Towards Fair Unsupervised Learning

Francois Buet-Golfouse
ucahfbu@ucl.ac.uk
University College London
London, United Kingdom

Islam Utyagulov
islam.utyagulov@gmail.com
Independent Researcher
London, United Kingdom

ABSTRACT

Bias-mitigating techniques are now well established in the supervised learning literature and have shown their ability to tackle fairness-accuracy, as well as fairness-fairness trade-offs. These are usually predicated on different conceptions of fairness, such as demographic parity or equal odds that depend on the available labels in the dataset. However, it is often the case in practice that unsupervised learning is used as part of a machine learning pipeline (for instance, to perform dimensionality reduction or representation learning via SVD) or as a standalone model (for example, to derive a customer segmentation via k-means). It is thus crucial to develop approaches towards fair unsupervised learning. This work investigates fair unsupervised learning within the broad framework of generalised low-rank models (GLRM). Importantly, we introduce the concept of fairness functional that encompasses both traditional unsupervised learning techniques and min-max algorithms (whereby one minimises the maximum group loss). To do so, we design straightforward alternate convex search or biconvex gradient descent algorithms that also provide partial debiasing techniques. Finally, we show on benchmark datasets that our fair generalised low-rank models ("fGLRM") perform well and help reduce disparity amongst groups while only incurring small runtime overheads.

CCS CONCEPTS

• **Theory of computation → Unsupervised learning and clustering**; • **Social and professional topics → User characteristics**.

KEYWORDS

Unsupervised Learning, PCA, Clustering, Fairness

ACM Reference Format:
Francois Buet-Golfouse and Islam Utyagulov. 2022. Towards Fair Unsupervised Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, **??** pages. https://doi.org/10.1145/3531146.3533197

1 INTRODUCTION

Using unsupervised learning algorithms, such as PCA, k-means or non-negative matrix factorisation – which is prevalent in recommender systems – without paying attention to fairness may lead to adverse outcomes in some particular demographic groups (e.g., customer segmentation or facial recognition). Indeed, we verify on a number of datasets that there are discrepancies, sometimes important, between the average cost (such as reconstruction error or distance to a centroid) born by a group versus another. Thus, the fairness metric that emerges in these applications is that of (average) cost parity amongst groups.

More broadly, recent advances have been made recently in the space of fair unsupervised learning, in particular by introducing fairlets [9], leading to fair PCA [36, 37], fair k-medoids [3, 17, 25] and fair spectral clustering [27]. Since fairness in this context is not obvious to tackle, there have been multiple attempts to define it. A recent overview of fair clustering has been given in [8]. On the other hand, in supervised learning, multiple technical definitions of fairness co-exist and have been reviewed in-depth [4, 23, 32, 40], bringing to the fore impossibility theorems [10, 24] that proved that these different acceptations of fairness cannot be satisfied at once. In addition, it has been shown that [2, 23, 34] fully debiased models could fail to generalise out-of-sample, which we would expect to also apply to the case of unsupervised learning.

Our contributions. Our main result is a framework that encompasses many applications such as fair PCA [36], fair k-medoids [17], fair non-negative matrix factorisation and other models whose standard versions can be expressed as generalised low-rank models [38], and provides added flexibility.

- First, we develop a general fair generalised low rank framework that reduces disparity across group-wise average cost in an unsupervised learning task (such as reconstruction error in PCA).
- Second, we show that a particular group functional, namely weighted Log-Sum Exponential, has interesting properties (such as convexity, differentiability, etc.) that make it particularly appropriate.
- Third, we build on [38] to develop generic algorithms that take advantage of *biconvexity*. This generality in specifying a fair GLRM model makes this framework a very flexible one, also including *partial debiasing*.
- Fourth, we apply our methodology to multiple datasets, benchmark it against fair PCA and fair k-means algorithms, and show its performance out-of-sample, thus highlighting the role of partial debiasing.

We also note that a number of extensions are possible by considering relative costs or outcome-based fairness.

2 GENERALISED LOW RANK MODELS

We start by recalling some concepts linked to generalised low rank models ("GLRMs"). A textbook exposition of GLRMs is given in [38]. The term itself, *generalised low rank models* refers –in general–

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533197

to the approximation of a data matrix by the product of two low-dimensional factors.

Definition 1. A generalised low rank model is defined based on the following elements:

(1) An $n \times p$ data matrix A and a (lower) dimension d;
(2) Element-wise (usually biconvex) loss functions $\ell_{i,j}$, (usually convex) penalty functions r_i and $\tilde{r}_j$ for all i, j, and an objective function $\tilde{\mathcal{L}}$ written as

$$\tilde{\mathcal{L}}(X, Y) = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \ell_{i,j}(x_i \cdot y_j, A_{i,j}) + \sum_{i=1}^{n} r_i(x_i) + \sum_{j=1}^{p} \tilde{r}_j(y_j), \quad (1)$$

where $x_i \in \mathbb{R}^d$ and $y_j \in \mathbb{R}^d$ for all i, j. The matrices X and Y correspond to the stacked vectors.

Remark 1. In this work, we understand the matrix Y made up of the row entries y_j ($j = 1, \cdots, p$) as the *dictionary* to be learnt and the matrix X made up of the row entries x_i ($i = 1, \cdots, n$) as the individual *weights*.

Let us now provide some examples:

(1) By choosing $\ell_{i,j}(u, a) = (u - a)^2$, one recovers PCA.
(2) Similarly, robust PCA can be obtained by picking $\ell_{i,j}(u, a) = |u - a|$ with $r(x) = \gamma/2\|x\|_2^2$ and $\tilde{r}(y) = \gamma/2\|y\|_2^2$.
(3) On the other hand, setting $r(x) = 0$ if $x \geq 0$ and $+\infty$ otherwise, with the same definition for $\tilde{r}$ leads to non-negative matrix factorisation ("NNMF").
(4) Finally, picking $r(x) = 0$ if $x = e_l$ for some $l \in \{1, \cdots d\}$ and $+\infty$ otherwise, while $\tilde{r}(y) = 0$, leads to the usual k-means clustering problem.

Many more applications (such as subspace clustering) can be shown to fit the generic GLRM form [38].

3 FAIRNESS IN UNSUPERVISED LEARNING

3.1 Literature review

The relationship between fairness in supervised and unsupervised learning is not straightforward. Many notions of fairness in supervised learning (such as classification and scoring [11, 19, 24]) focus on a single learning task, whereas unsupervised learning considers a generic transformation of the data. While fairness in unsupervised learning has recently become a major theme of research, fair PCA –for instance– can be seen directly in the line of earlier attempts to reduce the correlation between a protected (or sensitive) attribute [6, 41].

For the sake of brevity, we do not give an exhaustive account of fairness in unsupervised learning. It is tempting to consider [31] as an early attempt at introducing (invidual) fairness in clustering. By adapting the notion of disparate impact to clustering and introducing the notion of fairlets (i.e., minimal sets that satisfy fair representation while approximately preserving the clustering objective), [9] paved the way for much of the work in the field. [36, 37] explore PCA and dimensionality reduction with multiple constraints, with an application to fairness. In [27], the authors tackle the case of spectral clustering. [1] proposes a generic approach to fair clustering, including k-medoids and considers a minmax criterion across groups. On a different note, [7] tackles the issue of data summarisation via a determinantal measure of diversity.

Furthermore, work in fair recommender systems (which include some matrix factorisation techniques) has grown due to the better understanding of certain phenomena such as echo chambers or filter bubbles. In addition to biases linked to certain protected characteristics (such as poor performance of recommender systems to serve under-represented minority groups), specific issues have appeared such as *user under-representation* [29] and *item under-recommendation* (also known as popularity bias) [42].

3.2 Fairness criteria

As a result of unsupervised learning's diversity and breadth, multiple notions of fairness have been put forward for (or adapted to) unsupervised learning, including *social fairness*, *balance fairness* and *individual fairness*. Let us now give a brief account of these different criteria.

(1) The *social fairness* criterion was introduced in unsupervised learning in [17], where it was applied to a clustering problem and further developed in [1, 30]. In short, it requires that the average cost (e.g., reconstruction loss in PCA or distance to medoid in clustering) be the same across groups. This can be tackled by minimising the maximum of groups' average costs. Note that it has had a long history since it was introduced by philosopher John Rawls in his *Theory of Justice* [35] as a justice criterion ("maximin") applied to the usual utilitarian framework.
(2) Similarly, the principle according to which different groups should have the same distribution across clusters can be traced back to [9], where the authors posit the notion of *balance fairness*, which they attack through so-called "fairlets". A related concept is that of *bounded representation* [3], which requires that the proportion of a group in each cluster be between two pre-specified values. The *maximum fairness cost* [17] is the maximum of the sum of all deviations from the ideal proportion for each protected group in a cluster.
(3) Last, *individual fairness* compares the statistical distance between two points obtained from their inputs and the algorithm's output distribution and mandates that two similar inputs should have similar outputs. This paradigm was first used in [31] for clustering and further adapted by [26].

Many more concepts exist [8] (the reader is also referred to [32] for an overview of such concepts in supervised learning) but this paper focuses primarily on *social fairness*-type of metrics. Note, however, that –as pointed out in Section 6– the proposed framework can be easily adapted to other fairness notions.

4 GROUP FUNCTIONALS

In this Section, we introduce the key insight of our proposal, namely that of a group function. Suppose that there are K (distinct) groups, $k = 1, \cdots, K$, corresponding, say, to the K categories of the protected characteristic s. We can thus introduce the corresponding partition of Ω

$$\Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_K, \quad (2)$$

such that the intersection between any Ω_k and $\Omega_{k'}$ is empty.

4.1 Motivation

It is immediate to notice that Equation 1 is simply

$$\tilde{\mathcal{L}}(X,Y) = \sum_{k=1}^{K} \frac{|\Omega_k|}{|\Omega|} z_k(X,Y) + \sum_{i=1}^{n} r_i(x_i) + \sum_{j=1}^{p} \tilde{r}_j(y_j), \quad (3)$$

where we use the shorthand notation

$$z_k(X,Y) := \frac{\sum_{(i,j)\in\Omega_k} \ell_{i,j}(x_i y_j, A_{i,j})}{|\Omega_k|},$$

which can be rewritten as $\mathcal{L}(X,Y) = T(z_1(X,Y),\cdots,z_K(X,Y)) +$ Penalty terms, and $T(z_1,\cdots,z_K) = \sum_{k=1}^{K} \frac{|\Omega_k|}{|\Omega|} z_k$. In particular, T is simply a linear combination of the average cost in each group weighted by this group's proportion in the sample.

Fair GLRMs thus consist in finding functionals that are more suited to the task of reducing disparities. Finally, note that, by definition, this setup applies to any model that can be expressed as a generalised low rank model, and thus encompasses not only PCA and k-means, but also sparse PCA (sPCA), non-negative matrix factorisation (NMF), etc.

4.2 Fair GLRMs

Recalling the assumptions defining generalised low rank models, this leads us to the definition of *fair* GLRMs by modifying the loss function:

Definition 2. Suppose that $T : \mathbb{R}^K \to \mathbb{R}$ is a non-decreasing function in each of its arguments, then we define the fair generalised low rank model with respect to T as minimising the following objective function:

$$\mathcal{L}(X,Y) = T\left(z_1(X,Y),\cdots,z_K(X,Y)\right) + \sum_{i=1}^{n} r_i(x_i) + \sum_{j=1}^{p} \tilde{r}_j(y_j). \quad (4)$$

While this definition can apply to all functional T, the ones we consider aim at reducing disparities across groups. Let us now move to some concrete cases.

Remark 2. Last, note that some extreme cases are possible:

(1) If $K = 1$ (such that only one group is present), then the fGLRM is equivalent (up to a non-decreasing transformation) to the standard GLRM one.
(2) If $K = |i \in \Omega| = n$ (i.e., each row of the data matrix A is a group), then the fairness functional ensures that the maximal *individual* cost is minimised.

4.3 Examples

Let us introduce here some well-known fairness functionals, mostly stemming from the supervised learning literature.

Standard and reweighed loss function. Setting $T(z_1,\cdots,z_K) = \sum_{k=1}^{K} w_k z_k$, where $w_k \geq 0$ ad $\sum_{k=1}^{K} w_k = 1$. When $w_k = |\Omega_k|/|\Omega|$, we recover the usual generalized low rank model. If, on the other hand, $w_k = \frac{1}{K}$, then this corresponds to the *reweighed* GLRM. Reweighing is This is akin to some pre-processing techniques used in [22] for example.

Minmax. Choosing $T(z_1,\cdots,z_K) = \max_{k=1,\cdots,K} z_k = \|\mathbf{z}\|_\infty$ leads to a min-max problem. Note that this is the functional implicitly chosen in [36] (cf. Lemma 4.8 and Proof of Theorem 4.5 therein) to tackle fair PCA and in [17] to handle k-means. The minmax approach has the intuitive justification

(Weighted) L^p norms. A possible choice is a weighted L^p norm $T(z_1,\cdots,z_K) = \left(\sum_{k=1}^{K} w_k z_k^p\right)^{\frac{1}{p}} = \|\mathbf{z}\|_{p,\mathbf{w}}$. If the weights are uniform, then it follows that $T(z_1,\cdots,z_K) \propto \|\mathbf{z}\|_p$, which is –up to a simple transformation– a setup used in [28]. In particular, since the limit of the $\|\cdot\|_p$ norm is the $\|\cdot\|_\infty$ norm as $p \to +\infty$, one recovers the minmax formulation as an extreme case.

Penalised learning. Adding a term penalising unfairness and disparities is fairly common to (partially) debias supervised learning algorithms [13, 39] and leads to fairness functionals of the type $T(z_1,\cdots,z_K) = \sum_{k=1}^{K} w_k z_k + \lambda \sum_{k,k'} d(z_k, z_{k'})$, where d is a chosen distance (such as L^1 or L^2) and $\lambda > 0$ tunes the trade-off between the statistical loss and the disparity penalty term. One may avoid the double sum in the penalty term by considering instead $\sum_{k=1}^{K} w_k z_k + \lambda \sum_{k'} d\left(z_{k'}, \sum_{k=1}^{K} w_k z_k\right)$.

Building new fairness functionals from old ones. From J existing fairness functionals $T_1,\cdots,T_J$, one can create a new functional V

$$V(z_1,\cdots,z_K) = \mathcal{M}\left(T_1(z_1,\cdots,z_K)\cdots,T_J(z_1,\cdots,z_K)\right), \quad (5)$$

where $\mathcal{M} : \mathbb{R}^J \mapsto \mathbb{R}$ is a function that is non-decreasing in each of its components. The most straightforward example is to pick a convex combination of fairness functionals

$$V(z_1,\cdots,z_K) = \sum_{j=1}^{J} \lambda_j T_j(z_1,\cdots,z_K),$$

where $\lambda_j \geq 0$ and $\sum_{j=1}^{J} \lambda_j = 1$. For instance, one may consider functions V that "interpolate" between the usual average loss and the minmax case, as one may wish to control the fairness-accuracy control. For instance, one can pick $\gamma \in (0,1)$ such that

$$V(z_1,\cdots,z_K) = \lambda\left(\sum_{k=1}^{K} w_k z_k\right) + (1-\lambda)\max(z_1,\cdots,z_K) \quad (6)$$

However, in this work, we consider a specific group functional (but most considerations apply to any T).

4.4 Bayesian interpretation

The traditional GLRM framework offers a natural Bayesian interpretation, following [15]. Indeed, the minimisation of an fGLRM objective in Equation 4 can be seen as a *maximum a posteriori* problem, such that the hierarchical Bayesian model reads:

$$e^{-T(z_1(X,Y),\cdots,z_K(X,Y))} \cdot \prod_{i=1}^{n} e^{-r_i(x_i)} \cdot \prod_{j=1}^{p} e^{-\tilde{r}_j(y_j)}, \quad (7)$$

where the prior distributions on x_i and y_j have probability density functions proportional to $e^{-r_i(x_i)}$ and $e^{-\tilde{r}_j(y_j)}$ respectively. In the

case of $T(z_1, \cdots, z_K) = \sum_{k=1}^{K} w_k z_k$, we recover

$$e^{-T(z_1(X,Y),\cdots,z_K(X,Y))} = \prod_{(i,j)\in\Omega} e^{-\frac{w_{k(i)}}{|\Omega_{k(i)}|}\ell_{i,j}(x_i\cdot y_j; A_{ij})}.$$

In short, in this example, each entry $A_{i,j}$ is taken to be independent, but not necessarily identically distributed, depending on the values of the ratio $\frac{w_{k(i)}}{|\Omega_{k(i)}|}$.

However, generally, the product structure is *not* preserved and the group functional T introduces some dependence across observations, so that the observations A_{ij} are not independent anymore. Similarly, if one adopts the outcome-based version of fGLRMs in Equation 29 in Appendix E.3, then the prior distributions are not independent either. To summarise, fair GLRMs induce a *dependent* hierarchical Bayesian model.

4.5 Log-Sum Exponential Functional

Throughout this work, the main fairness functional that we use is the weighted Log-Sum Exponential (LSE) due to its many desirable properties and its ability to interpolate between the standard GLRM and the minmax programme.

4.5.1 Defining wLSE. We introduce the weighted (scaled) Log-Sum Exponential, that is a (small) generalisation of the usual Log-Sum Exponential, which is widely used in other machine learning applications (see [33]).

Definition 3. The weighted Log-Sum-Exponential ("wLSE") is defined as

$$T(z_1, \cdots, z_K) = \frac{1}{\alpha}\log\left(\sum_{k=1}^{K} w_k\, e^{\alpha z_k}\right), \tag{8}$$

where $\alpha > 0$, $w_k \geq 0$ for all k and $\sum_{k=1}^{K} w_k = 1$.

Remark 3. First, let us point out that the weight normalisation requirement is not strictly necessary but useful. Second, one can pick the natural choice $w_k = \frac{|\Omega_k|}{|\Omega|}$, but can also perform some sample reweighing simultaneously.

4.5.2 Properties. The wLSE has a number of properties of interest (both theoretically and practically).

Proposition 1. *Suppose that $\alpha > 0$, $w_k \geq 0$ for all k and $\sum_{k=1}^{K} w_k = 1$, then the weighted Log-Sum Exponential verifies the following properties:*

(1) *T is (jointly) convex in $(z_1, \cdots, z_K)$.*
(2) *The weighted average and the maximum functions are recovered as limiting cases:*

$$\lim_{\alpha\to 0} T(z_1, \cdots, z_K) = \sum_{k=1}^{K} w_k z_k \tag{9}$$

$$\lim_{\alpha\to+\infty} T(z_1, \cdots, z_K) = \max(z_1, \cdots, z_K) \tag{10}$$

(3) *A shift property holds for every $\overline{z} \in \mathbb{R}$: $T(z_1, \cdots, z_K) = \overline{z} + T(z_1 - \overline{z}, \cdots, z_K - \overline{z})$.*

Proof. See Appendix C.1. □

Remark 4. The shift property has a very natural explanation when thinking about fairness, as it decomposes the objective into the usual average cost, $\overline{z}$, (possibly reweighed), and a term that penalises disparity $T(z_1 - \overline{z}, \cdots, z_K - \overline{z})$. Based on this insight, one could tune the objective further: $T_\gamma(z_1, \cdots, z_K) = \overline{z} + \gamma T(z_1 - \overline{z}, \cdots, z_K - \overline{z})$, for $\gamma > 0$.

4.5.3 Choosing α and fairness implications. The hyper-parameter α enables one to "interpolate" between the traditional GLRM problem and its fair min-max formulation. Given that a number of articles use min-max formulations, it is worth justifying why one may wish to choose $\alpha \neq +\infty$. Indeed, $\alpha \neq +\infty$ introduces *partial debiasing* in unsupervised learning and helps relax assumptions of strict equal average costs amongst groups.

(1) wLSE is a soft maximum and enables modellers to approximate the maximum with a differentiable function [33], which is an advantage in many circumstances, including when gradient descent-type algorithms are used to minimise the objective function.
(2) Constraints may exist in the application of an algorithm, such as a minimal overall statistical performance, leading a modeller to debias an algorithm as much as possible while keeping the overall average loss below a given threshold.
(3) Partial debiasing was used in [23] to account for the presence of fairness-accuracy (or even fairness-fairness) trade-offs [24, 32]. The notion of a trade-off between an average statistical performance metric (such as an empirical average loss) and disparity metrics is illustrated empirically in Section 7.
(4) Issues regarding the out-of-sample performance of debiasing algorithms have been investigated [2, 12] in the context of supervised learning. [2], in particular, demonstrates the need to carefully tune a debiasing algorithm as "total" debiasing may lead to worse results out-of-sample. In other words, picking an intermediate value of α may lead to superior results on unseen data (such as the out-of-sample test set). In other words, it may not be optimal from a fairness point of view to choose $\alpha = +\infty$.

5 FITTING FAIR GLRMS

We now turn our focus to the minimisation of the objective function in Equation 4. At first glance, it may seem significantly more complex than in the case of standard GLRMs, but it turns out that –in most cases– the essential *biconvex* property of the objective function still holds.

5.1 Biconvexity of fGLRMs

The attractiveness of fGLRMs comes from the fact that under mild assumptions on the fairness functional T, they are biconvex functions in X and Y and thus fairly straightforward to minimise. Note that one cannot hope, in general, for something better than biconvexity since standard GLRMs are themselves biconvex. Note that not all GLRMs are biconvex, but most matrix factorisation techniques (such as SVD, PCA or NMF) are.

Proposition 2. *Under the assumptions that $T : \mathbb{R}^K \to \mathbb{R}$ is convex and is non-decreasing in each argument, and each individual loss*

function $\ell_{i,j}$ is biconvex in x_i and y_j, then the application

$$(X, Y) \mapsto T\left(z_1(X, Y), \cdots, z_K(X, Y)\right) \tag{11}$$

is biconvex in X and Y. If, in addition, each penalty function r_i or $\tilde{r}_j$ is convex, then the application

$$\begin{aligned}(X, Y) &\mapsto \mathcal{L}(X, Y) \\ &= T\left(z_1(X, Y), \cdots, z_K(X, Y)\right) + \sum_{i=1}^{n} r_i(x_i) + \sum_{j=1}^{p} \tilde{r}_j(y_j)\end{aligned} \tag{12}$$

is biconvex in X and Y.

Proof. See Appendix C.2. □

What this result shows is that the introduction of a group functional does not change the fundamental structure of a generalised low rank model. This has implications in terms of *optimisation*, as existing algorithms can simply be tweaked and reused. While one may use gradient descent algorithms and variants thereof on the non-convex objective function (Equation 4), more bespoke algorithms exist.

5.2 Alternating Minimisation (or Alternate Convex Search)

Alternating minimisation is a well-known algorithm that minimises the objective function one direction at the time. If the objective function is multi-convex (i.e., convex in each direction when the other ones are fixed), this is the same as alternate convex search. The reader is referred to [5, 18, 20].

Algorithm 1 Alternating Minimisation for fGLRM D.1

Require: Matrix $\mathbf{A}$, loss functions $\ell_{i,j}$ and penalty functions r_i and $\tilde{r}_j$.
Select initial values $\mathbf{X}^0$ and $\mathbf{Y}^0$
repeat
  for $i = 1, \cdots, n$ **do**
    $x_i \leftarrow \arg\min_x \mathcal{L}((\mathbf{X}_{-i}, x), \mathbf{Y}) + r_i(x)$
  end for
  for $j = 1, \cdots, p$ **do**
    $y_j \leftarrow \arg\min_y \mathcal{L}(\mathbf{X}, (\mathbf{Y}_{-j}, y) + \tilde{r}_j(y)$
  end for
until convergence
return $\mathbf{X}, \mathbf{Y}$

We have used the shorthand $\mathbf{X} = (\mathbf{X}_{-i}, x_i)$ for all $i = 1, \cdots, n$. This algorithm is the adaptation to the fair set-up of Algorithm 1 in [38].

Remark 5. Let us make a couple of practical remarks at this stage.

(1) The for loop $i = 1, \cdots, n$ in this algorithm may be replaced with a standard GLRM for loop if the penalty function r_i is a set indicator penalty (for instance in the case of clustering, $r_i(x) = 0$ if $x = e_l$ for some $l \in \{1, \cdots d\}$ and $+\infty$ otherwise).
(2) Due to overflow, it may sometimes be necessary to express the fairness functional slightly differently. For instance, $\|\mathbf{z}\|_{p,\mathbf{w}} = \|\mathbf{z}\|_\infty \left\| \frac{\mathbf{z}}{\|\mathbf{z}\|_\infty} \right\|_{p,\mathbf{w}}$. Similarly, $wLSE_\alpha(\mathbf{z}) = \|\mathbf{z}\|_\infty + wLSE_\alpha(z_1 - \|\mathbf{z}\|_\infty, \cdots z_K - \|\mathbf{z}\|_\infty)$.

5.3 Biconvex Gradient Descent

We suppose here that all functions are differentiable and that the penalty functions r_i and $\tilde{r}_j$ are convex. Then, thanks to the biconvexity of $\mathcal{L}$ in X and Y, one can derive the following expressions:

$$\begin{aligned}\frac{\partial \mathcal{L}(X, Y)}{\partial x_i} &= \frac{\partial T}{\partial z_{k(i)}} \frac{\partial z_{k(i)}}{\partial x_i} + \nabla r_i(x_i) \\ \frac{\partial \mathcal{L}(X, Y)}{\partial y_j} &= \sum_{k=1}^{K} \frac{\partial T}{\partial z_k} \frac{\partial z_k}{\partial y_j} + \nabla \tilde{r}_j(y_j)\end{aligned}$$

where $\frac{\partial z_{k(i)}}{\partial x_i} = \frac{1}{|\Omega_{k(i)}|} \sum_{j|(i,j)\in\Omega_{k(i)}} \nabla \ell_{i,j}(x_i \cdot y_j; A_{ij}) y_j$, $\frac{\partial z_k}{\partial y_j} = \frac{1}{|\Omega_k|} \left(\sum_{i|(i,j)\in\Omega_k} \nabla \ell_{i,j}(x_i \cdot y_j; A_{ij}) x_i \right)$, and $\frac{\partial T}{\partial z_k} = \frac{w_k e^{\alpha z_k}}{\sum_{k=1}^K w_k e^{\alpha z_k}}$. This leads to a biconvex gradient descent algorithm:

Algorithm 2 Biconvex Gradient Descent D.3

Require: Matrix $\mathbf{A}$, loss functions $\ell_{i,j}$ and penalty functions r_i and $\tilde{r}_j$, step sizes $(\alpha_t)_{t\geq 1}$.
Select initial values $\mathbf{X}^0$ and $\mathbf{Y}^0$
$t \leftarrow 1$
repeat
  for $i = 1, \cdots, n$ **do**
    $g_i^t \leftarrow \frac{\partial \mathcal{L}(\mathbf{X}^{t-1}, \mathbf{Y}^{t-1})}{\partial x_i^{t-1}}$
    $x_i^t \leftarrow x_i^{t-1} - \alpha_t g_i^t$
  end for
  for $j = 1, \cdots, p$ **do**
    $\tilde{g}_j^t \leftarrow \frac{\partial \mathcal{L}(\mathbf{X}^t, \mathbf{Y}^{t-1})}{\partial y_j^{t-1}}$
    $y_j^t \leftarrow y_j^{t-1} - \alpha_t \tilde{g}_j^t$
  end for
  $t \leftarrow t + 1$
until convergence
return $\mathbf{X}^t, \mathbf{Y}^t$

The main difference between GLRMs and fGLRMS comes from their particular gradient structure and the fact that an iterative weighing scheme has *implicitly* been introduced, similarly to boosting. Indeed, by denoting

$$\delta_k = \frac{\partial T}{\partial z_k} = \frac{w_k e^{\alpha z_k}}{\sum_{k=1}^K w_k e^{\alpha z_k}},$$

we obtain that $\delta_k \geq 0$ for all k's and $\sum_{k=1}^K \delta_k = 1$. When $\alpha \to 0$, we simply recover $\delta_k = w_k$, and δ_k does not change at each iteration. On the other hand, when α is non-zero, the weights are adaptive and over-weigh the groups with higher average cost in the previous iteration.

Remark 6. Some convergence properties of these algorithms are discussed in the Appendix D.

6 OUTCOME-BASED FAIRNESS AND OTHER EXTENSIONS TO FGLRMS

In this Section, we show how alternative notions of fairness can be included in the fGLRM framework. Importantly, the alternating minimisation algorithm can still be applied to these cases.

6.1 Group functional on outcome disparity

Let us now adopt an outcome-based viewpoint on unsupervised learning. Here, we consider that one wishes to apply notions of demographic parity to the output of the unsupervised learning algorithm, which we consider here to be $x_i \cdot y_j$. In recommender systems, for example, one may wish to ensure that all groups have the same (predicted) average rating or satisfaction. One way to tackle this issue is to penalise the disparity between each group's average output and the overall average and redefine the objective function as

$$\mathcal{L}^{O}(X,Y) := \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega} \ell_{i,j}(x_i \cdot y_j, A_{i,j}) + \gamma T\left(u_1(X,Y) - \overline{u}(X,Y), \cdots, u_K(X,Y) - \overline{u}(X,Y)\right) + \sum_{i=1}^{n} r_i(x_i) + \sum_{j=1}^{p} \tilde{r}_j(y_j), \quad (13)$$

where $u_k(X,Y) := \frac{\sum_{(i,j)\in\Omega_k} x_i \cdot y_j}{|\Omega_k|}$ is the average outcome in group k and $\overline{u}(X,Y) := \frac{\sum_{(i,j)\in\Omega} x_i \cdot y_j}{|\Omega|}$ is the (possibly reweighed) sample average. In the case of wLSE, thanks to its shift property, this can be rewritten as

$$\frac{1}{|\Omega|}\sum_{(i,j)\in\Omega} \left(\ell_{i,j}(x_i \cdot y_j, A_{i,j}) - \gamma(x_i \cdot y_j)\right) + \gamma T\left(u_1(X,Y), \cdots, u_K(X,Y)\right) + \sum_{i=1}^{n} r_i(x_i) + \sum_{j=1}^{p} \tilde{r}_j(y_j), \quad (14)$$

Importantly, this notion preserves the factorisation property of fGLRMS in the sense that non-penalty terms only depend on the dot product $x_i \cdot y_j$ and is still an fGLRM.

6.2 Integrating balanced notions of fairness

Let us focus here on additional notions of fairness [8], specific to clustering and show how to tackle them. In the below, we consider that we wish to cluster data points from K groups into C clusters.

6.2.1 Balance fairness [9]. The (reformulated) notion of balance can be implemented in a slightly generalised version of our framework. Indeed, the following objective encourages proportions of points in cluster l to be similar across groups:

$$\mathcal{L}^{\mathcal{B}}(X,Y) := \mathcal{L}(X,Y) + \gamma \sum_{l=1}^{C} T\left(\frac{\sum_{(i,j)\in\Omega_1} \mathbf{1}_{\{x_i=e_l\}}}{|\Omega_1|}, \cdots, \frac{\sum_{(i,j)\in\Omega_K} \mathbf{1}_{\{x_i=e_l\}}}{|\Omega_K|}\right). \quad (15)$$

6.2.2 Bounded representation [3]. Similarly, one can encode a notion such as bounded representation by introducing a new penalty term

$$\mathcal{L}^{\mathcal{R}}(X,Y) := \mathcal{L}(X,Y) + \gamma \sum_{k=1}^{K}\sum_{l=1}^{C} r_{k,l}\left(\frac{\sum_{(i,j)\in\Omega_k} \mathbf{1}_{\{x_i=e_l\}}}{|\Omega_k|}\right), \quad (16)$$

where $r_{k,l}(p)$ is worth 0 if $b \le p \le a$ and $+\infty$ otherwise. Note that one needs to be careful with the initialisation of an algorithm with bounded representation and may wish to perform stratified sampling per group and per cluster and/or use a smooth representation of $r_{k,l}(p)$.

Dataset	Reference	Binary	Multivariate
Adult	[14]	*sex*; female (16,192), male (32,650)	*race*; (white (41,762), black (4,685), other (406) asian-pac-islander (1,519), amer-indian-eskimo (470)
German Credit	[14]	*sex*; female (310), male (610)	*sex & marital status*; male : divorced/separated (50), female : divorced/separated/married (392), male : single (548); male : married/widowed (92)
Loan Defaults	[21]	*sex*; female (11,888), male (18,112)	-
LFW	[16]	*sex*; female (2,962), male (10,270)	-

Table 1: The details of binary and multivariate protected attribute in each dataset.

7 EXPERIMENTAL EVALUATION

In this section, we demonstrate the following results.

- **Multiple GLRMs.** We consider PCA, k-means and non-negative matrix factorisation (NMF) in our experiments.
- **Reproducibility and convergence.** Using our proposed wLSE functional with a large positive α (10^5), we are able to reproduce results from [17, 36] (which can also be recovered by simply picking $T = \max$ in our alternating minimisation approach).
- **Partial Debiasing.** By varying the α hyperparameter in the wLSE functional, we obtain a full spectrum of results ranging from the standard GLRM, to intermediate states and the min-max solution. This points to the usual fairness-accuracy trade-off as the overall average cost tends to increase as disparity decreases.
- **Generalisation.** For each level of α, we use the corresponding solution calibrated on the train set (i.e., we keep the y_j's fixed), and simply compute the new set of weights x_i for all i's in the test set. This is similar to online dictionary learning. We find that, first, the performance on the test set (expectedly) deteriorates and, second, that the completely fair solution may have become sub-optimal. This reinforces the the attractiveness of partial debiasing, which can thus be interpreted as a fair regularisation.

7.1 Data, models and approaches presented

7.1.1 Datasets. We have considered three datasets (whose details are indicated in Table 1):

- German Credit [14],
- Loan Default Credit [21],
- Adult [14].

Throughout our experiments, groups have been defined in terms of membership to a class defined thanks to a protected characteristic, as detailed in Table 1.

Remark 7. Results in Figures 1- 2 are based on the aforementioned datasets and in Table 1 we provide additional details on the protected attribute used, as well as unique values and their counts. In Appendix E.3, we provide some additional results that are based on LFW (Labeled Faces in the Wild) dataset [16].

7.1.2 fGLRMs under consideration. In this paper, we have implemented in the fGLRM framework the following objectives:

- k-Means (Figures 1-4),

- Principal Component Analysis (PCA) (Figures 5-6),
- Non-Negative Matrix factorization (NMF) (Figures 7-8 in Appendix E).

It has to be noted that the framework is flexible enough and allows for different modifications.

7.1.3 Fairness functionals and benchmarks. In addition to the wLSE functional, we have also considered a number of benchmarks:

- Standard GLRM (i.e., empirical average loss as in [38]);
- Reweighed GLRM (i.e., uniform weight $1/K$ for each group, similar to [22]), see Figure 8 in particular;
- Minmax approach (where $T = \max$) as in [17, 36];
- p-norm (or q-FFL) approach (in line with [28], see Figure 7);

We have shown that the proposed framework can handle cases with two (see Figures 1-2) and more protected groups (see Figures 3-4). Figures 5 - 6 demonstrate results when using group functional on outcome disparity presented in Section 6.1. For each aforementioned result we also illustrate the trade-off curve that is obtained when varying α parameter of wLSE functional[1].

Finally in Sections E.2 and E.3 we presented results based on supervised GLRM and outcome-based fairness incorporating a penalty term discussed in Sections B and E.3.1 respectively.

Remark 8. In experiments where a test set is needed, we have used a 70%-30% train-test split. We have used stratified sampling with respect to the protected attribute to ensure that both train and test sets have the same proportion of observations belonging to different groups. The algorithm used throughout these tests is the standard alternating minimisation. When considering a grid of values for α, we have considered the following values: $10^{-6}, 10^{-5}, 3*10^{-3}, 10^{-2}, 10^{-1}, 5*10^{-1}, 1, 5, 10^{1}, 2*10^{1}, 6*10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}$.

7.2 Main observations

Before delving into the precise results, we wish to summarise our key empirical findings.

First, considering a standard GLRM or a (partially) debiased one makes a difference both in terms of average statistical performance and disparity, thus indicating the presence of accuracy-fairness trade-offs in unsupervised learning too.

Second, "interpolating" techniques such as wLSE or q-FFL tend to offer similar results and converge to the minmax programme as their respective hyperparameter goes to ∞. In- and out-of-sample behaviour indicates that it is not always preferable to use the min-max formulation.

Third, using a reweighing scheme seems to help improve fairness in general (but not always), especially when α is small. However, as α increases, it becomes less relevant.

Fourth, the implementation within the fGLRM framework of the minmax approach or the adaptation of the q-FFL logic to unsupervised learning are straightforward and match results obtained via other algorithms. This underlines the interest in having a generic framework that accommodates easily many different functionals.

[1]When q-FFL functional used q parameter is varied instead.

7.3 Varying the hyperparameter α

Having established that the fGLRM approach can replicate previous results, we demonstrate some of its further benefits, such as its ability to interpolate between standard (i.e., no fairness considerations) and min-max solutions, with the degree of interpolation being controlled by a parameter α as shown in Equation 8. Figure 1 shows how both total loss and group disparity changes on the train set, as we vary parameter α through a grid of values. The larger values of α decrease group disparity, while smaller values bring solution closer to the standard solution.

In Figures 1-6, we use the following notations: "average" point is the solution of a standard algorithm, "min-max" point corresponds to the solution of wLSE with the largest $\alpha = 10^5$, while all other points are denoted as $wLSE_\alpha$ and correspond to the remaining values of α. The size of points is ordered according to the α used and the dotted line is a local regression line through the points.

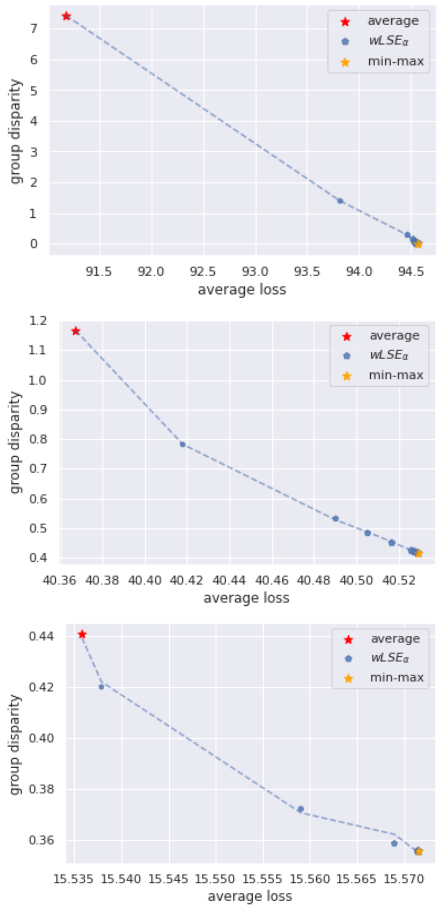

Figure 1: KMeans. Trade-off curve between the average loss and group disparity on the *train set* with $wLSE_\alpha$ functional on adult, german credit, and loan defaults' datasets. Each point corresponds to a different α.

Key takeaway. Results presented in Figure 1 are intuitive: decreasing the disparity amongst groups increases the overall loss,

and vice versa, which illustrates the fairness-accuracy trade-off [24] in the case of unsupervised learning.

7.4 Generalisation

To check whether debiasing generalises in fGLRMs, we consider out-of-sample behaviour and assess the performance of the fitted fGLRMs on a test set. The idea here is to keep "archetypes" Y learned on the train set fixed and solve for the best feature representations X_{test} of test set examples: $\frac{1}{|\Omega|}\sum_{j:(i,j)\in\Omega}\ell_{i,j}(x_i \cdot y_j^{train}, A_{i,j}^{test}) + \sum_{i=1}^{m} r_i(x_i)$. Once these are learned, computing average loss or group disparity is straightforward. For the sake of brevity, we only focus on k-means, and thus allocate test observations to the nearest centroid obtained during training. Despite a clear pattern on the train set, it is not always the case on the test set for different data sets as shown in Figure 2, in line with results pertaining to supervised learning [2].

Key takeaway. This suggests carefully choosing α and using cross-validation techniques, depending on the exact use case.

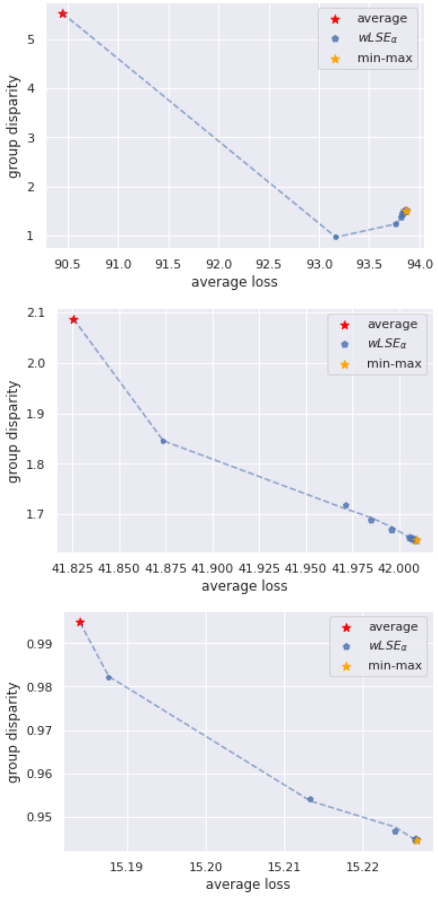

Figure 2: KMeans. Trade-off curve between average loss and group disparity on the *test set* with wLSE functional on adult, german credit, and loan defaults' datasets.

Remark 9. Additional results are presented in the Appendix E.

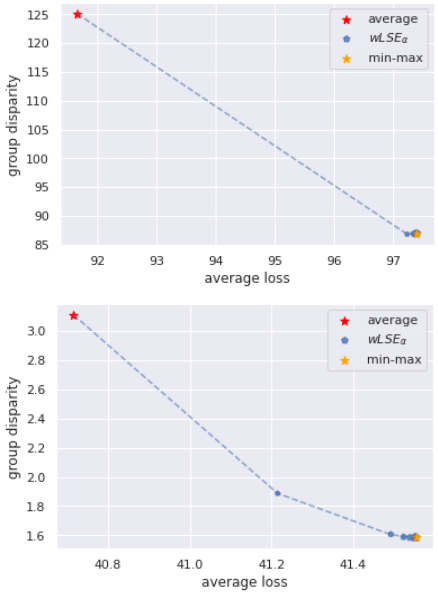

Figure 3: KMeans. Trade-off curve between average loss and group disparity on the *train set* with wLSE functional on adult and german credit datasets. Protected attribute is a multivariate feature (race in adult dataset, sex and marital status in german credit dataset respectively).

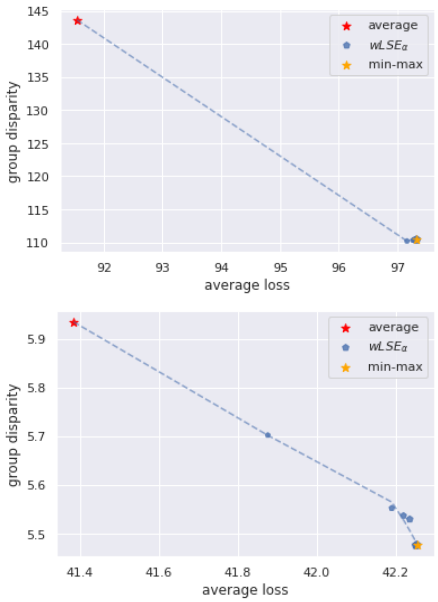

Figure 4: KMeans. Trade-off curve between average loss and group disparity on the *test set* with wLSE functional on adult and german credit datasets. Protected attribute is a multivariate feature (race in adult dataset, sex and marital status in german credit dataset respectively).

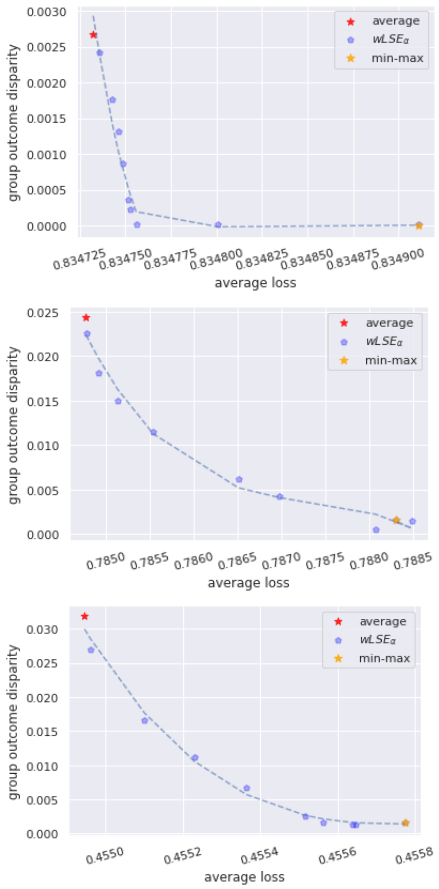

Figure 5: PCA. Trade-off curve between average loss and group outcome disparity on the *train set* with wLSE functional on adult, german credit, and loan defaults' datasets.

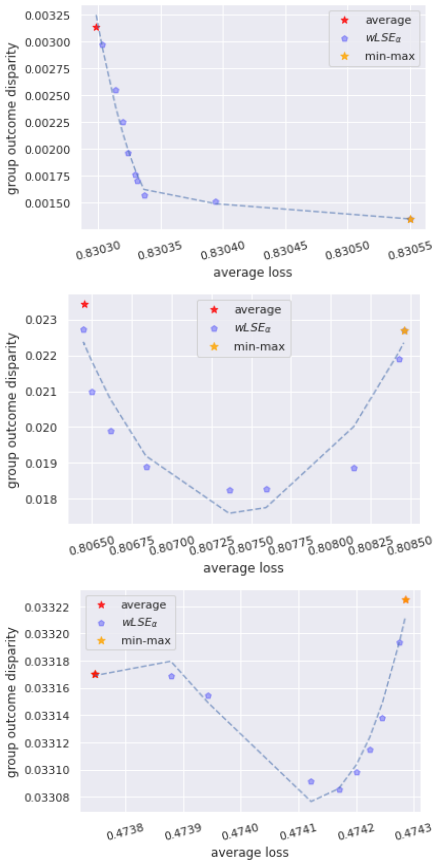

Figure 6: PCA. Trade-off curve between average loss and group outcome disparity on the *test set* with wLSE functional on adult, german credit, and loan defaults' datasets.

8 DISCUSSION

In this paper, we have introduced the notion of fair generalised low rank models by applying a fairness functional to group-wise average losses, leading to a reduction in cost disparity across groups. Building fair GLRMs has enabled us to devise a generic framework encompassing fair PCA and fair k-means, but also applicable to many other use cases, such as sparse PCA, non-negative matrix factorisation, subspace clustering and many more.

We have also specified a particular choice of such fairness functional, namely the weighted Log-Sum Exponential, which has many desirable properties. This permits users to select a hyper-parameter α that governs the fairness-accuracy trade-off. The importance of debiasing is emphasised by some of our out-of-sample results, showing that total debiasing in-sample may lead to very different results out-of-sample. In addition, we have shown that straightforward algorithms based on biconvexity (or variants thereof) could be efficiently leveraged to solve these fair objective functions. fGLRMs thus inherit some properties of GLRMs.

Finally, some extensions are straightforward, such as including orthogonality constraints between the learnt dictionary and the specified protected characteristic. However, further research is warranted to understand how to transpose multiple fairness definitions from classification or regression to unsupervised learning and how to assess out-of-sample performance.

ACKNOWLEDGMENTS

The authors did not receive funding for this project. This paper was prepared for informational purposes by the authors, and is not a product of any institution's Research Department. The views expressed therein are solely those of the authors and do not reflect those of any institution or employer, past and present. The authors make no representation and warranty whatsoever and disclaim all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

REFERENCES

[1] Mohsen Abbasi, Aditya Bhaskara, and Suresh Venkatasubramanian. 2021. Fair Clustering via Equitable Group Representations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 504–514. https://doi.org/10.1145/3442188.3445913

[2] Ashrya Agrawal, Florian Pfisterer, Bernd Bischl, Jiahao Chen, Srijan Sood, Sameena Shah, Francois Buet-Golfouse, Bilal A Mateen, and Sebastian Vollmer. 2020. Debiasing classifiers: is reality at variance with expectation?

[3] Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. 2019. Clustering without Over-Representation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 267–275. https://doi.org/10.1145/3292500.3330987

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* (Aug. 2018), 42 pages. https://doi.org/10.1177/0049124118782533 arXiv:1703.09207

[5] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge University Press.

[6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Long Beach, CA, 3992–4001. http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf

[7] Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth Vishnoi. 2018. Fair and Diverse DPP-Based Data Summarization. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 716–725. https://proceedings.mlr.press/v80/celis18a.html

[8] Anshuman Chhabra, Karina Masalkovaitė, and Prasant Mohapatra. 2021. An Overview of Fairness in Clustering. *IEEE Access* 9 (2021), 130698–130720. https://doi.org/10.1109/ACCESS.2021.3114099

[9] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair Clustering through Fairlets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 5036–5044.

[10] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (June 2017), 153–163. https://doi.org/10.1089/big.2016.0047 arXiv:1703.00056

[11] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047

[12] Ching-Yao Chuang and Youssef Mroueh. 2021. Fair Mixup: Fairness via Interpolation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=DNl5s5BXeBn

[13] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical Risk Minimization Under Fairness Constraints. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/83cdcec08fbf90370fcf53bdd56604ff-Paper.pdf

[14] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[15] William Fithian and Rahul Mazumder. 2018. Flexible Low-Rank Statistical Modeling with Missing Data and Side Information. *Statist. Sci.* 33, 2 (2018), 238 – 260. https://doi.org/10.1214/18-STS642

[16] Tamara Berg Gary B. Huang, Manu Ramesh and Erik Learned-Miller. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 49, 07 (2007).

[17] Mehrdad Ghadiri, Samira Samadi, and Santosh Vempala. 2021. Socially Fair K-Means Clustering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 438–448. https://doi.org/10.1145/3442188.3445906

[18] Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. 2007. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research* 66 (2007). https://doi.org/10.1007/s00186-007-0161-1

[19] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29 (Dec. 2016), 3323–3331. https://doi.org/10.5555/3157382.3157469 arXiv:1610.02413

[20] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

[21] & Che-hui Lien I-Cheng Yeh. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications* 2, 36 (2009), 2473–2480. https://doi.org/10.1016/J.ESWA.2007.12.020

[22] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[23] Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. 2020. Model-Agnostic Characterization of Fairness Trade-offs. In *Proceedings of the International Conference on Machine Learning*. Vienna, Austria / Online, 9339–9349. https://proceedings.icml.cc/paper/2020/hash/cf5530d9e441e0d7857435321437356 9

[24] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, Article 43, 23 pages. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[25] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. 2019. Fair k-Center Clustering for Data Summarization. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 3448–3457. https://proceedings.mlr.press/v97/kleindessner19a.html

[26] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. 2020. A Notion of Individual Fairness for Clustering. arXiv:2006.04960 [stat.ML]

[27] Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. 2019. Guarantees for Spectral Clustering with Fairness Constraints. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 3458–3467. https://proceedings.mlr.press/v97/kleindessner19b.html

[28] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2020. Fair Resource Allocation in Federated Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ByexElSYDr

[29] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-Oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 624–632. https://doi.org/10.1145/3442381.3449866

[30] Yury Makarychev and Ali Vakilian. 2021. Approximation Algorithms for Socially Fair Clustering. arXiv:2103.02512 [cs.DS]

[31] Geoffrey J McLachlan and Kaye E Basford. 1988. *Mixture models: Inference and applications to clustering*. Vol. 38. M. Dekker New York.

[32] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT* 18)*. New York, USA.

[33] Frank Nielsen and Ke Sun. 2016. Guaranteed Bounds on Information-Theoretic Measures of Univariate Mixtures Using Piecewise Log-Sum-Exp Inequalities. *Entropy* 18, 12 (2016). https://doi.org/10.3390/e18120442

[34] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., 5680–5689. http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf

[35] John Rawls. 1971. *A Theory of Justice* (1 ed.). Belknap Press of Harvard University Press.

[36] Samira Samadi, Uthaipon Tantipongpipat, Jamie Morgenstern, Mohit Singh, and Santosh Vempala. 2018. The Price of Fair PCA: One Extra Dimension. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 10999–11010.

[37] Uthaipon Tantipongpipat, Samira Samadi, Mohit Singh, Jamie H Morgenstern, and Santosh Vempala. 2019. Multi-Criteria Dimensionality Reduction with Applications to Fairness. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/2201611d7a08ffda97e3e8c6b667a1bc-Paper.pdf

[38] Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. 2016. Generalized Low Rank Models. *Foundations and Trends in Machine Learning* 9, 1 (2016). https://doi.org/10.1561/2200000055 arXiv:1410.0342 [stat-ml]

[39] US Congress. 2003. P. L. 108-159: Fair and Accurate Credit Transactions Act. https://www.gpo.gov/fdsys/pkg/PLAW-108publ159/pdf/PLAW-108publ159.pdf

[40] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Conference on Software Engineering*. ACM, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776

[41] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of Machine Learning Research*, Vol. 28. 1362–1370. http://proceedings.mlr.press/v28/zemel13.html

[42] Ziwei Zhu, Jianling Wang, and James Caverlee. 2020. Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 449–458. https:

//doi.org/10.1145/3397271.3401177