arXiv:1911.04931v1 [cs.LG] 12 Nov 2019

Efficient Fair Principal Component Analysis

Mohammad Mahdi Kamani† Farzin Haddadpour‡ Rana Forsati* Mehrdad Mahdavi‡

†College of Information Sciences and Technology
‡School of Electrical Engineering and Computer Science
The Pennsylvania State University
University Park, PA, USA
{mqk5591, fxh18, mzm616}@psu.edu

*Microsoft Bing
Bellevue, WA, USA
raforsat@microsoft.com

Abstract

The flourishing assessments of fairness measure in machine learning algorithms have shown that dimension reduction methods such as PCA treat data from different sensitive groups unfairly. In particular, by aggregating data of different groups, the reconstruction error of the learned subspace becomes biased towards some populations that might hurt or benefit those groups inherently, leading to an unfair representation. On the other hand, alleviating the bias to protect sensitive groups in learning the optimal projection, would lead to a higher reconstruction error overall. This introduces a trade-off between sensitive groups' sacrifices and benefits, and the overall reconstruction error. In this paper, in pursuit of achieving fairness criteria in PCA, we introduce a more efficient notion of Pareto fairness, cast the Pareto fair dimensionality reduction as a multi-objective optimization problem, and propose an adaptive gradient-based algorithm to solve it. Using the notion of Pareto optimality, we can guarantee that the solution of our proposed algorithm belongs to the Pareto frontier for all groups, which achieves the optimal trade-off between those aforementioned conflicting objectives. This framework can be efficiently generalized to multiple group sensitive features, as well. We provide convergence analysis of our algorithm for both convex and non-convex objectives and show its efficacy through empirical studies on different datasets, in comparison with the state-of-the-art algorithm.

1 Introduction

Recent advances in machine learning (ML) have vastly improved the capabilities of computational reasoning in complex domains. From tasks like image and video processing, game playing, text classification, to complex data analysis, machine learning is continually finding new applications and exceeding human-level performance in some cases. Nevertheless, when machine learning models are trained on real data, the existing societal inequalities in data are manifested on the systems built upon them that could misled models in ways that can have profound fairness implications such as being biased to sensitive features like race or gender. As more critical systems employ ML, such as financial systems, hiring and admissions, healthcare, and law, it is vitally important that we develop rigorous fair algorithms that are as accurate as possible.

Recently, the growing attention to the fairness problem in algorithmic decision-making systems has led to an unprecedented attempts to revisit machine learning models for supervised and unsupervised tasks to satisfy fairness constraints (Munoz et al., 2016). An expanding line of work dedicated to define different metrics for fairness problems and mechanisms to satisfy those measures in learning tasks such as (Hardt et al., 2016; Zafar et al., 2017b; Calders et al., 2009; Calders and Verwer, 2010; Kamishima et al., 2011; Agarwal et al., 2018; Zafar et al., 2017a, 2015). The work on this realm is focused on biased data or biased algorithms, however, using these biased algorithms in decision-making systems would lead into generating more biased data. This makes the causality of the fairness problem more complicated that exacerbates the problem even further (Barocas et al., 2017; Ghili et al., 2019).

Scheme	Time Complexity	Fair Dimension	Algorithm
Samadi et al. (2018)	$\mathcal{O}\left(d^{6.5}\log\left(\frac{1}{\epsilon}\right)\right)$	$r+k-1$	SDP + LP
Samadi et al. (2018)	$\mathcal{O}(\frac{d^3}{\epsilon^2})$	$r+k-1$	SDP via MW + LP
Morgenstern et al. (2019)	$\mathcal{O}\left(d^{6.5}\log\left(\frac{1}{\epsilon}\right)\right)$	$r+\lfloor\sqrt{2k+\frac{1}{4}}-\frac{3}{2}\rfloor$	SDP + LP
Morgenstern et al. (2019)	$\mathcal{O}(\frac{d^3}{\epsilon^2})$	$r+\lfloor\sqrt{2k+\frac{1}{4}}-\frac{3}{2}\rfloor$	SDP via MW + LP
This Work	$\boldsymbol{\mathcal{O}(\frac{r^2d}{\epsilon^2})}$	$\boldsymbol{r}$	GD

Table 1: Comparison of time complexity of different fair PCA algorithms to achieve an ϵ-fair subspace (please see Section 3 for definition). Here d denote the dimension of the original data, r is the target dimension, and k is the number of sensitive groups. Note that unlike previous studies that necessitates learning a subspace with larger dimension to guarantee fairness, our solution learns an exact r dimensional subspace by imposing additional constrains captured by a new notion of fairness proposed in this work to distinguish between local optimal fair subspaces (we used the following abbreviations above, SDP: Semi-Definite Programming, MW: Multiplicative Weight Algorithm, LP: Linear Programming, GD: Gradient Descent). Note that all the algorithms have an initial step of finding the optimal rank r subspace for each group, in which its time complexity is not included here.

Notwithstanding these flourishing efforts for fairness problem in supervised learning, fairness in unsupervised learning tasks has not been explored thoroughly. This is despite the fact that unsupervised learning tasks such as dimension reductions are mostly preceding those supervised ones, in the training procedures. Hence, having fair unsupervised learning models is as crucial as supervised ones. For instance, Principal Component Analysis (PCA) is widely used to reduce the dimension of the data before applying classification models. In addition to that, these unsupervised methods such as dimension reductions or clustering methods are commonly used for data visualizations, identifying common behaviors or trends, reduce the size of data, to name but a few. This ubiquitousness of unsupervised methods in machine learning models can affect decision-making systems if they unfairly treat different groups in data.

In this paper, we aim at defining a fairness measure for dimension reduction algorithms like PCA and propose an algorithm to enforce these fairness criteria. It is important to note that, despite supervised learning that fairness metrics are mostly focused on the beneficial outcome (usually the positive label), in an unsupervised task, there is no label to be used. Hence, we seek to find a subspace that is "good" enough for each protected group in the data. Indeed, when we apply PCA on a dataset, the resulting subspace found by a standard algorithm is different from what we achieve when only using the data of each group individually. This difference can be reflected as the difference between the reconstruction error of each group's data on both subspaces. Thus, when a dimension reduction algorithm is applied to the joint data, the reconstruction loss of some of the groups is degraded (from what they can achieve with their own data only), while others are benefiting from joint learning. In this regime, a fair algorithm is the one that can find a subspace with optimal trade-offs between these degradations and benefits.

An attempt to impose the fairness constraint on learning the optimal subspace for two protected groups has been made recently in Samadi et al. (2018); Olfat and Aswani (2018); Morgenstern et al. (2019), where the fair subspace learning is sought by minimizing the maximum *deviation of reconstruction error* suffered by any protected group (i.e., the difference of per group reconstruction error and joint reconstruction error). Interestingly, it has been shown that at any optimal local solution of optimization problem associated with learning such a fair subspace, all the groups suffer the same loss. Motivated by this observation, a semi-definite programming relaxation followed by a linear programming is proposed to find a fair subspace (Samadi et al., 2018; Olfat and Aswani, 2018; Morgenstern et al., 2019). In addition to the computational inefficiency of algorithms proposed by these works, the generalization of them to multiple group sensitive features is not conspicuous. Furthermore, since all optimal solutions do not assign same loss to all groups, extra dimensions are needed to ensure that the total loss of the projection remains at most the optimal objective in the original target dimension (in particular, $k-1$ extra dimensions are needed for k groups in Samadi et al. (2018) which is further tightened to $\sqrt{k}$ in a followup work Morgenstern et al. (2019)).

The overarching goal of this paper is to define a fairness metric for dimension reduction, dubbed as

pairwise disparity error, and propose a computationally efficient dimensional reduction algorithm to learn a fair subspace from multiple group sensitive features. Towards this end, we cast the problem of fairness in PCA dimension reduction algorithm as a multi-objective optimization problem and propose an adaptive gradient descent based approach to find the optimal trade-offs with provable convergence rates. Interestingly, the proposed framework is not bound to any specific notion of fairness metric and can be effortlessly applied to other metrics as well. Moreover, unlike the aforementioned prior works, no extra dimension is needed to ensure the loss suffered by each group matches the optimal fairness loss. The comparison of time complexity of exiting algorithms and current work is summarized in Table 1.

Contributions The main contributions of this paper can be summed up as follows:

- We introduce the notion of Pareto fair PCA to ponder conflicting objectives and achieve optimal trade-offs between them. Also, we introduce the notion of **pairwise disparity error** as a more efficient objective to learn fair subspaces. In addition, we provide conditions, under which a Pareto optimal solution exists.
- We propose a gradient descent algorithm to efficiently solve the obtained multi-objective optimization problem which is interesting by its own right, and provide theoretical guarantees on its convergence to optimal compromises or a Pareto stationary point.
- We empirically develop this algorithm and compare it to the state-of-the-art algorithm on two real-world datasets to demonstrate its efficacy that complements our theoretical results.

2 Related Work

The efforts to address fairness in algorithmic decision-making systems have roughly fallen into three different categories. Some scholars believe data itself could be biased, leading to unfair results; thus, they seek to solve this problem on data level and as a preprocessing step to the main learning task (Dwork et al., 2012; Feldman et al., 2015; Kamiran and Calders, 2009; Calders et al., 2009). The goal is achieved by either changing the value of sensitive feature or label data or find a subspace, where labels and sensitive features are independent. However, since the main objective of the learning is not involved in this process, the optimal solution for the main objective is not guaranteed. The second category includes methods that try to impose the fairness criteria after the learning, in order to attain a fair model (Hardt et al., 2016; Kamishima et al., 2011; Goh et al., 2016; Calders and Verwer, 2010). The third approach, includes methods that try to satisfy fairness constraint during the training procedure, usually by imposing them as a constraint to the main learning objective (Donini et al., 2018; Morgenstern et al., 2019; Zafar et al., 2015; Samadi et al., 2018; Pleiss et al., 2017). Some of these approaches treat the fairness problem similar to imbalanced data or rare event prediction (Yao and Huang, 2017; Kamani et al., 2019, 2018, 2016; Nikbakht and Papakonstantinou, 2019). While these approaches can achieve the state-of-the-art results in some problems, they still suffer from several issues. Solving a constrained optimization could be a very hard non-convex problem; hence, relaxation is needed to solve the problem that leads to sub-optimal solutions efficiently. Moreover, finding the optimal penalization parameter could be a difficult task, as discussed in Donini et al. (2018). Our approach belongs to the third category, yet, it differs from the prevailing trend of formulating the fairness problem as a constrained optimization. We will cast the fairness problem as a multi-objective optimization that can efficiently satisfy fairness objectives as well as the main learning objective and converge to a point with optimal compromises between objectives.

Fairness in dimension reduction algorithms is recently being vetted by Samadi et al. (2018), through which they propose a semi-definite programming and prove that its solution satisfies the proposed notion of fairness. Aside from the inefficiency of solving the SDP, they approach is developed for binary sensitive features and requires one extra dimension to guarantee fairness. To generalize it for multiple group sensitive features with k groups, they propose to add $k-1$ dimensions, which is impractical. The follow-up studies by Olfat and Aswani (2018); Morgenstern et al. (2019) are still in line with the previous one, trying to relax and solve

an SDP. We, on the other hand, propose an efficient gradient-based method to solve the aforementioned multi-objective optimization, with the capability of generalizing to multiple group sensitive features smoothly.

Although it has been asserted that fairness problems are multi-objective problems in nature (Kearns and Roth, 2019; Lipton et al., 2017; Morgenstern et al., 2019), as mentioned before, most of the existing works, apply different forms of relaxations and approximations to reduce the problem into a scalar-valued optimization problem. In this paper, we design the fairness problem at hand as a multi-objective optimization and solve it directly. Multi-objective or vector optimization is a well-studied problem in different domains for many years. The goal in this optimization is to achieve an optimal trade-off point between different objectives, known as Pareto optimal, named after Italian economist Vilfredo Pareto. We refer the reader to Miettinen (2012) and the references therein as a rich resource on multi-objective optimization. We will elaborate that directly solving the vector-valued problem associated with fair learning is appealing to reduction based counterparts (Ehrgott, 2006; Mahdavi et al., 2013) by being computationally efficient and providing provable guarantees on the fairness metric.

3 Problem Formulation

We start by mathematically defining the problem we ought to solve, and then discuss what is the notion of fairness in PCA algorithm, which could be quite different from what is known as fairness measures in supervised learning. In what follows we adapt the following notation. We use bold face upper case letters such as $\mathbf{X}$ to denote matrices and bold face lower case to denote vectors such as $\mathbf{f}$. The Frobenius norm and trace of a matrix $\mathbf{X}$ are denoted by $\|\mathbf{X}\|_{\mathrm{F}}$ and $\mathsf{tr}\,(\mathbf{X})$, respectively. The eigenvalues of a positive semi-definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ are denoted by $\gamma_{\max}(\mathbf{\Sigma}) = \gamma_1(\mathbf{\Sigma}) \geq \gamma_2(\mathbf{\Sigma}) \geq \ldots \geq \gamma_d(\mathbf{\Sigma}) = \gamma_{\min}(\mathbf{\Sigma})$. The set of integers, $\{1, 2, \ldots, m\}$, is represented by $[m]$.

3.1 PCA

The main objective of the PCA is to find the best representation of the data $\mathbf{X} \in \mathbb{R}^{n\times d}$ with n data points in d-dimensional space, in a lower dimension $r \leq d$ using a linear transformation, in order to have the minimum reconstruction error. This linear transformation can be represented by a projection matrix $\mathbf{U} \in \mathbb{R}^{d\times r}$. Thus, the objective of PCA is to find a projection matrix $\mathbf{U}$ and a recovery matrix $\mathbf{W} \in \mathbb{R}^{r\times d}$ to minimize this reconstruction error similar to Shalev-Shwartz and Ben-David (2014):

$$\underset{\mathbf{U}\in\mathbb{R}^{d\times r},\mathbf{W}\in\mathbb{R}^{r\times d}}{\arg\min} \|\mathbf{X} - \mathbf{X}\mathbf{U}\mathbf{W}\|_{\mathrm{F}}^2 \tag{1}$$

It can be proved that in the solution of (1), we have $\mathbf{W} = \mathbf{U}^\top$, and columns of $\mathbf{U}$ are orthonormal (i.e. $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_{r\times r}$). Therefore we can define the reconstruction loss for any PCA projection as follows:

Definition 1 (Reconstruction Loss)**.** *For any given dataset* $\mathbf{X}$ *and any projection matrix* $\mathbf{U}$*, the total reconstruction loss of* $\mathbf{X}$ *using* $\mathbf{U}$ *is defined as:*

$$\mathcal{L}(\mathbf{U}) \triangleq \ell(\mathbf{X}; \mathbf{U}) = \left\|\mathbf{X} - \mathbf{X}\mathbf{U}\mathbf{U}^\top\right\|_{\mathrm{F}}^2 \tag{2}$$

PCA minimizes the above loss, in order to find the optimal subspace with minimum reconstruction loss given $\mathbf{X}$.

3.2 Fair PCA

In this section, we will formally define the notion of fairness in dimension reduction algorithms such as PCA. As it was discussed before, the problem arises from having different reconstruction losses on different sensitive groups in a dataset. This means that finding an optimal projection matrix $\mathbf{U}^*$ by solving the minimization problem in (2), would have different reconstruction loss on data partitions from each sensitive group. However, in this problem, unlike supervised problems previously discussed, we are not able to reach equality between

these reconstruction losses for different groups. The reason for that is the subspace for each group's data is different, and so is the reconstruction error of that data for that projection. We note that while learning a separate (local) subspace for each individual group has the optimal reconstruction error, our focus here is to learn a single global subspace for all groups due to statistical and ethical concerns. In particular, from a statistical standpoint, since the number of training samples for some groups might be small for skewed data sets, joint learning to have more samples to learn a subspace is preferable. Ethically, as elaborated in (Lipton et al., 2017; Kannan et al., 2019), learning separate subspaces (having disparate treatment like in affirmative action) constructs no trade-offs, and it poses several ethical and legal concerns.

In order to quantify to what extent each group suffers or benefits from joint subspace learning, we should compare the subspaces learned from each group's data alone and the one with other groups' data included. Then, the idea of fairness is to reach a balance between these sacrifices and benefits of different groups. Formally, consider one of the d features of $\mathbf{X}$ as a sensitive feature with k different groups, $\mathcal{S} = \{s_1, \ldots, s_k\}$. We denote the matrix of each group's data points as $\mathbf{X}_i \in \mathbb{R}^{n_i \times d}$, where n_i is the number of samples belonging to the sensitive group s_i. Hence for any projection matrix $\mathbf{U} \in \mathbb{R}^{d \times r}$, the reconstruction loss for each group is defined as:

$$\mathcal{L}_i(\mathbf{U}) \triangleq \ell(\mathbf{X}_i; \mathbf{U}) = \left\| \mathbf{X}_i - \mathbf{X}_i \mathbf{U}\mathbf{U}^\top \right\|_{\mathrm{F}}^2, \quad 1 \leq i \leq k. \tag{3}$$

Then, if we only use the dataset $\mathbf{X}_i$ to learn the projection matrix, we can find the subspace represented by $\mathbf{U}_i^*$ that has the optimal reconstruction loss on that dataset, denoted by $\mathcal{L}_i(\mathbf{U}_i^*)$. Therefore, a fair dimension reduction algorithm is the one that can learn a global projection matrix $\mathbf{U}^*$ on all data points with having equal distance between each groups reconstruction loss on the subspace learned by the whole data with the subspace learned only by its own data. To formally define this fairness criteria, we introduce the notion of **disparity error** as follows:

Definition 2 (Disparity Error)**.** *Consider a dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$ with k sensitive groups with data matrix $\mathbf{X}_i, i = 1, 2, \ldots, k$ representing each sensitive group's data samples. Let $\mathbf{U}_i^* = \arg\min_{\mathbf{U}} \mathcal{L}_i(\mathbf{U})$ denote the projection matrix learned only based on $\mathbf{X}_i$. Then for any projection matrix $\mathbf{U}$ the disparity error for each sensitive group is defined as:*

$$\mathcal{E}_i(\mathbf{U}) = \mathcal{L}_i(\mathbf{U}) - \mathcal{L}_i(\mathbf{U}_i^*), \quad 1 \leq i \leq k. \tag{4}$$

This measure shows that how much reconstruction loss we are suffering or enjoying for any global projection matrix $\mathbf{U}$, with respect to the reconstruction loss of optimal projection matrix we can learn locally based on data points $\mathbf{X}_i$. Note that, calculating the optimal rank r subspace for each group in $\mathcal{L}_i(\mathbf{U}_i^*)$ has an one-time overhead to the algorithm's time complexity overall. However, we ignore this overhead, as did other algorithms we are comparing to and leave the joint learning of both local and global subspaces as a future work.

Using the Definition 2, we can define a fair PCA algorithm as follows:

Definition 3 (Fair PCA)**.** *A PCA algorithm with projection matrix $\mathbf{U}^*$ is called **fair**, if the disparity error among different groups are equal. That is:*

$$\mathcal{E}_1(\mathbf{U}^*) = \mathcal{E}_2(\mathbf{U}^*) = \ldots = \mathcal{E}_k(\mathbf{U}^*). \tag{5}$$

A subspace $\mathbf{U}^*$ that archives same disparity error for all groups is called a fair subspace.

4 Pareto Fair Subspace

In this section we discuss the key challenges in finding a fair subspace using relaxation methods and motivate our formulation of Pareto fair subspace followed by providing conditions sufficient to guarantee the existence of such subspaces.

4.1 Relaxation methods and their limitations

A major challenge to find a fair subspace as defined in Definition 3 is to solve the optimization problem that satisfies (5), which is essentially a multiple objective optimization problem by nature. To illustrate this and for ease of exposition, let us focus on binary sensitive feature ($k = 2$), i.e., there are only two groups in the sensitive feature of the data (e.g. male and female), in which the goal of fair PCA is to satisfy:

$$\mathcal{E}_1(\mathbf{U}^*) = \mathcal{E}_2(\mathbf{U}^*), \tag{6}$$

In Samadi et al. (2018), it has been shown that by casting the multiobjective optimization problem as a minmax problem of the form

$$\min_{\mathbf{U}\in\mathbb{R}^{d\times r}, \mathtt{rank}(\mathbf{U})\leq r} \max\{\mathcal{E}_1(\mathbf{U}), \mathcal{E}_2(\mathbf{U})\}, \tag{7}$$

and using an additional dimension for the projection, the optimal solution of minmax problem results in the same loss for both groups (i.e., $\mathcal{E}_1(\mathbf{U}^*) = \mathcal{E}_2(\mathbf{U}^*)$).

Motivated by this observation, a semi-definite relaxation to solve the optimization problem is proposed, which is not efficient for a large number of training samples. Also, to achieve their fairness criteria and ensure that the obtained local solution achieves the optimal fairness objective for all groups, the proposed solution requires adding an extra dimension for a binary sensitive feature and $k-1$ additional dimensions for a k-group sensitive feature, which is not reasonable for a large k. We note that in Morgenstern et al. (2019), the requirement of extra dimension is improved to $\lfloor\sqrt{2k+\frac{1}{4}}-\frac{3}{2}\rfloor$, but it still requires extra projection dimensions to satisfy the fairness constraint. Finally, the optimal trade-offs between fairness objectives and total reconstruction loss, in the case of the same target dimension, is not guaranteed, which would lead to a solution that sacrifices too much of the total reconstruction loss to achieve the fairness criteria. In fact, in order to guarantee that the solution of minmax optimization results in a rank r subspace with *optimal fairness* objective, one could choose the target dimension to be $r-s$, where s is the extra dimensions needed with $\mathcal{O}\left(\sqrt{k}\right)$ followed by a rounding to reach to the target r dimensional subspace as proposed in Morgenstern et al. (2019); but this remedy hurts the *optimal objective* value by a multiplicative factor of s/r. This issue becomes more concerning as the number of groups k, and hence s increases due to the fact that all local optimal solutions might not achieve the same loss for all groups. Consequently, any solution to fair PCA necessitates jointly minimizing the main objective, which is total reconstruction loss in (2), and fairness criteria to balance the trade-off between them.

An alternative solution to alleviate aforementioned issues which is explored in Donini et al. (2018) is to impose fairness constraints in minimizing the reconstruction loss in (1) as additional constraints, i.e.,

$$\begin{aligned} \min_{\mathbf{U}} \quad & \mathcal{L}(\mathbf{U}) \\ \text{subject to} \quad & \mathcal{E}_i(\mathbf{U}) \leq \epsilon,\ i \in [k]. \end{aligned} \tag{8}$$

which reduces the problem into an instance of *non-convex constrained* optimization problem to find a fair subspace to all sensitive groups. Relaxing the problem of finding the fair subspace as a constrained optimization similar to (8), apart from being a hard non-convex problem which is not evident to solve due to presence of non-convex constraints, requires the optimal constraint violation parameter, $\boldsymbol{\epsilon}$, to be decided heuristically which is a burden on the use and makes the problem even harder. Although using Lagrangian method we can turn the problem into a unconstrained non-convex optimization problem– a method known as scalarization relaxation for multi-objective optimization counterpart (e.g., please see Ehrgott (2006)), deciding the Lagrangian multipliers is as hard as solving the original problem and does not guarantee the optimally of obtained solution. Also, since the scale of the objectives might be different, it could lead to infeasiblity issues in the optimization problem, or some points from the Pareto frontier could not be attained.

To address challenges arise from the above reduction methods, and in order to achieve the optimal trade-offs between objectives and satisfy equality between disparity errors, we aim at directly solving the multi-objective programming (Miettinen, 2012). Towards this end, we note that the optimization problem in

(8) is a relaxation of the following generalized multi-objective optimization problem:

$$\arg\min_{\mathbf{U}} \left[\mathcal{L}(\mathbf{U}), \psi\left(\mathcal{E}_1\left(\mathbf{U}\right)\right), \ldots, \psi\left(\mathcal{E}_k\left(\mathbf{U}\right)\right)\right] \tag{9}$$

where $\psi(\cdot) : \mathbb{R} \to \mathbb{R}_+$ is any penalization function, such as $\psi(z) = |z|$, $\psi(z) = \frac{1}{2}z^2$, or $\psi(z) = e^{-z}$, however, for convergence analysis we will stick to squared or exponential penalization due to their smoothness. We will define the optimization problem in more detailed in the next section and then will introduce an adaptive gradient descent approach to solve it.

4.2 Pareto fair subspace

To characterize the solutions obtained by directly solving the multi-objective optimization problem in (9), we have to compare the objective vector of different solutions with each other, analogous to the what we do in a scalar or single-objective optimization problem. If we only have a single objective function $f(\mathbf{U})$, we can say the solution $\mathbf{U}_1$ is better than $\mathbf{U}_2$ if $f(\mathbf{U}_1) < f(\mathbf{U}_2)$. Similarly, in a multi-objective programming, we define the notion of dominance as follows:

Definition 4 (Dominance)**.** *Let* $\mathbf{f}(\mathbf{U}) = [f_1(\mathbf{U}), \ldots, f_m(\mathbf{U})]^\top$ *denote a vector-valued objective function with* m *objectives. We say the solution* $\mathbf{U}_1$ *dominates the solution* $\mathbf{U}_2$ *if* $f_i(\mathbf{U}_1) \leq f_i(\mathbf{U}_2)$ *for all* $i \in [m]$*, and* $f_j(\mathbf{U}_1) < f_j(\mathbf{U}_2)$ *for at least one* $j \in [m]$*. We denote this dominance as:*

$$\mathbf{f}(\mathbf{U}_1) \prec_p \mathbf{f}(\mathbf{U}_2). \tag{10}$$

The definition of dominance implies that when a solution cannot be dominated by any other solution in the search space, we cannot find any direction, to move to, from this solution without at least hurting one objective in the objective vector. Using this, now, we can define our notion of Pareto fair subspace as follows:

Definition 5 (Pareto Fair Subspace)**.** *Let* $\mathbf{f}(\mathbf{U}) = [\mathcal{L}(\mathbf{U}), f_1(\mathbf{U}), \ldots, f_{m-1}(\mathbf{U})]^\top$ *denote a vector-valued objective function with* m *objectives in the Fair PCA problem. Then, consider a set of fairness trade-off objectives* $f_i(\mathbf{U}), i \in [m-1]$*, (e.g.* $\psi\left(\mathcal{E}_i\left(\mathbf{U}\right)\right)$ *as in (9)) that ought to be minimized in addition to the main objective,* $\mathcal{L}(\mathbf{U})$*. The solution* $\mathbf{U}^*$ *is called Pareto fair subspace, if it is not dominated by any other feasible solution.*

The Pareto fair subspace is not unique, and the set of Pareto optimal solutions is called Pareto frontier (Miettinen, 2012). Thereupon, the ultimate goal of fair PCA solution reduces to finding a Pareto optimal solution via solving the problem (9), to fairly project the data points from all groups.

We can show that a Pareto fair solution always exists under some assumptions. Comparing to Lagrangian method of multipliers or other scalarization approaches, we aim at finding this Pareto fair frontier completely without any prior information such as weight for each objective. The following theorem establishes the conditions under which the set of Pareto optimal solutions is non-empty.

Theorem 1 (Existence)**.** *Consider the vector-valued optimization problem in (9). If the individual functions are convex and bounded, then the set of Pareto optimal solutions is non-empty.*

Proof. The proof is deferred to Appendix A. □

To guarantee the existence of a Pareto optimal solution, in Section 5 we convexify the objectives by properly regularizing them. Thereafter, we propose an efficient gradient based algorithm to find a subspace that is a Pareto stationary point of the fair PCA problem.

Although solving the optimization problem in (9) results in an efficient trade-off between different objectives, this does not reflect on balanced disparity errors among different groups, which is the ultimate goal of the fair PCA problem. As been asserted by Samadi et al. (2018), this issue would be exacerbated in problems with $k > 2$, that having a balanced disparity error among all groups is not always possible due to the fact that all optimal solutions will not assign the same loss to all groups. To alleviate this issue and ensure that the loss of each group remains at most the optimal fairness objective in original target dimension r, we introduce the notion of **pairwise disparity error**, that would address this issue.

Definition 6 (Pairwise Disparity Error)**.** *Consider the disparity errors for any projection matrix* $\mathbf{U}$ *and sensitive groups of* i *and* j *among* k *different groups, then the pairwise disparity error between these two groups is defined as:*

$$\Delta_{i,j} = \mathcal{E}_i(\mathbf{U}) - \mathcal{E}_j(\mathbf{U}), \quad i, j \in [k], i \neq j. \tag{11}$$

Thus, the optimization in (9) becomes:

$$\arg\min_{\mathbf{U}} \left[\mathcal{L}(\mathbf{U}), \psi\left(\Delta_{1,2}(\mathbf{U})\right), \ldots, \psi\left(\Delta_{k-1,k}(\mathbf{U})\right)\right] \tag{12}$$

where we have $\binom{k}{2}$ objectives in addition to the main objective. We will show the efficacy of pairwise disparity error over single disparity error in practice in Section 6.

5 Adaptive Optimization

In this section, we will develop a gradient based algorithm to solve the optimization problems in (9) or (12). To lay the groundwork for this algorithm, we first review how to solve the original PCA problem using gradient descent and then we propose our gradient based algorithm to solve the aforementioned multi-objective problem.

5.1 Gradient descent for PCA

To solve the PCA problem using the gradient descent approach, we need to iteratively update the projection matrix $\mathbf{U}$, based on the gradient of the total reconstruction loss with respect to it. Expanding the total reconstruction loss in (2) and removing the constant terms that will not affect the optimization, following Shalev-Shwartz and Ben-David (2014), we can write the optimization problem:

$$\arg\min_{\mathbf{U}\in\mathbb{R}^{d\times r}, \mathbf{U}^\top\mathbf{U}=\mathbf{I}} - \mathsf{tr}\left(\mathbf{U}^\top\mathbf{X}^\top\mathbf{X}\mathbf{U}\right), \tag{13}$$

Using (13), we can calculate the gradient of the total reconstruction loss with respect to $\mathbf{U}$ as follows:

$$\mathcal{G}(\mathbf{U}) = \frac{\partial\mathcal{L}(\mathbf{U})}{\partial\mathbf{U}} = -2\mathbf{X}^\top\mathbf{X}\mathbf{U}. \tag{14}$$

The projection can be learned using the gradient descent by iteratively updating an initial solution by:

$$\mathbf{U}_{t+1} = \Pi_{\mathcal{P}_r}\left(\mathbf{U}_t - \eta_t\mathcal{G}(\mathbf{U}_t)\right), \tag{15}$$

where η_t is the learning rate and $\Pi_{\mathcal{P}_r}(.)$ is the projection operator onto $\mathcal{P}_r = \left\{\mathbf{U}\in\mathbb{R}^{d\times r} \,\middle|\, \mathbf{U}^\top\mathbf{U} = \mathbf{I}_r\right\}$.

For a single-objective optimization like normal PCA, at each iteration we take a step toward the negative of the gradient at that point. However, when we are dealing with multiple objectives, the key question is what would be the best direction at each iteration to take, in order to decrease all the objectives. We answer this question in the next section by proposing an optimization problem to find such a descent direction.

5.2 Pareto fair PCA

In order to efficiently solve the multi-objective optimization problem in (9) or (12), we propose a gradient descent approach, that can guarantee convergence to a Pareto stationary point. For the ease of exposition, we consider the following general multi-objective problem with m objectives:

$$\mathbf{f}(\mathbf{U}) = [f_1(\mathbf{U}), \ldots, f_m(\mathbf{U})]$$

In a single-objective problem with gradient descent method, we always choose the opposite direction of the gradient on that point as the descent direction to decrease the objective function for the next iteration

Algorithm 1 Pareto Fair PCA

input The target projection dimension r, $\mathbf{X} = \mathbf{X}_1 \cup \mathbf{X}_1 \cup \ldots \cup \mathbf{X}_k$, $\mathbf{U}_0 \in \mathbb{R}^{d\times r}$, T
Find the optimal r-rank subspace for each group, $\mathcal{U}^* = \{\mathbf{U}_1^*, \ldots, \mathbf{U}_k^*\}$ ▷ e.g., using SVD or SGD (Shamir, 2015)
procedure ParetoFairPCA(r, $\mathbf{X}$, $\mathcal{U}^*$, $\mathbf{U}_0$, T)
    Form the objective vector $\mathbf{f}(\mathbf{U})$ using $\mathcal{U}^*$ and $\mathbf{X}$ from (9) or (12)
    for $t = 1, \ldots, T$ **do**
        Calculate the gradient of each objective, $\mathbf{G}_i^{(t)}$
        Find the descent direction $\mathbf{D}^{(t)}$ using (17)
        if $\mathbf{D}_t = \mathbf{0}$ **then** ▷ Pareto stationary point
            return $\mathbf{U}_t$
        Find minimum $p \in \mathbb{N}$ such that for $\eta_t = \frac{1}{2^p}$: ▷ Backtracking line search (`optional`)

$$f_i\left(\mathbf{U}_t + \eta_t \mathbf{D}_t\right) \leq f_i(\mathbf{U}_t) + \beta \eta_t \operatorname{tr}\left(\mathbf{D}_t^\top \mathbf{G}_i^{(t)}\right), \; i \in [m] \tag{16}$$

        $\mathbf{U}_{t+1} = \Pi_{\mathcal{P}_r}\left(\mathbf{U}_t + \eta_t \mathbf{D}_t\right)$
return $\mathbf{U}_{T+1}$

point. However, this notion in multi-objective programming is more complicated, as we have to find the direction that is a descent direction for all objectives based on their gradients on that point. In order to find a descent direction, let $\left[\mathbf{G}_1^{(t)}, \ldots, \mathbf{G}_m^{(t)}\right]$ denote the gradient of individual objectives at point $\mathbf{U}_t$. To find a descent direction with respect to all of the objectives at point $\mathbf{U}_t$, we solve the following minmax optimization problem (Fliege and Svaiter, 2000):

$$\mathbf{D}_t = \arg\min_{\mathbf{D}\in\mathbb{R}^{d\times r}} \left\{ \max_{i=1,\ldots,m} \operatorname{tr}\left(\mathbf{D}^\top \mathbf{G}_i^{(t)}\right) + \frac{1}{2}\|\mathbf{D}\|_{\mathrm{F}}^2 \right\}. \tag{17}$$

We note that for a single objective case, that is $m = 1$, the solution of above mimimax is the opposite of the gradient, i.e. $\mathbf{D}_t = -\mathbf{G}_1^{(t)}$. Using KKT optimally conditions, it is easy to show that the dual problem becomes a quadratic programming, and can be efficiently solved to identify a descent direction $\mathbf{D}_t$, for which all the objectives are non-increasing. The following lemma states this characteristic of the descent direction:

Lemma 1 (Descent Direction)**.** *The solution found in optimization problem (17) has one of the following two conditions. Either $\mathbf{D}_t = \mathbf{0}$, which means the point $\mathbf{U}_t$ is a Pareto stationary point, or $\mathbf{D}_t$ is a descent direction to all objectives, that is:*

$$\operatorname{tr}\left(\mathbf{D}_t^\top \mathbf{G}_i^{(t)}\right) \leq 0, \quad \forall\, 1 \leq i \leq m \tag{18}$$

Then, the obtained descent direction is in the form of $\mathbf{D}_t = -\sum_{i=1}^m \lambda_i^{(t)} \mathbf{G}_i^{(t)}$, where $\sum_{i=1}^m \lambda_i^{(t)} = 1$ and $\lambda_i^{(t)} \geq 0$ for $1 \leq i \leq m$.

Proof. The proof is provided in Appendix B. □

As elaborated in the proof in Appendix B, the theorem implies that the descent direction is the minimum norm matrix in the convex hull of the gradients of all objectives and is non-increasing direction with respect to each objective. Understanding this, the following corollary is palpable:

Corollary 1. *The first order Pareto stationary point holds for a solution $\mathbf{U}$ when the mentioned minimum norm is zero, i.e. there is no descent direction that is non-increasing for all objectives. In other words, there exists a $\boldsymbol{\lambda} \in \Delta_m$ such that $\mathbf{D} = -\sum_{i=1}^m \lambda_i \mathbf{G}_i = \mathbf{0}$ where $\mathbf{G}_i = \nabla f_i(\mathbf{U})$.*

Having a descent direction at hand, we can use it to decrease all the objectives in every iteration, similar to the procedure defined in Algorithm 1. Based on the first-order optimality condition of this problem, we know that at a Pareto optimal solution, the direction found in (17) should be $\mathbf{0}$, meaning, that it cannot further improve any objective without hurting others. Equipped with this descent direction and first-order optimality condition, we can iteratively update the initial solution in the direction of the descent direction, until it converges to a Pareto stationary point.

Remark 1. *One crucial step before finding the descent direction is to balance out the scale of different gradients. Since they are calculated based on very different and possibly contradictory objective functions, their Frobenius norm would vary a lot; hence, by a normalization step, we can avoid the dominance of the descent direction by some gradients with high Frobenius norm.*

Since the disparity errors, as well as the main PCA objective, are weakly convex functions, following Theorem 1, to guarantee the existence of Pareto optimal subspace, we add a regularization term to each objective to make them convex functions– with which we also stabilize the solutions and guarantee convergence. As a result, the optimization in (12) becomes:

$$\arg\min_{\mathbf{U}} \begin{bmatrix} \mathcal{L}(\mathbf{U}) + \alpha\|\mathbf{U}\|_{\mathrm{F}}^2 \\ \psi\left(\Delta_{1,2}\left(\mathbf{U}\right)\right) + \alpha\|\mathbf{U}\|_{\mathrm{F}}^2 \\ \vdots \\ \psi\left(\Delta_{k-1,k}\left(\mathbf{U}\right)\right) + \alpha\|\mathbf{U}\|_{\mathrm{F}}^2 \end{bmatrix}, \tag{19}$$

where α is the regularization parameter to make the Hessian matrices of objectives positive semi-definite and needs to be decided based on the maximum eigen-gap between covariance matrices of each pair of sensitive groups. Having k different groups, each with data matrix of $\mathbf{X}_i$, $i \in [k]$, we set $\gamma = \max_{i,j\in[k]} \gamma_d\left(\mathbf{X}_i^\top \mathbf{X}_i\right) - \gamma_1\left(\mathbf{X}_j^\top \mathbf{X}_j\right)$ to denote the maximum eigen-gap. Then, we should have $\alpha \geq \gamma$. We now turn to prove the convergence rate of Algorithm 1 for convex objectives as stated in the following theorem.

Theorem 2 (Convex Convergence)**.** *Let $\mathbf{f} = [f_1(\mathbf{U}), \ldots, f_m(\mathbf{U})]$ be convex component-wise Lipchitz continuous with constants $L_1, L_2, \ldots, L_m$. Then, for the sequence of the solutions $\mathbf{U}_1, \ldots, \mathbf{U}_T$ generated iteratively by Algorithm 1, and the sequence of $\hat{\boldsymbol{\lambda}}^{(1)}, \ldots, \hat{\boldsymbol{\lambda}}^{(T)}$ generated by (37) during T iterations, by setting $\eta = \frac{R}{L\sqrt{T}}$ and $\beta = \sqrt{T}/R$, we have:*

$$\sum_{i=1}^{m} \bar{\lambda}_i \left(f_i(\mathbf{U}_T) - f_i(\mathbf{U}^*)\right) \leq \frac{RL}{2\sqrt{T}}, \tag{20}$$

where $R^2 = \|\mathbf{U}_1 - \mathbf{U}^*\|_{\mathrm{F}}^2$, $L = \max_{i=1,\ldots,m} L_i$, $\bar{\lambda}_i = \frac{1}{T}\sum_{t=1}^{T} \hat{\lambda}_i^{(t)}$, and $\mathbf{U}^*$ is a Pareto efficient solution.

Proof. The detailed proof is deferred to Appendix C. □

Theorem 2 indicates that, using the Pareto descent direction, we can achieve an ϵ-accurate Pareto efficient solution with taking $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ gradient descent steps. Using (20), we can bound the average deviation of each objective from its respective value in the Pareto efficient solution of $\mathbf{U}^*$.

Remark 2. *We note that Algorithm 1 is guaranteed to converges to a single Pareto fair subspace, starting from fixed initial solution $\mathbf{U}_0$. Using different random starting points we can find different Pareto fair subspaces, and form the Pareto fair frontier of the problem. From an algorithmic point of view, we can not distinguish between different Pareto optimal subspaces, but as discussed by Kearns and Roth (2019), based on the preference of different objectives we can choose a desirable Pareto fair subspace from the frontier set.*

We note that when the regularization is not added to convexify the main objective, we have to deal with non-convex objectives in the optimization problem. In the following theorem, we investigate the convergence of Algorithm 1 for non-convex objectives that guarantees the gradient vanishes over iterations.

Theorem 3 (Nonconvex Convergence). *Let* $\mathbf{f}(\mathbf{U}) = [f_1(\mathbf{U}), \ldots, f_m(\mathbf{U})]$ *be the multi-objective function to be minimized to find a fair subspace with respect to* k *sensitive groups. Let* $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_T$ *be the sequence of solutions generated by Algorithm 1 updated using descent directions* $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_T$*. Then, if we choose the regularization parameter as* $\alpha \geq \gamma$*, we have the following:*

$$\min_{t=1,2,\ldots,T} \|\mathbf{D}_t\|_{\mathrm{F}} \leq \sqrt{\frac{\mathsf{M}_u - \mathsf{M}_l}{CT}} \tag{21}$$

where M_l *is a lower bound for the values of all objective functions,* M_u *is the maximum of the values of all functions at initial point, and* C *is a constant depending on the smoothness of objectives.*

Proof. The proof can be found in Appendix D. □

An immediate consequence of above theorem is that the gradient of Pareto descent directions vanishes and convergences to zero and thereby the solutions generated by the algorithm convergence to a stationary fair subspace. In particular, only $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ iterations are required to obtain an ϵ-close fair subspace. The analysis of Theorem 3 follows the standard analysis of gradient descent for non-convex smooth optimization where the obtained bound matches the known achievable convergence rate for the norm of the gradients. We would like to sketch another alternative method that results in the same rate with careful analysis. Specifically, observe that the descent direction can be considered as an inexact gradient from the viewpoint of individual functions with perturbation $\mathbf{D}_t - \mathbf{G}_i^{(t)}$. Noting that $\mathsf{tr}\left(\mathbf{D}_t^\top \mathbf{G}_j^{(t)}\right) \leq -\|\mathbf{D}_t\|_{\mathrm{F}}^2$ as shown in the proof of Lemma 1 and following the standard analysis of convergence of non-convex functions, we can show that norm of descent directions vanishes as algorithm proceeds, thereby the proposed algorithm can find a stationary point. However, the obtained solution is not guaranteed to be an optimal Pareto due to non-convexity of the objectives and might be a saddle point.

5.3 Comparison with other approaches

As it was discussed, one approach to solve a multi-objective optimization is to make it constrained optimization, in which we keep the main objective and change all other objectives to inequality constraints with parameters $\boldsymbol{\epsilon}$. Hence, constrained optimization is a relaxation of multi-objective optimization, where finding the best constraint parameter ($\boldsymbol{\epsilon}$) for each constraint could be very challenging as discussed in Donini et al. (2018). It also lacks theoretical guarantees due to the non-convex nature of constraints. Lagrangian method of multipliers is equivalent to constraint optimization problems, but not exactly to multi-objective counterpart. To see this, we note that by applying GD to Lagrangian function, the contribution of the gradient of each individual function, $\mathbf{G}_i^{(t)}$, is weighted by its Lagrangian multiplier, while in our case the weights are adaptively learned by finding a Pareto decent direction. We note that while Morgenstern et al. (2019) improves the requirement of extra dimensions over Samadi et al. (2018), it still needs $\lfloor\sqrt{2k + \frac{1}{4}} - \frac{3}{2}\rfloor$ extra dimensions for a k-group sensitive feature and has to solve an SDP which has the time complexity of $\mathcal{O}\left(d^{6.5}\log\left(\frac{1}{\epsilon}\right)\right)$ or $\mathcal{O}\left(d^3/\epsilon^2\right)$ with multiplicative weight update. On the other hand, our method enjoys the efficiency of GD with an overhead to solve the quadratic problem over the simplex for finding the descent direction. Also at each step, we need to project the solution to find the orthonormal bases for the updated solution which could be done using SVD with an overhead of $\mathcal{O}\left(r^2 d\right)$ or more efficiently using variance-reduced SGD (Shamir, 2015). Using vanilla SVD for per-iteration projection brings the overall time complexity of proposed algorithm to $\mathcal{O}\left(r^2 d/\epsilon^2\right)$. We note that the convex formulation in Olfat and Aswani (2018) also requires solving an SDP programming (e.g., ellipsoid method to interior point method), which suffers from high computational cost as well.

This is for the first time that we are solving the exact multi-objective problem, rather than its min-max relaxation using SDP in a fairness problem. Morgenstern et al. (2019) is suggesting that for $k = 2$ their approximation is exact, meaning their algorithm will find the fair representation in the exact r dimension they aim to reach. However, in practice, even for $k = 2$, we can show that our algorithm can achieve a smaller disparity error, as shown in Figure 2, which indicates that pairwise disparity error can achieve a

better subspace in terms of fairness. For $k > 2$, they are still solving an inefficient SDP problem to exact same problem we are proposing. Hence, the novelty of our approach lies in solving this problem using gradient descent and ensuring to reach a Pareto stationary point, which even does not require extra dimensions to satisfy fairness. This setting and its proposed gradient descent algorithm to solve it can be applied to other unsupervised and supervised fairness problems. Thus, it could open up new perspectives on all other fairness problems in learning tasks, by advocating optimal trade-offs between main learning objectives and fairness criteria using Pareto efficiency

6 Experiment

In this section we empirically examine the introduced algorithm for fair PCA with Adult dataset[1] and Credit dataset[2]. Adult dataset consists of census data to predict whether the income of a person exceeds 50K per year or not. The Credit dataset contains clients' credit history information to predict whether they would default in future or not. We will omit the label data and work with the rest of it, which contains 14 features for Adult dataset, including gender and race, which we consider them as sensitive features in this dataset. In the Credit dataset, we will have 23 features including sensitive features of sex and marriage. The gender feature in the Adult dataset and sex in the Credit dataset are binary features with two values, namely, Male and Female. Race from Adult dataset, on the other side, is a multiple group feature, with 5 different groups, including White, Asian-Pac-Islander, Amer-Indian-Eskimo, Black, and Other. Marriage in the Credit dataset is also a multiple group feature with 3 groups of Single, Married, and Other. For the Adult dataset, we use the training dataset which has $32,561$ number of samples, among which $10,548$ belongs to the Female group and $22,013$ to the Male group. The distribution of samples among race groups are as follows: Black $30,47$, White $27,994$, Asian-Pac-Islander 312, Amer-Indian-Eskimo 962, and Other 246. In the Credit dataset we have $30,000$ training samples, out of which there are $18,112$ Female and $11,888$ Male samples. The distribution of the Marriage feature is $13,659$ married, $15,964$ single, and 323 other samples. We first apply the fair PCA method to binary sensitive feature, in which we set the learning rate to $1/\sqrt{t}$, where t is the iteration number. This condition on learning rate satisfies the maximum decrease condition by backtracking line search in (16).

Binary sensitive feature In the Adult dataset, we observed that the Female group is benefiting from normal PCA on the whole dataset, while the Male group is sacrificing its reconstruction error. Hence, by applying Fair PCA algorithm, we can perfectly decrease these trade-offs, while suffering insignificant loss to the total reconstruction error, compared to normal PCA. The results are depicted in Figure 1, where the trade-offs and how Fair PCA is addressing them is noticeable. To compare the introduced Pareto fair PCA with algorithms using SDP in Samadi et al. (2018) and Morgenstern et al. (2019), we will use the average disparity errors across sensitive groups in both Adult and Credit datasets. Figure 2 shows the average disparity errors of Pareto fair PCA with single and pairwise disparity error objectives, SDP fair PCA, and normal PCA on binary features (gender and sex) of Adult and Credit datasets. First, it reveals that there is a huge gap between normal PCA and fair PCA algorithms in terms of disparity errors, which is indicating that these algorithms are decreasing this disparity error. Second, it shows the superiority of Pareto fair PCA over SDP relaxation methods in both datasets (especially with pairwise objectives), where Pareto fair PCA has a smaller average disparity error close to zero.

Multiple group sensitive feature The proposed Algorithm 1, can efficiently generalize to the multiple group sensitive features, by adding pairwise disparity errors of each pair of groups to the objective vector and minimize the overall vector to reach a Pareto optimal or stationary point. However, adding more objectives, introduces more trade-offs, makes the optimization over all objectives more difficult. First, in the Adult dataset with race as the sensitive feature, the reconstruction error of the Pareto fair PCA and normal PCA is shown in Figure 3, where the trade-offs between benefits and sacrifices of different groups are clearly noticeable.

[1] https://archive.ics.uci.edu/ml/datasets/Adult

[2] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

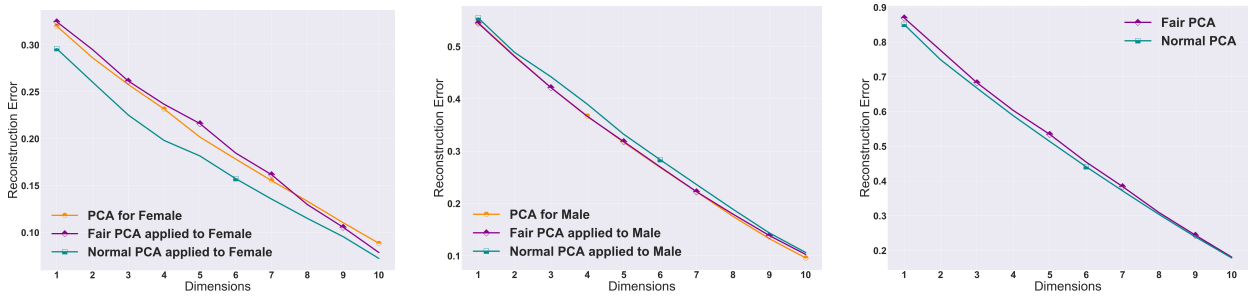
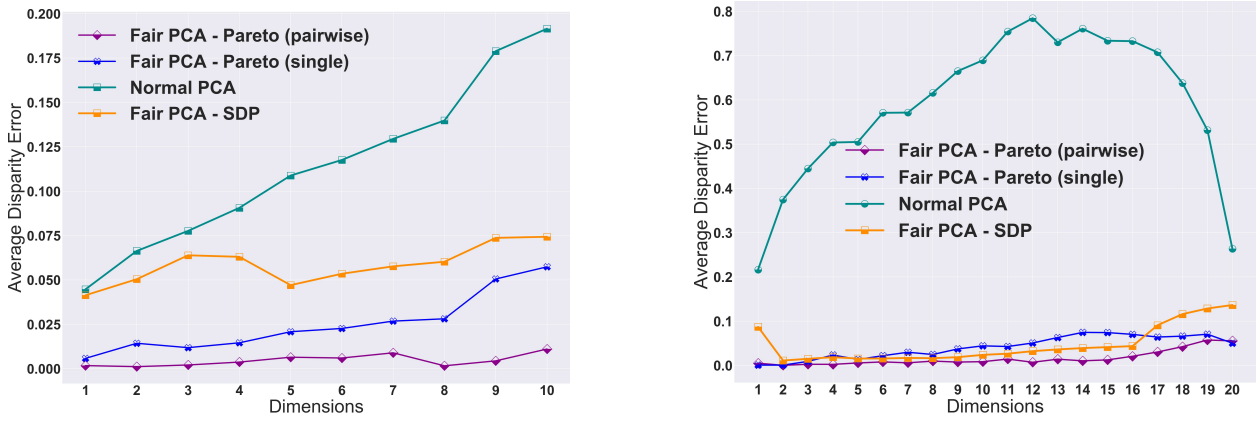


Figure 1: Applying normal PCA and fair PCA to the Adult dataset with gender as its sensitive feature. The first two figures show the reconstruction error of normal PCA (trained on all data) applied to each group, fair PCA (trained on all data) applied to each group, and normal PCA trained on the data of each group individually. The last figure reveals the difference between normal and fair PCA reconstruction loss on all data, which is very tiny and negligible.



(a) Adult dataset

(b) Credit dataset

Figure 2: Comparing Pareto fair PCA algorithm introduced in this paper (pairwise and single disparity error), with fair PCA algorithms using SDP relaxation introduced in Samadi et al. (2018) and Morgenstern et al. (2019). The experiment is on binary features of Adult and Credit datasets (gender and sex). The average disparity error of algorithms on Adult dataset clearly shows the superiority of the Pareto fair PCA with pairwise disparity error objectives and then the single disparity error objectives. In the Credit dataset Pareto fair PCA with pairwise objectives has a slightly better performance with respect to two other methods.

The fair PCA algorithm can superbly decrease these trade-offs for all but one group, with a negligible increase in overall reconstruction loss depicted in Figure 7a. Following the same step as in binary case, we show the disparity error of different groups in Figure 4, which reveals that fair PCA clearly outperforms normal PCA in most of the groups and in terms of average disparity error over different groups in Figure 7b.

As for the Credit dataset, we also test it on its multiple group sensitive feature, marriage, which has 3 different groups. The result of reconstruction error on Pareto fair PCA, normal PCA and normal PCA on each group's data individually is depicted in Figure 5, where it is clear that fair PCA is very close to each groups' PCA (except for other group, because the number of samples in that group is too low), while its reconstruction error is very close to that of normal PCA. Also, the disparity error and average disparity error of fair PCA versus normal PCA is shown in Figure 6, where the superiority of fair PCA is noticeable.

Figure 7 depicts the reconstruction loss and average disparity error of fair and normal PCA applied to Adult dataset with race as its sensitive feature. In this dataset, race has 5 categories, makes it a multiple

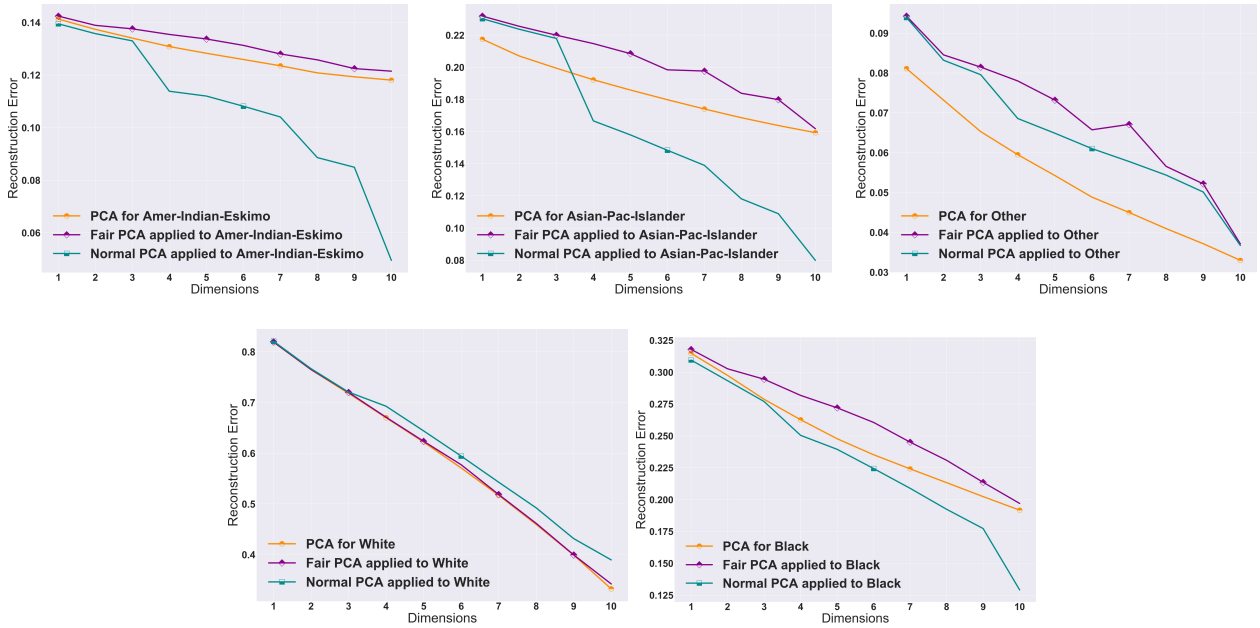


Figure 3: Applying normal and fair PCA on the Adult dataset with "race" as its sensitive feature. Each plot shows the reconstruction error of the normal PCA (trained on the whole data) applied to each group's data, fair PCA (trained on the whole data) applied to each group's data, and normal PCA trained on each group's data individually.

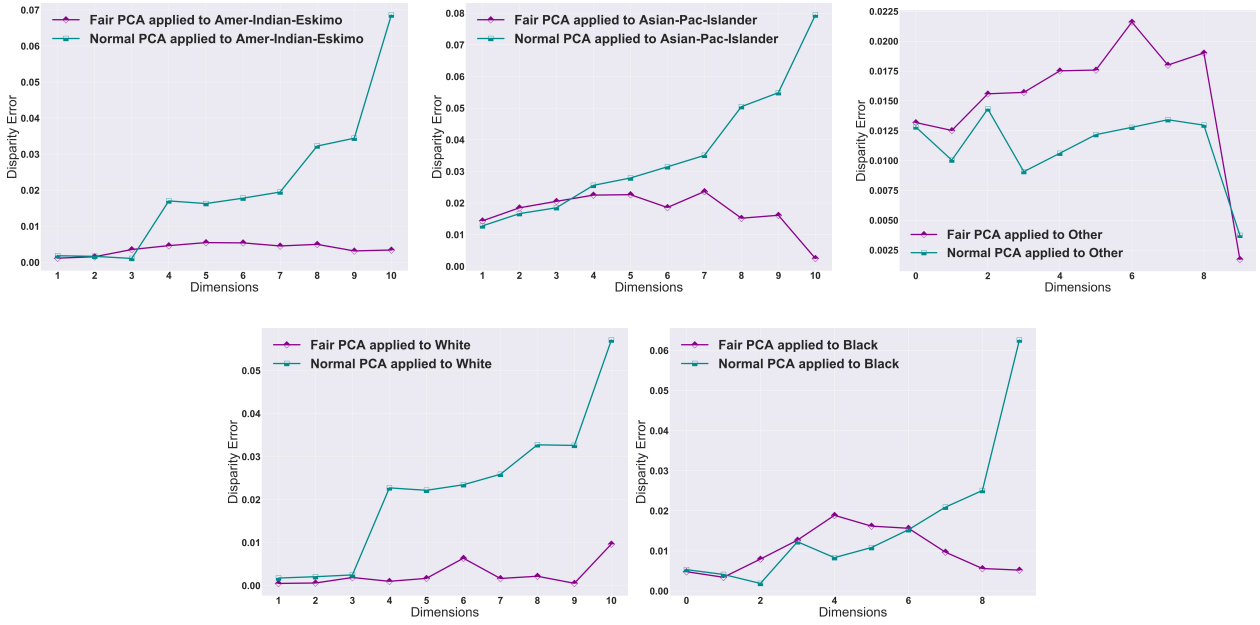


Figure 4: Disparity error of normal and fair PCA trained on the Adult dataset with "race" as its sensitive feature. Each plot depicts the disparity errors of different groups with normal and fair PCA.

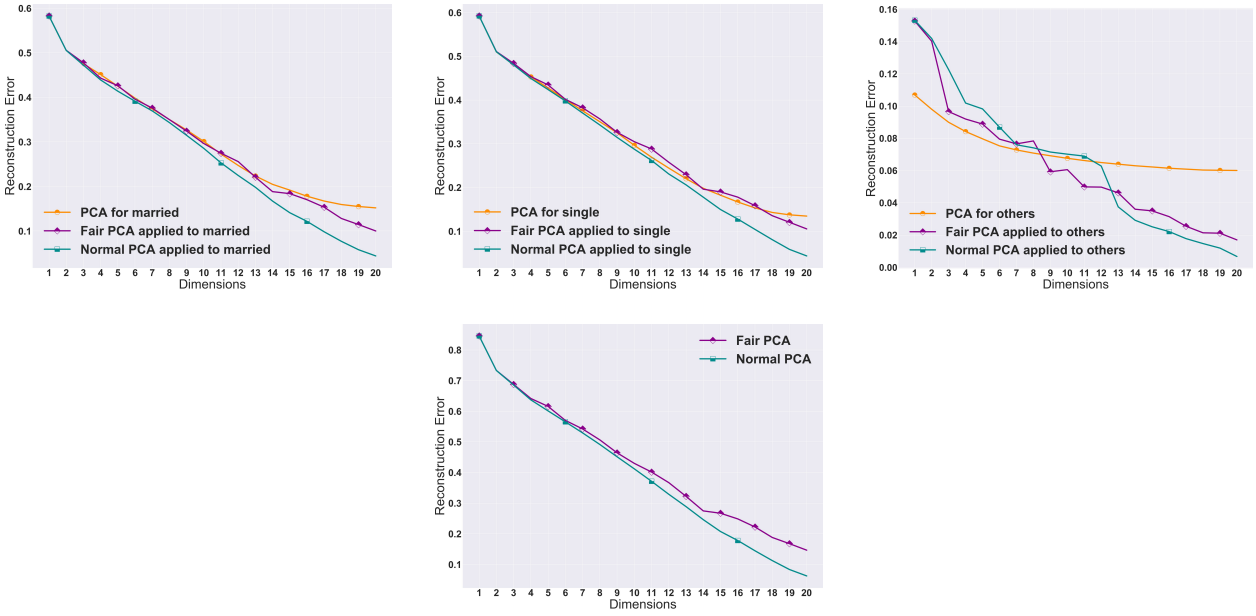


Figure 5: Applying normal and fair PCA on the Credit dataset with "Marriage" as its sensitive feature. Each plot shows the reconstruction error of the normal PCA (trained on the whole data) applied to each group's data, fair PCA (trained on the whole data) applied to each group's data, and normal PCA trained on each group's data individually. The last figure shows the reconstruction error of normal PCA and fair PCA on this dataset with multiple group sensitive feature of marriage.

group sensitive feature. The results indicates that even in a dataset with multiple group sensitive feature, the increase in the reconstruction loss of fair PCA compared to normal PCA is slim, while the gap between their disparity errors is huge. This means that, normal PCA, is unfairly treating different groups in its learned representation subspace.

7 Conclusion

In this paper, we cast the fairness problem in dimension reduction algorithms such as PCA as a multi-objective programming. Unlike supervised learning, there is not a clear definition of fairness in unsupervised learning tasks. Thus, we use the notion of balancing between sacrifices and benefits each sensitive group makes or enjoys to define a fairness metric for this problem. These sacrifices or benefits are the consequence of finding the optimal subspace over the whole data rather than using only each protected group's data. Hence, the notion of fairness is to have an equal contribution from each group to the overall reconstruction loss with respect to the reconstruction loss they have on the subspace learned by their own data. This introduces a trade-off between these contributions and overall reconstruction loss. We propose an efficient multi-objective optimization procedure that can guarantee the convergence to a Pareto stationary point, which has an efficient trade-off between these objectives. This paper also introduces some interesting problems worthy of future investigations. First, generalization of the proposed disparity error and pairwise disparity error as fairness metrics in other dimension reduction algorithms and, also, other unsupervised learning tasks. Moreover, it is interesting to investigate the stochastic version of the proposed algorithm and its convergence analysis since finding a descent direction where gradients are noisy might be a challenging task. Finally, as noted before, the existing methods including the one proposed in the present work require learning a local optimal projection subspace for each group before learning the global fair subspace. One interesting direction is to extend these works to efficiently learn all subspaces together while preserving the fairness of global subspace.

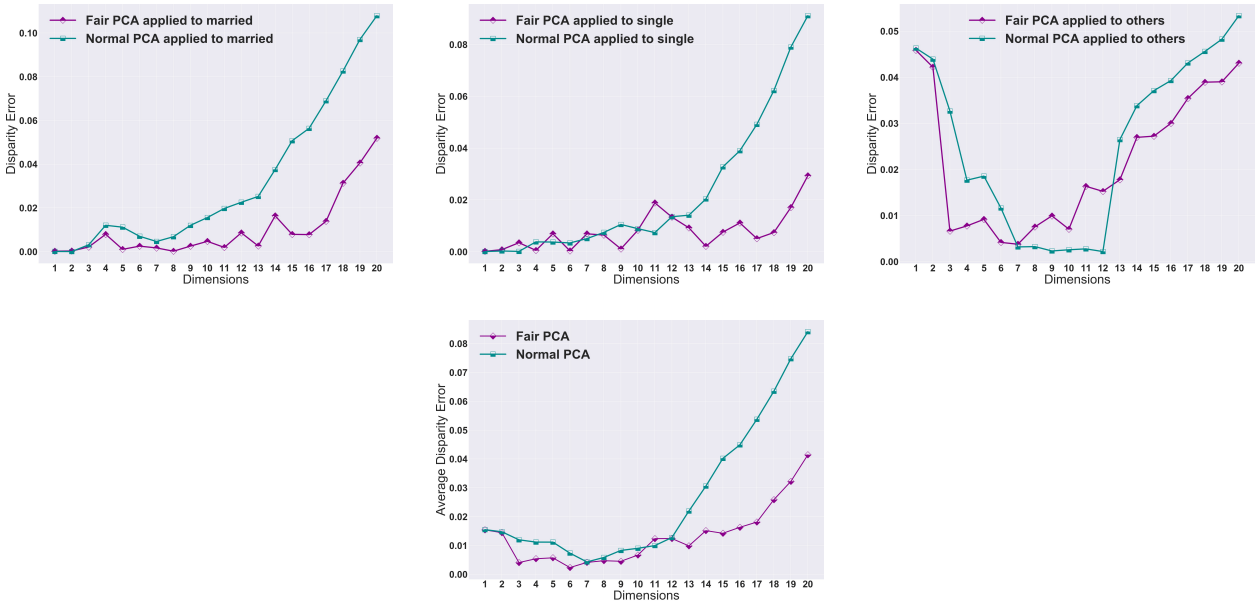
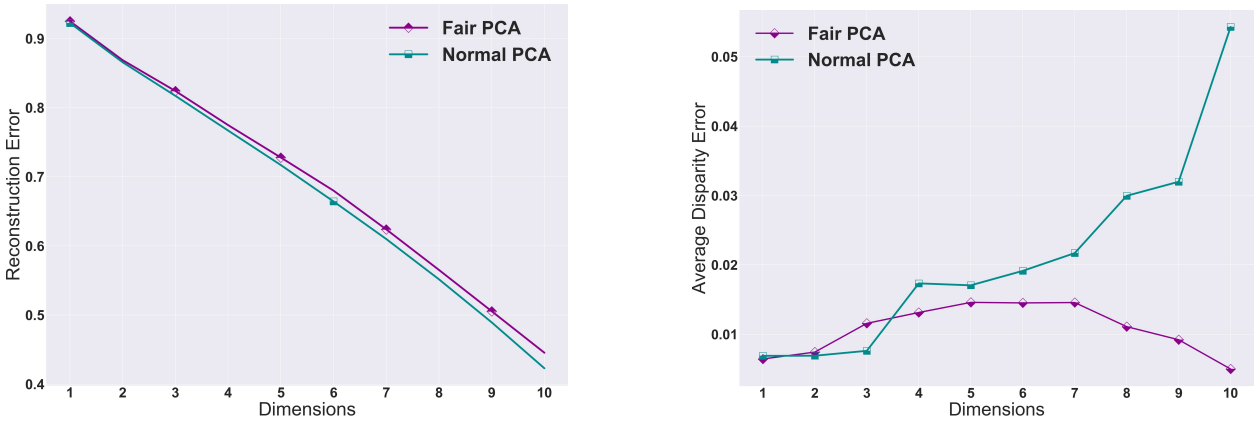


Figure 6: Disparity error of normal and fair PCA trained on the Credit dataset with "Marriage" as its sensitive feature. Each plot depicts the disparity errors of different groups with normal and fair PCA. The last figure shows the average of disparity errors across groups.



(a) Total Reconstruction Loss

(b) Average Disparity Error

Figure 7: Applying normal and fair PCA on the Adult dataset with "race" as its multiple group sensitive feature. Left shows the difference between reconstruction loss of fair and normal PCA, which is so small. On the other hand, the right plot shows their difference in terms of average disparity error, which is huge and demonstrating the efficacy of fair PCA in addressing fairness even in multiple group sensitive feature cases.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

Matthias Ehrgott. A discussion of scalarization techniques for multiple objective integer programming. *Annals of Operations Research*, 147(1):343–360, 2006.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

Jörg Fliege and Benar Fux Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical Methods of Operations Research*, 51(3):479–494, 2000.

Soheil Ghili, Ehsan Kazemi, and Amin Karbasi. Eliminating latent discrimination: Train then mask. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3672–3680, 2019.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

Mohammad Mahdi Kamani, Farshid Farhat, Stephen Wistar, and James Z Wang. Shape matching using skeleton context for automated bow echo detection. In *IEEE International Conference on Big Data*, pages 901–908, 2016.

Mohammad Mahdi Kamani, Farshid Farhat, Stephen Wistar, and James Z Wang. Skeleton matching with applications in severe weather detection. *Applied Soft Computing*, 70:1154–1166, 2018.

Mohammad Mahdi Kamani, Sadegh Farhang, Mehrdad Mahdavi, and James Z Wang. Targeted meta-learning for critical incident detection in weather data. *International Conference on Machine Learning, Workshop on "Climate Change: How Can AI Help?"*, 2019.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

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

Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 240–248. ACM, 2019.

Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019.

Zachary C Lipton, Alexandra Chouldechova, and Julian McAuley. Does mitigating ml's disparate impact require disparate treatment? *stat*, 1050:19, 2017.

Mehrdad Mahdavi, Tianbao Yang, and Rong Jin. Stochastic convex optimization with multiple objectives. In *Advances in Neural Information Processing Systems*, pages 1115–1123, 2013.

Kaisa Miettinen. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 2012.

Jamie Morgenstern, Samira Samadi, Mohit Singh, Uthaipon Tantipongpipat, and Santosh Vempala. Fair dimensionality reduction and iterative rounding for sdps. *arXiv preprint arXiv:1902.11281*, 2019.

Cecilia Munoz, Executive Office of the President, , Domestic Policy Council Director, Megan (US Chief Technology Officer Smith (Office of Science, Technology Policy)), DJ (Deputy Chief Technology Officer for Data Policy, Chief Data Scientist Patil (Office of Science, and Technology Policy)). *Big data: A report on algorithmic systems, opportunity, and civil rights*. Executive Office of the President, 2016.

Hamed Nikbakht and Konstantinos G Papakonstantinou. A direct hamiltonian mcmc approach for reliability estimation. *arXiv preprint arXiv:1909.03575*, 2019.

Matt Olfat and Anil Aswani. Convex formulations for fair principal component analysis. *arXiv preprint arXiv:1802.03765*, 2018.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair pca: One extra dimension. In *Advances in Neural Information Processing Systems*, pages 10976–10987, 2018.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Ohad Shamir. A stochastic pca and svd algorithm with an exponential convergence rate. In *International Conference on Machine Learning*, pages 144–152, 2015.

Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 2921–2930, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

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

Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017b.

A Proof of Theorem 1

Proof. Consider the following constrained optimization problem:

$$\begin{aligned} \sup \quad & \sum_{i\in[m]} \epsilon_i \\ \text{subject to} \quad & f_i\left(\mathbf{U}\right) + \epsilon_i = f_i\left(\widetilde{\mathbf{U}}\right),\ i \in [m], \\ & \epsilon_i \geq 0,\ i \in [m], \end{aligned} \tag{22}$$

where $\widetilde{\mathbf{U}}$ is any feasible subspace. By assuming that $f_i(.)$ is convex for $i \in [m]$, if there is no finite maximum value for this optimization, then the set of proper Pareto optimal solutions is empty. The main immediate implication of this theorem is that, if the objectives are bounded, then a Pareto optimal solution exist for this optimization problem. More specifically, if the solution of this optimization is the objective value of zero, then the $\widetilde{\mathbf{U}}$ is a Pareto optimal solution. To prove this theorem, we consider $\mathbf{U}^*$ to be a proper Pareto optimal solution to the problem (22), then there exists a vector $\boldsymbol{\lambda} \in \mathbb{R}^m_+$, such that the point $\mathbf{U}^*$ is a Pareto optimal solution to the problem:

$$\arg\min_{\mathbf{U}} \sum_{i\in[m]} \lambda_i f_i\left(\mathbf{U}\right) \tag{23}$$

Then, from the Pareto optimality we have for every feasible $\mathbf{U}$:

$$\sum_{i\in[m]} \lambda_i \left[f_i\left(\mathbf{U}\right) - f_i\left(\mathbf{U}^*\right)\right] \geq 0 \tag{24}$$

By setting $\mathbf{U} = \widetilde{\mathbf{U}}$, we can write:

$$\sum_{i\in[m]} \lambda_i \left[f_i\left(\widetilde{\mathbf{U}}\right) - f_i\left(\mathbf{U}^*\right)\right] = M^\dagger \geq 0 \tag{25}$$

Also, from the optimization problem (22) since there is not a finite maximum objective value available, for every $\widehat{M} \geq 0$ we can find a $\widehat{\mathbf{U}}$ such that:

$$\sum_{i\in[m]} \left[f_i\left(\widetilde{\mathbf{U}}\right) - f_i\left(\widehat{\mathbf{U}}\right)\right] \geq \widehat{M} \tag{26}$$

Then, if we set $\lambda_{\min} = \min\left\{\lambda_1, \ldots, \lambda_m\right\}$, we have:

$$\begin{aligned} \lambda_{\min}\widehat{M} &\leq \lambda_{\min} \sum_{i\in[m]} \left[f_i\left(\widetilde{\mathbf{U}}\right) - f_i\left(\widehat{\mathbf{U}}\right)\right] \\ &= \sum_{i\in[m]} \lambda_{\min} \left[f_i\left(\widetilde{\mathbf{U}}\right) - f_i\left(\widehat{\mathbf{U}}\right)\right] \\ &\leq \sum_{i\in[m]} \lambda_i \left[f_i\left(\widetilde{\mathbf{U}}\right) - f_i\left(\widehat{\mathbf{U}}\right)\right] \end{aligned} \tag{27}$$

If the $\widehat{\mathbf{U}}$ is chosen to satisfy $\lambda_{\min}\widehat{M} = M^\dagger$, then we have:

$$\begin{aligned} \sum_{i\in[m]} \lambda_i \left[f_i\left(\widetilde{\mathbf{U}}\right) - f_i\left(\mathbf{U}^*\right)\right] &\leq \sum_{i\in[m]} \lambda_i \left[f_i\left(\widetilde{\mathbf{U}}\right) - f_i\left(\widehat{\mathbf{U}}\right)\right] \\ \sum_{i\in[m]} \lambda_i f_i\left(\widehat{\mathbf{U}}\right) &\leq \sum_{i\in[m]} \lambda_i f_i\left(\mathbf{U}^*\right), \end{aligned} \tag{28}$$

which contradicts the assumption of Pareto optimality of $\mathbf{U}^*$, and hence, Pareto optimal set is empty. □

B Proof of Lemma 1

Proof. The proof is straightforward and directly follows from KKT optimally conditions for problem (29), however, we show the derivation here for completeness.

First, we note that the minmax optimization problem introduced in (18) to find the descent direction $\mathbf{D}_t$, can be rewritten as the following equivalent constrained optimization problem:

$$\begin{aligned}(\mathbf{D}_t, \epsilon_t) = \arg \min_{\mathbf{D}\in\mathbb{R}^{d\times r}, \epsilon\in\mathbb{R}_+} \quad & \epsilon + \frac{1}{2}\|\mathbf{D}\|_{\mathrm{F}}^2, \\ \text{s.t. } & \mathsf{tr}\left(\mathbf{D}^\top \mathbf{G}_i^{(t)}\right) \leq \epsilon, \quad \forall\, 1 \leq i \leq m.\end{aligned} \tag{29}$$

Forming the Lagrangian of the constrained problem as follows

$$\mathcal{L}\left(\mathbf{D}, \epsilon; \lambda_i\right) = \frac{1}{2}\|\mathbf{D}\|_{\mathrm{F}}^2 + \epsilon + \sum_{i=1}^{m} \lambda_i \left(\mathsf{tr}\left(\mathbf{D}^\top \mathbf{G}_i^{(t)}\right) - \epsilon\right), \tag{30}$$

and writing the KKT conditions gives:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{D}} = \mathbf{D} + \sum_{i=1}^{m} \lambda_i^{(t)} \mathbf{G}_i^{(t)} = 0 \tag{31}$$

$$\frac{\partial \mathcal{L}}{\partial \epsilon} = 1 - \sum_{i=1}^{m} \lambda_i = 0 \tag{32}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = \mathsf{tr}\left(\mathbf{D}^\top \mathbf{G}_i^{(t)}\right) - \epsilon = 0 \qquad \forall\, 1 \leq i \leq m \tag{33}$$

$$\lambda_i \left(\mathsf{tr}\left(\mathbf{D}^\top \mathbf{G}_i^{(t)}\right) - \epsilon\right) = 0 \qquad \forall\, 1 \leq i \leq m \tag{34}$$

$$\lambda_i \geq 0 \qquad \forall\, 1 \leq i \leq m \tag{35}$$

From (31) we have the following that holds for the descent direction:

$$\mathbf{D}_t = -\sum_{i=1}^{m} \lambda_i^{(t)} \mathbf{G}_i^{(t)}, \tag{36}$$

where $\boldsymbol{\lambda}^{(t)} = \left[\lambda_1^{(t)}, \ldots, \lambda_m^{(t)}\right]^\top$ belongs to Δ_m– the m-dimensional simplex. By plugging these conditions back to the main problem, the dual problem can be simplified as:

$$\hat{\boldsymbol{\lambda}}^{(t)} = \arg \min_{\boldsymbol{\lambda}\in\Delta_m} \frac{1}{2} \left\| \sum_{i=1}^{m} \lambda_i \mathbf{G}_i^{(t)} \right\|_{\mathrm{F}}^2. \tag{37}$$

By solving the dual problem which is a quadratic programming and using (36) we can find the descent direction from optimal dual variables.

Next, we need to show that the obtained direction is either $\mathbf{0}$ or a descent direction to all objectives. If the point $\mathbf{U}_t$ is a Pareto stationary point, then it means that we cannot find a direction that can decrease all the objectives, without increasing one. Hence, there is no such a $\mathbf{D}$ that $\mathsf{tr}\left(\mathbf{D}^\top \mathbf{G}_i^{(t)}\right) \leq 0$ for all $1 \leq i \leq m$, unless $\mathbf{D} = \mathbf{0}$. For points that are not Pareto stationary, consider the following quadratic optimization for every $1 \leq j \leq m$:

$$\arg \min_{\beta\in[0,1]} \frac{1}{2} \left\| (1-\beta)\mathbf{G}_j^{(t)} - \beta \mathbf{D}_t \right\|_{\mathrm{F}}^2. \tag{38}$$

We can see that this optimization problem is equivalent to the optimization problem in (37), with $\lambda_i = \beta\hat{\lambda}_i$ for $1 \le i \le m,\ i \neq j$, and $\lambda_j = 1 - \beta(1 - \hat{\lambda}_j)$. This means that the optimum of the quadratic optimization in (38) happens at $\beta = 1$. Then by using the first order optimally condition at optimum point we get:

$$\begin{aligned} 2\left\|\mathbf{G}_j^{(t)} + \mathbf{D}_t\right\|_{\mathrm{F}}^2 - 2\,\mathsf{tr}\left(\left(\mathbf{G}_j^{(t)}\right)^\top\left(\mathbf{G}_j^{(t)} + \mathbf{D}_t\right)\right) &\le 0 \\ 2\left\|\mathbf{G}_j^{(t)} + \mathbf{D}_t\right\|_{\mathrm{F}}^2 - 2\,\mathsf{tr}\left(\left(\mathbf{G}_j^{(t)} + \mathbf{D}_t - \mathbf{D}_t\right)^\top\left(\mathbf{G}_j^{(t)} + \mathbf{D}_t\right)\right) &\le 0 \\ \mathsf{tr}\left(\mathbf{D}_t^\top\left(\mathbf{G}_j^{(t)} + \mathbf{D}_t\right)\right) &\le 0 \\ \mathsf{tr}\left(\mathbf{D}_t^\top\mathbf{G}_j^{(t)}\right) &\le -\left\|\mathbf{D}_t\right\|_{\mathrm{F}}^2 \end{aligned} \tag{39}$$

which clearly shows that $\mathbf{D}_t$ is a descent direction for all objectives. □

C Proof of Theorem 2

To prove the Theorems 2 and 3, we first need to show that by properly choosing the regularization parameter α our objectives are smooth. Recall that, our goal is to solve the following multi-objective optimization problem with non-convex components:

$$\mathrm{f}(\mathbf{U}) = [f_1(\mathbf{U}), \ldots, f_m(\mathbf{U})] \tag{40}$$

where $m = 1 + \binom{k}{2}$ with k being the number of groups in the sensitive feature. Also, recall that in the case of fair PCA, we have $f_1(\mathbf{U}) = -\frac{1}{2}\,\mathsf{tr}\left(\mathbf{U}^\top\mathbf{X}^\top\mathbf{X}\mathbf{U}\right)$ is the overall reconstruction loss, and $f_i(\mathbf{U}), i = 2, 3, \ldots, m$ are disparity errors for pair of groups. In what follows we use $\|\cdot\|$ and $\|\cdot\|_{\mathrm{F}}$ to denote the spectral and Frobenius norms of a matrix, respectively.

To prove the theorem, we first show that all the the individual objective functions are smooth with bounded gradient (i.e., $\mathrm{f}(\cdot)$ is component-wise smooth), conditioned that the regularization parameter α satisfies $\alpha \ge \max\limits_{i,j\in[k]} \gamma_d\left(\mathbf{X}_i^\top\mathbf{X}_i\right) - \gamma_1\left(\mathbf{X}_j^\top\mathbf{X}_j\right)$ (recall that $\gamma_d(\cdot)$ is the smallest eigenvalue of input PSD matrix). To this end, we follow the definition of the smooth functions, i.e., $\|\nabla f(\mathbf{U}) - \nabla f(\mathbf{V})\|_{\mathrm{F}} \le L\|\mathbf{U} - \mathbf{V}\|_{\mathrm{F}}^2$.

In particular, for $f_1(\mathbf{U})$ we have:

$$\begin{aligned} \|\nabla f_1(\mathbf{U}) - \nabla f_1(\mathbf{V})\|_{\mathrm{F}} &= \|\mathbf{X}^\top\mathbf{X}\mathbf{U} - \mathbf{X}^\top\mathbf{X}\mathbf{V}\|_{\mathrm{F}} \\ &\overset{①}{\le} \|\mathbf{X}^\top\mathbf{X}\|\|\mathbf{U} - \mathbf{V}\|_{\mathrm{F}} \\ &\overset{②}{\le} \gamma_{\max}(\mathbf{X}^\top\mathbf{X})\|\mathbf{U} - \mathbf{V}\|_{\mathrm{F}}, \end{aligned}$$

where the first inequality ① follows from the fact that for any two matrices $\mathbf{A}$ and $\mathbf{B}$ it holds that $\|\mathbf{A}\mathbf{B}\|_{\mathrm{F}} \le \|\mathbf{A}\|\|\mathbf{B}\|_{\mathrm{F}}$, and ② follows from the definition of spectral norm. The above inequality indicates that the objective corresponding to the overall reconstruction error is smooth with parameter $\gamma_{\max}(\mathbf{X}^\top\mathbf{X})$.

To show the smoothness of disparity errors, for simplicity we only focus on one of the objectives between a single pair of sensitive features, say s_i, s_j, as the argument easily generalizes to other objectives/pairs due to symmetry. We also drop the subscript from function and use $f(\mathbf{U})$ to denote the regularized disparity error between groups s_i and s_j defined as

$$\begin{aligned} f(\mathbf{U}) &= \mathcal{E}_i(\mathbf{U}) - \mathcal{E}_j(\mathbf{U}) + \frac{\alpha}{2}\|\mathbf{U}\|_{\mathrm{F}}^2 \\ &= \mathcal{L}_i(\mathbf{U}) - \mathcal{L}_i(\mathbf{U}_i^*) - \mathcal{L}_j(\mathbf{U}) + \mathcal{L}_j(\mathbf{U}_j^*) + \frac{\alpha}{2}\|\mathbf{U}\|_{\mathrm{F}}^2 \\ &= -\frac{1}{2}\,\mathsf{tr}\left(\mathbf{U}^\top\mathbf{X}_i^\top\mathbf{X}_i\mathbf{U}\right) + \frac{1}{2}\,\mathsf{tr}\left(\mathbf{U}^\top\mathbf{X}_j^\top\mathbf{X}_j\mathbf{U}\right) + \frac{\alpha}{2}\|\mathbf{U}\|_{\mathrm{F}}^2 \\ &\quad + \frac{1}{2}\,\mathsf{tr}\left(\mathbf{U}_i^{*\top}\mathbf{X}_i^\top\mathbf{X}_i\mathbf{U}_i^*\right) - \frac{1}{2}\,\mathsf{tr}\left(\mathbf{U}_j^{*\top}\mathbf{X}_j^\top\mathbf{X}_j\mathbf{U}_j^*\right) \end{aligned} \tag{41}$$

Following the definition of smoothness, we have

$$\begin{aligned}
&\|\nabla f(\mathbf{U}) - \nabla f(\mathbf{V})\|_{\mathrm{F}} \\
&= \|\mathbf{X}_i^\top \mathbf{X}_i \mathbf{U} - \mathbf{X}_j^\top \mathbf{X}_j \mathbf{U} + \alpha \mathbf{U} - \mathbf{X}_i^\top \mathbf{X}_i \mathbf{V} + \mathbf{X}_j^\top \mathbf{X}_j \mathbf{V} - \alpha \mathbf{V}\|_{\mathrm{F}} \\
&= \left\| \left(\mathbf{X}_i^\top \mathbf{X}_i - \mathbf{X}_j^\top \mathbf{X}_j + \alpha \mathbf{I} \right) (\mathbf{U} - \mathbf{V}) \right\|_{\mathrm{F}} \\
&\leq \|\mathbf{X}_i^\top \mathbf{X}_i - \mathbf{X}_j^\top \mathbf{X}_j + \alpha \mathbf{I}\|_2 \|\mathbf{U} - \mathbf{V}\|_{\mathrm{F}}
\end{aligned}$$

Again, we can further upper bound the right hand side by using the definition of the spectral norm of a matrix:

$$\begin{aligned}
\|\mathbf{X}_i^\top \mathbf{X}_i - \mathbf{X}_j^\top \mathbf{X}_j + \alpha \mathbf{I}\|_2 &= \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} \mathbf{v}^\top \left(\mathbf{X}_i^\top \mathbf{X}_i - \mathbf{X}_j^\top \mathbf{X}_j + \alpha \mathbf{I} \right) \mathbf{v} \\
&\leq \gamma_{\max} \left(\mathbf{X}_i^\top \mathbf{X}_i \right) - \gamma_{\min} \left(\mathbf{X}_j^\top \mathbf{X}_j \right) + \alpha
\end{aligned} \tag{42}$$

where $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$ is the sphere in d dimensions.

As a result, as long as the regularization parameter α satisfies the following condition

$$\alpha > \max \left\{ 0, \gamma_{\max} \left(\mathbf{X}_j^\top \mathbf{X}_j \right) - \gamma_{\min} \left(\mathbf{X}_i^\top \mathbf{X}_i \right) \right\},$$

the disparity error objective between groups s_i and s_j is smooth with smoothness parameter $\gamma_{\max} \left(\mathbf{X}_i^\top \mathbf{X}_i \right) - \gamma_{\min} \left(\mathbf{X}_j^\top \mathbf{X}_j \right) + \alpha > 0$. By symmetry, we can derive the smoothness condition for other pairs of groups as well, which results in the following condition on the regularization parameter:

$$\alpha \geq \max_{i,j \in [k]} \gamma_{\min} \left(\mathbf{X}_i^\top \mathbf{X}_i \right) - \gamma_{\max} \left(\mathbf{X}_j^\top \mathbf{X}_j \right)$$

to satisfy the smoothness of all objectives $f_i(\cdot), i = 2, \ldots, m$. We note that one can use different regularization parameters for each pair depending on the eigen-gap between their covariance matrices as well.

We now turn to prove the convergence rate of the proposed algorithm to a Pareto fair subspace in general case as stated in (40), where we assume that the individual loss functions $f_i(\mathbf{U}), i = 1, 2, \ldots, m$ satisfy Lipschitz continuous gradient condition (smoothness) with smoothness parameters $L_i, i = 1, 2, \ldots, m$. We also use L to denote the maximum smoothness parameter, i.e., $L = \max_{i=1,2,\ldots,m} L_i$.

Proof. The proof begins by first bounding the difference in function values of each objective $f_i(\mathbf{U}_t) - f_i(\mathbf{U}^*)$, individually, following the convexity assumption:

$$f_i(\mathbf{U}_t) - f_i(\mathbf{U}^*) \leq \mathsf{tr}\left(\left(\mathbf{G}_i^{(t)} \right)^\top (\mathbf{U}_t - \mathbf{U}^*) \right)$$

Then we can multiply both sides by $\hat{\lambda}_i$ and sum over i:

$$\begin{aligned}
\sum_{i=1}^m \hat{\lambda}_i^{(t)} \left(f_i(\mathbf{U}_t) - f_i(\mathbf{U}^*) \right) &\leq \sum_{i=1}^m \hat{\lambda}_i^{(t)} \mathsf{tr}\left(\left(\mathbf{G}_i^{(t)} \right)^\top (\mathbf{U}_t - \mathbf{U}^*) \right) \\
&= \mathsf{tr}\left(\left(\sum_{i=1}^m \hat{\lambda}_i^{(t)} \mathbf{G}_i^{(t)} \right)^\top (\mathbf{U}_t - \mathbf{U}^*) \right) \\
&= -\mathsf{tr}\left(\mathbf{D}_t^\top (\mathbf{U}_t - \mathbf{U}^*) \right) \\
&= \frac{1}{\eta} \mathsf{tr}\left((\mathbf{U}_t - \mathbf{U}_{t+1})^\top (\mathbf{U}_t - \mathbf{U}^*) \right) \\
&= \frac{1}{2\eta} \left(\|\mathbf{U}_t - \mathbf{U}^*\|_{\mathrm{F}}^2 + \|\mathbf{U}_t - \mathbf{U}_{t+1}\|_{\mathrm{F}}^2 - \|\mathbf{U}_{t+1} - \mathbf{U}^*\|_{\mathrm{F}}^2 \right) \\
&= \frac{1}{2\eta} \left(\|\mathbf{U}_t - \mathbf{U}^*\|_{\mathrm{F}}^2 - \|\mathbf{U}_{t+1} - \mathbf{U}^*\|_{\mathrm{F}}^2 \right) + \frac{\eta}{2} \|\mathbf{D}_t\|_{\mathrm{F}}^2 \\
&\overset{①}{\leq} \frac{1}{2\eta} \left(\|\mathbf{U}_t - \mathbf{U}^*\|_{\mathrm{F}}^2 - \|\mathbf{U}_{t+1} - \mathbf{U}^*\|_{\mathrm{F}}^2 \right) + \frac{\eta L^2}{2}
\end{aligned}$$

where ① follows from the smoothness assumption and definition of L. By summing up above inequality for all iterations $t = 1, 2, \ldots, T$ gives:

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{m}\hat{\lambda}_i^{(t)}\left(f_i(\mathbf{U}_t) - f_i(\mathbf{U}^*)\right) &\leq \frac{1}{2\eta}\sum_{t=1}^{T}\left(\|\mathbf{U}_t - \mathbf{U}^*\|_{\mathrm{F}}^2 - \|\mathbf{U}_{t+1} - \mathbf{U}^*\|_{\mathrm{F}}^2 + \frac{\eta L^2}{2}\right)\\ &= \frac{1}{2\eta}\left(\|\mathbf{U}_1 - \mathbf{U}^*\|_{\mathrm{F}}^2 - \|\mathbf{U}_T - \mathbf{U}^*\|_{\mathrm{F}}^2\right) + \frac{\eta L^2 T}{2}\\ &\leq \frac{1}{2\eta}\|\mathbf{U}_1 - \mathbf{U}^*\|_{\mathrm{F}}^2 + \frac{\eta L^2 T}{2} \end{aligned} \tag{43}$$

For the left hand side, since the $f_i(\mathbf{U}_t)$ is a decreasing function by increasing t, we can bound it by:

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{m}\hat{\lambda}_i^{(t)}\left(f_i(\mathbf{U}_t) - f_i(\mathbf{U}^*)\right) &\geq \sum_{i=1}^{m}\left(\sum_{t=1}^{T}\hat{\lambda}_i^{(t)}\right)\left(f_i(\mathbf{U}_T) - f_i(\mathbf{U}^*)\right)\\ &= \sum_{i=1}^{m} T\cdot\bar{\lambda}_i\left(f_i(\mathbf{U}_T) - f_i(\mathbf{U}^*)\right), \end{aligned} \tag{44}$$

where $\bar{\lambda}_i = \frac{1}{T}\sum_{t=1}^{T}\hat{\lambda}_i^{(t)}$. By plugging (44) back into (43) we have:

$$\sum_{i=1}^{m}\bar{\lambda}_i\left(f_i(\mathbf{U}_T) - f_i(\mathbf{U}^*)\right) \leq \frac{1}{2\eta T}R^2 + \frac{\eta L^2}{2}, \tag{45}$$

where $\|\mathbf{U}_1 - \mathbf{U}^*\|_{\mathrm{F}}^2 = R^2$. By setting $\eta = \frac{R}{L\sqrt{T}}$, the convergence inequality reduces to:

$$\sum_{i=1}^{m}\bar{\lambda}_i\left(f_i(\mathbf{U}_T) - f_i(\mathbf{U}^*)\right) \leq \frac{RL}{2\sqrt{T}}, \tag{46}$$

We note that by setting $\beta = \sqrt{T}/R$, the sufficient decrease condition in (16) is satisfied if the backtacking is employed. □

D Proof of Theorem 3

Proof of Theorem 3. The proof proceeds using the smoothness condition. In particular, for a smooth function $f : \mathbb{R}^{d\times r} \mapsto \mathbb{R}$ with smoothness parameter L it holds that (descent lemma),

$$f(\mathbf{V}) \leq f(\mathbf{U}) + \mathsf{tr}\left(\nabla f(\mathbf{U}), \mathbf{V} - \mathbf{U}\right) + \frac{L}{2}\|\mathbf{V} - \mathbf{U}\|_{\mathrm{F}}^2.$$

From backtracking line search we can find the learning rate at each step that gives us the maximum decrease. To that end we will start from $\frac{1}{2}$ and decrease it each time by half until all the objective have a maximum decrease defined in (16). Thus, if an η satisfies the condition the one step before that, $2\eta^*$, there is at least one objective not satisfying the condition. For instance, we consider the ith objective does not satisfy the condition with $2\eta^*$:

$$f_i\left(\mathbf{U}_t + (2\eta^*)\,\mathbf{D}_t\right) \geq f_i(\mathbf{U}_t) + \beta\,(2\eta^*)\,\mathsf{tr}\left(\mathbf{D}_t^\top\mathbf{G}_i^{(t)}\right) \tag{47}$$

Now from the Lipschitz continuity of the function as:

$$\begin{aligned} f_i\left(\mathbf{U}_{t+1}\right) &\leq f_i(\mathbf{U}_t) + \mathsf{tr}\left(\nabla f_i(\mathbf{U}_t)^\top\left(\mathbf{U}_{t+1} - \mathbf{U}_t\right)\right) + \frac{L_i}{2}\left\|\mathbf{U}_{t+1} - \mathbf{U}_t\right\|_{\mathrm{F}}^2\\ f_i\left(\mathbf{U}_t + (2\eta^*)\,\mathbf{D}_t\right) &\leq f_i(\mathbf{U}_t) + (2\eta^*)\,\mathsf{tr}\left(\nabla f_i(\mathbf{U}_t)^\top\mathbf{D}_t\right) + \frac{L_i\,(2\eta^*)^2}{2}\|\mathbf{D}_t\|_{\mathrm{F}}^2, \end{aligned} \tag{48}$$

We proceed by combining (47) with this Lipschitz continuity (48) inequality which results in:

$$\mathsf{tr}\left(\mathbf{D}_t^\top \mathbf{G}_i^{(t)}\right) \geq \frac{-L_i \eta^*}{1-\beta} \|\mathbf{D}_t\|_{\mathrm{F}}^2, \tag{49}$$

Also, from (39), we note that the left hand side term has a upper bound of $-\|\mathbf{D}_t\|_{\mathrm{F}}^2$, implying

$$\eta^* \geq \frac{1-\beta}{L_i}, \tag{50}$$

which we can replace L_i with $L_{\max} = \max_{1 \leq i \leq m} L_i$, to obtain the lower bound on learning rate, that is $\eta_t \geq C_1 = \min\{1, \frac{1-\beta}{L_{\max}}\}$. Hence, by the choice of learning rate, we know that at every step, we have the maximum decrease for every objective $1 \leq i \leq m$:

$$\begin{aligned} f_i\left(\mathbf{U}_{t+1}\right) &\leq f_i(\mathbf{U}_t) + \beta \eta_t \, \mathsf{tr}\left(\mathbf{D}_t^\top \mathbf{G}_i^{(t)}\right) \\ &\overset{①}{\leq} f_i(\mathbf{U}_t) - \beta \eta_t \|\mathbf{D}_t\|_{\mathrm{F}}^2 \\ &\overset{②}{\leq} f_i(\mathbf{U}_t) - \beta C_1 \|\mathbf{D}_t\|_{\mathrm{F}}^2 \end{aligned}$$

where ① comes from Lemma 1 and (39), and ② follows from the bound on η_t in (50).

Summing up the last inequality for all iterations $t = 1, \ldots, T$ and setting $C = \beta C_1$, we obtain:

$$\sum_{t=1}^{T} C \|\mathbf{D}_t\|_{\mathrm{F}}^2 \leq \sum_{t=1}^{T} \left(f_i\left(\mathbf{U}_t\right) - f_i\left(\mathbf{U}_{t+1}\right)\right), \tag{51}$$

In (51), the left hand side is greater than $\min_{t=1,2,\ldots,T} CT \|\mathbf{D}_t\|_{\mathrm{F}}^2$; and the right hand telescopes and can be upper bounded by $\mathsf{M}_u - \mathsf{M}_l$, where $\mathsf{M}_u = \max_{i=1,2,\ldots,m} f_i(\mathbf{U}_1)$ is the maximum value among objective functions at starting point and M_l is the lower bound on all objectives. Then the inequality becomes:

$$\min_{t=1,2,\ldots,T} \|\mathbf{D}_t\|_{\mathrm{F}} \leq \sqrt{\frac{\mathsf{M}_u - \mathsf{M}_l}{CT}} \tag{52}$$

indicating that the proposed algorithm convergence to a Pareto stationary point (a point where the descent direction is $\mathbf{0}$ and none of the objectives can be further improved). □