Achieving Differential Privacy and Fairness in Logistic Regression

Depeng Xu
University of Arkansas
depengxu@uark.edu

Shuhan Yuan
University of Arkansas
sy005@uark.edu

Xintao Wu
University of Arkansas
xintaowu@uark.edu

ABSTRACT

Machine learning algorithms are used to make decisions in various applications. These algorithms rely on large amounts of sensitive individual information to work properly. Hence, there are sociological concerns about machine learning algorithms on matters like privacy and fairness. Currently, many studies focus on only protecting individual privacy or ensuring fairness of algorithms. However, how to meet both privacy and fairness requirements simultaneously in machine learning algorithms is under exploited. In this paper, we focus on one classic machine learning model, logistic regression, and develop differentially private and fair logistic regression models by combining functional mechanism and decision boundary fairness in a joint form. Theoretical analysis and empirical evaluations demonstrate our approaches effectively achieve both differential privacy and fairness while preserving good utility.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Supervised learning by classification**; • **Applied computing** → **Law, social and behavioral sciences**;

KEYWORDS

Differential Privacy; Fairness-aware Learning; Logistic Regression

ACM Reference Format:
Depeng Xu, Shuhan Yuan, and Xintao Wu. 2019. Achieving Differential Privacy and Fairness in Logistic Regression. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317584

1 INTRODUCTION

Nowadays, machine learning algorithms are being widely used to automatically make decisions, such as loan application and student admission, based on our individual information. It is important to address individuals' sociological concerns such as privacy and fairness and meet government laws and regulations (e.g., General Data Protection and Regulation on data protection and privacy, and Fair Credit Reporting Act or Equal Credit Opportunity Act on fairness) in training and deploying machine learning algorithms [2, 12].

Differential privacy has been established as a standard privacy model to achieve opt-out right of individuals [6]. Generally speaking, differential privacy guarantees the query results or the released model cannot be exploited by attackers to derive whether one particular record is present or absent in the underlining dataset. Various mechanisms have been proposed to achieve differential privacy [4]. For example, the Laplace mechanism works by injecting random noise into the released results such that the inclusion or exclusion of a single individual record from the dataset makes no statistical difference to the results found [6]. For prediction models, objective perturbation [1] and functional mechanism [16], which add noise to the objective function rather than parameters of built models, have been shown great success.

Meanwhile, fairness-aware learning is increasingly receiving attention in the machine learning field. Many studies have shown that classification models have biased performance against the *protected group*, since the classifiers are only trained to maximize the prediction accuracy. Current research to achieve fair classification can be mainly categorized into two groups: in-processing methods which incorporate fairness constraints into the classification models [11, 15], and pre/post-processing methods which modify the training data and/or the potentially unfair predictions made by the classifiers [5, 8, 10, 17, 18].

In this work, we focus on how to achieve both differential privacy and fairness in logistic regression – a widely-used classification model. It's challenging to achieve both requirements efficiently. The goal of differential privacy in a classification model is to make sure the classifier output is indistinguishable whether an individual record exists in the dataset or not. Its focus is on the individual level. The goal of fairness-aware learning is to make sure that predictions of the protected group are identical to those of the unprotected group, e.g., admission rate of female (as protected group) should be same to male (as unprotected group). Its focus is on the group level. There are few studies on achieving both privacy protection and fairness. Research in [5] proposed a notion of fairness that is a generalization of differential privacy. Research in [9] developed a pattern sanitization method that achieves k-anonymity and fairness. Most recently, the position paper [7] argued for integrating recent research on fairness and non-discrimination to socio-technical systems that provide privacy protection. However, there is no formal study on how to achieve both differential privacy and fairness in classification models.

We develop two methods to achieve differential privacy and fairness in logistic regression. Our simple method incorporates the decision boundary fairness constraint into the objective function of the logistic regression as a penalty term and then applies the functional mechanism to the whole constrained objective function to achieve differential privacy. The decision boundary fairness

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317584

constraint of logistic regression is defined as the covariance between the users' protected attribute and the signed distance from the users' unprotected attribute vectors to the decision boundary, and can be further formulated as the signed distance between the centroids of the protected and unprotected groups. To achieve differential privacy, the functional mechanism brings randomness to the polynomial coefficients of the constrained objective function by introducing Laplace noise with zero mean. Because the penalty term contributes to the global sensitivity of objective function, this simple approach may inject too much noise to the objective function, which reduces the utility of the built logistic regression model. We further develop an enhanced model that injects Laplace noise with shifted mean to the objective function of logistic regression. Our idea is based on the connection between ways of achieving differential privacy and fairness. We notice that both the fairness constraint and functional mechanism perturb the polynomial coefficients of the original objective function. Hence, we can combine them as a single term. In fact, the decision boundary fairness constraint of logistic regression can be treated as a shift of the polynomial coefficients by the signed distance between the centroids of the protected and unprotected groups. As a result, we add noise from a Laplace distribution with non-zero mean that is derived from the fairness constraint. In this way, the fairness constraint is not a penalty term, so we can use privacy budget more efficiently and add less noise.

Our contributions are as follows: 1) To our best knowledge, this is the first work to study how to achieve both differential privacy and fairness in classification models. 2) We develop two methods to achieve differential privacy and fairness in logistic regression. In particular, our enhanced method, which adds Laplace noise with non-zero mean as equivalence to fairness constraint, can reduce the amount of added noise and hence better preserve utility. 3) We conduct evaluation on two real-world datasets and results show that our approaches meet both differential privacy and fairness requirements while achieving good utility.

2 PRELIMINARY

Let $D = \{\mathbf{X}, S, Y\}$ be a dataset with n tuples $t_1, t_2, \cdots, t_n$, where $\mathbf{X} = (X_1, X_2, \cdots, X_d)$ indicates d unprotected attributes; S denotes the protected attribute; Y is the decision. For each tuple $t_i = \{\mathbf{x}_i, s_i, y_i\}$, without loss of generality, we assume $x_{i(l)} \in [0, 1]$ for $l = (1, 2, \cdots, d)$, $s_i \in \{0, 1\}$, and $y_i \in \{0, 1\}$. Our objective is to build a classification model $\hat{y} = q(\mathbf{x}; \mathbf{w})$ with parameter $\mathbf{w}$ from D that achieves reasonable utility and meets both fairness and differential privacy requirements. To fit $\mathbf{w}$ to make accurate predictions, we have an objective function $f_D(\mathbf{w}) = \sum_{i=1}^{n} f(t_i; \mathbf{w})$ that takes t_i and $\mathbf{w}$ as input. The optimal model parameter $\bar{\mathbf{w}}$ is defined as: $\bar{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} f(t_i; \mathbf{w})$.

2.1 Differential Privacy

Differential privacy guarantees the output of a query q be insensitive to the presence or absence of any one individual record in a dataset. We use D and D' to denote two neighboring datasets which differ in exactly one record.

Definition 2.1. **Differential Privacy** [6]. A mechanism $\mathcal{M}$ satisfies ε-differential privacy, if for all neighboring datasets D and D' and all subsets Z of $\mathcal{M}$'s range:

$$\Pr(\mathcal{M}(D) \in Z) \leq \exp(\varepsilon) \cdot \Pr(\mathcal{M}(D') \in Z). \quad (1)$$

The parameter ε denotes the privacy budget (smaller values indicate a stronger privacy guarantee).

Definition 2.2. **Global Sensitivity** [6]. Given a query $q: D \to \mathbb{R}^d$, the global sensitivity Δ is defined as $\Delta = \max_{D, D'} ||q(D) - q(D')||_1$.

The global sensitivity measures the maximum possible change in $q(D)$ when one record in the dataset changes. The Laplace mechanism is a popular method to achieve differential privacy. It adds identical independent noise into each output value of $q(D)$.

Definition 2.3. **Laplace Mechanism** [6]. Given a dataset D and a query q, a mechanism $\mathcal{M}(D) = q(D) + \boldsymbol{\eta}$ satisfies ε-differential privacy, where $\boldsymbol{\eta}$ is a random vector drawn from $Lap(0, \frac{\Delta}{\varepsilon})$ [1].

Functional Mechanism. Functional mechanism [16] is a differentially private method designed for optimization based models. It achieves ε-differential privacy by injecting noise into the objective function of the model and returns privacy preserving parameter $\bar{\mathbf{w}}$ that minimizes the perturbed objective function.

Because the objective function $f_D(\mathbf{w})$ is a complicated function of $\mathbf{w}$, the functional mechanism exploits the polynomial representation of $f_D(\mathbf{w})$. The model parameter $\mathbf{w}$ is a vector that contains d values $w_1, w_2, \cdots, w_d$. Let $\phi(\mathbf{w})$ denote a product of $w_1, w_2, \cdots, w_d$, i.e., $\phi(\mathbf{w}) = w_1^{c_1} \cdot w_2^{c_2} \cdots w_d^{c_d}$ for some $c_1, c_2, \cdots, c_d \in \mathbb{N}$. Let $\boldsymbol{\Phi}_j$ ($j \in \mathbb{N}$) denote the set of all products of $w_1, w_2, \cdots, w_d$ with degree j, i.e., $\boldsymbol{\Phi}_j = \{w_1^{c_1} w_2^{c_2} \cdots w_d^{c_d} | \sum_{l=1}^{d} c_l = j\}$. For example, $\boldsymbol{\Phi}_1 = \{w_1, w_2, \cdots, w_d\}$, and $\boldsymbol{\Phi}_2 = \{w_i \cdot w_j | i, j \in [1, d]\}$.

Based on the Stone-Weierstrass Theorem [13], any continuous and differentiable function can be expressed in the polynomial representation. Hence, the objective function $f_D(\mathbf{w})$ can be expressed as a polynomial of $w_1, w_2, \cdots, w_d$, for some $J \in \mathbb{N}$:

$$f_D(\mathbf{w}) = \sum_{i=1}^{n} \sum_{j=0}^{J} \sum_{\phi \in \boldsymbol{\Phi}_j} \lambda_{\phi t_i} \phi(\mathbf{w}), \quad (2)$$

where $\lambda_{\phi t_i} \in \mathbb{R}$ denotes the coefficient of $\phi(\mathbf{w})$ in the polynomial.

Functional mechanism perturbs the objective function $f_D(\mathbf{w})$ by injecting Laplace noise into its polynomial coefficients $\bar{\lambda}_\phi = \sum_{i=1}^{n} \lambda_{\phi t_i} + Lap(0, \frac{\Delta}{\varepsilon})$, where $\Delta = 2 \max_t \sum_{j=1}^{J} \sum_{\phi \in \boldsymbol{\Phi}_j} ||\lambda_{\phi t}||_1$. Then the model parameter $\bar{\mathbf{w}}$ is derived to minimize the perturbed function $\bar{f}_D(\mathbf{w})$.

Applying Functional Mechanism on Logistic Regression. A logistic regression on D returns a function which predicts $\hat{y}_i = 1$ with probability:

$$\hat{y}_i = q(\mathbf{x}_i; \mathbf{w}) = \exp(\mathbf{x}_i^T \mathbf{w}) / (1 + \exp(\mathbf{x}_i^T \mathbf{w})). \quad (3)$$

The objective function of logistic regression is defined as:

$$f_D(\mathbf{w}) = \sum_{i=1}^{n} \left[\log(1 + \exp(\mathbf{x}_i^T \mathbf{w})) - y_i \mathbf{x}_i^T \mathbf{w} \right]. \quad (4)$$

As the polynomial form of $f_D(\mathbf{w})$ in Equation 4 contains terms with unbounded degrees, to apply the functional mechanism, Equation

[1] The Laplace distribution $Lap(\eta | \mu, b)$ with mean μ and scale b has probability density function $Lap(\eta | \mu, b) = \frac{1}{2b} \exp(\frac{|x - \mu|}{b})$. Its variance is $2b^2$.

4 is rewritten as the approximate polynomial representation based on Taylor expansion [16]:

$$f_D(\mathbf{w}) = \Big(\sum_{i=1}^{n} \sum_{j=0}^{2} \frac{f_1^{(j)}(0)}{j!} \left(\mathbf{x}_i^T \mathbf{w}\right)^j \Big) - \Big(\sum_{i=1}^{n} y_i \mathbf{x}_i^T \Big)\mathbf{w}, \quad (5)$$

where $f_1(\cdot) = \log(1 + \exp(\cdot))$.

When rewriting Equation 5 in the form of Equation 2, we have

$$\{\lambda_{\phi t_i}\}_{\phi \in \Phi_1} =: \lambda_{1t_i} = \Big(\tfrac{f_1^{(1)}(0)}{1!}\mathbf{x}_i\Big) - \Big(y_i \mathbf{x}_i\Big), \quad (6)$$

$$\{\lambda_{\phi t_i}\}_{\phi \in \Phi_2} =: \lambda_{2t_i} = \tfrac{f_1^{(2)}(0)}{2!} \left(\mathbf{x}_i\right)^2. \quad (7)$$

The global sensitivity of $f_D(\mathbf{w})$ is:

$$\begin{aligned} \Delta_f &= 2 \max_t \left(\left| \Big(\frac{f_1^{(1)}(0)}{1!} - y \Big) \sum_{l=1}^{d} x_{(l)} \right| + \left| \frac{f_1^{(2)}(0)}{2!} \sum_{l,m}^{d} x_{(l)} x_{(m)} \right| \right) \\ &\leq 2(\frac{d}{2} + \frac{d^2}{8}) = \frac{d^2}{4} + d. \end{aligned} \quad (8)$$

Thus, to achieve ε-differential privacy, the functional mechanism adds $Lap(0, \frac{\Delta_f}{\varepsilon})$ noise to the polynomial coefficients of the objective function.

2.2 Classification Fairness

There are many definitions or requirements of classification fairness. A common notion of group fairness is demographic parity, which requires that a decision Y is independent of the protected attribute S.

Definition 2.4. **Demographic Parity** [10]. A classification model $\hat{y} = q(\mathbf{x}; \mathbf{w})$ satisfies demographic parity if $\Pr(\hat{Y} = 1|S = 1) = \Pr(\hat{Y} = 1|S = 0)$.

The discrimination of the model can be quantified by risk difference (RD):

$$RD = |\Pr(\hat{Y} = 1|S = 1) - \Pr(\hat{Y} = 1|S = 0)|. \quad (9)$$

To achieve classification fairness, the in-processing approaches are to find parameter $\mathbf{w}$ that minimizes the objective function under a fairness constraint:

$$\begin{aligned} &\text{minimize} \quad f_D(\mathbf{w}) \\ &\text{subject to} \quad g_D(\mathbf{w}) \leq \tau,\ g_D(\mathbf{w}) \geq -\tau, \end{aligned} \quad (10)$$

where $g_D(\mathbf{w})$ is the constraint term; $\tau \in \mathbf{R}^+$ is the threshold of constraint. For example, in [15], the fairness constraint is defined as the covariance between the users' protected attribute and the signed distance from the users' unprotected attribute vectors to the decision boundary $\{d_{\mathbf{w}}(\mathbf{x}_i)\}_{i=1}^n$,

$$g_D(\mathbf{w}) = \mathbb{E}[(s - \bar{s}) d_{\mathbf{w}}(\mathbf{x})] - \mathbb{E}[(s - \bar{s})] d_{\mathbf{w}}(\mathbf{x}) \propto \sum_{i=1}^{n} (s_i - \bar{s}) d_{\mathbf{w}}(\mathbf{x}_i), \quad (11)$$

where $\bar{s}$ is the mean value of the protected attribute; $\mathbb{E}[(s - \bar{s})] = 0$. For linear classification models, like logistic regression or linear SVMs, the decision boundary is simply the hyperplane defined by $\mathbf{x}^T \mathbf{w} = 0$. Then, Equation 11 reduces to $g_D(\mathbf{w}) = \sum_{i=1}^{n} (s_i - \bar{s}) \mathbf{x}_i^T \mathbf{w}$. The decision boundary fairness is proven to be a notion of fairness that minimizes the surrogate risk difference [14].

3 DIFFERENTIALLY PRIVATE AND FAIR LOGISTIC REGRESSION

In this section, we first present a simple approach (PFLR) to achieve differentially private and fair logistic regression. Then we show it leads to an enhanced approach (PFLR*) that achieves the same level of differential privacy and fairness with more flexibility and less noise.

3.1 PFLR: A Simple Approach

One straightforward approach to achieve both differential privacy and fairness is to apply the functional mechanism to the objective function with fairness constraint $\tilde{f}_D(\mathbf{w})$. We consider the fairness constraint as a penalty term to the objective function. Then, the objective function ends up as:

$$\tilde{f}_D(\mathbf{w}) = f_D(\mathbf{w}) + \alpha |g_D(\mathbf{w}) - \tau|, \quad (12)$$

where α is a hyper-parameter to balance the trade-off between utility and fairness. We set $\alpha = 1, \tau = 0$ for ease of discussion. The theoretical analysis still holds if α and τ are set to other values.

For logistic regression, $f_D(\mathbf{w})$ is the objective function shown in Equation 4, and $g_D(\mathbf{w})$ indicates the decision boundary fairness constraint shown in Equation 11. We then rewrite $\tilde{f}_D(\mathbf{w})$ in the polynomial form based on Taylor expansion.

$$\tilde{f}_D(\mathbf{w}) = \Big(\sum_{i=1}^{n} \sum_{j=0}^{2} \frac{f_1^{(j)}(0)}{j!} \left(\mathbf{x}_i^T \mathbf{w}\right)^j \Big) - \Big(\sum_{i=1}^{n} y_i \mathbf{x}_i^T \Big)\mathbf{w} + \Big| \sum_{i=1}^{n} (s_i - \bar{s}) \mathbf{x}_i^T \mathbf{w} \Big|. \quad (13)$$

By transforming $\tilde{f}_D(\mathbf{w})$ in the form of Equation 2, we have $\tilde{\lambda}_{1t_i} = \lambda_{1t_i} + |(s_i - \bar{s})\mathbf{x}_i|$ and $\tilde{\lambda}_{2t_i} = \lambda_{2t_i}$, where λ_{1t_i} and λ_{2t_i} are defined in Equations 6 and 7, respectively.

The global sensitivity of $\tilde{f}_D(\mathbf{w})$ is:

$$\begin{aligned} \Delta_{\tilde{f}} &= 2 \max_t \left(\left| \Big(\frac{f_1^{(1)}(0)}{1!} - y + |s - \bar{s}| \Big) \sum_{l=1}^{d} x_{(l)} \right| + \left| \frac{f_1^{(2)}(0)}{2!} \sum_{l,m}^{d} x_{(l)} x_{(m)} \right| \right) \\ &\leq 2(\frac{3d}{2} + \frac{d^2}{8}) = \frac{d^2}{4} + 3d. \end{aligned} \quad (14)$$

The derived $\bar{\mathbf{w}}$ satisfies ε-differential privacy by applying Algorithm 1. Since the objective function contains the fairness constraint as a penalty term, the classification model also achieves fairness.

Algorithm 1 PFLR (Dataset D, objective function $f_D(\mathbf{w})$, fairness constraint $g_D(\mathbf{w})$, privacy budget ε)

1: Set $\tilde{f}_D(\mathbf{w})$ by Equation 12
2: Compute λ_{1t_i} and λ_{2t_i} by Equations 6 and 7
3: Set $\tilde{\lambda}_{1t_i} = \lambda_{1t_i} + |(s_i - \bar{s})\mathbf{x}_i|$ and $\tilde{\lambda}_{2t_i} = \lambda_{2t_i}$
4: Set Δ by Equation 14
5: Set $\bar{\lambda}_1 = \left(\sum_{i=1}^{n} \tilde{\lambda}_{1t_i} \right) + Lap(0, \frac{\Delta_{\tilde{f}}}{\varepsilon})$
6: Set $\bar{\lambda}_2 = \left(\sum_{i=1}^{n} \tilde{\lambda}_{2t_i} \right) + Lap(0, \frac{\Delta_{\tilde{f}}}{\varepsilon})$
7: Let $\bar{f}_D(\mathbf{w}) = \bar{\lambda}_1^T \mathbf{\Phi}_1 + \bar{\lambda}_2^T \mathbf{\Phi}_2$
8: Compute $\bar{\mathbf{w}} = \arg\min_{\mathbf{w}} \bar{f}_D(\mathbf{w})$
9: Return $\bar{\mathbf{w}}$

Theorem 3.1. *Algorithm 1 satisfies ε-differential privacy.*

Proof. Assume D and D' are two neighbouring datasets. Without loss of generality, D and D' differ in row t_r and t'_r. Δ is calculated by Equation 14. $\bar{f}(\mathbf{w})$ is the output of Line 7. We have

$$\begin{aligned}
\frac{\Pr\{\bar{f}(\mathbf{w})|D\}}{\Pr\{\bar{f}(\mathbf{w})|D'\}} &= \frac{\prod_{j=1}^{2}\prod_{\phi\in\mathbf{\Phi}_j}\exp\left(\frac{\varepsilon\left\|\sum_{t_i\in D}\tilde{\lambda}_{\phi t_i}-\bar{\lambda}_\phi\right\|_1}{\Delta_{\tilde{f}}}\right)}{\prod_{j=1}^{2}\prod_{\phi\in\mathbf{\Phi}_j}\exp\left(\frac{\varepsilon\left\|\sum_{t'_i\in D'}\tilde{\lambda}_{\phi t'_i}-\bar{\lambda}_\phi\right\|_1}{\Delta_{\tilde{f}}}\right)} \\
&\le \prod_{j=1}^{2}\prod_{\phi\in\mathbf{\Phi}_j}\exp\left(\frac{\varepsilon}{\Delta_{\tilde{f}}}\cdot\left\|\sum_{t_i\in D}\tilde{\lambda}_{\phi t_i}-\sum_{t'_i\in D'}\tilde{\lambda}_{\phi t'_i}\right\|_1\right) \\
&= \prod_{j=1}^{2}\prod_{\phi\in\mathbf{\Phi}_j}\exp\left(\frac{\varepsilon}{\Delta_{\tilde{f}}}\cdot\left\|\tilde{\lambda}_{\phi t_r}-\tilde{\lambda}_{\phi t'_r}\right\|_1\right) \\
&= \exp\left(\frac{\varepsilon}{\Delta_{\tilde{f}}}\cdot\sum_{j=1}^{2}\sum_{\phi\in\mathbf{\Phi}_j}\left\|\tilde{\lambda}_{\phi t_r}-\tilde{\lambda}_{\phi t'_r}\right\|_1\right) \\
&\le \exp\left(\frac{\varepsilon}{\Delta_{\tilde{f}}}\cdot 2\max_t\sum_{j=1}^{2}\sum_{\phi\in\mathbf{\Phi}_j}||\tilde{\lambda}_{\phi t}||_1\right)=\exp(\varepsilon).
\end{aligned} \tag{15}$$

□

3.2 PFLR*: An Enhanced Approach

We further enhance the simple approach by incorporating the fairness constraint into the Laplace noise.

In Equation 13, the fairness constraint $g_D(\mathbf{w})=\sum_{i=1}^{n}(s_i-\bar{s})\mathbf{x}_i^T\mathbf{w}$ can be considered as shifting the first degree polynomial coefficients of $\mathbf{\Phi}_1$ in the objective function by $\sum_{i=1}^{n}(s_i-\bar{s})\mathbf{x}_i$. Since $\sum_{i=1}^{n}(s_i-\bar{s})\mathbf{x}_i$ is the signed distance between the centroids of the protected and unprotected groups, the derived $\bar{\mathbf{w}}$ based on the shifted coefficients ensures that the centroids of the protected and unprotected groups have the same distance to the decision boundary. Thus, the decision boundary fairness is achieved.

Meanwhile, the functional mechanism adds Laplace noise to inject randomness to the polynomial coefficients of the objective function. Because $\Pr\{\bar{f}(\mathbf{w})|D\}$ depends on the probability of the noise distribution, the designed Laplace noise provides the property of differential privacy.

Following this observation, instead of applying the fairness constraint as a penalty term to the objective function, we design a new functional mechanism that incorporates fairness constraint into the Laplace noise. In particular, we shift the polynomial coefficients when adding Laplace noise. The shift is achieved by setting the mean of Laplace distribution. As $g_D(\mathbf{w})$ only affects $\mathbf{\Phi}_1$, we change the mean of Laplace distribution $\mu=\{\mu_{(l)}\}_{l=1}^{d}$ from 0 to $\sum_{i=1}^{n}(s_i-\bar{s})\mathbf{x}_i$ for the coefficients related with $\mathbf{\Phi}_1$, so it has the equivalent effect to the fairness constraint. Formally, we have $\mu_{(l)}=\sum_{i=1}^{n}(s_i-\bar{s})x_{(l)t_i}$. Because the fairness constraint is not a penalty term of the objective function, PFLR* has the same objective function as the regular logistic regression $f_D(\mathbf{w})$ (defined in Equation 5). The global sensitivity of PFLR* is $\Delta_f=d^2/4+d$ as shown in Equation 8.

Note that, given a dataset, $\mu=\sum_{i=1}^{n}(s_i-\bar{s})\mathbf{x}_i$ is fixed. As we also access data when calculating μ, a small part of privacy budget ε_g is used to calculate μ in a differentially private manner by Laplace mechanism (Algorithm 2 Line 2). The sensitivity of μ is

$$\Delta_g=2\max_t\left|\sum_{l=1}^{d}(s_{t_r}-\bar{s})x_{t_r(l)}\right|\le 2d. \tag{16}$$

We formalize our new functional mechanism with fairness constraint as Algorithm 2. We split the total privacy budget ε into two parts ε_f and ε_g. We first calculate the differentially private μ with the privacy budget ε_g (Lines 1-2). Then, we introduce Laplace noise $Lap(\mu,\frac{\Delta_f}{\varepsilon_f})$ to the polynomial coefficients of the objective function with the privacy budget ε_f (Lines 3-7). Note that we only add the shifted Laplace noise to coefficients with $\mathbf{\Phi}_1$. Finally, we derive the optimized $\bar{\mathbf{w}}$ according to $\bar{f}_D(\mathbf{w})$ (Line 8). Next we show PFLR* achieves ε-differential privacy.

Theorem 3.2. *Algorithm 2 satisfies ε-differential privacy.*

Proof. Assume D and D' are two neighbouring datasets. Without loss of generality, D and D' differ in row t_r and t'_r. Δ_f is calculated by Equation 8. $\bar{f}(\mathbf{w})$ is the output of Line 7. Adding Laplace noise with non-zero mean to coefficients still satisfies ε_f-differential privacy.

$$\begin{aligned}
&\frac{\Pr(\bar{f}(\mathbf{w})|D)}{\Pr(\bar{f}(\mathbf{w})|D')} \\
&= \frac{\prod_{\phi\in\mathbf{\Phi}_1}\exp\left(\frac{\varepsilon_f\left\|\sum_{t_i\in D}\lambda_{\phi t_i}-\bar{\lambda}_\phi-\mu\right\|_1}{\Delta_f}\right)\prod_{\phi\in\mathbf{\Phi}_2}\exp\left(\frac{\varepsilon_f\left\|\sum_{t_i\in D}\lambda_{\phi t_i}-\bar{\lambda}_\phi\right\|_1}{\Delta_f}\right)}{\prod_{\phi\in\mathbf{\Phi}_1}\exp\left(\frac{\varepsilon_f\left\|\sum_{t'_i\in D'}\lambda_{\phi t'_i}-\bar{\lambda}_\phi-\mu\right\|_1}{\Delta_f}\right)\prod_{\phi\in\mathbf{\Phi}_2}\exp\left(\frac{\varepsilon_f\left\|\sum_{t'_i\in D'}\lambda_{\phi t'_i}-\bar{\lambda}_\phi\right\|_1}{\Delta_f}\right)} \\
&\le \prod_{\phi\in\mathbf{\Phi}_1}\exp\left(\frac{\varepsilon_f}{\Delta_f}\cdot\left\|\sum_{t_i\in D}\lambda_{\phi t_i}-\sum_{t'_i\in D'}\lambda_{\phi t'_i}\right\|_1\right) \\
&\quad\cdot\prod_{\phi\in\mathbf{\Phi}_2}\exp\left(\frac{\varepsilon_f}{\Delta_f}\cdot\left\|\sum_{t_i\in D}\lambda_{\phi t_i}-\sum_{t'_i\in D'}\lambda_{\phi t'_i}\right\|_1\right) \\
&= \prod_{j=1}^{2}\prod_{\phi\in\mathbf{\Phi}_j}\exp\left(\frac{\varepsilon_f}{\Delta_f}\cdot\left\|\lambda_{\phi t_r}-\lambda_{\phi t'_r}\right\|_1\right) \\
&= \exp\left(\frac{\varepsilon_f}{\Delta_f}\cdot\sum_{j=1}^{2}\sum_{\phi\in\mathbf{\Phi}_j}\left\|\lambda_{\phi t_r}-\lambda_{\phi t'_r}\right\|_1\right) \\
&\le \exp\left(\frac{\varepsilon_f}{\Delta_f}\cdot 2\max_t\sum_{j=1}^{2}\sum_{\phi\in\mathbf{\Phi}_j}||\lambda_{\phi t}||_1\right)=\exp(\varepsilon_f)
\end{aligned} \tag{17}$$

Using Laplace mechanism, Line 2 satisfies ε_g-differential privacy on calculating μ. Since $\varepsilon_f+\varepsilon_g=\varepsilon$, Algorithm 2 satisfies ε-differential privacy. □

Algorithm 2 PFLR* (Database D, objective function $f_D(\mathbf{w})$, fairness constraint $g_D(\mathbf{w})$, privacy budget $\varepsilon_f,\varepsilon_g$)

1: Set Δ_g by Equation 16
2: Calculate $\mu=\{\mu_{(l)}\}_{l=1}^{d}$ by $\mu_{(l)}=\sum_{i=1}^{n}(s_i-\bar{s})x_{(l)t_i}+Lap(0,\frac{\Delta_g}{\varepsilon_g})$
3: Compute λ_{1t_i} and λ_{2t_i} by Equation 6 and 7
4: Set Δ_f by Equation 8
5: Set $\bar{\lambda}_1=\left(\sum_{i=1}^{n}\lambda_{1t_i}\right)+Lap(\mu,\frac{\Delta_f}{\varepsilon})$
6: Set $\bar{\lambda}_2=\left(\sum_{i=1}^{n}\lambda_{2t_i}\right)+Lap(0,\frac{\Delta_f}{\varepsilon})$
7: Let $\bar{f}_D(\mathbf{w})=\bar{\lambda}_1^T\mathbf{\Phi}_1+\bar{\lambda}_2^T\mathbf{\Phi}_2$
8: Compute $\bar{\mathbf{w}}=\arg\min_{\mathbf{w}}\bar{f}_D(\mathbf{w})$
9: Return $\bar{\mathbf{w}}$

Comparison between PFLR and PFLR*. In PFLR, the fairness constraint term $g_D(\mathbf{w})$ contributes to the sensitivity of the polynomial coefficients of the objective function. In PFLR*, the fairness constraint is achieved by adding Laplace noise with non-zero mean value, so the sensitivity of the polynomial coefficient

is not related to $g_D(\mathbf{w})$. PFLR* uses separate budgets on objective function and fairness constraint, so it's more flexible to find good trade-offs among privacy, fairness and utility. The fairness constraint has a much smaller sensitivity than the objective function ($\Delta_g \ll \Delta_f$). Hence, we can allocate a relatively small privacy budget on calculating the fairness constraint with Laplace mechanism, the utility is still satisfactory. Then, more privacy budget can be used to the objective function, resulting in a smaller scale of noise. More concretely, we compare the amount of noise that is introduced to the two proposed approaches. The variance of $\lambda_{\phi t_i} \in \{\bar{\lambda}_1, \bar{\lambda}_2\}$ in PFLR is $2(\Delta_{\tilde{f}}/\varepsilon)^2$. Thus, the variance of total noise added in PFLR is $2(d^2+d)(\Delta_{\tilde{f}}/\varepsilon)^2$. On the other hand, the variance of $\lambda_{\phi t_i} \in \bar{\lambda}_2$ in PFLR* is $Var(\bar{\lambda}_2) = 2(\Delta_f/\varepsilon_f)^2$. Because PFLR* injects Laplace noise to both $\bar{\lambda}_1$ and μ, based on the law of total variance, the variance of $\lambda_{\phi t_i} \in \bar{\lambda}_1$ in PFLR* is $Var(\bar{\lambda}_1) = \mathbb{E}[Var(\bar{\lambda}_1|\mu)] + Var(\mathbb{E}[\bar{\lambda}_1|\mu]) = 2(\Delta_f/\varepsilon_f)^2 + 0$. Thus, the variance of total noise added in PFLR* is $2(d^2+d)(\Delta_f/\varepsilon_f)^2$. If we set $\varepsilon_f \geq (\Delta_f/\Delta_{\tilde{f}})\varepsilon$, PFLR* injects less noise.

4 EXPERIMENTS

4.1 Experiment Setup

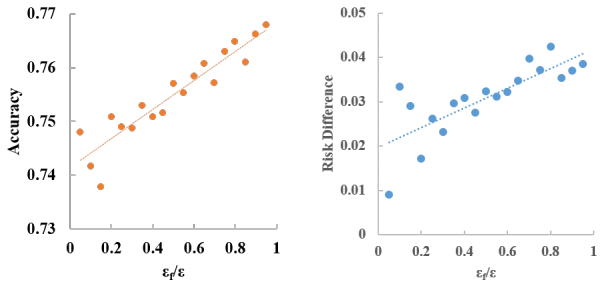
Dataset. We evaluate our methods on Adult [3] and Dutch [19]. For both datasets, we consider "Sex" as protected attribute and "Income" as decision. For unprotected attributes, we convert categorical attributes to one-hot vectors and normalize numerical attributes to $x \in [0, 1]$. The Adult dataset has 45222 records and 40 features. The Dutch dataset has 60420 records and 35 features.

Baselines. We compare the proposed differentially private and fair logistic regression models (**PFLR** and **PFLR***) with the following baselines: 1) a regular logistic regression model (**LR**); 2) a differentially private only LR using functional mechanism (**PrivLR**); 3) a fair only LR using Equation 11 as the fairness constraint (**FairLR**).

Metrics. We evaluate the performance of the proposed approaches and baselines on utility and fairness. We use *accuracy* as the utility metric and *risk difference (RD)* as the fairness metric. We run all models 10 times for each setting and report the mean and standard deviation of each metric.

4.2 Experimental Results

We first compare the performance of all five methods when $\varepsilon = 1$ ($\varepsilon_f = \varepsilon/2$ in PFLR*). As shown in Table 1, the regular logistic regression (LR) achieves the accuracy of 0.8380 on Adult and 0.8164 on Dutch, but it doesn't protect privacy nor achieves fairness (RD = 0.1577 and 0.1747, respectively). PrivLR has privacy protection but the accuracy decreases 11.42% on Adult and 17.52% on Dutch compared with LR as the result of the trade-off between privacy and utility. The risk difference of PrivLR is lower than LR, yet still larger than 0.05. The decrease of risk difference is mostly due to its low accuracy instead of fairness. FairLR achieves fairness (RD = 0.0095 on Adult and RD = 0.0299 on Dutch) as expected but it has no privacy guarantee. For PFLR and PFLR*, they both meet the privacy and fairness requirements. PFLR* has significantly higher accuracy than PFLR on both datasets (based on t-tests with p-values < 0.05). It indicates that PFLR* adds less noise to meet the same level of privacy guarantee.



(a) Accuracy **(b) Risk difference**

Figure 1: PFLR* with different privacy budget splits $\varepsilon_f/\varepsilon$ (Adult dataset, $\varepsilon = 10$)

Different Privacy Budgets. Table 2 shows how different settings of privacy budget ε affect our two methods and PrivLR. For PrivLR, its accuracy decreases dramatically when a stronger privacy requirement (smaller ε) is enforced. The risk difference of PrivLR decreases with the decrease of ε (the increase of noise). When ε is large, the accuracy is good and the risk difference is high. When ε is small, the accuracy is bad and the risk difference is low but with high variance.

For PFLR and PFLR*, when $\varepsilon = 0.1, 1$, PFLR* has significantly higher accuracy than PFLR on both Adult and Dutch (based on t-tests with p-values < 0.05). Especially, when $\varepsilon = 0.1$, PFLR's accuracy is only 0.6172 on Adult and 0.5069 on Dutch while PFLR*'s accuracy is 0.7491 on Adult and 0.6158 on Dutch. The accuracy of PFLR* is more consistent and relatively more resilient under different settings of privacy budget ε. When ε is relaxed to 100, PFLR and PFLR* have similar accuracy to FairLR (shown in Table 2). Overall, PFLR* outperforms PFLR especially when privacy budget is small.

Different Privacy Budget Splits $\varepsilon_f/\varepsilon$ for PFLR*. PFLR* splits the privacy budget ($\varepsilon = \varepsilon_f + \varepsilon_g$) into two parts: computing the fairness constraint (ε_g) and building the classification model (ε_f). Therefore, there is a trade-off between fairness and utility by controlling $\varepsilon_f/\varepsilon$. We further evaluate the performance of PFLR* in terms of accuracy and risk difference with various privacy budget splits by ranging the values of $\varepsilon_f/\varepsilon$ from 0.05 to 0.95 with an interval as 0.05. In Fig. 1a, we observe that with the increase of ε_f, the accuracy increases accordingly. This is because when ε_f keeps increasing, the privacy budget for the objective function becomes large, which reduces noise added to the classification model. For the risk difference, as shown in Fig. 1b, when ε_f increases, the risk difference increases. This is because PFLR* injects more noise to compute the fairness constraint. However, the risk difference is consistently smaller than 0.05 while increasing ε_f. Hence, as the result of a small sensitivity Δ_g, the utility of fairness constraint is well preserved even with a small ε_g.

5 CONCLUSION AND FUTURE WORK

In this work, we have developed two differentially private and fair logistic regression models, PFLR and PFLR*. PFLR is to apply the functional mechanism to the objective function with fairness

Table 1: Accuracy and risk difference (mean ± std.) of each method ($\varepsilon = 1$)

Method	Adult		Dutch	
	Accuracy	Risk Difference	Accuracy	Risk Difference
LR	0.8380 ± 0.0023	0.1577 ± 0.0064	0.8164 ± 0.0048	0.1747 ± 0.0033
PrivLR	0.7238 ± 0.0612	0.0502 ± 0.0581	0.6412 ± 0.0458	0.0739 ± 0.0574
FairLR	0.7739 ± 0.0521	0.0095 ± 0.0071	0.7673 ± 0.0064	0.0299 ± 0.0067
PFLR	0.7400 ± 0.0182	0.0213 ± 0.0258	0.6278 ± 0.0408	0.0206 ± 0.0204
PFLR*	0.7552 ± 0.0092	0.0053 ± 0.0070	0.6482 ± 0.0188	0.0430 ± 0.0265

Table 2: Accuracy and risk difference with different privacy budgets ε

	ε	PrivLR		PFLR		PFLR*	
		Accuracy	Risk Difference	Accuracy	Risk Difference	Accuracy	Risk Difference
Adult	0.1	0.6263 ± 0.1480	0.0883 ± 0.0805	0.6172 ± 0.1187	0.0351 ± 0.0493	0.7491 ± 0.0040	0.0028 ± 0.0039
	1	0.7238 ± 0.0612	0.0502 ± 0.0581	0.7400 ± 0.0182	0.0213 ± 0.0258	0.7552 ± 0.0092	0.0053 ± 0.0070
	10	0.7270 ± 0.0877	0.1459 ± 0.0798	0.7631 ± 0.0155	0.0338 ± 0.0255	0.7632 ± 0.0093	0.0204 ± 0.0140
	100	0.8295 ± 0.0032	0.1624 ± 0.0116	0.7835 ± 0.0318	0.0332 ± 0.0243	0.7913 ± 0.0200	0.0234 ± 0.0189
Dutch	0.1	0.5241 ± 0.0396	0.0317 ± 0.0187	0.5069± 0.0459	0.0441 ± 0.0245	0.6158 ± 0.0239	0.0516 ± 0.0204
	1	0.6412 ± 0.0458	0.0739 ± 0.0574	0.6278 ± 0.0408	0.0206 ± 0.0204	0.6482 ± 0.0188	0.0460 ± 0.0265
	10	0.7239 ± 0.0902	0.1346 ± 0.0563	0.7282 ± 0.0493	0.0211 ± 0.0152	0.7080 ± 0.0329	0.0220 ± 0.0208
	100	0.8154 ± 0.0042	0.1687 ± 0.0054	0.7681 ± 0.0054	0.0301 ± 0.0085	0.7618 ± 0.0144	0.0250 ± 0.0128

constraint as a penalty term. Our enhanced model, PFLR*, takes advantage of the connection between ways of achieving differential privacy and fairness and adds the Laplace noise with non-zero mean. The experimental results on two datasets demonstrate the effectiveness of two approaches and show the superiority of PFLR*. In this work, we consider logistic regression as the classification model and the covariance between decision boundary and the protected attribute as the fairness constraint. In future work, we plan to extend our methods to other classification models and other fairness constraints. Another research direction is to study allocation strategies of privacy budget, e.g., adding different amount of noise to coefficients containing different attributes.

ACKNOWLEDGMENTS

This work was supported in part by NSF 1646654, 1502273 and 1523115.

REFERENCES

[1] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. 2011. Differentially Private Empirical Risk Minimization. *J. Mach. Learn. Res.* 12 (2011), 1069–1109.
[2] Bart Custers. 2013. Data Dilemmas in the Information Society: Introduction and Overview. In *Discrimination and Privacy in the Information Society - Data Mining and Profiling in Large Databases*. 3–26.
[3] Dua Dheeru and Efi Karra Taniskidou. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[4] Cynthia Dwork. 2011. A Firm Foundation for Private Data Analysis. *Commun. ACM* 54, 1 (2011), 86–95.
[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science*. 214–226.
[6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography, Third*. 265–284.
[7] Michael D. Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. 2018. Privacy for All: Ensuring Fair and Equitable Privacy Protections. In *Conference on Fairness, Accountability and Transparency*. 35–47.
[8] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *KDD*.
[9] Sara Hajian, Josep Domingo-Ferrer, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. 2015. Discrimination- and privacy-aware patterns. *Data Min. Knowl. Discov.* 29, 6 (2015), 1733–1782.
[10] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*.
[11] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware Learning through Regularization Approach. In *ICDMW*.
[12] Ugo Pagallo. 2012. On the Principle of Privacy by Design and its Limits: Technology, Ethics and the Rule of Law. In *European Data Protection: In Good Health?* 331–346.
[13] Walter Rudin. 1953. *Principles of mathematical analysis*. McGraw-Hill.
[14] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. On Convexity and Bounds of Fairness-aware Classification. In *WWW*.
[15] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*.
[16] Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. 2012. Functional mechanism: regression analysis under differential privacy. *PVLDB* 5, 11 (2012), 1364–1375.
[17] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *IJCAI*.
[18] Lu Zhang, Yongkai Wu, and Xintao Wu. 2018. Achieving Non-Discrimination in Prediction. In *IJCAI*.
[19] I. Žliobaite, F. Kamiran, and T. Calders. 2011. Handling Conditional Discrimination. In *ICDM*.