A Distributed Fair Machine Learning Framework with Private Demographic Data Protection

Hui Hu, Yijun Liu, Zhen Wang, Chao Lan
Department of Computer Science, University of Wyoming, WY, USA
Email: {hhu1, yliu20, zwang10, clan}@uwyo.edu

***Abstract*—Fair machine learning has become a significant research topic with broad societal impact. However, most fair learning methods require direct access to personal demographic data, which is increasingly restricted to use for protecting user privacy (e.g. by the EU General Data Protection Regulation).**

In this paper, we propose a distributed fair learning framework for protecting the privacy of demographic data. We assume this data is privately held by a third party, which can communicate with the data center (responsible for model development) without revealing the demographic information. We propose a principled approach to design fair learning methods under this framework, exemplify four methods and show they consistently outperform their existing counterparts in both fairness and accuracy across two real-world data sets. We theoretically analyze the framework, and prove it can learn models with high fairness or high accuracy, with their trade-offs balanced by a threshold variable.

I. Introduction

It is reported machine learning models are giving unfair predictions on minority people when applied to assist consequential decision makings: they are biased against black defendants in recidivism prediction [3], female applicants in job hiring [1] and female employees in facial verification [23]. How to learn fair prediction model has become a pressing problem for government [20], industry [13], [33] and academia [6], [10]; many solutions are developed, from label processing [25], [37], feature processing [14], [36], to model regularization [12], [35] and model post-processing [15], [18].

We note that most fair learning methods require direct access to individuals' demographic data, e.g., they need race data to mitigate racial bias. However, such data are increasingly restricted to use for privacy protection. In 2018, Europe launches a General Data Protection Regulation (GDPR), which prohibits 'processing of personal data revealing racial or ethnic original' and allows users to request 'erasure of personal data' from the data controller. Besides, the privacy community has been hiding sensitive personal data from analysis [2], [28].

We thus see fairness and privacy are running in a dilemma, i.e., most fair learners need access to demographic data while these data are restricted to use for privacy protection. Debates are arising [34], [38]: should law permit the use of private demographic data for the sake of fair learning? is it technically necessary to have direct access to such data? Very few scientific studies are done to address these questions.

In this paper, we propose a distributed fair machine learning framework that does not require direct access to demographic data. Our key insight is as follows: we assume there is a third party that privately holds demographic data of the individuals, to learn fair models, data center first constructs a random but fair hypothesis space through private communication with the third party; then it learns an accurate hypothesis in this space. Our rational is that model fairness is ensured as the hypothesis space is fair, and model accuracy is promised by random projection theory [4], [16]. In this paper, we exemplify how to redesign four existing fair learning methods: fair ridge regression [7], fair kernel regression [30], fair logistic regression [21] and fair PCA [29], [32]. We show the redesigned methods consistently outperform their counterparts in both fairness and accuracy across two real-world data sets.

We prove theoretical properties of the proposed distributed and private fair learning framework. Under proper conditions, we prove the learned model is both fair and accurate, and their trade-off is indirectly controlled by a threshold. Our result also implies one can learn a fair model from a population with balanced demographic distribution. For all the proofs and more empirical studies, please see our extended paper on arXiv.

II. Related Work

A. Fairness Measure

Several fairness notions have been proposed in the literature, such as statistical disparity [14], equal odds [18], individual fairness [12], causal fairness [24] and envy-free fairness [5]. We focus on statistical disparity, since it is most common and perhaps most refutable.

In this paper, we propose to measure model fairness using covariance between prediction and demographic variable, as we find it extremely easy to use while giving very efficient accuracy-fairness trade-off. Similar measures have been used in the literature, such as mutual information [21], correlation [30] or independence [36] between these two variables. But none of them provide theoretical analysis on the used measure. In this paper, we theoretically analyze the covariance measure.

B. Fair Learning with Restricted Access to Demographic Data

Several lines of studies are related to the restricted access of demographic data, but do not directly address the problem.

A traditional fair learning method is to simply remove demographic feature from the model, However, this approach does not guarantee fairness due to the redlining effect [9]. Some studies use demographic data in other ways, such as, [25] uses k-NN to detect unfair labels; they do not use

2374-8486/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDM.2019.00131

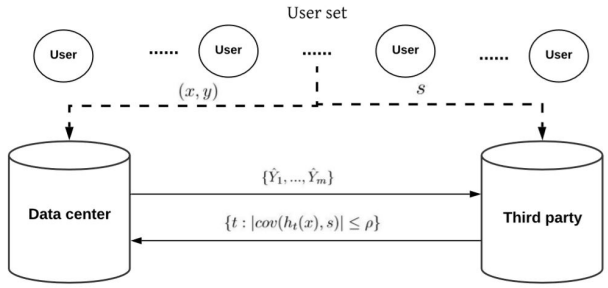


Fig. 1: A Distributed and Private Fair Learning Framework

demographic data to measure instance similarity, but still use it to measure label disparity in neighborhoods.

Specific discussions on the restricted use of demographic data appears in [34], [38]; but there lacks scientific investigations or solutions. Recently, Kilbertus et al [22] propose to encrypt demographic data before learning. This is a promising solution, but encryption also comes with extra cost of time and protocols. Our framework seeks another direction based on random projection; it is cheaper and easier to implement. Hashimoto et al [19] propose a fair learning method which automatically infers group membership and minimizes disparity across it; this method is also promising as it requires no access to demographic data. However, that study focuses on a less common fairness notion named distributive justice and on-line setting. Comparatively, we focus on a most common fairness measure named disparity and off-line setting. Besides, we hypothesize that having restricted access to demographic data would give fairer models than having no access at all.

III. Notations

In this section, we introduce the basic notations that will be used throughout the paper. More will be introduced later.

We will describe a random individual 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 label. For example, when studying gender bias in hiring, s will be an applicant's gender, x is the non-sensitive feature vector (e.g. education, working hours) and y indicates if the applicant is hired or not. We will index observed individuals by subscript, e.g., (x_i, s_i, y_i) is the i_{th} individual in a (training) sample set.

Let $f : \{x\} \to \{y\}$ be a prediction model, which does not take s as input but can use s for training.

IV. A Distributed Fair Learning Framework

In this section, we propose a distributed fair learning framework that can protect privacy of sensitive demographic data.

We assume the scenario in Fig 1, i.e., there is a data center and a third party, over which a training set $\{(x_i, s_i, y_i)\}_{i=1,...,n}$ is distributed. The center has $\{(x_i, y_i)\}$ and focuses on learning fair model f; the party has $\{s_i\}$ and can assist learning via private communications with the center that reveal no s.

To design fair learners, our key insight is to construct a *random but fair hypothesis space*. We will show such space can be constructed at the center via private communications, and any model learned in this space will be fair.

Algorithm 1 A Distributed Fair Learning Framework

Input: training set $\{(x_i, y_i)\}_{i=1,...,n}$, hypothesis set $\mathcal{H}$, number of generated hypotheses m, generator variance σ, fairness threshold ρ, data center (DC) and third party (TP).

Output: A prediction model f at DC.

1: DC randomly generate m hypotheses $h_1, \ldots, h_m \in \mathcal{H}$ with each parameter i.i.d. drawn from $\mathcal{N}(0, \sigma^2)$.
2: DC applies each h_t on $\{x_i\}$ to get a predicted label set $\hat{Y}_t = \{h_t(x_1), h_t(x_2), \ldots, h_t(x_n)\}$.
3: DC sends $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_m$ to TP.
4: TP estimates $\text{cov}(h_t(x), s)$ from $\hat{Y}_t$ and $\{s_i\}$ for each t, and returns t to DC if $|\text{cov}(h_t(x), s)| \leq \rho$.
5: DC receives a set of returned indices $r_1, r_2, \ldots, r_k$, and trains a prediction model f on $\{(x_i, y_i)\}$ assuming that

$$f = \alpha_1 h_{r_1} + \alpha_2 h_{r_2} + \ldots + \alpha_k h_{r_k}, \tag{1}$$

where $\vec{\alpha} = [\alpha_1, \ldots, \alpha_k]^T$ is unknown coefficient to learn.

Alg 1 elaborates our design strategy. Steps 1 to 4 construct a random and fair hypothesis space spanned by $h_{r_1}, \ldots, h_{r_k}$, and step 5 learns an accurate model in it. Note the communications do not reveal s, and thus protect its privacy.

In the sequel, we exemplify how to apply Alg 1 to redesign four existing non-private fair learners into private ones. (Other fair learners may be redesigned in similar ways.) We will write $X = [x_1, \ldots, x_n]^T$ as a sample matrix, $Y = [y_1, \ldots, y_n]^T$ as the associated label vector and $H = [h_{r_1}, \ldots, h_{r_k}]$ as a matrix of returned hypotheses. Since $\vec{\alpha} = [\alpha_1, \ldots, \alpha_k]^T$, we can write

$$f = \sum\nolimits_{t=1}^{k} \alpha_t h_{r_t} = H\vec{\alpha}. \tag{2}$$

A. Distributed Fair Ridge Regression (DFRR)

Calders et al [7] developed a fair ridge regression (FRR). It minimizes squared loss on training sample, while additionally penalizing prediction disparity across demographic groups. Let I_1, I_2 be the index sets of two demographic groups (e.g. female and male) respectively. Their objective function is

$$J_{FRR}(f) = \sum\nolimits_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \cdot \text{MD}(f),$$

where $\text{MD}(f) = \frac{1}{|I_1|} \sum_{i \in I_1} f(x_i) - \frac{1}{|I_2|} \sum_{i \in I_2} f(x_i)$ is the prediction disparity. We see $\min J(f)$ requires simultaneous access to (x, y) and s; thus this method cannot be directly applied in our private learning framework.

We propose a distributed fair ridge regression (DFRR) based on Algorithm 1. Our objective function is

$$\begin{aligned} J_{DFRR}(f) &= \sum\nolimits_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda ||f||^2, \\ &= \sum_{i=1}^{n} \left(\sum_{t=1}^{k} \alpha_t h_{r_t}(x_i) - y_i \right)^2 + \lambda \left\| \sum_{t=1}^{k} \alpha_t h_{r_t} \right\|^2. \end{aligned} \tag{3}$$

Minimizing the above objective for α_t's gives

$$\vec{\alpha} = (H^T X^T X H + \lambda I)^{-1} (H^T X^T Y). \tag{4}$$

B. Distributed Fair Kernel Ridge Regression (DFKRR)

Perez-Suay et al [30] developed fair kernel ridge regression (FKRR). It minimizes squared loss in RKHS while additionally penalizing the correlation between prediction and demographic feature. Its objective function is

$$J_{FKRR}(f) = \sum\nolimits_{i=1}^{n}(f(\phi(x_i)) - y_i)^2 + \lambda\Omega(f) + \mu I(f;s),$$

where $I(f;s) = \sum_{i=1}^{n}(\bar{f}(x_i)\cdot\bar{s}_i)$ is the correlation between prediction and demographic and $\bar{f}$ and $\bar{s}$ are centered variables. This method also needs simultaneous access to (x, y) and s.

We present a distributed fair kernel regression (DFKRR) method based on Algorithm 1. Our high-level objective is

$$J_{DFKRR}(f) = \sum_{i=1}^{n}(f(\phi(x_i)) - y_i)^2 + \lambda||f||^2. \quad (5)$$

Unlike the standard assumption that f is expressed by $\phi(x_i)$'s, we first assume f is expressed by h_{r_t}'s as in (1) and each h_{r_t} is linearly expressed by $\phi(x_i)$'s, i.e.,

$$h_{r_t} = \sum\nolimits_{i=1}^{n} c_{ti}\phi(x_i), \quad (6)$$

where c_{ti}'s are random coefficients associated with h_{r_t}. This is similar to the argument in [17].

Based on (5), we can generate a random hypothesis (and its predicted label set) by randomly generating a set of associated coefficients. Note the coefficients c_{t_i}'s are known and α_t's are unknown. Minimizing $J(f)$ gives

$$\vec{\alpha} = [C^T(K+\lambda I)^T(K+\lambda I)C]^{-1}C^T(K+\lambda I)^T Y. \quad (7)$$

where K is the Gram matrix and C is an n-by-k matrix with c_{ti} being its element at the i_{th} row and t_{th} column.

C. Distributed Fair Logistic Regression (DFGR)

Kamishima et al [21] developed a fair logistic regression (FGR). It maximizes the likelihood of label while additionally penalizing mutual information between model prediction and demographic feature. Its objective function is

$$J_{FGR}(f) = -\sum\nolimits_i \ln p(y_i \mid x_i, s_i, f) + \frac{\lambda}{2}||f||^2 + \mathrm{R}(f),$$

where $\mathrm{R}(f) = \mathrm{E}\, p(f(x), s)\ln\frac{p(f(x),s)}{p(f(x))p(s)}$ measures the mutual information and can be estimated from data. This method also requires simultaneous access to (x, y) and s.

We propose a distributed fair logistic regression (DFGR) based on Algorithm 1. Our high-level objective function is

$$J_{DFGR}(f) = -\sum_{i=1}^{n}\ln p(y_i \mid x_i, f) + \lambda||f||^2, \quad (8)$$

where $p(y_i|x_i, f)$ is constructed in the same way as logistic regression, with an additional assumption f has the form (1). Optimizing the objective by Newton's method, we have an update rule $\vec{\alpha} = \vec{\alpha} - (J^{''}(f))^{-1}(J^{'}(f))$, where

$$J^{'}(f) = H^T X^T(Y - \vec{p}) + 2\lambda H^T H\vec{\alpha}, \quad (9)$$

and

$$J^{''}(f) = H^T X^T M X H + 2\lambda H^T H, \quad (10)$$

with $\vec{p} = [p(f(x_1) = 1|x_1; f), \ldots, p(f(x_n) = 1|x_n; f)]^T$ and diagonal matrix M with $M_{ii} = p(f(x_i) = 1|x_i; f)\cdot p(f(x_i) = 0|x_i; f)$ – both are standard quantities in logistic regression.

D. Distributed Fair PCA

Samadi et al [32] developed a fair PCA which minimizes reconstruction error while equalizing this error across demographic groups. Let $X_1 \in \mathbb{R}^{n_1\times p}$ be the sample matrix of n_1 instances in one group, $X_2 \in \mathbb{R}^{n_2\times p}$ be the sample matrix of n_2 instances in another group, and $V \in \mathbb{R}^{p\times q}$ be the projection matrix. Their objective (to minimize) is

$$\max\left\{\frac{1}{n_1}loss(X_1, X_1VV^T),\ \frac{1}{n_2}loss(X_2, X_2VV^T)\right\}.$$

where $loss$ measures reconstruction error. Authors show the optimal V gives equal reconstruction errors across groups.

Matt Olfat et al [29] proposed another fair PCA method that minimizes prediction disparity in the projected space, i.e.,

$$\min_{w,V}\sup_t |p[w^TV^Tx \le t|s = 1] - p[w^TV^Tx \le t|s = 0]|,$$

where w is the prediction model and V is the project matrix.

Both methods need simultaneous access to (x, y) and s.

We propose a distributed fair PCA (DFPCA) method based on Algorithm 1. Let v be a projection vector. Our optimization probelm is the same as PCA, i.e.,

$$\max_v v^T\Sigma_x v, \quad \text{s.t.}||v|| = 1. \quad (11)$$

where Σ_x is the covariance matrix. Our additional assumption is that v is linearly expressed by fair random vectors h_{r_t}, i.e.,

$$v = \alpha_1 h_{r_1} + \ldots + \alpha_k h_{r_k} = H\vec{\alpha}. \quad (12)$$

Solving problem (11) for $\vec{\alpha}$ gives

$$H^T\Sigma_x H\vec{\alpha} = \lambda H^T H\vec{\alpha}, \quad (13)$$

which implies $\vec{\alpha}$ is the leading eigenvector.

V. Theoretical Analysis

Here we present the theoretical properties of Algorithm 1.

A. Preliminaries

Let (x, s) be a random instance. We say a hypothesis f is ρ-fair with respect to s if $|\mathrm{cov}[f(x), s]| \le \rho$. Note it means, in Algorithm 1, all returned hypotheses $h_{r_1}, \ldots, h_{r_k}$ are ρ-fair.

We will show ρ-fairness implies a popular fairness measure called *statistical parity* (SP) [27], defined as

$$\mathrm{SP}(f) = |p(f(x) = 1|s = 1) - p(f(x) = 1|s = 0)|. \quad (14)$$

To establish the implication, we will employ the following generalized covariance inequality [26, Theorem 2].

Lemma 1. *Let X, Y be two positively or negatively quadrant dependent random integers. Let $F_{X,Y}(x, y)$ be their joint CDF and $F_X(x)$, $F_Y(y)$ be their marginal CDF's respectively. Let*

$$cov_H(X, Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Delta F_{X,Y}(x, y)dxdy, \quad (15)$$

be their Hoeffding covariance, where

$$\Delta F_{X,Y}(x,y) = F_{X,Y}(x,y) - F_X(x)F_Y(y). \quad (16)$$

If $cov_H(X,Y)$ is bounded, then

$$\sup_{x,y}|\Delta F_{X,Y}(x,y)| \leq |cov_H(X,Y)|. \quad (17)$$

In the following, we will first present theoretical properties on model fairness and then on model error. Note that all results are presented in the context of Algorithm 1.

B. Theoretical Properties on Model Fairness

Our first result shows that ρ-fair implies statistical parity.

Lemma 2. *If $f(x)$ and s are positively or negatively quadrant dependent*[1]*, and if f is ρ-fair w.r.t. s, then SP$(f) \leq \rho/s_0 s_1$, where $s_0 = p(s=0)$ and $s_1 = p(s=1)$.*

Our second result suggests that a hypothesis spanned by fair hypotheses remains fair – this is the insight that motivates the study. More specifically, in (1), we show that f is ρ-fair because it is spanned by ρ-fair hypotheses $h_{r_1}, \ldots, h_{r_k}$.

Lemma 3. *In (1), f is $(\sqrt{k}\,||\vec{\alpha}||\rho)$-fair w.r.t. s.*

Combining the above result, we immediately have

Theorem 4. *In Algorithm 1, if $f(x)$ and s are positively or negatively quadrant dependent, then SP$(f) \leq \sqrt{k}||\vec{\alpha}||\rho/s_0 s_1$.*

This theorem implies one can obtain a fair model through several paths. First, we can choose a small threshold ρ, which will reduce prediction disparity at a rate of $O(\rho)$. Another way is to choose a small k but it does not seem very efficient as (i) it has a lower reduction rate $O(\sqrt{k})$ and (ii) it can be implied by choosing a small ρ (thus returning fewer hypotheses).

One may also choose a small $||\vec{\alpha}||$. In our proposed methods, this is done indirectly via regularizing $||f||$. In experiments, we observe this is more effective than directly regularizing $\vec{\alpha}$.

Finally, we see a model may be more fair if the demographic distribution is more balanced, i.e., the upper bound of SP(f) is minimized when $s_0 = s_1 = 0.5$. However, such distribution is typically formed by nature and cannot be easily modified.

Our following result gives more insight on the number of returned hypotheses k, and suggests it shall not be too small.

Lemma 5. *Let h be a random hypothesis. Then*

$$E[k] \geq m \cdot (1 - E[cov(h(x), s)]/\rho^2), \quad (18)$$

where both expectations are taken over the randomness of h, and the covariance is defined over the randomness of (x, s). Further, if h is linear and generated from $\mathcal{N}(0, \sigma^2 I)$, then

$$E[k] \geq m \cdot (1 - \sigma^2||\vec{cov}(x,s)||^2/\rho^2), \quad (19)$$

where $\vec{cov}(x,s)$=$\sum_{j=1}^{p} cov(x_j, s)$ and x_j is j_{th} entry of x.

Lemma 5 suggests that $E[k]$ will increase as ρ increases, at a rate of $O(1/\rho^2)$. In particular, when ρ approaches infinity, $E[k] \geq m$ which means all hypotheses will be returned. The lemma also suggests $E[k]$ will increase as σ^2 decreases.

[1]Quadrant dependence is a common assumption e.g., [11], [31]. Later we will show empirical evidence that our assumption holds in most cases.

C. Theoretical Properties on Model Generalization Error

To derive an error bound for the algorithm, our backbone technique is the random projection theory [16]. It states that data distance is likely to be preserved in a randomly projected space and thus a model's prediction error (dependent on such distance) is also likely to be preserved.

To apply the theory, we assume f, h are linear and interpret the returned hypotheses as basis of a randomly projected space, i.e., $h_{r_k}(x)$ is the k_{th} feature of x in the projected space.

We also assume Step 4 applies a *soft* threshold policy. Let h_* be a hypothesis satisfying cov$(h_*(x), s) = 0$. The soft policy will return t of any h_t with probability $\mathcal{N}(\vec{h}_*(x), \sigma_2^2 I)$, where $\vec{h}_*(x) = [h_*(x_1), \ldots, h_*(x_n)]^T$ and σ_2 is constant. As such, each returned hypothesis h_r in (1) is first drawn from a zero-mean Gaussian (Step 1) and then selected by a $\vec{h}_*$-mean Gaussian (Step 4). Therefore, we can say each h_r in (1) is generated from a Gaussian centered at $\vec{h}_*$. Without loss of generality, we assume this Gaussian has a unit variance.

Our first result extends the data distortion bound in [4] from zero-mean Gaussian to non-zero mean Gaussian.

Lemma 6. *Let x be any point and $H = [h_{r1}, \ldots, h_{r_k}]$ be a projection matrix with each projection vector h_{r_t} taken from a normal distribution $\mathcal{N}(h_*, I)$. Let $\tilde{x} = \frac{1}{\sqrt{k}}(H^T x)$ be the projection of x by H. We have for $0 \leq c < 1$,*

$$\Pr\{|\,||\tilde{x}||^2 - ||x||^2\,| \geq c||x||^2\} \leq g(x) \cdot e^{-\frac{c^2 k}{8}}, \quad (20)$$

where $g(x) = e^{(ck\langle h_*, x\rangle^2)/(4-2c)} + e^{-ck\langle h_*, x\rangle^2)/(2+2c)}$.

Compared to the original bound, our new bound has an additional term $g(x)$. It is smaller when $||h_*||$ is smaller; if $h_* = 0$, then $g(x) = 2$ and we recover the original bound.

Based on Lemma 6, we derive the following error bound.

Theorem 7. *Suppose Algorithm 1 adopts the soft threshold policy. Let $er(f)$ and $\hat{er}(f)$ be the expected and empirical error of f respectively. If f is linear and $||f|| = ||x|| = 1$, then with probability at least $1 - 4\delta$,*

$$er(h) \leq \hat{er}(h) + T + \frac{4}{n\delta}\sum_{i=1}^{n} g(x_i) e^{\frac{-k\langle f, x_i\rangle^2}{8(2+||\langle f, x_i\rangle||)^2}}, \quad (21)$$

where $T = 2\sqrt{[(k+1)\log(en/(k+1)) + \log 1/\delta]/n}$ and

$$g(x_i) = e^{(c_i k\langle h_*, x_i\rangle^2)/(4-2c_i)} + e^{-(c_i k\langle h_*, x_i\rangle^2)/(2+2c_i)} \quad (22)$$

with $c_i = |\langle f, x_i\rangle|/(2 + |\langle f, x_i\rangle|)$.

An important parameter in the error bound is k. To facilitate discussion, we can loosen the bound and have

Remark 8. *In Theorem 7, if*

$$\left(\langle h_*, x_i\rangle^2 - 1/4\right)(||\langle f, x_i\rangle|| - 2)^2 + 1 \leq 0, \quad (23)$$

then there exist positive constants c_1 and c_2 such that

$$er(h) \leq \hat{er}(h) + c_1 + O(e^{-c_2 k}). \quad (24)$$

We see the error bound decreases exponentially as k increases, suggesting one choose large k to get accurate models.

Note this is opposite to Theorem 4, which suggests choosing small k to get fair models. So we see a trade-off between accuracy and fairness is established (and controlled) via parameter k. In practice, we can adjust k by adjusting the threshold ρ.

VI. Experiment

A. Experiment Data

We experimented on two public data sets: the Community Crime data set and the COMPAS data set. The former contains 1993 communities described by 128 features; community crime rate is the label; we treated a community as minority if its fraction of African-American residents is above 0.5. The latter contains 18317 records described by 40 features; risk of recidivism is the label. We selected 16000 records and 15 non-empty numerical features. Similar to [8], we treated race as the sensitive feature.

B. Experiment Design

On each data set, we randomly chose 75% instances for training and the rest for testing. We evaluated each method for 50 random trials and reported its average results.

We compared each proposed distributed fair learner with its existing non-distributed counterpart, i.e., DFRR with fair ridge regression (FRR) [7], DFKRR with fair kernel regression (FKRR) [30]; DFGR with fair logistic regression (FGR) [21] and DFPCA with two fair PCAs (FPCA) [29], [32]. We also compared with a popular fair learner LFP [36].

We used five evaluation metrics: statistical parity (SP) [27], normed disparate (ND) [27], classifier error, error parity and error disparate. Let $er(f|s=1)$, $er(f|s=0)$ be the classifier errors in two demographic groups respectively. We define

$$\text{Error Parity}(f) = |er(f|s=1) - er(f|s=0)|. \tag{25}$$

and error disparate as

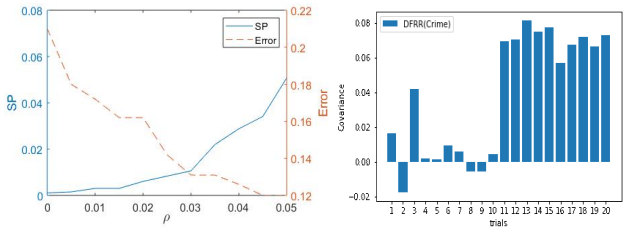
$$\text{Error Disparte}(f) = \left| \frac{er(f|s=1)}{er(f|s=0)} - 1 \right|. \tag{26}$$

C. Comparison Results and Discussions

Our experimental results on the two data sets are presented in Table I and II respectively. Since results in both tables are similar, our discussion will focus on Table I.

Our first observation is the proposed distributed fair learning methods consistently outperform their non-distributed counterparts. Take ridge regression as an example, DFRR not only achieves much lower SP than FRR (0.05 vs 0.31), but also simultaneously achieves lower classifier error (0.106 vs 0.110) and error parity (0.17 vs 0.23). Another example is PCA, where DFPCA achieves lower SP than FPCA's (0.03 vs 0.08), lower classifier error (0.14 vs 0.15) and lower error parity (0.15 vs 0.19). Similar comparisons can be observed for most methods on COMPAS data set. These observations imply that, for fair learning, our proposed ρ-fairness measure is more efficient than existing ones and our proposed learning framework is more effective.

Our second observation is that the performance gap between distributed and non-distributed methods is larger for linear models (ridge regression and PCA) and sometimes smaller for nonlinear ones (e.g. logistic and kernel). The former is further backed up by the desirable theoretical guarantees we proved for linear models. As to why our framework occasionally gives less performance improvement on non-linear base models, we do not have a principled hypothesis at the moment.



(a) (b)

Fig. 2: (a) Performance versus ρ and (b) $\text{cov}(f(x), s)$ of 20 random trials on the Community Crime data set

Finally, we observe existing Fair PCA methods do not lead to fair classification results (high SP and error). Our proposed distributed fair PCA significantly reduces SP and error, making itself a competitive method for classification tasks.

D. Sensitivity Analysis

Fig 2 (a) shows classifier error and SP of fair logistic regression versus ρ. We see that, as ρ decreases, error increases and SP decreases, This means the model is fairer but less accurate, which is consistent with the implications of Theorems 4 and 7. Fig 2(b) shows that, in most cases, the covariance is positive which implies $f(x)$ and s satisfy the PQD/PND assumption.

VII. Conclusion

In this paper, we propose a distributed fair machine learning framework for protecting the privacy of demographic data. We propose a simple and effective approach to design fair learning methods under this framework. We apply this approach to redesign four existing fair learning methods, and show our redesigns consistently outperform their counterparts on real-world data sets. We theoretically analyze the framework and prove its output model is both fair and accurate.

References

[1] "Amazon reportedly killed an ai recruitment system because it couldn't stop the tool from discriminating against women," in *Fortune*, 2018.

[2] R. Agrawal and R. Srikant, *Privacy-preserving data mining*. ACM, 2000, vol. 29, no. 2.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. and its's biased against blacks." in *ProPublica*, 2016.

[4] R. I. Arriaga and S. Vempala, "An algorithmic theory of learning: Robust concepts and random projection," *Machine Learning*, 2006.

[5] M.-F. Balcan, T. Dick, R. Noothigattu, and A. D. Procaccia, "Envy-free classification," 2018.

[6] S. Barocas and M. Hardt, "Fairness in machine learning," *NIPS Tutorial*, 2017.

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

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

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

TABLE I: Classification Performance on the Community Crime Data Set

Method	Statistical Parity	Normed Disparate	Classifier Error	Error Parity	Error Disparate
FRR [7]	.3062±.0452	.2457±.0128	.1102±.0128	.2260	.7321
DFRR	.0466±.0117	.1691±.1081	.1064±.0092	.1727	.6866
FKRR [30]	.0968±.0722	.1274±.0105	.1208±.0054	.1250	.2515
DFKRR	.0695±.0181	.1060±.0081	.1216±.0143	.1152	.2510
FGR [21]	.0898±.0971	.1154±.0308	.1166±.0189	.1424	.5723
DFGR	.0650±.0198	.1097±.0872	.1202±.0690	.1212	.5190
FPCA1 [32]	.0859±.0479	.3546±.0225	.1731±.0089	.1895	.5557
FPCA2 [29]	.0755±.0293	.3319±.0186	.1476±.0122	.1851	.6091
DFPCA	.0289±.0502	.2263±.0306	.1351±.0111	.1502	.6507
LFR [36]	.0738±.0377	.2240±.0194	.1264±.0068	.1319	.5431

TABLE II: Classification Performance on the COMPAS Data Set

Method	Statistical Parity	Normed Disparate	Classifier Error	Error Parity	Error Disparate
FRR [7]	.0515±.0042	.2361±.0414	.2276±.0040	.0317	.1081
DFRR	.0078±.0041	.1758±.0987	.2302±.0045	.0139	.0543
FKRR [30]	.0041±.0013	.1194±.0237	.2190±.0089	.0027	.0122
DFKRR	.0034±.0015	.1147±.0688	.2152±.0093	.0017	.0078
FGR [21]	.0408±.0162	.2842±.0319	.2428±.0917	.0222	.0865
DFGR	.0374±.0645	.1852±.0973	.2617±.0509	.0104	.0385
FPCA1 [32]	.2806±.0182	.3028±.0232	.3204±.1032	.0429	.1190
FPCA2 [29]	.1719±.0317	.2901±.1027	.2390±.0278	.0394	.1472
DFPCA	.0081±.0046	.2019±.1011	.2279±.0046	.0167	.0690
LFR [36]	.0182±.0211	.2201±.0318	.2496±.0044	.0044	.0190

[10] R. Courtland, "Bias detectives: the researchers striving to make algorithms fair," *Nature*, vol. 558, no. 7710, pp. 357–357, 2018.

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

[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *ACM Innovations in Theoretical Computer Science Conference*, 2012.

[13] J. V. E. Corbett, "Microsoft improves biased facial recognition technology," *Fortune*, 2018.

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

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

[16] A. Garg, S. Har-Peled, and D. Roth, "On generalization bounds, projection profile, and margin distribution," in *ICML*, 2002.

[17] Y. Grandvalet and S. Canu, "Adaptive scaling for feature selection in svms," in *NIPS*, 2002.

[18] M. Hardt, E. Price, N. Srebro *et al.*, "Equality of opportunity in supervised learning," in *NIPS*, 2016.

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

[20] W. House, "Preparing for the future of artificial intelligence," *Executive Office of the President*, 2016.

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

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

[23] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on Information Forensics and Security*, 2012.

[24] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *NIPS*, 2017.

[25] B. T. Luong, S. Ruggieri, and F. Turini, "k-nn as an implementation of situation testing for discrimination discovery and prevention," in *KDD*, 2011.

[26] P. Matula, "On some inequalities for positively and negatively dependent random variables with applications," *PUBLICATIONES MATHEMATICAE-DEBRECEN*, vol. 63, no. 4, pp. 511–522, 2003.

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

[28] N. Mohammed, R. Chen, B. Fung, and P. S. Yu, "Differentially private data release for data mining," in *KDD*, 2011.

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

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

[31] J. S. Racine, "Mixed data kernel copulas," *Empirical Economics*, 2015.

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

[33] T. Sloane, "Ibm helps eliminate bias in facial recognition training, but other faults may remain," *PaymentsJournal*, 2018.

[34] 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.

[35] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *AISTATS*, 2017.

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

[37] L. Zhang and X. Wu, "Anti-discrimination learning: a causal modeling-based framework," *Int. J. Data Science and Analytics*, 2017.

[38] 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.