Journal of Machine Learning Research 24 (2023) 1-57 Submitted 5/21; Revised 3/23; Published 5/23

Generalized Linear Models in Non-interactive Local Differential Privacy with Public Data

Di Wang[*] DI.WANG@KAUST.EDU.SA
CEMSE
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Lijie Hu LIJIE.HU@KAUST.EDU.SA
CEMSE
King Abdullah University of Science and Technology
Thuwal, Saudi Arabia

Huanyu Zhang HZ388@CORNELL.EDU
Meta
New York, NY, USA

Marco Gaboardi GABOARDI@BU.EDU
Department of Computer Science
Boston University
Boston, MA 02215, USA

Jinhui Xu JINHUI@BUFFALO.EDU
Department of Computer Science and Engineering
University at Buffalo, SUNY
Buffalo, NY 14260, USA

Editor: Moritz Hardt

Abstract

In this paper, we study the problem of estimating smooth Generalized Linear Models (GLMs) in the Non-interactive Local Differential Privacy (NLDP) model. Unlike its classical setting, our model allows the server to access additional public but unlabeled data. In the first part of the paper, we focus on GLMs. Specifically, we first consider the case where each data record is i.i.d. sampled from a zero-mean multivariate Gaussian distribution. Motivated by the Stein's lemma, we present an (ϵ, δ)-NLDP algorithm for GLMs. Moreover, the sample complexity of public and private data for the algorithm to achieve an ℓ_2-norm estimation error of α (with high probability) is $O(p\alpha^{-2})$ and $\tilde{O}(p^3\alpha^{-2}\epsilon^{-2})$ respectively, where p is the dimension of the feature vector. This is a significant improvement over the previously known exponential or quasi-polynomial in α^{-1}, or exponential in p sample complexities of GLMs with no public data. Then we consider a more general setting where each data record is i.i.d. sampled from some sub-Gaussian distribution with bounded ℓ_1-norm. Based on a variant of Stein's lemma, we propose an (ϵ, δ)-NLDP algorithm for GLMs whose sample complexity of public and private data to achieve an ℓ_∞-norm estimation error of α is $O(p^2\alpha^{-2})$ and $\tilde{O}(p^2\alpha^{-2}\epsilon^{-2})$ respectively, under some mild assumptions and if α is not too small (*i.e.*, $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$). In the second part of the paper, we extend our idea to the

[*] The first two authors contributed equally to this paper. An abstract version of this paper was presented at The 32nd International Conference on Algorithmic Learning Theory (ALT 2021) (Wang et al., 2021).

problem of estimating non-linear regressions and show similar results as in GLMs for both multivariate Gaussian and sub-Gaussian cases. Finally, we demonstrate the effectiveness of our algorithms through experiments on both synthetic and real-world datasets. To our best knowledge, this is the first paper showing the existence of efficient and effective algorithms for GLMs and non-linear regressions in the NLDP model with unlabeled public data.

Keywords: Differential Privacy, Generalized Linear Models, Local Differential Privacy

1. Introduction

Generalized Linear Model (GLM) is one of the most fundamental models in statistics and machine learning. It generalizes the ordinary linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. GLM was introduced as a way of unifying various statistical models, including linear, logistic, and Poisson regressions, and it has a wide range of applications in various domains, such as social sciences (Warne, 2017), genomics research (Takada et al., 2017), finance (McNeil and Wendin, 2007) and medical research (Lindsey and Jones, 1998). The model can be formulated as follows.

GLM: Let $y \in [0, 1]$ be the response variable that belongs to an exponential family with natural parameter ψ. [1] That is, its probability density function can be written as $p(y|\psi) = \exp(\psi y - \Phi(\psi))h(y)$, where Φ is the *cumulative generating function*. Given observations $y_1, \cdots, y_n$ such that $y_i \sim p(y_i|\psi_i)$ for $\psi = (\psi_1, \cdots, \psi_n)$, the maximum likelihood function can be written as $p(y_1, y_2, \cdots|\psi) = \exp(\sum_{i=1}^n y_i\psi_i - \Phi(\psi_i))\Pi_{i=1}^n h(y_i)$. In GLM, we assume that ψ is modeled by linear relations, *i.e.,* $\psi_i = \langle x_i, w^*\rangle$ for some $w^* \in \mathbb{R}^p$ and feature vector x_i. Thus, finding the maximum likelihood estimator (MLE) is equivalent to minimizing $\frac{1}{n}\sum_{i=1}^n[\Phi(\langle x_i, w\rangle) - y_i\langle x_i, w\rangle]$. The goal is to find w^*, which is equivalent to minimizing its population version

$$w^* = \arg\min_{w\in\mathbb{R}^p} \mathbb{E}_{(x,y)}[\Phi(\langle x, w\rangle) - y\langle x, w\rangle]. \tag{1}$$

One often encountered challenge for using GLMs in real-world applications is how to handle sensitive data, such as those in social science and medical research. As a commonly-accepted technique for preserving privacy, Differential Privacy (DP) (Dwork et al., 2006) provides provable protection against re-identification attacks and is resilient to arbitrary auxiliary information that might be available to attackers. It allows for rich statistical and machine learning analysis and is becoming a *de facto* notion for private data analysis.

As a popular way of achieving DP, Local Differential Privacy (LDP) has received considerable attention in recent years and has been adopted in industry (Ding et al., 2017; Erlingsson et al., 2014; Tang et al., 2017). In LDP, each individual manages his/her proper data and discloses them to a server through some DP mechanisms. The server collects each individual's (now private) data and combines them into a resulting data analysis. Information exchange between the server and individuals could be either only once or multiple times. Correspondingly, protocols for LDP are called non-interactive LDP (NLDP) or in-

[1] For simplicity in this paper we assume y is in $[0, 1]$. We will leave the case where y could be unbounded as future research.

teractive LDP. Due to its ease of implementation (*e.g.* no need to deal with the network latency issue), NLDP is often preferred in practice.

While there are many results on estimating GLMs in the DP and interactive LDP models such as (Chaudhuri et al., 2011; Bassily et al., 2014; Jain and Thakurta, 2014; Kasiviswanathan and Jin, 2016), estimating GLMs in NLDP is still not well-understood due to the limitation of the number of interaction round in the privacy model. Recently (Smith et al., 2017; Wang et al., 2018; Zheng et al., 2017) and (Wang et al., 2019b) provided comprehensive studies on this problem. However, all of these results are on the negative side. More specifically, they showed that to achieve an error of α, the sample complexity needs to be quasi-polynomial or exponential in α^{-1} (based on different assumptions) (Wang et al., 2019b; Zheng et al., 2017), or exponential in the dimension p (Smith et al., 2017; Wang et al., 2018) (see Related Work section for more details). Recently, (Dagan and Feldman, 2020) showed that an exponential lower bound (either in p or α^{-1}) on the number of samples for solving the standard task of learning a large-margin linear separator in the NLDP model. Due to these negative results, there is no study on the practical performance of these algorithms.

To address this high sample complexity issue of estimating GLMs in NLDP, a possible way is to use some recent developments in the central DP model. Quite a few results (Bassily and Nandi, 2019; Hamm et al., 2016; Papernot et al., 2016, 2018; Bassily et al., 2018; Liu et al., 2021) have suggested that by allowing the server to access some public but unlabeled data in addition to the private data, it is possible to further reduce the sample complexity in the central DP model, under the assumption that these public data samples have the same marginal distribution as the private ones. It has also been shown that such a relaxed setting is likely to enable better practical performance for various problems such as Empirical Risk Minimization (ERM) and Deep Neural Networks (Hamm et al., 2016; Papernot et al., 2016). Thus, it would be interesting to know whether the relaxed setting on public unlabeled data can also help to reduce the sample complexity of GLMs in the NLDP model.

With this thinking, our main questions now become the following. **Can we further reduce the sample complexity of GLMs in the NLDP model if the server has additional public but unlabeled data? Moreover, is there any efficient algorithm for this problem?**

In this paper, we provide positive answers to the above two questions, see Table 1 for our results. Specifically, our contributions can be summarized as follows:

1. Firstly, motivated by Stein's lemma (Lemma 5), we show that when the covariate (feature vector) x follows an (unknown) zero-mean multivariate Gaussian distribution, *i.e.,* $x \sim \mathcal{N}(0, \Sigma)$ with some $\Sigma \in \mathbb{R}^{p\times p}$, there exists an (ϵ, δ)-NLDP algorithm for GLMs. Moreover, the sample complexity of public and private data for the algorithm to achieve an ℓ_2-norm estimation error of α (with high probability), is $O(p\alpha^{-2})$ and $\tilde{O}(p^3\alpha^{-2}\epsilon^{-2})$ (with other terms omitted), respectively. We note that this is the first result that achieves a **fully polynomial** sample complexity for a general class of loss functions in the NLDP model with public unlabeled data.

2. We then consider a more general case where the covariate x in GLMs is sub-Gaussian with bounded ℓ_1-norm. Based on a variant of Stein's lemma, we propose an (ϵ, δ)-

Methods	Sample Complexity	Measure	Loss Function	With public data?	Data
(Smith et al., 2017)	$O(p\epsilon^{-2}\alpha^{-2})$	Excess Risk	Linear Regression	No	ℓ_2-norm Bounded
(Smith et al., 2017)	$\tilde{O}(4^p\alpha^{-(p+2)}\epsilon^{-2})$	Excess Risk	Lipschitz	No	ℓ_2-norm Bounded
(Smith et al., 2017)	$\tilde{O}(2^p\alpha^{-(p+1)}\epsilon^{-2})$	Excess Risk	Lipschitz and Convex	No	ℓ_2-norm Bounded
(Wang et al., 2018)	$\tilde{O}\big((c_0p^{\frac{1}{4}})^p\alpha^{-(2+\frac{p}{2})}\epsilon^{-2}\big)$	Excess Risk	$(8, T)$-smooth	No	ℓ_2-norm Bounded
(Wang et al., 2018)	$\tilde{O}(4^{p(p+1)}D_p^2\epsilon^{-2}\alpha^{-4})$	Excess Risk	(∞, T)-smooth	No	ℓ_2-norm Bounded
(Wang et al., 2019b, 2020)	$p\cdot\big(\frac{C}{\alpha^3}\big)^{O(1/\alpha^3)}/\epsilon^{O(\frac{1}{\alpha^3})}$	Excess Risk	Lipschitz Convex GLM	No	ℓ_2-norm Bounded
(Zheng et al., 2017)	$p(\frac{8}{\alpha})^{O(\log\log(\frac{1}{\alpha}))}(\frac{4}{\epsilon})^{O(\log(\frac{1}{\alpha}))}$	Excess Risk	Convex ∞-Smooth GLM	No	ℓ_2-norm Bounded
This paper	$O(p^3\alpha^{-2}\epsilon^{-2})$	ℓ_2-norm Error	Smooth GLM (with additional assumptions)	Yes	Gaussian
This paper	$O(p^2\alpha^{-2}\epsilon^{-2})$ for $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$	ℓ_∞-norm Error	Smooth GLM (with additional assumptions)	Yes	ℓ_1-norm Bounded and Sub-Gaussian

Table 1: Comparisons on the sample complexities (of private data) for achieving error α under different measurements for GLMs in the non-interactive LDP model, where c_0, C are constants, and D_p is a function of dimension p. For bounded norm case we assume that $\|x_i\| \leq 1$ for every $i \in [n]$. For multivariate Gaussian case we assume $x_i \sim \mathcal{N}(0, \Sigma)$ with some unknown Σ.

NLDP algorithm for GLMs. Moreover, under some mild assumptions, the sample complexity of private and public data to achieve an ℓ_∞-norm error of α is $\tilde{O}(p^2\epsilon^{-2}\alpha^{-2})$ and $\tilde{O}(p^2\alpha^{-2})$ (with other terms omitted) respectively, if α is not too small (*i.e.,* $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$).

3. We then extend our idea to the problem of estimating non-linear regressions. Using Stein's lemma and the zero-bias transformation (Goldstein et al., 1997), we propose (ϵ, δ)-NLDP algorithms in both cases where x is multivariate Gaussian and sub-Gaussian with bounded ℓ_1-norm. Moreover, we show similar estimation errors as in the GLMs case.

4. Finally, we provide an extensive experimental study of our algorithms on synthetic and real-world datasets. The experimental results suggest that our methods are efficient, effective, and consistent with our theoretical analysis. Moreover, based on these results, we also find some aspects that need further theoretical investigation.

2. Related Work

Private learning with public unlabeled data has been studied previously in (Hamm et al., 2016; Papernot et al., 2016, 2018; Bassily et al., 2018; Liu et al., 2021; Su et al., 2022). These results differ from ours in quite a few ways. Firstly, all of them consider either the multiparty setting or the central DP model and cannot be extended to the NLDP model. Consequently, none of them can be used to solve our problems. Specifically, Hamm et al.

(2016) considered the multiparty setting where each party possesses several data records. Their method needs each party to use its data to get a classifier. However, this approach could not be extended to the local DP model since in our case each party only has one data sample, and it is impossible to get any useful classifier based on one data sample only. Papernot et al. (2016, 2018) considered training some Deep Neural Networks in the DP model using the subsample-and-aggregate framework in (Nissim et al., 2007). However, there is no provable sample complexity for their methods. Bassily et al. (2018) studied DP-ERM in the central model, which is later extended by (Liu et al., 2021). Their method is based on combining the function of distance to instability and the sparse vector technique. However, both the subsample-and-aggregate framework and the sparse vector technique cannot be used in the local DP model. Secondly, public data samples in those methods are also used quite differently from ours. Specifically, the above approaches use private data to get private classifiers. Based on these classifiers, they label the public data and conduct the learning process on the public data (now with pseudo labels), while in this paper we use the public data to approximate some crucial constants. Finally, all of the previous methods rely on the known model or the explicit form of the loss function, while in our algorithms the loss functions could be unknown to users; also, the server could estimate multiple different GLMs with the same sample complexity.

The problems considered in this paper can be viewed as specific cases of the ERM problem in the NLDP model. Due to its challenging nature, ERM in NLDP has only been considered in a few papers, such as (Smith et al., 2017; Wang et al., 2018, 2019b; Zheng et al., 2017; Daniely and Feldman, 2018; Wang and Xu, 2019), see Table 1 for a summary. Smith et al. (2017) gave the first result on convex ERM in NLDP and provided an algorithm with a sample complexity of $O(2^p\alpha^{-(p+1)}\epsilon^{-2})$. They showed that the exponential dependency on the dimension p is unavoidable in the worst case. Later, Wang et al. (2018) showed that when the loss function is smooth enough, the exponential term of $\alpha^{-\Omega(p)}$ can be reduced to polynomial. However, there is still another exponential term in their sample complexity. Recently, Wang et al. (2019b, 2020) further showed that the sample complexity for any 1-Lipschitz convex GLM can be reduced to only linear in p and exponential in α^{-1}, which extends a result in (Zheng et al., 2017) whose sample complexity is linear in p and quasi-polynomial in α^{-1} for smooth GLMs. In this paper, we show, for the first time, that the sample complexity of GLMs can be reduced to polynomial with the help of some public but unlabeled data under some mild assumptions. There are also some results for specific loss functions. For example, (Wang and Xu, 2019) studied the high dimensional sparse linear regression problem, and (Daniely and Feldman, 2018; Su et al., 2022) considered the problem of PAC learning halfspaces with polynomial samples. Since these results are only for some special loss functions (instead of a family of functions), they are incomparable with ours.

As we mentioned, there is a long list of work studies GLMs in the central DP and the interactive LDP models. In the central DP model, Jain and Thakurta (2014) provided the first study and showed that to achieve an error α of the excess population risk, there is an (ϵ, δ)-DP algorithm with sample complexity $\tilde{O}(\epsilon^{-2}\alpha^{-2})$. Recently, Song et al. (2021) showed a sharper sample complexity bound of $\tilde{O}(\sqrt{\text{rank}}\epsilon^{-1}\alpha^{-1})$, where rank is the rank of the feature matrix formed by stacking the feature vectors as columns, which always holds that rank $\leq n$. Bassily et al. (2021) provided an algorithm that runs in (nearly) linear time instead of superlinear in the previous work, and its sample complexity is $\tilde{O}(\max\{\sqrt{\text{rank}}\epsilon^{-1}\alpha^{-1}, \alpha^{-2}\})$. They also extended from the ℓ_2-norm Lipschitz case and the convex setting to the ℓ_1-norm Lipschitz case and the weakly-convex setting. Arora et al. (2022b) studied DP-GLM where the loss is smooth and non-negative but not necessarily Lipschitz. They showed a near optimal sample complexity, which is $\tilde{O}(\max\{\alpha^{-2}, \alpha^{-\frac{3}{2}}\epsilon^{-1}, \sqrt{p}\epsilon^{-1}\alpha^{-1}\})$ (if $\|w^*\|_2 \leq 1$). Besides convex loss functions, Arora et al. (2022a) recently studied non-convex GLMs in the DP model and showed that to achieve an error α of the ℓ_2-norm of the gradient of the population risk function, there is an $\epsilon, \delta)$-DP algorithm whose sample complexity is $\tilde{O}(\max\{\alpha^{-2}, \sqrt{\text{rank}}\epsilon^{-1}\alpha^{-1}, \alpha^{-\frac{5}{2}}\epsilon^{-1}\})$. Cai et al. (2020) studied DP-GLM under statistical settings. In the low dimensional case where the covariate x satisfies $\|x\|_2 \leq \sqrt{p}$, to achieve an α ℓ_2-norm estimation error, they provided an algorithm with a near-optimal sample complexity of $\tilde{O}(\max\{p\alpha^{-1}, p\epsilon^{-1}\alpha^{-\frac{1}{2}}\})$. Moreover, under the high dimensional sparse setting, in the case where $\|x\|_\infty \leq 1$ and with some additional assumptions, they presented an algorithm with sample complexity $\tilde{O}(\max\{s^*\epsilon^{-1}, s^*\epsilon^{-1}\alpha^{-\frac{1}{2}}\})$, where s^* is the underlying sparsity of w^*. (Hu et al., 2022) recently generalized these results to the case where the covariates are heavy-tailed. In the interactive LDP model, Duchi et al. (2013) provided the first study on ERM in the sequentially interactive LDP model and showed the (nearly) optimal minimax rate of sample complexity should be $\tilde{O}(p\epsilon^{-2}\alpha^{-2})$ to achieve an error α of the excess population risk when the loss function is ℓ_2-norm Lipschitz.

3. Preliminaries

Since this paper mainly focuses on multivariate Gaussian and sub-Gaussian covariates, we first recall some definitions. More details can be found in (Vershynin, 2018).

Definition 1 (Sub-Gaussian) *For a given constant κ, a random variable $x \in \mathbb{R}$ is said to be sub-Gaussian if it satisfies $\sup_{m\geq 1} \frac{1}{\sqrt{m}}\mathbb{E}[|x|^m]^{\frac{1}{m}} \leq \kappa$. The smallest such κ is the* **sub-Gaussian norm** *of x and it is denoted by $\|x\|_{\psi_2}$. A random vector $x \in \mathbb{R}^p$ is called a sub-Gaussian vector if there exists a constant κ such that for any unit vector v, we have $\|\langle x, v\rangle\|_{\psi_2} \leq \kappa$.*

For sub-Gaussian data, we need the following assumptions on its distribution throughout the paper.

Assumption 1 *For a random vector x that is sub-Gaussian with zero mean and covariance matrix Σ, we assume the following conditions hold*

- *Its distribution is supported on a ℓ_1-norm ball of radius r.*
- *For the matrix Σ, its corresponding $\Sigma^{\frac{1}{2}}$ is diagonally dominant, where $\Sigma^{\frac{1}{2}}$ is the square root of matrix Σ.* [2]

[2] A square matrix is said to be diagonally dominant if, for every row of the matrix, the magnitude of the diagonal entry in a row is larger than or equal to the sum of the magnitudes of all the other (non-diagonal) entries in that row. For a semi-definite positive matrix $M \in \mathbb{R}^{p\times p}$, let its SVD composition be $M = U^T\Sigma U$, where $\Sigma = \text{diag}(\lambda_1, \cdots, \lambda_p)$, then $M^{\frac{1}{2}}$ is defined as $M^{\frac{1}{2}} = U^T\Sigma^{\frac{1}{2}}U$, where $\Sigma^{\frac{1}{2}} = \text{diag}(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_p})$.

- *Let $v = \Sigma^{-\frac{1}{2}}x$ be the whitened random vector of x, each v_i has constant first and second conditional moments,* i.e., *$\forall j \in [p]$ and $\tilde{w} = \Sigma^{\frac{1}{2}} w^*$, $\mathbb{E}[v_{ij} | \sum_{k \neq j} \tilde{w} v_{ik}] = O(1)$ and $\mathbb{E}[v_{ij}^2 | \sum_{k \neq j} \tilde{w} v_{ik}] = O(1)$.*

In Assumption 1 there are three terms. The first one is natural as it has also been used in previous studies on DP-GLM. For the other two terms, we note that they are crucial for Lemma 10 and Theorem 22, which are only used in utility analysis. Thus, even though these two assumptions do not hold, we still have privacy guarantees. Moreover, it is straightforward to observe that when the whitened covariates v have independent, but not necessarily identical entries, these two terms hold. We leave it as an open problem to further relax these assumptions.

Differential Privacy (DP): In DP, we have data universe $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}$, and a dataset $D \in (\mathcal{X} \times \mathcal{Y})^n$ whose size is n and the dataset is stored in some trusted curator. Each data record $(x, y) \in \mathcal{D}$ sampled from some distribution $\mathcal{P}$, where $x \in \mathbb{R}^p$ is the feature vector and $y \in \mathbb{R}$ is the label of response. We say that two datasets $D, D' \subseteq \mathcal{X}$ are neighbors if they differ by only one data record, denoted as $D \sim D'$.

Definition 2 (Differential Privacy (Dwork et al., 2006)) *We call a randomized algorithm Q is (ϵ, δ)-differentially private (DP) if for all neighboring datasets D, D' and for all events E in the output space of Q, the following holds*

$$\mathbb{P}(Q(D) \in E) \leq e^{\epsilon} \mathbb{P}(Q(D') \in E) + \delta.$$

When $\delta = 0$, $\mathcal{A}$ is ϵ-DP.

Local Differential Privacy (LDP): Instead of the trusted curator, in LDP model (Kasiviswanathan et al., 2011), each player (data provider) perturb his/her private data record locally via some DP algorithms before sending it to the curator. Specifically, there are n players with each holding a private data record $(x, y) \in \mathcal{X} \times \mathcal{Y}$ sampled from some distribution $\mathcal{P}$, and a server that is in charge of coordinating the protocol. An LDP protocol proceeds in T rounds. In each round, the server sends a message, which is often called a query, to a subset of the players, requesting them to run a particular algorithm. Based on the query, each player i in the subset selects an algorithm Q_i, runs it on her/his own data, and sends the output back to the server.

Definition 3 (Local Differential Privacy (Kasiviswanathan et al., 2011)) *A randomized algorithm Q is (ϵ, δ)-locally differentially private (LDP) if for all pairs $x, x' \in \mathcal{D}$, and for all events E in the output space of Q, we have*

$$\mathbb{P}(Q(x) \in E) \leq e^{\epsilon} \mathbb{P}(Q(x') \in E) + \delta.$$

When $\delta = 0$, $\mathcal{A}$ is ϵ-LDP. A multi-player protocol is $(\epsilon, \delta)/\epsilon$-LDP if for all possible inputs and runs of the protocol, the transcript of player i's interaction with the server is $(\epsilon, \delta)/\epsilon$-LDP. If $T = 1$, we say that the protocol is $(\epsilon, \delta)/\epsilon$ ***non-interactive LDP (NLDP)****.*

In this paper, we will mainly focus on (ϵ, δ)-NLDP and we will mainly use the Gaussian mechanism (Dwork et al., 2006) to guarantee (ϵ, δ)-LDP.

Lemma 4 (Gaussian Mechanism (Dwork et al., 2006)) *Given any function $q : (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{R}^d$, the Gaussian mechanism is defined as $\mathcal{M}_G(D, q, \epsilon) = q(D) + Y$, where Y is drawn from Gaussian Distribution $\mathcal{N}(0, \sigma^2 I_d)$ with $\sigma \geq \frac{\sqrt{2\ln(1.25/\delta)}\Delta_2(q)}{\epsilon}$. Here $\Delta_2(q)$ is the ℓ_2-sensitivity of the function q, i.e., $\Delta_2(q) = \sup_{D \sim D'} ||q(D) - q(D')||_2$. Gaussian mechanism preserves (ϵ, δ)-differential privacy.*

Our Model: Different from the above classical NLDP model where only one private dataset $D = \{(x_i, y_i)\}_{i=1}^n$ exists, the NLDP model in our setting allows the server to have an additional public but unlabeled dataset $D' = \{x_j\}_{j=n+1}^{n+m} \subset \mathcal{X}^m$, where each x_j is sampled from $\mathcal{P}_x$, which is the marginal distribution of $\mathcal{P}$ (*i.e.,* it has the same distribution as each x_i).

4. Privately Estimating Generalized Linear Models

In this section, we study GLMs in our privacy model and we aim to privately estimate w^* in (1) by using both private data $\{(x_i, y_i)\}_{i=1}^n$ and public unlabeled data $\{x_j\}_{j=n+1}^{n+m}$. Our goal is to achieve a fully polynomial sample complexity for n and m, *i.e.,* $n, m = \text{Poly}(p, \frac{1}{\epsilon}, \frac{1}{\alpha}, \log\frac{1}{\delta})$, such that there is an (ϵ, δ)-NLDP algorithm with estimation error less than α (with high probability).

4.1 Gaussian Case

We first consider a simpler case that each data record is sampled from some unknown Gaussian distribution $\mathcal{N}(0, \Sigma)$. The idea of our method is motivated by the following result, which is derived from Stein's lemma (Brillinger, 2012).

Lemma 5 ((Brillinger, 2012)) *If $x \sim \mathcal{N}(0, \Sigma)$, then w^* in (1) can be written as $w^* = c_\Phi \times w^{ols}$, where c_Φ is the fixed point of $z \mapsto (\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle z)])^{-1}$ (if we assume that $\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle z)] \neq 0$) and $w^{ols} = \Sigma^{-1}\mathbb{E}[xy]$ is the Ordinary Least Squares (OLS) vector.* [3]

From Lemma 5, we can see that to estimate w^*, it is sufficient to estimate w^{ols} and its corresponding constant c_Φ. Specifically, to estimate w^{ols} in a non-interactive local differentially private manner, a direct way is to let each player perturb her/his sufficient statistics, *i.e.,* $x_i x_i^T$ and $y_i x_i$. After receiving the private OLS estimator $\hat{w}^{ols}$,[4] the server can then estimate the constant c_Φ by using the public unlabeled data and $\hat{w}^{ols}$. From the definition, it is easy to see that c_Φ is independent of the label y. Thus, c_Φ can be estimated by using the empirical version of $\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle z)]$. That is, find the root of the function $1 - \frac{c}{m}\sum_{j=n+1}^{n+m} \Phi^{(2)}(c\langle x_j, \hat{w}^{ols}\rangle)$. Several methods are available for finding roots, and in our algorithms, we will use Newton's method which has a quadratic convergence rate.

However, there is a challenge to this approach. That is, Lemma 5 needs to assume x is Gaussian, which implies that the sensitivity of the terms $\|x_i x_i^T\|_F$ and $\|y_i x_i\|_2$ could be unbounded. To address this issue, we will use the concentration inequality on the ℓ_2-norm

[3] $\Phi^{(2)}$ is the second order derivative function of function Φ, similar for $\Phi^{(3)}$ in the later sections.

[4] Note that when n is large enough we can show $\hat{w}^{ols}$ is well defined, see Appendix for details.

of Gaussian distributions, and clip each x_i to let it has bounded ℓ_2-norm. Specifically, we are motivated by the following lemma:

Lemma 6 (Gaussian case of (Hsu et al., 2012)) *Let $x \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^p$. For all $t > 0$,*

$$\mathbb{P}(\|x\|_2^2 \geq \mathit{trace}(\Sigma) + 2\sqrt{\mathit{trace}(\Sigma^2)t} + 2\|\Sigma\|_2 t) \leq e^{-t}. \tag{2}$$

Since $\text{trace}(\Sigma) \leq p\|\Sigma\|_2$ and $\text{trace}(\Sigma^2) \leq (\text{trace}(\Sigma))^2$, from Lemma 6 we have with probability at least $1 - \frac{1}{n^2}$, $\|x\|_2 \leq r \equiv \sqrt{10p\|\Sigma\|_2 \log n}$. Motivated by this we clip each x_i to $\bar{x}_i = x_i \min\{1, \frac{r}{\|x_i\|_2}\}$ and now the terms $\|\bar{x}_i\bar{x}_i^T\|_F$ and $\|y_i\bar{x}_i\|_2$ are bounded. However, we can see the clipping threshold depends on the term of $\|\Sigma\|_2$, which is unknown in advance. To estimate this term, we can use the empirical covariance matrix of the public data $\{x_j\}_{j=n+1}^{n+m}$. See Algorithm 1 for details.

Algorithm 1 Non-interactive LDP for smooth GLMs with public data (Gaussian)

```
Input: Private data $\{(x_i, y_i)\}_{i=1}^n \in (\mathbb{R}^p \times [0,1])^n$, where $|y_i| \leq 1$, $\{x_i\}_{j=1}^{n+m} \sim \mathcal{N}(0, \Sigma)$ for some unknown $\Sigma$ and $\{x_j\}_{j=n+1}^{n+m}$ are public, loss function $\Phi : \mathbb{R} \mapsto \mathbb{R}$, privacy parameters $\epsilon, \delta$, and initial value $c \in \mathbb{R}$.
for The server do
    Calculate $\Sigma_m = \frac{1}{m}\sum_{j=n+1}^{n+m} x_j x_j^T$ and send it to each user.
end for
for Each user $i \in [n]$ do
    Let $\bar{x}_i = x_i \min\{1, \frac{r}{\|x_i\|_2}\}$, where $r \equiv \sqrt{20p\|\Sigma_m\|_2 \log n}$.
    Release $\widehat{x_i x_i^T} = \bar{x}_i\bar{x}_i^T + E_{1,i}$ and $\widehat{x_i y_i} = \bar{x}_i y_i + E_{2,i}$, where $E_{1,i} \in \mathbb{R}^{p\times p}$ is a symmetric matrix and each entry of the upper triangle matrix is sampled from $\mathcal{N}(0, \frac{32r^4 \log\frac{2.5}{\delta}}{\epsilon^2})$ and $E_{2,i} \in \mathbb{R}^p$ is sampled from $\mathcal{N}(0, \frac{32r^2 \log\frac{2.5}{\delta}}{\epsilon^2} I_p)$.
end for
for The server do
    Let $\widehat{X^T X} = \sum_{i=1}^n \widehat{x_i x_i^T}$ and $\widehat{X^T y} = \sum_{i=1}^n \widehat{x_i y_i}$. Calculate $\hat{w}^{ols} = (\widehat{X^T X})^{-1}\widehat{X^T y}$.
    Calculate $\tilde{y}_j = x_j^T \hat{w}^{ols}$ for each $j = n+1, \cdots, n+m$.
    Find the root $\hat{c}_\Phi$ such that $1 = \frac{\hat{c}_\Phi}{m}\sum_{j=n+1}^{n+m} \Phi^{(2)}(\hat{c}_\Phi \tilde{y}_j)$ by using Newton's root-finding method (or other methods):
    for $t = 1, 2, \cdots$ until convergence do
        $c = c - \frac{c\frac{1}{m}\sum_{j=n+1}^{n+m}\Phi^{(2)}(c\tilde{y}_j) - 1}{\frac{1}{m}\sum_{j=n+1}^{n+m}\{\Phi^{(2)}(c\tilde{y}_j) + c\tilde{y}_j\Phi^{(3)}(c\tilde{y}_j)\}}$.
    end for
end for
return $\hat{w}^{glm} = \hat{c}_\Phi \cdot \hat{w}^{ols}$.
```

Theorem 7 *For any $0 < \epsilon, \delta < 1$, Algorithm 1 is (ϵ, δ) non-interactive LDP.*

Next, we will show the estimation error bound of the output in Algorithm 1, before that we need the following assumptions for loss functions.

Assumption 2 *We assume*

- *$|\Phi^{(2)}(\cdot)| \leq L$ and $\Phi^{(3)}(\cdot)$ is G-Lipschitz.*
- *There exist constants $\bar{c}$ and $\tau > 0$, the function $f(c) = c\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle c)]$ satisfies the condition of $f(\bar{c}) \geq 1 + \tau$, where w^{ols} is in Lemma 5 and $x \sim \mathcal{N}(0, \Sigma)$.*
- *The derivative of f in the interval $[0, \bar{c}]$ does not change the sign,* i.e., *its absolute value is lower bounded by some constant $M > 0$.*

Note that the first condition ensures that $\Phi^{(1)}$ is Lipschitz, and the second and the last conditions are to ensure that the function $f - 1$ has a root and $\hat{c}_\Phi$ close to c_Φ for large enough m, see Theorem 16 and 17 for some concrete instances that satisfy the assumption.

Theorem 8 *Let $x_1, \cdots, x_n \in \mathbb{R}^p$ be i.i.d realizations of a random vector $x \sim \mathcal{N}(0, \Sigma)$. Moreover, under Assumption 2, for sufficiently large m, n such that*

$$n \geq \tilde{\Omega}\big(\frac{p^3\|w^*\|_2^2 \log\frac{1}{\delta}\log\frac{1}{\xi}}{\epsilon^2}\big) \tag{3}$$

$$m \geq \Omega\big(\|w^*\|_2^2 p\big). \tag{4}$$

Then for any $\zeta \in (0,1)$, with probability at least $1 - \exp(-\Omega(p)) - \xi$

$$\|\hat{w}^{glm} - w^*\|_2 \leq \tilde{O}\big(\frac{p^{\frac{3}{2}}\|w^*\|_2^2\sqrt{\log\frac{1}{\delta}\log\frac{1}{\xi}}}{\epsilon\sqrt{n}} + \frac{\sqrt{p}\|w^*\|_2^2}{\sqrt{m}}\big),$$

where Big-$\tilde{O}$ and Big-$\tilde{\Omega}$ notations omit the terms of $\|\Sigma\|_2, G, L, M, \bar{c}, \tau, c_\Phi$, $\lambda_{\min}(\Sigma)$ (the smallest eigenvalue of Σ) and other logarithmic factors (see Appendix for the explicit form of m and n).

Theorem 8 suggests that if $\|w^*\|_2 = O(1)$, then for any given error α, there is an (ϵ, δ)-NLDP algorithm whose sample complexity of private (n) and public unlabeled (m) data, to achieve the ℓ_2-norm error of α, is $\tilde{O}(p^3\epsilon^{-2}\alpha^{-2})$ and $O(p\alpha^{-2})$, respectively. We note that $m \leq n$, which means that the sample complexity of the public data is less than that of the private data. We also note that the sample complexity of the public data is independent of the privacy parameters ϵ and δ.

Actually, there is one possible way to improve the practical performance of Algorithm 1 (and all other algorithms in the paper). The key observation is that, in the procedure of estimating the OLS estimator, the empirical covariance matrix $\widehat{X^TX}$ does not depend on labels. Thus, we can further use those public unlabeled data to give a more precise estimator of the covariance matrix. That is, we can let $\widehat{X^TX} = \frac{1}{m+n}(\sum_{i=1}^n \widehat{x_ix_i^T} + \sum_{j=n+1}^{n+m} x_jx_j^T)$ and $\widehat{X^Ty} = \frac{1}{n}\sum_{i=1}^n \widehat{x_iy_i}$. However, by using a similar proof as in the proof of Theorem 8, we can see that the upper bound of error will be asymptotically the same as the bound in Theorem 8 (and all other theorems in the paper). In the experimental section, we will adopt this improved approach.

Remark 9 *It is notable that the public dataset is only used in line 11-15 of Algorithm 1 (similar to other algorithms), where we use it to find a root of some function and to estimate*

$\|\Sigma\|_2$ in the Gaussian case. Actually, we can adjust our idea to a 2-round LDP algorithm in the canonical model, i.e., *there is no public unlabeled data (suppose we already know $\|\Sigma\|_2$ for the Gaussian case). That is, in the first round we get $\hat{w}^{ols}$ by using half the privacy budget, and the server sends it to all the users. In the second round, each user uses another half privacy budget to compute a noisy version of $\tilde{y}_j = x_j^T \hat{w}_{ols}$ and sends it to the server. Then the server uses these noisy versions of $\tilde{y}_j$ to estimate the constant of c_Φ. We note that due to the noise added in the second round for each term of $\tilde{y}_j$, there could be a large amount of error when using $\hat{c}_\Phi$ to estimate c_Φ, and this will cause the private estimator has a large error. In the experiments section, we will practically show that this approach will lead to worse performance.*

4.2 Sub-Gaussian Case

The main weakness of the previous result is that due to Lemma 5, Theorem 8 only holds for Gaussian distributions. Fortunately, recently Erdogdu et al. (2019) generalized Stein's lemma to bounded sub-Gaussian random vectors. Compared with the Gaussian case, in this case there is an additional additive error of $O(\frac{\|w^*\|_\infty^2}{\sqrt{p}})$. Formally, we have the following lemma.

Algorithm 2 Non-interactive LDP for smooth GLMs with public data (General)

```
Input: Private data $\{(x_i, y_i)\}_{i=1}^n \subset (\mathbb{R}^p \times [0,1])^n$, where $\|x_i\|_1 \leq r$ and $|y_i| \leq 1$, public unlabeled data $\{x_j\}_{j=n+1}^{n+m}$, loss function $\Phi : \mathbb{R} \mapsto \mathbb{R}$, privacy parameters $\epsilon, \delta$, and initial value $c \in \mathbb{R}$.
for Each user $i \in [n]$ do
    Release $\widehat{x_i x_i^T} = x_i x_i^T + E_{1,i}$, where $E_{1,i} \in \mathbb{R}^{p \times p}$ is a symmetric matrix and each entry of the upper triangle matrix is sampled from $\mathcal{N}(0, \frac{32r^4 \log \frac{2.5}{\delta}}{\epsilon^2})$.
    Release $\widehat{x_i y_i} = x_i y_i + E_{2,i}$, where $E_{2,i} \in \mathbb{R}^p$ is sampled from $\mathcal{N}(0, \frac{32r^2 \log \frac{2.5}{\delta}}{\epsilon^2} I_p)$.
end for
for The server do
    Let $\widehat{X^T X} = \sum_{i=1}^n \widehat{x_i x_i^T}$ and $\widehat{X^T y} = \sum_{i=1}^n \widehat{x_i y_i}$. Calculate $\hat{w}^{ols} = (\widehat{X^T X})^{-1} \widehat{X^T y}$.
    Calculate $\tilde{y}_j = x_j^T \hat{w}^{ols}$ for each $j = n+1, \cdots, n+m$.
    Find the root $\hat{c}_\Phi$ such that $1 = \frac{\hat{c}_\Phi}{m} \sum_{j=n+1}^{n+m} \Phi^{(2)}(\hat{c}_\Phi \tilde{y}_j)$ by using Newton's root-finding method (or other methods):
    for $t = 1, 2, \cdots$ until convergence do
        $c = c - \frac{c \frac{1}{m} \sum_{j=n+1}^{n+m} \Phi^{(2)}(c\tilde{y}_j) - 1}{\frac{1}{m} \sum_{j=n+1}^{n+m} \{\Phi^{(2)}(c\tilde{y}_j) + c\tilde{y}_j \Phi^{(3)}(c\tilde{y}_j)\}}$.
    end for
end for
return $\hat{w}^{glm} = \hat{c}_\Phi \cdot \hat{w}^{ols}$.
```

Lemma 10 ((Erdogdu et al., 2019)) *Let $x_1, \cdots, x_n \in \mathbb{R}^p$ be i.i.d realizations of a random vector x that is zero-mean sub-Gaussian with covariance matrix Σ and satisfies Assumption 1. Let $v = \Sigma^{-\frac{1}{2}} x$ be the whitened random vector of x and denote $\|v\|_{\psi_2} = \kappa_x$. If the function $\Phi^{(2)}$ is Lipschitz with constant G, then for $c_\Phi = \frac{1}{\mathbb{E}[\Phi^{(2)}(\langle x_i, w^* \rangle)]}$ (assuming*

$\mathbb{E}[\Phi^{(2)}(\langle x_i, w^*\rangle)] \neq 0$*), the following holds for GLM in (1)*

$$\|\frac{1}{c_\Phi} \cdot w^* - w^{ols}\|_\infty \leq O(Gr\kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}), \tag{5}$$

where ρ_q *for* $q = \{2, \infty\}$ *is the conditional number of* Σ *in* ℓ_q*-norm,* i.e., $\rho_q = \|\Sigma\|_q \|\Sigma^{-1}\|_q$ *where* $\|A\|_q = \sup_{x\neq 0} \frac{\|Ax\|_q}{\|x\|_q}$ *for matrix* A*, and* $w^{ols} = \Sigma^{-1}\mathbb{E}[xy]$ *is the OLS vector.*

Lemma 10 indicates that we can use the same idea as in the previous section to estimate w^*. Note that the forms of constant c_Φ in Lemma 5 and 10 are different while one depends on w^{ols} and the other one depends on w^*. However, since by (5) we know w^* and $c_\Phi w^{ols}$ are close. Thus, intuitively we can still use $\frac{1}{\mathbb{E}[\Phi^{(2)}(\langle x_i, w^{ols}\rangle \bar{c}_\Phi])}$ to approximate c_Φ, where $\bar{c}_\Phi$ is the root of $c\mathbb{E}[\Phi^{(2)}(\langle x_i, w^{ols}\rangle c)] - 1$ which could be approximated by using public unlabeled data. Combining these ideas, we present Algorithm 2.

Theorem 11 *For any* $0 < \epsilon, \delta < 1$*, Algorithm 2 is* (ϵ, δ) *non-interactive LDP.*

The following theorem shows the sample complexity of the bounded sub-Gaussian distributions. Like Assumption 2, we need the following assumptions for loss functions.

Assumption 3 *We assume*

- $|\Phi^{(2)}(\cdot)| \leq L$ *and* $\Phi^{(3)}(\cdot)$ *is G-Lipschitz.*
- *For some constant* $\bar{c}$ *and* $\tau > 0$*, the function* $f(c) = c\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle c)]$ *satisfies the condition of* $f(\bar{c}) \geq 1+\tau$*, where* w^{ols} *is in Lemma 10 and the distribution of* x *satisfies Assumption 1.*
- *The derivative of* f *in the interval* $[0, \max\{\bar{c}, c_\Phi\}]$ *does not change the sign (*i.e., *its absolute value is lower bounded by some constant* $M > 0$*), where* c_Φ *is in Lemma 10.*

It seems that Assumption 3 is almost the same as Assumption 2. However, since these two assumptions rely on the underlying distribution of (x, y), which are different in these two cases. Thus, the two assumptions are different. Moreover, the third conditions in Assumption 3 and Assumption 2 are different due to different intervals and forms of c_Φ.

Theorem 12 *Under Assumption 1 and 3, for sufficiently large* m, n *such that*

$$\begin{aligned} &m \geq \Omega\big(\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} p^2\big), \\ &n \geq \tilde{\Omega}\big(\frac{p^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \log\frac{1}{\delta}\log\frac{1}{\xi}}{\epsilon^2}\big). \end{aligned} \tag{6}$$

Then for any $\zeta \in (0,1)$*, with probability at least* $1 - \exp(-\Omega(p)) - \xi$*, the output* $\hat{w}^{glm}$ *in Algorithm 2 satisfies*

$$\begin{aligned} \|\hat{w}^{glm} - w^*\|_\infty \leq \tilde{O}\big(&\frac{\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} p}{\sqrt{m}} \\ &+ \frac{\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} p \sqrt{\log\frac{1}{\delta}\log\frac{1}{\xi}}}{\epsilon\sqrt{n}} + \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}}\big), \end{aligned} \tag{7}$$

where Big-$\tilde{O}$ and Big-Ω notations omit the terms of $\|\Sigma\|_2, \rho_2, \rho_\infty, G, L, \tau, M, \bar{c}, r, \kappa_x, c_\Phi$ and $\lambda_{\min}(\Sigma)$, and other logarithmic factors (see Appendix for the explicit forms of m and n).

Remark 13 *Similar to the Gaussian case, Theorem 12 suggests that if we omit all the other terms and assume that $\|w^*\|_\infty = O(1)$, then for any given error $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$, there is an (ϵ, δ)-NLDP algorithm whose sample complexity of private data (n) and public unlabeled data (m) to achieve an estimation error of α (in ℓ_∞-norm), is $\tilde{O}(p^2\epsilon^{-2}\alpha^{-2})$ and $O(p^2\alpha^{-2})$, respectively. While compared with the Gaussian case, here we need larger m. However, as we will see in the experiments section, in practice we do not need such a large size for public data.*

Compared with the complexity for private data in the Gaussian case, it seems that the complexity in the sub-Gaussian case is less. However, due to different measure of estimation error (ℓ_2-norm v.s. ℓ_∞-norm) and different assumptions ($\|w^*\|_2 = O(1)$ v.s. $\|w^*\|_\infty = O(1)$), these two results are incomparable.

Compared with the previous work on linear regression in the NLDP model. Our sample complexities for the general GLMs seem worse than the previous results. However, these results are incomparable due to different settings and assumptions. Specifically, when $\|x_i\|_2 \leq 1$ and $\|w^*\|_2 \leq 1$, Smith et al. (2017) proposed an algorithm with a sample complexity of $\tilde{O}(p\alpha^{-2}\epsilon^{-2})$ for the optimization error. While in this paper, we mainly focus on the estimation error. Moreover, in the Gaussian case, we have each $\|x_i\|_2 \leq O(\sqrt{p})$ with high probability and in the sub-Gaussian case we assume $\|w^*\|_\infty \leq O(1)$, these assumptions are different with the assumptions in (Smith et al., 2017). Zheng et al. (2017) proposed an algorithm whose sample complexity is $\tilde{O}(\alpha^{-4}\epsilon^{-2})$ for the optimization error, under the assumptions of $\|x_i\|_1 \leq 1$ and $\|w^*\|_1 \leq 1$, which are also different with ours. Recently, Wang and Xu (2019) also considered the ℓ_2-norm statistical error. It relies on assumptions that w^* is 1-sparse, which is unnecessary in our setting. Besides these differences, we also have to mention that in this paper we need some additional assumptions (i.e., Assumption 1) on the data distribution compared with those previous results.

Remark 14 *Algorithm 1 and 2 have several advantages over the existing approaches. Firstly, different from the approaches that are based on (Stochastic) Gradient Descent methods to solve DP-ERM* (e.g., *(Wang et al., 2017)), our algorithm is parameter-free, i.e., we do not need to choose a specific step size, an iteration number or initial vectors. Secondly, compared with some previous work on GLM in the NLDP model such as (Zheng et al., 2017; Smith et al., 2017; Wang et al., 2019b), all of our above results do not need to assume the loss function $\Phi(\cdot)$ is convex. Thirdly, since the private data only contributes to obtaining the OLS estimator, and only the constant $\hat{c}_\Phi$ depends on the loss function Φ, these indicate that with probability at least $1 - T\exp(-\Omega(p)) - \xi$, our algorithm can simultaneously be implemented on T different loss functions to achieve the same error α for each loss with almost the same sample complexity as in Theorem 12 or Theorem 8 (if they all satisfy the corresponding assumption). This implies that we can answer at most $O(\exp(O(p))$ number of GLM queries with constant probability to achieve error α for each query with the same sample complexity as in Theorem 12 (Theorem 8). To our best knowledge, this is the first result that can answer multiple non-linear queries in the NLDP model with polynomial sample complexity. Previous results are either for linear queries (Blasiok et al., 2019; Bassily,* *2018) or in the central DP model (Ullman, 2015). Moreover, we can see when the dimension p increases, we can answer more GLMs queries. It sounds counter-intuitive that one can handle more loss functions with a larger dimension. However, we note that in this case we also need more data samples to achieve the fixed error α.*

Note that in Theorem 12, $\Phi^{(2)}$ is assumed to be bounded. Although this is a commonly-used assumption in previous work such as (Wang et al., 2018, 2019a), this condition can be further relaxed to the condition that $\Phi^{(2)}(\langle x, w\rangle)$ is sub-Gaussian in some range of w.

Assumption 4 *For a random vector x that is sub-Gaussian with zero mean and covariance matrix Σ, we assume the following conditions hold*

- $\sup_{w:\|w-\Sigma^{\frac{1}{2}}w^{ols}\|_2\leq 1}\|\Phi^{(2)}(\langle x, w\rangle)\|_{\psi_2} \leq \kappa_g$ *for some constant κ_g and $\Phi^{(3)}(\cdot)$ is G-Lipschitz.*
- *For some constant $\bar{c}$ and $\tau > 0$, the function $f(c) = c\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle c)]$ satisfies the condition of $f(\bar{c}) \geq 1+\tau$, where w^{ols} is in Lemma 10 and the distribution of x satisfies Assumption 1.*
- *The derivative of f in the interval $[0, \max\{\bar{c}, c_\Phi\}]$ does not change the sign (*i.e., *its absolute value is lower bounded by some constant $M > 0$), where c_Φ is in Lemma 10.*

Theorem 15 *Under Assumption 1 and 4, for sufficiently large m, n such that*

$$m \geq \tilde{\Omega}\Big(\frac{\epsilon^2 np}{(\mathbb{E}[\|x\|_2])^2\|w^{ols}\|_2^2}\Big), \tag{8}$$

$$n \geq \tilde{\Omega}\Big(\frac{p^2\|w^*\|_\infty^2\max\{1, \|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{1}{\xi}}{\epsilon^2}\Big). \tag{9}$$

Then the following holds with probability at least $1-\exp(-\Omega(p))-\xi$,

$$\|\hat{w}^{glm}-w^*\|_\infty \leq \tilde{O}\Big(\frac{p\|w^*\|_\infty\max\{1, \|w^*\|_\infty^3\}\sqrt{\log\frac{1}{\delta}\log\frac{1}{\xi}}}{\epsilon\sqrt{n}}+$$
$$\frac{\|w^*\|_\infty^2\max\{1, \|w^*\|_\infty^2\}}{\sqrt{p}}+\|w^*\|_\infty\max\{1, \|w^*\|_\infty\}\frac{1}{\mathbb{E}[\|x\|_2]}\frac{p}{\sqrt{m}}\Big), \tag{10}$$

where the Big-$\tilde{O}$ and Big-$\tilde{\Omega}$ notations omit the terms of $\rho_2, \rho_\infty, \|\Sigma\|_2$, $\lambda_{\min}(\Sigma)$, r, κ_x, κ_g, $G, M, \tau, \bar{c}$, c_Φ and other logarithmic factors (see Appendix for the explicit forms of m and n).

From the above theorem, we can see that even with more relaxed assumptions, to achieve the ℓ_∞-norm error α, the sample complexities in Theorem 15 is asymptotically the same as the ones in Theorem 12 up to some logarithmic factors (if we omit other terms and m, n satisfying (8) and (9)).

A not-so-desirable issue of Theorem 8, 12, and 15 is that they need quite a few assumptions/conditions. Although some of them commonly appear in some related work, the assumptions on function f seem to be a little weird. Fortunately, this is not a big issue in both practice and theory. For the theory side, in the following, motivated by (Erdogdu et al., 2019), we will provide two examples that satisfy Assumption 2. Moreover, for the practical side, as we will see later, our experiments show that the algorithm performs quite well for many loss functions that may not satisfy these assumptions (such as the cubic function). Also, we note that the error bounds in Theorem 12 and 15 are dependent on the ℓ_1-norm of x_i, while the previous results only depend on the ℓ_2-norm bound (Smith et al., 2017; Zheng et al., 2017). We leave the problem of relaxing/lifting these assumptions for future research.

Theorem 16 (Logistic Loss) *Consider the model (1) where the function $\Phi(z) = \log(1+e^z)$ (then $|\Phi^{(2)}(\cdot)| \leq 1$ and $\Phi^{(2)}(\cdot)$ is 1-Lipschitz), $x \sim \mathcal{N}(0, \frac{1}{p}I_p)$, $\|w^*\|_2 = \frac{\sqrt{p}}{4}$ and $\|w^{ols}\|_2 = \frac{\sqrt{p}}{20}$. Then when $\bar{c} = 6$ and $\tau = 0.22$, the function $f(c) = c\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle c)] > 1+\tau$. Moreover, $f'(z)$ is bounded by constant $M = 0.1$ on $[0, \bar{c}]$ from below and $c_\Phi < \bar{c}$.*

Theorem 17 (Boosting Loss) *Consider the model (1) where the function $\Phi(z) = \frac{z}{2} + \sqrt{1+\frac{z^2}{4}}$ (then $|\Phi^{(2)}(\cdot)| \leq \frac{1}{4}$ and $\Phi^{(2)}(\cdot)$ is $\frac{3}{16}$-Lipschitz), $x \sim \mathcal{N}(0, \frac{1}{p}I_p)$, $\|w^*\|_2 = \frac{\sqrt{p}}{4}$ and $\|w^{ols}\|_2 = \frac{\sqrt{p}}{20}$. Then when $\bar{c} = 6$ and $\tau = 0.22$, the function $f(\bar{c}) = \bar{c}\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle \bar{c})] > 1+\tau$. Moreover, $f'(z)$ is bounded by constant $M = 0.1$ on $[0, \bar{c}]$ from below and $c_\Phi < \bar{c}$.*

5. Privately Estimating Non-linear Regressions

In this section, we extend our ideas in the previous section to the problem of estimating non-linear regressions in the NLDP model with public unlabeled data. Specifically, we assume that there is an underlying vector $w^* \in \mathbb{R}^p$ with $\|w^*\|_2 \leq 1$ such that

$$y = f(\langle x, w^* \rangle) + \sigma, \tag{11}$$

where x is the feature vector sampled from some distribution (for simplicity, we assume its mean is zero) and y is the response. σ is a zero-mean noise which is independent of x and is bounded by some constant $C = O(1)$ (*i.e.*, $\sigma \in [-C, C]$). f is some known differentiable link function with $f(0) \neq \infty$ [5]. Notably, these assumptions have also been used in some previous work such as (Wang and Xu, 2019; Duchi and Ruan, 2018) in other privacy models. In our model, the goal is to obtain some estimator w^{priv} of w^*, based on the private dataset $D = \{(x_i, y_i)\}_{i=1}^n$ and the public unlabeled dataset $D' = \{x_j\}_{j=n+1}^{n+m}$ via some NLDP algorithm.

5.1 Gaussian Case

Similar to the previous section, we first consider the case where $x \sim N(0, \Sigma)$ with some unknown $\Sigma \in \mathbb{R}^{p\times p}$. Motivated by Lemma 5, we first show the following result via Stein's lemma.

Theorem 18 *If $x \sim \mathcal{N}(0, \Sigma)$, then w^* in (11) can be written as $w^* = c_f \times w^{ols}$, where c_f is the fixed point of $z \mapsto (\mathbb{E}[f'(\langle x, w^{ols}\rangle z)])^{-1}$ (if we assume that $\mathbb{E}[f'(\langle x, w^{ols}\rangle z)] \neq 0$) and $w^{ols} = \Sigma^{-1}\mathbb{E}[xy]$ is the OLS vector.*

[5] This assumption can be relaxed to ”there is a point x such that $f(x) \neq 0$”.

We can see that the result in Theorem 18 is similar to Lemma 5 where we replace the function $\Phi^{(2)}(\cdot)$ by function $f'(\cdot)$. Thus, based on the idea of Algorithm 1, we have Algorithm 3.

Algorithm 3 Non-interactive LDP for smooth non-linear regression with public data (Gaussian)

```
Input: Private data $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ with $\{x_i\}_{j=1}^{n+m} \sim \mathcal{N}(0, \Sigma)$ for some unknown $\Sigma$ and $\{x_j\}_{j=n+1}^{n+m}$ are public, link function $f : \mathbb{R} \mapsto \mathbb{R}$, privacy parameters $\epsilon, \delta$, and initial value $c \in \mathbb{R}$.
for The server do
    Calculate $\Sigma_m = \frac{1}{m}\sum_{j=n+1}^{n+m} x_j x_j^T$ and send it to each user.
end for
for Each user $i \in [n]$ do
    Let $\bar{x}_i = x_i \min\{1, \frac{r}{\|x_i\|_2}\}$, where $r \equiv \sqrt{20p\|\Sigma_m\|_2 \log n}$.
    Release $\widehat{x_i x_i^T} = \bar{x}_i \bar{x}_i^T + E_{1,i}$, where $E_{1,i} \in \mathbb{R}^{p\times p}$ is a symmetric matrix, and each entry of the upper triangle matrix is sampled from $\mathcal{N}(0, \frac{32r^4 \log \frac{2.5}{\delta}}{\epsilon^2})$.
    $\widehat{x_i y_i} = \bar{x}_i y_i + E_{2,i}$, where the vector $E_{2,i} \in \mathbb{R}^p$ is sampled from $\mathcal{N}(0, \frac{32r^2(Lr+|f(0)|+C)^2 \log \frac{2.5}{\delta}}{\epsilon^2} I_p)$.
end for
for The server do
    Denote $\widehat{X^T X} = \sum_{i=1}^n \widehat{x_i x_i^T}$ and $\widehat{X^T y} = \sum_{i=1}^n \widehat{x_i y_i}$. Calculate $\hat{w}^{ols} = (\widehat{X^T X})^{-1}\widehat{X^T y}$.
    Calculate $\tilde{y}_j = x_j^T \hat{w}^{ols}$ for each $j = n+1, \cdots, n+m$.
    Find the root $\hat{c}_f$ such that $1 = \frac{\hat{c}_f}{m}\sum_{j=n+1}^{n+m} f'(\hat{c}_f \tilde{y}_j)$ using Newton's root finding method:
    for $t = 1, 2, \cdots$ until convergence do
        $c = c - \frac{c\frac{1}{m}\sum_{j=n+1}^{n+m} f'(c\tilde{y}_j)-1}{\frac{1}{m}\sum_{j=n+1}^{n+m}\{f'(c\tilde{y}_j)+c\tilde{y}_j f^{(2)}(c\tilde{y}_j)\}}$.
    end for
end for
return $\hat{w}^{nlr} = \hat{c}_f \cdot \hat{w}^{ols}$.
```

As in the previous section, we need the following assumptions for function $f(\cdot)$.

Assumption 5 *We assume*

- $|f'(\cdot)| \leq L$ *and* $f^{(2)}(\cdot)$ *is G-Lipschitz.*
- *For some constant* $\bar{c}$ *and* $\tau > 0$*, the function* $\ell(c) = c\mathbb{E}[f'(\langle x, w^{ols}\rangle c)]$ *satisfies the condition of* $\ell(\bar{c}) \geq 1 + \tau$*, where* w^{ols} *is in Theorem 18.*
- *The derivative of* ℓ *in the interval* $[0, \bar{c}]$ *does not change the sign,* i.e., *its absolute value is lower bounded by some constant* $M > 0$*.*

Theorem 19 *For any* $0 < \epsilon, \delta < 1$*, Algorithm 3 is* (ϵ, δ) *non-interactive LDP. Moreover, let* $x_1, \cdots, x_n \in \mathbb{R}^p$ *be i.i.d realizations of a random vector* $x \sim \mathcal{N}(0, \Sigma)$*, under Assumption*

5, for sufficiently large m, n such that

$$n \geq \tilde{\Omega}\big(\frac{\|\Sigma\|_2^3 p^3 \|w^*\|_2^2 \log\frac{1}{\delta}\log\frac{1}{\xi}}{\epsilon^2}\big) \tag{12}$$

$$m \geq \Omega\big(\|w^*\|_2^2 p\big). \tag{13}$$

Then for any $\zeta \in (0,1)$, with probability at least $1-\exp(-\Omega(p))-\xi$ we have

$$\|\hat{w}^{nlr} - w^*\|_2 \leq \tilde{O}\big(\frac{p^{\frac{3}{2}}\|w^*\|_2^2 \log\frac{1}{\xi}\sqrt{\log\frac{1}{\delta}\log\frac{1}{\xi}}}{\epsilon\sqrt{n}} + \|w^*\|_2^2\sqrt{\frac{p}{m}}\big),$$

where Big-$\tilde{O}$ and Big-$\tilde{\Omega}$ notations omit the terms of $C, \|\Sigma\|_2, c_f, \tau, G, L, M, \bar{c}$ and $\lambda_{\min}(\Sigma)$, and other logarithmic factors (see Appendix for the explicit form of m and n).

Remark 20 *We can see the sample complexity of public and private data to achieve an ℓ_2-norm estimation error of α is $O(p\alpha^{-2})$ and $\tilde{O}(p^3\alpha^{-2}\epsilon^{-2})$, respectively (if we omit other terms). Compared with Theorem 8, they are asymptotically the same as the bounds in the GLM case. However, it is notable that non-linear regression models differ from GLMs as their conditional density functions of the response y cannot be written in exponential forms. The main reason for this similarity is the similar conclusions in Theorem 18 and Lemma 5. And this is because both of GLMs and non-linear regressions satisfy the property of $\mathbb{E}[xy] = \mathbb{E}[xg(\langle x, w^*\rangle)]$ for some function g (where $g(\cdot) = f(\cdot)$ in non-linear regressions and $g(\cdot) = \Phi'(\cdot)$ in GLMs). Thus, Theorem 18 could be considered the non-linear regression version of Stein's lemma, which may be used in other machine learning and statistics problems.*

5.2 Sub-Gaussian Case

We then consider estimating non-linear regressions where x is sub-Gaussian. We will first use the zero-bias transformation (Goldstein et al., 1997) and the techniques in (Erdogdu et al., 2019) to get a lemma that is similar to Lemma 10 and could be considered as a generalization of Theorem 19 to the sub-Gaussian covariates.

Definition 21 (Zero-bias Transformation) *Let z be a random variable with mean 0 and variance σ^2. Then, there exists a random variable z^* that satisfies $\mathbb{E}[zf(z)] = \sigma^2\mathbb{E}[f'(z^*)]$ for all differentiable functions f. The distribution of z^* is called the z-zero-bias distribution.*

Note that when z is Gaussian, then $z^* = z$, which is just Stein's lemma.

Theorem 22 *Let $x_1, \cdots, x_n \in \mathbb{R}^p$ be i.i.d realizations of a random vector x that is zero-mean sub-Gaussian with covariance matrix Σ and satisfies Assumption 1. Let $v = \Sigma^{-\frac{1}{2}}x$ be the whitened random vector of x and denote $\|v\|_{\psi_2} = \kappa_x$. If each v_i has constant first and second conditional moments and function f' is Lipschitz continuous with constant G, then for $c_f = \frac{1}{\mathbb{E}[f'(\langle x_i, w^*\rangle)]}$, the following holds, where w^{ols} is the OLS vector.*

$$\|\frac{1}{c_f}\cdot w^* - w^{ols}\|_\infty \leq O(Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty^2}{\sqrt{p}}).$$

Algorithm 4 Non-interactive LDP for smooth non-linear regression with public data (General)

```
Input: Private data $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ with $\|x_i\|_1 \leq r$, public unlabeled data $\{x_j\}_{j=n+1}^{n+m}$, link function $f : \mathbb{R} \mapsto \mathbb{R}$, privacy parameters $\epsilon, \delta$, and initial value $c \in \mathbb{R}$.
for Each user $i \in [n]$ do
    Release $\widehat{x_i x_i^T} = x_i x_i^T + E_{1,i}$, where $E_{1,i} \in \mathbb{R}^{p\times p}$ is a symmetric matrix and each entry of the upper triangle matrix is sampled from $\mathcal{N}(0, \frac{32r^4 \log \frac{2.5}{\delta}}{\epsilon^2})$.
    Release $\widehat{x_i y_i} = x_i y_i + E_{2,i}$, where the vector $E_{2,i} \in \mathbb{R}^p$ is sampled from $\mathcal{N}(0, \frac{32r^2(Lr+|f(0)|+C)^2 \log \frac{2.5}{\delta}}{\epsilon^2} I_p)$.
end for
for The server do
    Denote $\widehat{X^T X} = \sum_{i=1}^n \widehat{x_i x_i^T}$ and $\widehat{X^T y} = \sum_{i=1}^n \widehat{x_i y_i}$. Calculate $\hat{w}^{ols} = (\widehat{X^T X})^{-1} \widehat{X^T y}$.
    Calculate $\tilde{y}_j = x_j^T \hat{w}^{ols}$ for each $j = n+1, \cdots, n+m$.
    Find the root $\hat{c}_f$ such that $1 = \frac{\hat{c}_f}{m} \sum_{j=n+1}^{n+m} f'(\hat{c}_f \tilde{y}_j)$ using Newton's root finding method:
    for $t = 1, 2, \cdots$ until convergence do
        $c = c - \frac{c\frac{1}{m}\sum_{j=n+1}^{n+m} f'(c\tilde{y}_j) - 1}{\frac{1}{m}\sum_{j=n+1}^{n+m}\{f'(c\tilde{y}_j) + c\tilde{y}_j f^{(2)}(c\tilde{y}_j)\}}$
    end for
end for
return $\hat{w}^{nlr} = \hat{c}_f \cdot \hat{w}^{ols}$.
```

From Theorem 22, we can see that it shares the same phenomenon as in Lemma 10 (*i.e.,* the OLS vector with some constant could approximate w^* well). Thus, a similar idea to Algorithm 2 can be used to solve this problem, which gives us Algorithm 4 and the following theorem. Like the previous section, we need the following assumptions for function $f(\cdot)$.

Assumption 6 *We assume the following conditions hold:*

- $|f'(\cdot)| \leq L$ *and* $f^{(2)}(\cdot)$ *is G-Lipschitz.*
- *For some constant* $\bar{c}$ *and* $\tau > 0$*, the function* $\ell(c) = c\mathbb{E}[f'(\langle x, w^{ols}\rangle c)]$ *satisfies the condition of* $\ell(\bar{c}) \geq 1 + \tau$*, where* w^{ols} *is in Theorem 22.*
- *The derivative of* ℓ *in the interval* $[0, \max\{\bar{c}, c_f\}]$ *does not change the sign (*i.e., *its absolute value is lower bounded by some constant* $M > 0$*), where* c_f *is in Theorem 22.*

Theorem 23 *For any* $0 < \epsilon, \delta < 1$*, Algorithm 4 is* (ϵ, δ) *non-interactive LDP. Under the assumptions of Theorem 22, and if the link function* f *satisfies Assumption 6, then for sufficiently large* m, n *such that*

$$m \geq \Omega\big(\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} p^2\big), \tag{14}$$

$$n \geq \tilde{\Omega}\big(\frac{p^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \log \frac{1}{\delta} \log \frac{1}{\xi}}{\epsilon^2}\big). \tag{15}$$

Then for any $\zeta \in (0,1)$, with probability at least $1-\exp(-\Omega(p))-\xi$, the output of Algorithm 4 satisfies

$$\|\hat{w}^{nlr} - w^*\|_\infty \leq \tilde{O}\Big(\frac{\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}p}{\sqrt{m}} + \frac{\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}p\sqrt{\log\frac{1}{\delta}\log\frac{1}{\xi}}}{\epsilon\sqrt{n}} + \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}}\Big), \quad (16)$$

where Big-$\tilde{O}$ and Big-$\tilde{\Omega}$ notations omit the terms of $\lambda_{\min}(\Sigma)$, $\|\Sigma\|_2$, ρ_2, ρ_∞, G, L, τ, M, $\bar{c}$, r, κ_x, C and c_f, and other logarithmic factors (see Appendix for the explicit form of m and n).

Remark 24 *Similar to Theorem 15, we can see the sample complexity of public and private data for Algorithm 4 to achieve an ℓ_∞-norm estimation error of α is $O(p^2\alpha^{-2})$ and $\tilde{O}(p^2\alpha^{-2}\epsilon^{-2})$, respectively, if α is not too small (i.e., $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$). Such similarity is due to the similar conclusions in Theorem 22 and Lemma 10. Consider the simplest case where the covariate vector x has i.i.d. entries with mean 0 and variance 1. Then by using the zero-bias transformation to the j-th coordinate of $\mathbb{E}[yx]$ we have*

$$\mathbb{E}[yx_j] = \mathbb{E}[x_j f(\langle x, \hat{w}^*\rangle)] = w_j^*\mathbb{E}[f'((x_j^* - x_j)w_j^* + x_j w_j^* + \sum_{i\neq j} x_i w_i^*)]. \quad (17)$$

If w^* is well spread, it turns out that taken together with $j = 1, \cdots, p$, the right-hand side in (17) behaves similarly to the Gaussian case, where the proportionality relationship given in Theorem 18 holds. As we mentioned in Remark 20, Theorem 18 is similar to Lemma 5. Thus, Theorem 22 behaves similarly to Lemma 10.

In the following, we will provide an instance of $f(\cdot)$ and x that satisfies the assumptions in Theorem 19.

Theorem 25 (Sigmoid Link Function) *Consider the model (11) where the link function $f(z) = \frac{1}{1+e^{-z}}$, $x \sim \mathcal{N}(0, \frac{1}{p}I_p)$, $\|w^*\|_2 = \frac{\sqrt{p}}{4}$ and $\|w^{ols}\|_2 = \frac{\sqrt{p}}{20}$. Then when $\bar{c} = 6$ and $\tau = 0.22$, the function $\ell(c) = c\mathbb{E}[f'(\langle x, w^{ols}\rangle c)] > 1 + \tau$. Moreover, $\ell'(z)$ is bounded by constant $M = 0.1$ on $[0, \bar{c}]$ from below and $c_f \leq \bar{c}$.*

6. Experiments

In this section, we will evaluate the performance of our methods on both synthetic and real-world datasets. The experiments demonstrate the previous utility results of our algorithms and suggest that they are efficient. Moreover, we will show that our algorithms only need a few public unlabeled data to achieve outstanding performance.

6.1 Experimental Settings

Link functions. In this paper, we mainly study estimating GLMs and non-linear regressions, and we will use variants of loss functions and link functions for each algorithm. Specifically,

- For Algorithm 1 we consider the binary logistic regression where we have $\Phi(\langle x, w\rangle) = \ln(1+\exp(\langle x, w\rangle))$ in (1).
- For Algorithm 2, in addition to binary logistic regression, we also consider the exponential regression where $\Phi(\langle x, w\rangle) = e^{\langle x, w\rangle}$ in (1).
- For Algorithm 3 we consider the sigmoid link function, i.e., $f(\langle x, w\rangle) = \frac{1}{1+e^{\langle x, w\rangle}}$ in (11).
- For Algorithm 4 we consider the cubic link function where $f(\langle x, w\rangle) = \frac{1}{3}\langle x, w\rangle^3$ and the logistic function where $f(\langle x, w\rangle) = \log(1+e^{-\langle x, w\rangle})$ in (11).

Synthetic data generation. In this paper, we assume the distribution of the data feature vector is either Gaussian or sub-Gaussian with bounded ℓ_1-norm. Specifically, for each case, we consider the following procedure for feature vector generation.

- For Gaussian distributions, we consider two cases for the covariance matrix: (1) the covariance matrix is diagonal, i.e., $\Sigma = \sigma^2 I_p$ where σ is sampled from the uniform distribution in $[0, 1]$; (2) the covariance matrix is non-diagonal, and we assume $\Sigma = U + I_p$, where $U \in \mathbb{R}^{p\times p}$ is a random orthogonal matrix.
- In the sub-Gaussian case, each entry of each feature vector is generated independently from a Bernoulli distribution $\Pr\left(x_{i,j} = \pm\frac{1}{p}\right) = 0.5$.

After getting feature vectors, we generate the underlying parameter w^*, which is a random unit vector. After that, we generate the responses $\{y_i\}_{i=1}^n$ as follows.

- For GLMs, each response is generated according to its underlying model in (1). Specifically, for the logistic regression we generate $y_i = \frac{e^{\langle w, x_i\rangle}}{1+e^{\langle w, x_i\rangle}}$. For exponential regression we generate $y_i = e^{\langle x_i, w^*\rangle}$. Since we only consider the exponential loss for the sub-Gaussian case where $\langle x, w^*\rangle \leq 1$. Thus, in both cases $\{y_i\}_{i=1}^n$ are bounded.
- For non-linear regressions, each y_i is generated according to the model (11) where σ is bounded by $C = 0.05$. Notably, $\{y_i\}_{i=1}^n$ are bounded for all link functions we considered.

Experimental settings for synthetic data. Motivated by the results in previous sections, for data with Gaussian features vectors, we will use the (squared) relative ℓ_2-norm error $\frac{\|\hat{w}-w^*\|_2^2}{\|w^*\|_2^2}$ as the performance metric. Otherwise, we will use the (squared) relative ℓ_∞-norm error $\frac{\|\hat{w}-w^*\|_\infty^2}{\|w^*\|_\infty^2}$. For privacy parameters, we will choose ϵ between 4 to 20 and set $\delta = \frac{1}{n^{1.1}}$.[6] For dimension p, we choose from the set $\{5, 10, 15, 20, 25, 30, 40, 50, 60\}$. For different experiments, we will vary different private sample sizes n. However, we will always set the size of public unlabeled data m to be smaller than n. Specifically, without specification, we will always set $m = \lfloor \frac{n}{p^2} \rfloor$. For each experiment above, we repeated 20 times and reported the errors' average and variance.

[6] Note that in the studies on LDP ERM such as (Bhowmick et al., 2018), ϵ is always chosen as a large value. Moreover, we can use the shuffling technique in (Erlingsson et al., 2019) for privacy amplification.

Experimental settings for real-world data: We conduct experiments on binary logistic regression for GLMs on the Covertype dataset (Dua and Graff, 2017), the SUSY dataset (Baldi et al., 2014) and the Skin Segmentation dataset (Dua and Graff, 2017).

For the Covertype dataset, before running our algorithm, we first normalize the data and remove some co-related features. After the pre-processing, the dataset contains 581,012 samples and 44 features. There are seven possible values for the label. Here we consider a weaker test to classify whether the label is Lodgepole Pine (type 2) or not. We divide the data into training data and test data, where $n_{\text{training}} = 350,000$ and $n_{\text{testing}} = 200,000$ (other data will be used as the public unlabeled data). We randomly choose the sample size $n \in 10^4 \cdot \{10, 15, 20, 25, 30, 35\}$ from the training data and set the size of public unlabeled data as $m = 10^4$. Regarding the privacy parameter, we take $\delta = \frac{1}{n^{1.1}}$ and let ϵ take value from 4 to 15. We measure the performance by prediction accuracy. For each experiment above, we repeated 20 times and reported the errors' average and variance.

For the SUSY dataset, the task is to classify whether the class label is signal or background. After pre-processing and sampling, the dataset contains $500,000$ samples and 18 features. Then we divide the data into training data and test data, where $n_{\text{training}} = 450,000$ and $n_{\text{testing}} = 30,000$ (other data will be used as the public unlabeled data). We randomly choose the sample size $n \in 10^4 \cdot \{10, 15, \cdots, 45\}$ from the training data and set the size of public unlabeled data as $m = 10^4$. Regarding the privacy parameter, we take $\delta = \frac{1}{n^{1.1}}$ and let ϵ takes value from 2 to 10. We measure the performance by prediction accuracy. For each experiment, we repeated it 20 times.

The task for the Skin Segmentation dataset is to classify whether the class label is a Skin or Nonskin image. After pre-processing, the dataset contains $245,057$ samples and 3 features. We divide the data into training data and test data, where $n_{\text{training}} = 180,000$ and $n_{\text{testing}} = 5,000$ (other data will be used as the public unlabeled data). We randomly choose the sample size $n \in 10^4 \cdot \{2, 4, \cdots, 18\}$ from the training data and set the size of public unlabeled data as $m = 5,000$. We take $\delta = \frac{1}{n^{1.1}}$ for the privacy parameter. As the dimension of the feature vector is only 3, here we consider the high privacy regime and let ϵ take value from 0.2 to 1. We measure the performance by prediction accuracy. For each experiment, we repeated it 20 times.

Baseline and other methods. For synthetic data, as we mentioned previously, there is no previous work that provides efficient methods. Thus, we will use the Logistic Regression classifier in the scikit-learn library (Pedregosa et al., 2011) as the (non-private) baseline. Otherwise, we use the standard gradient descent as the baseline method. For real-world data, we use the Logistic Regression classifier as the baseline.

Besides the non-private baseline, as we mentioned in Remark 9, we can adapt our idea to design a 2-round LDP algorithm without using public unlabeled data (suppose we already know $\|\Sigma\|_2$ for the Gaussian case). In the first round, we get $\hat{w}^{ols}$ by using half the privacy budget, and the server sends it to all the users. In the second round, each user uses another half privacy budget to compute $\tilde{y}_j = x_j^T \hat{w}_{ols}$, then performs the clipping step to $\tilde{y}_j$ to project $\tilde{y}_j$ on the range of y and adds Gaussian noise to the clipped value. Finally, each user sends the noisy version of $\tilde{y}_j$ to the server. Then the server uses these perturbed $\tilde{y}_j$ to estimate the constant of c_Φ. We provide the details of the algorithm for the Gaussian covariates case in GLMs as an example in Appendix E; the other algorithms are similar. We call such algorithms 2-round algorithms. We will compare our methods with these 2-round algorithms. Note that the primary purpose for such comparison is not to claim there is no efficient 2-round LDP algorithm for ERM. We aim to demonstrate that the direct 2-round extensions of our methods will lead to worse performance, as mentioned in Remark 9.

6.2 Experimental Results

We aim to answer the following questions through experiments: (1) When the dimension p and the privacy budget ϵ are fixed, for synthetic data, what is the trend of the relative ℓ_2-norm or ℓ_∞-norm error with different private data size n? (2) What is the accuracy trend when n or ϵ increases for real-world data? What is the difference between our private estimator's accuracy and the non-private method's accuracy? (3) When the private data size n and private parameter ϵ are fixed. How will the dimension p affect the utility? (4) How will the number of public unlabeled data size m affect the (relative) error and the accuracy? (5) While those 2-round algorithms are heuristic, do they perform well? Moreover, compared to the 2-round algorithms, do our non-interactive methods perform better?

We conduct experiments for each method to answer the above questions; see Figure 1-18 for details. Specifically, in Figure 1-4 we consider the performance of Algorithm 1 on Gaussian data whose covariance matrix is either diagonal or non-diagonal. In Figure 5-8, we consider the performance of Algorithm 2 for Bernoulli data where the loss function could be either the exponential loss or the logistic loss. The results for Algorithm 3 are presented in Figure 9-10, where the link function is the sigmoid function and the covariance matrix of Gaussian is diagonal. For Algorithm 4, its experimental results in the case where the link function is either sigmoid or logistic are shown in Figure 11-14. Besides synthetic data, in Figure 15 and 16, we show the performance of Algorithm 2 for binary logistic regression on several real-world datasets. Moreover, in Figure 17, we show the error curves w.r.t the privacy parameter ϵ with different sample sizes for all methods. Finally, in Figure 18, we compare our algorithms with their corresponding 2-round LDP algorithms.

From (a), (b), and (c) in Figure 1, 3, 5, 7, 9, 11 and 13, we can first see that with different link functions, data distributions, and covariance matrices, when the dimension p is fixed, except for some cases, the (squared) relative (ℓ_2-norm or ℓ_∞) error will decrease when n becomes larger in general, which means the private estimator will be sufficiently closed to the baseline and the underlying parameter. Moreover, when n gets larger, the error will almost remain unchanged. This is because besides the private data size n, the error also theoretically depends on the public data size m. Secondly, from the above results, we also observe that the error will decrease when ϵ becomes larger. Moreover, from Figure 17, we can observe for all our methods, the relative error is (approximately) proportional to $\frac{1}{\epsilon^2}$, which matches our theoretical results. However, we can also see that when in the low privacy regime, i.e., when ϵ is large (e.g., $\epsilon = 10$), the relative error decreases slightly, and its curve becomes flat when n becomes larger. Our previous theoretical results show that the error will be dominated by the term related to m instead of n and ϵ in this case. Besides the relative error, in (a), (b), and (c) of Figure 15, we compare the classification accuracy on test data. Here we can get similar conclusions as in the synthetic data case. Furthermore, we can see that when the private data size n and the privacy parameter ϵ are large enough, the accuracy of our private estimator will be close to the accuracy of the non-private method. For example, for Covertype data, the accuracy of the non-private logistic regression is about 75% where our private estimator could achieve about 72.5% accuracy when $\epsilon = 15$ and $n = 3.5 \times 10^5$.

In (a) of Figure 2, 4, 6, 8, 10, 12 and 14 we present the results of relative error w.r.t different n and dimension p. From all the figures, we can see that the relative error increases as the dimension increases. However, it may seem a little weird that the relative error is not linear in the dimension, which was shown in the previous sections theoretically. We note that since the error depends on many terms theoretically; thus, when the dimension p changes, some other parameters, for example, the l_2 norm of the covariance matrix and $\|w^*\|_\infty$ will also change, which bring other effects to the relative error. Moreover, we can see from some results, such as $p = 40$ in Figure 14, even when $n = 7 \times 10^5$, the error is still unsatisfactory. This is because, in theory, the number of efficient samples is $\sqrt{n}$ since the dependency on n is $\frac{1}{\sqrt{n}}$ in error bound, which means the efficient sample size is only about 836. However, as we mentioned in the Related Work section (Section 2), even in the interactive LDP model, the dependency on n is also $\frac{1}{\sqrt{n}}$, and this is optimal (Duchi et al., 2013). Thus, large-scale data is essential for the LDP model, not only for our algorithms. Finally, from all the above results, we can see that when the dimension p is larger, or the sample size n and privacy level ϵ are smaller, the error variance becomes larger. This is mainly because, in these cases, the noise added to each sample becomes larger.

Next, we consider the effect of public unlabeled data. As mentioned earlier, in the experiments for synthetic data, we always set $m = \lfloor \frac{n}{p^2} \rfloor$, which means m is far less than n. Thus, from (a), (b), and (c) in Figure 1, 3, 5, 7, 9, 11 and 13 we can see even smaller public data size m can already achieve outstanding performance, i.e., there is no need to use a large amount of public data as our theoretical result requires to guarantee good performance. Thus, we conjecture that we can theoretically improve the bound on m, and we will leave it as future research. Moreover, we evaluate the performance of our algorithms with different m; see (b) and (c) of Figure 2, 4, 6, 8, 10, 12, 14 and 16 for details. Unlike the conclusions in the previous paragraphs, we can see that the error trend becomes complicated with different sizes of public data. Specifically, in the case when ϵ is large (such as $\epsilon = 10$), we can see it is sufficient to use a size smaller than $\lfloor \frac{n}{p^2} \rfloor$ for public data to achieve good performance. Moreover, the effect of the public data size is limited when m is large, as all curves are quite flat. For example, in (b) and (c) of Figure 2, we can see that when $m = 500$, the algorithm could achieve similar performance as in the case when $m = 3000$. However, when ϵ is small, we can see the trend of the relative error becomes more unstable when m increases. In some cases, larger m may decrease the relative error (such as $\epsilon = 5$ in (b) of Figure 4) while in some cases larger m may even increase the error (such as $\epsilon = 4$ in (c) of Figure 8). Such instability is mainly because when ϵ is smaller, the error variance becomes larger as the noise is more significant. However, we can also see that no matter whether increasing m will increase or decrease the error, the effect of such error change is relatively tiny unless m is sufficiently small.

Finally, we compare our algorithms with the above 2-round LDP algorithms in Figure 18. From all of those four figures, we can see that in most cases, the relative error of the 2-round LDP algorithm is relatively large compared with our methods, and its curve is quite unstable. In Figure 18(d), we can see the performance of the 2-round algorithm becomes acceptable under the setting of Algorithm 4 for cubic link function with $p = 40$ for Bernoulli data. However, our method still significantly outperforms the the 2-round algorithm in this case.

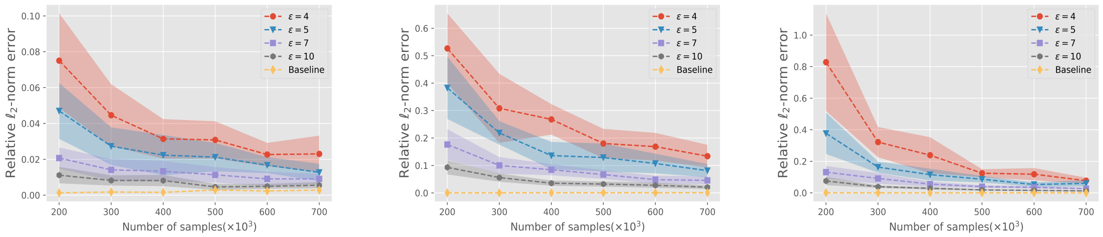
7. Conclusion and Open Problems

In this paper, motivated by Stein's lemma and its variants, we proposed the first efficient algorithm with polynomial sample complexity for Generalized Linear Models estimation in the Non-interactive Local Differential Privacy model with some public unlabeled data. The main idea of our algorithm is to use the OLS (Ordinary Least Square) estimator to approximate the underlying one. The critical observation is that, after multiplying the OLS vector by some constant, we can get a new estimator that is sufficiently close to the underlying estimator. Thus, in our approach, we use private data to estimate the OLS vector and public unlabeled data to estimate the constant. Moreover, we adopted similar ideas to the problem of estimating non-linear regressions and showed similar theoretical results. Finally, we provided intensive experiments of our methods on synthetic and real-world data. Most of the results support our theoretical analysis and show the effectiveness of our methods.

Besides the open problems we mentioned in the previous sections, many other problems remain. First, this paper mainly focused on the low dimensional case, where $n \gg p$. How to generalize to the high dimensional sparse case, i.e., $n \ll p$ and $\|w^*\|_0 \leq k$? In this case, since Stein's lemma will not be held, so we need new techniques. Second, from the experimental results, we can see that even if the loss function and the dataset do not satisfy our assumptions, they still perform well. Thus, how to relax these assumptions and reduce the sample complexity of public unlabeled data in our theoretical results? Finally, for the sub-Gaussian case in both GLMs and non-linear regressions, our estimators are biased, and the error is $\Omega(\frac{1}{\sqrt{p}})$. Can we get unbiased and consistent estimators?

Acknowledgments

Di Wang and Lijie Hu are supported in part by the baseline funding BAS/1/1689-01-01, funding from the CRG grand URF/1/4663-01-01, FCC/1/1976-49-01 from the Computational Bioscience Research Center (CBRC) and funding from the AI Initiative REI/1/4811-10-01 of King Abdullah University of Science and Technology (KAUST). Jinhui Xu was partly supported by the National Science Foundation (NSF) under Grant No. CCF-1716400 and IIS-1919492. he research of Jinhui Xu was supported in part by NSF through grants IIS-1910492 and CCF-2200173 and by KAUST CRG10-4663.2. Part of the work was done when Di Wang and Marco Gaboardi were visiting the Simons Institute of the Theory for Computing.

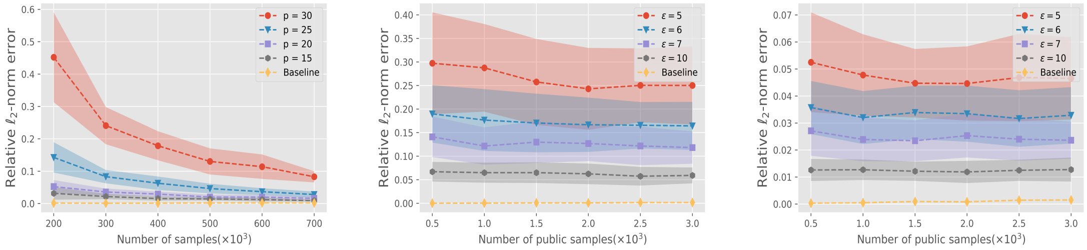


(a) $p = 15$

(b) $p = 20$

(c) $p = 30$

Figure 1: Algorithm 1 for logistic regression where the covariance matrix of Gaussian distribution is diagonal with different dimension p.

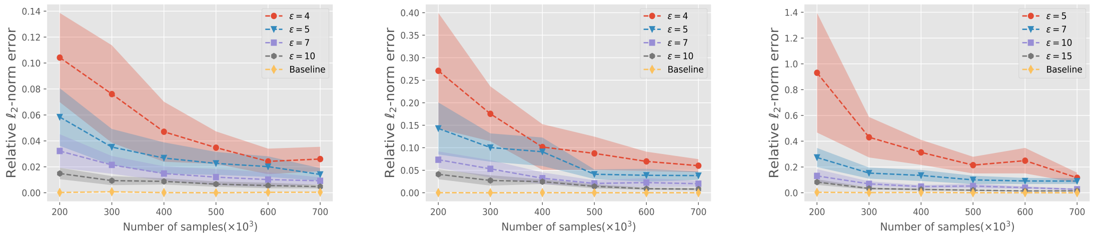


(a) $\epsilon = 5$

(b) $n = 10^5$ and $p = 20$

(c) $n = 4 \times 10^5$ and $p = 20$

Figure 2: Algorithm 1 for logistic regression where the covariance matrix of Gaussian distribution is diagonal. The left plot shows the relative error under different dimension p. The middle and the right plots show the relative error with different size of public data m when n and p are fixed.

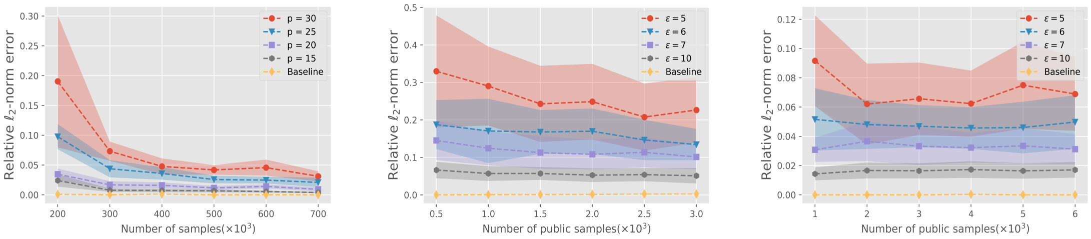


(a) $p = 15$

(b) $p = 20$

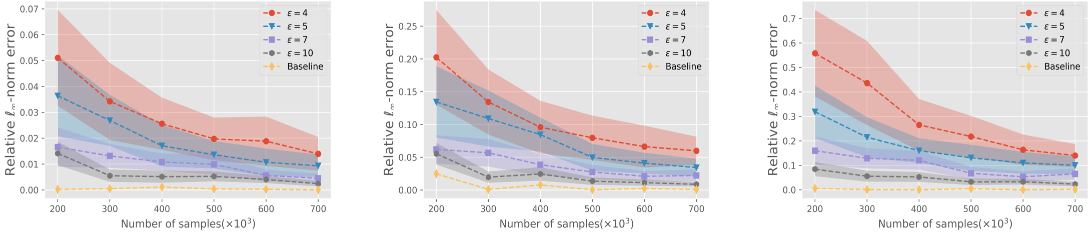
(c) $p = 30$

Figure 3: Algorithm 1 for logistic regression where the covariance matrix of Gaussian distribution is non-diagonal under different dimension p.



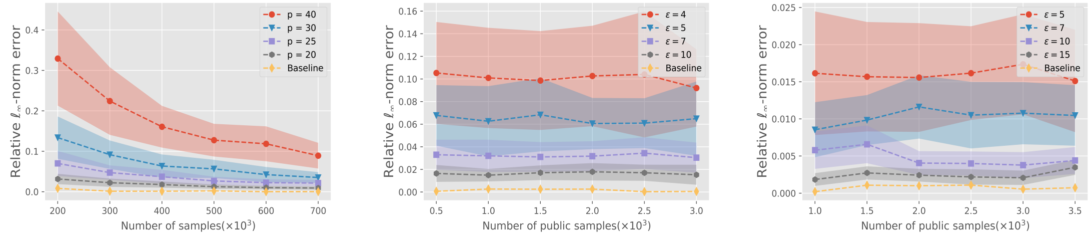
(a) $\epsilon = 5$ (b) $n = 5 \times 10^4$ and $p = 20$ (c) $n = 4 \times 10^5$ and $p = 20$

Figure 4: Algorithm 1 for logistic regression where the covariance matrix of Gaussian distribution is non-diagonal. The left plot shows the relative error with different dimension p. The middle and the right plots show the relative error with different size of public data m when n and p are fixed.



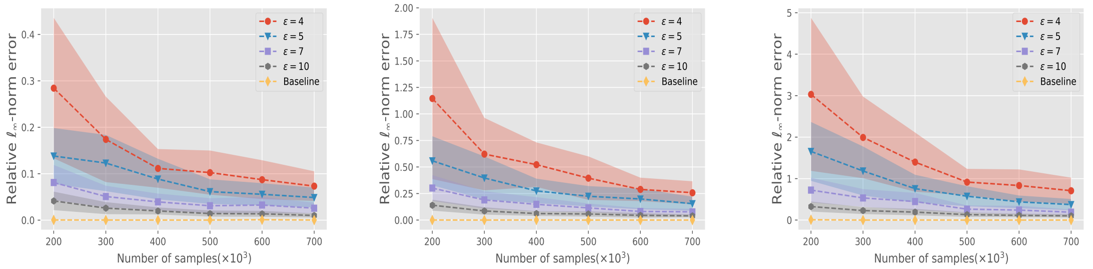
(a) $p = 20$ (b) $p = 30$ (c) $p = 40$

Figure 5: Algorithm 2 for exponential regression with Bernoulli data with different dimension p.



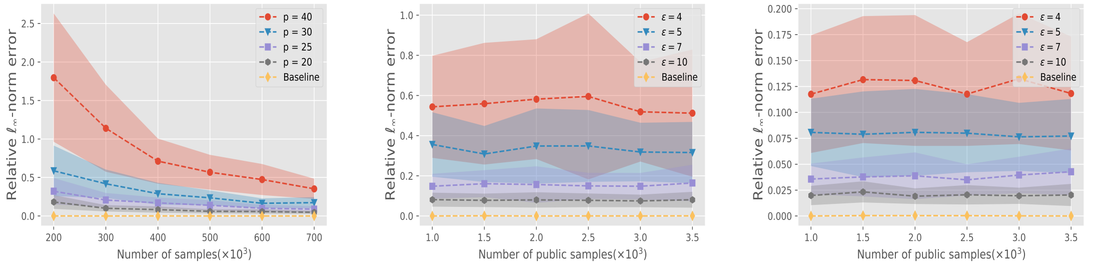
(a) $\epsilon = 5$ (b) $n = 10^5$ and $p = 20$ (c) $n = 4 \times 10^5$ and $p = 20$

Figure 6: Algorithm 2 for exponential regression with Bernoulli data. The left plot shows the relative error with different dimension p. The middle and the right plots show the relative error with different size of public data m when n and p are fixed.



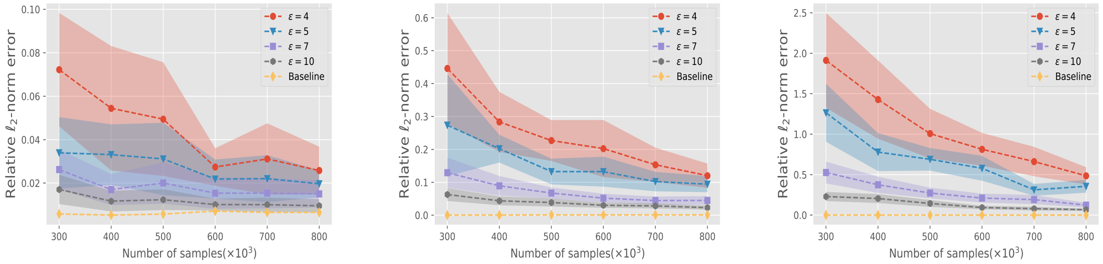
(a) $p = 20$ (b) $p = 30$ (c) $p = 40$

Figure 7: Algorithm 2 for logistic regression with Bernoulli data with different dimension p.



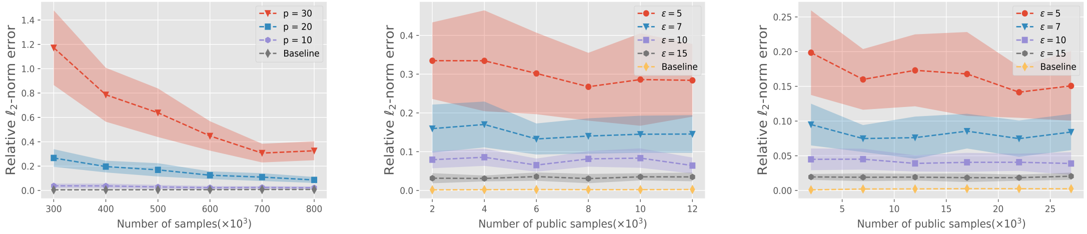
(a) $\epsilon = 5$ (b) $n = 10^5$ and $p = 20$ (c) $n = 4 \times 10^5$ and $p = 20$

Figure 8: Algorithm 2 for logistic regression with Bernoulli data. The left plot shows the relative error with different dimension p. The middle and the right plots show the relative error with different sizes of public data m when n and p are fixed.



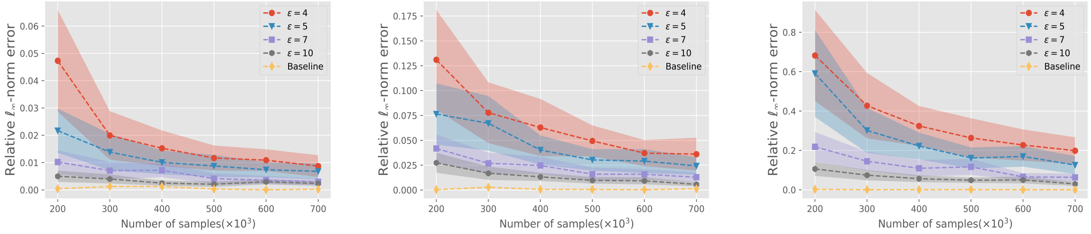
(a) $p = 10$ (b) $p = 20$ (c) $p = 30$

Figure 9: Algorithm 3 with sigmoid link function where the covariance matrix of Gaussian distribution is diagonal with different dimension p.



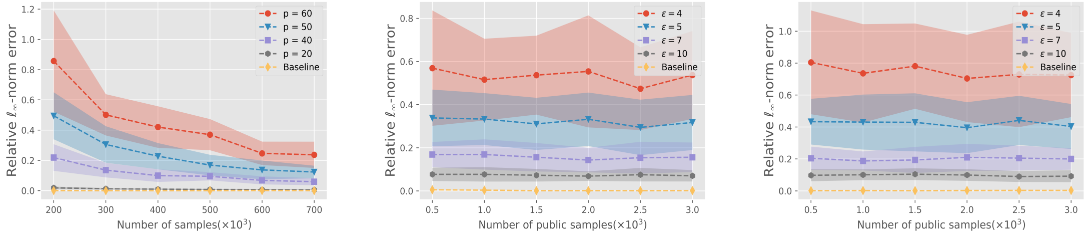
(a) $\epsilon = 5$ (b) $n = 2 \times 10^5$ and $p = 20$ (c) $n = 4 \times 10^5$ and $p = 20$

Figure 10: Algorithm 3 with sigmoid link function where the covariance matrix of Gaussian distribution is diagonal. The left plot shows the relative error with different dimension p. The middle and the right plots show the relative error with different size of public data m when n and p are fixed.



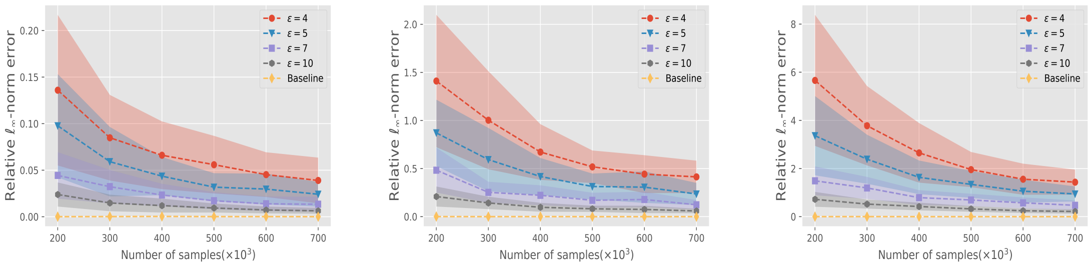
(a) $p = 20$ (b) $p = 30$ (c) $p = 50$

Figure 11: Algorithm 4 for cubic link function with Bernoulli data with different dimension p.



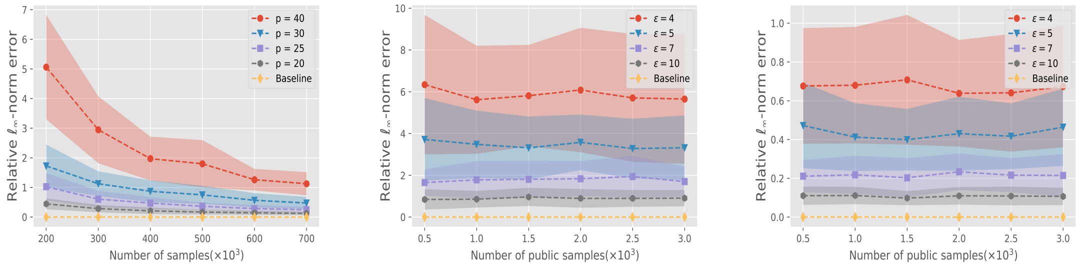
(a) $\epsilon = 5$ (b) $n = 5 \times 10^4$ and $p = 30$ (c) $n = 10^5$ and $p = 40$

Figure 12: Algorithm 4 for cubic link function with Bernoulli data. The left plot shows the relative error with different dimension p. The middle and the right plots show the relative error with different size of public data m when n and p are fixed.



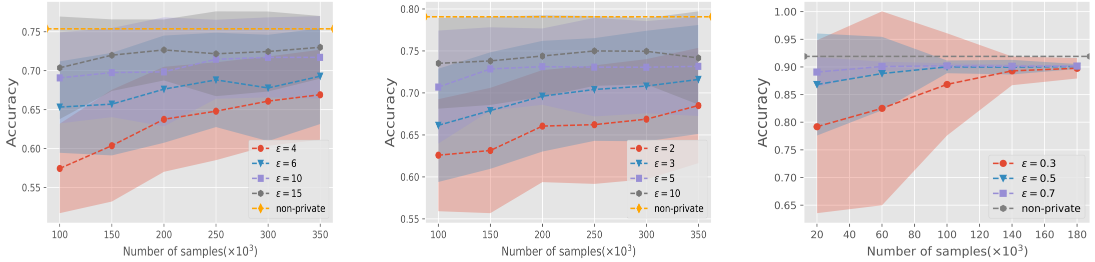
(a) $p = 10$ (b) $p = 20$ (c) $p = 30$

Figure 13: Algorithm 4 for logistic link function with Bernoulli data with different dimension p.



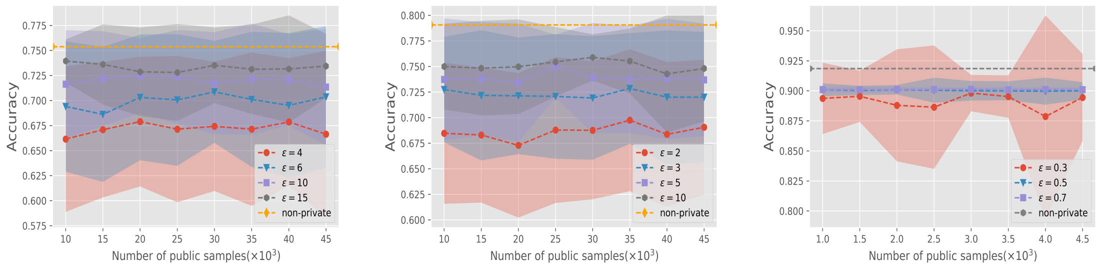
(a) $\epsilon = 7$ (b) $n = 5 \times 10^4$ and $p = 20$ (c) $n = 4 \times 10^5$ and $p = 40$

Figure 14: Algorithm 4 for logistic link function with Bernoulli data. The left plot shows the relative error with different dimension p. The middle and the right plots show the relative error with different size of public data m when n and p are fixed.



(a) Covertype (b) SUSY (c) Skin Segmentation

Figure 15: Algorithm 2 for logistic regression on different RealData.

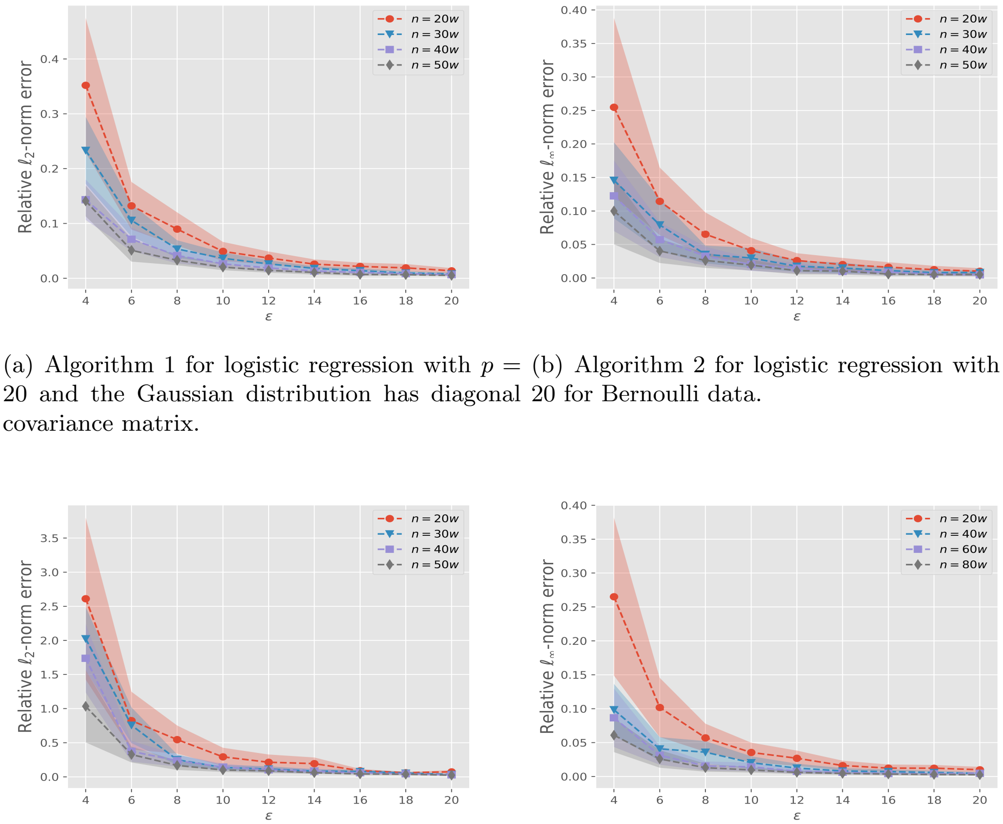


(a) Covertype

(b) SUSY

(c) Skin Segmentation

Figure 16: Algorithm 2 for logistic regression on different RealData with different size of public data m.



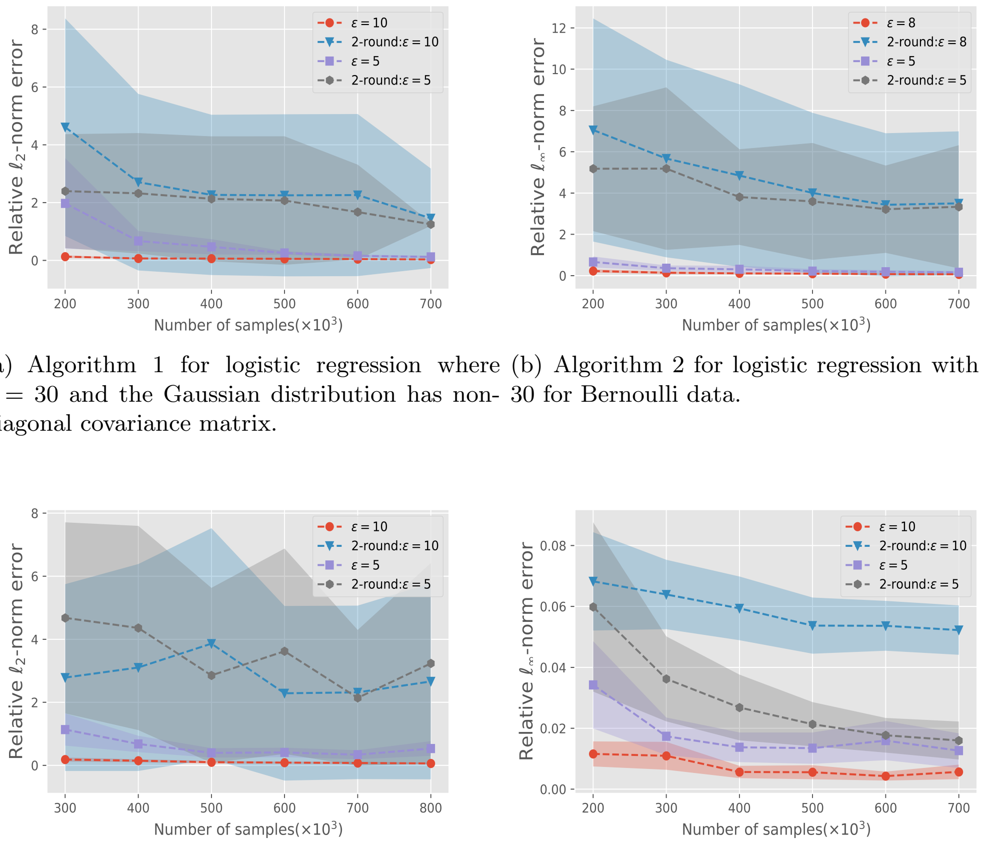
(a) Algorithm 1 for logistic regression with $p = 20$ and the Gaussian distribution has diagonal covariance matrix.

(b) Algorithm 2 for logistic regression with $p = 20$ for Bernoulli data.

(c) Algorithm 3 for sigmoid link function where $p = 20$ and the Gaussian distribution has non-diagonal covariance matrix.

(d) Algorithm 4 for cubic link function with $p = 30$ for Bernoulli data.

Figure 17: Comparison of our methods with different privacy parameter ϵ.



(a) Algorithm 1 for logistic regression where $p = 30$ and the Gaussian distribution has non-diagonal covariance matrix.

(b) Algorithm 2 for logistic regression with $p = 30$ for Bernoulli data.

(c) Algorithm 3 for sigmoid link function where $p = 20$ and the Gaussian distribution has non-diagonal covariance matrix.

(d) Algorithm 4 for cubic link function with $p = 40$ for Bernoulli data.

Figure 18: Comparison of our methods with their corresponding 2-round LDP algorithms.

References

Raman Arora, Raef Bassily, Tomás González, Cristóbal Guzmán, Michael Menart, and Enayat Ullah. Faster rates of convergence to stationary points in differentially private optimization. *arXiv preprint arXiv:2206.00846*, 2022a.

Raman Arora, Raef Bassily, Cristóbal Guzmán, Michael Menart, and Enayat Ullah. Differentially private generalized linear models revisited. *arXiv preprint arXiv:2205.03014*, 2022b.

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5:4308, 2014.

Raef Bassily. Linear queries estimation with local differential privacy. *arXiv preprint arXiv:1810.02810*, 2018.

Raef Bassily and Anupama Nandi. Privately answering classification queries in the agnostic pac model. *arXiv preprint arXiv:1907.13553*, 2019.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

Raef Bassily, Abhradeep Guha Thakurta, and Om Dipakbhai Thakkar. Model-agnostic private learning. In *Advances in Neural Information Processing Systems*, pages 7102–7112, 2018.

Raef Bassily, Cristóbal Guzmán, and Michael Menart. Differentially private stochastic optimization: New results in convex and non-convex settings. *Advances in Neural Information Processing Systems*, 34, 2021.

Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

Jaroslaw Blasiok, Mark Bun, Aleksandar Nikolov, and Thomas Steinke. Towards instance-optimal private query release. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2480–2497. Society for Industrial and Applied Mathematics, 2019.

David R Brillinger. A generalized linear model with "gaussian" regressor variables. In *Selected Works of David Brillinger*, pages 589–606. Springer, 2012.

T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. *arXiv preprint arXiv:2011.03900*, 2020.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

Yuval Dagan and Vitaly Feldman. Interaction is necessary for distributed learning with privacy or communication constraints. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 450–462, 2020.

Amit Daniely and Vitaly Feldman. Learning without interaction requires separation. *arXiv preprint arXiv:1809.09165*, 2018.

Paramveer Dhillon, Yichao Lu, Dean P Foster, and Lyle Ungar. New subsampling algorithms for fast least squares regression. In *Advances in neural information processing systems*, pages 360–368, 2013.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580, 2017.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

John C Duchi and Feng Ruan. The right complexity measure in locally private estimation: It is not the fisher information. *arXiv preprint arXiv:1806.05756*, 2018.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Murat A Erdogdu, Mohsen Bayati, and Lee H Dicker. Scalable approximations for generalized linear problems. *The Journal of Machine Learning Research*, 20(1):231–275, 2019.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.

Larry Goldstein, Gesine Reinert, et al. Stein's method and the zero bias transformation with application to simple random sampling. *The Annals of Applied Probability*, 7(4): 935–952, 1997.

Jihun Hamm, Yingjun Cao, and Mikhail Belkin. Learning privately from multiparty data. In *International Conference on Machine Learning*, pages 555–563, 2016.

Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. High dimensional differentially private stochastic optimization with heavy-tailed data. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 227–236, 2022.

Prateek Jain and Abhradeep Guha Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pages 476–484. PMLR, 2014.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

Shiva Prasad Kasiviswanathan and Hongxia Jin. Efficient private empirical risk minimization for high-dimensional learning. In *International Conference on Machine Learning*, pages 488–497, 2016.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

James K Lindsey and Bradley Jones. Choosing among generalized linear models applied to medical data. *Statistics in medicine*, 17(1):59–68, 1998.

Chong Liu, Yuqing Zhu, Kamalika Chaudhuri, and Yu-Xiang Wang. Revisiting model-agnostic private learning: Faster rates and active learning. In *International Conference on Artificial Intelligence and Statistics*, pages 838–846. PMLR, 2021.

Alexander J McNeil and Jonathan P Wendin. Bayesian inference for generalized linear mixed models of portfolio credit risk. *Journal of Empirical Finance*, 14(2):131–149, 2007.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 58–77. IEEE, 2017.

Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private glms. In *International Conference on Artificial Intelligence and Statistics*, pages 2638–2646. PMLR, 2021.

G. W. Stewart. Matrix perturbation theory, 1990.

Jinyan Su, Jinhui Xu, and Di Wang. On pac learning halfspaces in non-interactive local privacy model with public unlabeled data. *arXiv preprint arXiv:2209.08319*, 2022.

Yasuaki Takada, Ryutaro Miyagi, Aya Takahashi, Toshinori Endo, and Naoki Osada. A generalized linear model for decomposing cis-regulatory, parent-of-origin, and maternal effects on allele-specific gene expression. *G3: Genes, Genomes, Genetics*, 7(7):2227–2234, 2017.

Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and XiaoFeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *CoRR*, abs/1709.02753, 2017.

Terence Tao. Topics in random matrix theory. *Graduate Studies in Mathematics*, 132, 2011.

Jonathan Ullman. Private multiplicative weights beyond linear queries. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 303–312. ACM, 2015.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, June 9-15, 2019*, 2019.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, pages 2722–2731, 2017.

Di Wang, Marco Gaboardi, and Jinhui Xu. Empirical risk minimization in non-interactive local differential privacy revisited. In *Advances in Neural Information Processing Systems*, pages 965–974, 2018.

Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, June 9-15, 2019*, 2019a.

Di Wang, Adam Smith, and Jinhui Xu. Noninteractive locally private learning of linear models via polynomial approximations. In *Algorithmic Learning Theory*, pages 897–902, 2019b.

Di Wang, Marco Gaboardi, Adam Smith, and Jinhui Xu. Empirical risk minimization in the non-interactive local model of differential privacy. *J. Mach. Learn. Res.*, 21:200:1–200:39, 2020.

Di Wang, Huangyu Zhang, Marco Gaboardi, and Jinhui Xu. Estimating smooth glm in non-interactive local differential privacy model with public unlabeled data. In *Algorithmic Learning Theory*, pages 897–902, 2021.

Russell T. Warne. *Statistics for the Social Sciences: A General Linear Model Approach.* Cambridge University Press, 2017. doi: 10.1017/9781316442715.

Kai Zheng, Wenlong Mou, and Liwei Wang. Collect at once, use effectively: Making non-interactive locally private learning possible. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4130–4139. JMLR. org, 2017.

Appendix A. Background and Auxiliary Lemmas

Notations For a positive semi-definite matrix $M \in \mathbb{R}^{p\times p}$, we define the M-norm for a vector w as $\|w\|_M^2 = w^T M w$. $\lambda_{\min}(A)$ is the minimal singular value of the matrix A.

Lemma 26 (Non-communicative Matrix Bernstein inequality (Vershynin, 2010)) *Consider a finite sequence X_i of independent centered symmetric random $p \times p$ matrices. Assume we have for some numbers K and σ that*

$$\|X_i\|_2 \leq K, \| \sum_i \mathbb{E}[X_i^2]\|_2 \leq \sigma^2.$$

Then, for every $t \geq 0$ we have

$$Pr(\| \sum_i X_i\|_2 \geq t) \leq 2p \exp(-\frac{t^2/2}{\sigma^2 + Kt/3}).$$

Lemma 27 (Hoeffding type inequality for norm-subGaussian (Jin et al., 2019)) *If the random vectors $X_i \in \mathbb{R}^p$ satisfy*

$$Pr(\|X_i - \mathbb{E}X\|_2 \geq t) \leq \exp(-\frac{t^2}{2\sigma^2})$$

for $i = 1, 2, \cdots n$ with some σ and any $t > 0$. Then there exists an absolute constant c such that with probability at least $1-\delta$ for any $\delta > 0$:

$$\| \sum_{i=1}^n X_i\|_2 \leq c\sqrt{n\sigma^2 \log \frac{2d}{\delta}}.$$

Lemma 28 (Weyl's Inequality (Stewart, 1990)) *Let $X, Y \in \mathbb{R}^{p\times p}$ be two symmetric matrices, and $E = X - Y$. Then, for all $i = 1, \cdots, p$, we have*

$$|\sigma_i(X) - \sigma_i(Y)| \leq \|E\|_2,$$

where $\sigma_i(M)$ is the i-th eigenvalue of the matrix M.

Lemma 29 *Let $w \in \mathbb{R}^p$ be a fixed vector and E be a symmetric Gaussian random matrix where the upper triangle entries are i.i.d Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Then, with probability at least $1 - \xi$, the following holds for a fixed positive semi-definite matrix $M \in \mathbb{R}^{p\times p}$*

$$\|Ew\|_M^2 \leq \sigma^2 \, Tr(M) \|w\|^2 \log \frac{2p^2}{\xi}.$$

Proof [Proof of Lemma 29] Let $M = U^T \Sigma U$ denote the eigenvalue decomposition of M. Then, we have

$$\|Ew\|_M^2 = w^T E^T U^T \Sigma U E w = \sum_{i=1}^p \sigma_i \sum_{j=1}^p [UE]_{ij}^2 w_i^2.$$

Note that $[UE]_{i,j} = \sum_{k=1}^{p} U_{i,k}E_{j,k}$ where $E_{i,j}$ is Gaussian. Since U is orthogonal, we know that $[UE]_{i,j} \sim \mathcal{N}(0, \sigma^2)$. Using the Gaussian tail bound for all $i, j \in [d]^2$, we have

$$\mathbb{P}(\max_{i,j\in[p]^2} |[UE]_{i,j}| \geq \sqrt{\sigma^2 \log \frac{2p^2}{\xi}}) \leq \xi.$$

■

Lemma 30 (Theorem 4.7.1 in (Vershynin, 2018)) *Let x be a random vector in $\mathbb{R}^p$ that is sub-Gaussian with covariance matrix Σ and $\|\Sigma^{-\frac{1}{2}}x\|_{\psi_2} \leq \kappa_x$. Then, with probability at least $1 - \exp(-p)$, the empirical covariance matrix $\frac{1}{n}X^TX = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$ satisfies*

$$\|\frac{1}{n}X^TX - \Sigma\|_2 \leq C\kappa_x^2\sqrt{\frac{p}{n}}\|\Sigma\|_2.$$

Lemma 31 (Corollary 2.3.6 in (Tao, 2011)) *Let $M \in \mathbb{R}^{p\times p}$ be a symmetric matrix whose entries m_{ij} are independent for $j > i$, have mean zero, and are uniformly bounded in magnitude by 1. Then, there exists absolute constants $C_2, c_1 > 0$ such that with probability at least $1 - \exp(-C_2c_1p)$, the following inequality holds $\|M\|_2 \leq C\sqrt{p}$.*

Below we introduce some concentration lemmas given in (Erdogdu et al., 2019).

Lemma 32 *Let $\mathbb{B}^\delta(\tilde{w})$ denote the ball centered at $\tilde{w}$ and with radius δ (*i.e.*, $\mathbb{B}^\delta(\tilde{w}) = \{w : \|w - \tilde{w}\|_2 \leq \delta\}$). For $i = 1, 2\cdots, n$, let $x_i \in \mathbb{R}^p$ be i.i.d isotropic sub-Gaussian random vectors with $\|x_i\|_{\psi_2} \leq k_x$, and $\tilde{\mu} = \frac{\mathbb{E}[\|x\|_2]}{\sqrt{p}}$. For any given function $g : \mathbb{R} \mapsto \mathbb{R}$ that is Lipschitz continuous with G and satisfies $\sup_{w\in\mathbb{B}^\delta(\tilde{w})} \|g(\langle x, w\rangle)\|_{\psi_2} \leq \kappa_g$, with probability at least $1 - 2\exp(-p)$, the following holds for $np > 51\max\{\chi, \chi^2\}$*

$$\sup_{w\in\mathbb{B}^\delta(\tilde{w})} |\frac{1}{m}\sum_{i=1}^{m} g(\langle x_i, w\rangle) - \mathbb{E}[g(\langle x, w\rangle)]| \leq c(\kappa_g + \frac{\kappa_x}{\tilde{u}})\sqrt{\frac{p\log m}{m}},$$

where $\chi = \frac{(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2}{c\delta^2G^2\tilde{\mu}^2}$. c is some absolute constant.

Lemma 33 *Let $\mathbb{B}^\delta(\tilde{w})$ be the ball centered at $\tilde{w}$ and with radius δ (*i.e.*, $\mathbb{B}^\delta(\tilde{w}) = \{w : \|w - \tilde{w}\|_2 \leq \delta\}$). For $i = 1, 2\cdots, n$, let $x_i \in \mathbb{R}^p$ be i.i.d sub-Gaussian random vectors with covariance matrix Σ. For any given function $g : \mathbb{R} \mapsto \mathbb{R}$ that is uniformly bounded by L and Lipschitz continuous with G, the following holds with probability at least $1 - \exp(-p)$*

$$\sup_{w\in\mathbb{B}^\delta(\tilde{w})} |\frac{1}{m}\sum_{i=1}^{m} g(\langle x_i, w\rangle) - \mathbb{E}[g(\langle x, w\rangle)]| \leq 2\{G(\|\tilde{w}\|_2 + \delta)\|\Sigma\|_2 + L\}\sqrt{\frac{p}{m}}.$$

The following lemma shows that the private estimator $\hat{w}^{ols}$ is close to the unperturbed one.

Lemma 34 *Let $X = [x_1^T; x_2^T; \cdots ; x_n^T] \in \mathbb{R}^{n\times d}$ be a matrix such that X^TX is invertible, and $x_1, \cdots , x_n$ are realizations of a sub-Gaussian random variable x whose ℓ_2 norm is bounded by r. Moreover if x satisfies the condition of $\|\Sigma^{-\frac{1}{2}}x\|_{\psi_2} \leq \kappa_x = O(1)$ and $\Sigma = \mathbb{E}[xx^T]$ is the the population covariance matrix. Let $\tilde{w}^{ols} = (X^TX)^{-1}X^Ty$ denote the empirical linear regression estimator. Then, for sufficiently large $n \geq \Omega(\frac{\kappa_x^4\|\Sigma\|_2^2pr^4\log\frac{1}{\delta}}{\epsilon^2\lambda^2_{\min}(\Sigma)})$, the following holds with probability at least $1 - \exp(-\Omega(p)) - \xi$,*

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 = O\big(\frac{pr^2(1 + r^2\|\tilde{w}^{ols}\|_2^2)\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2n\lambda^2_{\min}(\Sigma)}\big), \tag{18}$$

where $\|x_i\|_2 \leq r$ is sampled from some bounded distribution.

Proof [Proof of Lemma 34] It is obvious that $\widehat{X^TX} = X^TX + E_1$, where E_1 is a symmetric Gaussian matrix with each entry sampled from $\mathcal{N}(0, \sigma_1^2)$ and $\sigma_1^2 = O(\frac{nr^4\log\frac{1}{\delta}}{\epsilon^2})$. $\widehat{X^Ty} = X^Ty + E_2$, where E_2 is a Gaussian vector sampled from $\mathcal{N}(0, \sigma_2^2I_p)$ and $\sigma_2^2 = O(\frac{nr^2\log\frac{1}{\delta}}{\epsilon^2})$.

We first show that $\widehat{X^TX}$ is invertible with high probability under our assumption.

It is sufficient to show that $X^TX + E_1 \succ \frac{X^TX}{2}$, *i.e.*, $\|E_1\|_2 \leq \frac{\lambda_{\min}(X^TX)}{2}$. By Lemma 31, we can see that with probability $1 - \exp(-\Omega(p))$,

$$\|E_1\|_2 \leq O(\frac{r^2\sqrt{pn\log\frac{1}{\delta}}}{\epsilon}).$$

Also, by Lemma 30 and Lemma 28 we know that with probability at least $1 - \exp(-\Omega(p))$,

$$\lambda_{\min}(X^TX) \geq n\lambda_{\min}(\Sigma) - O(\kappa_x^2\|\Sigma\|_2\sqrt{pn}).$$

Thus, it is sufficient to show that $n\lambda_{\min}(\Sigma) \geq O(\frac{\kappa_x^2\|\Sigma\|_2r^2\sqrt{pn\log\frac{1}{\delta}}}{\epsilon})$, which is true under the assumption of $n \geq \Omega(\frac{\kappa_x^4\|\Sigma\|_2^2pr^4\log\frac{1}{\delta}}{\epsilon^2\lambda^2_{\min}(\Sigma)})$. Thus, with probability at least $1 - \exp(-\Omega(p))$, it is invertible. In the following we will always assume that this event holds.

By direct calculation we have

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2 = -(X^TX + E_1)^{-1}E_1\tilde{w}^{ols} + (X^TX + E_1)^{-1}E_2.$$

Thus, by Cauchy-Schwartz inequality we get

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 = O\big(\|E_1\tilde{w}^{ols}\|^2_{(X^TX+E_1)^{-2}} + \|E_2\|^2_{(X^TX+E_1)^{-2}}\big).$$

Since we already assume that $X^TX + E_1 \succ \frac{X^TX}{2}$, by Lemma 29 we can obtain the following with probability at least $1 - \xi$

$$\|E_1\tilde{w}^{ols}\|^2_{(X^TX+E_1)^{-2}} \leq O\big(\frac{nr^4\log\frac{1}{\delta}}{\epsilon^2}\|\tilde{w}^{ols}\|_2^2\mathrm{Tr}((X^TX)^{-2})\log\frac{4p^2}{\xi}\big)$$

$$\|E_2\|^2_{(X^TX+E_1)^{-2}} \leq O\big(\frac{nr^2\log\frac{1}{\delta}}{\epsilon^2}\mathrm{Tr}((X^TX)^{-2})\frac{4p}{\xi}\big).$$

Thus, we have

$$\|\hat{w}^{ols}-\tilde{w}^{ols}\|_2^2 \leq C_1 n \cdot \frac{r^2(1+r^2\|\tilde{w}^{ols}\|_2^2)\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2}\text{Tr}((X^TX)^{-2}).$$

For the term of $\text{Tr}((X^TX)^{-2})$, we get

$$\text{Tr}((X^TX)^{-2}) \leq (\text{Tr}((X^TX)^{-1}))^2 \leq p\|(X^TX)^{-2}\|_2^2 = \frac{p}{\lambda_{\min}^2(X^TX)} \leq O(\frac{p}{n^2\lambda_{\min}^2(\Sigma)}),$$

where the last inequality is due to the fact that $\lambda_{\min}(X^TX) \geq n\lambda_{\min}(\Sigma) - O(\kappa_x^2\|\Sigma\|_2\sqrt{pn}) \geq \frac{1}{2}n\lambda_{\min}(\Sigma)$ (by the assumption on n). This completes the proof. ■

Let $w^{ols} = (\mathbb{E}[xx^T])^{-1}\mathbb{E}[xy]$ denote the population linear regression estimator. The following lemma bounds the estimation error between $\tilde{w}^{ols}$ and w^{ols}. The proof could be found in (Erdogdu et al., 2019) or (Dhillon et al., 2013).

Lemma 35 (Prop. 7 in (Erdogdu et al., 2019)) *Assume that* $\mathbb{E}[x_i]=0$, $\mathbb{E}[x_ix_i^T]=\Sigma$, *and* $\Sigma^{-\frac{1}{2}}x_i$ *and* y_i *are sub-Gaussian with norms* κ_x *and* γ*, respectively. If* $n \geq \Omega(\kappa_x\gamma p)$*, the following holds*

$$\|\tilde{w}^{ols}-w^{ols}\|_2 \leq O\big(\gamma\kappa_x\sqrt{\frac{p}{n\lambda_{\min}(\Sigma)}}\big),$$

with probability at least $1-3\exp(-p)$.

Appendix B. Proofs of LDP

The LDP proof of Algorithm 1 and 2 follows from the Gaussian mechanism (Lemma 4) and the post-processing property of DP.

For Algorithm 4, it is (ϵ,δ)-LDP due to the ℓ_2-norm bound on $\|x_iy_i\|_2 = \|x_i\|_2\|f(\langle x, w^*\rangle)+\sigma_i\|_2 \leq \|x_i\|_2(L\|x\|_2+|f(0)|+C)$, where the last inequality is due to the fact that f' is L-bounded and $\|w^*\|_2\leq 1$. That is, $|f(\langle x,w^*\rangle)-f(0)| \leq L|\langle x,w^*\rangle - 0| \leq L\|x\|_2\|w^*\|_2$. The proof is similar to Algorithm 3.

Appendix C. Proofs in Section 4

Since Theorem 15 is the most complicated one, we will first prove it and then prove Theorem 12. Finally we will proof Theorem 8.

C.1 Proof of Theorem 15

In the following proof we denote $\tilde{\mu} = \frac{\mathbb{E}[\|x\|_2]}{\sqrt{p}}$.

Since $r=O(1)$ (by assumption), combining this with Lemmas 34 and 35, we have that with probability at least $1-\exp(-\Omega(p))-\xi$ and under the assumption on n, there is a constant $C_3>0$ such that

$$\|\hat{w}^{ols}-w^{ols}\|_2 \leq C_3\frac{\kappa_x\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}. \tag{19}$$

Lemma 36 *Let $\Phi^{(2)}$ be a function that is Lipschitz continuous with constant G, and $f : \mathbb{R} \times \mathbb{R}^p \mapsto \mathbb{R}$ be another function such that $f(c, w) = c\mathbb{E}[\Phi^{(2)}(\langle x, w\rangle c)]$ and its empirical one is*

$$\hat{f}(c, w) = \frac{c}{m}\sum_{j=1}^{m} \Phi^{(2)}(\langle x, w\rangle c).$$

Let $\mathbb{B}^{\delta}(\bar{w}^{ols}) = \{w : \|w - \bar{w}^{ols}\|_2 \leq \delta\}$, where $\bar{w}^{ols} = \Sigma^{\frac{1}{2}} w^{ols}$. Under the assumptions in Lemma 34 and Eq. (19), if further assume that $\|\Sigma^{-\frac{1}{2}}x\|_{\psi_2} \leq \kappa_x$, $\sup_{w\in\mathbb{B}^{\delta}(\bar{w}^{ols})} \|\Phi^{(2)}(\langle x, w\rangle)\|_{\psi_2} \leq \kappa_g$, and there exist $\bar{c} > 0$ and $\tau > 0$ such that $f(\bar{c}, w^{ols}) \geq 1 + \tau$, then there is $\bar{c}_\Phi \in (0, \bar{c})$ such that $1 = f(\bar{c}_\Phi, w^{ols})$. Also, for sufficiently large n and m such that

$$m \geq \Omega\big((\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2 \max\{p \log m\tau^{-2}, \frac{1}{G^2\tilde{\mu}^2}\frac{\epsilon^2 n}{pr^4\|w^{ols}\|_2^2 \log\frac{1}{\delta}\log\frac{p^2}{\xi}\|\Sigma\|_2}\}\big), \tag{20}$$

$$n \geq \Omega(\kappa_x^4 G^2 \bar{c}^4 \|\Sigma\|_2 \frac{pr^4\|w^{ols}\|_2^2 \log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\big), \tag{21}$$

*with probability at least $1 - 2\exp(-p)$, there exists a $\hat{c}_\Phi \in [0, \bar{c}]$ such that $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$. Furthermore, if the derivative of $c \mapsto f(c, w^{ols})$ is bounded below in the absolute value (*i.e.*, does not change sign) by $M > 0$ in the interval $c \in [0, \bar{c}]$, then the following holds*

$$|\hat{c}_\Phi - \bar{c}_\Phi| \leq O\big(M^{-1}\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}} + M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\big). \tag{22}$$

Proof [**Proof of Lemma 36**] We divide the proof into three parts.

Part 1: Existence of $\bar{c}_\Phi$: From the definition, we know that $f(0, w^{ols}) = 0$ and $f(\bar{c}, w^{ols}) > 1$. Since f is continuous, we known that there exists a constant $\bar{c}_\Phi \in (0, \bar{c})$ which satisfies $f(\bar{c}_\Phi, w^{ols}) = 1$.

Part 2: Existence of $\hat{c}_\Phi$: For simplicity, we use the following notations.

$$\delta = C_3\frac{\kappa_x\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}, \delta' = \frac{\|\Sigma\|_2^{\frac{1}{2}}\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}, \tag{23}$$

where C_3 is the one in (19). Thus, $\|\Sigma^{\frac{1}{2}}\hat{w}^{ols} - \Sigma^{\frac{1}{2}}w^{ols}\|_2 \leq \delta'$.

Now consider the term of $|\hat{f}(c, \hat{w}^{ols}) - f(c, \hat{w}^{ols})|$ for $c \in [0, \bar{c}]$. We have

$$\sup_{c\in[0,\bar{c}]} |\hat{f}(c, \hat{w}^{ols}) - f(c, \hat{w}^{ols})| \leq \sup_{c\in[0,\bar{c}]}\sup_{w\in\mathbb{B}_{\Sigma}^{\delta'}(w^{ols})} |\hat{f}(c, w) - f(c, w)|, \tag{24}$$

where $\mathbb{B}_{\Sigma}^{\delta'}(w^{ols}) = \{w : \|\Sigma^{\frac{1}{2}}w - \Sigma^{\frac{1}{2}}w^{ols}\|_2 \leq \delta'\}$.

Note that for any x, we have $\langle x, w\rangle = \langle v, \Sigma^{\frac{1}{2}}w\rangle$, where $v = \Sigma^{-\frac{1}{2}}x$ follows an isotropic sub-Gaussian distribution. Also, by definition we know that $w \in \mathbb{B}_{\Sigma}^{\delta'}(w^{ols})$ is equivalent to

$\Sigma^{\frac{1}{2}}w \in \mathbb{B}^{\delta'}(\bar{w}^{ols})$. Thus, we have

$$
\begin{aligned}
&\sup_{c\in[0,\bar{c}]} \sup_{w\in\mathbb{B}_{\Sigma}^{\delta'}(w^{ols})} |\hat{f}(c,\hat{w}^{ols}) - f(c,\hat{w}^{ols})| \\
&\leq \bar{c} \sup_{c\in[0,\bar{c}]} \sup_{w\in\mathbb{B}_{\Sigma}^{\delta'}(w^{ols})} |\frac{1}{m}\sum_{j=1}^{m} \Phi^{(2)}(\langle v_i, \Sigma^{\frac{1}{2}}w\rangle c) - \mathbb{E}\Phi^{(2)}(\langle v, \Sigma^{\frac{1}{2}}w\rangle c)| \\
&= \bar{c} \sup_{c\in[0,\bar{c}]} \sup_{\Sigma^{\frac{1}{2}}w\in\mathbb{B}^{\delta'}(\bar{w}^{ols})} |\frac{1}{m}\sum_{j=1}^{m} \Phi^{(2)}(\langle v_i, \Sigma^{\frac{1}{2}}w\rangle c) - \mathbb{E}\Phi^{(2)}(\langle v, \Sigma^{\frac{1}{2}}w\rangle c)| \\
&= \bar{c} \sup_{w'\in\mathbb{B}^{\bar{c}\delta'}(\bar{w}^{ols})} |\frac{1}{m}\sum_{j=1}^{m} \Phi^{(2)}(\langle v_i, w'\rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w'\rangle)|. \qquad (25)
\end{aligned}
$$

By Lemma 32, we know that when $mp \geq 51\max\{\chi, \chi^{-1}\}$, where

$$
\chi = \frac{(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2}{c\delta'^2 G^2\tilde{\mu}^2} = \Theta\big(\frac{(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2}{G^2\tilde{\mu}^2}\frac{\epsilon^2 n\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}\|\Sigma\|_2}\big),
$$

the following holds with probability at least $1-2\exp(-p)$

$$
\sup_{w'\in\mathbb{B}^{\bar{c}\delta}(\bar{w}^{ols})} |\frac{1}{m}\sum_{j=1}^{m} \Phi^{(2)}(\langle v_i, w'\rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w'\rangle)| \leq O((\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}). \qquad (26)
$$

By the Lipschitz property of $\Phi^{(2)}$, we have that for any w_1 and w_2,

$$
\begin{aligned}
\sup_{c\in[0,\bar{c}]} |f(c,w_1) - f(c,w_2)| &\leq G\bar{c}^2\mathbb{E}[\langle v, \Sigma^{\frac{1}{2}}(w_1-w_2)\rangle] \\
&\leq \kappa_x G\bar{c}^2\|\Sigma^{\frac{1}{2}}(w_1-w_2)\|_2. \qquad (27)
\end{aligned}
$$

Taking $w_1 = \hat{w}^{ols}$ and $w_2 = w^{ols}$, we have

$$
\sup_{c\in[0,\bar{c}]} |f(c,\hat{w}^{ols}) - f(c,w^{ols})| \leq O\big(\kappa_x G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}\big).
$$

Combining this with (25), (26), (27), and taking δ as in (23), we get

$$
\sup_{c\in[0,\bar{c}]} |\hat{f}(c,\hat{w}^{ols}) - f(c,w^{ols})| \leq O\big(\bar{c}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}} + G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\kappa_x^2\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\big). \qquad (28)
$$

Let B denote the RHS of (28). If $c=\bar{c}$, we have $\hat{f}(c,\hat{w}^{ols}) \geq 1+\tau-B$. Thus, if $B\leq\tau$, there must exist a $\hat{c}_{\Phi}\in[0,\bar{c}]$ such that $\hat{f}(\hat{c}_{\Phi},\hat{w}^{ols}) = 1$.

To ensure that $B\leq\tau$ holds, it is sufficient to have

$$
O(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}) \leq \frac{\tau}{2}
$$

and

$$O(G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\kappa_x^2\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}})\leq\frac{\tau}{2}.$$

This means that

$$m\geq\Omega\big(\bar{c}^2(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})^2p\log m\tau^{-2}\big),$$
$$n\geq\Omega(\kappa_x^4G^2\bar{c}^4\|\Sigma\|_2\frac{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}),$$

which are assumed in the lemma.

Part 3: Estimation Error: So far, we know that $\hat{f}(\hat{c}_\Phi,\hat{w}^{ols})=f(\bar{c}_\Phi,w^{ols})=1$ with high probability. By (24), (25) and (26), we have

$$|1-f(\hat{c}_\Phi,\hat{w}^{ols})|=|\hat{f}(\hat{c}_\Phi,\hat{w}^{ols})-f(\hat{c}_\Phi,\hat{w}^{ols})|\leq O(\bar{c}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}).$$

By the same argument for (28), we have

$$|f(\hat{c}_\Phi,\hat{w}^{ols})-f(\hat{c}_\Phi,w^{ols})|\leq G\kappa_x\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}.$$

Thus, using Taylor expansion on $f(c,w^{ols})$ around c_Φ and by the assumption of the bounded derivative of f, we have

$$\begin{aligned}M|\hat{c}_\Phi-\bar{c}_\Phi|&\leq|f(\hat{c}_\Phi,w^{ols})-f(\bar{c}_\Phi,w^{ols})|\\&\leq|f(\hat{c}_\Phi,w^{ols})-f(\hat{c}_\Phi,\hat{w}^{ols})|+|f(\hat{c}_\Phi,\hat{w}^{ols})-1|\\&\leq O\big(\bar{c}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}+G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\big).\end{aligned}$$

■

Next, we prove our main theorem.

Proof [**Proof of Theorem 15**] By definition, we have

$$\begin{aligned}\|\hat{w}^{glm}-w^*\|_\infty&\leq\|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty+\|\bar{c}_\Phi w^{ols}-w^*\|_\infty\\&\leq\|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty+\|\bar{c}_\Phi w^{ols}-c_\Phi w^{ols}\|_\infty+\|c_\Phi w^{ols}-w^*\|_\infty.\end{aligned}\tag{29}$$

We first bound the term of $|\bar{c}_\Phi-c_\Phi|$. Since $\bar{c}_\Phi\mathbb{E}[\Phi^{(2)}(\langle x,w^{ols}\rangle\bar{c}_\Phi)]=1$ and $c_\Phi\mathbb{E}[\Phi^{(2)}(\langle x,w^*\rangle)]=1$ (by definition), we get

$$\begin{aligned}|f(\bar{c}_\Phi,w^{ols})-f(c_\Phi,w^{ols})|&=|c_\Phi\mathbb{E}[\Phi^{(2)}(\langle x,w^*\rangle)]-f(c_\Phi,w^{ols})|\\&\leq c_\Phi|\mathbb{E}[\Phi^{(2)}(\langle x,w^*\rangle)-\Phi^{(2)}(\langle x,w^{ols}\rangle c_\Phi)]\\&\leq c_\Phi G|\mathbb{E}[\langle x,(w^*-c_\Phi w^{ols})\rangle]\\&\leq c_\Phi G\|(w^*-c_\Phi w^{ols})\|_\infty\mathbb{E}\|x\|_1\\&\leq c_\Phi Gr\|c_\Phi w^{ols}-w^*\|_\infty,\end{aligned}$$

where the last inequality is due to the assumption that $\|x\|_1 \leq r$.

Thus, by the assumption of the bounded deviation of $f(c, w^{ols})$ on $[0, \max\{\bar{c}, c_\Phi\}]$, we have

$$M|\bar{c}_\Phi - c_\Phi| \leq |f(\bar{c}_\Phi, w^{ols}) - f(c_\Phi, w^{ols})| \leq c_\Phi G r \|c_\Phi w^{ols} - w^*\|_\infty.$$

By Lemma 10, we have

$$|\bar{c}_\Phi - c_\Phi| \leq 16 M^{-1} c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}. \tag{30}$$

Thus, the second term of (29) is bounded by

$$\begin{aligned}
\|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty &\leq 16 M^{-1} c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}} \|w^{ols}\|_\infty \\
&\leq 16 M^{-1} c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty^3}{\sqrt{p}} \Big(\frac{1}{c_\Phi} + 16 G r \kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}}\Big) \\
&= O\big(M^{-1} r^3 \kappa_x^6 G^3 \rho_2 \rho_\infty^2 \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}} \max\{1, c_\Phi\}\big),
\end{aligned} \tag{31}$$

where the last inequality is due to Lemma 10.

By Lemma 10, the third term of (29) is bounded by $16 c_\Phi G r \kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}$.

For the first term of (29), by (19) and Lemma 36 we have

$$\begin{aligned}
&\|\hat{c}_\Phi \hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty \leq |\hat{c}_\Phi| \cdot \|\hat{w}^{ols} - w^{ols}\|_\infty + |\hat{c}_\Phi - \bar{c}_\Phi| \cdot \|w^{ols}\|_\infty \\
&\leq O\big(\bar{c} \frac{\kappa_x \sqrt{p} r^2 \|w^{ols}\|_2 \sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \\
&+ \|w^{ols}\|_\infty \Big(M^{-1}\bar{c}\big(\kappa_g + \frac{\kappa_x}{\tilde{\mu}}\big)\sqrt{\frac{p \log m}{m}} + M^{-1} G \kappa_x^2 \bar{c}^2 \|\Sigma\|_2^{\frac{1}{2}} \frac{\sqrt{p} r^2 \|w^{ols}\|_2 \sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\Big)\big).
\end{aligned} \tag{32}$$

For the first term of (32), we have

$$\begin{aligned}
&\bar{c}\frac{\kappa_x \sqrt{p} r^2 \|w^{ols}\|_2 \sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \leq \bar{c}\frac{\kappa_x p r^2 \|w^{ols}\|_\infty \sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \\
&\leq \bar{c}\frac{\kappa_x p r^2 \|w^*\|_\infty \sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\Big(\frac{1}{c_\Phi} + 16 G r \kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}}\Big) \\
&= O\big(\bar{c}\frac{p \kappa_x^4 \sqrt{\rho_2}\rho_\infty G r^3 \|w^*\|_\infty \max\{1, \|w^*\|_\infty\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}\big).
\end{aligned} \tag{33}$$

For the second term of (32), we have

$$\begin{aligned}
&\|w^{ols}\|_\infty M^{-1}\bar{c}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\\
&\leq \bar{c}\|w^*\|_\infty(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}(\frac{1}{c_\Phi}+16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})\\
&\leq O\big(Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\bar{c}\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\max\{1,\frac{1}{c_\Phi}\}\big).
\end{aligned} \tag{34}$$

For the third term of (32), we have

$$\begin{aligned}
&\|w^{ols}\|_\infty M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}})\\
&\leq M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{pr^2\|w^*\|_\infty^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}(\frac{1}{c_\Phi}+16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})^2\\
&\leq O\big(M^{-1}G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\frac{pr^4\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\big).
\end{aligned} \tag{35}$$

Thus, the first term of (29) is bounded by (since $m\geq\Omega(n)$)

$$\begin{aligned}
&\|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty\leq O\big(\bar{c}\frac{p\kappa_x^4\sqrt{\rho_2}\rho_\infty Gr^3\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}\\
&+Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\bar{c}\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\max\{1,\frac{1}{c_\Phi}\}+\\
&M^{-1}G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\frac{pr^4\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\\
&=O\big(M^{-1}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\\
&\times\frac{pr^4\|w^*\|_\infty\max\{1,\|w^*\|_\infty^3\}\sqrt{\log m\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\big).
\end{aligned}$$

Putting all the bounds together, we have

$$\begin{aligned}
&\|\hat{w}^{glm}-w^*\|_\infty\leq\tilde{O}\big(M^{-1}G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\\
&\times\frac{pr^4\|w^*\|_\infty\max\{1,\|w^*\|_\infty^3\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\\
&+M^{-1}r^3\kappa_x^6c_\Phi G^3\rho_2\rho_\infty^2\frac{\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}}{\sqrt{p}}\max\{1,\frac{1}{c_\Phi}\}+\\
&Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\bar{c}\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\max\{1,\frac{1}{c_\Phi}\}\big).
\end{aligned} \tag{36}$$

Next, we bound the probability. We assume that Lemma 34, 35 and 36 hold with probability at least $1-\exp(-\Omega(p))-\rho$. They hold when

$$m \geq \Omega\big((\kappa_g+\frac{\kappa_x}{\tilde{\mu}})^2 \max\{p\log m\tau^{-2}, \frac{1}{G^2\tilde{\mu}^2}\frac{\epsilon^2 n}{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}\}\big), \tag{37}$$

$$n \geq \Omega(\max\{\kappa_x^4 G^2\bar{c}^4\|\Sigma\|_2\frac{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}, \frac{\kappa_x^4\|\Sigma\|_2^2 pr^4\log\frac{1}{\delta}}{\epsilon^2\lambda_{\min}^2(\Sigma)}\}\big). \tag{38}$$

Since $\|w^{ols}\|_2 \leq \sqrt{p}\|w^*\|_\infty(\frac{1}{c_\Phi}+16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})$, it suffices for n

$$n \geq \Omega\big(G^4\bar{c}^4\|\Sigma\|_2^2\frac{p^2r^6\kappa_x^{10}\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\big). \tag{39}$$

■

C.2 Proof of Theorem 12

Lemma 37 *Let $\bar{c}_\Phi, \bar{c}, \tau, f, \hat{f}$ be defined the same as in Lemma 36. If further assume that $|\Phi^{(2)}(\cdot)| \leq L$ for some constant $L>0$ and is Lipschitz continuous with constant G, then, under the assumptions in Lemma 34 and (19), with probability at least $1-4\exp(-p)$ there exists a constant $\hat{c}_\Phi \in [0,\bar{c}]$ such that $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols})=1$. Furthermore, if the derivative of $c \mapsto f(c, w^{ols})$ is bounded below in absolute value (*i.e., *does not change the sign) by $M>0$ in the interval $c\in[0,\bar{c}]$, then with probability at least $1-4\exp(-p)$, the following holds*

$$|\hat{c}_\Phi-\bar{c}_\Phi| \leq O\big(\frac{M^{-1}GL\bar{c}^2\kappa_x^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}+M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big) \tag{40}$$

for sufficiently large m, n such that

$$n \geq \Omega\big(\frac{LG^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\big) \tag{41}$$

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2p\tau^{-2}\big). \tag{42}$$

Proof [Proof of Lemma 37] The main idea of this proof is almost the same as the one for Lemma 36. The only difference is that instead of using Lemma 32 to get (26), we use here Lemma 33 to obtain the following with probability at least $1-\exp(-p)$

$$\begin{aligned}
&\sup_{w'\in\mathbb{B}^{\bar{c}\delta'}(\bar{w}^{ols})}|\frac{1}{m}\sum_{j=1}^m\Phi^{(2)}(\langle v_i,w'\rangle)-\mathbb{E}\Phi^{(2)}(\langle v,w'\rangle)| \\
&\leq O\big((G(\|\bar{w}^{ols}\|_2+\bar{c}\delta')\|I\|_2+L)\sqrt{\frac{p}{m}} \\
&\leq O\big((G\|\Sigma\|_2^{\frac{1}{2}}(\|w^{ols}\|_2+\bar{c}\frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)})+L)\sqrt{\frac{p}{m}}\big).
\end{aligned} \tag{43}$$

Thus, by (25), (27) and (43), we have

$$\sup_{c\in[0,\bar{c}]} |\hat{f}(c,\hat{w}^{ols}) - f(c,w^{ols})| \leq O\big(G\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}+ \\ \frac{G\kappa_x\bar{c}\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{p}r^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\sqrt{\frac{p}{mn}}+L\sqrt{\frac{p}{m}}\big). \quad (44)$$

Let D denote the RHS of (44), we have

$$\hat{f}(\bar{c},\hat{w}^{ols}) \geq 1+\tau-D.$$

It is sufficient to show that $\tau > D$, which holds when

$$O(G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\kappa_x^2\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}) \leq \frac{\tau}{2}$$

and

$$O(\frac{G\kappa_x\bar{c}\|\Sigma\|_2^{\frac{1}{2}}L\|w^{ols}\|_2\sqrt{p}r^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\sqrt{\frac{p}{mn}}) \leq \frac{\tau}{2}.$$

That is,

$$n \geq \Omega\big(\frac{G^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\big) \quad (45)$$

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2p\tau^{-2}\big). \quad (46)$$

Then, there exists $\hat{c}_\Phi \in [0,\bar{c}]$ such that $\hat{f}(\hat{c}_\Phi,\hat{w}^{ols}) = 1$. We can easily get

$$\begin{aligned} M|\hat{c}_\Phi - \bar{c}_\Phi| &\leq |f(\hat{c}_\Phi,w^{ols}) - f(\bar{c}_\Phi,w^{ols})| \\ &\leq O\big(\frac{G\bar{c}^2\kappa_x^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}} \\ &+\frac{G\kappa_x\bar{c}\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{p}r^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\sqrt{\frac{p}{mn}}+LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big) & (47) \\ &\leq O\big(\frac{GL\bar{c}^2\kappa_x^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}+LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big). & (48) \end{aligned}$$

■

Proof [**Proof of Theorem 12**] The proof is almost the same as the one for Theorem 15. By definition, we have

$$\begin{aligned} \|\hat{w}^{glm} - w^*\|_\infty &\leq \|\hat{c}_\Phi\hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols} - w^*\|_\infty \\ &\leq \|\hat{c}_\Phi\hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty + \|c_\Phi w^{ols} - w^*\|_\infty. & (49) \end{aligned}$$

The second term of (49) is bounded by

$$\|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty \leq O\big(M^{-1} r^2 \kappa_x^7 c_\Phi G^3 \rho_2 \rho_\infty^2 \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}} \max\{1, \frac{1}{c_\Phi}\}\big). \quad (50)$$

By Lemma 10, the third term of (49) is bounded by $16 c_\Phi G r \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}}$. The first term is bounded by

$$\begin{aligned}&\|\hat{c}_\Phi \hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty \leq \\ &O\big(\frac{M^{-1} G^3 L \bar{c}^2 \kappa_x^8 r^4 \rho_2 \rho_\infty^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \|\Sigma\|_2^{\frac{1}{2}} p \sqrt{\log \frac{1}{\delta} \log \frac{p}{\xi^2}}}{\epsilon \lambda_{\min}^{\frac{1}{2}}(\Sigma) \min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\} \sqrt{n}} \times \max\{\frac{1}{c_\Phi}, 1\}^2 \\ &\qquad + \frac{M^{-1} G^3 L \bar{c}^2 \kappa_x^6 r^2 \rho_2 \rho_\infty^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \|\Sigma\|_2^{\frac{1}{2}} p}{\sqrt{m}} \times \max\{\frac{1}{c_\Phi}, 1\}^2\big). \quad (51)\end{aligned}$$

Thus, in total we have

$$\begin{aligned}\|\hat{w}^{glm} - w^*\|_\infty \leq O\big(&\frac{M^{-1} G^3 L \bar{c}^2 \kappa_x^6 r^2 \rho_2 \rho_\infty^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \|\Sigma\|_2^{\frac{1}{2}} p}{\sqrt{m}} \times \max\{\frac{1}{c_\Phi}, 1\}^2 \\ &+ \frac{G^3 L \bar{c}^2 \kappa_x^6 r^4 \rho_2 \rho_\infty^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \|\Sigma\|_2^{\frac{1}{2}} p \sqrt{\log \frac{1}{\delta} \log \frac{p}{\xi^2}}}{\epsilon \lambda_{\min}^{\frac{1}{2}}(\Sigma) \min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\} \sqrt{n}} \max\{\frac{1}{c_\Phi}, 1\}^2 \\ &+ M^{-1} r^2 \kappa_x^7 c_\Phi G^3 \rho_2 \rho_\infty^2 \|\Sigma^{\frac{1}{2}}\|_\infty \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}} \max\{1, \frac{1}{c_\Phi}\}\big). \quad (52)\end{aligned}$$

The probability of success is at least $1 - \exp(-\Omega(p)) - \xi$. The sample complexity should satisfy

$$m \geq \Omega\big(G^2 L^2 \|\Sigma\|_2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} G^2 r^2 \kappa_x^6 \rho_2 \rho_\infty^2 p^2 \tau^{-2} \max\{1, \frac{1}{c_\Phi}\}^2\big) \quad (53)$$

$$n \geq \Omega\big(\frac{\rho_2 \rho_\infty^2 G^4 \tau^{-2} \bar{c}^4 \|\Sigma\|_2^2 \kappa_x^{10} p^2 \|w^*\|_\infty^2 r^6 \max\{1, \|w^*\|_\infty^2\} \log \frac{1}{\delta} \log \frac{p^3}{\xi}}{\epsilon^2 \lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}^2\big). \quad (54)$$

■

C.3 Proof of Theorem 8

Proof To prove the result, we first focus on the term of $\|\hat{w}^{ols} - w^{ols}\|_2$ where $w^{ols} = \Sigma^{-1}\mathbb{E}(xy)$. First, note that by Lemma 30 we have with probability at least $1 - \exp(-p)$,

$$\|\Sigma_m\|_2 \geq \|\Sigma\|_2 - O(k_x^2 \sqrt{\frac{p}{m}} \|\Sigma\|_2).$$

Thus, when $m \geq \Omega(k_x^4 p)$ we have $\frac{3\|\Sigma\|_2}{2} \geq \|\Sigma_m\|_2 \geq \frac{\|\Sigma\|_2}{2}$. In the following we will always assume the inequality holds. We denote $\hat{\Sigma} = \mathbb{E}(\bar{x}\bar{x}^T)$ where $x \sim \mathcal{N}(0, \Sigma)$ and

$\bar{x} = x \min\{1, \frac{r}{\|x\|_2}\}$. Next we show the lemma of bounding the term $\|\hat{\Sigma} - \Sigma\|_2$ and $\|\mathbb{E}(\bar{x}y) - \mathbb{E}(xy)\|_2$:

Lemma 38 *We have* $\|\hat{\Sigma} - \Sigma\|_2 \leq O(\frac{\|\Sigma\|_2^2}{n})$ *and* $\|\mathbb{E}(\bar{x}y) - \mathbb{E}(xy)\|_2 \leq O(\frac{\sqrt{p\|\Sigma \log n}}{n})$.

Proof [Proof of Lemma 38] By the definitions we have

$$\|\hat{\Sigma} - \Sigma\|_2 \leq \|\mathbb{E}[(\bar{x}\bar{x}^T - xx^T)\mathbb{I}_{\|x\|_2 \geq r}]\|_2.$$

For any unit vector $v \in \mathbb{R}^p$ we have

$$\begin{aligned} v^T\mathbb{E}[(xx^T - \bar{x}\bar{x}^T)\mathbb{I}_{\|x\|_2 \geq r}]v &= \mathbb{E}[((v^Tx)^2 - (v^T\bar{x})^2)\mathbb{I}_{\|x\|_2 \geq r}] \\ \leq \mathbb{E}[(v^Tx)^2\mathbb{I}_{\|x\|_2 \geq r}] &\leq \sqrt{\mathbb{E}[(v^Tx)^4]\Pr[\|x\|_2 \geq r]} \leq O(\frac{\|\Sigma\|_2^2}{n}), \end{aligned}$$

where the last inequality is due to that $\Pr[\|x\|_2 \geq r] \leq \Pr[\|x\|_2 \geq \sqrt{10p\|\Sigma\|_2 \log n}] \leq \frac{1}{n^2}$.

For $\|\mathbb{E}(\bar{x}y) - \mathbb{E}(xy)\|_2$ we have

$$\begin{aligned} &\|\mathbb{E}(\bar{x}y) - \mathbb{E}(xy)\|_2 = \|\mathbb{E}(\bar{x} - x)y\mathbb{I}_{\|x\|_2 \geq r}\|_2 \\ &\leq \sqrt{\mathbb{E}\|(\bar{x} - x)y\|_2^2 \Pr(\|x\|_2 \geq r)} \leq O(\frac{r + \sqrt{p}\|\Sigma\|_2}{n}). \end{aligned}$$

■

By Lemma 28 and 38 we have $\lambda_{\min}(\hat{\Sigma}) \geq \frac{\lambda_{\min}(\Sigma)}{2}$ when $n \geq \frac{\|\Sigma\|_2^2}{\lambda_{\min}(\Sigma)}$.

In the following we will bound the term $\|\hat{w}^{ols} - w^{ols}\|_2$. For simplicity we denote $\overline{XX^T} = \sum_{i=1}^n \bar{x}_i\bar{x}_i^T$ and $\overline{X^Ty} = \sum_{i=1}^n \bar{x}_iy_i$. Then we have

$$\begin{aligned} &\|\hat{w}^{ols} - w^{ols}\|_2 \leq \|\hat{w}^{ols} - (\overline{XX^T})^{-1}\overline{X^Ty}\|_2 + \|(\overline{XX^T})^{-1}\overline{X^Ty} - \Sigma^{-1}\mathbb{E}(xy)\|_2 \\ &\leq \|\hat{w}^{ols} - (\overline{XX^T})^{-1}\overline{X^Ty}\|_2 + \|(\overline{XX^T})^{-1}\overline{X^Ty} - \hat{\Sigma}^{-1}\mathbb{E}(\bar{x}y)\|_2 + \|\hat{\Sigma}^{-1}\mathbb{E}(\bar{x}y) - \Sigma^{-1}\mathbb{E}(xy)\|_2. \end{aligned} \tag{55}$$

We then bound each term in (55). We first bound the second term:

$$\begin{aligned} &\|(\frac{1}{n}\overline{XX^T})^{-1}(\frac{1}{n}\overline{X^Ty}) - \hat{\Sigma}^{-1}\mathbb{E}(\bar{x}y)\|_2 \\ &\leq \|(\frac{1}{n}\overline{XX^T})^{-1} - \hat{\Sigma}^{-1}\|_2\|\frac{1}{n}\overline{X^Ty}\|_2 + \|\hat{\Sigma}^{-1}\|_2\|\frac{1}{n}\overline{X^Ty} - \mathbb{E}(\bar{x}y)\|_2 \\ &\leq \|\hat{\Sigma}^{-1}\|_2\|(\frac{1}{n}\overline{XX^T})^{-1}\|_2\|\frac{1}{n}\overline{XX^T} - \hat{\Sigma}\|_2\|\frac{1}{n}\overline{X^Ty}\|_2 + \|\hat{\Sigma}^{-1}\|_2\|\frac{1}{n}\overline{X^Ty} - \mathbb{E}(\bar{x}y)\|_2 \end{aligned}$$

Below we consider two lemmas:

Lemma 39 *If* $n \geq \tilde{\Omega}(p\|\Sigma\|_2)$*, with probability at least* $1 - \zeta$

$$\|\frac{1}{n}\overline{XX^T} - \hat{\Sigma}\|_2 \leq O(\frac{\sqrt{p\|\Sigma\|_2 \log n \log\frac{p}{\zeta}}}{\sqrt{n}}).$$

Proof Note that $\|\bar{x}\bar{x}^T - \hat{\Sigma}\|_2 \leq \|\bar{x}\bar{x}^T\|_2 + \|\hat{\Sigma}\|_2 \leq 2r^2$. And for any unit vector $v \in \mathbb{R}^p$ we have the following if we denote $\bar{X} = \bar{x}\bar{x}^T$

$$\mathbb{E}(v^T \bar{X}^T \bar{X} v) = \mathbb{E}[\|\bar{x}\|_2^2 (v^T \bar{x})^2] \leq O(r^4).$$

Thus we have $\|\mathbb{E}[\bar{X}^T \bar{X}]\|_2 \leq O(r^2)$. Since $\|\mathbb{E}(\bar{X})^T \mathbb{E}(\bar{X})\|_2 \leq \|\mathbb{E}(\bar{X})\|_2^2 \leq r^2$, we have $\|\mathbb{E}[\bar{X} - \mathbb{E}\bar{X}]^T \mathbb{E}[\bar{X} - \mathbb{E}\bar{X}]\|_2 \leq O(r^2)$. Thus, by the Non-communicative Bernstein inequality (Lemma 26) we have for some constant $c > 0$:

$$\Pr(\|\frac{1}{n}\overline{XX^T} - \hat{\Sigma}\|_2 > t) \leq 2p \exp(-c \min(\frac{nt^2}{r^2}, \frac{nt}{r^2})).$$

Thus we have with probability at least $1 - \zeta$ and the definition of r we have,

$$\|\frac{1}{n}\overline{XX^T} - \hat{\Sigma}\|_2 \leq O(\frac{\sqrt{p\|\Sigma\|_2 \log n \log \frac{p}{\zeta}}}{\sqrt{n}}).$$

■

Since each $\|\bar{x}_i y_i - \mathbb{E}[\bar{x}_i y_i]\| \leq 2r$, by Lemma 27 we have

Lemma 40 *With probability at least* $1 - \zeta$, $\|\frac{1}{n}\overline{X^T y} - \mathbb{E}(\bar{x}y)\|_2 \leq O(\frac{r\sqrt{\log \frac{p}{\zeta}}}{\sqrt{n}})$.

Next we bound the term of $\|\hat{\Sigma}^{-1}\|_2, \|(\frac{1}{n}\overline{XX^T})^{-1}\|_2$. By Lemma 38 we can see we have $\|\hat{\Sigma}^{-1}\|_2 = \frac{1}{\lambda_{\min}(\hat{\Sigma})} \leq \frac{2}{\lambda_{\min}(\Sigma)}$. By Lemma 39 we have if $n \geq \tilde{\Omega}(\frac{p\|\Sigma\|_2}{\lambda_{\min}(\Sigma)})$ then we have $\lambda_{\min}(\frac{1}{n}\overline{XX^T}) \geq \frac{\lambda_{\min}(\hat{\Sigma})}{2} \geq \frac{\lambda_{\min}(\Sigma)}{4}$. Thus, in total we have

$$\begin{aligned}
&\|\hat{\Sigma}^{-1}\|_2 \|(\frac{1}{n}\overline{XX^T})^{-1}\|_2 \|\frac{1}{n}\overline{XX^T} - \hat{\Sigma}\|_2 \|\frac{1}{n}\overline{X^T y}\|_2 + \|\hat{\Sigma}^{-1}\|_2 \|\frac{1}{n}\overline{X^T y} - \mathbb{E}(\bar{x}y)\|_2 \\
&\leq O(\frac{p\|\Sigma\|_2 \log n \log \frac{p}{\zeta}}{\lambda_{\min}^2(\Sigma)\sqrt{n}} + \frac{\sqrt{p\|\Sigma\|_2 \log n \log \frac{p}{\zeta}}}{\lambda_{\min}(\Sigma)\sqrt{n}}).
\end{aligned}$$

Next we consider the third term of (55)

$$\begin{aligned}
&\|\hat{\Sigma}^{-1}\mathbb{E}(\bar{x}y) - \Sigma^{-1}\mathbb{E}(xy)\|_2 \\
&\leq \|\hat{\Sigma} - \Sigma\|_2 \|\hat{\Sigma}^{-1}\|_2 \|\Sigma^{-1}\|_2 \|\mathbb{E}(\bar{x}y)\|_2 + \|\Sigma^{-1}\|_2 \|\mathbb{E}(\bar{x}y) - \mathbb{E}(xy)\|_2 \\
&\leq O(\frac{\sqrt{p\|\Sigma\|_2 \log n}\|\Sigma\|_2^2}{\lambda_{\min}^2(\Sigma)n} + \frac{\sqrt{p\|\Sigma\|_2 \log n}}{\lambda_{\min}(\Sigma)n}).
\end{aligned}$$

For the first term of (55), by using a similar proof as in Lemma 34 we have if $n \geq \Omega(\frac{\|\Sigma\|_2^2 pr^4 \log \frac{1}{\delta}}{\epsilon^2 \lambda_{\min}^2(\Sigma)})$ then with probability at least $1 - \exp(-\Omega(p)) - \xi$ (if we denote $\bar{w}^{ols} = (\overline{XX^T})^{-1}\overline{X^T y}$)

$$\|\hat{w}^{ols} - \bar{w}^{ols}\|_2^2 = O(\frac{pr^2(1 + r^2\|\bar{w}^{ols}\|_2^2) \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\epsilon^2 n \lambda_{\min}^2(\Sigma)}). \tag{56}$$

By the previous proof we can see that

$$\begin{aligned}\|\bar{w}^{ols} - w^{ols}\|_2 &\leq O\left(\frac{p\|\Sigma\|_2 \log n \log \frac{p}{\zeta}}{\lambda^2_{\min}(\Sigma)\sqrt{n}} + \frac{\sqrt{p\|\Sigma\|_2 \log n \log \frac{p}{\zeta}}}{\lambda_{\min}(\Sigma)\sqrt{n}} + \frac{\sqrt{p\|\Sigma\|_2 \log n}\|\Sigma\|_2^2}{\lambda^2_{\min}(\Sigma) n} + \frac{\sqrt{p\|\Sigma\|_2 \log n}}{\lambda_{\min}(\Sigma) n}\right) \\ &= O\big(\frac{p\|\Sigma\|_2 \log n \log \frac{p}{\zeta}}{\sqrt{n}\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\big).\end{aligned}$$

In total when $n \geq \tilde{\Omega}(p^2\|\Sigma\|_2/\lambda^4_{\min}(\Sigma))$ we have

$$\|\hat{w}^{ols} - \bar{w}^{ols}\|_2^2 = O\big(\frac{pr^2(1 + r^2\|w^{ols}\|_2^2) \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\epsilon^2 n \lambda^2_{\min}(\Sigma)}\big). \tag{57}$$

Thus, combine all the previous results we have with probability at least $1 - \exp(-\Omega(p)) - \xi$ we have

$$\|\hat{w}^{ols} - w^{ols}\|_2 \leq O\big(\frac{\sqrt{p}r^2\|w^{ols}\|_2 \log \sqrt{\frac{1}{\delta}\log \frac{p}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}(\Sigma)} + \frac{p\|\Sigma\|_2 \log n \log \frac{p}{\zeta}}{\sqrt{n}\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\big)$$

Thus, there is a constant $C_3 > 0$ such that

$$\|\hat{w}^{ols} - w^{ols}\|_2 \leq C_3 \frac{\sqrt{p^3}\|\Sigma\|_2\|w^{ols}\|_2 \log n \sqrt{\log \frac{1}{\delta} \log \frac{p}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}. \tag{58}$$

The same Lemma 37, we have the following lemma.

Lemma 41 *Let $\bar{c}_\Phi, \bar{c}, \tau, f, \hat{f}$ be defined the same as in Lemma 36. If further assume that $|\Phi^{(2)}(\cdot)| \leq L$ for some constant $L > 0$ and is Lipschitz continuous with constant G, then, under the assumptions in Lemma 34 and (19), with probability at least $1 - 4\exp(-p)$ there exists a constant $\hat{c}_\Phi \in [0, \bar{c}]$ such that $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$. Furthermore, if the derivative of $c \mapsto f(c, w^{ols})$ is bounded below in absolute value (*i.e., *does not change the sign) by $M > 0$ in the interval $c \in [0, \bar{c}]$, then with probability at least $1 - 4\exp(-p)$, the following holds (note that for the Gaussian case $c_\Phi = \bar{c}_\Phi$)*

$$|\hat{c}_\Phi - c_\Phi| \leq O\big(\frac{M^{-1}GL\bar{c}^2\|\Sigma\|_2^{\frac{3}{2}}p^{\frac{3}{2}}\|w^{ols}\|_2 \log n \sqrt{\log \frac{1}{\delta}\log \frac{p}{\xi}}}{\epsilon\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}\sqrt{n}} + M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big) \tag{59}$$

for sufficiently large m, n such that

$$\begin{aligned} n &\geq \Omega\big(\frac{LG^2\tau^{-2}\bar{c}^4\|\Sigma\|_2^3 p^3\|w^{ols}\|_2^2 \log^2 n \log \frac{1}{\delta}\log \frac{p}{\xi}}{\epsilon^2\lambda^2_{\min}(\Sigma)\min\{\lambda^2_{\min}(\Sigma), 1\}}\big) \qquad (60) \\ m &\geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2 p\tau^{-2}\big). \qquad (61)\end{aligned}$$

Next we bound $\|\hat{w}^{glm} - w^*\|_2 = \|\hat{c}_\Phi\hat{w}^{ols} - c_\Phi w^{ols}\|_2$. We have

$$\|\hat{c}_\Phi\hat{w}^{ols} - c_\Phi w^{ols}\|_2 \leq |\hat{c}_\Phi - c_\Phi|\|\hat{w}^{ols}\|_2 + c_\Phi\|\hat{w}^{ols} - w^{ols}\|_2. \tag{62}$$

For the second term of (62), by (58) we have

$$c_\Phi\|\hat{w}^{ols}-w^{ols}\|_2 \leq O(\frac{\bar{c}p^{\frac{3}{2}}\|\Sigma\|_2\|w^{ols}\|_2\log n\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}).$$

For the first term of (62), by Lemma 41 and (58) we have

$$|\hat{c}_\Phi-c_\Phi|\|\hat{w}^{ols}\|_2 \leq O(\frac{M^{-1}GL\bar{c}^2\|\Sigma\|_2^{\frac{3}{2}}p^{\frac{3}{2}}\|w^{ols}\|_2^2\log n\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi}}}{\epsilon\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}\sqrt{n}}+M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2^2\sqrt{\frac{p}{m}})$$

Take $w^{ols}=\frac{w^*}{c_\Phi}$ we can get the proof. ■

C.4 Proof of Theorem 16

Proof We can see that

$$\Phi^{(2)}(z)=\frac{e^z}{(1+e^z)^2},\Phi^{(3)}(z)=\frac{e^z-e^{2z}}{(1+e^z)^3},\Phi^{(4)}(z)=\frac{e^z(1-4e^z+e^{2z})}{(1+e^z)^4}$$

We can see $|\Phi^{(2)}(\cdot)|\leq 1$ and $\Phi^{(2)}(\cdot)$ is 1-Lipschtitz, and $\Phi^{(2)}$ and $\Phi^{(4)}$ are even functions. Using the local convexity for $z\geq 0$ around $z=2.5$ we have

$$\Phi^{(2)}(z)\geq a-bz,$$

where $a=\Phi^{(2)}(2.5)-2.5\Phi^{(3)}(2.5)\approx 0.22$ and $b=-\Phi^{(3)}(2.5)\approx 0.06$. Denote $W\sim\mathcal{N}(0,1)$, ϕ as the density function of W and ζ as the cumulative distribution function of W, we have

$$\begin{aligned}f(z)=z\mathbb{E}[\Phi^{(2)}(\langle x,w^{ols}\rangle z)]&=z\mathbb{E}[\Phi^{(2)}(\frac{Wz}{20})]\\&=2z\int_0^\infty\Phi^{(2)}(\frac{wz}{20})\phi(w)dw\geq 2z\int_0^{\frac{20a}{bz}}(a-b\frac{wz}{20})\phi(w)dw\\&=2z(a\zeta(\frac{20a}{bz})-\frac{a}{2}-\frac{bz}{20\sqrt{2\pi}}(1-e^{\frac{-200a^2}{b^2z^2}})).\end{aligned}$$

Thus take $\bar{c}=6$ we have $f(\bar{c})>1+0.22$.

Next we will show $c_\Phi\leq\bar{c}$. Recall that $c_\Phi=\frac{1}{\mathbb{E}[\Phi^{(2)}(\langle x_i,w^*\rangle)]}$, thus we need to proof

$$\mathbb{E}[\Phi^{(2)}(\langle x_i,w^*\rangle)]>\frac{1}{6}.$$

This is because

$$\begin{aligned}\mathbb{E}[\Phi^{(2)}(\langle x,w^*\rangle)]&=\mathbb{E}[\Phi^{(2)}(\frac{W}{4})]\\&=2\int_0^\infty\Phi^{(2)}(\frac{w}{4})\phi(w)dw\geq 2\int_0^{\frac{4a}{b}}(a-b\frac{w}{4})\phi(w)dw\\&=2(a\zeta(\frac{4a}{b})-\frac{a}{2}-\frac{b}{4\sqrt{2\pi}}(1-e^{\frac{-8a^2}{b^2}}))>\frac{1}{6}.\end{aligned}$$

Finally, we will show that $f'(z)$ is bounded by constant $M = 0.19$ on $[0, \bar{c}]$ from below. Since x follows the Gaussian distribution, by Stein's lemma (Definition 21) we have

$$f'(z) = \mathbb{E}[\Phi^{(2)}(\frac{Wz}{20})] + \frac{z^2}{20^2}\mathbb{E}[\Phi^{(4)}(\frac{Wz}{20})].$$

Thus

$$\begin{aligned} f'(z) &\geq \mathbb{E}[\Phi^{(2)}(\frac{Wz}{20})] - \frac{9}{100}|\Phi^{(4)}| \\ &\geq 2(a\zeta(\frac{20a}{bz}) - \frac{a}{2} - \frac{bz}{20\sqrt{2\pi}}(1 - e^{\frac{-200a^2}{b^2z^2}})) - \frac{9}{800} > 0.1 \end{aligned}$$

■

C.5 Proof of Theorem 17

Proof By simple calculation we can see that

$$\Phi^{(2)}(z) = \frac{1}{4}(1+\frac{z^2}{4})^{-\frac{3}{2}}, \Phi^{(3)}(z) = -\frac{3}{16}z(1+\frac{z^2}{4})^{-\frac{5}{2}}, \Phi^{(4)} = \frac{3}{64}\frac{5z^2(1+\frac{z^2}{4})^{-2}-4}{(1+\frac{z^2}{4})^{\frac{5}{4}}},$$

we can see that $|\Phi^{(2)}(\cdot)| \leq \frac{1}{4}$, $|\Phi^{(2)}(\cdot)|$ is $\frac{3}{16}$-Lipschitz and these two functions are even. Using the local convexity for $z \geq 0$ around $z = 2$ we have

$$\Phi^{(2)}(z) \geq a - bz,$$

where $a = \Phi^{(2)}(2) - 2\Phi^{(3)}(2) \approx 0.22$ and $b = -\Phi^{(3)}(2) \approx 0.066$. Denote $W \sim \mathcal{N}(0,1)$, ϕ as the density function of W and ζ as the cumulative distribution function of W, we have

$$\begin{aligned} f(z) &= z\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols}\rangle z)] = z\mathbb{E}[\Phi^{(2)}(\frac{Wz}{20})] \\ &= 2z\int_0^\infty \Phi^{(2)}(\frac{wz}{20})\phi(w)dw \geq 2z\int_0^{\frac{20a}{bz}}(a - b\frac{wz}{20})\phi(w)dw \\ &= 2z(a\zeta(\frac{20a}{bz}) - \frac{a}{2} - \frac{bz}{20\sqrt{2\pi}}(1 - e^{\frac{-200a^2}{b^2z^2}})). \end{aligned}$$

Thus take $\bar{c} = 6$ we have $f(\bar{c}) > 1 + 0.22$.

Next we will show $c_\Phi \leq \bar{c}$. Recall that $c_\Phi = \frac{1}{\mathbb{E}[\Phi^{(2)}(\langle x_i, w^*\rangle)]}$, thus we need to proof

$$\mathbb{E}[\Phi^{(2)}(\langle x_i, w^*\rangle)] > \frac{1}{6}.$$

This is because

$$\begin{aligned} \mathbb{E}[\Phi^{(2)}(\langle x, w^*\rangle)] &= \mathbb{E}[\Phi^{(2)}(\frac{W}{4})] \\ &= 2\int_0^\infty \Phi^{(2)}(\frac{w}{4})\phi(w)dw \geq 2\int_0^{\frac{4a}{b}}(a - b\frac{w}{4})\phi(w)dw \\ &= 2(a\zeta(\frac{4a}{b}) - \frac{a}{2} - \frac{b}{4\sqrt{2\pi}}(1 - e^{\frac{-8a^2}{b^2}})) > \frac{1}{6}. \end{aligned}$$

Finally, we will show that $f'(z)$ is bounded by constant $M = 0.1$ on $[0, \bar{c}]$ from below. Since x follows the Gaussian distribution, by Stein's lemma we have

$$f'(z) = \mathbb{E}[\Phi^{(2)}(\frac{Wz}{20})] + \frac{z^2}{20^2}\mathbb{E}[\Phi^{(4)}(\frac{Wz}{20})].$$

Thus

$$\begin{aligned} f'(z) &\geq \mathbb{E}[\Phi^{(2)}(\frac{Wz}{20})] - \frac{9}{100}|\Phi^{(4)}| \\ &\geq 2(a\zeta(\frac{20a}{bz}) - \frac{a}{2} - \frac{bz}{20\sqrt{2\pi}}(1 - e^{\frac{-200a^2}{b^2z^2}})) - \frac{27}{1600} > 0.1 \end{aligned}$$

■

Appendix D. Proofs in Section 5

D.1 Proof of Theorem 18

Proof [**Proof of Theorem 18**] Denote $\phi(\cdot, \Sigma)$ as the multivariate normal density with mean 0 and covariance matrix Σ, by simple calculation we have $\frac{d\phi(x,\Sigma)}{dx} = -\Sigma^{-1}x\phi(x, \Sigma)$. By the setting of (11) we have.

$$\begin{aligned} \mathbb{E}[xy] = \mathbb{E}[xf(\langle x, w^*\rangle)] &= \int xf(\langle x, w^*\rangle)\phi(x, \Sigma)dx \\ &= -\Sigma \int f(\langle x, w^*\rangle)\frac{d\phi(x, \Sigma)}{dx}dx \\ &= \Sigma w^* \mathbb{E}[f'(\langle x, w^*\rangle)], \end{aligned}$$

where the last equation is deduced by integration by part. Thus

$$w^* = \frac{1}{\mathbb{E}[f'(\langle x, w^*\rangle)]}w^{ols}.$$

■

D.2 Proof of Theorem 22

The idea of the proof follows the one in (Erdogdu et al., 2019).
Proof [**Proof of Theorem 22**] By assumption, we have

$$\mathbb{E}[xy] = \mathbb{E}[xf(\langle x, w^*\rangle)] = \Sigma^{\frac{1}{2}}\mathbb{E}[vf(\langle v, \hat{w}^*\rangle)],$$

where $\hat{w}^* = \Sigma^{\frac{1}{2}}w^*$. Now, consider each coordinate $j \in [p]$ for the term $\mathbb{E}[vf(\langle v, \hat{w}^*\rangle)]$. Let v_j^* denote the zero-bias transformation of v_j conditioned on $V_j = \langle v, \hat{w}^*\rangle - v_j\hat{w}_j^*$. Then, we have

$$\begin{aligned}\mathbb{E}[v_j f(\langle v, \hat{w}^*\rangle)] &= \mathbb{E}\mathbb{E}[v_j f(v_j \hat{w}_j^* + V_j)|V_j] \\ &= \hat{w}_j^* \mathbb{E}\mathbb{E}[f'(v_j^* \hat{w}_j^* + V_j)|V_j] \\ &= \hat{w}_j^* \mathbb{E}\mathbb{E}[f'((v_j^* - v_j)\hat{w}_j^* + \langle v, \hat{w}^*\rangle)|V_j] \\ &= \hat{w}_j^* \mathbb{E}[f'((v_j^* - v_j)\hat{w}_j^* + \langle v, \hat{w}^*\rangle)].\end{aligned}$$

Thus, we have $w^{ols} = \Sigma^{-\frac{1}{2}} D \Sigma^{\frac{1}{2}} w^*$, where D is a diagonal matrix whose i-th entry is $\mathbb{E}[f'((v_j^* - v_j)\hat{w}_j^* + \langle v, \hat{w}^*\rangle)]$.

By the Lipschitz condition, we have

$$|\mathbb{E}[f'((v_j^* - v_j)\hat{w}_j^* + \langle v, \hat{w}^*\rangle)] - \mathbb{E}[f'(\langle v, \hat{w}^*\rangle)]| \leq G|\hat{w}_j^*|\mathbb{E}|(v_j^* - v_j)|.$$

By the same argument given in (Erdogdu et al., 2019), we have

$$\mathbb{E}|(v_j^* - v_j)| \leq 1.5\mathbb{E}[|v_j|^3].$$

Using the bound of the third moment induced by the sub-Gaussian norm, we have

$$L|\hat{w}_j^*|\mathbb{E}|(v_j^* - v_j)| \leq 8G\kappa_x^3 \max_{j\in[p]} |\hat{w}_j^*| \leq 8G\kappa_x^3 \|\Sigma^{\frac{1}{2}} w^*\|_\infty.$$

Thus, we get

$$\max_{j\in[d]} |D_{jj} - \frac{1}{c_f}| \leq 8G\kappa_x^3 \|\Sigma^{\frac{1}{2}} w^*\|_\infty.$$

This means that

$$\begin{aligned}\|w^{ols} - \frac{1}{c_f} w^*\|_\infty &= \|\Sigma^{-\frac{1}{2}}(D - \frac{1}{c_f} I)\Sigma^{\frac{1}{2}} w^*\|_\infty \\ &\leq \max_{j\in[p]} |D_{jj} - \frac{1}{c_f}| \|\Sigma^{-\frac{1}{2}}\|_\infty \|\Sigma^{\frac{1}{2}}\|_\infty \|w^*\|_\infty \\ &\leq 8L\kappa_x^3 \rho_\infty L \|\Sigma^{\frac{1}{2}}\|_\infty \|w^*\|_\infty^2.\end{aligned}$$

Due to the diagonal dominance property we have

$$\|\Sigma^{\frac{1}{2}}\|_\infty = \max_i \sum_{j=1}^{p} |\Sigma_{ij}^{\frac{1}{2}}| \leq 2\max \Sigma_{ii}^{\frac{1}{2}} \leq 2\|\Sigma\|_2^{\frac{1}{2}}.$$

Since we have $\|x\|_2 \leq r$, we write

$$r^2 \geq \mathbb{E}[\|x\|_2^2] = \text{Trace}(\Sigma) \geq p\|\Sigma \geq \frac{p\|\Sigma\|_2}{\rho_2}.$$

Thus we have $\|\Sigma^{\frac{1}{2}}\|_\infty \leq 2r\sqrt{\frac{\rho_2}{p}}$. ■

D.3 Proof of Theorem 23

By the same argument in the proof of Theorem 8, we can show that if $n \geq \Omega\big(\frac{\|\Sigma\|_2^2 p r^4 \log\frac{1}{\delta}}{\epsilon^2 \lambda_{\min}^2(\Sigma)\min\{\lambda_{\min}^2(\Sigma),1\}}\big)$ then with probability at least $1-\exp(-\Omega(p))-\xi$

$$\|\hat{w}^{ols}-\tilde{w}^{ols}\|_2^2 = O\big(\frac{pC^2r^2(L^2r^2+C^2+r^2\|\tilde{w}^{ols}\|_2^2)\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2 n\lambda_{\min}^2(\Sigma)}\big). \tag{63}$$

Thus, by Lemma 35 we have

$$\|\hat{w}^{ols}-w^{ols}\|_2 \leq O\big(\frac{CL\kappa_x\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\big). \tag{64}$$

In the following, we will always assume that (64) holds. By the same argument given in Lemma 37, we have the following Lemma, which can be proved in the same way as Lemma 37.

Lemma 42 *Let f' be a function that is Lipschitz continuous with constant G and $|f'(\cdot)| \leq L$, and $g: \mathbb{R}\times\mathbb{R}^p \mapsto \mathbb{R}$ be another function such that $g(c,w) = c\mathbb{E}[f'(\langle x,w\rangle c)]$ and its empirical one is*

$$\hat{g}(c,w) = \frac{c}{m}\sum_{j=1}^{m} f'(\langle x,w\rangle c).$$

*Let $\mathbb{B}^{\delta}(\bar{w}^{ols}) = \{w : \|w-\bar{w}^{ols}\|_2 \leq \delta\}$, where $\bar{w}^{ols} = \Sigma^{\frac{1}{2}}w^{ols}$. Then, under the assumptions in Lemma 34 and Eq. (64), with probability at least $1-4\exp(-p)$, there exists a constant $\hat{c}_f \in [0,\bar{c}]$ such that $\hat{g}(\hat{c}_f,\hat{w}^{ols}) = 1$. Furthermore, if the derivative of $c \mapsto g(c,w^{ols})$ is bounded below in absolute value (*i.e., *does not change the sign) by $M>0$ in the interval of $c\in[0,\bar{c}]$, then with probability at least $1-4\exp(-p)$, the following holds*

$$|\hat{c}_f-\bar{c}_f| \leq O\big(\frac{M^{-1}CGL\bar{c}^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\log\frac{1}{\delta}\log\frac{p}{\xi^2}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}+M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big) \tag{65}$$

for sufficiently large m,n such that

$$n \geq \Omega\big(\frac{LG^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\big) \tag{66}$$

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2p\tau^{-2}\big). \tag{67}$$

where $r = \max_{i\in[n]}\|x_i\|_2$.

D.4 Proof of Theorem 19

The proof is almost the same as the proof of Theorem 8. We know that when $n \geq \Omega\big(\frac{\|\Sigma\|_2^2 p r^4 \log\frac{1}{\delta}}{\epsilon^2 \lambda_{\min}^2(\Sigma)\min\{\lambda_{\min}^2(\Sigma),1\}}\big)$, with probability at least $1-\exp(-\Omega(p))-\xi$, there is a constant $C_3>0$ such that

$$\|\hat{w}^{ols}-w^{ols}\|_2 \leq C_3\frac{\sqrt{p^3}\|\Sigma\|_2\|w^{ols}\|_2\log n\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}. \tag{68}$$

Similar to Lemma 41, we have the following lemma.

Lemma 43 *Let f' be a function that is Lipschitz continuous with constant G and $|f'(\cdot)| \leq L$, and $g : \mathbb{R} \times \mathbb{R}^p \mapsto \mathbb{R}$ be another function such that $g(c, w) = c\mathbb{E}[f'(\langle x, w\rangle c)]$ and its empirical one is*

$$\hat{g}(c, w) = \frac{c}{m}\sum_{j=1}^{m} f'(\langle x, w\rangle c).$$

*Let $\mathbb{B}^{\delta}(\bar{w}^{ols}) = \{w : \|w - \bar{w}^{ols}\|_2 \leq \delta\}$, where $\bar{w}^{ols} = \Sigma^{\frac{1}{2}} w^{ols}$. Then, under Eq. (68), with probability at least $1 - 4\exp(-p)$, there exists a constant $\hat{c}_f \in [0, \bar{c}]$ such that $\hat{g}(\hat{c}_f, \hat{w}^{ols}) = 1$. Furthermore, if the derivative of $c \mapsto g(c, w^{ols})$ is bounded below in absolute value (*i.e., *does not change the sign) by $M > 0$ in the interval of $c \in [0, \bar{c}]$, then with probability at least $1 - 4\exp(-p)$, the following holds*

$$|\hat{c}_f - c_f| \leq O\big(\frac{M^{-1}GL\bar{c}^2\|\Sigma\|_2^{\frac{3}{2}} p^{\frac{3}{2}}\|w^{ols}\|_2 \log n\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi}}}{\epsilon\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}\sqrt{n}} + M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big) \tag{69}$$

for sufficiently large m, n such that

$$n \geq \Omega\big(\frac{LG^2\tau^{-2}\bar{c}^4\|\Sigma\|_2^3 p^3\|w^{ols}\|_2^2 \log n \log\frac{1}{\delta}\log\frac{p}{\xi}}{\epsilon^2\lambda_{\min}^2(\Sigma)\min\{\lambda_{\min}^2(\Sigma), 1\}}\big) \tag{70}$$

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2 p\tau^{-2}\big). \tag{71}$$

Next we bound $\|\hat{w}^{nlr} - w^*\|_2 = \|\hat{c}_f\hat{w}^{ols} - c_f w^{ols}\|_2$. We have

$$\|\hat{c}_f\hat{w}^{ols} - c_f w^{ols}\|_2 \leq |\hat{c}_f - c_f|\|\hat{w}^{ols}\|_2 + c_f\|\hat{w}^{ols} - w^{ols}\|_2. \tag{72}$$

For the second term of (72), by (68) we have

$$c_f\|\hat{w}^{ols} - w^{ols}\|_2 \leq O\big(\frac{\bar{c}p^{\frac{3}{2}}\|\Sigma\|_2\|w^{ols}\|_2 \log n\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\big).$$

For the first term of (72), by Lemma 43 and (68) we have

$$|\hat{c}_f - c_f|\|\hat{w}^{ols}\|_2 \leq O\big(\frac{M^{-1}GL\bar{c}^2\|\Sigma\|_2^{\frac{3}{2}} p^{\frac{3}{2}}\|w^{ols}\|_2^2 \log n\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi}}}{\epsilon\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}\sqrt{n}} + M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2^2\sqrt{\frac{p}{m}}\big)$$

Take $w^{ols} = \frac{w^*}{c_f}$ we can get the proof.

D.5 Proof of Theorem 25

Proof We can easily see that $f'(\cdot)$ is just the function $\Phi^{(2)}(\cdot)$ in Theorem 16 for the logistic loss function. Thus the function f' satisfies the assumptions in Theorem 23, which was showed in the Theorem 16. ■

Appendix E. A 2-Round LDP Algorithm for Algorithm 1

Algorithm 5 2-round LDP for smooth GLMs with public data (Gaussian)

```
Input: Private data $\{(x_i, y_i)\}_{i=1}^n \in (\mathbb{R}^p \times [0,1])^n$, where $|y_i| \leq 1$, $\{x_i\}_{j=1}^n \sim \mathcal{N}(0, \Sigma)$ for some unknown $\Sigma$, loss function $\Phi : \mathbb{R} \mapsto \mathbb{R}$, privacy parameters $\epsilon, \delta$, $\|\Sigma\|_2$, and initial value $c \in \mathbb{R}$.
In the first round:
for Each user $i \in [n]$ do
    Let $\bar{x}_i = x_i \min\{1, \frac{r}{\|x_i\|_2}\}$, where $r \equiv \sqrt{10p\|\Sigma\|_2 \log n}$.
    Release $\widehat{x_i x_i^T} = \bar{x}_i \bar{x}_i^T + E_{1,i}$ and $\widehat{x_i y_i} = \bar{x}_i y_i + E_{2,i}$, where $E_{1,i} \in \mathbb{R}^{p \times p}$ is a symmetric matrix and each entry of the upper triangle matrix is sampled from $\mathcal{N}(0, \frac{128r^4 \log \frac{2.5}{\delta}}{\epsilon^2})$ and $E_{2,i} \in \mathbb{R}^p$ is sampled from $\mathcal{N}(0, \frac{128r^2 \log \frac{2.5}{\delta}}{\epsilon^2} I_p)$.
end for
for The server do
    Let $\widehat{X^T X} = \sum_{i=1}^n \widehat{x_i x_i^T}$ and $\widehat{X^T y} = \sum_{i=1}^n \widehat{x_i y_i}$. Calculate $\hat{w}^{ols} = (\widehat{X^T X})^{-1} \widehat{X^T y}$.
    Send $\hat{w}^{ols}$ to all users.
end for
In the second round:
for Each user $i \in [n]$ do
    Calculate $\bar{y}_i = \langle x_i, \hat{w}^{ols} \rangle$.
    Project $\bar{y}_i$ onto the interval $[0,1]$ and denote it as $\hat{y}_i$.
    Send $\tilde{y}_i = \hat{y}_i + \mathcal{N}(0, \frac{8 \log \frac{2.5}{\delta}}{\epsilon^2})$ to the server.
end for
for The server do
    Find the root $\hat{c}_\Phi$ such that $1 = \frac{\hat{c}_\Phi}{n} \sum_{j=1}^n \Phi^{(2)}(\hat{c}_\Phi \tilde{y}_j)$ by using Newton's root-finding method (or other methods):
    for $t = 1, 2, \cdots$ until convergence do
        $c = c - \frac{c \frac{1}{n} \sum_{j=1}^n \Phi^{(2)}(c\tilde{y}_j) - 1}{\frac{1}{n} \sum_{j=1}^n \{\Phi^{(2)}(c\tilde{y}_j) + c\tilde{y}_j \Phi^{(3)}(c\tilde{y}_j)\}}$.
    end for
end for
return $\hat{w}^{glm} = \hat{c}_\Phi \cdot \hat{w}^{ols}$.
```