Computer Security Symposium 2020
26 - 29 October 2020

Toward Locally Private Logistic Regression with Missing Data

Hajime Ono[1,a)] Kazuhiro Minami[2]

Abstract: Missing values are prominent in many realistic datasets. In particular, this is a serious issue in a dataset that collects sensitive information on users, since the unwillingness of a user to answer a sensitive question often leads to missing piece of information in the dataset. In this paper, we study a privacy-preserving method based on local differential privacy (LDP) for performing logistic regression analysis on a dataset with some missing values. Since most previous research on LDP-based data analysis assumes that missing data does not exist in data, we consider two LDP logistic regression methods that are capable of handling missing values. We start with a naive approach of deleting records with missing values before conducting regression analysis, and then develop a two-phase method for obtaining unbiased estimates of regression coefficients. Our preliminary experiments with simulated data show that the latter method outperforms the naive approach when the missing rate of data and the privacy budget for LDP are relatively large.

Keywords: missing data, local differential privacy, logistic regression

1. Introduction

Logistic regression is one of the most popular methods for data analysis and is usually a starting point of any econometric analysis. In regression analysis, we study the relationship between explanatory variable $X \in \mathbb{R}$ and dependent variable $Y \in \{0, 1\}$. The target is to find parameters β_0^* and β_1^* such that

$$\Pr(Y = 1|X = x) = \frac{1}{-\exp(\beta_0 + \beta_1 x) + 1}.$$

Since β_1 explains how Y depends on X, learning the regressors β_0 and β_1 is equivalent to understand the relationship between X and Y. To find regressors for a given dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$, we choose $\hat{\beta}$ that minimizes cross entropy loss below:

$$\frac{1}{n}\sum_{i=1}^{n} -\Big(y_i \log(\sigma(\langle\beta, x_i'\rangle)) + (1 - y_i)\log(1 - \sigma(\langle\beta, x_i'\rangle)\Big),$$

where $x_i' = (1, x_i)$ for $i = 1, \ldots, n$. The minimizer of the cross entropy asymptotically approaches to the true regressor β^*. In this paper, we thus consider to perform the minimization of the cross entropy loss.

In this paper, we assume that an aggregator collects records from users (i.e., respondents) and performs regression analysis on them and that the users do not trust the aggregator not to disclose their sensitive attributes in their records. Thus, we employ local differential privacy(LDP) [1], which is a local version of differential privacy, to protect each user's sensitive information from an untrusted aggregator. There are several previous research on LDP-based regression [1], [2], [10] studying the lower bound of minimax errors of estimated regressors.

However, when a dataset contains missing values due to various reason, such as physical error and answer rejection [5], to perform regression analysis on such an incomplete dataset under LDP constraints raises a new technical challenge. Just not sending records with missing values to an aggregator is not an acceptable solution since missingness in a record implies the unwillingness of the user to provide sensitive information; that is, the denial of a query allows the aggregator to infer the user's sensitive information in the record.

To ensure LDP, each user needs to perturb his record values by adding a random noise to the original values before sending it to the aggregator. Therefore, if a user's record contains a missing variable, that user who wants to pretend to have the complete record, needs to pick some default value to be randomized. However, this approach has a risk of obtaining a poor biased estimator of the regressor, which does not converges to a true regressor β^*. The bias tends to be introduced when the probability of missingness in explanatory variable X depends on a value in dependent variable Y.

For example, suppose that Y is a categorical variable about the marital status of a person (i.e., a single or not) and that X is the amount of expenditures of that person. If singles are more likely be absent from home than non-singles [7], the singles are more likely to fail to participate

[1] School of Multidisciplinary Science Department of Statistical Science, SOKENDAI
[2] The Institute of Statistical Mathematics
[a)] hono@ism.ac.jp

in a public survey conducted by investigators visiting each household. If each user sets 0 for missing X and Y by default, which corresponds to a skewed sampling of weighing non-singles with relatively high expenditures, leads to a biased regressor.

Our goal, therefore, is to perform linear regression of incomplete data with local differential privacy by considering missingness in X's depending on Y to avoid obtaining biased estimators. We modify the stochastic gradient descent (SGD) algorithm [1] to support logistic regression while guaranteeing local differential privacy. In this scheme, each user iteratively provides an aggregator with the stocastic gradient of the logictic loss function in empirical cross entropy. Therefore, it is suitable to hide the fact that a user's record has missing values from the aggregator.

We first develop a simple *dummy* submission algorithm, in which a user sets a dummy value of 0 to the stocastic gradient when the data is missing. Although this scheme protects the privacy of users including those with missing values with LDP, the aggregator obtains an unbiased regressor with the dataset where records with missign values are removed. Therfore, we develop the *two-phase* regression algorithm, which models the mechanism for missingness explicitly. This algorithm consists of the preparation phase and main phase. We estimate parameters for the mechanism of missingness in the preparation phase, and then estimate unbiased regressors using the technique of inverse probability weighting (IPW) [6] in the main phase.

We compare the two algorithms theoretically and experimentally. We analyze utility without assuming any concrete algorithm. In the analysis, we derive an upper bound of excess risk, $\mathbb{E}\left[\mathcal{L}_{\mathrm{CE}}(\beta)\right] - \mathcal{L}_{\mathrm{CE}}(\beta^*)$, where the β is an output of an algorithm. Our theorem shows that the upper bound consist of the two terms which correspond to the variation and the bias of perturbed stochastic gradients, respectively and that the performance of the two algorithms varies depending on the parameter space in privacy budget for LDP and the size of a dataset. Our preliminary experiments with simulated data show that the two-phase regression method outperforms the approach of dummy submission when the missing rate of data and the privacy budget for LDP are relatively large.

2. Local Differential Privacy

Local differential privacy (LDP) [1], [3], [4] is a rigorous privacy definition for safe data collection. We suppose there are n users, each with private datum (i.e., record) R_i, $i = 1, \ldots, n$, and each user i communicates perturbed view $Z_i \in \mathcal{Z}$ of X_i. Communication between users and an aggregator is performed in T rounds. In the t-th round, user i communicates $Z_i^{(t)}$, which may depend on all previous communications. Let $Z_{\leq n} = (Z_1, \ldots, Z_n)$. The local differential privacy is defined as a property of the perturbed views.

Definition 1 (ϵ-LDP [3])**.** *The output $\mathbb{Z} \in \mathcal{Z}^{nT}$ is ϵ-LDP: for each $S \subset \mathcal{Z}^{nT}$ and pair of samples $r_{\leq n}, r'_{\leq n} \in X^n$ differing in at most a single element,*

$$\frac{\Pr(\mathbb{Z} \in S | R_{\leq n} = r_{\leq n})}{\Pr(\mathbb{Z} \in S | R_{\leq n} = r'_{\leq n})} \leq e^{\epsilon}. \tag{1}$$

Algorithm 1: Private Sampling [1] $\mathcal{Q}_{\mathrm{ps}}(\mathbf{v}; G, \epsilon)$

```
Input: vector $\mathbf{v} \in \mathbb{R}^d$, radius $G > 0$ and privacy parameter $\epsilon$
$B := r \frac{e^{\epsilon}+1}{e^{\epsilon}-1} \frac{\sqrt{\pi}}{2} \frac{d\Gamma(\frac{d-1}{2}+1)}{\Gamma(\frac{d}{2}+1)}$
$\bar{\mathbf{v}} := \begin{cases} +r\frac{\mathbf{v}}{\|\mathbf{v}\|_2} & \text{with probability } \frac{1}{2} + \frac{\|\mathbf{v}\|_2}{2r} \\ -r\frac{\mathbf{v}}{\|\mathbf{v}\|_2} & \text{with probability } \frac{1}{2} - \frac{\|\mathbf{v}\|_2}{2r} \end{cases}$
Sample $T \sim$ Bernoulli $\left(\frac{e^{\epsilon}}{e^{\epsilon}+1}\right)$
if $T = 1$ then
    Sample $Z \sim \mathrm{Unif}(\{z \in \mathbb{R}^d : \|z\|_2 = B, \langle \bar{v}, z \rangle > 0\})$
else
    Sample $Z \sim \mathrm{Unif}(\{z \in \mathbb{R}^d : \|z\|_2 = B, \langle \bar{v}, z \rangle \leq 0\})$
end
Output: $Z$
```

The definition says that any datum tied with one user is indistinguishable from the other candidate datum with any other users' data.

We remark that there are some variations of local differential privacy definition. Definition 1 is called fully interactive local differential privacy. In this paper, we just call it local differential privacy.

To satisfy LDP constraints, we choose *private sampling* [1] for vector submissions. Algorithm 1 is its pseudocode. The algorithm takes d-dimensional vector $\mathbf{v}$ and privacy parameter ϵ and assumes that $\|\mathbf{v}\| \leq G$ for some constant $G > 0$. The outputs are sampled from the hyper sphere whose radius is B. such that the inner product of the input and the output is positive in high probability. The inner product is negative with low probability,

We use the private sampling because of privacy and utility reasons. As for privacy, the private sampling satisfies the following inequality: for any subset S of the output domain and v, v' such that $\|\mathbf{v}\| \leq G$ and $\|v'\| \leq G$,

$$\frac{\Pr(\mathcal{Q}_{\mathrm{ps}}(\mathbf{v}; G, \epsilon) \in S)}{\Pr(\mathcal{Q}_{\mathrm{ps}}(\mathbf{v}'; G, \epsilon) \in S)} \leq e^{\epsilon}.$$

This inequality ensures that the private sampling outputs similar value for any input. This property helps us to construct an LDP algorithm.

In terms of utility, private sampling ensures unbiased outputs and affinity with convex optimization. The output $\mathcal{Q}_{\mathrm{ps}}(\mathbf{v}; G, \epsilon)$ satisfies $\mathbb{E}\left[\mathcal{Q}_{\mathrm{ps}}(\mathbf{v}; G, \epsilon)\right] = \mathbf{v}$. That is, if we take expectation of the output over the randomness for privacy, we obtain the input. Thanks to this property, some terms become 0. Thus, the unbiassed property makes utility analyses simpler. The affinity with convex optimization is that an optimizer using the private sampling achieves minimax optimality with $\epsilon \downarrow 0$ in convex optimizations [1]. Since logistic regression is a convex optimization problem, we expects that private sampling achieves better utility than the other randomizing mechanisms.

3. Problem Formulation

We formulate our problem of logistic regression with incomplete data under the LDP constraints.

We suppose there are n users and one aggregator. Each user i possesses datum $(x_i, y_i) \in [-1, +1] \times \{0, 1\}$ and wants to contribute to the data analysis but to hide the datum from the other parties. We assume that $\{x_i\}_{i=1}^n$ are sampled i.i.d., and each y_i is randomly generated with probability function

$$p(y_i|x_i) = \begin{cases} \sigma(\langle \beta^*, x_i' \rangle) & \text{if } y_i = 1, \\ 1 - \sigma(\langle \beta^*, x_i' \rangle) & \text{if } y_i = 0, \end{cases} \tag{2}$$

where $x_i' = (1 \quad x_i)^\top$, $\sigma : \mathbb{R} \to (0, 1); z \mapsto 1/(\exp(-z) + 1)$ and $\beta^* \in \mathbb{B}_{\sqrt{2}} = \{b \in \mathbb{R}^2 : \|b\|_2 \leq \sqrt{2}\}$ is an unknown parameter. For simplicity, we denote $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$. (x_i, y_i) and $\mathcal{D}$ correspond to r_i and $R_{\leq n}$ in Definition 1.

The aggregator aims to find the true regressor β^*. We utilize the property that β^* minimizes the cross entropy error $\mathcal{L}_{\mathrm{CE}}(\beta)$ defined as

$$\mathcal{L}_{\mathrm{CE}}(\beta) = \mathbb{E}_{X,Y}\left[\ell_{\mathrm{CE}}(\beta; X, Y)\right] \tag{3}$$

$$\text{where } \ell_{\mathrm{CE}}(\beta; x, y) = -\Big(y \log(\sigma(\langle \beta, x' \rangle)) + (1 - y) \log(1 - \sigma(\langle \beta, x' \rangle))\Big), \tag{4}$$

Since the true distributions of X and Y are unknown, we cannot evalute the cross entropy error. One of approaches to find β^* is to minimize the empirical cross entropy defined as follows:

$$\hat{\mathcal{L}}_{\mathrm{CE}}(\beta; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n \ell_{\mathrm{CE}}(\beta; x_i, y_i). \tag{5}$$

We also suppose that some of $\{x_i\}_{i=1}^n$ are missing. We use a variable m_i to indicate whether x_i for $i = 1, \ldots, n$ is missing such that:

$$m_i = \begin{cases} 0 & \text{if } x_i \text{ is observed,} \\ 1 & \text{if } x_i \text{ is missing.} \end{cases} \tag{6}$$

We focus on the case that m_i depends on y_i. Concretely, we assume that m_i follows the probability function $p(m_i|y_i)$ below.

$$\Pr(m_i = 1|y_i) = \sigma(\langle \alpha^*, y_i' \rangle). \tag{7}$$

where $\alpha^* \in \mathbb{B}_{\sqrt{2}}$ is an unknown parameter. The major difference between our problem and the standard logistic regression is that no one is able to evaluate the empirical cross entropy. The difference comes from the following two reasons. (i) some of $\{x_i\}_{i=1}^n$ are missing and (ii) the data are perturbed to satisfy the LDP constraint. We address the above two issues in this paper.

4. LDP-based Logistic Regression

We propose two algorithms for generating estimators β^* with incomplete data under the LDP constraints. In this section, we describe two algorithms and evaluate the biases of the estimators produced by the algorithms. We cover the utility analyses of the algorithms in Section 5.

4.1 Dummy Submission

A naive way to handle incomplete data is to ignore the data datum missing. Such a strategy is called *complete-case* analysis and is sometimes used for incomplete-data analyses. Here, we implement the complete-case analysis while satisfying the LDP constraints.

Since the LDP constraints require that a user's datum is indistinguishable from any possible candidate datum, the missingness in the datum also should be protected. To satisfy the LDP constraints, we consider the dummy submission strategy that a user submits dummy information if that user's datum contains missingness. Based on this idea, we consider a modified stochastic gradient descent (SGD) for incomplete data under the LDP constraints. Algorithm 2 is the pseudo-code. In the protocol, each user i computes stochastic gradient $\mathbf{g}_i$ if her datum is complete: otherwise she prepares the zero vector as a dummy. In the protocol, we denote the prepared vector as $\hat{\mathbf{g}_i}$. Then, the user makes $\tilde{\mathbf{g}}_i$ by perturbing $\hat{\mathbf{g}_i}$ via a randomizing mechanism of privacy sampling $\mathcal{Q}$. The aggregator updates the estimation using submitted $\tilde{\mathbf{g}}_i$ as follows:

$$\beta^{(i)} := \mathop{\arg\min}_{\beta \in \mathbb{B}_{\sqrt{2}}} \left\{ \eta_i \langle \tilde{\mathbf{g}}_i, \beta \rangle + \|\beta - \beta^{(i-1)}\|_2^2 / 2 \right\}. \tag{8}$$

This update rule realizes the projected stochastic gradient descent (SGD), which is one of the standard convex-optimization algorithms. At the end of the algorithm, the aggregator obtains $\beta^{(n)}$, an estimator of β^*. This optimization scheme is extension of the LDP convex optimizer proposed in [1] for our setting.

We confirm that Algorithm 2 ensures privacy for every user. To see that the privacy requirements of LDP are satisfied, it is sufficient to show, for any perturbed stochastic gradient $\tilde{\mathbf{g}}$ and any pair of stochastic gradients $\mathbf{g}, \mathbf{g}'$,

$$\frac{p(\tilde{\mathbf{g}}_i = \tilde{\mathbf{g}} | \mathbf{g}_i = \mathbf{g})}{p(\tilde{\mathbf{g}}_i = \tilde{\mathbf{g}} | \mathbf{g}_i = \mathbf{g}')} \leq e^{\epsilon}$$

holds. This inequality is derived directory from the property of private sampling.

We now analyze the property of the output of Algorithm 2. First, we show that $\beta^{(n)}$ is a biased estimator of β^*. That is, in Algorithm 2, $\beta^{(n)}$ does not converge to β^* with $n \to +\infty$. $\beta^{(n)}$ converge to the global minimum of $\mathcal{L}_{\mathrm{DS}}(\beta)$ defined as follows:

$$\begin{aligned} \mathcal{L}_{\mathrm{DS}}(\beta) &\equiv \int (1 - m) \ell_{\mathrm{CE}}(\beta; x, y)\, p(x, y) dx dy \\ &= \mathcal{L}_{\mathrm{CE}}(\beta) - \int m \ell_{\mathrm{CE}}(\beta; x, y)\, p(x, y) dx dy \\ &= \mathcal{L}_{\mathrm{CE}}(\beta) - \int \Pr(m = 1|y) \ell_{\mathrm{CE}}(\beta; x, y)\, p(x, y) dx dy. \end{aligned}$$

To see that the above function has the different global minimum from that in Eq. (5), it is sufficient to show

Algorithm 2: Dummy Submission

```
Input: privacy parameter $\epsilon$, learning rate $\eta$
Initialize $\beta^{(0)}$
for $i = 1$ to $n$ do
    // begin user local process
    if $m_i == 0$ then
        Compute $\ell_{\mathrm{CE}}\left(\beta^{(i-1)}; x_i, y_i\right)$
        $\hat{\mathbf{g}_i} := \nabla_\beta \ell_{\mathrm{CE}}\left(\beta^{(i-1)}; x_i, y_i\right)$
    else
        Set zero vector to $\hat{\mathbf{g}_i}$
    end
    $\tilde{\mathbf{g}}_i := \mathcal{Q}(\hat{\mathbf{g}_i}; \sqrt{2}, \epsilon)$
    Submit $\tilde{\mathbf{g}}_i$ to aggregator
    // end local process
    $\beta^{(i)} := \arg\min_{\beta \in \mathbb{B}_{\sqrt{2}}} \{\eta \langle \tilde{\mathbf{g}}_i, \beta \rangle + \|\beta - \beta^{(i-1)}\|_2^2/2\}$
end
Output: $\beta^{(n)}$
```

$\nabla_\beta \mathcal{L}_{\mathrm{DS}}(\beta^*) \neq 0$.

$$\begin{aligned}&\nabla_\beta \mathcal{L}_{\mathrm{DS}}\left(\beta^*\right)\\ =&\nabla_\beta \mathcal{L}_{\mathrm{CE}}\left(\beta^*\right) - \int \Pr(m=1|y) \nabla_\beta \ell_{\mathrm{CE}}\left(\beta; x, y\right) dxdy\\ =&-\int \Pr(m=1|y) \nabla_\beta \ell_{\mathrm{CE}}\left(\beta; x, y\right) dxdy.\end{aligned}$$

It also holds that

$$\begin{aligned}\nabla_\beta \mathcal{L}_{\mathrm{DS}}\left(\beta^*\right) &= \int (1-m) \nabla_\beta \ell_{\mathrm{CE}}\left(\beta^*; x, y\right) p(x,y) dxdy\\ &= \int \Pr(m=0|y) \nabla_\beta \ell_{\mathrm{CE}}\left(\beta^*; x, y\right) p(x,y) dxdy.\end{aligned}$$

In order for $\nabla_\beta \mathcal{L}_{\mathrm{DS}}(\beta^*) = 0$ to be true for any distribution of X, $\Pr(m=1|y) = \Pr(m=0|y)$ must hold. Thus, if $\alpha_1^* \neq 0$, there is a distribution of X such that $\nabla_\beta \mathcal{L}_{\mathrm{DS}}(\beta^*) \neq 0$. We note that this property holds independently of any randomizing mechanism; Even if we use an alternative mechanism instead of private sampling, the estimator is still biased.

4.2 Two-phase Regression

To reduce a bias in estimator β^*, we consider an alternative algorithm. The reason that the dummy submission algorithm makes the estimator biased is ignoring the missingness mechanism that describes the probabilistic dependence of missingness in explanatory variable X on dependent variable Y. Thus, we propose the method of explicitly modeling the missingness mechanism before estimating β^*. In this approach, the algorithm estimate α^*, which decides the missing probability, before β^*. we can reduce bias in estimator β^* by utilizing the estimated missingness mechanism α^*.

We develop *two-phase* regression algorithm, which consists of preparation phase and main phase. In the preparation phase, the algorithm estimates α^*. The algorithm estimates α^* by logistic regression with the objective function

$$\begin{aligned}&\hat{\mathcal{L}}_{\mathrm{CE}}\left(\alpha; \{y_i\}_{i=1}^n, \{m_i\}_{i=1}^n\right)\\ =&\sum_{i=1}^n m_i \log(\sigma(\langle \alpha, y_i' \rangle))\\ &+ (1-m_i)\log(1-\sigma(\langle \alpha, y_i' \rangle)).\end{aligned}$$

In the main phase, the algorithm constructs estimator β^* utilizing the estimated α^*. As an implementation of the main phase, we employ inverse probability weighting(IPW) [6], which is one of the standard methods used for incomplete data analyses [5].

The IPW method solves the weighted empirical loss minimization whose objective function is

$$\sum_{i=1}^n \frac{1-m_i}{p(m_i=0|y_i,\hat{\alpha})} \ell_{\mathrm{CE}}\left(\beta; x_i, y_i\right) \qquad (9)$$

where $p(m_i=0|y_i,\hat{\alpha})$ is the estimated probability estimated in the preparation phase. In this objective function, for each $i \in [n]$, datum (x_i, y_i) is weighted by the inverse of estimated observed probability $p(m_i=0|y_i)$ if $m_i=0$. Otherwise, datum (x_i, y_i) is ignored. We assume $0 < p(m_i=0|y_i,\hat{\alpha}) < 1$ for all $i = 1, \ldots, n$.

We implement the two-phase-regression algorithm with local differential privacy. Algorithm 3 is the pseudo-code. To guarantee ϵ-LDP, in each phase, each user consumes only $\epsilon/2$ of the privacy budget, respectively. In the preparation phase, the algorithm performs logistic regression to estimate α^* by SGD with randomizing mechanism $\mathcal{Q}$ consuming privacy parameter $\epsilon/2$. After the preparation phase, estimation $\alpha^{(n)}$ of α^* is obtained. Using $\alpha^{(n)}$, each user estimate missing probability $\Pr(m_i|y_i,\alpha^*)$ as $p(m_i|y_i,\alpha^{(n)})$. Then, the algorithm solves the minimization of the weighted empirical loss function defined in Eq. (9) to estimate β^* by SGD. To perform this minimization problem, each user i weights her stochastic gradient by inverse of estimated missingness probability $1/p(m_i=0|y_i,\alpha^{(n)})$, if $m_i=0$. Otherwise, she prepares the zero vector. The weighted stochastic gradient is perturbed by $\mathcal{Q}$ consuming privacy parameter $\epsilon/2$ and is submitted to the aggregator. Using the submitted noisy stochastic gradients, the aggregator obtains an estimation of β^*.

Before utility analysis, we confirm that output $\beta^{(n)}$ of Algorithm 3 guarantees ϵ-LDP. To see that, it is sufficient to show that it holds, for each $i \in [n]$, any preparation-phase output $\tilde{\mathbf{h}}$ and main-phase output $\tilde{\mathbf{g}}$, any stochastic gradients $\mathbf{h}, \mathbf{h}', \mathbf{g}, \mathbf{g}'$,

$$\begin{aligned}&\frac{p(\tilde{\mathbf{h}}_i = \tilde{\mathbf{h}}, \tilde{\mathbf{g}}_i = \tilde{\mathbf{g}} | \mathbf{h}_i = \mathbf{h}, \mathbf{g}_i = \mathbf{g})}{p(\tilde{\mathbf{h}}_i = \tilde{\mathbf{h}}, \tilde{\mathbf{g}}_i = \tilde{\mathbf{g}} | \mathbf{h}_i = \mathbf{h}', \mathbf{g}_i = \mathbf{g}')}\\ =&\frac{p(\tilde{\mathbf{h}}_i = \tilde{\mathbf{h}} | \mathbf{h}_i = \mathbf{h})}{p(\tilde{\mathbf{h}}_i = \tilde{\mathbf{h}} | \mathbf{h}_i = \mathbf{h}')} \times \frac{p(\tilde{\mathbf{g}}_i = \tilde{\mathbf{g}} | \tilde{\mathbf{h}}_i = \tilde{\mathbf{h}}, \mathbf{h}_i = \mathbf{h}, \mathbf{g}_i = \mathbf{g})}{p(\tilde{\mathbf{g}}_i = \tilde{\mathbf{g}} | \tilde{\mathbf{h}}_i = \tilde{\mathbf{h}}, \mathbf{h}_i = \mathbf{h}', \mathbf{g}_i = \mathbf{g}')}\\ =&\frac{p(\tilde{\mathbf{h}}_i = \tilde{\mathbf{h}} | \mathbf{h}_i = \mathbf{h})}{p(\tilde{\mathbf{h}}_i = \tilde{\mathbf{h}} | \mathbf{h}_i = \mathbf{h}')} \times \frac{p(\tilde{\mathbf{g}}_i = \tilde{\mathbf{g}} | \mathbf{g}_i = \mathbf{g})}{p(\tilde{\mathbf{g}}_i = \tilde{\mathbf{g}} | \mathbf{g}_i = \mathbf{g}')}\\ \leq& e^{\epsilon/2} e^{\epsilon/2} = e^{\epsilon}.\end{aligned}$$

The last inequality holds because of property of private sampling.

Algorithm 3: Two-phase Regression

```
Input: privacy parameter $\epsilon$, learning rate $\eta_p, \eta_m$
Split $\epsilon$ into $\epsilon_1$ and $\epsilon_2$ such that $\epsilon_1 + \epsilon_2 == \epsilon$
// Preparation phase
Initialize $\alpha^{(0)}$
for $i = 1$ to $n$ do
    // begin user local process
    Compute $\ell_{\mathrm{CE}}\left(\alpha^{(i-1)}; y_i, m_i\right)$
    $\mathbf{h}_i := \nabla_\alpha \ell_{\mathrm{CE}}\left(\alpha^{(i-1)}; y_i, m_i\right)$
    $\tilde{\mathbf{h}}_i := \mathcal{Q}(\mathbf{h}_i; \sqrt{2}, \epsilon_1)$
    Submit $\tilde{\mathbf{h}}_i$ to aggregator
    // end local process
    $\alpha^{(i)} := \arg\min_{\alpha \in \mathcal{A}} \left\{\eta_p \langle \mathbf{h}_i, \alpha \rangle + \|\alpha - \alpha^{(i-1)}\|_2^2/2\right\}$
end
$p_{\min} := \min_{y \in [-1,+1]} 1 - \frac{1}{\exp(-\alpha_0^{(n)} - \alpha_1^{(n)} y)}$
// Main phase
Initialize $\beta^{(0)}$
for $i = 1$ to $n$ do
    // begin user local process
    if $m_i == 0$ then
        Compute $\ell_{\mathrm{CE}}\left(\beta^{(i-1)}; x_i, y_i\right)$
        $\hat{\mathbf{g}_i} := \nabla_\beta \ell_{\mathrm{CE}}\left(\beta^{(i-1)}; x_i, y_i\right) / \sigma(-\left\langle \alpha^{(n)}, y'_n \right\rangle)$
    else
        $\hat{\mathbf{g}_i} := 0$
    end
    $\tilde{\mathbf{g}}_i := \mathcal{Q}(\hat{\mathbf{g}_i}; \sqrt{2}/p_{\min}, \epsilon_2)$
    // end local process
     $\beta^{(i)} := \arg\min_{\beta \in \mathcal{B}} \left\{\eta_m \langle \tilde{\mathbf{g}}_i, \beta \rangle + \|\beta - \beta^{(i-1)}\|_2^2/2\right\}$
end
Output: $\beta^{(n)}$
```

We analyze the properties of $\beta^{(n)}$. The most important property is that $\beta^{(n)}$ can converge to β with $n \to +\infty$, if $\alpha^{(n)} = \alpha^*$. We show that β^* is the global minimum of $\mathcal{L}_{\mathrm{TP}}()$. To see that it is sufficient to show

$$\mathbb{E}_{M,Y}\left[\frac{1-M}{p(m=0|y,\alpha^*)} \nabla \ell_{\mathrm{CE}}\left(\beta^*; X, y\right) \middle| X = x\right] = 0$$

since $\ell_{\mathrm{CE}}(;)$ is convex. We can see that the above equality holds as follows.

$$\begin{aligned}
&\mathbb{E}\left[\frac{1-m}{p(m=0|y)} \nabla \ell_{\mathrm{CE}}\left(\beta^*; x, y\right) \middle| x\right] \\
=& \int \frac{1-m}{p(m=0|y,\alpha^*)} \nabla \ell_{\mathrm{CE}}\left(\beta^*; x, y\right) f(y|x, m=0) dy \\
=& \int \frac{f(m=0)}{f(m=0|y)} \nabla \ell_{\mathrm{CE}}\left(\beta^*; x, y\right) \frac{f(m=0|y) f(y|x)}{f(m=0)} dy \\
=& \int \nabla \ell_{\mathrm{CE}}\left(\beta^*; x, y\right) p(y|x) dy = 0.
\end{aligned}$$

This property is a significant advantage of this algorithm over the dummy submission algorithm. We note that the unbiased property is guaranteed only when $\alpha^{(n)} = \alpha^*$.

5. Utility Analysis

Here, we analyze the utilities of the proposals. Our target is to characterize the two estimators outputted by the proposals. We compare the excess risk defined as $\mathbb{E}_{\mathcal{D},\mathbf{q}}\left[\mathcal{L}_{\mathrm{CE}}(\beta)^{(n)}\right] - \mathcal{L}_{\mathrm{CE}}(\beta^*)$ where $\mathbf{q} = (q_1, \ldots, q_n)$ is the randomness for the privacy protection. We upper bound the excess risk. Before going into the individual analyses, we show a lemma that is commonly used in both analyses. The lemma is an extension of a standard SGD convergence analysis for convex objective functions.

To derive a general utility guarantee, we consider a general locally private SGD procedure:

$$\begin{aligned}
\beta^{(i)} :=& \Pi_{\mathcal{B}}\left(\beta^{(i-1)} - \eta_i \tilde{\mathbf{g}}_i\right) & (10) \\
=& \arg\min_{\beta \in \mathcal{B}} \left\{\eta_i \langle \beta, \tilde{\mathbf{g}}_i \rangle + \frac{1}{2}\|\beta - \beta^{(i-1)}\|^2\right\}. & (11)
\end{aligned}$$

This procedure contains Algorithm 2 and Algorithm 3. We set learning rate $\eta_i = c/\sqrt{i}$ with some constant c as standard SGD for convex objectives. In this update rule, $\tilde{\mathbf{g}}_i$ is some perturbed $\mathbf{g}_i$. As discussed in the above subsection, it does not necessary hold $\mathbb{E}_{q_i}[\tilde{\mathbf{g}}_i] = \mathbf{g}_i$.

For the general LDP SGD, we show the following utility theorem.

Theorem 1. *We assume $\|\tilde{\mathbf{g}}_i\| \leq B$, $\|\mathcal{B}\|_2 \leq D$, and $\|\mathbb{E}_{x_i,y_i}[\mathbb{E}_{q_i}[\tilde{\mathbf{g}}_i] - \mathbf{g}_i]\| < b$ for $i = 1, \ldots, n$. Using update rule (11), after n-times update, we have*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D},\mathbf{q}}\left[\mathcal{L}_{CE}\left(\beta^{(n)}\right) - \mathcal{L}_{CE}\left(\beta^*\right)\right] & (12) \\
\leq& \left(\frac{D^2}{c} + cB^2\right) \frac{2 + \log(n)}{\sqrt{n}} + bD\log(n+1). & (13)
\end{aligned}$$

The upper bound consist of two terms, and we refer to the first and the second terms by the variance term and the bias term, respectively. The variance term is determined by the noise scale B of perturbation or the scale D of the parameter set and shrinks by $1/\sqrt{n}$. The bias term is determined by the parameter set scale D and the bias b of the perturbed stochastic gradient. Unlike the variance term, the bias term does not shrink by n. These fact implies that unbiased estimators of β^* is not always better than biased estimators. Even if the estimator is unbiased, the variance term can make the estimator poor. In our setting, since the sample size is finite and we should satisfy the LDP constraint, it is difficult to make both the variance and the bias terms small simultaneously. Thus, there is a trade-off relationship between the variance and the bias terms.

Now, we show Theorem 1. We first prove the lemma below, which we will use repeatedly to obtain the theorem.

Lemma 1. *For some $i = [n-1]$,*

$$\begin{aligned}
&\sum_{i=n-k}^{n}\left(\mathbb{E}\left[\mathcal{L}_{CE}\left(\beta^{(i)}\right)\right] - \mathcal{L}_{CE}(\beta)\right) & (14) \\
\leq& \frac{1}{2\eta_{n-k+1}} \mathbb{E}\left[\|\beta^{(n-k)} - \beta\|^2\right] & (15) \\
&+ \sum_{i=n-k+1}\left(\frac{1}{2\eta_{i+1}} - \frac{1}{2\eta_i}\right) \mathbb{E}\left[\|\beta^{(i+1)} - \beta\|^2\right] & (16) \\
&+ \frac{B^2}{2} \sum_{i=n-k}^{n} \eta_{i+1}. & (17)
\end{aligned}$$

Proof. The proof relies on the convexity of $\mathcal{L}_{\mathrm{CE}}()$. From the convexity, for any $\beta \in \mathcal{B}$ and each $i \in [n]$, we have the following inequality:

$$\begin{aligned} &\mathcal{L}_{\mathrm{CE}}\left(\beta^{(i)}\right) - \mathcal{L}_{\mathrm{CE}}\left(\beta\right) \\ \le &\left\langle \nabla \mathcal{L}_{\mathrm{CE}}\left(\beta^{(i)}\right), \beta^{(i)} - \beta \right\rangle \\ =&\mathbb{E}\left[\left\langle \mathbf{g}_{i+1}, \beta^{(i)}-\beta\right\rangle\right] \\ =&\mathbb{E}\left[\left\langle \tilde{\mathbf{g}}_{i+1}, \beta^{(i)}-\beta\right\rangle\right] - \mathbb{E}\left[\left\langle \tilde{\mathbf{g}}_{i+1}-\mathbf{g}_{i+1}, \beta^{(i)}-\beta\right\rangle\right]. \end{aligned} \tag{18}$$

The first term of (18) is the origin of the variance term, and the second term is the origin of the bias term. This inequality is still hard to interpret. We upper bound the two term in inequality (18).

First, we upper bound $\mathbb{E}\left[\left\langle \tilde{\mathbf{g}}_{i+1}, \beta^{(i)}-\beta\right\rangle\right]$. By convexity of $\mathbb{B}_{\sqrt{2}}$, for any $\beta \in \mathbb{B}_{\sqrt{2}}$, we have the following inequality:

$$\begin{aligned} &\mathbb{E}\left[\|\beta_{i+1}-\beta\|^2\right] \\ =&\mathbb{E}\left[\|\Pi_{\mathbb{B}_{\sqrt{2}}}(\beta_i - \eta_{t+1}\tilde{\mathbf{g}}_{i+1}) - \beta\|^2\right] \\ \le&\mathbb{E}\left[\|\beta_i - \eta_{t+1}\tilde{\mathbf{g}}_{i+1} - \beta\|^2\right] \\ =&\mathbb{E}\left[\|\beta_i - \beta\|^2\right] - 2\eta_{t+1}\mathbb{E}\left[\langle \tilde{\mathbf{g}}_{i+1}, \beta_i - \beta\rangle\right] \\ &+ \eta_{t+1}^2\mathbb{E}\left[\|\tilde{\mathbf{g}}_{i+1}\|^2\right]. \end{aligned}$$

Rearranging the above inequality, we obtain the next inequality:

$$\begin{aligned} &\mathbb{E}\left[\langle \tilde{\mathbf{g}}_{i+1}, \beta_i - \beta\rangle\right] \\ \le&\frac{1}{2\eta_{t+1}}\mathbb{E}\left[\|\beta_i - \beta\|^2\right] - \frac{1}{2\eta_{t+1}}\mathbb{E}\left[\|\beta_{i+1} - \beta\|^2\right] \\ &+ \eta_{t+1}^2\mathbb{E}\left[\|\tilde{\mathbf{g}}_{i+1}\|^2\right] \\ \le&\frac{1}{2\eta_{t+1}}\mathbb{E}\left[\|\beta_i - \beta\|^2\right] - \frac{1}{2\eta_{t+1}}\mathbb{E}\left[\|\beta_{i+1} - \beta\|^2\right] \\ &+ \frac{\eta_{i+1}}{2}B^2. \end{aligned} \tag{19}$$

Second, we bound $-\mathbb{E}\left[\left\langle \tilde{\mathbf{g}}_{i+1}-\mathbf{g}_{i+1}, \beta^{(i)}-\beta\right\rangle\right]$. By the Cauchy–Schwarz inequality, we have

$$\begin{aligned} &-\mathbb{E}\left[\left\langle \tilde{\mathbf{g}}_{i+1}-\mathbf{g}_{i+1}, \beta^{(i)}-\beta\right\rangle\right] \\ =&-\left\langle \mathbb{E}\left[\tilde{\mathbf{g}}_{i+1}-\mathbf{g}_{i+1}\right], \beta^{(i)}-\beta\right\rangle \\ \le&\|\mathbb{E}\left[\tilde{\mathbf{g}}_{i+1}-\mathbf{g}_{i+1}\right]\|\|\beta^{(i)}-\beta\| \le bD. \end{aligned} \tag{20}$$

Plugging (19) and (20) into (18),

$$\begin{aligned} &\mathcal{L}_{\mathrm{CE}}\left(\beta^{(i)}\right) - \mathcal{L}_{\mathrm{CE}}\left(\beta\right) \\ \le&\frac{1}{2\eta_{t+1}}\mathbb{E}\left[\|\beta_i - \beta\|^2\right] - \frac{1}{2\eta_{t+1}}\mathbb{E}\left[\|\beta_{i+1} - \beta\|^2\right] \\ &+ \frac{\eta_{i+1}}{2}B^2. \end{aligned} \tag{21}$$

Summing (21) up over $i = n-k, \dots, n$, we have

$$\begin{aligned} &\sum_{i=n-k}^{n}\left(\mathbb{E}\left[\mathcal{L}_{\mathrm{CE}}\left(\beta^{(i)}\right)\right] - \mathcal{L}_{\mathrm{CE}}\left(\beta\right)\right) \\ \le&\frac{1}{2\eta_{n-k+1}}\mathbb{E}\left[\|\beta^{(n-k)} - \beta\|^2\right] \\ &+ \sum_{i=n-k+1}\left(\frac{1}{2\eta_{i+1}} - \frac{1}{2\eta_i}\right)\mathbb{E}\left[\|\beta^{(i+1)} - \beta\|^2\right] \\ &+ \frac{B^2}{2}\sum_{i=n-k}^{n}\eta_{i+1}. \end{aligned}$$

□

Here, we back to the proof of Theorem 1. Substituting $\eta_{i+1} = c/\sqrt{i}$ and $\beta = \beta^{(n-k)}$ into inequality (17),

$$\begin{aligned} &\mathbb{E}\left[\sum_{i=n-k}^{n}\mathcal{L}_{\mathrm{CE}}\left(\beta^{(i)}\right) - \mathcal{L}_{\mathrm{CE}}\left(\beta^{(n-k)}\right)\right] \\ \le&\left(\frac{D^2}{2c} + cB^2\right)\left(\sqrt{n} - \sqrt{n-k-1}\right) + \sum_{i=n-k+1}^{n} bB \\ =&\left(\frac{D^2}{2c} + cB^2\right)\frac{k+1}{\sqrt{n} + \sqrt{n-k-1}} + kbD \\ \le&\left(\frac{D^2}{2c} + cB^2\right)\frac{k+1}{\sqrt{n}} + kbD. \end{aligned} \tag{22}$$

Let $S_k = \frac{1}{k+1}\sum_{i=n-k}^{n}\mathbb{E}\left[\mathcal{L}_{\mathrm{CE}}\left(\beta^i\right)\right]$. Then,

$$\begin{aligned} &\frac{1}{k+1}\mathbb{E}\left[\sum_{i=n-k}^{n}\mathcal{L}_{\mathrm{CE}}\left(\beta^{(i)}\right) - \mathcal{L}_{\mathrm{CE}}\left(\beta^*\right)\right] \\ =&S_k - \mathbb{E}\left[\mathcal{L}_{\mathrm{CE}}\left(\beta^{(n-k)}\right)\right]. \end{aligned}$$

Combining this equation and inequality (22), we have

$$-\mathbb{E}\left[\mathcal{L}_{\mathrm{CE}}\left(\beta^{(n-k)}\right)\right] \le -S_k + \frac{D^2/2c + cB^2}{\sqrt{n}} + \frac{kbD}{k+1}.$$

From the above inequality and the definition of S_k, we derive the following inequality.

$$\begin{aligned} kS_{k-1} =&(k+1)S_k - \mathbb{E}\left[\mathcal{L}_{\mathrm{CE}}\left(\beta^{(n-k)}\right)\right] \\ \le&(k+1)S_k - S_k + \frac{D^2/2c + cB^2}{\sqrt{n}} + \frac{kbD}{k+1} \\ \le&kS_k + \frac{D^2/2c + cB^2}{\sqrt{n}} + \frac{kbD}{k+1}. \end{aligned} \tag{23}$$

Dividing both sides by k,

$$S_{k-1} \le S_k + \frac{D^2/2c + cB^2}{k\sqrt{n}} + \frac{bD}{k+1}.$$

Expanding this inequality repeatedly,

$$\begin{aligned} \mathbb{E}\left[\mathcal{L}_{\mathrm{CE}}\left(\beta^{(n)}\right)\right] = S_0 \le&S_{n-1} + \frac{D^2/2c + cB^2}{\sqrt{n}}\sum_{k=1}^{n-1}\frac{1}{k} \\ &+ bD\sum_{k=1}^{n-1}\frac{1}{k+1}. \end{aligned} \tag{24}$$

$$S_{n-1} - \mathcal{L}_{\mathrm{CE}}\left(\beta^*\right) = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} \mathcal{L}_{\mathrm{CE}}\left(\beta^{(i)} - \mathcal{L}_{\mathrm{CE}}\left(\beta^*\right)\right)\right] \tag{25}$$

$$\leq \frac{D^2/c + cB^2}{\sqrt{n}}. \tag{26}$$

Since $\sum_{k=1}^{n-1} 1/k \leq (1+\log(n))$, we complete the proof of Theorem 1.

5.1 Variance Terms

We compare the variance terms of the two algorithms in Section 4. Since the variance terms are identical except for B, it is sufficient to compare B for the discussion. An algorithm with greater B has a greater variance term.

In Algorithm 2, B is immediately obtained from the property of private sampling as follows:

$$\|\tilde{\mathbf{g}}_i\| = G\frac{e^\epsilon + 1}{e^\epsilon - 1}\frac{\sqrt{\pi}}{2}\frac{d\Gamma(\frac{d-1}{2}+1)}{\Gamma(\frac{d}{2}+1)} = B, \tag{27}$$

where G is a constant such that $\|\mathbf{g}_i\| \leq G$ for any $\mathbf{g}_i$. In Algorithm 3, B is obtained in sequence with the above:

$$\begin{aligned}&\|\tilde{\mathbf{g}}_i\| \leq G\frac{e^{\epsilon/2} + 1}{e^{\epsilon/2} - 1}\frac{\sqrt{\pi}}{2}\frac{d\Gamma(\frac{d-1}{2}+1)}{\Gamma(\frac{d}{2}+1)}\times\\ &\max\{\frac{1}{\Pr(m_i=0|y_i=0,\alpha^{(n)})}, \frac{1}{\Pr(m_i=0|y_i=1,\alpha^{(n)})}\}\\ &= B^2\end{aligned}$$

Comparing these B, we can see that this variance term in the two-phase regression is always greater than that of the dummy submission. This relationship is immediately derived from the following two reasons. One is the monotonic diminution of $\frac{e^\epsilon+1}{e^\epsilon-1}$. Another is that the out of the max operator is greater than or equal to 1.

5.2 Bias Terms

We next compare the bias terms of the proposals. Since the bias terms are identical except for b, it is sufficient to compare b.

In Algorithm 2, b is derived as

$$\begin{aligned}\|\mathbb{E}\left[\tilde{\mathbf{g}}_i - \mathbf{g}_i\right]\| &= \|\int \Pr(m_i=1|y_i,\alpha^*)\mathbf{g}_i p(x_i,y_i)dxdy\|\\ &\leq G\max\{\Pr(m_i=1|y_i=-1,\alpha^*),\\ &\qquad \Pr(m_i=1|y_i=+1,\alpha^*)\} \qquad (28)\\ &= b.\end{aligned}$$

This upper bound is always greater than 0 independent of ϵ and n. In Algorithm 3, b is derived as

$$\begin{aligned}&\|\mathbb{E}\left[\tilde{\mathbf{g}}_i - \mathbf{g}_i\right]\|\\ &\quad = \|\int\left(\frac{\delta(m_i=0)}{\Pr(m_i=0|y_i,\alpha^{(n)})} - 1\right)\mathbf{g}_i p(x_i,y_i)dxdy\|\\ &\quad = \|\int\left(\frac{\Pr(m_i=0|y_i,\alpha^*)}{\Pr(m_i=0|y_i,\alpha^{(n)})} - 1\right)\mathbf{g}_i p(x_i,y_i)dxdy\|\\ &\leq G\max\left\{\left|\frac{\Pr(m_i=0|y_i=+1,\alpha^*)}{\Pr(m_i=0|y_i=+1,\alpha^{(n)})} - 1\right|,\right.\\ &\qquad\left.\left|\frac{\Pr(m_i=0|y_i=-1,\alpha^*)}{\Pr(m_i=0|y_i=-1,\alpha^{(n)})} - 1\right|\right\}\\ &\quad = b.\end{aligned}$$

Unlike (28), the bias term of the two-phase regression can be 0. This is one of the advantages of the two phase regression. Since b of the two phase regression depends on the accuracy of estimator $\alpha^{(n)}$ of α^*, the b can be significantly greater than 1 when $\alpha^{(n)}$ is poor. Thus, if ϵ or n is sufficiently small, b in the two-phase regression can be larger than b in the dummy submission.

Summarizing the analyses of the variance and the bias terms, we obtain the following findings. (i) When ϵ or n is small, the dummy submission has less cross entropy loss than the two-phase regression. (ii) When ϵ and n are large, you should use the dummy submission. We evaluate this intuitions by numerical observations in Section 6.

6. Numerical Evaluation

Here, we evaluate our findings that, when ϵ or n is small, the dummy submission is better and that, when ϵ and n is large, the two-phase regression is better. We consider two measures for $\beta^{(n)}$: ℓ_2 norm $\|\beta^{(n)} - \beta^*\|$ and empirical excess risk $\hat{\mathcal{L}}_{\mathrm{CE}}\left(\beta^{(n)};\mathcal{D}\right) - \hat{\mathcal{L}}_{\mathrm{CE}}\left(\beta^*;\mathcal{D}\right)$. We randomly generate $\mathcal{D}$ ten times for each set of parameters and compute the average of the measures over ten trials.

We perform numerical experiments in the low missing rate case and the high missing rate case where we set $\alpha^* = (1,1)$ and $(0,3)$, respectively. Variable x_i has a missing value when $y_i = 1$ with higher probability in the high missing rate case than in the low missing rate case. Comparing the two cases, we also observe how α^* affects the estimations of β^*, The other parameters are set to the same values in both cases: $n = 100,000, \epsilon \in \{0.1, 1, 10\}, \beta^* = (0, 1.)$. We choose the sufficiently large n and large and small values for ϵ. We consider that these parameter set is suitable for evaluating our hypothesis.

First, we perform experiments in the low missing rate case. Table 1 shows the mean with the standard deviation of $\|\beta^{(n)} - \beta^*\|$. Table 2 shows the mean with the standard deviation of observed with the same parameters as in Table 1. In these tables, the differences between the two algorithms are less than the standard deviations. So, we cannot determine which algorithm is better in this settings. This does not agree with our intuition.

Next, we perform experiments in the high missing rate case. Table 3 shows the mean with the standard deviation of $\|\beta^{(n)} - \beta^*\|$. Table 4 shows the mean with the standard deviation of observed with the same parameters as in Table 3. With $\epsilon = 1$ or 10, the two phase regression is significantly better than the dummy submission on the both measures. This results agree with our intuitions.

Summarizing those results in the two cases, we developed a new hypothesis that the scale of α^* strongly affects the estimation accuracy. The large value for ϵ_1^* may cause a significant difference between the estimators produced by the dummy submission and the two-phase regression.

7. Related Work

We discuss related work on regression of incomplete data, regression with local differential privacy, and incomplete data handling with local differential privacy, clarifying the difference between these studies and our work.

Regression with missing data is an important topic, which has been studied over long time [5]. Previous research covers various mechanisms for missingness. In this paper, we consider the case that the missingness in explanatory variable X probabilistically depends on the value in a dependent variable Y.

Regression with local differential privacy attracts attention in [1], [2], [10]. Duchi et al. show minimax lower bounds of regression problems with sequential-interactive local differential privacy [1], [2]. Wang et al. study sparse regression with local differential privacy. Since their theory does not assume that the data are incomplete [10], it is not applicable to the utility analysis in our setting.

There exist some LDP studies to obtain the mean of incomplete data [8], [9]. Sun et al. proposed a scheme based on the techinique of locally private matrix factorization [8]. Sun et al. propose the algorithm to obtain mean ignoring missing values [9]. However, they implicitly assumes that missingness occurs independent from data. Thus, we expect that their proposals induce bias into estimators in our setting where the missingness occurs depending on data.

method	$\epsilon = 0.1$	$\epsilon = 1.0$	$\epsilon = 10.0$
dummy	0.084 ± 0.251	0.012 ± 0.038	0.022 ± 0.066
two phase	0.088 ± 0.265	0.016 ± 0.049	0.016 ± 0.049

Table 1 Averages of $\|\beta^{(n)} - \beta^*\|$ over ten trials, with low missing rate data.

method	$\epsilon = 0.1$	$\epsilon = 1.0$	$\epsilon = 10.0$
dummy	0.075 ± 0.044	0.011 ± 0.014	0.002 ± 0.002
two phase	0.085 ± 0.074	0.020 ± 0.024	0.004 ± 0.005

Table 2 Averages of $\hat{\mathcal{L}}_{\mathrm{CE}}\left(\beta^{(n)}; \mathcal{D}\right) - \hat{\mathcal{L}}_{\mathrm{CE}}\left(\beta^*; \mathcal{D}\right)$ over ten trials, with low missing rate data.

method	$\epsilon = 0.1$	$\epsilon = 1.0$	$\epsilon = 10.0$
dummy	0.090 ± 0.270	0.057 ± 0.173	0.056 ± 0.170
two phase	0.117 ± 0.352	0.019 ± 0.056	0.004 ± 0.013

Table 3 Averages of $\|\beta^{(n)} - \beta^*\|$ over ten trials, with high missing rate data.

method	$\epsilon = 0.1$	$\epsilon = 1.0$	$\epsilon = 10.0$
dummy	0.060 ± 0.030	0.038 ± 0.021	0.032 ± 0.016
two phase	0.083 ± 0.061	0.021 ± 0.032	0.004 ± 0.005

Table 4 Averages of $\hat{\mathcal{L}}_{\mathrm{CE}}\left(\beta^{(n)}; \mathcal{D}\right) - \hat{\mathcal{L}}_{\mathrm{CE}}\left(\beta^*; \mathcal{D}\right)$ over ten trials, with hign missing rate data.

8. Conclusion

In this paper, we consider the problem of performing logistic regression on incomplete data while preserving local differential privacy. We propose two algorithms, dummy submission and two-phase regression, and derive the upper bounds of their utilities. The obtained upper bounds imply which algorithm is better varies in the parameter space of the problem. We theoretically find privacy buduget for local differential privacy and the size of dataset important factors. In addition, our experimental results show that the missing rate is another important factor.

As future work, we plan to evaluate the tightness of the upper bounds in in Theorem 1 and also derive lower bounds of the excess risks.

References

[1] Duchi, J. C., Jordan, M. I. and Wainwright, M. J.: Local Privacy and Statistical Minimax Rates, *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438 (2013).

[2] Duchi, J. C., Jordan, M. I. and Wainwright, M. J.: Minimax Optimal Procedures for Locally Private Estimation, *Journal of the American Statistical Association*, Vol. 113, No. 521, pp. 182–201 (online), DOI: 10.1080/01621459.2017.1389735 (2018).

[3] Joseph, M., Mao, J., Neel, S. and Roth, A.: The Role of Interactivity in Local Differential Privacy, *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 94–105 (2019).

[4] Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S. and Smith, A.: What Can We Learn Privately?, *SIAM Journal on Computing*, Vol. 40, No. 3, pp. 793–826 (online), DOI: 10.1137/090756090 (2011).

[5] Little, R. J. A. and Rubin, D. B.: *Statistical Analysis with Missing Data, Third Edition*, John Wiley & Sons (2019).

[6] Robins, J., Rotnitzky, A. and Zhao, L. P.: Estimation of Regression Coefficients When Some Regressors are not Always Observed, *Journal of the American Statistical Association*, Vol. 89, No. 427, pp. 846–866 (online), DOI: 10.1080/01621459.1994.10476818 (1994).

[7] Statistics Bureau of Japan: Problems and Issues in Conducting the 2005 Census (in Japanese), `https://www.stat.go.jp/info/kenkyu/kokusei/pdf/problems.pdf` (2006).

[8] Sun, H., Dong, B., Wang, H., Yu, T. and Qin, Z.: Truth Inference on Sparse Crowdsourcing Data with Local Differential Privacy, *2018 IEEE International Conference on Big Data (Big Data)*, pp. 488–497 (2018).

[9] Sun, L., Ye, X., Zhao, J., Lu, C. and Yang, M.: [This paper will appears in DASFAA2020 proceedings] Bisample: Bidirectional sampling for handling missing data with local differential privacy, *arXiv preprint arXiv:2002.05624* (2020).

[10] Wang, D. and Xu, J.: On Sparse Linear Regression in the Local Differential Privacy Model, *Proceedings of the 36th International Conference on Machine Learning* (Chaudhuri, K. and Salakhutdinov, R., eds.), Proceedings of Machine Learning Research, Vol. 97, Long Beach, California, USA, PMLR, pp. 6628–6637 (online), available from ⟨http://proceedings.mlr.press/v97/wang19m.html⟩ (2019).