# Fast and Efficient Fair PCA via Optimization over Stiefel Manifold

Junghyun Lee[1], Gwangsu Kim[2], Matt Olfat[3,4], Mark Hasegawa-Johnson[5], Chang D. Yoo[2]

[1]Graduate School of AI, KAIST

[2]School of Electrical Engineering, KAIST

[3]UC Berkeley

[4]Citadel

[5]Department of Electrical and Computer Engineering, UIUC

September 29, 2021

1. Introduction
2. Review of FPCA
3. Problems with FPCA
    - Problems with the definition
    - Problems with the SDP
4. New Definition: $\Delta$-fairness
5. New Framework: MbF-PCA
6. Experiments
7. Conclusion/Future works
8. References

# Outline

1. Introduction
2. Review of FPCA
3. Problems with FPCA
   - Problems with the definition
   - Problems with the SDP
4. New Definition: $\Delta$-fairness
5. New Framework: MbF-PCA
6. Experiments
7. Conclusion/Future works
8. References

# Fair Machine Learning

- An active area of research with enormous societal impact
- Machine learning algorithms should not be dependent on specific (sensitive) variables such as gender, age, race...etc.
- There are multiple frameworks on how to do this:
  - **Fair supervised learning**
  - Fair unsupervised learning
  - **Fair representation learning**
  - Fair data preprocessing
  - ...etc.

# Fair Supervised Learning

- We briefly review three of the most widely-used definitions of fairness in supervised learning, as formulated in [MCPZ18].
- $(Z, Y, A) \in \mathbb{R}^d \times \{0,1\} \times \{0,1\}$: joint distribution of the dimensionality-reduced data, (downstream task) label, and protected attribute.
- $g : \mathbb{R}^d \to \{0,1\}$: classifier that outputs prediction $\hat{Y}$ for $Y$ from $Z$.
- $D_s$: probability measure of $Z_s \triangleq Z|A = s$ for $s \in \{0,1\}$
- $D_{s,y}$: probability measure of $Z_s \triangleq Z|A = s, Y = y$ for $s, y \in \{0,1\}$.

# Fair Supervised Learning

### Definition ([FFM⁺15])

$g$ is said to satisfy **demographic parity (DP) up to** $\Delta_{DP}$ w.r.t. $A$ with $\Delta_{DP} \triangleq |\mathbb{E}_{x\sim D_0}[g(x)] - \mathbb{E}_{x\sim D_1}[g(x)]|$.

### Definition ([HPS16])

$g$ is said to satisfy **equalized opportunity (EOP) up to** $\Delta_{EOP}$ w.r.t. $A$ and $Y$ with $\Delta_{EOP} \triangleq \left|\mathbb{E}_{x\sim D_{0,1}}[g(x)] - \mathbb{E}_{x\sim D_{1,1}}[g(x)]\right|$.

### Definition ([HPS16])

$g$ is said to satisfy **equalized odds (EOD) up to** $\Delta_{EOD}$ w.r.t. $A$ and $Y$ with $\Delta_{EOD} \triangleq \max_{y\in\{0,1\}} \left|\mathbb{E}_{x\sim D_{0,y}}[g(x)] - \mathbb{E}_{x\sim D_{1,y}}[g(x)]\right|$.

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

# Fair Representation Learning

- “Representation learning is a promising approach for implementing algorithmic fairness” [CK19]
- In this framework, a modular separation between roles can be made:
    - data regulator
    - data producer
    - data user
- This has several positive implications:
    - centralize fairness constraints, by moving the fairness responsibility from the data user to the data regulator
    - simplify and centralize the task of fairness auditing
    - can be constructed to satisfy multiple fairness measures simultaneously
    - simplify the task of evaluating the fairness/performance tradeoff, e.g., using performance bounds

# Outline

1. Introduction

2. Review of FPCA

3. Problems with FPCA
   - Problems with the definition
   - Problems with the SDP

4. New Definition: $\Delta$-fairness

5. New Framework: MbF-PCA

6. Experiments

7. Conclusion/Future works

8. References

# Two Notions of Fair PCA

- There are two branches of works on ensuring fairness in PCA, each of which considers a completely *different* definition of fairness:
  - The reconstruction errors of two (or more) sensitive groups should be similar i.e. one group's reconstruction error should not be significantly higher than others. [STM+18, TSS+19]
  - The distribution of two sensitive groups should be similar i.e. an adversary should not be able to distinguish between the two groups, after the PCA [OA19, LKO+21]

# Two Notions of Fair PCA

- It turns out that these two definitions conflict with one another:

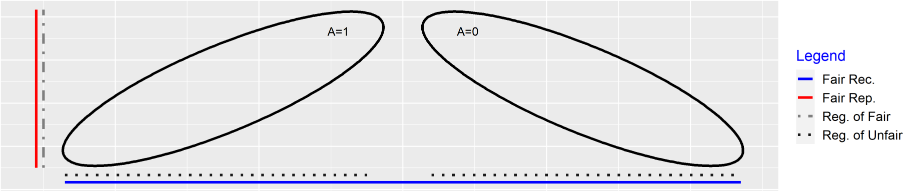


Figure: A toy example demonstrating how the two different definitions of fair PCA conflict with one another [LKO+21]

# Adversarial Definition: FPCA

- Here, we consider the latter notion regarding the distribution similarity after PCA.
- To motivate our work, we first review the work by Olfat & Aswani [OA19] that first considered this problem of fair PCA:

**Definition ($\Delta_A$-fairness, [OA19])**

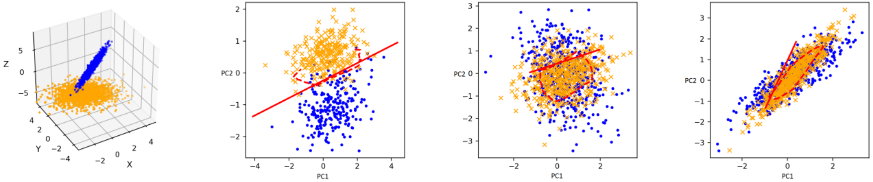
Consider a fixed classifier $h(u, t) : \mathbb{R}^d \times \mathbb{R} \to \{0, 1\}$ that inputs features $u \in \mathbb{R}^d$ and a threshold $t$, and predicts the protected class $z \in \{0, 1\}$. Then, a dimensionality reduction $\Pi : \mathbb{R}^p \to \mathbb{R}^d$ is $\Delta_A(h)$-fair if

$$\left| \mathbb{P}\Big[ h(\Pi(x), t) = 1 | z = 1 \Big] - \mathbb{P}\Big[ h(\Pi(x), t) = 1 | z = 0 \Big] \right| \leq \Delta_A(h), \ \ \forall t \in \mathbb{R}. \tag{1}$$

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

# Adversarial Definition: FPCA

- The classifiers in the definition are **adversarial**; they try to classify the protected class from the dimensionality-reduced data.
- Them performing poorly is an indicator of being fair.



(a) Original data (b) PCA (c) FPCA - Mean con. (d) FPCA - Both con.

Figure 1: Comparison of PCA and FPCA on synthetic data. In each plot, the thick red line is the optimal linear SVM separating by color, and the dotted line is the optimal Gaussian kernel SVM.

Figure: When vanilla PCA is applied to (a), an unfair dimensionality-reduced representation (b) is obtained. By constraining the PCA with appropriate fairness constraints, we obtain a fair representation (d). (From [OA19])

# Estimator for $\Delta(\mathcal{F})$

- With a slight abuse of notation, let

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

and

$$\Delta_A(\mathcal{F}) := \sup_{h\in\mathcal{F}} \Delta(h)$$

- For the actual computation, they proposed the following estimator:

$$\widehat{\Delta}(h) = \sup_t \left| \frac{1}{|P|} \sum_{i\in P} I_i(\Pi, h_t) - \frac{1}{|N|} \sum_{i\in N} I_i(\Pi, h_t) \right|, \quad \widehat{\Delta}(\mathcal{F}) = \sup_{h\in\mathcal{F}} \widehat{\Delta}(h)$$

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

# Fairness Constraints

- For PCA, $\Pi(x) = V^\intercal x$ for some $V \in \mathbb{R}^{p \times d}$ such that $V^\intercal V = \mathbb{I}$.
- Under **Gaussian assumption**, [OA19] derived the following fairness constraints:
  - *Mean constraint*:

$$h_{mean}(V) := \| V^\intercal (\mu_1 - \mu_0) \| = 0$$

  - *Covariance constraint*:

$$h_{cov}(V) := \| V^\intercal (\Sigma_1 - \Sigma_0) V \|_2 = 0$$

- The derivation is, however, **complicated (not straightforward from Definition 4)** in the sense that several inequalities (e.g. Pinsker's inequality) had to be utilized.

# Fairness Constraints

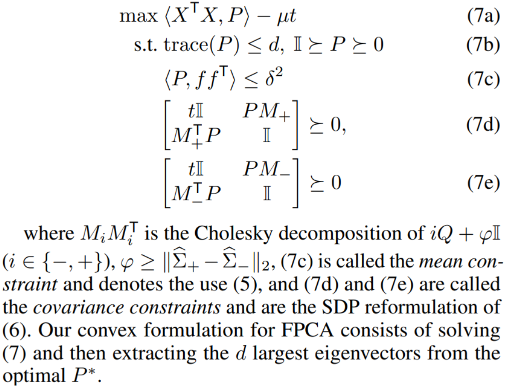
- For simplicity, let us denote $f := \mu_1 - \mu_0$ and $Q := \Sigma_1 - \Sigma_0$.
- Then the fair PCA can be written as a *constrained optimization* :

$$\begin{aligned} \underset{V}{\text{maximize}} \quad & \left\langle \frac{1}{n} X^\intercal X, VV^\intercal \right\rangle \\ \text{subject to} \quad & V^\intercal V = \mathbb{I}_d, \\ & h_{mean}(V) = 0, \\ & h_{cov}(V) = 0. \end{aligned} \tag{2}$$

  - $\frac{1}{n} X^\intercal X$: (total) covariance matrix of the original data matrix $X \in \mathbb{R}^{n \times p}$
  - $\left\langle \frac{1}{n} X^\intercal X, VV^\intercal \right\rangle$: *explained variance* of $X$ after applying (linear) PCA using $V$.

# SDP Formulation of Fair PCA

- [OA19] provided a SDP formulation of above optimization[1]:



$$
\begin{aligned}
&\max \ \langle X^\intercal X, P\rangle - \mu t && \text{(7a)}\\
&\text{s.t. } \operatorname{trace}(P) \le d,\ \mathbb{I} \succeq P \succeq 0 && \text{(7b)}\\
&\langle P, ff^\intercal\rangle \le \delta^2 && \text{(7c)}\\
&\begin{bmatrix} t\mathbb{I} & PM_+ \\ M_+^\intercal P & \mathbb{I} \end{bmatrix} \succeq 0, && \text{(7d)}\\
&\begin{bmatrix} t\mathbb{I} & PM_- \\ M_-^\intercal P & \mathbb{I} \end{bmatrix} \succeq 0 && \text{(7e)}
\end{aligned}
$$

where $M_i M_i^\intercal$ is the Cholesky decomposition of $iQ + \varphi\mathbb{I}$ $(i \in \{-, +\})$, $\varphi \ge \|\widehat{\Sigma}_+ - \widehat{\Sigma}_-\|_2$, (7c) is called the *mean constraint* and denotes the use (5), and (7d) and (7e) are called the *covariance constraints* and are the SDP reformulation of (6). Our convex formulation for FPCA consists of solving (7) and then extracting the $d$ largest eigenvectors from the optimal $P^*$.

Figure: $\delta$: bound for mean difference, $\mu$: bound for covariance difference

[1] This was heavily inspired from the SDP formulation of vanilla PCA [ACS13].

# Outline

1. Introduction
2. Review of FPCA
3. Problems with FPCA
    - Problems with the definition
    - Problems with the SDP
4. New Definition: Δ-fairness
5. New Framework: MbF-PCA
6. Experiments
7. Conclusion/Future works
8. References

# Problems with FPCA

- There are two parts to this section.
- First subsection is devoted to show the limitation of the $\Delta_A$-fairness definition itself.
- Second subsection is devoted to show the limitation of the numerical algorithm, namely the SDP.

# Outline

1. Introduction
2. Review of FPCA
3. Problems with FPCA
    - Problems with the definition
    - Problems with the SDP
4. New Definition: Δ-fairness
5. New Framework: MbF-PCA
6. Experiments
7. Conclusion/Future works
8. References

# Computational Inefficiency

- Recall the definition of $\widehat{\Delta}_A$:

$$\widehat{\Delta}_A(\mathcal{F}_c) = \sup_{h \in \mathcal{F}_c} \sup_{t} \left| \frac{1}{|P|} \sum_{i \in P} l_i(\Pi, h_t) - \frac{1}{|N|} \sum_{i \in N} l_i(\Pi, h_t) \right|$$

- Computing above requires considering all possible classifiers in the designated family $\mathcal{F}_c$, and all possible thresholds $t \in \mathbb{R}$.
- 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.

# Asymptotic Inconsistency

- $\widehat{\Delta}_A$ is known to satisfy the following bound:

**Proposition ([OA19])**

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

$$\left|\Delta_A(\mathcal{F}_c) - \widehat{\Delta}_A(\mathcal{F}_c)\right| \leq 8\sqrt{\frac{VC(\mathcal{F}_c)}{m+n}} + \delta \tag{3}$$

*where $VC(\cdot)$ is the VC dimension.*

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

# Outline

1. Introduction
2. Review of FPCA
3. Problems with FPCA
   - Problems with the definition
   - Problems with the SDP
4. New Definition: Δ-fairness
5. New Framework: MbF-PCA
6. Experiments
7. Conclusion/Future works
8. References

# SDP Relaxation of Fair PCA

- The orthogonality constraint $V^\intercal V = \mathbb{I}_d$ has to be relaxed to trace bound and matrix inequalities.
- Without the fairness constraints, it can be proven that the SDP formulation exactly outputs the optimal solution to the optimization problem.
- But with the additional constraints, theoretical guarantee (for optimality) is **not** available!

# Inscalability to High Dimensions

- Recall that the SDP is solved w.r.t. a new variable, $P \in \mathbb{R}^{p \times p}$.
- Size of the variable scales *quadratically* to the original data's dimension $p$, *independent of the dimension* $d$ *to which we are reducing to*.
- Empirically we've shown that the time complexity of FPCA grows very large in $p$, even for datas with simple block-structured covariance matrices!
- In comparison, our to-be introduced manifold optimization based approach scales very well in high dimensions.

# Inscalability to High Dimensions

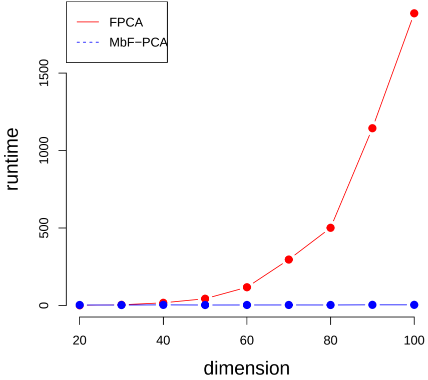
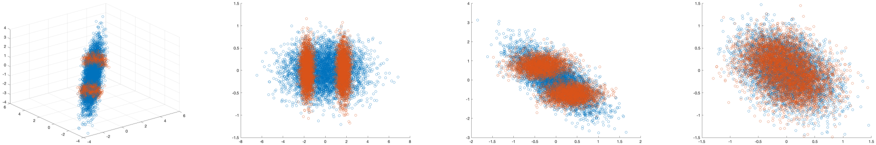


Figure: FPCA represents the SDP algorithm for fair PCA, and MbF-PCA represents our to-be introduced manifold-based algorithm for fair PCA. [LKO+21]

# “Counterexample” to FPCA

- Recall how the **Gaussian assumption** was required for deriving the actual numerical algorithm.
- FPCA, thus, cannot cover the case when two sensitive distributions, that are different, have the same first two moments (mean, covariance):



(a) Original data (b) PCA (c) FPCA [OA19] (d) MbF-PCA (ours)

Figure: Synthetic data #1: Comparison of PCA, FPCA, and MbF-PCA on data composed of two groups with same mean and covariance, but different distributions. Blue and orange represent different protected groups.

# Outline

1 Introduction

2 Review of FPCA

3 Problems with FPCA
- Problems with the definition
- Problems with the SDP

4 New Definition: $\Delta$-fairness

5 New Framework: MbF-PCA

6 Experiments

7 Conclusion/Future works

8 References

# Quick Review of MMD

- One way to define distance between two probability measures is as follows:

## Definition ([GBR$^+$07])

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:

$$MMD_k(\mu, \nu) := \sup_{f \in \mathcal{H}_k} \left| \int_{\mathbb{R}^d} f \; d(\mu - \nu) \right|$$

- $\mathcal{H}_k$ is the unit ball in the RKHS generated by some given kernel $k$.
- $\mathcal{P}_d$ is the set of all possible probability measures defined on $\mathbb{R}^d$.

# Quick Review of MMD

- As our fairness constraint involves exactly matching the considered distributions using *MMD*, we require the property of $MMD_k(\mu, \nu) = 0$ implying $\mu = \nu$:

**Definition ([FGSS08])**

If $MMD_k$ metrizes $\mathcal{P}_d$, then the kernel $k$ is said to be **characteristic** to $\mathcal{P}_d$.

- Sriperumbudur et al [SGF$^+$08] defined and characterized *stationary* characteristic kernels and identified that well-known kernels such as RBF and Laplace are characteristic.
- From hereon, we set $k$ to be the RBF kernel $k_{rbf}(x, y) := \exp\left(-\|x - y\|^2/2\sigma^2\right)$.

# Benefits of MMD

- It can act as a distance between distributions with different, or even disjoint, supports. (unlike, for instance, KL-divergence)
    - This is especially crucial as the empirical distributions are often discrete and completely disjoint.
- Since many problems in fairness involve comparing two distributions, *MMD* has already been used in much of the fairness literature as a metric [MCPZ18, AVGW19] and as an explicit constraint/penalty [QS17, LSL$^+$16, PQC$^+$19, ODL$^+$20, JLPM21], among other usages.

# Δ-fairness

- Motivated from previous discussions, we propose a new definition for fair PCA based on MMD:

### Definition (Δ-fairness, [LKO+21])

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

- Indeed, as every fair representation should satisfy, our definition also ensures that any *unconstrained* downstream tasks will also be fair:

### Proposition ([ODL+20], informal)

*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* $g$.

# Computational Efficiency

- We consider the following estimator:

$$\widehat{\Delta} := MMD(\hat{Q}_0, \hat{Q}_1) \tag{4}$$

  where $\hat{Q}_s$ is the usual empirical distribution, defined as the mixture of Dirac measures on the samples.
- Unlike $\widehat{\Delta}_A$, $\widehat{\Delta}$ can be computed exactly and efficiently:

### Lemma ([GBR$^+$07])

*$\widehat{\Delta}$ is computed as follows:*

$$\widehat{\Delta} = \left[ \frac{1}{m^2} \sum_{i,j=1}^{m} k(X_i, X_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(Y_i, Y_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(X_i, Y_j) \right]^{1/2}. \tag{5}$$

# Asymptotic Consistency

- Unlike $\widehat{\Delta}_A$, $\widehat{\Delta}$ is asymptotic convergent, with the rate depending only on $m$ and $n$:

**Theorem ([GBR$^+$07])**

*For any $\delta > 0$, with probability at least $1 - 2\exp\left(-\frac{\delta^2 mn}{2(m+n)}\right)$ the following holds:*

$$\left|\Delta - \widehat{\Delta}\right| \leq 2\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) + \delta \tag{6}$$

# Relation to $\Delta_A$-fairness (*)

- It can be argued that, for some choice of $\mathcal{F}_c$, the adversarial definition [OA19] and our definition are equivalent: in effect, that these are dual notions.
- Recognizing this, we proceed with ours, as it has two main advantages in the context of our work:
  - It ties more directly and intuitively into our optimization formulation, as we'll see later.
  - It can be represented non-variationally which allows for tighter statistical guarantees.

# Outline

1. Introduction
2. Review of FPCA
3. Problems with FPCA
    - Problems with the definition
    - Problems with the SDP
4. New Definition: $\Delta$-fairness
5. New Framework: MbF-PCA
6. Experiments
7. Conclusion/Future works
8. References

# Fair PCA as Manifold Optimization

- All the problems mentioned previously can be resolved by optimizing **directly** for $V$!
- But then, this becomes a non-convex optimization problem over $\mathbb{R}^{p\times d}$ with 2 constraints!
  - Orthogonality constraint:
  $$V^\intercal V = \mathbb{I}_d$$
  - Fairness constraint:
  $$h(V) := MMD^2(\hat{Q}_0, \hat{Q}_1) = 0$$

# Fair PCA as Manifold Optimization

- The set of all $V$'s with $V^\intercal V = \mathbb{I}_d$ has the intrinsic geometric structure of a *(matrix) manifold*:

**Definition**

For $p \geq d$, 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. $V^\intercal V = \mathbb{I}_d$ for all $V \in St(p, d)$.

- Main idea: instead of regarding $V^\intercal V = \mathbb{I}$ as a constraint, let's optimize **on** $St(p, d)$!!
- This is the basic idea behind the framework of manifold optimization.

# Quick Intuition behind Manifold Optimization

- Consider $\mathcal{M}$, an embedded Riemannian sub-manifold of $\mathbb{R}^{p \times d}$.
- Suppose we want to minimize some function $f : \mathbb{R}^{p \times d} \to \mathbb{R}$ *over* $\mathcal{M}$.
- If $\mathcal{M}$ is simply viewed as a subset of $\mathbb{R}^{p \times d}$, then this is a constrained optimization problem:

$$\begin{aligned} \underset{V}{\text{minimize}} \quad & f(V) \\ \text{subject to} \quad & V \in \mathcal{M}. \end{aligned} \tag{7}$$

- In this case, the optimization algorithm will make use of the canonical gradients and Hessians of $\mathbb{R}^{p \times d}$.

# Quick Intuition behind Manifold Optimization

- If $\mathcal{M}$ is "all there is", then this problem is an unconstrained optimization problem over $\mathcal{M}$.
    - Consider an ant living on $\mathcal{M}$. From the universe ($\mathbb{R}^{p \times d}$), the ant is constrained on $\mathcal{M}$. But from the ant's perspective, $\mathcal{M}$ is all they have i.e. he/she would feel *unconstrained*!
- In this case, the optimization algorithm will make use of the *Riemannian* gradients and Hessians of $\mathcal{M}$.
- By making use of the intrinsic geometry of $\mathcal{M}$, the optimization becomes much more efficient!

# Quick Intuition behind Manifold Optimization

- A very straightforward way to think of this is by considering the simplest Riemannian manifold[2], $\mathbb{R}^{p\times d}$.
- When we write the optimization as

$$\begin{aligned} \underset{V}{\text{minimize}} \quad & f(V) \\ \text{subject to} \quad & V \in \mathbb{R}^{p\times d}, \end{aligned} \tag{8}$$

  technically this is a "constrained" optimization because we're "constraining" $V$ to be in $\mathbb{R}^{p\times d}$.
- However, gradients and Hessian (and other geometric concepts) are derived directly from the intrinsic geometry of $\mathbb{R}^{p\times d}$ i.e. $V \in \mathbb{R}^{p\times d}$ **isn't considered as a constraint.**

[2] inner product is the Frobenius product: $\langle X, Y\rangle := \mathrm{tr}(X^\intercal Y)$

# MbF-PCA

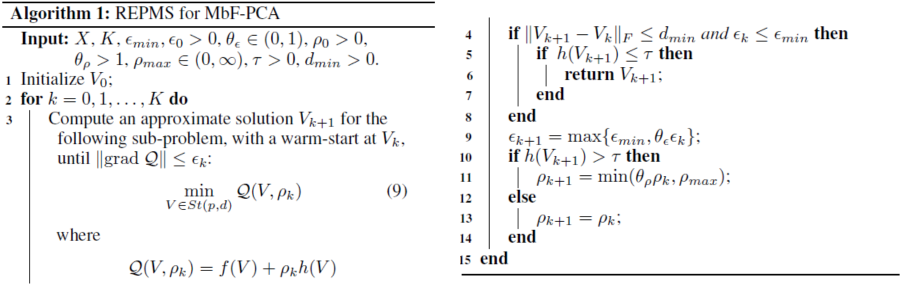
- Thus fair PCA can be formulated as a constrained manifold optimization problem, which we refer to as MbF-PCA:

$$\begin{aligned} \underset{V \in St(p,d)}{\text{maximize}} \quad & \left\langle \frac{1}{n} X^\intercal X, VV^\intercal \right\rangle \\ \text{subject to} \quad & h(V) := MMD^2(\hat{Q}_0, \hat{Q}_1) = 0. \end{aligned} \tag{9}$$

- With this formulation, we now have only one constraint.
- Moreover, leveraging kernel trick, an explicit form of $\nabla_V h(V)$ can be found; see Section G of the SP [LKO$^+$21].

# REPMS for MbF-PCA

- To solve the optimization, we use REPMS [LB19].



```
Algorithm 1: REPMS for MbF-PCA
Input: $X, K, \epsilon_{min}, \epsilon_0 > 0, \theta_\epsilon \in (0,1), \rho_0 > 0,$
       $\theta_\rho > 1, \rho_{max} \in (0, \infty), \tau > 0, d_{min} > 0.$
Initialize $V_0$;
for $k = 0, 1, \dots, K$ do
    Compute an approximate solution $V_{k+1}$ for the
      following sub-problem, with a warm-start at $V_k$,
      until $\|\text{grad } \mathcal{Q}\| \le \epsilon_k$:
          $\min_{V \in St(p,d)} \mathcal{Q}(V, \rho_k)$          (9)
      where
          $\mathcal{Q}(V, \rho_k) = f(V) + \rho_k h(V)$
    if $\|V_{k+1} - V_k\|_F \le d_{min}$ and $\epsilon_k \le \epsilon_{min}$ then
        if $h(V_{k+1}) \le \tau$ then
            return $V_{k+1}$;
        end
    end
    $\epsilon_{k+1} = \max\{\epsilon_{min}, \theta_\epsilon \epsilon_k\}$;
    if $h(V_{k+1}) > \tau$ then
        $\rho_{k+1} = \min(\theta_\rho \rho_k, \rho_{max})$;
    else
        $\rho_{k+1} = \rho_k$;
    end
end
```

Figure: Pseudocode

# Practical Consideration

- For practical concerns, we've set the fairness tolerance level, $\tau$, to be a fixed and sufficiently small, non-negative value.
- Accordingly, we consider the following definition:

### Definition

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**.

# New Theoretical Guarantees

- Our problem is non-convex in $V$, which naturally brings up the question of convergence and optimality guarantees.
- First, we theoretically motivate the following assumption, which is to the best of our knowledge, new:

**Assumption (informal; locality assumption)**

*Each $V_{k+1}$ is sufficiently close to a local minimum of Eq. (9).*

- Also, we consider the following auxiliary optimization problem:

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

# New Theoretical Guarantees

- Our problem is non-convex in $V$, which naturally brings up the question of convergence and optimality guarantees.
- First, under *ideal* hyperparameter setting, we provide an exact theoretical optimality guarantee:

## Theorem

*Let* $K = \infty$, $\rho_{max} = \infty$, $\epsilon_{min} = \tau = 0$, $\{V_k\}$ *be the sequence generated by Alg. 6 under Assumption 1, and* $\overline{V}$ *be any limit point of* $\{V_k\}$*, whose existence is guaranteed. Then the following holds:*

- $\overline{V}$ *is a local minimizer of Eq.* (10)*, which is a necessary condition for* $\overline{V}$ *to be fair.*
- *If* $\overline{V}$ *is fair, then* $\overline{V}$ *is a local minimizer of Eq.* (9)

# New Theoretical Guarantees

- The desired optimality guarantee is in the second bullet point, but it requires the assumption of $\overline{V}$ being fair.
- Such assumption is at least partially justified in the first bullet point in the following sense:
  - The ideal hyperparameter setting of $\rho_{max} = \infty, \tau = 0, \epsilon_{min} = 0$ implies the *exact* local minimality of $\overline{V}$ for Eq. (10), which is in turn a *necessary condition* for $\overline{V}$ to be fair.

# New Theoretical Guarantees

## Theorem

*Let $K = \infty$, $\rho_{max} < \infty$, $\epsilon_{min}, \tau > 0$. Then for any sufficiently small $\epsilon_{min}$ and $\tilde{r} = \tilde{r}(\epsilon_{min}) > 0$, the following hold:*

- $\overline{V}$ *is an approximate local minimizer of Eq. (10) in the sense that*

$$h(\overline{V}) \leq h(V) + \beta \|V - \overline{V}\| + (\beta + L_h) g(\epsilon_{min}) \tag{11}$$

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

  - $0 < \beta \leq \frac{2\|\Sigma\|}{\rho_0}$
  - $\beta(\rho_{max}, \tau)$ *is increasing in $\rho_{max}$ and decreasing in $\tau$.*

# New Theoretical Guarantees

## Theorem

- *If $\overline{V}$ is fair, then it is an approximate local minimizer of Eq. (9) in the sense that it satisfies*

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

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

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

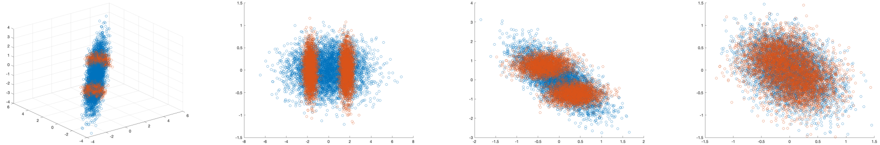
# Novelty of the guarantees (*)

- Existing optimality guarantee of REPMS (Proposition 4.2; [LB19]) states that when $\epsilon_{min} = 0$, $\rho$ 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 [YZS14].
- 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
  - by *introducing* a reasonable, yet novel locality assumption, we go beyond the existing KKT conditions and prove the *local minimality* of the limit point.
  - we provide a partial justification of the feasibility assumption in the first bullet point by proving a necessary condition.

# Outline

1. Introduction
2. Review of FPCA
3. Problems with FPCA
   - Problems with the definition
   - Problems with the SDP
4. New Definition: $\Delta$-fairness
5. New Framework: MbF-PCA
6. Experiments
7. Conclusion/Future works
8. References

# Synthetic data #1



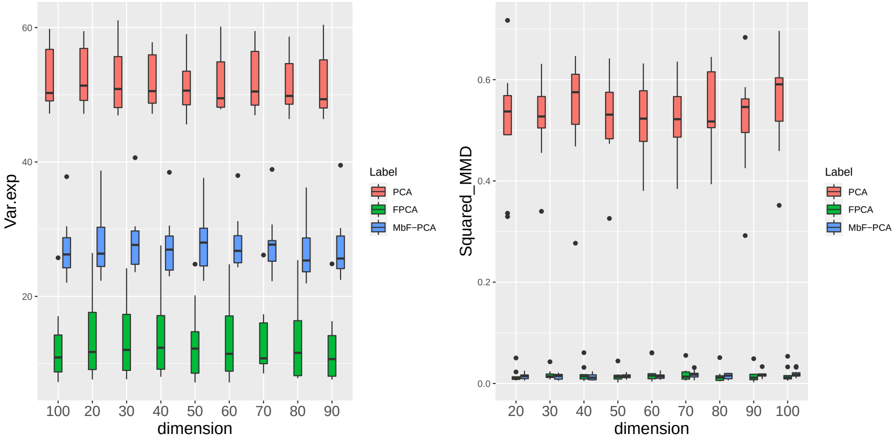
(a) Original data (b) PCA (c) FPCA [OA19] (d) MbF-PCA (ours)

Figure: Synthetic data #1: Comparison of PCA, FPCA, and MbF-PCA on data composed of two groups with same mean and covariance, but different distributions. Blue and orange represent different protected groups.

# 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 $p \in \{20, 30, \ldots, 100\}$; see Section H of the SP [LKO$^+$21] for a full description of the setting.

# Synthetic data #2



(a) Variance explained (%)

(b) $MMD^2$

Figure: Synthetic data #2: Comparison of PCA, FPCA, and MbF-PCA on the synthetic datasets of increasing dimensions.

# Synthetic data #2

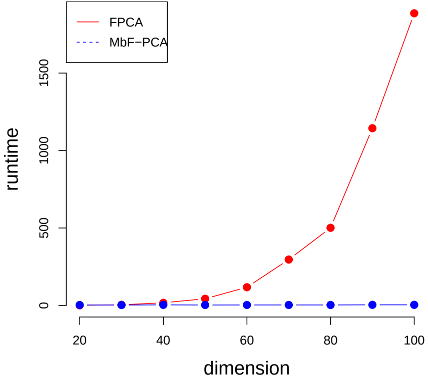


Figure: FPCA represents the SDP algorithm for fair PCA, and MbF-PCA represents our to-be introduced manifold-based algorithm for fair PCA. [LKO$^+$21]

# UCI Datasets

- We consider 3 datasets from the UCI Repository [DG17] (for pre-processing, we used AI Fairness 360 built-in functionalities [BDH+18]):
    - COMPAS dataset [KLMA16]
        - $n = 2468$, $p = 11$
        - sensitive feature: “race”
        - downstream classification: “crime again? (recidivism)”
    - German credit dataset
        - $n = 1000$, $p = 57$
        - sensitive feature: “age”
        - downstream classification: “good credit class?”
    - Adult income dataset
        - $n = 2261$, $p = 99$
        - sensitive feature: “gender”
        - downstream classification: “income larger than 50K?”

# UCI Datasets

- We compare our manifold-based MbF-PCA with the SDP-based approach and PCA, both in explained variance and fairness.
  - We emphasize that the **only** comparable approach for this problem is PCA and FPCA!
- For fairness, we consider our proposed MMD-metric *and* downstream task fairness.

# UCI Datasets

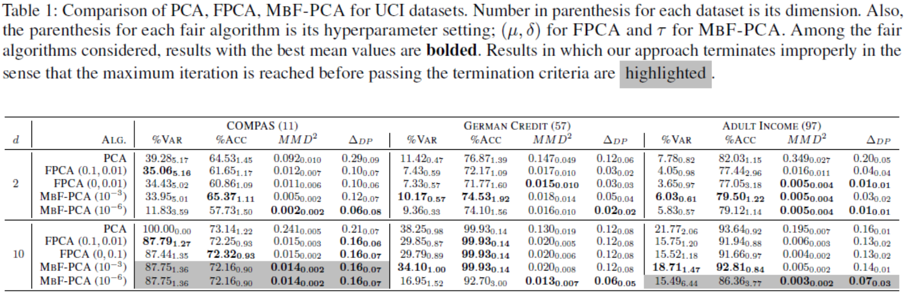


Table 1: Comparison of PCA, FPCA, MBF-PCA for UCI datasets. Number in parenthesis for each dataset is its dimension. Also, the parenthesis for each fair algorithm is its hyperparameter setting; $(\mu, \delta)$ for FPCA and $\tau$ for MBF-PCA. Among the fair algorithms considered, results with the best 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.

| | | COMPAS (11) | | | | German Credit (57) | | | | Adult Income (97) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $d$ | Alg. | %Var | %Acc | $MMD^2$ | $\Delta_{DP}$ | %Var | %Acc | $MMD^2$ | $\Delta_{DP}$ | %Var | %Acc | $MMD^2$ | $\Delta_{DP}$ |
| 2 | PCA | $39.28_{5.17}$ | $64.53_{1.45}$ | $0.092_{0.010}$ | $0.29_{0.09}$ | $11.42_{0.47}$ | $76.87_{1.39}$ | $0.147_{0.049}$ | $0.12_{0.06}$ | $7.78_{0.82}$ | $82.03_{1.15}$ | $0.349_{0.027}$ | $0.20_{0.05}$ |
| | FPCA $(0.1, 0.01)$ | **$35.06_{5.16}$** | $61.65_{1.17}$ | $0.012_{0.007}$ | $0.10_{0.07}$ | $7.43_{0.59}$ | $72.17_{1.09}$ | $0.017_{0.010}$ | $0.03_{0.02}$ | $4.05_{0.98}$ | $77.44_{2.96}$ | $0.016_{0.011}$ | $0.04_{0.04}$ |
| | FPCA $(0, 0.01)$ | $34.43_{5.02}$ | $60.86_{1.09}$ | $0.011_{0.006}$ | $0.10_{0.06}$ | $7.33_{0.57}$ | $71.77_{1.60}$ | **$0.015_{0.010}$** | $0.03_{0.03}$ | $3.65_{0.97}$ | $77.05_{3.18}$ | **$0.005_{0.004}$** | **$0.01_{0.01}$** |
| | MBF-PCA $(10^{-3})$ | $33.95_{5.01}$ | **$65.37_{1.11}$** | $0.005_{0.002}$ | $0.12_{0.07}$ | **$10.17_{0.57}$** | **$74.53_{1.92}$** | $0.018_{0.014}$ | $0.05_{0.04}$ | **$6.03_{0.61}$** | **$79.50_{1.22}$** | **$0.005_{0.004}$** | $0.03_{0.02}$ |
| | MBF-PCA $(10^{-6})$ | $11.83_{3.59}$ | $57.73_{1.50}$ | **$0.002_{0.002}$** | **$0.06_{0.08}$** | $9.36_{0.33}$ | $74.10_{1.56}$ | $0.016_{0.010}$ | **$0.02_{0.02}$** | $5.83_{0.57}$ | $79.12_{1.14}$ | **$0.005_{0.004}$** | **$0.01_{0.01}$** |
| 10 | PCA | $100.00_{0.00}$ | $73.14_{1.22}$ | $0.241_{0.005}$ | $0.21_{0.07}$ | $38.25_{0.98}$ | $99.93_{0.14}$ | $0.130_{0.019}$ | $0.12_{0.08}$ | $21.77_{2.06}$ | $93.64_{0.92}$ | $0.195_{0.007}$ | $0.16_{0.01}$ |
| | FPCA $(0.1, 0.01)$ | **$87.79_{1.27}$** | $72.25_{0.93}$ | $0.015_{0.003}$ | **$0.16_{0.06}$** | $29.85_{0.87}$ | **$99.93_{0.14}$** | $0.020_{0.005}$ | $0.12_{0.08}$ | $15.75_{1.20}$ | $91.94_{0.88}$ | $0.006_{0.003}$ | $0.13_{0.02}$ |
| | FPCA $(0, 0.1)$ | $87.44_{1.35}$ | **$72.32_{0.93}$** | $0.015_{0.002}$ | **$0.16_{0.07}$** | $29.79_{0.89}$ | **$99.93_{0.14}$** | $0.020_{0.006}$ | $0.12_{0.08}$ | $15.52_{1.18}$ | $91.66_{0.97}$ | $0.004_{0.002}$ | $0.13_{0.02}$ |
| | MBF-PCA $(10^{-3})$ | $87.75_{1.36}$ | $72.16_{0.90}$ | **$0.014_{0.002}$** | **$0.16_{0.07}$** | **$34.10_{1.00}$** | **$99.93_{0.14}$** | $0.020_{0.008}$ | $0.12_{0.08}$ | **$18.71_{1.47}$** | **$92.81_{0.84}$** | $0.005_{0.002}$ | $0.14_{0.01}$ |
| | MBF-PCA $(10^{-6})$ | $87.75_{1.36}$ | $72.16_{0.90}$ | **$0.014_{0.002}$** | **$0.16_{0.07}$** | $16.95_{1.52}$ | $92.70_{3.00}$ | **$0.013_{0.007}$** | **$0.06_{0.05}$** | $15.49_{6.44}$ | $86.36_{3.77}$ | **$0.003_{0.002}$** | **$0.07_{0.03}$** |

- Across all considered datasets, MBF-PCA is shown to outperform FPCA in terms of fairness ($MMD^2$ and $\Delta_{DP}$) with low enough $\tau$.
- For German Credit and Adult Income, MBF-PCA shows a clear trade-off between explained variance and fairness; by relaxing $\tau$, we see that MBF-PCA outperforms FPCA in terms of explained variance and downstream task accuracy.

# UCI Datasets

- Orthogonality of the PCs allows for us to *interpret* the PCs.
  - This is part of a big set of techniques for interpretable ML, called **exploratory data analysis (EDA)**
- Communality of a feature is its variance contributed by the PCs [JW08].
  - This is computed as the sum of squares of the loadings of the considered feature.
- It can be seen that the PCs resulting from MBF-PCA have the least correlations with age, the protected attribute.

# UCI Datasets

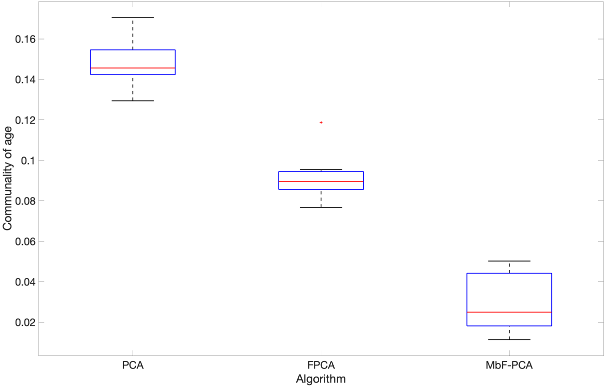
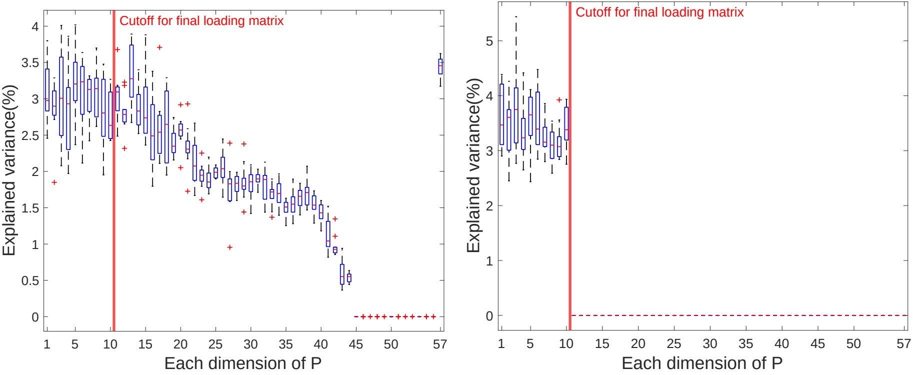


Figure: Comparison of communality of "age" of German credit dataset for PCA, FPCA, and MBF-PCA.

# Low Explained Variance of FPCA (*)

- We analyze why this occurs using GERMAN CREDIT DATASET.
- Empirically, we observe that the leakage of explained variance occurs due to the relaxed rank constraint.
- Theoretically, under Gaussian assumption, we show that this is due to the relaxed rank constraint and the complicated effect of covariance constraint. (See Section C of the SP [LKO$^+$21] for details)

# Low Explained Variance of FPCA (*)



(a) FPCA [OA19]

(b) MBF-PCA (ours)

Figure: 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 $P$)

# Outline

1. Introduction
2. Review of FPCA
3. Problems with FPCA
    - Problems with the definition
    - Problems with the SDP
4. New Definition: $\Delta$-fairness
5. New Framework: MbF-PCA
6. Experiments
7. Conclusion/Future works
8. References

# Conclusion

- **New definition** for fair PCA based on MMD.
- Theoretical discussions on why our definition is more desirable than the previous one [OA19].
- MbF-PCA, a **new manifold optimization framework** for fair PCA, along with theoretical discussions on why this approach is more favorable than the previous SDP-based approach [OA19].
- Provide new theoretical(optimality) guarantees of REPMS in both ideal and practical hyperparameter setting, extending previous results.
- Empirically, we show the efficacy of our algorithm on synthetic and UCI datasets in explained variance, fairness, and runtime.

# Future works

- Statistical characterizations of our fair PCA in asymptotic regime, as well as incorporation of sparsity [JLN$^+$09]
- Incorporating stochastic optimization-type modifications [Sha15, RLWS21]
  - This particularly important because our current approach scales quadratically in the number of data points as the computation of $MMD^2$ requires the full kernel gram matrix.
- Further exploration into the interpretation of fair loading matrix.

*Thank you for your attention! Any questions?*

# Outline

1. Introduction
2. Review of FPCA
3. Problems with FPCA
    - Problems with the definition
    - Problems with the SDP
4. New Definition: $\Delta$-fairness
5. New Framework: MbF-PCA
6. Experiments
7. Conclusion/Future works
8. References

# References I

Raman Arora, Andy Cotter, and Nati Srebro, *Stochastic optimization of pca with capped msg*, Advances in Neural Information Processing Systems (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), vol. 26, Curran Associates, Inc., 2013, pp. 1815–1823.

Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller, *One-network adversarial fairness*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 2412–2420.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang, *AI Fairness 360: An extensible toolkit for detecting,*

# References II

*understanding, and mitigating unwanted algorithmic bias*, October 2018.

Moustapha Cisse and Sanmi Koyejo, *Nips 2019 tutorial: Fairness and representation learning*, 2019.

Dheeru Dua and Casey Graff, *UCI machine learning repository*, 2017.

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

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf, *Kernel measures of conditional dependence*, Advances in Neural Information Processing Systems (J. Platt, D. Koller, Y. Singer, and S. Roweis, eds.), vol. 20, Curran Associates, Inc., 2008.

# References III

- Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola, *A kernel method for the two-sample-problem*, Advances in Neural Information Processing Systems (B. Schölkopf, J. Platt, and T. Hoffman, eds.), vol. 19, MIT Press, 2007.

- Moritz Hardt, Eric Price, and Nati Srebro, *Equality of opportunity in supervised learning*, Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016 (Barcelona, Spain), 2016, pp. 3315–3323.

- Iain M. Johnstone, Arthur Yu Lu, Boaz Nadler, Daniela M. Witten, Trevor Hastie, Robert Tibshirani, and James O. Ramsay, *On consistency and sparsity for principal components analysis in high dimensions [with comments]*, Journal of the American Statistical Association **104** (2009), no. 486, 682–703.

# References IV

- Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon, *Fair feature distillation for visual recognition*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, pp. 12115–12124.

- Richard A. Johnson and Dean W. Wichern, *Applied multivariate statistical analysis*, 6 ed., Pearson, 2008.

- L Kirchner, J. Larson, S. Mattu, and J. Angwin, *Machine bias*, ProPublica, 2016.

- Changshuo Liu and Nicolas Boumal, *Simple algorithms for optimization on riemannian manifolds with constraints*, Applied Mathematics and Optimization **82** (2019), 949–981.

- Junghyun Lee, Gwangsu Kim, Matt Olfat, Mark Hasegawa-Johnson, and Chang D. Yoo, *Fast and efficient mmd-based fair pca via optimization over stiefel manifold*, 2021.

# References V

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel, *The variational fair autoencoder*, 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (Yoshua Bengio and Yann LeCun, eds.), 2016.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel, *Learning adversarially fair and transferable representations*, Proceedings of the 35th International Conference on Machine Learning (Jennifer Dy and Andreas Krause, eds.), Proceedings of Machine Learning Research, vol. 80, PMLR, 10–15 Jul 2018, pp. 3384–3393.

Matt Olfat and Anil Aswani, *Convex formulations for fair principal component analysis*, The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, 2019, pp. 663–670.

# References VI

- Luca Oneto, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, and Massimiliano Pontil, *Exploiting mmd and sinkhorn divergences for fair and transferable representation learning*, Advances in Neural Information Processing Systems (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, Curran Associates, Inc., 2020, pp. 15360–15370.

- Flavlen Prost, Hal Qian, Qiuwen Chen, Ed H. Chi, Jilin Chen, and Alex Beutel, *Toward a better trade-off between performance and fairness with kernel-based distribution matching*, NeurIPS 2019 Workshop on Machine Learning with Guarantees (Vancouver, BC, Canada), 2019.

# References VII

- Novi Quadrianto and Viktoriia Sharmanska, *Recycling privileged learning and distribution matching for fairness*, Advances in Neural Information Processing Systems (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

- Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh, *Fairbatch: Batch selection for model fairness*, 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021.

- Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert R.G. Lanckriet, and Bernhard Schölkopf, *Injective hilbert space embeddings of probability measures*, Proceedings of the 21st Annual Conference on Learning Theory (COLT) (Rocco Servedio and Tong Zhang, eds.), 2008, pp. 111–222.

# References VIII

Ohad Shamir, *A stochastic pca and svd algorithm with an exponential convergence rate*, Proceedings of the 32nd International Conference on Machine Learning (Lille, France) (Francis Bach and David Blei, eds.), Proceedings of Machine Learning Research, vol. 37, PMLR, 07–09 Jul 2015, pp. 144–152.

S. Samadi, U. T. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. S. Vempala, *The price of fair PCA: one extra dimension*, Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018 (Montréal, Canada), 2018, pp. 10999–11010.

# References IX

- U. Tantipongpipat, S. Samadi, M. Singh, J. H. Morgenstern, and S. S. Vempala, *Multi-criteria dimensionality reduction with applications to fairness*, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019 (Vancouver, BC, Canada), 2019, pp. 15135–15145.

- Wei Hong Yang, Lei-Hong Zhang, and Ruyi Song, *Optimality conditions for the nonlinear programming problems on riemannian manifolds*, Pacific Journal of Optimization **10** (2014), 415–434.