Estimating smooth GLM in non-interactive local differential privacy model with pub...

Proceedings of Machine Learning Research vol 132:1–43, 2021

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
School of Electrical and Computer Engineering
Cornell University
Ithaca, NY

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

Jinhui Xu JINHUI@BUFFALO.EDU
Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY

Abstract

In this paper, we study the problem of estimating smooth Generalized Linear Models (GLM) in the Non-interactive Local Differential Privacy (NLDP) model. Different from its classical setting, our model allows the server to access some additional public but unlabeled data. Firstly, motived by Stein's lemma, we show that if each data record is i.i.d. sampled from zero-mean Gaussian distribution, we show that there exists an (ϵ, δ)-NLDP algorithm for GLM. The sample complexity of the public and private data, for the algorithm to achieve an α estimation error (in ℓ_2-norm) with high probability, is $O(p\alpha^{-2})$ and $O(p^3\alpha^{-2}\epsilon^{-2})$, respectively. This is a significant improvement over the previously known exponential or quasi-polynomial in α^{-1}, or exponential in p sample complexity of GLM with no public data. Then, by a variant of Stein's lemma, we show that there is an (ϵ, δ)-NLDP algorithm for GLM (under some mild assumptions), if each data record is i.i.d sampled from some sub-Gaussian distribution with bounded ℓ_1-norm. Then the sample complexity of the public and private data, for the algorithm to achieve an α estimation error (in ℓ_∞-norm) with high probability, is $O(p^2\alpha^{-2})$ and $O(p^2\alpha^{-2}\epsilon^{-2})$, respectively, if α is not too small (*i.e.,* $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$), where p is the dimensionality of the data. We also extend our idea to the non-linear regression problem and show a similar phenomenon for it. 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 GLM and non-linear regression in the NLDP model with public unlabeled data.

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

[*] The first three authors contributed equally to this paper.

[†] Part work of this paper was presented at The 32nd International Conference on Algorithmic Learning Theory (ALT 2021) (Wang et al., 2021).

1. Introduction

Generalized Linear Model (GLM) is one of the most fundamental models in statistics and machine learning. It generalizes 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. 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. 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 estimator (MLE) 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, maximizing 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 GLM in real world applications is how to handle sensitive data, such as those in social science and medical research. As a commonly-accepted approach for preserving privacy, Differential Privacy (DP) (Dwork et al., 2006) provides provable protection against identification and is resilient to arbitrary auxiliary information that might be available to attackers.

As a popular way of achieving DP, Local Differential Privacy (LDP) has received considerable attention in recent years and 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 the (now private) data of each individual and combines them into a resulting data analysis. Information exchange between the server and each individual could be either only once or multiple times. Correspondingly, protocols for LDP are called non-interactive LDP (NLDP) or interactive LDP. Due to its ease of implementation (*e.g.* no need to deal with the network latency problem), NLDP is often preferred in practice.

While there are many results on GLM in the DP and interactive LDP models (Chaudhuri et al., 2011; Bassily et al., 2014; Jain and Thakurta, 2014; Kasiviswanathan and Jin, 2016), GLM in NLDP is still not well understood due to the limitation of the model. (Smith et al., 2017; Wang et al., 2018; Zheng et al., 2017) and (Wang et al., 2019b) comprehensively studied 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 dimensionality 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.

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 (for private data) for achieving error α under different measurements, where c_0, C is a constant and D_p is a function of p. For bounded norm case we assume that $\|x_i\| \leq 1$ for every $i \in [n]$, for Gaussian case we assume $x_i \sim \mathcal{N}(0,\Sigma)$ with some unknown Σ.

To address this high sample complexity issue of NLDP, a possible way is to make use of some recent developments on 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) have suggested that by allowing the server to access some public but unlabeled data in addition to the private data, it is possible to reduce the sample complexity in the central DP model, under the assumption that these public data have the same marginal distribution as the private ones. It has also shown that such a relaxed setting is likely to enable better practical performance for problems like Empirical Risk Minimization (ERM) (Hamm et al., 2016; Papernot et al., 2016). Thus, it would be interesting to know whether the relaxed setting can also help reduce sample complexity in the NLDP model.

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

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, motived by Stein's lemma, we show that when the feature vector x is some (unknown) Gaussian distribution with zero mean, *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 GLM, the sample complexity of the public and private data for the algorithm to achieve an α estimation error (in ℓ_2-norm) with high probability, is $O(p\alpha^{-2})$ and $O(p^3\alpha^{-2}\epsilon^{-2})$ (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. Then we show that when the feature vector x of GLM is sub-Gaussian with bounded ℓ_1-norm, there is an (ϵ, δ)-NLDP algorithm for GLM (under some mild assumptions) whose sample complexities of the private and public data, for achieving an error of α (in ℓ_∞-norm), are $O(p^2\epsilon^{-2}\alpha^{-2})$ and $O(p^2\alpha^{-2})$ (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 non-linear regression problem. By using Stein's lemma and the zero-bias transformation (Goldstein et al., 1997), we show that when x is either Gaussian or sub-Gaussian with bounded ℓ_1-norm, it exhibits the same phenomenon as GLM.

4. Finally, we provide extensive experimental study of our algorithms on both synthetic and real world datasets. The experimental results suggest that our methods are efficient and effective, which is 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). These results differ from ours in quite a few ways. Firstly, all of them consider either the multiparty setting or the centralized 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 and each party uses their data to get a classifier, however, this approach could not be extended to local DP model since now each party only has just one data point and it is impossible to get any useful classifier based one data point. (Papernot et al., 2016, 2018) investigated the DP model, used sub-sample and aggregate to train some deep learning models, but provided no provable sample complexity. (Bassily et al., 2018) also studied the DP model by combining the distance to instability and the sparse vector techniques, and showed some theoretical guarantees. However, both the sub-sample/aggregate and the sparse vector methods cannot be used in the NLDP model. Moreover, public data in their methods are also used quite differently from ours. Secondly, all of the above results use the private data to label the public data and conduct the learning process on the public data, while we use the public data to approximate some crucial constants. Finally, all of the previous methods rely on the known model or loss functions, while in our algorithms the loss functions could be unknown to the users; also the server could use multiple loss functions with the same sample complexity.

The problems considered in this paper can be viewed as restricted versions of the general 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 dimensionality p is not avoidable for general loss functions. 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, but not the other exponential terms. Recently, (Wang et al., 2019b, 2020) further showed that the sample complexity

for any 1-Lipschitz convex GLM can be reduced to linear in p and exponential in α^{-1}, which extends the work in (Zheng et al., 2017), whose sample complexity is linear in p and quasi-polynomial in α^{-1} for smooth GLM. In this paper, we show, for the first time, that the sample complexity of GLM can be reduced to fully polynomial with the help of some public but unlabeled data and some mild assumptions on GLM. There are also works on some special loss functions. For example, (Wang and Xu, 2019) studied the high dimensional sparse linear regression problem and (Daniely and Feldman, 2018) considered the problem of 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.

3. Preliminaries

Since in this paper we mainly focus on sub-Gaussian distribution, we first recall its definition, 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 assumption 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*

- *For the matrix Σ, its corresponding $\Sigma^{\frac{1}{2}}$ is diagonally dominant.* [1]
- *Its distribution is supported on a ℓ_1-norm ball of radius r.*
- *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 $\tilde{w} = \Sigma^{\frac{1}{2}}w^*$, $\mathbb{E}[v_{ij}|\sum_{k\neq j}\tilde{w}v_{ik}]$ and $\mathbb{E}[v_{ij}^2|\sum_{k\neq j}\tilde{w}v_{ik}]$ are deterministic).*

Differential Privacy (DP): In DP, we have data universe $\mathcal{X}$ and $\mathcal{Y}$, 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, which is 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.

1. 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.

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, 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 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 Gaussian mechanism (Dwork et al., 2006) 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 $\{(x_i, y_i)\}_{i=1}^n$ 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.,* they have the same distribution as $\{x_i\}_{i=1}^n$).

4. Privately Estimating Generalized Linear Models

In this section, we study GLM in our model and privately estimate w^* in (1) by using both of the private data $\{(x_i, y_i)\}_{i=1}^n$ and the 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 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.* [2]

From Lemma 5, we can see that to obtain w^*, it is sufficient to estimate w^{ols} and the underlying 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 sufficient statistics, *i.e.,* $x_i x_i^T$ and $y_i x_i$. After receiving the private OLS estimator $\hat{w}^{ols}$,[3] 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 Algorithm 2 we use the Newton's method which has a quadratic convergence rate.

However, there is a difficulty of this approach. That is, Lemma 5 needs x to be 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 solve this issue, we will use the concentration bound for Gaussian distribution, this can help us filter some 'outliers' and keep other records bounded. Specifically, we will use 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 $1 - \zeta$, $\|x\|_2 \leq \sqrt{5p\|\Sigma\|_2 \log\frac{1}{\zeta}}$ for a fixed x. That is, with probability at least $1 - \zeta$, we have $\|x_i\|_2 \leq \sqrt{5p\|\Sigma\|_2 \log\frac{n}{\zeta}}$. Thus to make the sensitivity of the term $\|x_i\|_2$ bounded we can check whether each $\|x_i\|_2$ is upper bounded by $\sqrt{5p\|\Sigma\|_2 \log\frac{n}{\zeta}}$, if this is true, then we just use the Gaussian mechanism, otherwise we will filter it. However, we can see this upper bound 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.

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

The following theorem shows the 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.*
- *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 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 $M > 0$), where c_Φ is in Lemma 5.*

2. $\Phi^{(2)}$ is the second order derivative of function Φ, the similar to $\Phi^{(3)}$ and $\Phi^{(1)}$.
3. Note that when n is large enough we can show $\hat{w}^{ols}$ is well defined, see Appendix for details.

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

Input: Private data $\{(x_i, y_i)\}_{i=1}^n \in (\mathbb{R}^p \times [0,1])^n$, where $|y_i| \le 1$, $\{x_i\}_{j=1}^{n+m} \sim \mathcal{N}(0, \Sigma)$ for some unknown Σ and $\{x_j\}_{j=n+1}^{n+m}$ are public. loss function $\Phi : \mathbb{R} \mapsto \mathbb{R}$, privacy parameters ϵ, δ, and initial value $c \in \mathbb{R}$, failure probability ζ.

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 *Each user* $i \in [n]$ **do**

Check whether $\|x_i\|_2 \le \sqrt{5\|\Sigma_m\|_2 p \log\frac{n}{\zeta}}$; If not, release $\perp$;

Otherwise denote $r = \sqrt{5\|\Sigma_m\|_2 p \log\frac{n}{\zeta}}$, release $\widehat{x_i x_i^T} = x_i x_i^T + E_{1,i}$ and $\widehat{x_i y_i} = x_i y_i + E_{2,i}$, 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 *The server* **do**

Let $\widehat{X^TX} = \sum \widehat{x_i x_i^T}$ and $\widehat{X^Ty} = \sum \widehat{x_i y_i}$. Calculate $\hat{w}^{ols} = (\widehat{X^TX})^{-1}\widehat{X^Ty}$.

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**

$c = c - \frac{c\frac{1}{m}\sum_{j=n+1}^{n+m+1}\Phi^{(2)}(c\tilde{y}_j) - 1}{\frac{1}{m}\sum_{j=n+1}^{n+m+1}\{\Phi^{(2)}(c\tilde{y}_j) + c\tilde{y}_j\Phi^{(3)}(c\tilde{y}_j)\}}$.

end

end

return $\hat{w}^{glm} = \hat{c}_\Phi \cdot \hat{w}^{ols}$.

Note that the first condition means $\Phi^{(1)}$ is Lipschitz. The second and the last condition are to ensure that the function $f - 1$ has a root and $\hat{c}_\Phi$ is close to c_Φ for large enough m, see Remark 18 for more comments and 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*

$$n \ge \Omega\Big(\frac{\tau^{-2}\bar{c}^4\|\Sigma\|_2^3 p^3 \|w^*\|_2^2 \log^2\frac{n}{\xi}\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{c_\Phi^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\Big) \tag{3}$$

$$m \ge \Omega\big(\|\Sigma\|_2\|w^*\|_2^2 p\tau^{-2}c_\Phi^{-2}\big), \tag{4}$$

with probability at least $1 - \exp(-\Omega(p)) - \xi$

$$\|\hat{w}^{glm} - w^*\|_2 \le O\Big(\frac{\|\Sigma\|_2^{\frac{3}{2}} p^{\frac{3}{2}}\|w^*\|_2^2 \log\frac{n}{\xi}\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{c_\Phi^2\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\}\sqrt{n}} + c_\Phi^{-2}\|\Sigma\|_2^{\frac{1}{2}}\|w^*\|_2^2\sqrt{\frac{p}{m}}\Big)$$

where $G, L, M, \bar{c}$ are assumed to be $O(1)$ and thus omitted in the Big-O and Big-Ω notations (see Appendix for the explicit form of m and n).

Corollary 9 *Theorem 8 suggests that if we omit all the other terms and assume that $\|w^*\|_2 = O(1)$, then for any given error α, there is an (ϵ, δ)-LDP algorithm whose sample complexity of private (n) and public unlabeled (m) data, to achieve an estimation error of α (in ℓ_2-norm), is $O(p^3\epsilon^{-2}\alpha^{-2})$ 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, δ. All these are quite reasonable in practice.*

Remark 10 *It is notable that public unlabeled data is only used in several steps in Algorithm 1 (and all other algorithms in the paper), where we use it to find a root of some function. It is still an open problem whether we use can use the same idea to the original LDP model (*i.e., *there is no public unlabeled data). One possible way is to adopt our idea to a 2-round LDP algorithm. That is, in the first round we get $\hat{w}^{ols}$ by using half of the data and the server send it to all the users. In the second round, each user j in the left group computes a noisy version of $\tilde{y}_j = x_j^T \hat{w}_{ols}$ and sends it to the server, which then uses it to estimate the constant of c_Φ. We note that due to the noise we added in the second round for each term of $\tilde{y}_j$, there could be a large amount of error for the term of $\hat{c}_\Phi$ to estimate c_Φ, which may cause the private estimator to have large error.*

Remark 11 *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 covariance matrix $\hat{X^T X}$ does not depend on labels. Thus, we can use those public unlabeled data to give a more precise estimator of the covariance matrix, that is we can let $\widehat{X^T X} = \frac{1}{m+n}(\sum_{i=1}^{n} \widehat{x_i x_i^T} + \sum_{j=n+1}^{n+m} x_j x_j^T)$ and $\widehat{X^T y} = \frac{1}{n}\sum_{i=1}^{n} \widehat{x_i y_i}$. However, we note that here the upper bounds of error will be asymptotically the same as the bound in Theorem 8 (and all other theorems in the paper), sowe will omit the details of this improved approach. In the experiments part we will use this improved method.*

4.2. Sub-Gaussian case

We note that Lemma 5 is only for Gaussian distribution. The following lemma extends Lemma 5 to bounded sub-Gaussian with an additional additive error of $O(\frac{\|w^*\|_\infty^2}{\sqrt{p}})$. We fist give the assumptions for the data distribution.

Lemma 12 ((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 and 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 and $w^{ols} = (\mathbb{E}[xx^T])^{-1}\mathbb{E}[xy]$ is the OLS vector.

Lemma 12 indicates that we can use the same idea as above to estimate w^*. Note that the forms of c_Φ in Lemmas 5 and 12 are different. However, due to the closeness of w^* and w^{ols} in (5), 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 $c\mathbb{E}[\Phi^{(2)}(\langle x_i, w^{ols}\rangle c)] - 1$

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

Input: Private data $\{(x_i, y_i)\}_{i=1}^n \subset (\mathbb{R}^p \times [0,1])^n$, where $\|x_i\|_1 \le r$ and $|y_i| \le 1$, public unlabeled data $\{x_j\}_{j=n+1}^{n+m}$, loss function $\Phi : \mathbb{R} \mapsto \mathbb{R}$, privacy parameters ϵ, δ, 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 *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**

$c = c - \frac{c\frac{1}{m}\sum_{j=n+1}^{n+m+1} \Phi^{(2)}(c\tilde{y}_j) - 1}{\frac{1}{m}\sum_{j=n+1}^{n+m+1}\{\Phi^{(2)}(c\tilde{y}_j) + c\tilde{y}_j \Phi^{(3)}(c\tilde{y}_j)\}}$.

end

end

return $\hat{w}^{glm} = \hat{c}_\Phi \cdot \hat{w}^{ols}$.

and it could be approximated by using the public unlabeled data. Combining these ideas, we have Algorithm 2.

Theorem 13 *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 case. Just as Assumption 2, we need the following assumptions for loss function.

Assumption 3 *We assume*

- $|\Phi^{(2)}(\cdot)| \le L$ *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* $f(\bar{c}) \ge 1 + \tau$, *where* w^{ols} *is in Lemma 12 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 12.*

It seem like Assumption 3 is similar to Assumption 2. However, since these two assumption rely on the underlying distribution of (x, y), which are different in these two assumptions. Thus, these two assumptions are quite different. Moreover, the third conditions in Assumption 3 and Assumption 2 are different due to the different interval and the different form of c_Φ.

Theorem 14 *Under Assumption 1 and 3, for sufficiently large* m, n *such that*

$$m \ge \Omega\big(\|\Sigma\|_2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}\rho_2 \rho_\infty^2 p^2\big),$$
$$n \ge \Omega\Big(\frac{\rho_2\rho_\infty^2 \|\Sigma\|_2^2 p^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \log\frac{1}{\delta}\log\frac{p}{\xi}}{\epsilon^2 \lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\Big), \tag{6}$$

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

$$\|\hat{w}^{glm}-w^*\|_\infty \le O\Big(\frac{\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p}{\sqrt{m}}$$
$$+\frac{\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}+\frac{\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_\infty\|w^*\|_\infty^3\max\{1,\|w^*\|_\infty\}}{\sqrt{p}}\Big), \tag{7}$$

where $G, L, \tau, M, \bar{c}, r, \kappa_x, \frac{1}{c_\Phi}$ *are assumed to be* $O(1)$ *and thus omitted in the Big-O notations (see Appendix for the explicit form of* m *and* n*).*

Corollary 15 *Similar to Corollary 9, Theorem 14 suggests that if we omit all the other terms and assume that* $\|w^*\|_\infty = O(1)$*, then for any given error* $\alpha \ge \Omega(\frac{1}{\sqrt{p}})$*, there is an* (ϵ,δ)*-LDP algorithm whose sample complexity of private* (n) *and public unlabeled* (m) *data, to achieve an estimation error of* α *(in* ℓ_∞*-norm), is* $O(p^2\epsilon^{-2}\alpha^{-2})$ *and* $O(p^2\alpha^{-2})$*, respectively. We will also see that in practice we do not need large amount of public data (see Section 6.1 in Appendix for details).*

Compared with the complexity in the Gaussian case, it seems like 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 assumption (*$\|w^*\|_2 = O(1)$ *v.s.* $\|w^*\|_\infty = O(1)$*), these two results are incomparable.*

There are also some previous work on LDP linear regression. It seems that our sample complexities for the more general GLM is worse than theirs. However, these results are not really comparable due to their different settings. Specifically, when $\|x_i\|_2 \le 1$ and $\|w^*\|_2 \le 1$ (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, for the Gaussian case we assume $\|x_i\|_2 \le O(\sqrt{p})$ and for the sub-Gaussian case we assume $\|w^*\|_\infty \le O(1)$. (Zheng et al., 2017) proposed an algorithm whose sample complexity is $O(\alpha^{-4}\epsilon^{-2}\log p)$ for the optimization error, under the assumptions of $\|x_i\|_1 \le 1$ and $\|w^*\|_1 \le 1$, which are different with ours in the paper. Recently, (Wang and Xu, 2019) also considered the ℓ_2-norm statistical error, it relies on assumptions that w^* is 1-sparse, which is not needed in ours. However, we also have to say that, in this paper, we need some additional assumptions (i.e., Assumption 1) on the data distribution compared with the those previous results.

Remark 16 *Algorithm 1 and 2 have several advantages over existing techniques. Firstly, different from the approach of using Gradient Descent methods to solve DP-ERM* (e.g., *(Wang et al., 2017)), our algorithm is parameter-free. That is, we do not need to choose a specific step size, an iteration number or initial vectors. Secondly, comparing with some previous work such as (Zheng et al., 2017; Smith et al., 2017; Wang et al., 2019b), all of our above results do not need to assume that the loss function is convex. Thirdly, since the private data contributes only to obtaining the OLS estimator, and only the constant* $\hat{c}_\Phi$ *depends on the loss function* Φ*, this means that with probability at least* $1-T\exp(-\Omega(p))-\xi$*, our algorithm can simultaneously use* T *different loss functions to achieve the same errors and with the same sample complexity. 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 14 (Theorem 8). To our best knowledge, this is the*

first result which 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 dimensionality p increases, we could answer more GLMs queries. It sounds unintuitive that with more dimensions, one can handle more losses. However, we note that here we also need more data samples to achieve a fixed error of α.

Note that in Theorem 14, $\Phi^{(2)}$ is assumed to be bounded. This is a quite common assumption in related works such as (Wang et al., 2018, 2019a). Actually, this condition can be relaxed by only assuming 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 12 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 12.*

Theorem 17 *Under Assumption 1 and 4, for sufficiently large m,n such that*

$$m\geq\tilde{\Omega}\big(\frac{1}{\tilde{\mu}^2}\epsilon^2 n\big), n\geq\Omega\big(\frac{p^2\rho_2\rho_\infty^2\|\Sigma\|_2^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\big), \tag{8}$$

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

$$\|\hat{w}^{glm}-w^*\|_\infty\leq O\big(\frac{p\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\|w^*\|_\infty\max\{1,\|w^*\|_\infty^3\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}+$$
$$\frac{\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}}{\sqrt{p}}+\sqrt{\rho_2}\rho_\infty\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}\frac{1}{\tilde{\mu}}\sqrt{\frac{p^2\log m}{m}}\big), \tag{9}$$

where $\tilde{\mu}=\frac{\mathbb{E}[\|x\|_2]}{\sqrt{p}}$, the terms of $r,\kappa_x,\kappa_g,G,M,\tau,\bar{c},\frac{1}{c_\Phi}$ are assumed to be constants, and thus omitted in the Big-O notations (see Appendix for the explicit forms of m and n).

From the above theorem, we can see that with more relaxed assumptions, the sample complexity in Theorem 17 increases by a factor of $O(\log m)$ to achieve an upper bound on the statistical error (in ℓ_∞-norm) that is asymptotically the same as the one in Theorem 14.

Remark 18 *A not so desirable issue of Theorems 8, 14 and 17 is that they need quite a few assumptions/conditions. Although almost all of them commonly appear in some related work, the assumptions on function f seem to be a little weird. Fortunately, this is a not big issue in both practice and theory. In the following, motivated by (Erdogdu et al., 2019), we will provide two*

examples which satisfy Assumption 2. Moreover, as we will see later, our experiments show that the algorithm actually 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 14 and 17 are dependent on the ℓ_1-norm of the upper bound of x_i, while such a dependency is on the ℓ_2-norm in previous work such as (Smith et al., 2017; Zheng et al., 2017). We leave the problem of relaxing/lifting these assumptions to future research.

Theorem 19 (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 20 (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 regression in the NLDP model with public unlabeled data. We assume that there is an underlying vector $w^* \in \mathbb{R}^p$ with $\|w^*\|_2 \leq 1$ such that

$$y = f(\langle x, w^* \rangle) + \sigma, \tag{10}$$

where x is the feature vector sampled from some distribution (for simplicity, we assume that the mean is zero) and y is the response. σ is the zero-mean noise which is independent of x and bounded by some constant $C = O(1)$ (*i.e.,* $\sigma \in [-C, C]$). f is some known differentiable link function with $f(0) \neq \infty$ [4]. We note that these assumptions are quite common in related work such as (Wang and Xu, 2019; Duchi and Ruan, 2018). In our model, the goal is to obtain some estimator w^{priv} of w^*, based on the private dataset $\{(x_i, y_i)\}_{i=1}^n$ and the public unlabeled dataset $\{x_j\}_{j=n+1}^{n+m+1}$ via some NLDP algorithms.

5.1. Gaussian Case

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

Theorem 21 *If $x \sim \mathcal{N}(0, \Sigma)$, then w^* in (10) 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.*

We observe that the result in Theorem 21 is similar as Theorem Lemma 5. Thus, similar to Algorithm 1 we have Algorithm 3 when x is Gaussian.

Just as in the previous section, we need the following assumption for link function.

4. This assumption can be relaxed to "there is a point x such that $f(x) \neq 0$".

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 Σ and $\{x_j\}_{j=n+1}^{n+m}$ are public. Link function $f : \mathbb{R} \mapsto \mathbb{R}$, privacy parameters ϵ, δ, 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 *Each user* $i \in [n]$ **do**

Check whether $\|x_i\|_2 \le \sqrt{5\|\Sigma_m\|_2 p \log\frac{n}{\zeta}}$; If not, release $\perp$;

Denote $\sqrt{5\|\Sigma_m\|_2 p \log\frac{n}{\zeta}}$. 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 *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**

$c = c - \frac{c\frac{1}{m}\sum_{j=n+1}^{n+m+1} f'(c\tilde{y}_j) - 1}{\frac{1}{m}\sum_{j=n+1}^{n+m+1}\{f'(c\tilde{y}_j) + c\tilde{y}_j f^{(2)}(c\tilde{y}_j)\}}$.

end

end

return $\hat{w}^{nlr} = \hat{c}_f \cdot \hat{w}^{ols}$.

Assumption 5 *We assume*

- $|f'(\cdot)| \le L$ *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* $\ell(\bar{c}) \ge 1 + \tau$*, where* w^{ols} *is in Theorem 21.*
- *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* $M > 0$*), where* c_f *is in Theorem 21.*

Theorem 22 *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*

$$n \ge \Omega\Big(\frac{\tau^{-2}\bar{c}^4 \|\Sigma\|_2^3 p^3 \|w^*\|_2^2 \log^2\frac{n}{\xi}\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{c_f^2 \epsilon^2 \lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\Big) \tag{11}$$

$$m \ge \Omega\big(\|\Sigma\|_2 \|w^*\|_2^2 p \tau^{-2} c_f^{-2}\big), \tag{12}$$

with probability at least $1 - \exp(-\Omega(p)) - \xi$

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

where $G, L, M, \bar{c}$ *are assumed to be* $O(1)$ *and thus omitted in the Big-O and Big-*Ω *notations (see Appendix for the explicit form of* m *and* n*).*

5.2. Sub-Gaussian Case

To solve this problem, we first use the zero-bias transformation (Goldstein et al., 1997) and the techniques in (Erdogdu et al., 2019) to get a lemma similar to Lemma 12.

Definition 23 (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. When* z *is Gaussian, then* $z^* = z$*, this is just Stein's Lemma.*

Theorem 24 *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.*

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

From Theorem 24, we can see that it shares the same phenomenon as Lemma 12 (*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 for the bounded sub-Gaussian case, which gives us Algorithm 4 and the following theorem. The same as in the previous section, we need the following assumptions for link function.

Assumption 6 *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 24.*
- *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 24.*

Theorem 25 *For any* $0 < \epsilon, \delta < 1$*, Algorithm 4 is* (ϵ, δ) *non-interactive LDP. Under the assumptions of Theorem 24, and if the link function* f *satisfies Assumption 6, then for sufficiently large* m, n *such that*

$$m \geq \Omega\big(\|\Sigma\|_2\|w^*\|_\infty^2\max\{1, \|w^*\|_\infty^2\}\rho_2\rho_\infty^2 p^2\big), \tag{13}$$

$$n \geq \Omega\Big(\frac{\rho_2\rho_\infty^2\|\Sigma\|_2^2 p^2\|w^*\|_\infty^2\max\{1, \|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)}\Big), \tag{14}$$

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 ϵ, δ, 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 *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**

$c = c - \frac{c\frac{1}{m}\sum_{j=n+1}^{n+m+1} f'(c\tilde{y}_j) - 1}{\frac{1}{m}\sum_{j=n+1}^{n+m+1}\{f'(c\tilde{y}_j) + c\tilde{y}_j f^{(2)}(c\tilde{y}_j)\}}$.

end

end

return $\hat{w}^{nlr} = \hat{c}_f \cdot \hat{w}^{ols}$.

with probability at least $1 - \exp(-\Omega(p)) - \xi$*, the output of Algorithm 4 satisfies*

$$\|\hat{w}^{nlr} - w^*\|_\infty \leq O\Big(\frac{\rho_2 \rho_\infty^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}} p}{\sqrt{m}} + \frac{\rho_2 \rho_\infty^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}} p\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon \lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\}\sqrt{n}} + \frac{\rho_2 \rho_\infty^2 \|\Sigma^{\frac{1}{2}}\|_\infty \|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}}\Big), \tag{15}$$

where the terms of $G, L, \tau, M, \bar{c}, r, \kappa_x, C, \frac{1}{c_f}$ *are assumed to be* $O(1)$ *and thus omitted in the Big-O notations (see Appendix for the explicit form of* m *and* n*).*

Just as in the Theorem 19 and 20, in the following we will provide an example that satisfies the assumptions in Theorem 25.

Theorem 26 (Sigmoid Link Function) *Consider the model (10) 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 evaluate the performance of our methods on both synthetic and real world datasets. The experiments demonstrate the convergence results of our algorithms and suggest that they are practically efficient.

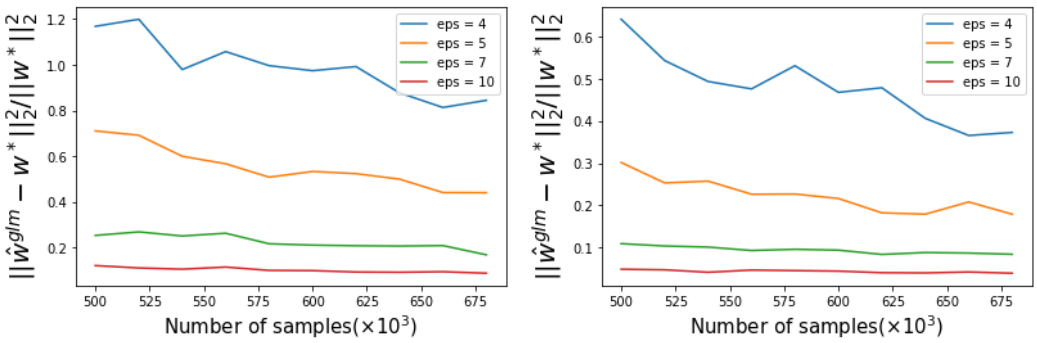
Link Functions: In this paper, we mainly study estimating GLMs and non-linear regression. For GLM, we consider the problem of binary logistic regression *i.e.,* $\Phi(\langle x, w\rangle) = \ln(1 + \exp(\langle x, w\rangle))$ and Poisson regression *i.e.,* $\Phi(\langle x, w\rangle) = e^{\langle x,w\rangle}$ in (1). For non-linear regression, we consider the case where the link function is either cubic *i.e.,* $f(x) = \frac{1}{3}x^3$ or logistic loss *i.e.,* $f(x) = \log(1+e^{-x})$ in (10).

Synthetic Data Generation: In this paper, we assume the data feature distribution is either Gaussian or sub-Gaussian with bounded ℓ_1-norm. When the distribution is Gaussian, we also consider two cases where the covariance matrix is either diagonal or not. For the case where the covariance matrix is diagonal, we sample each diagonal entry from the uniform distribution of $[0, 1]$. For general covariance case, we will randomly generate an orthogonal matrix. In the sub-Gaussian case, the features are generated independently from a Bernoulli distribution $\Pr\left(x_{i,j} = \pm\frac{1}{p}\right) = 0.5$. For GLM, the label is generated according to its definition in (1). In non-linear regression model, the label is generate according to (10) where σ is bounded by $C = 0.001$.

Experimental Setting for Synthetic Data: For data with Gaussian features, we will use squared ℓ_2-norm relative error $\frac{\|\hat{w}-w^*\|_2^2}{\|w^*\|_2^2}$ to measure performance, otherwise we will use squared ℓ_∞-norm relative error $\frac{\|\hat{w}-w^*\|_\infty^2}{\|w^*\|_\infty^2}$. We will first study the relative error with respect to different privacy parameters $\epsilon \in \{10, 5, 3, 2\}$ with $\delta = \frac{1}{n^{1.1}}$. In these experiments, we estimate the relative error with the fixed dimensionality $p = 10$ and the population parameter $w^* = (1, 1, ..., 1)/\sqrt{p}$. The sample size n is chosen from the set $10^4 \cdot \{1, 3, 5, ..., 29\}$. We assume that the same amount of public unlabeled data is available. For each problem we then evaluate the impact of the dimensionality. In these experiments, we fix the privacy parameters $\epsilon = 10$, $\delta = \frac{1}{n^{1.1}}$, and tune the dimensionality $p \in \{5, 10, 12, 15\}$.[5] w^*s are the same as above. The sample size takes values from $n \in 10^4 \cdot \{10, 12, 14, ..., 48\}$ and the same amount of public unlabeled data is assumed. . For each experiments above, we run 100 times and take the average of the errors.

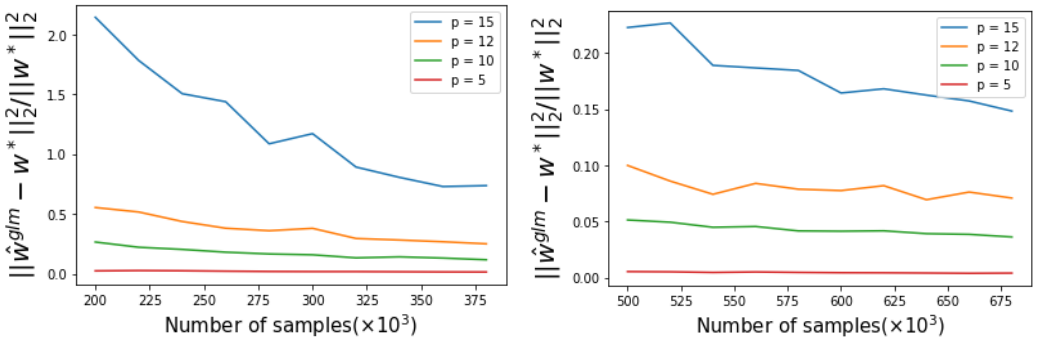
Experimental Setting for Real-world Data: We conduct experiment for GLM with logistic loss on the Covertype dataset (Dua and Graff, 2017). Before running our algorithm, we first normalize the data and remove some co-related features. After the pre-processing, the dataset contains 581012 samples and 44 features. There are seven possible values for the label. Since multinomial logistic regression can not be regarded as a Generalized Linear Model, we consider a weaker test, which is to classify whether the label is Lodgepole Pine (type 2) or not. The chosen algorithm is still binary logistic regression. We divide the data into training and testing, where $n_{\text{training}} = 406708$ and $n_{\text{testing}} = 174304$ and randomly choose the sample size $n \in 10^4 \cdot \{1, 2, 3, ..., 39\}$ from the training data and use the same amount of public data. Regarding the privacy parameter, we take $\delta = \frac{1}{n^{1.1}}$ and let ϵ take value from $\{20, 10, 5\}$. We measure the performance by the prediction accuracy. For each combination of ϵ and n, the experiment is repeated 1000 times.

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



(*a*) (*b*)

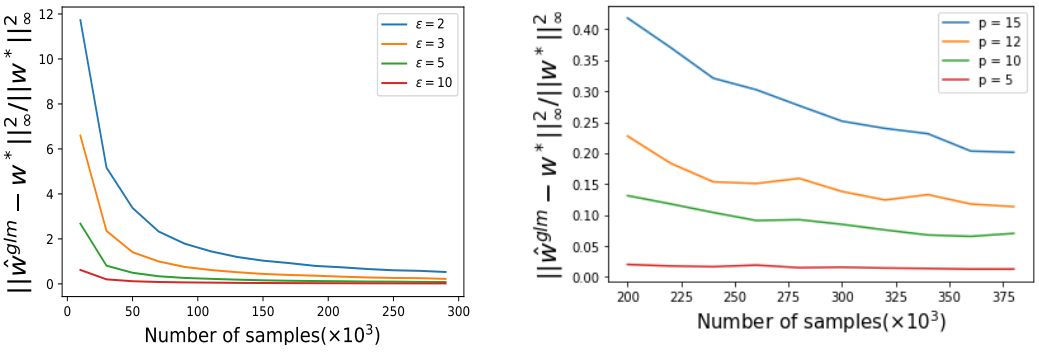
Figure 1: GLM under Gaussian design with different levels of privacy. The left plot show the squared ℓ_2-norm relative error of logistic regression where the covariance matrix is diagonal. The right plot show the squared ℓ_2-norm relative error of Poisson regression, where the covariance matrix is a random orthogonal matrix.



(*a*) (*b*)

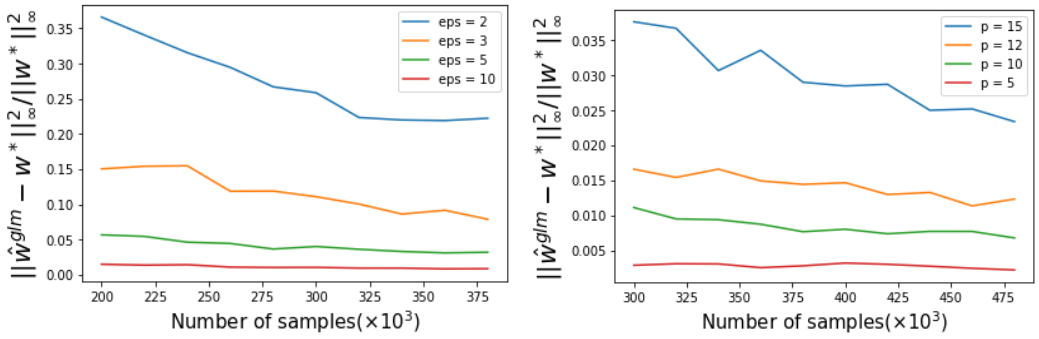
Figure 2: GLM under Gaussian design with different dimensionality p. The left plot show the squared ℓ_2-norm relative error of logistic regression where the covariance matrix is diagonal. The right plot show the squared ℓ_2-norm relative error of Poisson regression where the covariance matrix is a random orthogonal matrix.

We also conduct experiment for GLM with logistic loss on the SUSY dataset (Baldi et al., 2014). The task is to classify whether the class label is signal or background. After the pre-processing and sampling, the dataset contains 500000 samples and 18 features. Then we divide the data into training and testing, where $n_{\text{training}} = 350000$ and $n_{\text{testing}} = 150000$ and randomly choose the sample size $n \in 10^4 \cdot \{1, 3, \cdots, 33\}$ from the training data and use the same amount of public data. Regarding the privacy parameter, we take $\delta = \frac{1}{n^{1.1}}$ and let ϵ take value from $\{20, 10, 5\}$. We measure the performance by the prediction accuracy. For each combination of ϵ and n, the experiment is repeated 1000 times.



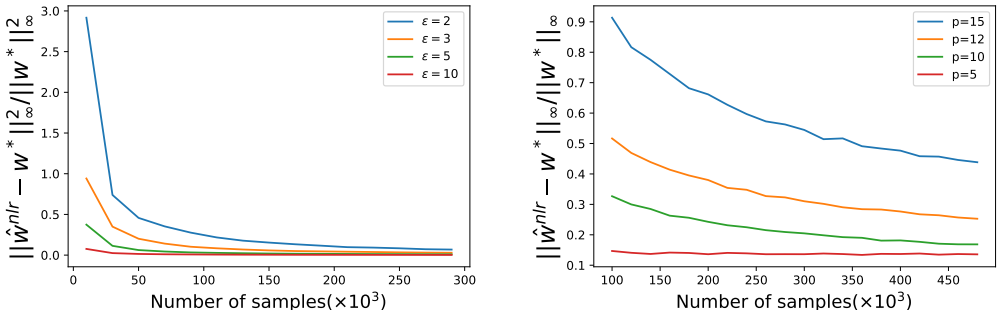
(*a*) (*b*)

Figure 3: GLM with logistic loss under i.i.d Bernoulli design. The left plot shows the squared relative error under different levels of privacy. The right one shows relative error under different dimensionality.



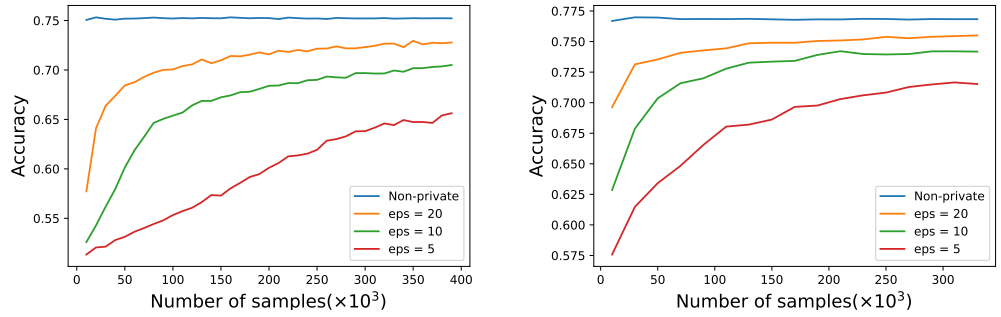
(*a*) (*b*)

Figure 4: Poisson regression under i.i.d Bernoulli design. The left plot shows the squared relative error under different levels of privacy. The right one shows relative error under different dimensionality.



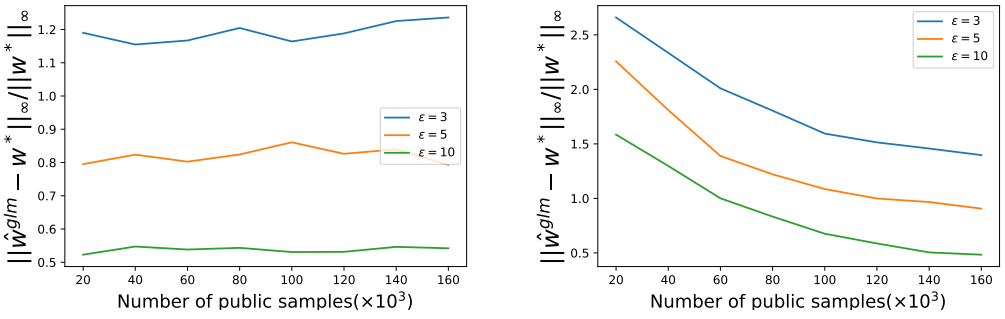
(*a*) (*b*)

Figure 5: Cubic regression with i.i.d Bernoulli design. The left plot shows the squared relative error under different level of privacy. The right one shows relative error under different dimensionality.



(*a*) (*b*)

Figure 6: GLM with logistic loss on real dataset. Left is for Covertype and right is for SUSY.



(*a*) (*b*)

Figure 7: The effect of the number of public unlabeled samples. The left plot shows the relative error of GLM with logistic loss. The right one shows the relative error of cubic regression.

6.1. Experimental Results

Evaluation on synthetic data: Figure 1, 2 are the results for Gaussian feature vectors, and Figure 3, 4, 5 are the results for Bernoulli feature vectors. We can see that the square of relative error is inversely proportional to the number of samples n. In other words, in order to achieve relative error α, we only need the number of private samples $n \sim \frac{1}{\alpha^2}$ if we omit the dependency on the other parameters. Besides, we also observe that the square of relative error is proportional to $\frac{1}{\epsilon^2}$, which matches our theoretical result. Moreover, we can see that the relative error increases as the dimensionality increases. It may seem a little weird that it is not linear in the dimensionality. We note that as the dimensionality p changes, some other parameters, for example, the l_2 norm of the covariance matrix and w^*_∞ also change, which bring other effects to the relative error.

Evaluation on real data From Figure 6 we can observe that when ϵ takes a reasonable value, the performance is approaching to the non-private case, provided that the size of private dataset is large enough. Thus, our algorithm is practical and is comparable to the non-private one.

The effect of public unlabeled data We use similar setting as our synthetic experiments in Section 6.1. For GLM we consider the problem of binary logistic loss *i.e.,* $\Phi(\langle x, w\rangle) = \ln\left(1 + \exp\left(\langle x, w\rangle\right)\right)$ in (1) while for non-linear regression we will set $f(x) = \frac{1}{3}x^3$ in (10). We compare relative error $\frac{\|\hat{w}-w^*\|_\infty}{\|w^*\|_\infty}$ with respect to different privacy parameters $\epsilon \in \{10, 5, 3\}$ with $\delta = \frac{1}{n}$. In these experiments, we fix dimensionality $p = 10$ and the population parameter $w^* = (1, 1, ..., 1)/\sqrt{p}$. We also fix the private sample size $n = 200000$ and the public data size is chosen from the set $10^4 \cdot \{2, 4, ..., 16\}$. We assume that the same amount of public unlabeled data is available. The features are generated independently from a Bernoulli distribution $\Pr\left(x_{i,j} = \pm\frac{1}{p}\right) = 0.5$ and the label is generated according to the logistic model or the model (10). In non-linear regression model, σ is bounded by $C = 0.001$. The results are shown in Figure 7(*a*) and 7(*b*). We can see that sometimes there is no need to use as large amount of public data as our theoretical result requires to guarantee a good performance, as is shown by Figure 7(*a*).

7. Conclusion and Open Problems

In this paper, motivated by Stein's lemma and its variants, we propose the first efficient algorithm with polynomial sample complexity for Generalized Linear Model estimation in the Non-interactive Local Differential Privacy model with some public unlabeled data. The main idea of our algorithm is to use OLS (Ordinary Least Square) estimator to approximate the underlying one. The key observation is that, after multiplying the OLS vector some constant, we can get a new estimator which can approximate the underlying estimator very well. Thus, we use the private data to estimate the OLS vector and the public unlabeled data to get the constant. Moreover, we use the same technique to the non-linear regression problem and show the same phenomenon.

There are still many open problems left. First, in this paper we mainly focused on the low dimensional case, where $n \gg p$. How to generalize to the high dimensional sparse case, that is $n \ll p$ and $\|w^*\|_0 \leq k$? Here since the Stein's lemma will not be hold, 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 will still have good performance. Thus, how to relax the assumptions and reduce the sample complexity of public unlabeled data in our theoretical results? Finally, for

the sub-Gaussian case, our estimator is biased and the error is $\Omega(\frac{1}{\sqrt{p}})$, can we get an unbiased and consistent estimator?

Acknowledgments

Di Wang and Lijie Hu were support in part by the baseline funding of King Abdullah University of Science and Technology (KAUST). Huanyu Zhang supported was in part by the National Science Foundation (NSF) under Grant No. 1815893 and 1704443. Jinhui Xu was supported in part by the National Science Foundation (NSF) under Grant No. CCF-1716400 and IIS-1919492. Part of the work was done when Di Wang and Marco Gaboardi were visiting the Simons Institute of the Theory for Computing.

References

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.

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.

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.

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.

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

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.

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.

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.

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.

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

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. *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. 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 $\Sigma^{\frac{1}{2}} = \text{diag}(\sqrt{\lambda_1}, \cdots, \sqrt{\lambda_p})$.

Lemma 27 (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 28 *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 28] 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 29 (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 30 (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 31 *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 32 *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 33 *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 $n \geq \Omega(\frac{\kappa_x^4\|\Sigma\|_2^2pr^4\log\frac{1}{\delta}}{\epsilon^2\lambda_{\min}^2(\Sigma)})$, the following holds with probability at least $1-\exp(-\Omega(p))-\xi$,*

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 = O(\frac{pr^2(1+r^2\|\tilde{w}^{ols}\|_2^2)\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2n\lambda_{\min}^2(\Sigma)}), \tag{16}$$

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

Proof [Proof of Lemma 33] 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 $\mathcal{N}(0, \sigma_2^2 I_p)$ 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 30, 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 29 and Lemma 27 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 $n\lambda_{\min}(\Sigma) \geq O(\frac{\kappa_x^2 \|\Sigma\|_2 r^2 \sqrt{pn \log \frac{1}{\delta}}}{\epsilon})$, which is true under the assumption of $n \geq \Omega(\frac{\kappa_x^4 \|\Sigma\|_2^2 pr^4 \log \frac{1}{\delta}}{\epsilon^2 \lambda_{\min}^2(\Sigma)})$. 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

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

Since we already assume that $X^TX + E_1 \succ \frac{X^TX}{2}$, by Lemma 28 we can obtain the following with probability at least $1 - \xi$

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

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

Thus, we have

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

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

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

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 34 (Prop. 7 in (Erdogdu et al., 2019)) *Assume that* $\mathbb{E}[x_i] = 0$, $\mathbb{E}[x_i x_i^T] = \Sigma$, *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 Gaussian mechanism and the composition property of DP.

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

Appendix C. Proofs and Comments in Section 4

Since Theorem 17 is the most complicated one, we will first prove it and then Theorem 14.

C.1. Proof of Theorem 17

Since $r = O(1)$ (by assumption), combining this with Lemmas 33 and 34, 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

$$\|\hat{w}^{ols} - w^{ols}\|_2 \leq C_3 \frac{\kappa_x \sqrt{p} r^2 \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}. \tag{17}$$

Lemma 35 *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 33 and Eq. (17), 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*

$$m \geq \Omega\big((\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2 \max\{p \log m\tau^{-2}, \frac{1}{G^2\tilde{\mu}^2} \frac{\epsilon^2 n}{pr^4 \|w^{ols}\|_2^2 \log \frac{1}{\delta} \log \frac{p^2}{\xi} \|\Sigma\|_2}\}\big), \tag{18}$$

$$n \geq \Omega(\kappa_x^4 G^2 \bar{c}^4 \|\Sigma\|_2 \frac{pr^4 \|w^{ols}\|_2^2 \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\tau^2 \epsilon^2 \lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}}), \tag{19}$$

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*

$$|\hat{c}_\Phi - \bar{c}_\Phi| \leq O\big(M^{-1}\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}} + M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\big). \tag{20}$$

Proof [Proof of Lemma 35] 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.

$$\delta = C_3\frac{\kappa_x\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}, \delta' = \frac{\|\Sigma\|_2^{\frac{1}{2}}\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}, \tag{21}$$

where C_3 is the one in (17). 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

$$\sup_{c\in[0,\bar{c}]}|\hat{f}(c,\hat{w}^{ols}) - f(c,\hat{w}^{ols})| \leq \sup_{c\in[0,\bar{c}]}\sup_{w\in\mathbb{B}_\Sigma^{\delta'}(w^{ols})}|\hat{f}(c,w) - f(c,w)|, \tag{22}$$

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

$$\begin{aligned}
&\sup_{c\in[0,\bar{c}]}\sup_{w\in\mathbb{B}_\Sigma^{\delta'}(w^{ols})}|\hat{f}(c,\hat{w}^{ols}) - f(c,\hat{w}^{ols})| \\
&\leq \bar{c}\sup_{c\in[0,\bar{c}]}\sup_{w\in\mathbb{B}_\Sigma^{\delta'}(w^{ols})}|\frac{1}{m}\sum_{j=1}^m\Phi^{(2)}(\langle v_i, \Sigma^{\frac{1}{2}}w\rangle c) - \mathbb{E}\Phi^{(2)}(\langle v, \Sigma^{\frac{1}{2}}w\rangle c)| \\
&= \bar{c}\sup_{c\in[0,\bar{c}]}\sup_{\Sigma^{\frac{1}{2}}w\in\mathbb{B}^{\delta'}(\bar{w}^{ols})}|\frac{1}{m}\sum_{j=1}^m\Phi^{(2)}(\langle v_i, \Sigma^{\frac{1}{2}}w\rangle c) - \mathbb{E}\Phi^{(2)}(\langle v, \Sigma^{\frac{1}{2}}w\rangle c)| \\
&= \bar{c}\sup_{w'\in\mathbb{B}^{\bar{c}\delta'}(\bar{w}^{ols})}|\frac{1}{m}\sum_{j=1}^m\Phi^{(2)}(\langle v_i, w'\rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w'\rangle)|.
\end{aligned} \tag{23}$$

By Lemma 31, we know that when $mp \geq 51\max\{\chi,\chi^{-1}\}$, where

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

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

$$\sup_{w' \in \mathbb{B}^{\bar{c}\delta}(\bar{w}^{ols})} |\frac{1}{m}\sum_{j=1}^{m} \Phi^{(2)}(\langle v_i, w'\rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w'\rangle)| \leq O((\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}). \tag{24}$$

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

$$\begin{aligned}\sup_{c\in[0,\bar{c}]} |f(c, w_1) - f(c, w_2)| &\leq G\bar{c}^2\mathbb{E}[\langle v, \Sigma^{\frac{1}{2}}(w_1 - w_2)\rangle] \\ &\leq \kappa_x G\bar{c}^2\|\Sigma^{\frac{1}{2}}(w_1 - w_2)\|_2. \end{aligned}\tag{25}$$

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 (23), (24), (25), and taking δ as in (21), we get

$$\sup_{c\in[0,\bar{c}]} |\hat{f}(c, \hat{w}^{ols}) - f(c, w^{ols})| \leq O\big(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}} + G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}} \frac{\kappa_x^2\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\big). \tag{26}$$

Let B denote the RHS of (26). 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

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

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 (22), (23) and (24), 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 (26), we have

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

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

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

■

Next, we prove our main theorem.

Proof [**Proof of Theorem 17**] By definition, we have

$$\begin{aligned} \|\hat{w}^{glm} - w^*\|_\infty &\le \|\hat{c}_\Phi \hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols} - w^*\|_\infty \\ &\le \|\hat{c}_\Phi \hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty + \|c_\Phi w^{ols} - w^*\|_\infty. \end{aligned} \tag{27}$$

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

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

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

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

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

By Lemma 12, we have

$$|\bar{c}_\Phi - c_\Phi| \le 16 M^{-1} c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}. \tag{28}$$

Thus, the second term of (27) is bounded by

$$\begin{aligned} &\|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty \le 16 M^{-1} c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}} \|w^{ols}\|_\infty \\ &\le 16 M^{-1} c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty^3}{\sqrt{p}} \big(\frac{1}{c_\Phi} + 16 G r \kappa_x^3 \sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}}\big) \\ &= O\big(M^{-1} r^3 \kappa_x^6 G^3 \rho_2 \rho_\infty^2 \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}} \max\{1, c_\Phi\}\big), \end{aligned} \tag{29}$$

where the last inequality is due to Lemma 12.

By Lemma 12, the third term of (27) is bounded by $16c_\Phi Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty^2}{\sqrt{p}}$.

For the first term of (27), by (17) and Lemma 35 we have

$$\begin{aligned}
&\|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty \le |\hat{c}_\Phi|\cdot\|\hat{w}^{ols}-w^{ols}\|_\infty+|\hat{c}_\Phi-\bar{c}_\Phi|\cdot\|w^{ols}\|_\infty\\
&\le O\big(\bar{c}\frac{\kappa_x\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\\
&+\|w^{ols}\|_\infty\big(M^{-1}\bar{c}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}+M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\big)\big).
\end{aligned}\tag{30}$$

For the first term of (30), we have

$$\begin{aligned}
&\bar{c}\frac{\kappa_x\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\le\bar{c}\frac{\kappa_x pr^2\|w^{ols}\|_\infty\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\\
&\le\bar{c}\frac{\kappa_x pr^2\|w^*\|_\infty\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}(\frac{1}{c_\Phi}+16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})\\
&=O\big(\bar{c}\frac{p\kappa_x^4\sqrt{\rho_2}\rho_\infty Gr^3\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}\big).
\end{aligned}\tag{31}$$

For the second term of (30), we have

$$\begin{aligned}
&\|w^{ols}\|_\infty M^{-1}\bar{c}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\\
&\le\bar{c}\|w^*\|_\infty(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}(\frac{1}{c_\Phi}+16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})\\
&\le O\big(Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\bar{c}\|w^*\|_\infty\max\{1,\|w^*\|_\infty\}(\kappa_g+\frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\max\{1,\frac{1}{c_\Phi}\}\big).
\end{aligned}\tag{32}$$

For the third term of (30), we have

$$\begin{aligned}
&\|w^{ols}\|_\infty M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}})\\
&\le M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{pr^2\|w^*\|_\infty^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}(\frac{1}{c_\Phi}+16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})^2\\
&\le O\big(M^{-1}G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2\frac{pr^4\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\big).
\end{aligned}\tag{33}$$

Thus, the first term of (27) is bounded by (since $m \geq \Omega(n)$)

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

Putting all the bounds together, we have

$$\|\hat{w}^{glm} - w^*\|_\infty \leq \tilde{O}\big(M^{-1}G^3\kappa_x^8\bar{c}^2\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_2$$
$$\times\frac{pr^4\|w^*\|_\infty\max\{1, \|w^*\|_\infty^3\}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\max\{1, \frac{1}{c_\Phi}\}^2$$
$$+ M^{-1}r^3\kappa_x^6 c_\Phi G^3\rho_2\rho_\infty^2\frac{\|w^*\|_\infty^2\max\{1, \|w^*\|_\infty^2\}}{\sqrt{p}}\max\{1, \frac{1}{c_\Phi}\}+$$
$$Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\bar{c}\|w^*\|_\infty\max\{1, \|w^*\|_\infty\}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p\log m}{m}}\max\{1, \frac{1}{c_\Phi}\}\big). \tag{34}$$

Next, we bound the probability. We assume that Lemma 33, 34 and 35 hold with probability at least $1 - \exp(-\Omega(p)) - \rho$. They hold when

$$m \geq \Omega\big((\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2\max\{p\log m\tau^{-2}, \frac{1}{G^2\tilde{\mu}^2}\frac{\epsilon^2 n}{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}\}\big), \tag{35}$$

$$n \geq \Omega(\max\{\kappa_x^4 G^2\bar{c}^4\|\Sigma\|_2\frac{pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}, \frac{\kappa_x^4\|\Sigma\|_2^2 pr^4\log\frac{1}{\delta}}{\epsilon^2\lambda_{\min}^2(\Sigma)}\}\big). \tag{36}$$

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

$$n \geq \Omega\big(G^4\bar{c}^4\|\Sigma\|_2^2\frac{p^2r^6\kappa_x^{10}\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1, \|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\max\{1, \frac{1}{c_\Phi}\}^2\big). \tag{37}$$

■

C.2. Proof of Theorem 14

Lemma 36 *Let $\bar{c}_\Phi, \bar{c}, \tau, f, \hat{f}$ be defined the same as in Lemma 35. 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 33 and (17), 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*

$$|\hat{c}_\Phi - \bar{c}_\Phi| \leq O\Big(\frac{M^{-1}GL\bar{c}^2\kappa_x^2 r^2 \|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\}\sqrt{n}} + M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\Big) \tag{38}$$

for sufficiently large m, n such that

$$n \geq \Omega\Big(\frac{LG^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4 p r^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\Big) \tag{39}$$

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2 p\tau^{-2}\big). \tag{40}$$

Proof [Proof of Lemma 36] The main idea of this proof is almost the same as the one for Lemma 35. The only difference is that instead of using Lemma 31 to get (24), we use here Lemma 32 to obtain the following with probability at least $1 - \exp(-p)$

$$\begin{aligned}
&\sup_{w' \in \mathbb{B}^{\bar{c}\delta'}(\bar{w}^{ols})} |\frac{1}{m}\sum_{j=1}^{m}\Phi^{(2)}(\langle v_i, w'\rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w'\rangle)| \\
&\leq O\big((G(\|\bar{w}^{ols}\|_2 + \bar{c}\delta')\|I\|_2 + L)\sqrt{\frac{p}{m}} \\
&\leq O\big((G\|\Sigma\|_2^{\frac{1}{2}}(\|w^{ols}\|_2 + \bar{c}\frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}) + L)\sqrt{\frac{p}{m}}\big).
\end{aligned} \tag{41}$$

Thus, by (23), (25) and (41), we have

$$\begin{aligned}
\sup_{c\in[0,\bar{c}]}|\hat{f}(c, \hat{w}^{ols}) - f(c, w^{ols})| \leq O\big(&G\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}} + \\
&\frac{G\kappa_x\bar{c}\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{p}r^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\sqrt{\frac{p}{mn}} + L\sqrt{\frac{p}{m}}\big).
\end{aligned} \tag{42}$$

Let D denote the RHS of (42), 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

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

That is,

$$n\geq\Omega\Big(\frac{G^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\Big) \tag{43}$$

$$m\geq\Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2p\tau^{-2}\big). \tag{44}$$

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

$$\begin{aligned}
&M|\hat{c}_\Phi-\bar{c}_\Phi|\leq|f(\hat{c}_\Phi,w^{ols})-f(\bar{c}_\Phi,w^{ols})|\\
&\leq O\Big(\frac{G\bar{c}^2\kappa_x^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}\\
&+\frac{G\kappa_x\bar{c}\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{p}r^2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}\sqrt{\frac{p}{mn}}+LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\Big) &(45)\\
&\leq O\Big(\frac{GL\bar{c}^2\kappa_x^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}+LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\Big). &(46)
\end{aligned}$$

■

Proof [**Proof of Theorem 14**] The proof is almost the same as the one for Theorem 17. By definition, we have

$$\begin{aligned}
\|\hat{w}^{glm}-w^*\|_\infty&\leq\|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty+\|\bar{c}_\Phi w^{ols}-w^*\|_\infty\\
&\leq\|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty+\|\bar{c}_\Phi w^{ols}-c_\Phi w^{ols}\|_\infty+\|c_\Phi w^{ols}-w^*\|_\infty. &(47)
\end{aligned}$$

The second term of (47) is bounded by

$$\|\bar{c}_\Phi w^{ols}-c_\Phi w^{ols}\|_\infty\leq O\big(M^{-1}r^2\kappa_x^7c_\Phi G^3\rho_2\rho_\infty^2\frac{\|w^*\|_\infty^3\max\{1,\|w^*\|_\infty\}}{\sqrt{p}}\max\{1,\frac{1}{c_\Phi}\}\big). \tag{48}$$

By Lemma 12, the third term of (47) is bounded by $16c_\Phi Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}}$. The first term is bounded by

$$\begin{aligned}
&\|\hat{c}_\Phi\hat{w}^{ols}-\bar{c}_\Phi w^{ols}\|_\infty\leq\\
&O\Big(\frac{M^{-1}G^3L\bar{c}^2\kappa_x^8r^4\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}\times\max\{\frac{1}{c_\Phi},1\}^2\\
&+\frac{M^{-1}G^3L\bar{c}^2\kappa_x^6r^2\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p}{\sqrt{m}}\times\max\{\frac{1}{c_\Phi},1\}^2\Big). &(49)
\end{aligned}$$

Thus, in total we have

$$\|\hat{w}^{glm} - w^*\|_\infty \le O\Big(\frac{M^{-1}G^3L\bar{c}^2\kappa_x^6 r^2\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p}{\sqrt{m}} \times \max\{\frac{1}{c_\Phi},1\}^2$$
$$+\frac{G^3L\bar{c}^2\kappa_x^6 r^4\rho_2\rho_\infty^2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}\|\Sigma\|_2^{\frac{1}{2}}p\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}\max\{\frac{1}{c_\Phi},1\}^2$$
$$+M^{-1}r^2\kappa_x^7 c_\Phi G^3\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_\infty\frac{\|w^*\|_\infty^3\max\{1,\|w^*\|_\infty\}}{\sqrt{p}}\max\{1,\frac{1}{c_\Phi}\}\Big). \quad (50)$$

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

$$m \ge \Omega\big(G^2L^2\|\Sigma\|_2\|w^*\|_\infty^2\max\{1,\|w^*\|_\infty^2\}G^2r^2\kappa_x^6\rho_2\rho_\infty^2p^2\tau^{-2}\max\{1,\frac{1}{c_\Phi}\}^2\big) \quad (51)$$

$$n \ge \Omega\Big(\frac{\rho_2\rho_\infty^2G^4\tau^{-2}\bar{c}^4\|\Sigma\|_2^2\kappa_x^{10}p^2\|w^*\|_\infty^2r^6\max\{1,\|w^*\|_\infty^2\}\log\frac{1}{\delta}\log\frac{p^3}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\max\{1,\frac{1}{c_\Phi}\}^2\Big). \quad (52)$$

■

C.3. Proof of Theorem 8

Proof By Lemma 4 we know that with probability $1-\xi$, $\|x_i\|_2 \le \sqrt{5p\|\Sigma_m\|_2\log\frac{n}{\xi}}$ for each $i\in[n]$. Next we will bound the term of $\|\Sigma_m\|_2$. By Lemma 29 we can see that when $m\ge\Omega(p)$, with probability at least $1-\exp(-p)$, $\|\Sigma_m\|_2 \le (1+C_1\sqrt{\frac{p}{m}})\|\Sigma\|_2 \le 2\|\Sigma\|_2$. Thus $\|x_i\|_2 \le \sqrt{10p\|\Sigma\|_2\log\frac{n}{\xi}}$ with probability at least $1-\zeta-\exp(-p)$. In the following we will assume this is true.

Combining this with Lemmas 33 and 34, 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

$$\|\hat{w}^{ols}-w^{ols}\|_2 \le C_3\frac{\sqrt{p^3}\|\Sigma\|_2\|w^{ols}\|_2\log\frac{n}{\xi}\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma),1\}}. \quad (53)$$

The same Lemma 36, we have the following lemma.

Lemma 37 *Let $\bar{c}_\Phi,\bar{c},\tau,f,\hat{f}$ be defined the same as in Lemma 35. If further assume that $|\Phi^{(2)}(\cdot)|\le L$ for some constant $L>0$ and is Lipschitz continuous with constant G, then, under the assumptions in Lemma 33 and (17), 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 $c_\Phi=\bar{c}_\Phi$)*

$$|\hat{c}_\Phi-c_\Phi| \le O\Big(\frac{M^{-1}GL\bar{c}^2\|\Sigma\|_2^{\frac{3}{2}}p^{\frac{3}{2}}\|w^{ols}\|_2\log\frac{n}{\xi}\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma),1\}\sqrt{n}}+M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\Big) \quad (54)$$

for sufficiently large m, n such that

$$n \geq \Omega\big(\frac{LG^2\tau^{-2}\bar{c}^4\|\Sigma\|_2^3 p^3\|w^{ols}\|_2^2\log^2\frac{n}{\xi}\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma),1\}}\big) \tag{55}$$

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2 p\tau^{-2}\big). \tag{56}$$

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

$$\|\hat{c}_\Phi\hat{w}^{ols} - c_\Phi w^{ols}\|_2 \leq |\hat{c}_\Phi - c_\Phi|\|\hat{w}^{ols}\|_2 + c_\Phi\|\hat{w}^{ols} - w^{ols}\|_2. \tag{57}$$

For the second term of (57), by (53) we have

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

For the first term of (57), by Lemma 37 and (53) we have

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

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

C.4. Proof of Theorem 19

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 23) 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 20

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 21

Proof [Proof of Theorem 21] 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 (10) 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 24

The idea of the proof follows the one in (Erdogdu et al., 2019).

Proof [Proof of Theorem 24] 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 25

By the same argument in the proof of Lemma 33, we can show that when $n \geq \Omega(\frac{\kappa_x^4 \|\Sigma\|_2^2 p r^4 \log \frac{1}{\delta}}{\epsilon^2 \lambda_{\min}^2(\Sigma)})$, with probability at least $1 - \exp(-\Omega(p)) - \xi$, the following holds

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 = O\big(\frac{pC^2r^2(L^2r^2 + C^2 + r^2\|\tilde{w}^{ols}\|_2^2)\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2 n \lambda_{\min}^2(\Sigma)}\big). \tag{58}$$

Thus, by Lemma 34 we have

$$\|\hat{w}^{ols} - w^{ols}\|_2 \leq O\big(\frac{CL\kappa_x\sqrt{p}r^2\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\big). \tag{59}$$

In the following, we will always assume that (59) holds. By the same argument given in Lemma 36, we have the following Lemma, which can be proved in the same way as Lemma 36.

Lemma 38 *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 33 and Eq. (59), 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*

$$|\hat{c}_f - \bar{c}_f| \leq O\big(\frac{M^{-1}CGL\bar{c}^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\log\frac{1}{\delta}\log\frac{p}{\xi^2}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\}\sqrt{n}} + M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\big) \tag{60}$$

for sufficiently large m, n such that

$$n \geq \Omega\big(\frac{LG^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4 p r^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\big) \tag{61}$$

$$m \geq \Omega\big(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2 p\tau^{-2}\big). \tag{62}$$

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

D.4. Proof of Theorem 22

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

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

where $r = \sqrt{10p\|\Sigma\|_2 \log \frac{n}{\xi}}$. Thus, by Lemma 34 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

$$\|\hat{w}^{ols} - w^{ols}\|_2 \leq C_3 \frac{\sqrt{p^3}\|\Sigma\|_2 \|w^{ols}\|_2 \log \frac{n}{\xi} \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}. \tag{64}$$

The same Lemma 37, we have the following lemma.

Lemma 39 *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. (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*

$$|\hat{c}_f - c_f| \leq O\Big(\frac{M^{-1} G L \bar{c}^2 \|\Sigma\|_2^{\frac{3}{2}} p^{\frac{3}{2}} \|w^{ols}\|_2 \log \frac{n}{\xi} \sqrt{\log \frac{1}{\delta} \log \frac{p}{\xi^2}}}{\epsilon \lambda_{\min}^{\frac{1}{2}}(\Sigma) \min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\} \sqrt{n}} + M^{-1} L G \|\Sigma\|_2^{\frac{1}{2}} \|w^{ols}\|_2 \sqrt{\frac{p}{m}}\Big) \tag{65}$$

for sufficiently large m, n such that

$$n \geq \Omega\Big(\frac{L G^2 \tau^{-2} \bar{c}^4 \|\Sigma\|_2^3 p^3 \|w^{ols}\|_2^2 \log^2 \frac{n}{\xi} \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\epsilon^2 \lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}}\Big) \tag{66}$$

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

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

$$\|\hat{c}_f \hat{w}^{ols} - c_f w^{ols}\|_2 \leq |\hat{c}_f - c_f| \|\hat{w}^{ols}\|_2 + c_f \|\hat{w}^{ols} - w^{ols}\|_2. \tag{68}$$

For the second term of (68), by (64) we have

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

For the first term of (68), by Lemma 39 and (53) we have

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

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

D.5. Proof of Theorem 26

Proof We can easily see that $f'(\cdot)$ is just the function $\Phi^{(2)}(\cdot)$ in Theorem 19 for the logistic loss function. Thus the function f' satisfies the assumptions in Theorem 25, which was showed in the Theorem 19. ■