Private Stochastic Convex Optimization and Sparse Learning with Heavy-tailed Data Revisited

Youming Tao[1] , **Yulian Wu**[2] , **Xiuzhen Cheng**[1] and **Di Wang**[2]*
[1]School of Computer Science, Shandong University
[2]CEMSE, KAUST
di.wang@kaust.edu.sa

Abstract

In this paper, we revisit the problem of Differentially Private Stochastic Convex Optimization (DP-SCO) with heavy-tailed data, where the gradient of the loss function has bounded moments. Instead of the case where the loss function is Lipschitz or each coordinate of the gradient has bounded second moment studied previously, we consider a relaxed scenario where each coordinate of the gradient only has bounded $(1+v)$-th moment with some $v \in (0,1]$. Firstly, we start from the one dimensional private mean estimation for heavy-tailed distributions. We propose a novel robust and private mean estimator which is optimal. Based on its idea, we then extend to the general d-dimensional space and study DP-SCO with general convex and strongly convex loss functions. We also provide lower bounds for these two classes of loss under our setting and show that our upper bounds are optimal up to a factor of $O(\text{Poly}(d))$. To address the high dimensionality issue, we also study DP-SCO with heavy-tailed gradient under some sparsity constraint (DP sparse learning). We propose a new method and show it is also optimal up to a factor of $O(s^*)$, where s^* is the underlying sparsity of the constraint.

1 Introduction

As one of the most fundamental problems in machine learning and statistics, Stochastic Convex Optimization (SCO) [Vapnik, 1999] with its empirical form, Empirical Risk Minimizataion (ERM), has been widely studied. Both SCO and ERM have found numerous applications in many areas such as medicine, finance, genomics and social science. However, due to the widespread concerns on privacy, how to handle sensitive data, such as biomedical datasets, has become a big hurdle for successful implementations of SCO in practice. To address the privacy issue, Differential Privacy (DP) [Dwork *et al.*, 2006] has established itself as a canonical privacy notation for privacy-preserving data analysis.

The study of SCO and ERM under DP constraint (i.e., DP-SCO and DP-ERM) has received significant attentions over the past decade. A long list of works have studied the problem from different perspectives: [Bassily *et al.*, 2014; Bassily *et al.*, 2019; Feldman *et al.*, 2020] studied the problems in the low dimensional case and the central DP model, [Cai *et al.*, 2020] considered the problems in the high dimensional sparse case and the central DP model, [Duchi *et al.*, 2018] focused on the problems in the local DP model.

Even though there are numerous works on DP-SCO, a critical issue in most existing results is that loss function has to be assumed to satisfy the $O(1)$-Lipschitz property, or the underlying data distribution is assumed to be sub-Gaussian or even bounded. Despite simplifying the procedure of designing DP algorithms, such assumptions are unrealistic and may not always hold when dealing with real-world datasets, especially those from biomedicine and finance, as it has been observed that they are often heavy-tailed [Woolson and Clarke, 2011]. The heavy-tailed data could lead to unbounded gradients and thus break the Lipschitz assumption, which implies that previous algorithms may fail to provide DP guarantee. Recently, to tackle this issue, there have been several works studying DP-SCO with heavy-tailed data [Wang *et al.*, 2020b; Kamath *et al.*, 2021; Hu *et al.*, 2021] or private mean estimation for heavy-tailed distributions [Barber and Duchi, 2014; Kamath *et al.*, 2020; Liu *et al.*, 2021]. However, all these results still need to assume that the distribution of each coordinate of the gradient of the loss function has bounded second moment, which implies the data is still well-behaved to some extent. Thus, a natural question is,

Can we further relax the bounded second moment condition to model the data distribution that are more heavy-tailed? And what are the theoretical behaviors of DP-SCO with more extremely heavy-tailed data?

In this paper, we revisit the problem of DP-SCO with heavy-tailed data under more relaxed assumptions. For the first time, we consider the case with more extremely heavy-tailed data such that the distribution of each coordinate of the loss gradient has only bounded $(1+v)$-th moment for some $v \in (0,1]$. Our contributions can be summarized as follows.

- First, we consider one-dimensional private mean estimation for heavy-tailed distributions. We propose a novel robust and (ϵ, δ)-DP estimator based on truncating

*Contact Author

the data, which achieves an error of $O\Big(\big(\frac{\sqrt{\log\frac{1}{\delta}}}{n\epsilon}\big)^{\frac{v}{1+v}}\Big)$, where n is the number of data samples. We then show that our proposed estimator is optimal by providing the matching lower bound on the estimation error.

- Based on the idea in the one-dimensional case, we then extend to estimate the mean of heavy-tailed distribution in a general d-dimensional space and use it to DP-SCO. Specifically, we consider both strongly convex and general convex loss functions for DP-SCO, and propose (ϵ, δ)-DP algorithms that achieve an error of $\widetilde{O}\Big(\frac{d^{\frac{1+4v}{1+v}}}{(\epsilon n)^{\frac{2v}{1+v}}}\Big)$ and $\widetilde{O}\Big(\frac{d^{\frac{1+4v}{2+2v}}}{(\epsilon n)^{\frac{v}{1+v}}}+\frac{d^{\frac{3+12v}{2+2v}}}{(\epsilon n)^{\frac{3v}{1+v}}}\Big)$ respectively, if we omit other terms. We also provide the lower bounds under our setting in both (ϵ, δ)-DP and ϵ-DP, and show that our upper bound is optimal up to a factor of $\text{Poly}(d)$.
- Finally, to mitigate the high dimensionality issue, we also study DP-SCO with heavy-tailed data under the sparsity constraint, *i.e.,* DP sparse learning. Based on the previous ideas, we propose a new method under our setting and show that it is possible to achieve an error of $\widetilde{O}\Big((s^*)^{\frac{1+2v}{1+v}}\big(\frac{\log d}{n\epsilon}\big)^{\frac{2v}{1+v}}\Big)$, where s^* is the underlying sparsity. We also proof that the bound is optimal up to a factor of $O(s^*)$.

Due to space limit, all the proofs are included in Appendix.

2 Related Work

As mentioned in Section 1, there are a vast number of works on DP-SCO and DP-ERM. Due to space limit, here we just discuss the results that are the most related to ours. For DP-SCO with heavy-tailed data, [Wang *et al.*, 2020b] first studies the problem by proposing three methods based on different assumptions. However, all the three methods need to assume the distribution of gradient of the loss is sub-exponential or has at least bounded second moment. [Kamath *et al.*, 2021] recently revisits the problem under the same assumption as in [Wang *et al.*, 2020b] and improves the (expected) excess population risk for both convex and strongly convex loss functions. It also provides the lower bounds in both (ϵ, δ)-DP and ϵ-DP models. Note that although some ideas of our algorithms and proofs are the same as theirs, their methods cannot be used in our relaxed setting and there are several differences. See Remark 1 and 2 for details.

DP sparse learning has been studied previously [Wang and Gu, 2019; Wang and Xu, 2019; Wang and Xu, 2021]. However, all of the previous methods need either the loss function be Lipschitz, or the data distribution be sub-Gaussian or even bounded [Cai *et al.*, 2019]. [Hu *et al.*, 2021] recently extends to the heavy-tailed case where each coordinate of the gradient has bounded second moment. However, due to their private estimator, it is impossible to use their methods to the bounded $(1+v)$-th moment case. See Remark 3 for details.

3 Preliminaries

Definition 1 (Differential Privacy (DP)[Dwork *et al.*, 2006])**.** A randomized algorithm $\mathcal{M} : \mathcal{X}^n \mapsto \mathcal{Y}$ satisfies (ϵ, δ)-differential privacy if for every pair of neighbouring datasets $X, X' \in \mathcal{X}^n$ (i.e., datasets that differ in exactly one entry), it holds for $\forall Y \subseteq \mathcal{Y}$ that

$$\mathbb{P}(\mathcal{M}(X) \in Y) \leq e^{\epsilon} \cdot \mathbb{P}(\mathcal{M}(X') \in Y) + \delta.$$

When $\delta = 0$, we call $\mathcal{M}$ as ϵ-DP.

Lemma 1 (Adaptive Composition Theorem)**.** Given target privacy parameters $0 < \epsilon < 1$ and $0 < \delta < 1$, to ensure (ϵ, δ)-DP over m mechanisms, it suffices that each mechanism is (ϵ', δ')-DP, where $\epsilon' = \frac{\epsilon}{2\sqrt{2m\ln(2/\delta)}}$ and $\delta' = \frac{\delta}{2m}$.

Lemma 2 (Laplacian Mechanism)**.** Given a dataset $D \in \mathcal{X}^n$ and a function $q : \mathcal{X}^n \to \mathbb{R}^d$, the Laplacian Mechanism is defined as $q(D) + (Y_1, Y_2, \cdots, Y_d)$, where each Y_i is i.i.d. sampled from the Laplacian Distribution $\text{Lap}(\frac{\Delta_1(q)}{\epsilon})$, where $\Delta_1(q)$ is the ℓ_1-sensitivity of the function q, *i.e.,* $\Delta_1(q) = \sup_{D \sim D'} ||q(D) - q(D')||_1$. The density of the Laplacian distribution with parameter λ is $\text{Lap}(\lambda)(x) = \frac{1}{2\lambda}\exp(-\frac{x}{\lambda})$. Laplacian mechanism preserves ϵ-DP.

Lemma 3 (Gaussian Mechanism)**.** Given a dataset $D \in \mathcal{X}^n$ and a function $q : \mathcal{X}^n \to \mathbb{R}^d$, the Gaussian mechanism is defined as $q(D) + \xi$ where $\xi \sim \mathcal{N}(0, \frac{2\Delta_2^2(q)\log(1.25/\delta)}{\epsilon^2}\mathbb{I}_d)$, where $\Delta_2(q)$ is the ℓ_2-sensitivity of the function q, *i.e.,* $\Delta_2(q) = \sup_{D \sim D'} \|q(D) - q(D')\|_2$. Gaussian mechanism preserves (ϵ, δ)-DP.

Definition 2 (DP-SCO [Bassily *et al.*, 2014])**.** Let $\mathcal{D}$ be some unknown distribution over the data universe $\mathcal{X}$ and $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}^n$ be i.i.d samples from the distribution $\mathcal{D}$. Given a convex constraint set $\mathcal{W} \subseteq \mathbb{R}^d$ and a convex loss function $\ell : \mathcal{W} \times \mathcal{X} \mapsto \mathbb{R}$. Differentially Private Stochastic Convex Optimization (DP-SCO) is to find w^{priv} so as to minimize the population risk, *i.e.,* $L_{\mathcal{D}}(w) = \mathbb{E}_{x\sim\mathcal{D}}[\ell(w, x)]$ with the guarantee of being differentially private. The utility of an algorithm $\mathcal{A}$ for DP-SCO is measured by the *expected excess population risk*, which is defined as follows:

$$err_{\mathcal{D}}(w^{\text{priv}}) = \mathbb{E}_{X\sim\mathcal{D}^n,\mathcal{A}}\left[L_{\mathcal{D}}(w^{\text{priv}}) - \min_{w\in\mathcal{W}} L_{\mathcal{D}}(w)\right].$$

4 One-dimensional Private Mean Estimation for Heavy-tailed Distributions

Before studying DP-SCO with heavy-tailed data, we start from the one-dimensional private mean estimation for heavy-tailed distributions to illustrate the idea of our following methods more clearly. Here we assume the data distribution has bounded $(1+v)$-th moment with some $v \in (0, 1]$. Formally, we are given a dataset $X = \{x_1, \ldots, x_n\}$ with each $x_i \in \mathbb{R}$ sampled from the one dimension distribution $\mathcal{D}$ such that $\mathbb{E}_{x\sim\mathcal{D}}[|x|^{1+v}] \leq u = O(1)$. [1] We aim to privately estimate the mean of the distribution, *i.e.,* $\mu = \mathbb{E}_{x\sim\mathcal{D}}[x]$. Note that here we use the raw moment, which also implies that its central moment is bounded, *i.e.,* $\mathbb{E}_{x\sim\mathcal{D}}[|x - \mathbb{E}_{x\sim\mathcal{D}}[x]|^{1+v}] \leq O(1)$, and vice versa. See Lemma 13 in Appendix for details.

We propose a private and robust estimator based on truncation. Specifically, for each data sample x_i, if its magnitude is within the designed threshold B then we will keep it, otherwise we set it be 0. After the preprocessing, each sample now

[1] Throughout the whole paper we assume that v and u are known.

Algorithm 1 Truncation Based DP Mean Estimator: $\text{DPODME\_T}_{\epsilon,\delta,\xi}(X)$

```
Input: Data samples $X = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}$. Parameters $\epsilon$, $\delta$, $\xi$
Output: A private mean estimator $\widetilde{\mu} \in \mathbb{R}$
1: Truncate each data sample $x_i$ by $x_i' \leftarrow x_i \cdot \mathbf{1}_{|x_i| \le B}$, where
       $B = \left( \frac{un\epsilon}{\log\frac{1}{\xi}\sqrt{\log\frac{1.25}{\delta}}} \right)^{\frac{1}{1+v}}$.
2: Get the robust mean estimator $\widehat{\mu} \leftarrow \frac{1}{n}\sum_{i=1}^n x_i'$.
3: return $\widetilde{\mu} \leftarrow \widehat{\mu} + \nu$, where $\nu \sim \mathcal{N}(0, \frac{8B^2}{n^2\epsilon^2}\log\frac{1.25}{\delta})$.
```

is bounded in $[-B, B]$. Thus we can add Gaussian noise to the mean of the truncated data. See Algorithm 1 for details.

Theorem 1. For any $0 < \epsilon, \delta \le 1$, Algorithm 1 $\text{DPODME\_T}_{\epsilon,\delta,\xi}(X)$ satisfies (ϵ, δ)-DP. Moreover, given any failure probability ξ, with probability at least $1 - \xi$, the output $\widetilde{\mu}$ satisfies

$$|\widetilde{\mu} - \mu| \le O\left(u^{\frac{1}{1+v}} \left(\frac{\log\frac{1}{\xi}\sqrt{\log\frac{1}{\delta}}}{n\epsilon} \right)^{\frac{v}{1+v}} \right).$$

Remark 1. When $v = 1$, the error becomes $O\left(\frac{1}{\sqrt{n\epsilon}}\right)$ (if we omit other terms), which matches the bound in [Wang *et al.*, 2020a] and is optimal [Kamath *et al.*, 2020]. However, previous results are for the case where the distribution has bounded second moment. Here we relax it to the case where the distribution only has its $(1 + v)$-th moment bounded. Thus, our method can be thought as a generalization of the previous results. Recently [Tao *et al.*, 2021] also considers a similar problem. However, there are several differences: First, [Tao *et al.*, 2021] focuses on the online setting in the ϵ-DP model while we consider the offline setting and the (ϵ, δ)-DP model. Secondly, the methods in [Tao *et al.*, 2021] are based on the tree mechanism and Laplacian mechanism, while we mainly use the Gaussian mechanism. Thus, our error bound is lower. Thirdly, besides the upper bound, we also show that the bound of $O((u^{\frac{1}{v+1}}(\frac{1}{n\epsilon})^{\frac{v}{1+v}}))$ in Theorem 1 is optimal by showing its lower bound in the Theorem 2 below, which has not been studied in [Tao *et al.*, 2021].

Theorem 2. There exists a distribution $\mathcal{D}$ with mean μ and its $(1+v)$-th raw moment is bounded by u. For any (ϵ, δ)-DP algorithm, its output $\widetilde{\mu}$ satisfies the following with at least a constant probability

$$|\widetilde{\mu} - \mu| \ge \Omega\left(u^{\frac{1}{v+1}} \left(\frac{1}{n\epsilon} \right)^{\frac{v}{1+v}} \right).$$

When $v = 1$ and $u = 1$, our result will be equivalent to the lower bound in [Kamath *et al.*, 2020]. Thus, Theorem 2 is a generalization of the previous result on the lower bound of private mean estimation for heavy-tailed distributions. Moreover, besides the (ϵ, δ)-DP model, the lower bound also holds for any ϵ-DP algorithm.

5 DP-SCO with Heavy-Tailed Data

In this section, based on the previous one dimensional private mean estimator, we provide our methods for DP-SCO with heavy-tailed data. Before that, we provide the assumptions that will be used throughout the section.

Definition 3 (Lipschitz)**.** A function $f : \mathcal{W} \to \mathbb{R}$ is L-Lipschitz if for all $w_1, w_2 \in \mathcal{W}$ we have $|f(w_1) - f(w_2)| \le L\|w_1 - w_2\|_2$.

Definition 4 (Strong convexity)**.** A function f is α-strongly convex on $\mathcal{W}$ if for all $w_1, w_2 \in \mathcal{W}$ we have $f(w_1) \ge f(w_2) + \langle \nabla f(w_2), w_1 - w_2 \rangle + \frac{\alpha}{2}\|w_1 - w_2\|_2^2$.

Definition 5 (Smoothness)**.** A function f is β-smooth on $\mathcal{W}$ if for all $w_1, w_2 \in \mathcal{W}$ we have $f(w_1) \le f(w_2) + \langle \nabla f(w_2), w_1 - w_2 \rangle + \frac{\beta}{2}\|w_1 - w_2\|_2^2$

Assumption 1. We make the following assumptions:

1. The parameter space $\mathcal{W}$ is convex and bounded with diameter Δ, i.e., for $\forall w_1, w_2 \in \mathcal{W}$, $\|w_1 - w_2\|_2 \le \Delta$.
2. The loss function $\ell(w, x)$ is non-negative and differentiable for all $w \in \mathcal{W}$ and $x \in \mathcal{D}$.
3. The population risk $L_{\mathcal{D}}(\cdot)$ is β-smooth over $\mathcal{W}$. For any $w \in \mathcal{W}$, the gradient of the population risk function satisfies $\|\nabla L_{\mathcal{D}}(w)\|_2 \le R = O(1)$. Moreover, the optimal solution $w^* = \arg\min_{w \in \mathcal{W}} L_{\mathcal{D}}(w)$ satisfies that $\nabla L_{\mathcal{D}}(w^*) = 0$.
4. For any $w \in \mathcal{W}$, the distribution of each coordinate of the gradient of the loss function has bounded $(1 + v)$-th (raw) moment with some $v \in (0, 1]$, i.e., there is a constant $u > 0$ such that $\mathbb{E}_{x\sim\mathcal{D}}[|\nabla_j \ell(w, x)|^{1+v}] \le u$ for all $j \in [d]$.

There are several notes on the terms in Assumption 1. Firstly, the first three terms in Assumption 1 are commonly used in the previous works on DP-SCO with heavy-tailed data [Kamath *et al.*, 2021; Wang *et al.*, 2020b]. The fourth condition assumes that the gradient of the loss is heavy-tailed, which is a commonly used assumption in the study of robust learning with heavy-tailed data. However, as mentioned previously, most of those works only assume that the gradient has at least bounded second moment while here we relax to the $(1+v)$-th moment. Moreover, we can see it is a relaxation of the Lipschitz condition that $\|\nabla L_{\mathcal{D}}(w)\|_2 \le L$ for all w.

In Algorithm 2, we propose our framework. The idea of the algorithm is quite straightforward: we use the private version of Projected Gradient Descent (PGD). Specifically, in each iteration t, we first privately estimate the vector $\nabla L_{\mathcal{D}}(w_{t-1})$ by using gradients $\{\nabla \ell(w_{t-1}, x_i)\}_{i=1}^n$, where w_{t-1} is the current parameter. Then we update the parameter via the PGD. In the classical setting where the data is regular or the loss function is Lipschitz, we can use the Gaussian mechanism to the average of the gradients $\frac{\sum_{i=1}^n \nabla \ell(w_{t-1}, x_i)}{n}$ to get a private estimation of $\nabla L_{\mathcal{D}}(w_{t-1})$. However, due to the heavy-tailed assumption on gradients in Assumption 1, in our problem we cannot use the same approach directly as now the ℓ_2-norm sensitivity maybe infinite. Thus the main difficulty is designing private estimator for $\nabla L_{\mathcal{D}}(w) = \mathbb{E}_{x\sim\mathcal{D}}[\nabla \ell(w, x)]$,

Algorithm 2 DP-SCO$_{\epsilon,\delta,\eta,T,\tau}$

```
Input: Data samples $X = \{x_i\}_{i=1}^n \subseteq \mathbb{R}^d$, parameters $\epsilon, \delta, \eta, T, \tau$.
Output: Private minimizer $w^{\text{priv}}$.
for $t \leftarrow 1, \cdots, T$ do
    if $\ell$ is convex then
        Set $X_t = X, \epsilon' \leftarrow \frac{\epsilon}{2\sqrt{2T\log(2/\delta)}}, \delta' \leftarrow \frac{\delta}{2T}$.
    else if $\ell$ is strongly convex then
        Set $X_t = \{x_{(t-1)n/T+1}, \cdots, x_{tn/T}\}$.
        Set $\epsilon' \leftarrow \epsilon, \delta' \leftarrow \delta$.
    end if
    $\nabla\widetilde{L}_{\mathcal{D}}(w_{t-1}) \leftarrow \text{DPHDME}_{\epsilon',\delta',\tau}(\{\nabla\ell(w_{t-1}, x)\}_{x\in X})$
    $w_t \leftarrow \text{Proj}_{\mathcal{W}}(w_{t-1} - \eta\nabla\widetilde{L}_{\mathcal{D}}(w_{t-1}))$
end for
if $L_{\mathcal{D}}(\cdot)$ is strongly convex then
    Set $w^{\text{priv}} \leftarrow w_T$
else if $L_{\mathcal{D}}(\cdot)$ is convex then
    Set $w^{\text{priv}} \leftarrow \frac{1}{T}\sum_{t\in[T]} w_t$
end if
return $w^{\text{priv}}$
```

Algorithm 3 DP High-Dimension Mean Estimator DPHDME$_{\epsilon,\delta,\tau}(X)$

```
Input: Data samples $X = \{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$. Parameters $\epsilon$, $\delta$, $\tau$
Output: A DP mean estimator $\widetilde{\mu} \in \mathbb{R}^d$
$m \leftarrow 4\log\left(\frac{2d}{\xi}\right)$.
for $j \leftarrow 1, \cdots, d$ do
    for $k \leftarrow 1, \cdots, m$ do
        for $i \leftarrow (k-1)\cdot\frac{n}{m} + 1, \cdots, k\cdot\frac{n}{m}$ do
            $[x'_i]_j \leftarrow [x_i]_j \cdot \mathbf{1}_{|[x_i]_j| \leq \tau}$
        end for
        $\widehat{\mu}_j^k \leftarrow \frac{m}{n}\sum_{i=1}^n [x'_i]_j$
    end for
    $\widehat{\mu}_j \leftarrow \text{median}(\widehat{\mu}_j^1, \widehat{\mu}_j^2, \cdots, \widehat{\mu}_j^m)$
end for
$\widehat{\mu} \leftarrow (\widehat{\mu}_1, \widehat{\mu}_2, \cdots, \widehat{\mu}_d)$
return $\widetilde{\mu} \leftarrow \widehat{\mu} + \nu$, where $\nu \sim \mathcal{N}\left(0, \frac{8\tau^2 m^2 d \ln\frac{1.25}{\delta}}{\epsilon^2 n^2} I_d\right)$
```

which could be seen as an instance of the private mean estimation in the d-dimensional space.

In Section 4 we considered the case where $d = 1$. Now we will use its idea in the general high dimensional case. Our estimator is presented in Algorithm 3. For each coordinate, we first partite the whole dataset into m subgroups. Then in each sub-dataset, for each coordinate, we truncate the data and calculate the mean of the truncated data. Finally, we use the traditional Median of Means (MoM) method in each coordinate, *i.e.,* for each coordinate, we calculate the median among the means of these m subgroups, and add Gaussian noise to ensure (ϵ, δ)-DP.

Theorem 3. For any $0 < \epsilon, \delta < 1$, Algorithm 3 is (ϵ, δ)-DP. Moreover, assume each data $x_i \sim \mathcal{D}$ where the distribution $\mathcal{D}$ satisfies: (1) $\mathcal{D}$ has the mean $\mu \in \mathbb{R}^d$ and $\|\mu\|_2 \leq R = O(1)$; (2) $\mathbb{E}_{x_i\sim\mathcal{D}}|[x_i]_j|^{1+v} \leq u$ for each $j \in [d]$. Then for any given truncation parameter $\tau \in \mathbb{R}$ and failure probability ξ, with probability at least $1 - \xi$, Algorithm 3 outputs a private mean estimator $\widetilde{\mu} \in \mathbb{R}^d$ such that,

$$\|\widetilde{\mu} - \mu\|_2 \leq O\Bigg(\sqrt{d}\Big(u^{\frac{1}{1+v}}\Big(\frac{\log\frac{d}{\xi}}{n}\Big)^{\frac{v}{1+v}} + \frac{u}{\tau^v}\Big) + \frac{\tau\sqrt{d}\Big(\sqrt{d} + \sqrt{\log\frac{1}{\xi}}\Big)\log\left(\frac{d}{\xi}\right)\sqrt{\ln\frac{1}{\delta}}}{\epsilon n}\Bigg),$$

where the Big-O notation omits the term of R.

Remark 2. To privately estimate the mean of heavy-tailed distributions with bounded second moments, in general there are three approaches: The first one is directly using one dimensional private estimator to each coordinate [Wang *et al.*, 2020b; Wang *et al.*, 2020a]. However, the bound of this approach is only sub-optimal. [Kamath *et al.*, 2020] proposes a method which aggressively truncate the distribution around a point, and compute the noisy empirical mean. However, their theoretical guarantee only holds with constant probability. Instead of these two approaches, here we adopt the idea of the third one, which is a private version of the MoM method and was recently proposed by [Kamath *et al.*, 2021]. However, there are two crucial differences: First, the truncation step is quite different, [Kamath *et al.*, 2021] truncates each data into an interval $[a, b]$, *i.e.,* for the sample x, if $x > b$ then we will let $x' = b$ and if $x < a$ then we will let $x' = a$. However, our approach could be seen as thresholding, *i.e.,* when $|x| > \tau$ then $x' = 0$. Second, to get the theoretical guarantee, [Kamath *et al.*, 2021] needs Lemma 4.4 in [Kamath *et al.*, 2020] (or Lemma A.2 in [Kamath *et al.*, 2021]), which only holds for data distributions that have at least bounded second moment. Thus, we cannot adopt their approach in our setting. To overcome the challenge we provide the following lemma on the concentration of heavy-tailed distributions, which is the key to prove Theorem 3.

Theorem 4. Let $x_1, x_2, \cdots, x_n \in \mathbb{R}$ be i.i.d. random variables with bounded $(1+v)$-th moment, i.e., for $\forall i \in [n]$, we have $\mathbb{E}[|x_i|^{1+v}] \leq M$ for some constant $v \in (0, 1]$. Let μ be $\mathbb{E}[x_i], \forall i \in [n]$. Then

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n x_i - \mu\right| \geq t\right) \leq \frac{8M}{n^v}\frac{1}{t^{1+v}}. \qquad (1)$$

Based on Theorem 3 and the convergent rate of PGD, we give the accuracy guarantees of Algorithm 2 by considering both general convex and strongly convex loss functions.

Theorem 5. For any $0 < \epsilon, \delta < 1$, Algorithm 2 is (ϵ, δ)-DP.

Theorem 6 (General Convex Case)**.** Suppose we have a DP-SCO problem satisfying Assumption 1. Taking $T = \frac{R^2\epsilon^2n^2}{\tau^2d^4}$, $\eta = \frac{\Delta}{R\sqrt{T}}$ and $\tau = \left(\frac{\epsilon n}{d^{3/2}}\right)^{\frac{1}{1+v}}$ in Algorithm 2, then the output $w^{\text{priv}} = \frac{1}{T}\sum_{t\in[T]} w^t$ satisfies

$$err_{\mathcal{D}}(w^{\text{priv}}) \leq \widetilde{O}\left(\Delta u^2\left(\frac{d^{\frac{1+4v}{2+2v}}}{(\epsilon n)^{\frac{v}{1+v}}} + \frac{d^{\frac{3+12v}{2+2v}}}{(\epsilon n)^{\frac{3v}{1+v}}}\right)\right),$$

where the Big-$\widetilde{O}$ notation omits all the logarithmic terms and R.

Theorem 7 (Strongly Convex Case)**.** Suppose we have a DP-SCO problem satisfying Assumption 1, and additionally the loss function $\ell(\cdot, x)$ is α-strongly convex for every $x \in \mathcal{X}$. Taking parameters $T = \log\left(\frac{(\alpha+\beta)G}{\alpha\beta}\right) / \log\left(\frac{\alpha^2+\beta^2+\alpha\beta}{(\alpha+\beta)^2}\right)$, $\eta = \frac{1}{\alpha+\beta}$ and $\tau = \left(\frac{u\epsilon n}{d^{\frac{3}{2}}\sqrt{T}}\right)^{\frac{1}{1+v}}$ in Algorithm 2, then the output $w^{\text{priv}} = w^T$ satisfies

$$err_{\mathcal{D}}(w^{\text{priv}}) \leq \widetilde{O}\left(\frac{(\Delta+1)^2(\alpha+\beta)^2}{\alpha^2\beta} u^{\frac{2}{1+v}} \frac{d^{\frac{1+2v}{1+v}}}{(\epsilon n)^{\frac{2v}{1+v}}}\right),$$

where the Big-$\widetilde{O}$ notation omits all the logarithmic terms and R.

When $v = 1$, the rate becomes $\widetilde{O}\Big(\frac{d^{5/4}}{(n\epsilon)^{\frac{1}{2}}} + \frac{d^{15/4}}{(n\epsilon)^{\frac{3}{2}}}\Big)$ and $\widetilde{O}\Big(\frac{d^{3/2}}{n\epsilon}\Big)$ for convex and strongly convex case respectively. These bounds are consistent with the best known result in [Kamath *et al.*, 2021]. However, the methods in [Kamath *et al.*, 2021] cannot be extended to the case where $v \in (0, 1)$. In the following we show that the term of $O\Big(\frac{1}{(\epsilon n)^{\frac{v}{1+v}}}\Big)$ and $O\Big(\frac{1}{(\epsilon n)^{\frac{2v}{1+v}}}\Big)$ is optimal for convex and strongly convex loss in (ϵ, δ)-DP respectively. Moreover, we provide lower bounds in the ϵ-DP model.

Theorem 8 (Lower Bound of Strongly Convex Loss)**.** Assume $\mathcal{W}$ is the unit norm ball. For any $v \in (0, 1]$, there exists a strongly convex and smooth loss function $\ell : \mathcal{W} \times \mathbb{R}^d \mapsto \mathbb{R}$ such that, for any ϵ-DP algorithm $\mathcal{A}$, there is a distribution $\mathcal{D}$ over $\mathbb{R}^d$ such that for any w, $\sup_{j\in[d]} \mathbb{E}_{x\sim\mathcal{D}}[|\nabla_j \ell(w, x)|^{1+v}] \leq u$, the output w^{priv} of $\mathcal{A}$ satisfies the following if $n \geq \Omega(u^{\frac{1}{v}} d^{\frac{1+3v}{2v}}/\epsilon)$

$$err_{\mathcal{D}}(w^{\text{priv}}) \geq \Omega\left(u^{\frac{2}{1+v}} d \left(\frac{d}{\epsilon n}\right)^{\frac{2v}{1+v}}\right).$$

For any (ϵ, δ)-DP algorithm $\mathcal{A}$ with $\epsilon \ll \log\frac{1}{\delta}$, there is a distribution $\mathcal{D}$ over $\mathbb{R}^d$ such that for any w, $\sup_{j\in[d]} \mathbb{E}_{x\sim\mathcal{D}}[|\nabla_j \ell(w, x)|^{1+v}] \leq u$, its output w^{priv} satisfies the following if $n \geq \Omega(u^{\frac{1}{v}} \sqrt{\log\frac{1}{\delta}} d^{\frac{1+2v}{2v}}/\epsilon)$

$$err_{\mathcal{D}}(w^{\text{priv}}) \geq \Omega\left(u^{\frac{2}{1+v}} d \left(\frac{\sqrt{d\log\frac{1}{\delta}}}{\epsilon n}\right)^{\frac{2v}{1+v}}\right).$$

Theorem 9 (Lower Bound of Convex Loss)**.** Assume $\mathcal{W}$ is the unit norm ball. For any $v \in (0, 1]$, there exists a convex and smooth loss function $\ell : \mathcal{W} \times \mathbb{R}^d \mapsto \mathbb{R}$ such that, for any ϵ-DP algorithm $\mathcal{A}$, there is a distribution $\mathcal{D}$ over $\mathbb{R}^d$ such that for any w, $\sup_{j\in[d]} \mathbb{E}_{x\sim\mathcal{D}}[|\nabla_j \ell(w, x)|^{1+v}] \leq u$, its output w^{priv} satisfies the following when $n \geq \Omega(d/\epsilon)$

$$err_{\mathcal{D}}(w^{\text{priv}}) \geq \Omega\left(u^{\frac{1}{1+v}} \sqrt{d} \left(\frac{d}{\epsilon n}\right)^{\frac{v}{1+v}}\right).$$

For any (ϵ, δ)-DP algorithm $\mathcal{A}$ with $\epsilon \ll \log\frac{1}{\delta}$, there is a distribution $\mathcal{D}$ over $\mathbb{R}^d$ such that for any w, $\sup_{j\in[d]} \mathbb{E}_{x\sim\mathcal{D}}[|\nabla_j \ell(w, x)|^{1+v}] \leq u$, its output w^{priv} satisfies the following when $n \geq \Omega(\sqrt{d\log\frac{1}{\delta}}/\epsilon)$

$$err_{\mathcal{D}}(w^{\text{priv}}) \geq \Omega\left(u^{\frac{1}{1+v}} \sqrt{d} \left(\frac{\sqrt{d\log\frac{1}{\delta}}}{\epsilon n}\right)^{\frac{v}{1+v}}\right).$$

When $v = 1$, all the rates in Theorem 8 and 9 match the lower bounds in [Kamath *et al.*, 2021]. Thus, our results can be seen as extensions of the previous results. In Theorem 8, the gap between the rates in ϵ-DP and (ϵ, δ)-DP is $O(d^{\frac{v}{1+v}})$, while it is $O\big(d^{\frac{v}{2(1+v)}}\big)$ in Theorem 9. This is quite different with the case when the loss is Lipschitz [Bassily *et al.*, 2014]. To prove the lower bounds, we first reduce the problem to mean estimation, and then we use the private version of Fano's lemma in [Acharya *et al.*, 2021; Kamath *et al.*, 2021], based on the packing of distributions in [Barber and Duchi, 2014].

6 Differentially Private Sparse Learning with Heavy-tailed Data

In the previous section, we studied the general case of DP-SCO under the assumption that the distribution of each coordinate of the loss gradient has $(1 + v)$-th moment. However, one weakness of our previous results is that, all the error bounds are in the form of $O(\text{Poly}(d, \frac{1}{n}, \frac{1}{\epsilon}))$, which indicates that the error will be large in the high dimensional case where $d \gg n$. Moreover, we also showed that in general these polynomial dependencies are unavoidable. Thus, to address the high dimensionality issue, in this section, we focus on some special cases. Specifically, we will study the problem of DP-SCO under sparsity constraints, which is also called DP sparse learning, *i.e.,* $\mathcal{W}$ is defined as $\mathcal{W} = \{w : \|w\|_0 \leq s^*\}$. We note that such a formulation encapsulates several important problems such as the ℓ_0-constrained linear/logistic regression [Bahmani *et al.*, 2013]. In this section, unlike the previous results on DP sparse learning which need strong assumptions on data distribution, we study the problem under the assumption that the gradient has only $(1 + v)$-th moments. We first introduce some definitions to the loss functions, which are commonly used in previous research on sparse learning.

Definition 6 (Restricted Strong Convexity, RSC)**.** A differentiable function $f(x)$ is restricted μ_r-strongly convex with parameter r if for any x, x' with $\|x - x'\|_0 \leq r$, we have $f(x) - f(x') - \langle \nabla f(x'), x - x' \rangle \geq \frac{\mu_r}{2}\|x - x'\|_2^2$.

Definition 7 (Restricted Strong Smoothness, RSS)**.** A differentiable function $f(x)$ is restricted γ_s-strongly smooth with parameter r if for any x, x' with $\|x - x'\|_0 \leq r$, we have $f(x) - f(x') - \langle \nabla f(x'), x - x' \rangle \leq \frac{\gamma_r}{2}\|x - x'\|_2^2$.

Note that RSC and RSS are weaker than the strong convexity and smoothness. Next we propose the assumptions that will be used in this section.

Assumption 2. We assume that the objective function $L_{\mathcal{D}}(\cdot)$ is μ_r-RSC and $\ell(w, x)$ is γ_r-RSS with parameter $r = 2s + s^*$,

Algorithm 4 Peeling($v, s, \epsilon, \delta, \lambda$)[Cai *et al.*, 2019]

```
Input: A vector $v \in \mathbb{R}^d$ of a dataset $X$, sparsity $s$, privacy parameter $\epsilon, \delta$, and noise scale $\lambda$.
Initialize $S = \emptyset$.
for $i \leftarrow 1, \cdots, s$ do
    Generate $w_i \in \mathbb{R}^d$ with $w_{i,1}, \cdots, w_{i,d} \sim \mathrm{Lap}(\frac{4\lambda\sqrt{2s\log\frac{1}{\delta}}}{\epsilon})$.
    Append $j^* = \arg\max_{j\in[d]\setminus S} |v_j| + w_{i,j}$ to $S$.
end for
Generate $\widetilde{w} \in \mathbb{R}^d$ with $\widetilde{w}_1, \cdots, \widetilde{w}_d \sim \mathrm{Lap}(\frac{4\lambda\sqrt{2s\log\frac{1}{\delta}}}{\epsilon})$.
return $v_S + \widetilde{w}_S$.
```

where $s = O((\frac{\gamma_r}{\mu_r})^2 s^*)$. We also assume for any $w \in \mathcal{W}'$, the distribution of each coordinate of the gradient of the loss function has bounded $(1+v)$-th (raw) moment with some $v \in (0,1]$, *i.e.,* for each $j \in [d]$, $\mathbb{E}[|\nabla_j \ell(w,x)|^{1+v}] \leq u$, where $\mathcal{W}' = \{w | \|w\|_0 \leq s\}$.

There are many problems satisfying Assumption 2, e.g., mean estimation and ℓ_2-norm regularized generalized linear loss where $L_{\mathcal{D}}(w) = \mathbb{E}[\ell(\langle w, x\rangle)] + \frac{\lambda}{2}\|w\|_2^2$. If $|\ell'(\cdot)| \leq O(1)$, $|\ell''(\cdot)| \leq O(1)$ (such as the logistic loss) and x_j has bounded $(1+v)$-th moment, then we can see that it satisfies Assumption 2.

Our method can be found in Algorithm 5, which is built upon the ideas of our previous private one dimensional mean estimator for heavy-tailed distributions and the Iterative Hard Thresholding method [Blumensath and Davies, 2009]. In detail, in each iteration, we first perform the truncation step to each coordinate of the gradient of the loss to get one-dimensional mean estimator. Next, unlike the one dimensional private mean estimator in Algorithm 1 where we add Gaussian noise to the mean of the truncated gradients, here we privately select top s indices via the Peeling mechanism (shown in Algorithm 4). This is due to that if we use Algorithm 1 to each coordinate, then the magnitude of the noise we add will depend on polynomial of d, which is large. However, using the Peeling mechanism will introduce an error that only depends on polynomial of s and $\log d$. In the following we show the theoretical guarantee of our algorithm.

Theorem 10. For any $0 < \epsilon, \delta < 1$, Algorithm 5 is (ϵ, δ)-DP. Moreover, under Assumption 2, given any failure probability ξ, if we set $T = \widetilde{O}(\frac{\gamma_r}{\mu_r}\log n)$, $s = O((\frac{\gamma_r}{\mu_r})^2 s^*)$, $\eta_0 = \frac{2}{3\gamma_r}$ and $B = O\Big(\Big(\frac{\gamma_r u n \epsilon}{T\log\frac{dT}{\xi}\sqrt{s\log\frac{1}{\delta}}}\Big)^{\frac{1}{1+v}}\Big)$, then with probability at least $1-\xi$,

$$err_{\mathcal{D}}(w^{\text{priv}}) \leq O\left((s^*)^{\frac{1+2v}{1+v}} u^{\frac{2}{1+v}} \left(\frac{\log n \log\frac{d}{\xi}\sqrt{\log\frac{1}{\delta}}}{n\epsilon}\right)^{\frac{2v}{1+v}}\right), \tag{2}$$

where the Big-O notation omits γ_r and μ_r.

Compared with the results in Section 5, we can see that in Theorem 10 the bound is only logarithmic in d and polynomial in s^*, $\frac{1}{\epsilon}$ and $\frac{1}{n}$, which means it is more suitable to the high dimensional case.

Algorithm 5 Heavy-Tailed Private Sparse Optimization

```
Input: Data samples $X = \{x_i\}_{i=1}^n \subseteq \mathbb{R}^d$, parameters $s, T, \eta$, initial $s$-sparse parameter $w^1$, privacy parameter $\epsilon, \delta$.
Output: Private minimizere $w^{\text{priv}}$
Split $X$ into $T$ parts $\{X_t\}_{t=1}^T$ with $|X_t| = m = \frac{n}{T}$.
for $t \leftarrow 1, \cdots, T$ do.
    for each dimension $j \in [d]$ do
        for each data sample $x \in X_t$ do
            $\nabla_j \ell'(w^t, x) \leftarrow \nabla_j \ell(w^t, x) \mathbb{1}_{|\nabla_j \ell(w^t,x)| \leq B}$
        end for
    end for
    Get the robust gradient estimator $\widetilde{g}^t(w^t, X_t)$:
        $$\left[\widetilde{g}^t(w^t, X_t)\right]_j \leftarrow \frac{1}{m}\sum_{x\in X_t} \nabla_j \ell'(w^t, x).$$
    Denote $w^{t+0.5} \leftarrow w^t - \eta_0 \widetilde{g}^t(w^t, X_t)$
    Let $w^{t+1} \leftarrow \mathrm{Peeling}(w^{t+0.5}, s, \epsilon, \delta, \frac{2B\eta_0}{m})$.
end for
return $w^{\text{priv}} \leftarrow w^{T+1}$.
```

Remark 3. For DP sparse learning with Lipschitz loss or regular data, [Wang and Xu, 2019] provided an upper bound of $\widetilde{O}(\frac{s^*}{n^2\epsilon^2})$. Moreover, for high dimensional sparse mean estimation and Generalized Linear Model (GLM) with the Lipschitz loss and sub-Gaussian data, [Cai *et al.*, 2020; Cai *et al.*, 2019] provided optimal rates of $\widetilde{O}\big(\frac{s^*\log d}{n} + \frac{(s^*\log d)^2}{(n\epsilon)^2}\big)$. We can see that compared with these results, the error bound now becomes $\widetilde{O}\big(\frac{(s^*)^{\frac{1+2v}{1+v}} u^{\frac{2}{1+v}}}{(n\epsilon)^{\frac{2v}{1+v}}}\big)$ due to data irregularity. When $v = 1$, the error bound now becomes to $\widetilde{O}\Big(\frac{(s^*)^{\frac{3}{2}} u}{(n\epsilon)}\Big)$, which matches the result in [Hu *et al.*, 2021]. Thus, our result can be seen as a generalization of the previous ones.

One open question is whether we can further improve the rate of error in Theorem 10. In the following we show that the bound is optimal up to a factor of $\widetilde{O}(s^*)$.

Theorem 11. For $\mathcal{W} = \{w \in \mathbb{R}^d : \|w\|_0 \leq s^*\}$ and any $v \in (0,1]$, there exists a strongly convex and smooth loss function $\ell : \mathcal{W} \times \mathbb{R}^d \mapsto \mathbb{R}$ such that, for any ϵ-DP algorithm $\mathcal{A}$, there is a distribution $\mathcal{D}$ over $\mathbb{R}^d$ such that for any w, $\sup_{j\in[d]} \mathbb{E}_{x\sim\mathcal{D}}[|\nabla_j \ell(w,x)|^{1+v}] \leq u$, its output w^{priv} satisfies the following when $n \geq \Omega(s^*\log d/\epsilon)$

$$err_{\mathcal{D}}(w^{\text{priv}}) \geq \Omega\left(u^{\frac{2}{1+v}}\left(\frac{s^*\log d}{\epsilon n}\right)^{\frac{2v}{1+v}}\right).$$

For any (ϵ, δ)-DP algorithm $\mathcal{A}$ with $\epsilon \ll \log\frac{1}{\delta}$, there is a distribution $\mathcal{D}$ over $\mathbb{R}^d$ such that for any w, $\sup_{j\in[d]} \mathbb{E}_{x\sim\mathcal{D}}[|\nabla_j \ell(w,x)|^{1+v}] \leq u$, its output w^{priv} satisfies the following when $n \geq \Omega(\sqrt{s^*\log d\log\frac{1}{\delta}}/\epsilon)$

$$err_{\mathcal{D}}(w^{\text{priv}}) \geq \Omega\left(u^{\frac{2}{1+v}}\left(\frac{\sqrt{s^*\log d\log\frac{1}{\delta}}}{\epsilon n}\right)^{\frac{2v}{1+v}}\right).$$

Acknowledgments

Di Wang and Yulian Wu were support in part by the baseline funding BAS/1/1689-01-01, funding from the CRG grand URF/1/4663-01-01 and funding from the AI Initiative REI/1/4811-10-01 of King Abdullah University of Science and Technology (KAUST). Xiuzhen Cheng and Youming Tao were partially supported by National Key R&D Program of China (No. 2019YFB2102600).

References

[Acharya *et al.*, 2021] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *ALT*, volume 132 of *Proceedings of Machine Learning Research*, pages 48–78. PMLR, 2021.

[Bahmani *et al.*, 2013] Sohail Bahmani, Bhiksha Raj, and Petros T Boufounos. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14(Mar):807–841, 2013.

[Barber and Duchi, 2014] Rina Foygel Barber and John C. Duchi. Privacy and statistical risk: Formalisms and minimax bounds, 2014.

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

[Bassily *et al.*, 2019] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradee Thakurta. Private stochastic convex optimization with optimal rates. In *NeurIPS*, 2019.

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

[Cai *et al.*, 2019] T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019.

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

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

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

[Feldman *et al.*, 2020] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *STOC*, pages 439–449, 2020.

[Hu *et al.*, 2021] Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. High dimensional differentially private stochastic optimization with heavy-tailed data. *arXiv preprint arXiv:2107.11136*, 2021.

[Kamath *et al.*, 2020] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory*, pages 2204–2235. PMLR, 2020.

[Kamath *et al.*, 2021] Gautam Kamath, Xingtu Liu, and Huanyu Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. *arXiv preprint arXiv:2106.01336*, 2021.

[Liu *et al.*, 2021] Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. *arXiv preprint arXiv:2102.09159*, 2021.

[Tao *et al.*, 2021] Youming Tao, Yulian Wu, Peng Zhao, and Di Wang. Optimal rates of (locally) differentially private heavy-tailed multi-armed bandits. *arXiv preprint arXiv:2106.02575*, 2021.

[Vapnik, 1999] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[Wang and Gu, 2019] Lingxiao Wang and Quanquan Gu. Differentially private iterative gradient hard thresholding for sparse learning. In *28th International Joint Conference on Artificial Intelligence*, 2019.

[Wang and Xu, 2019] Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 6628–6637. PMLR, 2019.

[Wang and Xu, 2021] Di Wang and Jinhui Xu. On sparse linear regression in the local differential privacy model. *IEEE Trans. Inf. Theory*, 67(2):1182–1200, 2021.

[Wang *et al.*, 2020a] Di Wang, Jiahao Ding, Lijie Hu, Zejun Xie, Miao Pan, and Jinhui Xu. Differentially private (gradient) expectation maximization algorithm with statistical guarantees. *arXiv preprint arXiv:2010.13520*, 2020.

[Wang *et al.*, 2020b] Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pages 10081–10091. PMLR, 2020.

[Woolson and Clarke, 2011] Robert F Woolson and William R Clarke. *Statistical methods for the analysis of biomedical data*, volume 371. John Wiley & Sons, 2011.