On Sparse Linear Regression in the Local Differential Privacy Model

Di Wang, *Student Member, IEEE*, and Jinhui Xu, *Member, IEEE*

***Abstract*—In this paper, we study the sparse linear regression problem under the Local Differential Privacy (LDP) model. We first show that polynomial dependency on the dimensionality p of the space is unavoidable for the estimation error in both non-interactive and sequential interactive local models, if the privacy of the whole dataset needs to be preserved. Similar limitations also exist for other types of error measurements and in the relaxed local models. This indicates that differential privacy in high dimensional space is unlikely achievable for the problem. With the understanding of this limitation, we then present two algorithmic results. The first one is a sequential interactive LDP algorithm for the low dimensional sparse case, called Locally Differentially Private Iterative Hard Thresholding (LDP-IHT), which achieves a near optimal upper bound. This algorithm is actually rather general and can be used to solve quite a few other problems, such as (Local) DP-ERM with sparsity constraints and sparse regression with non-linear measurements. The second one is for the restricted (high dimensional) case where only the privacy of the responses (labels) needs to be preserved. For this case, we show that the optimal rate of the error estimation can be made logarithmically dependent on p (i.e., $\log p$) in the local model, where an upper bound is obtained by a label-privacy version of LDP-IHT. Experiments on real world and synthetic datasets confirm our theoretical analysis.**

***Index Terms*—Sparse linear regression, local differential privacy.**

I. Introduction

LINEAR regression is a fundamental and classical tool for data analysis, and finds numerous applications in social sciences [2], genomics research [3] and signal recovery [4]. One frequently encountered challenge for such a technique is how to deal with the high dimensionality of the dataset, such as those in genomics, educational and psychological research. A commonly adopted strategy for dealing with such an issue is to assume that the unknown regression vector is sparse.

Another often encountered challenge for linear regression is how to handle sensitive data, such as those in social science. As a commonly-accepted approach for preserving privacy, differential privacy [5] provides provable protection against identification and is resilient to arbitrary auxiliary information that might be available to attackers. Methods to guarantee differential privacy have been widely studied, and recently adopted in industry [6]–[8].

Two main user models have emerged for differential privacy: the central model and the local one. In the central model, data are managed by a trusted central entity which is responsible for collecting them and for deciding which differentially private data analysis to perform and to release. A classical application of this model is the one of census data. In the local model instead, each individual manages his/her proper data and discloses them to a server through some differentially private mechanisms. The server collects the (now private) data of each individual and combines them into a resulting data analysis. A classical example of this model is the one aiming at collecting statistics from user devices like in the case of Google's Chrome browser [7], and Apple's iOS-10 [6], [8].

Despite being used in industry, the local model has been much less studied than the central one. Part of the reason for this is that there are intrinsic limitations in what one can do in the local model. As a consequence, many basic questions, that are well studied in the central model, have not been completely understood in the local model, yet.

To advance our understanding on the local model, we study, in this paper, the locally differentially private version of the sparse linear regression problem, where each user $i \in [n]$ holds a data record $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$. There are two commonly used ways for measuring the performance of this problem, which correspond to two different settings, the statistical learning and the statistical estimation settings. For the first setting, the measurement is based on the optimization error, *i.e.* $F(\theta^{\text{priv}}) - \min_{\theta \in \mathcal{C}} F(\theta)$, where $F(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{P}}(\langle x, \theta\rangle - y)^2$, and $\mathcal{P}$ is an unknown distribution. For the second setting, y is assumed to be $y = \langle x, \theta^*\rangle + \sigma$, where $x \sim \mathcal{D}$, $\mathcal{D}$ is a known distribution, σ is a random noise, and $\theta^* \in \mathbb{R}^p$ is the to-be-estimated vector that satisfies the condition of $\|\theta^*\|_0 \leq s$. The estimation error for this setting is represented by the loss of the squared ℓ_2 norm, *i.e.*, $\|\theta^{priv} - \theta^*\|_2^2$. In this paper, we will focus on the latter setting, and assume that $x \sim \text{Uniform}\{+1, -1\}^p$.

Manuscript received June 27, 2019; revised June 12, 2020; accepted October 27, 2020. Date of publication November 25, 2020; date of current version January 21, 2021. The work of Di Wang and Jinhui Xu was supported in part by the National Science Foundation (NSF) under Grant CCF-1716400 and Grant IIS-1919492. This article was presented at the 36th International Conference on Machine Learning (ICML 2019). *(Corresponding author: Di Wang.)*

Di Wang is with the Division of Computer, Electrical, and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia, and also with the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14260 USA.

Jinhui Xu is with the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14260 USA (e-mail: dwang45@buffalo.edu; jinhui@buffalo.edu).

Communicated by K. Chaudhuri, Associate Editor for Statistical Learning.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2020.3040406.

Digital Object Identifier 10.1109/TIT.2020.3040406

Our contributions can be summarized as follows:

- We first present a negative result which suggests that the ϵ non-interactive private minimax risk of $\|\theta^{\text{priv}} - \theta^*\|_2^2$ is lower bounded by $\Omega(\frac{p \log p}{n\epsilon^2})$ if the privacy of the whole dataset $\{(x_i, y_i)\}_{i=1}^n$ needs to be preserved. This indicates that it is impossible to obtain any non-trivial error bound in high dimensional space (*i.e.* $p \gg n$). The private minimax risk is still lower bounded by $\Omega(\frac{p}{n\epsilon^2})$, even in the sequentially interactive local model. Our proofs are based on a locally differentially private version of the Fano and Le Cam method [9]–[11]. We further reveal that this polynomial dependency on p cannot be avoided even if the measurement of the loss function or definitions of differential privacy is relaxed.
- With the understanding of this limitation, we then propose an ϵ-sequential interactive LDP algorithm for the low dimensional sparse case, called Locally Differentially Private Iterative Hard Thresholding (LDP-IHT), which achieves a near optimal upper bound. Furthermore, we show that the idea of DP-IHT is actually rather general and can be used to achieve differential privacy for quite a few other problems. Specifically, it can be applied to the (Locally) Differentially Private Empirical Risk Minimization (DP-ERM) problem with sparsity constraints, and achieves an upper bound that depends only logarithmically on p (i.e., $\log p$) and the sparsity parameter of the optimal estimator, making it suitable for applications in high dimensions. To our best knowledge, this is the first paper studying DP-ERM with non-convex constraint set. Another application of LDP-IHT is the sparse regression problem with non-linear measurements [12], [13].
- We also give a positive result for high dimensions. Particularly, we consider the restricted case where only the responses (labels) are required to be private, *i.e.,* the dataset $\{x_i\}_{i=1}^n$ is assumed to be public and $\{y_i\}_{i=1}^n$ is private (note that this is a valid assumption as shown in [14], [15]). For this case, we propose a general algorithm which achieves an upper bound of $O(\frac{s \log p}{n\epsilon^2})$ for the estimation error. We show that this bound is actually optimal, as the ϵ non-interactive private minimax risk can also be lower bounded by $\Omega(\frac{s \log p}{n\epsilon^2})$.
- Finally, we perform our algorithms on both synthetic and real world datasets. Experimental results also support our theoretical analysis.

II. Related Work

There is a vast number of existing results studying the differentially private linear regression problem (or more generally, DP-ERM) from different perspectives, such as [16]–[22]. Below, we focus only on those with theoretical guarantees on the error.

For the central model, [18] recently conducted a comprehensive study, from both theoretical and practical points of views, on the differentially private linear regression problem. The author gave upper bounds of the optimization error in the statistical learning setting and the estimation error in the statistical estimation setting, as well as a general lower bound of the optimization error. There are also other works on this problem (we refer the reader to the Related Work section in [18] for more details). But all these results are only for the low dimensional case (*i.e.* the dimensionality p is a small constant number). Contrarily, we study mainly, in this paper, the high dimensional sparse case under the statistical estimation setting and provide both upper and lower bounds of the estimation error for the non-interactive and sequentially interactive models. A couple of results also exist for the high dimensional sparse linear regression problem in the central model [20], [23]; but all of them consider only the optimization error. [24] studied the problem of Bayesian linear regression, which is incomparable to our problem. Reference [19] focused the confidence interval of Ordinary Linear Regression while we mainly focus on the estimation error. It is notable that recently [25] studied the optimal rates of the estimation error of linear regression in both low dimension and high dimensional sparse settings. Specifically, for (ϵ, δ)-DP, they showed that in the low dimension setting, the near optimal rate of estimation error is $\tilde{O}(\sqrt{\frac{p}{n}} + \frac{p\sqrt{\log 1/\delta}}{n\epsilon})$, while in the high dimensional setting it is $\tilde{O}(\sqrt{\frac{s \log p}{n}} + \frac{s \log p \sqrt{\log 1/\delta}}{n\epsilon})$, here $\tilde{O}$-term omits $\log n$ factor. We will show more details in Remark 2 for the comparison between sparse linear regression in the central model and the local model.

Unlike the central model where tremendous progresses have been made, linear regression in the local model is still not well understood. The only known results are [9], [10], [21], [26]. Reference [9] studied the low dimensional, non-interactive private minimax risk of the estimation error for the restricted case of keeping the responses private, while we consider the high dimensional case of the problem in the interactive local model. Reference [21] gave the optimal lower bound of the optimization error, $\Theta(\sqrt{\frac{p}{n\epsilon^2}})$, for the low dimensional case which was later improved to $O((\frac{\log p}{n\epsilon^2})^{\frac{1}{4}})$ by [26], [27] in the case where the constraint set is a unit ℓ_1 norm ball. However, their settings are different from ours since they all assume that the norm of x_i is bounded by 1, *i.e.* $\|x_i\|_2 \leq 1$, while in our statistical setting, $\|x_i\|_2 = \sqrt{p}$. Thus, our results are incomparable with theirs.

DP-ERM has been studied in [9], [27]–[32] under different settings. However, none of these considered the non-convex constraint case.

To proof the low bounds in this paper, we mainly use private version of the Fano and Le Cam method, which are initially given by [9]–[11]. Based on different settings or problems, there are different versions of private Fano and Le Cam method. For example, [33] proposed a generalized private Assouad method to deal with the lower bounds of some matrix estimation problems in the local differential privacy model. Reference [34] proposed private Fano, Le Cam and Assouad method under central differential privacy. Reference [35] proved lower bounds for various testing and estimation problems under local differential privacy using a notion of chi-squared contractions based on Le Cam's method and Fano's inequality.

III. Preliminaries

In this section, we introduce some definitions that will be used throughout the paper. More details can be found in Section A of Appendix or [10].

A. Classical Minimax Risk

Since all of our lower bounds are in the form of private minimax risk, we first introduce the classical statistical minimax risk before discussing the locally private version.

Let $\mathcal{P}$ be a class of distributions over a data universe $\mathcal{X}$. For each distribution $p \in \mathcal{P}$, there is a deterministic function $\theta(p) \in \Theta$, where Θ is the parameter space. Let $\rho : \Theta \times \Theta :\mapsto \mathbb{R}_+$ be a semi-metric function on the space Θ and $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a non-decreasing function with $\Phi(0) = 0$ (in this paper, we assume that $\rho(x, y) = |x-y|$ and $\Phi(x) = x^2$ unless specified otherwise). We further assume that $\{X_i\}_{i=1}^n$ are n i.i.d observations drawn according to some distribution $p \in \mathcal{P}$, and $\hat{\theta} : \mathcal{X}^n \mapsto \Theta$ be some estimator. Then the minimax risk in metric $\Phi \circ \rho$ is defined by the following saddle point problem:

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) := \inf_{\hat{\theta}} \sup_{p \in \mathcal{P}} \mathbb{E}_p[\Phi(\rho(\hat{\theta}(X_1, \cdots, X_n), \theta(p))],$$

where the supremum is taken over distributions $p \in \mathcal{P}$ and the infimum over all estimators $\hat{\theta}$.

B. Local Differential Privacy and Private Minimax Risk

Since we will consider the sequential interactive and non-interactive local models in this paper, we follow the definitions in [9].

We assume that $\{Z_i\}_{i=1}^n$ are the private observations transformed from $\{X_i\}_{i=1}^n$ through some privacy mechanisms. We say that the mechanism is **sequentially interactive**, when it has the following conditional independence structure:

$$\{X_i, Z_1, \cdots, Z_{i-1}\} \mapsto Z_i, Z_i \perp\!\!\!\perp X_j \mid \{X_i, Z_1, \cdots, Z_{i-1}\}$$

for all $j \neq i$ and $i \in [n]$, where $\perp\!\!\!\perp$ means independent relation. The full conditional distribution can be specified in terms of conditionals $Q_i(Z_i \mid X_i = x_i, Z_{1:i-1} = z_{1:i-1})$. The full privacy mechanism can be specified by a collection $Q = \{Q_i\}_{i=1}^n$.

When Z_i is depending only on X_i, the mechanism is called **non-interactive** and in this case we have a simpler form for the conditional distributions $Q_i(Z_i \mid X_i = x_i)$. We now define local differential privacy by restricting the conditional distribution Q_i.

Definition 1 [9]: For given privacy parameters $\epsilon > 0, \delta \geq 0$, the random variable Z_i is an (ϵ, δ) sequentially locally differentially private view of X_i if for all $z_1, z_2, \cdots, z_{i-1}$ and $x, x' \in \mathcal{X}$ we have the following for all the events S:

$$Q_i(Z_i \in S \mid X_i = x_i, Z_{1:i-1} = z_{1:i-1}) \leq$$
$$e^{\epsilon} Q_i(Z_i \in S \mid X_i = x_i', Z_{1:i-1} = z_{1:i-1}) + \delta.$$

If $\delta = 0$, we will omit the term of δ (the same for other definitions).

We say that the random variable Z_i is an (ϵ, δ) non-interactively locally differentially private view of X_i if

$$Q_i(Z_i \in S \mid X_i = x_i) \leq e^{\epsilon} Q_i(Z_i \in S \mid X_i = x_i') + \delta.$$

We say that the privacy mechanism $Q = \{Q_i\}_{i=1}^n$ is (ϵ, δ)-sequentially (non-interactively) locally differentially private (LDP) if each Z_i is a sequentially (non-interactively) locally differentially private view.

For a given privacy parameter $\epsilon > 0$, let $\mathcal{Q}_\epsilon$ be the set of conditional distributions that have the ϵ-LDP property. For a given set of samples $\{X_i\}_{i=1}^n$, let $\{Z_i\}_{i=1}^n$ be the set of observations produced by any distribution $Q \in \mathcal{Q}_\epsilon$. Then, our estimator will be based on $\{Z_i\}_{i=1}^n$, that is, $\hat{\theta}(Z_1, \cdots, Z_n)$. This yields a modified version of the minimax risk:

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, Q) := \inf_{\hat{\theta}} \sup_{p \in \mathcal{P}} \mathbb{E}_p[\Phi(\rho(\hat{\theta}(Z_1, \cdots, Z_n), \theta(p))].$$

From the above definition, it is natural for us to seek the mechanism $Q \in \mathcal{Q}_\epsilon$ that has the smallest value for the minimax risk. This allows us to define functions that characterize the optimal rate of estimation in terms of privacy parameter ϵ.

Definition 2: Given a family of distributions $\theta(\mathcal{P})$ and a privacy parameter $\epsilon > 0$, the ϵ sequential private minimax risk in the metric $\Phi \circ \rho$ is:

$$\mathcal{M}_n^{\text{Int}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) := \inf_{Q \in \mathcal{Q}_\epsilon} \mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, Q),$$

where $\mathcal{Q}_\epsilon$ is the set of all ϵ sequentially locally differentially private mechanisms. Moreover, the ϵ non-interactive private minimax risk in the metric $\Phi \circ \rho$ is:

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) := \inf_{Q \in \mathcal{Q}_\epsilon} \mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, Q),$$

where $\mathcal{Q}_\epsilon$ is the set of all ϵ non-interactively locally differentially private mechanisms.

IV. Problem Set-Up

The focus of this paper is the sparse linear regression problem. In this problem, we have n pair of observations $\{(x_i, y_i)\}_{i=1}^n$, where each $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$. Moreover, there is some unknown parameter vector $\theta^* \in \mathbb{R}^p$ that links each pair (x_i, y_i) by the standard linear model

$$y_i = \langle x_i, \theta^* \rangle + \sigma_i,$$

where $|\sigma_i| \leq C$ is observation noise and $C > 0$ is some constant. Here θ^* satisfies the sparsity constraint, meaning that θ^* has no more than $s \ll p$ non-zero entries. The goal is to estimate the unknown vector θ^* based on these n observations while also under the local differential privacy constraint. Specifically, we want to find an estimator θ^{priv} via some locally differentially private algorithm to make its estimation error $\|\theta^{priv} - \theta^*\|_2^2$ be as small as possible. Specifically, in this paper we will focus on the following collection of samples $(x, y) \in \{+1, -1\}^p \times \mathbb{R}$:

$$\begin{aligned}\mathcal{P}_{s,p,C} = \{P_{\theta,\sigma} \mid x \sim \text{Uniform}\{+1, -1\}^p, y = \langle \theta, x \rangle + \sigma,\\ \text{where } \sigma \text{ is the random noise s.t } \mathbb{E}[\sigma|x] = 0, |\sigma| \leq C\\ \text{for some constant } C > 0, \|\theta\|_2 \leq 1, \|\theta\|_0 \leq s\}. \quad (1)\end{aligned}$$

In the above definition, σ is sampled from a bounded stochastic noise domain such as uniform distribution and could depend on x.

It is notable that in the non-private setting, [36] showed the following optimal minimax rate $\mathcal{M}_n(\theta(\mathcal{P}_{s,p,C}), \|\cdot\|_2^2) = \Theta(\frac{C^2 s \log \frac{p}{s}}{n})$.

It is worth noting that there is some difference between our model (1) and the sub-Gaussian linear model, which is a classic model in statistics [36]. That is, here x is assumed to follow a uniform distribution (which is an often adopted assumption in estimating lower bounds in differential privacy [37]) in our model, while it is often sampled from general sub-Gaussian distribution in a sub-Gaussian model. Even though the uniform distribution can be viewed as a sub-Gaussian distribution, the way of using it in our paper is different.

V. Keeping the Whole Dataset Private

A. Lower Bounds of Private Minimax Risk

In this section, we investigate the private minimax risk in the case where the whole dataset $\{(x_i, y_i)\}_{i=1}^n$ needs to be locally private, and show that even if the parameter vector θ^* is 1-sparse, the polynomial dependence on the dimensionality p in the estimation error cannot be avoided. This implies that achieving ϵ-LDP for the high dimensional sparse linear regression problem is unlikely.

To show the limitations of the problem with respect to the private minimax risk, we first give some intuition. Consider a raw data record (x_i, y_i) which is sampled from some $P_{\theta,\sigma} \in \mathcal{P}_{1,p,C}$, where $\mathcal{P}_{1,p,C}$ has the form as in (1). Suppose that we want to use a Gaussian or Laplacian mechanism on (x_i, y_i) in order to make the algorithm locally differentially private. Then, due to sensitivity, the ℓ_1 or ℓ_2 norm of (x_i, y_i) is a polynomial of p. The scale of the added random noise will also be a polynomial of p, which makes the final estimation error large.

The following theorem indicates that for some fixed privacy parameter $\epsilon \in (0, 1)$, the optimal rate of the ϵ non-interactive private minimax risk is lower bounded by $\Omega(\min\{1, \frac{p \log p}{n\epsilon^2}\})$.

Theorem 1: For a given fixed privacy parameter $\epsilon \in (0, \frac{1}{2}]$, the ϵ non-interactive private minimax risk (measured by the $\|\cdot\|_2^2$ metric) of the 1-sparse high dimensional sparse linear regression problem $\mathcal{P}_{1,p,2}$ needs to satisfy the following inequality,

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}_{1,p,2}), \|\cdot\|_2^2, \epsilon) \geq \Omega(\min\{1, \frac{p \log p}{n\epsilon^2}\}). \quad (2)$$

With the above theorems, our question now is to determine whether there are other factors in the local model that might allow us to avoid the polynomial dependency on p in the estimation error.

We first consider the necessity of interaction in the model, since for some problems, such as convex Empirical Risk Minimization (ERM), there exists a large gap in the estimation error between the interactive and non-interactive local models [21]. The following theorem suggests that even if sequential interaction is allowed in the local model, the polynomial dependence on p is still unavoidable. Note that sequential interaction is a commonly used model in LDP [9], [21].

Theorem 2: For a given fixed privacy parameter $\epsilon \in (0, \frac{1}{2}]$, the ϵ sequential private minimax risk (measured by the $\|\cdot\|_2^2$ metric) of the 1-sparse high dimensional sparse linear regression problem $\mathcal{P}_{1,p,2}$ needs to satisfy the following inequality,

$$\mathcal{M}_n^{\text{int}}(\theta(\mathcal{P}_{1,p,2}), \|\cdot\|_2^2, \epsilon) \geq \Omega(\min\{1, \frac{p}{n\epsilon^2}\}). \quad (3)$$

Remark 1: Since the lower bound of the non-private minimax risk is $O(\frac{\log p}{n})$ [36], we conjecture that the lower bound in Theorem 2 is not tight and the tightest bound should be $O(\frac{p \log p}{n\epsilon^2})$, which is the same as Theorem 1. Later, we will propose a near optimal algorithm (compared with (3)) in Section V-B and leave the problem of finding a tighter lower bound as future research.

Corollary 1: Recently, [38] proposed a general framework which could transfer any k-compositional *fully* interactive LDP algorithm to sequentially interactive LDP algorithm with an $O(k)$ blowup in the same complexity. Combining with Theorem 2, we can claim that even in the $O(p)$-compositional fully interactive LDP model, the dependence on the polynomial of the dimensionality p still cannot be avoided.

Remark 2: Recently [25] studied the lower bound of linear regression with statistical error in both low and high dimensional case under central (ϵ, δ)-DP model. Specifically, they show that for s-sparse high dimensional case, the private minimax risk under the ℓ_2 norm measurement is lower bound by $\Omega(\sqrt{\frac{s \log p}{n}} + \frac{s \log p \sqrt{\log 1/\delta}}{n\epsilon})$ while for the low dimensional case it is lower bounded by $\Omega(\sqrt{\frac{p}{n}} + \frac{p\sqrt{\log 1/\delta}}{n\epsilon})$, all of these bounds are optimal up to factors of $\text{Poly}(\log n)$. From Theorem 1 and 2, we can see that for sparse linear regression problem, LDP and DP are quite different.

Then, we investigate whether the loss function in the estimation error is too strong. For example, if let $\theta^* = e_j$ and the private estimator $\theta^{\text{priv}} = e_i$ for some $i \neq j$, then by the squared ℓ_2 norm loss, we have $\|\theta^{\text{priv}} - \theta^*\|_2^2 = 2$. Since it is possible to get $|\langle 1, \theta^{\text{priv}} - \theta^* \rangle| = 0$, this seems to suggest that relaxing the loss function could possibly lower the dependency on p. However, our next theorem gives a negative answer.

Theorem 3: Consider the loss function $L : \Theta \times \Theta \mapsto \mathbb{R}_+$, where $L(\theta, \theta') = |1^T(\theta - \theta')|$. Then, for any fixed $\epsilon \in (0, \frac{1}{2}]$, the ϵ sequential private minimax risk of the loss function L in the 1-sparse high dimensional sparse linear regression problem $\mathcal{P}_{1,p,2}$ needs to satisfy the following inequality,

$$\mathcal{M}_n^{\text{int}}(\theta(\mathcal{P}_{1,p,2}), L, \epsilon) \geq \Omega(\min\{1, \sqrt{\frac{p}{n\epsilon^2}}\}). \quad (4)$$

Finally, we consider the possibility of lowering the dependence of p by relaxing the definition of ϵ local differential privacy. This is motivated by the following fact in the central model, where there is a big difference between ϵ and (ϵ, δ)-differential privacy for a number of problems, such as the Empirical Risk Minimization [39] and the 1-way marginal [37]. However, as shown in a recent study [40], any non-interactive (ϵ, δ)-LDP protocol can be transformed to an ϵ-LDP protocol. This implies that relaxing to (ϵ, δ) LDP cannot avoid the polynomial dependence.

To further investigate the problem, we consider other types of relaxation for LDP, such as Local Rényi Differential Privacy

(LRDP) [41] and Local Zero-Concentrated Differential Privacy (LzCDP) [42]. The following theorem shows that the lower bounds on the minimax risk of the $(2, \log(1+\epsilon^2))$ sequential LRDP and (κ, ρ) sequential LzCDP still have polynomial dependence on p.

We first recall the definitions of Rényi Differential Privacy and Zero-Concentrated Differential Privacy and then extend them to the sequentially interactive model. For any $\alpha \geq 1$, we denote the Rényi divergence of distribution P and Q as

$$D_\alpha(P \| Q) = \frac{1}{\alpha - 1} \log \int (\frac{dP}{dQ})^\alpha dQ.$$

For $\alpha = 1$, it is just the KL-divergence.

Definition 3: Similar to the Definition of local differential privacy, a random variable Z_i is a (κ, ρ) locally zero-concentrated differentially private view of X_i if for all $\alpha > 1$, $z_1, z_2, \cdots, z_{i-1}$ and $x, x' \in \mathcal{X}$,

$$D_\alpha(Q_i(Z_i \in S \,|\, x_i, z_{1:i-1}) \| Q_i(Z_i \in S \mid x'_i, = z_{1:i-1})) \leq \kappa + \rho\alpha$$

holds for all events S. Similar to the locally differentially private case, we have (κ, ρ) local zero-concentrated differential privacy (LzCDP) and (κ, ρ) sequential zero-concentrated differential private minimax risk (sequential zCDP minimax risk).

Definition 4: Similarly, we have (α, ϵ) local Rényi differential privacy and (α, ϵ) (sequential) Renyi differential private minimax risk (called sequential RDP minimax risk) if

$$D_\alpha(Q_i(Z_i \in S \mid x_i, z_{1:i-1}) \| Q_i(Z_i \in S \mid x'_i, z_{1:i-1})) \leq \epsilon.$$

Theorem 4: For given fixed privacy parameters $0 < \epsilon \leq 1, \kappa, \rho > 0$, the (κ, ρ) sequential zCDP minimax risk (under the $\| \cdot \|_2^2$ metric) of the 1-sparse high dimensional sparse linear regression problem $\mathcal{P}_{1,p,2}$ needs to satisfy the following inequality,

$$\mathcal{M}_n^{\text{int}}(\theta(\mathcal{P}_{1,p,2}), \| \cdot \|_2^2, (\kappa, \rho)) \geq \Omega(\min\{1, \frac{p}{n(e^{\kappa + 2\rho} - 1)}\}).$$

The $(2, \log(1+\epsilon^2))$ sequential RDP minimax risk (under the $\| \cdot \|_2^2$ metric) of the 1-sparse high dimensional sparse linear regression problem $\mathcal{P}_{1,p,2}$ needs to satisfy :

$$\mathcal{M}_n^{\text{int}}(\theta(\mathcal{P}_{1,p,2}), \| \cdot \|_2^2, (2, \log(1+\epsilon^2))) \geq \Omega(\min\{1, \frac{p}{n\epsilon^2}\}).$$

B. Near Optimal Upper Bound for Sequential Interactive Local Model

With the understanding of the limitation in high dimensions, we focus, in this section, on the low dimensional sparse case (i.e., $n \geq \Omega(\frac{p}{\epsilon^2})$) and propose an ϵ sequential interactive LDP algorithm that achieves a near optimal upper bound on the estimation error (compared with (3)). Instead of considering the 1-sparse case as in Theorem 2, we study here the general case, that is, $\{(x_i, y_i)\}_{i=1}^n \sim P_{\theta^*, \sigma}$, where $P_{\theta^*, \sigma} \in \mathcal{P}_{s^*, p, C}$, and assume that some upper bound of s^* is already known.

Our method is called Locally Differentially Private Iterative Hard Thresholding (LDP-IHT), which is a locally differentially private version of the traditional Iterative Hard Thresholding method [43]. We consider the following more general optimization problem, with the intention to extend it to other problems (see Section VII),

$$\min L(\theta; D) = \frac{1}{2n} \sum_{i=1}^n (\langle x_i, \theta \rangle - y_i)^2$$
$$s.t. \|\theta\|_2 \leq 1, \|\theta\|_0 \leq s. \tag{5}$$

The key ideas for solving (5) in our Algorithm 1 are the follows. First, we partition the users into T groups $\{S_t\}_{t=1}^T$ (where the value of T will be specified later). Then, in the i-th iteration, each user receives the current estimator θ_{i-1}, and all users in group S_i conduct the ϵ-LDP randomizer procedure [10] on their current gradients $x_i^T(\langle x_i, \theta_{i-1} \rangle - y_i)$ (see below for the definition of the Randomizer). After receiving the noisy version of the gradient from each user, the server runs the iterative hard thresholding algorithm and produces a new estimator. That is, it executes first a gradient descent step, and then a truncation step $\theta'_{t+1} = \text{Trunc}(\tilde{\theta}_{t+1}, s)$, where the truncation function simply keeps the largest s entries of $\tilde{\theta}_{t+1}$ (in terms of the magnitude) and converts the rest of the entries to zero. This can be done by first sorting $\{|\tilde{\theta}_{t+1,j}|\}_{j=1}^p$, where $\tilde{\theta}_{t+1,j}$ is the j-th coordinate of the vector, then keeping the s-largest ones, and making the entries of all other coordinates 0. Finally, the algorithm projects θ'_{t+1} onto the unit ℓ_2 norm ball $\mathbb{B}_1$.

a) Randomizer $\mathcal{R}_\epsilon^r(\,\cdot\,)$ [10]: On input $x \in \mathbb{R}^p$, where $\|x\|_2 \leq r$, the randomizer $\mathcal{R}_\epsilon(x)$ does the following. It first sets $\tilde{x} = \frac{brx}{\|x\|_2}$ where $b \in \{-1, +1\}$ a Bernoulli random variable $\text{Ber}(\frac{1}{2} + \frac{\|x\|_2}{2r})$. We then sample $T \sim \text{Ber}(\frac{e^\epsilon}{e^\epsilon + 1})$ and outputs $O(r\sqrt{p})\mathcal{R}_\epsilon(x)$, where

$$\mathcal{R}_\epsilon(x) = \begin{cases} \text{Uni}(u \in \mathbb{S}^{p-1} : \langle u, \tilde{x} \rangle > 0) \text{ if } T = 1 \\ \text{Uni}(u \in \mathbb{S}^{p-1} : \langle u, \tilde{x} \rangle \leq 0) \text{ if } T = 0 \end{cases} \tag{6}$$

Using the same proof as in [21] we can show that each coordinate of the the randomizer $\mathcal{R}_\epsilon^r(x)$ is sub-Gaussian.

Lemma 1 [21]: Given any vector $x \in \mathbb{R}^p$, where $\|x\|_2 \leq r$, each coordinate of the randomizer $\mathcal{R}_\epsilon^r(x)$ defined above is a sub-Gaussian random vector with variance $\sigma^2 = O(\frac{r^2}{\epsilon^2})$ and $\mathbb{E}[\mathcal{R}_\epsilon(x)] = x$.

Before giving the theoretical analysis of Algorithm 1, we first show the assumption of the partitioned datasets $\{X_{S_t}\}_{t=1}^T$.

Assumption 1: $\{X_{S_t}\}_{t=1}^T$ satisfies the Restricted Isometry Property (RIP) with parameter $2s + s^*$, where $s = 8s^*$. That is, for any $v \in \mathbb{R}^p$ with $\|v\|_0 \leq 2s + s^*$, there exists a constant Δ which satisfies $(1 - \Delta)\|v\|^2 \leq \frac{1}{|S_t|}\|X_{S_t} v\|_2^2 \leq (1 + \Delta)\|v\|_2^2$ for any $t \in [T]$.

Note that for an $m \times p$ matrix $X = (x_1^T, \cdots, x_m^T)^T \sim \text{Uniform}\{+1, -1\}^{m \times p}$, it satisfies the RIP condition (with parameter s^*) with probability at least $1 - \epsilon$ if $m \geq c\Delta^{-2}(s^* \log p + \ln(1/\epsilon))$ for some universal constant c (see Theorem 2.12 in [44]). Thus, with probability at least $1 - \xi$, $\{X_{S_t}\}_{t=1}^T$ satisfies Assumption 1 if $n \geq \Omega(\Delta^{-2}(Ts^* \log p \log \frac{T}{\xi}))$. Later, we will see that $T = O(\log n)$. Thus, in order to ensure that Assumption 1 and $n \geq \Omega(\frac{p}{\epsilon^2})$ hold, we need to assume that $\frac{n}{\log n} \geq \Omega(\frac{ps^* \log p}{\epsilon^2})$.

Algorithm 1 LDP-IHT

```
Input: Private data records $\{(x_i, y_i)\}_{i=1}^n \sim P_{\theta^*,\sigma}$, where $P_{\theta^*,\sigma} \in \mathcal{P}_{s^*,p,C}$, iteration number $T$, privacy parameter $\epsilon$, step size $\eta$. Set $\theta_0 = 0$. $s = 8s^*$.
For $t = 1, \cdots, T$, define the index set $S_t = \{(t-1)\lfloor \frac{n}{T} \rfloor, \cdots, t\lfloor \frac{n}{T} \rfloor - 1\}$; if $t = T$, then $S_t = S_t \bigcup \{t\lfloor \frac{n}{T} \rfloor, \cdots, n\}$.
for $t = 1, 2, \cdots, T$ do
    The server sends $\theta_{t-1}$ to all the users. Every use $i$, $i \in S_t$, conducts the following operation: let $\nabla_i = x_i^T(\langle \theta_{t-1}, x_i \rangle - y_i)$, compute $z_i = \mathcal{R}_\epsilon^r(\nabla_i)$, where $\mathcal{R}_\epsilon^r$ is the randomizer defined above with $r = O(C\sqrt{p})$ and send back to the server.
    The server compute $\tilde{\nabla}_{t-1} = \frac{1}{|S_t|}\sum_{i \in S_t} z_i$.
    Perform the gradient descent updating $\tilde{\theta}_t = \theta_{t-1} - \eta\tilde{\nabla}_{t-1}$.
    $\theta'_t = \text{Trunc}(\tilde{\theta}_{t-1}, s)$.
    $\theta_t = \arg_{\theta \in \mathbb{B}_1} \|\theta - \theta'_t\|_2^2$.
end for
Return $\theta_T$
```

Theorem 5: For any $\epsilon > 0$, Algorithm 1 is ϵ sequentially interactive LDP. Moreover, under Assumption 1 with $\Delta = O(1)$ and $\frac{n}{\log n} \geq \Omega(\frac{ps^* \log p}{\epsilon^2})$, if $\{(x_i, y_i)\}_{i=1}^n \sim P_{\theta^*,\sigma}$, where $P_{\theta^*,\sigma} \in \mathcal{P}_{s^*,p,C}$, then by taking $s = 8s^*$ and $\eta = O(1)$, the output θ_T of the algorithm satisfies

$$\|\theta_T - \theta^*\|_2 \leq (\frac{1}{2})^T\|\theta^*\|_2 + O(\frac{C\sqrt{p\log p}\sqrt{T}\sqrt{s^*}}{\sqrt{n}\epsilon}), \tag{7}$$

with probability at least $1 - \frac{2T}{p^c}$ for some constant $c > 0$.

Note that Theorem 5 shows that if $s^* = 1$, $T = O(\log \frac{n\epsilon^2}{p \log p})$, then $\|\theta_T - \theta^*\|_2^2 = O(\frac{p \log p \log n}{n\epsilon^2})$. Compared with the lower bound in Theorem 2, it is an optimal upper bound up to a factor of $\sqrt{\log p}$.

We notice that recently [45] also used IHT to distributed DP-sparse PCA. However, compared with theirs, our method is ϵ-sequentially LDP while theirs is (ϵ, δ)-fully interactive LDP. Thus, the algorithms are quite different.

VI. Keeping the Responses Private

In this section, we consider a restricted case where only the responses or labels (*i.e.,* $\{y_i\}_{i=1}^n$) are required to be locally differentially private and all the observations $\{x_i\}_{i=1}^n$ are assumed to be public. Preserving the privacy of the labels has been studied in [14], [15] for private PAC learning. We also note that keeping the responses private is related to some issues of physical sensory data and the sparse recovery problem, which has been studied in [46]. In this case, we can actually assume that $\{x_i\}_{i=1}^n \sim \text{Uniform}(\{+1, -1\}^p)^n$ are public, and the collection of probability $\mathcal{P}_{s,p,C}$ in (1) is now reduced to the following model:

$$\mathcal{P}'_{s,p,C} = \{P_{\theta,\sigma}(y_1, \cdots, y_n) \mid y_i = \langle \theta^*, x_i \rangle + \sigma_i, \text{ where } \|\theta\|_0 \leq s, \|\theta\|_2 \leq 1 \text{ and the random noise } |\sigma_i| \leq C\}. \tag{8}$$

Algorithm 2 Label-LDP-IHT

```
Input: Public dataset $\{x_i\}_{i=1}^n$, private $\{y_i\}_{i=1}^n \in P_{\theta^*,\sigma}$, where $P_{\theta^*,\sigma} \in \mathcal{P}'_{s^*,p,C}$, $\epsilon, \delta$ are privacy parameters, $T$ is the number of iteration, $\eta$ is the step size, and $s = 8s^*$. Set $\theta_0 = 0$.
for Each $i \in [n]$ do
    Denote $\tilde{y}_i = y_i + z_i$, where $z_i \sim \mathcal{N}(0, \tau^2)$, $\tau^2 = \frac{32C^2 \ln(1.25/\delta)}{\epsilon^2}$.
end for
for $t = 0, 1, \cdots, T-1$ do
    $\tilde{\theta}_{t+1} = \theta_t - \eta(\frac{1}{n}\sum_{i=1}^n(\tilde{y}_i - \langle x_i, \theta_t \rangle)x_i^T)$.
    $\theta'_{t+1} = \text{Trunc}(\tilde{\theta}_{t+1}, s)$.
    $\theta_{t+1} = \arg_{\theta \in \mathbb{B}_1} \|\theta - \theta'_{t+1}\|_2^2$.
end for
Return $\theta_T$.
```

The following theorem shows that, for every set of data $\{(x_i, y_i)\}_{i=1}^n$, if only $\{y_i\}_{i=1}^n$ needs to be private, then there is an (ϵ, δ) non-interactively locally differentially private algorithm DP-IHT, which yields a non-trivial upper bound on the squared ℓ_2 norm of the estimation error (see Algorithm 2). More specifically, the algorithm first perturbs each y_i by Gaussian noise to ensure that it is (ϵ, δ)-LDP. Then, it performs the classical IHT procedure on the server side. Note that we can combine our algorithm with the protocol in [40] to obtain an ϵ non-interactive LDP algorithm.

Assumption 2: $X = (x_1^T, \cdots, x_n^T)^T \in \{-1, +1\}^{n \times p}$ satisfies the Restricted Isometry Property (RIP) with parameter $2s + s^*$, where $s = 8s^*$. That is, for any $v \in \mathbb{R}^p$ with $\|v\|_0 \leq 2s + s^*$, there exists a constant Δ which satisfies $(1 - \Delta)\|v\|^2 \leq \frac{1}{n}\|Xv\|_2^2 \leq (1 + \Delta)\|v\|_2^2$.

Theorem 6: For any $0 < \epsilon \leq 1$ and $0 < \delta < 1$, Algorithm 2 is (ϵ, δ) (non-interactively) locally differentially private for $\{y_i\}_{i=1}^n$. Moreover, if X satisfies Assumption 2 with $0 < \Delta \leq \frac{2}{7}$, then by setting $s = 8s^*$ in Algorithm 2, there is an $\eta = \eta(\Delta)$ which ensures that the output θ_T satisfies the following inequality with probability at least $1 - \exp(-n) - \frac{2}{p^c}$

$$\|\theta_T - \theta^*\|_2 \leq (\frac{1}{2})^T\|\theta^*\|_2 + O(\frac{C\log(1/\delta)\sqrt{s^* \log p}}{\sqrt{n}\epsilon}). \tag{9}$$

Note that if $T = O(\log \frac{\sqrt{n}\epsilon}{C\sqrt{s^* \log p}})$ in (9), we have $\|\theta_T - \theta^*\|_2^2 \leq O(C^2 \frac{s \log p}{n\epsilon^2})$. Compared with the bounds in Theorem 1 and 2, the dependency on p is reduced from polynomial to logarithmic, which makes it suitable for handling high dimensional data. We note that the term $O(\frac{s \log p}{n})$ also appears in the optimal minimax rate of the high dimensional sparse sub-Gaussian linear model [36].

Also note that after obtaining $\{(x_i, \tilde{y}_i)\}_{i=1}^n$, we can get another private estimator, which has the same upper bound of $O(\frac{s \log p}{n\epsilon^2})$, by performing Lasso $\theta^{\text{priv}} \in \arg_{\theta \in \mathbb{R}^p}\{\frac{1}{2n}\sum_{i=1}^n(\tilde{y}_i - \langle \theta, x_i \rangle)^2 + \lambda\|\theta\|_1\}$, for some $\lambda = O(\sqrt{\frac{\log p}{n\epsilon^2}})$ [47]. However, we would like to point out that our algorithm is more practical and can be extended to the case of non-linear measurements.

With the above theorem, a natural question is to determine whether the upper bound in Theorem 6 can be further improved. The following theorem (adopted from [36]) suggests that it is actually tight as the ϵ non-interactive local private minimax risk (under the $\|\cdot\|^2$ metric) is lower bounded by $\Omega(\frac{C^2 s^* \log p}{n\epsilon^2})$.

Theorem 7: Under Assumption 2 and for a given fixed privacy parameter $\epsilon \in (0, \frac{1}{2}]$, the ϵ non-interactive local private minimax risk (under the $\|\cdot\|^2$ metric) satisfies the following inequality if only $\{y_i\}_{i=1}^n$ needs to be kept locally private

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}'_{s,p,C}), \|\cdot\|_2^2, \epsilon) \geq \Omega\big(\min\{1, \frac{C^2 s \log \frac{p}{s}}{n\epsilon^2(1+\Delta)}\}\big).$$

VII. Extension to Other Problems

As mentioned earlier, the (Local) DP-IHT method is actually quite general for achieving differential privacy. In this section, we extend it to other problems. Specifically, we use it to the DP-ERM problem [1] under some sparsity constraint and the sparse regression problem with non-linear monotone measurements.

A. ERM With Sparsity Constraint

We start with reviewing some definitions of DP-ERM.

Definition 5 (DP-ERM [48]): Given a dataset $D = \{z_1, \cdots, z_n\}$ from a data universe $\mathcal{X}$, a loss function $\ell(\cdot, \cdot)$ and a constraint set $\mathcal{C} \subseteq \mathbb{R}^p$, DP-ERM is to find x^{priv} so as to minimize the empirical risk, *i.e.* $L(x; D) = \frac{1}{n}\sum_{i=1}^n \ell(x, z_i)$ with the guarantee of being differentially private [5]. The utility of the algorithm is measured by the expected excess empirical risk, that is $\mathbb{E}_{\mathcal{A}}[L(x^{\text{priv}}; D)] - \min_{x\in\mathcal{C}} L(x; D)$, where the expectation of $\mathcal{A}$ is taking over all the randomness of the algorithm.

Here, in the Differential Privacy (DP) model, data are managed by a trusted central entity which is responsible for collecting them and for deciding which differentially private data analysis to perform and to release.

Definition 6 (Differential Privacy [5]): Given a data universe $\mathcal{X}$, we say that two datasets $D, D' \subseteq \mathcal{X}$ are neighbors if they differ by only one entry, which is denoted as $D \sim D'$. A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private (DP) if for all neighboring datasets D, D' and for all events S in the output space of $\mathcal{A}$, the following holds

$$\mathbb{P}(\mathcal{A}(D) \in S) \leq e^{\epsilon}\mathbb{P}(\mathcal{A}(D') \in S) + \delta.$$

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

In this section, we consider the sparsity-constrained (ϵ, δ) DP-ERM problem. That is, the constraint set $\mathcal{C}$ is defined as $\mathcal{C} = \{x : \|x\|_0 \leq k\}$, where $\|x\|_0$ denotes the number of non-zero entries in vector x. We note that such a formulation encapsulates several important problems such as the ℓ_0-constrained linear/logistic regression [49].

We first introduce some assumptions to the loss function, which are commonly used in the research of ERM under the sparsity-constrained optimization.

Algorithm 3 DP-IHT

```
Input: Initial point $x_0$, learning rate $\eta$, empirical risk $L(x; D)$, privacy parameters $1 > \epsilon, \delta > 0$, and iteration number $T$.
for $t = 0, 1, \cdots, T-1$ do
    Let $\tilde{x}_{t+1} = x_t - \eta(\nabla L(x_t; D) + z_t)$, where $z_t \sim \mathcal{N}(0, \sigma^2 I_p)$, $\sigma^2 = \frac{cT \log\frac{1}{\delta} G^2}{n^2\epsilon^2}$ for some constant $c$.
    Let $x_{t+1} = \text{Trun}(\tilde{x}_{t+1}, k)$.
end for
Return $x_T$.
```

Definition 7 (Restricted Strong Convexity, RSC): A differentiable function $f(x)$ is restricted ρ_s-strongly convex with parameter s if there exists a constant $\rho_s > 0$ such that for any x, x' with $\|x - x'\|_0 \leq s$, we have

$$f(x) - f(x') - \langle \nabla f(x'), x - x' \rangle \geq \frac{\rho_s}{2}\|x - x'\|_2^2.$$

Definition 8 (Restricted Strong Smoothness, RSS): A differentiable function $f(x)$ is restricted ℓ_s-strong smooth with parameter s if there exists a constant $\ell_s > 0$ such that for any x, x' with $\|x - x'\|_0 \leq s$, we have

$$f(x) - f(x') - \langle \nabla f(x'), x - x' \rangle \leq \frac{\ell_s}{2}\|x - x'\|_2^2.$$

Assumption 3: Denote $x^* = \arg\min_{x\in\mathcal{C}} L(x; D)$ and $\|x^*\|_0 = k^*$. We assume that the objective function $L(x; D)$ is ρ_s-RSC and $\ell(x, z)$ is ℓ_s-RSS for all $z \in \mathcal{X}$ with parameter $s = 2k + k^*$. We also assume that $\ell(x, z)$ is G-Lipshitz w.r.t ℓ_2 norm for all $z \in \mathcal{X}$.

For the sparsity-constrained DP-ERM problem, we follow the idea in Algorithm 1 to solve the optimization problem (5). That is, we first execute a DP-Gradient Descent step and then perform a hard thresholding operation (see Algorithm 3 for details).

Theorem 8: Under Assumption 3, for any $1 \geq \epsilon, \delta > 0$, there exists a constant $c > 0$ which makes Algorithm 3 (ϵ, δ)-DP. Moreover, if the sparsity level $k \geq (1 + 64\kappa_s^2)k^*$, where $\kappa_s = \frac{\ell_s}{\rho_s}$, then by setting $\eta = \frac{1}{2\ell_s}$ and $T = O(\kappa_s \log\frac{n^2\epsilon^2}{k^*})$, we have

$$\mathbb{E}L(x_T; D) - L(x^*; D) \leq O(\frac{\log n \log pk^* \log\frac{1}{\delta}}{n^2\epsilon^2}), \qquad (10)$$

where the big O-notation omits the terms of G, ρ_s and ℓ_s.

Remark 3: We note that the upper bound in (10) depends only logarithmically on p (i.e., $logp$), rather than polynomially (i.e., Poly(p)) as in general DP-ERM with (strongly) convex loss functions [31], [48]. This means that we have obtained a non-trivial upper bound for the high dimensional case ($p \gg n$) of the problem. Recently, [23], [50] also studied the case of high dimensional DP-ERM with specified constraint set. However, there are considerable differences. Firstly, the [23] paper considers only linear regression and ℓ_1-norm Lipshitz with the constraint set restricted to an ℓ_1-norm ball. Secondly, the [50] paper shows that its upper bound depends only on the Gaussian width of the underlying constraint set, which could has sub-linear dependence on p (e.g., for the case of

[1] It is easy to extend to LDP model

the unit ℓ_1-norm ball, it is logarithmic in p). However, their algorithm is based on the mirror descent method, which needs the constraint set to be convex. But it is non-convex in our problem. Thus, these previous results are not comparable with ours.

It would be interesting to find a general condition on the constraint set such that the upper bound of the problem can be independent of Poly(p). Also, we note that to achieve the bound in (10), the gradient complexity of Algorithm 3 needs to be $\tilde{O}(n\kappa_s)$, which is quite large. We leave it as an open problem to make it more practical.

B. Non-Linear Regression

We now study a model with non-linear non-convex measurement: $y_i = f(\langle \theta^*, x_i \rangle) + \sigma$, where f is some known function and θ^* is sparse. This model has recently been studied in [12], [13]. Note that when f is the identity function, it reduces to the sparse linear regression model. In this paper, we focus on a special class of functions called (a, b) monotone:

Definition 9: A function $f : \mathbb{R} \mapsto \mathbb{R}$ is (a, b) monotone for some $0 < a \leq b$ if f is differentiable and $f'(x) \in [a, b]$ for all $x \in \mathbb{R}$.

Like in the linear model, we also consider the cases of keeping the whole dataset and only the responses $\{y_i\}_{i=1}^n$ locally differentially private.

1) Keeping the Whole Dataset Private: Same as in the linear model case, we consider the following distribution collection of samples $(x, y) \in \{+1, -1\}^p \times \mathbb{R}$:

$$\begin{aligned}\mathcal{P}_{s,p,C,f,a,b} = \{P_{\theta,\sigma} \,|\, x \sim \text{Uniform}\{+1,-1\}^p, y = f(\langle \theta, x \rangle) + \sigma,\\ \text{where } \sigma \text{ is the random noise } |\sigma| \leq C, C > 0\\ \text{is some constant } \|\theta\|_2 \leq 1, \|\theta\|_0 \leq s, f \text{ is } (a, b) \text{ monotone } \}. \end{aligned} \tag{11}$$

We note that when $f(x) = x$, it reduces to (1).

To obtain an upper bound of the empirical risk, we can easily extend Algorithm 1 to the non-linear measurement case (see Algorithm 4) to solve the following problem

$$\begin{aligned}\min L(\theta; D) = \frac{1}{n}\sum_{i=1}^{n}(f(\langle x_i, \theta \rangle) - y_i)^2\\ s.t. \|\theta\|_2 \leq 1, \|\theta\|_0 \leq s. \end{aligned} \tag{12}$$

Theorem 9: For any $\epsilon > 0$, Algorithm 4 is ϵ sequential interactive LDP. Moreover, if $\{X_{S_t}\}$ satisfies Assumption 1 with $0 \leq \delta' \leq \frac{9a^2-5b^2}{14}$ in Section 4.2 and $\frac{n}{\log n} \geq \Omega(\frac{ps^* \log p}{\epsilon^2})$, and $\{(x_i, y_i)\}_{i=1}^n \sim P_{\theta^*,\sigma}$, where $P_{\theta^*,\sigma} \in \mathcal{P}_{s^*,p,C,f,a,b}$ (we assume $\frac{a^2}{b^2} \geq \frac{5}{9}$), then after taking $s = 8s^*$ and $\eta = \eta(a, b)$, the output θ_T satisfies

$$\|\theta_T - \theta^*\|_2 \leq (\frac{1}{2})^T \|\theta^*\|_2 + O(\frac{\sqrt{p\log p}\sqrt{T}\sqrt{s}}{\sqrt{n}\epsilon}), \tag{13}$$

with probability at least $1 - \frac{2T}{p^c}$ for some constant $c > 0$.

Algorithm 4 LDP-IHT

Input: Private data records $\{(x_i, y_i)\}_{i=1}^n \sim P_{\theta^*,\sigma}$, where $P_{\theta^*,\sigma} \in \mathcal{P}_{s,p,C,f,a,b}$, T is the Iteration number, ϵ is the privacy parameter, and η is the step size. Set $\theta_0 = 0$. s is a parameter to be specified later.

```
For $t = 1, \cdots, T$, define the index set $S_t = \{(t-1)\lfloor \frac{n}{T} \rfloor, \cdots, t\lfloor \frac{n}{T} \rfloor - 1\}$, if $t = T$, then $S_t = S_t \bigcup \{t\lfloor \frac{n}{T} \rfloor, \cdots, n\}$.
**for** $t = 1, 2, \cdots, T$ **do**
    The server sends $\theta_{t-1}$ to all the users. Every use $i$ which $i \in S_t$ does the following operation: let $\nabla_i = x_i^T f'(\langle \theta_{t-1}, x_i \rangle)(f(\langle \theta_{t-1}, x_i \rangle) - y_i)$, compute $z_i = \mathcal{R}_\epsilon^r(\nabla_i)$, where $\mathcal{R}_\epsilon^r$ is the randomizer defined in the previous section with $r = O(bC\sqrt{p})$ and send back to the server.
    The server compute $\tilde{\nabla}_{t-1} = \frac{1}{|S_t|}\sum_{i \in S_t} z_i$.
    Do the gradient descent updating $\tilde{\theta}_t = \theta_{t-1} - \eta\tilde{\nabla}_{t-1}$.
    $\theta'_t = \text{Trunc}(\tilde{\theta}_t, s)$.
    $\theta_t = \arg_{\theta \in \mathbb{B}_1} \|\theta - \theta'_t\|_2^2$.
**end for**
Return $\theta_T$
```

2) Keeping the Labels Private: For a fixed $X = (x_1^T, \cdots, x_n^T)^T \in \{+1, -1\}^{n \times p}$, we consider the following collection of distributions:

$$\begin{aligned}\mathcal{P}'_{s,p,C,f,a,b} = \{P_{\theta,\sigma}(\{y_i\}_{i=1}^n) \mid y_i = f(\langle \theta^*, x_i \rangle) + \sigma_i,\\ \text{where } \|\theta\|_0 \leq s, \|\theta\|_2 \leq 1, \text{ the random noise}\\ |\sigma_i| \leq C \text{ for some constant } C > 0, \text{and } f \text{ is } (a, b) \text{ monotone}\}.\end{aligned}$$

The following theorem shows the lower bound of the private minimax risk (under the $\| \cdot \|_2^2$ metric) with respect to the above collection of distributions, which is similar to the one in Theorem 6.

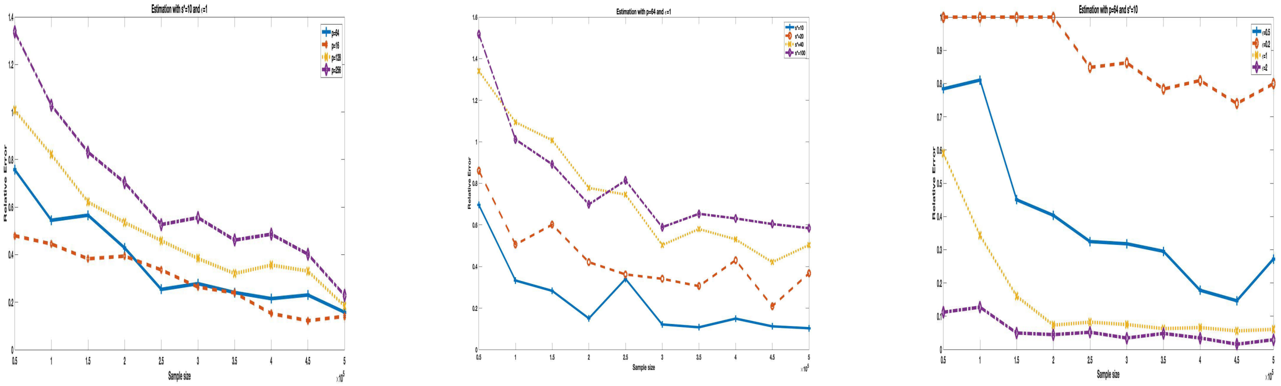
Theorem 10: Under Assumption 2 and for a given fixed privacy parameter $\epsilon \in (0, \frac{1}{2}]$, the ϵ non-interactive local private minimax risk (under the $\| \cdot \|^2$ metric) in the case of keeping $\{y_i\}_{i=1}^n$ locally private satisfies the following inequality

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}'_{s,p,C,f,a,b}), \|\cdot\|_2^2, \epsilon) \geq \Omega(\min\{1, C^2 \frac{s \log \frac{p}{s}}{nb^2\epsilon^2(1+\Delta)}\}).$$

Comparing to the lower bound in Theorem 6 in the previous section, we can see that there is an additional factor of b^2 in Theorem 10, which is due to the fact that the model is more complicated.

For the upper bound, we adopt a similar approach as in DP-IHT for linear regression. Particularly, we let $L(\theta) = \frac{1}{2n}\sum_{i=1}^n (\tilde{y}_i - \langle x_i, \theta \rangle)^2$ and then apply the ideas of IHT.

Theorem 11: For any $0 < \epsilon \leq 1$ and $0 < \delta < 1$, Algorithm 5 is (ϵ, δ) (non-interactively) locally differentially private for $\{y_i\}_{i=1}^n$. Moreover, if $\{y_i\}_{i=1}^n \in P_{\theta^*,\sigma}$ (where $P_{\theta^*,\sigma} \in \mathcal{P}'_{s^*,p,C,f,a,b}$ with $1 \geq \frac{a}{b} > \frac{\sqrt{5}}{3}$) and X satisfies Assumption 1 with $0 < \Delta \leq \frac{9a^2-5b^2}{14}$, then by setting $s = 8s^*$ in Algorithm 5, there is an $\eta = \eta(\Delta)$ which ensures that the

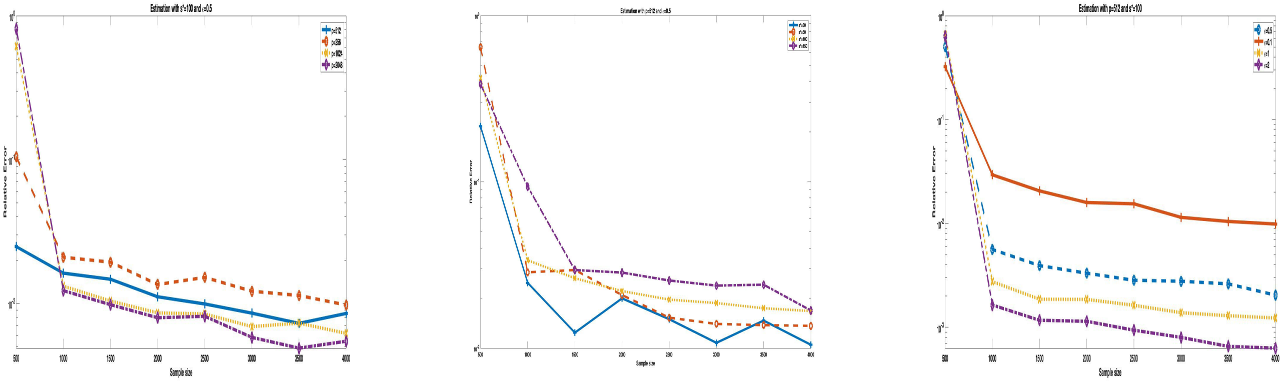


(a) Relative error w.r.t dimensionality

(b) Relative error w.r.t sparsity level

(c) Relative error w.r.t privacy level

Fig. 1. Experimental results on sparse linear regression under LDP while keeping the whole dataset private (Algorithm 1).



(a) Relative error w.r.t dimensionality

(b) Relative error w.r.t sparsity level

(c) Relative error w.r.t privacy level

Fig. 2. Experimental results on sparse linear regression under LDP while keeping the labels private (Algorithm 2).

Algorithm 5 General DP-Iterative Hard Thresholding

```
Input: Public dataset $\{x_i\}_{i=1}^n$, private $\{y_i\}_{i=1}^n \in P_{\theta^*,\sigma}$, where $P_{\theta^*,\sigma} \in \mathcal{P}_{s^*,p,C,f,a,b}$, $\epsilon, \delta$ are privacy parameters, $T$ is the number of iteration, $\eta$ is the step size, and $s$ is a parameter to be specified. Set $\theta_0 = 0$.
for Each $i \in [n]$ do
    Denote $\tilde{y}_i = y_i + z_i$, where $z_i \sim \mathcal{N}(0, \tau^2)$, $\tau^2 = \frac{32C^2 \ln(1.25/\delta)}{\epsilon^2}$.
end for
for $t = 0, 1, \cdots, T-1$ do
    $\tilde{\theta}_{t+1} = \theta_t - \eta \nabla L(\theta_t)$.
    $\theta'_{t+1} = \text{Trunc}(\tilde{\theta}_{t+1}, s)$.
    $\theta_{t+1} = \arg_{\theta \in \mathbb{B}_1} \|\theta - \theta'_{t+1}\|_2^2$.
end for
Return $\theta_T$.
```

output θ_T satisfies the following inequality

$$\|\theta_T - \theta^*\|_2 \leq (\frac{1}{2})^T \|\theta^*\|_2 + O(\frac{bC\log(1/\delta)\sqrt{s^* \log p}}{\sqrt{n}\epsilon}),$$

with probability at least $1 - T\exp(-n) - \frac{2T}{p^c}$.

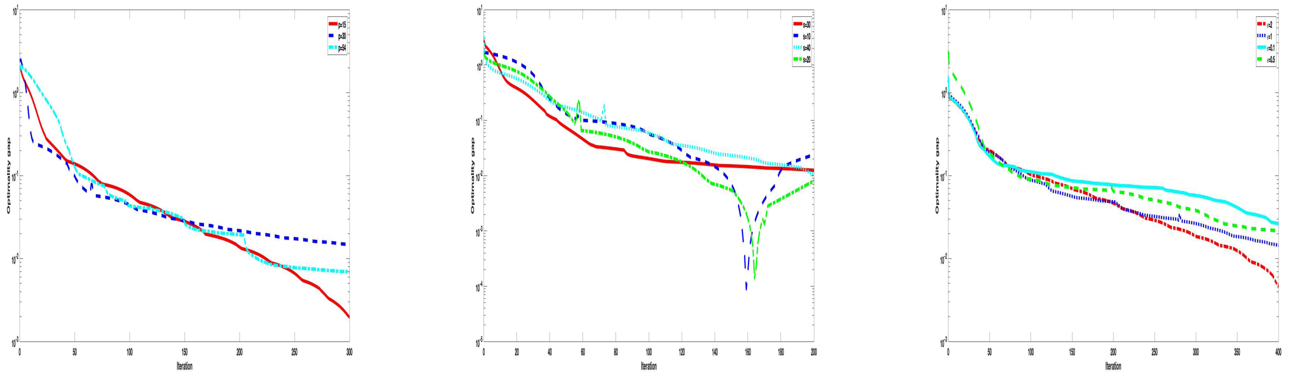
VIII. Experiments

A. Experiments on Sparse Linear Regression

a) Data Generation: Our data generation process is similar to the one in [32]. We first fix a parameter vector θ^* by randomly choosing s^* coordinates, with each of them sampled independently from a uniform distribution in interval $[0, 1]$, and setting the remaining coordinates/entries to zero. Then, we generate the data samples using equation $y_i = \langle x_i, \theta^* \rangle + \sigma_i$, where $x_i \in \text{Uniform}\{-1, +1\}^p$ and $\sigma_i \in \text{Uniform}[-C, C]$. We assume $C = 0.05$ in our experiment.

b) Experiment Results: We compare the relative error, *i.e.* $\frac{\|\theta_T - \theta^*\|_2}{\|\theta^*\|_2}$, with the sample size n in three different settings, *i.e.,* under varying dimensionality, sparsity and privacy level, respectively. We run algorithms Label-LDP-IHT with $\eta = 0.2$ or $\eta = 0.1$, $s = s^*$, $T = \lceil \log \frac{n}{p} \rceil$, $\delta = 10^{-3}$ and a random normal Gaussian vector as the initial point to obtain θ_T. For each experiment, we run the algorithm 10 times and take the one with the lowst relative error as the final value.

Figure 1 and 2 depict the results of Algorithm 1 and 2, respectively. From Figure 1, we can see that when the dimensionality and the sparsity level increase or the privacy parameter ϵ decreases, the relative error increases, especially when

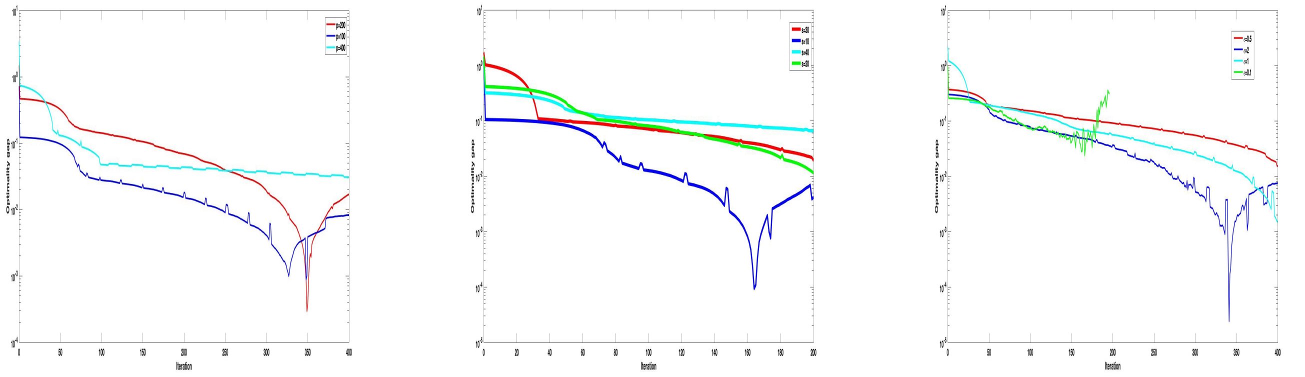


(a) Optimality gap w.r.t dimensionality with fixed $s = 10$ and $\epsilon = 2$.

(b) Optimality gap w.r.t sparsity level with fixed $p = 54$ and $\epsilon = 2$

(c) Optimality gap w.r.t privacy level with fixed $p = 54$ and $s = 10$

Fig. 3. Experimental results on Covertype dataset [52] for ℓ_0-constrained logistic regression under (ϵ, δ)-DP (Algorithm 3).

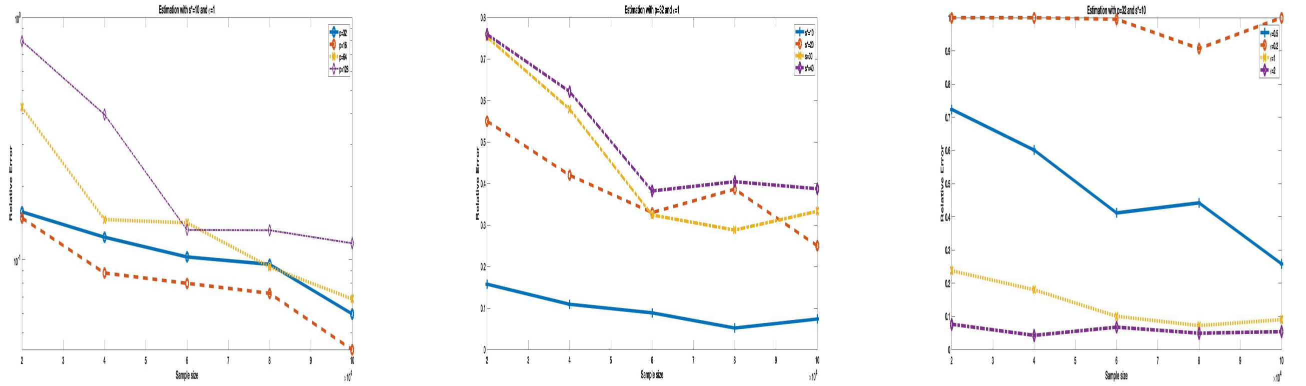


(a) Optimality gap vs dimensionality with fixed $s = 10$ and $\epsilon = 2$.

(b) Optimality gap vs sparsity level with fixed $p = 200$ and $\epsilon = 2$

(c) Optimality gap vs privacy level with fixed $p = 200$ and $s = 10$

Fig. 4. Experimental results on rcv1 dataset [51] for ℓ_0-constrained logistic regression under (ϵ, δ)-DP (Algorithm 3).



(a) Relative error vs dimensionality.

(b) Relative error vs sparsity level.

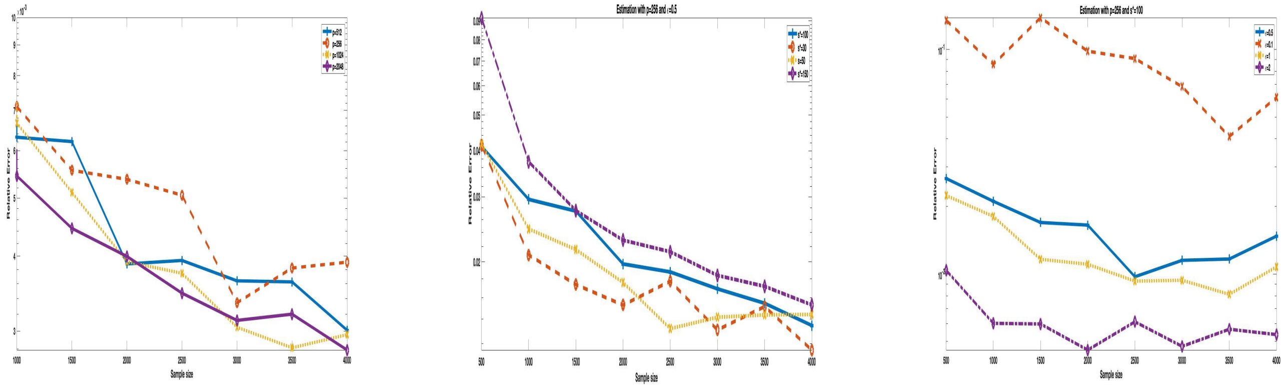
(c) Relative error vs privacy level.

Fig. 5. Experimental results for sparse regression with non-linear measurement under LDP when keeping the whole dataset private (Algorithm 4).

the sample size n is small. When the sample size increases, the relative error will decreases. From Figure 2, we can learn that when the dimensionality p increases, unlike Figure 1, it does not cause the relative error to change significantly. This can be explained by the fact that the error bound is only logarithmically depending on p. Moreover, when the privacy parameter increases, the relative error decreases. These results confirm our theoretical claims.

B. Experiments on Sparsity-Constrained DP-ERM

In this section, we test Algorithm 3 on real world datasets Covertype and rcv1 [51]. Particularly, we study



(a) Relative error vs dimensionality.

(b) Relative error vs sparsity level.

(c) Relative error vs privacy level.

Fig. 6. Experimental results for sparse regression with non-linear measurement under LDP when keeping the label private (Algorithm 5).

the sparsity-constrained logistic regression problem with $\ell(w,z) = \log(1+\exp(-y_i\langle w, x_i\rangle)) + \frac{\lambda}{2}\|w\|_2^2$, where y_i is the label of x_i. As pre-processing, the data is first normalized. Since there is no ground truth on real data, we run the algorithm in [32] sufficiently long until $\|w_t - w_{t+1}\|_2/\|w_t\|_2 \leq 10^{-4}$ and then use the output $L(w_t; D)$ as the approximate optimal value. With this, we can calculate the optimality gap of our estimator. In the experiments, we set $\lambda = 10^{-3}$, $\eta = 0.1$ and $\delta = 10^{-3}$, and use zCDP [42] to achieve the (ϵ, δ)-DP.

From Figure 3 and 4, we can see that when the dimensionality p increases, the optimality gap does not change too much, which is due to the fact that the error bound is only logarithmically depending on p. Also, when the sparsity level increases or ϵ decreases, the optimality gap increases. Clearly, all these experimental results are consistent with Theorem 8.

C. Tests on Synthetic Datasets for Linear Regression With Non-Linear Measurements

Our data generation process is similar to the one in [32]. We first fix a parameter vector θ^* by randomly choosing s^* coordinates, with each of them sampled independently from a uniform distribution in interval $[0,1]$, and setting the remaining coordinates/entries to zero. For the case of non-linear measurements, we assume that $y_i = f(\langle x_i, \theta^*\rangle) + \sigma_i$, where $f(x) := 8x + \cos x$ where $x_i \in \text{Uniform}\{-1,+1\}^p$ and $\sigma_i \in \text{Uniform}[-C, C]$ so that it satisfies the assumptions in Theorem 9. The results are shown in Figure 6 and 5. We can see that these results are almost the same as in Figure 1 and 2, respectively.

IX. Conclusion

In this paper, we comprehensively studied the sparse linear regression problem in the non-interactive and sequential interactive local differential privacy models. Specifically, we first showed that polynomial dependency on the dimensionality p of the space is unavoidable for the estimation error in both non-interactive and sequential interactive local models if the privacy of the whole dataset needs to be preserved, even if we allow relaxed privacy models and relaxed measurements of error. This is quite different compared with both of the non-private case and the problem in the central (ϵ, δ) Differential Privacy model. However, in a restricted (high dimensional) case where only the privacy of the responses (labels) needs to be preserved. We showed that the optimal rate of the error estimation can be made logarithmically dependent on p (i.e., $\log p$) in the local model, which is quite similar as in the non-private case. Second, we proposed a general method which is called Differentially Private Iterative Hard Thresholding whose output achieve an optimal rate up to a $\sqrt{\log n}$ factor. Moreover, we used this method to solve some other problems, such as (Local) DP-ERM with sparsity constraints and sparse regression with non-linear measurements.

Appendix

A. Technical Lemmas

For the estimation error, we first give some definitions and lemmas.

Definition 10: A random variable X is said to be sub-Gaussian with σ^2 if $\mathbb{E}(X) = 0$ and

$$\mathbb{E}[\exp(sX)] \leq \exp(\frac{\sigma^2 s^2}{2}), \forall s \in \mathbb{R}.$$

For the case that X is a d-dimensional random vector, it is sub-Gaussian with σ^2 if for any unit vector $u \in \mathbb{S}^{d-1}$, $u^T X$ is sub-Gaussian with σ^2.

It is well known that if $X_1, X_2, \cdots, X_n$ are all sub-Gaussian with σ^2, then $a_1 X_1 + \cdots + a_n X_n$ is sub-Gaussian with $(\sum_{i=1}^n a_i^2)\sigma^2$.

We can easily see that if $x \sim \text{Uniform}\{+1,-1\}^d$, x is sub-Gaussian with $\sigma^2 = 1$.

Lemma 2 [53]: Let $X_1, X_2, \cdots, X_n$ be n random variables such that each X_i is sub-Gaussian with σ^2. Then the following holds

$$\Pr[\max_{i\in n} X_i \geq t] \leq n e^{-\frac{t^2}{2\sigma^2}},$$

$$\Pr[\max_{i\in n} |X_i| \geq t] \leq 2n e^{-\frac{t^2}{2\sigma^2}}.$$

Lemma 3 [32]: For any $\theta \in \mathbb{R}^k$ and an integer $s \leq k$, if $\theta_t = \text{Trunc}(\theta, s)$ then for any $\theta^* \in \mathbb{R}^k$ with $\|\theta^*\|_0 \leq s$, we have $\|\theta_t - \theta\|_2 \leq \frac{k-s}{k-s^*}|\theta^* - \theta\|_2^2$.

Lemma 4: Let $\mathcal{K}$ be a convex body in $\mathbb{R}^p$, and $v \in \mathbb{R}^p$. Then for every $u \in \mathcal{K}$, we have

$$\|\mathcal{P}_{\mathcal{K}}(v) - u\|_2 \leq \|v - u\|_2,$$

where $\mathcal{P}_{\mathcal{K}}$ is the operator of projection onto $\mathcal{K}$.

The following theorem says that when $X \in \text{Uniform}\{+1,-1\}^{n\times p}$, with high probability it satisfies the Restricted Isometry Property if n is sufficiently large.

Lemma 5 (Theorem 2.12 in [44]): Let $X \in \{+1,-1\}^{n\times p}$ be a Bernoulli Random Matrix and $\xi, \Delta \in (0,1)$. Assume that

$$n \geq C\Delta^{-2}(s\log(p/s) + \log(1/\xi)).$$

Then with probability at least $1-\xi$, X satisfies the Restricted Isometry Property (RIP) with sparsity level s and parameter Δ, that is, for every $\|v\|_0 \leq s$,

$$(1-\Delta)\|v\|^2 \leq \frac{1}{n}\|Xv\|_2^2 \leq (1+\Delta)\|v\|_2^2.$$

Note that if X satisfies the Restricted Isometry Property (RIP) with sparsity level s and parameter Δ, it means that

$$\Delta = \max_{\|x\|_2=1, \|x\|_0 \leq s} \|(\frac{1}{n}X^TX - I_{p\times p})x\|_2.$$

Lemma 6 [54]: If $z \sim \chi_n^2$, where χ_n^2 is the Chi-square distribution with parameter n, then

$$\Pr[z - n \geq 2\sqrt{nx} + 2x] \leq \exp(-x).$$

B. Private Fano and Le Cam Method

Our lower bounds are basic on the locally private version Fano and Le Cam method [10], [11]. Given a finite set $\mathcal{V}$, a family of distributions $\{P_v, v \in \mathcal{V}\}$ with $P_v \in \mathcal{P}$ is 2δ-separated in a metric ρ if $\rho(\theta(P_v), \theta(P_{v'})) \geq 2\delta$ for all distinct pairs $v, v' \in \mathcal{V}$. Given any 2δ-separated set, the private Fano's method for the ϵ non-interactive private minimax risk can be summarized by the following lemma.

Lemma 7 (Prop. 2 in [10]): Given any 2δ-separated set $\{P_v, v \in \mathcal{V}\}$, and $\alpha \in (0, \frac{1}{2}]$, the ϵ non-interactive private minimax risk satisfies the following inequality

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi\circ\rho, \epsilon) \geq \frac{\Phi(\delta)}{2}\big(1 - \frac{n\alpha^2\mathcal{C}_\infty^{Nint}(\{P_v\}_{v\in\mathcal{V}}) + \log 2}{\log|\mathcal{V}|}\big),$$

where $\mathcal{C}_\infty^{Nint}(\{P_v\}_{v\in\mathcal{V}}) = \frac{1}{|\mathcal{V}|}\sup_{\gamma\in\mathbb{B}_\infty}\sum_{v\in\mathcal{V}}(\psi_v(\gamma))^2$, $\mathbb{B}_\infty$ is the 1-ball of the supremum norm $\mathbb{B}_\infty = \{\gamma \in L^\infty(\mathcal{X}) \mid \|\gamma\|_\infty \leq 1\}$, and $L^\infty(\mathcal{X}) = \{f : \mathcal{X} \mapsto \mathbb{R} \mid \|f\|_\infty < \infty\}$ is the space of uniformly bounded functions with the supremum norm $\|f\|_\infty = \sup_x |f(x)|$. Also, for each $v \in \mathcal{V}$, $\psi_v : L^\infty(\mathcal{X}) \mapsto \mathbb{R}$ is a linear function defined by

$$\psi_v(\gamma) = \int_{\mathcal{X}} \gamma(x)dP_v(x) - d\bar{P}(x),$$

where $\bar{P}$ is the mixture distribution $\bar{P} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} P_v^n$.

A useful corollary is the following:

Lemma 8 (Corollaries 2 and 4 in [9]): Let V be randomly and uniformly distributed in $\mathcal{V}$. Assume that given $V = v$, X_i is sampled independently according to the distribution of $P_{v,i}$ for $i = 1, \cdots, n$. Then, there is a universal constant $c < 19$ such that for $\alpha \in (0, \frac{1}{2}]$,

$$I(Z_1, Z_2, \cdots, Z_n; V) \leq c\epsilon^2 \sum_{i=1}^n \frac{1}{|\mathcal{V}|^2} \sum_{v,v'\in\mathcal{V}} \|P_{v,i} - P_{v',i}\|_{TV}^2.$$

The ϵ non-interactive private minimax risk satisfies

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi\circ\rho, \epsilon) \geq \frac{\Phi(\delta)}{2}\big(1 - \frac{I(Z_1, \cdots, Z_n; V) + \log 2}{\log|\mathcal{V}|}\big).$$

Now we introduce the generalized private Le Cam method. Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be two collections of distributions in $\mathcal{P}$. We say that $\mathcal{P}_0$ and $\mathcal{P}_1$ are δ-separated for loss function L if $d_L(P_0, P_1) \geq \delta$ for all $P_0 \in \mathcal{P}_0$ and $P_1 \in \mathcal{P}_0$, where $d_L(P_0, P_1) = \inf_{\theta\in\Theta}\{L(\theta, \theta(P_0)) + L(\theta, \theta(P_1))\}$. Then we have the following lemma.

Lemma 9 (Theorem 2 in [11]): Consider a set of distributions $\mathcal{P}$, a collection of distributions on $\mathcal{X}$, $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$, indexed by $v \in \mathcal{V}$, as well as a distribution $P_0 \in \mathcal{P}$. For each of these distributions, we have i.i.d. observations X_i, that is, samples from the product with density

$$dP_v^n = \Pi_{i=1}^n dP_v(x_i).$$

We also define the marginal distributions $M_v^n(\cdot) = \int Q(\cdot|x_{1:n})dP_v^n(x_{1:n})$ and $\bar{M}^n = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} M_v^n$, where Q is a private channel. For any $\epsilon \in (0, \frac{1}{2}]$, the ϵ sequential private minimax risk in the loss function L satisfies the following inequality

$$\mathcal{M}_n^{\text{Int}}(\theta(\mathcal{P}), L, \epsilon) \geq \frac{1}{2}\min_{v\in\mathcal{V}} d_L(P_0, P_v)\big(1 - \frac{1}{2}\sqrt{D_{kl}(M_0^n\|\bar{M}^n)}\big),$$

where

$$D_{kl}(M_0^n\|\bar{M}^n) \leq \frac{n\epsilon^2}{4}\mathcal{C}_\infty(\{P_v\}_{v\in\mathcal{V}})\min\{e^\epsilon, \max_{v\in\mathcal{V}}\|\frac{dP}{dP_v}\|_\infty\}$$

for any distribution P supported on $\mathcal{X}$. Here

$$\mathcal{C}_\infty(\{P_v\}_{v\in\mathcal{V}}) = \inf_{\text{supp}P^*\in\mathcal{X}} \sup_\gamma\{\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\phi_v(\gamma)^2 | \|\gamma\|_{L^\infty(P^*)}\}.$$

Where the linear functional $\phi_v(f)$ is defined as

$$\phi_v(f) := \int f(x)(dP_0(x) - dP_v(x)).$$

C. Proofs in Sections V and VI

Proof of Theorem 1: The main idea of the proof is :

- Find an index set $\mathcal{V}$ which corresponds to a 2δ-separated set $\{P_{v,\sigma_v}, v \in \mathcal{V}\}$.
- Obtain an upper bound on $\mathcal{C}_\infty(\{P_{v,\sigma_v}\}_{v\in\mathcal{V}})$, use Lemma 7 to specify δ, and then get an lower bound.

We consider $\mathcal{V}$ as the set of $\{\pm e_j, j \in [p]\}$, where $\{e_j\}_{j=1}^n$ is the standard basis of $\mathbb{R}^p$. Let $\theta_v = \delta v$ for some $\delta < 1$ and every $v \in \mathcal{V}$. Then for each θ_v, we define the distribution P_{θ_v,σ_v} as

$$P_{\theta_v,\sigma_v} = \big\{x \in \text{Uniform}\{+1,-1\}^p; p_{\theta_v}(y\,|\,x,\sigma) = \langle x, \theta_v\rangle + \sigma; \text{where } \sigma = \begin{cases} 1 - \langle x, \theta_v\rangle \text{ w.p.} \frac{1+\langle x,\theta_v\rangle}{2} \\ -1 - \langle x, \theta_v\rangle \text{ w.p.} \frac{1-\langle x,\theta_v\rangle}{2} \end{cases} \big\}. \quad (14)$$

It is easy to see that $P_{\theta_v,\sigma_v} \in \mathcal{P}_{1,p,2}$ since the noise $|\sigma_v| \leq 1+|\langle x, \theta_v\rangle| \leq 2$. Note that the distribution in (14) is equivalent to

$$p_{\theta_v,\sigma_v}((x,y)) = \frac{1+y\langle x,\theta_v\rangle}{2^{p+1}} \text{ for } (x,y) \in \{+1,-1\}^{p+1}. \tag{15}$$

Also for every fixed $(x,y) \in \{+1,-1\}^{p+1}$, we have $\bar{p}((x,y)) := \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} p_{\theta_v,\sigma_v}((x,y)) = \frac{1}{2^{p+1}}$.

Now we show our main lemma used in the proof. For convenience we denote $P_v = P_{\theta_v,\sigma_v}$ (the same for later theorems).

Lemma 10: The term $\mathcal{C}_\infty^{Nint}(\{P_v\}_{v\in\mathcal{V}})$ satisfies the following inequality

$$\mathcal{C}_\infty^{Nint}(\{P_v\}_{v\in\mathcal{V}}) \leq \frac{\delta^2}{p}. \tag{16}$$

Proof of Lemma 10: By definition, for each $v \in \mathcal{V}$ we have

$$\begin{aligned}\psi_v(\gamma) &= \sum_{(x,y)\in\{+1,-1\}^{p+1}} \gamma(x,y)[p_v((x,y)) - \bar{p}((x,y))]\\ &= \frac{\delta}{2^{p+1}} \sum_{(x,y)\in\{+1,-1\}^{p+1}} \gamma(x,y)y\langle x,v\rangle\\ &= \frac{\delta}{2^{p+1}} \sum_{x\in\{+1,-1\}^p} [\gamma(x,1)\langle x,v\rangle - \gamma(x,-1)\langle x,v\rangle]\end{aligned}$$

Thus, we can get

$$\begin{aligned}\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \psi_v^2(\gamma) &\leq 2\times\frac{1}{2p}\sum_{v\in\mathcal{V}}[\big(\frac{\delta}{2^{p+1}}\sum_{x\in\{+1,-1\}^p}\gamma(x,1)\langle x,v\rangle\big)^2\\ &\quad + \big(\frac{\delta}{2^{p+1}}\sum_{x\in\{+1,-1\}^p}\gamma(x,-1)\langle x,v\rangle\big)^2]\\ &= \frac{\delta^2}{p4^{p+1}}\sum_{v\in\mathcal{V}}\sum_{x_1,x_2\in\{+1,-1\}^p}[\big(\gamma(x_1,1)\gamma(x_2,1)\\ &\quad + \gamma(x_1,-1)\gamma(x_2,-1)\big)x_1^Tvv^Tx_2]\\ &= \frac{2\delta^2}{p4^{p+1}}\sum_{x_1,x_2\in\{+1,-1\}^p}\big(\gamma(x_1,1)\gamma(x_2,1)x_1^Tx_2\\ &\quad + \gamma(x_1,-1)\gamma(x_2,-1)x_1^Tx_2\big),\end{aligned}$$

where the last equation is due to $\sum_{v\in\mathcal{V}} vv^T = 2I_{p\times p}$. Thus by the definition of $\mathcal{C}_\infty^{Nint}(\{P_v\}_{v\in\mathcal{V}})$ we have

$$\begin{aligned}\mathcal{C}_\infty^{Nint}(\{P_v\}_{v\in\mathcal{V}}) &\leq \frac{1}{2}\frac{\delta^2}{p4^p}[\sup_{\gamma\in\mathbb{B}_\infty}\sum_{x_1,x_2\in\mathcal{X}}\gamma(x_1,1)\gamma(x_2,1)x_1^Tx_2\\ &\quad + \sup_{\gamma\in\mathbb{B}_\infty}\sum_{x_1,x_2\in\mathcal{X}}\gamma(x_1,-1)\gamma(x_2,-1)x_1^Tx_2]\\ &= \frac{\delta^2}{2p}[\sup_{\gamma\in\mathbb{B}_\infty}\|\mathbb{E}_{P_0}[\gamma(X,1)X]\|^2 + \sup_{\gamma\in\mathbb{B}_\infty}\|\mathbb{E}_{P_0}[\gamma(X,-1)X]\|^2],\end{aligned}$$

where P_0 is the uniform distribution on $\{+1,-1\}^p$. Note that since $\|a\|_2^2 = \sup_{\|v\|\leq 1}\langle v,a\rangle^2$ for any vector a, by Cauchy-Schwartz inequality we have

$$\begin{aligned}&\sup_{\gamma\in\mathbb{B}_\infty}\|\mathbb{E}_{P_0}[\gamma(X,1)X]\|^2\\ &= \sup_{\gamma\in\mathbb{B}_\infty,\|v\|_2\leq 1}(\mathbb{E}_{P_0}[\gamma(X,1)v^TX])^2\\ &\leq \sup_{\gamma\in\mathbb{B}_\infty}\mathbb{E}_{P_0}[\gamma(X,1)^2]\times\sup_{\|v\|_2\leq 1}\mathbb{E}_{P_0}[(v^TX)^2]\\ &\leq \sup_{\|v\|_2\leq 1} v^T\sum_{x\in\{-1,1\}^p}\frac{xx^T}{2^p}v \leq 1,\end{aligned}$$

where the second inequality is due to the definition of X and γ. Similarly, we can bound the term $\sup_{\gamma\in\mathbb{B}_\infty}\|\mathbb{E}_{P_0}[\gamma(X,-1)X\|^2] \leq 1$. This completes the proof. □

By Lemma 7 and Lemma 10 , we can get

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}_{1,p,2}),\Phi\circ\rho,\alpha) \geq \frac{\delta^2}{2}\big(1 - \frac{n\epsilon^2\frac{\delta^2}{p} + \log 2}{\log 2p}\big).$$

If we take $\delta^2 = \Omega(\min\{1, \frac{p\log 2p}{n\epsilon^2}\})$, we can get the proof of the lower bound in Theorem 1. □

Proof of Theorem 2: Now we use the squared loss as the loss function $L(\theta,\theta') = \|\theta-\theta'\|_2^2$. Then, $d_L(P_0,P_1) = \frac{1}{2}\|\theta(P_0) - \theta(P_1)\|_2^2$. Define $P_0 \in \mathcal{P}_{1,p,C}$ as the uniform distribution on $\{+1,-1\}^p \times \{+1,-1\}$, that is,

$$P_0 = \big\{x \in \text{Uniform}\{+1,-1\}^p; p_{\theta_v}(y \mid x,\sigma) = \langle x,0\rangle + \sigma;\\ \text{where } \sigma = \begin{cases}1-\langle x,0\rangle \text{ w.p. } \frac{1+\langle x,0\rangle}{2}\\ -1-\langle x,0\rangle \text{ w.p.} \frac{1-\langle x,0\rangle}{2}\end{cases}\big\}.$$

Thus, $\theta(P_0) = 0$.

Define the set of distributions $\{P_v, v\in\mathcal{V}\}$ in the same way as in the proof of Theorem 1. Then, we have $d_L(P_0,P_1) = \frac{1}{2}\delta^2$. As in Lemma 9, we have M_0^n and $\bar{M}^n$. For the KL-divergence D_{kl} between M_0^n and $\bar{M}^n$, by Lemma 9 we have

$$D_{kl}(M_0^n\|\bar{M}^n) \leq \frac{n\epsilon^2}{4}\mathcal{C}_\infty(\{P_v\}_{v\in\mathcal{V}})\min\{e^\epsilon, \max_{v\in\mathcal{V}}\|\frac{dP}{dP_v}\|_\infty\}.$$

We can easily see that for each $\gamma\in\mathbb{B}_\infty$ and $v\in\mathcal{V}$, we have that $\psi_v(\gamma)$ in the proof of Lemma 10 is equivalent to $\phi_v(\gamma)$ in Lemma 9 for our construction. Thus, by Lemma 10 we have $\mathcal{C}_\infty(\{P_v\}_{v\in\mathcal{V}}) \leq \frac{\delta^2}{p}$. Taking $P = P_0$, we get $\max_{v\in\mathcal{V}}\|\frac{dP}{dP_v}\|_\infty = \frac{1}{1-\delta}$. Thus, if choosing $\delta^2 = \Omega(\min\{1, \frac{p}{n\epsilon^2}\})$, we have

$$D_{kl}(M_0^n\|\bar{M}^n) \leq \frac{n\epsilon^2\delta^2(1+\delta)}{8p}.$$

By Lemma 9, we can get

$$\mathcal{M}_n^{\text{Int}}(\theta(\mathcal{P}_{1,p,2}),\Phi\circ\rho,\alpha) \geq \frac{\delta^2}{4}(1 - \sqrt{\frac{n\epsilon^2\delta^2(1+\delta)}{8p}}).$$

Thus, if taking $\delta^2 = \Omega(\min\{1, \frac{p}{n\epsilon^2}\})$, we have the proof. □

Proof of Theorem 3: Now consider the case of $L(\theta,\theta') = |1^T(\theta-\theta')|$. We can easily obtain $d_L(P_1,P_2) \geq |1^T(\theta(P_2) - \theta(P_1))|$. Consider the same distributions $P_0, \{P_v, v\in\mathcal{V}\}$ as in the proof of Theorem 2, we have $\min_{v\in\mathcal{V}} d_L(P_0,P_v) \geq \delta$.

Since $D_{kl}(M_0^n\|\bar{M}^n) \leq \frac{n\epsilon^2\delta^2(1+\delta)}{8p}$ for $\delta^2 = \Omega(\min\{1, \frac{p}{n\epsilon^2}\})$, we have

$$\mathcal{M}_n^{\text{int}}(\theta(\mathcal{P}_{1,p,2}), L, \alpha) \geq \frac{\delta}{2}(1 - \sqrt{\frac{n\epsilon^2\delta^2(1+\delta)}{8p}}).$$

Thus, we have the proof if set $\delta^2 = \Omega(\min\{1, \frac{p}{n\epsilon^2}\})$. □

Proof of Theorem 4: Before the proof, let us recall the definition of χ^2-local differential privacy [11]:

For any convex function f on $\mathbb{R}_+$ with $f(1) = 0$, the f-divergence of distributions P and Q is

$$D_f(P\|Q) := \int f(\frac{dP}{dQ})dQ.$$

Definition 11: Let $f(x) = (x-1)^2$. Following the above definitions, we have ϵ^2-χ^2-divergence local differential privacy and ϵ-χ^2-divergence (sequentially) private minimax risk if

$$D_f(Q_i(Z_i \in S \mid x_i, z_{1:i-1})\|Q_i(Z_i \in S \mid x_i', = z_{1:i-1})) \leq \epsilon^2.$$

From the above definitions, it is easy to see that if a channel Q is (κ, ρ) sequentially locally zero-concentrated differentially private, it is ($\epsilon^2 = e^{\kappa+2\rho} - 1$)-χ^2-divergence sequentially locally differentially private. Also, since $(2, \log(1+\epsilon^2))$ local Renyi differential privacy is equivalent to ϵ^2-χ-divergence local differential privacy, to prove Theorem 4, we only need to show the lower bound of ϵ^2-χ^2-divergence sequential local private minimax risk, which is denoted as $\mathcal{M}_{n,\chi^2}^{\text{Int}}(\theta(\mathcal{P}), L, \epsilon^2)$. To do that, we need the following lemma.

Lemma 11 (Theorem 2 in [11]): For any $\epsilon \in (0, 1]$, the ϵ^2-χ^2-divergence sequential private minimax risk in the loss function L satisfies the following inequality

$$\mathcal{M}_{n,\chi^2}^{\text{Int}}(\theta(\mathcal{P}), L, \epsilon^2) \geq \frac{1}{2}\min_{v\in\mathcal{V}} d_L(P_0, P_v) \times (1 - \frac{1}{2}\sqrt{D_{kl}(M_0^n\|\bar{M}^n)}),$$

where

$$D_{kl}(M_0^n\|\bar{M}^n) \leq n\epsilon^2\mathcal{C}_2(\{P_v\}_{v\in\mathcal{V}})\min\{e^\epsilon, \max_{v\in\mathcal{V}}\|\frac{dP_v}{dP}\|_\infty\}$$

for any distribution P supported on $\mathcal{X}$, and $\mathcal{C}_2(\{P_v\}_{v\in\mathcal{V}}) = \frac{1}{|\mathcal{V}|}\inf_{\text{supp}P\subset\mathcal{X}}\sup_\gamma\{\sum_{v\in\mathcal{V}}(\phi_v(\gamma))^2 \mid \|\gamma\|_{L^2(P)} \leq 1\}$, where $\phi(\gamma)$ is defined in Lemma 9.

Now, we will proof Theorem 4.

The construction of P_0 and $\{P_v, v \in \mathcal{V}\}$ is the same as in the proof of Theorem 3. Thus, by Lemma 11, we only need to bound $\mathcal{C}_2(\{P_v\}_{v\in\mathcal{V}})$, instead of $\mathcal{C}_\infty(\{P_v\}_{v\in\mathcal{V}})$. From the proof of Lemma 10, we can see that if taking P as a uniform distribution, then for any γ with $\|\gamma\|_{L^2(P_0)} \leq 1$, we always have $\mathbb{E}_{P_0}[\gamma(X,1)^2] \leq 1$. This means that $\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}(\psi_v(\gamma))^2 \leq \frac{\delta^2}{p}$. Thus, we have $\mathcal{C}_2(\{P_v\}_{v\in\mathcal{V}}) \leq \frac{\delta^2}{p}$. The remaining part of the proof is the same as the one in the proof of Theorem 2. □

Proof of Theorem 5: Follow from the fact that the linear model is a special case of the non-linear measurement. See the proof of Theorem 9 in Section VII-B for the case $f(x) = x$ and $a = b = 1$. □

Proof of Theorem 6: Follow from the fact that the linear model is a special case of the non-linear measurement. See the proof of Theorem 11 in Section VII-B for the case $f(x) = x$ and $a = b = 1$. □

Proof of Theorem 7: Follow from the fact that the linear model is a special case of the non-linear measurement. See the proof of Theorem 10 in Section VII-B for the case $f(x) = x$ and $a = b = 1$. □

D. Proofs in Section VII-A

Proof of Theorem 8: For the guarantee of (ϵ, δ)-DP, it follows from the Moment accountant and composition theorem, see [31], [55] for details.

Let $\mathcal{I} = \mathcal{I}^{t+1}\bigcup\mathcal{I}^t\bigcup\mathcal{I}^*$, where $\mathcal{I}^* = \text{supp}(x^*)$, $\mathcal{I}^t = \text{supp}(x_t)$ and $\mathcal{I}^{t+1} = \text{supp}(x_{t+1})$, and $g_t = \nabla L(x_t) + z_t$. Since $\|x_{t+1} - x_t\|_0 \leq 2k$. By the assumption of RSS, we have

$$\begin{aligned} L(x_{t+1}) &\leq L(x_t) + \langle\nabla L(x_t), x_{t+1} - x_t\rangle + \frac{\ell_s}{2}\|x_{t+1} - x_t\|^2 \\ &\leq L(x_t) + \langle(g_t)_\mathcal{I}, (x_{t+1} - x_t)_\mathcal{I}\rangle + \frac{\ell_s}{2}\|x_{t+1} - x_t\|^2 \\ &\quad + \|z_{t,\mathcal{I}}\|\|(x_{t+1} - x_t)_\mathcal{I}\|_2 \\ &= L(x_t) + \frac{1}{2\eta}\|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \frac{\eta\|g_{t,\mathcal{I}}\|^2}{2} \\ &\quad - \frac{1-\eta\ell_s}{2\eta}\|x_{t+1} - x_t\|^2 + \|z_{t,\mathcal{I}}\|\|(x_{t+1} - x_t)_\mathcal{I}\|_2 \\ &= L(x_t) + \frac{1}{2\eta}(\|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \eta^2\|g_{t,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|^2) \\ &\quad - \frac{\eta\|g_{t,\mathcal{I}^t\bigcup\mathcal{I}^*\|^2}}{2} + \|z_{t,\mathcal{I}}\|\|(x_{t+1} - x_t)_\mathcal{I}\|_2, \end{aligned} \tag{17}$$

where the second inequality is due to $x_{t+1} - x_t = x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}}$.

We now bound the term of $\|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \eta^2\|g_{t,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|^2$ by the idea in [32]. Since $\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*) = \mathcal{I}^{t+1}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*) \subseteq \mathcal{I}^{t+1}$, we have

$$x_{t+1,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)} = x_{t,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)} - \eta g_{t,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)}.$$

Also, since $x_{t,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)} = 0$, this means that $x_{t+1,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)} = -\eta g_{t,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)}$. Next, we choose a set $\mathcal{R} \subseteq \mathcal{I}^t\backslash\mathcal{I}^{t+1}$ such that $|\mathcal{R}| = |\mathcal{I}^{t+1}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)|$. Note that such $\mathcal{R}$ can be found since $|\mathcal{I}^{t+1}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)| = |\mathcal{I}^t\backslash\mathcal{I}^{t+1}| - |(\mathcal{I}^{t+1}\bigcap\mathcal{I}^*)\backslash\mathcal{I}^t|$ (which is a consequence of $|\mathcal{I}^t| = |\mathcal{I}^{t+1}|$). Thus, we have

$$\begin{aligned} \eta^2\|g_{t,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|^2 &= \|x_{t+1,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|^2 \\ &\geq \|x_{t,\mathcal{R}} - \eta g_{t,\mathcal{R}}\|^2. \end{aligned} \tag{18}$$

With (18) and the fact that $x_{t+1,\mathcal{R}} = 0$, we have

$$\begin{aligned} &\|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \eta^2\|g_{t,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|^2 \\ &\leq \|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \|x_{t+1,\mathcal{R}} - x_{t,\mathcal{R}} + \eta g_{t,\mathcal{R}}\|^2 \\ &= \|x_{t+1,\mathcal{I}\backslash\mathcal{R}} - x_{t,\mathcal{I}\backslash\mathcal{R}} + \eta g_{t,\mathcal{I}\backslash\mathcal{R}}\|^2. \end{aligned} \tag{19}$$

We then bound the size of $|\mathcal{I}\backslash\mathcal{R}|$ as $|\mathcal{I}\backslash\mathcal{R}| \leq |\mathcal{I}^{t+1}| + |(\mathcal{I}^t\backslash\mathcal{I}^{t+1})\backslash\mathcal{R}| + |\mathcal{I}^*| \leq k + |(\mathcal{I}^{t+1}\bigcap\mathcal{I}^*)\backslash\mathcal{I}^t| + k^* \leq k + 2k^*$. Also, since $\mathcal{I}^{t+1} \subseteq (\mathcal{I}\backslash\mathcal{R})$, we have $x_{t+1,\mathcal{I}\backslash\mathcal{R}} = \text{Trun}(x_{t,\mathcal{I}\backslash\mathcal{R}} - \eta g_{t,\mathcal{I}\backslash\mathcal{R}}, k)$. Thus, by (18) and Lemma 3 we have

$$\begin{aligned}
&\|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \eta^2\|g_{t,\mathcal{I}\setminus(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|^2\\
&\leq \|x_{t+1,\mathcal{I}\setminus\mathcal{R}} - x_{t,\mathcal{I}\setminus\mathcal{R}} + \eta g_{t,\mathcal{I}\setminus\mathcal{R}}\|^2\\
&\leq \frac{2k^*}{k+k^*}\|x^*_{\mathcal{I}\setminus\mathcal{R}} - x_{t,\mathcal{I}\setminus\mathcal{R}} + \eta g_{t,\mathcal{I}\setminus\mathcal{R}}\|^2\\
&\leq \frac{2k^*}{k+k^*}\|x^*_{\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2\\
&= \frac{2k^*}{k+k^*}(\|x^* - x^t\|^2 + 2\eta\langle g_{t,\mathcal{I}}, (x^*-x_t)_{\mathcal{I}}\rangle + \eta^2\|g_{t,\mathcal{I}}\|^2)\\
&= \frac{2k^*}{k+k^*}(\|x^* - x^t\|^2 + 2\eta\langle \nabla L(x_t), (x^*-x_t)\rangle + \eta^2\|g_{t,\mathcal{I}}\|^2)\\
&\quad + \frac{4k^*}{k+k^*}\langle z_{t,\mathcal{I}}, (x^*-x_t)_{\mathcal{I}}\rangle\\
&\leq \frac{2k^*}{k+k^*}[\|x^* - x^t\|^2 + 2\eta(L(x^*) - L(x_t) - \frac{\rho_s}{2}\|x^*-x_t\|^2)\\
&\quad + \eta^2\|g_{t,\mathcal{I}}\|^2] + \frac{4k^*}{k+k^*}\langle z_{t,\mathcal{I}}, (x^*-x_t)_{\mathcal{I}}\rangle\\
&= \frac{4\eta k^*}{k+k^*}(L(x^*) - L(x_t)) + \frac{2(1-\eta\rho_s)k^*}{k+k^*}\|x^*-x_t\|^2\\
&\quad + \frac{2\eta^2 k^*}{k+k^*}\|g_{t,\mathcal{I}\setminus(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|^2 + \frac{2\eta^2 k^*}{k+k^*}\|g_{t,(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|^2\\
&\quad + \frac{4k^*}{k+k^*}\langle z_{t,\mathcal{I}}, (x^*-x_t)_{\mathcal{I}}\rangle.
\end{aligned}$$

Plugging this into (17), we get

$$\begin{aligned}
&L(x_{t+1}) \leq L(x_t) + \frac{2k^*}{k+k^*}(L(x^* - L(x_t))\\
&+ \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)}\|x^*-x_t\|^2 + \frac{\eta k^*}{k+k^*}\|g_{t,\mathcal{I}\setminus(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|^2\\
&+ (\frac{\eta k^*}{k+k^*} - \frac{\eta}{2})\|g_{t,\mathcal{I}^t\bigcup\mathcal{I}^*}\|^2 + \frac{2k^*}{\eta(k+k^*)}\langle z_{t,\mathcal{I}}, (x^*-x_t)_{\mathcal{I}}\rangle\\
&+ \|z_{t,\mathcal{I}}\|\|(x_{t+1}-x_t)_{\mathcal{I}}\|_2 - \frac{1-\eta\ell_s}{2\eta}\|x_{t+1}-x_t\|^2\\
&\leq L(x_t) + \frac{2k^*}{k+k^*}(L(x^* - L(x_t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)}\|x^*-x_t\|^2 -\\
&(\frac{1-\eta\ell_s}{2\eta} - \frac{k^*}{\eta(k+k^*)})\|x_{t+1}-x_t\|^2 + (\frac{\eta k^*}{k+k^*} - \frac{\eta}{2})\|g_{t,\mathcal{I}^t\bigcup\mathcal{I}^*}\|^2\\
&\quad + \frac{2k^*}{\eta(k+k^*)}\langle z_{t,\mathcal{I}}, (x^*-x_t)_{\mathcal{I}}\rangle + \|z_{t,\mathcal{I}}\|\|(x_{t+1}-x_t)_{\mathcal{I}}\|_2
\end{aligned} \tag{20}$$

$$\begin{aligned}
&\leq L(x_t) + \frac{2k^*}{k+k^*}(L(x^* - L(x_t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)}\|x^*-x_t\|^2\\
&\quad + (\frac{\eta k^*}{k+k^*} - \frac{\eta}{2})\|g_{t,\mathcal{I}^t\bigcup\mathcal{I}^*}\|^2 + \frac{2k^*}{\eta(k+k^*)}\langle z_{t,\mathcal{I}}, (x^*-x_t)_{\mathcal{I}}\rangle\\
&\quad + \frac{\eta(k+k^*)}{2((1-\eta\ell_s)k - (1+\eta\ell_s)k^*)}\|z_{t,\mathcal{I}}\|^2,
\end{aligned} \tag{21}$$

where the second inequality is due to the fact that $\|x_{t+1} - x_t\| \geq \eta\|g_{t,\mathcal{I}\setminus(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|$ and the third inequality is due to the fact that $ab \leq \frac{a^2}{4c} + cb^2$ for any $c > 0$.

For the term $\|x_t - x^*\|^2$, we have the following lemma:

Lemma 12:

$$\|x_t - x^*\|^2 \leq \frac{4}{\rho}[L(x^*) - L(x_t)] + \frac{8}{\rho_s^2}\|g_{t,\mathcal{I}^t\bigcup\mathcal{I}^*}\|^2 + \frac{8}{\rho_s^2}\|z_{t,\mathcal{I}}\|^2. \tag{22}$$

Proof: From RSC, we have

$$\begin{aligned}
&L(x^*) \geq L(x_t) + \langle\nabla L(x_t), x^*-x_t\rangle + \frac{\rho_s}{2}\|x^*-x_t\|^2\\
&= L(x_t) + \langle\nabla_{\mathcal{I}^t\bigcup\mathcal{I}^*}L(x_t) - g_{t,\mathcal{I}^t\bigcup\mathcal{I}^*} + g_{t,\mathcal{I}^t\bigcup\mathcal{I}^*}, x^*-x_t\rangle\\
&\quad + \frac{\rho_s}{2}\|x^*-x_t\|^2\\
&\geq L(x_t) - \frac{2}{\rho_s}\|z_{t,\mathcal{I}}\|^2 - \frac{2}{\rho_s}\|g_{t,\mathcal{I}^t\bigcup\mathcal{I}^*}\|^2 + \frac{\rho_s}{4}\|x^*-x_t\|^2,
\end{aligned}$$

where the last inequality is due to $ab \leq \frac{a^2}{4c} + cb^2$. □

With this lemma, we get

$$\begin{aligned}
&L(x_{t+1}) \leq L(x_t) + \frac{2k^*}{k+k^*}(1 + \frac{2(1-\eta\rho_s)}{\eta\rho_s})(L(x^*) - L(x_t))\\
&- (\frac{\eta}{2} - \frac{(\eta^2\rho_s^2 + 8(1-\eta\rho_s))k^*}{\eta\rho_s^2(k+k^*)})\|g_{t,\mathcal{I}^t\bigcup\mathcal{I}^*}\|^2\\
&+ \frac{2k^*}{\eta(k+k^*)}\langle z_{t,\mathcal{I}}, (x^*-x_t)_{\mathcal{I}}\rangle + (\frac{\eta(k+k^*)}{2((1-\eta\ell_s)k - (1+\eta\ell_s)k^*)}\\
&+ \frac{8(1-\eta\rho_s)k^*}{\eta\rho_s^2(k+k^*)})\|z_{t,\mathcal{I}}\|^2.
\end{aligned} \tag{23}$$

Taking $\eta = \frac{1}{2\ell_s}$ and $k \geq (1 + \frac{64\ell_s^2}{\rho_s^2})k^*$, we further get

$$\begin{aligned}
L(x_{t+1}) &\leq L(x_t) + \frac{\rho_s}{8\ell_s}(L(x^*) - L(x_t))\\
&+ \frac{4k^*\ell_s}{(k+k^*)}\langle z_{t,\mathcal{I}}, (x^*-x_t)_{\mathcal{I}}\rangle + \frac{37\ell_s}{\rho_s^2}\|z_{t,\mathcal{I}}\|^2.
\end{aligned} \tag{24}$$

Lemma 13: For $x \sim \mathcal{N}(0, \sigma^2 I_p)$

$$\mathbb{E}|x|_\infty^2 \leq O(\sigma^2\log p)$$

Proof: By definition of expectation, we have

$$\begin{aligned}
&\mathbb{E}|x|_\infty^2 = \int_0^\infty \Pr[|x|_\infty^2 \geq t]dt\\
&= \int_0^{O(\sigma^2\log p)} \Pr[|x|_\infty^2 \geq t]dt + \int_{O(\sigma^2\log p)}^\infty \Pr[|x|_\infty^2 \geq t]dt\\
&\leq O(\sigma^2\log p) + \int_{O(\sigma^2\log p)}^\infty 2p\exp(-\frac{t}{2\sigma^2})dt\\
&\leq O(\sigma^2\log p) + 2\sqrt{2}p\sigma^2\exp(-O(\log p)) = O(\sigma^2\log p).
\end{aligned}$$

□

Note that $\mathbb{E}\langle z_{t,\mathcal{I}}, (x^*-x_t)_{\mathcal{I}}\rangle = \mathbb{E}\langle z_t, x^*-x_t\rangle = 0$. Taking the expectation w.r.t z_t and by the fact that $\|z_{t,\mathcal{I}}\|^2 \leq |I||z_t|_\infty^2$ (from the above lemma), we have

$$\begin{aligned}
\mathbb{E}L(x_{t+1}) &\leq L(x_t) + \frac{\rho_s}{8\ell_s}(L(x^*) - L(x_t))\\
&\qquad + O(\frac{\kappa_s k^* G^2 \log\frac{1}{\delta}\log pT}{\rho_s n^2\epsilon^2}).
\end{aligned} \tag{25}$$

That is

$$\begin{aligned}
\mathbb{E}[L(x_{t+1}) - L(x^*)] &\leq (1 - \frac{\rho_s}{8\ell_s})\mathbb{E}[L(x_t) - L(x^*)]\\
&\qquad + O(\frac{\kappa_s k^* G^2 \log p\log\frac{1}{\delta}T}{\rho_s n^2\epsilon^2}).
\end{aligned} \tag{26}$$

Thus, taking $T = O(\kappa_s\log(\frac{n^2}{k^*}))$, we get the theorem. □

E. Proofs in Section VII-B

Proof of Theorem 9: We first show that each stochastic gradient $\|x_i^T f'(\langle x_i,\theta_{t-1}\rangle)(f(\langle x_i,\theta_{t-1}\rangle)-y_i)\|_2 \leq O(bC\sqrt{p})$, this is due to that

$$\begin{aligned}&\|x_i^T f'(\langle x_i,\theta_{t-1}\rangle)(f(\langle x_i,\theta_{t-1}\rangle)-y_i)\|_2\\ &\leq b\|x_i^T\|_2(f(\langle x_i,\theta_{t-1}\rangle)-y_i)\\ &\leq b\sqrt{p}(f(1)-y_i)\leq O(bC\sqrt{p}),\end{aligned}$$

where the second inequality is due to that $\langle x_i,\theta_{t-1}\rangle \leq \|x_i\|_\infty\|\theta_{t-1}\|_2\leq 1$, f is monotone and $|y_i| = |f\langle\theta^*,x_i\rangle+\sigma_i|\leq O(C)$.

W.o.l.g we assume that each $|S_t| = \frac{n}{T}$. From the randomizer $\mathcal{R}_\epsilon(\cdot)$ and Lemma 1, we can see that $\tilde{\nabla}_t = \frac{T}{n}\sum_{i\in S_t} x_i^T f'(\langle x_i,\theta_{t-1}\rangle)(f(\langle x_i,\theta_{t-1}\rangle)-y_i)+\zeta_t$, where each coordinate of ζ_t is a sub-Gaussian vector with $\sigma^2 = O(\frac{bCpT}{n\epsilon^2})$.

Let $\mathcal{S}^* = \text{supp}(\theta^*)$ denote the support of θ^*, and $s^* = |\mathcal{S}^*|$. Similarly, we define $\mathcal{S}^t = \text{supp}(\theta_t)$, and $\mathcal{F}^{t-1} = \mathcal{S}^{t-1}\cup\mathcal{S}^t\cup\mathcal{S}^*$. Thus, we have $|\mathcal{F}^{t-1}|\leq 2s+s^*$.

We let $\tilde{\theta}_{t-\frac{1}{2}}$ denote the following

$$\tilde{\theta}_{t-\frac{1}{2}} = \theta_{t-1}-\eta\tilde{\nabla}_{t-1,\mathcal{F}^{t-1}},$$

where $v_{\mathcal{F}^{t-1}}$ means keeping v_i for $i\in\mathcal{F}^{t-1}$ and converting all other terms to 0. By the definition of $\mathcal{F}^{t-1}$, we have $\theta_t' = \text{Trunc}(\tilde{\theta}_{t-\frac{1}{2}}, s)$. Denote by Δ_t the difference of $\theta_t-\theta^*$. We have the following

$$\|\tilde{\theta}_{t-\frac{1}{2}}-\theta^*\|_2 = \|\Delta_{t-1}-\eta([\nabla L_t(\theta_{t-1})+\zeta_t]_{\mathcal{F}^{t-1}})\|_2,$$

where $\nabla L_t(\theta_{t-1}) = \frac{T}{n}\sum_{i\in S_t}(f(\langle x_i,\theta_{t-1}\rangle)-y_i)f'(\langle x_i,\theta_{t-1}\rangle)x_i^T$. Taking $y_i = \langle x_i,\theta^*\rangle+\sigma_i$ and by the triangle inequality we can get

$$\begin{aligned}&\|\tilde{\theta}_{t-\frac{1}{2}}-\theta^*\|_2\leq\|\Delta_{t-1}-\\ &\eta[\frac{T}{n}\sum_{i\in S_t}(f(\langle x_i,\theta_{t-1}\rangle)-f(\langle x_i,\theta^*\rangle))f'(\langle x_i,\theta_{t-1}\rangle)x_i^T]_{\mathcal{F}^{t-1}}\|_2\\ &\quad+\eta\sqrt{|\mathcal{F}^{t-1}|}[|\frac{T}{n}\sum_{i\in S_t}f'(\langle x_i,\theta_{t-1}\rangle)\sigma_i x_i^T|_\infty+|\zeta_t|_\infty].\end{aligned}$$

We denote the followings:

$$\begin{aligned}A^{t-1} &= \|\Delta_{t-1}-\eta[\frac{T}{n}\sum_{i\in S_t}(f(\langle x_i,\theta_{t-1}\rangle)-f(\langle x_i,\theta^*\rangle))\\ &\quad\times f'(\langle x_i,\theta_{t-1}\rangle)x_i^T]_{\mathcal{F}^{t-1}}\|_2 &(27)\\ B^{t-1} &= \eta\sqrt{|\mathcal{F}^{t-1}|}|\frac{T}{n}\sum_{i\in S_t}f'(\langle x_i,\theta_{t-1}\rangle)\sigma_i x_i^T|_\infty &(28)\\ C^{t-1} &= \eta\sqrt{|\mathcal{F}^{t-1}|}|\zeta_t|_\infty &(29)\end{aligned}$$

We first bound B^{t-1}. Since each $x_i\in\text{Uniform}\{+1,-1\}^p$, which is sub-Gaussian with 1, we know that for each coordinate $j\in[p]$, $\frac{T}{n}\sum_{i\in S_t}f'(\langle x_i,\theta_{t-1}\rangle)\sigma_i x_{i,j}$ is sub-Gaussian with $\sigma^2 = \frac{T^2}{n^2}\sum_{i\in S_t}f'^2(\langle x_i,\theta_{t-1}\rangle)\sigma_i^2 \leq \frac{Tb^2C^2}{n}$. Thus, by Lemma 2 we have

$$\Pr[|\frac{1}{n}\sum_{i=1}^n f'(\langle x_i,\theta_t\rangle)\sigma_i x_i^T|_\infty \leq O(\frac{\sqrt{T\log p}bC}{\sqrt{n}})]\geq 1-\frac{1}{p^c}.$$

This means that with probability at least $1-\frac{1}{p^c}$, we have

$$B^t\leq\eta\sqrt{2s+s^*}O(\frac{\sqrt{T\log p}bC}{\sqrt{n}}). \tag{30}$$

For the term C^{t-1}, by Lemma 1 and 2 and since each coordinate $\zeta_{t,i}$ is sub-Gaussian, we have $C^{t-1}\leq\eta\sqrt{2s+s^*}O(\frac{\sqrt{TpbC\log p}}{\sqrt{n\epsilon^2}})$ with probability at least $1-\frac{1}{p^c}$ for some constant $c>0$.

Finally, we bound the term A^{t-1}. By the mean value theorem, we know that there exists a $\theta_{t-1,i}$ line between θ_{t-1} and θ^* which satisfies the equation $f(\langle x_i,\theta_{t-1}\rangle)-f(\langle x_i,\theta^*\rangle) = f'(\langle x_i,\theta_{t-1,i}\rangle)\langle x_i,\theta_{t-1}-\theta^*\rangle)$. Hence, we have

$$\frac{T}{n}\sum_{i\in S_t}(f(\langle x_i,\theta_{t-1}\rangle)-f(\langle x_i,\theta^*\rangle))f'(\langle x_i,\theta_{t-1}\rangle)x_i^T = D^{t-1}\Delta_{t-1},$$

where $D^{t-1} = \frac{T}{n}\sum_{i\in S_t}f'(\langle x_i,\theta_{t-1,i}\rangle)f'(\langle x_i,\theta_{t-1}\rangle)x_i x_i^T\in\mathbb{R}^{p\times p}$.

Since $\text{Supp}(D^{t-1}\Delta_{t-1})\subset\mathcal{F}^{t-1}$ (by assumption), we have $A^{t-1} = \|\Delta_{t-1}-\eta D^{t-1}_{\mathcal{F}^{t-1},\cdot}\Delta_{t-1}\|_2\leq\|(I-\eta D^{t-1}_{\mathcal{F}^{t-1},\mathcal{F}^{t-1}})\|_2\|\Delta_{t-1}\|_2$. Now we bound the term $\|(I-\eta D^{t-1}_{\mathcal{F}^{t-1},\mathcal{F}^{t-1}})\|_2$, where I is the $|\mathcal{F}^{t-1}|$-dimensional identity matrix.

By the RIP property of X and $|\mathcal{F}^{t-1}|\leq 2s+s^*$, we can easily get the following for any $|\mathcal{F}^{t-1}|$-dimensional vector v

$$a^2[1-\Delta(2s+s^*)]\|v\|_2^2\leq v^T D^{t-1}_{\mathcal{F}^{t-1},\mathcal{F}^{t-1}}v\leq b^2[1+\Delta(2s+s^*)].$$

Thus, $\|(I-\eta D^{t-1}_{\mathcal{F}^{t-1},\mathcal{F}^{t-1}})\|_2\leq\max\{1-\eta a^2[1-\Delta(2s+s^*)],\eta b^2[1+\Delta(2s+s^*)]-1\}$.

This means that if we can find an η satisfying the condition of

$$\frac{5}{7}\frac{1}{a[1-\Delta(2s+s^*)]}\leq\eta\leq\frac{9}{7}\frac{1}{b^2[1+\Delta(2s+s^*)]},$$

then we have $\|(I-\eta D^{t-1}_{\mathcal{F}^{t-1},\mathcal{F}^{t-1}})\|_2\leq\frac{2}{7}$. Note that such an η can indeed be found if $\Delta(2s+s^*)\leq\frac{5a^2-9b^2}{14}$. This means that $\frac{a}{b}>\frac{\sqrt{5}}{3}$.

Thus, in total we have the following with probability at least $1-\frac{2}{p^c}$

$$\|\tilde{\theta}_{t-\frac{1}{2}}-\theta^*\|_2\leq\frac{2}{7}\|\Delta_{t-1}\|_2+O(\frac{\sqrt{Tp(2s+s^*)\log p}bC}{\sqrt{n}\epsilon}).$$

Our next task is to bound $\|\theta_t'-\theta^*\|_2$ by $\|\tilde{\theta}_{t-\frac{1}{2}}-\theta^*\|_2$ by Lemma 3.

Thus, we have $\|\theta_t'-\tilde{\theta}_{t-\frac{1}{2}}\|_2^2\leq\frac{|\mathcal{F}^{t-1}|-s}{|\mathcal{F}^{t-1}|-s^*}\|\tilde{\theta}_{t-\frac{1}{2}}-\theta^*\|_2^2\leq\frac{s+s^*}{2s}\|\tilde{\theta}_{t-\frac{1}{2}}-\theta^*\|_2^2$.

Taking $s=8s^*$, we get

$$\|\theta_t'-\tilde{\theta}_{t-\frac{1}{2}}\|_2\leq\frac{3}{4}\|\tilde{\theta}_{t-\frac{1}{2}}-\theta^*\|_2$$

and

$$\begin{aligned}\|\theta_t'-\theta^*\|_2&\leq\frac{7}{4}\|\tilde{\theta}_{t-\frac{1}{2}}-\theta^*\|_2\\ &\leq\frac{1}{2}\|\Delta_{t-1}\|_2+O(\frac{\sqrt{Tps^*\log p}bC}{\sqrt{n}\epsilon}).\end{aligned}$$

Finally, we need to show that $\|\Delta_t\|_2 = \|\theta_t - \theta^*\|_2 \leq \|\theta'_t - \theta^*\|_2$, which is due to the Lemma 4.

Putting all together, we have the following with probability at least $1 - \frac{2}{p^c}$,

$$\|\Delta_t\| \leq \frac{1}{2}\|\Delta_{t-1}\|_2 + O(\frac{\sqrt{Tps^* \log p} bC}{\sqrt{n}\epsilon}).$$

Thus, we get with probability at least $1 - \frac{2T}{p^c}$,

$$\|\Delta_T\|_2 \leq (\frac{1}{2})^T \|\theta^*\|_2 + O(\frac{\sqrt{Tps^* \log p} bC}{\sqrt{n}\epsilon}).$$

□

Proof of Theorem 10: Our proof is inspired by the ones in [9], [13] and [36]. Since it is reduced to the linear model when $f(x) \equiv x$, we only need to consider the general case. Similar to the proof of Theorem 1, we first construct a packing set $\{P_v : v \in \mathcal{V}\}$ and then bound $\mathcal{C}_\infty(\{P_v\})$. To do so, we need the following lemma.

Lemma 14 [36]: For any $s \in [p]$, define the set

$$\mathcal{H}(s) := \{z \in \{-1, 0, +1\}^d \mid \|z\|_0 = s\}$$

with Hamming distance $\rho_H(z, z') = \sum_{i=1}^d 1[z_j \neq z'_j]$ between the vectors z and z'. Then, there exists a subset $\tilde{\mathcal{H}} \subset \mathcal{H}$ with cardinality $|\tilde{\mathcal{H}}| \geq \exp(\frac{s}{2} \log \frac{p-s}{s/2})$ such that $\rho_H(z, z') \geq \frac{s}{2}$ for all $z, z' \in \tilde{\mathcal{H}}$.

Now consider the rescaled version of $\tilde{\mathcal{H}}$, $\sqrt{\frac{2}{\delta}}\tilde{\mathcal{H}}$, for some $\delta \leq \frac{1}{\sqrt{2}}$. For any two $\theta, \theta' \in \tilde{\mathcal{H}}$, we have

$$8\delta^2 \geq \|\theta - \theta'\|_2^2 \geq \delta^2. \tag{31}$$

Then, $\sqrt{\frac{2}{\delta}}\tilde{\mathcal{H}}$ is a δ packing in ℓ_2 norm with $M = |\tilde{\mathcal{H}}|$ elements, denoted as $\{\theta_1, \theta_2, \cdots, \theta_M\}$. For each θ_i, let σ_i denote the uniform distribution on the interval $[-C, C]$. Thus, we have P_{θ_i}, which can be easily verified that $P_{\theta_i} \in \mathcal{P}'_{s,p,C,f,a,b}$.

Our idea is to use Lemma 8. Thus, our goal is to bound the sum of the Total Variance $\sum_{v,v' \in \mathcal{V}} \|P_{v,i} - P_{v',i}\|_{TV}^2$. Now consider the case of $P_{\theta,i}$ and $P_{\theta',i}$, where (due to our construction) $P_{\theta,i}$ is the uniform distribution on the interval of $[f(\langle x_i, \theta\rangle - C, f(\langle x_i, \theta\rangle + C]$. Thus, we have

$$\|P_{\theta,i} - P_{\theta',i}\|_{TV} = \frac{1}{2}\int |p_{\theta,i}(y) - p_{\theta',i}(y)| dy$$
$$\leq \frac{1}{2C}|f(\langle \theta, x_i\rangle) - f(\langle \theta', x_i\rangle)| \leq \frac{b}{2C}|\langle \theta - \theta', x_i\rangle|,$$

where the last inequality is due to the assumption on f. Hence, we have

$$\sum_{i=1}^n \frac{1}{|\mathcal{V}|^2} \sum_{v,v' \in \mathcal{V}} \|P_{v,i} - P_{v',i}\|_{TV}^2$$
$$\leq \sum_{i=1}^n \frac{b^2}{4C^2} \sum_{v,v \in \mathcal{V}} (\theta_v - \theta_{v'})^T x_i x_i^T (\theta_v - \theta_{v'})$$
$$= \frac{b^2}{4C^2} \frac{1}{|\mathcal{V}|^2} \sum_{v,v \in \mathcal{V}} (\theta_v - \theta_{v'}) X^T X (\theta_v - \theta_{v'})$$
$$\leq 8\frac{b^2(1+\Delta)}{4C^2}\delta^2 = \frac{2b^2(1+\Delta)\delta^2}{C^2},$$

where the last inequality is due to the fact that for every pair (v, v') with $\|\theta_v - \theta_{v'}\|_0 \leq 2s$, $(\theta_v - \theta_{v'}) X^T X (\theta_v - \theta_{v'}) \leq n(1+\Delta)$ holds (by Assumption 1).

Thus by Lemmas 14 and 8, we have

$$\frac{\Phi(\delta)}{2} \geq \frac{\delta^2}{8}(1 - \frac{2cn\epsilon^2\delta^2 \frac{b^2(1+\Delta)}{C^2} + \log 2}{\frac{s}{2}\log \frac{p-s}{s/2}}).$$

Taking $\delta^2 = \Omega(\min\{1, \frac{s \log p/sC^2}{(1+\Delta)b^2 n\epsilon^2}\})$, we get the result. □

Proof of Theorem 11: For the guarantee of (ϵ, δ) locally differentially private, it is due to the fact that x_i is known and each $y_i \in [\langle x_i, \theta^*\rangle - C, \langle x_i, \theta^*\rangle - C]$ (since the random noise σ_i is bounded by C). Thus, by the Gaussian Mechanism [5], we can see that it is locally differentially private.

Now we prove Theorem the upper bound.

Let $\mathcal{S}^* = \text{supp}(\theta^*)$ denote the support of θ^*, and $s^* = |\mathcal{S}^*|$. Similarly, we define $\mathcal{S}^{t+1} = \text{supp}(\theta_{t+1})$, and $\mathcal{F}^t = \mathcal{S}^t \cup \mathcal{S}^{t+1} \cup \mathcal{S}^*$. Thus, we have $|\mathcal{F}^t| \leq 2s + s^*$.

We let $\tilde{\theta}_{t+\frac{1}{2}}$ denote the following

$$\tilde{\theta}_{t+\frac{1}{2}} = \theta_t - \eta\nabla_{\mathcal{F}^t} L(\theta_t),$$

where $v_{\mathcal{F}^t}$ means keeping v_i for $i \in \mathcal{F}^t$ and making all other terms 0. By the definition of $\mathcal{F}^t$, we have $\theta'_{t+1} = \text{Trunc}(\tilde{\theta}_{t+\frac{1}{2}}, s)$. Denote by Δ_{t+1} the difference of $\theta_{t+1} - \theta^*$. We have the following

$$\|\tilde{\theta}_{t+\frac{1}{2}} - \theta^*\|_2 = \|\Delta_t - \eta\nabla_{\mathcal{F}^t} L(\theta_t)\|_2,$$

where $\nabla_{\mathcal{F}^t} L(\theta_t) = [\frac{1}{n}\sum_{i=1}^n (f(\langle x_i, \theta_t\rangle) - \tilde{y}_i) f'(\langle x_i, \theta_t\rangle) x_i^T]_{\mathcal{F}^t}$. Plugging $\tilde{y}_i = f(\langle \theta^*, x_i\rangle) + \sigma_i + z_i$, where $z_i \sim \mathcal{N}(0, \tau^2)$, and $\tau^2 = \frac{32C^2 \log(1.25/\delta)}{\epsilon^2}$ into the above equality, we get

$$\|\tilde{\theta}_{t+\frac{1}{2}} - \theta^*\|_2 \leq$$
$$\|\Delta_t - \eta[\frac{1}{n}\sum_{i=1}^n (f(\langle x_i, \theta_t\rangle) - f(\langle x_i, \theta^*\rangle)) f'(\langle x_i, \theta_t\rangle) x_i^T]_{\mathcal{F}^t}\|_2 +$$
$$\eta\sqrt{|\mathcal{F}^t|}[|\frac{1}{n}\sum_{i=1}^n f'(\langle x_i, \theta_t\rangle)\sigma_i x_i^T|_\infty + |\frac{1}{n}\sum_{i=1}^n f'(\langle x_i, \theta_t\rangle) z_i x_i^T|_\infty].$$

Define the following terms

$$A^t = \|\Delta_t - \eta[\frac{1}{n}\sum_{i=1}^n (f(\langle x_i, \theta_t\rangle) - f(\langle x_i, \theta^*\rangle)) f'(\langle x_i, \theta_t\rangle) x_i^T]_{\mathcal{F}^t}\|_2$$
$$B^t = \eta\sqrt{|\mathcal{F}^t|}|\frac{1}{n}\sum_{i=1}^n f'(\langle x_i, \theta_t\rangle)\sigma_i x_i^T|_\infty,$$
$$C^t = \eta\sqrt{|\mathcal{F}^t|}|\frac{1}{n}\sum_{i=1}^n f'(\langle x_i, \theta_t\rangle) z_i x_i^T|_\infty.$$

We first bound B^t. Since each $x_i \in \text{Uniform}\{+1, -1\}^p$, which is sub-Gaussian with 1, we know that for each coordinate $j \in [p]$, $\frac{1}{n}\sum_{i=1}^n f'(\langle x_i, \theta_t\rangle)\sigma_i x_{i,j}$ is sub-Gaussian with $\sigma^2 = \frac{1}{n^2}\sum_{i=1}^n f'^2(\langle x_i, \theta_t\rangle)\sigma_i^2 \leq \frac{b^2 C^2}{n}$. Thus, by Lemma 2 we have

$$\Pr[|\frac{1}{n}\sum_{i=1}^n f'(\langle x_i, \theta_t\rangle)\sigma_i x_i^T|_\infty \leq O(\frac{\sqrt{\log p} bC}{\sqrt{n}})] \geq 1 - \frac{1}{p^c}.$$

This means that with probability at least $1-\frac{2}{p^c}$, we have

$$B^t \leq O(\eta\sqrt{2s+s^*}\frac{\sqrt{\log p}bC}{\sqrt{n}}). \tag{32}$$

Similarly, for C^t we have that with probability at least $1-\frac{1}{p^c}$, the following holds

$$|\frac{1}{n}\sum_{i=1}^{n} f'(\langle x_i, \theta_t\rangle) z_i x_i^T|_\infty \leq O(\frac{b\sqrt{\log p}\sqrt{\sum_{i=1}^{n} z_i^2}}{n}).$$

Since z_i is Gaussian with variance τ^2, we know that $\sum_{i=1}^{n} z_i^2 = \tau^2 \sum_{i=1}^{n} r_i^2$, where $\sum_{i=1}^{n} r_i^2$ is a χ^2-distribution with parameter n.

By the above concentration bound for χ^2-distribution and Lemma 6, we have $\sum_{i=1}^{n} z_i^2 \leq 5\tau^2\, n$ with probability at least $1-\exp(-n)$. Thus,

$$C^t \leq \eta\sqrt{2s+s^*}O(\frac{b\sqrt{\log p}\tau}{\sqrt{n}}) \tag{33}$$

with probability at least $1-\frac{1}{p^c}-\exp(-n)$.

For the term of A^t, the proof is the same as the one for A^{t-1} in the proof of Theorem 9, and thus we omit it from here.

By (32) and (33) and plugging $\tau^2 = \frac{32C^2\log(1.25/\delta)}{\epsilon^2}$ into (33), we have the following with probability at least $1-\frac{2}{p^c}-\exp(-n)$

$$\|\tilde{\theta}_{t+\frac{1}{2}} - \theta^*\|_2 \leq \frac{2}{7}\|\Delta_t\|_2 + O(\frac{\sqrt{(2s+s^*)\log p}\log(1/\delta)bC}{n\epsilon}).$$

Putting all together, we have the following with probability at least $1-\frac{2}{p^c}-\exp(-n)$,

$$\|\Delta_{t+1}\| \leq \frac{1}{2}\|\Delta_t\|_2 + O(\frac{\sqrt{s^*\log p}\log(1/\delta)bC}{n\epsilon}).$$

Thus, we get the bound in Theorem 11 with probability at least $1-\frac{2T}{p}-T\exp(-n)$. For the linear case, since $f' \equiv 1$, (32) and (33) will be the same in each iteration, the probability for the linear case becomes $1-\frac{2}{p^c}-\exp(-n)$. □

References

[1] D. Wang and J. Xu, "On sparse linear regression in the local differential privacy model," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, in Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds. Long Beach, CA, USA: PMLR, Jun. 2019, pp. 6628–6637.

[2] L. A. Marascuilo and R. C. Serlin, *Statistical Methods for the Social and Behavioral Sciences*. San Francisco, CA, USA: Freeman, 1988.

[3] P. Bužková, "Linear regression in genetic association studies," *PLoS ONE*, vol. 8, no. 2, Feb. 2013, Art. no. e56976.

[4] P. Bühlmann and S. Van De Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Cham, Switzerland: Springer, 2011.

[5] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptogr. Conf.* Cham, Switzerland: Springer, 2006, pp. 265–284.

[6] J. Near, "Differential privacy at scale: Uber and Berkeley collaboration," in *Proc. Enigma (Enigma)*. Santa Clara, CA, USA: USENIX Association, 2018.

[7] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2014, pp. 1054–1067.

[8] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang, "Privacy loss in apple's implementation of differential privacy on MacOS 10.12," 2017, *arXiv:1709.02753*. [Online]. Available: https://arxiv.org/abs/1709.02753

[9] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. IEEE 54th Annu. Symp. Found. Comput. Sci.*, Oct. 2013, pp. 429–438.

[10] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Minimax optimal procedures for locally private estimation," *J. Amer. Stat. Assoc.*, vol. 113, no. 521, pp. 182–201, Jan. 2018.

[11] J. C. Duchi and F. Ruan, "The right complexity measure in locally private estimation: It is not the Fisher information," 2018, *arXiv:1806.05756*. [Online]. Available: http://arxiv.org/abs/1806.05756

[12] K. Zhang, Z. Yang, and Z. Wang, "Nonlinear structured signal estimation in high dimensions via iterative hard thresholding," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 258–268.

[13] Z. Yang, Z. Wang, H. Liu, Y. Eldar, and T. Zhang, "Sparse nonlinear regression: Parameter estimation under nonconvexity," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2472–2481.

[14] K. Chaudhuri and D. Hsu, "Sample complexity bounds for differentially private learning," in *Proc. 24th Annu. Conf. Learn. Theory*, 2011, pp. 155–186.

[15] A. Beimel, K. Nissim, and U. Stemmer, "Private learning and sanitization: Pure vs. approximate differential privacy," in *APPROX*. Cham, Switzerland: Springer, 2013, pp. 363–378.

[16] Y. Chen, A. Machanavajjhala, J. P. Reiter, and A. F. Barrientos, "Differentially private regression diagnostics," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 81–90.

[17] A. F. Barrientos, J. P. Reiter, A. Machanavajjhala, and Y. Chen, "Differentially private significance tests for regression coefficients," *J. Comput. Graph. Statist.*, vol. 28, no. 2, pp. 1–24, Jun. 2019.

[18] Y. Wang, "Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain," in *Proc. 34th Conf. Uncertainty Artif. Intell. (UAI)*, Monterey, CA, USA, Aug. 2018, pp. 93–103.

[19] O. Sheffet, "Differentially private ordinary least squares," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3105–3114.

[20] D. Kifer, A. Smith, and A. Thakurta, "Private convex empirical risk minimization and high-dimensional regression," in *Proc. Conf. Learn. Theory*, 2012, pp. 1–25.

[21] A. Smith, A. Thakurta, and J. Upadhyay, "Is interaction necessary for distributed private learning?" in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 58–77.

[22] A. G. Thakurta and A. Smith, "Differentially private feature selection via stability arguments, and the robustness of the lasso," in *Proc. Conf. Learn. Theory*, 2013, pp. 819–850.

[23] K. Talwar, A. G. Thakurta, and L. Zhang, "Nearly optimal private lasso," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3025–3033.

[24] G. Bernstein and D. R. Sheldon, "Differentially private Bayesian linear regression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 523–533.

[25] T. Tony Cai, Y. Wang, and L. Zhang, "The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy," 2019, *arXiv:1902.04495*. [Online]. Available: http://arxiv.org/abs/1902.04495

[26] K. Zheng, W. Mou, and L. Wang, "Collect at once, use effectively: Making non-interactive locally private learning possible," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 4130–4139.

[27] D. Wang, M. Gaboardi, and J. Xu, "Empirical risk minimization in non-interactive local differential privacy revisited," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montréal, QC, Canada, Dec. 2018, pp. 973–982.

[28] M. Huai, D. Wang, C. Miao, J. Xu, and A. Zhang, "Pairwise learning with differential privacy guarantees," in *Proc. 34th AAAI Conf. Artif. Intell.*, New York, NY, USA, Feb. 2020, pp. 694–701.

[29] D. Wang, A. Smith, and J. Xu, "Noninteractive locally private learning of linear models via polynomial approximations," in *Proc. Algorithmic Learn. Theory*, 2019, pp. 897–902.

[30] D. Wang and J. Xu, "Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI)*, Honolulu, HI, USA, Jan./Feb. 2019, pp. 1182–1189.

[31] D. Wang, M. Ye, and J. Xu, "Differentially private empirical risk minimization revisited: Faster and more general," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 2719–2728.

[32] P. Jain, A. Tewari, and P. Kar, "On iterative hard thresholding methods for high-dimensional m-estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 685–693.

[33] D. Wang and J. Xu, "Lower bound of locally differentially private sparse covariance matrix estimation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4788–4794.

[34] J. Acharya, Z. Sun, and H. Zhang, “Differentially private Assouad, Fano, and Le Cam,” 2020, *arXiv:2004.06830*. [Online]. Available: http://arxiv.org/abs/2004.06830

[35] J. Acharya, C. L. Canonne, and H. Tyagi, “Inference under information constraints I: Lower bounds from chi-square contraction,” 2018, *arXiv:1812.11476*. [Online]. Available: http://arxiv.org/abs/1812.11476

[36] G. Raskutti, M. J. Wainwright, and B. Yu, “Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls,” *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6976–6994, Oct. 2011.

[37] M. Bun, J. Ullman, and S. Vadhan, “Fingerprinting codes and the price of approximate differential privacy,” *SIAM J. Comput.*, vol. 47, no. 5, pp. 1888–1938, Jan. 2018.

[38] M. Joseph, J. Mao, S. Neel, and A. Roth, “The role of interactivity in local differential privacy,” 2019, *arXiv:1904.03564*. [Online]. Available: http://arxiv.org/abs/1904.03564

[39] R. Bassily, A. Smith, and A. Thakurta, “Differentially private empirical risk minimization: Efficient algorithms and tight error bounds,” 2014, *arXiv:1405.7085*. [Online]. Available: http://arxiv.org/abs/1405.7085

[40] M. Bun, J. Nelson, and U. Stemmer, “Heavy hitters and the structure of local privacy,” in *Proc. 37th ACM SIGMOD-SIGACT-SIGAI Symp. Princ. Database Syst.*, May 2018, pp. 435–447.

[41] I. Mironov, “Rényi differential privacy,” in *Proc. IEEE 30th Comput. Secur. Found. Symp. (CSF)*, Aug. 2017, pp. 263–275.

[42] M. Bun and T. Steinke, “Concentrated differential privacy: Simplifications, extensions, and lower bounds,” in *Proc. Theory Cryptogr. Conf.* Cham, Switzerland: Springer, 2016, pp. 635–658.

[43] T. Blumensath and M. E. Davies, “Iterative hard thresholding for compressed sensing,” *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, Nov. 2009.

[44] H. Rauhut, “Compressive sensing and structured random matrices,” *Theor. Found. Numer. Methods Sparse Recovery*, vol. 9, pp. 1–92, Jan. 2010.

[45] J. Ge, Z. Wang, M. Wang, and H. Liu, “Minimax-optimal privacy-preserving sparse PCA in distributed systems,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 1589–1598.

[46] A. Mcmillan and A. C. Gilbert, “Local differential privacy for physical sensor data and sparse recovery,” in *Proc. 52nd Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2018, pp. 1–6.

[47] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, “A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers,” *Stat. Sci.*, vol. 27, no. 4, pp. 538–557, Nov. 2012.

[48] R. Bassily, A. Smith, and A. Thakurta, “Private empirical risk minimization: Efficient algorithms and tight error bounds,” in *Proc. IEEE 55th Annu. Symp. Found. Comput. Sci.*, Oct. 2014, pp. 464–473.

[49] S. Bahmani, B. Raj, and P. Boufounos, “Greedy sparsity-constrained optimization,” *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 807–841, Jan. 2013.

[50] K. Talwar, A. Thakurta, and L. Zhang, “Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry,” 2014, *arXiv:1411.5417*. [Online]. Available: http://arxiv.org/abs/1411.5417

[51] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, Jan. 2011. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[52] D. Dheeru and E. K. Taniskidou. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[53] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” 2010, *arXiv:1011.3027*. [Online]. Available: http://arxiv.org/abs/1011.3027

[54] B. Laurent and P. Massart, “Adaptive estimation of a quadratic functional by model selection,” in *Proc. Ann. Statist.*, 2000, pp. 1302–1338.

[55] M. Abadi *et al.*, “Deep learning with differential privacy,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.