Privacy-Preserving Fair Machine Learning Without Collecting Sensitive Demographic Data

Hui Hu
Department of Computer Science
University of Wyoming
Laramie, USA
hhu1@uwyo.edu

Mike Borowczak
Department of Computer Science
University of Wyoming
Laramie, USA
mike.borowczak@uwyo.edu

Zhengzhang Chen
Department of Electrical Engineering and Computer Science
Northwestern University
Evanston, USA
zzc472@eecs.northwestern.edu

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9534017

***Abstract*—With the rising concerns over privacy and fairness in machine learning, privacy-preserving fair machine learning has received tremendous attention in recent years. However, most existing fair models still need to collect sensitive demographic data, which may be impossible given privacy regulations. To address the dilemma between model fairness and sensitive data collection, we propose DicPF, a distributed and privacy-preserving fair learning framework that operates without collecting sensitive demographic data. In particular, DicPF assumes multiple local agents and a modeler are distributed, and sensitive demographic data is separately held by multiple local agents. To assist fair learning at the modeler, each agent learns a fair local dictionary and send it to the modeler. The modeler learns a fair model based on an aggregated dictionary. Under DicPF framework, we propose a private z-Sparse Fair Learner. Extensive experiments on three real-world datasets demonstrate the efficiency of the proposed model. Compared with the state-of-the-art fair learners, the proposed z-Sparse Fair Learner achieves superior fairness performance by lowering prediction disparity. We also develop a privacy inference model to demonstrate the excellent privacy-preserving performance of DicPF. Finally, we theoretically analyze z-Sparse Fair Learner and prove upper bounds on its model fairness and accuracy.**

***Index Terms*—fair machine learning, privacy-preserving, dictionary learning, sparse representation theory, inference attack**

I. Introduction

Machine learning is widely used to make important and life-changing decisions from helping us decide who to hire to assessing violence risk in prisons [1], [2]. It is crucial to ensure that the decisions are not based on prior discriminatory behaviors toward certain groups or populations. Thus, building fair models is extremely important and it is part of the latest national AI R&D strategy plan [3].

Many fair models have been proposed, but most of them require direct or indirect access to private data. In practice, many situations arise where it is impossible to collect sensitive demographic data for decision-making. The main reasons are from two aspects: On one hand, data privacy protection is being forced in regulations such as the Europe General Data Protection Regulation (GDPR) and the latest California Consumer Privacy Act (CCPA) [3]; On the other hand, individuals are not willing to disclose their private data to the modeler [4]. We thus see fair machine learning and privacy protection are running into a dilemma.

To address this problem, a few solutions have been proposed in the literature, which can be categorized into multi-party computation [4], [5], demographic proxy [6] and demographic data query with cost [7]. While they have achieved promising results, each direction has its own limits. One common limit of most solutions is that they need to collect users' private data. In practice, this may be impossible given privacy regulations. For example, hospitals have a large number of patient cases, a data analysis of these cases helps the doctors make accurate diagnoses. However, each hospital can not share its patient cases with research institutions as the information about the patients is extremely private [8]. In this situation, the aforementioned technical solutions may be impractical.

In this paper, we propose **DicPF**, a novel privacy-preserving fair machine learning framework that doesn't need to collect sensitive demographic data. **DicPF** assumes users' sensitive demographic data is privately and separately held by multiple local agents. To avoid disclosing users' private data to the modeler, each local agent learns a fair and accurate dictionary via dictionary learning technique, then sends the learned local dictionary to the modeler. The modeler firstly aggregates the dictionaries sent by multiple local agents; then learns a fair model only with the non-private data and the aggregated dictionary. Under this framework, we propose a private **z-Sparse Fair Learner**, which learns a fair model by sparsely selecting atoms from the aggregated dictionary. Our insight is that (i) model fairness is promised by sparse fair atom selection and (ii) model accuracy is guaranteed by sparse representation theory [9], [10]. The experimental results show **z-Sparse Fair Learner** outperforms most existing non-private and semi-private fair learners across three real-world datasets. Next, we examine the privacy-preserving performance of **DicPF** framework under an inference attack, the experimental results show the proposed framework has better privacy-preserving performance than non-private and semi-private models [5]. Finally, we theoretically analyze **z-Sparse Fair Learner** and prove upper bounds on its model fairness and accuracy.

To sum up, the contributions of this work include:

- Propose a privacy-preserving and distributed fair machine learning framework **DicPF**, which learns a fair model without collecting sensitive demographic data;
- Propose a private **z-Sparse Fair Learner** under **DicPF**

978-0-7381-3366-9/21/$31.00 ©2021 IEEE

framework, which outperforms most state-of-the-art fair learners in model fairness;
- Develop a privacy inference model to demonstrate the better privacy-preserving performance of **DicPF** framework compared with non-private and semi-private models;
- Provide theoretical analysis on **z-Sparse Fair Learner** and prove upper bounds on its model fairness and accuracy.

The remainder of this paper is organized as follows: Section II introduces the related work; Section III formalizes the problem statement; Section IV presents the proposed framework; Section V describes the design of z-sparse fair learner; Section VI presents an inference attack; Section VII presents theoretical analysis; Section VIII shows experimental results; Section IX presents conclusion and future work; And Appendix contains all proofs.

II. Related Work

A. Fair Learning with Direct Access to Private Data

Many fair models with direct access to sensitive demographic data have been developed. Feature processing [11]–[13] assumes a model is fair if it is built on a fair data representation. They first learn fair features and then learn a fair model from them. Label processing [14] assumes that there are unfair labels in training data. They detect and correct these labels before model learning. Model in-processing [15]–[18] modifies the learning algorithms during the training process by incorporating changes into the objective function or imposing a fairness constraint. Model post-processing [19], [20] learns a standard model and modifies its predictions to make them fair. Model ensemble [21] supposes an ensemble of standard models is fair as unethical biases in these models may be averaged out through bagging. In [22], the authors present an interesting framework to mitigate bias via adversarial learning technique. In this framework, they maximize accuracy of the predictor on y, and at the same time minimize the ability of the adversary to predict the sensitive variable. Even though some fair models have achieved good performance, they need to use sensitive demographic data directly.

B. Fair Learning with Restricted Access to Private Data

Specific discussions on the restricted use of private data appears in [23], [24], but there lacks scientific solutions.The natural solution to protect privacy in fair learning is removing the demographic feature from the model, but this approach can not guarantee fairness due to the redlining effect [25]. Some studies do not use demographic data as a feature of the model, but use it in other ways during learning. For example, [26] uses k-NN to detect unfair labels. They do not use demographic data to measure instance similarity, but still use it to measure label disparity in neighborhoods. Recently, Kilbertus *et al.* [4] propose an interesting solution by employing the cryptographic tool of secure multiparty computation. This is a promising solution, but encryption comes with extra cost of time and protocols. Hashimoto *et al.* [27] propose a fair learner that automatically infers group membership and minimizes disparity across it. However, this method focuses on a less common fairness notion called distributive justice and on-line learning, In contrast, we focus on the common disparity measure and off-line setting. In [6], the authors use proxy sensitive features to mitigate model bias, but its fairness performance is decided by the accuracy of proxy features. H. Hu *et al.* [5] propose a distributed fair learning framework, where sensitive demographic data is collected by a trusted third party and the modeler communicates with the third party for fair learning. Nevertheless, this framework is vulnerable to inference attacks as an attacker can infer sensitive information with high accuracy [28]. In addition, in reality, it is not always realistic to find a highly trusted third party to hold private data and participate in model training. Preethi Lahoti *et al.* [29] propose an interesting approach to mitigate bias without demographics, which is adversarially reweighted learning. Their method hypothesizes that non-sensitive features and task labels are valuable for identifying fairness issues. Y. Liu *et al.* [7] adopt active learning technique to train a fair model, but their assumption is that sensitive demographic data can be collected.

TABLE I: Summary of Notations

Notations	Description
W_i	The i-th local agent
$\mathcal{W}$	The local agent set
$\mathcal{M}$	The modeler
f	The prediction model
f_z	z-sparse prediction model
$\mathbf{X}$	Non-private data
$\mathbf{X_i}$	Local data at i-th local agent
$\mathbf{S_i}$	Private data at i-th local agent
Y	True label vector
$\mathbf{Z}$	Sparse matrix
$\mathbf{D}_{ia}$	An accurate dictionary related to $\mathbf{X}_i$
$\mathbf{D}_i$	The local dictionary at the i-th agent
$\mathbf{D}$	The aggregated dictionary at modeler
d_i	The i-th atom
$\vec{\beta}$	Coefficient vector
$\lVert\cdot\rVert_1$, $\lVert\cdot\rVert_2$	L_1 norm, L_2 norm
λ_0, λ	Hyper-parameters
ρ	Fairness threshold
r	The number of local agents

III. Problem Statement

In this section, we present the preliminaries and problem definition of our work.

The symbols used in the paper are summarized in Table I. The bold capital letters denote matrix. The local agent set is defined as $\mathcal{W} := \{W_1, \ldots, W_r\}$, where W_i $(i \in [1, r])$ denotes the i-th local agent. Let $\mathcal{M}$ denote the modeler. A random instance is described by a triple (x, s, y), where $s \in \mathbb{R}$ is a sensitive demographic feature, $x \in \mathbb{R}^p$ is a vector of p non-sensitive features, and $y \in \mathbb{R}$ is the true label. Let $\mathbf{X} = \{\sum_{i=1}^{r}(x_{i1}, y_{i1}), ..., (x_{im}, y_{im})\}$ denote the non-private data of all users, and $\mathbf{X}_i = \{(x_{i1}, s_{i1}, y_{i1}), ..., (x_{im}, s_{im}, y_{im})\}$ denote the local data set at the i-th agent, which includes private data $S_i = \{s_{i1}, \ldots, s_{im}\}$. For ease of discussion, we will write $\mathbf{X} = [x_1, \ldots, x_n]^T$ as a sample matrix, $Y = [y_1, \ldots, y_n]^T$

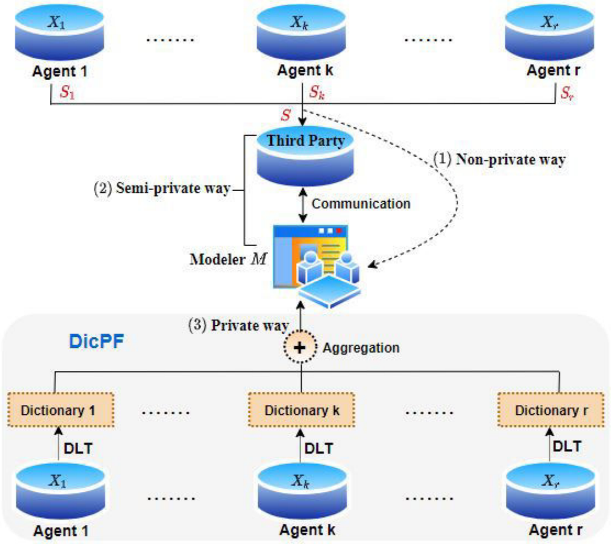


Fig. 1: Comparison between **DicPF** workflow and existing fair learning workflow. (1) *Non-private fair learning* requires direct access to private data and (2) *Semi-private fair learning* requires indirect access to private data [5]. (3) *Private fair learning* **DicPF** trains a fair model f **without collecting private data** from multiple local agents. (Red color shows private data; DLT denotes dictionary learning technique.)

as the associated label vector. Let $\mathbf{D}_i \in \mathbb{R}^{p\times k_i}$ be a learned local dictionary at the i-th agent. The aggregated dictionary at modeler $\mathcal{M}$ is denoted as $\mathbf{D} = \{\mathbf{D}_1, \ldots, \mathbf{D}_r\} = [d_1, \ldots, d_k]$, where $k = k_1 + \ldots + k_r$. Each column in a dictionary is called an atom, *i.e.*, each $d_i \in \mathbf{D}$ is an atom. Let $f : \{x\} \to \{y\}$ be a linear prediction model.

With the aforementioned notations, the problem we aim to study in this work can be formally defined as follows:

Given a local agent set $\mathcal{W} = \{W_1, \ldots, W_r\}$, where each W_i ($i \in [1, r]$) privately holds users' private data S_i and the modeler $\mathcal{M}$ only holds users' non-private data X and true labels Y, then our goal is to learn a fair model f at the modeler $\mathcal{M}$ without collecting users' private data S_i from each $W_i \in \mathcal{W}$.

IV. DicPF Framework

In this section, we present the proposed **DicPF** framework. As Figure 1 shows: assume there are r local agents and one modeler $\mathcal{M}$, private data of users is separately and privately held by multiple local agents, non-private data is distributed on modeler $\mathcal{M}$ and all local agents. **DicPF** framework has two advantages:

① Different from the existing non-private and semi-private fair learning frameworks (as shown in Figure 1 (1) and (2)), the modeler $\mathcal{M}$ in **DicPF** only separately obtains r accurate and fair dictionaries from multiple local agents instead of collecting users' private data.

② The modeler $\mathcal{M}$ trains a fair model f with an aggregated dictionary $\mathbf{D}$ and non-private data $\mathbf{X}$, which doesn't exchange information with the holders of private data.

Algorithm 1 DicPF Framework

```
Input: Modeler $\mathcal{M}$, non-private training set $\mathbf{X} \in \mathbb{R}^{p\times n}$, local agent
    set $\mathcal{W}$, and fairness threshold $\rho$, $\lambda_0$.
Output: A prediction model $f$ at $\mathcal{M}$.
1: $\forall W_i \in \mathcal{W}$ separately learns an accurate dictionary $\mathbf{D}_{ia} \in \mathbb{R}^{p\times k_{ic}}$ with local data $\mathbf{X}_i$ via
       $\min_{\mathbf{D}_{ia}, \mathbf{Z}} ||\mathbf{X}_i - \mathbf{D}_{ia}\mathbf{Z}||_2^2 + \lambda_0 ||\mathbf{Z}||_1,$    (1)
   where $\mathbf{Z} \in \mathbb{R}^{k_{ic}\times m}$ is a sparse matrix and $\lambda_0$ is a hyperparameter.
2: $W_i$ applies each $d_i \in \mathbf{D}_{ia}$ on $\{x_{ij} \in \mathbf{X}_i\}$ to get a predicted label set $\hat{Y}_i = \{d_i(x_{i1}), \ldots, d_i(x_{im})\}$.
3: $W_i$ estimates $\mathrm{cov}(d_i(x), s)$ and put $d_i$ into matrix $\mathbf{D}_i$ if $|\mathrm{cov}(d_i(x), s)| \leq \rho$.
4: $W_i$ sends $\mathbf{D}_i$ to $\mathcal{M}$.
5: $\mathcal{M}$ receives $\mathbf{D_i}(i = 1, \ldots, r)$ and trains a fair prediction (or classification) model $f$ on $\mathbf{X}$ assuming that
       $f = \beta_1 d_1 + \beta_2 d_2 + \ldots + \beta_k d_k,$    (2)
   where $k = k_1 + \ldots + k_r$, $d_i \in \mathbf{D}(i \in [1, k])$ and $\vec{\beta} = [\beta_1, \ldots, \beta_k]^T$ is unknown coefficients to learn.
```

Next, we will elaborate the design details of **DicPF** framework. The strategy design is shown in Algorithm 1. It has two phases: (1) Steps 1 to 4 construct a fair and accurate dictionary $\mathbf{D}$ aggregated by $\mathbf{D_1}, \ldots, \mathbf{D_r}$; (2) Step 5 learns a fair model based on the aggregated dictionary $\mathbf{D}$ and non-private data $\mathbf{X}$.

Specifically, **Step 1**: Each local agent $W_i \in \mathcal{W}$ separately learns an accurate dictionary based on local data $\mathbf{X_i}$ via Equation (1) [10]; **Step 2**: W_i applies all atoms in the accurate dictionary on the local data $\mathbf{X_i}$ to get predictions; **Step 3**: W_i estimates correlation between its sensitive demographic data and each atom's prediction. If a correlation is small enough, then the corresponding atom is fair and put it into $\mathbf{D_i}$; **Step 4**: W_i sends $\mathbf{D_i}$ to the modeler $\mathcal{M}$; **Step 5**: The modeler $\mathcal{M}$ learns a fair prediction model f with the aggregated dictionary $\mathbf{D}$ and non-private data $\mathbf{X}$.Throughout the process, sensitive demographic data at each agent is not revealed to the modeler.

V. z-Sparse Fair Learner

Under **DicPF** framework, we propose a private ***z*-Sparse Fair Learner**. The motivation for our design is that sparsity can ensure model fairness and accuracy performance simultaneously based on Algorithm 1 (see Section VII for theoretical analysis). To elaborate the design of ***z*-Sparse Fair Learner**, the following terms need to be defined first.

Definition 1: (*z-Sparsity* [9]). Let $\mathbf{A} \in R^{N\times N}$, $x \in R^{N\times 1}$, and $x = \mathbf{A}v$, where $v \in R^{N\times 1}$ is the column vector of weighting coefficients. If only $z(z << N)$ elements of v are nonzero and the rest elements in v are zeros, we call x is z-sparse.

Definition 2: (*ρ-Fair Atom*). In Algorithm 1, an atom $d_i \in \mathbb{R}^p$ is ρ-fair ($\rho > 0$) if

$$|\mathrm{cov}[d_i(x), s]| \leq \rho, \tag{3}$$

where $\mathrm{cov}(\cdot)$ is covariance and $|\cdot|$ denotes absolute value.

Definition 3: (*ρ-Fair Dictionary*). In Algorithm 1, if $\forall d_i \in \mathbf{D}$ satisfies $|\mathrm{cov}[d_i(x), s]| \leq \rho$, then $\mathbf{D} \in \mathbb{R}^{p\times k}$ is a ρ-fair dictionary, *i.e.*, a ρ-fair dictionary is composed of k ρ-fair atoms.

Based on Definition 1, we say f is z-sparse linear prediction model if only z coefficients are non-zeros in vector $\vec{\beta}$, which is denoted as f_z.

To obtain an optimal fair model f_z in Algorithm 1, the objective function of **z-Sparse Fair Learner** is designed as follows:

$$\begin{aligned} J(f_z) &= \min_{\vec{\beta}} \sum_{i=1}^{n} (f_z(x_i) - y_i)^2 + 2\lambda ||\vec{\beta}||_1, \\ &= \min_{\vec{\beta}} \sum_{i=1}^{n} (\sum_{t=1}^{k} \beta_t d_t(x_i) - y_i)^2 + 2\lambda \sum_{t=1}^{k} |\beta_t|, \end{aligned} \tag{4}$$

where λ is a hyperparameter.

By equation (4), we learn a z-sparse $\vec{\beta}$ such that the model loss is minimal based on a ρ-fair dictionary $\mathbf{D}$.

Since L_1 norm is not differentiable, we apply coordinate descent algorithm [30] to solve Equation (4). This numerical method iteratively updates $\vec{\beta}$. In each iteration, it optimizes a random element $\beta_j (j \neq 0)$ while fixing the rest. We rewrite $J(f_z)$ as

$$\begin{aligned} J(f_z) &= ||\mathbf{X}\mathbf{D}_{:j}\beta_j + \sum_{t \neq j} \mathbf{X}\mathbf{D}_{:t}\beta_t - Y||_2^2 + 2\lambda \sum_{t=1}^{k} |\beta_t| \\ &= \sum_{i=1}^{n} (x_i^T \mathbf{D}_{:j}\beta_j + \mathbf{A}_i^{(j)})^2 + 2\lambda|\beta_j| + B^{(j)} \end{aligned} \tag{5}$$

where $\mathbf{A}^{(j)} = \sum_{t \neq j} \mathbf{X}\mathbf{D}_{:t}\beta_t - Y$, $B^{(j)} = 2\lambda \sum_{t \neq j} |\beta_t|$.

Because $|\beta_j|$ is not differentiable, we remove the absolute value by case-studying β_j and apply critical point method to obtain β_j.

Case 1: $\beta_j > 0$**.**

$$\frac{\partial J(f_z)}{\partial \beta_j} = \sum_{i=1}^{n} 2x_i^T \mathbf{D}_{:j}(x_i^T \mathbf{D}_{:j}\beta_j + \mathbf{A}_i^{(j)}) + 2\lambda. \tag{6}$$

Setting the right-hand-side to zero and solving for β_j, we have

$$\beta_j = \frac{-2\lambda - \sum_{i=1}^{n} 2x_i^T \mathbf{D}_{:j} \mathbf{A}_i^{(j)}}{\sum_{i=1}^{n} 2(x_i^T \mathbf{D}_{:j})^2}. \tag{7}$$

Since $\beta_j > 0$, then $-2\lambda > \sum_{i=1}^{n} 2x_i^T \mathbf{D}_{:j} \mathbf{A}_i^{(j)}$.

Case 2: $\beta_j < 0$**.** Similar to case 1, we have

$$\beta_j = \frac{2\lambda - \sum_{i=1}^{n} 2x_i^T \mathbf{D}_{:j} \mathbf{A}_i^{(j)}}{\sum_{i=1}^{n} 2(x_i^T \mathbf{D}_{:j})^2}. \tag{8}$$

Since $\beta_j < 0$, then $2\lambda < \sum_{i=1}^{n} 2x_i^T \mathbf{D}_{:j} \mathbf{A}_i^{(j)}$.

Case 3: $\beta_j = 0$**.** In this case, $2\lambda \geq |\sum_{i=1}^{n} 2x_i^T \mathbf{D}_{:j} \mathbf{A}_i^{(j)}|$.

Summarizing the three cases, we update $\beta_j (j \in [1, k])$ via

$$\beta_j = \begin{cases} \frac{-2\lambda - Q}{R} & \text{if} \quad Q < -2\lambda \\ \frac{2\lambda - Q}{R} & \text{if} \quad Q > 2\lambda \\ 0 & \text{if} \quad |Q| \leq 2\lambda \end{cases}, \tag{9}$$

where $Q = \sum_{i=1}^{n} 2x_i^T \mathbf{D}_{:j} \mathbf{A}_i^{(j)}$ and $R = \sum_{i=1}^{n} 2(x_i^T \mathbf{D}_{:j})^2$.

VI. Attack Model: Privacy Inference

In **DicPF** framework, we assume the modeler obeys privacy regulation and has no malicious attempt. However, in reality, the modeler may be an adversary who attempts to infer users' private information based on all accessible information. Her intention may be malicious, e.g., to conduct discriminatory decisions in some applications. Therefore, it is necessary to examine the vulnerability of the proposed framework under privacy inference attack. In this section, we first define privacy loss, then present an attack model for privacy inference.

Definition 4: (*Privacy Loss*). Given the original sensitive feature vector $S = \{s_1, \ldots, s_n\}$ and the inference vector of an adversary $\hat{S} = \{\hat{s}_1, \cdots, \hat{s}_n\}$, the privacy loss of S is defined as

$$PL_S(\mathbf{X}, \mathbf{D}, Y) = 1 - \frac{1}{n} \sum_{i=1}^{n} (s_i - \hat{s}_i)^2, \tag{10}$$

where $s_i, \hat{s}_i \in \{0, 1\}$ and $\hat{s}_i$ is the inferred sensitive feature value of individual x_i.

We see the privacy loss depends on the inference accuracy of adversary. The inference is more accurate, $PL_S(\mathbf{X}, \mathbf{D}, Y)$ is bigger.

The adversary's goal is to obtain an accurate inferred $\hat{S}$ based on all accessible information. In **DicPF** framework, The known background knowledge of the adversary includes: (i) original non-sensitive feature $\mathbf{X}$; (ii) the aggregated dictionary $\mathbf{D}$ and (iii) true labels of all individuals Y. However, the adversary does not know (i) original sensitive feature values S; (ii) The fairness strategies to obtain $\mathbf{D_i}$ at each agent.

Assume sensitive feature is binary, based on the known knowledge, the inference model can be formulated as

$$\begin{aligned} &\min\nolimits_{\hat{S}} \; SP(f_z), \\ &s.t. \; \hat{s}_i \in \{0, 1\}, \qquad i = 1, \ldots, n, \end{aligned} \tag{11}$$

where $SP(f_z)$ denotes the statistical parity of model f_z.

By equation (11), the adversary tries to find a binary vector such that the *statistical parity* of prediction is minimal.

VII. Theoretical Analysis

In this section, we present the theoretical properties of Algorithm 1.

A. Theoretical Properties on Model Fairness

We evaluate model fairness using a popular measure of *statistical parity* (SP) [31]. Let $p(f(x) = 1|s(x) = 1)$, $p(f(x) = 1|s = 0)$ be the probabilities of positive classification in two demographic groups, respectively,

$$\text{SP}(f) = |p(f(x) = 1|s(x) = 1) - p(f(x) = 1|s(x) = 0)|. \tag{12}$$

Our first result shows f_z is fair if it is spanned by a set of ρ-fair atoms.

Lemma 5: In Algorithm 1, if $f_z(x)$ is spanned by z ρ-fair atoms, then

$$\text{cov}[f_z(x), s] \leq \sqrt{z} ||\vec{\beta_z}|| \rho, \tag{13}$$

where $\vec{\beta_z}$ is the vector of non-zero sparse coefficients and z is the number of non-zero coefficients.

To prove z-sparsity on ρ-fair dictionary $\mathbf{D}$ implies statistical parity, we employ the arguments in [5, Lemma 2], then we have

Theorem 6: If a sparse model $f_z(x)$ and s are positively or negatively quadrant dependent, with a ρ-fair dictionary D, then

$$\text{SP}(f_z) \leq t * \sqrt{z}||\vec{\beta_z}||, \tag{14}$$

where $t = \rho / p(s=0)p(s=1)$.
This theorem implies when ρ is fixed, one can obtain a fair model through two paths: (i) Choose a small z, which means the model is sparser and will be fairer; (ii) Choose a small $||\vec{\beta_z}||$.

B. Theoretical Properties on Model Loss

To derive a loss bound for Algorithm 1, our backbone technique is dictionary learning [9], [10]. It is a sparse representation learning technique which aims at expressing a given sample x as a sparse linear combination of atoms. Intuitively, the classifier result is more reliable when the reconstruction error of samples is smaller [32].

Given any sample x_i, the first result shows the upper bound of reconstruction error based on an accurate dictionary $\mathbf{D}_{ia}$, which is learned via Equation (1).

Lemma 7: Assuming $||x_i|| \leq r$, $||\mathbf{D}_{ia} e_j|| \leq \gamma$ where $\{e_j | 1 \leq j \leq k_{ic}\}$ is the orthonormal basis of $\mathbb{R}^{k_{ic}}$, let $||z_i|| \leq 1$, then we have

$$||x_i - \mathbf{D}_{ia} z_i||^2 \leq r^2 + k_{ic}^2 \gamma^2. \tag{15}$$

This lemma implies that the reconstruction error of sample x_i has the worst-case upper bound on k_{ic} of $\mathcal{O}(k_{ic}^2)$.

Our second result shows sparsity implies smaller upper bound of distance between the original sample and the projected sample.

Lemma 8: Let x be any sample and $\mathbf{D}_z \subset \mathbf{D}(z << k)$ be the projection matrix. Let $\widetilde{x} = \mathbf{D}_z^T x$ be the projection of x. Assume $||x|| \leq r$, we have

$$|||x||^2 - ||\widetilde{x}||^2| \leq r^2 + \sum_{i=1}^{z} < d_i, x >^2 . \tag{16}$$

To prove the prediction loss of $f_{\underset{\sim}{z}}$, we extend the arguments in [33, Lemma 1] to f_z. Assume $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{D}$, then we have

Theorem 9: Let $Y_1 = \widetilde{\mathbf{X}} \beta^*_{\widetilde{\mathbf{X}}} + \epsilon_1$, $Y_2 = \mathbf{X} \beta^*_{\mathbf{X}} + \epsilon_2$, If $\lambda_1 \geq 2||\widetilde{\mathbf{X}}^T \epsilon_1||_\infty$, $\lambda_2 \geq 2||\mathbf{X}^T \epsilon_2||_\infty$, constants v_1 and v_2 satisfy the following two conditions:

(1) $0 < v_1 \leq \frac{\sqrt{s_1}||\widetilde{\mathbf{X}}\eta_1||_2}{\sqrt{n}||\eta_1 S_1||_1}$ for all $\eta_1 \in \mathbb{R}^k$ such that $||\eta_1 S_1^c||_1 \leq 3||\eta_1 S_1||_1$,

(2) $0 < v_2 \leq \frac{\sqrt{s_2}||\mathbf{X}\eta_2||_2}{\sqrt{n}||\eta_2 S_2||_1}$ for all $\eta_2 \in \mathbb{R}^p$ such that $||\eta_2 S_2^c||_1 \leq 3||\eta_2 S_2||_1$, then the prediction error of f_z satisfies

$$||\widetilde{\mathbf{X}} \widetilde{\beta}_{\widetilde{\mathbf{X}}}(\lambda_1) - \mathbf{X}\beta^*_{\mathbf{X}}||_2^2 \leq \frac{16 s_1 \lambda_1^2}{v_1^2 n} + \frac{16 s_2 \lambda_2^2}{v_2^2 n} + d(\beta^*_{\widetilde{\mathbf{X}}}, \widetilde{\beta}_{\mathbf{X}}(\lambda_2)), \tag{17}$$

where $d(\beta^*_{\widetilde{\mathbf{X}}}, \widetilde{\beta}_{\mathbf{X}}(\lambda_2)) := ||\widetilde{\mathbf{X}}\beta^*_{\widetilde{\mathbf{X}}} - \mathbf{X}\widetilde{\beta}_{\mathbf{X}}(\lambda_2)||_2^2$, $S_1 := supp(\beta^*_{\widetilde{\mathbf{X}}})$ is the index set of the non-zero entries of $\beta^*_{\widetilde{\mathbf{X}}}$, S_1^c is the complement of S_1; $S_2 := supp(\beta^*_{\mathbf{X}})$ is the index set of the non-zero entries of $\beta^*_{\mathbf{X}}$ and S_2^c is the complement of S_2.

This theorem implies model loss is related with different parameters. However, if we fix hyper-parameters λ_1, λ_2, sample size n, and sparsity levels s_1, s_2, then the important factor in the error bound is $\widetilde{\mathbf{X}}$, which is decided by dictionary $\mathbf{D}$.

VIII. Experiment

In this section, we evaluate the performance of z**-Sparse Fair Learner** under the proposed **DicPF** framework on three benchmark datasets. To encourage reproducibility, we make our data and code publicly available [34].

A. Datasets

We conduct experiments on three popular datasets, which are commonly used for evaluating algorithm fairness: the Community Crime data, the Credit Card data and the COMPAS data [34].

The Community Crime data contains 1,993 communities described by 101 features with community crime rate as its label. We treat the 'fraction of African-American residents' as the sensitive feature. And a community is 'minority' if the fraction is above 0.5 and 'majority' otherwise. The Credit Card data contains 20,000 users described by 23 features with default payment as its label. Similar to [17], we select education degree as the sensitive feature. The COMPAS data contains 16,000 records described by 15 features after removing the incomplete data. The class label is the risk of recidivism. Similar to [35], we treat race as the sensitive feature.

B. Experimental Settings

In the experiment, considering the sizes of the data sets, we set up three local agents $\{W_1, W_2, W_3\}$ and one modeler $\mathcal{M}$. The three agents holds their local data $\{X_1, X_2, X_3\}$ with private data $\{S_1, S_2, S_3\}$, respectively. $\mathcal{M}$ only holds the non-private data X and true labels. We randomly split each dataset into three parts and assign to $\{W_1, W_2, W_3\}$ for learning local dictionaries; At modeler $\mathcal{M}$, we choose 75% of the instances for training and use the rest for testing. We evaluate our learner for 50 random trials and report its averaged performance.

Baselines. We compare the proposed private fair learner with the following six non-private baselines and four semi-private fair learners:

(1) Non-private fair learners: Fair Ridge Regression (FRR) [18], Fair Kernel Regression (FKR) [36], Fair Logistic Regression (FGR) [37], two Fair PCAs (FPCA1 [17], FPCA2 [38]), and Fair Representation Learning (LFR) [11].

(2) Semi-private fair learners: Four distributed fair learners proposed in [5], includes Distributed Fair Ridge Regression(DFRR), Distributed Fair Kernel Regression (DFKRR), Distributed Fair Logistic Regression (DFGR), and Distributed Fair PCA (DFPCA).

Hyper-parameter Settings. For baselines, we use their default hyper-parameters (or grid-search from the default candidate values). For the proposed learner, similar to [5], we set ρ to 0.01, 0.1, and 0.25 on the three datasets, respectively. The maximal iteration in coordinate descent algorithm is $1,500$.

TABLE II: Parameter Settings for Privacy Inference (DL denotes Dictionary Learning)

Data	Inference Size	GA Parameters					DL Parameters	
		initial population size step size is 100	crossover rate step size is 0.1	mutation rate	selection	iteration	inaccuracy tolerance	iteration
Crime	all samples	$200 \sim 500$	$0.1 \sim 0.3$	0.01	rank	to converge	10^{-6}	1000
Credit	all samples	$200 \sim 500$	$0.1 \sim 0.3$	0.01				

For the hyperparameter λ, we grid-search its optimal value in $1 \sim 10^3$ and the step size is 50.

Evaluation metrics. We use *statistical parity (SP)* [31] to measure model fairness, which is defined in Equation (12). And we use *classifier error* to measure model error. A smaller SP implies a fairer model, while a smaller classifier error implies a more accurate model.

Privacy Loss (PL) is used to measure the model privacy-preserving performance, which is defined in Equation (10). A smaller PL value indicates better privacy-preserving capability.

C. *Results and Discussions*

The experimental results on the three datasets are presented in Table III, IV, and V, respectively. We will discuss the results from two aspects: (i) Non-private vs. Private and (ii) Semi-private vs. Private.

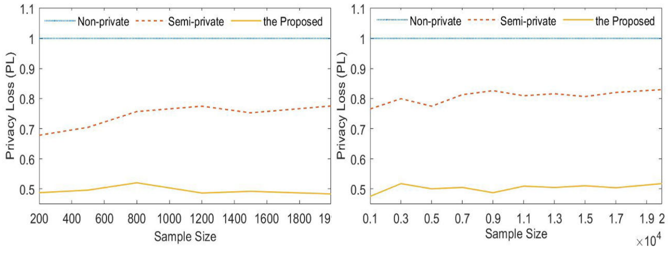
The first observation is that the fairness performance of z**-Sparse Fair Learner** outperforms most non-private and semi-private baselines across all three datasets. Take the non-private fair learner FGR as an example, the proposed method achieves much lower SP than FGR (0.0898 vs. 0.0189; 0.0779 vs. 0.0076; and 0.0408 vs. 0.0067). Similar observations can be found when we compare z**-Sparse Fair Learner** with the semi-private learners. Take DFRR as an example, z**-Sparse Fair Learner** decreases SP from 0.0466 to 0.0189 on the Crime data, from 0.0118 to 0.0076 on the Credit data, and from 0.0078 to 0.0067 on the COMPAS data. The two observations imply that the proposed private learner is very effective for learning a fair model.

The second observation is that the accuracy of z**-Sparse Fair Learner** is also better compared with some baselines. For example, on the Crime data, compared with FPCA1, FPCA2, and DFPCA, our algorithm not only achieves a much lower SP, but also achieves a lower classifier error. On the Credit data, similar observations can be found for FGR. This implies our method achieves a more efficient trade-off between fairness and accuracy than some compared fair learners.

The third observation is that we notice the superior fairness of z**-Sparse Fair Learner** is not achieved without any cost. It has a slightly higher classification error rate than several baselines. Nevertheless, we argue the loss of model accuracy is relatively small compared with the increase in fairness. For example, on the Credit data, we choose DFRR for comparison. z**-Sparse Fair Learner** lowers prediction disparity by 35.59%= (0.0118-0.0076)/0.0118 but only increases classification error by 4.60% = (0.2388-0.2283)/0.2283.

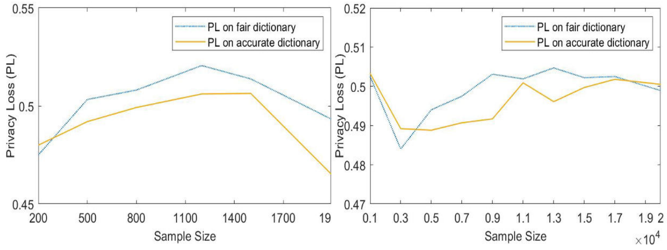
D. *Privacy Inference Analysis*

We examine the model privacy-preserving performance on the Crime and Credit data sets with different sample sizes. The inference model is as Equation (11) shows and we solve it by using Genetic Algorithm (GA). On the two data sets, we infer the sensitive features (minority/majority; education degree) respectively. The parameter settings on each data set are as Table II shows.



(a) (b)

Fig. 2: Privacy loss comparison between the proposed and non-private models / semi-private models in [5] on the (a) Crime data set and (b) Credit Card data set.



(a) (b)

Fig. 3: Privacy loss comparison on fair dictionary and accurate dictionary ((a) Crime data set and (b) Credit Card data set).

We compare the proposed private model with the non-private and semi-private models [5]. For semi-private method [5], we set the number of random hypotheses $m = 100$, because if m is too big, integer programming (IP) may not return any feasible solutions in finite iterations. For the proposed model, because the GA solutions depend on its search space, we report results averaged over 20 random trials.

As Figure 2 shows, the proposed model decreases privacy loss significantly compared with non-private and semi-private models. For non-private model, the sensitive demographic data is available to the adversary, therefore, the privacy loss is 1. For semi-private model, the adversary can obtain an optimal solution via multiple constraints, the privacy loss is also high (> 0.6) when the sensitive data is binary. However, the privacy loss of the proposed method is smaller than 0.6 on two data sets, this observation implies that the adversary can not infer sensitive demographic data accurately by using equation (11). The main reason is that the adversary does not know the true SP value for accurate inference.

TABLE III: Classification Performance on the Crime Data

	Method	Statistical Parity	Classifier Error
Non-private	FRR	.3062±.0452	.1102±.0128
	FKR	.0968±.0722	.1208±.0054
	FGR	.0898±.0971	.1166±.0189
	FPCA1	.0859±.0479	.1731±.0089
	FPCA2	.0755±.0293	.1476±.0122
	LFR	.0738±.0377	.1264±.0068
Semi-private	DFRR	.0466±.0117	.1064±.0092
	DFKRR	.0695±.0181	.1216±.0143
	DFGR	.0650±.0198	.1202±.0690
	DFPCA	.0289±.0502	.1351±.0111
Private	***z*-Sparse Fair Learner**	**.0189 ± .0151**	.1299 ±.0614

TABLE IV: Classification Performance on the Credit Card Data

	Method	Statistical Parity	Classifier Error
Non-private	FRR	.0994±.0016	.2340±.0058
	FKR	.0079±.0011	.2001±.0054
	FGR	.0779±.0571	.2412±.0469
	FPCA1	.1716±.0149	.4025±.0382
	FPCA2	.0981±.0164	.3224±.0045
	LFR	.0288±.0132	.2835±.0051
Semi-private	DFRR	.0118±.0006	.2283±.0062
	DFKRR	.0085±.0015	.1823±.0092
	DFGR	.0494±.0601	.2244±.0382
	DFPCA	.0344±.0061	.2304±.0041
Private	***z*-Sparse Fair Learner**	**.0076± .0063**	.2388 ± .0206

TABLE V: Classification Performance on the COMPAS Data

	Method	Statistical Parity	Classifier Error
Non-private	FRR	.0515±.0042	.2276±.0040
	FKR	.0041±.0013	.2190±.0089
	FGR	.0408±.0162	.2428±.0917
	FPCA1	.2806±.0182	.3204±.1032
	FPCA2	.1719±.0317	.2390±.0278
	LFR	.0182±.0211	.2496±.0044
Semi-private	DFRR	.0078±.0041	.2302±.0045
	DFKRR	.0034±.0015	.2152±.0093
	DFGR	.0374±.0645	.2617±.0509
	DFPCA	.0081±.0046	.2279±.0046
Private	***z*-Sparse Fair Learner**	**.0067±.0049**	.2337±.0271

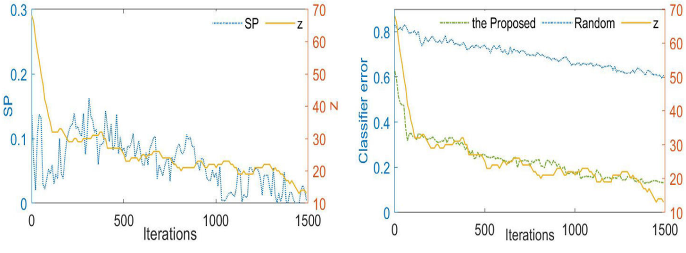
To study the privacy loss on a fair dictionary and accurate dictionary separately, we test two cases: (i) Privacy loss on a fair dictionary; (ii) Privacy loss on an accurate dictionary. The size of two dictionaries is $k = 200$. The results are shown in Figure 3. We see the privacy loss on the fair dictionary is slightly higher in most cases. This means the inference of adversary may be more accurate with a fair dictionary based on Equation (11).

E. Model Sensitivity Analysis

We examine the performance of z**-Sparse Fair Learner** on the Crime data with different configurations and report results on testing samples averaged over 20 random trials.

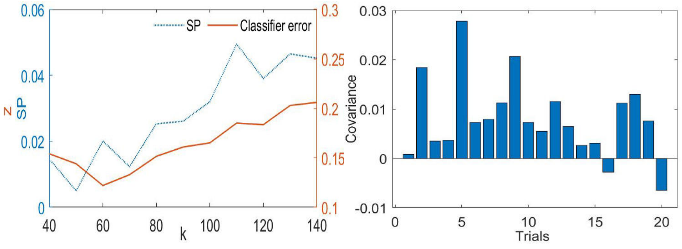
First, we examine the model fairness performance with different numbers of the selected atoms (z). There are three observations from Figure 4(a): (i) The curve of SP shows the worst fairness performance is smaller than 0.166, since the aggregated dictionary $\mathbf{D}$ is a fair space. (ii) As z decreases, SP decreases. This implies the model is fairer with smaller z; (iii) The decrease rate of SP is slow because the reduction rate is $\mathcal{O}(\sqrt{z})$. These observations are consistent with Theorem 6.

Then, we examine the model accuracy performance with different numbers of the selected atoms (z). Figure 4(b) shows if the dictionary is randomly generated, the accuracy of the sparse model is worse (convergence error ≥ 0.6). This is because the small size of a random fair space can not ensure model accuracy. However, the accuracy of the proposed method can converge to $[0.1, 0.2]$, which means sparse representation theory guarantees the model accuracy.



(a) (b)

Fig. 4: (a) SP versus z (b) Classifier error versus z.



(a) (b)

Fig. 5: (a) Classifier error and SP versus k (b) $cov(f_z(x), s)$ of 20 random trials on the Crime data.

Next, we examine the performance of the proposed model with different sizes of the aggregated dictionary $\mathbf{D}$ (k). Fix $\lambda = 50$, we increase k from 40 to 140. The performance is shown in Figure 5(a). We observe that as k increases, the SP increases. This means sparse fair atom selection promises model fairness, which is consistent with the implication of Theorem 6. However, the classifier error slightly increases, this observation also implies model accuracy is ensured by sparse representation theory.

Finally, we examine the PQD/NQD assumption [39], [40] in Theorem 6. Figure 5(b) shows $\text{cov}(f_z(x), s)$ of z**-Sparse Fair Learner** over 20 random trials on the Crime data. We observe that the covariance is positive in most cases, which implies $f_z(x)$ and s are PQD/NQD.

IX. Conclusion and Future Work

In this paper, we addressed an important and challenging problem of private fair machine learning. We proposed **DicPF**, a privacy-preserving fair learning framework that doesn't need to collect sensitive demographic data, and demonstrated its superior privacy-preserving performance under inference attack. We proposed a z**-Sparse Fair Learner** under this framework, then theoretically analyzed z**-Sparse Fair Learner** and proved upper bounds on its model fairness and accuracy. The experimental results on three real-world data sets demonstrated the effectiveness of the proposed private learner.

In this study, we introduce sparsity to improve model fairness, however, model accuracy is slightly lower than some compared models. Overcoming this limitation is our future work.

X. Appendix

A. Proof of Lemma 5

We will prove that f_z is fair with a ρ-fair $\mathbf{D}$. Let $\vec{\beta}_z$ denote the vector of non-zero coefficients, by the linear property of covariance,

$$\begin{aligned}\text{cov}[f_z(x), s] &= \text{cov}\left[\sum\nolimits_{t=1}^{k} \beta_t d_t(x),\ s\right] \\ &= \text{cov}[\sum\nolimits_{\beta_t \neq 0} \beta_t d_t(x) + \sum\nolimits_{\beta_t = 0} 0 * d_t(x), s] \\ &= \sum_{\beta_t \neq 0} \beta_t \text{cov}[d_t(x), s] \\ &\leq ||\vec{\beta}_z|| \cdot ||\, c\vec{o}v[d_z^{(s)},\ s]\,||,\end{aligned} \tag{18}$$

where $c\vec{o}v[d_z^{(s)}, s] = [\,\text{cov}[d_1(x), s], \ldots, \text{cov}[d_z(x), s]\,]$ and the last inequality is by the Cauchy–Schwarz inequality.

$$||c\vec{o}v[d_z^{(s)}, s]||^2 = \sum_{t=1}^{z} \text{cov}[d_t(x), s]^2 \leq \sum_{t=1}^{z} \rho^2 = z\rho^2, \tag{19}$$

Combining (18) and (19), Lemma 5 is proved.

B. Proof of Lemma 7

First, we quantify the reconstruction error of sample x_i based on the accurate dictionary $\mathbf{D}_{ia}$ as

$$g_{\mathbf{D}_{ia}}(x_i) = ||x_i - \mathbf{D}_{ia} z_i||^2. \tag{20}$$

Recall $||x_i|| \leq r$, $||\mathbf{D}_{ia} e_i|| \leq \gamma$ and $||z_i|| \leq 1$, then

$$\begin{aligned}||x_i - \mathbf{D}_{ia} z_i||^2 &\leq ||x_i||^2 + ||\mathbf{D}_{ia} z_i||^2 \\ &\leq r^2 + \sum_{r,j}^{k_{ic}} < z_{ir} \mathcal{D}_{ia} e_r, z_{ij} \mathbf{D}_{ia} e_j > \\ &\leq r^2 + \sum_{r,j}^{k_{ic}} ||z_{ir} \mathbf{D}_{ia} e_r|| ||z_{ij} \mathbf{D}_{ia} e_j|| \\ &\leq r^2 + k_{ic}^2 \gamma^2,\end{aligned} \tag{21}$$

where the third step is by the Cauchy-Schwarz Inequality.

C. Proof of Theorem 9

Let $\vec{a} = \widetilde{\mathbf{X}} \widetilde{\beta}_{\widetilde{\mathbf{X}}}(\lambda_1)$, $\vec{b} = \widetilde{\mathbf{X}} \beta^*_{\widetilde{\mathbf{X}}}$, $\vec{c} = \mathbf{X} \widetilde{\beta}_{\mathbf{X}}(\lambda_2)$, $\vec{d} = \mathbf{X} \beta^*_{\mathbf{X}}$, $A = \frac{16 s_1 \lambda_1^2}{v_1^2 n}$, $B = \frac{16 s_2 \lambda_2^2}{v_2^2 n}$, then we need to show

$$||\vec{a} - \vec{d}||_2^2 \leq A + B + ||\vec{b} - \vec{c}||_2^2, \tag{22}$$

with the two known conditions: (i) $||\vec{a} - \vec{b}||_2^2 \leq A$ and (ii) $||\vec{c} - \vec{d}||_2^2 \leq B$.

We relax this problem from n dimensions to two dimensions. Define point $a := (x_1, y_1)$, $b := (x_2, y_2)$, $c := (x_3, y_3)$, $d := (x_4, y_4)$, then we have

$$\begin{aligned}||a - b||_2^2 &= (x_1 - x_2)^2 + (y_1 - y_2)^2 \leq A \\ ||c - d||_2^2 &= (x_3 - x_4)^2 + (y_3 - y_4)^2 \leq B\end{aligned} \tag{23}$$

We see formula (23) are two circle areas as the Figure 6 shows, then we will prove that when a and d are two further points of intersections with line l_{bc}, the distance between a and d is maximal.

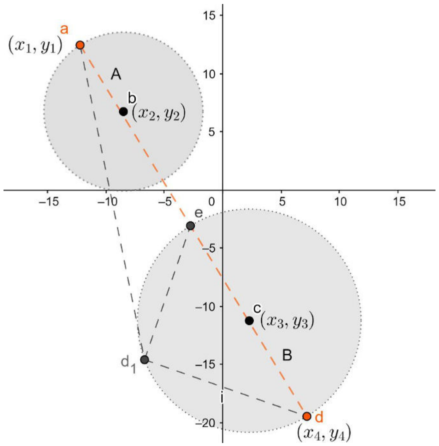


Fig. 6: $||a-b||_2^2 \leq A$ and $||c-d||_2^2 \leq B$

Let $h(x, y)$ be the distance between point x and point y. Assuming there is another point d_1 on the bottom circle edge such that $h(a, d_1) > h(a, d)$, then we will show such point d_1 does not exist. By triangle inequality, we have $h(a, d_1) \leq h(a, e) + h(d_1, e)$ (①). By Pythagorean Theorem, we have $h^2(d_1, e) + h^2(d_1, d) = h^2(d, e)$ (②), then we obtain $h(d_1, e) < h(d, e)$ (③). Plugging (③) back to (①), then $h(a, d_1) \leq h(a, e) + h(d, e) = h(a, d)$(④). This contradicts with our assumption. Therefore, there is no another point d_1 on the bottom circle edge such that $h(a, d_1) > h(a, d)$. Finally, we extend the dimension into n and replace the a, b, c, d with $\widetilde{\mathbf{X}}\widetilde{\beta}_{\widetilde{\mathbf{X}}}(\lambda_1)$,$\widetilde{\mathbf{X}}\beta^*_{\widetilde{\mathbf{X}}}$,$\mathbf{X}\widetilde{\beta}_{\mathbf{X}}(\lambda_2)$, $\mathbf{X}\beta^*_{\mathbf{X}}$ respectively, this theorem is proved.

References

[1] M. Hoffman, L. B. Kahn, and D. Li, "Discretion in hiring," *The Quarterly Journal of Economics*, vol. 133, no. 2, pp. 765–800, 2018.

[2] M. D. Cunningham and J. R. Sorensen, "Actuarial models for assessing prison violence risk: revisions and extensions of the risk assessment scale for prison (rasp)," *Assessment*, vol. 13, no. 3, pp. 253–265, 2006.

[3] H. Hu, *Links.*, https://nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf;https://oag.ca.gov/privacy/ccpa.

[4] N. Kilbertus, A. Gascon, M. Kusner, M. Veale, K. P. Gummadi, and A. Weller, "Blind justice: Fairness with encrypted sensitive attributes," in *ICML*, 2018.

[5] H. Hu, Y. Liu, Z. Wang, and C. Lan, "A distributed fair machine learning framework with private demographic data protection," in *ICDM*. IEEE, 2019, pp. 1102–1107.

[6] M. Gupta, A. Cotter, M. M. Fard, and S. Wang, "Proxy fairness," *arXiv preprint arXiv:1806.11212*, 2018.

[7] Y. Liu and C. Lan, "Active query of private demographic data for learning fair models," in *CIKM*, 2020, pp. 2129–2132.

[8] C. Li, P. Zhou, and e. a. Xiong, "Differentially private distributed online learning," *in TKDE*, vol. 30, no. 8, pp. 1440–1453, 2018.

[9] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: algorithms and applications," *IEEE access*, vol. 3, pp. 490–530, 2015.

[10] Y. Xu and e. a. j. v. p. y. p. Li, Zhengming, "A survey of dictionary learning algorithms for face recognition."

[11] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *ICML*, 2013.

[12] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *KDD*, 2015.

[13] A. Okray, H. Hu, and C. Lan, "Fair kernel regression via fair feature embedding in kernel space," in *ICTAI*. IEEE, 2019, pp. 1417–1421.

[14] L. Zhang and X. Wu, "Anti-discrimination learning: a causal modeling-based framework," *International Journal of Data Science and Analytics*, vol. 4, no. 1, pp. 1–16, 2017.

[15] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[16] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*, 2017, pp. 962–970.

[17] S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala, "The price of fair pca: One extra dimension," in *NIPS*, 2018.

[18] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling attribute effect in linear regression," in *ICDM*, 2013.

[19] B. Fish, J. Kun, and Á. D. Lelkes, "A confidence-based approach for balancing fairness and accuracy," in *SDM*, 2016, pp. 144–152.

[20] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *NIPs*, 2016, pp. 3315–3323.

[21] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller, "On fairness, diversity and randomness in algorithmic decision making," *arXiv preprint arXiv:1706.10208*, 2017.

[22] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[23] M. Veale and R. Binns, "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data," *Big Data & Society*, vol. 4, no. 2, p. 2053951717743530, 2017.

[24] I. Žliobaitė and B. Custers, "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models," *Artificial Intelligence and Law*, vol. 24, no. 2, pp. 183–201, 2016.

[25] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *CoRR*, 2018.

[26] e. a. Luong, Binh Thanh, "k-nn as an implementation of situation testing for discrimination discovery and prevention," in *KDD*, 2011.

[27] T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *ICML*, 2018.

[28] H. Hu and C. Lan, "Inference attack and defense on the distributed private fair machine learning framework."

[29] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. H. Chi, "Fairness without demographics through adversarially reweighted learning," *arXiv preprint arXiv:2006.13114*, 2020.

[30] S. J. Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.

[31] D. McNamara, C. S. Ong, and R. C. Williamson, "Provably fair representations," *CoRR*, 2017.

[32] S. Liu, G. Song, and W. Huang, "Real-time transportation prediction correction using reconstruction error in deep learning," *TKDD*, vol. 14, no. 2, pp. 1–20, 2020.

[33] J. Bien, I. Gaynanova, J. Lederer, and C. L. Müller, "Prediction error bounds for linear regression with the trex," *Test*, vol. 28, no. 2, pp. 451–474, 2019.

[34] H. Hu, https://github.com/HuiHu1.

[35] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[36] A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls, "Fair kernel learning," in *ECMLPKDD*, 2017.

[37] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *ECMLPKDD*, 2012.

[38] M. Olfat and A. Aswani, "Convex formulations for fair principal component analysis," *CoRR*, 2018.

[39] Z. Cai and G. G. Roussas, "Efficient estimation of a distribution function under quadrant dependence," *Scandinavian Journal of Statistics*, vol. 25, no. 1, pp. 211–224, 1998.

[40] M. Denuit and O. Scaillet, "Nonparametric tests for positive quadrant dependence," *Journal of Financial Econometrics*, 2004.