Supplemental Material For "On Sparse Linear Regression in the Local Differential Privacy Model"

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

A. Background

A.1. Private Le Cam and Fano Method

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 1 (Prop. 2 in (Duchi et al., 2018))**.** *Given any 2δ-separated set $\{P_v, v \in \mathcal{V}\}$, and $\alpha \in (0, \frac{23}{35}]$, the ϵ non-interactive private minimax risk satisfies the following inequality*

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

where $C_\infty^{Nint}(\{P_v\}_{v\in\mathcal{V}}) = \frac{1}{|\mathcal{V}|}\sup_{\gamma\in\mathbb{B}_\infty}\sum_{v\in\mathcal{V}}(\psi_v(\gamma))^2$, $\mathbb{B}_\infty$ 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 2 (Corollaries 2 and 4 in (Duchi et al., 2013))**.** *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 $\alpha \in (0, \frac{23}{35}]$,*

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

The ϵ non-interactive private minimax risk satisfies

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

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.

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

Proceedings of the 36^{th} *International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

Lemma 3 (Theorem 2 in (Duchi & Ruan, 2018))**.** *For any $\epsilon \in (0, \frac{23}{35}]$, the ϵ sequential private minimax risk in the loss function L satisfies the following inequality*

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

where

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

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

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

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

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

A.2. Technical Lemmas

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

Definition 1. *A random variable X is said to be sub-Gaussian with σ^2 if $\mathbb{E}(X) = 0$ and*

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

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 $a_1X_1 + \cdots + a_nX_n$ 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 4 ((Vershynin, 2010))**.** *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*

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

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

Lemma 5 ((Jain et al., 2014))**.** *For any $\theta \in \mathbb{R}^k$ and an integer $s \leq k$, if $\theta_t = Trunc(\theta, s)$ 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 6. *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 7 (Theorem 2.12 in (Rauhut, 2010))**.** *Let $X \in \{+1, -1\}^{n\times p}$ be a Bernoulli Random Matrix and $\epsilon, \delta \in (0, 1)$. 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

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

Lemma 8 ((Laurent & Massart, 2000))**.** *If $z \sim \chi^2_n$, then*

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

We now review the randomizer $\mathcal{R}_\epsilon(\cdot)$ in (Smith et al., 2017; Duchi et al., 2018) and its properties .

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

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

Lemma 9 ((Smith et al., 2017))**.** *Given any vector $x \in \mathbb{R}^p$, the randomizer $\mathcal{R}_\epsilon(x)$ defined above is a sub-Gaussian random vector with variance $\sigma^2 = O(\frac{p}{\epsilon^2})$ and $\mathbb{E}(\mathcal{R}_\epsilon(x)) = x$.*

B. Some Omitted Proofs

Proof of Theorem 1. We first prove the lower bound.

The main idea of the proof is :

- Find an index set $\mathcal{V}$ which corresponds to a 2δ-separated set $\{P_v, v \in \mathcal{V}\}$.
- Obtain an upper bound on $\mathcal{C}_\infty(\{P_v\}_{v\in\mathcal{V}})$, use Lemma 1 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} as

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

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

$$p_{\theta_v}((x, y)) = \frac{1 + y\langle x, \theta_v\rangle}{2^{p+1}} \text{ for } (x, y) \in \{+1, -1\}^{p+1}. \tag{3}$$

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

Now we show our main lemma used in the proof.

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

$$C^{Nint}_\infty(\{P_v\}_{v\in\mathcal{V}}) \leq \frac{\delta^2}{p}. \tag{4}$$

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

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

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

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

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

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

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 1 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.

Next, we prove the upper bound. We propose an (ϵ,δ) non-interactive LDP algorithm. Note that by using the protocol in (Bun et al., 2018a) we can transform it into ϵ non-interactive LDP algorithm.

The idea of the proof is that each user perturbs the data by using Gaussian mechanism and sends it to the server, and then the server performs the following LASSO:

$$\min_{\theta\in\mathbb{R}^p}\frac{1}{2n}\sum_{i=1}^n(\langle\tilde{x}_i,\theta\rangle-\tilde{y}_i)^2+\frac{\lambda}{2}\|\theta\|_1$$

for some λ, see Algorithm 1.

Theorem 1. *For any $0<\epsilon,\delta<1$, Algorithm 1 is (ϵ,δ) non-interactive LDP. Furthermore, if $n\geq\Omega(\frac{p\log p}{\epsilon^2})$ and $X\in\{-1,1\}^{n\times p}$ satisfies the Assumption 1 with some sparsity and parameter $\delta' = O(1)$ (by Lemma 7 we know this satisfies with* *probability at least* $1-\exp(\Omega(-p)))$*, then by setting* $\lambda = O(\frac{\sqrt{p\ln\frac{1}{\delta}}C}{\epsilon}\sqrt{\frac{\log p}{n}})$*, with high probability (at least* $1-\exp(-\Omega(p)))$*, the output* $\tilde{\theta}$ *satisfies:*

$$\|\tilde{\theta}-\theta^*\|_2^2 \leq O(\frac{p\ln\frac{1}{\delta}C^2}{n\epsilon^2}).$$

Proof. By the construction of $\{\tilde{x}_i\}_{i=1}^n$ and $\{\tilde{y}_i\}_{i=1}^n$, we know

$$\tilde{y}_i = \langle x_i, \theta^*\rangle + \langle z_i, \theta^*\rangle + s_i + w_i. \tag{5}$$

From the definition and the fact that $\|\theta^*\|_2 \leq 1$, we know that the term $\langle z_i, \theta^*\rangle + s_i + w_i$ can be seen as a sub-Gaussian noise with the variance $\sigma^2 = O(\frac{p\ln\frac{1}{\delta}C^2}{\epsilon^2})$. By the definition and the assumption of X and Corollary 2 in (Negahban et al., 2012), we can the get the proof. □

Algorithm 1 Local Gaussian-LASSO

Input: Private data records $\{(x_i, y_i)\}_{i=1}^n \sim P_{\theta^*,\sigma}$, where $P_{\theta^*,\sigma} \in \mathcal{P}_{1,p,C}$, privacy parameters ϵ, δ.

```
1: for Each $i \in [n]$ do
2:     Let $\tilde{x} = x_i + z_i$, where $\sigma \sim \mathcal{N}(0, \sigma^2 I_p)$ for $\sigma^2 = \frac{32p\ln\frac{1.25}{\delta}}{\epsilon^2}$. And $\tilde{y}_i = y_i + s_i$, where $s_i \sim \mathcal{N}(0, \sigma_1^2)$, $\sigma_1^2 = \frac{32(C+1)^2\ln(1.25/\delta)}{\epsilon^2}$.
3: end for
4: for The server do
5:     Solve and return
           $$\tilde{\theta} = \arg\min_{\theta\in\mathbb{R}^p} \frac{1}{2n}\sum_{i=1}^n(\langle\tilde{x}_i, \theta\rangle - \tilde{y}_i)^2 + \frac{\lambda}{2}\|\theta\|_1.$$
       Where $\lambda = O(\frac{\sqrt{p\ln\frac{1}{\delta}}C}{\epsilon}\sqrt{\frac{\log p}{n}})$.
6: end for
```

□

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,

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

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 $d_L(P_0, P_1) = \frac{1}{2}\delta^2$. For the KL-divergence D_{kl} between M_0^n and $\bar{M}^n$, since

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

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 3 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 3, 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 Theorems 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. Follow from the fact that the linear model is a special case of the non-linear measurement. See the proof of Theorem 3 in Section D for the case $f(x) = x$ and $a = b = 1$.

□

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 4 in Section D 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 5 in Section D for the case $f(x) = x$ and $a = b = 1$. □

Proof of Theorem 7. For the guarantee of (ϵ, δ)-DP, it follows from the Moment accountant and composition theorem, see (Abadi et al., 2016; Wang et al., 2017) 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

$$\begin{aligned}
L(x_{t+1}) &\leq L(x_t) + \langle \nabla L(x_t), x_{t+1} - x_t\rangle + \frac{\ell_s}{2}\|x_{t+1} - x_t\|^2 \\
&\leq L(x_t) + \langle (g_t)_{\mathcal{I}}, (x_{t+1} - x_t)_{\mathcal{I}}\rangle + \frac{\ell_s}{2}\|x_{t+1} - x_t\|^2 + \|z_{t,\mathcal{I}}\|\|(x_{t+1} - x_t)_{\mathcal{I}}\|_2 \\
&= L(x_t) + \frac{1}{2\eta}\|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \frac{\eta\|g_{t,\mathcal{I}}\|^2}{2} - \frac{1-\eta\ell_s}{2\eta}\|x_{t+1} - x_t\|^2 + \|z_{t,\mathcal{I}}\|\|(x_{t+1} - x_t)_{\mathcal{I}}\|_2 \\
&= L(x_t) + \frac{1}{2\eta}(\|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \eta^2\|g_{t,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|^2) - \frac{\eta\|g_{t,\mathcal{I}^t\bigcup\mathcal{I}^*\|^2}}{2} - \frac{1-\eta\ell_s}{2\eta}\|x_{t+1} - x_t\|^2 \\
&+ \|z_{t,\mathcal{I}}\|\|(x_{t+1} - x_t)_{\mathcal{I}}\|_2, \qquad (6)
\end{aligned}$$

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 (Jain et al., 2014). 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

$$\eta^2\|g_{t,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|^2 = \|x_{t+1,\mathcal{I}\backslash(\mathcal{I}^t\bigcup\mathcal{I}^*)}\|^2 \geq \|x_{t,\mathcal{R}} - \eta g_{t,\mathcal{R}}\|^2. \qquad (7)$$

With (7) and the fact that $x_{t+1,\mathcal{R}} = 0$, we have

$$\begin{aligned}\|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \eta^2 \|g_{t,\mathcal{I}\backslash(\mathcal{I}^t \bigcup \mathcal{I}^*)}\|^2 &\leq \|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \|x_{t+1,\mathcal{R}} - x_{t,\mathcal{R}} + \eta g_{t,\mathcal{R}}\|^2 \\ &= \|x_{t+1,\mathcal{I}\backslash\mathcal{R}} - x_{t,\mathcal{I}\backslash\mathcal{R}} + \eta g_{t,\mathcal{I}\backslash\mathcal{R}}\|^2. \end{aligned} \tag{8}$$

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 (7) and Lemma 5 we have

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

Plugging this into (6), we get

$$\begin{aligned}L(x_{t+1}) &\leq L(x_t) + \frac{2k^*}{k+k^*}(L(x^* - L(x_t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)}\|x^* - x_t\|^2 + \frac{\eta k^*}{k+k^*}\|g_{t,\mathcal{I}\backslash(\mathcal{I}^t \bigcup \mathcal{I}^*)}\|^2 + (\frac{\eta k^*}{k+k^*} - \frac{\eta}{2})\|g_{t,\mathcal{I}^t \bigcup \mathcal{I}^*}\|^2 \\ &\quad + \frac{2k^*}{\eta(k+k^*)}\langle z_{t,\mathcal{I}}, (x^* - x_t)_{\mathcal{I}}\rangle + \|z_{t,\mathcal{I}}\|\|(x_{t+1} - x_t)_{\mathcal{I}}\|_2 - \frac{1-\eta\ell_s}{2\eta}\|x_{t+1} - x_t\|^2 \end{aligned} \tag{9}$$

$$\begin{aligned}&\leq L(x_t) + \frac{2k^*}{k+k^*}(L(x^* - L(x_t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)}\|x^* - x_t\|^2 - (\frac{1-\eta\ell_s}{2\eta} - \frac{k^*}{\eta(k+k^*)})\|x_{t+1} - x_t\|^2 + \\ &(\frac{\eta k^*}{k+k^*} - \frac{\eta}{2})\|g_{t,\mathcal{I}^t \bigcup \mathcal{I}^*}\|^2 + \frac{2k^*}{\eta(k+k^*)}\langle z_{t,\mathcal{I}}, (x^* - x_t)_{\mathcal{I}}\rangle + \|z_{t,\mathcal{I}}\|\|(x_{t+1} - x_t)_{\mathcal{I}}\|_2 \end{aligned} \tag{10}$$

$$\begin{aligned}&\leq L(x_t) + \frac{2k^*}{k+k^*}(L(x^* - L(x_t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)}\|x^* - x_t\|^2 + (\frac{\eta k^*}{k+k^*} - \frac{\eta}{2})\|g_{t,\mathcal{I}^t \bigcup \mathcal{I}^*}\|^2 \\ &+ \frac{2k^*}{\eta(k+k^*)}\langle z_{t,\mathcal{I}}, (x^* - x_t)_{\mathcal{I}}\rangle + \frac{\eta(k+k^*)}{2((1-\eta\ell_s)k - (1+\eta\ell_s)k^*)}\|z_{t,\mathcal{I}}\|^2, \end{aligned} \tag{11}$$

where the second inequality is due to the fact that $\|x_{t+1} - x_t\| \geq \eta\|g_{t,\mathcal{I}\backslash(\mathcal{I}^t \bigcup \mathcal{I}^*)}\|$ 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 11.

$$\|x_t - x^*\|^2 \leq \frac{4}{\rho}[L(x^*) - L(x_t)] + \frac{8}{\rho_s^2}\|g_{t,\mathcal{I}^t \bigcup \mathcal{I}^*}\|^2 + \frac{8}{\rho_s^2}\|z_{t,\mathcal{I}}\|^2. \tag{12}$$

Proof. From RSC, we have

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

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

With this lemma, we get

$$\begin{aligned} L(x_{t+1}) \le L(x_t) + \frac{2k^*}{k+k^*}(1+\frac{2(1-\eta\rho_s)}{\eta\rho_s})(L(x^*)-L(x_t)) - (\frac{\eta}{2} - \frac{(\eta^2\rho_s^2+8(1-\eta\rho_s))k^*}{\eta\rho_s^2(k+k^*)})\|g_{t,\mathcal{I}^t\bigcup\mathcal{I}^*}\|^2 \\ + \frac{2k^*}{\eta(k+k^*)}\langle z_{t,\mathcal{I}},(x^*-x_t)_{\mathcal{I}}\rangle + (\frac{\eta(k+k^*)}{2((1-\eta\ell_s)k-(1+\eta\ell_s)k^*)} + \frac{8(1-\eta\rho_s)k^*}{\eta\rho_s^2(k+k^*)})\|z_{t,\mathcal{I}}\|^2. \end{aligned} \tag{13}$$

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

$$L(x_{t+1}) \le L(x_t) + \frac{\rho_s}{8\ell_s}(L(x^*)-L(x_t)) + \frac{4k^*\ell_s}{(k+k^*)}\langle z_{t,\mathcal{I}},(x^*-x_t)_{\mathcal{I}}\rangle + \frac{37\ell_s}{\rho_s^2}\|z_{t,\mathcal{I}}\|^2. \tag{14}$$

Lemma 12. *For $x \sim \mathcal{N}(0,\sigma^2 I_p)$*

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

Proof. By definition of expectation, we have

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

□

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 $\|z_{t,\mathcal{I}}\|^2 \le |\mathcal{I}||z_t|_\infty^2$ (from the above lemma), we have

$$\mathbb{E}L(x_{t+1}) \le L(x_t) + \frac{\rho_s}{8\ell_s}(L(x^*)-L(x_t)) + O(\frac{\kappa_s k^* G^2 \log\frac{1}{\delta}\log pT}{\rho_s n^2\epsilon^2}). \tag{15}$$

That is

$$\mathbb{E}[L(x_{t+1})-L(x^*)] \le (1-\frac{\rho_s}{8\ell_s})\mathbb{E}[L(x_t)-L(x^*)] + O(\frac{\kappa_s k^* G^2\log p\log\frac{1}{\delta}T}{\rho_s n^2\epsilon^2}). \tag{16}$$

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

C. Discussion on Relaxation of ϵ-LDP

In Section 4 of the main paper, we show that even if sequential interaction and some relaxation on the loss function are allowed, the polynomial dependence on p still cannot be avoided. We now consider the possibility of lowering the dependence 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 (Bassily et al., 2014) and the 1-way marginal (Bun et al., 2018b). However, as shown in a recent study (Bun et al., 2018a), 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) (Mironov, 2017) and Local Zero-Concentrated Differential Privacy (LzCDP) (Bun & Steinke, 2016). 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.

Theorem 2. *For a given fixed privacy parameter $0 < \epsilon \leq 1$, 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,*

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

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 :

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

To prove Theorem 2, we first recall the definitions of Rényi Differential Privacy, zero-Concentrated Differential Privacy and χ^2-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 2. *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 $x, x' \in \mathcal{X}$, $D_\alpha(Q_i(Z_i \in S \mid x_i, z_{1:i-1})\|Q_i(Z_i \in S \mid x_i', = z_{1:i-1})) \leq \kappa + \rho\epsilon$ 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 3. *Similarly, we have (α, ϵ) local Rényi differential privacy and (α, ϵ) (sequential) Renyi differential private minimax risk (called sequential RDP minimax risk) if*

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

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 4. *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 2) 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 13. *[Theorem 2 in (Duchi & Ruan, 2018)] 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*

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

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 $\mathcal{C}_2(\{P_v\}_{v\in\mathcal{V}}) = \frac{1}{|\mathcal{V}|}\inf_{supp P\subset\mathcal{X}}\sup_\gamma\{\sum_{v\in\mathcal{V}}(\phi_v(\gamma))^2 \mid \|\gamma\|_{L^2(P)} \leq 1\}$, where $\phi(\gamma)$ is defined in Lemma 3.

Proof of Theorem 2 in Section C. The construction of P_0 and $\{P_v, v \in \mathcal{V}\}$ is the same as in the proof of Theorem 2 in Section B. Thus, by Lemma 13, we only need to bound $C_2(\{P_v\}_{v\in\mathcal{V}})$, instead of $C_\infty(\{P_v\}_{v\in\mathcal{V}})$. 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 $C_2(\{P_v\}_{v\in\mathcal{V}}) \leq \frac{\delta^2}{p}$. The remaining part of the proof is the same as the one in the proof of Theorem 2. □

D. Extending Linear to Non-linear Measurements

Our methods in Algorithm 1 and 2 are actually quite general which can be extended to 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 (Zhang et al., 2018; Yang et al., 2016). 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 5. *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.

D.1. Keeping the Whole Dataset Private

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

$$\mathcal{P}_{s,p,C,f,a,b} = \{P_{\theta,\sigma} \mid x \sim \text{Uniform}\{+1,-1\}^p, y = f(\langle\theta, x\rangle)+\sigma,\ \text{where } \sigma \text{ is the random noise satisfying the condition of } |\sigma| \leq C, C > 0 \text{ is some constant } \|\theta\|_2 \leq 1, \|\theta\|_0 \leq s, f \text{ is } (a,b) \text{ monotone } \}. \quad (17)$$

We note that when $f(x) = x$, it reduces to (1) in Section 4.1.

To obtain an upper bound of the empirical risk, we can easily extend Algorithm 1 in Section 4 to the non-linear measurement case (see Algorithm 2) to solve the following problem

$$\begin{aligned} &\min L(\theta; D) = \frac{1}{n}\sum_{i=1}^{n}(f(\langle x_i, \theta\rangle) - y_i)^2 \\ &s.t. \|\theta\|_2 \leq 1, \|\theta\|_0 \leq s. \end{aligned} \quad (18)$$

Algorithm 2 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, $\epsilon$ is the privacy parameter, and $\eta$ 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 $z_i = \mathcal{R}_\epsilon(\nabla_i)$, where $\mathcal{R}_\epsilon$ is the randomizer defined in (Smith et al., 2017) or (Duchi et al., 2018) and send back to the server.
    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 $\theta_T$
```

Theorem 3. *For any $\epsilon > 0$, Algorithm 2 is ϵ sequential interactive LDP. Moreover, if $\{X_{S_t}\}$ satisfoes 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*

$$\|\theta_T - \theta^*\|_2 \leq (\frac{1}{2})^T \|\theta^*\|_2 + O(\frac{\sqrt{p \log p}\sqrt{T}\sqrt{s}}{\sqrt{n}\epsilon}), \tag{19}$$

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

Proof of Theorem 3. W.o.l.g we assume that each $|S_t| = \frac{n}{T}$. From the randomizer $\mathcal{R}_\epsilon(\cdot)$ and Lemma 9, we can see that $\tilde{\nabla}_t = \frac{T}{n}\sum_{i\in S_t} x_i^T f'(\langle x_i, \theta_{t-1}\rangle)(f(\langle x_i, \theta_{t-1}\rangle) - y) + \zeta_t$, where ζ_t is a sub-Gaussian vector with $\sigma^2 = O(\frac{pT}{n\epsilon^2})$. By the definition of sub-Gaussian vector we know that each coordinate of σ_t is a sub-Gaussian random variable.

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

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

We denote the followings:

$$A^{t-1} = \|\Delta_{t-1} - \eta[\frac{T}{n}\sum_{i\in S_t}(f(\langle x_i, \theta_{t-1}\rangle) - f(\langle x_i, \theta^*\rangle))f'(\langle x_i, \theta_{t-1}\rangle)x_i^T]_{\mathcal{F}^{t-1}}\|_2 \tag{20}$$

$$B^{t-1} = \eta\sqrt{|\mathcal{F}^{t-1}|}|\frac{T}{n}\sum_{i\in S_t} f'(\langle x_i, \theta_{t-1}\rangle)\sigma_i x_i^T|_\infty \tag{21}$$

$$C^{t-1} = \eta\sqrt{|\mathcal{F}^{t-1}|}|\zeta_t|_\infty \tag{22}$$

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 4 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

$$B^t \leq \eta\sqrt{2s + s^*}O(\frac{\sqrt{T\log p}bC}{\sqrt{n}}). \tag{23}$$

For the term C^{t-1}, by Lemma 9 and 4 and since each coordinate $\zeta_{t,i}$ is sub-Gaussian, we have $C^{t-1} \leq \eta\sqrt{2s+s^*}O(\frac{\sqrt{Tp\log p}}{\sqrt{n\epsilon^2}})$ 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 5.

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

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

Finally, we need to show that $\|\Delta_t\|_2 = \|\theta_t - \theta^*\|_2 \leq \|\theta'_t - \theta^*\|_2$, which is due to the Lemma 6.

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 $1-\frac{2T}{p^c}$,

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

□

D.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:

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

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 4. *Under Assumption 3 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*

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

Comparing to the lower bound in Theorem 6 in the main paper, we can see that there is an additional factor of b^2 in Theorem 4, which is due to the fact that the model is more complicated.

Proof. Our proof is inspired by the ones in (Duchi et al., 2013; Yang et al., 2016) and (Raskutti et al., 2011). 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 ((Raskutti et al., 2011))**.** *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

$$8\delta^2\geq\|\theta-\theta'\|_2^2\geq\delta^2. \tag{24}$$

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 2. 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

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

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

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.

Algorithm 3 General DP-Iterative Hard Thresholding

Input: Public dataset $\{x_i\}_{i=1}^n$, private $\{y_i\}_{i=1}^n \in P_{\theta^*,\sigma}$, where $P_{\theta^*,\sigma} \in \mathcal{P}_{s^*,p,C,f,a,b}$, ϵ, δ are privacy parameters, T is the number of iteration, η 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 $z_i \sim \mathcal{N}(0,\sigma_1^2)$, $\sigma_1^2 = \frac{32C^2\ln(1.25/\delta)}{\epsilon^2}$.
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 $\theta_T$.
```

Theorem 5. *For any* $0 < \epsilon \leq 1$ *and* $0 < \delta < 1$*, Algorithm 3 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 3, there is an* $\eta = \eta(\delta')$ *which ensures that the 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 $1 - T\exp(-n) - \frac{2T}{p^c}$.

Proof. For the guarantee that Algorithm 2 is (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 (Dwork et al., 2006), 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 $\sigma^2 = \frac{1}{n^2}\sum_{i=1}^n f'^2(\langle x_i, \theta_t\rangle)\sigma_i^2 \leq \frac{b^2C^2}{n}$. Thus, by Lemma 4 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

$$B^t \leq O(\eta\sqrt{2s + s^*}\frac{\sqrt{\log p}bC}{\sqrt{n}}). \tag{25}$$

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 8, we have $\sum_{i=1}^n z_i^2 \leq 5\tau^2 n$ with probability at least $1-\exp(-n)$. Thus,

$$C^t \leq \eta\sqrt{2s+s^*}O(\frac{b\sqrt{\log p}\tau}{\sqrt{n}}) \tag{26}$$

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 3, and thus we omit it from here.

By (25) and (26) and plugging $\tau^2=\frac{32C^2\log(1.25/\delta)}{\epsilon^2}$ into (26), 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 5 with probability at least $1-\frac{2T}{p}-T\exp(-n)$. For the linear case, since $f'\equiv 1$, (25) and (26) will be the same in each iteration, the probability for the linear case becomes $1-\frac{2}{p^c}-\exp(-n)$. □

E. Experiments

E.1. Tests on Real World Dataset for Keeping the Lable Private

We apply our algorithms (Algorithms 2) to an image reconstruction problem. Note that the ϵ-LDP of high dimensional sparse linear regression is essentially a private sparse signal recovery problem. The construction of the sparse signals follows the one in (Zhang et al., 2018). Let $\mathcal{I}\in\mathbb{R}^{h\times w}$ be the image with height h and width w with $h\leq w$. Let $\mathcal{I}=\sum_{i=1}^h \lambda_i u_i v_i^T$ be the singular value decomposition of $\mathcal{I}$, where $\lambda_1,\cdots,\lambda_h$ are the eigenvalues in descending order. For a fixed integer s^*, let $\tilde{\mathcal{I}}=\sum_{i=1}^{s^*}\lambda_i u_i v_i^T$ be the best rank-s^* approximation of $\mathcal{I}$. We choose $b=(\sigma_1,\sigma_2,\cdots,\sigma_{s^*},0,\cdots,0)^T\in\mathbb{R}^h$ and $\theta^*=\frac{b}{\|b\|_2}$, and fix $\{(u_i,v_i)\}_{i\in[h]}$ and $\alpha=\|b\|_2$. Given an estimator $\theta^T\in\mathbb{R}^h$, we reconstruct an image by $\hat{\mathcal{I}}=\sum_{i=1}^h \alpha\theta_{T,i}u_i v_i^T$, which is an estimator of the image $\tilde{\mathcal{I}}$.

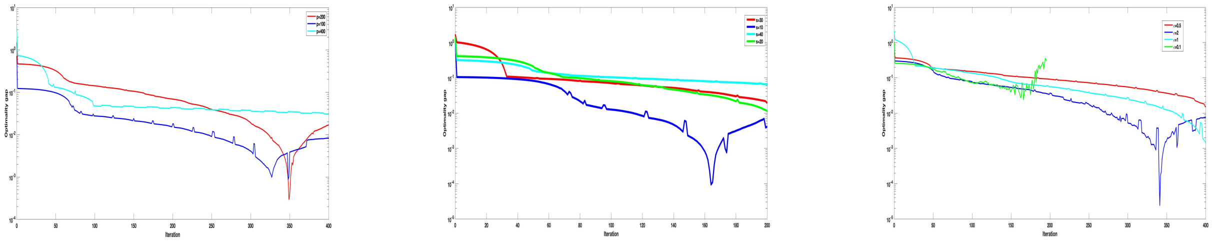
Experiment Results In the experiment, we let $\mathcal{I}$ be an image with $h=2448, w=3264$. The sparse signal θ^* is constructed with $s^*=100$. We sample $n=2s^*\log h$ i.i.d observations of the regression model $y=f(\langle x,\theta^*\rangle)+\sigma$ for $f(x)=x$ or $f(x)=8x+\cos x$, and apply our algorithms with $\eta=0.2$, $T=1000$, and $s=100$. Figures 3 and 4 are the reconstructed images using the two algorithms at different levels of privacy (see supplemental material for more details). Comparing the original image with the reconstructed ones, we can see that there is very little visual difference even when $\epsilon=0.5$. This confirms the effectiveness of our algorithms. Also, it is worth noting that the relative error is about 10% when $\epsilon=0.5$, and decreases to 7% and 2.5% as ϵ increases to 1 and 2, respectively.

E.2. More Results For Sparsity-constrained ERM

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318. ACM, 2016.

Bassily, R., Smith, A., and Thakurta, A. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *arXiv preprint arXiv:1405.7085*, 2014.



(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 $s = 10$

Figure 1. Experimental results on rcv1 dataset (Chang & Lin, 2011) for ℓ_0-constrained logistic regression under (ϵ, δ)-DP).

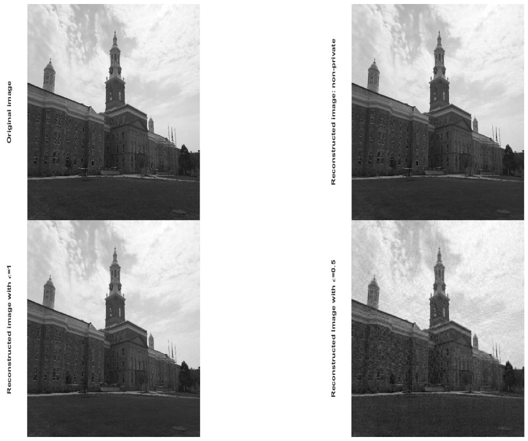

Original image
Reconstructed image: non-private
Reconstructed image with ϵ=1
Reconstructed image with ϵ=0.5

Figure 2. Reconstructed image for high dimensional sparse linear regression. The upper left one is the original image, the upper right one is for $\epsilon = 2$; the lower left one is for $\epsilon = 1$, and the lower right one is for $\epsilon = 0.5$.

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.

Chang, C.-C. and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

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.

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.

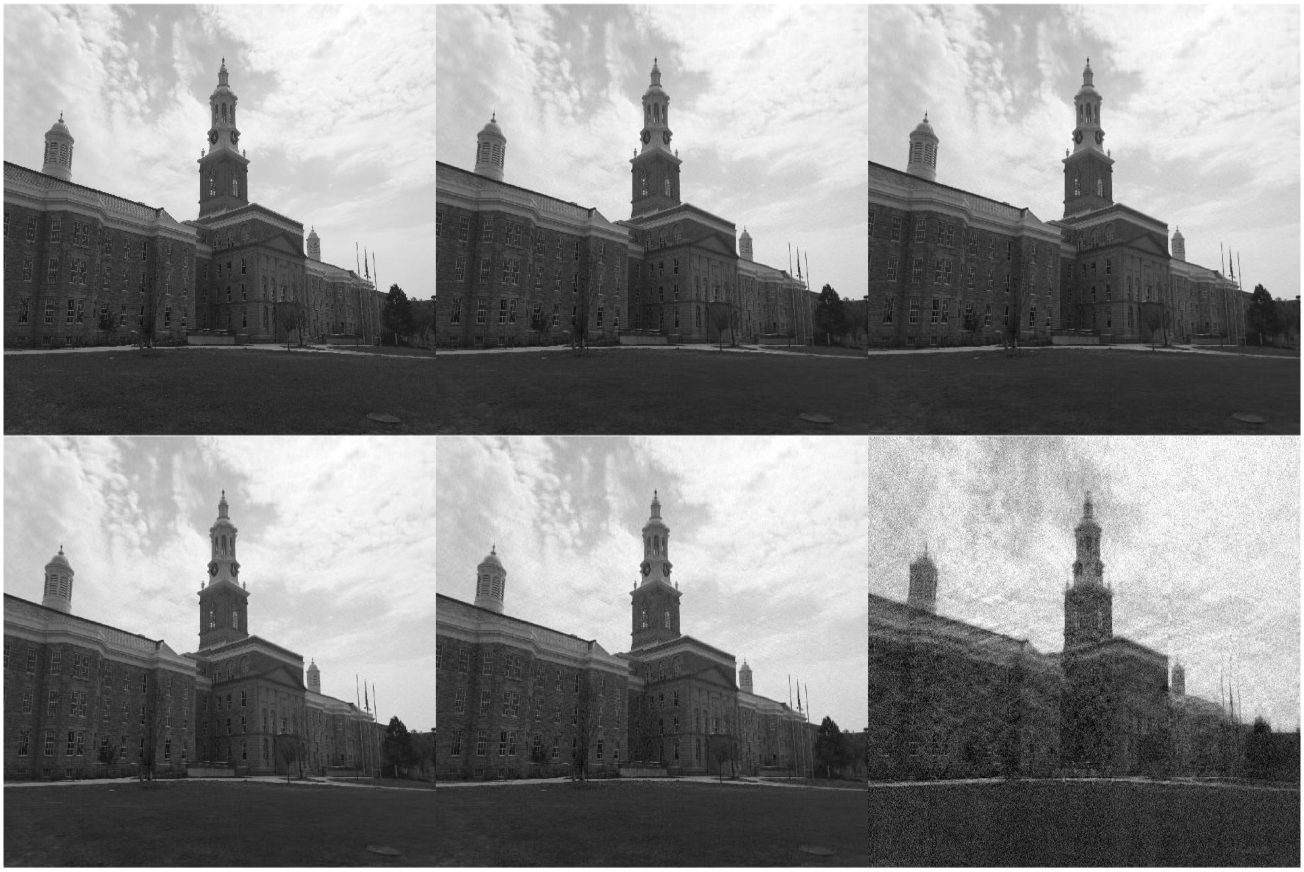


Figure 3. Reconstructed image using high dimensional sparse regression with non-linear measurement $f(x) = 8x + \cos x$. The upper left one is the original image, the upper middle one is the non-private reconstructed image, the upper right one is for $\epsilon = 2$, the lower left one is for $\epsilon = 1$, the lower middle one is for $\epsilon = 0.5$, and the lower right one is for $\epsilon = 0.1$.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Mironov, I. Renyi differential privacy. In *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*, pp. 263–275. IEEE, 2017.

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.

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.

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

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.

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.