Fast and Efficient MMD-based Fair PCA via Optimization over Stiefel Manifold

Junghyun Lee,1 Gwangsu Kim,2 Matt Olfat,3,4 Mark Hasegawa-Johnson,5 Chang D. Yoo2
1 Graduate School of AI, KAIST, Seoul, Republic of Korea
2 School of Electrical Engineering, KAIST, Daejeon, Republic of Korea
3 UC Berkeley IEOR, Berkley, CA, USA
4 Citadel, Chicago, IL, USA
5 Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, IL, USA
{jh_lee00, s88012}@kaist.ac.kr, molfat@berkeley.edu, jhasegaw@illinois.edu, cd_yoo@kaist.ac.kr

Abstract
This paper defines fair principal component analysis (PCA) as minimizing the maximum mean discrepancy (MMD) between dimensionality-reduced conditional distributions of different protected classes. The incorporation of MMD naturally leads to an exact and tractable mathematical formulation of fairness with good statistical properties. We formulate the problem of fair PCA subject to MMD constraints as a non-convex optimization over the Stiefel manifold and solve it using the Riemannian Exact Penalty Method with Smoothing (REPMS; Liu and Boumal, 2019). Importantly, we provide local optimality guarantees and explicitly show the theoretical effect of each hyperparameter in practical settings, extending previous results. Experimental comparisons based on synthetic and UCI datasets show that our approach outperforms prior work in explained variance, fairness, and runtime.

1 Introduction
It has become increasingly evident that many widely-deployed machine learning algorithms are biased, yielding outcomes that can be discriminatory across key groupings such as race, gender and ethnicity (Mehrabi et al. 2019). As the applications of these algorithms proliferate in protected areas like healthcare (Karan et al. 2012), hiring (Chien and Chen 2008) and criminal justice (Kirchner et al. 2016), this creates the potential for further exacerbating social biases. To address this, there has recently been a surge of interest in ensuring fairness in resulting machine learning algorithms.

Working in high-dimensional spaces can be undesirable as the curse of dimensionality manifests in the form of data sparsity and computational intractability. Various dimensionality reduction algorithms are deployed to resolve these issues, and Principal Component Analysis (PCA) (Pearson 1901; Hotelling 1933), is one of the most popular methods (Jolliffe and Cadima 2016). One particular advantage of PCA is that there’s no need to train a complex neural network.

In this work, fair PCA is defined as doing PCA while minimizing the difference in the conditional distributions of projections of different protected groups. Here, the projected data can be considered as a dimension-reduced fair representation of the input data (Zemel et al. 2013). We answer the questions of 1) how fairness should be defined for PCA and 2) how to (algorithmically) incorporate fairness into PCA in a fast and efficient manner. This work takes a different approach from prior studies on PCA fairness (Samadi et al. 2018; Olfat and Aswani 2019), which is discussed in Section 3 and 7.

Our main contributions are as follows:

• We motivate a new mathematical definition of fairness for PCA using the maximum-mean discrepancy (MMD), which can be evaluated in a computationally efficient manner from the samples while guaranteeing asymptotic consistency. Such properties were not available in the previous definition of fair PCA (Olfat and Aswani 2019). This is discussed in detail in Section 3 and 4.

• We formulate the task of performing MMD-based fair PCA as a constrained optimization over the Stiefel manifold and propose using REPMS (Liu and Boumal 2019). For the first time, we prove two general theoretical guarantees of REPMS regarding the local minimality and feasibility. This is discussed in detail in Section 5 and 6.

• Using synthetic and UCI datasets, we verify the efficacy of our approach in terms of explained variance, fairness, and runtime. Furthermore, we verify that using fair PCA does indeed result in a fair representation, as in (Zemel et al. 2013). This is discussed in detail in Section 7.

2 Preliminaries

2.1 Notations
For \( b \geq 1 \), let \( \mathcal{P}_b \) be the set of all Borel probability measures defined on \( \mathbb{R}^b \). For some measurable function \( \Pi : \mathbb{R}^p \to \mathbb{R}^d \) and a measure \( P \in \mathcal{P}_p \), the push-forward measure of \( P \) by \( \Pi \) is the probability measure \( \Pi_\# P \in \mathcal{P}_d \), defined as \( \Pi_\# P(S) = P(\Pi^{-1}(S)) \) for any Borel set \( S \). Let 0 and 1 denote matrices (or vectors) of zeros and ones of appropriate size, respectively. In this work, we focus on binary cases, i.e., we assume that the protected attribute \( A \) and outcome \( Y \) are binary \( (A, Y) \in \{0, 1\} \). We abbreviate demographic parity, equalized opportunity, and equalized odds as DP, EOP, and EOD, respectively.

2.2 Maximum Mean Discrepancy (MMD)
Let \( k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \) be a positive-definite kernel function, and \( H_k \) be a unit-ball in the RKHS generated by \( k \). We impose some regularity assumptions on \( k \):
**Assumption 1.** $k$ is measurable, and bounded i.e. $K := \sup_{x,y} k(x,y) < \infty$.

Then one can pseudo-metrize $\mathcal{P}_d$ by the following distance:

**Definition 1** (Gretton et al. [2007]). Given $\mu, \nu \in \mathcal{P}_d$, their maximum mean discrepancy (MMD), denoted as $MMD_k(\mu, \nu)$, is a pseudo-metric on $\mathcal{P}_d$, defined as follows:

$$MMD_k(\mu, \nu) := \sup_{f \in H_s} \left| \int \! f \, d(\mu - \nu) \right|$$ (1)

As our fairness constraint involves exactly matching the considered distributions using MMD, we require the property of $MMD_k(\mu, \nu) = 0$ implying $\mu = \nu$. This is formalized as follows:

**Definition 2** (Fukumizu et al. [2008]). If $MMD_k$ metrizes $\mathcal{P}_d$, then the kernel $k$ is said to be characteristic to $\mathcal{P}_d$.

Sriperumbudur et al. [2008] defined and characterized stationary characteristic kernels and identified that well-known kernels such as RBF and Laplace are characteristic. Based on this fact, we set $k$ to be the RBF kernel $k_{rbf}(x,y) := \exp(-\|x - y\|^2 / 2\sigma^2)$.

For the choice of bandwidth $\sigma$, the median of the set of pairwise distances of the samples after vanilla PCA is considered following the median heuristic of Schölkopf, Smola, and Müller [1998]. For simplicity, we refer to $MMD_{k_{rbf}}$ as $MMD$.

**Benefits of MMD** There are several reasons for using MMD as the distance on a space of probability measures. First, it can act as a distance between distributions with different, or even disjoint, supports. This is especially crucial as the empirical distributions are often discrete and completely disjoint. Such a property is not generally true, one prominent example being the KL-divergence. Second, since many problems in fairness involve comparing two distributions, $MMD$ has already been used in much of the fairness literature as a metric [Madras et al. 2018, Adel et al. 2019] and as an explicit constraint/penalty [Quadrianto and Sharmanska 2017, Louizos et al. 2016, Frost et al. 2019, Oneto et al. 2020, Jung et al. 2021], among other usages.

### 2.3 Fairness for supervised learning

The fair PCA discussed above should ultimately lead to fairness in supervised learning tasks based on the dimensionality-reduced data with minimal loss in performance. Let us now review three of the most widely-used definitions of fairness in supervised learning, as formulated in Madras et al. [2018].

Let $(Z, Y, A) \in \mathbb{R}^d \times \{0, 1\} \times \{0, 1\}$ be the joint distribution of the dimensionality-reduced data, (downstream task) label, and protected attribute. Furthermore, let $g : \mathbb{R}^d \rightarrow \{0, 1\}$ be a classifier that outputs prediction $\hat{Y}$ for $Y$ from $Z$. We want to determine the fairness of a well-performing classifier $g$ w.r.t. protected attribute $A$.

First, let $D_s$ be the probability measure of $Z_s \triangleq Z|A = s$ for $s \in \{0, 1\}$.

**Definition 3** (Feldman et al. [2015]). $g$ is said to satisfy demographic parity (DP) up to $\Delta_{DP}$ w.r.t. $A$ with $\Delta_{DP} \triangleq |E_{x \sim D_0}[g(x)] - E_{x \sim D_1}[g(x)]|$.

Now, let $D_{s,y}$ be the probability measure of $Z_{s,y} \triangleq Z|A = s, Y = y$ for $s, y \in \{0, 1\}$.

**Definition 4** [Hardt, Price, and Srebro 2016]. $g$ is said to satisfy equalized opportunity (EOP) up to $\Delta_{EOP}$ w.r.t. $A$ and $Y$ with $\Delta_{EOP} \triangleq |E_{x \sim D_0,y}[g(x)] - E_{x \sim D_1,y}[g(x)]|$.

**Definition 5** [Hardt, Price, and Srebro 2016]. $g$ is said to satisfy equalized odds (EOD) up to $\Delta_{EOD}$ w.r.t. $A$ and $Y$ with $\Delta_{EOD} \triangleq \max_{y \in \{0, 1\}} |E_{x \sim D_{0,y}}[g(x)] - E_{x \sim D_{1,y}}[g(x)]|$.

From hereon, we refer to such $\Delta_f(g)$ as the fairness metric of $f \in \{DP, EOP, EOD\}$ w.r.t. $g$, respectively.

### 3 New definition of fairness for PCA

For $p > d$, let $\mathbb{R}^d$ be the space onto which data will be projected. A dimensionality reduction is a map $\Pi : \mathbb{R}^p \rightarrow \mathbb{R}^d$, and PCA is defined as $\Pi(x) = V^T x$ for some $V \in \mathbb{R}^{p \times d}$ satisfying $\|V^T V - I_d\|_2$ i.e. PCA is a linear, orthogonal dimensionality-reduction. From hereon, we denote a linear PCA as the mapping $\Pi$. The definition for fairness that we will be following throughout is given as follows.

**Definition 6** ($\Delta$-fairness). Let $P_d$ be the probability measure of $X_s \triangleq X|A = s$ for $s \in \{0, 1\}$, and let $Q_s := \Pi_{\#}P_d \in \mathcal{P}_d$. Then $\Pi$ is said to be $\Delta$-fair with $\Delta := MMD(Q_0, Q_1)$, and we refer to $\Delta$ as the fairness metric.

In other words, for lower $\Delta$-fairness, the discrepancy between the dimensionality-reduced conditional distributions of different protected classes, measured in a non-parametric manner using $MMD$, while retaining as much variance as possible, should be minimized.

Furthermore, Definition 6 ensures that a downstream classification task using $\Delta$-fair dimensionality-reduced data will be fair, as formalized below.

**Proposition 1** [Oneto et al. 2020]. Up to a constant factor, $MMD(Q_0, Q_1)$ bounds the $MMD$ of the push-forward measures of $Q_0, Q_1$ via the weight vector of any given downstream task classifier $g$.

**Remak 1.** The above discussions easily generalize to equal opportunity and equalized odds.

#### 3.1 Relation with other definitions of fair PCA

The notion of fairness proposed by [Olfat and Aswani 2019] is similar to ours in that it measures the predictability of protected group membership in dimensionality-reduced data. However, unlike ours, their definition is explicitly adversarial, which can be a problem.

**Definition 7** ($\Delta_A$-fairness; [Olfat and Aswani 2019]). Consider a fixed classifier $h(u, t) : \mathbb{R}^d \times \mathbb{R} \rightarrow \{0, 1\}$ that inputs features $u \in \mathbb{R}^d$ and a threshold $t$, and predicts the protected class $s \in \{0, 1\}$. Then, $\Pi$ is $\Delta_A(h)$-fair if

$$\sup_{t \in \mathbb{R}} \left| P[h(\Pi(x), t) = 1|s = 1] - P[h(\Pi(x), t) = 1|s = 0] \right| \leq \Delta_A(h)$$ (2)

1The benefits of pursuing orthogonality in the loading matrix, and thus the resulting PCs, are already well-studied; for example, see Qi, Luo, and Zhao [2013], Bendelis et al., 2016.

2See Lemma 3 of Oneto et al. [2020] for the precise statement.
Moreover, for a family of classifiers $\mathcal{F}_c$, if $\Pi$ is $\Delta_A(h)$-fair for $\forall h \in \mathcal{F}_c$, we say that $\Pi$ is $\Delta_A(\mathcal{F}_c)$-fair.

**Remark 2.** It can be argued that, for some choice of $\mathcal{F}_c$, Definition 6 and 7 are equivalent: in effect, that these are dual notions. Recognizing this, we proceed with Definition 6 as it has two main advantages in the context of our work:

- It ties more directly and intuitively into our optimization formulation; see Section 5
- It can be represented non-variationally which allows for tighter statistical guarantees.

### 4 Statistical properties of $\Delta$

#### 4.1 Consistent and efficient estimation of $\Delta$

As defined in Definition 6 let $Q_0, Q_1 \in \mathcal{P}_d$ be the probability measures with respect to the samples of which we want to estimate $\text{MMD}(\cdot, \cdot)$. Let $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^n$ be these samples, respectively. Accordingly, we consider the following estimator:

$$\hat{\Delta} := \text{MMD}(\hat{Q}_0, \hat{Q}_1)$$

where $\hat{Q}_p$ is the usual empirical distribution, defined as the mixture of Dirac measures on the samples.

Unlike other statistical distances (e.g. total variation distance), $\hat{\Delta}_k$ has several theoretical properties that have important practical implications; see [Sriperumbudur et al. 2010] for more details.

First, it can be computed exactly and efficiently:

**Lemma 1** [Gretton et al. 2007]. $\hat{\Delta}$ is computed as follows:

$$\hat{\Delta} = \left[ \frac{1}{m^2} \sum_{i,j=1}^{m} k(X_i, X_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(Y_i, Y_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(X_i, Y_j) \right]^{1/2}.$$  \hfill (4)

Moreover, it is asymptotically consistent with a convergence rate, depending only on $m$ and $n$:

**Theorem 1** [Gretton et al. 2007]. For any $\delta > 0$, with probability at least $1 - 2 \exp\left(-\frac{\delta^2}{2(m+n)}\right)$ the following holds:

$$|\Delta - \hat{\Delta}| \leq 2 \left( \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}} \right) + \delta$$  \hfill (5)

#### 4.2 Advantages over $\Delta_A$

In their work, [Olfat & Aswani 2019] propose the following estimator for $\Delta_A$:

$$\hat{\Delta}_A(\mathcal{F}_c) := \sup_{h \in \mathcal{F}_c} \sup_{h \in P} \sum_{i \in P} I_i(\Pi, h) - \frac{1}{N} \sum_{i \in N} I_i(\Pi, h_k)$$  \hfill (6)

where $\{x_i\}_{i=1}^n$ are the data points, $(P, N)$ is a partition of the index set $\{1, 2, \ldots, n\}$ into two protected groups, $I_i(\Pi, h_k) = 1(h(\Pi(x_i), t) = +1)$, and $1(\cdot)$ is the indicator function.

Such estimator is known to satisfy the following high probability bound:

**Proposition 2** [Olfat & Aswani 2019]. Consider a fixed family of classifiers $\mathcal{F}_c$. Then for any $\delta > 0$, with probability at least $1 - \exp\left(-\frac{\delta^2}{2(n+m)^2}\right)$ the following holds:

$$|\Delta_A(\mathcal{F}_c) - \hat{\Delta}_A(\mathcal{F}_c)| \leq 8 \sqrt{\frac{\text{VC}(\mathcal{F}_c)}{m+n}} + \delta$$  \hfill (7)

where $\text{VC}(\cdot)$ is the VC dimension.

**Remark 3.** If $\mathcal{F}_c$ is too expressive in terms of VC-dimension, then the above bound may become void. This is the case, for instance, when $\mathcal{F}_c$ is the set of RBF-kernel SVMs.

In addition, computing $\hat{\Delta}_A$ requires considering all possible classifiers in the designated family $\mathcal{F}_c$. This is computationally infeasible, and it forces one to use another approximation (e.g. discretization of $\mathcal{F}_c$), which incurs additional error that may further inhibit asymptotic consistency. As exhibited in the previous subsection, our $\text{MMD}$-based approach suffers from none of these issues.

### 5 Manifold optimization for MbF-PCA

#### 5.1 Improvements over FPCA

[Olfat & Aswani 2019] proposed FPCA, an SDP formulation of fair PCA in which matching the first and second moments of the protected groups after dimensionality-reduction are approximated as convex constraints. However, this has several shortcomings, which we discuss here and empirically exhibit in a later section.

First, there are cases in which matching the mean and covariance alone is not enough. The simplest “counterexample” would be when two protected groups have the same mean and covariance, yet they have different distributions. This is illustrated in Figure 1. While this previous point can be countered by the application of the kernel trick to FPCA, this raises a second issue: their formulation requires solving a $p \times p$-dimensional SDP, motivated by the reparameterization $P = V V^\top$ [Arora, Cotter, and Srebro 2013]. Since SDP is known to become inefficient (or even computationally infeasible) in high dimensions, this quickly becomes intractable for high-dimensional data (for linear or polynomial kernels) or for any moderate to large size datasets (for the RBF kernel). Finally, their approach involves a relaxation of a rank constraint ($\text{rank}(P) \leq d$) to a trace constraint ($\text{tr}(P) \leq d$), yielding sub-optimal outputs in presence of (fairness) constraints, even to substantial order in some cases. In Section C of the supplementary material (SP), we discuss in detail why FPCA may lead to such degraded explained variance.

#### 5.2 Formulating MbF-PCA

Observing that the shortcomings of FPCA stem from the reparametrization of $P = V V^\top$, we propose a new formulation of fair PCA that solves directly for $V$. This allows for an effective and efficient approach.

We start by noting that the set of all $V$ with orthonormal columns has the intrinsic geometric structure of a manifold:

---

3See Section B of the supplementary material for its precise description.

4In their approach, the final solution $V$ is obtained by taking the first $d$ eigenvectors of $P$. 


Definition 8. For $p > d$, the Stiefel manifold, denoted as $St(p, d)$, is an embedded Riemannian sub-manifold of $\mathbb{R}^{p \times d}$ such that each element of $St(p, d)$ has orthonormal columns $i.e. V^TV = I_d$ for all $V \in St(p, d)$.

$St(p, d)$ has several desirable properties such as compactness, completeness and smoothness, which we present in Section C of the SP. As $St(p, d)$ is prevalent in various fields of machine learning (most notably PCA), much work has been done that focuses on exploiting this geometric structure for efficient optimization (Han and Mangasarian 1979), (Boumal, 2019). Note that smoothing tricks (Liu and Boumal 2019). This is due to our choice of estimator for $MMD$ inducing a non-negative estimate of $MMD^2$; see Section 2 of Gretton et al. (2012) for more detailed discussions.

For practical concerns that will be addressed in the following subsection, we’ve set the fairness tolerance level, $\tau$, to be a fixed and sufficiently small, non-negative value. Formally, we consider the following definition:

**Definition 9.** For fixed $\tau \geq 0$, $V \in St(p, d)$ is $\tau$-approximate fair if it satisfies $h(V) \leq \tau$. If $\tau = 0$, we simply say that $V$ is fair.
6.2 New theoretical guarantees for Algorithm 1

We start by observing that Eq. (9) in Algorithm 1 is smooth, unconstrained manifold optimization problem, which can be solved using conventional algorithms; these include first-order methods like line-search methods (Absil, Mahony, and Sepulchre 2007), or second-order methods like the Riemannian Trust Region (RTR; Absil, Baker, and Gallivan 2007) method. It is known that, pathological examples excluded, most conventional unconstrained manifold optimization solvers produce iterates whose limit points are local minima, and not other stationary points such as saddle point or local maxima; see (Absil, Baker, and Gallivan 2007; Absil Mahony, and Sepulchre 2007) for more detailed discussions.

Motivated by this, we consider the following assumption:

Assumption 2 (informal; locality assumption). Each $V_{k+1}$ is sufficiently close to a local minimum of Eq. (9).

Lastly, we consider the following auxiliary optimization problem:

$$\min_{V \in St(p,d)} h(V) \tag{10}$$

The following theorem, whose proof is deferred to Section 6.2.1 of the SP, provides an exact theoretical convergence guarantee of MBF-PCA under the ideal hyperparameter setting:

**Theorem 2.** Let $K = \infty$, $\rho_{\text{max}} = \infty$, $\epsilon_{\text{min}} = \tau = 0$, $\{V_k\}$ be the sequence generated by Alg. 1 under Assumption 2 and $V$ be any limit point of $\{V_k\}$, whose existence is guaranteed. Then the following holds:

(A) $V$ is a local minimizer of Eq. (10), which is a necessary condition for $V$ to be fair.

(B) If $V$ is fair, then $V$ is a local minimizer of Eq. (9).

The assumption of $V$ being fair, which is used in (B), is at least partially justified in (A) in the following sense: the ideal hyperparameter setting of $\rho_{\text{max}} = \infty$, $\tau = 0$, $\epsilon_{\text{min}} = 0$ implies the exact local minimality of $V$ for Eq. (10), which is in turn a necessary condition for $V$ to be fair.

The next theorem, whose proof is also deferred to Section 6.2.1 of the SP, asserts that with small $\tau$, $\epsilon_{\text{min}}$ and large $\rho_{\text{max}}$, the above guarantee can be approximated in rigorous sense:

**Theorem 3.** Let $K = \infty$, $\rho_{\text{max}} < \infty$, $\epsilon_{\text{min}}$, $\tau > 0$, $\{V_k\}$ be the sequences generated by Alg. 1 under Assumption 2 and $V$ be any limit point of $\{V_k\}$, whose existence is guaranteed. Then for any sufficiently small $\epsilon_{\text{min}}$ and $\tilde{r} = \tilde{r}(\epsilon_{\text{min}}) > 0$, the following hold:

(A) $V$ is an approximate local minimizer of Eq. (10) in the sense that

$$h(V) \leq h(V) + \beta\|V - V\|^2 + (\beta + L_k)g(\epsilon_{\text{min}}) \tag{11}$$

for all $V \in B_\tilde{r}(V) \cap St(p,d)$, where $\beta = \beta(\rho_{\text{max}}, \tau)$ is a function that satisfies the following:

• $0 < \beta \leq \frac{2\|\Sigma\|}{\rho_0}$

• $\beta(\rho_{\text{max}}, \tau)$ is increasing in $\rho_{\text{max}}$ and decreasing in $\tau$.

(B) If $V$ is fair, then it is an approximate local minimizer of Eq. (9) in the sense that

$$f(V) \leq f(V) + 2\|\Sigma\|g(\epsilon_{\text{min}}) \tag{12}$$

for all fair $V \in B_\tilde{r}(V) \cap St(p,d)$.

In both (A) and (B), $g$ is some continuous, decreasing function that satisfies $g(0) = 0$, and $\tilde{r}(\epsilon_{\text{min}}) = r - g(\epsilon_{\text{min}})$ for some fixed constant $r > 0$.

Existing optimality guarantee of REPS (Proposition 4.2; Liu and Boumal 2019) states that when $\epsilon_{\text{min}} = 0$, $\rho$ is not updated (i.e. line 10-14 is ignored), and the resulting limit point is feasible, then that limit point satisfies the KKT condition (Yang, Zhang, and Song 2014). Comparing Theorem 2 and 3 to the previous result, we see that ours extend the previous result in several ways:

• Our theoretical analyses are much closer to the actual implementation, by incorporating the $\rho$-update step (line 11) and the practical hyperparameter setting.

• Our theoretical analyses are much more stronger in the sense that 1) by introducing a reasonable, yet novel locality assumption, we go beyond the existing KKT conditions and prove the local minimality of the limit point, and 2) we provide a partial justification of the feasibility assumption in (A) by proving a necessary condition for it.

6.3 Practical implementation

In line 4 in Algorithm 1, we implemented the termination criteria: sufficiently small distance between iterates and sufficiently small tolerance for solving Eq. (9). However, such a heuristic may return some point $V$ that is not $\tau$-approximate fair for user-defined level $\tau$ in practical hyperparameter setting. To overcome this issue, we’ve additionally implemented line 5 that forces the algorithm to continue on with the loop until the desired level of fairness is achieved.

7 Related Work

7.1 Fairness in ML

A large body of work regarding fairness in the context of supervised learning (Feldman et al. 2015; Calders, Kamiran, and Pechenizkiy 2009; Dwork et al. 2012; Hardt, Price, and Srebro 2016; Zafar et al. 2017) has been published. This includes key innovations in quantifying algorithmic bias, notably the concepts of demographic parity and equalized odds (opportunity) that have become ubiquitous in fairness research (Barocas and Selbst 2016; Hardt, Price, and Srebro 2016). More recently, fair machine learning literatures have branched out into a variety of fields, including deep learning (Beutel et al. 2017), regression (Calders et al. 2013), and even hypothesis testing (Olfat et al. 2020).

Among these, one line of research has focused on learning fair representations (Kamiran and Calders 2011; Zemel et al. 2013; Feldman et al. 2015; Calmon et al. 2017), which aims to learn a representation of the given data on which various fairness definitions are ensured for downstream modeling. A growing number of inquiries have been made into highly specialized algorithms for specific unsupervised learning problems like clustering (Chierichetti et al. 2017; Kleindessner, Awasthi, and Morgenstern 2019; Bera et al. 2019), but these lack the general applicability of key dimensionality reduction algorithms such as PCA (Pearson 1901; Hotelling 1933).

To the best of our knowledge, Olfat & Aswani (2019) is the only work on incorporating fair representation to PCA,
making it the sole comparable approach to ours. Another line of work (Bian and Tao 2011, Samadi et al. 2018, Tan-tipongpipat et al. 2019, Zalcberg and Wiesel 2021) considers a completely orthogonal definition of fairness for PCA: minimizing the discrepancy between reconstruction errors over protected attributes. This doesn’t ensure the fairness of downstream tasks, rendering it incomparable to our definition of fairness; see Section A of the SP for more details.

7.2 Manifold Optimization

A constrained problem over Euclidean space can be transformed to an unconstrained problem over a manifold (or at least manifold optimization with less constraints). Many algorithms for solving Euclidean optimization problems have direct counterparts in manifold optimization problems that includes Riemannian gradient descent and Riemannian BFGS. By making use of the geometry of lower dimensional manifold structure, often embedded in potentially very high dimensional ambient space, such Riemannian counterparts are much more computationally efficient than algorithms that do not make use of manifold structure. This is shown in numerous literatures (Liu and Boumal 2019, Alsharif et al. 2021, Meng, Chakraborty, and Singh 2021), including this work. We refer interested readers to the standard textbooks (Absil, Mahony, and Sepulchre 2007) on this field, along with a survey by Hu et al. (2020).

8 Experiments

MbF-PCA was implemented using ROPTLIB (Huang et al. 2018), a state-of-the-art manifold optimization framework on MATLAB. For solving Eq. (9), we use the cautious Riemannian BFGS method (RBFGS; Huang, Absil, and Galván 2018), a quasi-Newton method that is much more memory-efficient. We’ve set $K = 100, \epsilon_{\text{min}} = 10^{-6}, \epsilon_0 = 10^{-1}, \theta_\epsilon = (\epsilon_{\text{min}}/\epsilon_0)^{1/5}, \rho_{\text{max}} = 10^{10}, \theta_\rho = 2, d_{\text{min}} = 10^{-6}$. For FPCA, we use the same Python MOSEK(ApS 2021) implementation as provided by (Olfat and Aswani 2019). $(\mu, \delta)$ are the hyperparameters of FPCA; see Section B of the SP. Codes are available in our Github repository.

All data is pre-processed to be standardized such that each covariate has zero mean and unit variance. For all experiments, we considered 10 different 70 – 30 train-test splits.

8.1 Synthetic data #1

We consider synthetic data composed of two groups, each of size $n = 150$; one is sampled from $N_3(0, 0.1I_3 + 1)$ and one is sampled from a (balanced) mixture of $N_3(1, 0.1I_3)$ and $N_3(-1, 0.1I_3)$. Note how the two groups follow different distributions, yet have the same mean and covariance. Thus, we expect FPCA to project in a similar way as vanilla PCA, while MbF-PCA should find a fairer subspace such that the projected distributions are exactly the same. Hyperparameters are set as follows: $\delta = 0, \mu = 0.01$ for FPCA and $\tau = 10^{-5}$ for MbF-PCA. We’ve set $d = 2$ and Figure 1 displays the results of each algorithm using the top two principal components. Indeed, only MbF-PCA successfully

8.2 Synthetic data #2

We consider a series of synthetic datasets of dimension $p$. For each $p$, the dataset is composed of two groups, each of size $n = 240$ and sampled from two different $p$-variate normal distributions. We vary $p \in \{20, 30, \ldots, 100\}$, which approximates the three asymptotic regimes: $p < n, p \gg n$, and $p > n$. See Section C of the SP for a full description of the setting. For the hyperparameters, we’ve set $\delta = 0, \mu = 0.01$ for FPCA and $\tau = 10^{-5}$ for MbF-PCA.

Figure 2 plots the explained variance and fairness metric values. Observe how MbF-PCA achieves better explained variance, while achieving similar level of fairness. In addition, Figure 3 shows a clear gap in runtime between FPCA and MbF-PCA; the runtime of FPCA explodes for even moderate problem sizes, while MbF-PCA scales well. For higher dimensions, conventional computing machine will not be able to handle such computational burden.

https://github.com/nick-jhlee/fair-manifold-pca
Table 1: Comparison of PCA, FPCA, MBF-PCA for UCI datasets. Number in parenthesis for each dataset is its dimension. Also, the parenthesis for each fair algorithm is its hyperparameter setting: (µ, δ) for FPCA and τ for MBF-PCA. Among the fair algorithms considered, results with the best mean values are bolded. Results in which our approach terminates improperly in the sense that the maximum iteration is reached before passing the termination criteria are highlighted.

<table>
<thead>
<tr>
<th>d</th>
<th>ALC</th>
<th>PCA</th>
<th>FPCA (0.1, 0)</th>
<th>FPCA (0.1, 0.01)</th>
<th>M-BF-PCA (10⁻²)</th>
<th>M-BF-PCA (10⁻¹)</th>
<th>M-BF-PCA (10⁻²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>%VAR</td>
<td>%ACC</td>
<td>MMD²</td>
<td>ΔDP</td>
<td>%VAR</td>
<td>%ACC</td>
<td>MMD²</td>
</tr>
<tr>
<td>10</td>
<td>9.02%</td>
<td>64.53%</td>
<td>0.092%</td>
<td>0.27%</td>
<td>11.42%</td>
<td>76.87%</td>
<td>0.147%</td>
</tr>
</tbody>
</table>

Figure 4: Comparison of communality of “age” of German credit dataset for PCA, FPCA, and MBF-PCA.

8.3 UCI datasets

For target dimensions d ∈ {2, 10}, we compare the performance of FPCA and MBF-PCA on 3 datasets from the UCI Machine Learning Repository (Dua and Graff 2017): COMPAS dataset (Kirchner et al. 2016; Adult income dataset, and German credit dataset. See Section I of the SP for complete description of the pre-processing steps. For both algorithms, we consider two different hyperparameter settings, such that one simulates the relaxed fairness while the other simulates a stricter fairness constraints. For computing ΔDP(τ), we trained a RBF SVM to be the downstream task classifier that best classifies the target attribute in the dimensionality-reduced data. Table 1 displays the results, from which several observations can be made:

- Across all considered datasets, MBF-PCA is shown to outperform FPCA in terms of fairness (both MMD² and ΔDP) with low enough τ.
- For GERMAN CREDIT and ADULT INCOME, MBF-PCA shows a clear trade-off between explained variance and fairness: by relaxing τ, we see that MBF-PCA outperforms FPCA in terms of explained variance and downstream task accuracy.

In addition, to see how correlated are the PCs with the protected attribute, we examine the communalities. For clarity of exposition, we consider the German credit dataset, whose protected attribute is age, and d = 10. Here, we again consider PCA, FPCA (0, 0.01), and MBF-PCA (10⁻³). For PCA, communality of a feature is its variance contributed by the PCs (Johnson and Wichern 2008), which is computed as the sum of squares of the loadings of the considered feature. Larger value of communality implies that the correlations between the considered feature and the PCs are strong. Figure 4 displays the boxplot of communality of considered 10 splits. Indeed the amount of variance in age that is accounted for from the loadings of MBF-PCA is much smaller than that of PCA or FPCA. In other words, the PCs resulting from MBF-PCA have the least correlations with age, the protected attribute. The complete results (PC biplots and communality box plots) for the considered UCI datasets are presented in Section I of the SP.

9 Conclusion and Future Works

We present a MMD-based definition of fair PCA, and formulate it as a constrained optimization over the Stiefel manifold. Through both theoretical and empirical discussions, we show that our approach outperforms the previous approach (Olfat and Aswani 2019) in terms of explained variance, fairness, and runtime. Many avenues remain for future research. Statistical characterizations of our fair PCA in asymptotic regime, as well as incorporation of sparsity (Johnstone et al. 2009) are important open questions. Incorporating stochastic optimization-type modifications (Shamir 2015; Roh et al. 2021) is also an important direction, as such modifications are expected to result in better scalability and performance.

References


A Relation to (Samadi et al. 2018)

There is another branch work (Samadi et al. 2018; Tantipongpipat et al. 2019; Mahdi Kamani et al. 2019; Pelegrina et al. 2020) in which the task of solving fair PCA is defined as finding a (linear) subspace in which the reconstruction errors for each protected class are approximately equal. Here, we emphasize again that their fair PCA is incomparable with our considered fair PCA.

Consider the task of performing PCA on a two-dimensional data as shown in Figure 5, where $A = \{0, 1\}$ denotes a protected attribute and the target dimension is set to $d = 1$. We assume that the data distributions for both $A = 0$ and $A = 1$ take an elliptical density, the only difference being their orientation and position.

Fairness in reconstruction error can be achieved by projecting the data onto blue or red subspace. However, fair PCA defined by [Samadi et al. 2018] aims to minimize the reconstruction error with the fairness constraint, and thus the blue subspace is chosen. Under our MMD-based definition of fair PCA, projecting onto the blue subspace is considered unfair because the distributions conditioned on each protected attribute $A$ are not the same; our fair PCA would choose the red subspace.

One important observation is that while our definition of fairness ensures that the downstream tasks are fair w.r.t many existing fairness definitions (see Proposition 1), former definition has no such guarantees.

B Brief Description of FPCA

This section follows the discussion in [Olfat and Aswani 2019], and by “they”, we refer to its authors. First recall their proposed adversarial definition of fairness for PCA:

**Definition 7 ($\Delta_A$-fairness; Olfat and Aswani 2019).** Consider a fixed classifier $h(u, t): \mathbb{R}^d \times \mathbb{R} \rightarrow \{0, 1\}$ that inputs features $u \in \mathbb{R}^d$ and a threshold $t$, and predicts the protected class $s \in \{0, 1\}$. Then, $\Pi$ is $\Delta_A(h)$-fair if

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}[h(\Pi(x), t) = 1|s = 1] - \mathbb{P}[h(\Pi(x), t) = 1|s = 0] \right| \leq \Delta_A(h)$$

(2)

Moreover, for a family of classifiers $F_c$, if $\Pi$ is $\Delta_A(h)$-fair for $\forall h \in F_c$, we say that $\Pi$ is $\Delta_A(F_c)$-fair.

B.1 Fairness Constraints

By choosing $F_c$ to be $F_{lin} = \{h(u, t) = 1(w \tau u - t) \leq 0 : w \in \mathbb{R}^d\}$, $\Delta_A(F_{lin})$ becomes the Kolmogorov distance between $X_s' \triangleq w \tau V \tau X|A = s$, which is upper bounded by their total variation distance. Applying Pinsker’s inequality [Massart 2007], we have that

$$\Delta_A(F_{lin}) \leq \sqrt{\frac{1}{2} KL(X_s'||X'_s)}$$

In order to obtain a tractable form of fairness constraints, they make normality assumptions on the data as follows: $X_s = X|A = s \sim N_p(\mu_s, \Sigma_s)$. Then an upper bound of closed form can be obtained as follows:

$$\Delta_A(F_{lin}) \leq \frac{1}{\sqrt{2}} \sqrt{\frac{1}{|\Sigma_s|} + \text{tr}(\Sigma_s^{-1}\Sigma_0) - d + (\mu_1 - \mu_0)^{\tau}\Sigma_s^{-1}(\mu_1 - \mu_0)},$$

and thus, the sufficient condition for the RHS to be 0 is that $\mu_0 = \mu_1$ and $\Sigma_0 = \Sigma_1$.

Denoting $f = \mu_1 - \mu_0$ and $Q = \Sigma_1 - \Sigma_0$, they introduced two constraints:

- Mean constraint: $h_m(V) := \|V^{\tau}f\| = 0$
- Covariance constraint: $h_c(V) = \|V^{\tau}QV\|_2 = 0$

Based on these, the optimization considered in their work is as follows:

$$\text{minimize} \quad f(V) = -\langle \Sigma, VV^{\tau} \rangle$$

subject to $V^{\tau}V = I_d$, $h_m(V) = 0$, $h_c(V) = 0$. (13)

B.2 SDP Formulation

Standard convex relaxation techniques (Boyd et al. 1994) were then used to derive an SDP w.r.t. a new variable $P = VV^{\tau}$:

$$\text{max} \quad \langle X^{\tau}X, P \rangle - \mu t$$

s.t. $\text{tr}(P) \leq d$ (14b)
$$I \succeq P \succeq 0$$
$$\langle P, f f^{\tau} \rangle \leq \delta^2$$

$$\begin{bmatrix} tI & PM^+ \tfrac{1}{2} \\ PM^- & I \end{bmatrix} \succeq 0$$

(14e)

(14f)

where $M_sM_s^{\tau}$ is the Cholesky decomposition of $(-1)^sQ + \varphi_{\lambda_d} \varphi_{\lambda_d}^{\tau}$, $\varphi_{\lambda_d} \geq \|Q\|_2$, and $\delta, \mu \geq 0$ are the fairness tolerance levels, for mean and covariance constraints, respectively. The final loading matrix $V$ is obtained by extracting the top $d$ eigenvectors from the resulting $P^*$. 
C Low Explained Variance of FPCA

C.1 Empirical Exploration

As seen in the Experiments, FPCA generally results in low explained variance than MBF-PCA, when same(similar) level of fairness is achieved. For clarity of exposition, we consider GERMAN CREDIT DATASET with $d = 10$ and the following hyperparameter settings: $\tau = 10^{-3}$ for MBF-PCA and $(\delta, \mu) = (0, 0.1)$ for FPCA. The reason for choosing a rather relaxed hyperparameter setting for MBF-PCA is to ensure a fair comparison i.e. to compare the two algorithms under the setting in which they achieve the same level of fairness.

To analyze this phenomenon in a detailed manner, we consider the variance explained of each eigenvector of $P^*$ in order of the eigenvalues.

C.2 Theoretical Exploration

We now give a sketch of theoretical argument of why such phenomenon of misalignment in ordering of eigenvalues of $P^*$ and explained variance of each eigenvector of $P^*$ may occur.

Let $\hat{\Sigma}$ be positive definite, and let $\hat{\Sigma} = \sum_{k=1}^{p} \lambda_k w_k w_k^T$ be its spectral decomposition with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ being the eigenvalues and $w_k$’s being the corresponding orthonormal eigenvectors.

Also,

$$\text{tr} \left( V^T \hat{\Sigma} V \right) = \sum_{l=1}^{d} v_l^T \left\{ \sum_{k=1}^{p} \lambda_k w_k w_k^T \right\} v_l \quad (15)$$

$$= \sum_{l=1}^{d} \sum_{k=1}^{p} \lambda_k < v_l, w_k >^2 \quad (16)$$

$$= \sum_{l=1}^{d} \sum_{k=1}^{p} \lambda_k \cos^2 \theta_{v_l, w_k} \quad (17)$$

where $V^T V = I_d$, $v_l$ is the $l$-th column of $V$, and $\cos \theta_{u,v} := \langle u, v \rangle$. Eq. (15) shows that the maximization of $\text{tr} \left( V^T \hat{\Sigma} V \right)$ can be thought of as a optimization problem w.r.t. orthonormal $v_1, \ldots, v_d$, given $\lambda_k$’s and $w_k$’s. For example, with no constraint, the $\{v_l\}_{l=1}^{d}$ such that $\cos \theta_{v_l, w_l} = 1$ achieve the maximization of $\text{tr} \left( V^T \hat{\Sigma} V \right)$ and the variances explained are ordered by the $\{v_l\}_{l=1}^{d}$, which is precisely Eckart-Young theorem (Eckart and Young [1936]).

When the covariance constraint $\|VQV^T\|_2 = 0$ is imposed on the Eq. (15), we will show the effect of this quadratic constraint to the optimal $V^*$. The feasible space on the constraint $\|VQV^T\|_2 = 0$ can be rewritten (Ofat and Aswani [2019]): for any $\varphi \geq \|VQV\|_2$,

$$\max \left\{ \|VQV + \varphi I_d\|_2, \|VQV - \varphi I_d\|_2 \right\} - \varphi = 0. \quad (18)$$

Note that $VQV$ is a matrix that is orthgonal in $V$, $\{Q \pm \varphi I_d\}$ has $Q$ as $VQV$, $Q$ is $\sum_{k=1}^{p} q_k s_k s_k^T$, $q_k \in \mathbb{R}^p$. Here the $q_k$ and $s_k$ are the eigenvalues and eigenvectors from the spectral decomposition of

Figure 6: Explained variance of each eigenvector of $P^*$ for GERMAN CREDIT DATASET, over the considered 10 train-test splits. Note how in FPCA’s case, the there’s significant “leakage” of explained variance in the latter part (i.e. starting from 11-th eigenvector of $P$).
Q. Note that $q_k$’s are all real numbers due to the symmetry of $Q$.

Thus Eq. (18) can be rewritten as

$$\varphi = \max_k \left\{ \max_{1 \leq l \leq d} |q_k + \varphi| \sum_{l=1}^d \cos^2 \theta_{vl,sk} \right\},$$

$$\max_k |q_k - \varphi| \sum_{l=1}^d \cos^2 \theta_{vl,sk} \right\}.$$

In certain dataset, there may exist some $q_k$’s and $\varphi$ such that $\max_k \max(\varphi/|q_k + \varphi|, \varphi/|q_k - \varphi|) \in (0, 1)$. If that is the case, then there must exist some $\delta \in (0, 1)$ such that

$$\sum_{l=1}^d \cos^2 \theta_{vl,sk} \leq 1 + \delta.$$  \hspace{1cm} (19)

With $d \geq 2$, the restriction on $\theta$ as expressed by Eq. (19) may cause severe deviation from the optimal $\theta$’s obtained without such restriction. In turn, this may cause a misalignment in the ordering of the explained variance of each loading vector. In other words, the ordering of the explained variance may be substantially different from the ordering of the eigenvalues in the presence of the covariance (quadratic) constraint.

**D Theoretical Minimum**

In this section, we provide (minimum) required background for understanding the proofs of propositions/theorems presented in the main paper. All proofs of the results here are deferred to the respective references and therein.

**D.1 A Primer on (Matrix) Manifolds, and Stiefel Manifold**

**Preliminaries** Let $\mathcal{M}$ be a smooth, finite-dimensional Riemannian manifold, endowed with a Riemannian metric $\langle \cdot, \cdot \rangle_x$ on each tangent space $T_x \mathcal{M}$ and $x \in \mathcal{M}$. Let $\|\cdot\|_x$ be the associated norm. We omit $x$ when it is clear from context. Let $\exp_x$ be the exponential map at $x$, and let dist be the Riemannian distance.

Let $f: \mathcal{M} \rightarrow \mathbb{R}$ be a smooth function. We first consider the gradient of $f$, defined on manifolds:

**Definition 10** [Absil, Mahony, and Sepulchre 2007]. The Riemannian gradient at $x \in \mathcal{M}$ of $f$, denoted as $\nabla f(x)$, is the unique element of $T_x \mathcal{M}$ such that

$$\forall \xi \in T_x \mathcal{M}, \langle \nabla f(x) \xi, \xi \rangle = Df(x)[\xi],$$

where $Df(x)[v]$ is the directional derivative of $f$ at $x$ along $v$.

Finding a closed-form of grad may be difficult in some cases. However, if $f$ can be extended to the ambient Euclidean space $\mathbb{R}^{p \times d}$, then grad can be computed as follows:

$$\nabla f(x) = \mathbf{P}_{T_x \mathcal{M}}(\nabla f(x)) \quad \hspace{1cm} \text{(20)}$$

where $\nabla$ is the usual Euclidean gradient, and $\mathbf{P}_{T_x \mathcal{M}}$ is the projection operator onto $T_x \mathcal{M}$.

**Matrix Manifolds** Matrix is a mathematical object of fundamental importance. Often in applications in machine learning and optimization, we are given the task of finding some matrix satisfying certain conditions, such as orthornormality, while optimizing some real-valued function of the matrices [Hu et al. 2020]. In the perspective of optimization, it is natural to see whether an intrinsic geometric structure can be found in a set of matrices with certain common properties. It turns out that such structures can be characterized by manifold, hence the term matrix manifold.

To start, let us consider $\mathbb{R}^{p \times d}$, a set of all possible real-valued $p \times d$ matrices. Along with the Frobenius inner product, defined as $\langle X, Y \rangle := \text{tr}(X^T Y)$, $\mathbb{R}^{p \times d}$ can be regarded as a Euclidean space, and thus a Riemannian manifold.

Many of the well-known matrix manifolds are actually embedded Riemannian submanifold of $\mathbb{R}^{p \times d}$ i.e. its subspace topology is exactly the given topology on $\mathbb{R}^{p \times d}$. We shall discuss about one specific matrix manifold that is crucial to our MbF-PCA in the next subsection.

**Stiefel Manifold** We are mainly concerned with the Stiefel manifold, defined as follows:

$$St(p, d) = \{ V \in St(p, d) | V^T V = I_d \} \quad \hspace{1cm} (21)$$

By considering the mapping $F(V) = V^T V - I_d$, we can prove the following result:

**Theorem 4** [Absil, Mahony, and Sepulchre 2007]. $St(p, d)$, defined as above, is an embedded (Riemannian) submanifold of $\mathbb{R}^{p \times d}$.

Now let us look at several properties of $St(p, d)$, many of which prove itself to be effective in both theoretical and practical perspective.

**Theorem 5.** $St(p, d)$ is $(p - d - 1)$-connected.

**Corollary 5.1.** When $p > d$, $St(p, d)$ is always at least path-connected, which implies connectedness.

**Theorem 6.** $St(p, d)$ is compact.

**Lemma 2.** Projection of any $X \in \mathbb{R}^{p \times d}$ onto $T_V St(p, d)$ is given as

$$\mathbf{P}_{T_V St(p, d)}(X) = X - V \text{sym}(V^T X) \quad \hspace{1cm} \text{(22)}$$

where $\text{sym}(X) = \frac{1}{2}(X + X^T)$.

**E Feasibility of Fair PCA**

Here, we briefly discuss the feasibility problem of fair PCA.

Let us consider Eq. (13), which is a special case of our formulation of fair PCA under Gaussian assumption. For our analysis, we consider Eq. (13) as a constrained manifold optimization over $St(p, d)$; geometric properties of manifolds turn out to be crucial in obtaining to-be-presented results.

When only the mean constraint is in place i.e. assume that $Q = O$, then the feasibility of Eq. (13) is always guaranteed:

**Proposition 3.** If $Q = O$, then Eq. (13) is always feasible.
**Proof.** We start by observing that the mean constraint, \( V \top f = 0 \), is a linear hyperplane passing through the origin \( V = 0 \), and the manifold constraint \( V \top V = 0 \), is a manifold that is at least path-connected. Now suppose that there exists some \( f \) such that the linear hyperplane does not intersect \( St(p,d) \). Let \( V \in St(p,d) \) be on the “left” side of the hyperplane. Since \( St(p,d) \) is symmetric about the origin \((V \top V = I_d) \rightarrow (-V)\top(-V) = I_d \), \(-V\) is on the “right” side of the hyperplane. Then there must exist some path on \( St(p,d) \) connected \( V \) and \(-V\), which should intersect the hyperplane, a contradiction.

When covariance constraint is also in-place, the problem becomes more complex. However, especially in high dimensions, it is expected that the feasibility assumption is not so far-off. As such discussion is not limited to Gaussian case of Eq. (15), from hereon and forth we consider the following assumption:

**Assumption 3.** Eq. (15) is feasible.

\[ \]
Lemma 6. There exists some $L_h > 0$ such that $h : St(p, d) \to \mathbb{R}$ is $L_h$-Lipschitz.

Proof. Suppose not. Then for any $L_h > 0$, there exists some $V, W \in St(p, d)$ such that $|h(V) - h(W)| > L_h |V - W|$. Especially for $W \neq V$, we have that $L_h < \frac{|h(V) - h(W)|}{|V - W|}$. Letting $W \to V$, we have that $L_h < \lim_{W \to V} \frac{|h(V) - h(W)|}{|V - W|}$. This implies that $\nabla_x h$ is unbounded on $St(p, d)$, a contradiction since the continuity of $\nabla_x h$ and compactness of $St(p, d)$ implies that $\nabla_x h$ is bounded by the Extreme Value Theorem.

As done in proof of Lemma 4 let us define $\rho = \lim_{k \to \infty} \rho_k$. We now show how $\rho_{max}, \rho$ affect $\rho$.

Lemma 7. $\rho = \rho(\rho_{max}, \rho)$ is characterized as follows: there exists some function $K := K(\tau) \in \mathbb{N}$, satisfying $K < \infty$ for $\tau > \min\{\tau(\psi)\} = \min\{\rho(\rho_{max}, \rho)\}$. Moreover, $K$ is decreasing in $\tau$ and increasing in $\rho_{max}$.

Proof. Define $K(\tau) := \min\{|k \in \mathbb{N} : h(V_k) > \tau\}$, which is the number of $\rho$-updates in the algorithm. From the algorithm, Eq. (31) is directly implied. Now it remains to characterize the function $K(\tau)$.

Let us first consider the case of $\tau > h(V)$. Since $h(V_k) \to h(V)$ and $h$ is continuous, it must be that $K(\tau) < \infty$ by the algorithm, regardless of the other hyperparameters. Now suppose that $\tau \leq h(V)$. Then there are two cases to consider:

1. If $h(V_M) \leq \tau$ for some $M < \infty$, then from the algorithm it is clear that $K(\tau) \leq M$.
2. If $h(V_k) > \tau$ for all $k$, then again, from the algorithm it is clear that $K(\tau) = \infty$.

By construction, $K(\tau)$ is decreasing in $\tau$, and thus, $\rho$ is decreasing in $\tau$ and increasing in $\rho_{max}$.

Corollary 6.1. If no $k$ satisfies $h(V_k) \leq \tau$, then $\rho = \infty$ if and only if $\rho_{max} = \infty$, $\tau = 0$.

Denote $C := St(p, d) \cap h^{-1}(0)$ as the set of feasible points. By compactness of $St(p, d)$ and continuity of $h$, we have that $C$ is also compact. By Assumption 5, $C$ is non-empty. Lastly, recall from Eq. (24) $\tau(\psi) = \min\{\rho(\rho_{max}, \rho)\}$. The last lemma asserts that after sufficiently many iterations have passed, the sequences $\{V_k\}$ and $\{W_k\}$ are always close to $C$.

Lemma 8. If $\nabla V$ is fair and $\epsilon_{min}$ is sufficiently small in the sense that $g(\epsilon_{min}) < r$, then there exists some $K > 0$ such that for all $k > K$, $B(\epsilon_{min}) \cap C \neq \emptyset$ and $B(\epsilon_{min}) \cap C \neq \emptyset$.

Proof. First, suppose that for all $k$, $B(\epsilon_{min}) \cap C \neq \emptyset$. Taking the limit of $k \to \infty$ on both sides, we have that $V \in B(\epsilon_{min}) \cap C \neq \emptyset$ by the given assumption, a contradiction. Thus there exists some $K_1$ such that for all $k > K_1$, $B(\epsilon_{min}) \cap C \neq \emptyset$.

Now suppose that for all $k$, $B(\epsilon_{min}) \cap C \neq \emptyset$. Recall from Lemma 3 that

$$\|W - V\| \leq g(\epsilon_{min}) < r$$

Again taking the limit of $k \to \infty$, we have that $V \in C \cap B(\epsilon_{min}) = \emptyset$, a contradiction. Thus there exists some $K_2$ such that for all $k > K_2$, $B(\epsilon_{min}) \cap C \neq \emptyset$.

Set $K = \max(K_1, K_2)$, and we are done.

F.3 Proof of Theorem 2

(The proof is inspired from the optimality guarantee for the Euclidean quadratic penalty method; see Theorem 17.1 of [Nocedal and Wright 2006]. The proof is divided into 3 parts.

Existence of limit point of $\{W_k\}$

By Lemma 4, there exists some limit point $W$ of $\{W_k\}$ such that $\|V - W\| \leq g(\epsilon_{min})$. By considering appropriate subsequence, let $W_k \to W$.

$\nabla$ is approximate local minimizer of $\psi$

By Assumption 4, we have that for all $W \in B(\epsilon_{min}) \cap St(p, d)$,

$$Q(W_k, \rho_k) \leq Q(W, \rho_k)$$

i.e.

$$f(W_k) + \rho_k h(W_k) \leq f(W) + \rho_k h(W).$$

Rearranging gives

$$h(W_k) \leq h(W) + \frac{1}{\rho_k} (f(W) - f(W_k))$$

Taking the limit $k \to \infty$ and denoting $1/\rho_k = 0$ gives

$$h(W) \leq h(W) + \frac{2\|\nabla\|}{\rho} \|W - W_k\|,$$

for all $W \in B(\epsilon_{min}) \cap St(p, d)$.

Specifically, since $B(\epsilon_{min}) \cap St(p, d) \subset B(\epsilon_{min}) \cap St(p, d)$ by Lemma 3, we have that

$$h(W) \leq h(W) + |h(W) - h(W)|$$

$$\leq h(W) + \frac{2\|\nabla\|}{\rho} \|W - W\| + L_h \|W - W\|$$

$$\leq h(W) + \frac{2\|\nabla\|}{\rho} \|W - W\| + \left(L_h \frac{2\|\nabla\|}{\rho} \right) g(\epsilon_{min})$$

for all $W \in B(\epsilon_{min}) \cap St(p, d)$.

Define $\beta(\rho_{max}, \tau) := \frac{2\|\nabla\|}{\rho(\rho_{max}, \tau)}$. By Lemma 7, $\beta$ indeed satisfies

$$\lim_{\rho_{max} \to \infty} \beta(\rho_{max}, 0) \geq 0$$

and

$$\lim_{\rho_{max} \to \infty} \beta(\rho_{max}, \tau) > 0$$

for $\tau > 0$. Also, by the same lemma, $\beta$ is indeed decreasing in $\tau$ and increasing in $\rho_{max}$.

(Note how this particular part of the proof does not rely on the assumption that $\nabla$ is feasible!)

If $\nabla$ is fair, then $\nabla$ is an approximate local minimizer of Eq. (8)

By Lemma 8, we may only consider $k$’s such that $B(\epsilon_{min}) \cap C \neq \emptyset$ and $B(\epsilon_{min}) \cap C \neq \emptyset$. By Assumption 4, we have that for all $W \in B(\epsilon_{min}) \cap C$,

$$Q(W_k, \rho_k) \leq Q(W, \rho_k)$$

i.e.

$$f(W_k) \leq f(W_k) + \rho_k h(W_k) \leq f(W) + \rho_k h(W) = f(W),$$

where the last equality follows from $W \in C$. 


Thus we have that
\[
 f(\mathbf{v}) \leq f(\mathbf{w}) + |f(\mathbf{v}) - f(\mathbf{w})| \\
 \leq \lim_{k \to \infty} f(W_k) + 2\|\mathbf{v} - \mathbf{w}\| \\
 \leq f(\mathbf{w}) + 2\|\mathbf{v}\|g(\epsilon_{\min}),
\]
for all \(W \in B_r(W_k) \cap C\). Specifically, since \(B_r(\mathbf{v}) \cap \text{St}(p, d) \subset B_r(\mathbf{w}) \cap \text{St}(p, d)\) by Lemma 4, we have that
\[
 f(\mathbf{v}) \leq f(\mathbf{w}) + 2\|\mathbf{v}\|g(\epsilon_{\min}),
\]
for all \(V \in B_r(\mathbf{v}) \cap C\).

G Closed form of \(\nabla_V h(V)\)

In this section, we derive the closed form of the constraint gradient.

First, we recall the lemma that provides a closed form of \(h(V) = \text{MMD}^2(\cdot, \cdot)\):

**Lemma 1** [Gretton et al., 2007], \(\hat{\Delta} \text{ is computed as follows:}\)

\[
\hat{\Delta} = \left[ \frac{1}{m^2} \sum_{i,j=1}^{m} k(X_i, X_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(Y_i, Y_j) \right. \\
- \left. \frac{2}{mn} \sum_{i,j=1}^{m,n} k(X_i, Y_j) \right]^{1/2}.
\]

After squaring both sides, let us refer to the three terms as \(h_1, h_2, \text{ and } h_3\) i.e. \(h(V) = \Delta^2 = h_1(V) + h_2(V) - 2h_3(V)\) where
\[
h_1(V) = \frac{1}{m^2} \sum_{i,j=1}^{m} k(X_i, X_j), \\
h_2(V) = \frac{1}{n^2} \sum_{i,j=1}^{n} k(Y_i, Y_j), \text{ and } \\
h_3(V) = \frac{1}{mn} \sum_{i,j=1}^{m,n} k(X_i, Y_j).
\]

G.1 Closed form of \(\nabla_V h_1(V)\)

Denote
\[
g_{ij}(V) = \operatorname{tr}((X_i - X_j)^\top VV^\top(X_i - X_j)), \\
K_{ij}(V) = \exp(-g_{ij}(V)/(2\sigma^2)), \text{ and } \\
K_{XY}(V) = \{K_{ij}(V)\}_{ij} \in \mathbb{R}^{m \times n},
\]
and \(X\) be the data matrix whose \(i\)-th row is \(X_i^\top\). By chain rule, we have that
\[
\nabla_V h_1(V) = -\frac{1}{2m^2\sigma^2} \sum_{i,j=1}^{m} K_{ij}(V) \nabla_V g_{ij}(V) \\
= -\frac{1}{m^2\sigma^2} \sum_{i,j=1}^{m} K_{ij}(V)(X_i - X_j)(X_i - X_j)^\top V
\]

Now denote \(H_i(V) = \sum_{j=1}^{m} K_{ij}(V)\) and \(H_{XY}(V) = \text{diag}(H_1(V), H_2(V), \ldots, H_m(V))\), where ‘\text{diag}’ is the block-diagonal operator. We have that
\[
\sum_{i,j=1}^{m} K_{ij}(V)X_iX_j^\top = \sum_{i=1}^{m} \left( \sum_{j=1}^{m} K_{ij}(V) \right) X_iX_i^\top \\
= X^\top H_X(V)X,
\]
and
\[
\sum_{i,j=1}^{m} K_{ij}(V)X_iX_i^\top = X^\top K_X(V)X.
\]

Since \(K_X(V)\) is symmetric i.e. \(K_{ij}(V) = K_{ji}(V), \text{ and } \\
\sum_{i,j=1}^{m} K_{ij}(V)X_iX_i^\top = \sum_{i=1}^{m} K_{ji}(V)X_jX_i^\top = X^\top H_X(V)X, \text{ and } \\
\sum_{i,j=1}^{m} K_{ij}(V)X_iX_j^\top = \sum_{i=1}^{m} K_{ji}(V)X_jX_i^\top = X^\top K_X(V)X.
\]

Thus we have that
\[
\nabla_V h_1(V) = -\frac{2}{m^2\sigma^2} X^\top (H_X(V) - K_X(V)) XV.
\]

G.2 Closed form of \(\nabla_V h_2(V)\)

The computation is almost exactly the same as above, and thus we only show the final result:
\[
\nabla_V h_2(V) = -2\frac{1}{n^2\sigma^2} Y^\top (H_Y(V) - K_Y(V)) YV,
\]
where \(Y, H_Y(V), \text{ and } K_Y(V)\) are defined analogously.

G.3 Closed form of \(\nabla_V h_3(V)\)

Let us overload the notation a bit and now denote
\[
g_{ij}(V) = \operatorname{tr}((X_i - Y_j)^\top VV^\top(X_i - Y_j)), \\
K_{ij}(V) = \exp(-g_{ij}(V)/(2\sigma^2)), \text{ and } \\
K_{XY}(V) = \{K_{ij}(V)\}_{ij} \in \mathbb{R}^{m \times n},
\]
Again by chain rule, we have that
\[
\nabla_V h_3(V) = -\frac{1}{2mn\sigma^2} \sum_{i,j=1}^{n} K_{ij}(V) \nabla_V g_{ij}(V) \\
= -\frac{1}{mn\sigma^2} \sum_{i,j=1}^{m} K_{ij}(V)(X_i - Y_j)(X_i - Y_j)^\top V.
\]

Now, again with a slight notation overload, denote
\[
H_i(V) = \sum_{j=1}^{m} K_{ij}(V), \\
H_{XY}(V) = \text{diag}(H_1(V), H_2(V), \ldots, H_m(V)), \\
\hat{H}_i(V) = \sum_{i=1}^{m} K_{ij}(V), \\
\hat{H}_{XY}(V) = \text{diag}(\hat{H}_1(V), \hat{H}_2(V), \ldots, \hat{H}_n(V)).
\]
We have that
\[\sum_{i,j=1}^{m,n} K_{ij}(V)X_iX_i^T = \sum_{i=1}^{m} \left( \sum_{j=1}^{n} K_{ij}(V) \right) X_iX_i^T = X^TH_XY(V)X,\]
and
\[\sum_{i,j=1}^{m,n} K_{ij}(V)Y_jY_j^T = \sum_{j=1}^{n} \left( \sum_{i=1}^{m} K_{ij}(V) \right) Y_jY_j^T = Y^TH_XY(V)Y.\]

Next, we have that
\[\sum_{i,j=1}^{m,n} K_{ij}(V)X_iY_j^T = X^TK_{XY}(V)Y,\]
and
\[\sum_{i,j=1}^{m,n} K_{ij}(V)Y_jX_i^T = Y^TK_{XY}(V)^TX.\]

Thus we have that
\[\nabla_{V} h_3(V) = -\frac{1}{mnd^2} \left( X^TH_XY(V)X + Y^TH_XY(V)Y - X^TK_{XY}(V)Y - Y^TK_{XY}(V)^TX \right)V.\]

\section{Description of 2nd Part of Experiment 1}

With \(p_0 = 1000\) and \(n = 250\), we first constructed two different \(p_0\)-variate Gaussian distributions \(N_{p_0}(\mu^{(s)}_{p_0},\Sigma^{(s)}_{p_0})\) and sampled \(n\) points for each \(s \in \{0,1\}\). Then for each considered \(p\), we projected the samples by multiplying the constructed data matrices with a Gaussian random matrix of size \(p_0 \times p\). This creates \(18\) pairs of \(p\)-dimensional distributions. This allowed for our model the situation in which two protected groups have different non-Gaussian distributions.

We now describe how we’ve set \(\mu^{(s)}_{p_0}\) and \(\Sigma^{(s)}_{p_0}\) for each \(s\). The mean difference is set as \(f_{p_0} = \mu^{(1)}_{p_0} - \mu^{(0)}_{p_0}\) where
\[2 + \frac{f_{RAW}^{RAW}}{\|f_{RAW}^{RAW}\|}\] and \(f_{RAW}^{RAW} = 1.0\). Then \(\mu^{(0)}_{p_0}\) is set as \(0\) and \(\mu^{(1)}_{p_0}\) is set as \(\mu^{(0)}_{p_0} + f_{p_0}\).

The covariance difference is set as \(Q^{RAW} = A^{(1)}_p - A^{(0)}_p\),
\[
\Sigma^{(0)}_p = \text{diag} \left( AR_{p/5}(0.99), AR_{p/5}(0.98), AR_{p/5}(0.97), AR_{p/5}(0.98), AR_{p/5}(0.95) \right),
\]
and
\[
\Sigma^{(1)}_p = \text{diag} \left( AR_{p/5}(0.99), AR_{p/5}(0.98), AR_{p/5}(0.97), AR_{p/5}(0.98), AR_{p/5}(0.99) \right).
\]
Here, ‘\(\text{diag}\)’ is the block-diagonal operator, and
\[
(AR_{p/5}(r))_{ij} = \frac{r^{|i-j|}}{1 - r^2}, \quad i,j \in \{1,\ldots,p/5\}\] is the covariance matrix of a \(p/5\)-variate Gaussian AR(1) process \cite{shumway2017time}. The \(r \in (0,1)\) is a parameter that controls the strength of correlation. Finally, \(\Sigma^{(0)}_p\) is set as \(A^{(s)}_p / \|Q_p\|\).

Due to the scaling, \(\|f_p\| = 2\) and \(\|Q_p\| = 1\) for all \(p\). This is to ensure that \(\|f_p\|\) and \(\|Q_p\|\) do not explode as \(p\) increases, allowing for us to see the effect of high dimensions more clearly under the “same” fairness setting.

\section{Description of Experiment 2}

Here, we describe the full pre-processing steps for the three UCI datasets considered in this paper. All pre-processing steps were done using AI Fairness 360 \cite{bellamy2018ai}, built-in functionalities. In our Github repository\textsuperscript{[1]}, we include the links for downloading the raw datasets, along with the pre-processing codes.

\subsection{COMPAS dataset}

The raw data was randomly reduced to 40\%, resulting in 2468 samples. This was to avoid computational burden in computing the \(MMD^2\) metric, as it scales quadratically in number of samples. ‘Race’ was set as the protected attribute, and the task is to classify the bank account holders into credit class ‘Good or Bad’. The features ‘sex’ and ‘e_charge_desc’ were dropped, resulting in total of 11 features.

\subsection{German credit dataset}

The total number of samples was 1000. ‘Age’ was set as the protected attribute as done in \cite{kamiran2011data}, and the task is to classify whether the individual (who served jail time) will commit the crime again (i.e. recidivism). The features ‘sex’ and ‘personal_status’ were dropped, resulting in total of 57 features.

\subsection{Adult income dataset}

The raw data was randomly reduced to 5\%, resulting in 2261 samples; this was due to the same reason as the COMPAS dataset. ‘Sex’ was set as the protected attribute as done in \cite{kamisima2011fairness}, and the task is to indicate whether the individual’s income is larger than 50K dollars. The features ‘fnlwgt’ and ‘race’ were dropped, resulting in total of 99 features.

\section{Full results for EDA of considered UCI datasets}

We show two things:
\begin{itemize}
  \item Boxplots of communality of the protected attribute of each considered UCI dataset, for both \(d = 2\) and \(d = 10\).
  \item PC biplot of each considered UCI dataset, for \(d = 2\).
\end{itemize}

Particularly for biplots, one train-test split was chosen, and based on that, only the top 10 features with the highest communalities (+ the sensitive attribute, if it is not included) were chosen for plots.

\textsuperscript{[1]}\url{https://github.com/nick-jhlee/fair-manifold-pca}
Figure 7: Boxplots of communalities of race for COMPAS dataset

(a) $d = 2$

(b) $d = 10$

Figure 8: PC biplot of PCA for COMPAS dataset
Figure 9: PC biplot of FPCA for COMPAS dataset

Figure 10: PC biplot of MbF-PCA for COMPAS dataset
Figure 11: Boxplots of communalities of age for German credit dataset

(a) $d = 2$  
(b) $d = 10$

Figure 12: PC biplot of PCA for German credit dataset
Figure 13: PC biplot of FPCA for German credit dataset

Figure 14: PC biplot of MBF-PCA for German credit dataset
Figure 15: Boxplots of communalities of gender(sex) for Adult income dataset

Figure 16: PC biplot of PCA for Adult income dataset
Figure 17: PC biplot of FPCA for Adult income dataset

Figure 18: PC biplot of MBF-PCA for Adult income dataset