Differentially Private and Fair Classification via Calibrated Functional Mechanism

Jiahao Ding,[1] **Xinyue Zhang,**[1] **Xiaohuan Li,**[2] **Junyi Wang,**[2] **Rong Yu,**[3] **Miao Pan**[1]
[1]University of Houston
[2]Guilin University of Electronic Technology
[3]Guangdong University of Technology
{jding7, xzhang67, mpan2}@uh.edu, {lxhguet, wangjy}@guet.edu.cn, yurong@gdut.edu.cn

Abstract

Machine learning is increasingly becoming a powerful tool to make decisions in a wide variety of applications, such as medical diagnosis and autonomous driving. Privacy concerns related to the training data and unfair behaviors of some decisions with regard to certain attributes (e.g., sex, race) are becoming more critical. Thus, constructing a fair machine learning model while simultaneously providing privacy protection becomes a challenging problem. In this paper, we focus on the design of classification model with fairness and differential privacy guarantees by jointly combining functional mechanism and decision boundary fairness. In order to enforce ϵ-differential privacy and fairness, we leverage the functional mechanism to add different amounts of Laplace noise regarding different attributes to the polynomial coefficients of the objective function in consideration of fairness constraint. We further propose an utility-enhancement scheme, called relaxed functional mechanism by adding Gaussian noise instead of Laplace noise, hence achieving (ϵ, δ)-differential privacy. Based on the relaxed functional mechanism, we can design (ϵ, δ)-differentially private and fair classification model. Moreover, our theoretical analysis and empirical results demonstrate that our two approaches achieve both fairness and differential privacy while preserving good utility and outperform the state-of-the-art algorithms.

Introduction

In this big data era, machine learning has been becoming a powerful technique for automated and data-driven decision making processes in various domains, such as spam filtering, credit ratings, housing allocation, and so on. However, as the success of machine learning mainly rely on a vast amount of individual data (e.g., financial transactions, tax payments), there are growing concerns about the potential for privacy leakage and unfairness in training and deploying machine learning algorithms (Fredrikson, Jha, and Ristenpart 2015; Datta, Tschantz, and Datta 2015). Thus, the problem of fairness and privacy in machine learning has attracted considerable attention.

Fairness-aware learning has received growing attentions in the machine learning field due to the social inequities and unfair behaviors observed in classification models. For example, a classification model of automated job hiring system is more likely to hire candidates from certain racial or gender groups (Giang 2018; Wachter-Boettcher 2018). Hence, substantial effort has centered on developing algorithmic methods for designing fair classification models and balancing the trade-off between accuracy and fairness, mainly including two groups: pre/post-processing methods (Dwork et al. 2012; Feldman et al. 2015; Hardt et al. 2016) and in-processing methods (Kamishima, Akaho, and Sakuma 2011; Zafar et al. 2017b). Pre/post-processing methods achieve fairness by directly changing values of the sensitive attributes or class labels in the training data. As pointed out in (Zafar et al. 2017b), pre/post-processing methods treat the learning algorithm as a black box, which can result in unpredictable loss of the classification utility. Thus, in-processing methods, which introduce fairness constraints or regularization terms to the objective function to remove the discriminatory effect of classifiers, have been shown a great success.

At the same time, differential privacy (Dwork and Roth 2014) has emerged as the de facto standard for measuring the privacy leakage associated with algorithms on sensitive databases, which has recently received considerable attentions by large-scale corporations such as Google (Erlingsson, Pihur, and Korolova 2014) and Microsoft (Ding, Kulkarni, and Yekhanin 2017), etc. Generally speaking, differential privacy ensures that there is no statistical difference to the output of a randomized algorithm whether a single individual opts in to, or out of its input. A large class of mechanisms has been proposed to ensure differential privacy. For instance, the Laplace mechanism is employed by introducing random noise drawn from the Laplace distribution to the output of queries such that the adversary will not be able to confirm a single individual is in the input with high confidence (Dwork et al. 2006b). To design private machine learning models, more complicated perturbation mechanisms have been proposed like objective perturbation (Chaudhuri, Monteleoni, and Sarwate 2011) and functional mechanism (Zhang et al. 2012), which inject random noise into the objective function rather than model parameters.

Thus, in this paper, we mainly focus on achieving classification models that simultaneously provide differential pri-

vacy and fairness. As pointed out in recent study (Xu, Yuan, and Wu 2019), achieving both requirements efficiently is quite challenging, due to the different aims of differential privacy and fairness. Differential privacy in a classification model focuses on the individual level, i.e., differential privacy guarantees that the model output is independent of whether any individual record presents or absents in the dataset, while fairness in a classification model focuses on the group level, i.e., fairness guarantees that the model predictions of the protected group (such as female group) are same to those of the unprotected group (such as male group). Lots of researches have emerged in achieving both privacy protection and fairness. Specifically, in (Dwork et al. 2012), Dwork et al. gave a new definition of fairness that is an extended definition of differential privacy. In (Hajian et al. 2015), Hajian et al. imposed fairness and k-anonymity via a pattern sanitization method. Moreover, Ekstrand et al. in (Ekstrand, Joshaghani, and Mehrpouyan 2018) put forward a set of questions about whether fairness are compatible with privacy. However, only Xu et al. in (Xu, Yuan, and Wu 2019) studied how to meet the requirements of both differential privacy and fairness in classification models by combining functional mechanism and decision boundary fairness together. Therefore, how to simultaneously meet the requirements of differential privacy and fairness in machine learning algorithms is under exploited.

In this paper, we propose **P**urely and **A**pproximately **D**ifferential private and **F**air **C**lassification algorithms, called PDFC and ADFC, respectively, by incorporating functional mechanism and decision boundary covariance, a novel measure of decision boundary fairness. As shown in (Kamiran and Calders 2012), due to the correlation between input features (attributes), the discrimination of classification still exists even if removing the protected attribute from the dataset before training. Hence, different from (Xu, Yuan, and Wu 2019), which adds same scale of noise in each attribute, in PDFC, we consider a calibrated functional mechanism, i.e., injecting different amounts of Laplace noise regarding different attributes to the polynomial coefficients of the constrained objective function to ensure ϵ-differential privacy and reduce effects of discrimination. To further improve the model accuracy, in ADFC, we propose a relaxed functional mechanism by inserting Gaussian noise instead of Laplace noise and leverage it to perturb coefficients of the polynomial representation of the constrained objective function to enforce (ϵ, δ)-differential privacy and fairness. Our salient contributions are listed as follows.

- We propose two approaches PDFC and ADFC to learn a logistic regression model with differential privacy and fairness guarantees by applying functional mechanism to a constrained objective function of logistic regression that decision boundary fairness constraint is treated as a penalty term and added to the original objective function.
- For PDFC, different magnitudes of Laplace noise regarding different attributes are added to the polynomial coefficients of the constrained objective function to enforce ϵ-differential privacy and fairness.
- For ADFC, we further improve the model accuracy by proposing the relaxed functional mechanism based on Extended Gaussian mechanism, and leverage it to introduce Gaussian noise with different scales to perturb objective function.
- Using real-world datasets, we show that the performance of PDFC and ADFC significantly outperforms the baseline algorithms while jointly providing differential privacy and fairness.

The rest of paper is organized as follows. We first give the problem statement and background in differential privacy and fairness. Next, we present our two approaches PDFC and ADFC to achieve DP and fair classification. Finally, we give the numerical experiments based on real-world datasets and draw conclusion remarks. Due to the space limit, we leave all the proofs in the supplemental materials.

Problem Statement

This paper considers a training dataset $\mathscr{D}$ that includes n tuples $t_1, t_2, \cdots, t_n$. We also denote each tuple $t_i = (\boldsymbol{x}_i, y_i)$ where the feature vector $\boldsymbol{x}_i$ contains d attributes, i.e., $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})$, and y_i is the corresponding label. Without loss of generality, we assume $\sqrt{\sum_{j=1}^{d} x_{ij}^2} \leq 1$ where $x_{ij} \geq 0$, and $y_i \in \{0, 1\}$ for binary classification tasks. The objective is to construct a binary classification model $\rho(\boldsymbol{x}, w)$ with model parameters $w = (w_1, w_2, \cdots, w_d)$ that taken $\boldsymbol{x}$ as input, can output the prediction $\hat{y}$, by minimizing the empirical loss on the training dataset $\mathscr{D}$ over the parameter space w of ρ.

In general, we have the following optimization problem.

$$w^* = \arg\min_{w} f(\mathscr{D}, w) = \arg\min_{w} \sum_{i=1}^{n} f(t_i, w) \quad (1)$$

where f is the loss function. In this paper, we consider logistic regression as the loss function, i.e., $f(\mathscr{D}, w) = \sum_{i=1}^{n} [\log(1+exp(\boldsymbol{x}_i^T w)) - y_i \boldsymbol{x}_i^T w]$. Thus, the classification model has the form $\rho(\boldsymbol{x}, w^*) = \frac{exp(\boldsymbol{x}^T w^*)}{1+exp(\boldsymbol{x}^T w^*)}$.

Although there is no need to share the dataset during the training procedure, the risk of information leakage still exists when we release the classification model parameter w^*. For example, the adversary may perform model inversion attack (Fredrikson, Jha, and Ristenpart 2015) over the release model w^* together with some background knowledge about the training dataset to infer sensitive information in the dataset.

Furthermore, if labels in the training dataset are associated with a protected attribute z_i (note that we denote $\boldsymbol{x}_i$ as unprotected attributes), like gender, the classifier may be biased, i.e., $P(\hat{y}_i = 1 | z_i = 0) \neq P(\hat{y}_i = 1 | z_i = 1)$, where we assume the protected attribute $z_i \in \{0, 1\}$. According to (Pedreshi, Ruggieri, and Turini 2008), even if the protected attribute is not used to build the classification model, this unfair behavior may happen when the protected attribute is correlated with other unprotected attributes.

Therefore, in this paper, our objective is to learn a binary classification model, which is able to guarantee differential privacy and fairness while preserving good model utility.

Background

In this section, we first introduce some background knowledge of differential privacy, which helps us to build private classification models. Then we present fairness definition, which helps us to enforce classification fairness.

Differential Privacy

Differential privacy is introduced to guarantee that the ability of an adversary to obtain additional information about any individual is independent of whether any individual record presents or absents in the dataset.

Definition 1 (**ϵ-Differential Privacy).** *A randomized Mechanism $\mathcal{A}$ is enforced by ϵ-differential privacy, if for any two neighboring datasets $\mathscr{D}, \mathscr{D}' \in \mathbb{D}$, i.e., differing at most one single data sample, and for any possible output s in the output space of $\mathcal{A}$, it holds that* $\Pr(\mathcal{A}(\mathscr{D}) = s) \leq e^{\epsilon} \Pr(\mathcal{A}(\mathscr{D}') = s)$.

The privacy parameter ϵ controls the strength of the privacy guarantee. A smaller value indicates a stronger privacy protection. Though differential privacy provides very strong guarantee, in some cases it may be too strong to have a good data utility. We then introduce a relaxation, (ϵ,δ)-differential privacy, that has been proposed in (Dwork et al. 2006a).

Definition 2 (**(ϵ,δ)-Differential Privacy).** *A randomized Mechanism $\mathcal{A}$ is enforced by (ϵ,δ)-differential privacy, if for any two neighboring datasets $\mathscr{D}, \mathscr{D}' \in \mathbb{D}$ differing at most one single data item, and for any possible output s in the output space of $\mathcal{A}$, it holds that $Pr(\mathcal{A}(\mathscr{D}) = s) \leq e^{\epsilon} Pr(\mathcal{A}(\mathscr{D}') = s) + \delta$.*

Laplace mechanism (Dwork and Roth 2014) and Extended Gaussian mechanism (Phan et al. 2019) are common techniques for achieving differential privacy, both of which add random noise calibrated to the sensitivity of the query function q.

Theorem 1 (**Laplace Mechanism).** *Given any function $q : \mathcal{X}^n \to \mathbb{R}^d$, the Laplace mechanism defined by*

$$\mathcal{M}_L(\mathscr{D}, q, \epsilon) = q(\mathscr{D}) + (\mathscr{Y}_1, \mathscr{Y}_2, \cdots, \mathscr{Y}_d)$$

preserves ϵ-differential privacy, where $\mathscr{Y}_i$ are i.i.d. random variables drawn from $Lap(\Delta_1 q/\epsilon)$ and l_1-sensitivity of the query q is $\Delta_1 q = \sup_{\mathscr{D},\mathscr{D}'} \|q(\mathscr{D}) - q(\mathscr{D}')\|_1$ taken over all neighboring datasets $\mathscr{D}$ and $\mathscr{D}'$.

Theorem 2 (**Extended Gaussian Mechanism).** *Given any function $q : \mathcal{X}^n \to \mathbb{R}^d$ and for any $\epsilon > 0$, $\delta \in (0, 1)$, the Extended Gaussian mechanism defined by*

$$\mathcal{M}_G(\mathscr{D}, q, \epsilon) = q(\mathscr{D}) + (\mathscr{Y}_1, \mathscr{Y}_2, \cdots, \mathscr{Y}_d)$$

preserves (ϵ, δ)-differential privacy, where $\mathscr{Y}_i$ are i.i.d drawn from a Gaussian distribution $\mathcal{N}(0, \sigma^2 I_d)$ with $\sigma \geq \frac{\sqrt{2}\Delta_{2q}}{2\epsilon}(\sqrt{\log(\sqrt{\frac{2}{\pi}}\frac{1}{\delta})} + \sqrt{\log(\sqrt{\frac{2}{\pi}}\frac{1}{\delta}) + \epsilon})$ and l_2-sensitivity of the query q is $\Delta_2 q = \sup_{\mathscr{D},\mathscr{D}'} \|q(\mathscr{D}) - q(\mathscr{D}')\|_2$ taken over all neighboring datasets $\mathscr{D}$ and $\mathscr{D}'$.

Functional Mechanism. Functional mechanism, introduced by (Zhang et al. 2012), as an extension of the Laplace mechanism is designed for regression analysis. To preserve ϵ-differential privacy, functional mechanism injects differentially private noise into the objective function $f(\mathscr{D}, w)$ and then publishs a noisy model parameter $\hat{w}$ derived from minimizing the perturbed objective function $\hat{f}(\mathscr{D}, w)$ rather than the original one. As a result of the objective function being a complex function of w, in functional mechanism, $f(\mathscr{D}, w)$ is represented in polynomial forms trough Taylor Expansion. The model parameter w is a vector consisting of several values $w_1, w_2, \cdots, w_d$. We denote $\phi(w)$ as a product of $w_1, w_2, \cdots, w_d$, namely, $\phi(w) = w_1^{c_1} w_2^{c_2} \cdots w_d^{c_d}$ for some $c_1, c_2, \cdots, c_d \in \mathbb{N}$. We also denote $\Phi_j (j \in \mathbb{N})$ as the set of all products of $w_1, w_2, \cdots, w_d$ with degree j, i.e., $\Phi_j = \{w_1^{c_1} w_2^{c_2} \cdots w_d^{c_d} | \sum_{l=1}^{d} c_l = j\}$.

According to the Stone-Weierstrass Theorem (Rudin and others 1964), any continuous and differentiable function can always be expressed as a polynomial form. Therefore, the objective function $f(\mathscr{D}, w)$ can be written as follows

$$f(\mathscr{D}, w) = \sum_{i=1}^{n} \sum_{j=0}^{J} \sum_{\phi \in \Phi_j} \lambda_{\phi t_i} \phi(w), \tag{2}$$

where $\lambda_{\phi t_i}$ represents the coefficient of $\phi(w)$ in polynomial.

To preserve ϵ-differential privacy, the objective function $f(\mathscr{D}, w)$ is perturbed by adding Laplace noise into the polynomial coefficients, i.e., $\lambda_\phi = \sum_{i=1}^{n} \lambda_{\phi t_i} + Lap(\Delta_1/\epsilon)$, where $\Delta_1 = 2 \max_t \sum_{j=1}^{J} \sum_{\phi \in \Phi_j} \|\lambda_{\phi t}\|_1$. And then the model parameter $\hat{w}$ is obtained by minimizing the noisy objective function $\hat{f}(\mathscr{D}, w)$. The sensitivity of logistic regression is given in the following lemma

Lemma 1 (**l_1-Sensitivity of Logistic Regression).** *Let $f(\mathscr{D}, w)$ and $f(\mathscr{D}', w)$ be the logistic regression on two neighboring datasets $\mathscr{D}$ and $\mathscr{D}'$, respectively, and denote their polynomial representations as $f(\mathscr{D}, w) = \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{\phi \in \Phi_j} \lambda_{\phi t_i} \phi(w)$ and $f(\mathscr{D}', w) = \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{\phi \in \Phi_j} \lambda_{\phi t_i'} \phi(w)$. Then, we have the following inequality*

$$\begin{aligned} \Delta_1 &= \sum_{j=1}^{2} \sum_{\phi \in \Phi_j} \| \sum_{t_i \in \mathscr{D}} \lambda_{\phi t_i} - \sum_{t_i' \in \mathscr{D}'} \lambda_{\phi t_i'} \|_1 \\ &\leq 2 \max_t \sum_{j=1}^{2} \sum_{\phi \in \Phi_j} \|\lambda_{\phi t}\|_1 \leq \frac{d^2}{4} + d, \end{aligned}$$

where t_i, t_i' or t is an arbitrary tuple.

Classification Fairness

The goal of classification fairness is to find a classifier that minimizes the empirical loss while guaranteeing certain fairness requirements. Many fairness definitions have been proposed for in the literature including mistreatment parity (Zafar et al. 2017a), demographic parity (Pedreshi, Ruggieri, and Turini 2008), etc.

Demographic parity, the most widely-used fairness definition in the classification fairness domain, requires the decision made by the classifier is not dependent on the protected attribute z, for instance, sex or race.

Definition 3. *(***Demographic Parity in a Classifier)** *Given a classification model $\hat{y} = \rho(\boldsymbol{x}, w)$ and a labeled dataset $\mathscr{D}$, the property of demographic parity in a classifier is defined by $\Pr(\hat{y} = 1|z = 1) = \Pr(\hat{y} = 1|z = 0)$ where $z \in \{0, 1\}$ is the protected attribute.*

Moreover, demographic parity is quantified in terms of the risk difference (RD) (Pedreschi, Ruggieri, and Turini 2012), i.e., the difference of the positive decision made in between the protected group and unprotected group. Thus, the risk difference produced by a classifier is defined as $RD = |\Pr(\hat{y} = 1|z = 1) - \Pr(\hat{y} = 1|z = 0)|$.

One of the in-processing methods, called decision boundary fairness (Zafar et al. 2017b), to ensure classification fairness is to find a model parameter w that minimizes the loss function $f(\mathscr{D}, w)$ under a fairness constraint. Thus, the fair classification problem is formulated as follows,

$$\begin{aligned} &minimize \quad f(\mathscr{D}, w) \\ &subject\ to \quad g(\mathscr{D}, w) \leq \tau, g(\mathscr{D}, w) \geq -\tau, \end{aligned} \tag{3}$$

where $g(\mathscr{D}, w)$ is a constraint term, and τ is the threshold. For instance, Zafar et al. (Zafar et al. 2017b) have proposed to adopt the decision boundary covariance to define the fairness constraint, i.e.,

$$\begin{aligned} g(\mathscr{D}, w) &= \mathbb{E}[(z - \bar{z})d(\boldsymbol{x}, w)] - \mathbb{E}[z - \bar{z}]d(\boldsymbol{x}, w) \\ &\propto \sum_{i=1}^{n}(z_i - \bar{z})d(\boldsymbol{x}_i, w), \end{aligned} \tag{4}$$

where $\{d(\boldsymbol{x}_i, w)\}_{i=1}^n$ is decision boundary, $\bar{z}$ is the average of the protected attribute and $\mathbb{E}[z - \bar{z}] = 0$. For logistic regression classification models, the decision boundary is defined by $\boldsymbol{x}^T w$. The decision boundary covariance (4) then reduces to $g(\mathscr{D}, w) = \sum_{i=1}^{n}(z_i - \bar{z})\boldsymbol{x}_i^T w$.

Differentially Private and Fair Classification

In this section, we first present our approach PDFC to achieve fair logistic regression with ϵ-differentially private guarantee. Then we propose a relaxed functional mechanism by injecting Gaussian noise instead of Laplace noise to provide (ϵ, δ)-differential privacy. By leveraging the relaxed functional mechanism, we will show that our second approach ADFC can jointly provide (ϵ, δ)-differential privacy and fairness.

Purely DP and Fair Classification

In order to meet the requirements of ϵ-differential privacy and fairness, motivated by (Xu, Yuan, and Wu 2019), we consider to combine the functional mechanism and decision boundary fairness. We first consider to transform the constrained optimization problem (3) into unconstrained problem by treating the fairness constraint as a penalty term, where the fairness constraints are shifted to the original objective function $f(\mathscr{D}, w)$. Then, we have the new objective function $\tilde{f}_{\mathscr{D}}(w)$ defined as $\tilde{f}(\mathscr{D}, w) = f(\mathscr{D}, w) + \alpha_1|g(\mathscr{D}, w) - \tau|$, where we consider α_1 as a hyperparameter to optimize the trade-off between model utility and fairness. For convenience of discussion, we set $\tau = 0$ and choose suitable values to make $\alpha_1 = 1$. Note that our theoretical results still hold if we choose other values of α_1 and τ. By equation (4), we have

$$\begin{aligned} \tilde{f}(\mathscr{D}, w) = &\sum_{i=1}^{n}[\log(1 + exp(\boldsymbol{x}_i^T w)) - y_i \boldsymbol{x}_i^T w] \\ &+ \left|\sum_{i=1}^{n}(z_i - \bar{z})\boldsymbol{x}_i^T w\right|. \end{aligned} \tag{5}$$

To apply functional mechanism, we first write the approximate objective function $\bar{f}(\mathscr{D}, w)$ based on (2) as follows.

$$\begin{aligned} \bar{f}(\mathscr{D}, w) &= \sum_{i=1}^{n}\sum_{j=0}^{2}\frac{f_1^{(j)}(0)}{j!}(\boldsymbol{x}_i^T w)^j - \left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i^T\right) w \\ &\quad + \left|\sum_{i=1}^{n}(z_i - \bar{z})\boldsymbol{x}_i^T w\right| \\ &= \sum_{i=1}^{n}\sum_{j=0}^{2}\sum_{\phi \in \Phi_j}\bar{\lambda}_{\phi t_i}\phi(w), \end{aligned} \tag{6}$$

where $\bar{\lambda}_{\phi t_i}$ denotes the coefficient of $\phi(w)$ in the polynomial of $\bar{f}(t_i, w)$ and $f_1(\cdot) = \log(1 + \exp(\cdot))$.

The attributes involving in the dataset may not be independent from each other, which means some unprotected attributes in $\boldsymbol{x}$ are quite correlated with the protected attribute z. For instance, the protected attribute, like gender, may be correlated with the attribute, marital status. Thus, to reduce the discrimination between the protected attribute z and the labels y, it is important to weaken the correlation between these most correlated attributes and protected attribute z. However, it is often impossible to determine the degree of relation between an unprotected attribute and the protected attribute. Therefore, we randomly select an unprotected attribute x_s and leverage functional mechanism to add noise with large scale to the corresponding polynomial coefficients of the monomials involving w_s. Interestingly, this approach not only helps to reduce the correlation between attributes x_s and z, but also improve the privacy on attribute x_s to prevent model inversion attacks, as shown in (Wang, Si, and Wu 2015).

The key steps of PDFC are outlined in Algorithm 1. We first set two different privacy budgets, ϵ_s and ϵ_n, for attribute x_s and the rest of attributes $\{\boldsymbol{x} \setminus x_s\}$. Before injecting noise to the coefficients, all coefficients ϕ should be separated into two groups Φ_s and Φ_n by considering whether w_s involves in the corresponding monomials (i.e., whether their the coefficients contain attribute x_s). We then add Laplace noises drawn from $Lap(\Delta_1/\epsilon_s)$ and $Lap(\Delta_1/\epsilon_n)$ to the coefficients of $\phi \in \Phi_s$ and $\phi \in \Phi_n$ respectively to reconstruct the differentially private objective function $\hat{f}(\mathscr{D}, w)$, where Δ_1 can be found in Lemma 2. Finally, the differentially private model parameter $\hat{w}$ is obtained by minimizing $\hat{f}(\mathscr{D}, w)$. Note that $\hat{w}$ also ensures classification fairness due to the objective function involving fairness constraint.

Lemma 2. *Let $\mathscr{D}$ and $\mathscr{D}'$ be any two neighboring datasets differing in at most one tuple. Let $\bar{f}(\mathscr{D}, w)$ and $\bar{f}(\mathscr{D}', w)$ be*

Algorithm 1 Purely DP and Fair Classification (PDFC)

Input: Dataset $\mathscr{D}$; The objective function $f(\mathscr{D}, w)$; The fairness constraint $g(\mathscr{D}, w)$; The privacy budget ϵ_s for unprotected attribute x_s; The privacy budget ϵ_n for other unprotected attributes $\{\boldsymbol{x} \setminus x_s\}$; l_1-sensitivity Δ_1.
Output: $\hat{w}$, ϵ.
Set the approximate function $\bar{f}(\mathscr{D}, w)$ by equation (6).
Set two sets $\Phi_s = \{\}$, $\Phi_n = \{\}$.
for $1 \leq j \leq 2$ **do**
  for each $\phi \in \Phi_j$ **do**
    if ϕ includes w_s for a particular attribute x_s **then**
      Put ϕ into Φ_s.
    else
      Put ϕ into Φ_n.
    end if
  end for
end for
for $1 \leq j \leq 2$ **do**
  for each $\phi \in \Phi_j$ **do**
    if $\phi \in \Phi_s$ **then**
      Set $\hat{\lambda}_\phi = \sum_{i=1}^{n} \bar{\lambda}_{\phi t_i} + Lap(\Delta_1/(\epsilon_s))$.
    else
      Set $\hat{\lambda}_\phi = \sum_{i=1}^{n} \bar{\lambda}_{\phi t_i} + Lap(\Delta_1/(\epsilon_n))$.
    end if
  end for
end for
Let $\hat{f}(\mathscr{D}, w) = \sum_{j=1}^{2} \sum_{\phi \in \Phi_j} \hat{\lambda}_\phi \phi(w)$.
Compute $\hat{w} = \arg\min_w \hat{f}(\mathscr{D}, w)$.
Compute $\epsilon = \epsilon_s/d + \epsilon_n(d-1)/d$.
return: $\hat{w}$, ϵ.

the approximate objective function on $\mathscr{D}$ and $\mathscr{D}'$, then we have the following inequality,

$$\Delta_1 = \sum_{j=1}^{2} \sum_{\phi \in \Phi_j} \| \sum_{i=1}^{n} \bar{\lambda}_{\phi t_i} - \sum_{i=1}^{n} \bar{\lambda}_{\phi t_i'} \|_1 \leq \frac{d^2}{4} + 3d.$$

The following theorem shows the privacy guarantee of PDFC.

Theorem 3. *The output model parameter $\hat{w}$ in PDFC (Algorithm 1) preserves ϵ-differential privacy, where $\epsilon = \frac{1}{d}\epsilon_s + \frac{d-1}{d}\epsilon_n$.*

Approximately DP and Fair Classification

We now focus on using the relaxed version of ϵ-differential privacy, i.e., (ϵ, δ)-differential privacy to further improve the utility of differentially private and fair logistic regression. Hence, in order to satisfy (ϵ, δ)-differential privacy, we propose the relaxed functional mechanism by making use of Extended Gaussian mechanism. As shown in Theorem 2, before applying Extended Gaussian mechanism, we first calculate the sensitivity of a query function, i.e., the objective function of logistic regression $f(\mathscr{D}, w) = \sum_{i=1}^{n} [\log(1 + exp(\boldsymbol{x}_i^T w)) - y_i \boldsymbol{x}_i^T w]$, given in the following lemma.

Algorithm 2 Relaxed Functional Mechanism

Input: Dataset $\mathscr{D}$; The objective function $f(\mathscr{D}, w) = \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{\phi \in \Phi_j} \lambda_{\phi t_i} \phi(w)$; The privacy parameters ϵ, δ.
Output: $\hat{w}$
Set Δ_2 according Lemma 3.
for $1 \leq j \leq J$ **do**
  for each $\phi \in \Phi_j$ **do**
    Set $\lambda_\phi = \sum_{i=1}^{n} \lambda_{\phi t_i} + \mathcal{N}(0, \sigma^2)$, where $\sigma = \frac{\sqrt{2}\Delta_2}{2\epsilon}(\sqrt{\log(\sqrt{\frac{2}{\pi}}\frac{1}{\delta})} + \sqrt{\log(\sqrt{\frac{2}{\pi}}\frac{1}{\delta}) + \epsilon})$.
  end for
end for
Let $\hat{f}(\mathscr{D}, w) = \sum_{j=1}^{J} \sum_{\phi \in \Phi_j} \lambda_\phi \phi(w)$.
Compute $\hat{w} = \arg\min_w \hat{f}(\mathscr{D}, w)$.
return: $\hat{w}$.

Lemma 3 (l_2**-Sensitivity of Logistic Regression).** *For polynomial representations of logistic regression, two $f(\mathscr{D}, w)$ and $f(\mathscr{D}', w)$ given in Lemma 1, we have the following inequality*

$$\Delta_2 = \|\mathscr{A}_1 - \mathscr{A}_2\|_2 \leq \sqrt{\frac{d^2}{16} + d},$$

where we denote $\mathscr{A}_1 = \{\sum_{i=1}^{n} \lambda_{\phi t_i}\}_{\phi \in \cup_{j=1}^{J} \Phi_j}$ and $\mathscr{A}_2 = \{\sum_{i=1}^{n} \lambda_{\phi t_i'}\}_{\phi \in \cup_{j=1}^{J} \Phi_j}$ as the set of polynomial coefficients of $f(\mathscr{D}, w)$ and $f(\mathscr{D}', w)$. And we denote t_i or t_i' as an arbitrary tuple.

We then perturb $f(\mathscr{D}, w)$ by injecting Gaussian noise drawn from $\mathcal{N}(0, \sigma^2)$ with $\sigma = \frac{\sqrt{2}\Delta_2}{2\epsilon}(\sqrt{\log(\sqrt{\frac{2}{\pi}}\frac{1}{\delta})} + \sqrt{\log(\sqrt{\frac{2}{\pi}}\frac{1}{\delta}) + \epsilon})$ into its polynomial coefficients, and obtain the differentially private model parameter $\hat{w}$ by minimizing the noisy function $\hat{f}(\mathscr{D}, w)$, as shown in Algorithm 2. Finally, we provide a privacy guarantee of proposed relaxed functional mechanism by the following theorem.

Theorem 4. *The relaxed functional mechanism in Algorithm 2 guarantees (ϵ, δ)-differential privacy.*

Our second approach called, ADFC, applies the relaxed functional mechanism into the objective function with decision boundary fairness constraint to enforce (ϵ, δ)-differential privacy and fairness. As shown in Algorithm 3, we first derive the polynomial representation $\bar{f}(\mathscr{D}, w)$ according to (6), and employ random Gaussian noise to perturb the objective function $\bar{f}(\mathscr{D}, w)$, i.e., injecting Gaussian noise into its polynomial coefficients. Furthermore, we also allocate differential privacy parameters, (ϵ_s, δ_s) and (ϵ_n, δ_n) for a particular unprotected attribute x_s and the rest of unprotected attributes $\{\boldsymbol{x} \setminus x_s\}$ to improve the privacy on attribute x_s and reduce the correlation between attributes x_s and z. Hence, we add random noise drawn from $\mathcal{N}(0, \sigma_s^2)$ to polynomial coefficients of $\phi \in \Phi_s$. For polynomial coefficients in Φ_n, we inject noise drawn from $\mathcal{N}(0, \sigma_n^2)$.

Algorithm 3 Approximately DP and Fair Classification (ADFC)

Input: Dataset $\mathscr{D}$; The objective function $f(\mathscr{D}, w)$; The fairness constraint $g(\mathscr{D}, w)$; The privacy parameters ϵ_s, δ_s for unprotected attribute x_s; The privacy parameters ϵ_n, δ_n for other unprotected attributes $\{\boldsymbol{x} \backslash x_s\}$.
Output: $\hat{w}$, ϵ and δ.
Set the approximate function $\bar{f}(\mathscr{D}, w)$ by equation (6).
Set two sets $\Phi_s = \{\}$, $\Phi_n = \{\}$.
for $1 \leq j \leq 2$ **do**
  for each $\phi \in \Phi_j$ **do**
    if ϕ includes w_s for a particular attribute x_s **then**
      Put ϕ into Φ_s.
    else
      Put ϕ into Φ_n.
    end if
  end for
end for
Set l_2-sensitivity Δ_2' by Lemma 4.
for $1 \leq j \leq 2$ **do**
  for each $\phi \in \Phi_j$ **do**
    if $\phi \in \Phi_s$ **then**
      Set $\hat{\lambda}_\phi = \sum_{i=1}^n \bar{\lambda}_{\phi t_i} + \mathcal{N}(0, \sigma_s^2)$, where $\sigma_s = \frac{\sqrt{2}\Delta_2'}{2\epsilon_s}(\sqrt{\log(\sqrt{\frac{2}{\pi}}\frac{1}{\delta_s})} + \sqrt{\log(\sqrt{\frac{2}{\pi}}\frac{1}{\delta_s}) + \epsilon_s})$.
    else
      Set $\hat{\lambda}_\phi = \sum_{i=1}^n \bar{\lambda}_{\phi t_i} + \mathcal{N}(0, \sigma_n^2)$, where $\sigma_n = \frac{\sqrt{2}\Delta_2'}{2\epsilon_n}(\sqrt{\log(\sqrt{\frac{2}{\pi}}\frac{1}{\delta_n})} + \sqrt{\log(\sqrt{\frac{2}{\pi}}\frac{1}{\delta_n}) + \epsilon_n})$.
    end if
  end for
end for
Let $\hat{f}(\mathscr{D}, w) = \sum_{j=1}^2 \sum_{\phi \in \Phi_j} \hat{\lambda}_\phi \phi(w)$.
Compute $\hat{w} = \arg\min_w \hat{f}(\mathscr{D}, w)$.
Compute $\epsilon = \frac{1}{d}\epsilon_s + \frac{d-1}{d}\epsilon_n$ and $\delta = 1-(1-\delta_s)(1-\delta_n)$.
return: $\hat{w}$, ϵ and δ.

Lemma 4. *Let $\mathscr{D}$ and $\mathscr{D}'$ be any two neighboring datasets differing in at most one tuple. Let $\bar{f}(\mathscr{D}, w)$ and $\bar{f}(\mathscr{D}', w)$ be the approximate objective function on $\mathscr{D}$ and $\mathscr{D}'$, then we have the following inequality,*

$$\Delta_2' = \|\mathscr{A}_1' - \mathscr{A}_2'\|_2 \leq \sqrt{\frac{d^2}{16} + 9d}.$$

where we denote $\mathscr{A}_1' = \left\{\sum_{i=1}^n \bar{\lambda}_{\phi t_i}\right\}_{\phi \in \cup_{j=1}^2 \Phi_j}$ and $\mathscr{A}_2' = \left\{\sum_{i=1}^n \bar{\lambda}_{\phi t_i'}\right\}_{\phi \in \cup_{j=1}^2 \Phi_j}$ as the set of polynomial coefficients of $\bar{f}(\mathscr{D}, w)$ and $\bar{f}(\mathscr{D}', w)$. And we denote t_i or t_i' as an arbitrary tuple.

Finally, by minimizing the differentially private objective function $\hat{f}(\mathscr{D}, w)$, we derive the model parameter $\hat{w}$, which achieves differential privacy and fairness at the same time. We now show that ADFC satisfies (ϵ, δ)-differential privacy in the following theroem.

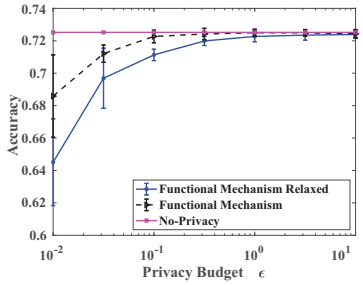


Figure 1: Compare accuracy under different privacy budgets on *US*. ($\delta = 10^{-3}$)

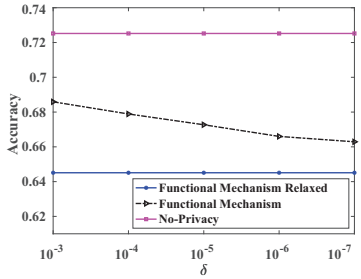


Figure 2: Compare accuracy under different values of δ on *US*.

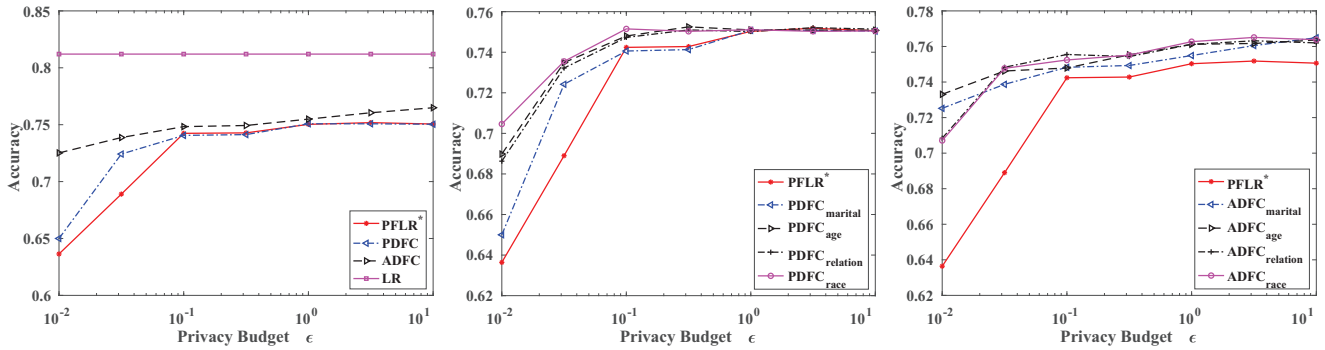
Theorem 5. *The output model parameter $\hat{w}$ in ADFC (Algorithm 3) guarantees (ϵ, δ)-differential privacy, where $\epsilon = \frac{1}{d}\epsilon_s + \frac{d-1}{d}\epsilon_n$ and $\delta = 1 - (1-\delta_s)(1-\delta_n)$.*

Performance Evaluation

Simulation Setup

Data preprocessing We evaluate the performance on two datasets, *Adult* dataset and *US* dataset. The *Adult* dataset from UCI Machine Learning Repository (Dheeru and Karra Taniskidou 2017) contains information about 13 different features (e.g., work-class, education, race, age, sex, and so on) of 48,842 individuals. The label is to predict whether the annual income of those individuals is above 50K or not. The *US* dataset is from Integrated Public Use Microdata Series (Center 2018) and consists of 370,000 records of census microdata, which includes features like age, sex, education, family size, etc. The goal is to predict whether the income is over 25K a year. In both datasets, we consider sex as a binary protected attribute.

Baseline algorithms In our experiments, we compare our approaches, PDFC, and ADFC against several baseline algorithms, namely, LR and PFLR*. LR is a logistic regression model. PFLR* (Xu, Yuan, and Wu 2019) is a differentially private and fair logistic regression model that injects Laplace noise with shifted mean to the objective function of logistic regression with fairness constraint. Moreover, we compare



(a) (b) (c)

Figure 3: Compare accuracy under different privacy budgets on *Adult* ($\delta = 10^{-3}$).

our relaxed functional mechanism against the original functional mechanism proposed in (Zhang et al. 2012) and No-Privacy, which is the original functional mechanism without injecting any noise to the polynomial coefficients.

Evaluation The utility of algorithms is measured by *Accuracy*, defined as follows,

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made},$$

which demonstrates the quality of a classifier. The fairness of classification models is qualified by *risk difference (RD)*

$$RD = |\Pr(\hat{y} = 1|z = 1) - \Pr(\hat{y} = 1|z = 0)|,$$

where z is the protected attribute. We consider a random 80-20 training-testing split and conduct 10 independent runs of algorithms. We then record the mean values and standard deviation values of *Accuracy* and *RD* on the testing dataset. For the parameters of differential privacy, we consider $\epsilon = \{10^{-2}, 10^{-1.5}, 10^{-1}, 10^{0}, 10^{0.5}, 10^{1}\}$, and $\delta = \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$.

Results and Analysis

In Figure 1, we show the accuracy of each algorithm, functional mechanism, relaxed functional mechanism and No-Privacy, as a function of the privacy budget with fixed $\delta = 10^{-3}$. We can see that the accuracy of No-Privacy remains unchanged for all values of ϵ, as it does not provide any differential privacy guarantee. Our relaxed functional mechanism exhibits quite higher accuracy than functional mechanism in high privacy regime, and the accuracy of relaxed functional mechanism is the same as No-Privacy baseline when $\epsilon > 10^{-1}$. Figure 2 studies the accuracy of each algorithm under different values of δ with fixed $\epsilon = 10^{-2}$. Relaxed functional mechanism incurs lower accuracy when δ decreases, as a smaller δ requires a larger scale of noise to be injected in the objective function. But the accuracy of functional mechanism remains considerably lower than relaxed functional mechanism in all cases.

Figure 3a studies the accuracy comparison among PFLR*, LR, PDFC and ADFC on *Adult* dataset with the particular unprotected attribute x_s denoted by marital status. We can observe that ADFC continuously achieves better accuracy than PFLR* in all privacy regime, and PDFC only outperforms PFLR* when ϵ is small. We also evaluate the effect of choosing different attributes as x_s by performing experiments on *Adult* dataset. As shown in Figure 3b and Figure 3c, choosing different attributes, marital status, age, relation and race, has different effects on the accuracy of PDFC and ADFC. However, PDFC and ADFC still outperform PFLR* under varying values of ϵ. As expected, as the value of ϵ increases, the accuracy of each algorithm becomes higher in above three figures.

Table 1 shows how different privacy budgets affect the risk difference of LR, PFLR*, PDFC and ADFC on two datasets. Note that we consider the attribute x_s as race on *Adult* dataset, and work on *US* dataset. It is clear that PDFC and ADFC produce less risk difference compared to PFLR* in most cases of ϵ. The key reason is that adding different amounts of noise regarding different attributes indeed reduces the correlation between unprotected attributes and protected attributes.

Table 1: Risk difference with different privacy budgets ϵ on two datasets ($\delta = 10^{-3}$).

Data	ϵ	LR	PFLR*	PDFC	ADFC
Adult	0.01	0.187 ± 0.049	0.045 ± 0.095	0.048 ± 0.108	0.146 ± 0.131
	0.1	0.187 ± 0.049	0.004 ± 0.009	0.005 ± 0.022	0.068 ± 0.028
	1	0.187 ± 0.049	0.022 ± 0.088	0.002 ± 0.011	0.045 ± 0.027
	10	0.187 ± 0.049	0.003 ± 0.001	0.035 ± 0.041	0.019 ± 0.003
US	0.01	0.191 ± 0.014	0.037 ± 0.038	0.003 ± 0.034	0.004 ± 0.007
	0.1	0.191 ± 0.014	0.078 ± 0.021	0.001 ± 0.006	0.008 ± 0.003
	1	0.191 ± 0.014	0.069 ± 0.007	0.022 ± 0.047	0.031 ± 0.004
	10	0.191 ± 0.014	0.067 ± 0.003	0.022 ± 0.031	0.045 ± 0.002

Conclusion

In this paper, we have introduced two approaches, PDFC and ADFC, to address the discrimination and privacy concerns in logistic regression classification. Different from existing techniques, in both approaches, we consider leveraging functional mechanism to the objective function with decision boundary fairness constraints, and adding noise with different magnitudes into the coefficients of different attributes to further reduce the discrimination and improve the

privacy protection. Moreover, for ADFC, we utilize the proposed relaxed functional mechanism that is built upon Extended Gaussian mechanism, to further improve the model accuracy. By performing extensive empirical comparisons with state-of-the-art methods for differentially private and fair classification, we demonstrated the effectiveness of proposed approaches.

Acknowledgments

The work of J. Ding, X. Zhang, and M. Pan was supported in part by the U.S. National Science Foundation under grants US CNS-1350230 (CAREER), CNS-1646607, CNS-1702850, and CNS-1801925. The work of X. Li was supported in part by the Programs of NSFC under Grant 61762030, in part by the Guangxi Natural Science Foundation under Grant 2018GXNSFDA281013, and in part by the Key Science and Technology Project of Guangxi under Grant AA18242021.

References

Center, M. P. 2018. Integrated public use microdata series, international: Version 7.0.

Chaudhuri, K.; Monteleoni, C.; and Sarwate, A. D. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research* 12:1069–1109.

Datta, A.; Tschantz, M. C.; and Datta, A. 2015. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies* 2015(1):92–112.

Dheeru, D., and Karra Taniskidou, E. 2017. UCI machine learning repository.

Ding, B.; Kulkarni, J.; and Yekhanin, S. 2017. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*.

Dwork, C., and Roth, A. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9(3–4):211–407.

Dwork, C.; Kenthapadi, K.; McSherry, F.; Mironov, I.; and Naor, M. 2006a. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006b. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*.

Ekstrand, M. D.; Joshaghani, R.; and Mehrpouyan, H. 2018. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*.

Erlingsson, Ú.; Pihur, V.; and Korolova, A. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *21st ACM Conference on Computer and Communications Security*.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*.

Giang, V. 2018. The potential hidden bias in automated hiring systems.

Hajian, S.; Domingo-Ferrer, J.; Monreale, A.; Pedreschi, D.; and Giannotti, F. 2015. Discrimination-and privacy-aware patterns. *Data Mining and Knowledge Discovery* 29(6):1733–1782.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*.

Kamiran, F., and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33(1):1–33.

Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2012. A study of top-k measures for discrimination discovery. In *The 27th Annual ACM Symposium on Applied Computing*.

Pedreshi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In *14th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Phan, N.; Vu, M. N.; Liu, Y.; Jin, R.; Dou, D.; Wu, X.; and Thai, M. T. 2019. Heterogeneous gaussian mechanism: Preserving differential privacy in deep learning with provable robustness. *arXiv preprint arXiv:1906.01444*.

Rudin, W., et al. 1964. *Principles of mathematical analysis*. McGraw-hill New York.

Wachter-Boettcher, S. 2018. Ai recruiting tools do not eliminate bias.

Wang, Y.; Si, C.; and Wu, X. 2015. Regression model fitting under differential privacy and model inversion attack. In *24th International Joint Conference on Artificial Intelligence*.

Xu, D.; Yuan, S.; and Wu, X. 2019. Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of The 2019 World Wide Web Conference*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017a. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *The 26th International Conference on World Wide Web*.

Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2017b. Fairness constraints: Mechanisms for fair classification. In *The 20th International Conference on Artificial Intelligence and Statistics*.

Zhang, J.; Zhang, Z.; Xiao, X.; Yang, Y.; and Winslett, M. 2012. Functional mechanism: regression analysis under differential privacy. *Proceedings of the VLDB Endowment* 5(11):1364–1375.