On Sparse Linear Regression in the Local Differential Privacy Model

Di Wang [1] **Jinhui Xu** [1]

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 depending 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.

1. Introduction

Linear regression is a fundamental and classical tool for data analysis, and finds numerous applications in social sciences (Marascuilo & Serlin, 1988), genomics research (Bůžková, 2013) and signal recovery (Bühlmann & Van De Geer, 2011). 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 (Dwork et al., 2006) 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 (Near, 2018; Erlingsson et al., 2014; Near, 2018; Tang et al., 2017).

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 (Erlingsson et al., 2014), and Apple's iOS-10 (Near, 2018; Tang et al., 2017).

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 differ-

[1]Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, USA. Emails: *{dwang45,jinhui}@buffalo.edu*. Correspondence to: Di Wang <dwang45@buffalo.edu>.

Proceedings of the 36^{th} *International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019.

ent settings, the statistical learning and the statistical estimation settings. For the first setting, the measurement is based on the optimization error, *i.e.* $F(\theta^{\text{priv}}) - \min_{\theta \in C} F(\theta)$, 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 $x \sim \text{Uniform}\{+1, -1\}^p$.

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 (Duchi et al., 2013; 2018; Duchi & Ruan, 2018). 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 (Zhang et al., 2018; Yang et al., 2016).

- 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 (Chaudhuri & Hsu, 2011; Beimel et al., 2013)). 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 $\Omega(\frac{s \log p}{n\epsilon^2})$.

Due to space limit, some background, all the proofs and additional experiments are left to the Supplemental Material.

2. 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 (Chen et al., 2016; Barrientos et al., 2019; Wang, 2018; Sheffet, 2017; Kifer et al., 2012; Smith et al., 2017; Thakurta & Smith, 2013). Below, we focus only on those with theoretical guarantees on the error.

For the central model, (Wang, 2018) 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. However, the lower bound of the estimation error is still unknown. There are also other works on this problem (we refer the reader to the Related Work section in (Wang, 2018) 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 (Kifer et al., 2012; Talwar et al., 2015); but all of them consider only the optimization error.

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 (Smith et al., 2017; Zheng et al., 2017; Duchi et al., 2018; 2013). (Duchi et al., 2013) 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. (Smith et al., 2017) 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 (Zheng et al., 2017; Wang et al., 2018) 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 (Wang et al., 2019; Wang & Xu, 2019; Wang et al., 2017; 2018; Duchi et al., 2013; Jain et al., 2014) under different settings. However, none of these considered the non-convex constraint case.

3. Preliminaries

In this section, we introduce some definitions that will be used throughout the paper. More details can be found in Section A of Supplemental Material or (Duchi et al., 2018).

3.1. 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 $\hat{\theta}$.

3.2. 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 (Duchi et al., 2013).

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 $Q_i(Z_i \mid X_i = x_i, Z_{1:i} = z_{i:i})$. 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 ((Duchi et al., 2013))**.** *For a given privacy parameter $\epsilon > 0$, the random variable Z_i is an ϵ 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:*

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

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

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

We say that the privacy mechanism $Q = \{Q_i\}_{i=1}^n$ is ϵ-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:*

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

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:

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

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

4. Keeping the Whole Dataset Private

4.1. 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.

We focus on the following collection of samples $(x, y) \in \{+1, -1\}^p \times \mathbb{R}$:

$$\mathcal{P}_{s,p,C} = \{P_{\theta,\sigma} \mid x \sim \text{Uniform}\{+1,-1\}^p, y = \langle\theta, x\rangle + \sigma, \text{ where } \sigma \text{ is the random noise satisfying the condition of } |\sigma| \le C \text{ for some constant } C > 0, \|\theta\|_2 \le 1, \|\theta\|_0 \le s\}. \tag{1}$$

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

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 (Raskutti et al., 2011). That is, here x is assumed to follow a uniform distribution (which is an often adopted assumption in estimating lower bounds in differential privacy (Bun et al., 2018b)) 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.

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 $P_{\theta,\sigma} \in \mathcal{P}_{1,p,C}$. 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 $\Theta(\min\{1, \frac{p\log p}{n\epsilon^2}\})$.

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,*

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

Moreover, there exists an (ϵ, δ) non-interactive LDP algorithm whose output achieves an upper bound of $O(\min\{1, \frac{p\log p}{n\epsilon^2}\})$ under the squared ℓ_2-norm with probability at least $1 - \exp(-\Omega(p))$.

With the above theorem, 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 (Smith et al., 2017). 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 (Duchi et al., 2013; Smith et al., 2017).

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,*

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

Remark 1. *Since the lower bound of the non-private minimax risk is $O(\frac{\log p}{n})$ (Raskutti et al., 2011), 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 4.2 and leave the problem of finding a tighter lower bound as future research.*

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 $i \ne j$, 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,*

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

4.2. 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 (Blumensath & Davies, 2009). We consider the following more general optimization problem, with the intention to extend it to other problems (see Section 6),

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

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 (Duchi et al., 2018) on their current gradients $x_i^T(\langle x, \theta_{i-1}\rangle - y)$ (see Supplemental Material for details). 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 $\mathbb{B}_1$.

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 $n \geq c\delta'^{-2}(s^* \log p + \ln(1/\epsilon))$ for some universal constant c (see Theorem 2.12 in (Rauhut, 2010)). 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 ϵ, step size η. Set $\theta_0 = 0$. $s = 8s^*$.

1: 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 $S_t = S_t \bigcup \{t\lfloor \frac{n}{T} \rfloor, \cdots, n\}$.
2: **for** $t = 1, 2, \cdots, T$ **do**
3: The server sends θ_{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 $z_i = \mathcal{R}_\epsilon(\nabla_i)$, where $\mathcal{R}_\epsilon$ is the randomizer defined in (Smith et al., 2017) or (Duchi et al., 2018) and send back to the server.
4: The server compute $\tilde{\nabla}_{t-1} = \frac{1}{|S_t|}\sum_{i\in S_t} z_i$.
5: Perform the gradient descent updating $\tilde{\theta}_t = \theta_{t-1} - \eta\tilde{\nabla}_{t-1}$.
6: $\theta'_t = \text{Trunc}(\tilde{\theta}_{t-1}, s)$.
7: $\theta_t = \arg_{\theta\in\mathbb{B}_1}\|\theta - \theta'_t\|_2^2$.
8: **end for**
9: Return θ_T

Theorem 4. *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*

$$\|\theta_T - \theta^*\|_2 \leq (\frac{1}{2})^T\|\theta^*\|_2 + O(\frac{C\sqrt{p\log p}\sqrt{T}\sqrt{s^*}}{\sqrt{n\epsilon}}), \quad (6)$$

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

Note that Theorem 4 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 a near optimal upper bound.

We notice that recently Ge et al. (2018) 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.

5. 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 (Chaudhuri & Hsu, 2011; Beimel et al., 2013) for private PAC. 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 (McMillan & Gilbert, 2018). In this case, we can actually assume that $\{x_i\}_{i=1}^n \in (\{+1,-1\}^p)^n$ are fixed, and the collection of probability $\mathcal{P}_{s,p,C}$ in (1) is now reduced to the following model:

$$\mathcal{P}'_{s,p,C} = \{P_{\theta,\sigma}(y_1,\cdots,y_n) \mid y_i = \langle\theta^*, x_i\rangle + \sigma_i, \text{where } \|\theta\|_0 \le s, \|\theta\|_2 \le 1 \text{ and the random noise } |\sigma_i| \le C\}. \quad (7)$$

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 (Bun et al., 2018a) to obtain an ϵ non-interactive LDP algorithm.

Algorithm 2 Label-LDP-IHT

Input: Public dataset $\{x_i\}_{i=1}^n$, private $\{y_i\}_{i=1}^n \in P_{\theta^*,\sigma}$, where $P_{\theta^*,\sigma} \in \mathcal{P}'_{s^*,p,C}$, ϵ,δ are privacy parameters, T is the number of iteration, η is the step size, and $s = 8s^*$. Set $\theta_0 = 0$.

1: **for** Each $i \in [n]$ **do**
2: Denote $\tilde{y}_i = y_i + z_i$, where $z_i \sim \mathcal{N}(0,\sigma_1^2)$, $\sigma_1^2 = \frac{32C^2\ln(1.25/\delta)}{\epsilon^2}$.
3: **end for**
4: **for** $t = 0,1,\cdots,T-1$ **do**
5: $\tilde{\theta}_{t+1} = \theta_t - \eta(\frac{1}{n}\sum_{i=1}^n(\tilde{y}_i - \langle x_i,\theta_t\rangle)x_i^T)$.
6: $\theta'_{t+1} = \text{Trunc}(\tilde{\theta}_{t+1}, s)$.
7: $\theta_{t+1} = \arg_{\theta\in\mathbb{B}_1}\|\theta - \theta'_{t+1}\|_2^2$.
8: **end for**
9: Return θ_T.

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 $\|v\|_0 \le 2s+s^*$, there exists a constant δ' which satisfies $(1-\delta')\|v\|^2 \le \frac{1}{n}\|Xv\|_2^2 \le (1+\delta')\|v\|_2^2$.*

Theorem 5. *For any $0<\epsilon\le 1$ 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 $0<\delta'\le\frac{2}{7}$, 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 $1-\exp(-n)-\frac{2}{p^c}$*

$$\|\theta_T - \theta^*\|_2 \le (\frac{1}{2})^T\|\theta^*\|_2 + O(\frac{C\log(1/\delta)\sqrt{s^*\log p}}{\sqrt{n}\epsilon}). \quad (8)$$

Note that if $T = O(\log\frac{\sqrt{n}\epsilon}{C\sqrt{s^*\log p}})$ in (8), we have $\|\theta_T - \theta^*\|_2^2 \le O(C^2\frac{s\log p}{n\epsilon^2})$. 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 (Raskutti et al., 2011).

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 $\lambda = O(\sqrt{\frac{\log p}{n\epsilon^2}})$ (Negahban et al., 2012). However, we would like to point out that our algorithm is more practical and can be extended to the case of non-linear measurements (see Section D of the Supplemental Material).

With the above theorem, a natural question is to determine whether the upper bound in Theorem 5 can be further improved. The following theorem (adopted from (Raskutti et al., 2011)) suggests that it is actually tight as the ϵ non-interactive local private minimax risk (under the $\|\cdot\|^2$ metric) is lower bounded by $\Omega(\frac{C^2s^*\log p}{n\epsilon^2})$.

Theorem 6. *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*

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

6. 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 under some sparsity constraint and the sparse regression problem with non-linear monotone measurements. Due to space limit, we discuss here only DP-ERM. Details on other problems can be found in Section D of the Supplemental Material. We start with reviewing some definitions of DP-ERM.

Definition 3 (DP-ERM (Bassily et al., 2014))**.** *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 $C \subseteq \mathbb{R}^p$, 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 (Dwork et al., 2006). The utility of the algorithm is measured by the expected excess empirical risk, that is $\mathbb{E}_{\mathcal{A}}[L(x^{priv}; D)] - \min_{x\in C} L(x; D)$, where the expectation of $\mathcal{A}$ is taking over all the randomness of the algorithm.*

In this section, we consider the sparsity-constrained (ϵ, δ) DP-ERM problem. That is, the constraint set C is defined as $C = \{x : \|x\|_0 \le k\}$, 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 (Bahmani et al., 2013).

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

Definition 4 (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 $\|x - x'\|_0 \le s$, we have $f(x) - f(x') - \langle \nabla f(x'), x - x' \rangle \ge \frac{\rho_s}{2}\|x - x'\|_2^2$.*

Definition 5 (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 $\|x - x'\|_0 \le s$, we have $f(x) - f(x') - \langle \nabla f(x'), x - x' \rangle \le \frac{\ell_s}{2}\|x - x'\|_2^2$.*

Assumption 3. *Denote $x^* = \arg\min_{x\in C} L(x; D)$ 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 7. *Under Assumption 3, for any $1 \ge \epsilon, \delta > 0$, there exists a constant $c > 0$ which makes Algorithm 3 (ϵ, δ)-DP. Moreover, if the sparsity level $k \ge (1 + 64\kappa_s^2)k^*$, 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*

$$\mathbb{E}L(x_T; D) - L(x^*; D) \le O(\frac{\log n \log pk^* \log \frac{1}{\delta}}{n^2\epsilon^2}), \quad (9)$$

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

Remark 2. *We note that the upper bound in (9) depends only logarithmically on p (i.e., $\log p$), rather than polynomially (i.e., Poly(p)) as in general DP-ERM with (strongly) convex loss functions (Wang et al., 2017; Bassily et al., 2014). This means that we have obtained a non-trivial upper bound for the high dimensional case ($p \gg n$) of the problem. Recently, (Talwar et al., 2015; 2014) also studied the case of high dimensional DP-ERM with specified constraint set. However, there are considerable differences. Firstly, the (Talwar et al., 2015) paper considers only linear regression and ℓ_1-norm Lipshitz with the constraint set restricted to an ℓ_1-norm ball. Secondly, the (Talwar et al., 2014) paper shows that its upper bound depends only on the Gaussian width of the underlying constraint set, instead of 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.*

Algorithm 3 DP-IHT

Input: Initial point x_0, learning rate η, empirical risk $L(x; D)$, privacy parameters $1 > \epsilon, \delta > 0$, and iteration number T.

1: **for** $t = 0, 1, \cdots, T-1$ **do**
2: 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.
3: Let $x_{t+1} = \text{Trun}(\tilde{x}_{t+1}, k)$.
4: **end for**
5: Return x_T.

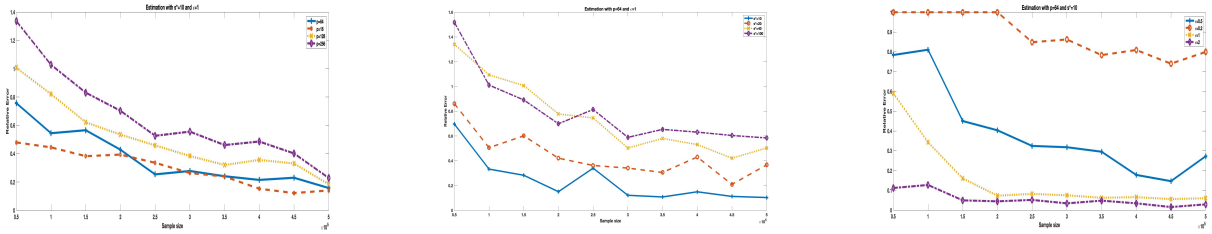
It would be interesting to find a general condition on the constraint set such that the upper bound of the problem can be independent of Poly(p). Also, we note that to achieve the bound in (9), the gradient complexity of Algorithm 3 needs to be $\tilde{O}(n\kappa_s)$, which is quite large. We leave it as an open problem to make it more practical.

7. Experiments

7.1. Experiments on Sparse Linear Regression

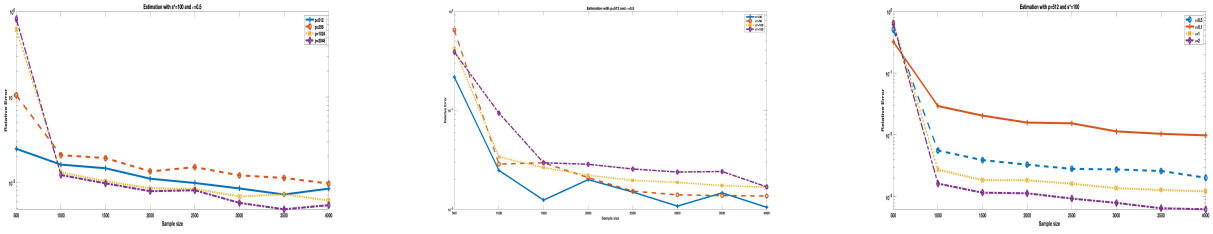
Data Generation Our data generation process is similar to the one in (Jain et al., 2014). We first fix a parameter vector θ^* by randomly choosing s^* coordinates, with each of them sampled independently from a uniform distribution in interval $[0, 1]$, and setting the remaining coordinates/entries to 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.

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



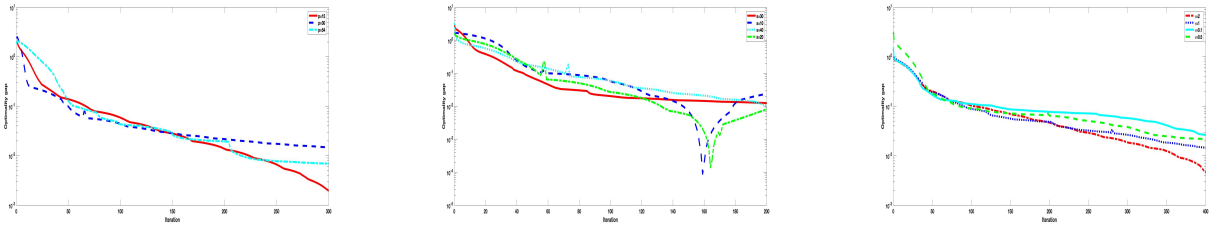
(a) Relative error w.r.t dimensionality (b) Relative error w.r.t sparsity level (c) Relative error w.r.t privacy level

Figure 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

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



(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 $s = 10$

Figure 3. Experimental results on ℓ_0-constrained logistic regression under (ϵ, δ)-DP (Algorithm 3).

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 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.

7.2. Experiments on Sparsity-constrained DP-ERM

In this section, we test Algorithm 3 on a real world dataset Covertype. Particularly, we study 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 (Jain et al., 2014) 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 (Bun & Steinke, 2016) to achieve the (ϵ, δ)-DP.

From Figure 3, 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 7.

Acknowledgements

The research of this work was supported in part by NSF through grant CCF-1716400. Part of this research was done

while Di Wang was visiting Boston University and Harvard University's Privacy Tools Project. We are grateful to Adam Smith for valuable discussions in the early stages of this work.

References

Bahmani, S., Raj, B., and Boufounos, P. T. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14(Mar):807–841, 2013.

Barrientos, A. F., Reiter, J. P., Machanavajjhala, A., and Chen, Y. Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, pp. 1–24, 2019.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 464–473. IEEE, 2014.

Beimel, A., Nissim, K., and Stemmer, U. Private learning and sanitization: Pure vs. approximate differential privacy. In *APPROX*, pp. 363–378. Springer, 2013.

Blumensath, T. and Davies, M. E. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

Bühlmann, P. and Van De Geer, S. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–658. Springer, 2016.

Bun, M., Nelson, J., and Stemmer, U. Heavy hitters and the structure of local privacy. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 435–447. ACM, 2018a.

Bun, M., Ullman, J., and Vadhan, S. Fingerprinting codes and the price of approximate differential privacy. *SIAM Journal on Computing*, 47(5):1888–1938, 2018b.

Bŭžková, P. Linear regression in genetic association studies. *PLoS One*, 8(2):e56976, 2013.

Chaudhuri, K. and Hsu, D. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 155–186, 2011.

Chen, Y., Machanavajjhala, A., Reiter, J. P., and Barrientos, A. F. Differentially private regression diagnostics. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pp. 81–90. IEEE, 2016.

Duchi, J. C. and Ruan, F. The right complexity measure in locally private estimation: It is not the fisher information. *CoRR*, abs/1806.05756, 2018. URL http://arxiv.org/abs/1806.05756.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pp. 429–438. IEEE, 2013.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Erlingsson, Ú., Pihur, V., and Korolova, A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067. ACM, 2014.

Ge, J., Wang, Z., Wang, M., and Liu, H. Minimax-optimal privacy-preserving sparse pca in distributed systems. In *International Conference on Artificial Intelligence and Statistics*, pp. 1589–1598, 2018.

Jain, P., Tewari, A., and Kar, P. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pp. 685–693, 2014.

Kifer, D., Smith, A., and Thakurta, A. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pp. 25–1, 2012.

Marascuilo, L. A. and Serlin, R. C. *Statistical methods for the social and behavioral sciences.* WH Freeman/Times Books/Henry Holt & Co, 1988.

McMillan, A. and Gilbert, A. C. Local differential privacy for physical sensor data and sparse recovery. In *Information Sciences and Systems (CISS), 2018 52nd Annual Conference on*, pp. 1–6. IEEE, 2018.

Near, J. Differential privacy at scale: Uber and berkeley collaboration. In *Enigma 2018 (Enigma 2018)*, Santa Clara, CA, 2018. USENIX Association.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Raskutti, G., Wainwright, M. J., and Yu, B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57 (10):6976–6994, 2011.

Rauhut, H. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9:1–92, 2010.

Sheffet, O. Differentially private ordinary least squares. In *International Conference on Machine Learning*, pp. 3105–3114, 2017.

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

Talwar, K., Thakurta, A., and Zhang, L. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*, 2014.

Talwar, K., Thakurta, A. G., and Zhang, L. Nearly optimal private lasso. In *Advances in Neural Information Processing Systems*, pp. 3025–3033, 2015.

Tang, J., Korolova, A., Bai, X., Wang, X., and Wang, X. Privacy loss in apple's implementation of differential privacy on macos 10.12. *CoRR*, abs/1709.02753, 2017.

Thakurta, A. G. and Smith, A. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pp. 819–850, 2013.

Wang, D. and Xu, J. Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view. *Thirty-Third AAAI Conference on Artificial Intelligence, (AAAI-19), Honolulu, Hawaii, USA, January 27-February 1, 2019*, 2019.

Wang, D., Ye, M., and Xu, J. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 2719–2728, 2017.

Wang, D., Gaboardi, M., and Xu, J. Empirical risk minimization in non-interactive local differential privacy revisited. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pp. 973–982, 2018.

Wang, D., Smith, A., and Xu, J. Noninteractive locally private learning of linear models via polynomial approximations. In *Algorithmic Learning Theory*, pp. 897–902, 2019.

Wang, Y. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pp. 93–103, 2018.

Yang, Z., Wang, Z., Liu, H., Eldar, Y., and Zhang, T. Sparse nonlinear regression: Parameter estimation under nonconvexity. In *International Conference on Machine Learning*, pp. 2472–2481, 2016.

Zhang, K., Yang, Z., and Wang, Z. Nonlinear structured signal estimation in high dimensions via iterative hard thresholding. In *International Conference on Artificial Intelligence and Statistics*, pp. 258–268, 2018.

Zheng, K., Mou, W., and Wang, L. Collect at once, use effectively: Making non-interactive locally private learning possible. In *International Conference on Machine Learning*, pp. 4130–4139, 2017.