdot q_i \mid q_i \in Q\}$. We can then define the overlap between P and Q along v as the length of $I_P(v) \cap I_Q(v)$. Alternatively, we can define the overlap along v with the cost function used by soft-margin SVMs, which is designed for data that is not linearly separable (see [4] for more details). The overlap between two point sets P and Q is then defined as the minimum overlap over all (unit) vectors v. More precisely, for a given (projected) point set P, along with (implicit) property a_1, we use the function $g(P, v)$ to describe the overlap of a_1 along the vector v, and we refer to g as the *overlap function*. The overlap of a_1 is then defined as $\min_v g(P, v)$. Our goal is to find the projection that maximizes this overlap after projection. More precisely, if we use $P' = \pi_w(P)$ to denote the projection of a point set P along the unit vector w, then the goal is to maximize the function $f(P, w) = \min_v g(\pi_w(P), v)$ over all separability/separation preserving projection vectors w. We consider the following two overlap functions $g(P, v)$ (although other options are possible):

Interval. For a point set P and unit vector v, let $I_- = CH(\{v \cdot p_i \mid p_i \in P_-\})$ and $I_+ = CH(\{v \cdot p_i \mid p_i \in P_+\})$. Then $g_{\text{int}}(P, v) = |I_- \cap I_+|$.

SVM. The goal of the soft-margin SVM optimization is to minimize $g_{\text{svm}}(P, v) = \lambda\|v\|^2 + \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - a_1(p_i)(v \cdot p_i - b))$. Note that g_{svm} also requires a parameter $b \in \mathbb{R}$, but we will often omit that dependence (we can assume that the overlap function minimizes over all $b \in \mathbb{R}$). Furthermore, $\lambda > 0$ is a parameter that can be set for g_{svm}. Finally, note that v does not need to be a unit vector.

We first consider the variant of the problem that aims to find the optimal separability preserving projection. The vector v that minimizes $g(P, v)$ for a given point set is typically computed using convex programming (in particular for SVMs, see [21]). Note that convex programming heavily relies on the fact that there exists only one local optimum (which hence must be the global optimum). We show that, for the problem of finding the optimal separability preserving projection, there may be multiple local optima for $f(P, w)$. This eliminates the hope of finding a convex programming formulation for this problem.

▶ **Theorem 12.** *There exists a point set P in $\mathbb{R}^3$ with 2 properties a_1, a_2 such that $f(P, w)$ with $g = g_{svm}$ has two local maxima when restricted to all separability preserving projection vectors w.*

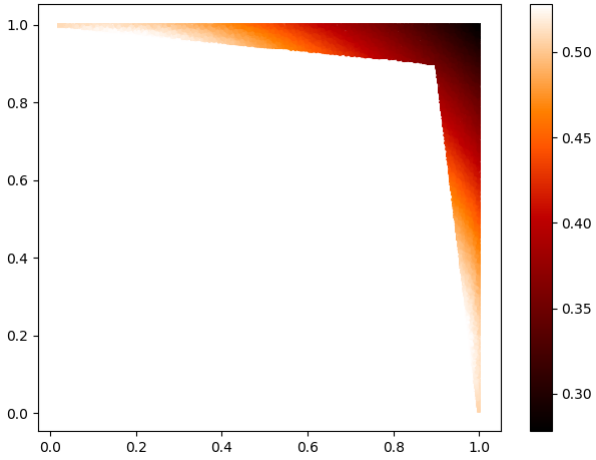

Figure 8 Illustration for Theorem 12: The domain for projections (x, y) with $\epsilon = 0.2$. Higher values in the overlap function are indicated with lighter colors. We can see two distinct local maxima.

Proof. The set P mostly consists of the vertices of a unit cube with side lengths 2 centered at the origin. We also add an extra point $p^* = (1-\epsilon, 1-\epsilon, 1)$ for some $\epsilon > 0$ (a point slightly moved inward from the point $p_8 = (1, 1, 1)$). Thus, P consists of nine points $\{p_1, \ldots, p_8, p^*\}$. For property a_1 we choose that $a_1(p) = z(p)$ for all $p = (x(p), y(p), z(p)) \in P$. For property a_2 we have that $a_2(p) = a_1(p)$ for all $p \in P \setminus \{p_8\}$, and $a_2(p_8) = -1$. Now we limit and encode the space of possible projection vectors w to $\mathbb{R}^2$ as follows. For $w = (x(w), y(w), z(w))$ with $z(w) = 0$ it is clear that a projection along w will keep a_1 linearly separated, so we may encode all possible projections as $(x, y) = (x(w)/z(w), y(w)/z(w))$. Now consider the effect of using a projection (x, y) on P: we may assume that the x- and y-coordinates of points $p \in P$ with $z(p) = 1$ do not change and that for the other points we obtain a shifted square: $(x(p'), y(p')) = (x(p) + 2x, y(p) + 2y)$ for all $p \in P$ with $z(p) = -1$. Let $p_1 = (-1, -1, -1)$ such that $p'_1 = (-1 + 2x, -1 + 2y)$. Furthermore, let A consist of the projections of all points $p \in P$ with $a_1(p) = 1$, and let B consist of the projections of all points $p \in P$ with $a_2(p) = 1$. Note that A forms a square and B forms a square with one of the corners pushed inwards. By Fact 1, a projection (x, y) can only be separability preserving if $p'_1 \notin CH(B)$. By the same observation, a projection (x, y) with $x \geq 0$ and $y \geq 0$ preserves the linear separability of a_1 if $p'_1 \notin CH(A)$. Thus, we require that $p'_1 = (-1 + 2x, -1 + 2y) \in CH(A) \setminus CH(B)$. Note that $CH(A) \setminus CH(B)$ is a thin and nonconvex shape. The same thus holds for the domain of the projections (x, y) as shown in Figure 8, and hence the optimization problem is not convex. Furthermore, by evaluating the overlap function g_{svm} (using $\lambda = 10$) on this domain, we can see that there are two distinct local maxima: one close to $(0, 1)$ and one close to $(1, 0)$. ◀

Theorem 12 demonstrates that the constraint on projections to be separability preserving is generally not convex. We now consider the special case that we have only one property a_1 (hence no separability preserving constraint), and analyze if we can then efficiently maximize $f(P, w)$. For that, we first put a restriction on the overlap function $g(P, v)$. We say that $g(P, v)$ is *projectionable* if there exists a function $h\colon \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ such that $g(P, v) = h(v \cdot P, v)$, where $v \cdot P = \{v \cdot p_i \mid p_i \in P\}$. In other words, g should only depend on P via the dot products of points in P with v. Note that both the Interval and SVM overlap functions are indeed projectionable. For projectionable overlap functions g we can redefine the optimization function f. In the following, let $v \perp w$ indicate that v and w are orthogonal.

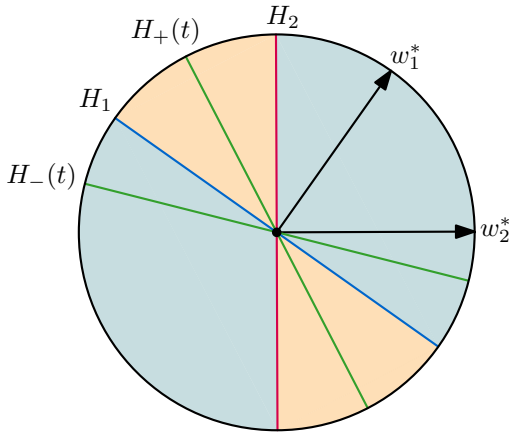

Figure 9 Illustration for Theorem 14 with $d = 2$: $H_+(t)$ can only intersect $R_{00} \cup R_{11}$ (orange, clipped to the unit disk) and $H_-(t)$ can only intersect $R_{01} \cup R_{10}$ (blue).

▶ **Lemma 13.** *If $g(P, v)$ is a projectionable overlap function, then $\max_w \min_v g(\pi_w(P), v) = \max_w \min_{v \perp w} g(P, v)$ for any point set $P \subset \mathbb{R}^d$.*

Proof. We will treat both w and v as a vector in $\mathbb{R}^d$. Since g is projectionable, there exists an equivalent function h that depends on v and the dot products between v and points in P. Now assume that $(v \cdot w) = 0$. Then we get

$$\begin{aligned} g(\pi_w(P), v) &= h(v \cdot \pi_w(P), v) \\ &= h(\{v \cdot (p_i - (w \cdot p_i)w) \mid p_i \in P\}, v) \\ &= h(\{v \cdot p_i \mid p_i \in P\}, v) \\ &= g(P, v). \end{aligned}$$

Since the vector v that minimizes $g(\pi_w(P), v)$ must be perpendicular to w, we obtain the desired equality. ◀

In the following we assume that the overlap function g is projectionable. Hence, by Lemma 13, we can rewrite f as $f(P, w) = \min_{v \perp w} g(P, v)$. This has the advantage that we can keep the point set P fixed while optimizing for w. We now aim to link properties of g to properties of f. As already discussed earlier, we can often find the vector v that minimizes $g(P, v)$ using convex programming. This implies that g has only one local minimum (for fixed P). We now use this fact to show that f has only one local maximum.

▶ **Theorem 14.** *If a function $g(P, v)$ has one local minimum for fixed P, then $f(P, w) = \min_{v \perp w} g(P, v)$ has one local maximum for fixed P.*

Proof. Note that $w \in \mathcal{S}^{d-1}$, where $\mathcal{S}^{d-1}$ is the unit $(d-1)$-sphere, and $f(P, w) = f(P, -w)$, so we will treat w and $-w$ as equivalent. Similarly, if the local minimum of $g(P, v)$ is at $v = v^*$, then $v = -v^*$ may also be a local minimum, and together they will be counted as a single local minimum. For the sake of contradiction, assume that $f(P, w)$ has two distinct local maxima, one at $w = w_1^*$ and one at $w = w_2^*$ (and also at $w = -w_1^*$ and $w = -w_2^*$). We do not require that w_1^* and w_2^* are strict local maxima, but we do require that there exists no path $\gamma\colon [0, 1] \to \mathcal{S}^{d-1}$ with $\gamma(0) = w_1^*$ and $\gamma(1) = \pm w_2^*$ such that $f(P, \gamma(t)) \geq \min(f(P, \gamma(0)), f(P, \gamma(1)))$ for all $0 \leq t \leq 1$. Now consider the hyperplanes $H_1 = \{v \mid (v \cdot w_1^*) = 0\}$ and $H_2 = \{v \mid (v \cdot w_2^*) = 0\}$. Furthermore, let $\gamma_+\colon [0, 1] \to \mathcal{S}^{d-1}$ denote

the shortest (hyper)spherical interpolation from w_1^* to w_2^*, and let $\gamma_- : [0,1] \to \mathcal{S}^{d-1}$ denote the shortest (hyper)spherical interpolation from w_1^* to $-w_2^*$. Note that γ_- and γ_+ are unique, since both w_1^* to w_2^* lie on a great circle on $\mathcal{S}^{d-1}$ and $w_1^* \neq -w_2^*$. The hyperplanes H_1 and H_2 split $\mathbb{R}^d$ into four parts (each hyperplane cuts $\mathbb{R}^d$ into two parts): R_{00}, R_{01}, R_{10}, and R_{11}. Now let $x^* = \min(f(P, w_1^*), f(P, w_2^*))$ and consider the sublevel set $S = \{v \mid g(P, v) < x^*\}$. By definition of f, H_1 and H_2 are disjoint from S. Now consider a vector $\gamma_+(t)$ for some $0 < t < 1$, and let $H_+(t) = \{v \mid (v \cdot \gamma_+(t)) = 0\}$ be the corresponding hyperplane. Similarly define $H_-(t)$ for $\gamma_-(t)$. It is easy to see that $H_+(t)$ intersects either $R_{01} \cup R_{10}$ or $R_{00} \cup R_{11}$, but not both, and $H_-(t)$ intersects only the other region (see Figure 9). By assumption, there exist values t_-^* and t_+^* such that $f(P, \gamma_-(t_-^*)) < x^*$ and $f(P, \gamma_+(t_+^*)) < x^*$. Thus, by the definition of f, there must be two non-opposite regions, say R_{00} and R_{01}, that contain a point in S. These points cannot be in the same connected component, as they are separated by either H_1 or H_2. Thus, S has multiple (non-opposite) connected components, and hence $g(P, v)$ must have at least two local minima. This contradicts our assumption, and hence $f(P, w)$ can have at most one local maximum. ◀

Following Theorem 14, we can use a hill-climbing approach to find the optimal projection vector w, if there is only one property a_1. This same approach can be applied to find the optimal separation preserving projection for k properties. In that case, the corresponding separating hyperplanes $H_2, \ldots, H_k$ each take away a degree of freedom, but otherwise do not bound the domain of w. More precisely, if there is only one property a_1, then $w \in \mathcal{S}^{d-1}$, where $\mathcal{S}^d$ is the unit d-sphere. If there are k properties, then $w \in \mathcal{S}^{d-1} \cap H_2 \cap \ldots \cap H_k = \mathcal{S}^{d-k}$. This reduction in dimensionality of the domain of w does not affect the proof of Theorem 14.

B Omitted proofs

▶ **Lemma 4.** *For all $k > 1$ and $d \geq k$, there exist point sets P in $\mathbb{R}^d$ with properties $a_1, \ldots, a_k$ using $2^k - 1$ labels such that any separability preserving projection along a unit vector w also keeps a_1 strictly linearly separable after projection.*

Proof. We first construct the point set P for arbitrary k and $d = k$. Consider the vertices of a $(k-1)$-dimensional hypercube C_ϵ with side length $\epsilon > 0$ centered at the origin, for which all nonzero coordinates lie in the first $k-1$ dimensions of $\mathbb{R}^d$. For each vertex p of C_ϵ, set

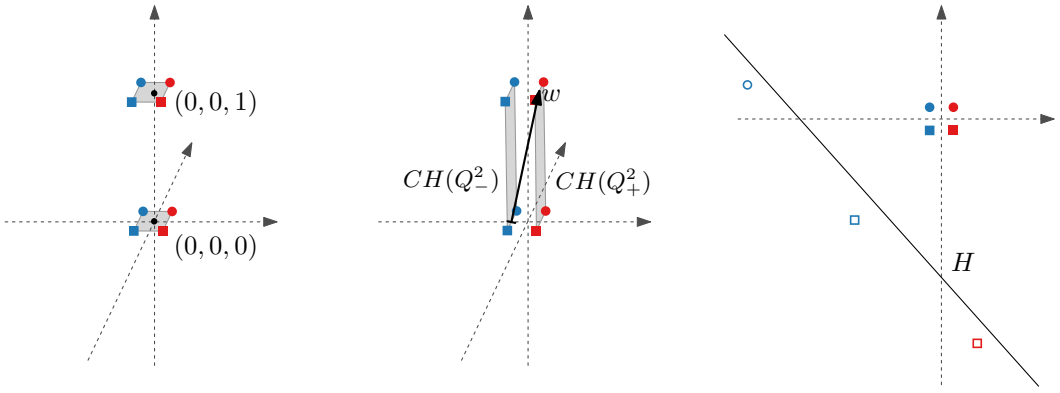

Figure 10 Illustration for Lemma 4 with $d = k = 3$ and properties fill (a_1), color (a_2), and shape (a_3). Left: Q consisting of two copies of C_ϵ. Middle: a separability preserving projection must be nearly orthogonal to the (x, y)-plane. Right: the flat H separating property a_1.

the properties of p based on its coordinates $(p^1, \ldots, p^d)$: $a_1(p) = 1$, and $a_i(p) = \text{sgn}(p^{i-1})$ for $2 \leq i \leq k$, where $\text{sgn}(x)$ is the sign function. Next, create a copy of C_ϵ (along with the assigned properties) and place it around the coordinate $(0, \ldots, 0, 1)$ (see Figure 10 left). Let the resulting point set be Q, and consider projecting Q along a unit vector w. Let $w = (w^1, \ldots, w^d)$ and assume w.l.o.g. that $|w^1| \geq |w^i|$ for all $2 \leq i < k$. If $|w^1| > \epsilon$, then there always exists a line ℓ parallel to w that intersects both $CH(Q^2_-)$ and $CH(Q^2_+)$. By Lemma 3 this would imply that a_2 is not strictly linearly separable after projection along w, so we may assume that $|w^1| \leq \epsilon$ for any separability preserving projection (see Figure 10 middle).

Now consider a $(k-2)$-dimensional flat H with the following properties: (1) it is not parallel to one of the first $k-1$ axes, (2) it lies in the first $k-1$ dimensions of $\mathbb{R}^d$ (the other coordinates are zero), and (3) the distance from the origin to H is 1 (see Figure 10 right). Consider the orthants of the $(k-1)$-dimensional subspace A spanned by the first $k-1$ axes. Based on the labels of the vertices of C_ϵ, each orthant is associated with a label for the properties $a_2, \ldots, a_k$. Due to Property (1), H intersects all the first $k-1$ axes, either at the positive or the negative half-axis. Since there is exactly one orthant bounded by only the non-intersected half-axes, H intersects exactly $2^{k-1} - 1$ orthants. We now construct P by extending Q with an additional point in each of the intersected orthants, such that H separates this point from the origin. The label of each such point p has $a_1(p) = -1$ and is otherwise determined by the orthant. As a result, P uses $2^k - 1$ different labels.

Let v be the normal of H in the $(k-1)$-dimensional subspace A. The margin for P_- and P_+ along v is at least $1 - k\epsilon$ (rough bound). For any separability preserving projection along unit vector w, we have that $|(w \cdot v)| \leq \epsilon$. Now consider any point $p \in P$ and its projection $p' = p - (w \cdot p)w$. We have that $(p' \cdot v) = (p \cdot v) - (w \cdot p)(w \cdot v) = (p \cdot v) \pm O(\epsilon)$, where we use the fact that $(w \cdot p) = O(1)$. Thus, the margin for property a_1 can be reduced by at most $O(\epsilon)$ by the projection, and hence the projection keeps a_1 strictly linearly separable if we choose ϵ small enough.

If $d > k$, then we can construct a simplex with side lengths 1 in the last $d - k + 1$ dimensions, and place a copy of C_ϵ around each of its vertices (for $d = k$ this simplex is simply an edge, as used above). With this construction we can still enforce w.l.o.g. that $|w^1| \leq \epsilon$ for any separability preserving projection along unit vector w, and the rest of the argument follows. ◀